Scalable Methodologies and Analyses for Modality Bias and Feature Exploitation in Language-Vision Multimodal Deep Learning

THOMAS IAIN WINTERBOTTOM

Scalable Methodologies and Analyses for Modality Bias and Feature Exploitation in Language-Vision Multimodal Deep Learning

Thomas Winterbottom

A Thesis presented for the degree of
Doctor of Philosophy

Department of Computer Science
Durham University
United Kingdom
March 2023

Abstract

Multimodal machine learning benchmarks have exponentially grown in both capability and popularity over the last decade. Language-vision question-answering tasks such as Visual Question Answering (VQA) and Video Question Answering (video-QA) have —thanks to their high difficulty— become a particularly popular means through which to develop and test new modelling designs and methodology for multimodal deep learning. The challenging nature of VQA and video-QA tasks leaves plenty of room for innovation at every component of the deep learning pipeline: from dataset to modelling methodology. Such circumstances are ideal for innovating in the space of language-vision multimodality. Furthermore, the wider field is currently undergoing an incredible period of growth and increasing interest. I therefore aim to contribute to multiple key components of the VQA and video-QA pipeline, but specifically in a manner such that my contributions remain relevant, scaling with the revolutionary new benchmark models and datasets of the near future instead of being rendered obsolete by them. The work in this thesis: highlights and explores the disruptive and problematic presence of language bias in the popular TVQA video-QA dataset, and proposes a dataset-invariant method to identify subsets that respond to different modalities; thoroughly explores the suitability of bilinear pooling as a language-vision fusion technique in video-QA, offering experimental and theoretical insight, and highlighting the parallels in multimodal processing with neurological theories; explores the nascent visual equivalent of languague modelling ('visual modelling') in order to boost the power of visual features; and proposes a dataset-invariant neurolinguistically-inspired labelling scheme for use in multimodal question-answering. I explore the positive and negative results that my experiments across this thesis yield. I conclude by discussing the limitations of my contributions, and conclude with proposals for future directions of study in the areas I contribute to.

Declaration

The work in this thesis is based on research carried out at the Department of Computer Science, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

As with many who have come before me, this PhD has been both extremely challenging, and the experience of a lifetime. I thank the following people, in no uncertain terms, for their kindness, support, and friendship.

First and foremost, I thank Noura Al Moubayed for being the phenomenal supervisor, human, and friend that she is. Noura's tireless support and encouragement are what make her an S-tier human. In honour of her awesomeness, I have —at great personal cost— refrained from swearing throughout this thesis.

Thank you Bashar Awwad Shiekh Hassan for being a supportive and inspirational mentor when I was new to machine learning research.

Thank you to my old desk neighbours, *Dr* Grégoire Payen de La Garanderie, and *Dr* Ning Jia for making a very overwhelmed first-year me feel welcome. Cheers Greg for technical support and for fostering my enthusiasm for lower level computer systems.

Thank you G. Thomas Hudson for all the time we spent memeing, coding, and shilling vim (and for our shared Pavlovian response to gongs or the word Django).

Thank you to my friends across the Durham computer science department, shoutouts go to Jack, Abdul, Nina, David, Dan, James, Sean, Nour, Anza, Karl, Siani, Zakh, and Amit.

Thank you to Fish for being almost taller than my boyfriend, Nem for being a *slightly* better Ganondorf, Kam for his connoisseurship of creative sets, Liam for going fast, Pooky for our shared taste in references and his irrational hatred of salac berries, Marcus for our unholy resonance, Rob for his ongoing mercy on the eardrums of the world, Will for being an unholy enabler of fried chicken, James for calling me maybe, Harry for the best donut-ios storm-15 has yet managed, Tom H for his decorated career as a double-dash co-driver, Charlotte for her amazing sense of humour and skill at drawing cats on postcards, Ikit Claw for knowing who his boss is, Jack W for inspiring me to work hard and aim high, Kyle Q for not fighting hulkzilla, the fairy type for dabbing on the forces of edginess, Gonzap for being a sturdy birdy, Tom C and Taj for dancing skill only matched by their taste in

affordable alcohol, Alex for carrying team 2, Dan for top-tier lunch dates, The Wire for styling on the competition, and Pickaan and Ben for their disdain for calcium and pizza.

Thank you Ellie Littlewood for being a human of impeccable strength, love, and integrity.

A very special thank you to my mum, Liz White, for somehow being both the strongest and most supportive person in my life.

Finally, a small and passing thank you to my early twenties for the sacrifice they have made in service to this PhD.

Contents

List of Figures

List of Tables

Dedication

To anyone who is big even when they are little.
To anyone who makes the reads.
To anyone who swears ‘too much’.

CHAPTER 1

Introduction

As unimodal language and vision machine learning benchmarks rapidly approach and exceed human reasoning capabilities, there is growing interest in pushing machine learning to its limits through more complex tasks that require reasoning across *multiple* different input data modalities. Though the definition of such 'multimodal' machine learning covers *any* combination of 2 or more modalities, the overlap of language and vision remains by far the most explored and heavily resourced area of multimodal deep learning research. Language-vision multimodal tasks —of which visual question answering tasks are most prominent— have therefore become a popular testing ground for exploring multimodal processing methodology. Unfortunately, such language-vision benchmarks are plagued with a myriad of problems that diminish their ability to enable multimodal methodologies: datasets contain biases and shortcuts that can be exploited to achieve high performance *without using both language and vision input as intended*; visual contributions are often both underrepresented in datasets and under-exploited by models compared to the historically more dominant textual input modalities; techniques used in the models themselves for harmoniously combining the visual and textual features (*i.e. 'multimodal fusion techniques'*) are often overstated and only empirically justified, lacking strong

theoretical grounding for their claimed properties; and perhaps most esoterically, existing models of human cognition remain relatively unexplored as motivation for multimodal processing (*e.g.* compared to the inspiration of convolutional neural networks or 'CNNs' in the field of vision). This thesis represents my contributions in addressing these similar yet separate problems problems through four different areas united by the theme of *improving multimodal machine learning*.

1.1 Definitions

In this section, I explicitly define key terms and acronyms I use throughout this thesis. A familiarity with these definitions and their use in the field of machine learning (in the era of ∼2015-2023) serves as a strong starting point to engage with the work in this thesis.

- **'Scalable':** Has the capacity to itself be changed in size and scale, or applied to a model/dataset of increased scale. In this way a scalable methodology may be invariant or 'agnostic' to any specific dataset or model. Used descriptively to refer to my motivation behind the methodology I present.

- **Modality:** A mode or form in which information is experienced or expressed: *e.g.* the English language.

- **Medium:** The means by which information is stored and delivered. *e.g.* text is the **medium** through which the English Language **modality** is expressed.

- **Multimodal:** An adjective for a task, dataset, model, or other methodology that uses or requires information from more than one modality. Note that machine learning research generally does not yet distinguish between 'multi-medium' and 'multi-modality' scenarios.

- **Modality Bias:** Reviewed thoroughly in Section 2.3.2. The terms 'bias', 'modality bias', and 'language bias' are often used interchangeably in the field of deep learning. We say that *predictions* made by machine learning models suffer from 'bias' when we *believe* said predictions are caused by:

  - A propagation of 'spurious' correlations in the data.
  - An over-reliance on some inputs while ignoring other inputs.
  - A propagation of human assumptions or stereotypes present in the training text.

- **'Strengthening' Vision / Visual Contributions:** A methodology that increases the harmonious exploitation or 'power' of visual features by improving them *directly e.g.* projecting vision feature vectors into a vector space that improves their empirical performance on some task. In contrast to improving visual contributions *indirectly e.g.* increasing the contributions of visual features by altering other problematic non-vision representations. Elaborated on in Section 2.3.2.7.

- **VQA:** Visual Question Answering. A popular multimodal machine learning task that consists of question-answer pairs about an accompanying image or visual input.

- **Video-QA:** VQA with video/GIFs as the visual inputs.

- **Computer Vision:** A broad field of research that considers how computers represent and exploit understanding of the visual world. This thesis is concerned with the machine learning subfield of Computer Vision.

- **CNN:** Convolutional Neural Networks. A type of neural network that uses 'convolution' layers, which are inspired by undertanding of the human visual system. A popular model for computer vision tasks.

- **NLP:** Natural Language Processing. A broad field of research that considers the computational representation and exploitation of language. This thesis is concered with the machine learning subfield of NLP.

- **Language Modelling:** The practice or task of determining the probability of a given sequence of words occuring in a given sentence or context. Language modelling in deep learning often refers to:

  - Generative language modelling tasks: Popular and powerful training methodologies that predict words from sentences from a very large corpus of text.
  - Language Models: The machine learning models, pretrained or otherwise, that perform the task of language modelling.

- **Attention Mechanism:** A popular and widely used methodology often found in large and generatively trained neural networks. Attention mechanisms are intended to mediate the flow and sharing of information in a visualisable ways, and have been a backbone of state-of-the-art NLP and computer vision models since their widespread adoption in ∼2018.

- **Multimodal Fusion:** How information from multiple modalities is combined (here in the context of neural networks for multimodal machine learning tasks). I use this term in contexts where I compare different techniques to do this *i.e.* 'is concatenating features from two modalities sufficient?'; 'are there instead specific designs for neural network layers that would lead to better results?'

- **Pretraining** and **Finetuning: Pretraining** is the act of training a given machine learning model on a preliminary dataset (often very large and generally applicable). Such 'pretrained' models are often used as starting points for further training *i.e.* **'finetuning'** on another often more specific task. Finetuning intuitively seeks to slightly adjust (rather than overwrite) the more general understanding induced by pretraining.

- **'Downstream' Test-Task:** A more specific and targeted task used in finetuning. As such tasks come *after* pretraining, they are considered and deployed 'further down' the metaphorical 'stream' that is the life-cycle of a given machine learning application.

- **SL1:** Smooth L1 loss function.

- **SSIM:** The Structural SIMiliarity measure of comparison between images.

- **Labelling Scheme:** The manner in which ground truth labels are created for machine learning datasets *e.g.* one-hot labelling vs. multiclass labelling.

- **Answer Vocabulary:** The total collection of unique answers for a given machine learning task or dataset.

- A **Promising** idea: An idea: where the reasons that it *might* work are clear; that is realistically achievable with the resources available; and bluntly speaking, that I believe would be of interest to the researchers of the field. I state in my thesis that I aim for 'promising' ideas that are overlooked.

1.2 Motivation

Every individual that has ever considered machine learning has their own unique opinion on its exact *nature* and *purpose*: from business application, to experimental curiosity, to perhaps a desire to witness non-human entities display or surpass our own capacity for sentience and sapience. I posit that one of the most substantial principal components running through our collective desires and interpretations for machine learning is: *a useful —and ideally understandable— exploitation of potentially-informative stimuli.* Put in a deliberately informal way, we want machine learning agents *to* ***do*** *stuff that is* ***worth doing*** through the perspective of our own humanity. I therefore find it unsurprising that we prioritise tasks and data in two of the forms that humans have such affinity for: goal/object oriented visual reasoning, and the processing of natural language. We visually perceive meaning in the world around us, and are unprecedented[1] in our capacity to use language to nuance and enrich our visual beliefs, yielding valuable information which we freeze in time through the medium of our memories, and conveniently proliferate through the invention of text. As impressive as unimodal language models or computer vision benchmarks currently are, can we humans really be satisfied with our machine learning capacities so disconnected from our own multimodal reality? This question sufficiently approximates the *'gist'* of why I chose to contribute to multi-

[1] so far as we have yet discovered.

modal machine learning, and why I am motivated to play my part in harmonising our impressive unimodal text and vision machine learning benchmarks together: to unlock the symbiotic multimodal understanding we humans exhibit for our use, our curiosity, and our wonder.

The language-vision overlap remains the most advanced and heavily resourced field in multimodal machine learning. I therefore focus my experiments on some of the most complex, challenging, and flourishing language-vision state-of-the-art benchmarks: visual question answering (VQA) and video question answering (video-QA). As multimodal machine learning is currently experiencing an era of extreme growth —with new datasets and models released yearly that rapidly outgrow their predecessors— the methodology and analysis in this thesis is specifically motivated to remain relevant and applicable to future datasets and model designs *i.e.* to be 'scalable' or model/dataset-agnostic/invariant. I aim to avoid contributing to areas that are saturated with many similar approaches differentiated through minor changes. Such a series of targeted works *can* be beneficial to the wider field by exploring, in parallel, incremental changes and improvements to a new promising methodology. Such efforts can rapidly yield an optimised approach for wider use *e.g.* adversarial regularisation approaches for tackling language shortcuts in VQA (see Figure 2.1). While I recognise the genuine benefits of such contributions to the field, to be blunt, I am simply more interested in testing hypotheses that I consider intuitive and promising, but 'overlooked'. Nevertheless, though I have identified overlooked research gaps, my earlier chapters contribute to *topics* that are well established in multimodal deep learning: Chapter 3 focuses on the problem of language bias in the TVQA dataset, an established research area of VQA that is relatively underexplored in *video-QA*. Chapter 4 analyses the use of the multimodal fusion technique 'bilinear pooling' (BLP), experimentally and critically assessing its use as a multimodal fusion technique. I have allowed the scope of my later and larger chapters to grow progressively more ambitious, learning from the findings in my earlier work: Chapter 5 takes a more abstracted approach in harmonising visual and textual features; by exploring the nascent visual equivalent of language modelling —'visual modelling'— in the hopes of unlocking an improvement to visual

understanding not-unlike that which generative pretraining has brought to language modelling. My most recent work in Chapter 6 —taking advantage of the new improvements to modelling and bias-mitigation in VQA datasets— explores the use of neurolinguistic measures of similarity in order to instill a more 'human-like' logic to labelling for multimodal classification scenarios. Across the hypotheses tested in my thesis, I have found a mix of positive results (Chapters 3 and 5) and negative results (Chapters 4 and 6). I have aimed to thoroughly evidence my positive results and carefully explain their relevance. In turn, I have aimed to thoroughly explore potential causes for my negative results: qualitatively, quantifiably, experimentally, and with respect to the surrounding literature as appropriate.

There are limits to what any single PhD student can achieve alone. In machine learning however, there is currently an abundance of publicly available large datasets and an established culture of open-source code and libraries on GitHub and Python (from *most* of us) for which I am extremely grateful. Such conditions mean that those of us in the field of machine learning can afford to be more ambitious than might otherwise have been the case, even when armed with relatively modest computational resources (compared to the likes of Google and OpenAI).

1.3 Problems and Research Hypotheses

In this section, I outline the problems and research questions that guide my work in this thesis. I have organised my research questions into the chapters that explore them.

- **Chapter 3**
  - **Problem:** The presence of modality bias in our datasets allows models to shortcut learning multimodal interactions by using unintended unimodal priors as shortcuts. Learning to rely on such shortcuts causes models to generalise poorly to other multimodal scenarios.
  - **Question:** "Can modality bias be mitigated in a 'scalable' manner, applicable to future dataset and model designs?"

- **Chapter 4**
  - **Problem:** Multimodal fusion techniques used in VQA have led to empirical increases in task performance, with these improvements often attributed to an induction of 'richer multimodal features'. However, if these performance increases are not sustained when moving from VQA to video-QA, then we must reassess how we use and discuss these fusion techniques.
  - **Question:** "How do multimodal fusion techniques from VQA apply to video-QA?"
  - **Question:** "What are the problems with our analysis, and what do we overlook when we prioritise the empirical performance of our methodology?"
- **Chapter 5**
  - **Problem:** Generative pretraining in language modelling has revolutionised the field of natural language processing. Such generative pretraining could yield similar improvements for computer vision, but has not yet been explored near as thoroughly as language modelling.
  - **Question:** "What are the similarities, differences, and challenges of potential visual parallels to language modelling?"
  - **Question:** "What are the barriers to applying this methodology to multimodality now and in the future?"
- **Chapter 6**
  - **Problem:** The default one-hot labelling schemes typical of VQA datasets appear to enforce a problematic and oversimplified understanding of reality: that all 'incorrect' answers are equally incorrect. Though some datasets may attempt to alleviate this somewhat by gathering multiple potential answers to a given question, these auxiliary answers are limited in number and not necessarily applicable to other datasets. Similarity

scores of word pairs gathered through neurolinguistic study could be instrumental in augmenting this simplistic 'one-correct-answer' scenario in a way that is applicable to *all* VQA datasets, but as of yet remains unexplored.

- **Question:** "Can measures and metrics from human neurolinguistic theory induce desirable behaviour to modern multimodal deep learning benchmarks?"
- **Question:** "What must be achieved/overcome for neurolinguistic measures to be successfully applied to modern multimodal deep learning benchmarks?"

I return to my research questions in Section 7.3 to discuss how I have addressed them.

1.4 Thesis Contributions

The significant contributions of this thesis are:

- **Chapter 3** (On Modality Bias in the TVQA Dataset): A scalable, model-and-dataset-invariant methodological framework for detecting 'modality reliant' subsets with a proposed 'inclusion-exclusion' metric (IEM). This framework is applied to the popular large scale TVQA video-QA dataset, demonstrating the presence of harmful text biases therein.
- **Chapter 4** (Bilinear Pooling in Video-QA: Empirical Challenges and Motivational Drift from Neurological Parallels): An empirical, taxonomical, and theoretical analysis of the performance drop of bilinear pooling when applied to a video-QA context. I leverage my novel insights of the parallels between multimodal fusion and neurological theories (*i.e.* Dual Coding Theory and the Two-Stream Model of Vision) to propose several alternative neurologically-guided multimodal fusion techniques.
- **Chapter 5** (Visual Modelling: The Visual Parallel to Language Modelling Evaluated on Dynamic Simulations): 6 synthetic dynamic simulation datasets,

each of which come with physical-constant-regression tasks that together are designed to test the hypothesis: "does generative visual pretraining produce more powerful vision features". My results demonstrate the promising potential of my hypothesis, with generative visual pretraining increasing 'downstream' task performance for 7 of the 8 test tasks, yielding improvements as high as 80%.

- **Chapter 6** (Neurolinguistic Multiclass Labelling: Are Human Measures of Similarity Suitable for VQA?): A scalable and dataset/model-agnostic neurolinguistic multiclass labelling scheme leveraging human scores of semantic similarity. My extensive experiments allude to an incongruency bewteen the human measures of similarity and the VQA dataset.

1.5 Publications

The work in this thesis is published in, under-review by, or planned for imminent submission to peer-reviewed publications detailed below with myself —Tom Winterbottom, the author of this thesis— as the only first author. These publications correspond to the chapters indicated below:

- **On Modality Bias in the TVQA Dataset**
  - *Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed*
  - Proceedings of the British Machine Vision Conference (BMVC) 2020 :- 18/12/2022
  - Contributes to Chapter 3
- **Bilinear Pooling in Video-QA: Empirical Challenges and Motivational Drift from Neurological Parallels**
  - *Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed*
  - PeerJ Computer Science, 8 :- 18/4/2022
  - Contributes to Chapter 4

- **Visual Modelling: The Visual Parallel to Language Modelling Evaluated on Dynamic Simulations**
  - *Thomas Winterbottom, G Thomas Hudson, Daniel Kluvanec, Zheming Zuo, and Noura Al Moubayed*
  - Under review since December 2021 at Journal of Machine Learning Research (JMLR)
  - Contributes to Chapter 5
- **Neurolinguistic Multiclass Labelling: Are Human Measures of Similarity Suitable for VQA?**
  - *Thomas Winterbottom and Noura Al Moubayed*
  - Submission Imminent
  - Contributes to Chapter 6

1.6 Thesis Scope and Structure

The related works for each of my chapters in terms of *methodological* aim are relatively distinct from eachother. Each chapter therefore begins by reviewing the related works singularly relevant to itself. However, in Chapter 2, I review areas that are *shared* across the chapters of this thesis:

- Video-QA and VQA datasets (Chapters 3, and 4, and 6).
- Modality bias in both video-QA and VQA.

Each of my chapters aim to a contribute to one of the primary components of multimodal machine learning, but crucially, in a way that is 'scalable' *i.e.* applicable to the datasets and models of the future: Chapter 3 focuses on the problems in multimodal *datasets* through mitigating *language bias* in the TVQA video-QA dataset. Chapter 4 shifts focus to the problems with the application of multimodal *modelling techniques* through the use of bilinear pooling in video-QA benchmarks. Chapter 5 aims to take a step back and improve the power of often-under-exploited visual

features in a more general sense by bringing the paradigm of language modelling to video datasets. Chapter 6 seeks to improve *labelling practices* for multimodal tasks by exploring a neurolinguistically-guided labelling scheme for language-vision machine learning, which I evaluate on VQA.

1.7 Typographical Choices

I explicitly choose to liberally use colour and font options to highlight my intended meaning and help readers efficiently keep track of concepts I discuss. My intention is that these typographical choices help break up the walls of homogeneous text typical of academic writing, and help reduce the reader's frustration, boredom, and time spent reading on 'autopilot' (or shallow processing). I use these typographical choices to convey the following meaning:

- **Bold face** is used to grab attention, format headings, and quickly enumerating a series of key concepts in the text.
- *Italics* are used to emphasise the intended *subject* of the sentence, the intended

	Modality Bias in TVQA (Chapter 3)	**BLP in Video-QA (Chapter 4)**	**Visual Modelling (Chapter 5)**	**Neurolinguistic Multiclass Labelling (Chapter 6)**
Contribution Type				
Modality Bias	Subtitle+QA	Subtitle+QA	Static Frame Bias	Not Main Focus
Feature Exploitation	✗	✓	✓	✓
Scalable Methodology	Dataset Enrichment	Verifying BLP	Improving Vision	Improving Labelling
Task				
VQA	✗	✗	✗	✓
Video-QA	✓	✓	✗	✗
Visual Dynamics	✗	✗	✓	✗
Findings				
Positive Experimental Results	✓	-	✓	✗
Critique	✓	✓	-	-
Peer Review				
Pulblication Status	Published	Published	Under Review	Submission soon
Journal/Conference	BMVC 2020	PeerJ	JMLR	-

Table 1.1: The scope of this thesis' chapters.

key insights, or to highlight specific contrast *e.g.* Work A concludes B, however I argue *C*.

- Colour Coding: Within each individual chapter, colours are used across tables, figures, and text consistently, and aim to convey semantic meaning *e.g.* in Chapter 5 green, blue, and pink follow 3 different models and their results across figures and tables of that chapter.

- Orange and Blue text is used to help the reader follow points where two contrasting concepts or entities must be followed beyond a single sentence. Most notably, it is heavily used in Chapter 6 to keep track of concrete and abstract subjects and concepts.

- Magenta text is heavily used in my literature review (Chapter 2) to draw attetntion to where literature in the field *directly* links to or inspires my own work.

- There are times in my thesis where I must discuss and compare several different results. Some of these sections make for particularly dense reading, and must sometimes feature precise but 'wordy' language, most notably using 'double negatives': *e.g.* 'performance is degraded less substantially'. For such secnarios, I use green and red to highlight what I consider to be contrastively good and contrastively bad results respectively to help readers quickly digest the finer details of my results without getting lost in them.

- I aggressively partition my work into verbosely titled subsections —most notable in my discussion sections— to help readers locate the finer points of my arguments and findings.

CHAPTER 2

Literature Review

In this chapter, I review the literature relevant to the broader scope of my 4 contribution chapters. Though there are themes and elements shared between each of my contribution chapters, each individual chapter also demands a substantial amount of background material that is not required by other chapters. The elements that *are* shared across my contributions chapters cluster into 3 main areas: modality bias, VQA, and video-QA. These shared topics of using VQA datasets, modality bias mitigation, and evaluating methodology on video-QA datasets require a full *literature review*. However, the unique *background* to my work in each chapter is only relevant to it's respective chapter. In order to not separate such specific details from the experiments they contextualise, I therefore elect to split introducing the *background* to the respective chapters in which they're relevant. A familiarity with the concepts in Section 1.1 serves as a strong starting point to build on in the background of each contribution chapter. I instead elect in this section to conduct a *literature review* across that shared themes in my contribution chapters, with further literature reviews in each chapter as appropriate.

2.1 Video-QA

The term visual question answering or 'VQA' is ambiguous in multimodal deep learning and often refers to QA tasks with visual inputs, be they video, or single 'still images'. In thesis, I refer to image-QA tasks and datasets as 'VQA' because this was the term originally used for them before video-QA benchmarks were popularised. I instead choose to distinctly describe QA tasks with *video* inputs as video-QA. Though interest in video-QA benchmarks grew substantially *after* the success of VQA, I choose to review video-QA first in this chapter to reflect the chronological order of my research.

2.1.1 Relevance

There is a wealth of recent research in the field of video-QA focused on improving model designs and evaluation metrics. However, my work on video-QA in this thesis is instead focused on both the video-QA *datasets* themselves, and exploiting their properties to address 2 specific research gaps:

1. Chapter 3 analyses textual bias in the TVQA video-QA dataset, and I use this biased dataset to propose model/dataset-invariant dataset refinement methodologies.

2. Chapter 4 uses the known properties of 4 diverse video-QA datasets (TVQA [116], MSVD-QA [210], TGIF-QA [93], and Ego-VQA [53]) to explore the experimental shortcomings of the bilinear pooling modality fusion technique when it is extended to video-QA.

I review modality bias in both VQA and video-QA in the upcoming Section 2.3, and the specifics of the relevant modelling details in their respective chapters. As such, this section specifically focuses on video-QA *datasets* and the various design philosophies that aimed to endow them with properties I find useful for the multimodal research in this thesis. A full exploration of the many modelling designs and evaluation metrics used for the many variations of video-QA benchmarks are

beyond the scope of this thesis. A thorough review of the many aspects of video-QA can instead be found here [105, 156].

2.1.2 Review

Dataset	Year	# QA	# Videos	Domain	Answer Scheme
MovieQA [184]	2016	14,944	1075	Movies	Multiple Choice (5)
PororoQA [108]	2016	8,193	171	Cartoon	Multiple Choice (5)
TGIF-QA Count	2017	30,397	-	Varied	Multiple Choice (11)
TGIF-QA Action	2017	22,749	-	Varied	Multiple Choice (5)
TGIF-QA Trans	2017	58,936	-	Varied	Multiple Choice (5)
TGIF-QA Frame-QA	2017	53,083	-	Varied	Open Ended (1746)
TGIF-QA Total [93]	2017	165,165	71,741	Varied	-
TGIF Open [213]	2017	287,763	101,983	Varied	Open Ended
MovieFIB [136]	2017	348,998	128,085	Movie	Fill in the Blank
MarioQA [146]	2017	187,757	-	Game Footage	Open Ended (Compositional) †
MSVD-QA [210]	2017	50,505	1,970	Varied	Open Ended
MSRVTT-QA [210]	2017	243,680	10,000	Varied	Open Ended
YouTube2Text-QA [220]	2017	99,421	1,970	YouTube Videos	Multiple Choice (4)
TVQA [116]	2018	152,545	21,793	TV Shows	Multiple Choice (5)
Ego-VQA [53]	2019	610	16	Egocentric (1^{st} Person)	Multiple Choice (5)
Social-IQ [230]	2019	7,500	1,250	'Social Intelligence'	Open Ended (52,500)
AVSD [6]	2019	118,160	11,816	Multiple QA Rounds	Open Ended
LifeQA [23]	2020	2,326	275	'Day-to-day' life	Multiple Choice (4)
CLEVRER [222]	2020	305,280	20,000	Objects	Varied
VQuAD [80]	2022	1,359,999	7,000	Diagnostic Objects	Open Ended (Compositional)

Table 2.1: Breakdown of video-QA datasets. † = dataset is generated by user. Green rows indicate datasets used in the experiments in this thesis.

Video-QA datasets are relatively new, having gained substantial research intrigue over the past 6 years. One of the earliest practical video-QA datasets is MovieQA [184], which provided long video clips from movies but has a relatively low number of video clips (1075) and QA pairs (14,944). The questions in MovieQA are on topics such as movie events, actions, and plot developments. Questions are of '5w' style (*i.e.* who, what when, where, why, how), with a much larger number of 'what', and relatively fewer 'when' question. Though MovieQA represents an ambitious starting point for the size and scope of video-QA datasets, research has since demonstrated a QA language prior problem such that at least half the questions can be answered when ignoring the visual inputs [94, 217]. Such size and modality bias limitations encouraged me to use the newer video-QA datasets for my experiments. Successive video-QA datasets (see Table 2.1) distinguish themselves from the initial Movie-QA benchmark in 2 main ways:

1. By being larger with more rich/careful annotation.

2. By focusing on a new specialised 'domain' of video (*e.g.* cartoons, TV series, YouTube videos).

2.1.2.1 Increasing Size and Annotation Complexity

Movie-QA uses a relatively small number of longer video clips (1,075) for its ∼15,000 questions. TGIF-QA [93] instead uses a much larger number of shorter GIFs (∼71,000) as visual content for over 160,000 QA pairs, yielding a much larger and more diverse dataset. TGIF-QA is split into 4 subsets that each focus on a specific scenario of visual understanding: **Count**ing the number of times some process is repeated; classifying an **Action**; querying after a state **Trans**ition; and a general open-ended question-answering setup ('**Frame-QA**'). For these reasons, TGIF-QA is a widely used benchmark for video-QA, and as such I include it in my Chapter 4 experiments.

Like TGIF-QA, the TGIF Open dataset [213] is built from images in the TGIF dataset [123]. TGIF Open is an even larger scale question-answering dataset that focuses on an open-ended answer vocabulary scheme as opposed to multiple choice used in previous datasets.

MovieFIB [136] takes a different approach to achieve more complex answering behaviour through a 'fill-in-the-blank' annotation scheme. Though MovieFIB contains nearly 350,000 QA pairs on ∼128,000 video clips, its domain of 'movies' is thematically similar to the TV-shows used in the newer TVQA dataset.

The MSVD-QA and MSRVTT-QA datasets introduced by Xu et al. [210] were originally motivated by the lack of publicly available video-QA datasets at the time. MSVD-QA is built on the Microsoft Research Video Description corpus [26] used in video captioning, and functions as a smaller and 'less complex' benchmark for my experiments. Likewise, the MSR-VTT [211] dataset is used in the larger and more "complex" MSRVTT-QA dataset.

The TVQA dataset [116] aimed to address the major limitations of its predecessors: it is much larger than MovieQA, with ∼152,000 QA pairs on ∼22,000 video clips from popular TV shows; the visual content uses real humans interactions often

missing in the smaller domain-specific datasets; each video clip comes paired with time-stamped subtitles for use as a novel auxiliary input; and the questions and answers are annotated by humans using Amazon Mechanical Turk (AMT) to *specifically* require both text and vision inputs to answer *i.e.* 'multimodal reasoning'. Though such properties made TVQA an attractive starting-point for my research, I found through my work in Chapter 3 ([204]) that the multimodal reasoning criteria has not been realised due to subtitle and QA language priors allowing the majority of TVQA's questions to be answered with only the textual inputs. Yang et al. [217] have also since highlighted the QA priors (not subtitles) plaguing TVQA in their wider analysis of other video-QA datasets. Nevertheless, TVQA remains an important and widely used video-QA benchmark.

After my work in Chapters 3 and 4, I chose to move to VQA benchmarks for the methodology I propose in Chapter 6 thanks to research identifying and mitigating debilitating language priors being much more thoroughly developed for VQA than in video-QA. More ambitious video-QA datasets have since been released.

AlAmri et al. [6] introduce the audio visual scene-aware dialog (AVSD) task and dataset to push further beyond the basic video-QA framework. AVSD boasts 118,160 human annotated (AMT) QA pairs on 11,816 videos augmented with textual descriptions, video, and audio, and as such represents a substantial increase in the potential for multimodal processing. AVSD is unique in video-QA as each example features multiple successive rounds of question-answering which function to create a 'dialog'.

More recently, Yi et al. [222] design the CLEVRER video-QA dataset specifically to minimise bias —inspired by already-proven success from VQA (*i.e.* CLEVR [98])— by balancing question-answer distributions and using counterfactual examples to limit spurious correlations. The visual content in CLEVRER consists of clear and distinct objects of various sizes and shapes performing relatively simple visual dynamics. The Video Question Answering Diagnostic Dataset (VQuAD) [80] features a similar controlled visual dynamic environment with more complex movements, visual diversity, and over 1.3M QA pairs. Though the stated purpose of CLEVRER and VQuAD is to discourage video-QA bias, their careful and 'diagnos-

tic' design philosophy *i.e.* creating synthetic datasets of controlled environments is also important. In particular, my work on visual modelling in Chapter 5 features a synthetic visual-dynamics dataset carefully designed to test the visual reasoning capacity of models.

2.1.2.2 Domain-Specific Datasets

Though larger and more richly annotated datasets are instrumental for progress in video-QA, such datasets often neglect alternative domains of video content, and therefore motivates the creation of domain-specific datasets. The high cost of dataset collection and relative lack of domain-appropriate visual content often means such domain-specific datasets are relatively small. However, these domain-specific datasets fill a crucial research gap, often indicating how well benchmarks generalise to different video content.

Early 'specialised' datasets include: Pororo-QA [108], with 8,193 questions on the 171 episodes of the 'Pororo' cartoon; and MarioQA [146], designed with 187,757 questions about actions in the game Super Mario (relatively large for a specialised dataset, though much smaller than MovieFIB, CLEVRER, and VQuAD).

My experiments in Chapter 4 include YouTube2Text-QA [220] and Ego-VQA [53]. YouTube2Text-QA aims to capture "in-the-wild" action semantics through YouTube videos and enrich descriptions with 'web-scale' NLP databases and language models. Ego-VQA addresses the lack of 1^{st}-person video data with a very small 'egocentric' video-QA corpus of 16 video clips with 610 QA pairs. I use the small and specialised Ego-VQA dataset for my experiments in Chapter 4 by first pretraining on YouTube2Text-QA as recommended in Fan [53].

More recently, smaller specialised datasets focus on human socialising and interactions. The 7,500 questions and 1,250 videos in Social-IQ [230] aim to capture 'social intelligence' such as 'who is to blame for a situation' and 'what is the *mood* of the conversation'. Questions are annotated with 3 difficulty levels by 'hired and trained' undergraduate students. However, the design of Social-IQ does not explicitly outline any steps to mitigate the various question biases that have recently come into focus in video-QA. In contrast, Castro et al. [23] introduce the 'real-life' LifeQA

video-qa dataset (2,326 questions, 275 videos) and actively and transparently report accuracy for various benchmark models (including human performance) using combinations of their datasets inputs (video, questions, answers, and transcripts) similar to my breakdown of TVQA in Chapter 3. Their analysis indicates that an accuracy of 44% and 63.4% is achievable through machine learning or by humans respectively on LifeQA without using the visual input. I believe this honest admission of dataset bias should be encouraged, even celebrated, for the long term benefit of the field of multimodal machine learning.

2.2 VQA

The first 2 chapters of this thesis use video-QA benchmarks. However, as previously outlined, I return to the 'simpler' task of VQA for Chapter 6 because the VQA benchmark datasets have at this time gone further to mitigate language bias than their video-QA counterparts. Such datasets enable hypotheses for new multimodal methodology to be evaluated more reliably.

2.2.1 Relevance

Chapter 6 directly uses VQA datasets (VQA v1, VQA v2, VQA-CP v1/v2, and GQA) as experimental benchmarks to test my neurolinguistic multi-class labelling hypothesis. Though Chapter 4 uses a video-QA dataset instead of VQA for experiments, it analyses the nuances and consequences of extending the bilinear pooling techniques popularised in VQA to video-QA. I use VQA datasets as a 'means-to-an-end' in evaluating specific multimodal methodology. As both Chapters 4 and 6 review work similar in *methodology*, this section is therefore limited to VQA benchmark datasets and the properties that make them useful. A survey of VQA methodologies is beyond the scope of this thesis, but can be found here [207, 79, 241].

2.2.2 Review

See Table 2.2 for a breakdown of benchmark VQA datasets.

Dataset	Year	# Qs	# Imgs	Source	Research Gap
Flickr30k [225]	2014	158,915	31,783	Flickr	Early VQA benchmark dataset.
DAQUAR [137]	2014	12,468	1,449	NYU Dataset	Aims to have 'high complexity' questions.
COCO-VQA [166]	2015	~122,000	~120,000	MS-COCO	Aims to be larger than DAQUAR. Makes QA pairs from descriptions.
FM-IQA [59]	2015	316,193	158,392	MS-COCO	Chinese question-answer pairs with English translations.
Visual Madlibs [227]	2015	360,001	10,738	MS-COCO	A 'fill-in-the-blank' or 'finish-the-sentence' style VQA dataset.
VQA v1 [9]	2015	~0.76M	204,721	MS-COCO	New task pushing a better benchmark for 'open-ended' and multimodal reasoning.
Visual7W [240]	2016	327,939	47,300	MS-COCO/AMT	Push past 'loose global' text-image associations using 'semantic links' between objects in the images and descriptions.
Visual Genome [110]	2016	1,773,258	108,000	MS-COCO	Very large dataset richly annotated with object-attribute relationships.
Binary VQA on A.S. [234]	2016	22,055	-	AMT	Focuses on balancing binary 'yes'-'no' questions.
TDIUC [99]	2017	1,654,167	167,437	COCO-VQA/Visual Genome	Actively balances questions by 'category' (*e.g.* colour questions) for more nuanced analysis.
C-VQA [3]	2017	369,861	205,363	VQA v1	Enforce 'compositional' questions *i.e.* inferring an attribute about an object in *testing* when the attribute and object did not coincide during *training*.
VQA v2 [70]	2017	~1.1M	265,016	AMT	Improve VQA v1 by balancing QA priors with counterfactual images.
KB-VQA [199]	2017	2402	700	MS-COCO/AMT	Exploits object-attribute relations from external databases to augment questions.
CLEVR [98]	2017	999,968	100,000	Synthetic	Mitigates language bias with very precise compositional design and question balancing.
FigureQA [101]	2018	~1.55M	~120,000	Synthetic	Images are synthetically made graphs and plots displaying dataset.
VQA-CP v1 [4]	2018	~0.76M	204,721	VQA v1	Reshuffle VQA datasets such that QA priors are further minimised.
VQA-CP v2 [4]	2018	~1.1M	265,016	VQA v2	↑
DVQA [100]	2018	3,487,194	300,000	Synthetic	Similar to FigureQA. Large VQA dataset for understanding information in bar charters.
GQA (Full) [91]	2019	~22M	~113,000	AMT/Visual Genome	Much larger than VQA. Emphasises reducing statistical biases with question balancing and 'grounded' visual question design.
GQA (Balanced) [91]	2019	~1.7M	85,361	↑	↑
MUTANT [67]	2020	~679,000	~679,000	VQA v2/CP v2	Small 'mutations' of VQA (Q,I,A) triplets to mitigate statistical biases.

Table 2.2: Breakdown of VQA datasets. AMT = Amazon Mechanical Turk. Green rows indicate datasets used in the experiments in this thesis.

2.2.2.1 Early Datasets

The VQA task was itself was first explicitly formalised in Antol et al. [9]. There are however several functionally similar early benchmark datasets. The earliest datasets focus mainly on increasing the complexity of questions (DAQUAR [137]) and contributing a large dataset size (Flickr30k [225]).

Subsequent early datasets would use the large Microsoft Common Objects in Context dataset (MS-COCO) [128] for their images: COCO-VQA [166] aims specifically to be larger than DAQUAR while maintaining complex question design; FM-IQA uses 316,193 Chinese question-answer pairs (with English translations); Visual Madlibs contains a similarly large question-answer count (316,193) with a 'fill-in-the-blank'/'finish-the-sentence' answer style.

The largest early VQA dataset is the widely-used VQA v1 introduced in Antol et al. [9]. VQA v1 uses 204,721 images from MS-COCO with ∼760,000 questions which use a new diverse 'open-ended' answer scheme. I include VQA v1 in my experiments alongside its subsequent improved versions as it still remains an important and widely reported benchmark.

2.2.2.2 No Active Language Prior Mitigation: Additional Annotation

Though VQA benchmarks are specifically designed to be multimodal (*i.e.* require *both* vision and textual inputs to answer), datasets are often plagued by strong correlations between questions and answers that allow questions to be answered correctly while ignoring the visual information. Some few datasets however do not actively mitigate this in their design, instead focusing augmenting questions with additional annotations.

Visual7w [240] uses AMT workers to annotate bounding boxes for ∼327,000 questions on 47,300 MS-COCO images. Such bounding boxes function as auxiliary annotations that can be used to help models 'resolve the co-reference ambiguity problem between QA sentences and images'. Zhu et al. [240] *do* demonstrate that their LSTM model learns from from answers in the training set, but do not actively mitigate this in their design process.

KB-VQA [199] is a relatively small dataset of 700 images and 2,402 questions. AMT workers create object-attribute relations from templates as additional inputs to questions.

The Visual Genome dataset [110] stands out from its predecessors through *very* rich annotation. Images feature multiple bounding boxes indicating entities in relation to eachother paired with textual descriptions. Such descriptions form semantic

graphs representing the relationships between object. The authors outline many 'biases' found in their dataset *e.g.*: an over-representation of images of people, or skiing and surfing for images of sports. Though the authors try to encourage question diversity by asking annotators to make different question types ('what', 'when', 'where'), Visual Genome is not designed to be limited to the task of VQA and as such does not undergo QA-prior balancing.

2.2.2.3 Active Language Prior Mitigation

Over the past 6 years, actively mitigating such 'biases' during dataset design has been a top priority for new VQA datasets.

2.2.2.3.1 QA Balancing

Zhang et al. [234] create a VQA dataset using a clip-art library to generate images, and actively balance their binary ('yes/no') questions with 'counterfactual' examples *i.e.* minor changes to an image such that the answer changes. Though this design prevents models from gaining unintended boosts from over-represented 'yes' answers previously common in VQA datasets, it only addresses binary questions.

Kafle and Kanan [99] introduce the Task Driven Image Understanding Challenge (TDIUC) dataset with *questions* grouped into 12 categories that work alongside Mean-Per-(question)-Type metrics to demonstrate when models are failing to overcome statistical biases for certain question types. TDIUC functions as an analytic tool for detecting and subsequently preventing spurious correlations common in various question types, but is not itself ambitiously balanced to prevent language priors.

The VQA v2 dataset [70] builds on top of the VQA v1 dataset with extra counterfactual questions. The 'balanced' split features not one but *2* counterfactual examples for questions, yielding a much larger dataset (∼1.1M questions) that has a much better question-answer balance than VQA v1. As VQA v2 serves as a larger and improved contrast to the original VQA v1, I include it in my experiments in this thesis. Though the counterfactual design in VQA v2 is effective, correlations still exist in the training set that can be exploited as shortcuts to misleadingly high

testing accuracy.

2.2.2.3.2 Train-Test Split Reorganisation

Compositional VQA (C-VQA) [3] is a split of VQA v1 that aims to actively ensure that unseen scenarios for testing are *genuinely* unseen. This is achieved by rearranging the training and validation splits such that questions are 'compositionally novel'. For example, question-answer pairs such as ["What colour is the plate?", "Red"] and ["What is the colour of the plate?", "Red"] would not be allowed in *both* training and testing splits as the questions have been detected as compositionally similar and have the same answer despite the different phrasing. Instead ["What colour is the plate?", "Green"] would be a suitable counterpart to appear in the opposite split. I believe that this represents a more general case of the previously discussed 'counterfactuals'.

VQA under Changing Priors (VQA-CP) v1/v2 datasets introduce in Agrawal et al. [4] represent a more global decoupling of training and testing question-answer distributions. Specifically, the training and testing splits of the VQA v1/v2 datasets are reshuffled such that each question type (*e.g.* 'how many', 'what colour is') has a different answer distribution. No questions, answers, or images are altered in any way, simply redistributed between the training and testing splits. The authors demonstrate a very significant significant drop in performance for *all* benchmarks in their experiments (*all* models approximately achieving roughly *half* their original accuracy). Such unprecedented performance drops indicates that the design philosophy of VQA-CP is perhaps the most successful approach to mitigating question-answer priors. I use VQA-CP in my experiments as it acts as a natural comparison to *both* the VQA v1/v2 datasets I experiment on, *and* represents a more thorough decoupling of training and testing correlations than C-VQA.

2.2.2.3.3 Synthetic and Controlled Image Content

Preceding VQA datasets struggle to formulate questions that *require* the image to answer, motivating the creation of synthetic datasets with more 'controlled' and specific image content to ensure that crucial visual information is undiluted and clear.

FigureQA [101] is a large dataset of ∼1.55M questions on ∼120,000 images of programatically generated graphs and plots of data. Questions focus on specific detail in the plots graphs *i.e.* 'Does the Light Gold plot have the lowest value?/area-under-curve?', and 'Yes-no' answers for each question type and each *figure* are balanced. The DVQA dataset [100] focuses entirely on synthetic bar charts, and is over twice the size of the already-large FigureQA. DVQA's 'data retrieval' and 'reasoning' questions are similar to those in FigureQA, and they also similarly balanced their 'yes-no' answers.

CLEVR [98] uses a controlled environment where images feature several visually-distinct shapes with varying colours, and compositional questions focus on counting, comparing, or identifying objects and their relations to eachother. CLEVR creates actively balanced 'question families' to minimise an question-answer shortcuts.

I do not use any synthetic *VQA* datasets for the experiments in my thesis. I *do* however recognise that such datasets are powerful tools for targeting very specific research gaps thanks to the careful control designers have with the image content and that relatively large datasets can be programmatically generated at a fraction of the cost of full annotation for more open-domain datasets. Indeed, in Chapter 5, I generate several synthetic video datasets with specific properties to test the hypothesis: 'Does pretraining on generative tasks improve downstream performance for vision as it does in NLP?'.

2.2.2.3.4 Large Scale and General Purpose

More recently introduced VQA datasets distinguish themselves by combining the language-prior mitigation insights of their predecessors with a much larger scale of question design, complexity, and annotation.

The GQA dataset [91] combines and refines pivotal features of previous benchmarks: Images are associated with richly annotated 'scene graphs' that represent objects, attributes, and their relations from the Visual Genome dataset; Questions are associated with 'functional programs' *i.e.* a series of 'reasoning steps' that lead to the correct answer, with each answer augmented with textual and visual justifications (a multi-step and multimodal extension of the approach in KB-VQA, with

Visual Genome bounding boxes); Questions are 'compositionally' designed to disentangle coinciding attributes and objects to emphasise novel combinations (similar to C-VQA and CLEVR); Dataset annotation is done by humans through Amazon Mechnical Turk (in a more complex and thorough manner than other AMT annotated datasets *i.e.* VQA v2, Visual7w, etc); The full dataset reaches an unprecedented ∼22M questions on ∼113,000 images; A respectably sized (∼1.7M questions, 85,361 images) 'balanced' split of the dataset is made by balancing the answer distribution of question types; Such question type boundaries are derived from 'global' and 'local' labels in GQA (more extensive than the simpler 'what/when/where' question types categories balanced in previous datasets). I use the GQA dataset in my experiments as the combination of the substantial improvements in each of the discussed components sets GQA apart as a new benchmark in VQA.

The MUTANT training approach introduced in Gokhale et al. [67] augments image-question-answer triplets from the VQA v2 and VQA-CP v2 datasets, effectively generating a new VQA dataset. Small 'mutations' are made to either the question or the image in the VQA sample which results in a new changed answer, helping to balance question-answer distributions by yielding a total of ∼679,000 subtly unique questions and images (not unlike counterfactuals). The MUTANT training paradigm leads to noticably increased performance for the harder VQA-CP datasets, implying that the training paradigm successfully discourages the question-answer priors that VQA-CP punishes.

2.3 Modality Bias in Multimodal Question Answering Datasets

2.3.1 Relevance

Mulitmodal question-answering datasets have been used as a convenient platform to develop multimodal processing methodology over the past decade. As discussed in the previous sections of this review, these datasets no longer facilitate multimodal processing methodology if they contain shortcuts that can be answered without one

of the principal modalities, rather than *requiring* multiple modalities as intended. This motivates a guiding principal for this thesis: **Research into developing dataset and model-agnostic multimodal processing methodology therefore *first* requires addressing the modality bias in multimodal question-answering datasets *before* any hypothesis for improving multimodal processing can be properly tested.** This principal is reflected in the journey through my contributions: Chapter 3 highlights the presence of language bias in the large scale video-QA dataset TVQA. This study was predominantly carried out in 2019, and notably co-incides with expanding interest in languague-priors in the field of VQA. Chapter 4 discusses the identified language bias in TVQA as a potential factor in the poor video-QA performance of bilinear pooling. Chapter 5 aims to balance modalities by switching the focus from 'reducing language bias' to 'improving visual contributions' by applying the powerful generative pretraining methodology instrumental to the powerful generative pretraining paradigm of language models to vision, while highlighting the 'static frame bias' common in video benchmarks. Finally, Chapter 6 uses new VQA datasets with state-of-the-art language-bias mitigation to test my dataset and model-agnostic neurolinguistic multiclass labelling hypothesis for improving multimodal processing.

Modality bias is a pivotal theme as either *focus* or *support* for each of my contributions. I therefore review modality bias in the question-answering benchmarks relevant to this thesis. Details specific to each contribution are expanded upon in their respective chapters.

2.3.2 Review

The terms 'bias', 'modality bias', and 'language bias' are overloaded with different definitions in the field of deep learning. We most often say that *predictions* made by machine learning models suffer from 'bias' when we *believe* said predictions are caused by:

- A propagation of human assumptions or stereotypes present in the training text.

- A propagation of 'spurious' correlations in the data.
- An over-reliance on some inputs while ignoring other inputs.

These 3 common diagnoses often coincide, either causing eachother or present as symptoms of some underlying flaw in the dataset. The first diagnosis (human assumptions and stereotypes) is out of the scope of this thesis. See Figure 2.1 for a visual breakdown of the research addressing modality bias in the broader fields of visual QA.

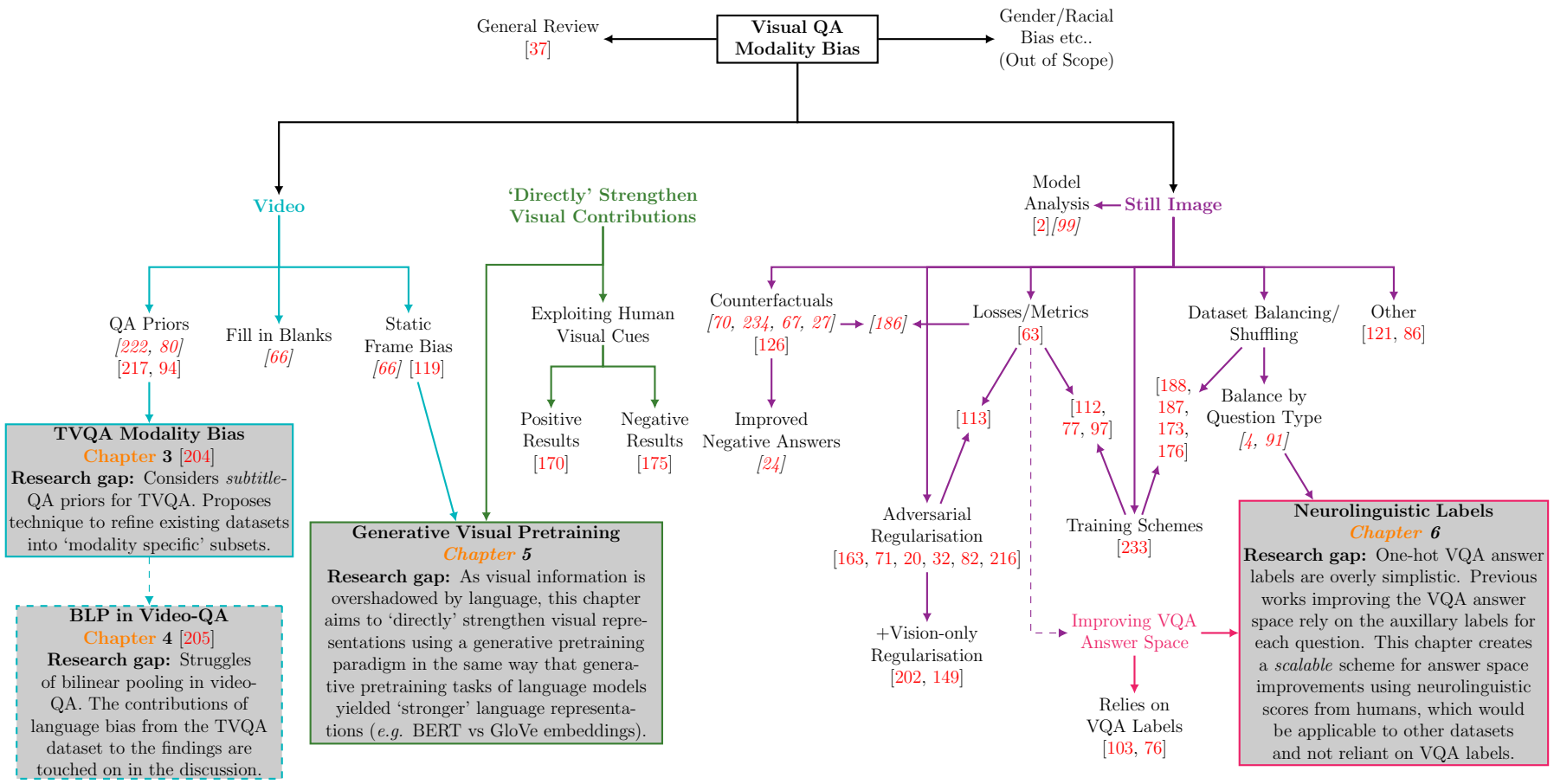


Figure 2.1: An overview of the literature of modality bias in VQA and Video-QA. The works displayed here do either one or both of bringing attention to modality bias or proposing direct/indirect solutions to modality bias. References in *italics* are studies that introduce substantial *datasets* alongside other contributions. Though Chapter 4 *does* consider modality bias in the discussion, it is *not* primarily motivated by modality bias.

2.3.2.1 Adversarial Regularisation

A successful method for mitigating language priors in VQA models is 'adversarial regularisation': to train a *purposefully* biased model that only sees the question (*i.e.* a 'question-only' model), and using the purposefully biased question-only predictions to alter the *real* model's training such that said biased predictions are discouraged. Such approaches have the significant benefit of being both datasets and model invariant.

One of the earliest of such training strategies is introduced in Ramakrishnan et al. [163] yields ∼4% increase in performance on VQA-CP v1.

Grand and Belinkov [71] propose a scheduling scheme that improves stability during training. The authors further provide a deeper analysis of adversarial regularisation with several key insights: it yields significant improvement for yes-no questions, but harms performance on more 'challenging' questions; linguistic cues in the question can often be ignored in favour of salient visual features; and the improved 'out-of-domain' performance comes at the cost of 'in-domain' performance.

The 'reducing unimodal bias' (RUBi) method [20] builds on [163] by training both question-only and 'real' models together, with the question-only votes directly applied to the gradients of the real model *during* training.

The work of Clark et al. [32] is distinguished from RUBi by using an already-pretrained question-only model during adversarial training. They find 'improved robustness in all settings', but do not provide the same thorough diagnostic breakdown of results as [71].

Yang et al. [216] posit that some 'good' language priors are worth exploiting (similar to [202]), arguing that 'reducing language bias weakens the ability of VQA models to learn context prior(s)'. The authors introduce a content and context with language bias (CCB) training scheme alongside a 'bias-only' model to regularise biased predictions.

Han et al. [82] argue that what has conventionally been called 'language bias' should be split into 'distribution bias' and 'shortcut bias'. Their Greedy Gradient Ensemble (GGE) method is designed to greedily overfit to 'distribution bias', thus forcing the full model to learn more thoroughly from examples that the biased models

cannot answer.

Lao et al. [113] introduce a language-prior based focal loss 'LP-Focal' —using votes from the biased question-only model to rescale the standard cross entropy— increasing VQA-CP v2 performance by ∼18% accuracy.

It is important to note that while these regularisation methods yield improved performance on datasets that are specifically designed to punish linguistic priors (VQA-CP), they often lead to *decreased* performance for datasets where such shortcuts exist to be exploited (VQA v1).

During my research, I found the field of adversarial regularisation to be saturated and thoroughly explored. Though I aimed to find more unique research gaps to address, I used adversarial regularisation as a tool to help test my hypotheses. My work in Chapter 3 demonstrates that the RUBi method *reduces* accuracy on the TVQA *video-QA* dataset, further implying that linguistic shortcuts exist in TVQA as suppressing the exploitation of such priors causes a small decrease in performance.

2.3.2.1.1 +Vision-only Regularisation Preceding adversarial regularisation techniques focus solely on a question-only model: unsurprising given the dominant corrupting effect of language priors in VQA. However, recent research efforts include a vision-only model in their adversarial regularisation.

Niu et al. [149] introduce a counterfactual adversarial regularisation technique designed to make the model aware of the chance for bias on a given question through *counterfactual* reasoning on *both* the image and question.

The D-VQA method Wen et al. [202] similarly includes a vision-only branch to find and reduce 'vision-to-answer' bias alongside that of 'question-to-answer'. However, the authors rather unconventionally argue that not all biases are undesirable to VQA models, as "some biases learnt from datasets represent natural rules that can help limit the range of answers".

It is important to note that some of the author's examples of 'good biases' — *e.g.* 'oranges (fruit) are orange (colour)— directly contradict the interpretations of previous works that denounce *functionally identical* bias: 'bananas are yellow'. **It is critical to the field of multimodality that such conflicting narratives are**

identified and reviewed in the future as the field continues to expand.

2.3.2.2 Counterfactuals

One of the most established methods for mitigating language priors is the use of counterfactual examples with similar visual content but different answers. In the preceding Section 2.2, I outline the design history of several counterfactual VQA *benchmark datasets*: VQA v2 [70], Zhang et al. [234], and MUTANT [67]. Though I take advantage of the counterfactual improvements in the VQA v2 and VQA-CP v2 datasets in my experiments in Chapter 6, there exists further research efforts more closely examining the effects of counterfactuals themselves.

2.3.2.2.1 Includes Adversarial Regularisation

A trend in more recent research with counterfactuals as their primarily contribution is to include adversarial regularisation in their experimental process and ablation.

Teney et al. [186] introduce a training scheme for generating counterfactuals by both: completely masking or otherwise removing objects from images; and editing questions in VQA v2. They further introduce a 'gradient supervision' loss function that aims to improve learning from counterfactual pairs.

Chen et al. [27] introduce a counterfactual training scheme very similar to that of Teney et al. [186]. A key insight in Chen et al. [27] is that such linguistic masking is hypothesised to make question-only models in the adversarial ensemble more robust against paraphrasing.

Liang et al. [126] propose a training methodology that creates factual and counterfactual examples from each training sample for use in a triplet loss.

As with adversarial regularisation, I found the 'counterfactual' subfield to be thoroughly explored and saturated. I instead use the useful properties counterfactual design gives the VQA v2 and VQA-CP v2 datasets to test the hypotheses in my work.

2.3.2.3 Improving Negative Answers

Chao et al. [24] demonstrate that the negative answers used in multiple choice VQA

datasets have a significant and often-overlooked impact on model behaviour. If the negative 'decoy' answers are too obviously incorrect for a given question, the authors demonstrate that models learn to ignore visual information altogether. The authors demonstrate that more carefully selected negative answers reduces vision-agnostic behaviour. However, it remains unclear what ramification this may have for VQA datasets with larger and open-ended answer vocabularies.

2.3.2.4 Other Training Schemes

This subsection highlights a more general and varied set of 'training scheme' contributions.

Zhang et al. [233] argue that some questions concerned with logical reasoning cause errors because 'only visual information is present' and is separated from 'general knowledge' that could give crucial context for a correct answer. They introduce a model that exploits an external knowledgebase of relations between objects not unlike that of Visual Genome.

2.3.2.4.1 +Losses and Metrics

Some training schemes introduce new losses and metrics alongside their methodology.

Lao et al. [112] present a curriculum learning strategy that starts with 'easy' questions ('more-biased examples'), then proceeds to 'harder' questions (less biased examples requiring multimodal reasoning) as determined by a proposed 'difficulty' metric. This metric is composed in part by a 'Visual Sensitive Coefficient' which aims to detect the 'difficulty driven by the language bias of each VQA sample'.

Guo et al. [77] consider the class-imbalance side of language priors in VQA, and their analyses indicate that VQA models often give a frequent yet wrong answer to a question whose correct answer is sparse in the training set. The authors introduce a method of 'loss rescaling' in order to weight answers based on 'training data statistics' such that easy mistakes don't contribute much to the overall loss whereas more challenging scenarios do.

Jiang et al. [97] propose a graph generative modelling-based training scheme

(X-GGM) to improve 'out-of-distribution' performance. Out-of-distribution examples are generated by varying the attributes or objects in the image by 'injecting perturbations' directly into the intermediate cross-modal representations. An accompanying 'gradient distribution consistency' loss function is designed to stabilise the training procedure.

2.3.2.4.2 +Dataset Unshuffling/Balancing

Other training strategies focus specifically on some form of dataset balancing or shuffling to yield improvements.

Si et al. [176] introduce a select and re-rank mechanism which proposes candidate answers at first, restricting the answer space, and then re-ranking the answers afterwards. This approach is designed to minimise spurious correlations in answer vocabularies with the aid of image features, but it is unclear if the subsequent re-ranking of answers would not themselves be spuriously correlated.

Teney et al. [188] improve the out-of-distribution performance of their experiments by splitting the dataset into multiple "well chosen" subsets and treating them as separate instances of training. Subsets are chosen either through question-types, or through K-means clustering on a binary bag-of-words representation of the questions.

Though VQA-CP has been demonstrated as a significant improvement over VQA v1/v2 by many of the studies discussed in this section, Teney et al. [187] show that it still contains "embarassingly simple" examples.

Sha et al. [173] extend the idea of balancing and shuffling subsets for superior training to work *across multiple datasets*.

2.3.2.5 QA Priors in Video

I discuss the diagnostic design and question-type balancing of the CLEVRER [222] and VQuAD [80] datasets in Section 2.1.2.1. However, there are other studies (aside from dataset design) that analyse question-answering priors that exist in the less-explored video-QA datasets.

Jasani et al. [94] highlight the existence of language bias in the MovieQA dataset by demonstrating that state-of-the-art results are achieved by using a text-only model without any visual inputs. They further find that even a simple fully-connected model using appropriately trained Word2Vec [140] vectors is sufficient to correctly answer over 40% of questions. The authors speculate that this bias is introduced by AMT workers either through: an over-representation of movie-relevant characters and plots in the correct answers, or not watching the movie itself before creating the questions.

Yang et al. [217] explore question-answer bias in both MovieQA and TVQA. They demonstrate that the RoBERTa language model has the ability to overfit even to the QA style of *individual annotators*. In particular, the authors show that 'why' and 'how' question types 'incur more biases that language models can exploit' which they attribute to the relative complexities of such questions compared to the more "factual and direct" 'what', 'who', or 'where' questions.

My work in Chapter 3 and its corresponding publication [204] was conducted around the same time as both the studies discussed above, and the topic of text bias in video-QA still remains relatively unexplored to date. I consider both textual bias in TVQA from *both* questions *and* subtitles.

2.3.2.6 Static Frame Bias

A more esoteric form of modality bias is the 'static frame bias' identified in video-lanaguage datasets *i.e.* models overly focus on information shortcuts from a single frame (or very few) of an input video, effectively ignoring the temporal information that *should* need to be exploited to succeed.

The diagnostic dataset CATER [66] builds the on design framework of CLEVR with 'compositional' action recognition and object location. CATER is designed to have 'fully observable and controllable scene bias' such that success on the dataset 'truly requires' temporal reasoning.

Lei et al. [119] demonstrate that models using only a *single* for the video-QA and text-to-video retrieval tasks in their experiments outperform previous approaches. The authors further introduce 2 new tasks, 'template retrieval' and 'label retrieval'

that they believe 'require temporal modelling'.

I qualitatively explore the tendency of video tasks to tend towards static positions through my visual modelling work in Chapter 5.

2.3.2.7 'Directly' Strengthen Visual Contributions

Though research discussed in the previous subsection is aimed at creating conditions such that the visual information is required and not ignored, such efforts strengthen visual contributions *indirectly* *i.e.* they do not 'increase the power' of the visual inputs to contribute. There *does* however exist some few works that aim to take a more *direct* approach to improve visual information exploitation:

The Human Importance-aware Network Tuning 'HINT' algorithm [170] uses small 'hints' of attention provided by humans to find out if such human-guided attention for visual inputs is an improvement versus allowing the machine learning model to learn as it normally would. The authors demonstrate that by providing human hints for just 6% of the training data yields a substantial improvement of 5% on the VQA-CP v1 test set. In contrast, Shrestha et al. [175] later find that the similar 'human grounding' methods are not responsible for improved accuracy, but rather give rise to regularisation effect that reduces language priors.

I find these conflicting results to be a fascinating component of the broader hypothesis 'are more human-like behaviours desirable to encourage for deep learning?'. My work in Chapter 6 tests this hypothesis through an augmented labelling scheme for multimodal tasks using neurolinguistic scores from humans for visual concepts.

My work in Chapter 5 aims to *directly* improve the power of visual contributions through a visual generative pretraining scheme that would yield stronger visual representations (what BERT embeddings are to GloVe) for use in downstream visual tasks *i.e.* VQA.

CHAPTER 3

On Modality Bias in the TVQA Dataset

This chapter highlights and analyses textual biases in the TVQA video-QA dataset. The contents of this chapter draw heavily from my BMVC paper: 'On Modality Bias in the TVQA Dataset' [204]. The work in this chapter was conducted in 2019, when the textual biases of *VQA* datasets were a very active area of research, but *video-QA* datasets had yet to be similarly explored. This chapter also discusses the benchmark *model* that was proposed alongside the dataset. This model is referred to as the 'TVQA model'. Readers should note that the term 'TVQA' will most often refer to the TVQA dataset, but also may contextually refer to the aforementioned TVQA model.

3.1 Introduction

Videos promise more raw visual content than still images used in VQA, and include temporal dependencies that models can exploit. Many notable video-QA datasets have been developed to facilitate multimodal and temporal methodologies for deep learning: MovieQA [184], MovieFIB [136], PororoQA [108], TGIF-QA [93],

YouTube2Text-QA [220], EgoVQA [53] and TVQA [116]. The TVQA[1] dataset in particular was designed to address shortcomings in previous datasets: It is relatively large; uses longer clips and realistic video content; and provides timestamps allowing the identification of the subtitles and video frames relevant to a given question. Most notably, the questions were specifically designed to encourage multimodal reasoning: *i.e.* models built to answer the questions should require both visual *and* textual cues simultaneously. To achieve this, Amazon Mechanical Turk workers were asked to design two-part compositional questions with a 'main' part ("What was House saying..") and a 'grounding' part ("..before he leaned over the bed?"). The authors claimed this will naturally produce questions that require both visual and language information to answer since "people often naturally use visual signals to ground questions in time". Despite these specific efforts to ensure that "questions require both vision understanding *and* language understanding", my work in this chapter shows that in practice this is not the case. I demonstrate that the subtitles are informative enough to answer the majority of questions in TVQA *without* requiring complementary visual information as intended. I show that 68% of the questions can be correctly answered using only the subtitles. Adding the visual information merely increases the accuracy to ∼72%, yet without subtitles this drops by ∼27%. The TVQA authors stress the importance of subtitles in video-QA "because it reflects the real world, where people interact through language". Though this argument has merit and subtitles substantially improves performance on TVQA, my work in this chapter finds their inclusion actively discourages multimodal reasoning, and that the subtitles *dominate* rather than *complement* the video features. The TVQA+ dataset [117] is not considered in this study. Despite TVQA+ providing improved timestamp annotations, it is a substantially smaller subset of TVQA. Furthermore, the 'visual concept words' collected for TVQA+ sample from the questions and the correct answers. This means that models trained on TVQA+ will be trained to use additional textual hints disproportionately from correct answers [2]. This defeats the purpose of video-QA models as it assumes the correct answer is known to the model

[1] http://tvqa.cs.unc.edu
[2] TVQA+: http://tvqa.cs.unc.edu/download_tvqa_plus.html

during feature extraction.

The main contributions of this chapter are:

1. An evaluation framework for multimodal datasets that is suitable for use in future datasets and models, but applied here to the TVQA dataset.

2. I introduce an 'inclusion-exclusion measure' (IEM) method of defining data subsets that are correctly answered by a single modality or a combination of modalities. This methodology is applicable to future datasets, but is introduced and applied here on the TVQA dataset.

3. A topical case study of the ways in which even a carefully designed multimodal dataset may still exhibit problematic modality biases.

4. Extensive analysis of the performance of the TVQA model and dataset per modality and feature type, including the relative contributions of each feature, notably finding that models trained with subtitles learn to suppress video feature contributions.

5. Demonstration of an inherent reliance in the questions on the subtitles rather than multimodal interactions as intended.

6. State of the art results[3] achieved using the baseline TVQA model by simply improving its textual reasoning with 'contextual' word embeddings from the BERT language model [45].

7. I demonstrate that the model-agnostic RUBi learning strategy [20] fails to improve TVQA performance, inline with other textually biased datasets.

The evaluation framework and proposed subsets are available on GitHub[4].

[3] At time of publication.
[4] Available at https://github.com/Jumperkables/tvqa_modality_bias

3.2 The TVQA Dataset

The TVQA dataset [116] is designed to address the shortcomings of previous video-QA datasets. It has substantially longer clip lengths than other datasets and is based on TV shows instead of cartoons, giving it realistic video content with simple coherent narratives. It contains over 150k QA pairs. Each question is labelled with timestamps for the relevant video frames and subtitles. The questions were gathered using AMT workers. Most notably, the questions were specifically designed to encourage multimodal reasoning by asking the workers to design two-part compositional questions. The first part asks a question about a 'moment' and the second part localises the relevant moment in the video clip *i.e.* [What/How/Where/Why/Who/...] — [when/before/after] —, *e.g. [What] was House saying [before] he leaned over the bed?*. The authors argue this facilitates questions that require both visual and language information since "people often naturally use visual signals to ground questions in time". The authors *do* identify certain biases in the dataset *e.g.* they find that the average length of correct answers are longer than incorrect answers. They analyse the performance of their proposed baseline model with different combinations of visual and textual features on different question types they have identified. However, they didn't note the substantial performance of their baseline model on either visual or textual features alone.

Other 'Type'	Example
Spelling Variation	'*Whom* did Roger say was following him after he made the drop?'
Typo	'**tWhat** was the reason House said they should do a brain biopsy when they were discussing options of what to do?'
Did/*Does*	'*Did* Joey walk into the room before or after Chandler?'
Double '*When*' Question	'*When* did Lucas say he made the video *when* he was showing to Beckett and Castle?'

Table 3.1: Example questions from 'other' question type category. The 'other' category makes up 1.1% of the validation set.

3.3 Experimental Framework

The evaluation framework I present here is built on the original TVQA model and is designed to assess the contributions of visual and textual information in a multi-modal dataset. However, inline with one of the guiding principals of this thesis, the general approach of this chapter is invariant to model *and* dataset. The goal is to identify any inherent biases towards either modality. As such, I focus the analysis of the model on the processing streams of the visual and textual features, and the use of 'context matching'. This provides a powerful tool to assess in isolation the contribution of the individual feature types and any combination of them.

3.3.1 Model Definition

The model takes as inputs:

- A question q (13.5 words on average)
- 5 potential answers $\{a_i\}_{i=0}^{4}$ (each between 7-23 words)
- A subtitle S
- A video-clip V (∼60-90s at 3FPS)

and outputs the predicted answer. As the model can either use the entire video-clip and subtitle or only the parts specified in the timestamp, I refer to the sections of video and subtitle used as segments from now on. Figure 3.1 demonstrates the textual and visual streams and their associated features in the model architecture.

3.3.2 ImageNet Features

Each frame is processed by a ResNet101 [83] pretrained on ImageNet [44] to produce a 2048-d vector. These vectors are then L2-normalised and stacked in frame order: $V^{img} \in \mathbb{R}^{f \times 2048}$ where f is the number of frames used in the video segment.

3.3.3 Regional Features

Each frame is processed by a Faster R-CNN [167] trained on Visual Genome [110] in order to detect objects. Each detected object in the frame is given a bounding box, and has an affiliated 2048-d feature extracted. Since there are multiple objects detected per frame (I cap it at 20 per frame), it is difficult to efficiently represent this in time sequences [116]. The model uses the top-K regions for all detected labels in the segment as in Anderson et al. [7] and Karpathy and Fei-Fei [102]. Hence the regional features are $V^{reg} \in \mathbb{R}^{n_{reg} \times 2048}$ where n_{reg} is the number of regional features used in the segment. Where you may consider ImageNet features as a representation of information for the entire frame, regional features are information representations of *specific objects* in that frame.

3.3.4 Visual Concepts

The classes or labels of the detected regional features are called 'Visual Concepts'. Yin and Ordonez [224] found that simply using detected labels instead of image features gives comparable performance on image captioning tasks. Importantly they argued that combining CNN features with detected labels outperforms either approach alone. Visual concepts are represented as either GloVe [158] or BERT [45] embeddings $V^{vcpt} \in \mathbb{R}^{n_{vcpt} \times 300}$ or $\mathbb{R}^{n_{vcpt} \times 768}$ respectively, where n_{vcpt} is the number of visual concepts used in the segment.

3.3.5 Text Features

In the evaluation framework, the model encodes the questions, answers, and subtitles using either GloVe ($\in \mathbb{R}^{300}$) or BERT embeddings ($\in \mathbb{R}^{768}$). Formally, $q \in \mathbb{R}^{n_q \times d}$, $\{a_i\}_{i=0}^{4} \in \mathbb{R}^{n_{a_i} \times d}$, $S \in \mathbb{R}^{n_s \times d}$ where n_q, n_{a_i}, n_s is the number of words in q, a_i, S respectively and $d = 300, 768$ for GloVe or BERT embeddings respectively.

3.3.6 Context Matching

Context matching refers to context-query attention layers recently adopted in machine comprehension [172, 226]. Given a context-query pair, context matching layers

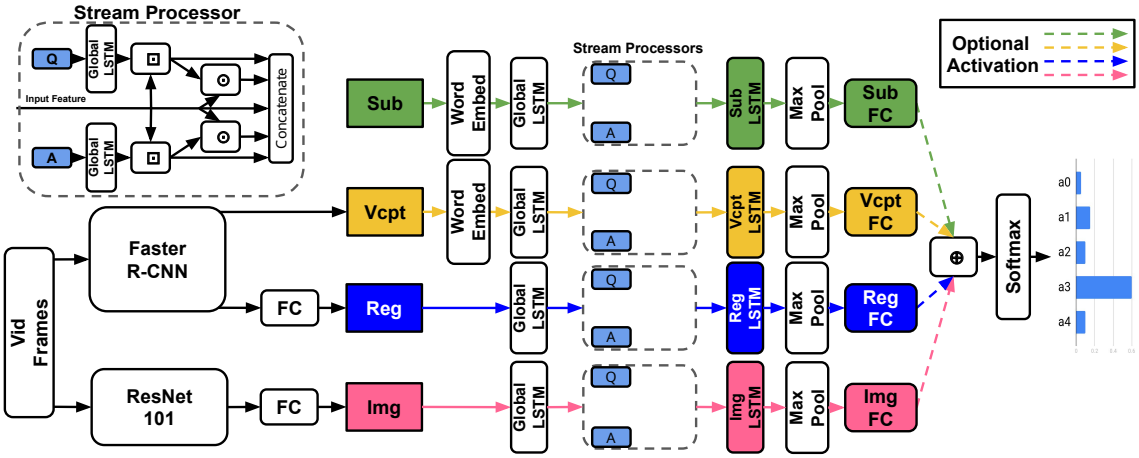


Figure 3.1: TVQA Model. $\odot$ = element-wise multiplication, $\oplus$ = element-wise addition, $\boxdot$ = context matching. FC = stacked fully-connected linear layers. Any of the feature streams may be enabled/disabled. Each model variation is trained end-to-end with each of the activated features.

return 'context aware queries'.

3.3.7 Bilinear Pooling: MCB and MFH

Bilinear pooling (BLP) refers to a family of operations recently developed for fusing features from different modalities predominantly for visual question answering (VQA) models. The various BLP techniques and their properties are the primary focus of Chapter 4. In contrast, the work in this chapter instead *applies* the following two BLP techniques simply in their capacity as examples of 'joint representations' [10]: multimodal compact bilinear pooling (MCB) [58], and multimodal factorised higher-order bilinear pooling (MFH) [229]. Where I functionally describe the practical use of MCB for this scope of this chapter here, I *formally and fully* introduce MCB and MFH alongside the other bilinear methods in Sections 4.2.3 and 4.2.7 respectively.

3.3.8 Model and Framework Details

In my evaluation framework, any combination of subtitles or visual features can be used. All features are mapped into word vector space through a tanh non-linear layer. They are then processed by a shared bi-directional LSTM [87, 72] ('Global LSTM' in Figure 3.1) of output dimension 300. Features are context-matched with

the question and answers. The original context vector is then concatenated with the context-aware question and answer representations and their combined element-wise product ('Stream Processor' in Figure 3.1, *e.g.* for subtitles S, the stream processor outputs $[F^{sub};A^{sub,q};A^{sub,a_{0-4}};\ F^{sub}\odot A^{sub,q};F^{sub}\odot A^{sub,a_{0-4}}]\in \mathbb{R}^{n_{sub}\times 1500}$ where $F^{sub}\in \mathbb{R}^{n_s\times 300}$. Each concatenated vector is processed by their own unique bi-directional LSTM of output dimension 600, followed by a pair of fully connected layers of output dimensions 500 and 5, both with dropout 0.5 and ReLU activation. The 5-dimensional output represents a vote for each answer. The element-wise sum of each activated feature stream is passed to a softmax producing the predicted answer ID. All features remain separate through the entire network, effectively allowing the model to choose the most useful features. This makes this model a strong tool in assessing the biases towards certain features in the dataset.

3.3.9 Further Experimental Setup Details

My experiments are on the TVQA dataset and I use and adapt the code provided by the authors[5]. Due to their size, the regional features are unavailable for download and I extract them myself following the author's instructions. The models are trained on an RTX 2080-ti GPU with batch size 32 and a rectified-Adam solver [130]. I use a pretrained, non-finetuned BERT embedding layer using the uncased base tokenizer[6]. When using regional features I use the top 20 detections per video segment. All further settings are as described in TVQA, most notably: I use 6B-300d GloVe embeddings and all word embedding layers are frozen during training. I use the timestamps annotations and train the model until improvements on the validation set accuracy is not made for 3 epochs. I check validation and training set accuracies every 400 iterations, except for the models that include regional features where I check every 800 iterations as these run substantially slower. In this study I control for the modality used in order to isolate its influence on the performance of the overall model.

[5]https://github.com/jayleicn/TVQA
[6]https://github.com/huggingface/transformers

3.4 Results and Discussion

3.4.1 Feature Contributions

Table 3.2 shows that all evaluated models trained with subtitles substantially outperform models trained without them.

Model	Text	Val Set	Train Set
V	GloVe	45.39%	60.82%
V	BERT	43.44%	52.76%
I	GloVe	44.86%	61.52%
I	BERT	44.44%	65.02%
R	GloVe	42.36%	54.83%
R	BERT	42.53%	53.85%
VIR	GloVe	46.72%	61.10%
VIR	BERT	44.61%	61.38%
S	GloVe	66.07%	76.42%
S	BERT	68.30%	80.77%
SI	GloVe	67.78%	78.78%
SI	**BERT**	**70.56%**	84.84%
SVI	GloVe	69.34%	78.90%
SVI	**BERT**	**72.13%**	86.84%
SVIR	GloVe	69.53%	80.08%
SVIR	**BERT**	**71.80%**	81.58%
STAGE [117]	GloVe	66.92%	-
STAGE [117]	BERT	70.50%	-
VSQA [218]	GloVe	67.70%	-
VSQA [218]	**BERT**	**72.41%**	-
Human	-	93.44%	-

Table 3.2: Each experiment is a separate end-to-end model. *E.g.* 'SI with BERT' is the submodel of subtitles and ImageNet features (green and pink in Figure 3.1) with BERT embeddings used for the subtitles, questions and answers (random choice accuracy is 20%). Models shown in **bold** surpass the SOTA (at time of publication [204]). Models use timestamp annotations except *STAGE* which instead uses 'temporal supervision'.

3.4.1.1 Models with Subtitles

Each BERT variation that includes subtitles gains at least 2% accuracy compared to GloVe, leading to the SI (*i.e.* subtitle+ImageNet features), SVI, and SVIR variations achieving state-of-the-art results. Models trained using only subtitles achieve

+20% accuracy using GloVe and +23% using BERT embeddings when compared to the best performance using any combination of video features. This implies that the subtitles are the most informative features in answering the majority of the questions.

3.4.1.2 Models without Subtitles

With GloVe embeddings, the most impactful video feature is the visual concepts, which increases performance by 0.5%, following a trend in image captioning [224]. Similarly, I find that using image features and visual concepts combined outperforms using either alone. However, using BERT with just visual concepts drops performance by ∼2%. I theorise this is due to the contextual nature of the BERT embeddings hindering the model by sequentially processing the intrinsically unordered visual concepts.

3.4.2 Subtitles Dominate Instead of Complement

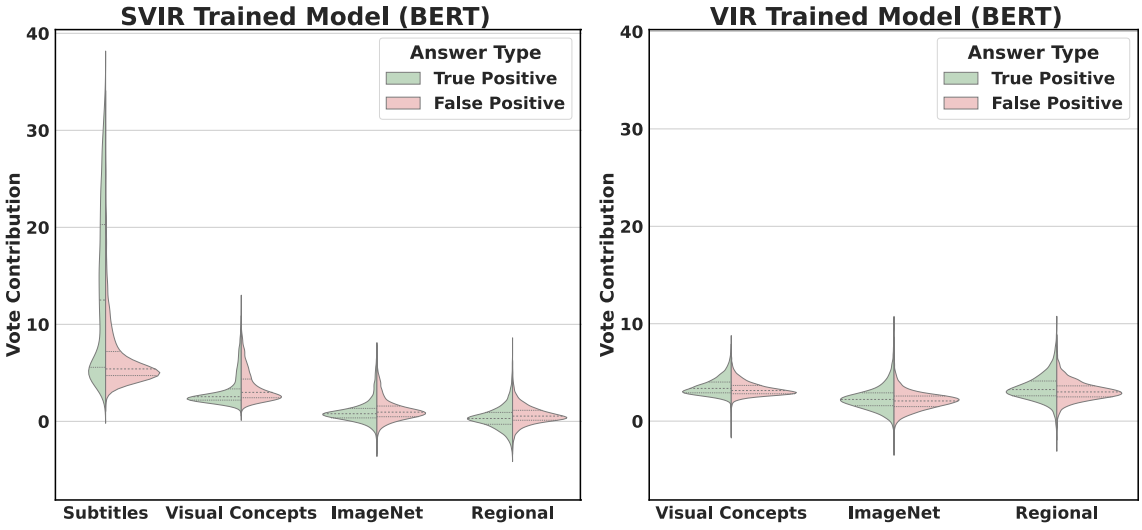


Figure 3.2: Pre-softmax vote contributions for answers in the validation set. Both are BERT models: VIR (left) vs SVIR (right). The dashed lines represent quartiles.

To further analyse the contributions of each feature, I plot the pre-softmax votes for answers between models trained with and without subtitles. Figure 3.2 shows the votes per feature for SVIR and VIR trained model with BERT embeddings, measured in true and false positive answers. In the VIR model (left side of Figure

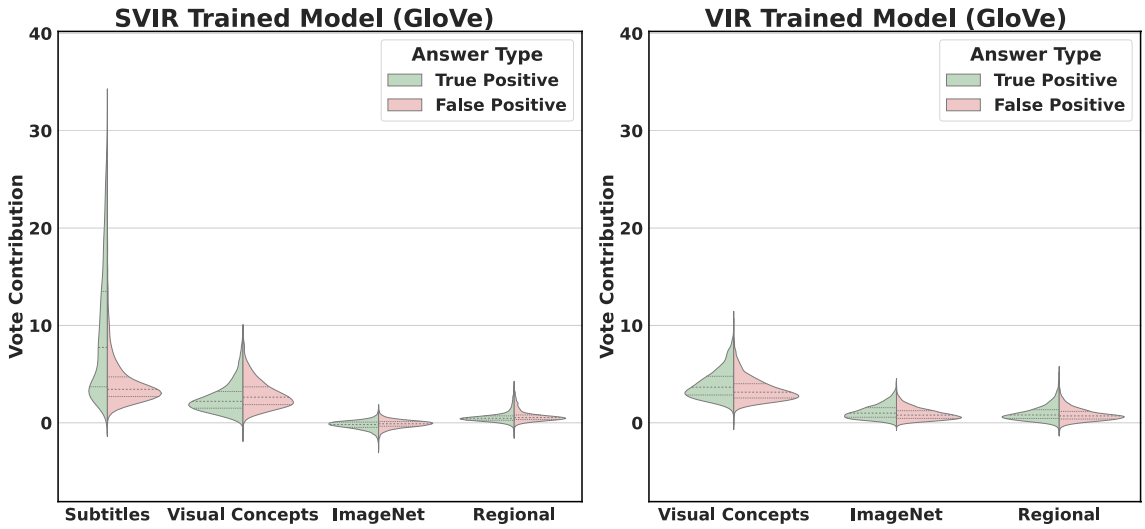


Figure 3.3: Pre-softmax vote contributions for answers in the validation set for the VIR (left) and SVIR (right) trained models with GloVe embeddings. This is the GloVe embedding counterpart to Figure 2.

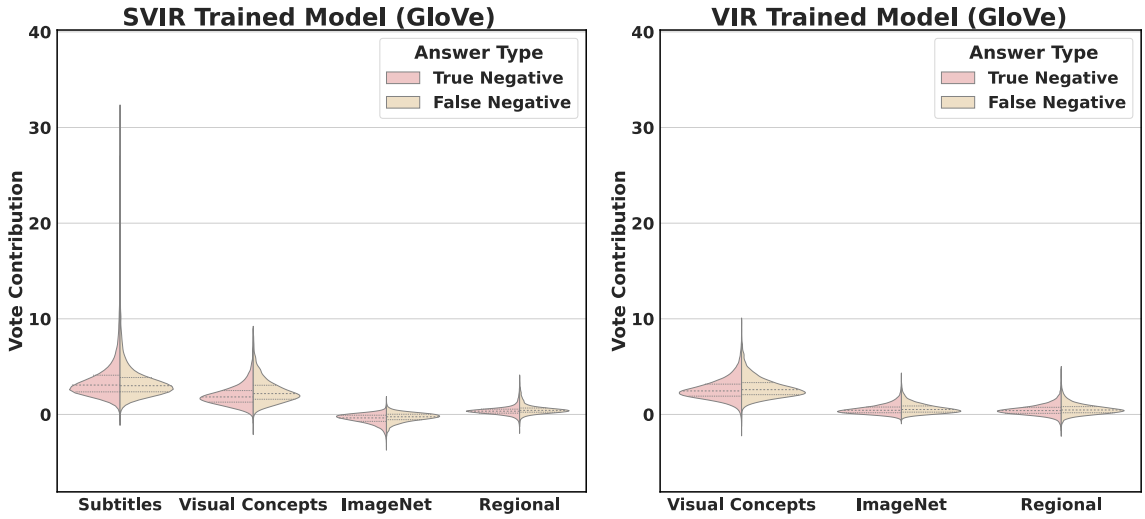


Figure 3.4: Pre-softmax vote contributions for answers in the validation set for the VIR (left) and SVIR (right) trained models with GloVe embeddings.

3.2) I find that all video features similarly contribute to answer votes. When averaged across correct predictions, each feature contributes positively to the correct answer *i.e.* true positives, and contributes less to the other incorrect answers. However, when trained with subtitles in the SVIR model, the subtitles overwhelmingly contribute to the correct answer. Furthermore, in the true positive case, each video feature actually contributes *less* on average than in the accompanying incorrect answers. This is shown in the SVIR model in Figure 3.2, where the quartiles of true positive votes in each video feature are *below* the votes for false positives. This shows

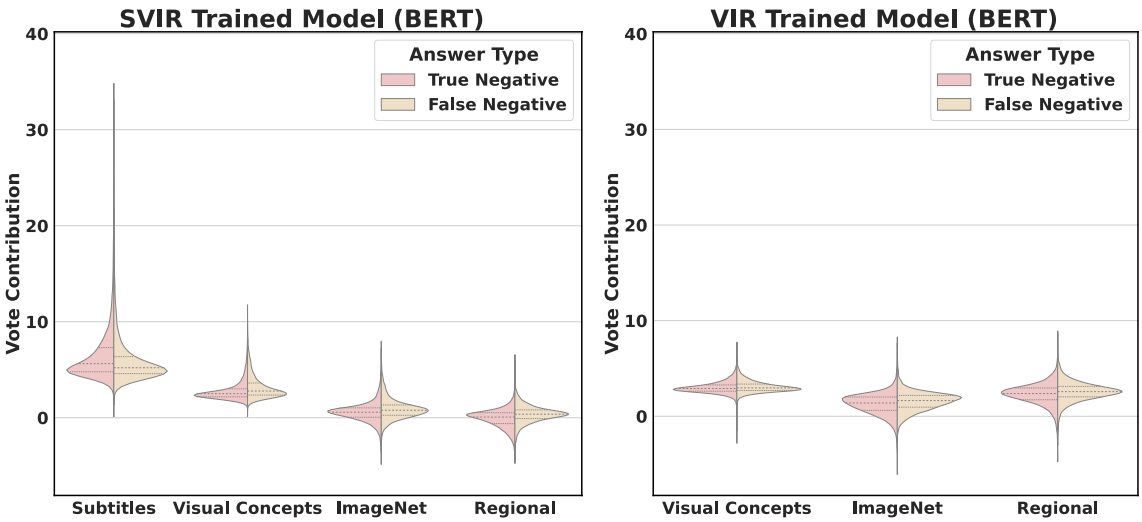


Figure 3.5: Pre-softmax vote contributions for answers in the validation set for the VIR (left) and SVIR (right) trained models with BERT embeddings.

that in case of correct predictions, models trained with subtitles learn, on-average, to suppress the video feature contribution, demonstrating a substantial bias towards subtitles in TVQA. I find similar results in the GloVe models (see Figures 3.8 and 3.9). Strictly speaking, video-QA models that can constructively use video information at all are, to an extent, multimodal as they interpret the video features with respect to the textual questions and answers. However in TVQA, using subtitles 'on-average' actively suppresses these multimodal contributions.

3.4.3 All You Need is BERT

The state-of-the-art STAGE model [117], proposed by the authors of TVQA, improves on the original TVQA model by exploiting spatio-temporal relationships and simultaneously replacing GloVe with BERT embeddings. BERT embeddings have been shown to empirically outperform GloVe embeddings on previous NLP tasks. This comes with a substantial increase of model complexity with over 14 additional layers and steps added to the original model. I show in Table 3.2 that simply upgrading GloVe to BERT embeddings in the relatively simple original model outperforms the more complex STAGE model. Yang et al. [218] present a detailed analysis of BERT on the TVQA dataset and propose a 'V+S+Q+A' model that is structurally similar to the simple TVQA baseline (*i.e.* separate visual and subtitle streams that

additively combines their contributions at the voting stage). Though they do not explore bias in TVQA, they too demonstrate a substantial boost in performance over STAGE by upgrading from GloVe to BERT on a simpler model. This implies that better modelling of the subtitles using BERT in TVQA leads to higher performance regardless of any improvement in modelling the video information. Furthermore, these results indicate that complex models focused on improving more abstract behaviours do not necessarily improve video-QA performance in TVQA. I theorise that these complex models are currently introducing unhelpful overhead and that, if the goal is to increase performance on TVQA, models are best served by exploiting the subtitles. These results also suggest that there is an imbalance in the information contributed by visual and textual modalities. I argue that the contextual nature of BERT embeddings makes them ideal for processing the sequential subtitles which, since TVQA is based on TV shows, often follow a structured narrative.

3.4.4 Dataset Analysis

3.4.4.1 Feature Distributions

I analyse the features that are most useful in answering each of the question types. Figure 3.6 shows that models trained without subtitles substantially underperform (relative to their own model accuracy) on 'which' and 'who' questions. This makes intuitive sense as names and named entities commonly appear in the subtitles. Subtitle models substantially overperform on 'why' and 'how' questions, at ∼82%. Intuitively these question types are harder because the answers are implied rather than concrete and often revolve around explanations that are best represented in language.

As an alternative set comparison measure, I consider the proportion of questions in the validation set that each pair of models answer the same, regardless if the answer is correct or incorrect.

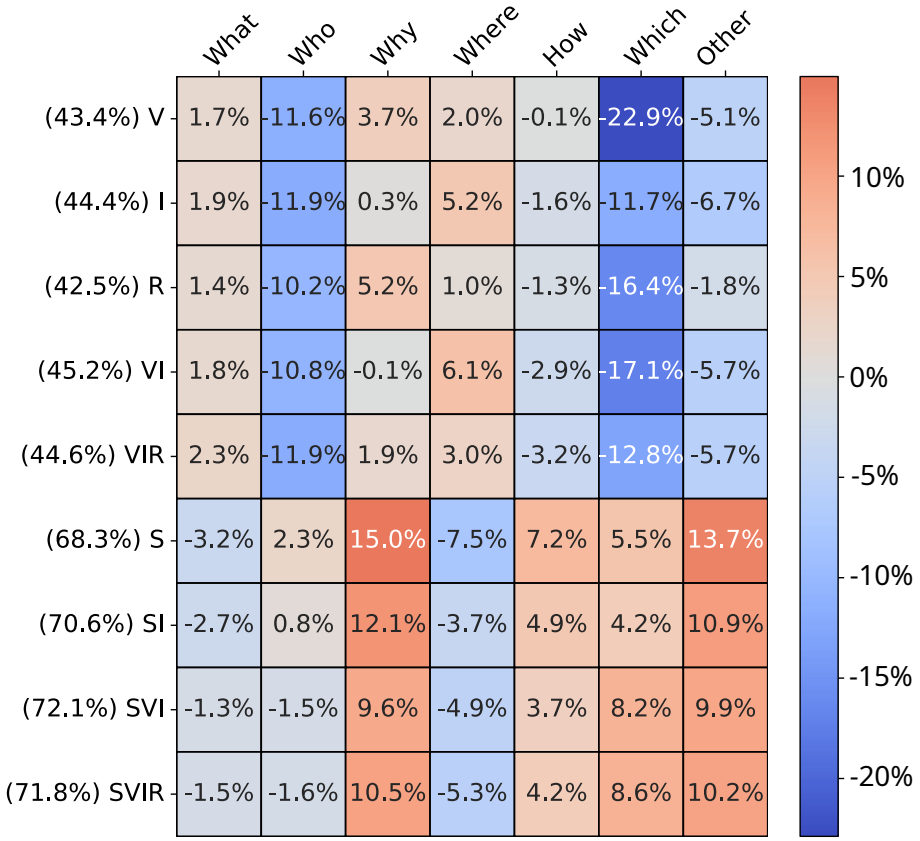


Figure 3.6: Performance of models on each question type (offset from each model's overall accuracy).

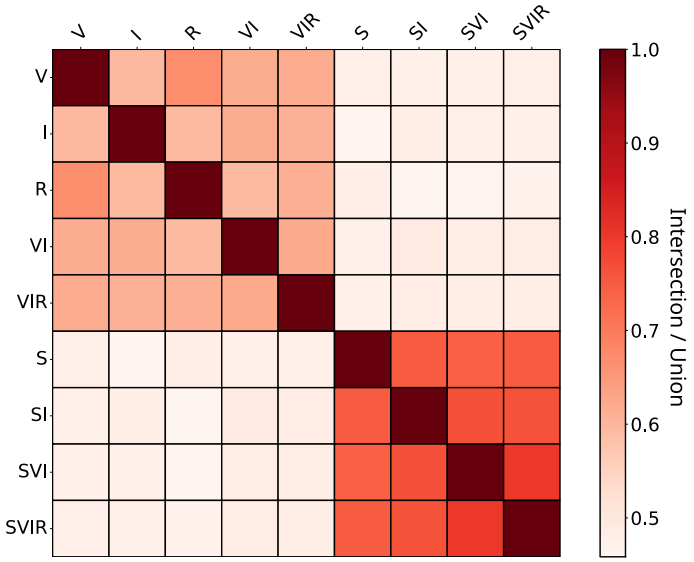


Figure 3.7: IoU of correct answers between models.

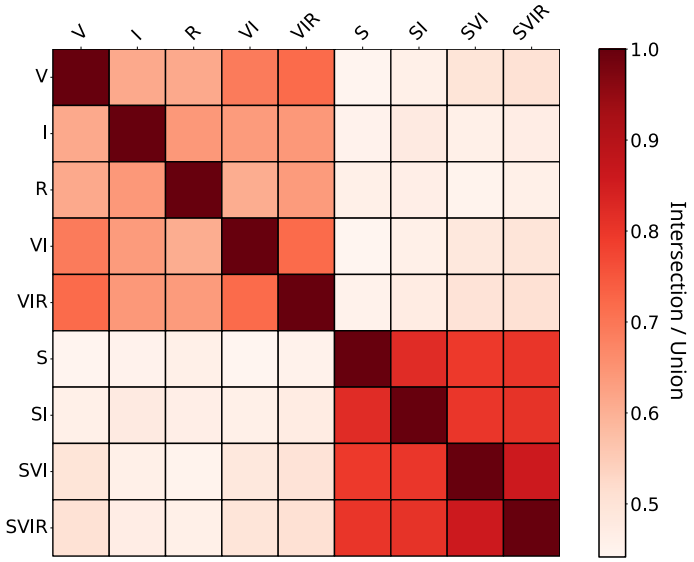


Figure 3.8: Intersection / Union (IoU) score for correct predictions in the validation set between GloVe models.

3.4.4.2 Modality Subsets

My work in this chapter thusfar has focused on carefully demonstrating the existence of language biases in the TVQA dataset. In this section, I instead aim to complement these findings with a proposed solution to the underlying problem of modality biases in datasets. By analysing the similarity between the outputs of the different models, I label each question with the modalities needed to answer it. I isolate subsets of the validation set that are answered correctly using each of the evaluated models.

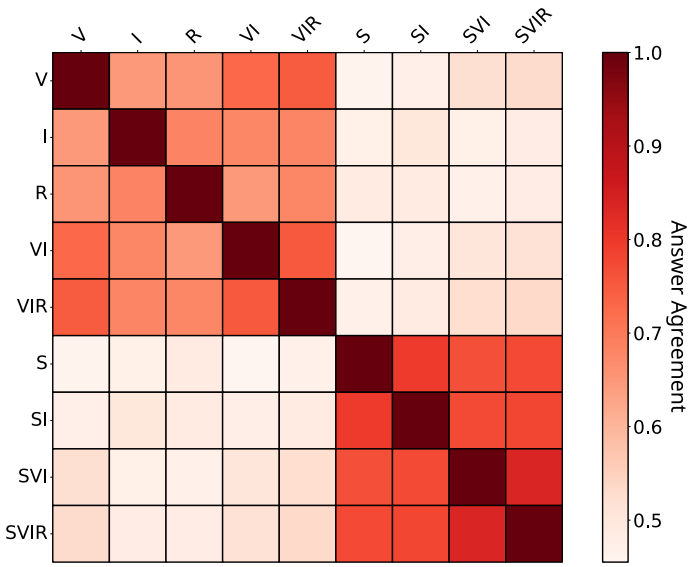


Figure 3.9: Proportion of the validation set that GloVe models answer the same.

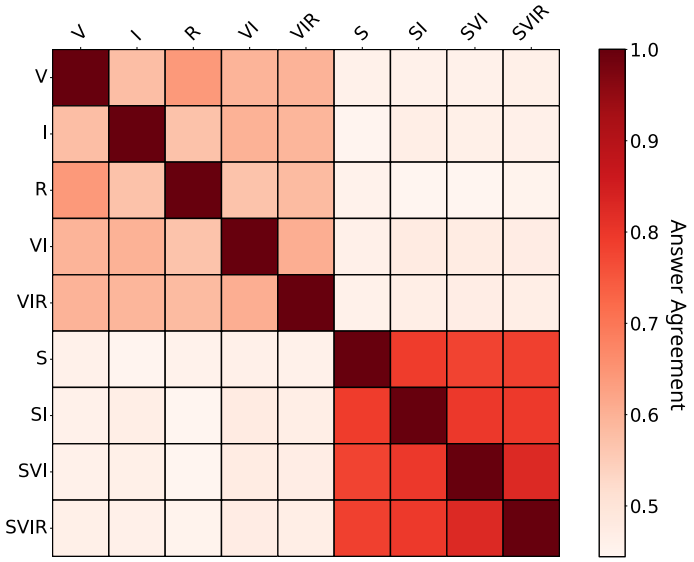


Figure 3.10: Proportion of the validation set that BERT models answer the same.

Figure 3.7 shows that the correctly answered questions of models trained without subtitles have a relatively low intersection over union (IoU) score, approximately 58-68%. Although the models have similar overall accuracies, they seem to perform well on different questions, implying they successfully use information from relatively separate feature types. The overall accuracies of subtitle models are substantially higher, giving higher IoU scores among those models. To inform my recommendation on how to introduce the data subsets, I run a comparative analysis of the outcomes of different groups of models, *i.e.* Group A and Group B, to identify the proportion

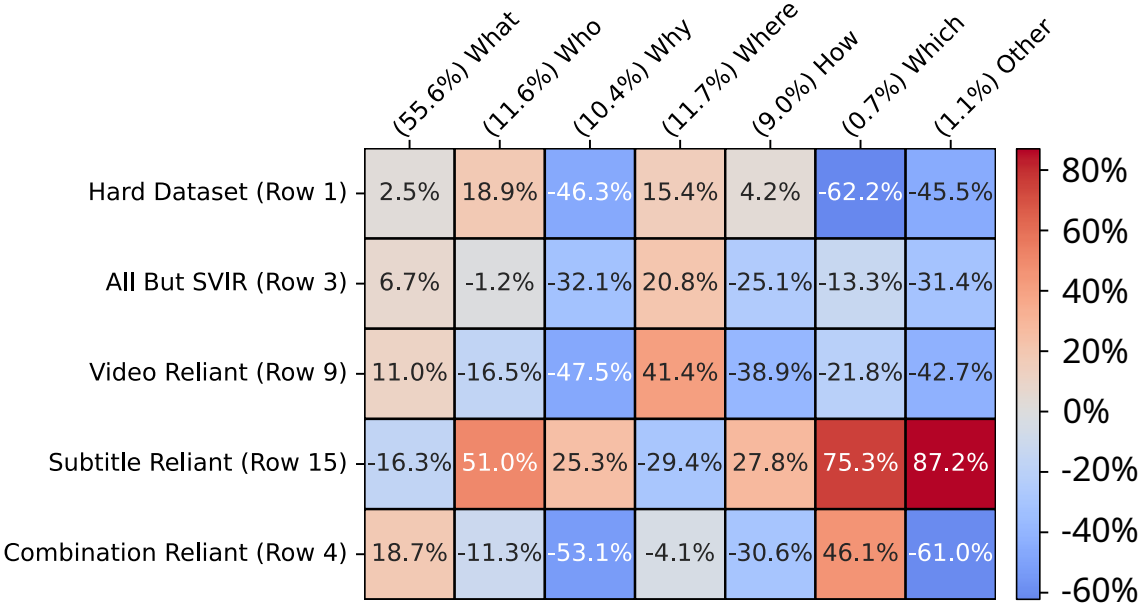


Figure 3.11: The percentage increase of each respective question type, in the specified IEM subset, compared to the overall question type distribution in the validation set. Each of the subsets analysed corresponds to a row in Table 3.3. The 'Hard dataset' alluded to in this figure is the *negation* of the set of answers in Row 1 of Table 3.3.

of the dataset that is answered correctly by models in Group A and incorrectly by all models in Group B. I refer to this measure as the Inclusion-Exclusion Measure (IEM). Table 3.3 summarises this analysis. Row '*All/-*' shows that the union of all correct answers of all models is 90.1%, whereas '*All/Non-Subtitle*', contrasts predictions of all models trained with subtitles versus those trained without subtitles.

This illustrates that 22.7% of the questions cannot be answered by non-subtitle models. '*Non-Subtitle/Subtitle*' shows that only 5.4% can be uniquely answered by the models that don't use subtitles. To identify multimodal reliant questions, I consider those that SVIR can answer correctly but that the unimodal models cannot. 'SVIR/S,V,I,R' shows that 3.79% of the validation set and 2.62% of the training set (see Table 3.4) fits this multimodal criteria (∼4.3k). IEM is a strict and minimal lower bounding method. A less strict method of partitioning the dataset is to consider 'popular vote', *i.e.* a set where the majority vote of the models in question agree on the answer. Though more restrictive than popular vote, IEM removes potential ambiguity, *i.e.* if a question cannot be correctly answered by any subtitle model, then subtitle content is not answering that question. Note that my proposed

Group A	Group B	BERT	GloVe
All	-	90.12%	87.68%
All	*Non-Subtitle*	22.68%	22.91%
All	SVIR	18.32%	18.16%
All	S, V, I, R	7.68%	5.40%
Subtitle	-	84.74%	80.56%
Subtitle	*Non-Subtitle*	22.68%	22.91%
Non-Subtitle	-	67.44%	64.77%
Non-Subtitle	*Subtitle*	5.38%	7.12%
Non-Subtitle	S	16.72%	18.50%
S, V, I, R	-	82.44%	82.28%
S, V, I, R	SVIR	14.44%	14.77%
SVIR	-	71.80%	69.52%
SVIR	S, V, I, R	3.79%	2.01%
S	-	68.30%	66.07%
S	*Non-Subtitle*	17.59%	19.80%
S	V, I, R	21.94%	22.27%
S	VIR	32.83%	30.78%

Table 3.3: The percentages of questions in the validation set that are correctly answered by models in Group A, but incorrectly answered by Group B. *Subtitle*= all models trained with subtitles. S = subtitle-only model. SVIR = model trained with all 4 features. S, V, I, R = group of 4 models each trained with one of the 4 features. The scenarios where Group B is '-' are those which do not consider models that *could not* answer: *e.g.* the top row of this table (*All* | -) simply shows the proportion which all models could answer.

.

subsets are inherently linked to the model. Using my IEM approach, I discount the large amount of questions answered by unimodal models as not multimodal (by definition), providing a valuable starting point. Including better models as they are developed in IEM would provide increasingly better subset splits. To provide insight into how TVQA question information is actually distributed, I present the relative abundance of each question type in my proposed IEM subsets in Figure 3.11. Most notably, 'who' and 'which' questions are more highly concentrated in the 'subtitle reliant' subset. This is unsurprising as the subtitles contain a wealth of named entities and nouns. Conversely, the 'video reliant' dataset contains more 'what' and 'where' questions. Despite the potential benefits of subsets derived from IEM, my results on TVQA imply such subsets are likely to be substantially smaller than the original size of the dataset, and are therefore subject to the drawbacks of

Group A	Group B	BERT Models	GloVe Models
All	-	96.77%	94.54%
All	*Non-Subtitle*	14.32%	14.56%
All	SVIR	15.19%	14.47%
Subtitle	-	94.80%	89.91%
Subtitle	*Non-Subtitle*	14.32%	14.56%
Non-Subtitle	-	82.45%	79.99%
Non-Subtitle	*Subtitle*	1.96%	4.63%
Non-Subtitle	S	12.34%	15.97%
S, V, I, R	-	91.11%	90.52%
S, V, I, R	SVIR	12.15%	12.03%
SVIR	-	81.58%	80.07%
SVIR	S, V, I, R	2.62%	1.58%
S	-	80.77%	76.41%
S	*Non-Subtitle*	10.67%	12.39%

Table 3.4: The percentages of the *training* set that are correctly answered by models in Group A, but incorrectly answered by Group B. *Subtitle models* = {S, SI, SVI, SVIR}, *Non-Subtitle models* = {V, I, R, VI, VIR}. *All models* = *Subtitle* + *Non-Subtitle*. Though considering responses of the training set is inherently flawed due to training bias, it provides a reasonable starting point and considerable size boost to my initially proposed IEM subsets. The scenarios where Group B is '-' are those which do not consider models that *could not* answer: *e.g.* the top row of this table (*All* | -) simply shows the proportion which all models could answer.

small datasets. Though this is not necessarily a problem with the IEM metric itself, it is nonetheless a consequence that must be acknowledged.

3.4.5 Further Experimental Findings

3.4.5.1 Joint Representations Appear Detrimental

Baltrusaitis et al. [10] consider representation as summarising multimodal data "in a way that exploits the complementarity and redundancy of multiple modalities". Joint representations (*e.g.* concatenation, bilinear pooling [62, 58, 109, 106, 13]) combine unimodal signals into the same representation space. However, they struggle to handle missing data [10] as they tend to preserve shared semantics while ignoring modality-specific information [75]. I explore how a joint representation in the TVQA model affects performance as another method of inferring potential unimodal reliances. I create the 'dual-stream' (Figure 3.12) model from the SI TVQA

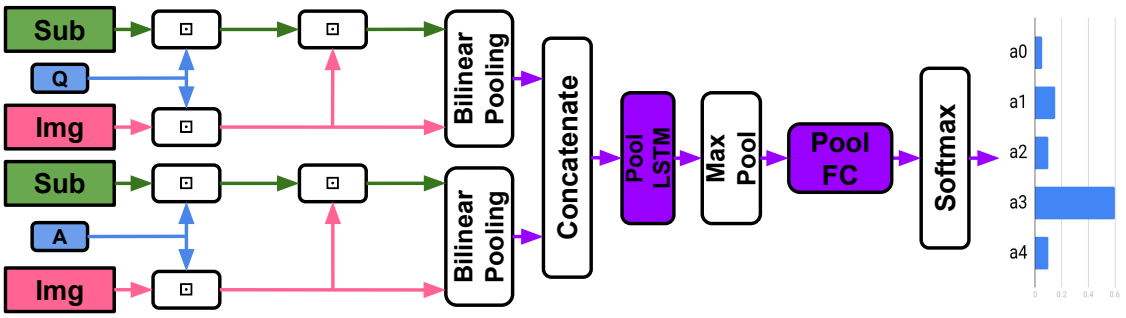


Figure 3.12: The dual-stream model. Both features are integrated into a single adapted 'stream processor'. ⊡ = context matching. BLP is used to fuse S and I features.

baseline model with as few changes as possible:

- I use context matching between subtitle and ImageNet features to allow bilinear pooling at each time step between both modalities.
- I use the new pooled feature as input for a single stream processor.

Table 3.5 shows that both my dual-stream models perform substantially worse than the baseline model. This implies that questions in TVQA do not effectively use a joint representation of its features, and potentially highlights:

- Information from either modality is consistently missing.
- Prioritising 'shared semantics' over 'modality-specific' information harms performance on TVQA.

Both of these possibilities would contradict TVQA's stated aim as a multimodal benchmark.

3.4.5.2 RUBi Doesn't Help

Strong unimodal language biases are prevelent in VQA. As discussed in Chapter 2 earlier, the VQA-CP v1/2 datasets [4] are rearrangements of the VQA v1/2 datasets [9, 70] such that certain kinds of identified QA priors appear exclusively in the training or test sets. Unable to rely on these priors, many VQA baseline model's performance substantially drops. The model-agnostic RUBi strategy [20] uses a text-only variant of a model during training (see this Figure 3.13) to reduce (increase) the

Model	Text	Val Acc
TVQA SI	GloVe	67.78%
TVQA SI	BERT	70.56%
Dual-Stream MCB	GloVe	63.46%
Dual-Stream MCB	BERT	60.63%
Dual-Stream MFH	GloVe	62.71%
Dual-Stream MFH	BERT	59.34%

Table 3.5: Dual-stream vs TVQA SI baseline. The hidden pooling dimension is 1500. Both use only the subtitle and ImageNet features (including questions and answers).

loss, and therefore importance, of highly-biased (visually dependent and difficult) training samples. Shown in Table 3.6, benchmark models using RUBi perform substantially better on the less-textually-biased VQA-CP dataset, implying RUBi successfully discourages models from relying on the now unhelpful text prior shortcuts. Conversely, RUBi harms performance on datasets with greater text biases (VQA v2 test-dev/val), implying RUBi's bias-averse behaviour is actually detrimental on

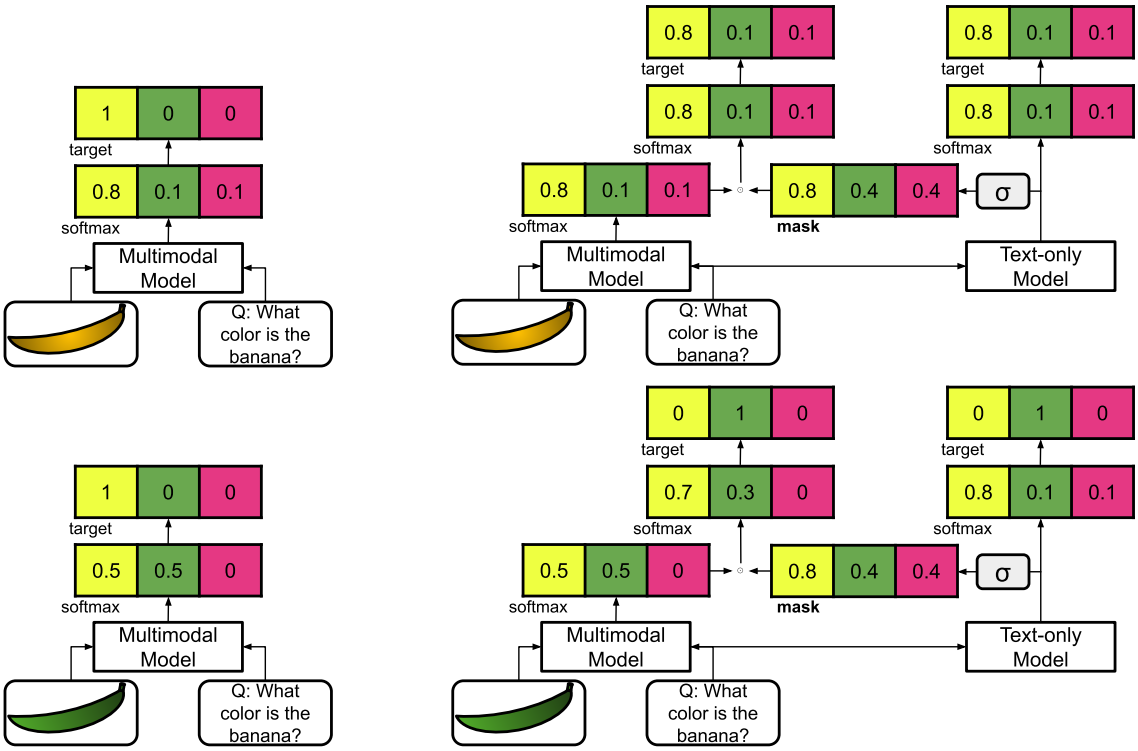


Figure 3.13: The RUBi (reducing unimodal bias) learning strategy used in VQA. The model-agnostic RUBi strategy [20] uses a text-only variant of a model during training to reduce (increase) the loss, and therefore importance, of highly-biased (visually dependent and difficult) training samples.

Model	Dataset	Baseline Acc	RUBi Acc
TVQA SI (GloVe)	TVQA	**67.78%**	67.67%
TVQA SI (BERT)	TVQA	**70.56%**	70.37%
RUBi Baseline [20]	VQA-CP v2 test	38.46%	**47.11%**
SAN [219]	VQA-CP v2 test	33.29%	**36.69%**
UpDn [7]	VQA-CP v2 test	41.17%	**44.23%**
RUBi Baseline [20]	VQA v2 test-dev	**63.10%**	61.16%
RUBi Baseline [20]	VQA v2 val	**64.75%**	63.18%

Table 3.6: TVQA SI model trained on the RUBi criterion [20]. As subtitles can also provide learned text-prior shortcuts, the TVQA text-only model in the RUBi strategy also includes subtitles.

datasets where these shortcuts exist. I find that RUBi fails to improve accuracy on TVQA and in fact *slightly* decreases performance on both BERT and GloVe models, implying that TVQA could benefit heavily as a multimodal benchmark by addressing its own textual priors. To the best of my knowledge, at time of experimentation, I was the first to apply RUBi to a video-QA dataset. I note that subtitles can also provide learned text-prior shortcuts, as such, the TVQA text-only model in the RUBi strategy also includes subtitles.

3.5 Conclusion

I develop a multimodal evaluation framework using the TVQA model that aims to not only assess potential dataset biases, but also circumvent them by isolating and removing problematic questions. I find that information needed to answer questions in the TVQA dataset is concentrated in the subtitles to the extent that video information is suppressed during training, contradicting the multimodal nature TVQA was specially designed to have. I provide an extensive analysis on which question types in TVQA require video or textual features and propose subsets of TVQA, in particular those which require specific features for multimodal reasoning. I achieve state-of-the-art results on the TVQA dataset by simply using BERT embeddings with the TVQA model. This demonstrates that the performance increase in the STAGE model is largely due to improved NLP embeddings. I find further evidence for TVQA's unimodal textual bias through my experiments with joint representations and the RUBi learning strategy. I show that multimodality is not always guar-

anteed in video-QA and suggest that it is challenging to design questions without introducing biases that discourage multimodality. My work in this chapter serves as a crucial reminder that great care should be taken in creating datasets for multimodal reasoning in video-QA. Despite the potential upsides of my proposed IEM metric, my results here on the TVQA dataset imply that the subsets it generates are likely to be substantially smaller than the original size of the dataset on current benchmarks. My work here is applicable to both existing and future multimodal datasets, able to evaluate these datasets on modality-reliant subsets and highlight how a given model performs under different modality conditions.

CHAPTER 4

Bilinear Pooling in Video-QA: Empirical Challenges and Motivational Drift from Neurological Parallels

Where Chapter 3 focuses on problems with the use of a popular multimodal *datasets*, this chapter shifts focus to the problems with the use of a popular multimodal *modelling techniques*. This chapter heavily draws from my published paper 'Bilinear pooling in video-QA: empirical challenges and motivational drift from neurological parallels' [205]. The work in this chapter was undertaken in 2020/2021.

4.1 Introduction

It is essential to develop modelling and learning strategies with the capacity to learn complex and nuanced multimodal relationships and representations. A particularly notorious solution to learning multimodal relationships in VQA is the family of bilinear pooling (BLP) operators [62, 106, 228, 12, 229, 13]. A bilinear (outer product) expansion is thought to encourage models to learn interactions between 2 feature spaces and has experimentally outperformed 'simpler' vector operations (*i.e.* concatenation and element-wise-addition/multiplication) on VQA benchmarks. Though successive BLP techniques focus on leveraging higher performance with lower com-

putational expense, which I wholeheartedly welcome, the context of their use has subtly drifted from application in earlier bilinear models *e.g.* where in Lin et al. [127] the bilinear mapping is learned between convolution maps (a tangible and visualisable parameter), from compact BLP [62] onwards the bilinear mapping is learned between indexes of deep feature vectors (a much less tangible unit of representation). Though such changes are not necessarily problematic and the improved VQA performance they have yielded is valuable, they represent a broader trend of the use of BLP methods in multimodal fusion being justified *only* by empirical success. This can be limited as a measure of objective success as we have seen that a higher accuracy does not always imply more desirable behaviour. Furthermore, despite BLP's history of success in text-image fusion in VQA, it has not yet gained such notoriety in video-QA. Though BLP methods have continued to perform well on video tasks when fusing vision and *non-textual* features [90, 237, 155, 212, 43, 198, 42, 181], BLP has recently been overshadowed by other vision and *textual* feature fusion techniques in video-QA [107, 122, 61, 129, 125]. In this chapter, I aim to add a new perspective to the empirical and motivational drift in BLP. In doing so, I aim to offer helpful insight to the usage of this popular modality fusion technique for use in the multimodal benchmark models and datasets of the future. My contributions include the following:

1. I carefully and experimentally ascertain the empirical strengths and limitations of BLP as a multimodal text-vision fusion technique on 2 models (TVQA baseline and heterogeneous-memory-enchanced 'HME' model) and 4 datasets (TVQA, TGIF-QA, MSVD-QA and Ego-VQA). To this end, my experiments include replacing feature concatenation in the existing models with BLP, and a modified version of the TVQA baseline to accommodate BLP that I name the 'dual-stream' model. Furthermore, I contrast BLP (classified as a 'joint' representation by [10]) with deep canonical cross correlation (a 'co-ordinated representation'). I find that my relatively simple integration of BLP does not increase, and mostly harms, performance on these video-QA benchmarks. I discuss how the decreased performance demonstrated by each *individual* experiment uniquely highlights the shortcomings of BLP as a video-QA modality

fusion methodology.

2. I discuss the motivational origins of BLP and share my observations of bilinearity in text-vision fusion.

3. By observing trends in recent work using BLP for multimodal video tasks and recently proposed theoretical multimodal fusion taxonomies, I offer insight into why BLP-driven performance gain for video-QA benchmarks may be more difficult to achieve than in earlier VQA models.

4. I identify temporal alignment and inefficiency (computational resources *and* performance) as key issues with BLP as a multimodal text-vision fusion technique in video-QA, and highlight concatenation and attention mechanisms as an ideal alternative for use in the models of the future.

5. In parallel with the *empirically justified* innovations driving BLP methods, I explore the often-overlooked similarities of bilinear and multimodal fusion to neurological theories *e.g.* Dual Coding Theory [152, 153] and the Two-Stream Model of Vision [69, 142], and propose several potential *neurologically justified* alternatives and improvements to existing text-image fusion. I highlight the latent potential already in existing video-QA datasets to exploit neurological theories by presenting a qualitative analysis of the occurrence of neurolinguistically 'concrete' words in the vocabularies of the textual components of the 4 video-QA datasets I experiment with.

4.2 Background: Bilinear Pooling

In this section I outline the development of BLP techniques, highlight how bilinear models parallel the two-stream model of vision, and discuss where bilinear models diverged from their original motivation.

4.2.1 Concatenation

Early works use Vector concatenation to project different features into a new joint feature space. Zhou et al. [236] use vector concatenation on the CNN image and text features in their simple baseline VQA model. Similarly, Lu et al. [135] concatenate image attention and textual features. Vector concatenation is a projection of both input vectors into a new 'joint' dimensional space. Vector concatenation as a multimodal feature fusion technique in VQA is considered a baseline and has historically been empirically outperformed in VQA by the following bilinear techniques.

$$\mathbf{a} = (a_0, a_1, ..., a_{m-1})$$
$$\mathbf{b} = (b_0, b_1, ..., b_{n-1})$$
$$Concatenation(\mathbf{a}, \mathbf{b}) = (a_0, a_1, ..., a_{m-1}, b_0, b_1, ..., b_{n-1})$$

where $\mathbf{a}, \mathbf{b}$ are vectors of dimension m, n respectively.

4.2.2 Bilinear Models

Working from the observations that "perceptual systems routinely separate 'content' from 'style' ", Tenenbaum and Freeman [185] proposed a bilinear framework on these 2 different aspects of purely visual inputs. They find that the multiplicative bilinear model provides "sufficiently expressive representations of factor interactions". The bilinear model in [127] is a 'two-stream' architecture where distinct subnetworks model temporal and spatial aspects. The bilinear interactions are between the outputs of 2 CNN streams, resulting in a bilinear vector that is effectively an outer product directly on convolution maps (features are aggregated with sum-pooling). This makes intuitive sense as individual convolution maps represent specific patterns. It follows that learnable parameters representing the outer product between these maps learn weightings between distinct and visualisable patterns directly. Interestingly, both [185, 127] are reminiscent of two-stream hypothesises of visual processing in the human brain [69, 143, 141, 68, 142] (discussed in detail later). Though these models focus on only visual content, their generalisable two-factor frameworks would later be inspiration to multimodal representations. Fully bilinear

representations using deep learning features can easily become impractically large, necessitating informed mathematical compromises to the bilinear expansion.

$$\mathbf{a} = (a_0, a_1, ..., a_{m-1})$$
$$\mathbf{b} = (b_0, b_1, ..., b_{n-1})$$
$$`Fully' Bilinear(\mathbf{a}, \mathbf{b}) = \mathbf{a} \otimes_{outer} \mathbf{b}$$

where $\mathbf{a}, \mathbf{b}$ are vectors of dimension m, n respectively, and $\otimes_{outer}$ is the outer product.

4.2.3 Compact Bilinear Pooling

Gao et al. [62] introduce 'Compact Bilinear Pooling', a technique combining the count sketch function [25] and convolution theorem [47] in order to 'pool' the outer product into a smaller bilinear representation. The count sketch function projects vector $\mathbf{v} \in \mathbb{R}^n$ to $\mathbf{y} \in \mathbb{R}^d$. Initialise 2 vectors $\mathbf{s} \in \{-1, 1\}^n$ and $\mathbf{h} \in \{1, ..., d\}^n$: Each coefficient of $\mathbf{s}$ is initialised to 1 or -1, and $\mathbf{h}$ maps each index i in input $\mathbf{v}$ to an index j in the output $\mathbf{y}$. $\mathbf{s}$ and $\mathbf{h}$ are randomly initialised from a normal distribution and are constant for all further count sketch calls. For each element of the input $\mathbf{v}[i]$, we find it's destination index in the output $\mathbf{y}$, $j = \mathbf{h}[i]$, and add $\mathbf{s}[i] \cdot \mathbf{v}[i]$ to output $\mathbf{y}[i]$. The count sketch function Ψ has the favourable property

$$E[\Psi(\mathbf{x}, \mathbf{h}, \mathbf{s}) \odot \Psi(\mathbf{y}, \mathbf{h}, \mathbf{s})] = \mathbf{x} \odot \mathbf{y} \text{ [25]}$$

where E is statistical expectation and $\odot$ is the dot product, *i.e.* the expectation of the dot product of $\Psi(\mathbf{x}, \mathbf{h}, \mathbf{s})$ and $\Psi(\mathbf{y}, \mathbf{h}, \mathbf{s})$ is the dot product of $\mathbf{x}$ and $\mathbf{y}$ (the count sketch algorithm is 'unbiased'). Pham and Pagh [160] show that

$$\Psi(\mathbf{x} \otimes \mathbf{y}, \mathbf{h}, \mathbf{s}) = \Psi(\mathbf{x}, \mathbf{h}, \mathbf{s}) * \Psi(\mathbf{y}, \mathbf{h}, \mathbf{s})$$

where $*$ is the convolution operation, *i.e.* the count sketch of the outer product of 2 vectors is the convolution of the individual count sketches of those vectors. Furthermore, by convolution theorem [47]:

- Convolution in the time domain is equivalent to element-wise multiplication in the frequency domain.

- One can apply the fast Fourier transform (FFT) [34] to the count-sketch of the input vectors to use convolution theorem and simplify the calculation needed from convolution to element-wise multiplication.

- One then applies the inverse fast Fourier transform to the resulting vector, obtaining the count-sketch of the outer product, the compact bilinear representation.

Fukui et al. [58] use compact BLP in their VQA model to learn interactions between text and images *i.e.* multimodal compact bilinear pooling (MCB). I note that for MCB, the learned outer product is no longer on convolution maps, but rather on the indexes of image and textual tensors. Intuitively, a given index of an image or textual tensor is more abstracted from visualisable meaning when compared to convolution map. As far as I am aware, no research has addressed the potential ramifications of this switch from distinct maps to feature indexes, and later usages of bilinear pooling methods continue this trend. Though MCB is significantly more efficient than full bilinear expansions, it still requires relatively large latent dimensions to perform well on VQA (d≈16000).

4.2.4 Multimodal Low-Rank Bilinear Pooling

To further reduce the number of needed parameters, Kim et al. [106] introduce multimodal low-rank bilinear pooling (MLB), which approximates the outer product weight representation W by decomposing it into 2 rank-reduced projection matrices:

$$\mathbf{z} = MLB(\mathbf{x}, \mathbf{y}) = (X^T\mathbf{x}) \odot (Y^T\mathbf{y})$$
$$\mathbf{z} = \mathbf{x}^T W \mathbf{y} = \mathbf{x}^T X Y^T \mathbf{y} = \mathbb{1}^T (X^T\mathbf{x} \odot Y^T\mathbf{y})$$

where $X \in \mathbb{R}^{m \times o}$, $Y \in \mathbb{R}^{n \times o}$, $o < min(m, n)$ is the output vector dimension, $\odot$ is element-wise multiplication of vectors or the Hadamard product, and $\mathbb{1}$ is the unity vector. MLB performs better than MCB in [151], but it is sensitive to hyperparameters and converges slowly. Furthermore, Kim et al. [106] suggest using *Tanh* activation on the output of $\mathbf{z}$ to further increase model capacity.

4.2.5 Multimodal Factorised Low Rank Bilinear Pooling

Yu et al. [228] propose multimodal factorised bilinear pooling (MFB) as an extension of MLB. Consider the bilinear projection matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$ outlined above, to learn output $\mathbf{z} \in \mathbb{R}^o$ one needs to learn $\mathbf{W} = [\mathbf{W}_0, ..., \mathbf{W}_{o-1}]$. I generalise output $\mathbf{z}$:

$$z_i = \mathbf{x}^T \mathbf{X}_i \mathbf{Y}_i^T \mathbf{y} = \sum_{d=0}^{k-1} \mathbf{x}^T a_d b_d^T \mathbf{y} = \mathbb{1}^T (\mathbf{X}_i^T \mathbf{x} \odot \mathbf{Y}_i^T \mathbf{y}) \tag{4.1}$$

Note that MLB is equivalent to MFB where k=1. MFB can be thought of as a 2-part process: features are 'expanded' to higher-dimensional space by $\mathbf{W}_\sigma$ matrices, then 'squeezed' into a "compact ouput". The authors argue that this gives "more powerful" representational capacity in the same dimensional space than MLB.

4.2.6 Multimodal Tucker Fusion

Ben-younes et al. [12] extend the rank-reduction concept from MLB and MFB to factorise the entire bilinear tensor using tucker decomposition [191] in their multimodal tucker fusion (MUTAN) model. I will briefly summarise the notion of rank and the mode-n product to describe the tucker decomposition model.

4.2.6.1 Rank and mode-n product

If $\mathbf{W} \in \mathbb{R}^{I_1 \times, ..., \times I_N}$ and $\mathbf{V} \in \mathbb{R}^{J_n \times I_n}$ for some $n \in \{1, ..., N\}$ then

$$\text{rank}(\mathbf{W} \otimes_n \mathbf{V}) \leq \text{rank}(\mathbf{W})$$

where $\otimes_n$ is the mode-n tensor product:

$$(\mathbf{W} \otimes_n \mathbf{V})(i_1, ..., i_{n-1}, j_n, i_{n+1}, ..., i_N) := \sum_{i_n=1}^{I_n} \mathbf{W}\ (i_1, ..., i_{n-1}, i_n, i_{n+1}, ..., i_N)\mathbf{V}(j_n, i_n)$$

In essence, the mode-n fibres (also known as mode-n vectors) of $\mathbf{W} \otimes_n \mathbf{V}$ are the mode-n fibres of $\mathbf{W}$ multiplied by $\mathbf{V}$ (proof here [74]). See Figure 4.1 for a visualisation of mode-n fibres. Each mode-n tensor product introduces an upper bound to the rank of the tensor. I note that conventionally, the mode-n fibres count from 1 instead of 0. I will follow this convention for the tensor product portion of this chapter to avoid confusion. The tucker decomposition of a real 3^{rd} order tensor

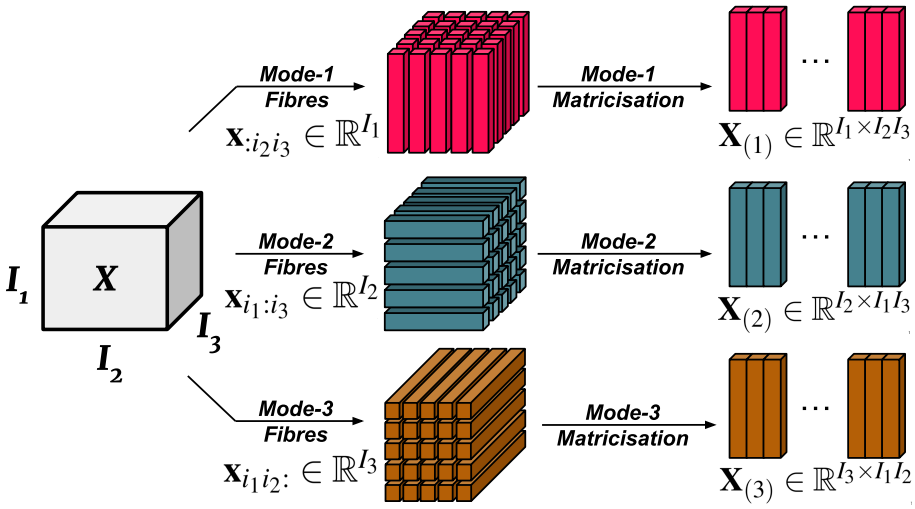


Figure 4.1: Visualisation of mode-n fibres and matricisation.

$\mathbf{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is:

$$\mathbf{T} = \tau \otimes_1 \mathbf{W}_1 \otimes_2 \mathbf{W}_2 \otimes_3 \mathbf{W}_3$$

where $\tau \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ *(core tensor)*, and $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3 \in \mathbb{R}^{d_1 \times d_1}, \mathbb{R}^{d_2 \times d_2}, \mathbb{R}^{d_3 \times d_3}$ *(factor matrices)* respectively.

4.2.6.2 MUTAN

The MUTAN model uses a reduced rank on the core tensor to constrain representational capacity, and the factor matrices to encode full bilinear projections of the textual and visual features, and finally output an answer prediction, i.e:

$$\mathbf{y} = ((\tau \otimes_1 (\mathbf{q}^T \mathbf{W}_q)) \otimes_2 (\mathbf{v}^T \mathbf{W}_v)) \otimes_3 \mathbf{W}_o$$

Where $\mathbf{y} \in \mathbb{R}^{|A|}$ is the answer prediction vector and $\mathbf{q}, \mathbf{v}$ are the textual and visual features respectively. A slice-wise attention mechanism is used in the MUTAN model to focus on the 'most discriminative interactions'. Multimodal tucker fusion is an empirical improvement over the preceeding BLP techniques on VQA, but it introduces complex hyperparameters to refine that are important for relatively its high performance ($\mathbf{R}$ and core tensor dimensions).

4.2.7 Multimodal Factorised Higher Order Bilinear Pooling

All the BLP techniques discussed up to now are 'second-order', *i.e.* take 2 functions as inputs. Yu et al. [229] propose multimodal factorised higher-order bilinear pooling (MFH), extending second-order BLP to 'generalised high-order pooling' by stacking multiple MFB units, i.e:

$$\mathbf{z}_{exp}^{i} = MFB_{exp}^{i}(\mathbf{I}, \mathbf{Q}) = \mathbf{z}_{exp}^{i-1} \odot Dropout(\mathbf{U}^{T}\mathbf{I} \odot \mathbf{V}^{T}\mathbf{Q})$$
$$\mathbf{z} = SumPool(\mathbf{z}_{exp})$$

for $i \in \{1, ..., p\}$ where $\mathbf{I}$, $\mathbf{Q}$ are visual and text features respectively. Similar to how MFB extends MLB, MFH is MFB where $p = 1$. Though MFH slightly outperforms MFB, there has been little exploration into the theoretical benefit in generalising to higher-order BLP.

4.2.8 Bilinear Superdiagonal Fusion

Ben-Younes et al. [13] proposed another method of rank restricted bilinear pooling: Bilinear Superdiagonal Fusion (BLOCK). I will briefly outline block term decomposition before describing BLOCK.

4.2.8.1 Block Term Decomposition

Introduced in a 3-part paper [39, 40, 41], block term decomposition reformulates a bilinear matrix representation as the sum of rank restricted matrix products (contrasting low rank pooling which is represented by only a single rank restricted matrix product). By choosing the number of decompositions in the approximated sum and their rank, block-term decompositions offer greater control over the approximated bilinear model. Block term decompositions are easily extended to higher-order tensor decompositions, allowing multilinear rank restriction for multilinear models in future research. A *block term decomposition* of a tensor $\mathbf{W} \in \mathbb{R}^{I_1 \times, ..., \times I_N}$ is a decomposition of the form:

$$\mathbf{W} = \sum_{r=1}^{R} \mathbf{S}_r \otimes_1 \mathbf{U}_r^1 \otimes_2 \mathbf{U}_r^2 \otimes_3, ..., \otimes_n \mathbf{U}_r^n$$

where $R \in \mathbb{N}^*$ and for each $r \in \{1, ..., R\}, \mathbf{S}_r \in \mathbb{R}^{R_1 \times, ..., \times R_n}$ where each $\mathbf{S}_r$ are 'core tensors' with dimensions $R_n \leq I_n$ for $n \in \{1, ..., N\}$ that are used to restrict the rank of the tensor $\mathbf{W}$. $\mathbf{U}_r^n \in \text{St}(R_n, I_n)$ are the 'factor matrices' that intuitively expand the n^{th} dimension of $\mathbf{S}$ back up to the original n^{th} dimension of $\mathbf{W}$. $\text{St}(a, b)$ here refers to the Stiefel manifold, *i.e.* $\text{St}(a, b)$:$\{\mathbf{Y} \in \mathbb{R}^{a \times b} : \mathbf{Y}^T\mathbf{Y} = \mathbf{I}_p\}$. Figure 4.2 visualises the block term decomposition process.

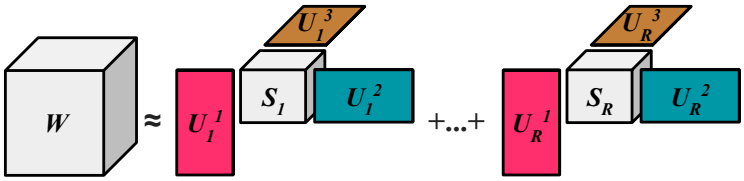


Figure 4.2: Block Term Decomposition (n=3).

4.2.8.2 Bilinear Superdiagonal Model

The BLOCK model uses block term decompositions to learn multimodal interactions. The authors argue that since BLOCK enables "very rich (full bilinear) interactions between groups of features, while the block structure limits the complexity of the whole model", that it is able to represent very fine grained interactions between modalities while maintaining powerful mono-modal representations. The bilinear model with inputs $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$ is projected into o dimensional space with tensor products:

$$\mathbf{z} = \mathbf{W} \otimes_1 \mathbf{x} \otimes_2 \mathbf{y}$$

where $\mathbf{z} \in \mathbb{R}^o$. The superdiagonal BLOCK model uses a 3 dimensional block term decomposition. The decomposition of $\mathbf{W}$ in rank (R_1, R_2, R_3) is defined as:

$$\mathbf{W} = \sum_{r=1}^{R} \mathbf{S}_r \otimes_1 \mathbf{U}_r^1 \otimes_2 \mathbf{U}_r^2 \otimes_3 \mathbf{U}_r^3$$

This can be written as

$$\mathbf{W} = \mathbf{S}^{bd} \otimes_1 \mathbf{U}^1 \otimes_2 \mathbf{U}^2 \otimes_3 \mathbf{U}^3$$

where $\mathbf{U}^1 = [\mathbf{U}_1^1, ..., \mathbf{U}_R^1]$, similarly with $\mathbf{U}^2$ and $\mathbf{U}^3$, and now $\mathbf{S}^{bd} \in \mathbb{R}^{RR^1 \times RR^2 \times RR^3}$. So $\mathbf{z}$ can now be expressed with respect to $\mathbf{x}$ and $\mathbf{y}$. Let $\hat{\mathbf{x}} = \mathbf{U}^1\mathbf{x} \in \mathbb{R}^{RR^1}$ and

$\hat{\mathbf{y}} = \mathbf{U}^2\mathbf{y} \in \mathbb{R}^{RR^2}$. These 2 projections are merged by the block-superdiagonal tensor $\mathbf{S}^{bd}$. Each block in $\mathbf{S}^{bd}$ merges together blocks of size R^1 from $\hat{\mathbf{x}}$ and of size R^2 from $\hat{\mathbf{y}}$ to produce a vector of size R^3:

$$\mathbf{z}_r = \mathbf{S}_r \otimes_x \hat{\mathbf{x}}_{rR^1:(r+1)R^1} \otimes_y \hat{\mathbf{y}}_{rR^2:(r+1)R^2}$$

where $\hat{\mathbf{x}}_{i:j}$ is the vector of dimension $j - i$ containing the corresponding values of $\hat{\mathbf{x}}$. Finally all vectors $\mathbf{z}_r$ are concatenated producing $\hat{\mathbf{z}} \in \mathbb{R}^{RR^3}$. The final prediction vector is $\mathbf{z} = \mathbf{U}^3, \hat{\mathbf{z}} \in \mathbb{R}^o$. Similar to tucker fusion, the block term decomposition based fusion in BLOCK theoretically allows more nuanced control on representation size and empirically outperforms previous techniques.

4.3 Related Works

4.3.1 Bilinear Pooling in Video-QA With Language-Vision Fusion

I aim to highlight and explore a broad shift away from BLP in favour of methods such as attention in video-QA benchmarks. Several video models have incorporated and contrasted BLP techniques to their own model designs for language-vision fusion tasks. Kim et al. [107] find various BLP fusions perform worse than their 'dynamic modality fusion' mechanism on TVQA [116] and MovieQA [184]. Li et al. [122] find MCB fusion performs worse on their model in ablation studies on TGIF-QA [93]. Chou et al. [31] use MLB as part of their baseline model proposed alongside their 'VQA 360°' dataset. Gao et al. [61] contrast their proposed 2-stream attention mechanism to an MCB model for TGIF-QA, demonstrating a substantial performance increase over the MCB model. Liu et al. [129] use MUTAN fusion between question and visual features to yield impressive results on TGIF-QA, though they are outperformed by an attention based model using element-wise multiplication [114]. The Focal Visual-Text Attention network (FVTA) [125] is a hierarchical model that aims to dynamically select from the appropriate point across both time and modalities that outperforms an MCB approach on Movie-QA.

4.3.2 Bilinear Pooling in Video Without Language-Vision Fusion

Where recent research in video-QA tasks (which includes textual questions as input) has moved away from BLP techniques, several video tasks that do *not* involve language have found success using BLP techniques. Zhou et al. [237] use a multilevel factorised BLP based model to fuse audio and visual features for emotion recognition in videos. Hu et al. [90] use compact BLP to fuse audio and 'visual long range' features for human action recognition. Pang et al. [155] use MLB as part of an attention-based fusion for audio and visual features for violence detection in videos. Xu et al. [212] use BLP to fuse visual features from different channels in RGBT tracking. Deng et al. [43] use compact BLP to fuse spatial and temporal representations of video features for action recognition. Wang et al. [198] fuse motion and appearance visual information together achieving state-of-the-art results on MSVD-QA. Sudhakaran et al. [181] draw design inspiration from bilinear processing of [127] and MCB to propose 'Class Activation Pooling' for video action recognition. Deb et al. [42] use MLB to process video features for video captioning.

4.4 Datasets

In this section, I describe in greater detail the video-QA datasets I use in my experiments.

4.4.1 MSVD-QA

Xu et al. [210] argue that simply extending VQA methods is "insufficient and sub-optimal" to conduce quality video-QA, and that instead the focus should be on the temporal structure of videos. Using an NLP method to automatically generate QA pairs from descriptions [84], Xu et al. [210] create the MSVD-QA dataset based on the Microsoft research video description corpus [26]. The dataset is made from 1970 video clips, with over 50k QA pairs in '5w' style *i.e.* ("what", "who", "how", "when", "where").

4.4.2 TGIF-QA

Jang et al. [93] speculate that the relatively limited progress in video-QA compared to image-QA is "due in part to the lack of large-scale datasets with well defined tasks". As such, they introduced the TGIF-QA dataset to 'complement rather than compete' with existing VQA literature and to serve as a bridge between video-QA and video understanding. To this end, they propose 3 subsets with specific video-QA tasks that aim to take advantage of the temporal format of videos:

- **Count:** Counting the number of times a specific action is repeated [120] *e.g.* "How many times does the girl jump?". Models output the predicted number of times the specified actions happened. (Over 30k QA pairs).
- **Action:** Identify the action that is repeated a number of times in the video clip. There are over 22k multiple choice questions *e.g.* "What does the girl do 5 times?".
- **Trans:** Identifying details about a state transition [92]. There are over 58k multiple choice questions *e.g.* "What does the man do after the goal post?".

TGIF-QA includes a 4^{th} subset that is of a more general multimodal question-answering design.

- **Frame-QA:** A VQA split using automatically generated QA pairs from frames and captions in the TGIF dataset [123] (over 53k multiple choice questions).

4.4.3 TVQA

As I have already introduce the TVQA dataset [116] across Chapters 2 and 3, I will not repeat the description here.

4.4.4 EgoVQA

Most video-QA datasets focus on video-clips from the 3^{rd} person. Fan [53] argue that 1^{st} person video-QA has more natural use cases that real-world agents would need. As such, they propose the egocentric video-QA dataset (EgoVQA) with 609

QA pairs on 16 first-person video clips. Though the dataset is relatively small, it has a diverse set of question types (*e.g.* 1^{st} & 3^{rd} person 'action' and 'who' questions, 'count', 'colour' etc..), and aims to generate hard and confusing incorrect answers by sampling from correct answers of the same question type. Models on EgoVQA have been shown to overfit due to its small size. To remedy this, Fan [53] pretrain the baseline models on the larger YouTube2Text-QA [220]. YouTube2Text-QA is a multiple choice dataset created from MSVD videos [26] and questions created from YouTube2Text video description corpus [73]. YouTube2Text-QA has over 99k questions in 'what', 'who' and 'other' style.

4.5 Models

In this section, I describe the models used in my experiments, built from the official TVQA [1] and HME-VideoQA [2] implementations.

4.5.1 TVQA Model

I use the same TVQA model as described in Section 3.3 in Chapter 3. Figure 4.3 shows the TVQA model including the redesigns to the multimodal fusion components relevant to the experiments in this chapter.

4.5.2 HME-VideoQA

To better handle semantic meaning through long sequential video data, recent models have integrated external 'memory' units [209, 182] alongside recurrent networks to handle input features [60, 232]. These external memory units are designed to encourage multiple iterations of inference between questions and video features, helping the model revise its visual understanding as new details from the question are presented. The heterogeneous memory-enhanced video-QA model (HME) [52] proposes several improvements to previous memory based architectures:

[1] https://github.com/jayleicn/TVQA
[2] https://github.com/fanchenyou/HME-VideoQA

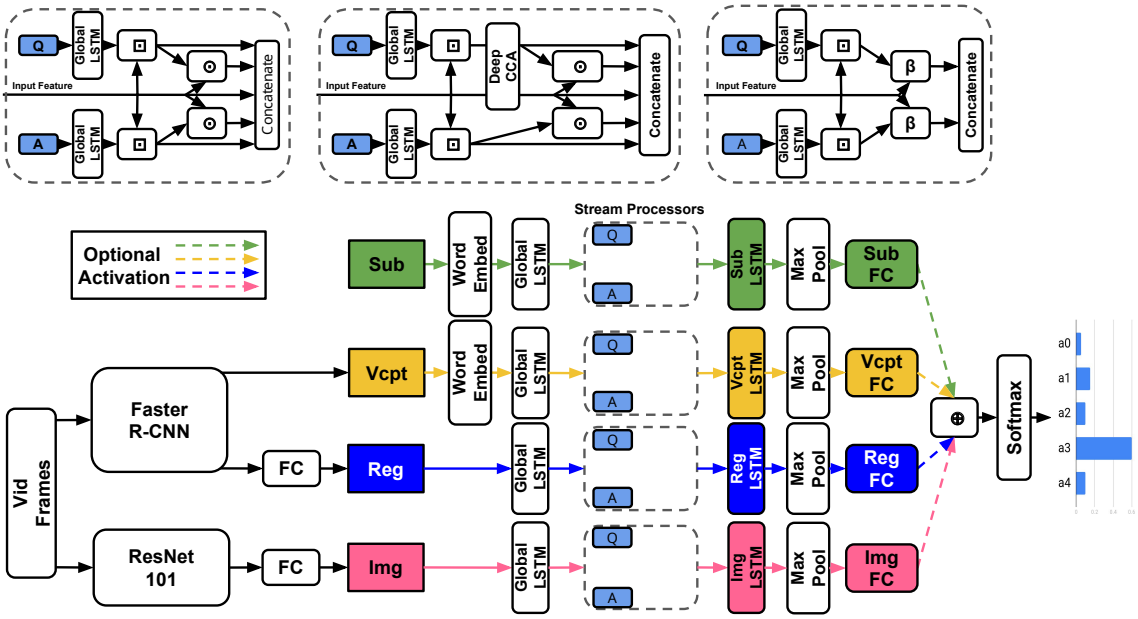


Figure 4.3: TVQA Model. $\odot/\oplus$ = Element-wise multiplication/addition, $\boxdot$ = context matching [172, 226], β = BLP. Any feature streams may be enabled/disabled.

4.5.2.1 Heterogeneous Read/Write Memory

The memory units in HME use an attention-guided read/write mechanism to read from/update memory units respectively (the number of memory slots used is a hyperparameter). The claim is that since motion and appearance features are heterogeneous, a 'straightforward' combination of them cannot effectively describe visual information. The video memory aims to effectively fuses motion (C3D [190]) and appearance (ResNet [83] and VGG [178]) features by integrating them in the joint read/write operations (visual memory in Figure 4.4).

4.5.2.2 Encoder-Aware Question Memory

Previous memory models used a single feature vector outputted by an LSTM or GRU for their question representation [60, 232, 209, 7]. HME uses an LSTM question encoder and question memory unit pair that augment each other dynamically (question memory in Figure 4.4).

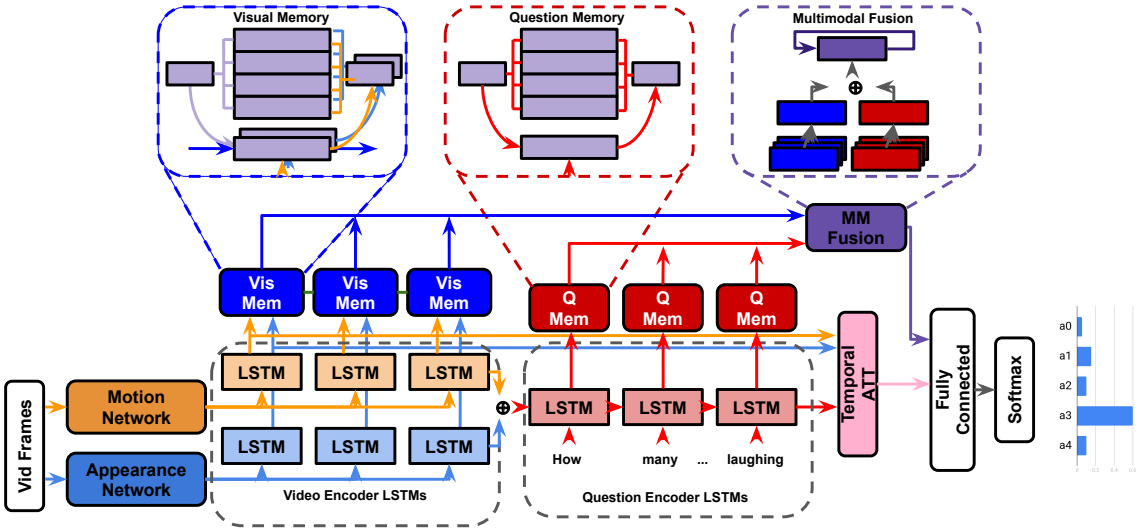


Figure 4.4: HME Model.

4.5.2.3 Multimodal Fusion Unit

The hidden states of the video and question memory units are processed by a temporal attention mechanism. The joint representation 'read' updates the fusion unit's own hidden state. The visual and question representations are ultimately fused by vector concatenation (multimodal fusion in Figure 4.5). My experiments will involve replacing this concatenation step with BLP techniques.

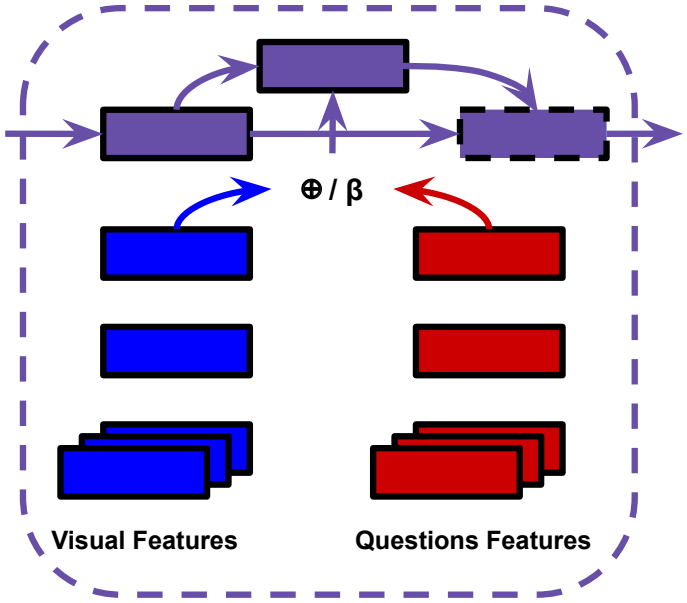


Figure 4.5: $\oplus$ = Concatenation, β = BLP.

4.6 Experiments and Results

In this section I outline my experimental setup and results. I save my insights for the discussion in the next section. See my GitHub repository[3] for both the datasets and code used in my experiments. Table 4.1 shows the benchmarks and SotA results for the datasets I consider in this paper.

Dataset	Benchmark	SoTA
TVQA (Val)	68.85% [116]	74.97% [104]
TVQA (Test)	68.48% [116]	72.89% [104]
EgoVQA (Val 1)	37.57% [53]	45.05%* [29]
EgoVQA (Test 2)	31.02% [53]	43.35%* [29]
MSVD-QA	32.00% [210]	40.30% [78]
TGIF-Action	60.77% [93]	84.70% [114]
TGIF-Count	4.28† [93]	2.19† [114]
TGIF-Trans	67.06% [93]	87.40% [171]
TGIF-FrameQA	49.27% [93]	64.80% [114]

Table 4.1: Dataset benchmark and SoTA results to the best of my knowledge at time of publication. † = Mean L2 loss. * = Results I replicated using the cited implementation.

4.6.1 Concatenation to BLP (TVQA)

As previously discussed, BLP techniques have outperformed feature concatenation on a number of VQA benchmarks. The baseline stream processor concatenates the visual feature vector with question and answer representations. Each of the 5 inputs to the final concatenation are 300-d. I replace the visual-question/answer concatenation with BLP (Figure 4.6). All inputs to the BLP layer are 300-d, the outputs are 750-d and the hidden size is 1600 (a smaller hidden state than normal, however, the input features are also smaller compared to other uses of BLP). I make as few changes as possible to accommodate BLP, *i.e.* I use context matching to facilitate BLP fusion by aligning visual and textual features temporally. My experiments include models with/without subtitles or questions (Table 4.2).

[3] https://github.com/Jumperkables/trying_blp

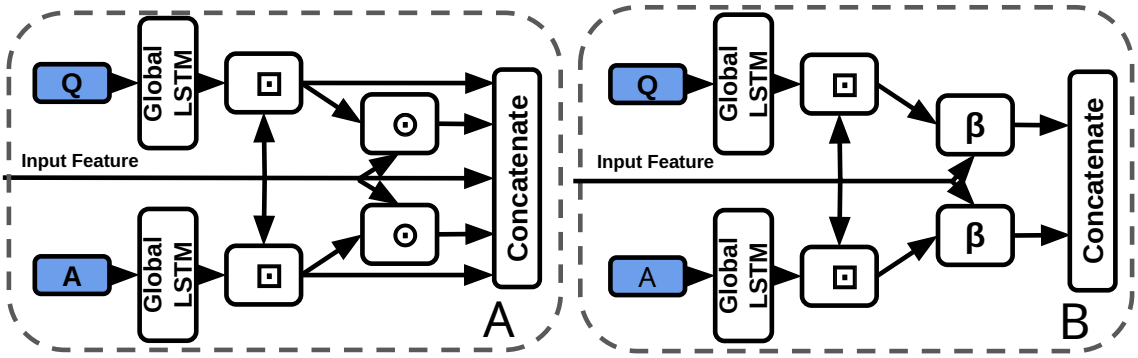


Figure 4.6: Baseline concatenation stream processor from TVQA model (left-A) vs my BLP stream processor (right-B). $\odot$ = Element-wise multiplication, β = BLP, $\boxdot$ = Context Matching.

4.6.2 Dual-Stream Model

I create my 'dual-stream' (Figure 4.7, Table 4.3) model from the SI TVQA baseline model for 2 main purposes:

1. To explore the effects of a joint representation on TVQA.

2. To contrast the concatenation-replacement experiment with a model restructured specifically with BLP as a focus. The baseline BLP model keeps subtitles and other visual features completely separate up to the answer voting step.

My aim here is to create a joint representation BLP-based model similar in essence to the baseline TVQA model that fuses subtitle and visual features. As before, I use context matching to temporally align the video and text features.

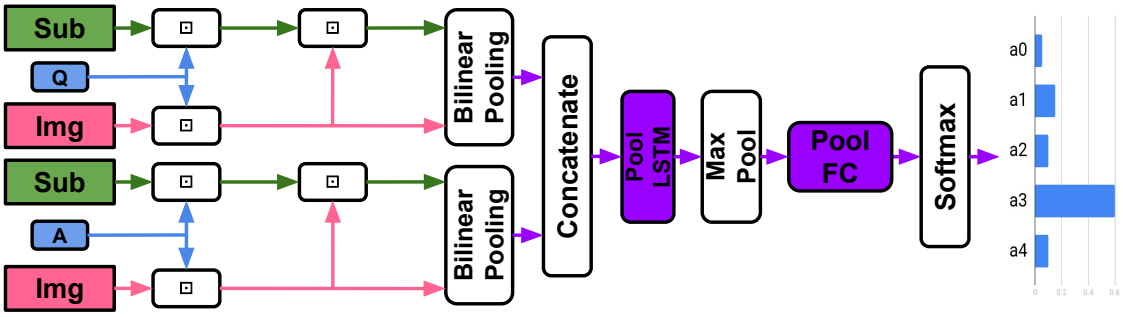


Figure 4.7: My Dual-Stream Model. $\boxdot$ = Context Matching.

Subtitles	Fusion Type	Accuracy	Baseline Offset
-	Concatenation	45.94%	-
GloVE	Concatenation	69.74%	-
BERT	Concatenation	72.20%	-
- (No Q)	Concatentation	45.58%	-0.36%
GloVE (No Q)	Concatentation	68.31%	-1.42%
BERT (No Q)	Concatentation	70.43%	-1.77%
-	MCB	**45.65%**	**-0.29%**
GloVE	MCB	**69.32%**	**-0.42%**
BERT	MCB	**71.68%**	**-0.52%**
-	MLB	41.98%	-3.96%
GloVE	MLB	69.30%	-0.44%
BERT	MLB	69.04%	-3.16%
-	MFB	41.82%	-4.12%
GloVE	MFB	68.87%	-0.87%
BERT	MFB	67.29%	-4.91%
-	MFH	44.44%	-1.5%
GloVE	MFH	68.43%	-1.31%
BERT	MFH	67.29%	-4.91%
-	Blocktucker	44.44%	-1.5%
GloVE	Blocktucker	67.95%	-1.79%
BERT	Blocktucker	67.04%	-5.16%
-	BLOCK	41.09%	-4.85%
GloVE	BLOCK	65.31%	-4.43%
BERT	BLOCK	66.94%	-5.26%

Table 4.2: Concatenation replaced with BLP in the TVQA model on the TVQA Dataset. All models use visual concepts and ImageNet features. 'No Q' indicates questions are not used as inputs *i.e.* answers rely purely on input features.

4.6.3 Deep CCA in TVQA

In contrast to joint representations, Baltrusaitis et al. [10] define 'co-ordinated representations' as a category of multimodal fusion techniques that learn "separated but co-ordinated" representations for each modality (under some constraints). Peng et al. [157] claim that since there is often an information imbalance between modalities, learning separate modality representations can be beneficial for preserving 'exclusive and useful modality-specific characteristics'. For example, given a question about the orientation of an object, it may happen that there are no hints in the text, and the visual inputs must exclusively be relied for that *kind* of information. I include one such representation, deep canonical correlation analysis (DCCA) [8], in my experiments to contrast with the joint BLP models.

4.6.3.1 CCA

Canonical cross correlation analysis (CCA) [89] is a method for measuring the correlations between 2 sets. Let $(\mathbf{X_0}, \mathbf{X_1}) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$ be random vectors with covariances $(\sum_{r=00}, \sum_{r=11})$ and cross-covariance $\sum_{r=01}$. CCA finds pairs of linear projections of the 2 views $(w_0'\mathbf{X_0}, w_1'\mathbf{X_1})$ that are maximally correlated:

$$\begin{aligned}\rho = (w_0^*, w_1^*) &= \underset{w_0, w_1}{argmax}\ corr(w_0'\mathbf{X_0}, w_1'\mathbf{X_1}) \\ &= \underset{w_0, w_1}{argmax} \frac{w_0' \sum_{01} w_1}{\sqrt{w_0' \sum_{00} w_0 w_1' \sum_{11} w_1}}\end{aligned}$$

where ρ is the correlation coefficient. As ρ is invariant to the scaling of w_0 and w_1, the projections are constrained to have unit variances, and can be represented as the following maximisation:

$$\underset{w_0, w_1}{argmax}\ w_0' \sum_{01} w_1\ s.t\ w_0' \sum_{00} w_0 = w_1' \sum_{11} w_1 = \mathbf{1}$$

However, CCA can only model linear relationships regardless of the underlying realities in the dataset. Thus, CCA extensions were proposed, including kernel CCA (KCCA) [5] and later DCCA.

4.6.3.2 DCCA

DCCA is a parametric method used in multimodal neural networks that can learn non-linear transformations for input modalities. Both modalities t, v are encoded in neural network transformations $H_t, H_v = f_t(t, \theta_t),\ f_v(v, \theta_v)$, and then the canonical correlation between both modalities is maximised in a common subspace (*i.e.* maximise cross-modal correlation between H_t, H_v).

$$\max corr(H_t, H_v) = \underset{\theta_t, \theta_v}{argmax}\ corr(f_t(t, \theta_t), f_v(v, \theta_v))$$

I use DCCA over KCCA to co-ordinate modalities in my experiments as it is generally more stable and efficient, learning more 'general' functions.

4.6.3.3 DCCA in TVQA

I use a 2-layer DCCA module to coordinate question and context (visual or subtitle) features (Figure 4.8, Table 4.4). Output features are the same dimensions as inputs.

Though DCCA itself is not directly related to BLP, it has recently been classified as a coordinated representation [75], which contrasts a 'joint' representation.

Model	Text	Val Acc
TVQA SI	GloVe	67.78%
TVQA SI	BERT	70.56%
Dual-Stream MCB	GloVe	63.46%
Dual-Stream MCB	BERT	60.63%
Dual-Stream MFH	GloVe	62.71%
Dual-Stream MFH	BERT	59.34%

Table 4.3: Dual-Stream Results Table. 'SI' for TVQA models indicates the model is using subtitle and ImageNet feature streams only, *i.e.* the green and pink streams in Figure 4.3

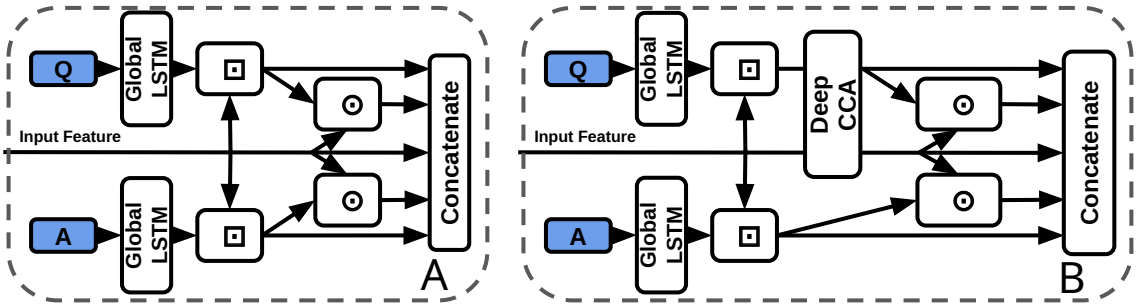


Figure 4.8: Baseline concatenation stream processor from TVQA model (left-A) vs my DCCA stream processor (right-B). $\odot$ = Element-wise multiplication, $\boxdot$ = Context Matching.

Model	Text	Baseline Acc	DCCA Acc
VI	GloVe	45.94%	45.00% (-0.94%)
VI	BERT	–	41.70%
SVI	GloVe	69.74%	67.91% (-1.83%)
SVI	BERT	72.20%	68.48% (-3.72%)

Table 4.4: DCCA in the TVQA Baseline Model.

4.6.4 Concatenation to BLP (HME-VideoQA)

As described in the previous section, I replace a concatenation step in the HME model between textual and visual features with BLP (Figure 4.5, corresponding to the multimodal fusion unit in Figure 4.4). The goal here is to explore if BLP can better facilitate multimodal fusion in aggregated memory features (Table 4.5). I

replicate the results from [52] with the HME on the MSVD, TGIF and EgoVQA datasets using the official github repository [29]. I extract my own C3D features from the frames in the TVQA.

Dataset	Fusion Type	Val	Test
TVQA (GloVE)	Concatenation	41.25%	N/A
EgoVQA-0	Concatenation	36.99%	37.12%
EgoVQA-1	Concatenation	48.50%	43.35%
EgoVQA-2	Concatenation	45.05%	39.04%
MSVD-QA	Concatenation	30.94%	33.42%
TGIF-Action	Concatenation	70.69%	73.87%
TGIF-Count	Concatenation	3.95†	3.92†
TGIF-Trans	Concatenation	76.33%	78.94%
TGIF-FrameQA	Concatenation	52.48%	51.41%
TVQA (GloVE)	MCB	41.09% (-0.16%)	N/A%
EgoVQA-0	MCB	No Convergence	No Convergence
EgoVQA-1	MCB	No Convergence	No Convergence
EgoVQA-2	MCB	No Convergence	No Convergence
MSVD-QA	MCB	30.85% (-0.09%)	33.78% (+0.36%)
TGIF-Action	MCB	73.56% (+2.87%)	73.00% (-0.87%)
TGIF-Count	MCB	3.95† (+0†)	3.98† (+0.06†)
TGIF-Trans	MCB	79.30% (+2.97%)	77.10% (-1.84%)
TGIF-FrameQA	MCB	51.72% (-0.76%)	52.21% (+0.80%)

Table 4.5: HME-VideoQA Model. The default fusion technique is concatenation. † refers to minimised L2 loss.

4.7 Discussion

This section discusses both my experimental results and their significance with respect to the trends in the literature that I have identified. Though the empirical struggles and proposed alternatives to BLP in video-QA is one of the major intended insights of my work in this chapter, each of the following subsections discusses —in detail— the finer points and findings that each individual experiment raises.

4.7.1 TVQA Experiments

4.7.1.1 No BLP Improvements on TVQA

On the HME concat-to-BLP substitution model (Table 4.5), MCB barely changes model performance at all. I find that none of my TVQA concat-to-BLP substitutions

(Table 4.2) yield any improvements at all, with almost all of them performing worse overall (0.3-5%) than even the questionless concatenation model. Curiously, MCB scores the highest of all BLP techniques. The dual-stream model performs worse still, dropping accuracy by between 5-10% vs the baseline (Table 4.3). Similarly, I find that MCB performs best despite being known to require larger latent spaces to work on VQA.

4.7.1.2 BERT Impacted the Most

For the TVQA BLP-substitution models, I find that the GloVe, BERT and, 'no-subtitle' variations all degrade by roughly similar margins, with BERT models degrading more most often. This slight discrepancy is unsurprising as the most stable BERT baseline model is the best, and thus may degrade more on the inferior BLP variations. However, BERT's relative degradation is much more pronounced on the dual-stream models, performing 3% worse than GloVe. I theorise that here, the significant and consistent drop is potentially caused by BERT's more contextual nature is no longer helping, but actively obscuring more pronounced semantic meaning learned from subtitles and questions.

4.7.1.3 Blame Smaller Latent Spaces?

Naturally, bilinear representations of time series data across multiple frames or subtitles are highly VRAM intensive. Thus I can only explore relatively small hidden dimensions (*i.e.* 1600). However, I cannot simply conclude my poor results are due to my relatively small latent spaces because:

1. MCB is my best performing BLP technique. However, MCB has been outperformed by MFH on previous VQA models *and* it has been shown to require much larger latent spaces to work effectively in the first place [58] (16000).

2. My vector representations of text and images are also much smaller (300-d) compared to the larger representation dimensions conventional in previous benchmarks (*e.g.* 2048 in [58]). I note that $16000/2048 \approx 1600/300$, and so my latent-to-input size ratio is not substantially different to previous works.

4.7.1.4 Unimodal Biases in TVQA and Joint Representation

Another explanation may come from works exploring textual biases inherent in TVQA to textual modalities [204]. BLP has been categorised as a 'joint representation'. Baltrusaitis et al. [10] consider representation as summarising multimodal data "in a way that exploits the complementarity and redundancy of multiple modalities". Joint representations combine unimodal signals into the same representation space. However, they struggle to handle missing data [10] as they tend to preserve shared semantics while ignoring modality-specific information [75]. The existence of unimodal text bias in TVQA implies BLP may perform poorly on TVQA as a joint representation of its features because:

1. Information from either modality is consistently missing.

2. Prioritising 'shared semantics' over 'modality-specific' information harms performance on TVQA.

Though concatenation could also be classified as a joint representation, I argue that this observation still has merit. Theoretically, a concatenation layer can still model modality specific information (see Figure 4.9), but a bilinear representation would seem to inherently entangle its inputs which would make modality specific information more challenging to learn since each parameter representing one modality is by definition weighted with the other. This may explain why my simpler BLP substitutions perform better than my more drastic 'joint' dual-stream model.

4.7.1.5 What About DCCA?

Table 4.4 shows my results on the DCCA augmented TVQA models. I see a slight but noticable performance degradation with this relatively minor alteration to the stream processor. As previously mentioned, DCCA is in some respects an opposite approach to multimodal fusion than BLP, *i.e.* a 'coordinated representation'. The idea of coordinated representations is to learn a separate representation for each modality , but with respect to the other. In this way, it is thought that multimodal interactions can be learned while still preserving modality-specific information that a joint representation may otherwise overlook [75, 157]. DCCA specifically maximises

cross-modal correlation. Without further insight from surrounding literature, it is difficult to conclude what TVQA's drop in performance using both joint *and* coordinated representations could mean. I will revisit this when I discuss the role of attention in multimodal fusion.

4.7.1.6 Does Context Matching Ruin Multimodal Integrity?

The context matching technique used in the TVQA model is the birdirectional attention flow (BiDAF) module introduced in [172]. It is used in machine comprehension between a textual context-query pair to generate query-aware context representations. BiDAF uses a 'memoryless' attention mechanism where information from each time step does not directly affect the next, which is thought to prevent early summarisation. BiDAF considers different input features at different levels of granularity. The TVQA model uses bidirectional attention flow to create context aware (visual/subtitle) question and answer representations. BiDAF can be seen as a coordinated representation in some regards, but it does project questions and answers representations into a new space. I use this technique to prepare my visual and question/answer features because it temporally aligns both features, giving them the same dimensional shape, conveniently allowing us to apply BLP at each time step. Since the representations generated are much more similar than the original raw features and there is some degree of information exchange, it may affect BLP's representational capacity. Though it is worth considering these potential shortcomings, I cannot immediately assume that BiDAF would cause serious issues as earlier bilinear technique were successfully used between representations in the same modality [185, 62]. This implies that multimodal interactions can still be learned between the more similar context-matched representations, provided the information is still present. Since BiDAF does allow visual information to be used in the TVQA baseline model, it is reasonable to assume that some of the visual information is in fact intact and exploitable for BLP. However, it is still currently unclear if context matching is fundamentally disrupting BLP and contributing to the poor results I find. I note that in BiDAF, 'memoryless' attention is implemented to avoid propagating errors through time. I argue that though this may be true and help in

some circumstances, conversely, this will not allow some useful interactions to build up over time steps.

4.7.2 The Other Datasets on HME

4.7.2.1 BLP Has No Effect

My experiments on the EgoVQA, TGIF-QA and MSVD-QA datasets are on concat-to-BLP substitution HME models. These results are inconclusive. There is virtually no variation in performance between the BLP and concatenation implementations. Interestingly, EgoVQA consistently does not converge with this simple substitution. I cannot comment for certain on why this is the case. There seems to be no intuitive reason why its 1^{st} person content would cause this. Rather, I believe this is symptomatic of overfitting in training, as EgoVQA is very small *and* pretrained on a different dataset, and various BLP techniques have been shown to have difficulties converging during training.

4.7.2.2 Does Better Attention Explain the Difference?

Attention mechanisms have been shown to improve the quality of text and visual interactions. Yu et al. [228] argue that methods without attention are 'coarse joint-embedding models' which use global features that contain noisy information unhelpful in answering fine-grained questions commonly seen in VQA and video-QA. This provides strong motivation for implementing attention mechanisms alongside BLP, so that the theoretically greater representational capacity of BLP is not squandered on less useful noisy information. The TVQA model uses the previously discussed BiDAF mechanism to focus information from both modalities. However, the HME model integrates a more complex memory-based multi-hop attention mechanism. This difference may potentially highlight why the TVQA model suffers more substantially integrating BLP than the HME one.

4.7.3 BLP in Video-QA: Problems and Recommendations

I have experimented with BLP in 2 video-QA models and across 4 datasets. My experiments show that the BLP fusion techniques popularised in VQA has not extended to increased performance to video-QA. In the preceding sections, I have supported this observation with experimental results which I contextualise by surveying the surrounding literature for BLP for multimodal video tasks. In this section, I condense my observations into a list of problems that BLP techniques pose to video-QA, and my proposal for alternatives and solutions:

4.7.3.1 Inefficient and Computationally Expensive Across Time

BLP as a fusion mechanism in video-QA can be exceedingly expensive due to added temporal relations. Though propagating information from each time step through a complex text-vision multimodal fusion layer is an attractive prospect, my experiments imply that modern BLP techniques simply do not empirically perform in such a scenario. I recommend avoiding computationally expensive fusion techniques like BLP for text-image fusion *throughout* timesteps, and instead simply concatenate features at these points to save computational resources for other stages of processing (*e.g.* attention). Furthermore, I note that any prospective fusion technique used across time will quickly encounter memory limitations that could force the hidden-size used sub-optimally low. Though summarising across time steps into condensed representations may allow more expensive BLP layers to be used on the resultant text and video representations, I instead recommend using state-of-the-art and empirically proven multimodal attention mechanisms [118, 215]. Attention mechanisms are pivotal in VQA for reducing noise and focusing on specific fine-grained details [228]. The sheer increase in feature information when moving from still-image to video further increases the importance of attention in video-QA. My experiments show the temporal-attention based HME model performs better when it is not degraded by BLP. My findings are in line with that of Long et al. [134] as they consider multiple different fusion methods for video classification, *i.e.* LSTM, probability, 'feature' and attention. 'Feature' fusion is the direct connection of each modality within each local time interval, which is effectively what context matching

does in the TVQA model. Long et al. [134] finds temporal feature based fusion sub-par, and speculates that the burden of learning multimodal *and* temporal interactions is too heavy. My experiments lend further evidence that for video tasks, attention-based fusion is the ideal choice.

4.7.3.2 Problem with Alignment of Text and Video

As I highlight in Subsection 4.3.2, BLP has yielded great performance in video tasks where it fuses the visual features with *non-textual* features. Audio and visual feature fusion demonstrates impressive performance on action recognition [90], emotion recognition [237], and violence detection [155]. Likewise, different visual representations have thrived in RGBT tracking [212], action recognition [43] and video-QA on MSVD-QA [198]. On the other hand, I notice that several recent video-QA works (Section 4.3.1) have found in ablation that BLP fusion which specifically fuse visual and *textual* features give poor results [107, 122, 61, 129, 125]. My observations and my experimental results highlight a pattern of poor performance for BLP in text-video fusion specifically. I demonstrate poor performance using BLP to fuse both 'BiDAF-aligned' (TVQA) and 'raw' (HME) text and video features *i.e.* temporally aligned and unaligned respectively. As the temporally-aligned modality combinations of video-video and video-audio BLP fusion continue to succeed, I believe that the 'natural alignment' of modalities is a significant contributing factor to this performance discrepancy in video. To the best of my knowledge, I am the first to draw attention to this trend. Attention mechanisms continue to achieve state-of-the-art in video-language tasks and have been demonstrated (with visualisable attention maps) to focus on relevant video and question features. I therefore recommend using attention mechanisms for their strong performance and relatively interpretable behaviour, and avoiding BLP for specifically video-text fusion.

4.7.3.3 Empirically Justified on VQA

Successive BLP techniques have helped drive increased VQA performance in recent years, as such they remain an important and welcome asset to the field of multimodal machine learning. I stress that these improvements, welcome as they are, are

only justified by their empirical improvements in the tasks they are applied to, and lack strong theoretical frameworks which explain their superior performance. This is entirely understandable given the infamous difficulty in interpreting how neural networks *actually* make decisions or exploit their training data. However, it is often claimed that such improvements are the result of some intrinsic property of the BLP operator, *e.g.* creating 'richer multimodal representations': Fukui et al. [58] *hypothesise* that concatenation is not as expressive as an outer product of visual and textual features. Kim et al. [106] claim that "bilinear models provide rich representations compared with linear models". Ben-younes et al. [12] claim MUTAN "focuses on modelling fine and rich interactions between image and text modalities". Yu et al. [229] claim that MFH significantly improves VQA performance "*because* they achieve more effective exploitation of the complex correlations between multimodal features". Ben-Younes et al. [13] carefully demonstrate that the extra control over the dimensions of components in BLOCK fusion can be leveraged to achieve yet higher VQA performance, however this is attributed to its ability "to represent very fine interactions between modalities while maintaining powerful mono-modal representations". In contrast, Yu et al. [228] carefully assess and discuss the *empirical* improvements their MFH fusion offers on VQA. My discussions and findings highlight the importance of being measured and nuanced when discussing the theoretical nature of multimodal fusion techniques and the benefits they bring.

4.8 Theoretically Motivated Observations and Neurologically Guided Proposals

BLP techniques effectively exploit mathematical innovations on bilinear expansions represented in neural networks. As previously discussed, it remains unclear *why* any bilinear representation would be intrinsically superior for multimodal fusion to alternatives *e.g.* a series of non-linear fully connected layers or attention mechanisms. In this section, I share my thoughts on the properties of bilinear functions, and how they relate to neurological theories for multimodal processing in the human brain. I provide qualitative analysis of the distribution of neurolinguistic norms present

in the video-QA datasets used in my experiments with which, through the lens of 'Dual Coding Theory' and the 'Two-Stream' model of vision, I propose neurologically motivated multimodal processing methodologies.

4.8.1 Observations: Bilinearity in BLP

4.8.1.1 Nonlinearities in Bilinear Expansions

As previously mentioned in my description of MLB, Kim et al. [106] suggest using *Tanh* activation on the output of vector $\mathbf{z}$ to further increase model capacity. Strictly speaking, I note that adding the the non-linearity means the representation is **no longer bilinear** as it is not linear with respect to either of its input domains. It is instead the 'same kind of non-linear' in both the input domains. I suggest that an alternative term such as 'bi-nonlinear' would more accurately described such functions. Bilinear representations are not the most complex functions with which to learn interactions between modalities. As explored by Yu et al. [229], I believe that higher-order interactions between features would facilitate a more realistic model of the world. The non-linear extension of bilinear or higher-order functions is a key factor to increase representational capacity.

4.8.1.2 Outer Product Forces Multimodal Interactions

The motivation for using bilinear methods over concatenation in VQA and video-QA was that it would enable learning more 'complex' or 'expressive' interactions between the textual and visual inputs. I note however that concatenation of inputs features should theoretically allow both a weighted multimodal combination of textual and visual units, *and* allow unimodal units of input features. As visualised in Figure 4.9, weights representing a bilinear expansion in a neural network each represent a multiplication of input units from each modalitiy. This appears to, in some sense, *force* multimodal interactions where it could possibly be advantageous to allow some degree of separation between the text and vision modalities. As discussed earlier, it is thought that 'joint' representations [10] preserve shared semantics while ignoring modality-specific information [75]. Though it is unclear if concatenation could effec-

tively replicate bilinear processing while also preserving unimodal processing, it also remains unclear how *exactly* bilinear representations learn. For now, the successes and struggles of bilinear representations across VQA and video-QA remain justified by empirical performance on datasets.

4.8.2 Proposals: Neurological Parallels

I have recommended that video-QA models prioritise attention mechanisms over BLP given my own experimental results and my observations of the current state-of-the-art trends. I *can* however still explore how bilinear models in deep learning are related to 2 key areas of relevant neurological research, *i.e.* the Two-Stream model of vision [69, 142] and Dual Coding Theory [152, 153].

4.8.2.1 Two-Stream Vision

Introduced in Goodale and Milner [69], the current consensus on primate visual processing is that it is divided into 2 networks or streams: the 'ventral' stream which mediates transforming the contents of visual information into 'mental furniture' that guides memory, conscious perception, and recognition; and the 'dorsal' stream which mediates the visual guidance of action. There is a wealth of evidence showing that these 2 subsystems are not mutually insulated from each other, but rather interconnect and contribute to one another at different stages of processing

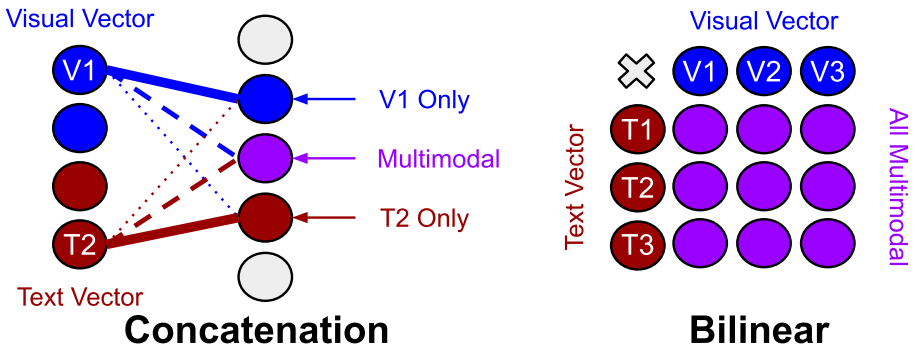


Figure 4.9: Visualisation of the differences between concatenation and bilinear representations for unimodal processing. Concatenation (left) can theoretically allow unimodal features from text or vision to process *independently* of the other modality by reducing its weighted contribution (see 'V1 Only'). Bilinear representations (right) *force* multimodal interactions.

[142, 95]. In particular, Jeannerod and Jacob [95] argue that valid comparisons between visual representation must consider the direction of fit, direction of causation, and the level of conceptual content. They demonstrate that visual subsystems and behaviours inherently rely on aspects of both streams. Recently, Milner [142] consider 3 potential ways these cross-stream interactions could occur:

1. Computations along the 2 pathways are independent and combine at a 'shared terminal' (the independent processing account).

2. Processing along the separate pathways is modulated by feedback loops that transfer information from 'downstream' brain regions, including information from the complementary stream (the feedback account).

3. Information is transferred between the 2 streams at multiple stages and locations along their pathways (the continuous cross-talk account).

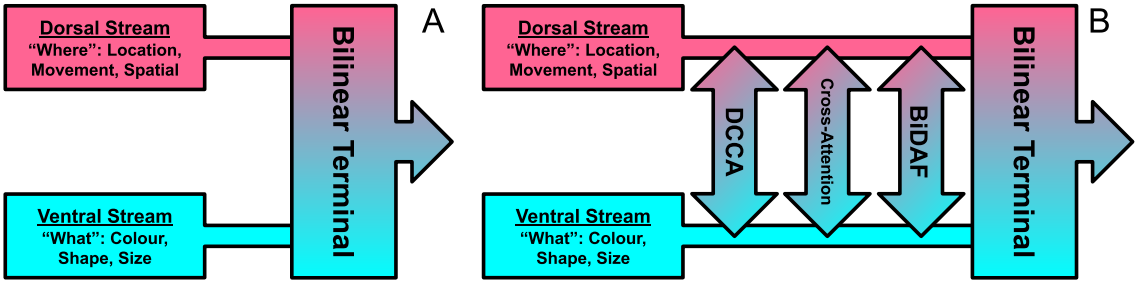


Figure 4.10: Visualisation of the 1^{st} and 3^{rd} cross-stream scenarios for the 2-stream model of vision described by Milner [142]. The early bilinear model proposed by Tenenbaum and Freeman [185] strikingly resembles the 1^{st} (left-A). The 3^{rd} and more recently favoured scenario features a continuous exchange of information across streams at multiple stages, and can be realised by introducing 'cross-talking' of deep learning features (right-B).

Though Milner [142] focus mostly on the 'continuous cross-talk' idea, they believe that a unifying theory would include aspects from each of these scenarios. The vision-only deep bilinear models proposed in [185, 127] are strikingly reminiscent to the 1^{st} 'shared-terminal' scenario (see Figure 4.10). The bilinear framework proposed in [185] focuses on splitting up 'style' and 'content', and is designed to be applied to any 2-factor task. Lin et al. [127] note but do not explore the similarities between their proposed network and the two-stream model of vision. Their bilinear

CNN model aims to processes 2 subnetworks separately, 'what' (ventral) and 'where' (dorsal) streams, and later combine in a bilinear 'terminal'. BLP methods developed from these baselines would later focus on multimodal tasks between language and vision. As Milner [142] focus mainly on their 3^{rd} scenario (right), subsequent bilinear models that draw inspiration from the two-stream model of vision could realise the 'cross-talk' mechanism *i.e.* using co-attention or 'co-ordinated' DCCA.

4.8.2.2 Dual Coding Theory

Dual coding theory (DCT) [152] broadly considers the interactions between the verbal and non-verbal systems in the brain (recently surveyed in [153]). DCT considers verbal and non-verbal interactions by way of 'logogens' and 'imagens' respectively, *i.e.* units of verbal and non-verbal recognition. Imagens may be multimodal, *i.e.* haptic, visual, smell, taste, motory etc. We should appreciate the distinction between medium and modality: image is both medium and modality and videos are an image based modality. Similarly, text is the medium through which the natural language modality is expressed. I can see parallels in multimodal deep learning and dual coding theory, with textual features as logogens and visual (or audio) features as visual (or auditory) imagens. There are many insights from DCT that could guide and drive multimodal deep learning:

1. Logogens and imagens are discrete units of recognition and are often related to tangible concepts (*e.g.* 'pictogens' [145]). By drawing inspiration from pictogen/imagen style of information representation, it could be hypothesised that multimodal models should additionally focus on deriving more tangible features (*i.e.* discrete convolution maps previously used in vision-only bilinear models [127]) as opposed to more abstracted 'ImageNet-style' feature vectors more commonly used in recent BLP models (see Figure 4.11) are a more ideal way to represent features.

2. Bezemer and Kress [15] explore the differences in student understanding when text information is presented alongside other modalities. They argue that when meaning is moved from one medium to another semiotic relations are

redefined. This paradigm could be emulated to control how networks learn concepts in relation to certain modal information.

3. Imagens (and potentially logogens) may be a function of many modalities, *i.e.* one may recognise something as a function of haptic and auditory experiences alongside visual ones. I believe this implies that non-verbal modalities (vision/sound etc..) should be in some way grouped or aggregated, and that while DCT remains widely accepted, multimodal research should consider 'verbal vs non-verbal' interactions as a whole instead of focusing too intently on 'case-by-case' interactions, *i.e.* text-vs-image and text-vs-audio. This text/non-text insight may be related to the apparent difference in text-vision video task performance previously discussed.

4. Multimodal cognitive behaviours in people can be improved by providing cues.

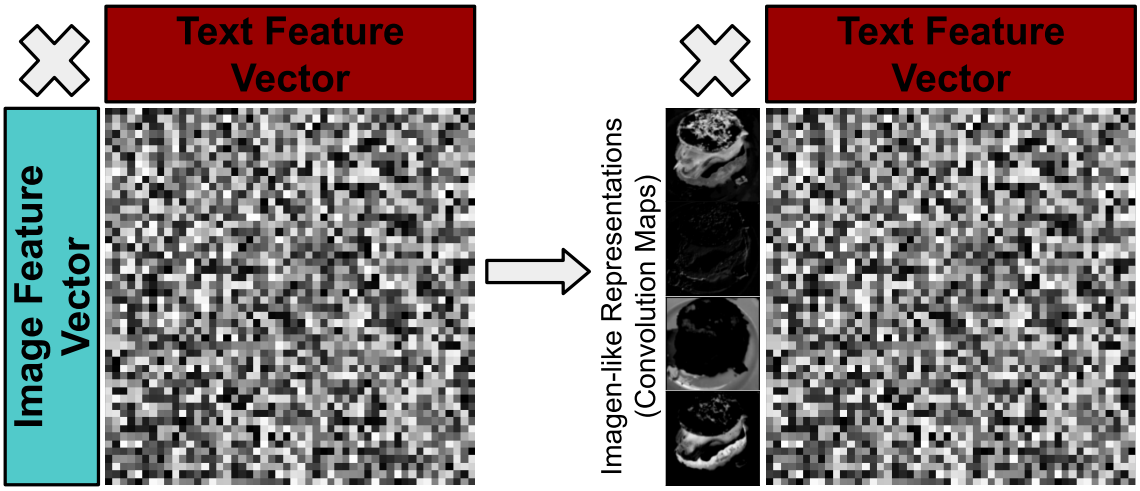


Figure 4.11: Visualisation of moving from less tangible visual features to more 'imagen-like' visual features *e.g.* convolution maps of an image.

For example, referential processing (naming an object or identifying an object from a word) has been found to additively affect free recall (recite a list of items), with the memory contribution of non-verbal codes (pictures) being twice that of verbal codes [154]. Begg [11] find that free recall from 'concrete phrases' (can be visualised) of their constituent words is roughly twice that of 'abstract' phrases. However, this difference increased 6-fold for concrete phrases when cued with one of the phrase words, yet using cues for abstract phrases did not help at all. This was named

the 'conceptual peg' effect in DCT, and is interpreted as memory images being re-activated by 'a high imagery retrieval cue'. Given such apparent differences in human cognitive processing for 'concrete' and 'abstract' words, it may similarly be beneficial for multimodal text-vision tasks to explicitly exploit the neurolinguistic 'concreteness' word norm. Leveraging existing neurolinguistic word-norm datasets, I identify the relative abundance of concrete words in textual components of the video-QA datasets I experiment with (see Figure 4.12). As the various word-norm datasets use various scoring systems for concreteness (*e.g.* MTK40 uses a Likert scale 1-7), I rescale the scores for each dataset such that the lowest score is 0 (highly abstract), and the highest score is 1 (highly concrete). Though I cannot find a concreteness score for every word in each dataset component's vocabulary, I see that the 4 video-QA datasets I experiment with have more concrete than abstract words overall. Furthermore, I see that answers are on-average significantly more concrete than they are abstract, and that (as intuitively expected) visual concepts from TVQA are even more concrete. Taking inspiration from human processing through DCT, it could be hypothesised that multimodal machine learning tasks could benefit from explicitly learning relations between 'concrete' words and their constituents, whilst treating 'abstract' words and concepts differently. Recently proposed computational models of DCT have had many drawbacks [153], I believe that neural networks can be a natural fit for modelling neural correlates explored in DCT and should be considered as a future modelling option.

I personally find the exploitation of abstract and concrete concepts to be the most promising avenue for neurological inspiration in machine learning thanks to the abundance of concreteness norms in multimodal QA datasets. My work in Chapter 6 uses this part of dual coding theory to propose an improved neurolinguistic multiclass labelling scheme for use in VQA.

4.9 Conclusion

In light of BLP's empirical success in VQA, I have experimentally explored their use in video-QA on 2 models and 4 datasets. I find that switching from vector concate-

nation to BLP through simple substitution on the HME and TVQA models does not improve and in fact actively harms performance on video-QA. I find that a more substantial 'dual-stream' restructuring of the TVQA model to accommodate BLP significantly reduces performance on TVQA. My results and observations about the downturn in successful text-vision BLP fusion in video tasks imply that naively using BLP techniques can be very detrimental in video-QA. I caution against automatically integrating bilinear pooling in video-QA models and expecting similar empirical increases as in VQA. I offer several interpretations and insights of my negative results using surrounding multimodal and neurological literature and find my results inline with trends in VQA and video-classification. I take care to discuss the finer points raised by each of my *individual* experiments. To the best of my knowledge, I am the first to outline how important neurological theories *i.e.* dual coding theory and the two-stream model of vision relate to the history of (and journey to) modern multimodal deep learning practices. I offer a few experimentally and theoretically guided suggestions to consider for multimodal fusion in video-QA, most notably that attention mechanisms should be prioritised over BLP in text-vision fusion. I qualitatively show the potential for neurologically-motivated multimodal approaches in video-QA by identifying the relative abundance of neurolinguistically 'concrete' words in the vocabularies for the text components of the 4 video-QA datasets I experiment with. I would like to emphasise the importance of related neurological theories in deep learning and encourage researchers to explore Dual Coding Theory and the Two-Stream model of vision. My findings here on the abundance of concreteness in multimodal QA vocabularies motivates my neurolinguistic inspirations for my work in Chapter 6.

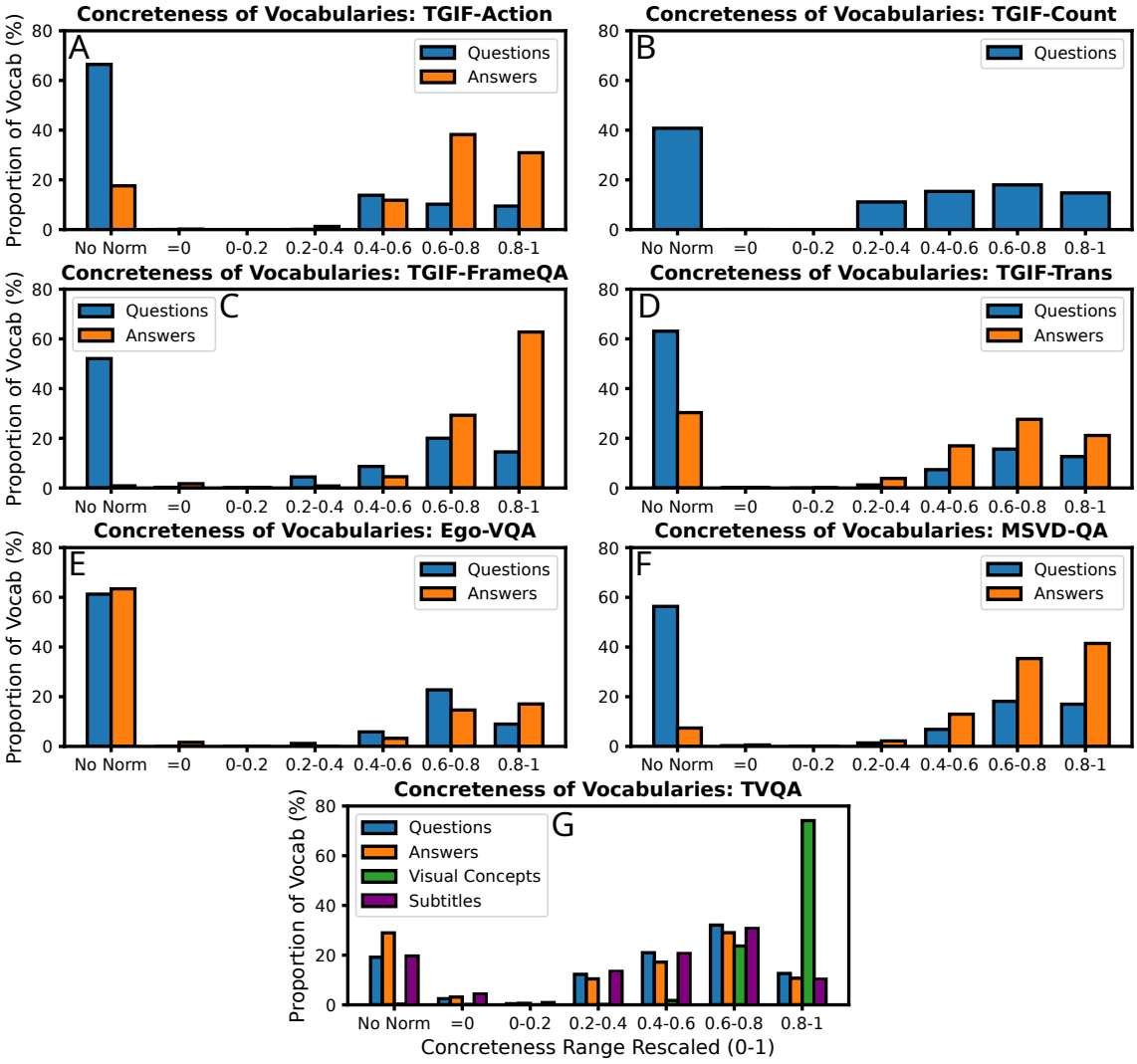


Figure 4.12: The relative abundance of the neurolinguistic 'concreteness' score in the *vocabularies* of each source of text in the video-QA datasets I experiment with. Stopwords are not included. Concreteness scores are taken from the following datasets: MT40k [18], USF [147], SimLex999 [85], Clark-Paivio [33], Toronto Word Pool [57], Chinese Word Norm Corpus [221], MEGAHR-Crossling [132], Glasgow Norms [169], [165], and [177]. The scores for each word are rescaled from 0-1 such that most abstract = 0 and most concrete = 1, and the result averaged if more than 1 dataset has the same word.

CHAPTER 5

Visual Modelling: The Visual Parallel to Language Modelling Evaluated on Dynamic Simulations

This chapter heavily draws from a paper currently under review at the Journal of Machine Learning Research (JMLR). As such, this chapter's contents are derived from a collaboration project. I —Tom Winterbottom, the author of this thesis— am the only first author of this paper and contributed: the underlying code and framework; the experimental design; the literature review; the gathering of external datasets; the design of the 6 proposed datasets; the analysis of experimental results; and the contents of the paper and this chapter. Descriptions in subsections 5.4.1, 5.4.2, and 5.3.2; Figures 5.1, 5.4, and 5.5; and both the dataset design and experimental design/analysis were created and undertaken in service to this project in collaboration with G. Thomas Hudson and Daniel Kluvanec.

5.1 Introduction

As I have repeatedly highlighted throughout this thesis so far, multimodal QA benchmarks fail to use visual data as effectively as its text. Chapter 2 extensively reviews research aiming to address such visual 'underperformance', the vast major-

ity of which seek to 'correct' the text information in some manner. In Chapter 3, I have similarly played my part in highlighting the presence and consequences of such language bias. I aimed to distinguish my work from the abundance of 'text-focused' research by introducing a method to isolate subsets responding to *any* particular modality, of which textual data is currently most problematic. However, such subsets currently remain unsuitably small for enabling multimodal research. Furthermore, my analysis in Chapter 4 demonstrates how textual biases can frustrate attempts to test specific and targeted hypotheses.

I am therefore motivated to shift my focus back to the pressing and overarching problem of imbalance of vision and lanaguage, **but in a manner such that I am not subject-to or hindered-by the language bias**. In this chapter, I look to address the relative under-exploitation of vision vs language by instead *improving the exploitation of visual information* using the same generative pretraining paradigm that has been instrumental in improving the scope of language capabilities in deep learning (language modelling).

Generative video prediction is a popular and active area of research [22, 150, 238] that has recently adopted transformer-based architectures [21, 162, 55, 56] which have led to great progress in language modelling. However, where other work focuses on the state-of-the-art performance a transformer-based model can bring to video generation [214, 56, 162], I instead explore the visual equivalent of the predictive language modelling training approach that transformers are well known for. I dub this generative pretraining 'visual modelling' (*i.e.* image-sequence-to-image). What are the similarities and differences of predictive modelling in vision and language? Under what circumstances could it be easier to predict a frame of a video than the next word of a sentence? Is it always easier to predict a few language tokens than it is to fully generate output pixels? Is language a more information dense modality? Or is a picture worth 1,000 (or 16×16 [50]) words? In parallel with similarly motivated work [164, 28], I seek to push this conversation to the forefront of the field. However, the remarkable successes of predictive language modelling casts a large shadow. Where language models can generate paragraphs of text comparable to human quality [17], instead video generation models are comparatively primitive.

I identify 3 major barriers to closing this research gap:

1. Generating videos instead of language tokens is fundamentally more challenging. Instead of generating confidence scores for a word from a fixed language token vocabulary, video requires precisely predicting values for clusters of pixels, or even entire images.

2. Video datasets have a much higher memory and storage overhead than language datasets of comparative scale. Even the most ambitious and well resourced video models down-sample frames to 64×64 [28, 214].

3. Though the first few frames of video predictions are impressive [28, 214], the quality of longer term predictions is lacking [*i.e.* 10, 20, 50 frames into the future, see Section 7 of 150]. This implies that these models do not have a strong understanding of the physical laws underpinning the video.

Together these barriers highlight the often underestimated complexity of video datasets (*e.g.* the simple action of walking involves simultaneous bends and rotations of various body parts) and poor predictive performance resulting from poorly understood visual laws. Motivated by this, I focus on simpler dynamic simulations that I can use to verify the visual understanding that visual modelling pretraining induces. I verify this understanding *qualitatively* in the observed properties of output frames, *quantitatively* with pre-established vision metrics, and *experimentally* with test-tasks that can only be solved if the model understands the appropriate visual laws. To this end, I propose 6 dynamic simulation datasets for video prediction pretraining (see Figure 5.1) with a total of 7 'probing tasks' defined amongst them (*e.g.* a 2D bouncing ball video dataset that also functions as a gravity regression task). I couple the Moving-MNIST (MMNIST) video dataset [180] with MNIST classification [115] for a total of 7 video datasets with 8 probing tasks. As I explore the themes of vision and language modelling, I evaluate these datasets on 3 appropriate models: A fully convolutional 2D *'CNN'* serving as a baseline inspired from vision, a language model style *'Image Transformer'* as a baseline inspired from language modelling, and a *'Patch Transformer'* using convolutions and transformer blocks serving as an overlap of both vision and natural language processing (NLP). I find

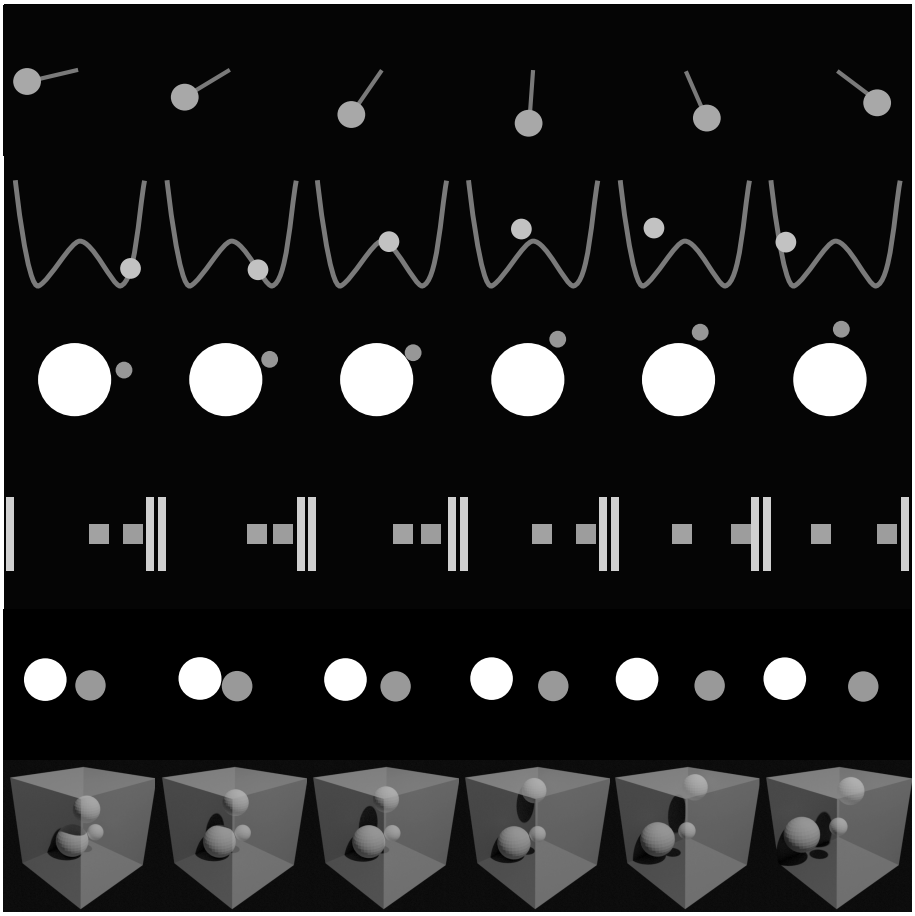


Figure 5.1: Example visualisations of 6 sequential frames from each of the 6 proposed dynamic simulation video datasets. From top to bottom: **Pendulum**, **Roller** coaster with flight, Mars **Moon**, Colliding **Blocks**, **2D Bouncing** balls, and **3D bouncing** balls.

that appropriately focused and simple datasets can demonstrate the potential of visual modelling pretraining for downstream vision tasks. I find both the convolution based models (in particular the patch transformer) outperform the image transformer in frame generation, highlighting the importance of convolutions in video models. I find that these models can generate physically reasonable simulations over 20 frames into the future (despite a buildup of small errors), thus demonstrating the potential of long-term video prediction when visual laws are properly understood. My probing experiments demonstrate that pretraining on the visual modelling task induces features that are directly useful to downstream tasks. Furthermore, I find that finetuning models on the proposed test-tasks that has been pretrained on visual modelling either does not substantially harm, and often greatly, improves test-task

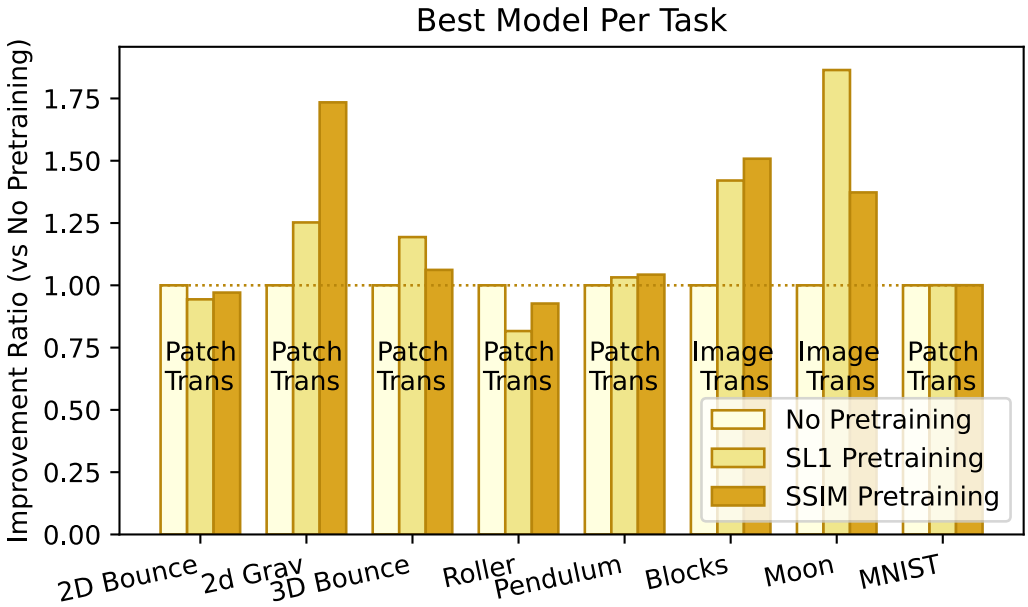


Figure 5.2: The improvement ratio in scores for each task when pretrained on visual modelling, compared with no pretraining, for the best performing model of each task.

performance. 4 of these tasks improve by 20-80% (see Figure 5.2). This demonstrates the potential for predictive pretraining in vision tasks, which represents a small but sure step towards bringing the relative 'power' of vision features in line with those in lanaguage. The visual modelling pretraining paradigm I introduce in this chapter is designed to apply to any appropriate visual task or feature that future models and datasets may leverage. The implementation is available on GitHub[1].

5.2 Related Work

I give an overview of work exploring the overlapping themes of vision and language modelling (*i.e. visual modelling*) and highlight my unique contributions. As I use modern deep learning model architectures to predict videos with visual dynamics and learn physical laws, I briefly summarise both the recent history of *video generation* models, and the state of *visual physics modelling* in deep learning.

[1] https://github.com/Visual-modelling/visual_modelling

5.2.1 Visual Modelling

Though I coin the term 'visual modelling', other works have explored visual parallels to language modelling. Ranzato et al. [164] present an early baseline for unsupervised video feature learning, and demonstrate its use in representing motion and deformation. They find poor long-term prediction results as elements tend to a static position. van den Oord et al. [195] introduce a contrastive learning strategy for unsupervised feature learning in 4 domains (including image), and the BigBiGAN model introduced by [48], achieves state-of-the-art image generation and representation learning results on ImageNet [44]. However, neither of these approaches consider video data. The work in this chapter is most similar in motivation to that of Chen et al. [28] who explore unsupervised representation learning for images using a minimal adaption of transformers and vector-quantised Variational Autoencoder (VAE) [194] architectures. By leveraging a very large amount of computational resources, (a model comparable to GPT-2 *i.e.* 48 layers and ∼1.4B parameters [161]) and using auto-regressive next-pixel prediction or masked-pixel prediction training strategies, Chen et al. [28] present an extensive probing study of the learned representational capacity of layers in their models. They further demonstrate that unsupervised image pretraining leads to state-of-the-art performance on downstream tasks *that increases with model scale*. Unlike the previously discussed approaches, I aim specifically to *pair the visual pretraining datasets with quantifiable and observable test-tasks*, allowing me to strongly argue the benefits of pretraining that the experiments demonstrate. Where Ranzato et al. [164] found poor long-term prediction quality, I now find that modern vision architectures and more focused datasets can improve both long-term prediction and downstream task performance. I show that visual modelling pretraining can lead to substantial performance increases on test-tasks *even without* the large-scale computational resources necessary [28] to try and approach the scale of modern language models.

5.2.2 Video Generation

The recurrent CNN proposed by Ranzato et al. [164] for unsupervised frame prediction and filling drew inspiration from early language models [14] and RNNs [139]. Model predictions often tend towards still images after a few frames allegedly due to the local spatial and temporal stationary assumption made by the model. The authors note that predicting beyond a few frames inevitably invokes the curse of dimensionality and argue it could necessitate moving from pixel-wise prediction to higher-level pixel cluster features. Dosovitskiy and Brox [49] propose a family of deep perceptual similarity metrics 'DeepSiM' to avoid the 'over-smoothed' results of pixel-wise predictions by instead computing distances between image feature vectors. van Amersfoort et al. [193] propose a convolutional network to generate future frames by predicting a transformation based on previous frames and constructing the future frames accordingly, leading to sharper images and simultaneously avoiding the curse of high dimension predictions. Wang et al. [201] propose an integrated Bayesian framework to cope with uncertainties caused by noisy observations (*i.e.* perceptual) and forward modelling process (*i.e.* dynamics). Yilmaz and Tekalp [223] use deformable convolutions [36] to try and exploit a larger and more adaptive receptive field as opposed to normal convolutions. The recently released VideoGPT [214] is a video generation model which combines vector-quantised VAE and transformer designs with large-scale training setups similar to those used by [28]; *i.e.* similar in scale to Image-GPT and trained on up to 8 Quadro RTX 6000 GPUs. Video-GPT yields very high quality frame predictions on the UCF-101 [179] and TGIF [123] datasets . However, due to the inherent difficulty of modelling complex real-world long-term videos, errors in motion still build up.

I acknowledge the difficulties that even richly resourced models encounter and echo Yan et al. [214]: videos are just simply a "hard modelling challenge". I instead focus my efforts on demonstrating what *is* possible when the visual laws underpinning the video data are kept appropriately simple. See the work of both Oprea et al. [150] and Zhou et al. [238] for a thorough review and survey of video generation.

5.2.3 Visual Physics Modelling

Wu et al. [206] collect the 'Physics 101' dataset which facilitates models explicitly learning physical properties of objects in videos (*e.g.* mass, acceleration, and friction). Where they focus on encoding physical laws into neural networks, I additionally explore if generative visual modelling is a sufficient or desirable method to induce a quantifiable understanding of these laws. Neural networks have successfully modelled a variety of dynamic systems using images: *e.g.* fluid flow [189], Lyapunov functions [138], motion flow [38] and precipitation nowcasting [174]. Li et al. [124] propose a novel Fourier neural operator that can learn Burger's equations, Darcy flow, and Navier-Stokes with differing input image resolutions. Recently, Wang et al. [200] push for more generalisable physical modelling with their proposed multi-task DyAd approach.

5.3 Models and Configurations

To explore the visual parallels of language modelling, I focus on both CNNs (due to their long history of state-of-the-art success in computer vision) and Transformers (because of their well established dominance in language modelling). To this end, I experiment on 3 models:

1. A fully convolutional *'CNN'* model with skip connections as a baseline model from vision (Figure 5.3).

2. A multi-head attention transformer typical of language modelling with minimal redesigns to accommodate video prediction that serves as a candidate from language models. I call this the *'image transformer'* (Figure 5.4).

3. The recently proposed SegFormer [208] transformer minimally adapted from semantic segmentation to video generation as a transformer model directly designed for use in vision. I call this adaptation the *'patch transformer'* (Figure 5.5).

Each of these models has been designed or adapted to take a sequence of video

frames as input, and output predictions for the next frame (further described in Section 5.5).

5.3.1 Fully Convolutional 2D CNN

Introduced by Long et al. [133], fully convolutional neural networks (FCN) use 'deconvolution' layers [231] *i.e.* convolutional layers with fractional strides. Deconvolution layers can be used to 'reverse' the convolution layers and generate a full sized output. Note that upsampling between layers can be learned or can be fixed (*e.g.* bilinear upsampling). As such, FCNs offer relatively inexpensive forward and backward computation, and a reversal of convolution layers back up to the original input dimensions for outputs. My CNN baseline model (depicted in Figure 5.3) is a U-Net style [168] FCN with skip connections [83]. Given a sequence of m frames of a video, the model takes as input the m 64×64 grayscale input frames in temporal order as inputs for m channels into the first of a series of U-Net style double-convolution units. Each subsequent step halves the image resolution and doubles the number of channels from a starting factor of 64. The upwards pass uses bilinear upsampling and halves the number of channels, mirroring the downward pass in reverse, resulting in the final output frame.

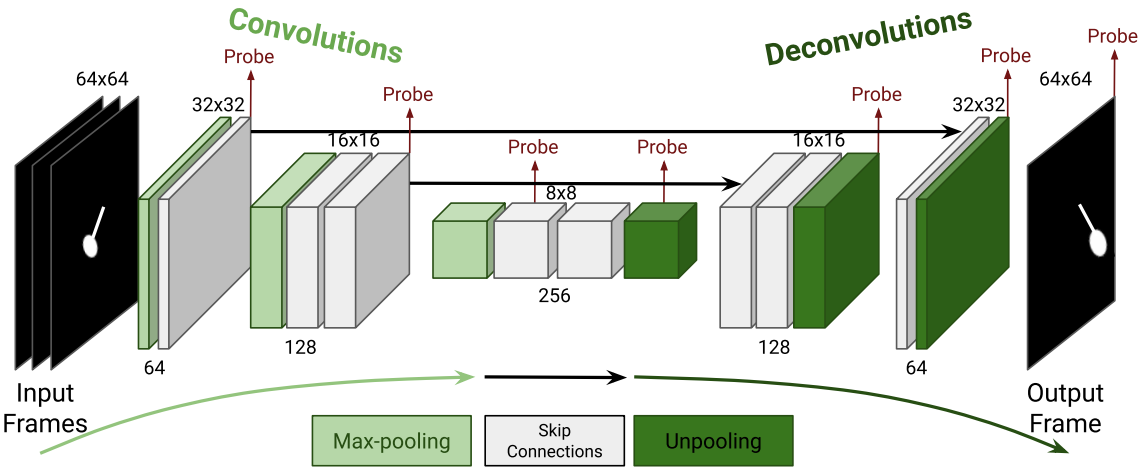


Figure 5.3: Fully Convolutional 2D **CNN** model. Each convolution unit is made from 2 convolution layers, *i.e.* an initial convolution layer that changes the input resolution, followed by another of kernel size 1×1 that does not. The arrows labelled 'Probe' indicate which points in the network are extracted to form linear probes used in Section 5.5.2.2.

5.3.2 Image Transformer

I adapt the model introduced by Vaswani et al. [197] and deploy it on sequences of images that form a video instead of word embeddings that form language (Figure 5.4). I use every 64×64 pixel image as a single token consisting of 4096 features. Unlike the original transformer that was trained on a language translation task, here the output is only conditioned on its previous video frames. I choose not to use an encoder and only use the decoder to predict the next frame output frame because the input and output image sequences are not synchronised over the same time span (there are m input frames translated to a single output frame; *i.e.* $m \neq 1$). Since the tokens represent images, I use a pixel regression layer at the end of the architecture. Using a dedicated pixel regression layer allows the output tokens to represent images while alleviating this requirement from the transformer blocks. Although I can train using batches with m input frames and predict one output frame, as in the CNN and patch transformer, I can instead train the model with the entire sequence at once. This can be done by predicting the next frame for every input using a masked multi-head attention where every output is only conditioned on the past frames. I find negligible difference in performance between the 2 approaches, and therefore decide to use the fixed m input and one output for a more direct comparison with

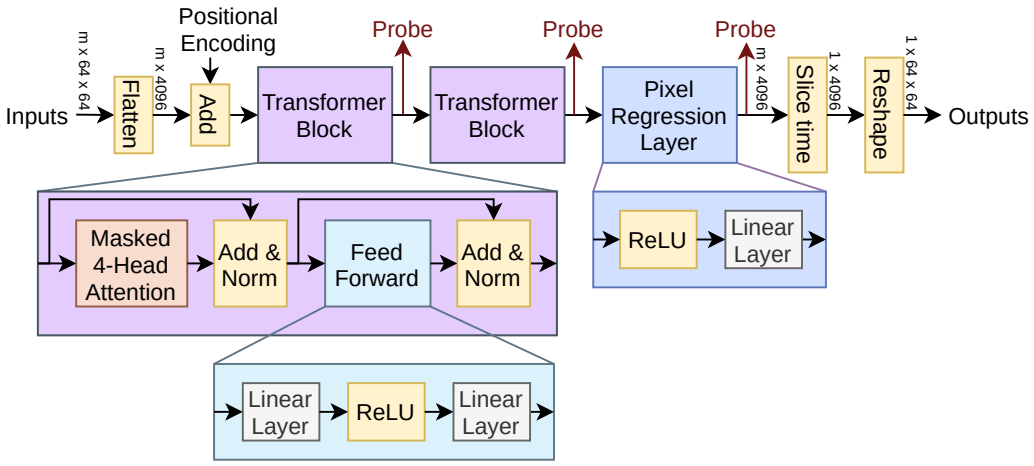


Figure 5.4: The **Image Transformer** model serves as a candidate from language modelling. I use 64×64 images as tokens. The arrows labelled 'Probe' indicates which points in the network are extracted to form linear probes used in Section 5.5.2.2.

the other 2 models. I found negligible difference in performance when using more than 2 transformer blocks. Similarly, I found a smaller number of heads in the multi-head attention block to be beneficial, and thus only used 4.

5.3.3 Patch Transformer

I use a transformer-based semantic segmentation model and modify it for visual modelling (Figure 5.5). Instead of feeding the model RGB images with 3 channels, I use the sequence of video frames as the input channels and predict the following frame as the output. Unlike the Image Transformer that applies the attention layers across the time sequence of video frames, the patch transformer applies the attention layers across patches of the images. I base the patch transformer architecture on the SegFormer model [208] with the following light modifications. Since the visual modelling task requires the same resolution of the input and output images and the SegFormer outputs at 1/4 resolution, I balance this by predicting 16 times as many channels and fold each pixel with 16 channels into a 4×4 patch with one channel. When compared to the original SegFormer study, the smaller size of images I use

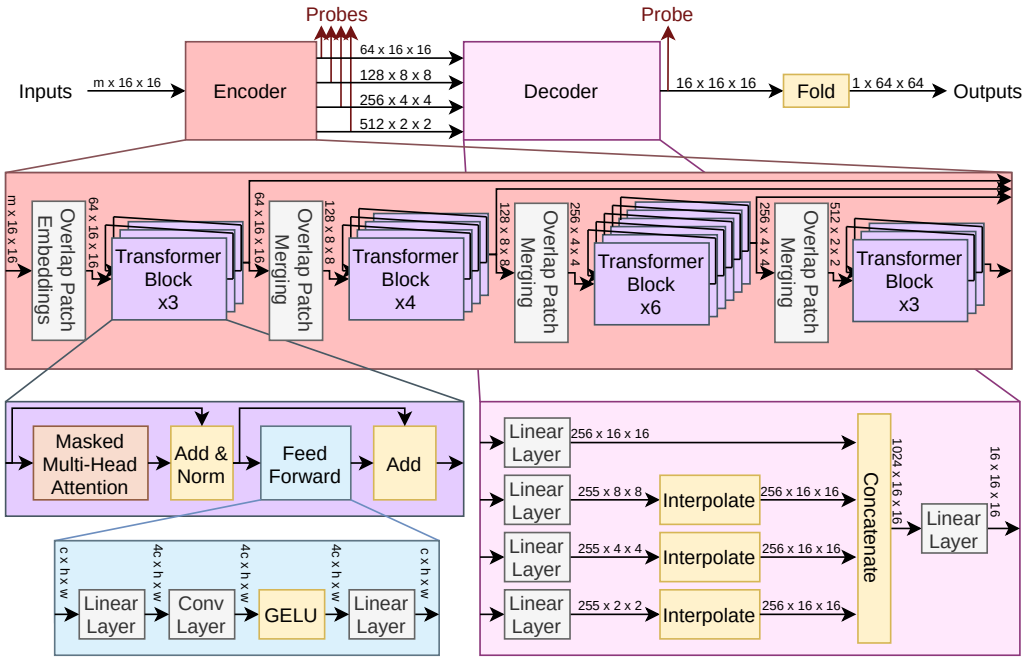


Figure 5.5: The **Patch Transformer** model, adapted from the SegFormer [208] for video generation. The arrows labelled 'Probe' indicate which points in the network are extracted to form linear probes used in Section 5.5.2.2.

allows me to use a reduction ratio 'R' of 1 in the efficient attention layer (*i.e.* not compromising the representational capacity). This makes it functionally equivalent to a standard multi-head attention layer.

5.4 Datasets

To explore the potential of visual modelling, given the previously discussed limitations of the field, I focus my experiments on datasets that are both:

1. Simple dynamic simulations that can be used for video prediction.
2. Naturally affiliated with a classification or regression 'downstream' task accompanying the modelling.

See Table 5.1 for examples of the proposed datasets and details on the test-tasks they pair with. I further ablate on 2 real world datasets, CMU Motion Capture 'MOCAP' dataset[2] and the human motion database 'HMDB-51' [111] in order to help demonstrate the limitations of the 3 models.

5.4.1 2D and 3D Bouncing Balls

The 2D bouncing dataset consists of videos with 1-3 balls which can collide with both each other and the borders of the image. In this explicit Euler simulation, I vary: the number of balls, ball radius (per ball), initial position (per ball), initial velocity (per ball), gravity strength, gravity direction, background colour, and ball colour (per ball). I define 2 downstream test-tasks on this 2D dataset: *total number of bounces prediction* and *y-directional gravity prediction*. Where the 2D version represents the balls as simple circles, I extend this approach to 3D, rendering the balls using realistic lighting from a single light source. This 3D scenario is designed to be more a challenging dataset as balls occlude each other and cast shadows on both the environment boundary and on other balls. I define one downstream test-task for this 3D bouncing dataset: *total number of bounces prediction*.

[2] http://mocap.cs.cmu.edu/

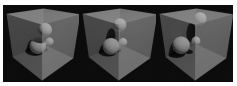
Dataset	Example	# Videos	Vid Length	Affiliated Test-Task(s)
2D Bouncing		20,000	60	**Bounce Regression:** *59 input frames.* Count total bounces demonstrated in the input video. Both ball-to-ball and ball-to-wall bounces, capped at a maximum of 50. **Gravity Regression:** *5 input frames.* Predict the gravity demonstrated in the given 5 frames. Gravity is in the y axis with 7 potential values [-3e-4, -2e-4, $\cdots$, 3e-4].
3D Bouncing		10,000	100	**Bounce Regression:** *99 input frames.* Count total bounces demonstrated in the input video. Both ball-to-ball and ball-to-wall bounces, capped at a maximum of 50.
Roller		10,000	100	**Gravity Regression:** *5 input frames.* Predict the gravity demonstrated in the given 5 frames. Gravity is in the y axis with 201 potential values [0, 0.5, 1, $\cdots$,100].
Pendulum		10,000	100	**Gravity Regression:** *5 input frames.* Predict the gravity demonstrated in the given 5 frames. Gravity is in the y axis with 41 potential values [0, 0.5, 1.0, $\cdots$, 20].
Blocks		10,000	100	**Block Mass Difference Regression:** *49 input frames.* 2 blocks of different masses move towards each other on a smooth surface and collide. Predict the difference of the masses between the blocks, positive or negative (positive direction is fixed). Block 1 is always of mass 10, block 2 takes 39 different masses [0.5, 1.0, $\cdots$, 19.5].
Moon		10,000	100	**Moon Mass Regression:** *5 input frames.* Predict the gravity demonstrated in the given 5 frames. Gravity acting on the small moon towards the centre of a planet. Masses have 26 potential values [70, 75, 80, $\cdots$,195].
MMNIST		5,200	100	**MNIST Classification:** *1 input frame.* Input MNIST frame. Copied 5 times for a model pretrained on 5-1 input-ouput for example.

Table 5.1: Further details of the datasets and their affiliated test-tasks. The constants for predictions for all test-tasks (aside from MNIST classification) are normalised such that the standard deviation of constants across each individual dataset is 1.

5.4.2 Myphysicslab

Myphysicslab[3] is a series of open-source animated physics simulations involving pendulums, springs, collisions, and more. These simulations are calculated using the Runge-Kutta method [19]. I experiment with the following simulations:

5.4.2.1 Mars Moon

A simplified simulation of an asteroid orbiting a moon using a rigid body simulation. I vary the initial velocity, moon radius, moon mass, and asteroid radius. The test-task is to *predict the mass of the moon.*

5.4.2.2 Colliding Blocks

Simulates 2 blocks that move along a single axis colliding with both the boundary walls and each other. I vary the masses of *one* of the blocks (leaving the other block mass fixed), the starting positions, and starting velocities. The test-task is to *predict the difference in the masses of the 2 blocks.*

5.4.2.3 Pendulum

A single pendulum modeled as a point mass at the end of a massless rod. I vary the initial angle, gravity strength, pendulum length, and pendulum mass. The test-task is to *predict the gravity acting on the pendulum.*

5.4.2.4 Roller Coaster with Flight

A ball of mass M is released down a curved track under gravity g following $F = M \cdot g \cdot \cos(\theta)$. The ball can switch to free flight when the acceleration normal to the curve is greater than v^2/k (where v is the velocity of the ball, and k is the radius of curvature at the current point along the curve). I vary the gravity strength and track position. The test-task is to *predict the strength of gravity acting on the ball.*

[3] https://www.myphysicslab.com

5.4.3 Moving MNIST

Moving MNIST (MMNIST[4]) refers to a video dataset [180] where one or more MNIST [115] digit(s) are moving around a black background (overlapping each other and bouncing off the frame boundaries). The digits move at fixed constant velocity with no friction or gravity acting on them. The original dataset was 10,000 sequences with 2 digits in each. I adapt code[5] to generate an MMNIST dataset containing videos with 1-3 digit(s). MMNIST serves as a simple dynamics dataset that can be naturally considered alongside a downstream vision task: MNIST classification.

5.4.4 CMU Motion Capture

The CMU Motion Capture dataset is comprised of videos focusing on the motion of humans and objects. I extract images from the 937 videos[6] at 10 frames-per-second. I convert the coloured frames to grayscale and then downscale and crop them to a resolution of 64×64, in line with the other datasets.

5.4.5 HMDB-51

Motivated to challenge the high performance of models on the relatively simple action datasets of the time, the Human Motion Recognition database (HMDB-51) [111] is a large action video dataset with 51 action categories. HMDB-51 aims to better capture the "richness and complexity of human actions", with videos including more challenging clutter and occlusion not typical of action benchmarks at the time. Categories include 'brush hair', 'sword exercise', and 'jumping'. I extract images at 10 frames-per-second and convert and crop the images to grayscale in 64×64 resolution.

[4] http://www.cs.toronto.edu/~nitish/unsupervised_video/
[5] https://gist.github.com/tencia/afb129122a64bde3bd0c
[6] http://mocap.cs.cmu.edu/allmpg/

5.5 Experiments

My experiments are carried out on the same 3 model architectures but are comprised of 2 separate instances of training:

- **Modelling tasks**, a generative training strategy intended to mirror language modelling and to induce an understanding of visual dynamics.
- These are followed by **Test-tasks**, a classification or regression task paired with their appropriate modelling counterpart that function as downstream vision tasks.

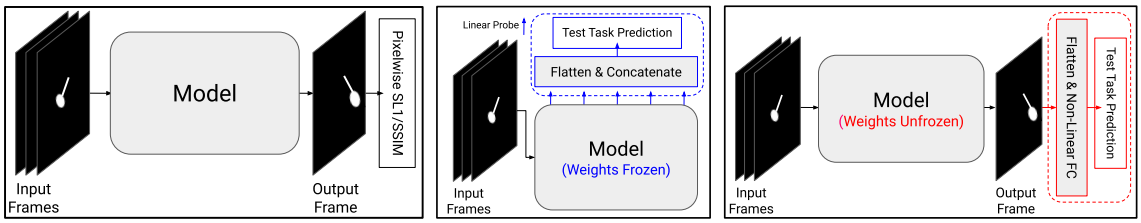
Together these tasks allow me to explore which laws of dynamics current vision benchmarks can model, what information and understanding visual modelling (*i.e.* next-frame prediction) *induces*, and to what extent this *induced understanding* is desirable as a starting point for downstream vision tasks.

5.5.1 Modelling Tasks

Given a sequence of $m+1$ frames of a video, the model takes as input the first m 64×64 grayscale input frames in temporal order and predicts as output the next frame. This predicted frame is compared against the final (ground truth) frame in the $m+1$ sequence. A value is predicted for each pixel (*i.e.* dense prediction) and the sigmoid function is used as activation for the outputs of the final layer, which are then multiplied by 255 to create the resulting output grayscale image. I use either smooth-$L1$ (SL1) or Structural Similarity (SSIM) as loss functions. As SSIM takes values between -1 and 1, and should be maximised, I reformulate it as a minimisation problem (as in Equation 5.1) in order to use it as a loss function:

$$\mathcal{L}_{\text{SSIM}} = 1 - \frac{1 + SSIM}{2} \in [0, 1]. \tag{5.1}$$

Using the 3 models introduced in Section 5.3, I perform modelling experiments (see Figure 5.6a) on the 9 datasets introduced in Section 5.4. The values of m in my experiments (*i.e.* number of input frames) depends on the downstream test-task it will be paired with. For example, the 2D bouncing, roller, pendulum, and



(a) Modelling (b) Linear Probes (c) Direct Learning

Figure 5.6: The 3 different experimental setups. All 3 of these model architectures can be adapted to visual modelling (*i.e.* video prediction; Figure 5.6a), and the accompanying test-tasks with either linear probing of frozen models (Figure 5.6b) or directly learning and finetuning on the test-task (Figure 5.6c).

moon datasets all have downstream test-tasks involving gravity, thus 5 input frames should be sufficient to demonstrate gravity. The 2D and 3D bouncing datasets have 'counting bounces' tasks associated with them, thus $m = 59$ and 99 for 2D and 3D datasets respectively (*i.e.* the length of the video clip; it is necessary to to see the entire clip to predict the total number of bounces). See Table 5.1 for further details. To assess the of performance the modelling task, I consider the Peak Signal-to-Noise Ratio (PSNR) [88], Structural Similarity (SSIM) [239], and $L1$ scores between the predicted frame and the ground truth frame. I do not consider metrics such as Learned Perceptual Image Patch Similarity (LPIPS) [235] and Fréchet Video Distance (FVD) [192]. LPIPS is a deep model based metric that is only well defined on RGB images, and FVD requires an image resolution of at least 224×224.

5.5.2 Test-Tasks

Each of the modelling datasets is designed with a complimentary test-task in mind (except 2D bouncing which has 2), and I further pair the MMNIST dataset with the test-task of MNIST classification for a total of 8 test-tasks. I apply a cross entropy loss for MNIST classification and a smooth-$L1$ loss with $\beta = 0.01$ for all other tasks. In this subsection I describe my 3 categories of test-task experiments: random baselines, frozen model probing, and direct training or finetuneing.

5.5.2.1 Random Scores for Tasks

To contextualise the scores for each task, I ablate with 3 scenarios designed to give a lower bound on performance (seen on the first, second, and third lines respectively of each section in Tables 5.3 and 5.4):

1. Constant Output: I instantiate a layer of biases and optimise them directly on the loss of each task. This scenario takes no inputs but *is* given the ground truth. This should theoretically learn to mimic the average output of values that the ground truth alone would induce.

2. Image + Linear Layer: I flatten the input images and pass them into a trainable linear layer.

3. Frozen Random Model: I randomly instantiate the full model and freeze the weights. I pass the loss function through trainable linear probes in the same way as in the probing experiments.

5.5.2.2 Probing Frozen Models

I seek to ascertain if training on a given modelling dataset induces an understanding of the appropriate physical laws (*e.g.* does pretraining on bouncing balls dataset induce a verifiable understanding of gravity?). To this end, I freeze the weights of a pretrained model and repurpose it for the appropriate test-task (see Figure 5.6b) by flattening and concatenating the outputs of each substantial layer throughout the network (see 'Probe' arrows in Figures 5.3, 5.4 and 5.5) and passing them through a trainable linear layer and into the appropriate loss function. It has been shown that CNN and transformer-based language models 'learn representations that vary with network depth' [159], and that there is varying transferability of representations in different layers of language models [131]. This motivates me to form the linear probe from the output of all substantial blocks of the networks. Though the model as a whole may contain the information needed to solve a test-task, only considering the outputs of the final layer can be insufficient as that layer is perhaps instead for example focused on solving the pixel regression.

5.5.2.3 Finetuning for Downstream Test-Tasks

The proposed test-tasks allow me to verify if pretraining provides an advantageous starting point for downstream vision tasks when compared with random initialisation (*e.g.* will I notice superior performance on gravity prediction after pretraining the model to understand bouncing?). Given a network that is either pretrained on visual modelling or randomly instantiated: I unfreeze the weights, flatten the outputs of the final layer, and pass them through a non-linear (GELU) fully connected unit with dropout and batch normalisation. Lastly, the outputs of the fully connected unit are passed to the appropriate loss function (see Figure 5.6c).

5.6 Results and Discussion

In this section I discuss the results of my visual modelling and test-task experiments.

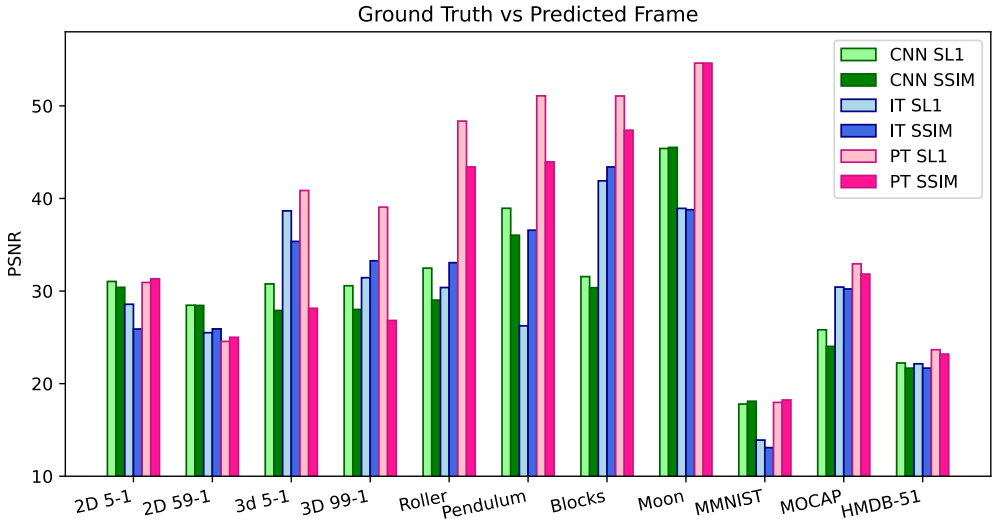
5.6.1 Modelling Quality

I consider the quality of the generated frames by computing PSNR, SSIM and $L1$ scores between the predicted and ground truth frame (Figure 5.7). I provide an alternative tabularised version for closer inspection of these scores in Table 5.2. In the following subsections, I discuss the differences in modelling quality with respect to the different datasets, models, and losses.

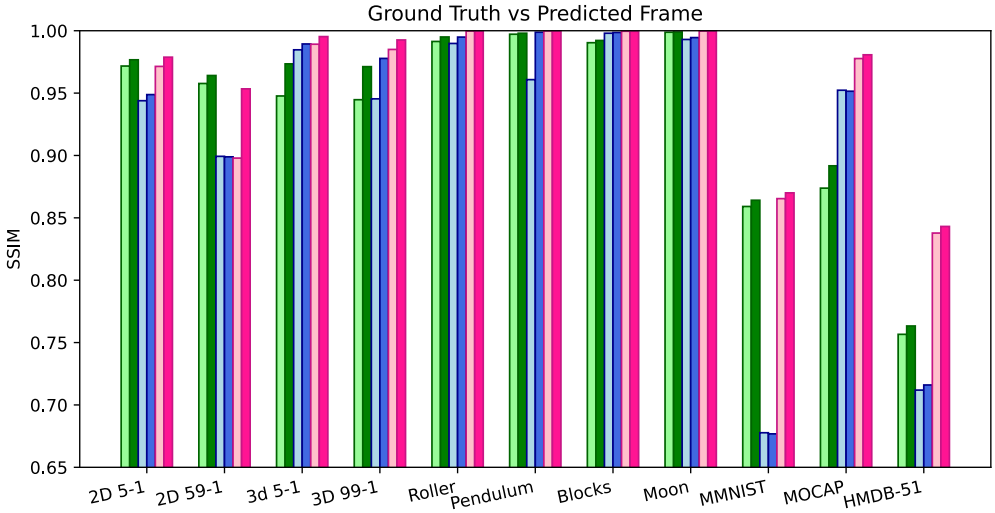
5.6.1.1 Modelling Quality by Dataset

The results show that the modelling tasks yielding the highest scores are roller, pendulum, blocks, and moon. This is most obviously seen in the PSNR values from Figure 5.7a (over 8 PSNR higher than the other datasets for the best model). This trend is echoed with the very high SSIM ($\sim$0.998; see Figure 5.7b) and lower $L1$ scores (see Figure 5.7c). I believe this is because the background for these tasks is always black, and there are less structural variations than the other datasets. Next are the 2D and 3D bouncing datasets which score slightly lower than the top performing 4 datasets. Though I expected the 3D bouncing dataset to be more

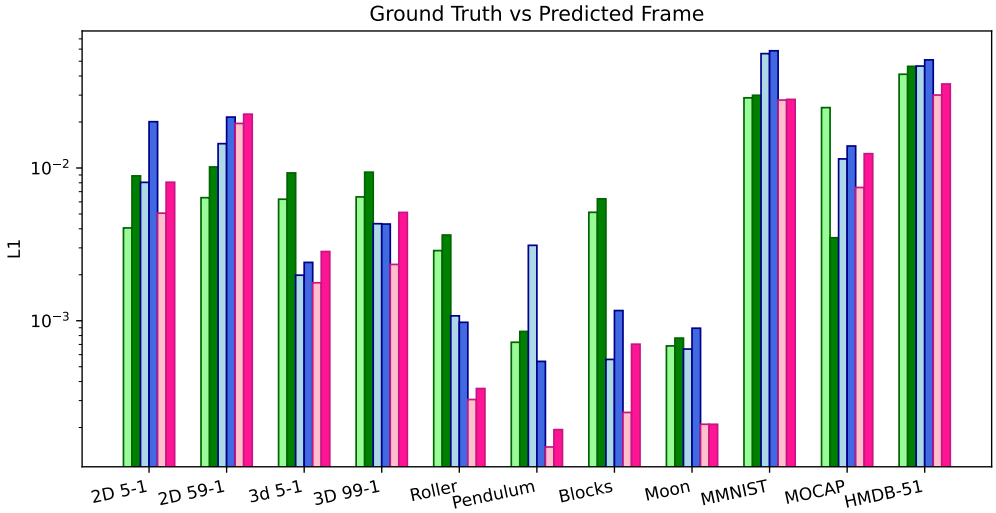
complicated than the 2D version because of the added potential for occlusion and z-dimensional movement, the 3D bouncing dataset yields slightly higher scores than the 2D bouncing dataset (∼2-10 PSNR higher for the Image and Patch Transformers). I argue that although 3D bouncing must model motion in an extra dimension, the fixed background and colour scheme of the 3D dataset allow an extra boost to metric scores when compared to the 2D bouncing dataset which has varying the background and ball colours. I find that increasing the number of input frames (m) for the 2D and 3D bouncing datasets from 5 to 59 and 99 respectively causes a very minor but consistent decrease in all 3 scores (by comparing the 1^{st}-2^{nd}, and 3^{rd}-4^{th} entries of each sub-figure in Figure 5.7 respectively), implying there is little to gain from increasing the input context for these modelling tasks. The real world MOCAP and HMDB-51 datasets score lower than all other datasets except MMNIST. Though the models explored in this chapter fail to capture the far more complex motion laws dictating these real world datasets, the single predicted frame from these real world datasets does not change substantially from the inputs. I argue that the movements in these videos, despite being underpinned by complex motion, are relatively small as the entities in them are small. This means that predicted movements are often smaller in terms of absolute pixel variations when compared to the movements of the high-contrast shapes seen in the simulation datasets I propose, causing smaller penalties in the 3 metrics than might be expected. I discuss the inability of my models to model *long-term* real-world movements in Section 5.6.2.2. The MMNIST modelling task scores lowest on the 3 metrics of all datasets. Though intuitively MMNIST should score higher than the real world datasets, I believe this result is unsuprising for several reasons. MMNIST features a black background contrasted with multiple white digits, meaning that a prediction with a white digit in a slightly incorrect location will yield very high absolute differences in raw pixel values and thus lower metric scores *e.g.* consider the difference in raw pixel values of a human (mostly grey pixels) jumping slightly vs. a fully white shape of the same relative size moving across a black background. The shape and structure of digits is more complex than the circles and spheres typical of the other datasets. Furthermore, the holes and lines tend to fill and fade, and each digit is moving in different direc-



(a) PSNR↑



(b) SSIM↑



(c) L1↓

Figure 5.7: Metrics calculated between the ground truth and the predicted frame on each modelling dataset. ↑ (↓) indicates that a higher (lower) score is better.

Metric	CNN		Img Trans		Patch Trans	
	SL1	SSIM	SL1	SSIM	SL1	SSIM
2D Bouncing $m = 5$						
PSNR↑	31.036	30.400	28.572	25.900	30.924	31.328
SSIM↑	0.9717	0.9767	0.9439	0.9488	0.9714	0.9788
L1↓	4.051e-3	8.879e-3	8.053e-3	2.009e-2	5.062e-3	8.069e-3
2D Bouncing $m = 59$						
PSNR↑	28.468	28.446	25.504	25.915	24.564	25.007
SSIM↑	0.9577	0.9641	0.8993	0.8989	0.8979	0.9534
L1↓	6.393e-3	1.016e-2	1.443e-2	2.154e-2	1.957e-2	2.251e-2
3D Bouncing $m = 5$						
PSNR↑	30.772	27.903	38.661	35.631	40.859	28.138
SSIM↑	0.9477	0.9734	0.9847	0.9894	0.9892	0.9953
L1↓	6.247e-3	9.284e-3	1.988e-3	2.410e-3	1.773e-3	2.839e-3
3D Bouncing $m = 99$						
PSNR↑	30.571	28.013	31.443	33.269	39.060	26.823
SSIM↑	0.9447	0.9712	0.9454	0.9778	0.9850	0.9926
L1↓	6.468e-3	9.388e-3	4.318e-3	4.297e-3	2.337e-3	5.120e-3
Roller $m = 5$						
PSNR↑	32.475	29.023	30.381	33.063	48.358	43.404
SSIM↑	0.9914	0.9950	0.9898	0.9949	0.9998	0.9998
L1↓	2.879e-3	3.642e-3	1.077e-3	9.768e-4	3.051e-4	3.600e-4
Pendulum $m = 5$						
PSNR↑	38.941	36.031	26.246	36.574	51.085	43.948
SSIM↑	0.9972	0.9980	0.9608	0.9988	0.9998	0.9998
L1↓	7.231e-4	8.512e-4	3.117e-3	5.420e-4	1.492e-4	1.938e-4
Blocks $m = 49$						
PSNR↑	31.561	30.365	41.905	43.402	51.069	47.371
SSIM↑	0.9904	0.9922	0.9980	0.9986	0.9996	0.9996
L1↓	5.128e-3	6.285e-3	5.585e-4	1.166e-3	2.510e-4	7.025e-4
Moon $m = 5$						
PSNR↑	45.399	45.515	38.928	38.785	54.618	54.618
SSIM↑	0.9988	0.9990	0.9930	0.9945	0.9998	0.9998
L1↓	6.844e-4	7.699e-4	6.529e-4	8.937e-4	2.100e-4	2.100e-4
MMNIST $m = 5$						
PSNR↑	17.794	18.097	13.904	13.089	17.975	18.237
SSIM↑	0.8591	0.8641	0.6777	0.6768	0.8654	0.8700
L1↓	2.880e-2	2.992e-2	5.607e-2	0.05853	2.784e-2	2.812e-2
MOCAP $m = 5$						
PSNR↑	25.816	24.024	30.431	30.226	32.943	31.835
SSIM↑	0.8738	0.8917	0.9523	0.9515	0.9777	0.9807
L1↓	2.485e-2	3.495e-3	1.147e-2	1.393e-2	7.456e-3	1.241e-2
HDMB51 $m = 5$						
PSNR↑	22.229	21.675	22.141	21.673	23.664	23.191
SSIM↑	0.7566	0.7633	0.7119	0.7160	0.8378	0.8431
L1↓	4.109e-2	4.627e-2	4.649e-2	5.102e-2	3.004e-2	3.549e-2

Table 5.2: Metrics between the first generated image and its respective ground truth. All metrics are reported from the best epoch of the models respective loss. The L1 metric is calculate with mean reduction. ↑ (↓) indicates higher (lower) is better.

tions. As I explore later in Section 5.6.2, the long-term predictions for MMNIST are more physically realistic than those generated from the real world datasets despite these lower first-frame prediction metrics. This stresses how crucial it is to comple-

ment metrics for first-frame prediction with long-term predictive analysis in video generation.

5.6.1.2 Modelling Quality by Model

The patch transformer consistently has the best PSNR, SSIM, and $L1$ metric scores for all but one of my experiments (2D bouncing with $m = 59$). The CNN and image transformer alternate as the second best model across both the easier and harder tasks. The patch transformer gives moderately higher scores for 2D bouncing ($m = 5$), 3D bouncing ($m = 5$), MMNIST, and the real world datasets ($\sim$1-3 higher PSNR scores). On the 4 datasets with best scores overall (roller, pendulum, blocks, and moon), the patch transformer demonstrates considerably higher scores when compared with the other 2 models. This implies that the patch transformer excels at replicating the more controlled conditions in these datasets, indicated by its very high SSIM scores ($\sim$0.999).

5.6.1.3 Modelling Quality by Loss

Models trained with the SSIM-based loss demonstrate higher SSIM scores on their predictions compared to those pretrained on SL1. This can be seen in Figure 5.7b, where the higher values of the dark bars of each colour represent SSIM training for each model variant ($\sim$0.01-0.05 increase consistently). Conversely, models trained with the mean pixelwise SL1 loss have a lower (*i.e.* better) mean pixelwise $L1$ score as seen in the lower lighter bars representing SL1 training in Figure 5.7c. This improvement in each metric of models trained with that metric's respective loss counterpart further highlights how limited any single metric is in demonstrating the quality of generated images. Despite each score's preference for its own loss function, I find that SL1 trained models almost always give a higher PSNR than their SSIM counterparts (Figure 5.7a). This more pronounced covariance of PSNR and SL1 scores is expected behaviour as PSNR tends to infinity as mean squared error (and hence SL1) tends to zero. This suggests that PSNR is optimised for by an SL1 loss.

5.6.2 Long-Term Self-Output Prediction

As previously described, the models are trained using subsets of clips of size $(m + 1)$ where m is the number of input frames and the final frame serves as ground truth for the prediction. In order to gauge the model's capacity for long-term video generation, I consider 'self-output' experiments, *i.e.* for a video in the test set:

1. Generate the predicted frame from the first m frames.
2. Create, as new inputs, the starting m frames with the first frame removed and the newly generated frame added to the end.

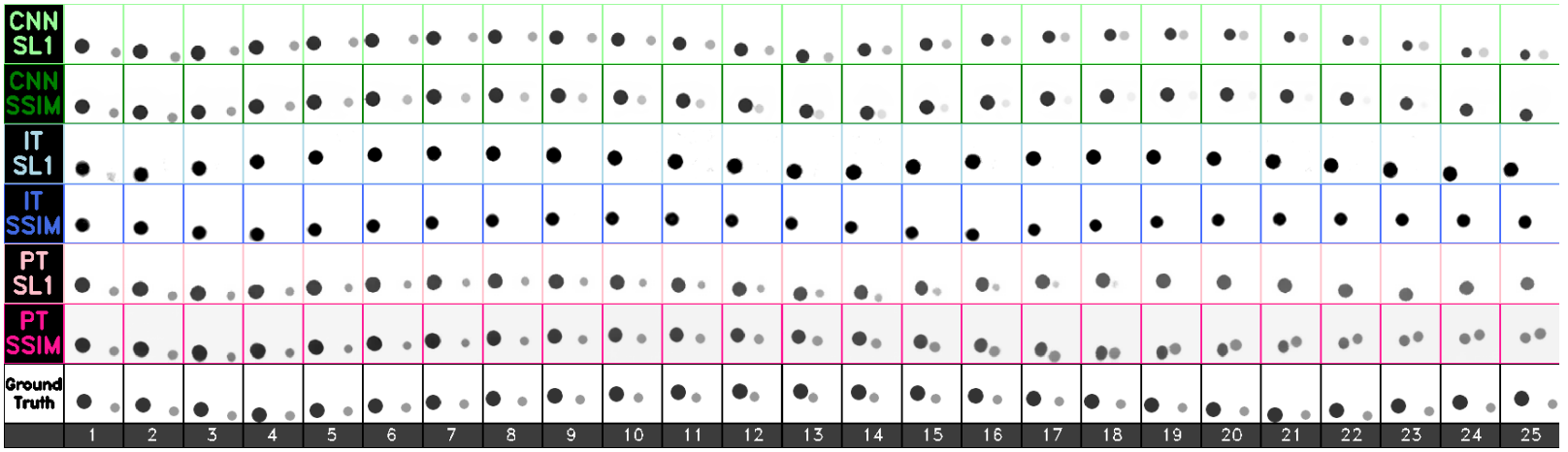
Algorithm 1 Self-Output Visualisation.

Input: *video*, m

```
totalFrames ← length(video)
inputs ← video[0 : m]                          ▷ Python-style slicing.
for i ← m to totalFrames do
    output ← model(inputs)
    groundTruth ← video[i]
    plotTogether(groundTruth, output)          ▷ See Figure 5.8.
    inputs ← concatenate(inputs[1 : m], output)   ▷ Shift one frame into the
future.
end for
```

This effectively tests the model's capacity to continue generating the video from the initial m frames by using the predicted frame to shift the next inputs into the future one frame at a time. Using the process described in Algorithm 1, I create a side-by-side comparison of the predicted video frames and their ground truth counterparts in Figures 5.8, 5.9, 5.10, 5.11, and 5.12. As it would be intractable to discuss detailed behaviours from each model on each dataset and with both loss functions, I focus on the major trends and properties I can observe in the generated videos. I invite readers to explore the complete set of self-output videos and metrics for each test set [7]. In the following subsections, I discuss the differences in self-output predictions with respect to the different models and losses.

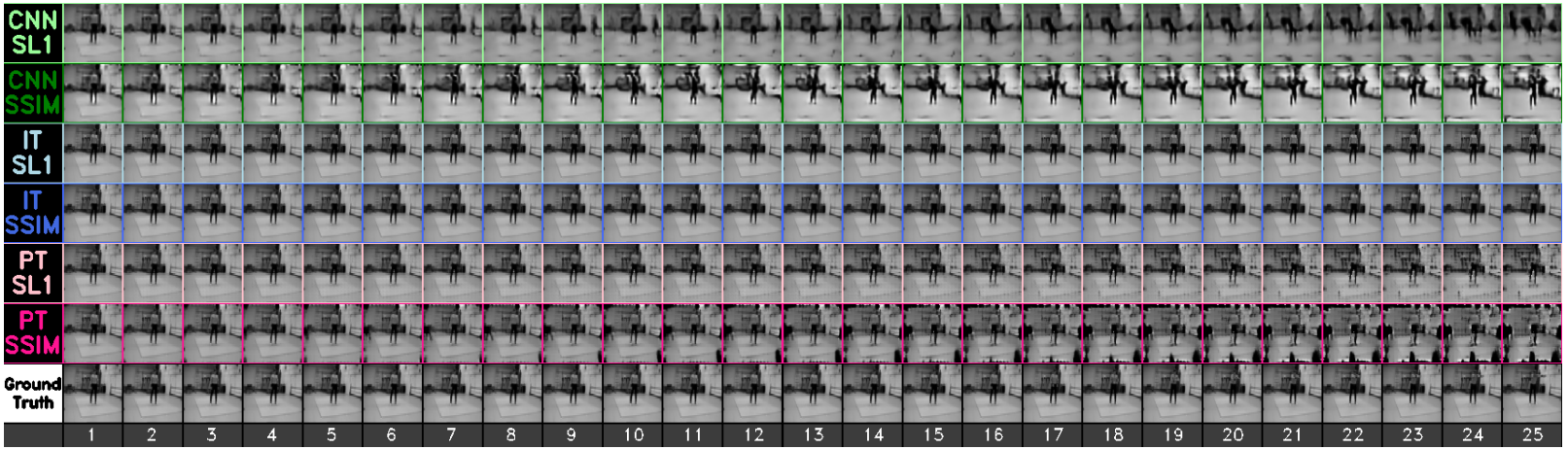
[7] https://github.com/Visual-modelling/visual_modelling#all-self-output-gifs



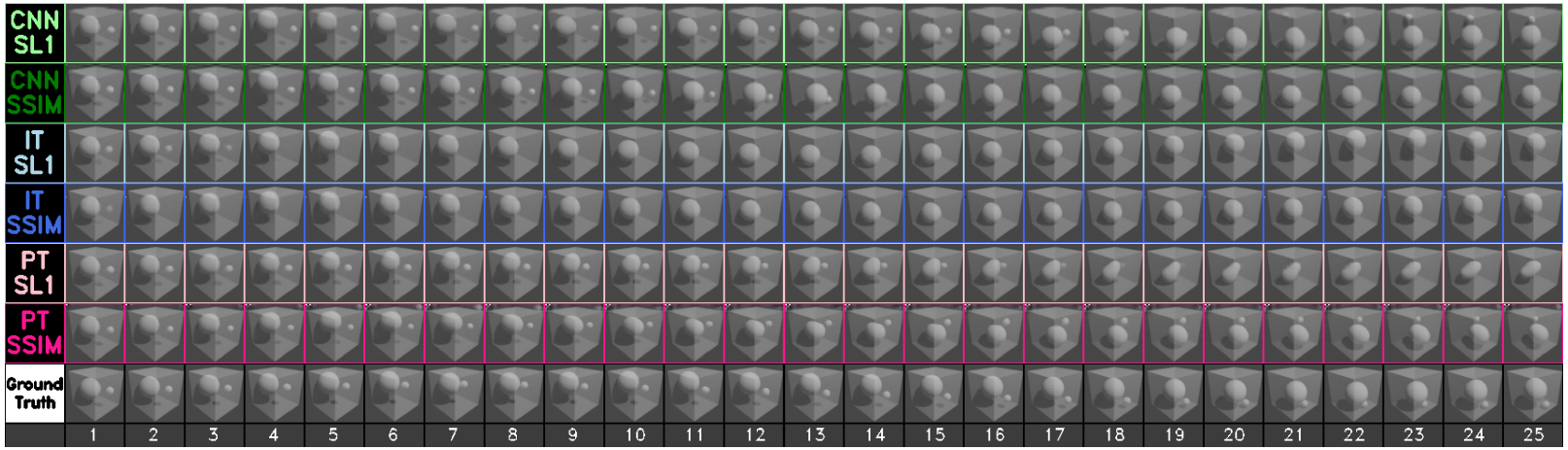
(a) 2D Bouncing

(b) MMNIST

Figure 5.8: Comparison of the first 25 generated frames of each model ($m = 5$) visualised alongside the ground truth.



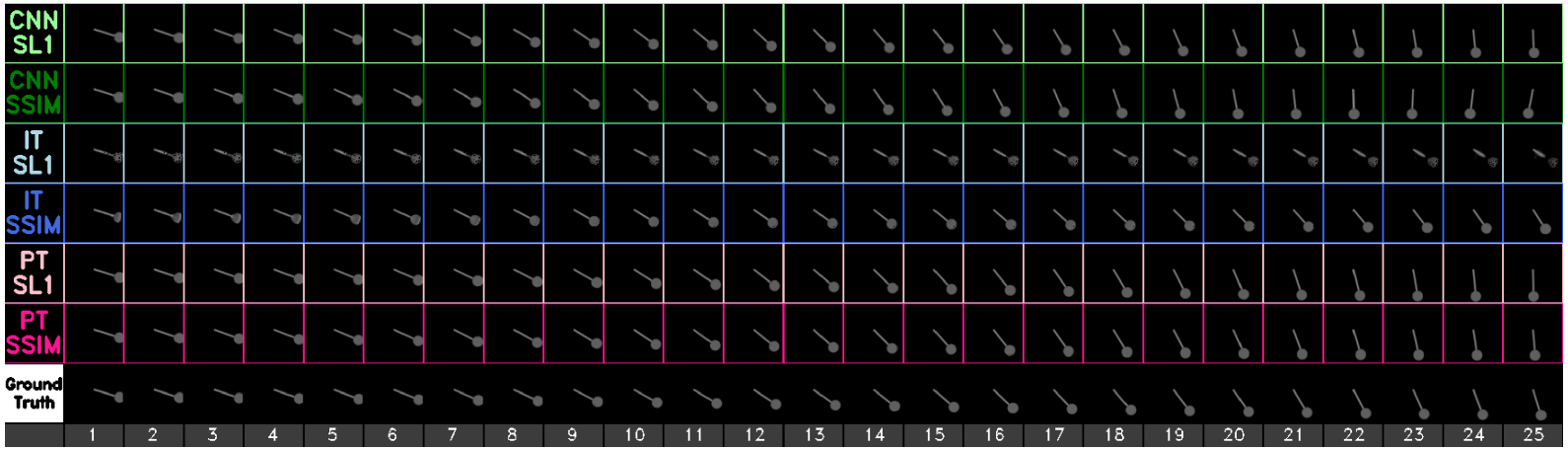
(a) MOCAP



(b) 3D Bouncing

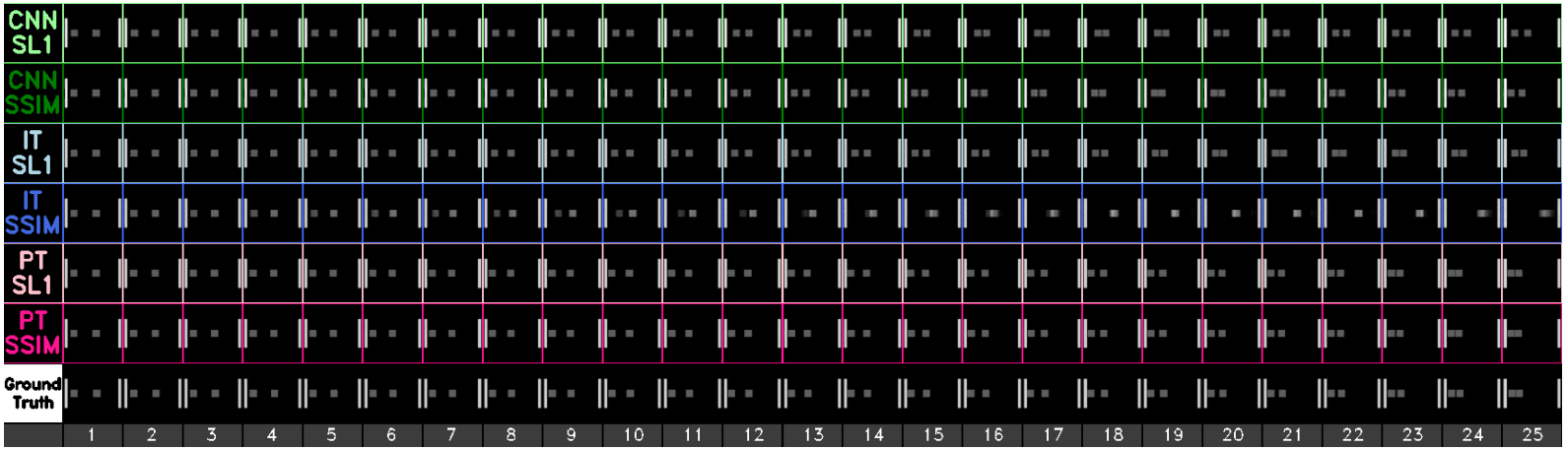
Figure 5.9: Comparison of the first 25 generated frames of each model ($m = 5$) across the datasets vs the ground truth.

(a) Roller

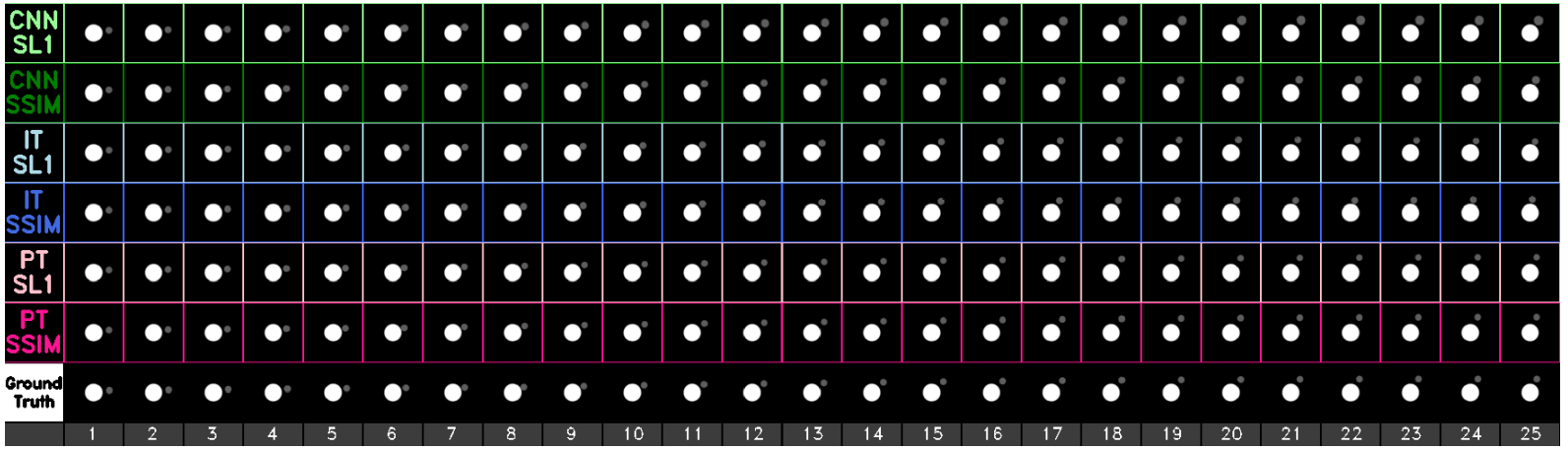


(b) Pendulum

Figure 5.10: Comparison of the first 25 generated frames of each model ($m = 5$) across the datasets vs the ground truth.

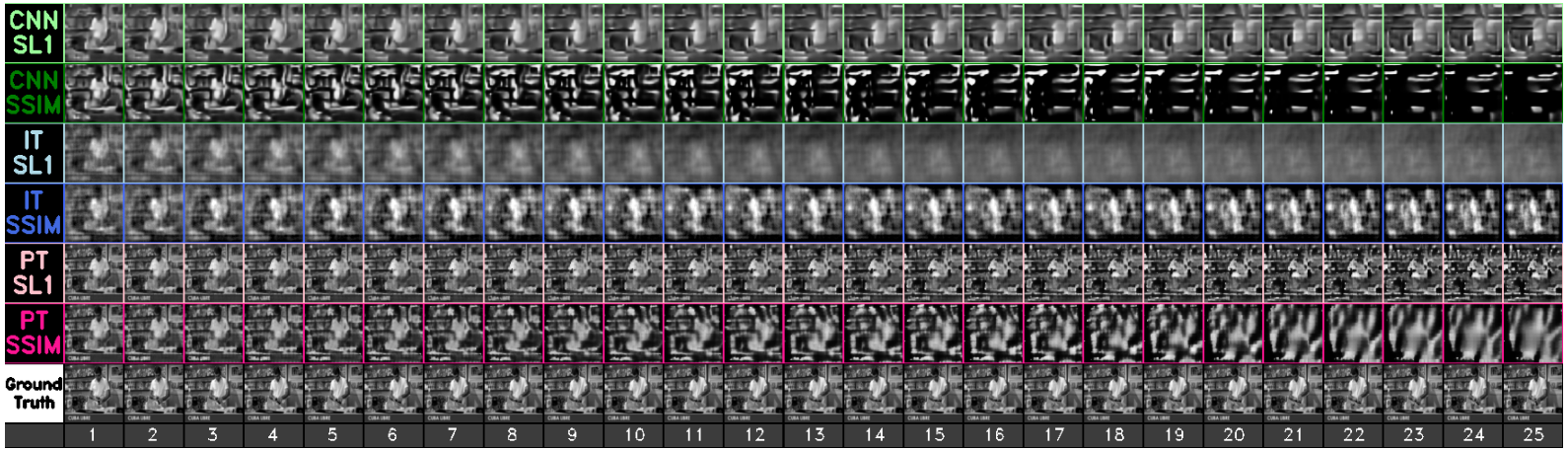


(a) Blocks



(b) Moon

Figure 5.11: Comparison of the first 25 generated frames of each model ($m = 5$) across the datasets vs the ground truth.



(a) HMDB-51

Figure 5.12: Comparison of the first 25 generated frames of each model ($m = 5$) across the datasets vs the ground truth.

5.6.2.1 Self-Output By Models

The patch transformer generates the most physically accurate long-term predictions (*i.e.* more closely obeys the laws underpinning the video) on the 3D bouncing, roller, pendulum, blocks, and moon datasets. The CNN model performs best on the 2D bouncing (Figure 5.8a) and MMNIST (Figure 5.8b) datasets. The image transformer is less physically accurate than both of the other models on all datasets. The CNN struggles to smooth out lower resolution graininess. There are occasional oddities with the CNN that imply it is overly relying on local spatial information *e.g.* a ball will sometimes accelerate off the screen in the roller dataset. I speculate that this is because the model has learned that a ball above a rail at a certain angle should be moving upwards. Although it is thought that pure CNN architectures struggle to properly model inter-frame variations in video sequences [150], my CNN model's self-output videos challenge this assumption by demonstrating an understanding of gravity and collision physics (Figure 5.8a). The image transformer struggles to generalise to different shapes. This can be observed in MMNIST where digits degrade immediately within the first 5 frames (Figure 5.8b). Furthermore, the image transformer often accumulates black 'dead' pixels. Though the patch transformer appears to be the best model, it too demonstrates architecture specific artefacts (*e.g.* the resolution of the patches forming the outputs are visible in some examples from MMNIST).

5.6.2.2 Self-Output By Loss

When any of the 3 models are trained with SSIM loss, their predictions eventually begin to distort the otherwise constant background colour. This can be seen in the SSIM rows of Figure 5.8a as the background shifts from static white to grey. This phenomenon implies a gradual buildup of errors in the background which I argue is due to the insensitivity of SSIM function to changes in the background, and thus the model is not forced to carefully maintain the background. Such background distortion does not happen in SL1 outputs, indicating that the pixelwise-SL1 loss is particularly sensitive to these artifacts and prioritises minimising them. My self-output results challenge assumptions about the effects of different types of loss

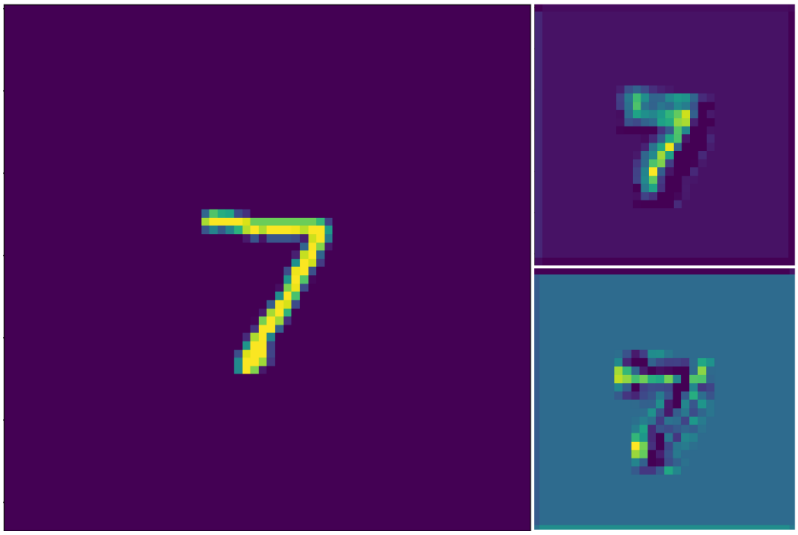
functions on generated videos. In their recent survey paper of video generation techniques, Oprea et al. [150] argue that SL1 is a 'deterministic' pixelwise loss that learns to accommodate uncertainty by blurring its predictions. It is argued that this allows "plausible predictions in deterministic scenarios" (*i.e.* synthetic datasets), and yet struggle with more uncertain video data. However, my self-output videos demonstrate immediate and substantial blurring on the realistic MOCAP data for *both* the SL1 and SSIM trained models (Figure 5.9a). Furthermore, the simpler simulation data (MMNIST in Figure 5.8b, 2D bouncing in Figure 5.8a) shows almost no blurring for *either* loss function. These results imply that, in some cases, blurring can be caused by the difficulty of the dataset and not inherent properties of the SL1 loss *i.e.* if the model/loss function is unable to induce understanding about a specific movement, it may instead be optimised by smoothing that region into the average pixel value. Despite the 'synthetic' MMNIST dataset scoring lower PSNR, $L1$ and SSIM metrics compared to the real-world MOCAP for *first-frame prediction* (discussed in Section 5.6.1.1), the superior *long-term prediction stability* of MMNIST compared to MOCAP stresses the importance of complementing frame-to-frame metrics and loss scores with qualitative long-term prediction comparisons.

5.6.3 Test-Task Performance

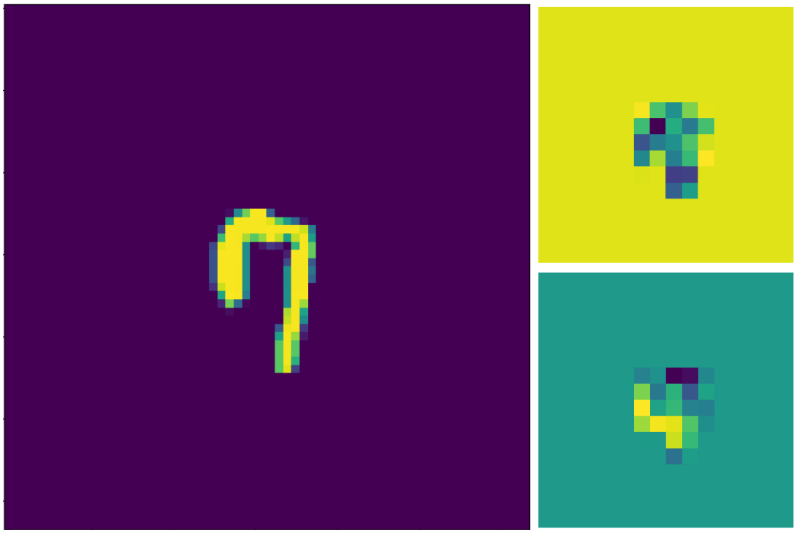
To make performance between my experiments more comparable, I normalise the values for all predictions (excluding MNIST classification) such that the dataset has a standard deviation of 1. In the following subsections, I will discuss the results of the 4 variations of my downstream test-task experiments: random scores as lower bounds on performance, models trained on test-tasks *without* pretraining, linear probing for frozen models *with* pretraining, and models trained on test-tasks *with* pretraining (*i.e. finetuning*).

5.6.3.1 Random Lower Bound

The results for random and frozen models may initially appear too strong to be random baselines (*i.e.* each 3rd row of Tables 5.3 and 5.4; 4th row for MNIST only). However, I note that random and frozen Conv2D PyTorch layers allow through a



(a)



(b)

Figure 5.13: Matplotlib visualisation demonstrating that randomly initiated pytorch Conv2D layers (with bias) can allow substantial image information to leak through. Left: A visualisation of a 64x64 grayscale image from the MNIST dataset. Right: Visualisations of 2 different channels (16x16) from the outputs of the first layer in a randomly instantiated patch transformer/CNN model.

substantial amount of image information, even with trainable biases. This means substantial image information can still bleed through to the trainable linear probe layers as visualised in the MNIST examples in Figures 5.13a and 5.13b. I find this induces strong performance in the random frozen models, in particular scoring 98.58% accuracy on MNIST. This limits its use as a lower bound of task performance for the CNN and patch transformer. However as the image transformer does not

have Conv2D layers, it does not suffer such information leakage. This allows the random and frozen results from the image transformer to reasonably serve as an indicator for random task performance. For this reason, I specifically refer to the *frozen random score for the Image Transformer* when I discuss 'random' scores in the following sections (*i.e.* 3rd row of subsections of Tables 5.3 and 5.4; 4th row for MNIST).

5.6.3.2 Unfrozen Without Pretraining + Non-Linear MLP

The unfrozen model experiments with non-linear fully connected units serving as classifiers (see Figure 5.6c) are designed to show the potential of the models to learn visual laws directly from the task itself. The results of these experiments can be seen as the last grey row of each section in Tables 5.3 and 5.4. All 3 of the models can learn each task noticeably better than random (except the image transformer on 3D bouncing and blocks). The tasks with the best performance are 2D bounce counting, 2D gravity prediction, and roller gravity prediction. As these tasks all represent gravity working directly on 2D balls in freefall, I believe this style of simulation is the easiest to learn for these models. The 2D bounce counting task performance implies these models are capable of effectively tracking and counting collisions throughout the input frames. I note in particular the improved scores of the 3 following tasks. The bounce counting task for 2D bouncing scores 0.09786, which is a drastic improvement compared to its random score of 0.6457. The 2D bouncing gravity prediction task improves even further, scoring 0.02402 from a random score of 0.6442. The roller gravity prediction improves from a random score of 0.8393 down to 0.0722. The patch transformer is strongest model on 2D bounce prediction and roller gravity prediction by a margin of $\sim$0.01-0.03. However, the CNN scores about half the patch transformer's loss on 2D bounce prediction. As MNIST classification is an easy vision dataset, it is unsurprising that each model scores above 99%. The moon and pendulum prediction tasks appear to be the most difficult, improving from random scores of $\sim$0.89 to 0.325 and 0.2266 respectively. Though the patch transformer is most often the best scoring model, the CNN scores are almost as good. The CNN scores slightly better on the 3D bounce prediction task ($\sim$0.01), and

Model Details		Baseline	CNN		Image Transformer		Patch Transformer	
Pretraining	Task	Task L1	Modelling Loss	Task L1	Modelling Loss	Task L1	Modelling Loss	Task L1
Modelling Dataset = 2D Bounce	**Task = Counting Bounces in 59 Frames**							
-	Constant Output	0.6525	-	-	-	-	-	-
-	Input Image + Linear Layer	0.7077	-	-	-	-	-	-
None	Frozen Model + Linear Probes	-	-	0.5193	-	0.6457	-	0.6070
None	Unfrozen Model + Non-Linear MLP	-	-	0.1007	-	0.2080	-	9.786e-2
$m = 59$ SL1	Frozen Model + Linear Probes	-	5.003e-3	0.2830	1.207e-2	0.6569	1.557e-2	0.5532
$m = 59$ SSIM	Frozen Model + Linear Probes	-	1.797e-2	0.2943	5.057e-2	0.6563	2.328e-2	0.3839
$m = 59$ SL1	Unfrozen Model + Non-Linear MLP	-	5.003e-3	0.1247	1.207e-2	0.2148	1.557e-2	0.1037
$m = 59$ SSIM	Unfrozen Model + Non-Linear MLP	-	1.797e-2	0.1165	5.057e-2	0.2737	2.328e-2	0.1008
Modelling Dataset = 2D Bounce	**Task = Gravity, from 5 frames**							
-	Constant Output	0.8676	-	-	-	-	-	-
-	Input Image + Linear Layer	0.8687	-	-	-	-	-	-
None	Frozen Model + Linear Probes	-	-	0.4272	-	0.6442	-	0.6076
None	Unfrozen Model + Non-Linear MLP	-	-	1.269e-2	-	7.697e-2	-	2.402e-2
$m = 5$ SL1	Frozen Model + Linear Probes	-	3.401e-3	0.1621	6.529e-3	0.7533	3.439e-3	0.4789
$m = 5$ SSIM	Frozen Model + Linear Probes	-	1.167e-2	0.1390	2.558e-2	0.7605	1.062e-2	0.3440
$m = 5$ SL1	Unfrozen Model + Non-Linear MLP	-	3.401e-3	1.891e-2	6.529e-3	7.146e-2	3.439e-3	1.918e-2
$m = 5$ SSIM	Unfrozen Model + Non-Linear MLP	-	1.167e-2	1.774e-2	2.558e-2	5.960e-2	1.062e-2	1.385e-2
Modelling Dataset = 3D Bounce	**Task = Counting Bounces in 99 frames**							
-	Constant Output	0.6651	-	-	-	-	-	-
-	Input Image + Linear Layer	0.4255	-	-	-	-	-	-
None	Frozen Model + Linear Probes	-	-	0.3756	-	0.6420	-	0.4832
None	Unfrozen Model + Non-Linear MLP	-	-	0.2662	-	0.6441	-	0.2713
$m = 99$ SL1	Frozen Model + Linear Probes	-	3.905e-3	0.3200	2.398e-3	0.5982	7.494e-4	0.3279
$m = 99$ SSIM	Frozen Model + Linear Probes	-	1.439e-2	0.3198	1.112e-2	0.5721	3.713e-3	0.3414
$m = 99$ SL1	Unfrozen Model + Non-Linear MLP	-	3.905e-3	0.2952	2.398e-3	0.4276	7.494e-4	0.2273
$m = 99$ SSIM	Unfrozen Model + Non-Linear MLP	-	1.439e-2	0.2711	1.112e-2	0.4897	3.713e-3	0.2555
Modelling Dataset = Roller	**Task = Gravity, from 5 frames**							
-	Constant Output	0.8645	-	-	-	-	-	-
-	Input Image + Linear Layer	0.8199	-	-	-	-	-	-
None	Frozen Model + Linear Probes	-	-	0.4305	-	0.8393	-	0.4699
None	Unfrozen Model + Non-Linear MLP	-	-	0.1034	-	0.1679	-	7.220e-2
$m = 5$ SL1	Frozen Model + Linear Probes	-	2.326e-3	0.1966	9.248e-4	0.8383	1.150e-4	0.3948
$m = 5$ SSIM	Frozen Model + Linear Probes	-	2.489e-3	0.1765	2.573e-3	0.8278	8.672e-5	0.2750
$m = 5$ SL1	Unfrozen Model + Non-Linear MLP	-	2.326e-3	0.1016	9.248e-4	0.1318	1.150e-4	8.844e-2
$m = 5$ SSIM	Unfrozen Model + Non-Linear MLP	-	2.489e-3	0.1018	2.573e-3	0.1279	8.672e-5	7.790e-2
Modelling Dataset = Pendulum	**Task = Gravity, from 5 frames**							
-	Constant Output	0.8878	-	-	-	-	-	-
-	Input Image + Linear Layer	0.8903	-	-	-	-	-	-
None	Frozen Model + Linear Probes	-	-	0.7194	-	0.8914	-	0.7238
None	Unfrozen Model + Non-Linear MLP	-	-	0.2837	-	0.3493	-	0.2266
$m = 5$ SL1	Frozen Model + Linear Probes	-	5.564e-4	0.3770	2.934e-3	0.8982	6.726e-5	0.3834
$m = 5$ SSIM	Frozen Model + Linear Probes	-	9.868e-4	0.3757	5.886e-4	0.9021	8.816e-5	0.3731
$m = 5$ SL1	Unfrozen Model + Non-Linear MLP	-	5.564e-4	0.3214	2.934e-3	0.3903	6.726e-5	0.2196
$m = 5$ SSIM	Unfrozen Model + Non-Linear MLP	-	9.868e-4	0.2898	5.886e-4	0.3358	8.816e-5	0.2173
Modelling Dataset = Blocks	**Task = Mass Ratio, from 49 frames**							
-	Constant Output	0.8880	-	-	-	-	-	-
-	Input Image + Linear Layer	0.8755	-	-	-	-	-	-
None	Frozen Model + Linear Probes	-	-	0.6372	-	0.8950	-	0.7296
None	Unfrozen Model + Non-Linear MLP	-	-	0.5571	-	0.7836	-	0.4871
$m = 49$ SL1	Frozen Model + Linear Probes	-	4.541e-3	0.6184	3.344e-4	0.8892	1.038e-4	0.7280
$m = 49$ SSIM	Frozen Model + Linear Probes	-	3.890e-3	0.6077	7.180e-4	0.8915	1.990e-4	0.6783
$m = 49$ SL1	Unfrozen Model + Non-Linear MLP	-	4.541e-3	0.5799	3.344e-4	0.5516	1.038e-4	0.5296
$m = 49$ SSIM	Unfrozen Model + Non-Linear MLP	-	3.890e-3	0.5585	7.180e-4	0.5196	1.990e-4	0.5219

Table 5.3: Performance on test-tasks without vs. with modelling pretraining. Results are generated using checkpoints from the best validation epoch for pretraining and test-task.

Model Details		Baseline	CNN		Image Transformer		Patch Transformer	
Pretraining	Task	Task L1	Modelling Loss	Task L1	Modelling Loss	Task L1	Modelling Loss	Task L1
Modelling Dataset = Moon	**Task = Mass, from 5 frames**							
-	Constant Output	0.8763	-	-	-	-	-	-
-	Input Image + Linear Layer	0.8512	-	-	-	-	-	-
None	Frozen Model + Linear Probes	-	-	0.7009	-	0.8751	-	0.8175
None	Unfrozen Model + Non-Linear MLP	-	-	0.3287	-	0.5829	-	0.3250
$m = 5$ SL1	Frozen Model + Linear Probes	-	4.781e-4	0.382	5.656e-4	0.8136	9.834e-5	0.6579
$m = 5$ SSIM	Frozen Model + Linear Probes	-	5.199e-4	0.4909	2.751e-3	0.8268	9.945e-5	0.6259
$m = 5$ SL1	Unfrozen Model + Non-Linear MLP	-	4.781e-4	0.3552	5.656e-4	0.3127	9.834e-5	0.3324
$m = 5$ SSIM	Unfrozen Model + Non-Linear MLP	-	5.199e-4	0.3822	2.751e-3	0.4246	9.945e-5	0.3528
Modelling Dataset = MMNIST	**Task = MNIST**							
-	Most Common Class	11.28%	-	-	-	-	-	-
-	Constant Output	11.28%	-	-	-	-	-	-
-	Input Image + Linear Layer	93.30%	-	-	-	-	-	-
None	Frozen Model + Linear Probes	-	-	98.58%	-	39.68%	-	97.60%
None	Unfrozen Model + Non-Linear MLP	-	-	99.44%	-	99.00%	-	99.52%
$m = 5$ SL1	Frozen Model + Linear Probes	-	2.828e-2	98.70%	5.549e-2	81.16%	2.732e-2	86.22%
$m = 5$ SSIM	Frozen Model + Linear Probes	-	6.797e-2	98.66%	1.616e-1	77.76%	6.501e-2	87.24%
$m = 5$ SL1	Unfrozen Model + Non-Linear MLP	-	2.828e-2	99.16%	5.549e-2	99.10%	2.732e-2	99.56%
$m = 5$ SSIM	Unfrozen Model + Non-Linear MLP	-	6.797e-2	99.42%	1.616e-1	99.02%	6.501e-2	99.56%

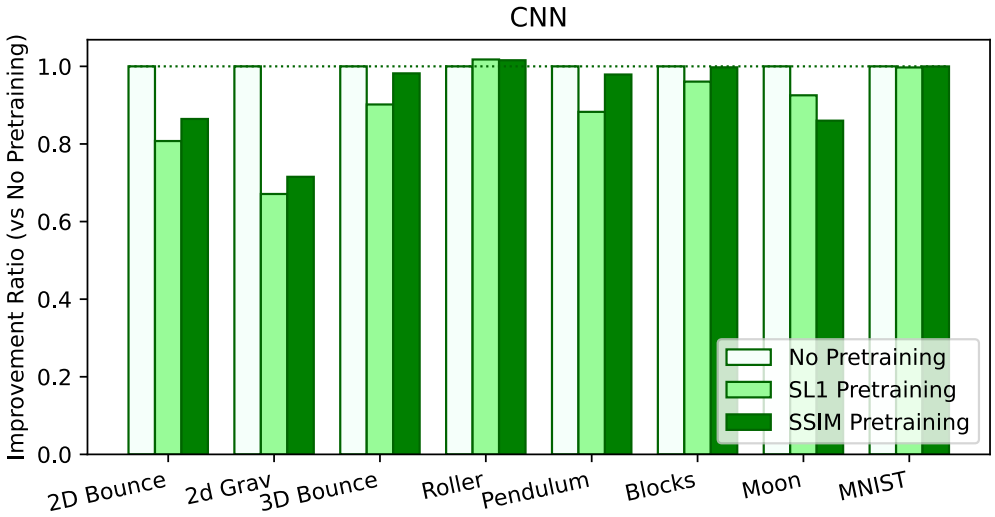
Table 5.4: Performance on test-tasks without vs. with modelling pretraining. Results are generated using checkpoints from the best validation epoch for pretraining and test-task.

substantially better on 2D gravity prediction as previously mentioned. The image transformer is substantially worse than both the CNN and patch transformer on every non-MNIST task, which stresses the strengths of convolutional based models in learning these test-tasks *without* pretraining.

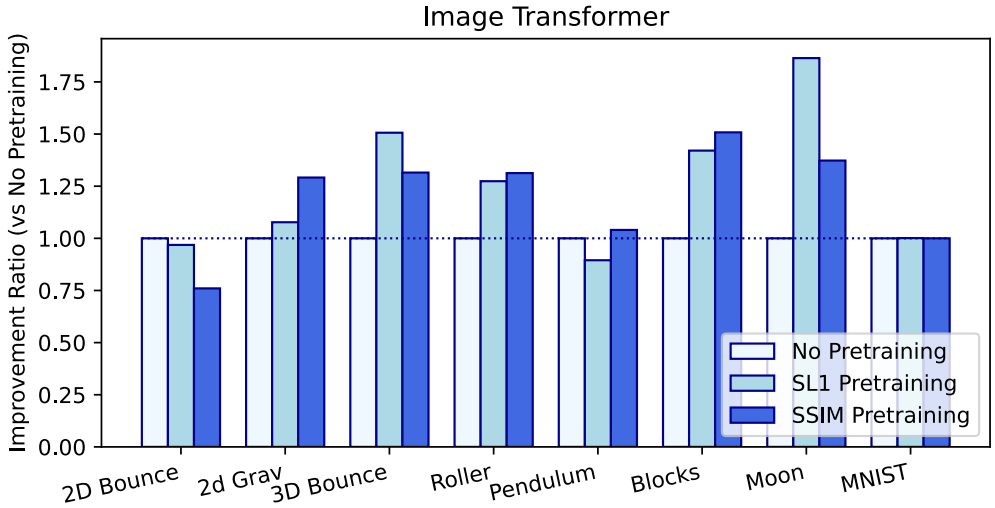
5.6.3.3 Linear Probes

The linear probing experiments (see Figure 5.6b) with **frozen** *and* **pretrained** models are designed to verify if modelling pretraining induces a transferable understanding of visual laws for downstream test-tasks (*i.e.* the top 2 white highlighted rows in each section of Tables 5.3 and 5.4). The CNN and patch transformer test-tasks on pretrained models consistently score better than random, yet do not perform near as well as the '*without* pretraining and unfrozen' experiments described in the previous subsection. The CNN model now outperforms the patch transformer on the the previous subsection's best performing experiments: 2D bounce prediction, 2D gravity prediction, and roller gravity prediction with ∼0.29 (CNN) vs. 0.553/0.3839 (PT), ∼0.15 (CNN) vs. 0.34/0.4789 (PT), ∼0.18 (CNN) vs. 0.395/0.2750 (PT) for each task respectively. However, I **cannot** directly conclude that all this information has been induced by pretraining because the leakage of image information through

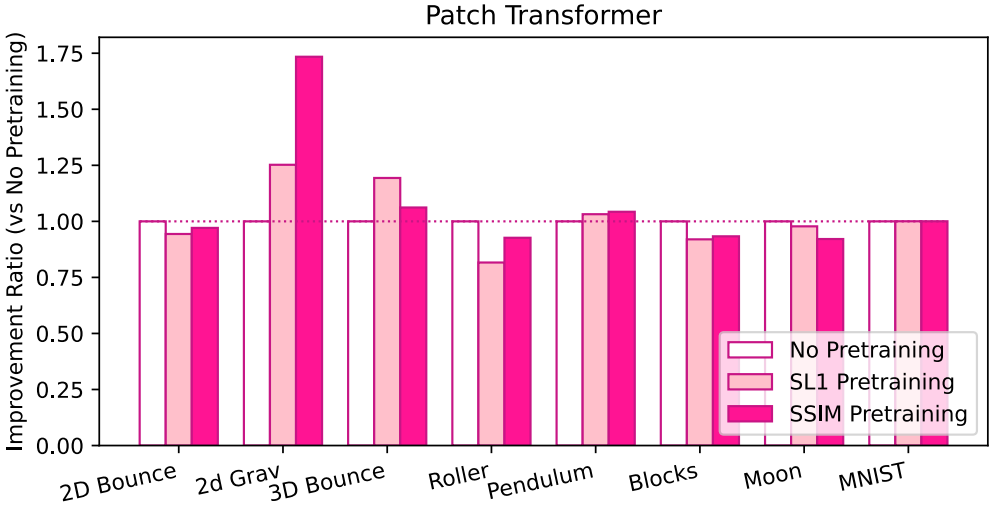
random and frozen Conv2D layers demonstrates that the convolution based models feed enough information to linear probes to achieve good scores. This highlights a unique difficulty in probing frozen convolutional models. Despite this information leakage, I can still verify the use of pretraining through linear probes on the CNN and patch transformer by comparing their probing scores with 'input image+linear layer' scores (see the second grey row of each section on Tables 5.3 and 5.4). This 'input image+linear layer' scenario shows how highly a trainable linear layer can score on the test-tasks with the images as input. Given that image information can leak through the frozen convolutions of the CNN and patch transformers and into the trainable linear probes, the 'input image+linear layer' experiment scores provide an upper bound to the performance the linear probes could have gained using *only* the leaked image information. I argue that the 'input image+linear layer' scores *are* upper bounds as *all* image information is provided, and images leaking through convolution layers will be at least partially distorted. As the probing scores are *still* much better than 'input image+linear layer' scores, this implies that modelling pretraining alone has encoded some understanding that is useful to my downstream test-tasks. Despite this, the linear probing models do not match 'unfrozen+without pretraining' performance, showing that although pretraining does induce some detectable understanding, it is still more useful to directly learn the task. As the image transformer does not suffer from random Conv2D leakage, it can more accurately demonstrate the induced understanding from pretraining. The image transformer is however the worst performing model on all tasks by a large margin, and its frozen probing tasks fail to exceed random scores on all but 3D bounce prediction and moon mass prediction (both which only improve by a 0.05). There are few noticeable differences between SL1 and SSIM pretraining for the CNN. SSIM is marginally better on 2D bouncing gravity prediction (by 0.03), and SL1 is 0.11 better for the moon task. The patch transformer however demonstrates improved performance for SSIM pretraining on all but the 3D bouncing bounce counting task, improving by between (0.03-0.15).



(a) CNN



(b) Image Transformer



(c) Patch Transformer

Figure 5.14: The improvement ratio in scores (losses) for each task when pretrained on visual modelling (vs. *without* pretraining) for each model architecture.

5.6.3.4 Unfrozen With Pretraining + Non-Linear MLP (Finetuning)

My 'finetuning' experiments are designed to verify if visual modelling pretraining can give an advantageous starting point for learning downstream vision tasks (vs. *without* pretraining). The results of these experiments are on the last 2 rows for each section on Tables 5.3 and 5.4. The results of directly training on tasks *with* versus *without* pretraining are visualised in Figure 5.14. I find that the CNN noticeably degrades in performance from both SL1 and SSIM pretraining in 3 of 8 tasks. The CNN degrades most substantially on the 2D bouncing gravity and 2D bounce counting tasks, with an improvement ratio of 0.8 and 0.7 respectively (Figure 5.14a). Overall, the CNN fails to benefit from pretraining in its performance on the downstream test-tasks. The image transformer however improves substantially on 5 of the 8 tasks when finetuning a modelling-pretrained model (Figure 5.14b), with improvement ratios varying between 1.3 to as high as 1.8. Visual modelling pretraining for the patch transformer is almost as good or substantially better compared to no pretraining for all tasks but roller gravity prediction (Figure 5.14c). Interestingly, the image and patch transformer models show the same preference for pretraining loss for each task. SSIM pretraining consistently gives higher scores compared to SL1 for 2D bouncing, 2D gravity, roller, pendulum, and blocks datasets. Conversely, SL1 outperforms SSIM for the 3D bouncing and moon datasets. When I consider the best performing model for each task (Figure 5.2), *i.e.* either the image or patch transformer, I find that pretraining on modelling tasks is around as good, or considerably better for downstream task performance for 7 of the experimented 8 tasks (improving by a ratio of 1.25 to 1.8) demonstrating the potential of the generative pretraining visual modelling paradigm.

5.6.4 General Discussion

I focus specifically on *pairing* the visual pretraining datasets with *quantifiable* and *observable* test-tasks and analysis, allowing me to strongly argue for the benefits of pretraining that my experiments demonstrate. My finetuning experiments show that visual modelling pretraining can yield substantially improved performance on

test-tasks while not impacting performance on the majority of other tasks that do not benefit from pretraining. This shows that there can be both little risk and large potential gain in generative pretraining for downstream vision tasks. The image and patch transformer architectures benefit most substantially from visual modelling pretraining. However, the CNN did not improve with pretraining on any of the test-tasks. This may imply that there is some inherent property in a transformer that the CNN lacks (*e.g.* self-attention) that can successfully facilitate such pretraining. When finetuning from pretraining, the image transformer scores the highest on the blocks and moon test-tasks, with the patch transformer performing best on the other 6 tasks. When considered alongside the first-frame prediction and self-output experiments, I find that the patch transformer performs best with visual modelling pretraining, and the image transformer benefits most substantially. Though it would be thematically fitting that a convolutional adaption of a transformer architecture would best serve the overlap of vision and language modelling, I cannot directly conclude this from my results. Future research focusing more specifically on architecture would help verify my findings. Further improving the size and scale of visual modelling pretraining is a promising path to realise even greater performance increase. Chen et al. [28] demonstrate that by leveraging very large computational resources (a model comparable to GPT-2 *i.e.* 48 layers and ∼1.4B parameters), unsupervised image pretraining leads to state-of-the-art downstream ImageNet and CIFAR-101 performance *that increases with model scale*. This trend of increasing downstream task performance as model and data scale increases highlights how promising both visual modelling pretraining and the datasets that can facilitate it can be to the future of downstream vision tasks.

5.6.5 Not as Directly Applicable to Multimodal Processing

The work in this chapter represents an attempt to avoid the problems of language bias in multimodality entirely and simply focus on improving methodology from a different perspective: 'improving the methodology behind visual features, and therefore their relative strength'. Though this distance from language bias has unshackled me from similar limitations as in Chapter 4 and I have demonstrated

positive experimental results, it is still the case however that the findings in this chapter are currently *not as directly* applicable to multimodal processing as my other contributions. This is not necessarily problematic, as the *goal* is to one day apply my visual modelling methodology to use *e.g.* 'improved' ImageNet features for multimodal processing. Nevertheless, there is still quite a long way to push this visual modelling paradigm —evaluated here on simpler and *controlled* visual dynamics— before it can be directly applied to the more complex behaviours and scenarios underpinning the larger-scale VQA and video-QA datasets. In order to be applicable to such VQA benchmarks, the datasets used for visual pretraining should contain as many different visual scenarios as training resources can allow (*e.g.* they should demonstrate colour, gravity, human interactions with objects and the environment etc...). Though training on such datasets would be orders-of-magnitude more expensive than on those in this chapter, it would be necessary to at least match the variety and complexity of the visual components of VQA datasets. Given sufficient resources to collect datasets at a similar scale to those used in the training of DALL-E 2[8], a VQA-compatible visual pretraining dataset could be collected by focusing on gathering captioned *GIFs and videos* (instead of just captioned images). Such a dataset would allow visual pretraining with expansive visual variety applicable to many smaller benchmark VQA datasets, and would itself be naturally paired with its own QA dataset that could be used to test the visual modelling hypothesis.

5.7 Conclusion

I explore the visual parallel to the powerful predictive language modelling paradigm, which I dub 'visual modelling'. I introduce 6 dynamic simulation video datasets that also function as physical property prediction datasets with a total of 7 test-tasks. In line with the theme of this chapter, I experiment with 3 models inspired from the fields of both vision and NLP. My modelling results highlight the importance of convolutional models in video generation. The patch transformer yielding the best

[8] https://openai.com/dall-e-2/

performance may imply a combination of transformer and convolutional designs is optimal for visual modelling. My self-output experiments show that current video generation models can sustain reasonable predictions over many frames into the future when datasets are sufficiently simplistic enough for the model to learn the physical laws underpinning them. My test-task experiments demonstrate the potential performance increases visual modelling pretraining can bring to vision. I show that visual modelling pretraining induces results on the downstream test-tasks that are as good or significantly better when compared to no pretraining for 7 of the 8 test-tasks. My work in this chapter represents an encouraging step towards 'improving the power of vision', however the visual modelling methodology will need further advancement before it is ready for *direct* application to state-of-the-art multimodal benchmarks.

CHAPTER 6

Neurolinguistic Multiclass Labelling: Are Human Measures of Similarity Suitable for VQA?

6.1 Introduction

The preceding chapters represent my contributions to the problems of language bias, multimodal processing, and the 'strengthening' of the exploitation of visual information. Though I have aimed to ensure that my contributions can 'scalably' apply to future datasets and models, they each contain drawbacks that limits their applicability to multimodal research *as it is right now*: The modality-subset methodology in Chapter 3 yields subsets an order-of-magnitude smaller than the original, leaving one vulnerable to the problems of small datasets in deep learning. The work to improve the relative 'strength' of vision compared to language through visual modelling in Chapter 5 is not yet at a stage to be immediately applicable to the multimodal benchmarks of our time. As such, for my final major project, I am therefore motivated to contribute to a part of the machine learning pipeline common across multimodal benchmarks regardless of dataset and model, but in a manner such that I can best adhere to the lessons from my previous chapters. **I argue that the labelling scheme used for multimodal tasks evaluated on VQA is a**

reasonable candidate for this objective as:

1. Multimodal tasks share a structure of 'predicting candidates from a vocabulary' regardless of if it happens in a single step (typical of the answer schemes of VQA datasets) or in multiple steps (for generative open-ended multimodal models that propose individual words one at a time), *i.e.* invariant to specific datasets and therefore applicable to future datasets.

2. Returning to multimodal question-answering tasks makes the methodology *immediately* applicable to multimodal processing, *i.e.* learning from Chapter 5.

3. A change to labelling would ideally mean no need to discard question-answer pairs as in Chapter 3.

4. Evaluating this methodology on VQA is ideal as state-of-the-art VQA datasets have gone further than video-QA datasets to control for language bias (as detailed in Chapter 2), *i.e.* learning from Chapter 4.

Furthermore, answering schemes —and therefore their respective loss functions— typically used in multimodal QA tasks seem to encourage a worrying oversimplification of reality by assuming one correct answer, *and functionally treating all other answers as equally incorrect* as visualised in Figure 6.1. Given that the consequences of this 'one-hot' limitation in multimodal labelling schemes are also surprisingly under-explored, it would represent an ideal research question for this thesis. However, despite the same probabilistic cross-entropy answering style being used across different datasets, the *answer vocabularies themselves* are different *i.e. different words in the answer vocabularies*. The aims of this chapter's contribution are therefore:

1. To develop a new labelling scheme.

2. A labelling scheme that induces a more 'realistic' understanding of the world than the scenario in Figure 6.1.

Figure 6.1: The limitations of typical answer schemes for VQA datasets.

3. To develop such a labelling scheme in a way that is invariant to the different answer vocabularies across datasets.

To achieve this, I propose augmenting VQA labelling schemes using numerical scores of similarity between words provided by neurolinguistic study, drawing from ‘dual coding theory’ discussed at the end of Chapter 4. For the typical ‘one-hot’ labelling scheme used in VQA tasks, the ground truth has a score of 1, and all other answers are treated as equally incorrect at 0. Instead, I propose increasing the label of specially selected *incorrect* answers to a score $\in (0, 1)$ based on how neurolinguistically ‘close’ it is to the *correct* answer (see Figure 6.2). In theory, this multiclass neurolinguistic answer scheme satisfies the aims of this chapter by:

1. Inducing a more realistic model of the world than the arguably over-simplistic ‘one-hot’ scenario;

2. Because the scores of neurolinguistic ‘closeness’ are for mutually exclusive pairs of words, and are assumed to be ‘human ground truths’, they can be applied to ***any*** VQA answer vocabulary (provided that it contains at least one of the word pairs that the ‘closeness’ score).

One might expect such an answer scheme to induce a more desirable and ‘better’ understanding of the world, potentially evidenced by some improved accuracy on the

VQA tasks as a sufficient proxy to real-world understanding. However, my results in this chapter show that this answering scheme in practice significantly *reduces* accuracy the vast majority of the time for all 5 VQA datasets —VQA v1, VQA v2, VQA-CP v1, VQA-CP v2, and GQA— and both VQA models —LXMERT [183], and METER [51]— in my experiments. I cannot immediately conclude that this decrease in VQA accuracy implies that the proposed labelling scheme has failed to induce either a more 'realistic' or desirable understanding of the world. The problem may instead lie with the VQA datasets themselves being an unsuitable test of 'realistic' understanding (*e.g.* dataset biases), or the metric of VQA accuracy itself being a poor measure of the potentially improved understanding induced. Though it has been shown that VQA datasets are not always a good representation of 'realistic' or 'desirable' behaviour, (see Figure 6.2, and Chapter 2), it is prudent to suspect that there are also non-trivial problems with my proposed approach. As such, this chapter represents a thorough exploration of the hypothesised labelling scheme across 2 VQA benchmark models and 5 benchmark datasets. I aim to identify and test potential causes of these initial negative results. Most notably:

1. I find that despite the large quantity of neurolinguistic word-pair scores I use, and the high number of words in dataset vocabularies with *concreteness scores* (as in Figure 4.12, repeated for convenience in Figure 6.5) in practice there are relatively few word pairs with scores in the answer vocabularies of the datasets I use, leaving my labelling scheme relatively sparse.

2. I find that the VQA accuracy is not quite as significantly impacted using my proposed loss function when using the *full* VQA datasets, which I suspect is caused by the aforementioned sparseness.

3. To circumvent the limitations of my 'sparse' labelling scheme, I isolate and experiment on subsets of the 5 VQA datasets that *only contain answers I have at least one score for* in order to more effectively test the loss function by 'concentrating' the dataset. I find that these more 'concentrated' datasets experience a significant drop in accuracy in almost every scenario.

4. As the 'concentrated' dataset splits (designed to verify the effectiveness of my

approach) are much smaller than the original splits, I 'expand' them by allowing word pair scores using another —less ideal— measure of neurolinguistic 'closeness'. I find that the drop in VQA performance persists in the larger 'expanded' splits.

5. I look beyond the simple VQA accuracy metric to detect potential benefits of my approach by considering top-2, top-3, top-5, and top-10 accuracy. Though it might be intuitively expected that a stronger understanding of the similarities of concepts in VQA may yield improved top-k accuracy through 'improved second guesses', in practice I find my labelling scheme *also* reduces top-k accuracies when compared to the typical one-hot labelling scheme.

6. Given the successes of parallel work using multiclass labelling for VQA [103], I conclude that the approach for multiclass labelling is itself promising, but that there is some disconnect between the neurolinguistic scores I use and the semantic space of the VQA labels in the 5 datasets I use.

I conclude this chapter with discussion offering further insight into my results. My code and appropriate directions to the relevant datasets are available at my github repository https://github.com/Jumperkables/a_vs_c.

6.2 Neurolinguistic Multiclass Labelling Scheme

The measures of neurolinguistic 'closeness' I use in this labelling scheme emulate measures 'categorical similarity' and 'association' discussed in the neurolinguistic model of 'dual coding theory'. A key insight from dual coding theory is the different

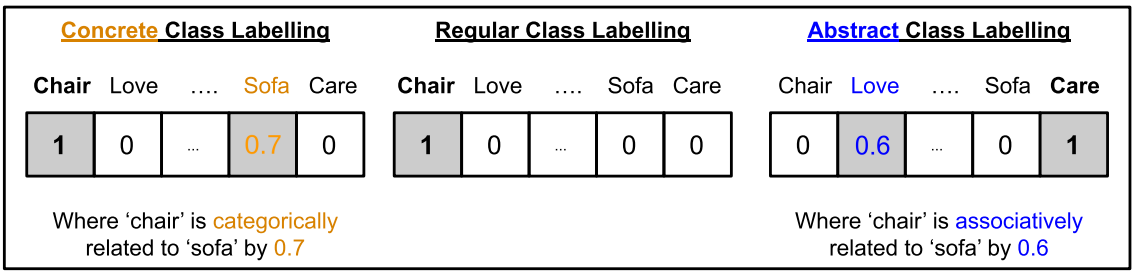


Figure 6.2: The proposed neurolinguistically-guided labelling scheme.

ways 'abstract' and 'concrete' concepts (non-imageable and imageable) are stored and accessed by the brain, and the differences in cognitive processing (*i.e.* free-recall) associated with either concept type. Most interestingly, Crutch and Warrington [35] find evidence that abstract and concrete concepts are stored in structurally different ways (irrespective of the type of information). Concrete concepts appear to be categorically organised *i.e.* stored in more rigid, semantically related networks. For example, a chair is categorically similar to a sofa in that they are both objects for sitting. In contrast, abstract concepts are represented in associative frameworks, near other concepts associated with it (but not necessarily similar in meaning). For examples, 'justice' and 'illegal' may be associated, but are not categorically similar in the same way that concrete concepts are. My experiments use both these similarity scores —gathered from human participants from previous neurolinguistic studies— in an effort to encourage more human-like behaviour. The hypothesis I test in this chapter (see Figure 6.2) considers the idea that more human-like behaviour is a desirable property to have for this VQA scenario. However, it is crucial to emphasise that even the most modern artificial neural networks used in deep learning models remain comparatively simple compared to sheer complexity of biological neural structures that cause human behaviours. Though my hypothesised labelling scheme may make intuitive sense, I stress that it is not currently self evident that it is beneficial or even desirable to emulate human neurology in the confines of current artificial neural networks.

6.2.1 Concreteness: How to Select Either Categorical or Associative Similarity

For a given VQA question-answer pair, my proposed labelling scheme determines which of the 2 potential similarity scores to use in populating the answer tensor by measuring how neurolinguistically 'concrete' the answer is. Concreteness (Figure 6.3) and abstractness (Figure 6.4) scores are on the same continuous spectrum, *i.e.* a lack of concreteness being referred to as abstractness. Note that . I collect neurolinguistic scores from the following datasets: MT40k [18], MRC [203], USF [147], Clark-Paivio [33], Glasgow Word Norms [169], Ljubešić et al. [132], Yee [221],

Sianipar et al. [177], Reilly and Kean [165], Toronto Word Pool [57], and SimLex-999 [85]. The scores in these datasets do not share a standardised scale and are instead mostly defined on various discrete likert scales *e.g.* 1-7 or 1-5. As such, I rescale the scores from each of these datasets individually such that the concreteness score $\in [0, 1]$, 1 being the highly concrete and 0 being highly abstract. For answers that have multiple scores from different dataset, I use the average of each of the rescaled scores. Naturally, there are answers in the VQA datasets that are too specific and esoteric to have had a concreteness score *e.g.* 'hazardous materials prohibited'. The vast majority of such answers *also* do not have either associative or categorical similarity scores, so words without scores default to the typical 'one-hot' scenario for their ground truth. Figure 6.5 (also discussed earlier in Chapter 4) visualises the relative abundance of concreteness scores to be exploited in video-QA datasets vocabularies along with those of VQA datasets that are the focus of this chapter.

Figure 6.3: Wordcloud of the most concrete words from across each of the various collected word norm datasets. Words displayed have concreteness > 0.95 from a scale of 0-1.

Figure 6.4: Wordcloud of the most abstract words (least concrete) from across each of the various collected word norm datasets. Words displayed have concreteness $<$ 0.05 from a scale of 0-1.

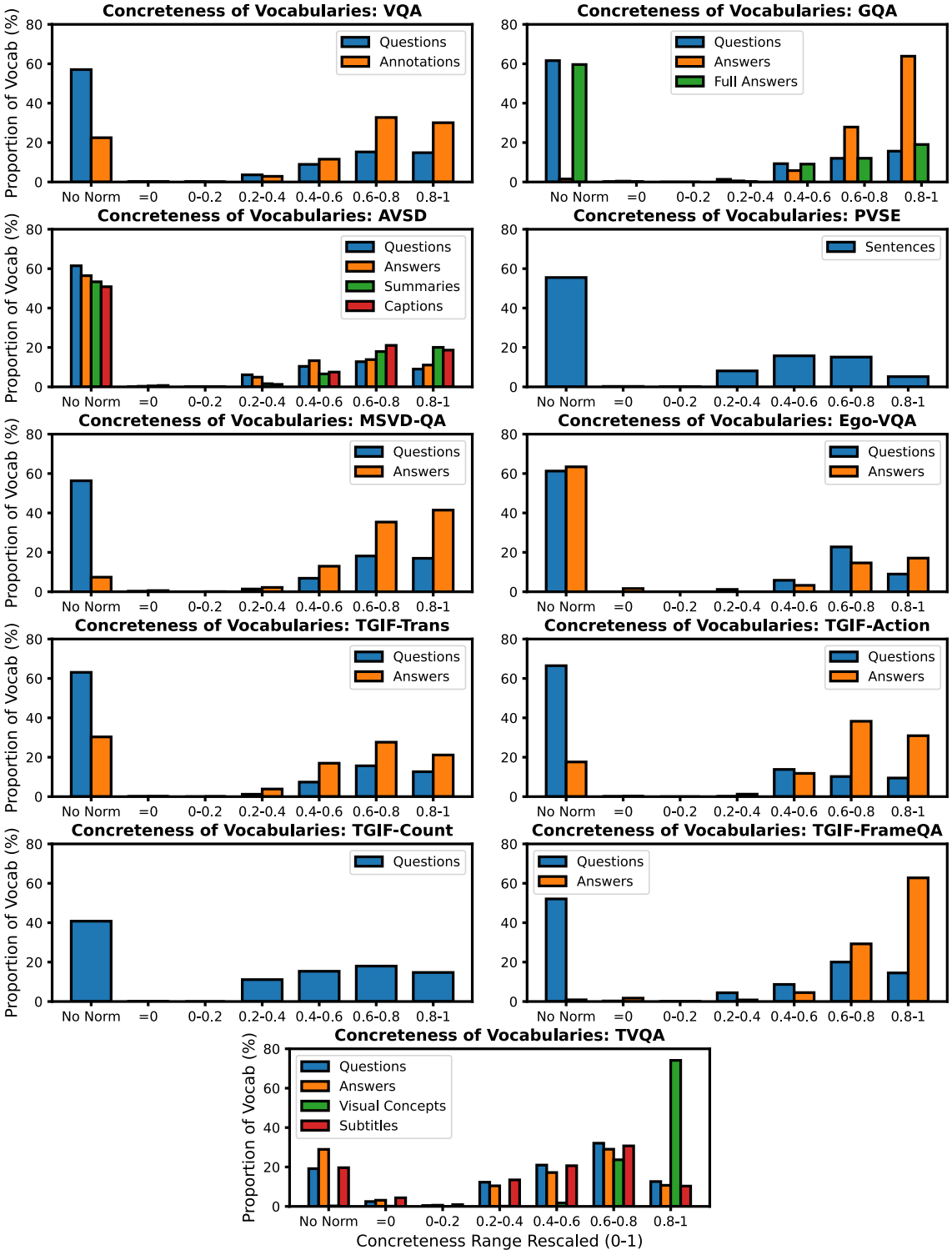


Figure 6.5: Concreteness of vocabularies for components of VQA and video-QA datasets.

6.2.2 Constructing The Labelling Scheme From Word Pairs Similarity Scores

I have thus far described the neurolinguistic motivation behind the 2 similarity scores I use and how I choose between them for my labelling scheme. This subsection describes exactly what the aforementioned categorical and associative similarity scores are and how they are used.

The scores for association come from 'free association' metric in the University of South Florida Free Association Norms (USF) dataset [147]. This metric was gathered by asking participants to write the first word that comes to mind given a presented 'prompt' word. The 'strength' of the association score is measured from the proportion of participants that gave a given word for a particular prompt. The scores for categorical similarity are taken from the SimLex-999 [85] dataset. The SimLex-999 score is designed specifically to "capture object *similarity* instead of *relatedness* or *association*"[1]. As with the concreteness scores, I scale the word similarity measures such that both association and SimLex-999 $\in [0, 1]$. For a given question-answer pair, I construct the VQA answer tensor such that the ground-truth answer retains its value of 1, with the index of any other word in the answer tensor replacing its default 0 with the appropriate similarity score (as visualised in Figure 6.2). This new answer tensor is used as the ground truth for a binary cross entropy loss function. The intuition here is that the ground truth answer will still retain the highest value in loss tensor and should therefore still be selected as the top answer for VQA top-1 accuracy, but that the model will learn to associate that answer with a neighbourhood of similar reasonable answers as dual coding theory implies humans do. The output logits for the answers of the models in my experiments still use softmax to bound the output space, thus the models predictions still sum to 1 wheras the new answer tensor will sum to a value greater than 1. I control for any odd behaviours this scenario may cause by running additional experiments with a 'scaled' version of this labelling scheme where the new answer tensor is rescaled such that it sums to 1 (see Figure 6.6) — inline with the output logits of the model.

[1]See example given here https://fh295.github.io/simlex.html.

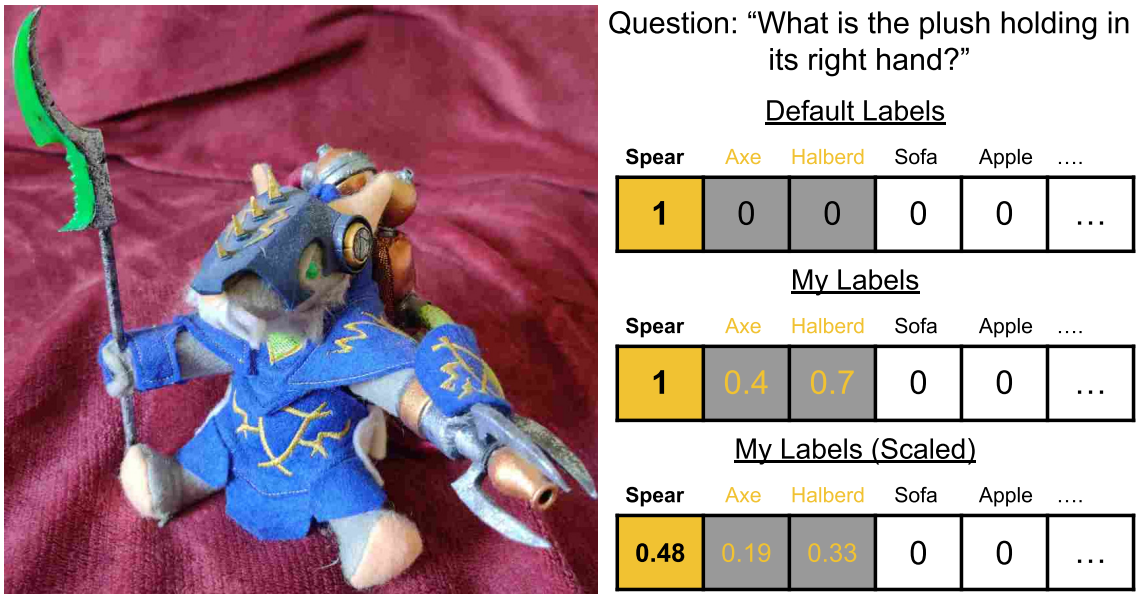


Figure 6.6: Visualisation of answer tensors under my proposed labelling scheme: both with and without scaling.

My initial experiments in Section 6.5 considers the full splits of the VQA datasets, where there are many question-answer pairs that I have no scores for. However, in order to overcome this lack of labelling to more precisely measure the effect of my labelling scheme, my later experiments in Sections 6.6 and 6.7 experiment on subsets that contain at least one score.

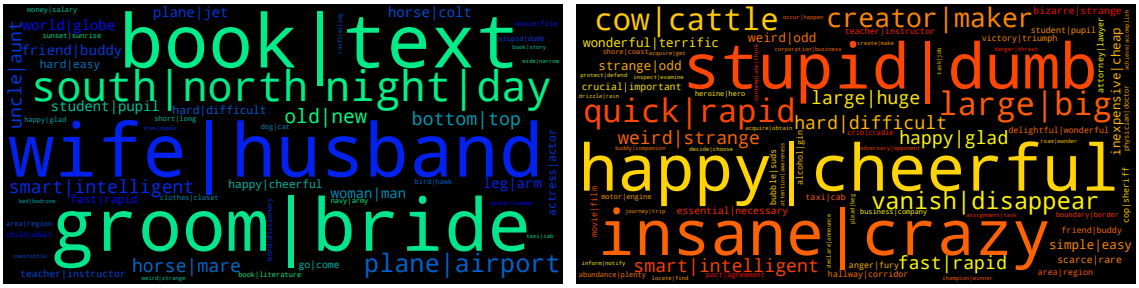
6.2.3 SimLex-999 and Association Measure Statistics

It is beyond the scope of this thesis to *qualitatively* verify if the association and SimLex-999 *truly* measure what they aim to, and if they are meaningfully different measures. However, as *each* of the word pair scores that have a SimLex-999 score

Similarity Measures between Assoc and SimLex word-pair scores			
	VQA v1	**VQA v2**	**GQA**
Cosine Similarity	0.503	0.518	0.491
Kolmogorov-Smirnov Test	0.671	0.667	0.682
Wasserstein Distance	0.041	0.036	0.039
Spearman Correlation	0.336	0.333	0.343

Table 6.1: Similarity measures between the SimLex-999 and USF Association scores. Note that VQA v1 and v2 statistics are the same as VQA-CP v1 and v2 respectively as they are identical dataset *overall* with a different train-test split.

also has a USF score, I can *quantitatively* compare these numerical measures on that full set of shared word pairs. Table 6.1 shows various similarity measures between the association and SimLex-999 scores that appear in each of the datasets in my experiments. These measures together overall imply that in reality there is both reasonable overlap *and* difference in the association and SimLex-999 measures. I provide word clouds of the word-pairs in the datasets used for my experiments (Figure 6.7) for the reader's qualitative consideration.



(a) Associated: USF Association word pairs. (SimLex-999 split only).

(b) Categorically related: SimLex-999 word pairs.

Figure 6.7: Word pairs from the SimLex-999 dataset of word norms. The SimLex-999 score is used as a measure of 'categorical relatedness' for the proposed loss function. The SimLex-999 dataset also contributes the USF association scores to complement their SimLex-999 score, which I use as a measure of 'association'. The size of word-pairs in this figure corresponds to the intensity of each respective score.

6.2.4 Further Nuances of Dual Coding Theory

In this subsection I highlight some additional details in dual coding theory that the reader may find interesting.

It is thought that all words are polysemous to some extent as their precise meaning changes in different contexts [81], inspiring speculation that a strict 'associative/categorical dichotomy' is overly simplistic [35]. Concepts of middling concreteness (*e.g.* nurse, chemistry) are thought to have both associative and categorical connections. Dhond et al. [46] and Binder et al. [16] find evidence implying that abstract and concrete words initially activate similar brain regions and then later separate ones, with concrete concepts activating regions associated with visualisation.

6.3 Related Works

The work in this chapter shares similarities with various 'multi-label' and 'soft-evaluation' learning approaches.

The authors of the DAQUAR dataset [137] introduce the 'WUPS' score, building 'fuzzy' labels for questions, with 'semantically close' answers yielding 'partial membership' scores not unlike the scenario in my proposed loss function. However, the WUPS score focuses on penalising 'naive solutions' where either too many or too few answers are proposed for a given question. Furthermore, in constrast to my approach, the proposed WUPS score is *not* designed as a loss function to be used directly in training.

Geng [64] include various multi-label learning scenarios in their proposed 'Label Distribution Learning' approach. The authors hypothesise a 'general case' of multi-label learning in which multiple labels are weighted differently corresponding to how 'correct' they are for a given question.

The evaluation metric for VQA introduced with the VQA v2 dataset [70] uses additional ground truth answers for each question, however this metric is not dataset-invariant by relying on the additional labels of the dataset.

The design of the VQA-CP datasets [4] specifically *do not* change the distribution of images between the train and test splits, instead changing the distribution of answers for each question type under the hypothesis that: 'it is reasonable to expect models that are answering questions for the right reasons to recognise *black* colour at test time even though *white* may be the most popular answer for that question type in the training set'. This hypothesis implies that it is advantageous for VQA to recognise answers that are semantically similar.

Jedoui et al. [96] highlight the high levels of uncertainty inherent in models trained on 'mutually exclusive' output spaces with one correct answer. Given this undesirable behaviour, the authors propose estimating answer uncertainty from the 'internal hidden space' of a model instead of its output layer. By using a triplet loss to encourage similar answers together in the visual semantic space, the authors demonstrate significantly improved performance on answers that are paraphrases or eachother in VQA.

The work in this chapter is most similar to that of Kervadec et al. [103]. Their work shares our motivation that the typical 'one-hot' labelling practice in VQA (as in Figure 6.1) to be overly simplistic. The authors propose 2 'proximity measures' and a loss function designed to account for the similarities between answers. The authors project their answers into a new semantic space that is designed to satisfy the nuanced proximity measures by using either:

1. Co-occurrences of the ground truth with the auxiliary labels in the VQA dataset.

2. GloVe vector representations of the ground truths.

I argue that, in theory, my approach offers desirable improvements compared to those proposed in Kervadec et al. [103]. Nevertheless, their approach demonstrates VQA accuracy *increases* whereas my results show VQA accuracy *decrease*. I further elaborate on the details of their approach and the potential causes of the discrepancy between our results in this chapter's discussion (Section 6.8.3).

6.4 Experimental Practices

I use the official training and validation splits of each dataset for training and testing respectively. All training and testing subsets proposed in Sections 6.6 and 6.7 are derived entirely from their original split by simply discarding examples. The original boundaries of the splits are respected and there is no mixing or merging of any kind. There *are* test splits for these datasets available for evaluation on submission to their respective leaderboard websites. I acknowledge the advantages (and perceived necessity) in the 'stewards'[2] of datasets *not* releasing the labels of held-back testing sets to the public. In the case of my particular experimental hypothesis however, the *point* is to look beyond top-1 accuracy (top-2, 3, 5, 10) and evaluate the behaviour of the models in neighbourhoods of labels and looking for any signs or benefit of predetermined 'humanlike' behaviour. As this is unfortunately not possible with the

[2]For lack of a more appropriate term.

official test splits of these datasets, I therefore use the validation splits for testing in my experiments.

6.5 Initial Experiments: Full Datasets

My initial experiments apply my proposed labelling scheme to the *full* splits of the following 5 datasets: VQA v1, VQA v2, GQA, VQA-CP v1, and VQA-CP v2 datasets. The model used is the LXMERT multimodal transformer introduced in Tan and Bansal [183]. See Section A.2 for further details of the training setup. Table 6.2 shows that both the scaled and unscaled versions of my proposed labelling

Dataset	Loss	Accuracy	Top-2 Acc	Top-3 Acc	Top-5 Acc	Top-10 Acc
LXMERT						
VQA v1 (Full)	Default	63.63	81.19	85.10	89.01	92.78
VQA v1 (Full)	Mine	-3.26 60.37	77.76	81.06	84.59	88.54
VQA v1 (Full)	Mine (Scaled)	-5.47 58.16	78.69	81.96	85.41	88.96
VQA v2 (Full)	Default	61.26	80.27	84.42	88.59	92.54
VQA v2 (Full)	Mine	-2.83 58.43	78.31	82.33	86.55	90.83
VQA v2 (Full)	Mine (Scaled)	-4.90 56.36	77.15	80.71	84.45	88.39
VQA-CP v1 (Full)	Default	43.03	62.52	70.07	76.31	82.95
VQA-CP v1 (Full)	Mine	-3.76 39.27	57.90	64.04	72.44	80.53
VQA-CP v1 (Full)	Mine (Scaled)	-5.96 37.07	57.46	63.95	73.23	82.05
VQA-CP v2 (Full)	Default	47.71	66.90	74.38	81.91	88.99
VQA-CP v2 (Full)	Mine	-8.23 39.48	51.76	60.51	69.97	80.64
VQA-CP v2 (Full)	Mine (Scaled)	-7.15 40.56	53.79	63.03	73.21	83.45
GQA (Full)	Default	65.89	83.47	87.48	91.68	95.43
GQA (Full)	Mine	-10.25 55.64	79.80	83.58	88.25	93.35
GQA (Full)	Mine (Scaled)	-11.59 54.30	80.33	84.35	88.82	93.78

Table 6.2: Accuracies for my initial experiments on the *full* splits of each of the 5 datasets using the LXMERT model.

scheme lead to a consistent and significant drop in accuracy for all 5 datasets, the worst of which is GQA with an approximate ∼10% decrease. I initially intuitively suspected that the benefits of my approach may instead be seen in the increase of top-k accuracies, as the similarity measures may induce a more robust understanding of a subject and therefore reasonable 'second best guesses'. Instead however, all datasets *also* demonstrate a performance decrease in each of the top-2, top-3, top-5, and top-10 accuracy metrics. The differences between my approaches and the default

setup generally decrease for the higher top-5 and top-10 accuracies: the differences in top-10 accuracies are no worse than ∼4% for all datasets but VQA-CP v2 (which demonstrates a more substantial ∼5% to ∼8% decrease). The generally 'easier' VQA v1/v2 datasets demonstrate the smallest accuracy reductions. Interestingly, though the gap between top-1 accuracies for GQA is large, the discrepancy between top-2 accuracies immediately reduces to ∼3%, and consistently remains closer for the top-3, 5, and 10 accuracy metrics.

6.6 A More Targeted Scenario: Discarding Answers Without Similarity Scores

My initial results in the previous section imply that my labelling scheme is unhelpful in the context of VQA. Before I can conclude this however, I need to explore and control for some of the potential problems that may cause these negative results. As I mention previously, there are many questions with answers for which I have no similarity scores. It could reasonably be expected that accuracy increases may *only* happen on the subset that I have similarity scores for, with the overall decrease resulting from answers without similarity scores weighing the success down. This section explores this potential cause by discarding questions from all datasets that I do *not* have a similarity score for. I re-emphasise that one of the principal aims of the work in this chapter is avoid the limitations of my work in Chapter 3: the need to discard data. Unfortunately, I cannot reasonably gather word similarity scores for the esoteric and specific answers in the VQA answer vocabularies. Regardless of my initial motivation, I can only test the effect of a more 'concentrated' dataset by discarding data and accepting the accompanying drawbacks. Table 6.3 shows the differences in the sizes of the full dataset splits, and those 'SimLex-only' splits that are the subject of this subsection. The first and fifth rows of both halves of Table 6.3 demonstrate that both the total number of answers and unique answers are an order of magnitude lower for the 'SimLex-only' splits. In exchange for this substantial sacrifice: all remaining questions have answers with similarity scores; and as the answer vocabulary is ∼10x smaller, the relative abundance of similarity

	VQA v1 (Full)	VQA v2 (Full)	GQA (Full)
Dataset statistics for answers with Assoc and Simlex word-pair scores			
Total Answers	335,738	609,939	1,075,062
Answers With Assoc Scores	26,825	51,806	184,468
Answers With Simlex Scores	26,893	51,950	190,722
Answers With Assoc or Simlex Scores	26,893	51,950	190,722
Total Unique Answers	2184	3128	1842
Unique Answers With Assoc	205	256	185
Unique Answers With Simlex	209	262	192
Unique Answers With Assoc or Simlex Score	209	262	192
Sum of Scores in Assoc Tensor (Total)	1.016	1.014	1.015
Sum of Scores in Assoc Tensor (Non-zero)	1.170	1.174	1.146
Avg # of Classes in Assoc Tensor (Total)	1.139	1.127	1.159
Avg # of Classes in Assoc Tensor (Non-zero)	2.483	2.547	2.578
Sum of Scores in Simlex Tensor	1.056	1.052	1.064
Sum of Scores in Simlex Tensor (Non-zero)	1.585	1.624	1.614
Avg # of Classes in Simlex Tensor	1.150	1.138	1.170
Avg # of Classes in Simlex Tensor (Non-zero)	2.569	2.649	2.635
	VQA v1 (SimLex-Only)	**VQA v2 (SimLex-Only)**	**GQA (SimLex-Only)**
Dataset statistics for answers with Assoc and Simlex word-pair scores			
Total Answers	26,893	51,950	190,451
Answers With Assoc Scores	26,825	51,806	184,197
Answers With Simlex Scores	26,893	51,950	190,451
Answers With Assoc or Simlex Scores	26,893	51,950	190,451
Total Unique Answers	209	262	192
Unique Answers With Assoc	205	256	185
Unique Answers With Simlex	209	262	192
Unique Answers With Assoc or Simlex Score	209	262	192
Sum of Scores in Assoc Tensor (Total)	1.166	1.170	1.141
Sum of Scores in Assoc Tensor (Non-zero)	1.170	1.174	1.146
Avg # of Classes in Assoc Tensor (Total)	2.455	2.511	2.521
Avg # of Classes in Assoc Tensor (Non-zero)	2.483	2.547	2.578
Sum of Scores in Simlex Tensor	1.585	1.624	1.614
Sum of Scores in Simlex Tensor (Non-zero)	1.585	1.624	1.614
Avg # of Classes in Simlex Tensor	2.569	2.649	2.635
Avg # of Classes in Simlex Tensor (Non-zero)	2.569	2.649	2.635

Table 6.3: Similarity measures of SimLex-999 and USF Assoc scores. Note that VQA v1 and v2 statistics are the same as VQA-CP v1 and v2 respectively as they are identical dataset *overall* with a different train-test split.

scores is ∼10x higher. It is particularly important that I emphasise that the subsets I propose in Sections 6.6 and 6.7 *do not include questions with 'yes' or 'no' answers.* I exclude these binary questions because I do not have similarity scores for either of the words 'yes' or 'no'.

6.6.1 Experimental Results

There are a few factors to consider when comparing the *expected* overall accuracy on my proposed data splits: The dataset splits are substantially smaller than the full splits, and it therefore might be expected they are easier to solve. As previously mentioned however, my subsets do not include binary 'yes-no' questions. Such questions are almost always the most easily solved subsets of VQA datasets, and I believe their exclusion from my subsets causes the accuracy on my smaller subsets to be lower than may otherwise be expected.

6.6.1.1 LXMERT Model on SimLex-Only Data

Broadly speaking, my both versions of my proposed labelling scheme reduces accuracy for the SimLex-only splits *slightly more* than it does on the full dataset splits. The differences in accuracies between the full splits in Table 6.2 and the SimLex-only splits in Table 6.4 vary for each different dataset:

- VQA-CP v1/v2: The overall scores are very similar to those of the full dataset, including the decrease in accuracy under my labelling scheme.

Dataset	Loss	Accuracy	Top-2 Acc	Top-3 Acc	Top-5 Acc	Top-10 Acc
LXMERT						
VQA v1 (SimLex)	Default	73.90	83.80	87.38	90.98	94.27
VQA v1 (SimLex)	Mine	-10.62 63.28	71.96	77.05	82.77	88.34
VQA v1 (SimLex)	Mine (Scaled)	-9.79 64.11	72.70	78.28	83.98	89.13
VQA v2 (SimLex)	Default	71.60	82.90	87.02	90.92	94.36
VQA v2 (SimLex)	Mine	-7.37 64.23	72.40	76.82	82.46	88.02
VQA v2 (SimLex)	Mine (Scaled)	-9.19 62.41	71.80	76.85	82.69	88.77
VQA-CP v1 (SimLex)	Default	43.03	62.52	70.07	76.31	82.95
VQA-CP v1 (SimLex)	Mine	-3.76 39.27	57.90	64.04	72.44	80.53
VQA-CP v1 (SimLex)	Mine (Scaled)	-5.96 37.07	57.46	63.95	73.23	82.05
VQA-CP v2 (SimLex)	Default	47.71	66.90	74.38	81.91	88.99
VQA-CP v2 (SimLex)	Mine	-8.21 39.50	51.58	59.41	68.19	78.13
VQA-CP v2 (SimLex)	Mine (Scaled)	-7.15 40.56	53.79	63.03	73.21	83.45
GQA (SimLex)	Default	70.15	84.63	91.02	95.63	98.59
GQA (SimLex)	Mine	-3.86 66.29	74.66	79.79	87.13	94.55
GQA (SimLex)	Mine (Scaled)	-3.91 66.24	74.28	78.70	86.18	93.33

Table 6.4: Accuracies for my more targeted experiments on the splits for which I have similarity scores between answers. Accuracies are reported for each of the 5 datasets using the LXMERT model.

- GQA: The overall top-1 accuracy for the SimLex-only GQA split are ∼5-10% *higher* than the full dataset splits. The impact of my labelling scheme has lessened from a ∼10% decrease on the full split to ∼4% decrease in top-1 accuracy. Interestingly, the difference in the top-k accuracies between the full and SimLex-only splits disappears as k increases.

- VQA v1: The accuracy of SimLex-only VQA v1 split under the default labelling has increased ∼10% compared to the full split. However, the negative impact of my labelling scheme on deepens from -3.26/-5.47% to -10.62/-9.79%.

- VQA v2: Similarly to VQA v1, default labelling top-1 accuracy increases by ∼10% and the negative impact of my labelling scheme now deepens from -2.83/-4.90% to -7.37/9.19%.

The top-2, 3, 5, and 10 accuracies for all datasets change relatively inline with the top-1 accuracy (other than those aforementioned in GQA). Overall, my proposed labelling scheme under the LXMERT model fails to show any benefit in any of the 5 accuracy metrics.

Dataset	Loss	Accuracy	Top-2 Acc	Top-3 Acc	Top-5 Acc	Top-10 Acc
METER						
VQA v1 (SimLex)	Default	77.06	85.26	88.23	91.24	93.94
VQA v1 (SimLex)	Mine	-1.60 75.46	81.83	85.34	89.28	92.72
VQA v1 (SimLex)	Mine (Scaled)	-2.46 74.60	81.13	84.96	89.19	92.61
VQA v2 (SimLex)	Default	63.06	73.71	78.69	83.79	88.91
VQA v2 (SimLex)	Mine	-0.75 62.31	69.60	74.39	80.23	86.24
VQA v2 (SimLex)	Mine (Scaled)	-1.30 61.76	69.56	74.54	80.41	86.70
VQA-CP v1 (SimLex)	Default	69.10	79.39	82.79	88.20	92.27
VQA-CP v1 (SimLex)	Mine	-8.55 60.55	73.88	80.32	86.61	92.08
VQA-CP v1 (SimLex)	Mine (Scaled)	-9.58 59.52	72.03	76.78	83.92	89.53
VQA-CP v2 (SimLex)	Default	59.75	72.28	77.59	83.21	88.76
VQA-CP v2 (SimLex)	Mine	-0.00 59.75	67.49	73.40	80.35	87.26
VQA-CP v2 (SimLex)	Mine (Scaled)	-0.78 58.97	67.15	73.32	80.19	87.24

Table 6.5: Accuracies for my more targeted experiments on the splits for which I have similarity scores between answers. Accuracies are reported for each of the 5 datasets using the METER framework.

6.6.1.2 METER Model on SimLex-Only Data

While the LXMERT model remains a strong benchmark for VQA, in the 3 years since its introduction a number of newer, higher-performing, larger-scale multimodal transformer models have been proposed. In this subsection, I experiment on the SimLex-only dataset splits instead using larger and more modern Multimodal End-to-End TransformER framework: 'METER' [51]. Overall, the METER framework yields a substantial increase in all accuracies for 3 of the 4 VQA datasets (VQA v1, VQA-CP v1/v2). However, the negative impact from my proposed labelling scheme is significantly reduced (but still present) for 3 of the 4 VQA datasets (VQA v1/v2, VQA-CP v2). A breakdown of the METER results in Table 6.5 versus the LXMERT results in Table 6.4 (both on the SimLex-only splits) is as follows:

- VQA v1: METER yields a ∼4% top-1 accuracy *increase* versus LXMERT for the default labelling scheme. The negative impact of my labelling scheme is substantially *improved* from -10.62/-9.79% to -1.60/-2.46%.
- VQA v2: METER yields a ∼8% top-1 accuracy *decrease* versus LXMERT for the default labelling scheme. The negative impact of my labelling scheme is substantially *improved* from -7.37/-9.19% to -0.75/-1.30%.
- VQA-CP v1: METER yields a substantial ∼26% accuracy *increase* versus LXMERT for the default labelling scheme. However, the negative impact of my labelling scheme is significantly *worsened* from -3.76/-5.96% to -8.55/-9.58%.
- VQA-CP v2: METER yields a substantial ∼23% accuracy *increase* versus LXMERT for the default labelling scheme. The negative impact of my labelling scheme is substantially *improved* from -8.21/-7.15% to 0,-0.78%.

Interestingly, the 3 datasets in Table 6.5 for which my labelling scheme causes a relatively minor decrease in top-1 accuracy *also* demonstrate a more significant negative difference in top-2 and top-3 accuracy. This implies that my intuition: 'improvements may be seen in better second guesses' is not correct for this VQA scenario.

6.7 Using Only the Highest Similarity Scores: Score Clipping

6.7.1 Motivation

My experiments in the previous subsection imply that the negative results are *not* alleviated by focusing only on questions for which I have similarity scores. Helpfully however, the slight *exacerbation* in negative results for such 'concentrated' dataset subsets implies that these 'norm-only' scenarios are ideal for experimentally identifying the effects of my labelling scheme. This subsection therefore uses these 'norm-only' subsets to explore another potential cause for the negative results: *'Is it destructively counter-productive to use low (but still **non-zero**) similarity scores?'*. The intuition here is that a low similarity score indicates 2 concepts are not related. However, since the scores for my labelling scheme are relatively sparse, these low scores between 2 concepts that I happen to have a score for would still *stand out* from amongst the many other label pairs that are 0. Worse still, there will be scenarios where 2 answers are relatively similar, but my labelling scheme will leave a similarity score of 0 for them simply because the neurolinguistic datasets do not have that particular pairing. This leads to scenarios where the labelling scheme implies that 2 very *unrelated* are more similar than more related answer pairs purely

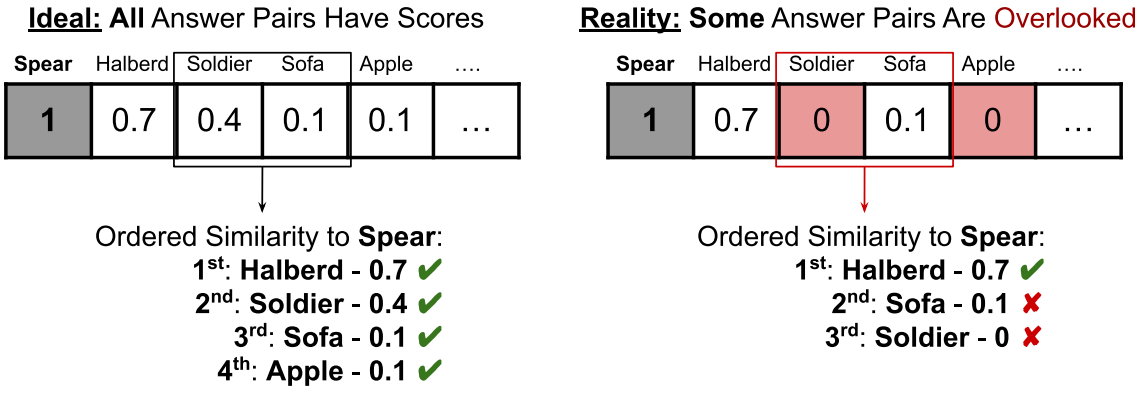


Figure 6.8: An example scenario of the problems that incomplete similarity scores can cause.

because the neurolinguistic dataset is incomplete and may lack certain similarity scores (see Figure 6.8).

6.7.2 Norm Clipping and Expanding Datasets

This section explores the previously described problems caused by similarity score omission as a potential culprit for the negative VQA results of my labelling scheme. To this end, the experiments in this section focus on 'clipping' the similarity scores (derived from the neurolinguistic word 'norms') in the answer tensor at 3 different thresholds: 0, 0.4, and 0.7 (*i.e.* replace any score below the given threshold with 0, see Figure 6.9). My experiments in the previous sections can be thought as a norm clipping threshold of 0 (*i.e.* no norm clipping). However, introducing norm clipping on the 'norm-only' splits would even further reduces the already small and sparse subsets. To offset this further decrease in dataset size, I augment the 'SimLex-only' splits from the previous section by expanding the 'gold standard' SimLex-999 and USF association similarity scores with the next most appropriate word-pair scores available respectively: the SimVerb dataset's 'similarity' score [65] and extra USF association scores beyond those for which a SimLex-999 score exists. See Figure 6.10 for wordclouds for these 'expanded' dataset splits.

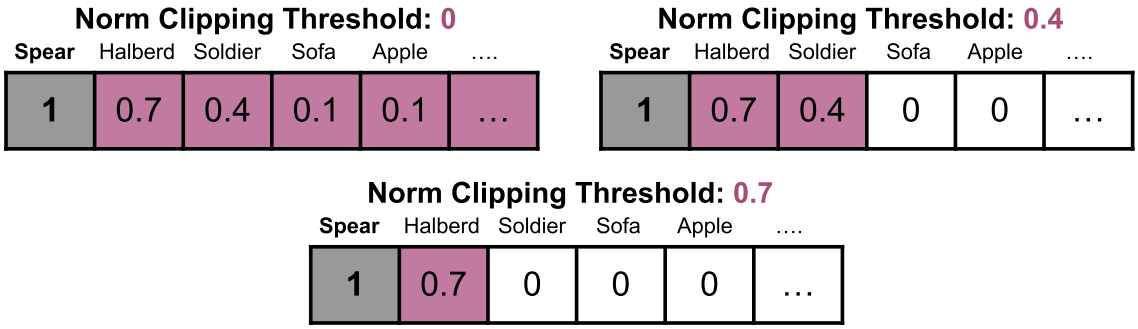
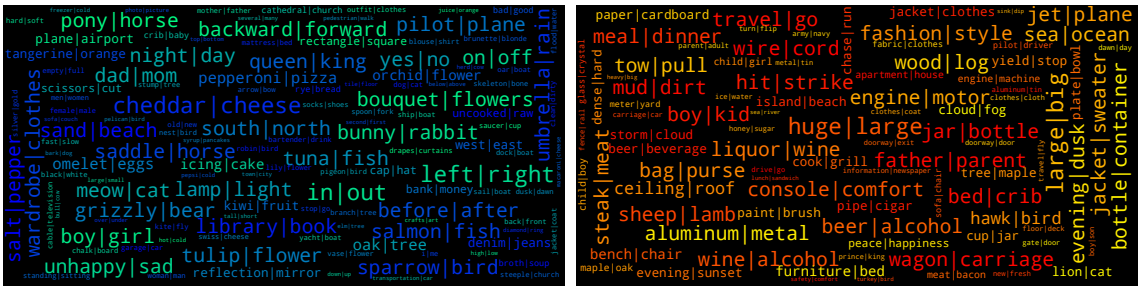


Figure 6.9: A visualisation of norm clipping: replacing the similarity scores (derived from the neurlinguistic word-pair norms) below a certain threshold with 0. Note trivially that the experiments in the previous sections use norm clipping threshold = 0.



(a) Associated: Association scores from the full USF dataset to augment the smaller split in the SimLex-999 dataset.

(b) Categorically related: SimLex-999 word pair scores augmented with the larger 'similarity' metric from the USF dataset.

Figure 6.10: The proposed expanded dataset of word pairs diluting the SimLex-999 scores with extra scores from the full USF dataset. Augmenting the USF Association scores that appear in the SimLex-999 dataset alongside their own SimLex-999 scores with further scores from the *full* USF represents a trade-off for more raw data at the cost of potential incongruence between the SimLex-999 measure of categorical relatedness and the newer 'sim' metric from SimVerb. The size of word-pairs in this figure corresponds to the intensity of each respective score.

6.7.3 LXMERT Model with Norm Clipping

My experiments thus far have followed convention in VQA and use only the questions the VQA splits whose answer occurs at least 9 times. In order to further offset the loss of data from norm clipping, in this section I experiment with relaxing this minimum answer occurrence to 3. Tables 6.6, 6.7, 6.8, and 6.9 detail the results for VQA v1, VQA v2, VQA-CP v1, and VQA-CP v2 respectively. As this subsection's experiments aim to control for dataset size while the impact of norm clipping is measured, I provide alternative representations for the results of Tables 6.6, 6.7, 6.8, and 6.9 that appropriately visualise the accuracies of my experiments alongside the size of the respective dataset splits in Figures 6.11, 6.12, 6.13, and 6.14.

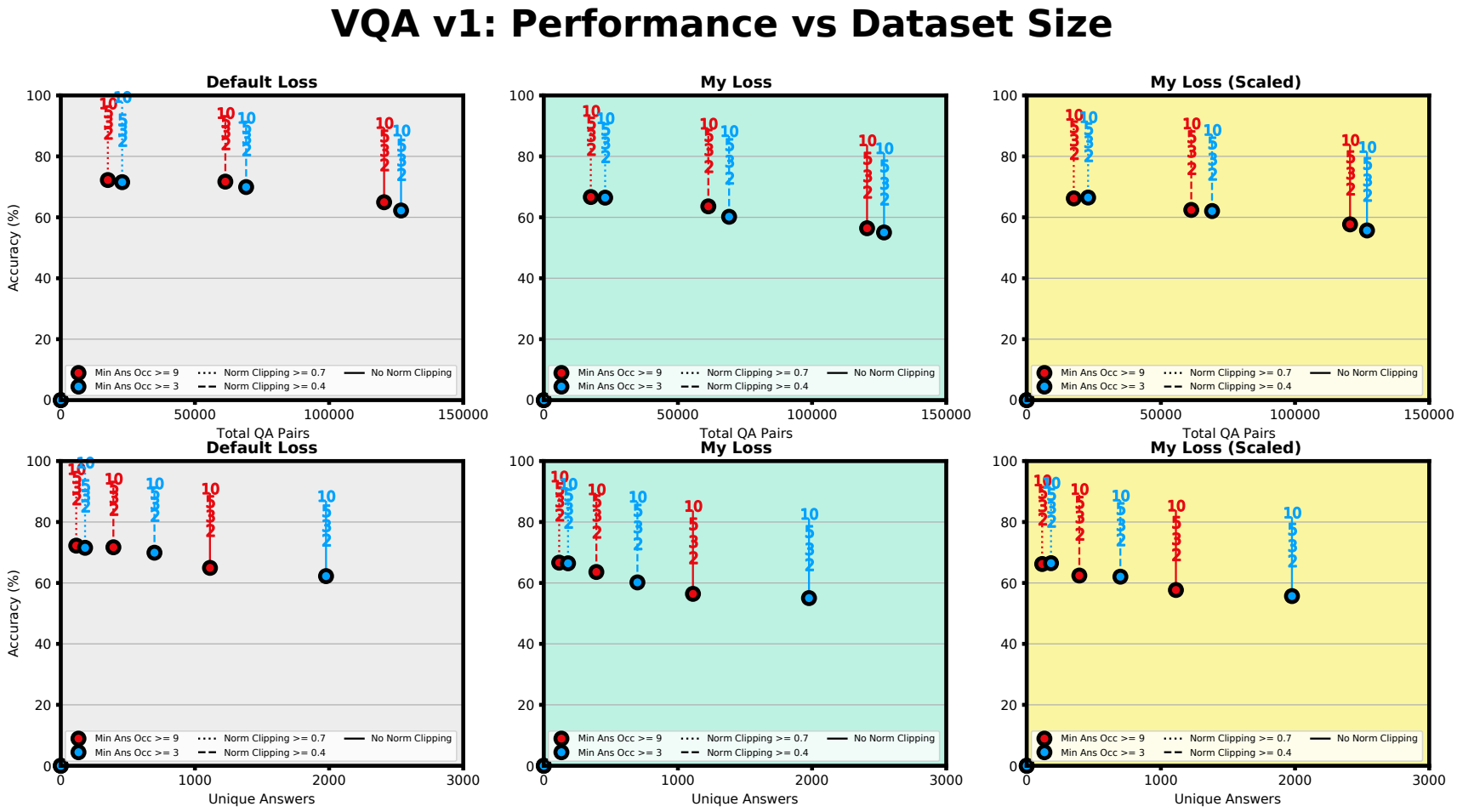


Figure 6.11: VQA v1: Accuracies vs dataset size for various norm clipping dataset setups. The outlined circle corresponds to top-1 accuracy, and the numerical symbols correspond to the appropriate top-k accuracy. The top 3 plots compare accuracies to the *total number of question-answer pairs*, and the bottom 3 plots compare these same accuracies to the number of *unique answers*.

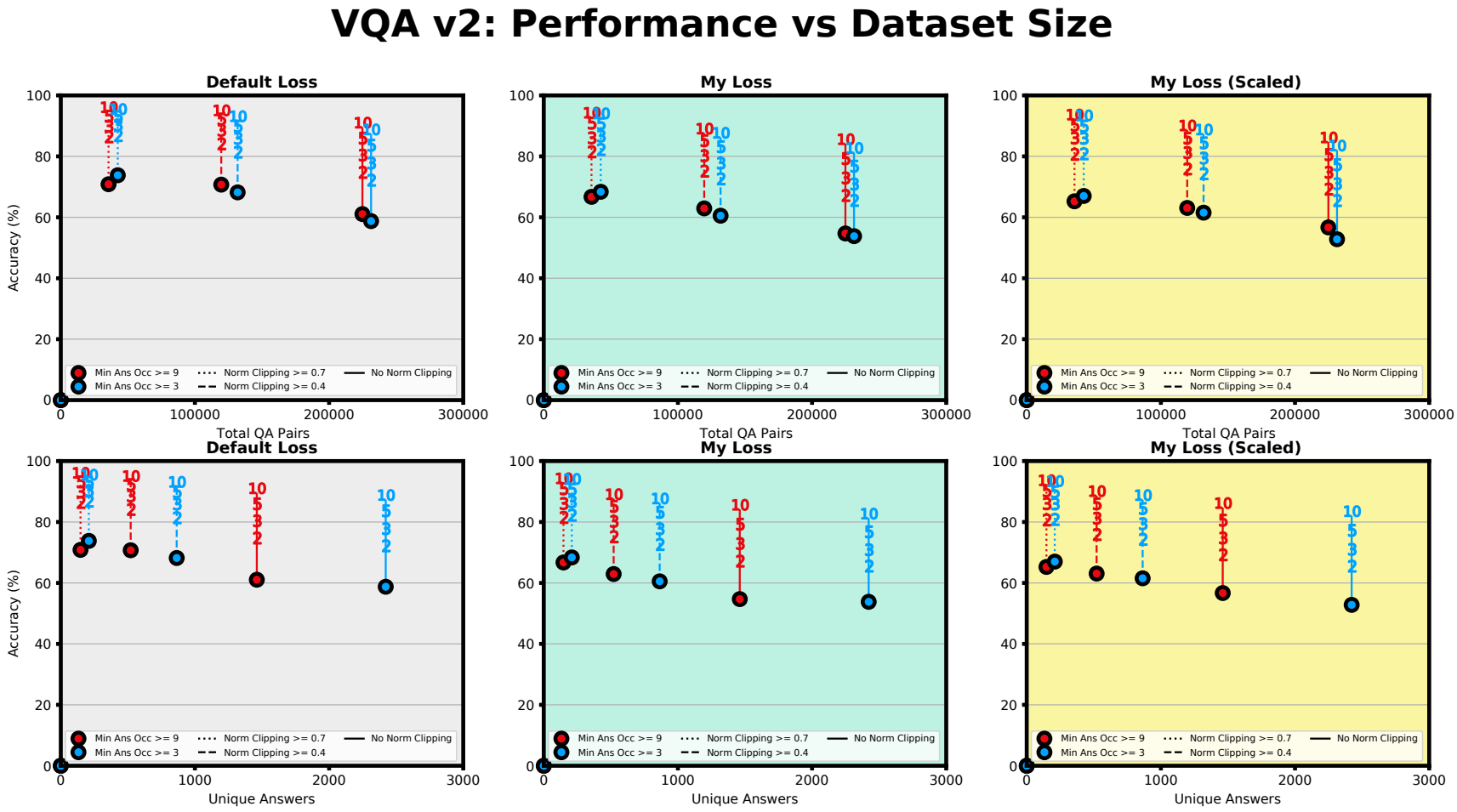


Figure 6.12: VQA v2: Accuracies vs dataset size for various norm clipping dataset setups. The outlined circle corresponds to top-1 accuracy, and the numerical symbols correspond to the appropriate top-k accuracy. The top 3 plots compare accuracies to the *total number of question-answer pairs*, and the bottom 3 plots compare these same accuracies to the number of *unique answers*.

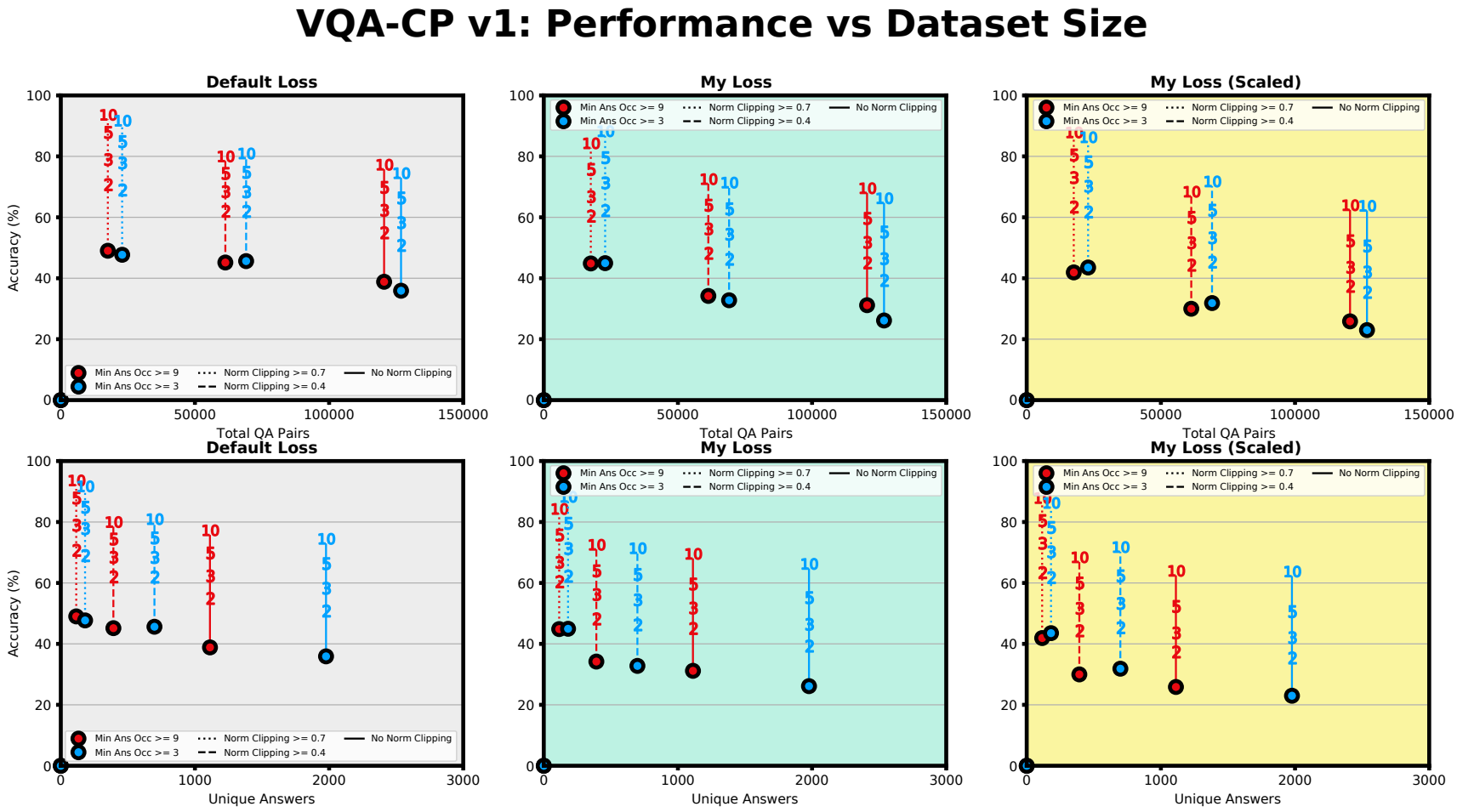


Figure 6.13: VQA-CP v1: Accuracies vs dataset size for various norm clipping dataset setups. The outlined circle corresponds to top-1 accuracy, and the numerical symbols correspond to the appropriate top-k accuracy. The top 3 plots compare accuracies to the *total number of question-answer pairs*, and the bottom 3 plots compare these same accuracies to the number of *unique answers*.

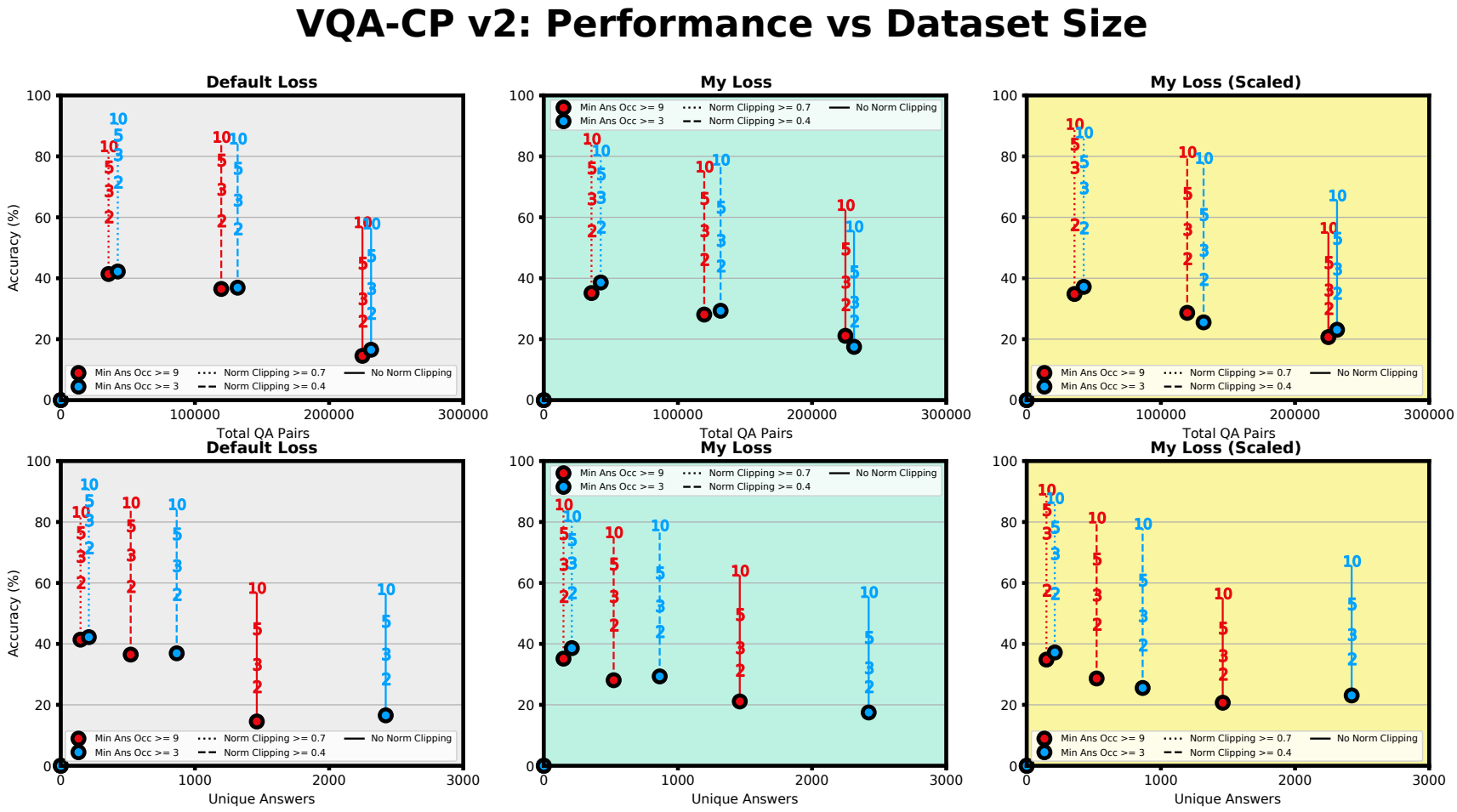


Figure 6.14: VQA-CP v2: Accuracies vs dataset size for various norm clipping dataset setups. The outlined circle corresponds to top-1 accuracy, and the numerical symbols correspond to the appropriate top-k accuracy. The top 3 plots compare accuracies to the *total number of question-answer pairs*, and the bottom 3 plots compare these same accuracies to the number of *unique answers*.

Dataset	Min Ans Occ	Norm Clip	Loss	Accuracy	Top-2 Acc	Top-3 Acc	Top-5 Acc	Top-10 Acc
LXMERT								
VQA v1	3	0	Default	62.24	74.08	78.79	83.63	88.48
VQA v1	3	0	Mine	-7.21 55.03	66.02	71.14	76.85	82.65
VQA v1	3	0	Mine (Scaled)	-6.55 55.69	67.25	72.23	77.64	83.06
VQA v1	3	0.4	Default	69.93	82.19	86.08	89.57	92.64
VQA v1	3	0.4	Mine	-9.75 60.18	72.87	78.40	83.83	88.28
VQA v1	3	0.4	Mine (Scaled)	-7.87 62.06	74.16	79.11	84.38	88.66
VQA v1	3	0.7	Default	71.54	85.27	88.71	91.91	99.49
VQA v1	3	0.7	Mine	-5.09 66.45	79.94	84.42	88.86	92.51
VQA v1	3	0.7	Mine (Scaled)	-5.05 66.49	80.27	84.78	89.05	92.86
VQA v1	9	0	Default	64.95	77.24	81.97	86.67	90.94
VQA v1	9	0	Mine	-8.53 56.42	68.28	73.58	79.36	85.25
VQA v1	9	0	Mine (Scaled)	-7.26 57.69	69.29	74.46	79.85	85.36
VQA v1	9	0.4	Default	71.71	84.17	88.12	91.40	94.20
VQA v1	9	0.4	Mine	-8.11 63.60	76.69	82.21	86.89	90.71
VQA v1	9	0.4	Mine (Scaled)	-9.28 62.43	76.03	81.65	86.43	90.79
VQA v1	9	0.7	Default	72.24	87.62	91.29	94.62	97.34
VQA v1	9	0.7	Mine	-5.58 66.66	82.31	86.88	90.90	94.91
VQA v1	9	0.7	Mine (Scaled)	-6.02 66.22	80.98	85.32	89.67	93.63

Table 6.6: VQA v1: Accuracies vs dataset size for various norm clipping dataset setups. Mirrors Figure 6.11.

Dataset	Min Ans Occ	Norm Clip	Loss	Accuracy	Top-2 Acc	Top-3 Acc	Top-5 Acc	Top-10 Acc
LXMERT								
VQA v2	3	0	Default	58.75	72.21	77.81	83.64	88.84
VQA v2	3	0	Mine	-4.94 53.81	65.56	70.90	76.75	82.74
VQA v2	3	0	Mine (Scaled)	-5.94 52.81	65.53	71.03	77.24	83.54
VQA v2	3	0.4	Default	68.19	81.42	85.94	89.90	93.17
VQA v2	3	0.4	Mine	-7.66 60.53	72.64	77.76	83.16	87.73
VQA v2	3	0.4	Mine (Scaled)	-6.66 61.53	74.34	79.45	84.41	88.81
VQA v2	3	0.7	Default	73.81	86.85	90.33	93.05	95.52
VQA v2	3	0.7	Mine	-5.43 68.38	82.25	86.67	90.71	94.13
VQA v2	3	0.7	Mine (Scaled)	-6.80 67.01	80.85	85.91	89.92	93.44
VQA v2	9	0	Default	61.06	74.77	80.40	85.94	91.06
VQA v2	9	0	Mine	-6.36 54.70	67.19	72.92	79.25	85.64
VQA v2	9	0	Mine (Scaled)	-4.37 56.69	69.27	74.76	80.56	86.20
VQA v2	9	0.4	Default	70.73	84.15	88.59	92.25	95.09
VQA v2	9	0.4	Mine	-7.80 62.93	75.17	80.20	85.13	89.11
VQA v2	9	0.4	Mine (Scaled)	-7.65 63.08	75.92	81.25	85.98	90.14
VQA v2	9	0.7	Default	70.86	86.43	89.89	93.17	96.14
VQA v2	9	0.7	Mine	-4.41 66.72	81.53	86.07	90.89	94.32
VQA v2	9	0.7	Mine (Scaled)	-5.60 65.26	80.70	85.99	90.25	93.70

Table 6.7: VQA v2: Accuracies vs dataset size for various norm clipping dataset setups. Mirrors Figure 6.12.

Dataset	Min Ans Occ	Norm Clip	Loss	Accuracy	Top-2 Acc	Top-3 Acc	Top-5 Acc	Top-10 Acc
LXMERT								
VQA-CP v1	3	0	Default	35.91	50.81	58.24	66.31	74.47
VQA-CP v1	3	0	Mine	-9.77 26.14	39.30	46.41	55.14	66.24
VQA-CP v1	3	0	Mine (Scaled)	-12.92 22.99	35.39	42.05	50.56	63.75
VQA-CP v1	3	0.4	Default	45.63	61.96	68.34	74.73	80.94
VQA-CP v1	3	0.4	Mine	-12.85 32.78	46.27	54.45	62.75	71.38
VQA-CP v1	3	0.4	Mine (Scaled)	-13.77 31.86	45.34	53.27	62.30	71.76
VQA-CP v1	3	0.7	Default	47.71	68.98	77.84	84.82	91.70
VQA-CP v1	3	0.7	Mine	-2.75 44.96	62.21	71.29	79.54	88.26
VQA-CP v1	3	0.7	Mine (Scaled)	-4.18 43.53	61.78	70.20	78.17	86.26
VQA-CP v1	9	0	Default	38.85	54.94	62.30	69.79	77.26
VQA-CP v1	9	0	Mine	-7.69 31.16	45.08	51.76	59.61	69.54
VQA-CP v1	9	0	Mine (Scaled)	-12.99 25.86	37.37	43.58	52.25	63.97
VQA-CP v1	9	0.4	Default	45.17	62.02	68.41	74.37	79.98
VQA-CP v1	9	0.4	Mine	-10.97 34.20	48.20	56.18	63.94	72.52
VQA-CP v1	9	0.4	Mine (Scaled)	-15.18 29.99	44.32	51.66	59.94	68.41
VQA-CP v1	9	0.7	Default	49.03	70.73	78.95	87.83	93.63
VQA-CP v1	9	0.7	Mine	-4.17 44.86	60.39	66.75	75.68	84.28
VQA-CP v1	9	0.7	Mine (Scaled)	-7.11 41.92	63.40	73.00	80.43	87.85

Table 6.8: VQA-CP v1: Accuracies vs dataset size for various norm clipping dataset setups. Mirrors Figure 6.13.

Dataset	Min Ans Occ	Norm Clip	Loss	Accuracy	Top-2 Acc	Top-3 Acc	Top-5 Acc	Top-10 Acc
LXMERT								
VQA-CP v2	3	0	Default	16.54	28.51	36.62	47.38	57.94
VQA-CP v2	3	0	Mine	+0.95 17.49	25.97	32.14	42.07	57.00
VQA-CP v2	3	0	Mine (Scaled)	+4.15 20.69	30.09	36.13	45.12	56.53
VQA-CP v2	3	0.4	Default	36.91	56.25	65.63	76.06	85.81
VQA-CP v2	3	0.4	Mine	-7.58 29.33	44.07	52.44	63.28	78.89
VQA-CP v2	3	0.4	Mine (Scaled)	-8.26 28.65	46.45	56.01	67.81	81.43
VQA-CP v2	3	0.7	Default	42.22	71.55	80.58	86.93	92.39
VQA-CP v2	3	0.7	Mine	-3.61 38.61	56.80	66.54	74.20	81.89
VQA-CP v2	3	0.7	Mine (Scaled)	-7.38 34.84	57.59	76.31	83.98	90.61
VQA-CP v2	9	0	Default	14.53	25.96	33.24	44.92	58.36
VQA-CP v2	9	0	Mine	+6.58 21.11	31.36	38.72	49.60	64.02
VQA-CP v2	9	0	Mine (Scaled)	+8.56 23.09	35.10	43.05	53.07	67.17
VQA-CP v2	9	0.4	Default	36.49	58.85	69.16	78.73	86.37
VQA-CP v2	9	0.4	Mine	-8.41 28.08	46.19	55.58	66.18	76.62
VQA-CP v2	9	0.4	Mine (Scaled)	-10.96 25.53	39.64	49.19	60.89	79.44
VQA-CP v2	9	0.7	Default	41.40	60.16	68.68	76.49	83.32
VQA-CP v2	9	0.7	Mine	-6.25 35.15	55.59	66.05	76.16	85.75
VQA-CP v2	9	0.7	Mine (Scaled)	-4.22 37.18	56.56	69.61	78.24	87.84

Table 6.9: VQA-CP v2: Accuracies vs dataset size for various norm clipping dataset setups. Mirrors Figure 6.14.

As detailed in 4^{th} row of Table 6.3, the SimLex-only splits for VQA v1/v2 contain 26,893 and 51,950 question-answer pairs respectively. Figures 6.6, 6.7 and 6.8, 6.9 show that expanded subsets of VQA v1/VQA-CP v1 and VQA v2/VQA-CP v2 without norm clipping contain ∼120,000 and ∼230,000 question-answer pairs. Loosening the required minimum answer occurrence from 9 to 3 only slightly increases the *total number of question-answer pairs*, but does significantly increase the number of *unique answers* (as can be seen by comparing the blue and red plots in the top and bottom halves of Figures 6.6, 6.7, 6.8, and 6.9). These same figures show that the norm clipping thresholds of 0.4 (0.7) yield dataset splits of size ∼60,000 (20,000) and ∼120,000 (40,000) for VQA v1/VQA-CP v1 and VQA v2/VQA-CP v2 respectively.

Generally speaking, the experiments in this subsection reveal that my labelling scheme *still* does not improve VQA accuracies under *any* of the norm clipping scenarios. Though my labelling scheme still leads to a reduction in all accuracies under norm clipping, the reductions are significantly lessened under the highest level of norm clipping. It is unclear at this time if this less significant decrease for higher norm clipping is caused by alleviating 'incomplete similarity' hypothesis detailed in Figure 6.8, or if it is simply because 'the less my labelling scheme is used, the better'. The overall accuracies decrease slightly but consistently when loosening the minimum answer occurrence requirement to 3. The overall accuracies increase significantly as the norm clipping thresholds raise from 0 to 0.4 and 0.4 to 0.7, which I believe is evidence that the overall reduction in dataset size in turn reduces the complexity/difficulty. Interestingly, though my labelling scheme causes the least decrease for the highest level of norm clipping (0.7), the medium level of norm clipping (0.4) consistently causes more of a decrease than no norm clipping at all (*e.g.* in Table 6.6 rows 5 and 6 represent a bigger decrease of -9.75 and -7.87 than rows 2 and 3 with -7.21 and -6.55). The top-k accuracies continue to reveal no advantage noticeable advantage in my proposed labelling scheme. A breakdown of the experimental results by dataset is as follows:

- VQA v1: The LXMERT experiments are in line with all of the aforementioned trends: *e.g.* the highest norm clipping level (0.7) yields the smallest decrease

in top-1 accuracy under my labelling scheme (-5.58/-6.02 for minimum answer occurrence of 9, and -5.09/5.05 for minimum answer occurrence of 3).

- VQA v2: As with VQA v1, the LXMERT experiments are in line with all of the aforementioned trends. As VQA v2 is generally a harder dataset, the overall accuracies for VQA v2 are lower than that of VQA v1. The decreases in performance of my labelling scheme are generally smaller than those in VQA v1.

- VQA-CP v1: As with VQA v1/v2, the LXMERT experiments are in line with the aforementioned trends. VQA-CP v1 is a harder dataset than VQA v1/v2, and thus its overall accuracies are lower still than those of VQA v2. The negative impact of my labelling scheme is more pronounced at norm clipping = 0.4: with -12.85/-13.77 and -10.97/-15.18 for minimum answer occurrence = 3 and 9 respectively. Conversely, the negative impact of my labelling scheme is comparatively lessened at norm clipping = 0.7: with -2.75/-4.18 and -4.17/-7.11 for minimum answer occurrence = 3 and 9 respectively.

- VQA-CP v2: The LXMERT experiments are in line with those of the previous 3 datasets for norm clipping = 0.4 and 0.7. For norm clipping = 0 however, my labelling scheme appears to give noticable increases over the default: +0.95/+4.15 and +6.58/+8.56 for minimum answer occurrence of 9 and 3 respectively. However, I believe these results are erroneous outliers caused by the higher difficulty of the norm clipping = 0 split. Note that the overall accuracies for norm clipping = 0 for VQA-CP v2 are very poor (∼15%). VQA-CP v2 is already the hardest dataset here, and recall that my dataset design excludes questions with 'yes'/'no' answers (which are generally the easier questions in this dataset). This combination of factors leads to poor training and sub-optimal convergence in labelling scenarios, and thus the supposedly positive results should not be considered reliable.

6.7.4 METER Model with Norm Clipping

The difference in accuracies between the minimum answer occurrence of 9 and 3 dataset scenarios in the LXMERT experiments of this section proved relatively negligible and uninformative: *i.e.* the increase in data it yields does not in practice lead to any significant differences. I therefore omit minimum answer occurrences of 3 from the more expensive and time-consuming METER experiments in this subsection. Tables 6.10, 6.11, 6.12, and 6.13 details the accuracies of the METER framework on the VQA v1, VQA v2, VQA-CP v1, and VQA-CP v2 datasets respectively.

Generally speaking, the negative impact of my labelling scheme on these expanded dataset splits is almost entirely eliminated for almost all METER experiments. The decrease in performance remains consistent, but in a much lower range of ∼0.2-2.0% (except the VQA-CP v1 experiments). Notably, norm clipping = 0.7 yields the biggest decrease in accuracy for 3 of the 4 datasets, in contrast to the LXMERT experiments in the previous subsection where norm clipping = 0.7 caused the least decrease in accuracy. A breakdown of the results by dataset for the METER model is as follows:

- VQA v1: The overall accuracies for lower norm clipping thresholds are noticeably lower than those of LXMERT. However, the norm clipping = 0.7 results are slightly higher than in LXMERT. The negative impact of my labelling scheme is much lower at all norm clipping thresholds than in LXMERT. The most negatively impacted norm clipping scenario is 0.7, with -2.14/-1.90% top-1 accuracy reductions. The norm clipping = 0 scenario even yields a top-1 accuracy increase versus default labelling (+0.22). However, this increase is not sustained in the accompanying top-2, 3, 5, and 10 accuracies.

- VQA v2: The overall accuracies for lower norm clipping thresholds are considerably lower than those of LXMERT. The negative impact of my labelling scheme is much lower at all norm clipping thresholds than in LXMERT. The impacts of my labelling scheme range from +0.15% to -0.67%, with norm clipping = 0.7 displaying the largest negative impact.

- VQA-CP v1: The VQA-CP v1 METER experiments are the 'odd ones out' that break the aforementioned general trends. The METER results are higher than those of LXMERT at each norm clipping threshold. Though norm clipping = 0 yields no significant change under my labelling scheme, the higher norm clipping thresholds display much more significant accuracy decreases than the other METER experiments in this subsection (up to -6.51/-8.31% at norm clipping = 0.7).

- VQA-CP v2: The VQA-CP v1 METER results are significantly better than the corresponding LXMERT results in the previous subsection. Each model reliably trains and converges at each norm clipping threshold. The most negative impact from my labelling scheme is experienced at norm clipping = 0.4 (-2.64/-3.37%).

Dataset	Min Ans Occ	Norm Clip	Loss	Accuracy	Top-2 Acc	Top-3 Acc	Top-5 Acc	Top-10 Acc
METER								
VQA v1	9	0	Default	48.09	59.26	64.57	70.40	77.82
VQA v1	9	0	Mine	+0.22 48.31	58.26	63.72	69.87	77.09
VQA v1	9	0	Mine (Scaled)	-0.25 47.84	58.45	63.96	70.07	77.24
VQA v1	9	4	Default	64.87	75.86	80.58	84.94	89.53
VQA v1	9	4	Mine	-0.49 64.38	74.21	78.89	83.64	88.61
VQA v1	9	4	Mine (Scaled)	-0.60 64.27	74.62	79.20	84.13	88.85
VQA v1	9	7	Default	74.88	88.59	91.57	94.64	96.39
VQA v1	9	7	Mine	-2.14 72.74	86.70	90.74	93.89	96.22
VQA v1	9	7	Mine (Scaled)	-1.90 72.98	86.96	90.65	93.93	96.36

Table 6.10: VQA v1: Accuracies vs dataset size for METER on various norm clipping dataset setups.

Dataset	Min Ans Occ	Norm Clip	Loss	Accuracy	Top-2 Acc	Top-3 Acc	Top-5 Acc	Top-10 Acc
METER								
VQA v2	9	0	Default	28.98	39.44	45.76	53.60	63.47
VQA v2	9	0	Mine	+0.15 29.13	38.93	44.94	52.38	62.39
VQA v2	9	0	Mine (Scaled)	-0.18 28.80	39.03	44.98	52.30	61.90
VQA v2	9	4	Default	46.63	59.04	65.67	72.59	80.47
VQA v2	9	4	Mine	-0.26 46.37	57.07	63.23	70.36	78.51
VQA v2	9	4	Mine (Scaled)	-0.58 46.05	57.78	64.06	71.18	79.45
VQA v2	9	7	Default	65.63	80.37	85.14	89.47	93.15
VQA v2	9	7	Mine	-0.35 65.28	78.95	84.11	88.48	92.58
VQA v2	9	7	Mine (Scaled)	-0.67 64.96	79.28	84.09	88.73	92.89

Table 6.11: VQA v2: Accuracies vs dataset size for METER on various norm clipping dataset setups.

Dataset	Min Ans Occ	Norm Clip	Loss	Accuracy	Top-2 Acc	Top-3 Acc	Top-5 Acc	Top-10 Acc
METER								
VQA-CP v1	9	0	Default	40.87	53.77	59.93	66.91	75.57
VQA-CP v1	9	0	Mine	+0.70 41.57	54.54	61.27	68.76	77.05
VQA-CP v1	9	0	Mine (Scaled)	-1.96 38.91	52.32	58.74	66.41	74.98
VQA-CP v1	9	4	Default	53.29	68.31	74.91	81.88	88.62
VQA-CP v1	9	4	Mine	-2.18 51.11	65.01	72.38	80.95	87.13
VQA-CP v1	9	4	Mine (Scaled)	-4.77 48.52	62.93	69.84	77.55	85.57
VQA-CP v1	9	7	Default	58.77	72.81	79.17	85.90	92.49
VQA-CP v1	9	7	Mine	-6.51 52.26	72.81	81.80	88.24	92.63
VQA-CP v1	9	7	Mine (Scaled)	-8.31 50.46	71.09	82.38	87.54	91.49

Table 6.12: VQA-CP v1: Accuracies vs dataset size for METER on various norm clipping dataset setups.

Dataset	Min Ans Occ	Norm Clip	Loss	Accuracy	Top-2 Acc	Top-3 Acc	Top-5 Acc	Top-10 Acc
METER								
VQA-CP v2	9	0	Default	26.51	36.85	43.08	50.97	61.02
VQA-CP v2	9	0	Mine	+0.09 26.60	36.34	42.44	50.29	60.94
VQA-CP v2	9	0	Mine (Scaled)	-0.81 25.70	35.79	42.51	50.59	61.21
VQA-CP v2	9	4	Default	41.84	56.53	64.10	72.41	80.16
VQA-CP v2	9	4	Mine	-2.64 39.20	51.27	58.43	66.65	75.91
VQA-CP v2	9	4	Mine (Scaled)	-3.37 38.47	53.37	61.52	69.95	78.62
VQA-CP v2	9	7	Default	50.56	70.59	78.67	85.56	91.80
VQA-CP v2	9	7	Mine	-0.79 49.77	66.72	77.67	85.08	91.64
VQA-CP v2	9	7	Mine (Scaled)	-1.61 48.95	65.27	76.76	85.44	92.11

Table 6.13: VQA-CP v2: Accuracies vs dataset size for METER on various norm clipping dataset setups.

6.8 Discussion

6.8.1 General Discussion

Overall, almost all of my experimental scenarios consistently demonstrate that my labelling scheme (in both scaled and unscaled forms) significantly reduces top-1, 2, 3, 5, and 10 accuracies on the VQA datasets in my experiments. This chapter's contributions therefore point towards a 'negative results' study. Though negative results can be important and useful [144, 54], it is *crucial* that such studies specifically prioritise exploring potential causes of the negative results, and discuss *why* efforts to alleviate them fail. I have approached the work in this chapter with that specific philosophy in mind.

After my initial negative results on the full dataset splits (Section 6.5), my first step in diagnosing the problem is to focus on the subset samples for which I have similarity scores (Section 6.6). These focused 'SimLex-only' dataset splits slightly exacerbated the negative impact of my labelling approach. As this magnification of negative results is caused by concentrating the abundance of my labelling scheme, then it is implied that labelling scheme is the direct source of the accuracy decrease. Furthermore, as the top-2, 3, 5, and 10 accuracies mirror the reduction in top-1 accuracy, then it follows my hypothesis that the benefits of 'a more humanlike understanding of similar concepts' my labelling scheme aspires to has failed *i.e.* models do not exhibit improved 'second guesses'.

After exploring the previous 2 avenues for diagnosing negative results, I identify the potential problem: 'incomplete answer tensors encourage models to incorrectly learn that 2 unrelated answers are more similar than other answers that are realistically more similar, but unfortunately lack a similarity score'. Section 6.7 seeks improved results for my labelling scheme by remedying this problem through 'norm clipping', but must relax the 'gold standard' of SimLex-999 categorical and associative similarity scores in order to regain more data and offset the further loss of data norm clipping creates. Though my results are mixed, these norm clipping experiments overall imply that removing the smaller similarity scores in the answer tensor only *reduces the negative impact* of my labelling scheme, but still fails to

render it beneficial to VQA accuracies. It is unclear at this stage if alleviating the 'incomplete score' problem causes this improvement, or if it is simply that 'the less my labelling scheme is used, the better'. It could be argued that norm clipping is indeed beneficial in its own right, as the highest norm clipping splits generate much smaller and more concentrated datasets through the significant reduction in both the total number of question-answer pairs and unique answers: since the highest norm clipping threshold (0.7) leads to a more 'concentrated dataset' with *less negative* results —in contrast to the 'concentrated dataset' producing *more negative* results in Section 6.6. It could therefore be argued that norm clipping is a partial success, but I believe these results alone are insufficient to argue this.

My most promising findings are the results of METER framework on the norm clipping subsets: the VQA v1/v2 and VQA-CP v2 datasets demonstrate either negligible change in accuracy, or a *slight* decrease in accuracy. The METER framework is more modern and larger scale than the older LXMERT model. It could be that when my labelling scheme is refined by norm clipping, the more powerful METER framework is nuanced enough to either (optimistically) adapt to my labelling function or (more cynically) compensate for it. This may imply that my approach would be worth re-visiting for use in the increasingly competent VQA models of the future.

6.8.2 Quality of Word Norms

Another potential explanation for my poor results may lie with the concreteness or word pair scores themselves. It could be that scores themselves are not sufficiently reliable for use in my labelling scheme: *i.e.* perhaps the SimLex-999 metric does not always successfully measure the categorical-similarity it aims to. Though I may find examples in the word cloud figures in this chapter that I find unintuitive, ultimately however it is beyond my area of expertise to critique the quality of the word norms I leverage. I can still quantifiably measure the scores I use (as in Table 6.1). However, I cannot draw any conclusions about my results with respect to the quality of the neurolinguistic norms I use.

6.8.3 Comparison to Similar Research

As previously outlined in Section 6.3, my work in this chapter is similar in motivation to Kervadec et al. [103]. The 2 'proximity measures' the authors define are used in a loss function designed to exploit the similarities between answers. The authors project their answers into a new semantic space that is designed to satisfy the nuanced proximity measures by using either:

1. Co-occurrences of the ground truth with the auxiliary labels in the VQA dataset.

2. GloVe vector representations of the ground truths.

I argue that my proposed labelling scheme would ideally offer the following advantages over their proposed approach:

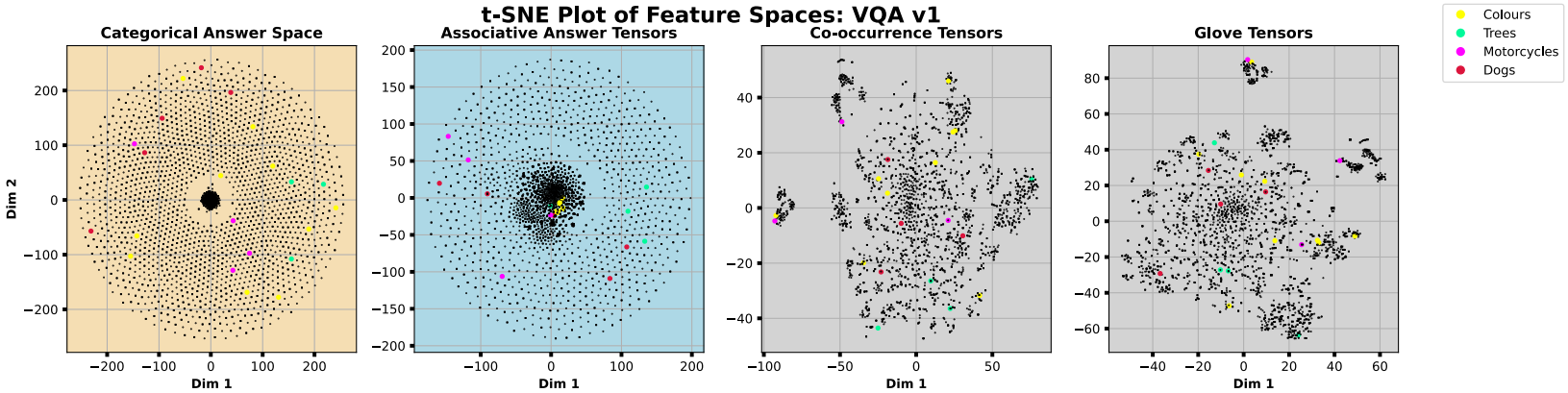
1. Their first co-occurrence based approach is reliant on the VQA auxillary labels to work, and thus is not guaranteed to be dataset-invariant, whereas my labelling scheme can be applied to any dataset that contains answers there are neurolinguistic scores for.

2. As the co-occurrence based approach work relies on the VQA auxillary labels, it risks introducing another avenue of dataset bias by inducing 'hints' for VQA specific answer trends.

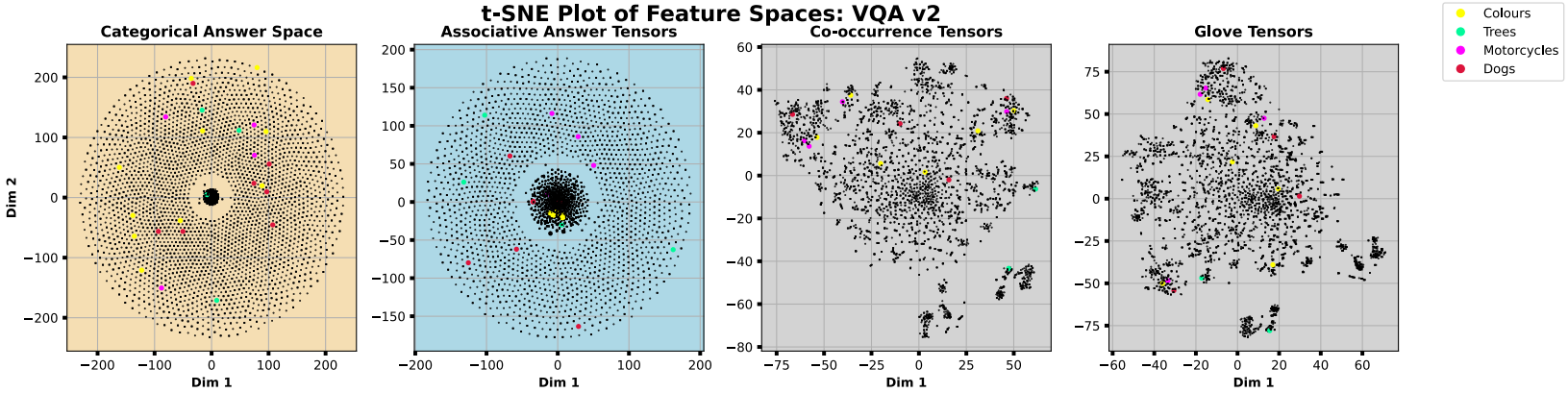
3. Their second approach using GloVe vectors *is* dataset invariant, however it is subject to the difficult-to-quantify biases that exist in the GloVe vectors, and as such its behaviour may be more difficult to understand. In contrast, I argue my proposed labelling scheme is intrinsically more easily interpreted because the similarity scores in the answer vectors are ready for inspection, and have been produced by rigorous neurolinguistic study.

I argue my approach would theoretically offer these desirable improvements compared to those proposed in Kervadec et al. [103]. Nevertheless, their approach demonstrate VQA accuracy *increases* whereas my results show VQA accuracy *decrease*. Their successes implies that the *methodology* of multiclass labelling based on

similarity is promising, but that my specific approach is not suitable for the VQA context. The authors qualitatively assess the structure of their 2 semantic spaces by visualising them in t-SNE [196] plots. I regenerate their feature spaces and qualitatively compare them to the answer space of my labelling scheme: *i.e.* the set of answer tensors (fully annotated with all similarity scores norm clipping = 0). I project these feature spaces with t-SNE in Figures 6.15a and 6.15b. Note that the VQA and VQA-CP versions both contain the same answers and data but reshuffled *i.e.* Figure 6.15a depicts both VQA v1 *and* VQA-CP v1.



(a) Features from VQA v1 (and VQA-CP v1).



(b) Features from VQA v2 (and VQA-CP v2).

Figure 6.15: t-SNE representations of the categorical and associative answer tensors in my experiments alongside the auxiliary answer co-occurrence and GloVe representations in Kervadec et al. [103]. The coloured marks indicate the locations of the same examples focused on in the qualitative analysis by Kervadec et al. [103]. Colours: 'orange', 'white', 'red', 'blue', 'green', 'gray', 'black', 'pink', 'black', 'yellow'. Trees: 'log', 'palm tree', 'tree branch', 'christmas tree'. Motorcycles: 'yamaha', 'kawasaki', 'harley', 'suzuki'. Dogs: 'puppy', 'golden', 'retriever', 'german shepherd', 'husky', 'terrier', 'labrador', 'rottweiler', 'corgi'.

The radial separation of clusters from the 2 answer spaces used in my approach reflect the fact that most of the similarity tensors for answers remain very sparse (despite my best efforts to expand them). However, Figure 6.15 shows that some of the very examples selected for visualisation in Kervadec et al. [103] similarly remain loosely proximal in the t-SNE projection of my answer space. Though I have argued that neurolinguistic human scores in theory offer attractive advantages, Figure 6.15 strongly imply that both the GloVe and answer co-occurrence spaces —in spite of their other drawbacks— are more heavily featured, nuanced, and complete than the similarity spaces used in my experiments. It could be argued that the comparative negative VQA results of my approach versus the successes of the auxiliary co-occurrence and GloVe representations are instead in some sense incongruent: *i.e.* there is some fundamental semantic difference between VQA auxiliary labels/GloVe vectors and 'real human neurolinguistic representations'. Indeed, my review of dataset bias in Chapter 2 emphasises how specific, biased, and unintuitive VQA question-answer semantics can be. However, my work in this chapter indicates that my proposed neurolinguistic similarity scores would need to be much more thoroughly annotated before this hypothesis can be properly tested. Considering all I have covered in this discussion, overall I believe that my approach would greatly benefit by achieving much more 'complete' (*i.e. less sparse*) similarity scores. Furthermore, the successes of similar approaches on VQA accuracy imply that the underlying methodology is promising.

6.9 Conclusion

Under the guiding principal of contributing dataset and model invariant methodology, in this chapter I propose a neurolinguistically-guided multiclass labelling scheme that aims to induce a more 'humanlike' behaviour in VQA frameworks. My initial results demonstrate deteriorating accuracies in two multimodal transformer models across 5 VQA datasets. My subsequent efforts to diagnose potential causes of this deterioration imply that 'the more intensely my labelling scheme is used, the more noticable the decrease in performance' in the vast majority of my experimental

scenarios. I do however find that applying 'norm clipping' to remove low similarity scores causes the more powerful and modern METER framework to deteriorate much less substantially under my labelling scheme, with some experimental scenarios demonstrating no significant accuracy deterioration. I qualitatively compare my approach to similar successful methodologies by visualising clusters of our respective answer spaces. I conclude that in principal my methodology is sound, but that the neurolinguistic similarity scores I leverage are too general to sufficiently populate the often esoteric and specific VQA answer space. In the context of the widely explored problem of VQA dataset bias, I discuss the possibility that there is a problematic incongruence between VQA answer spaces and 'human-like' neurolinguistic scores. However, I ultimately conclude that my labelling scheme would need to be more completely annotated —through extra neurolinguistic similarity scores or otherwise— before such a hypothesis can be properly entertained.

CHAPTER 7

Discussion and Conclusion

In each of my preceding contribution chapters, I have provided a full discussion and conclusion of the findings and results. I have explicitly chosen to do this so that the significance of all my findings and details can be thoroughly and fully contextualised right alongside their respective results subsections. To complement these existing conclusions, this chapter aims to condense them into a more general discussion: focusing on the findings, themes, and guiding questions of this thesis *as a whole*. To these concluding ends, this is chapter therefore structured as follows:

1. Section 7.1 summarises the contributions and conclusions of my work in this thesis.

2. Section 7.2 summarises the motivation and guiding principals that I have adhered to in my work.

3. Section 7.3 discusses the significance of my findings, and examines how I have answered the research questions of this thesis.

4. Section 7.4 discusses both the limitations of my work and outlines avenues for future work to consider.

7.1 Contributions

My literature review in **Chapter 2** provides an extensive analytical overview of areas share across my contribution chapters: VQA datasets, video-QA datasets, and modality bias in multimodal question-answering datasets.

Chapter 3 exhaustively demonstrates the catastrophic presence of language bias in the TVQA video-QA dataset, where most research in the field is focused instead on VQA. I demonstrate that state-of-the-art results are achievable on TVQA by focusing purely on the language features, rather than requiring the visual inputs as intended. I propose a multimodal evaluation framework that can isolate 'enriched' modality-reliant subsets of datasets through my proposed Inclusion-Exclusion Metric (IEM), with which I generate such subsets of the TVQA dataset. I emphasise that a truly multimodal dataset is not easily attained, even with design a philosophy specifically aspiring to it.

Chapter 4 experiments with applying the bilinear pooling operation —popularised in VQA— to video-QA. My results across 2 models and 4 datasets show that bilinear pooling harms performance in video-QA, contrasting its successes on VQA. I combine my experimental results with insights from the surrounding literature to offer explanations for the poor performance of bilinear pooling in a video-QA context, most notably: that language-vision video tasks underperform with bilinear pooling where other modality combinations succeed; and that bilinear operations are expensive and inefficient to train throughout multiple time steps. I draw parallels between the early use of bilinear techniques and two neurological theories: The Two-Stream Model of Vision, and Dual Coding Theory. I complement these observations by proposing several ideas for neurologically-inspired multimodal processing. I develop one of these proposals from dual coding theory in Chapter 6.

Chapter 5 takes a different approach in addressing the relative imbalance between text and vision utilisation by 'improving the power of vision'. This chapter endeavours to avoid the problems of language bias entirely by focusing on visual dynamics video datasets, and applying the powerful generative pretraining paradigm to vision —from its successes in language modelling— *i.e.* visual modelling. I achieve this by demonstrating the benefits of visual pretraining for 'downstream'

tasks. To this end, I introduce 6 dynamic simulations that are naturally affiliated with 7 'downstream tasks' on which I demonstrate that visual pretraining can lead to significant improvement in downstream task performance.

Chapter 6 returns to multimodal question-answering, using VQA instead of video-QA as the experimental platform thanks more thorough VQA research in mitigating language bias (VQA-CP and GQA). I realise one of my proposals in Chapter 4 and introduce a neurolinguistically-guided multiclass labelling scheme that aims to induce more humanlike behaviour through neurolinguistic similarity scores. My initial results show my proposed labelling scheme deteriorates VQA accuracy on 2 VQA models and 5 datasets. My subsequent experiments demonstrate that 'clipping' away the low similarity scores can, at the very best, bring my labelling scheme inline with that of the default 'one-hot' labelling. In light of my negative results, I qualitatively compare my approach to similar successful methodologies by visualising clusters of our respective answer spaces. I conclude that in principal my methodology is sound, but that the neurolinguistic similarity scores I leverage are too general to sufficiently populate the often esoteric and specific VQA answer spaces.

7.2 Motivation and Guiding Principals

The most fundamental motivation that has underpinned the research questions I tackle in this thesis is: "to improve the quality and capacity of multimodal processing in machine learning". As outlined in Chapter 1, I identified the overlap of language and vision to be both the most advanced and heavily resourced area of multimodal machine learning. In fact, interest and research output in multimodal language-vision machine learning tasks has continued to grow exponentially in the years I have dedicated to this thesis. Across the literature in Chapter 2 that I have studied, I have identified language-vision question-answering tasks (VQA and video-QA) as a suitably complex task, paired with the latest state-of-the-art modelling techniques of their times, from which to develop and test my own methodology. Recent history in the field has shown that the real *significant* leaps forward in the

performance and state-of-the-art of language-vision question-answering is primarily achieved by exponentially increasing the size of datasets and models, and that the conceptually 'uncomplicated' and somewhat 'pure' attention-based transformer architecture is sufficient (perhaps even optimal) to facilitate this. The players in this field with the resources sufficient to push the 'scale' of language-vision benchmarks provide appreciated and exciting new frontiers in multimodal processing and I wholeheartedly welcome such investment in this field. As a researcher with more modest resources, I have determined that I can best play my part in furthering multimodal progress by focusing on the major problems in the field that I *can* address:

- Modality bias and the imbalance of classifying 'power' of language and vision inputs.
- The sub-optimal exploitation of input features at both the modelling and dataset level.
- The rapid growth of the field rendering even recently published work irrelevant and outdated.

This incredible period of growth, interest, and topic volatility[1] in particular motivates one of my key guiding principals: to ensure my contributions remain relevant, 'scaling' with revolutionary new benchmark models and datasets of the near future. Given this path I have described: through the initial research question of my thesis, to the end of my final contribution in Chapter 6, I have therefore verbosely but precisely chosen to my thesis title to be: **Scalable Methodologies and Analyses for Modality Bias and Feature Exploitation in Language-Vision Multimodal Deep Learning**. My work in this thesis represents my cumulative efforts to contribute to key elements of language-vision multimodality in deep learning. I direct my work away from more 'saturated' areas of research towards those that are 'intuitive, yet overlooked'. I have aspired to be methodologically novel. This has led me to focus more on theoretical studies, exploring the properties of modern

[1]Fuelled in no small part by exponentially more expensive and out-of-reach aforementioned privately funded models.

methodology, and discussing their suitability for future use. I have not focused on application studies that —though of extreme import and interest to the field— are likely to be quickly outgrown by the next big transformer or stable-diffusion model.

7.3 Significance, and Answering my Research Questions

In the following subsections: I highlight the prominent research questions raised in each of my chapters; I discuss the extent to which I have addressed that research question; and I discuss the significance of the answers I yield to the wider field.

7.3.1 Chapter 3

My early work in this PhD prioritised finding appropriately modern and complex multimodal benchmark datasets and models to serve as an initial experimental platform. I settled on the newly-proliferated video-QA task, but quickly found that the state-of-the-art TVQA dataset I had chosen was seemingly plagued by textual shortcuts that allowed models to catastrophically undervalue its visual inputs. The most immediate (and to-date most focused on in citation) message from my work here is that: despite the best efforts of the authors of TVQA to ensure 'multimodality' in their dataset, hazardous language bias flaws still permeate this popular and widely used dataset. My findings should serve as a caution to be critical of the design of the datasets we use. With my guiding principals in mind, this raises the first research question of this chapter: "Can we mitigate modality bias in datasets in a 'scalable', future-proofed manner?". In short, yes but with some limitations. My IEM method *does* allow for modality-reliant subsets of a biased dataset to be isolated, but the subsets I can propose on TVQA are an order of magnitude smaller that the full dataset. The subset splits I propose are therefore likely to be less desirable for experimentation in future research as dataset size is a crucial factor in deep learning. However, the small size of IEM subsets are likely to grow conveniently large as future datasets increase in size. I argue that this will allow my IEM modailty-

subset methodology (already model/dataset-invariant) to be of relevant use for the datasets of the future. At time of my experiments, such textual-bias analyses were thoroughly analysed in still-image *VQA* datasets, yet surprisingly underexplored for *video-QA*. As such, Chapter 3 addresses the relevant research gap of modality bias in *video-QA*.

7.3.2 Chapter 4

My next area of focus shifted towards multimodal *modelling* methodology and the way it is discussed and critically analysed in the field. My initial work in Chapter 4 —exploring a seemingly-overlooked research gap in the application of 'bilinear pooling' from VQA to video-QA— grew from a curious experiment into a much longer discussion about how the focus on achieving incremental empirical improvements can overshadow other interesting findings. As for the immediate question: "how does bilinear pooling work as a multimodal fusion technique when applied to video-QA?", my experiments across 4 datasets and 2 models show that bilinear pooling yields consistent and significant accuracy decreases when substituted for a simple concatenation of visual+textual features. Though my findings alone are already of note for people wishing to 'plug-and-play' BLP in video models of their own, I have further found in my review of surrounding literature that it coincides with a notable gap in language-vision bilinear fusion for video tasks. Though I attribute the poor performance of BLP in languague-video fusion to problems of temporal alignment and sheer inefficiency-of-scale, I find the lack of work pointing to this relative under-adoption quite perplexing. Though it might be natural to silently decide 'What is the point of bilinear fusion when attention in transformer models work so well?', does this hint at a problem in the way that we view or discuss our methodology? Is it perhaps presumptive to claim that a bilinear operation (mathematically ingenious though it may be) increased VQA performance thanks to *'richer multimodal representation'*? Furthermore, if we are so focused on the empirical performance of our methodology on our datasets (even though this is quite naturally our main priority), does this attitude not sometimes cause us to overlook something interesting? Indeed, I have noticed parallels in the application of bilinear models and multimodal

fusion more generally to the 'two-stream' theory of vision, and dual coding theory. I therefore hope that my work draws attention to what can often goes unsaid in our collective mission for greater empirical performance on multimodal tasks.

7.3.3 Chapter 5

As I entered the latter half of my studies, I prioritised finding a more ambitious way to try and harmonise visual contributions with their language complements: I left the constraints of language-vision QA datasets behind for a time and focused on exploring the vision parallel to the highly-successful language modelling training paradigm. Naturally the most immediate questions I contend with in this work are: "what are the similarities, differences, and challenges of potential visual parallels to language modelling?". Perhaps the most obvious difference is that the prediction in visual generation here in generative learning is on a dense pixel space as opposed to a sequential series of text tokens in a vocabulary. I found from both my work and the surrounding literature (reviewed in Section 5.2.1) that this visual scenario is still challenging in more 'primitive' tasks, compared to the prediction quality in modern language modelling. I find that it is extremely difficult to sustain a realistic visual simulation even a few frames into the future without obvious indication that it is artificial *i.e.* failing some manner of visual Turing-test. Despite these obvious visual inconsistencies, I show that an understanding of the underpinning physical laws can be adequately induced *e.g.* gravity and bouncing is clearly understood even if shape of object might shimmer and distort slightly. Furthermore, my results show that this visual pretraining is beneficial on 'downstream' test-tasks, serving as a promising proof-of-concept that generative pretraining can yield more useful visual features on a variety of future tasks in the field. Though I am able to achieve these promising results by gaining distance from the dataset and modelling challenges in VQA and video-QA, such abstracted work is, by nature, less directly applicable to current language-vision tasks of interest. I believe that the answer to the question "what are the barriers to applying this methodology to multimodality in the future?" will be answered in the not-so-distant future by the very new multimodal diffusion models. I have used my modest resources to generate simple datasets on

reasonably sized CNN and transformer models, aiming to carefully but unequivocally demonstrate the promise of generative visual pretraining. However, given the massive text-image datasets and computational resources invested in training this diffusion process, I believe the fundamental step in diffusion decoder models —the incremental 'de-noising' of image representations— could be repurposed to work temporally instead, paralleling my generative work in visual modelling in a popular and powerful benchmark that has already proven highly effective in multimodal deep learning.

7.3.4 Chapter 6

My finishing work in Chapter 6 returns to multimodal QA —benefiting from the rapid advancement of the field during my previous study— aiming to unify the themes of my work with design decisions informed by my previous findings: trading the novelty and complexity of video-QA benchmarks for the more thoroughly improved text-bias mitigation in the more established VQA benchmarks (which conveniently has improved substantially since the beginning of my studies); leveraging state-of-the-art text-image transformer models to overcome modelling inadequacies as best I can; applying neurologically-inspired (initially considered in Chapter 4) and dataset/model-invariant labelling methodology. I believe that if my labelling scheme proposed in Chapter 6 could be refined to demonstrate consistent positive results, then it could have a significant impact on the field: the answer vocabulary for ***all*** such current and future classification datasets would be able to adopt it as the neurolinguistic scores of similarity are gathered from human responses to words, not linked with any specific dataset. However, the approaches that I have tried thus far have not yet managed to yield consistent VQA improvements. Nevertheless, I am content with my work in this chapter: I believe that it satisfyingly unifies the guiding principals and research themes of my thesis thusfar; I argue that the potential reward of improving multimodal classification labelling schemes (in a dataset-invariant manner) is worth the risk; my approach is significantly distinguished from the few existing similar approaches by leveraging neurolinguistic scores and not relying on correlations and bias of machine learning features or datasets;

and perhaps most importantly, there are many yet-untested avenues left to refine and improve my labelling approach in hope of its ultimate goal. In my opinion, the most immediate and promising next-step would be to secure more neurolinguistic word similarity scores. Collecting such scores could be reasonably achieved with multidisciplinary collaboration for even a modest research project, and would allow scores to be collected to specifically fill in common gaps in VQA answer vocabularies. Though I have done my best to improve relative quality of my labelling scheme, I needed to concede a significant portion of the dataset to do so. I would suggest cleverly filling in this labelling gap through targeted data collection to be the most promising next step of my approach.

7.4 Limitations and Future Work

Though I have explored limitations and future works in my contribution chapters in isolation, In the following sections, I discuss them more thoroughly and in the context the thesis as a whole.

7.4.1 Chapter 3: IEM Subsets Are Very Small

The subsets of TVQA that my IEM metric from Chapter 3 isolates are an order of magnitude smaller than the full dataset. Such a size reduction severely limits the applicability of the datasets generated. The small size subsets for any one dataset *could* be offset by combining IEM subsets from multiple different datasets. However, while the yielded subsets remain very small, it would require a large number of datasets to offset this problem.

7.4.2 Chapter 3: IEM Works Better on More Accurate Models

The IEM subsets should theoretically become more accurate if they are collected using a higher quality model. It therefore follows that the subsets I propose are likely to be at least somewhat inaccurate, as the random answer selections of the

imperfect TVQA model I use will inevitably corrupt the 'modality-reliant subsets'. In this way, IEM is even less useful if the answer vocabulary is small, as the chance of any specific corruption increases. Future work choosing to utilise my proposed subsets or IEM methodology should revisit datasets with increasingly more accurate models, and should be aware that a larger answer vocabulary may be beneficial for avoiding accidental corruptions.

7.4.3 Hard to Escape Modality Bias

While Chapter 4 aims to shift focus from dataset bias to multimodal fusion techniques, in practice I found it hard to test the hypothesis with the influence of modality bias present in the video-QA datasets. As explored in Chapter 2, modality bias in video-QA datasets is much less thoroughly explored than that in VQA. As interest grows in 'harder' video datasets and computational resources increase to meet this challenge, it is very important that thorough modality bias analysis is undertaken (*e.g.* a TVQA-CP dataset).

7.4.4 Neurological Inspirations Can Be Difficult to Realise

Future work may consider my propositions as detailed in Section 4.8.2: 'two-stream-style' cross talk for 'style' and 'content' processing; or grouping features as either 'verbal' or 'non-verbal' modalities to reflect the idea of logogens and imagens. It should be noted however that though neurological inspirations may be fine as initial motivation for new methodology, my difficulties in Chapter 6 imply that even relatively intuitive and seemingly well-resourced ideas can lack sufficient neurological ground truth to properly test.

7.4.5 Chapter 5: Scale

My visual modelling approach does not match the sheer scale of modern language modelling. Modern language models can train on a huge amount of data due to the comparative ease of collecting, storing, and tokenising raw language. Such language models are often trained with substantial computational resources. By directly

training on and outputting raw images, which are much larger representations than language tokens, I can't yet approach the scale of large pretrained modern language models with the resources available to me.

7.4.6 Chapter 5: Dataset Complexity

Larger and more complicated visual dynamics datasets are unexplored in Chapter 5. Though I *do* verify the pretraining on HMDB-51 and MOCAP, the vista between these results and the longer term predictive power of simpler datasets highlight the need for a more measured and careful probing approach to simpler visual tasks first.

7.4.7 Chapter 5: Long Term Visual Prediction

I don't propose a *direct* solution in Chapter 5 for the tendency of visual models to suffer in long term predictive power. Though there is work primarily aiming to address this problem (surveyed by [150] in their Section 2.4). I instead highlight that simpler physical simulations hold much longer under current loss functions and suggest that loss functions may not always be to blame for poor generative performance. I argue that sometimes the blame instead lies with capacity of current models to handle the quality and scale of available datasets.

7.4.8 Chapter 5: More Nuanced Predictive Training Strategy

Though I parallel language modelling, I only predict the final frame of a sequence, and do not parallel more intricate language modelling training strategies, *e.g.* bidirectional token prediction for words in the middle of a sentence. Although more nuanced training methods for video has been explored on techniques from several years ago [164], and more recently for generative pretraining with *images* [28], I am able to present strong pretraining results without using such nuanced approaches in this *video-based* study. Regardless, this is an interesting line of research for future work.

7.4.9 Chapter 6: More Complete Answer Similarity Vocabularies

My results indicate that my proposed neurolinguistic similarity scores need to be much more thoroughly 'complete' before my hypothesis 'does human-like behaviour benefit a VQA context' can be properly tested. Future research should look to 'complete' my sparse answer similarity tensors by either collecting more neurolinguistic scores themselves, or finding a more elegant way to give an 'overly specific' VQA answer a score *e.g.* by considering the concreteness of the *subject* of the answer.

7.4.10 Chapter 6: Better Suited to Better Models?

I find that the more powerful METER framework suffers much less performance degradation than the older LXMERT model. It *could* be that my labelling scheme performs better on more powerful models. Future research with access to better computational resources could test this hypothesis.

7.4.11 Chapter 6: Setting Highly Unrelated Answer Scores to *Below* 0

Given the chance to extend the work in this thesis myself, one of the two very next things I would try is extending the concept of 'norm clipping' by setting the similarity scores for highly dissimilar answers to *below 0* in an effort to test the alternative hypothesis 'is it helpful for VQA models to *not* behave as humans *do not*'.

7.4.12 The Next Frontier? Text-to-Image Models

In the time I have spent working on this thesis, some of the newest and most disruptive frontiers in multimodal research have been: breakthroughs in modality bias; and the rise of multimodal transformers. However, most recently, the largest and newest of these multimodal transformers have made monumental strides in the

previously challenging text-to-image generation task *e.g.* DALL-E 2[2], GLIDE [148], CM3 [1]. Given their unprecedented popularity and widespread appeal amongst the general public, the use of these models may come to dominate the interest of multimodal research in the near future. The priorities for bias in multimodality may experience a shift away from language bias as it has been addressed over the past few years, and towards the 'social' biases derived from human assumptions that these models exhibit [30].

7.4.13 Visual Modelling in Multimodal Diffusion Models?

Perhaps the most natural application of my work to these powerful new diffusion models would be to adapt the de-noising step in the decoder to work temporally, *i.e.* given an input text condition, instead of gradually removing noise from an image feature, instead train the model by gradually apply 'time steps' to an input image, potentially inducing an understanding of visual dynamics in a powerful and popular established multimodal benchmark. Were I to continue my work in this thesis myself, this would be the second of the two ideas I would explore next.

[2] https://openai.com/dall-e-2/

Bibliography

[1] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. Cm3: A causal masked multimodal model of the internet. *CoRR*, abs/2201.07520, 2022. URL https://arxiv.org/abs/2201.07520. 191

[2] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1203. URL https://aclanthology.org/D16-1203. 29

[3] Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset. *ArXiv*, abs/1704.08243, 2017. 21, 24

[4] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018. 21, 24, 29, 56, 150

[5] Shotaro Akaho. A kernel method for canonical correlation analysis. *Proceedings of the International Meeting of the Psychometric Society (IMPS 2001); Osaka*, 4, 10 2001. 79

[6] Huda AlAmri, Vincent Cartillier, Abhishek Das, Jue Wang, Stefan Lee, Pip Anderson, Irfan Essa, Devi Parikh, Dhruv Batra, Anoop Cherian, Tim K. Marks, and Chiori Hori. Audio visual scene-aware dialog. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7550–7559, 2019. 16, 18

[7] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image

captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 42, 58, 74

[8] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, 2013. 78

[9] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015. 21, 22, 56

[10] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:423–443, 2019. 43, 55, 61, 78, 83, 89

[11] Ian Begg. Recall of meaningful phrases. *Journal of Verbal Learning and Verbal Behavior*, 11(4):431–439, 1972. ISSN 0022-5371. doi: https://doi.org/10.1016/S0022-5371(72)80024-0. 93

[12] Hedi Ben-younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2631–2639, 2017. 60, 66, 88

[13] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8102–8109, 2019. 55, 60, 68, 88

[14] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, 2000. 103

[15] Jeff Bezemer and Gunther Kress. Writing in multimodal texts: A social semiotic account of designs for learning. *Written Communication - WRIT COMMUN*, 25:166–195, 04 2008. doi: 10.1177/0741088307313177. 92

[16] Jeffrey R. Binder, Chris F. Westbury, Kristen A. McKiernan, Edward T. Possing, and David A. Medler. Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, 17:905–917, 2005. 149

[17] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *NeurIPS*, volume 33, pages 1877–1901, 2020. 98

[18] Marc Brysbaert, Amy Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46, 10 2013. doi: 10.3758/s13428-013-0403-5. 96, 143

[19] John Charles Butcher. A history of runge-kutta methods. *Applied numerical mathematics*, 20(3):247–260, 1996. 110

[20] Rémi Cadène, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering. In *NeurIPS*, 2019. xix, 29, 30, 39, 56, 57, 58

[21] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. ISBN 978-3-030-58452-8. 98

[22] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In *ICCV*, 2019. 98

[23] Santiago Castro, Mahmoud Azab, Jonathan C. Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. Lifeqa: A real-life dataset for video question answering. In *LREC*, 2020. 16, 19

[24] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. In *NAACL-HLT*, 2018. 29, 32

[25] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In Peter Widmayer, Stephan Eidenbenz, Francisco Triguero, Rafael Morales, Ricardo Conejo, and Matthew Hennessy, editors, *Automata, Languages and Programming*, pages 693–703, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-45465-6. 64

[26] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL 2011*, 2011. 17, 71, 73

[27] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, 2020. 29, 32

[28] Mark Chen, Alec Radford, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020. 98, 99, 102, 103, 135, 189

[29] Fan Chenyou. Hme-videoqa, 2019. URL github.com/fanchenyou/HME-VideoQA. 76, 81

[30] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *CoRR*, abs/2202.04053, 2022. URL https://arxiv.org/abs/2202.04053. 191

[31] Shih-Han Chou, Wei-Lun Chao, Min Sun, and Ming-Hsuan Yang. Visual question answering on 360° images. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1596–1605, 2020. 70

[32] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural*

Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4069–4082, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1418. URL https://aclanthology.org/D19-1418. 29, 30

[33] James Clark and Allan Paivio. Extensions of the paivio, yuille, and madigan (1968) norms. https://link.springer.com/article/10.3758/BF03195584#SecESM1, 2004. 96, 143

[34] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965. 65

[35] Sebastian J. Crutch and Elizabeth K. Warrington. Abstract and concrete concepts have structurally different representational frameworks. *Brain : a journal of neurology*, 128 Pt 3:615–27, 2005. 143, 149

[36] Jifeng Dai, Haozhi Qi, Y. Xiong, Y. Li, Guodong Zhang, H. Hu, and Yichen Wei. Deformable convolutional networks. *ICCV*, pages 764–773, 2017. 103

[37] Anubrata Das, Samreen Anjum, and Danna Gurari. Dataset bias: A case study for visual question answering. *Proceedings of the Association for Information Science and Technology*, 56, 2019. 29

[38] Emmanuel de Bezenac, Arthur Pajot, and Patrick Gallinari. Deep learning for physical processes: Incorporating prior scientific knowledge. In *ICLR*, 2018. 104

[39] Lieven De Lathauwer. Decompositions of a higher-order tensor in block termspart i: Lemmas for partitioned matrices. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1022–1032, 2008. 68

[40] Lieven De Lathauwer. Decompositions of a higher-order tensor in block termspart ii: Definitions and uniqueness. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1033–1066, 2008. 68

[41] Lieven De Lathauwer and Dimitri Nion. Decompositions of a higher-order tensor in block termspart iii: Alternating least squares algorithms. *SIAM journal on Matrix Analysis and Applications*, 30(3):1067–1083, 2008. 68

[42] Tonmoay Deb, Akib Sadmanee, Kishor Kumar Bhaumik, Amin Masud Ali, M. Ashraful Amin, and A. K. M. Mahbubur Rahman. Variational stacked local attention networks for diverse video captioning. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2493–2502, 2022. 61, 71

[43] Haoyang Deng, Jun Kong, Min Jiang, and Tianshan Liu. Diverse features fusion network for video-based action recognition. *Journal of Visual Communication and Image Representation*, 77:103121, 2021. ISSN 1047-3203. doi: https://doi.org/10.1016/j.jvcir.2021.103121. 61, 71, 87

[44] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. 41, 102

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 39, 42

[46] Rupali P. Dhond, Thomas Witzel, Anders M. Dale, and Eric Halgren. Spatiotemporal cortical dynamics underlying abstract and concrete word reading. *Human brain mapping*, 28 4:355–62, 2007. 149

[47] A. Domnguez. A history of the convolution operation [retrospectroscope]. *IEEE Pulse*, 6(1):38–49, Jan 2015. ISSN 2154-2287. doi: 10.1109/MPUL.2014.2366903. 64

[48] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *NeurIPS*, 2019. 102

[49] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *NIPS*, 2016. 103

[50] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021. 98

[51] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2111.02387. 141, 157, 218

[52] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1999–2007, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. doi: 10.1109/CVPR.2019.00210. 73, 81

[53] Chenyou Fan. Egovqa - an egocentric video question answering benchmark dataset. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 15, 16, 19, 38, 72, 73, 76

[54] Daniele Fanelli. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904, 2012. 172

[55] H. Farazi and Sven Behnke. Frequency domain transformer networks for video prediction. *ArXiv*, abs/1903.00271, 2019. 98

[56] Hafez Farazi, Jan Nogga, and Sven Behnke. Local frequency domain transformer networks for video prediction. *arXiv preprint arXiv:2105.04637*, 2021. 98

[57] Michael Friendly, Patricia Franklin, David Hoffman, and David Rubin. The toronto word pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods and Instrumentation*, 14:375–399, 09 1982. doi: 10.3758/BF03203275. 96, 144

[58] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1044. URL https://www.aclweb.org/anthology/D16-1044. 43, 55, 65, 82, 88

[59] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. *CoRR*, abs/1505.05612, 2015. URL http://arxiv.org/abs/1505.05612. 21

[60] Jiyang Gao, Runzhou Ge, Kan Chen, and Ramakant Nevatia. Motion-appearance co-memory networks for video question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6576–6585, 2018. 73, 74

[61] Lianli Gao, Pengpeng Zeng, Jingkuan Song, Yuan-Fang Li, Wu Liu, Tao Mei, and Heng Tao Shen. Structured two-stream attention network for video question answering. In *AAAI*, 2019. 61, 70, 87

[62] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 317–326, 2016. 55, 60, 61, 64, 84

[63] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546. 29

[64] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016. doi: 10.1109/TKDE.2016.2545658. 150

[65] Daniela Gerz, Ivan Vulic, Felix Hill, Roi Reichart, and Anna Korhonen. Simverb-3500: A large-scale evaluation set of verb similarity. *CoRR*, abs/1608.00869, 2016. URL http://arxiv.org/abs/1608.00869. 159

[66] Rohit Girdhar and Deva Ramanan. CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. In *ICLR*, 2020. 29, 35

[67] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. In *EMNLP (1)*, pages 878–892, 2020. URL https://doi.org/10.18653/v1/2020.emnlp-main.63. 21, 26, 29, 32

[68] Melvyn A Goodale. How (and why) the visual control of action differs from visual perception. *Proceedings of the Royal Society B: Biological Sciences*, 281 (1785):20140337, 2014. 63

[69] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992. 62, 63, 90

[70] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 21, 23, 29, 32, 56, 150

[71] Gabriel Grand and Yonatan Belinkov. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. *Proceedings of the Second Workshop on Shortcomings in Vision and Language, Association for Computational Linguistics*, 2019. 29, 30

[72] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005. 43

[73] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. *2013 IEEE International Conference on Computer Vision*, pages 2712–2719, 2013. 73

[74] Lieven DE LATHAUWER Guillaume OLIKIER, Pierre-Antoine ABSIL. Tensor approximation by block term decomposition. *Dissertation*, 2017. 66

[75] W. Guo, J. Wang, and S. Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019. 55, 80, 83, 89

[76] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, Ji Zhang, and A. Bimbo. Adavqa: Overcoming language priors with adapted margin cosine loss. In *IJCAI*, 2021. 29

[77] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Qi Tian, and Min Zhang. Loss re-scaling vqa: Revisiting the language prior problem from a class-imbalance view. *IEEE Transactions on Image Processing*, 31:227–238, 2022. 29, 33

[78] Zhicheng Guo, Jiaxuan Zhao, Licheng Jiao, Xu Liu, and Lingling Li. Multi-scale progressive attention network for video question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021. 76

[79] Akshay Kumar Gupta. Survey of visual question answering: Datasets and techniques. *CoRR*, abs/1705.03865, 2017. URL http://arxiv.org/abs/1705.03865. 20

[80] Vikrant Gupta, Badri N. Patro, Hemant Parihar, and Vinay P. Namboodiri. Vquad: Video question answering diagnostic dataset. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 282–291, 2022. 16, 18, 29, 34

[81] John T. Guthrie. Comprehension and teaching : research reviews. In *Psychology*, 1981. 149

[82] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. Greedy gradient ensemble for robust visual question answering. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1564–1573, 2021. 29, 30

[83] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016. 41, 74, 105

[84] Michael Heilman and Noah A. Smith. Question generation via overgenerating transformations and ranking. In *CMU*, 2009. 71

[85] Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December 2015. doi: 10.1162/COLI_a_00237. URL https://aclanthology.org/J15-4004. 96, 144, 147

[86] Yusuke Hirota, Noa García, Mayu Otani, Chenhui Chu, Yuta Nakashima, Ittetsu Taniguchi, and Takao Onoye. A picture may be worth a hundred words for visual question answering. *ArXiv*, abs/2106.13445, 2021. 29

[87] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735. 43

[88] A. Hor and D. Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, pages 2366–2369, 2010. 113

[89] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4): 321–377, 1936. 79

[90] Feiyan Hu, Eva Mohedano, Noel E. O'Connor, and Kevin McGuinness. Temporal bilinear encoding network of audio-visual features at low sampling rates. In *VISIGRAPP*, 2021. 61, 71, 87

[91] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, 2019. 21, 25, 29

[92] Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1383–1391, 2015. 72

[93] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1359–1367, 2017. 15, 16, 17, 37, 70, 72, 76

[94] Bhavan Jasani, Rohit Girdhar, and Deva Ramanan. Are we asking the right questions in movieqa? *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1879–1882, 2019. 16, 29, 35

[95] Marc Jeannerod and Pierre Jacob. Visual cognition: a new look at the two-visual systems model. *Neuropsychologia*, 43(2):301–312, 2005. 91

[96] Khaled Jedoui, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Deep bayesian active learning for multiple correct outputs. *ArXiv*, abs/1912.01119, 2019. 150

[97] Jingjing Jiang, Zi yi Liu, Yifan Liu, Zhixiong Nan, and Nanning Zheng. X-ggm: Graph generative modeling for out-of-distribution generalization in visual question answering. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 29, 33

[98] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 18, 21, 25

[99] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *ICCV*, pages 1983–1991, 10 2017. doi: 10.1109/ICCV.2017.217. 21, 23, 29

[100] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5648–5656, 2018. doi: 10.1109/CVPR.2018.00592. 21, 25

[101] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1mz0OyDz. 21, 25

[102] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 42

[103] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Estimating semantic structure for the vqa answer space. *ArXiv*, abs/2006.05726, 2020. xviii, 29, 142, 151, 174, 176, 177

[104] Aisha Urooj Khan, Amir Mazaheri, N. Lobo, and M. Shah. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. In *FINDINGS*, 2020. 76

[105] Khushboo Khurana and Umesh Deshpande. Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: A comprehensive survey. *IEEE Access*, 9:43799–43823, 2021. 16

[106] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *The 5th International Conference on Learning Representations*, 2017. 55, 60, 65, 88, 89

[107] Junyeong Kim, Minuk Ma, Kyungsu Kim, S. Kim, and C. Yoo. Progressive attention memory network for movie story question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8329–8338, 2019. 61, 70, 87

[108] K Kim, C Nan, MO Heo, SH Choi, and BT Zhang. Pororoqa: Cartoon video series dataset for story understanding. In *Proceedings of NIPS 2016 Workshop on Large Scale Computer Vision System*, volume 19, 2016. 16, 19, 37

[109] Shu Kong and Charless C. Fowlkes. Low-rank bilinear pooling for fine-grained classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7025–7034, 2017. 55

[110] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016. 21, 22, 42

[111] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011. doi: 10.1109/ICCV.2011.6126543. 108, 111

[112] Mingrui Lao, Yanming Guo, Yu Liu, Wei Chen, Nan Pu, and Michael S. Lew. From superficial to deep: Language bias driven curriculum learning for visual question answering. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 29, 33

[113] Mingrui Lao, Yanming Guo, Yu Liu, and Michael S. Lew. A language prior based focal loss for visual question answering. *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 29, 31

[114] Hung Le, Doyen Sahoo, Nancy F. Chen, and S. Hoi. Bist: Bi-directional spatio-temporal reasoning for video-grounded dialogues. In *EMNLP*, 2020. 70, 76

[115] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. 99, 111

[116] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 15, 16, 17, 38, 40, 42, 70, 72, 76

[117] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.730. URL https://www.aclweb.org/anthology/2020.acl-main.730. 38, 45, 48

[118] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*, 2021. 86

[119] Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *ArXiv*, 2022. URL https://arxiv.org/pdf/2206.03428.pdf. 29, 35

[120] Ofir Levy and Lior Wolf. Live repetition counting. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3020–3028, 2015. 72

[121] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10312–10321, 2019. 29

[122] Xiangpeng Li, L. Gao, Xuanhan Wang, W. Liu, Xing Xu, Heng Tao Shen, and Jingkuan Song. Learnable aggregating net with diversity learning for video question answering. *Proceedings of the 27th ACM International Conference on Multimedia*, 2019. 61, 70, 87

[123] Yuncheng Li, Yale Song, Liangliang Cao, Joel R. Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4641–4650, 2016. 17, 72, 103

[124] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *ICLR*, 2021. 104

[125] Junwei Liang, Lu Jiang, L. Cao, Yannis Kalantidis, L. Li, and A. Hauptmann. Focal visual-text attention for memex question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1893–1908, 2019. 61, 70, 87

[126] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *EMNLP*, 2020. 29, 32

[127] T. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1457, Dec 2015. doi: 10.1109/ICCV.2015.170. 61, 63, 71, 91, 92

[128] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 22

[129] Fei Liu, Jing Liu, Richang Hong, and Hanqing Lu. Question-guided erasing-based spatiotemporal attention learning for video question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 61, 70, 87

[130] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019. 44

[131] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *NAACL*, 2019. 114

[132] Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proceedings of The Third Workshop on Representation Learning for NLP*, 2018. 96, 143

[133] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 105

[134] Xiang Long, Chuang Gan, Gerard de Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen. Multimodal keyless attention fusion for video classification. In *AAAI*, 2018. 86, 87

[135] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 289297. Curran Associates Inc., 2016. ISBN 9781510838819. 63

[136] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron C Courville, and Christopher Joseph Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Computer Vision and Pattern Recognition (CVPR)*,

2017. URL http://openaccess.thecvf.com/content_cvpr_2017/papers/Maharaj_A_Dataset_and_CVPR_2017_paper.pdf. 16, 17, 37

[137] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 16821690, Cambridge, MA, USA, 2014. MIT Press. 21, 150

[138] Gaurav Manek and J. Kolter. Learning stable deep dynamics models. *NeuIPS*, 2020. 104

[139] Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Černocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, 2010. 103

[140] Tomas Mikolov, Kai Chen, Gregory S Corrado, and Jeffrey A Dean. Computing numeric representations of words in a high-dimensional space, May 19 2015. US Patent 9,037,464. 35

[141] A David Milner and Melvyn A Goodale. Two visual systems re-viewed. *Neuropsychologia*, 46(3):774–785, 2008. 63

[142] Anthony David Milner. How do the two visual streams interact with each other? *Experimental Brain Research*, 235:1297 – 1308, 2017. xvi, 62, 63, 90, 91, 92

[143] David Milner and Mel Goodale. *The visual brain in action*, volume 27. OUP Oxford, 2006. 63

[144] Ana Mlinari, Martina Triplat Horvat, and Vesna upak Smoli. Dealing with the positive publication bias: Why you should really publish your negative results. *Biochemia Medica*, 27, 2017. 172

[145] John Morton. *Facilitation in Word Recognition: Experiments Causing Change in the Logogen Model*, pages 259–268. Springer US, Boston, MA, 1979. ISBN 978-1-4684-0994-9. doi: 10.1007/978-1-4684-0994-9_15. URL https://doi.org/10.1007/978-1-4684-0994-9_15. 92

[146] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. Marioqa: Answering questions by watching gameplay videos. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2886–2894, 2017. 16, 19

[147] Douglas Nelson, Cathy Mcevoy, and Thomas Schreiber. The university of south florida word association, rhyme, and word fragment norms. http://w3.usf.edu/FreeAssociation/, 1998. 96, 143, 147

[148] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 191

[149] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 29, 31

[150] Sergiu Oprea, P. Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, S. Orts-Escolano, J. García-Rodríguez, and Antonis A. Argyros. A review on deep learning techniques for video prediction. *IEEE TPAMI*, PP, 2020. 98, 99, 103, 126, 127, 189

[151] Ahmed Osman and Wojciech Samek. Drau: Dual recurrent attention units for visual question answering. *Computer Vision and Image Understanding*, 185: 24–30, 2019. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2019.05.001. 65

[152] Allan Paivio. *Imagery and verbal processes*. Psychology Press, 2013. 62, 90, 92

[153] Allan Paivio. Intelligence, dual coding theory, and the brain. *Intelligence*, 47: 141158, 11 2014. doi: 10.1016/j.intell.2014.09.002. 62, 90, 92, 94

[154] Allan Paivio and Wallace Lambert. Dual coding and bilingual memory. *Journal of Verbal Learning and Verbal Behavior*, 20(5):532–539, 1981. ISSN 0022-5371. doi: https://doi.org/10.1016/S0022-5371(81)90156-0. 93

[155] Wen-Feng Pang, Qian-Hua He, Yong-jian Hu, and Yan-Xiong Li. Violence detection in videos based on fusing visual and audio information. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2260–2264, 2021. doi: 10.1109/ICASSP39728.2021.9413686. 61, 71, 87

[156] Devshree Patel, Ratnam Parikh, and Yesha Shastri. Recent advances in video question answering: A review of datasets and methods. In *ICPR Workshops*, 2020. 16

[157] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing*, 27:5585–5599, 2018. 78, 83

[158] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162. 42

[159] Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP*, 2018. 114

[160] Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *KDD*, 2013. 64

[161] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019. 102

[162] Ruslan Rakhimov, Denis Volkhonskiy, A. Artemov, D. Zorin, and Evgeny Burnaev. Latent video transformer. In *VISIGRAPP*, 2021. 98

[163] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *NeurIPS*, 2018. 29, 30

[164] Marc'Aurelio Ranzato, Arthur Szlam, Joan Bruna, Michaël Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *ArXiv*, abs/1412.6604, 2014. 98, 102, 103, 189

[165] Jamie Reilly and Jacob Kean. Formal distinctiveness of high- and low-imageability nouns: Analyses and theoretical implications. *Cognitive science*, 31:157–68, 02 2007. doi: 10.1080/03640210709336988. 96, 144

[166] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *ArXiv*, abs/1505.02074, 2015. 21, 22

[167] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. 42

[168] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 105

[169] Graham G Scott, Anne Keitel, Marc Becirspahic, Patrick J ODonnell, and Sara C Sereno. The glasgow norms: Ratings of 5,500 words on 9 scales, Aug 2017. URL osf.io/42s65. 96, 143

[170] Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2591–2600, 2019. 29, 36

[171] Ahjeong Seo, Gi-Cheon Kang, J. Park, and Byoung-Tak Zhang. Attend what you need: Motion-appearance synergistic networks for video question answering. In *ACL/IJCNLP*, 2021. 76

[172] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *ICLR*, abs/1611.01603, 2017. URL http://arxiv.org/abs/1611.01603. 42, 74, 84

[173] Fei Sha, Hexiang Hu, and Wei-Lun Chao. Cross-dataset adaptation for visual question answering. In *CVPR*, pages 5716–5725, 06 2018. doi: 10.1109/CVPR.2018.00599. 29, 34

[174] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. Deep learning for precipitation nowcasting: A benchmark and a new model. In *NeurIPS*, volume 30, 2017. 104

[175] Robik Shrestha, Kushal Kafle, and Christopher Kanan. A negative case analysis of visual grounding methods for vqa. In *ACL*, 2020. 29, 36

[176] Qingyi Si, Zheng Lin, Mingyu Zheng, Peng Fu, and Weiping Wang. Check it again: Progressive visual question answering via visual entailment. In *ACL/IJCNLP (1)*, pages 4101–4110, 2021. URL https://doi.org/10.18653/v1/2021.acl-long.317. 29, 34

[177] Agnes Sianipar, Pieter Groenestijn, and Ton Dijkstra. Affective meaning, concreteness, and subjective frequency norms for indonesian words. *Frontiers in Psychology*, 7:1907, 12 2016. doi: 10.3389/fpsyg.2016.01907. 96, 144

[178] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 74

[179] Khurram Soomro, Amir Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, 12 2012. 103

[180] Nitish Srivastava, Elman Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015. 99, 111

[181] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Learning to recognize actions on objects in egocentric video with attention dictionaries. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2021. 61, 71

[182] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *NIPS*, 2015. 73

[183] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 141, 152, 218

[184] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 16, 37, 70

[185] Joshua B. Tenenbaum and William T. Freeman. Separating style and content with bilinear models. *Neural Comput.*, 12(6):1247–1283, June 2000. ISSN 0899-7667. doi: 10.1162/089976600300015349. URL http://dx.doi.org/10.1162/089976600300015349. 63, 84, 91

[186] Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 580–599, Cham, 2020. Springer International Publishing. 29, 32

[187] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *ArXiv*, abs/2005.09241, 2020. 29, 34

[188] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization in visual question answering. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1397–1407, 2021. 29, 34

[189] Jonathan Tompson, Kristofer Schlachter, Pablo Sprechmann, and Ken Perlin. Accelerating eulerian fluid simulation with convolutional networks. In *ICML*, page 34243433, 2017. 104

[190] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3d: Generic features for video analysis. *ArXiv*, abs/1412.0767, 2014. 74

[191] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966. 66

[192] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *ICLR*, 2019. 113

[193] Joost R. van Amersfoort, Anitha Kannan, Marc'Aurelio Ranzato, Arthur Szlam, Du Tran, and Soumith Chintala. Transformation-based models of video sequences. *ArXiv*, abs/1701.08435, 2017. 103

[194] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, 2017. 102

[195] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 102

[196] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 175

[197] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 106

[198] Jianyu Wang, Bingkun Bao, and Changsheng Xu. Dualvgr: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia*, 2021. 61, 71, 87

[199] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Henge. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 12901296. AAAI Press, 2017. ISBN 9780999241103. 21, 22

[200] Rui Wang, R. Walters, and R. Yu. Meta-learning dynamics forecasting using task inference. *ArXiv*, abs/2102.10271, 2021. 104

[201] Y. Wang, J. Wu, M. Long, and J. B. Tenenbaum. Probabilistic video prediction from noisy data with a posterior confidence. In *CVPR*, 2020. 103

[202] Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. Debiased visual question answering from feature and sample perspectives. In *NeurIPS*, 2021. 29, 30, 31

[203] Michael Wilson. Mrc psycholinguistic database. https://websites.psychology.uwa.edu.au/school/MRCDatabase/mrc2.html, 1988. 143

[204] T. Winterbottom, S. Xiao, A. McLean, and N. Al Moubayed. On modality bias in the tvqa dataset. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. 18, 29, 35, 37, 45, 83

[205] Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. Bilinear pooling in video-qa: empirical challenges and motivational drift from neurological parallels. *PeerJ Computer Science*, 8:e974, 2022. 29, 60

[206] Jiajun Wu, Joseph J. Lim, Hongyi Zhang, Joshua B. Tenenbaum, and William T. Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, 2016. 104

[207] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *Comput. Vis. Image Underst.*, 163:21–40, 2017. 20

[208] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021. xvii, 104, 107

[209] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016. 73, 74

[210] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. *Proceedings of the 25th ACM international conference on Multimedia*, 2017. 15, 16, 17, 71, 76

[211] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 17

[212] Qin Xu, Yiming Mei, Jinpei Liu, and Chenglong Li. Multimodal cross-layer bilinear pooling for rgbt tracking. *IEEE Transactions on Multimedia*, pages 1–1, 2021. doi: 10.1109/TMM.2021.3055362. 61, 71, 87

[213] Hongyang Xue, Zhou Zhao, and Deng Cai. Unifying the video and question attentions for open-ended video question answering. *IEEE Transactions on Image Processing*, 26(12):5656–5666, 2017. doi: 10.1109/TIP.2017.2746267. 16, 17

[214] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers, 2021. 98, 99, 103

[215] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1686–1697, October 2021. 86

[216] Chao Yang, Su Feng, Dongsheng Li, Huawei Shen, Guoqing Wang, and Bin Jiang. Learning content and context with language bias for visual question answering. *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 29, 30

[217] Jianing Yang, Yuying Zhu, Yongxin Wang, Ruitao Yi, Amir Zadeh, and Louis-Philippe Morency. What gives the answer away? question answering bias analysis on video qa datasets. *ArXiv*, abs/2007.03626, 2020. 16, 18, 29, 35

[218] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. Bert representations for video question answering. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1545–1554, 2020. 45, 48

[219] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–29, 2015. 58

[220] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. Video question answering via attribute-augmented attention network learning. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017. 16, 19, 38, 73

[221] Lydia Ting Sum Yee. Valence, arousal, familiarity, concreteness, and imageability ratings for 292 two-character chinese nouns in cantonese speakers in hong kong. *PLoS ONE*, 12, 2017. 96, 143

[222] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxYzANYDB. 16, 18, 29, 34

[223] M. Yilmaz and A. Tekalp. Dfpn: Deformable frame prediction network. *ArXiv*, abs/2105.12794, 2021. 103

[224] Xuwang Yin and Vicente Ordonez. Obj2text: Generating visually descriptive language from object layouts. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 177–187, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1017. URL https://www.aclweb.org/anthology/D17-1017. 42, 46

[225] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006. 21

[226] Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B14TlG-RW. 42, 74

[227] Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. Visual madlibs: Fill in the blank image generation and question answering. *ArXiv*, abs/1506.00278, 2015. 21

[228] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017. 60, 66, 85, 86, 88

[229] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29:5947–5959, 2018. 43, 60, 68, 88, 89

[230] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8799–8809, 2019. 16, 19

[231] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 105

[232] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. *ArXiv*, abs/1611.04021, 2017. 73, 74

[233] Liyang Zhang, Shuaicheng Liu, Donghao Liu, Pengpeng Zeng, Xiangpeng Li, Jingkuan Song, and Lianli Gao. Rich visual knowledge-based augmentation network for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4362–4373, 2021. doi: 10.1109/TNNLS.2020.3017530. 29, 33

[234] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5014–5022, 2016. doi: 10.1109/CVPR.2016.542. 21, 23, 29, 32

[235] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 113

[236] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *ArXiv*, abs/1512.02167, 2015. 63

[237] Hengshun Zhou, Jun Du, Yuanyuan Zhang, Qing Wang, Qing-Feng Liu, and Chin-Hui Lee. Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2617–2629, 2021. doi: 10.1109/TASLP.2021.3096037. 61, 71, 87

[238] Yufan Zhou, Haiwei Dong, and Abdulmotaleb El Saddik. Deep learning in next-frame prediction: A benchmark review. *IEEE Access*, 8:69273–69283, 2020. doi: 10.1109/ACCESS.2020.2987281. 98, 103

[239] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 113

[240] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004, 2016. 21, 22

[241] Yeyun Zou and Qiyu Xie. A survey on vqa: Datasets and approaches. *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, pages 289–297, 2020. 20

CHAPTER A

Appendix

A.1 Chapter 5: Further Training Details

All inputs are preprocessed to grayscale images of dimension 64x64. The pixel values are then normalised to $\in [0,1]$ by dividing all pixels by 255. A pixel-wise sigmoid activation is used to create the output image and calculate metrics and losses. The visualisations of outputs in this chapter are the result of multiplying the pixel values of the outputs by 255 and rounding them.

A.1.1 Fully Convolutional CNN

'Depth' refers to the number of broad convolution steps in both the downwards and upwards directions. The models are of depth 3 (see Figure 5.3) and the convolutions have both a kernel size of 3x3 and padding of 1. I found that varying the depth and kernel size did not yield much difference in results and I have accordingly fixed them at their highest performing values. The linear probes (for probing frozen models) are created from:

1. The outputs of the final layer.

2. The outputs of each double convolution layer in both the downward and upward direction.

A.1.2 Patch Transformer

As previously described, the patch transformer is formed from 2 objects from the SegFormer repository[1]:

1. As an encoder, the MixVisionTransformer object with: in_chans = number of input frames, img_size = 64, sr_ratios = [1,1,1,1] (scale reduction ratios).

2. As a decoder, the SegFormerHead object with: feature_strides = [4,8,16,32], in_channels = [64,128,256,512], channels = 128, num_classes = 16, in_index = [0, 1, 2, 3], decoder_params = {"embed_dim":256}, dropout_ratio = 0.1, align_corners = False. The linear probes (for probing frozen models) are created from: **1)** the outputs of the encoder, **2)** the resized and reshaped outputs of each of the linear projection layers in the decoder, **3)** the output 'linear_fuse' convolution module which takes as input the concatenation the outputs described in **2)**.

A.1.3 Learning Rates

This section shows the learning rates used (chosen through a parameter search) for the experimental results in this chapter.

[1]https://github.com/NVlabs/SegFormer

CNN		Img Trans		Patch Trans	
SL1	**SSIM**	**SL1**	**SSIM**	**SL1**	**SSIM**
2D Bouncing $m = 5$					
1e-3	1e-4	1e-5	1e-5	1e-3	1e-3
2D Bouncing $m = 59$					
1e-3	1e-3	1e-5	3e-6	1e-4	1e-4
3D Bouncing $m = 5$					
1e-4	1e-3	3e-6	1e-5	1e-3	1e-3
3D Bouncing $m = 99$					
1e-3	1e-3	3e-6	1e-5	1e-3	1e-3
Roller $m = 5$					
1e-3	1e-3	1e-5	3e-6	1e-3	1e-3
Pendulum $m = 5$					
1e-3	1e-4	3e-6	3e-6	1e-3	1e-4
Blocks $m = 49$					
1e-3	1e-3	3e-6	3e-6	1e-3	1e-3
Moon $m = 5$					
1e-2	1e-3	1e-5	3e-6	1e-3	1e-3
MMNIST $m = 5$					
1e-2	1e-2	3e-6	3e-6	5e-4	5e-4
MOCAP $m = 5$					
1e-4	1e-4	1e-5	1e-5	1e-4	1e-4
HDMB51 $m = 5$					
1e-4	1e-4	3e-6	3e-6	1e-4	1e-4

Table A.1: Learning rates used for modelling experiments.

Model Details		CNN	I-Trans	P-Trans
Modelling Dataset = 2D Bounce — Task = Counting Bounces in 59 Frames				
-	Constant Output			
-	Input Image + Linear Layer			
None	Frozen Model + Linear Probes	1e-4	3e-7	3e-6
None	Unfrozen Model + Non-Linear MLP	1e-3	1e-5	1e-4
$m = 59$ SL1	Frozen Model + Linear Probes	5e-6	3e-7	1e-5
$m = 59$ SSIM	Frozen Model + Linear Probes	1e-5	3e-7	1e-5
$m = 59$ SL1	Unfrozen Model + Non-Linear MLP	1e-3	1e-5	1e-3
$m = 59$ SSIM	Unfrozen Model + Non-Linear MLP	1e-3	1e-5	1e-3
Modelling Dataset = 2D Bounce — Task = Gravity, from 5 frames				
-	Constant Output			
-	Input Image + Linear Layer			
None	Frozen Model + Linear Probes	1e-4	1e-5	3e-6
None	Unfrozen Model + Non-Linear MLP	1e-4	3e-6	1e-4
$m = 5$ SL1	Frozen Model + Linear Probes	3e-6	3e-6	1e-6
$m = 5$ SSIM	Frozen Model + Linear Probes	3e-6	3e-6	3e-6
$m = 5$ SL1	Unfrozen Model + Non-Linear MLP	1e-3	1e-5	1e-3
$m = 5$ SSIM	Unfrozen Model + Non-Linear MLP	1e-4	1e-5	1e-3
Modelling Dataset = 3D Bounce — Task = Counting Bounces in 99 frames				
-	Constant Output			
-	Input Image + Linear Layer			
None	Frozen Model + Linear Probes	1e-3	3e-6	2e-5
None	Unfrozen Model + Non-Linear MLP	1e-5	1e-6	1e-4
$m = 99$ SL1	Frozen Model + Linear Probes	1e-5	1e-7	1e-5
$m = 99$ SSIM	Frozen Model + Linear Probes	1e-5	1e-6	1e-5
$m = 99$ SL1	Unfrozen Model + Non-Linear MLP	1e-5	1e-6	1e-4
$m = 99$ SSIM	Unfrozen Model + Non-Linear MLP	1e-4	1e-6	1e-4
Modelling Dataset = Roller — Task = Gravity, from 5 frames				
-	Constant Output			
-	Input Image + Linear Layer			
None	Frozen Model + Linear Probes	1e-3	1e-5	3e-5
None	Unfrozen Model + Non-Linear MLP	1e-4	1e-6	1e-4
$m = 5$ SL1	Frozen Model + Linear Probes	1e-5	1e-6	8e-6
$m = 5$ SSIM	Frozen Model + Linear Probes	1e-5	3e-6	8e-6
$m = 5$ SL1	Unfrozen Model + Non-Linear MLP	5e-3	3e-6	1e-3
$m = 5$ SSIM	Unfrozen Model + Non-Linear MLP	5e-3	1e-7	1e-3

Table A.2: Learning rates of each of the test-task experiments.

Model Details		CNN	I-Trans	P-Trans
Modelling Dataset = Pendulum — Task = Gravity, from 5 frames				
-	Constant Output			
-	Input Image + Linear Layer			
None	Frozen Model + Linear Probes	1e-3	1e-6	1e-5
None	Unfrozen Model + Non-Linear MLP	1e-4	3e-6	1e-4
$m = 5$ SL1	Frozen Model + Linear Probes	1e-5	1e-6	3e-6
$m = 5$ SSIM	Frozen Model + Linear Probes	1e-5	3e-7	3e-6
$m = 5$ SL1	Unfrozen Model + Non-Linear MLP	1e-4	3e-7	1e-3
$m = 5$ SSIM	Unfrozen Model + Non-Linear MLP	1e-3	1e-6	1e-3
Modelling Dataset = Blocks — Task = Mass Ratio, from 49 frames				
-	Constant Output			
-	Input Image + Linear Layer			
None	Frozen Model + Linear Probes	1e-3	3e-6	1e-5
None	Unfrozen Model + Non-Linear MLP	1e-4	3e-6	1e-4
$m = 49$ SL1	Frozen Model + Linear Probes	1e-5	1e-7	6e-6
$m = 49$ SSIM	Frozen Model + Linear Probes	1e-5	3e-7	6e-6
$m = 49$ SL1	Unfrozen Model + Non-Linear MLP	1e-3	3e-6	1e-3
$m = 49$ SSIM	Unfrozen Model + Non-Linear MLP	1e-3	3e-6	1e-3
Modelling Dataset = Moon — Task = Mass, from 5 frames				
-	Constant Output			
-	Input Image + Linear Layer			
None	Frozen Model + Linear Probes	1e-3	1e-6	3e-6
None	Unfrozen Model + Non-Linear MLP	1e-4	3e-7	1e-4
$m = 5$ SL1	Frozen Model + Linear Probes	3e-6	1e-5	6e-6
$m = 5$ SSIM	Frozen Model + Linear Probes	2e-5	3e-6	6e-6
$m = 5$ SL1	Unfrozen Model + Non-Linear MLP	1e-4	1e-6	1e-3
$m = 5$ SSIM	Unfrozen Model + Non-Linear MLP	2e-5	3e-7	1e-3
Modelling Dataset = MMNIST — Task = MNIST				
-	Constant Output			
-	Input Image + Linear Layer			
None	Frozen Model + Linear Probes	1e-4	3e-6	3e-6
None	Unfrozen Model + Non-Linear MLP	1e-4	1e-6	1e-4
$m = 5$ SL1	Frozen Model + Linear Probes	3e-6	1e-7	1e-6
$m = 5$ SSIM	Frozen Model + Linear Probes	3e-6	1e-5	1e-6
$m = 5$ SL1	Unfrozen Model + Non-Linear MLP	1e-3	1e-6	1e-3
$m = 5$ SSIM	Unfrozen Model + Non-Linear MLP	1e-3	3e-7	1e-3

Table A.3: Learning rates of each of the test-task experiments.

A.2 Chapter 6: Additional Implementation Details

Accuracies are reported from the *best* checkpoint at during training.

I build a codebase around the implementation of LXMERT [183] submitted to Hugging Face[2]. LXMERT models are trained with: a learning rate of 1e-6, an Adam optimiser, and trained for up to 300 epochs. All layers are unfrozen, and the only change made is to set the size of the output layer to fit each dataset's answer vocabulary.

My experiments using the METER [51] transformer use the METER codebase[3]. The models in our experiment are finetuned by starting from the provided pretrained checkpoint provided at the METER codebase. Experiments are trained for up to 20 epochs.

[2] https://huggingface.co/docs/transformers/model_doc/lxmert
[3] https://github.com/zdou0830/METER