THE UNIVERSITY *of* ADELAIDE

Developing a System for Free-Form Visual Question Answering

Violetta Shevchenko

Supervised by:

A/Prof. Anthony Dick

Prof. Anton van den Hengel

Dr. Damien Teney

A thesis submitted for the degree of

Doctor of Philosophy

The University of Adelaide

August, 2022

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of a scholarship from the ARC Centre of Excellence on Robotic Vision CE140100016.

Violetta Shevchenko

AUGUST 2022

Acknowledgements

First of all, I am deeply thankful to my supervisors for the greatest opportunity to become a part of the amazing research community. I am very grateful to my principal supervisor, Anthony Dick, for his constant guidance and support throughout my PhD journey. To my co-supervisor, Damien Teney, for all his help and invaluable advice that helped me to understand what a good research is. And to my co-supervisor, Anton van den Hengel, for giving me a chance to join his fantastic team, supporting and inspiring me through these years. I am an extremely lucky person because I had a chance to work with these people and I am grateful for all the time and effort they spent on me.

I would like to thank Ehsan Abbasnejad, Rafa Felix and Amin Parvaneh for all the research discussions we had. Their advice and insights motivated me through my studies and led to the ideas included in this thesis.

Many thanks to my lab mates Shin Fang Chng, Yutong Dai, Yang Zhao, Dzung Doan, Cuong Nguyen, Chee Kheng Chng, Huangying Zhan, Ming Cai and Yu Liu for all the support and joyful memories you brought to my life. Special thanks to the members of AIML ECR committee for organising all the memorable events where I could take a break and just have fun.

Finally, I am grateful to the staff of Australian Institute for Machine Learning and Australian Centre for Robotic Vision for all the resources and facilities their provided.

Contents

University of Adelaide

Abstract

Developing a System for Free-Form Visual Question Answering

In the past few years, we have witnessed significant advances in the field of visual question answering (VQA). This complex task connects the areas of computer vision and natural language processing research to build a step towards solving artificial intelligence (AI). A crucial feature of any AI-complete problem is the ability to scale for real-world applications. For VQA, in particular, it implies that a model should answer any question about any possible image. However, it seems unlikely that any model would be able to learn all the required knowledge from a single training set, so the use of external knowledge has become a promising direction for VQA research. In this thesis, we investigate techniques to exploit external information to improve the performance of visual question answering methods. First, we explore the benefits of unsupervised image pre-training for VQA. We create a dataset of simple images, where only a small fraction is annotated with VQA questions. We experiment with two self-supervised approaches and show that they can be used for VQA pre-training and generalise well from little annotated data. Next, we frame VQA as a multi-task problem and complement the traditional classification objective with an additional regression loss that aims to learn vector representations of answers. This novel learning branch allows a model to embed prior knowledge about answer semantics, and we show that the information captured in the relations between answer embeddings is important for VQA. This method not only shows clear improvements in accuracy and consistency over a range of different question types but also unlocks the potential for novel answer prediction. Finally, we implement a method that embeds information from external knowledge bases into vision-and-language transformers. This method proposes to optimise an additional objective that aligns learned word representations with the matching knowledge embeddings. We evaluate the applicability of various knowledge bases to multiple downstream tasks and show that the method brings clear improvement on knowledge-demanding and general visual reasoning datasets.

Publications

This thesis contains the following papers that were published or prepared for publication:

- **Violetta Shevchenko**, Damien Teney, Anthony Dick, and Anton van den Hengel. "Visual Question Answering with Prior Class Semantics." *arXiv preprint arXiv:2005.01239* (2020).

- **Violetta Shevchenko**, Damien Teney, Anthony Dick, and Anton van den Hengel. "Reasoning over Vision and Language: Exploring the Benefits of Supplemental Knowledge." In *Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pp. 1-18. 2021.

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence.

CNN Convolutional Neural Network.

CV Computer Vision.

EBM Energy-Based Model.

ECE Expected Calibration Error.

fPMC factorised Probabilistic Model of Compatibility.

GRU Gated Recurrent Unit.

KB Knowledge Base.

LSTM Long Short-Term Memory.

MAC Memory, Attention and Composition.

MCB Multimodal Compact Bilinear.

MCMC Markov Chain Monte Carlo.

MLP Multilayer Perceptron.

NLP Natural Language Processing.

NSM Neural State Machine.

QA Question Answering.

RAUs Recurrent Answering Units.

ReLU Rectified Linear Unit.

RNN Recurrent Neural Network.

V&L Vision and Language.

VQA Visual Question Answering.

Chapter 1

Introduction

This chapter provides the background and motivation for the research tasks addressed in this thesis. We also identify research gaps and outline the objectives and main contributions of this work.

1.1 Overview

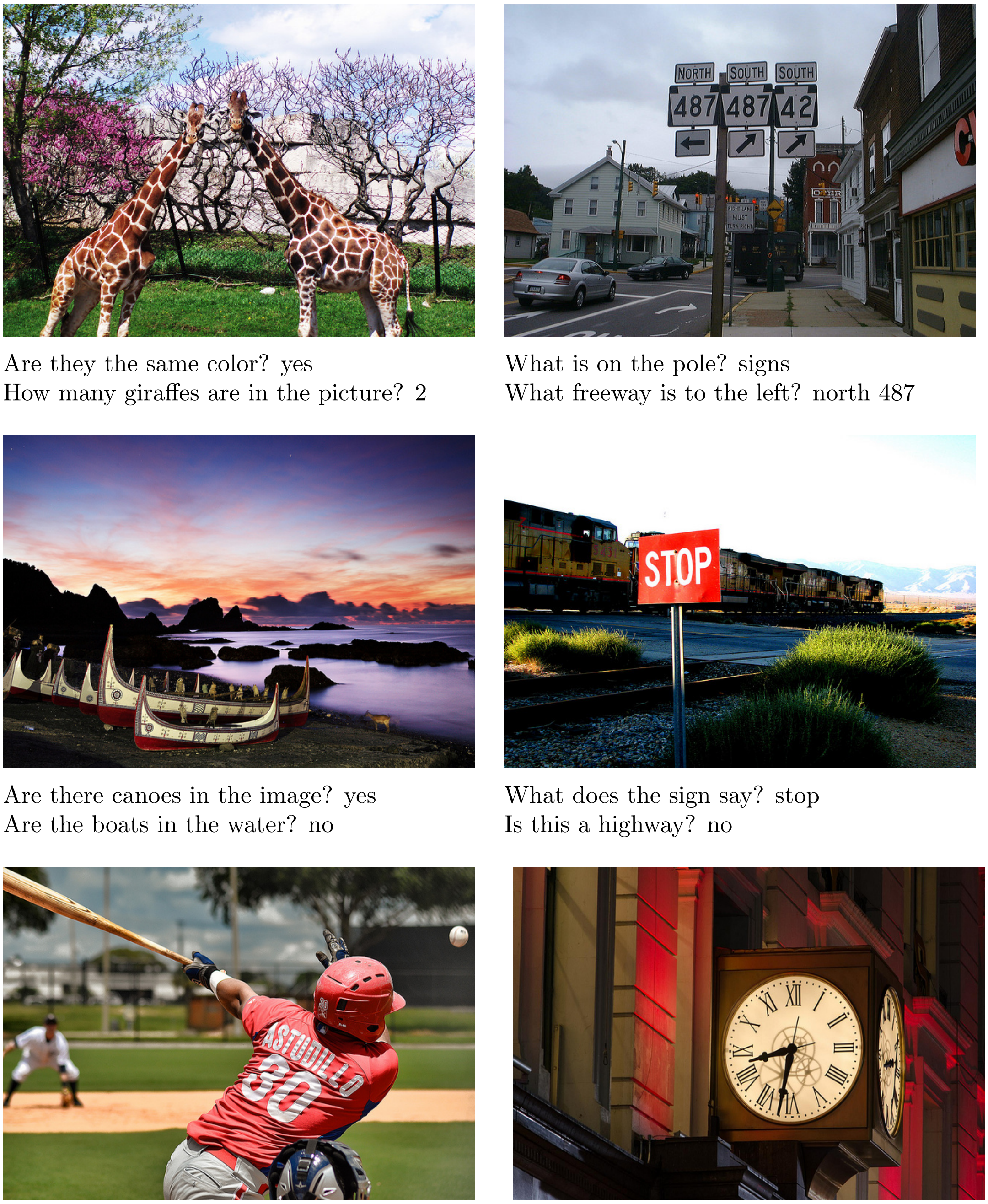
Creating an artificial intelligence (AI) that can reproduce or even exceed human abilities is the ultimate goal of many research efforts in the fields of machine learning and computer science. Such a comprehensive goal can be naturally divided into several sub-tasks typically representing basic human skills, specifically, learning, planning, information perception, reasoning, motion, *etc.* (Russell and Norvig, 2016). The past few years have seen remarkable advances in computer vision (CV) and natural language processing (NLP) fields due to the rise of deep learning. As a result, a multitude of low- and mid-level computer intelligence tasks like image classification (Russakovsky et al., 2015), segmentation (Everingham et al., 2010), object detection (T.-Y. Lin et al., 2014), sentiment analysis (L. Zhang et al., 2018), named entity recognition (Yadav and Bethard, 2018), *etc.*, can be solved with almost human-level accuracy. This progress enabled researchers to tackle more complex multi-discipline problems that challenge AI algorithms' high-level understanding and reasoning skills.

One of the tasks that can be considered AI-complete (*i.e.* the problem that requires human-level intelligence Yampolskiy, 2013) is visual question answering (VQA). The task of VQA has become a benchmark to evaluate joint progress in computer vision and natural language processing. This task, in its most general formulation, requires deep analysis of both visual and textual information in order to correctly answer a question, given an associated image (Figure 1.1).

Behind its simple formulation, VQA is an extremely complex task that offers a testbed for a multitude of capabilities required to develop strong AI systems.

VQA task can be regarded as an evaluation benchmark for robust visual and language understanding. To show visual understanding, a model should extract high-level information from visual input and perform reasoning over it. The natural language question here serves as a guide for the information extraction and reasoning process. For example, the question *What kind of bird is there?* requires object detection and recognition, the question *How many people are playing football?* tests activity recognition and counting skills, while the question *How long do these animals usually live?* requires knowledge-based reasoning in addition to typical vision tasks. Similarly, to demonstrate language understanding abilities, a model might need to perform named entity extraction (*Is this place in France or in England?*) or co-reference resolution (*There is a big box on the left. What colour is it?*). These natural language processing tasks are not trivial themselves, and a VQA model must perform them implicitly while solving a more complicated problem of question answering. Finally, the information extracted from two modalities needs a proper alignment in order to find an answer that depends on both an image and a question. Therefore, due to its multi-modal understanding and grounding requirements, VQA can serve as a visual Turing test (Geman et al., 2015) to assess the progress in AI.

Besides its scientific importance, VQA can be directly applied to various practical problems. A system that is able to recognise and analyse visual input through natural language communication can have vast applications for human-computer interaction. For example, a mobile app that answers users' questions about the taken photograph will immeasurably help visually impaired people to perceive the surrounding world and increase their independence. The systems used so far mostly relies on human volunteers who receive the questions online and give their answers in real time ("Be My Eyes - See the world together", n.d.; Bigham et al., 2010; Lasecki et al., 2013). Automation of this process with a VQA model will reduce its cost and latency while enhancing privacy (Gurari et al., 2018). Another potential application for VQA task are social or service robots that commonly interact with people (S. Cho et al., 2020). Such robots will have greater appeal if users are able to communicate with them freely using natural language rather than using pre-defined commands. Finally, VQA can be applied in a wide range of tasks including medical diagnostics, advertising, surveillance, and education (see Barra et al., 2021 for a survey of practical VQA applications).



Are they the same color? yes
How many giraffes are in the picture? 2

What is on the pole? signs
What freeway is to the left? north 487

Are there canoes in the image? yes
Are the boats in the water? no

What does the sign say? stop
Is this a highway? no

What color is the uniform? red
Is he bunting or swinging? swinging

What kind of numbers are these? roman
What time is it? 8:32

FIGURE 1.1. The task of VQA requires to answer the question about the related image. Examples are taken from VQA v2 dataset (Goyal et al., 2017).

Regardless of the exact purpose of VQA (whether scientific or application-driven), most researchers aim to build a VQA model capable of answering any possible question about any possible image. This implies that a model must capture diverse textual and visual semantics, incorporate additional knowledge, be able to work with novel visual and textual concepts there were not seen during training, and easily scale to more data. This challenging setting, called free-form VQA, conforms to real-world conditions and arouses the greatest interest. In this thesis, we will explore current approaches in the field and propose new methods to address existing problems in free-form VQA.

1.2 Background

The VQA task is a sub-field of a broader vision and language (V&L) research area. This area comprises a variety of multi-discipline tasks, including image captioning (X. Chen et al., 2015; Sharma et al., 2018; Young et al., 2014), visual reasoning (Suhr et al., 2017; Suhr et al., 2019), visual entailment (N. Xie et al., 2019), visual dialog (Das et al., 2017; De Vries et al., 2017), vision and language navigation (Anderson, Wu, et al., 2018), referring expression comprehension (Qiao et al., 2020) and text-to-image generation (Frolov et al., 2021). All these benchmarks were designed to assess joint vision and language understanding and can be regarded as a step towards true AI. However, they all have flaws that call into question their ability to solve the task. For example, tasks that include the generation process (*e.g.* image captioning or visual dialog) require complex evaluation, typically involving humans, to assess the quality and relevance of the generated text. But the presence of humans in a training loop does not really conform with the definition of AI. In other tasks, a model needs to choose an output among a small set of candidates (*e.g.* true/false labels for visual reasoning or object regions for referring expression comprehension). As a result, such models lack interpretability which makes it hard to measure their reasoning capabilities. Differently, in VQA, a model chooses between a large set of possible answers. This setting allows to compute accuracy across the different question and answer types and measure consistency and validity of answers. That provides an insight into the model's behaviour while preserving its compliance with automatic evaluation.

Regarding the problem definition, VQA has taken inspiration from the textual question answering (QA) task (Fader et al., 2014; Rajpurkar et al., 2016; Weston et al., 2015). While in VQA the answer is grounded into the image, in QA the answer should be extracted from text paragraphs (Abbasiantaeb and Momtazi, 2021) or knowledge bases (Fu et al., 2020), which is related to reading comprehension and information retrieval problems. A change towards visual context brought an additional challenge to the task because images represent a high-dimensional and unstructured source of information that cannot be easily parsed. Furthermore, VQA needs training data with aligned images and questions that are visually and linguistically diverse. This kind of data is not common in the wild, as opposed to publicly available text corpora ("Wikipedia: The free encyclopedia", 2004) frequently used in QA, and requires complex data collection and annotation.

Different from most QA and V&L tasks, VQA is usually treated as a classification problem. A common way to tackle VQA includes a model that has image and question encoders, a feature fusion module and a classifier. The architectures used for feature extraction range from simple Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) models that have been used originally (Antol et al., 2015; M. Ren et al., 2015; B. Zhou et al., 2015), to transformer encoders that have received broad attention recently (L. H. Li et al., 2019; X. Li et al., 2020; Tan and Bansal, 2019). The purpose of the fusion module is to combine two distinct representations into a joint multi-modal embedding which is then passed through the classifier. Finally, the classifier outputs probability scores for all possible answers in the training set and that with the highest score is returned as the predicted answer. Most of the existing VQA methods intrinsically build upon this core frame, although various advanced techniques have been proposed to improve feature extraction and fusion (we will discuss different architecture choices in Section 2.1).

1.3 Motivation

The VQA community has grown dramatically in the last few years. To track the progress made in the field, several challenges[1][2] are hosted every year to measure the current state-of-the-art performance on popular datasets (Goyal et al., 2017; Hudson and Manning, 2019a). Although recently proposed methods (W. Li et al., 2020; X. Li et al., 2020; Z. Wang et al., 2021) achieved nearly human performance on these benchmarks, the question of whether they actually reached human-like visual understanding remains open. Several studies (Agrawal et al., 2018; K. Kafle and Kanan, 2017a) have revealed major shortcomings of current

[1] VQA Challenge https://visualqa.org/challenge.html

[2] GQA Challenge https://cs.stanford.edu/people/dorarad/gqa/challenge.html

methods and datasets mostly associated with bias, inadequate evaluation and lack of generalisation (K. Kafle et al., 2019). In this thesis, we will address the following problems that, we believe, impede the development of deep visual understanding:

1. Need for external knowledge.

   Although the questions in VQA are designed to query images, the visual information solely may not be sufficient to derive the answer. When humans perceive the visual world, there is a whole lot of commonsense and prior factual knowledge that we unconsciously link to everything we see. For example, to answer *Is this pizza vegetarian?*, one should not only identify the pizza's ingredients but also know about the nutritional preferences of vegetarians. Typical VQA models can only learn the information present in the training data, but no single dataset can ever cover the whole world knowledge. For that reason, a line of work (Marino et al., 2021; P. Wang et al., 2017a; Q. Wu et al., 2016; Z. Zhu et al., 2020) has investigated the use of external knowledge sources to boost VQA performance. However, current methods show relatively low accuracy on knowledge-demanding tasks (Marino et al., 2019; P. Wang et al., 2017b), which indicates the need for better ways to incorporate external knowledge.

2. Dependency on annotated data.

   The performance of deep learning models scales with the amount of training data (Kaplan et al., 2020; Sun et al., 2017). In VQA, the collection of clean annotated data is an expensive and time-consuming process, so the scope of available data may not be enough to train the model in an end-to-end manner. To facilitate the task, multiple studies (Anderson, He, et al., 2018; Jabri et al., 2016; Teney, Anderson, et al., 2017) proposed to transfer knowledge from the models (K. He et al., 2016; S. Ren et al., 2015) that were pre-trained on common visual tasks like image classification or object detection. However, such an approach limits the performance on VQA samples if they contain novel concepts not present in pre-training data or if they come from a completely different domain. We believe that unsupervised pre-training can be a solution for visual feature learning when labelled VQA data is scarce.

3. Limitation of classification approach.

   In order to reduce the computation and evaluation costs of the task, VQA is commonly treated as a classification problem. However, this simplification leads to the loss of valuable information contained in the answer's words. When answers are treated as abstract class labels, a model handles them equally without considering their semantics. For example, if the ground truth answer to the question *What is the colour of the men's shirt?* is *light blue*, then the answer *blue* should be penalised less than the answer *orange*. Furthermore, synonymical and paraphrased answers can be often regarded as equivalent, but current models consider all answers that are not present in the annotation as incorrect.

4. Inability to generalise to novel answers.

   Another adverse consequence of the classification approach is the restriction of an answer set. In the real-world setting, a VQA model must be applied to open unlimited domains which, among other things, implies an ability to generalise to novel answers. Current methods, on the opposite, select a subset of the most common answers in the training data and run classification over it. It means that a new classifier must be trained every time we need to incorporate new answers. Given the scope of possible answers in the real, constantly expanding world, this approach seems impractical. This issue motivated researchers to design zero-shot VQA (Teney and van den Hengel, 2016) – a setting where test samples contain new concepts, including novel answers never seen during training. Despite the crucial importance of the zero-shot setting, little progress has been made in this direction.

Motivated by these shortcomings, we aim to explore current methods and develop new techniques to address the issues outlined above. The main goal of this thesis is to investigate how additional information can help to learn better representations of images (Chapter 3), answers (Chapter 4) and questions (Chapter 5), to make the task applicable to real-world conditions.

1.4 Contributions

The main contributions of this thesis are summarised as follows:

Chapter 3. We explore the applicability of unsupervised image feature pre-training for visual question answering. We experiment with two self-supervised approaches, namely, energy-based modelling and contrastive learning. In our setting, the data mostly consists of unlabelled images with a small fraction of VQA-annotated samples. We show that both methods, pre-trained on unlabelled images, can be efficiently fine-tuned on little VQA data and generalise to novel test instances never seen during fine-tuning. Moreover, contrastive learning method shows performance superior to the fully-supervised method that was pre-trained on a considerably larger ImageNet dataset (Russakovsky et al., 2015).

Chapter 4. We present a novel mechanism to embed prior answer knowledge in a model for visual question answering. We formulate VQA as a multi-task problem, where the model is trained not only with the classification objective but also learns to perform a regression in a vector space that represents answer semantics. We perform an extensive analysis of the model and various ablations. We demonstrate clear advantages on the GQA dataset (Hudson and Manning, 2019a) and show improvements in consistency and accuracy over a range of question types. An extensive study of learned representations reveals that important semantic information is captured in the *relations* between embeddings in the answer space. Experiments with novel answers, unseen during training, indicate the method's potential for open-set prediction.

Chapter 5. We implement a method that injects external information from knowledge bases (KBs) into a vision-and-language transformer. This technique is model-agnostic and can expand the applicability of any vision-and-language transformer with minimal architectural modifications and computational overhead. We empirically study the relevance of various KBs to multiple tasks and benchmarks. An extensive evaluation on four downstream tasks shows clear improvement on knowledge-demanding and general visual reasoning datasets. We perform probing experiments and show that the injection of additional knowledge regularises the space of embeddings, which improves the representation of lexical, semantic, and relational knowledge that is lacking in typical V&L models.

1.5 Thesis Outline

The rest of the thesis is organised as follows:

- In Chapter 2, we review the relevant literature in the visual question answering research field. We include an overview of the task's history and a summary of the most popular methods, datasets and evaluation metrics.
- In Chapter 3, we explore energy-based and contrastive learning for image feature pre-training and its applicability to VQA task.
- In Chapter 4, we introduce a method to embed prior semantic information about answers into VQA models and propose a technique to train VQA as a multi-task problem.
- In Chapter 5, we describe a method to embed information from external knowledge bases into vision-and-language transformer models.
- In Chapter 6, we summarise the contributions of the thesis, discuss the current limitations and outline promising directions for future research.

Chapter 2

Literature Review

The task of visual question answering has attracted considerable attention from natural language processing and computer vision research communities. Despite the noticeable advances made in recent years, the VQA task remains unsolved. In this chapter, we review the publications in the visual question answering field. We start with early works that first proposed the task and described simple baseline methods. We then describe more advanced techniques introduced to improve performance and address common shortcomings of visual question answering models. Finally, we survey the commonly used datasets and metrics designed for VQA evaluation.

2.1 Common Approaches

The general idea of image understanding through language had emerged long before the first VQA methods were proposed. Barnard et al., (2003) argued that visual and textual information complement each other and resolve the ambiguity inherent in separate images and text. They have studied methods that link words and images to perform image regions annotation. This task, where a model needs to describe visual content (Farhadi et al., 2010; Karpathy and Fei-Fei, 2015; Kulkarni et al., 2013; Socher et al., 2014), can be regarded as one of the first attempts to approach visual understanding. Similarly, the VQA task is designed to connect vision and language through questions that test visual reasoning.

2.1.1 Baseline Models

The first known attempt to solve visual question answering was made by Malinowski and Fritz, (2014). They proposed to use semantic image segmentation and semantic question parsing in a Bayesian algorithm that models spatial relationships between image objects. K. Tu et al., (2014) described a framework for answering template-based user queries about videos based on a joint video and text processing. Geman et al., (2015) presented a query engine that proposes questions about the given image. However, these methods were rather restricted by the range of possible questions and were trained on relatively small datasets which cannot be regarded as a solution for true visual understanding due to their limited abilities. The fundamental study that stimulated researchers to attend to VQA was the work of Antol et al., (2015) as they were the first who thoroughly formulated and described the VQA task. The authors introduced a free-form VQA setting together with an extensive dataset and a baseline model intended for establishing the foundation for the VQA problem.

The recent progress made in the field of deep learning has led to the majority of existing VQA approaches using deep neural networks in attempts to solve the problem. The common pipeline of a VQA algorithm consists of four parts: (1) image feature extraction, (2) question feature extraction, (3) feature combination, and (4) classification over the range of possible answers (Figure 2.1). Although such algorithms can use different strategies to extract and fuse features, they can be categorised as joint embedding methods as all of them exploit the same idea - to project visual and textual data in one joint space.

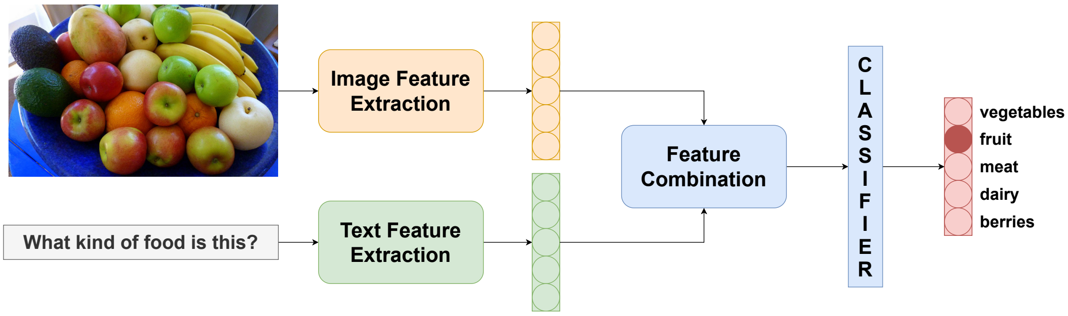


FIGURE 2.1. A common pipeline of baseline VQA models.

For image feature extraction a pre-trained Convolutional Neural Network (CNN) (*e.g.* ResNet K. He et al., 2016, VGGNet Simonyan and Zisserman, 2015) is usually used. It allows researchers to skip low-level image processing and exploit the resources of CNN pre-trained on large datasets. For question features extraction the common approach (Antol et al., 2015; Malinowski et al., 2015; M. Ren et al., 2015) is to use a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) or Gated Recurrent Unit (GRU) (K. Cho et al., 2014) cells. Other methods (K. Kafle and Kanan, 2016; B. Zhou et al., 2015) apply simpler strategies, like bag-of-words or skip-thought vectors (R. Kiros et al., 2015) to get question embeddings. Similar to the idea of pre-trained image features, some methods make use of pre-trained word embeddings like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) to get initial word representations.

Apart from various feature extraction processes, models can differ in the way they combine image and question vectors. The simplest forms of fusion are concatenation (B. Zhou et al., 2015), element-wise multiplication (Antol et al., 2015) or addition (Gao et al., 2015). Saito et al., (2017) described the idea of joining element-wise multiplication and addition in order to take advantage of both visual and textual information. A more complex approach consists in combining features by an outer product which can hypothetically increase an expressive capacity of a resultant vector. Fukui et al., (2016) proposed Multimodal Compact Bilinear (MCB) pooling - a method that approximates the outer product of multimodal features by projecting it to a lower-dimensional space. To decrease computational complexity of MCB pooling, J.-H. Kim et al., (2016) introduced a lower-rank bilinear pooling. This method, despite its reduced complexity, can achieve results comparable to MCB.

2.1.2 Attention-based Models

In spite of its effectiveness, the simple joint embedding approach has considerable limitations. For example, CNN extracts global image features that describe an image as a whole, although questions usually relate to the information contained only in certain parts of an image. Focusing on local regions instead can help to filter the noise and produce more relevant answers. Inspired by the success in other vision tasks (Ba et al., 2014; Tang et al., 2014; K. Xu et al., 2015), VQA models incorporated attention mechanisms (Figure 2.2), where the features representing unimportant parts of the input data are multiplied by lower weights (soft attention) or completely ignored (hard attention).

The simplest way to implement visual attention is to apply a uniform spatial grid to divide an image into separate blocks and compute the relevance of each block to the question asked. It is usually done by using the output of one of the last layers in CNN and multiplying it by computed attention weights. K. Chen et al., (2015) introduced a CNN model that generates question-guided attention maps to localise informative image regions. Y. Zhu et al., (2016) added spatial attention to LSTM model. Z. Yang et al., (2016) proposed Stacked Attention Network that consists of several attention layers able to perform progressive inference to iteratively locate the most relevant areas of an image. A similar

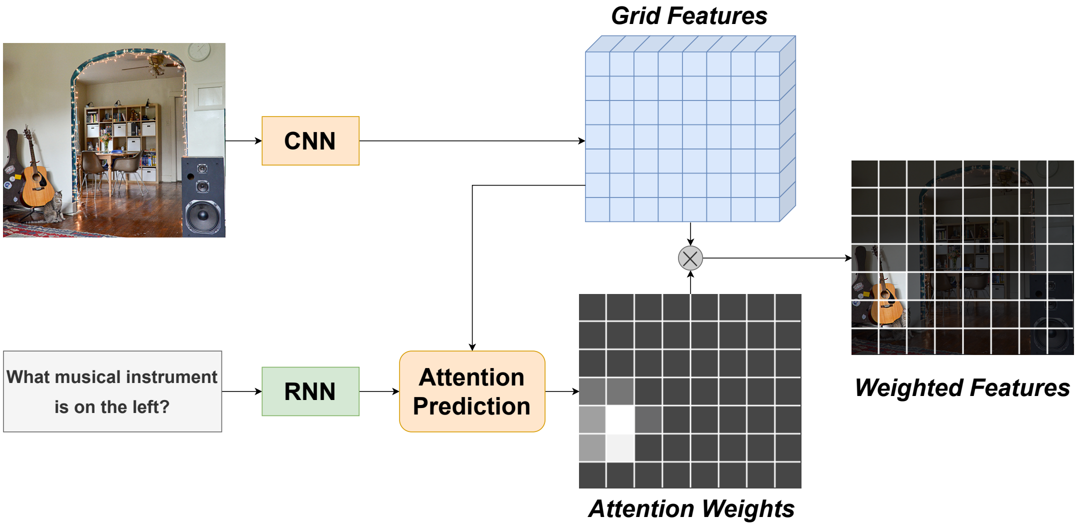


FIGURE 2.2. Visualisation of a simple question-guided visual attention.

model introduced by H. Xu and Saenko, (2016) uses a memory component to memorise region representations and perform multi-hop inference. In contrast to the major line of works, Malinowski et al., (2018) used hard attention that discards all irrelevant image patches. Some models incorporate additional high-level information to guide visual attention. For instance, D. Yu et al., (2017) trained a separate concept detector that extracts semantic concepts from images and uses them for learning semantic visual attention. Besides commonly used question features, Shi et al., (2018) utilised information about the question's type (*i.e.* whether it requires detecting objects and their attributes, performing counting, scene or action recognition, *etc.*).

The methods described above divide images into uniform equally-sized visual blocks which can restrict their ability to perform proper localisation. One may argue that the use of uniform spatial grids does not conform with the way humans attend to different parts of an image, as it is more natural for humans to break an image into a set of regions corresponding to individual objects. In (Shih et al., 2016) and (Ilievski et al., 2016) authors used automatically selected image regions located by Edge Boxes method (Zitnick and Dollár, 2014). Anderson, He, et al., (2018) introduced a new bottom-up and top-down attention mechanism that uses Faster R-CNN (S. Ren et al., 2015) to detect objects and extract region features and then uses question-driven attention to obtain weighted task-specific features. Similarly, P. Huang et al., (2019) exploited Faster R-CNN detections together with predicted object labels to perform multi-grained visual attention. However, a recent work of H. Jiang et al., (2020) has proven that the grid-based approach can achieve the same accuracy as the region-based one while significantly reducing the running time. This experiment shows that the semantic information carried by the visual features (*i.e.* the data that CNN was pre-trained on) is more important than the form of the regions used.

In addition to visual attention, some VQA models include a question attention component. Given that questions can contain redundant or noisy information, the ability to focus on the most relevant words is crucial for VQA algorithms. The Hierarchical Co-Attention model proposed by Lu et al., (2016) emphasises the importance of combined visual and textual attention by using both question-guided visual attention and image-guided question attention. Besides, the model hierarchically encodes questions and performs co-attention at three levels: word level, phrase level and question level, and then co-attention features are combined into the final representation. A similar idea was described in (Nam et al., 2017) where authors used a memory component to store attention results and recursively update them from both image and question information simultaneously. While in most methods the question or image is considered as a whole when used to compute attention weights, a line of work (J.-H. Kim et al., 2018; Nguyen and Okatani, 2018) introduced dense co-attention, where each word attends to each image region and vice versa. In (Z. Yu et al., 2017) question attention is computed without image features at all, while the visual attention still relies on question representations. Similarly, C. Yang et al., (2019) proposed to first use self-attention to identify the most important question words and then use them for image attention. Z. Yu et al., (2019) implemented a deep co-attention architecture by combining self- and guided attention. Another line of research that extensively uses self-attention is based on transformer architecture (Vaswani et al., 2017) and will be discussed in Section 2.1.5.

2.1.3 Compositional Models

Although the most common approach in VQA is to use a single neural network to perform all the manipulations over the input data, the effectiveness of this strategy is quite limited for certain types of questions. A simple unified algorithm is likely to lack the reasoning power needed to cover the diversity of all possible image-question pairs. For example, a question like *"How many horses are there?"* requires object detection and counting, while questions like *"What is next to the door?"* involve analysis of spatial relationships and object recognition. Conversely, questions *"What colour is the dog?"* and *"What colour is the cat?"* share the same semantic structure and require similar steps of reasoning, although they may refer to quite different images. So a number of works in VQA have investigated how the simple learning from separate image-question examples can be replaced with more complex strategies to better capture the compositional nature of questions and images.

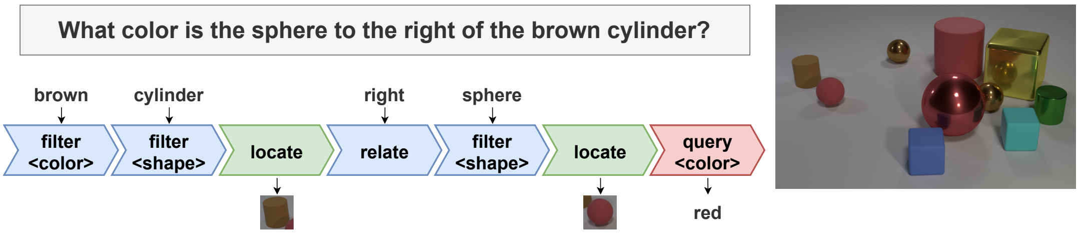


FIGURE 2.3. An example of a compositional approach where the question is parsed into tasks required for answer prediction.

To exploit the compositionality of the VQA task, Andreas et al., (2016b) proposed an architecture called Neural Module Networks. The main feature of the framework is the use of a semantic parser that divides questions into parts regarded as separate sub-tasks required for the reasoning process. Each sub-task determines the specified neural module that implements an associated action, such as attend, classify, measure, *etc*. The resulting model is composed of the selected neural modules and the general model structure, thus, is different for each question. In this way, the described framework carries out the concept of compositional models constructed from several parts in order to adapt and generalise to different tasks (see Figure 2.3). This technique is contrasted to the approach when a single universal network is used for all questions and is proven to show better performance on highly compositional questions. In a follow-up work (Andreas et al., 2016a) the authors suggested learning the model structure for each question instead of using manually-specified modules, that is the model is now able to dynamically choose the best architecture. Furthermore, R. Hu et al., (2017) refused to use an external semantic parser and attempted to learn the best network layout directly from the question. Johnson, Hariharan, van der Maaten, Hoffman, et al., (2017) developed a model that includes the program generator – an LSTM network that transforms a question into a functional program – and the execution engine that applies this program to an image. Yi et al., (2018) has extended this approach and included a scene parser to exploit the structural information of the image.

Another compositional model based on Recurrent Answering Units (RAUs) was presented by Noh and Han, (2016). The authors also argue the necessity to decompose the learning process into sub-tasks and emphasise that the range and the order of operations required for answering questions vary significantly for different instances. For solving this issue, the authors proposed a novel recurrent network architecture composed of successive RAUs– blocks that are capable of solving the whole task by themselves. By using the number of consecutive RAUs a model can perform progressive reasoning and implicitly solve different sub-problems. In contrast to the previously described methods, this model does not exploit any language parsers and does not specify different problem-solving modules, but instead it relies on the successive model structure and its ability to perform progressive reasoning and refinement by itself.

Recent studies have been focused on combining the classical end-to-end approach with explicit compositional reasoning. Hudson and Manning, (2018) proposed a new recurrent network consisting of Memory, Attention and Composition (MAC) cells. This recurrent network decomposes a question into a series of reasoning operations which are further applied to an image, and then stores all intermediate results in memory states. The model thus learns the required reasoning chains directly from the input data without relying on predefined neural modules. An extension of the MAC model, called the Neural State Machine (Hudson and Manning, 2019b), performs the reasoning over visual scene graphs instead of the raw image input, which improves its compositionality and generalisation skills compared to MAC. Despite the complete success on synthetic compositional VQA tasks (*e.g.* achieving almost 100% accuracy on CLEVR Johnson, Hariharan, van der Maaten, Fei-Fei, et al., 2017), such modular methods do not generalise well to real-world VQA datasets indicating a need for general multi-purpose methods that can generalise across these domains (Shrestha et al., 2019).

2.1.4 Models with External Knowledge

The ultimate goal of the VQA research is to implement an algorithm able to answer any possible question about any image. However, current VQA models are restricted by the scope of available training data and limited pre-defined answer sets. Typical VQA datasets, although comprising millions of image-question instances, cover a rather small fraction of existing knowledge. Firstly, the data collection for VQA often involves human annotators making this process quite expensive and time-consuming. Next, our knowledge about the world keeps growing so the datasets need to be updated constantly. Therefore, the models trained on these limited datasets lack the ability to generalise and cannot be freely applied in real-world conditions. This shortcoming has motivated researchers to search for alternative sources of information that can benefit the training of VQA models.

One of the data resources suitable for VQA is image web search. It is an unlimited source of raw visual information which is easy to collect but prone to noise, so the methods utilising data from the web must include some pre-processing and filtering techniques. Teney and van den Hengel, (2016) used an image search engine to retrieve images for each word in questions and answers. The images were then used during test time to obtain feature representations for unknown words never seen during training which allowed to apply this method to zero-shot VQA. Zheng et al., (2020) made use of images collected from the web together with their tags to generate additional training examples and compensate for insufficient coverage of the used dataset. Another possible source of external knowledge for VQA is Wikipedia. For example, Q. Wu et al., (2016) used image attributes to query DBpedia (Auer et al., 2007) – an ontology of Wikipedia data – and get related text passages that were used along with the question. Similarly, Marino et al., (2019) used multiple combinations of question words and image objects as queries to Wikipedia search API to retrieve the relevant article and find the answer in it.

Unlike the raw unstructured data described above, knowledge bases (KBs) offer an alternative with curated and categorised information. Commonly used KBs, like Freebase (Bollacker et al., 2008), YAGO (Mahdisoltani et al., 2014), ConceptNet (Speer et al., 2017), Wikidata (Vrandečić and Krötzsch, 2014), *etc.*, contain general knowledge about the world in a structured computer-readable format. This format is typically represented by facts, or triplets, describing different concepts and relationships between them. For example, the knowledge that *"dogs can be used to guide blind people"* is stored in the KB as a triplet (*'dog'*, *'is capable of'*, *'guide the blind'*). The use of such structured knowledge not only expands the range of answerable questions but also allows to perform explicit reasoning. When the model picks up the most relevant facts and uses them for answer prediction, one can use these facts to observe and explain the reasoning. This approach then opens up the direction towards more interpretable and robust VQA methods.

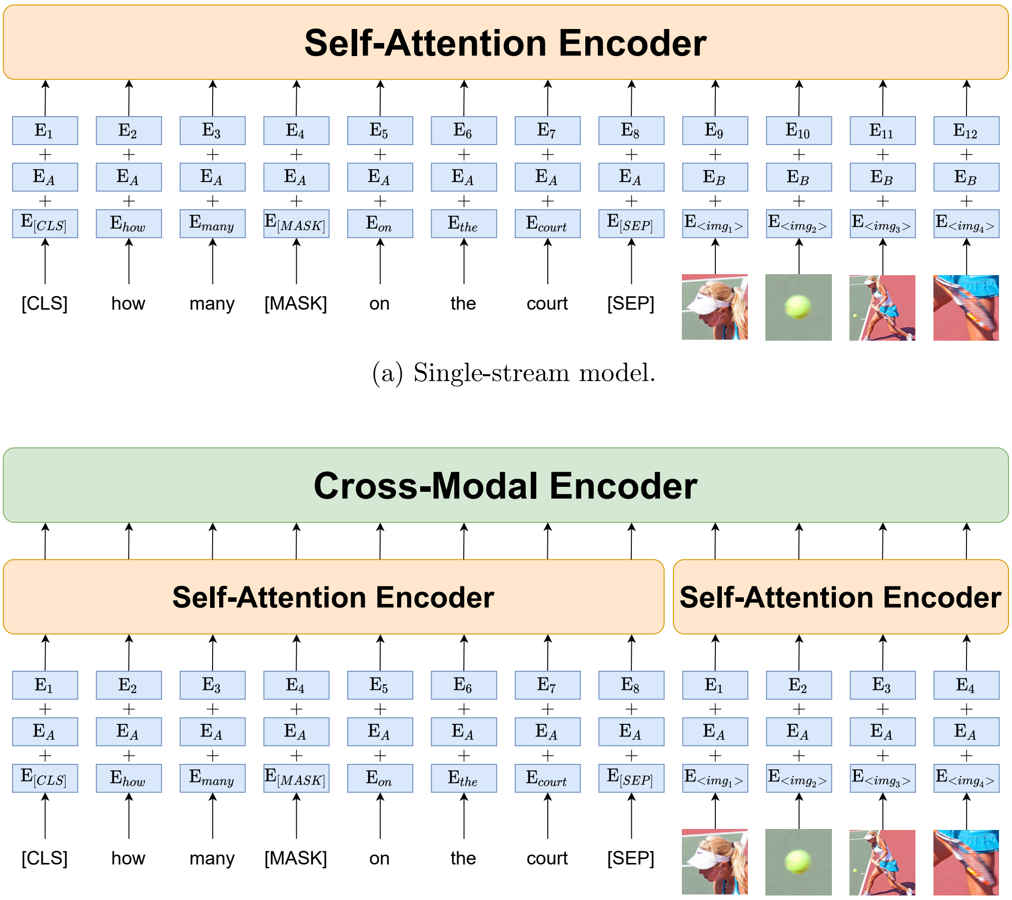
In an attempt to perform explicit reasoning P. Wang et al., (2017a) developed a model that can justify the answer choice by providing explanations. The model detects visual concepts (objects, scenes and attributes) and maps them to the related entities in a KB. The question is then processed to a query that selects desired facts from these entities. The selected facts not only help to derive the answer but also expose the reasoning process. In their further work P. Wang et al., (2017b) used predicted question categories to improve query mapping. It helped to filter out excessive knowledge and select only question-relevant facts. Narasimhan and Schwing, (2018) eliminated the querying step and instead learned a mapping model which helped to avoid synonyms and word disambiguation challenges that inhere in exact query mapping. While early methods mostly relied on external data, recent works (Gardères et al., 2020; Marino et al., 2021) have shown the benefit of combining explicit knowledge retrieved from KBs with the information implicitly learned through vision and/or language pre-training. This approach takes advantage of both a large amount of uni- and multi-modal training data and pre-processed structured knowledge information.

Differently from the methods described above, where the source of extra information lies outside of the training data, several works proposed to extract additional information directly from images. Singh et al., (2019) introduced the task of TextVQA, which requires a model to read the text present in images. Concurrently, Biten et al., (2019) introduced the ST-VQA dataset, which also requires a model to exploit textual cues for question answering. The methods that attempt to solve TextVQA usually incorporate an optical character recognition component (OCR) and use predicted OCR tokens as additional input (R. Hu et al., 2020; Mishra et al., 2019; Q. Zhu et al., 2020). Although these methods stand out from typical knowledge-based VQA methods, they help bridging the gap between current VQA solutions and real-world applications.

2.1.5 Transformer-based Models

Over the last few years, transformer-based models have dominated the VQA field. Transformer, proposed by Vaswani et al., (2017), is an encoder-decoder architecture that, unlike recurrent networks, relies solely on self-attention to perform sequence-to-sequence translation. Inspired by the efficiency of transformers, Devlin et al., (2019) introduced BERT – a transformer encoder pre-trained in a self-supervised way. The training of BERT consists of two phases: pre-training on large unsupervised textual data and fine-tuning on labelled task-specific datasets. This two-step approach allows to share the same architecture for massive language pre-training and transfer learning across different downstream tasks. BERT achieved state-of-the-art results on a multitude of natural language processing tasks and has become one of the most commonly used language representations models. Motivated by the success of BERT, recent

VQA models adapted the pre-training strategy and transformer architecture to embrace two modalities: visual and textual.


Self-Attention Encoder
E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12
EA EA EA EA EA EA EA EA EB EB EB EB
E[CLS] Ehow Emany E[MASK] Eon Ethe Ecourt E[SEP] E<img1> E<img2> E<img3> E<img4>
[CLS] how many [MASK] on the court [SEP]
Cross-Modal Encoder
Self-Attention Encoder
Self-Attention Encoder
E1 E2 E3 E4 E5 E6 E7 E8 E1 E2 E3 E4
EA EA EA EA EA EA EA EA EA EA EA EA
E[CLS] Ehow Emany E[MASK] Eon Ethe Ecourt E[SEP] E<img1> E<img2> E<img3> E<img4>
[CLS] how many [MASK] on the court [SEP]

(a) Single-stream model.

(b) Two-stream model.

FIGURE 2.4. Common architectures for two types of transformer-based models.

Transformer-based VQA models can be divided into two architecture categories: single-stream and two-stream. In single-stream models (Figure 2.4a), such as VL-BERT (Su et al., 2019), VLP (L. Zhou et al., 2020), Unicoder-VL (G. Li et al., 2020), VisualBERT (L. H. Li et al., 2019), UNITER (Y.-C. Chen et al., 2020), InterBERT (J. Lin et al., 2020), *etc*., both image and text features are processed by a single transformer. This approach allows learning unified vision-language representations through the early fusion of two modalities. On the contrary, two-stream models (Figure 2.4b), like LXMERT (Tan and Bansal, 2019), ViLBERT (Lu et al., 2019), ERNIE-ViL (F. Yu et al., 2021), *etc*., push back the inter-modal fusion into the deeper layers of the model. In these models, visual and textual inputs are passed into two separate encoders succeeded by one cross-modal encoder. Despite the clear differences in design, the choice between the two architectures is not obvious. Although single- and two-stream models encode deeper interactions between modalities in different layers (Cao et al., 2020), their overall performance is comparable (Bugliarello et al., 2021).

One of the features that distinguish transformer-based models is the way they process input data. For language input, most models follow the procedures adopted in BERT. First, the text is tokenised into sub-words and framed by special tokens [CLS] and [SEP] indicating segment boundaries. Then the tokens are passed through three embedding layers encoding token, segment and position information. The resulting embedding is then the sum of these layers' outputs. Some models (Y.-C. Chen et al., 2020; Tan and Bansal, 2019) omit segment embedding since the whole input belongs to one segment only. Others add extra layers like visual feature embedding in (Su et al., 2019). From the image side, a common practice is to utilise region features extracted with an object detector like Faster R-CNN (S. Ren et al., 2015) together with the bounding box coordinates (*i.e.* position features). Although the use of an object detector allows the model to enjoy the benefits of pre-training, the semantic coverage of extracted features is narrowed to the object categories used during the training. To address this issue recent works (Z. Huang et al., 2021; Z. Huang et al., 2020; W. Kim et al., 2021) proposed to exclude object detection and instead learn visual features directly from images.

The pre-training phase common to all transformer-based models mostly differs in the choice of data and objective functions. A standard pick for pre-training data is vision-and-language examples coming from image captioning (X. Chen et al., 2015; Sharma et al., 2018), VQA (Goyal et al., 2017; Hudson and Manning, 2019a) and alt-text (Jia et al., 2021) datasets. The current trend is to aggregate several pre-training datasets to expand data coverage, however, it has been shown (Singh et al., 2020) that the origin of the data (*i.e.* whether it comes from the same domain as the downstream task) can outweigh its size. During pre-training, multiple objectives are optimised simultaneously to facilitate the proper fusion of two modalities. These objectives can be divided into (1) uni–modal tasks, like masked language modelling (Y.-C. Chen et al., 2020; Lu et al., 2019; Tan and Bansal, 2019), masked object modelling (Y.-C. Chen et al., 2020; Tan and Bansal, 2019), and (2) multi-modal tasks, like image-text matching (Lu et al., 2019; Tan and Bansal, 2019), visual question answering (Tan and Bansal, 2019) and contrastive loss (W. Li et al., 2020; X. Li et al., 2020). However, Z. Wang et al., (2021) have demonstrated recently that even with the single language modelling objective the model can learn powerful joint representations and achieve state-of-the-art results.

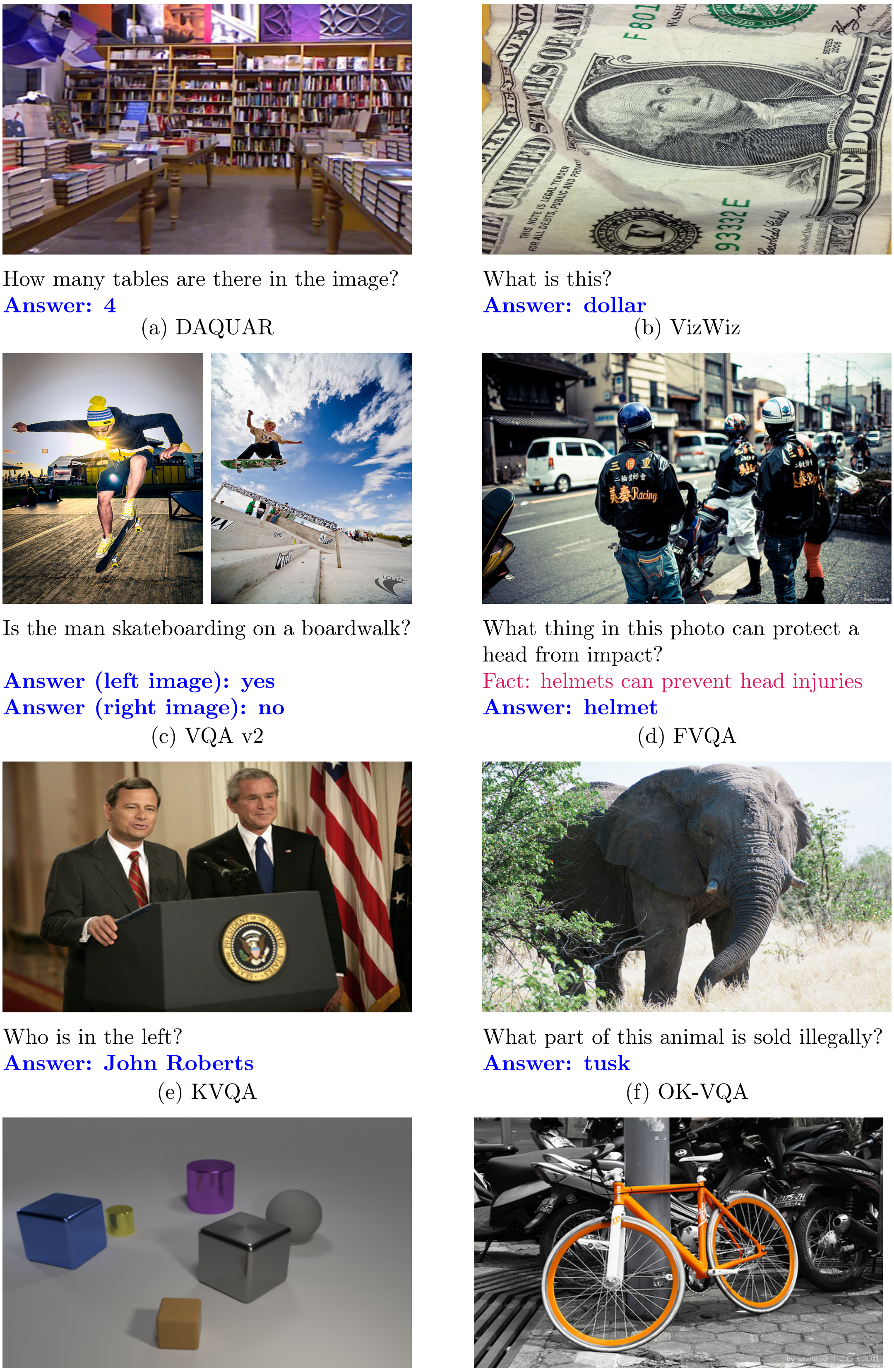
2.2 Evaluation

2.2.1 Datasets

As with any learning task, the choice of training data is one of the key aspects that boosts the performance of VQA models. Generally, a sample from a VQA dataset is composed of an image and a related question-answer pair (see examples in Figure 2.5). The common sources for VQA datasets are large bases of real images like MS-COCO (T.-Y. Lin et al., 2014) and YFCC100M (Thomee et al., 2016), manually created clipart (Antol et al., 2015), and rendered images (Andreas et al., 2016b; Johnson, Hariharan, van der Maaten, Fei-Fei, et al., 2017). Question-answer pairs are usually collected either manually through crowd-sourcing (Antol et al., 2015; Marino et al., 2019) or automatically (Hudson and Manning, 2019a; Johnson, Hariharan, van der Maaten, Fei-Fei, et al., 2017) (for example, generated from image captions or scene graphs). The generation of answers, in turn, is typically organised in two settings: multiple-choice and open-ended. In the former, each question is provided with a set of possible answers where only one of them is correct. In the latter, on the contrary, each question is labelled with the correct answer(s) only.

One of the first benchmarks released for VQA is DAQUAR dataset (Malinowski and Fritz, 2014). It is a relatively small dataset with low-quality images of indoor scenes and questions narrowly focused on colours, numbers and objects. DAQUAR was the first VQA dataset that has attracted considerable attention from the research community, but due to its small size and limited coverage, it is insufficient for a thorough evaluation of modern VQA models. A significantly larger COCO-QA dataset (M. Ren et al., 2015) includes real-world images from MS-COCO and questions automatically generated from captions. Although this dataset attempted to increase the scope of VQA data, an automated annotation significantly restricted the variety of questions.

The seminal dataset that enabled the large-scale research in VQA and remains one of the most popular benchmarks in the field is VQA dataset (Antol et al., 2015). It consists of two sets: natural images collected from MS-COCO (VQA-real) and abstract cartoon images (VQA-abstract). Generally, each image in the dataset has three related questions and ten answers per question collected from different annotators. The annotators were encouraged to provide the questions with varied types, difficulty and levels of knowledge required. Despite the well-defined procedures for careful and diverse data annotation, the analysis of the



How many tables are there in the image?
Answer: 4

(a) DAQUAR

What is this?
Answer: dollar

(b) VizWiz

Is the man skateboarding on a boardwalk?

Answer (left image): yes
Answer (right image): no

(c) VQA v2

What thing in this photo can protect a head from impact?
Fact: helmets can prevent head injuries
Answer: helmet

(d) FVQA

Who is in the left?
Answer: John Roberts

(e) KVQA

What part of this animal is sold illegally?
Answer: tusk

(f) OK-VQA

How many large things are either purple cylinders or cyan metal objects?
Answer: 1

(g) CLEVR

Are there both bikes and cars in this scene?
Answer: no

(h) GQA

Figure 2.5. Examples from eight VQA datasets.

data clearly reveals the presence of bias. For instance, the model can achieve almost 50% of accuracy by looking at the questions only (K. Kafle and Kanan, 2016), revealing the lack of visual grounding in image-question pairs. Moreover, it is possible to reach high performance just by answering "yes" to all binary questions (P. Zhang et al., 2016) which indicates a strong labelling bias. An updated version of the dataset, called VQA v2 (Goyal et al., 2017), was introduced in order to reduce these biases. Each question in this dataset is connected with two images that lead to two different answers. It addressed the issue of visual grounding deficiency, but the problem with the imbalanced distribution of question and answer types remained unsolved. For that reason, VQA-CP dataset (Agrawal et al., 2018) was proposed. It was created by re-organising training and validation sets of VQA v2.0 such that distributions of answers for each question type are different in training and test splits. A noticeable decline in performance between original VQA v2 and VQA-CP splits for all the models reported proves that they are prone to memorising the superficial correlations in the data and can not generalise well.

One of the crucial features of VQA data is the size since small datasets are usually inadequate to reflect the difficulty of the VQA task. Visual Genome QA dataset (Krishna et al., 2017), for example, includes over 1.7 million question-answer samples and is one of the largest in the field. It comprises questions divided into seven categories according to their first words: *what*, *where*, *when*, *who*, *why*, *how* and *which*. The main features of the dataset are high answer diversity (more than 200,000 unique answers), absence of binary questions and strong visual grounding. Moreover, each image is accompanied by the scene graphs and region descriptions that can be used for additional supervision or data augmentation. A newer GQA dataset (Hudson and Manning, 2019a), based on Visual Genome, includes 22 million questions requiring different reasoning skills. The main focus of this dataset is the strong visual grounding, spatial understanding and multi-hop inference. It is the first known attempt to build a large-scale dataset for compositional reasoning over real-world images, while all previous compositional datasets (*e.g.* SHAPES Andreas et al., 2016b and CLEVR Johnson, Hariharan, van der Maaten, Fei-Fei, et al., 2017) only include synthetic images of geometric objects.

Another line of works has been aimed at creating smaller but more challenging datasets that test models' ability to reason beyond the training data. In zero-shot VQA dataset (Teney and van den Hengel, 2016) the test set contains words never seen during training which helps to measure how the model generalises to novel concepts. FVQA (P. Wang et al., 2017b) dataset consists of questions requiring basic commonsense or factual knowledge. It also provides additional supporting facts collected from different knowledge sources to enable explicit reasoning. Similarly, in OK-VQA dataset (Marino et al., 2019) most of the questions depend on external knowledge although it is not restricted by any specific knowledge base and can be retrieved from any source. The largest knowledge-based dataset, KVQA (Shah et al., 2019), includes more than 180,000 questions but focuses only on facts about famous people obtained from Wikipedia. Overall, knowledge-demanding datasets are still one of the most difficult settings in VQA, so even current state-of-the-art models achieve relatively low performance (Marino et al., 2021; Z. Zhu et al., 2020).

While all the described datasets were designed for the research purpose primarily, VizWiz dataset (Gurari et al., 2018) is the first one designed with an applied goal in mind. The data for VizWiz was collected by blind people for the purpose of assisting visually impaired people in their everyday life. The most distinguishing feature of this dataset is that the data comes from a natural setting where images are taken with mobile phone cameras and questions are recorded in a spoken language. Such natural data collection process results in lower quality samples due to blur, bad lighting conditions, cropping, *etc.*, but at the same time induces the task to conform with real-world conditions. Therefore, VizWiz is one of the most challenging datasets that reveals the need for making current VQA models more suitable for practical applications.

2.2.2 Metrics

The evaluation of VQA approaches is a complex task itself. To determine whether the predicted answer is correct, one must take multiple aspects into consideration. Such language properties as synonyms, homonyms, paraphrasing and grammar make the evaluation of natural language sentences nontrivial due to the ambiguity. From this angle, the multiple-choice setting is the easiest for evaluation since simple accuracy can be calculated. The predicted answer is deemed correct if it exactly matches the ground truth label. The total accuracy is then calculated as the ratio of correct predictions to all predictions. With the open-ended setting, evaluation is not that straightforward because generated answers are not limited to any set. In practice, however, open-ended tasks are often treated as multiple-choice ones where the set of candidate answers for each question is equal to the set of all possible answers in the dataset. Such simplification allows for faster evaluation but reduces its quality and impedes thorough analysis.

Several metrics have been proposed to deal with ambiguity. Malinowski and Fritz, (2014) suggested using WUPS (Z. Wu and Palmer, 1994) metric to match the answers based on their semantic similarity. With that, the answer gets a high score if it is semantically or lexically close to the ground truth even though it is not an exact match. However, this metric also assigns high scores to semantically similar but quite different answers like *red*, *blue* and *white*, or even antonymous answers like *left* and *right*, *yes* and *no*. Moreover, WUPS can not be directly used for multiple-word answers which makes this metric inapplicable to modern VQA datasets. The authors of GQA dataset (Hudson and Manning, 2019a) proposed a set of new metrics to assess models' consistency and validity. These metrics can not replace the original accuracy metric but provide a deeper insight into models' behaviour – whether the model chooses reasonable answers or not. The main drawback of the proposed metrics is the need for collecting additional labels to mark plausible and valid answers which may not be feasible for large datasets with human annotators.

The most popular VQA metric is the modified accuracy proposed by Antol et al., (2015) for VQA v2 dataset. Each question in this dataset has ten ground-truth answers from different annotators, and the answer gets the full score if at least three annotators provided it. The accuracy is then calculated as

$$accuracy = \min\left(\frac{n}{3}\ ,\ 1\right)\ , \tag{2.1}$$

where n is the number of people who provided the same answer as the model. Although this metric helps to diminish the ambiguity problem, it is intrinsically dependent on the quality of annotations. The presence of noise in labels and low inter-human agreement can lead to the model achieving a high score for common but false answers and, on the contrary, getting a low score for accurate but rare answers. Despite its disadvantages, this metric has been widely adopted by other datasets and remains the dominant measure of performance for VQA.

2.3 Summary

In this chapter, we reviewed current literature in the VQA field that covers proposed methods, datasets and evaluation techniques. The first VQA methods used simple architectures that combine separate image and question feature encoders, a feature fusion module and a classifier. They, however, could not properly encode the fine-grained image information required for accurate question answering. For that reason, various attention mechanisms have been adopted to help models learn relevant local image features. While the majority of papers aim to boost VQA performance in general, a line of work focuses on solving specific tasks, including answering compositional questions and questions requiring external knowledge. Finally, in recent years, the whole VQA research area has been dominated by transformer-based architectures that have achieved tremendous success in computer vision and NLP tasks and have been successfully extended to the multi-modal VQA problem. Despite the variety of existing methods, current state-of-the-art accuracy on complex, compositional and knowledge-demanding benchmarks is far from human performance, which signifies the need for further advances in the VQA field.

The choice of architecture for a VQA model usually depends on the complexity of the task to be solved. In Chapter 3 we explore the potential of unsupervised pre-training with energy-based and contrastive learning. Current energy-based models can not be trained with modern deep learning components (Du, Li, Tenenbaum, et al., 2020). Therefore, we adopt a simple CNN-LSTM baseline architecture and create a dataset of plain images that can be efficiently learned with such architecture. In Chapter 4 we use a more advanced attention-based model (Y. Jiang et al., 2018) that was state of the art at the time of this work. Our main goal is to investigate how the use of answer semantics can boost VQA performance, so we choose a popular GQA dataset (Hudson and Manning, 2019a) with a wide range of metrics for the model's evaluation. In Chapter 5 we tackle a complex task of knowledge-demanding VQA, hence a more powerful transformer-based model (Tan and Bansal, 2019) is used as a backbone. We test our method on four knowledge-based and general visual reasoning datasets (Marino et al., 2019; Suhr et al., 2019; P. Wang et al., 2017b; N. Xie et al., 2019) to show the benefits of commonsense and factual knowledge embedding.

Chapter 3

Exploring Self-Supervised Pre-Training for Visual Question Answering

In this chapter, we explore unsupervised image feature pre-training for the visual question answering task. The availability of clean and diverse labelled data has always been one of the major driving forces in VQA research. However, the human annotation of multi-modal vision and language data is an expensive process that can become a bottleneck for VQA development. In this study, we adopt energy-based and contrastive learning – two popular unsupervised approaches – for image feature pre-training. We show that both methods can learn efficient image representations from unlabelled images and benefit the VQA task when the amount of annotated data is limited.

3.1 Introduction

The presence of large-scale diverse datasets has become one of the major driving forces of deep learning research in recent years. Datasets such as ImageNet (J. Deng et al., 2009) and Microsoft COCO (T.-Y. Lin et al., 2014) have significantly advanced computer vision and contributed to the development of widely-used image classification (Dosovitskiy et al., 2020; K. He et al., 2016; Simonyan and Zisserman, 2015; Szegedy et al., 2016; S. Xie et al., 2017) and object detection (K. He et al., 2017; Redmon et al., 2016; S. Ren et al., 2015) methods. In Natural Language Processing (NLP) research, datasets like WordNet (Miller, 1998), BookCorpus (Y. Zhu et al., 2015), WebText (Radford et al., 2019) and SQuAD (Rajpurkar et al., 2016) have facilitated the development of various NLP sub-fields. The task of visual question answering (VQA), which combines computer vision and NLP, consists of two distinct modalities and requires a large amount of aligned visual and textual training data. That complicates the collection of diverse data and thus impedes the progress of VQA.

The generation of natural language annotations in the form of questions and answers typically requires the involvement of human annotators and establishing accurate labelling procedures. Besides being an expensive and time-consuming operation, human annotation introduces biases and noise into data (K. Kafle and Kanan, 2017b; K. Kafle et al., 2019) which harms VQA training and evaluation. To speed up the process and reduce human error, several works (Hudson and Manning, 2019a; Johnson, Hariharan, van der Maaten, Fei-Fei, et al., 2017; M. Ren et al., 2015) have proposed the use of template-based question generation from image scene graphs or captions. Although this setting allows for controlled and compositional question generation, it essentially contradicts the definition of VQA which implies that the question is formulated freely with natural language. Given all the challenges of VQA data acquisition, researchers are now widely using transfer learning and pre-training to exploit the data from VQA-related domains.

Most VQA methods make use of pre-trained image models (Anderson, He, et al., 2018; K. He et al., 2016) and language features (Mikolov et al., 2013; Pennington et al., 2014) to ease the complexity of VQA training. As these features were trained separately on distinct computer vision and NLP tasks, they may be less suitable for combined vision and language tasks. Therefore, recent studies (Y.-C. Chen et al., 2020; L. H. Li et al., 2019; X. Li et al., 2020; Tan and Bansal, 2019) have explored joint pre-training with a large amount of aligned vision and language data. With this massive pre-training, the model is expected to learn general vision and language representations, and then it needs much less task-specific data to be fine-tuned for a downstream task. However, this setting still requires human annotations such as image captions or object labels and finding suitable pre-training data may become a bottleneck for VQA research. That motivates our search for unsupervised pre-training strategies.

A popular direction in unsupervised machine learning research is generative models (Goodfellow et al., 2014; Kingma and Welling, 2013). Although they are mostly used for purely generative tasks, recent advances with Energy-Based Models (EBMs) (Du and Mordatch, 2019; Grathwohl et al., 2019; Zhao et al., 2020) have shown that generative methods can benefit discriminative downstream tasks with improved calibration, robustness and out-of-distribution detection. In this work, we explore the applicability of EBMs to unsupervised image feature pre-training for VQA. We compare them with contrastive learning – another popular unsupervised method. We show that both these methods can learn efficient image representations from unlabelled data. Moreover, both models can be further fine-tuned with as little as 72 annotated VQA samples and still show a strong ability to generalise to unseen data.

3.2 Related Work

3.2.1 Energy-Based Models

Energy-Based Models have been used extensively for data modelling across different research fields (see LeCun et al., 2006 for a comprehensive review) due to their simple and nonrestrictive formulation. Despite these advantages, the complexity of EBMs training has prevented them from gaining much attention from the deep learning community. However, recent studies (Du and Mordatch, 2019; Nijkamp et al., 2020; Nijkamp et al., 2019) have investigated training techniques that enable the application of EBMs to high-dimensional data and improved the training stability. That allowed one to apply energy-based approaches in different fields of deep learning, including image generation (Arbel et al., 2020; Du, Li, and Mordatch, 2020; Du and Mordatch, 2019; Han et al., 2019; Xiao et al., 2020; J. Xie et al., 2018), graph generation (Suhail et al., 2021), image classification (Grathwohl et al., 2019), regression (F. Gustafsson et al., 2020; F. K. Gustafsson et al., 2020), continual learning (S. Li et al., 2020) and natural language processing (Y. Deng et al., 2019; T. He et al., 2021; L. Tu et al., 2020). In this study, we explore the benefits of energy-based training for visual question answering.

3.2.2 Contrastive Learning

Traditional supervised learning methods have dominated the field of computer vision since the rise of deep learning. However, their vital need for a large amount of annotated data has urged researchers to seek alternative approaches that do not require an expensive labelling process. Contrastive learning, a self-supervised discriminative approach, provides such an alternative where the data itself is a source of supervision and the model is trained to differentiate similar samples from dissimilar ones. Early attempts to apply contrastive methods to computer vision tasks (Bachman et al., 2019; K. He et al., 2020; Henaff, 2020; Oord et al., 2018; Tian et al., 2020; Z. Wu et al., 2018; Zhuang et al., 2019) showed promising results but could not compete with their supervised counterparts. More recent methods, like SimCLR (T. Chen et al., 2020) and SwAV (Caron et al., 2020), introduced novel data augmentations and architectural modifications that helped to significantly reduce the gap and achieve results comparable to supervised methods. In this work, we apply contrastive learning to the visual question answering task. Another VQA method that incorporates contrastive loss was proposed by Whitehead et al., (2021), where the model is trained on image-question pairs in a self-supervised manner. In contrast, in this work we utilise images without any additional annotations.

3.3 Methodology

In this work, we experiment with two techniques that enable self-supervised training, namely energy-based models and contrastive learning. We use a self-supervised strategy to pre-train a Convolutional Neural Network (CNN) on unlabelled images. These pre-trained weights are then used to initialise the image feature extractor of the VQA model, which is further fine-tuned on a small set of labelled data (*i.e.* image, question and answer triplets). With this setting, we aim to investigate how pre-training can improve the model's generalisation ability when the annotated data is limited.

3.3.1 Energy-Based Learning

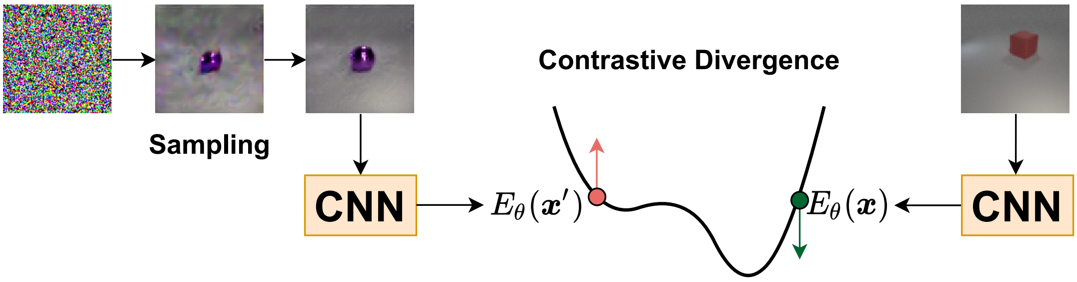


FIGURE 3.1. An illustration of energy-based model training.

Energy-Based Model is a statistical model that relies on energy function to capture dependencies between variables. The energy function maps each configuration of variables to a scalar energy value, such that correct values of variables have lower energy (and hence higher probability). With this formulation, the probability density for an input $\boldsymbol{x} \in \mathbb{R}^D$ can be represented as

$$p_\theta(\boldsymbol{x}) = \frac{\exp\left(-E_\theta(\boldsymbol{x})\right)}{Z(\theta)} , \tag{3.1}$$

where $E_\theta(\boldsymbol{x}) : \mathbb{R}^D \rightarrow \mathbb{R}$ is the energy function, $Z(\theta) = \int \exp\left(-E_\theta(\boldsymbol{x})\right) d\boldsymbol{x}$ is the partition function and θ is the model's parameters. In this work, we parameterise the energy function with a CNN that takes image as an input and returns a scalar.

In practice, the partition function $Z(\theta)$ is usually intractable to compute, which means that the standard maximum likelihood approach can not be applied directly to train the model. However, we can instead use the gradient-based optimisation approach, since the derivative of the log-likelihood of a data sample $\boldsymbol{x}$ does not require the partition function to be calculated:

$$\frac{\partial \log p_\theta(\boldsymbol{x})}{\partial \theta} = \mathbb{E}_{p_\theta(\boldsymbol{x}')}\left[\frac{\partial E_\theta(\boldsymbol{x}')}{\partial \theta}\right] - \frac{\partial E_\theta(\boldsymbol{x})}{\partial \theta} \,, \tag{3.2}$$

where $\boldsymbol{x}'$ is sampled from the model distribution. Sampling $\boldsymbol{x}'$ from p_θ is not a trivial task, but recent works (Du and Mordatch, 2019; Nijkamp et al., 2019) have proposed to use Markov Chain Monte Carlo (MCMC) sampling based on Langevin dynamics (Welling and Teh, 2011) that iterates

$$\boldsymbol{x}'_k = \boldsymbol{x}'_{k-1} - \frac{\lambda}{2}\frac{\partial E_\theta(\boldsymbol{x}'_{k-1})}{\partial \boldsymbol{x}'_{k-1}} + \omega_k \,, \quad k = 0, 1, ..., K \,, \tag{3.3}$$

where k numerates steps, λ is the step size, $\omega_k \sim \mathcal{N}(0, \lambda)$, and $\boldsymbol{x}_0$ is typically initialised with uniform random noise. This procedure defines a distribution q_θ and, as shown in (Welling and Teh, 2011), if $K \rightarrow \infty$ and $\lambda \rightarrow 0$ then $q_\theta \rightarrow p_\theta$. In practice, the sampling runs for the finite number of steps and we differentiate only through the last step to reduce the computational cost, as done in (Du, Li, Tenenbaum, et al., 2020). The model is then trained with contrastive divergence (Hinton, 2002) objective that aims to minimise the energy of the training data and, on the contrary, maximise the energy of the generated samples (see Figure 3.1 for the illustration of model's training).

3.3.2 Contrastive Learning

Contrastive learning is another self-supervised technique whose core idea is to learn data representations such that similar samples are grouped together and dissimilar ones are pushed apart. In the absence of ground-truth annotations, all samples in a dataset are considered to belong to different classes, while several augmentations (also called views) of one sample constitute one class.

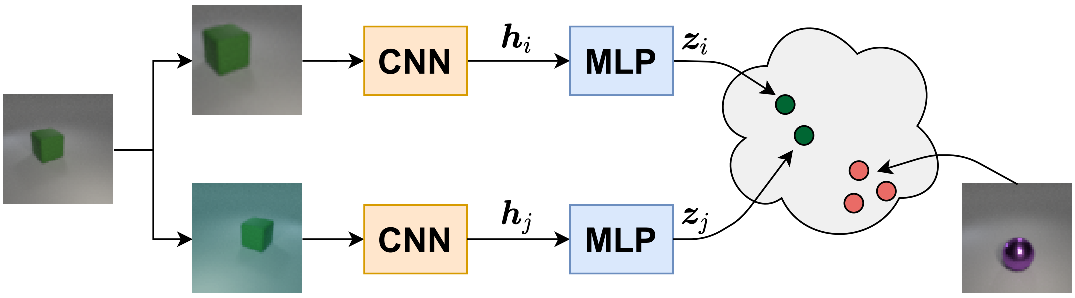


FIGURE 3.2. An illustration of contrastive learning.

Typically, the key components of contrastive learning include (1) data augmentation, (2) encoding, and (3) contrastive loss that maximises the similarity between encodings of one class and minimises inter-class similarity (Figure 3.2).

In the first step, an input image $\boldsymbol{x}$ is transformed with two sets of augmentations to get two correlated views $\widetilde{\boldsymbol{x}}_i$ and $\widetilde{\boldsymbol{x}}_j$. These views are passed through an encoding function $f(\cdot)$, which is typically a CNN, and obtained encodings are mapped into a common space with the projection head $g(\cdot)$, commonly implemented as a Multilayer Perceptron (MLP):

$$\begin{aligned} \boldsymbol{h}_i &= f(\widetilde{\boldsymbol{x}}_i)\ , \\ \boldsymbol{h}_j &= f(\widetilde{\boldsymbol{x}}_j)\ , \\ \boldsymbol{z}_i &= g(\boldsymbol{h}_i)\ , \\ \boldsymbol{z}_j &= g(\boldsymbol{h}_j)\ . \end{aligned} \tag{3.4}$$

The final representations $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are called a positive pair, while the combinations of views from different images constitute negative pairs. The model's objective is then to maximise the similarity of the positive pair while minimising it for all other negative pairs. A common choice for this objective is the normalised temperature-scaled cross-entropy loss with the cosine similarity:

$$\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j) \ = \ \frac{\boldsymbol{z}_i^\top \boldsymbol{z}_j}{||\boldsymbol{z}_i|| ||\boldsymbol{z}_j||}\ , \tag{3.5}$$

$$l(i, j) \ = \ -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau}{\sum_{k, k \neq i} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}\ , \tag{3.6}$$

where τ is a temperature parameter and k iterates over all negative samples, which are usually sampled from the the same batch. The final loss is calculated as the sum of losses for all positive pairs in a batch.

3.3.3 Fine-Tuning

During the pre-training stage, an image feature encoder f^I is trained with either energy-based or contrastive learning strategy. For the fine-tuning step, we combine the pre-trained image encoder with a question encoder f^Q which is a Long Short-Term Memory (LSTM) network initialised with random weights. These encoders process image $\boldsymbol{x}^I$ and question $\boldsymbol{x}^Q$ inputs to get visual $\boldsymbol{v} \in \mathbb{R}^{D_I}$ and textual $\boldsymbol{q} \in \mathbb{R}^{D_Q}$ representations respectively:

$$\begin{aligned} \boldsymbol{v} &= f^I(\boldsymbol{x}^I) \ , \\ \boldsymbol{q} &= f^Q(\boldsymbol{x}^Q) \ . \end{aligned} \tag{3.7}$$

Later, feature encodings $\boldsymbol{v}$ and $\boldsymbol{q}$ are concatenated into a single vector $\boldsymbol{p} = [\boldsymbol{v}, \boldsymbol{q}]$, $\boldsymbol{p} \in \mathbb{R}^{D_I + D_Q}$. The fused vector is passed through a classifier to obtain answer scores $\boldsymbol{y} = f^{CLS}(\boldsymbol{p})$, with $\boldsymbol{y} \in \mathbb{R}^A$ where A is the size of a set of candidate answers. The model is then trained to minimise cross-entropy loss:

$$\text{softmax}(y_i) = \frac{\exp(y_i)}{\sum_{j=1}^{A} \exp(y_j)} \ , \tag{3.8}$$

$$L_{CE} = -\sum_{i=1}^{A} \hat{a}_i \cdot \log(\text{softmax}(y_i)) \ , \tag{3.9}$$

where $\hat{\boldsymbol{a}} \in \{0, 1\}^A$ denotes the one-hot (multi-hot) vector of the ground-truth answer(s), and i indexes vector elements.

3.4 Experiments

3.4.1 Dataset

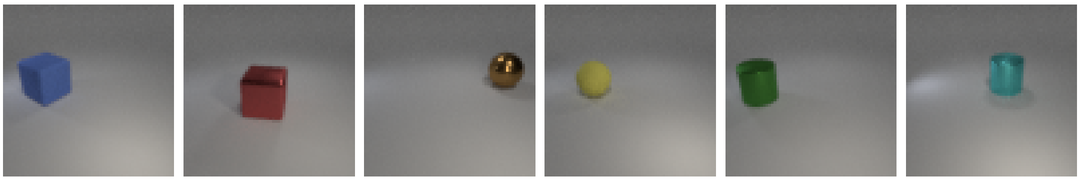


Figure 3.3. Example images from our dataset: simple objects with different shape-colour combinations (24 in total).

EBM that we use in our experiments includes a computationally expensive MCMC sampling process that impedes scaling to large image datasets. As a result, recent EBM approaches (Du, Li, and Mordatch, 2020; Grathwohl et al., 2019) have focused on datasets with small-size images of simple objects (*e.g.* MNIST L. Deng, 2012, CIFAR Krizhevsky, Hinton, et al., 2009 or CelebA Z. Liu et al., 2015). Current datasets designed for VQA (see Section 2.2.1) consist of large-scale images, typically with multiple objects presented, what makes them unusable for EBM training. Therefore, we generated our own dataset to run experiments in a suitable and well-controlled setting.

Following the procedure similar to the one described in (Atzmon et al., 2020), we generated a small dataset of easy objects. All images were rendered with Blender (Community, 2018) software using CLEVR framework (Johnson, Hariharan, van der Maaten, Fei-Fei, et al., 2017). Each image of size 64x64 contains one object (sphere, cube or cylinder) with rubber or metallic material of one of the eight possible colours: red, purple, yellow, blue, green, cyan, grey or brown (Figure 3.3). For each image, we generated a related question based on the templates provided in CLEVR. Questions are designed to query the object's shape, for example, *There is a blue object in the image; what is it?* with words *sphere*, *cube* and *cylinder* being possible answers.

We generated 2000 samples per each object-colour combination where the object's size, material, position and lightning are chosen randomly. We further divided the whole dataset into several splits used for pre-training, fine-tuning, validation and test phases. Since we aim to investigate how the model can generalise to unseen data when the labelled data is limited, we split the data in a way similar to the CLEVR-CoGenT (Johnson, Hariharan, van der Maaten, Fei-Fei, et al., 2017) dataset creation. Specifically, we defined three data types according to possible object-colour combinations (see Table 3.1 for the description of types). Thus, our pre-training split contains data of type A, the fine-tuning split has data of type B, the validation split includes both B and C types, and the test split has only type C data. Table 3.2 contains the details for each split in our dataset.

3.4.2 Experimental Setting

Recent studies (Du, Li, Tenenbaum, et al., 2020; Du and Mordatch, 2019) have shown that the use of complex deep learning components can cause instability when training EBMs. Therefore, we selected a simple CNN-RNN architecture

Type	Object	Colours
A	cylinder	any
	cube	any
	sphere	any
B	cube	grey, blue, brown, yellow
	sphere	red, green, purple, cyan
C	cube	red, green, purple, cyan
	sphere	grey, blue, brown, yellow

TABLE 3.1. The list of available object-colour combinations for each data type.

Split	Data Type	Images	Questions
Pre-training	A	12,000	n/a
Fine-tuning	B	3,600	3,600
Validation	B+C	800	800
Test	C	3,600	3,600

TABLE 3.2. A summary of the data splits.

for our VQA model – a common baseline for the task as discussed in Section 2.1.1. We adopted the modified ResNet (K. He et al., 2016) architecture from Du, Li, Tenenbaum, et al., (2020) and combined it with LSTM network. Note that the same core ResNet architecture is used in all pre-training and fine-tuning experiments to enable transfer learning. We only modify the last few layers for each pre-training task as will be discussed later.

For EBM pre-training, a CNN model is followed by a linear layer that maps visual feature vectors into scalar energy values. To facilitate the training we incorporated several changes proposed by Du, Li, Tenenbaum, et al., (2020). Concretely, we use a replay buffer to store previously generated samples that are randomly chosen to re-initialise the sampling chain (Equation 3.3) instead of the uniform noise. We also apply random data augmentations (*e.g.* cropping, horizontal flipping, blurring and colour distortion) to images sampled from the buffer as it was proved to improve the diversity and mixing of sampling chains. Finally, we included additional losses that further improve contrastive divergence training (we refer the reader to the source paper Du, Li, Tenenbaum, et al., 2020 for details).

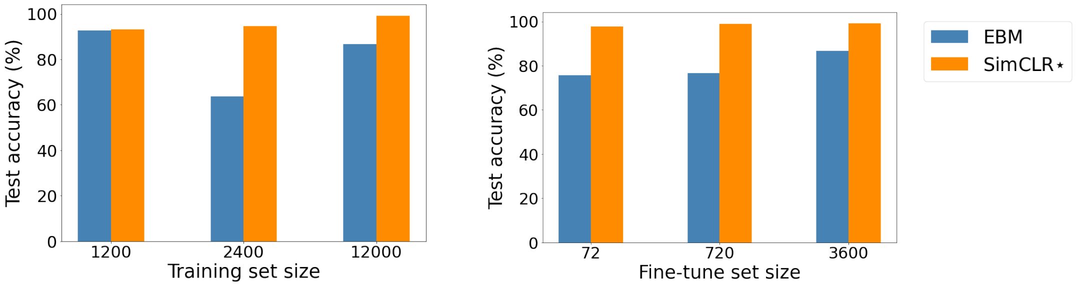
	Validation	Test		
		Total	Cube	Sphere
ResNet	99.43 ± 0.33	99.14 ± 0.20	98.35 ± 0.33	**99.92** ± **0.09**
Baseline	56.96 ± 6.08	12.78 ± 9.75	0.04 ± 0.06	25.52 ± 19.57
EBM	93.67 ± 3.18	86.74 ± 8.34	96.57 ± 1.99	76.91 ± 14.75
SimCLR⋆	**99.75** ± **0.33**	**99.26** ± **0.29**	**98.87** ± **0.27**	99.65 ± 0.32

Table 3.3. Results for validation and test splits. We report the average accuracy (%) ± one standard deviation over three random seeds. SimCLR⋆ model reaches highest accuracy on both validation and test sets.

For contrastive learning, we experiment with the popular SimCLR (T. Chen et al., 2020) model. SimCLR is a simple framework that uses ResNet-50 to extract visual representations, and a projection head to map them into a common space where contrastive loss is applied. We changed the CNN architecture as discussed above, but used the same projection head on top of it. This projection head is a MLP with one hidden layer and ReLU activation function. We denote this model as 'SimCLR⋆' in our experiments to emphasise the modified architecture. We also show results for our baseline CNN-LSTM model that was not pre-trained for any task but instead initialised with random weights (noted 'Baseline'). Finally, as a point of comparison, we include the results of ResNet-18 model pre-trained on ImageNet (J. Deng et al., 2009) for the classification task (noted 'ResNet'). That allows us to compare with the traditional supervised transfer learning approach commonly used for small labelled datasets.

3.4.3 Quantitative Results

The accuracy results for validation and test splits are shown in Table 3.3 (see Appendix A for additional results). SimCLR⋆ model achieves the highest accuracy on both splits. A minor difference between validation and test results (99.75% → 99.26%) indicates the model's ability to generalise to unseen data. EBM also shows decent generalisation performance (86.74% on the test set), but it comes with a high variance across different runs (standard deviation of 8.34 for the test set). It means that the model is more sensitive to random training parameters and produces less consistent results. SimCLR⋆ model, on the opposite, shows stable performance for all runs (standard deviation of 0.29).



(a) (b)

Figure 3.4. Test accuracy results for two unsupervised methods trained with different set sizes for (a) pre-training and (b) fine-tuning data.

Moreover, SimCLR⋆ scores on par with the fully supervised ResNet model, although ResNet contains 1.5 times more parameters and was pre-trained on a significantly larger image dataset. Finally, the baseline model trained from scratch shows the lowest accuracy (12.78%) in our experiments. Without any pre-training, the model quickly overfits the shape-colour combinations present in the fine-tuning set and uses colour attribute as a ground for shape prediction.

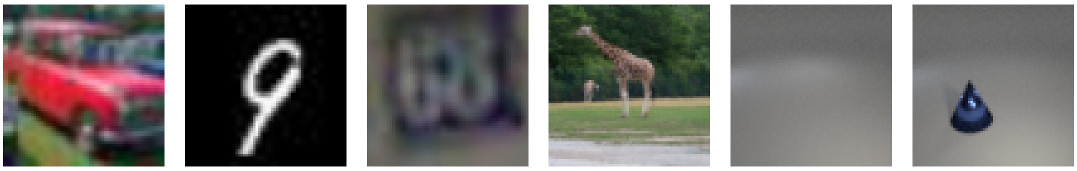
We want the model to learn the task with as little data as possible, so we study the effect of different pre-training (Figure 3.4a) and fine-tuning (Figure 3.4b) set sizes. The performance of SimCLR⋆ model only slightly degrades when the pre-training size is reduced five or even ten times, but it still outperforms EBM with all data sizes. Interestingly, EBM shows the highest accuracy with less pre-training data, while its performance first drops and then improves as the data size is growing. It has been shown that the training of EBM is unstable and highly depends on random parameters (Grathwohl et al., 2019). So the EBM's performance could be mainly influenced by the randomness in training, rather than by data size. As for fine-tuning step, both models can be trained with as little as 72 labelled samples. While the EBM's performance drops a bit with the reduced fine-tuning size, the accuracy of SimCLR⋆ model remains almost identical. Overall, both models do not require much data to learn useful image features and can be successfully fine-tuned to a downstream task with limited labelled data.

In conclusion, we note that although EBM shows competitive results, the difficulty and constraints of its training outweigh the possible benefits. The contrastive learning method, on the contrary, shows stable superior performance across all experiments and does not impose restrictions on image size and model architecture. Furthermore, contrastive learning proved to be a good alternative for traditional supervised transfer learning, since the latter may not be applicable when pre-training and downstream task domains do not intersect (for example, in medical VQA X. He et al., 2020; Lau et al., 2018).

3.4.4 Out-of-Distribution Detection

One of the main advantages of EBMs that has attracted researchers is their ability to detect out-of-distribution data (Elflein et al., 2021; W. Liu, Wang, et al., 2020). In VQA, models typically can not distinguish between in-distribution and out-of-distribution samples and can not tell if a particular question about a particular image is unanswerable. Although in the real-world setting, it is a common case when users provide images of low quality or irrelevant to the question, and a model is just unable to find the correct answer to the question asked (Bhattacharya et al., 2019; Chiu et al., 2020). We thus aim to test our models for their out-of-distribution detection abilities.

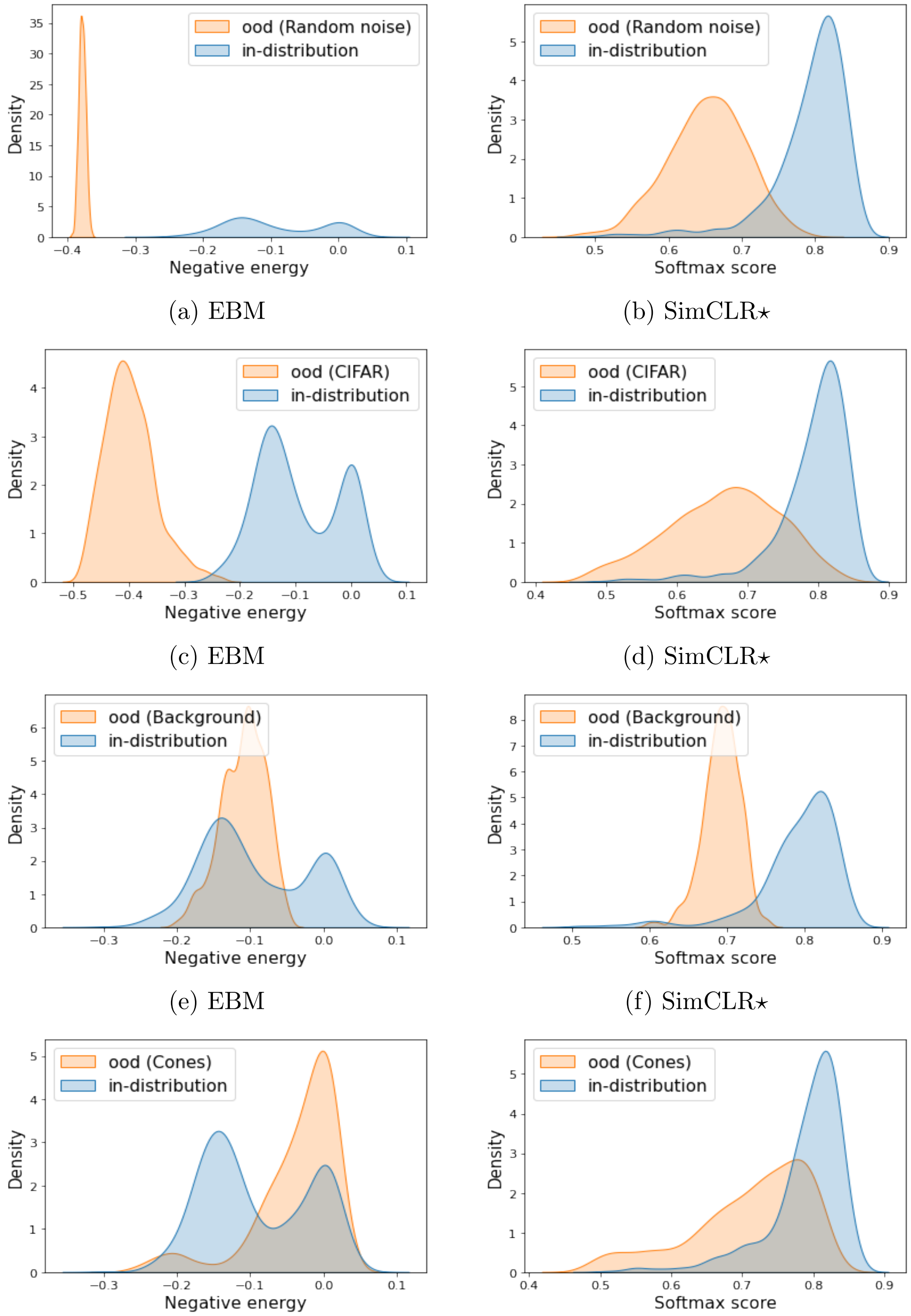
A simple way to detect out-of-distribution samples is to set a threshold on the softmax scores predicted by the model, such that all samples with low confidence scores are classified as out-of-distribution (Hendrycks and Gimpel, 2016). In energy-based models, each input is mapped to a scalar value, where training data is associated with lower energy values and unobserved data gets higher energy. Therefore, negative energy scores can be used instead of softmax confidence for out-of-distribution detection. In our experiments, we use four datasets: CIFAR (Krizhevsky, Hinton, et al., 2009), MNIST (L. Deng, 2012), SVHN (Netzer et al., 2011) and VQA v2 (Goyal et al., 2017); and the generated images of random noise, cone shapes and empty images of the background, to serve as an out-of-distribution data (examples are given in Figure 3.5).



CIFAR MNIST SVHN VQA v2 Background Cones

FIGURE 3.5. Example images from out-of-distribution datasets.

To compare out-of-distribution detection abilities of the models, we compute Area Under the ROC Curve (AUROC) values using softmax and negative energy scores for SimCLR$\star$ and EBM methods respectively (Table 3.4). We also visualise the distributions of scores in Figure 3.6. Overall, energy scores help to almost perfectly distinguish between training data and external datasets



(a) EBM

(b) SimCLR⋆

(c) EBM

(d) SimCLR⋆

(e) EBM

(f) SimCLR⋆

(g) EBM

(h) SimCLR⋆

Figure 3.6. Distribution of energy and softmax scores from EBM and SimCLR⋆ models respectively. In-distribution (blue) data comes from the pre-training dataset and out-of-distribution (orange) data is collected from public datasets or generated manually.

	Noise	CIFAR	MNIST	SVHN	VQA v2	Background	Cones
EBM	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.52	0.30
SimCLR$\star$	0.96	0.92	0.97	0.91	0.94	**0.94**	**0.82**

TABLE 3.4. AUROC scores for out-of-distribution detection. EBM ranks first for all external datasets, but fails to distinguish samples of alternated original data (background and cone images).

(AUROC of 1.0 in five experiments). However, when out-of-distribution data is visually similar to the training one (*e.g.* images of background and cones) EBM's performance falls far below its softmax-based counterpart. Therefore, energy scores can be used to filter out the samples that are clearly distinct from the training data but are not sufficient to spot fine-grained differences.

3.4.5 Supervised Training with Energy-Based Models

One of the reasons behind our choice to explore energy-based models is their increasing popularity in multiple computer vision and natural language processing applications. This recent interest is partially caused by the EBM's ability to solve generative and discriminative tasks simultaneously and improve uncertainty calibration. We aim to investigate whether EBMs can benefit VQA task in a similar way. We test two popular energy-based methods, namely JEM (Grathwohl et al., 2019) and Conditional EBM (CEBM) (Du, Li, Tenenbaum, et al., 2020; Du and Mordatch, 2019) with fully supervised training on the VQA task. The dataset for these EBMs is similar to the one used in the main experiments (Section 3.4.1). The training split contains 16,000 images (each with one corresponding question about the shape), where spheres are of any colour, cubes are either grey, blue, brown or yellow, and cylinders are either red, green, purple or cyan. Test split has 5,000 images with spheres of any colour, and cubes and cylinders having swapped colour sets. As a baseline, we use a simple classifier model that has the same CNN-LSTM architecture as JEM and CEBM in our experiments.

JEM method proposes to treat a discriminative classifier as an energy-based model and optimise two objectives simultaneously:

$$\log p_\theta(\boldsymbol{x}, y) \;=\; \log p_\theta(\boldsymbol{x}) + \log p_\theta(y|\boldsymbol{x}) \ , \tag{3.10}$$

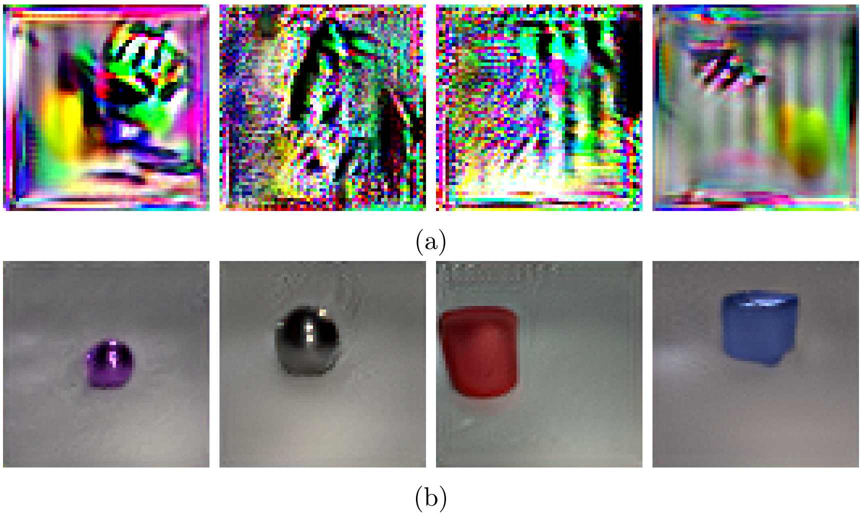


Figure 3.7. Examples of generated images with (a) high and (b) low energy.

where $\boldsymbol{x}$ is a data point, and y is a class label. This objective can be optimised with standard cross-entropy for classification part and log-likelihood (Equation 3.2) for energy learning. The energy function is defined as negative LogSumExp$(\cdot)$ function of the logits of the classifier:

$$E_\theta(\boldsymbol{x}) \;=\; -\log \sum_y \exp(f_\theta(\boldsymbol{x})[y]) \; . \tag{3.11}$$

During experiments, we found it difficult to balance both losses to make sure that the model learns to both correctly classify answers and generate new images. Typically, classification loss converges much faster and, to minimise energy loss, the model starts to generate non-realistic images that always give high energy. With this setting (denoted as 'JEM' in results), the inclusion of energy learning in the classifier becomes meaningless. To mitigate the dominance of classification loss, we multiplied it by the weight parameter (0.1 value showed the best results through parameter search). That slowed down the convergence rate which gave the model more time to learn to produce naturally-looking images. Nevertheless, we observed that at some point during training the model always starts to generate noisy images that are easily distinguishable from the real ones (examples in Figure 3.7a). We thus introduced early stopping that stops the training when energies of real and generated images diverge from each other too far (*i.e.* when $|E_\theta(\boldsymbol{x}') - E_\theta(\boldsymbol{x})| > 0.8$), which means that the model is no longer generating realistic images.

CEBM proposes to learn conditioned energy function $E_\theta(\boldsymbol{x}|c)$. It is built upon a CNN architecture with conditional gains and biases (Dumoulin et al., 2016) that was designed to generate images conditioned on style. Although CEBM is mainly designed for the generation task, it shows robust classification performance (Du and Mordatch, 2019) when the energy of images conditioned on class labels is used to predict the label:

$$y^{\star} \;=\; \arg\min_{y} E_{\theta}(\boldsymbol{x}|y) \;. \tag{3.12}$$

As CEBM is a generative model, there are no obvious stopping criteria for training. We used FID score (Heusel et al., 2017) to evaluate the model's performance and stopped the training when FID value became stable and generated images became similar to the real ones (by visual inspection of the results). However, we further observed that there is no correlation between the generative performance of the model measured by FID score and its classification results on the VQA task. The model's checkpoints from different epochs produce significantly different predictions. For a fair comparison, we report the results for the epoch that gives the best FID score.

Along with the accuracy performance, we measure the model's calibration as it has been shown that EBM training helps the model to achieve better calibration (Grathwohl et al., 2019; T. He et al., 2021). A model is considered well-calibrated if its confidence is aligned with the predicted accuracy. That is, the model is unsure about the wrong predictions and confident about the correct ones. We use a standard metric for calibration - Expected Calibration Error (ECE) (Guo et al., 2017). It splits the predictions into M equally-sized bins according to their confidence values and measures the weighted average of the difference between accuracy and confidence of each bin. The confidence is a probability score, or softmax output, for the predicted label:

$$\mathrm{acc}(B_m) \;=\; \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \;, \tag{3.13}$$

$$\mathrm{conf}(B_m) \;=\; \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \;, \tag{3.14}$$

$$\mathrm{ECE} \;=\; \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right| \;, \tag{3.15}$$

	Total	Sphere	Cube	Cylinder	ECE ↓
Baseline	28.46	**100.00**	0.00	0.00	71.54
CEBM	34.54	54.60	31.56	**21.10**	**9.24**
JEM	28.44	99.93	0.00	0.00	71.28
JEM (early stopping)	**45.86**	99.93	**39.23**	8.12	36.55

Table 3.5. Accuracy results and expected calibration error (ECE) for supervised energy-based training.

where B_m is a set of indices in a bin, n is the number of samples, $\hat{y}_i$ and $\hat{p}_i$ are ground truth label and probability score for the prediction y_i respectively. ECE values ranges from 0 to 100, where 0 indicates that the model is perfectly calibrated (*i.e.* probability score p_i is 1 for all correct predictions and 0 for all incorrect ones).

The results for supervised EBM training are given in Table 3.5. The simple classification baseline, unsurprisingly, learns shape-colour combinations seen during training and is therefore unable to predict novel combinations in the test set. That results in 100% accuracy for questions about spheres (present in the training set) and 0% accuracy for cubes and cylinders. ECE score is pretty high (71.54) since the model is confident in its wrong predictions. JEM shows similar results, although with a slightly smaller calibration error (71.28). As discussed above, in standard JEM training, classification objective quickly dominates energy learning which explains similar performance. JEM trained with classification loss weight and early stopping gives higher overall accuracy (45.86%) and, consequently, lower ECE (36.55). While the baseline classifier clearly overfits the training data, JEM can correctly answer some of the unseen samples, although it still often confuses cubes and cylinders with relatively high confidence. CEBM surpasses the baseline with 34.54% accuracy, but finer analysis suggests that this result is closer to random predictions. Both baseline and JEM show almost perfect accuracy for spheres, while CEBM achieves only 54.60% accuracy, meaning that the model is underfitted to the task. The lowest ECE score (9.24) is due to the model's low confidence for all the predictions and it does not imply good calibration. Overall, current EBMs are difficult to train and can be used with limited data and architectures which makes them unsuitable, in their current form, for such a complex task as VQA.

3.5 Conclusion

In this work, we explored the applicability of unsupervised image feature pre-training for visual question answering. We showed that two self-supervised methods, energy-based model and contrastive learning can learn image representations from unlabelled data sufficient for the downstream VQA task. Further, both methods are able to generalise to unseen test samples from a small annotated fine-tuning set. However, given the complexity of EBM training and its unstable results, we can conclude that contrastive learning is currently a more promising approach for unsupervised feature learning for the VQA task. Moreover, we found that a contrastive method performs on par with a larger fully-supervised model trained on a colossal ImageNet dataset.

Statement of Authorship

Title of Paper	Visual Question Answering with Prior Class Semantics
Publication Status	□ Published □ Accepted for Publication □ Submitted for Publication ☒ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Violetta Shevchenko, Damien Teney, Anthony Dick and Anton van den Hengel. "Visual Question Answering with Prior Class Semantics". arXiv preprint arXiv:2005.01239 (2020).

Principal Author

Name of Principal Author (Candidate)	Violetta Shevchenko		
Contribution to the Paper	Design and implementation of experiments, paper writing.		
Overall percentage (%)	70 %		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	18/12/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Damien Teney		
Contribution to the Paper	Discussions, paper writing.		
Signature		Date	21 Dec. 2021

Name of Co-Author	Anthony Dick		
Contribution to the Paper	Discussions, paper revision.		
Signature		Date	22 Dec 2021

Name of Co-Author	Anton van den Hengel		
Contribution to the Paper	Discussions, paper revision		
Signature		)ate	21/12/21

Chapter 4

Visual Question Answering with Prior Class Semantics

In this chapter we present a novel mechanism to embed prior knowledge in a model for visual question answering. The open-set nature of the task is at odds with the ubiquitous approach of training of a fixed classifier. We show how to exploit additional information pertaining to the semantics of candidate answers. We extend the answer prediction process with a regression objective in a semantic space, in which we project candidate answers using prior knowledge derived from pre-trained word embeddings. An extensive study of learned representations with the GQA dataset reveals that important semantic information is captured in the *relations* between embeddings in the answer space. The proposed method brings improvements in consistency and accuracy over a range of question types. Experiments with novel answers, unseen during training, indicate the method's potential for open-set prediction.

4.1 Introduction

Most recent developments in the field of visual question answering (VQA) have focused on the development of deep learning architectures that can be trained with end-to-end supervision (*i.e.* questions, images, and answers). However, even current large-scale datasets (Antol et al., 2015; Goyal et al., 2017; Hudson and Manning, 2019a) can only cover a limited fraction of all knowledge potentially useful for the task. The underlying reasons for this limitation are that (1) the collection of data with end-to-end annotations, *i.e.* questions/answers is expensive as it usually requires human resources, and (2) the desirable knowledge about the world is constantly expanding, and no single dataset can ever capture it all. Existing models trained once and for all on any of these datasets

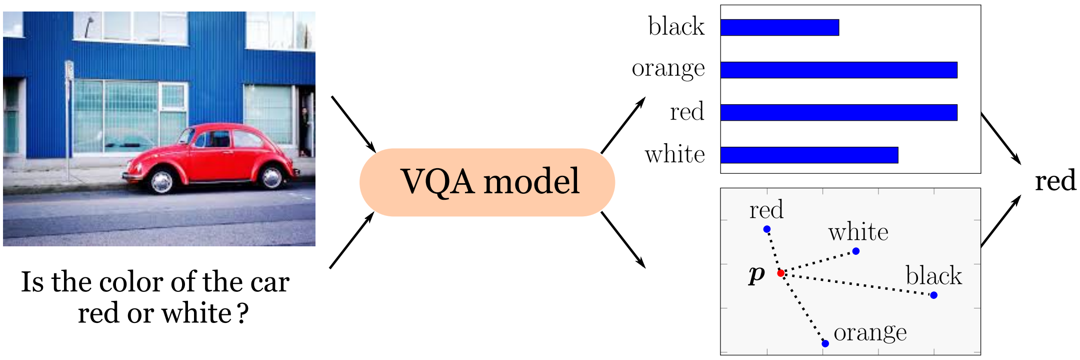


Figure 4.1. Existing models treat VQA as a classification task over predefined answers (upper branch). We supplement our model with a regression objective in a semantic answer space (lower branch). This allows incorporating additional prior knowledge about answer semantics. This improves its accuracy and consistency. In the above example, *red* and *orange* are similarly likely with the traditional objective. Our regression lands closer to the representation of *red* in the answer space. This resolves the ambiguity and *red* is chosen as the final answer.

lack the generalisation and adaptation capabilities desirable in real-world applications. These shortcomings motivate our search for alternative sources of information, and a method to exploit them in a VQA model.

A common approach to include existing knowledge in VQA models is to use pre-trained models to obtain image and question features. On the image side, pre-trained Convolutional Neural Network (CNN) or object detectors are ubiquitous (Anderson, He, et al., 2018) to extract representative image features. On the language side, pre-trained word embeddings like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) usually serve to encode the words of the question. The advantage of these techniques is to leverage knowledge learned from larger, non-VQA specific data (*e.g.* ImageNet J. Deng et al., (2009) and large text corpora). The benefit of these approaches has been widely demonstrated, which further motivates our quest for additional sources of usable knowledge and techniques to incorporate it.

Existing models for VQA follow the common blueprint of a two-stream embedding, followed by fusion and classification stages (Antol et al., 2015; Teney, Anderson, et al., 2017; Z. Yang et al., 2016). The typical setting in VQA consists of an image and a related question. The model takes this image-question pair and predicts the correct answer by solving a classification problem over the set of candidate answers that occur in the training data. This classification approach, in contrast to text generation (Gao et al., 2015; Q. Wu, Shen, et al., 2017), considerably simplifies the evaluation process, as the model can be assessed by its classification accuracy. However, treating VQA as a classification task has major drawbacks. The answers are treated as distinct class labels and answer words are abstracted from their meanings. This disregards semantic relations between related answers. Moreover, some questions contain possible answers in their wording (*e.g. Is this car red or white?*) and it seems natural to include mechanisms to explicitly represent the semantics of possible answers as it is done for question words. Guided by these observations, we develop an architecture that leverages prior knowledge about answers to improve the performance of a VQA model.

Our main technical contribution is to treat VQA as a multitask problem, where we both predict the answer label based on classification scores, and we additionally learn a mapping into an answer representation space that captures the semantics of these answers (Figure 4.1). We incorporate prior knowledge into the model by initialising the representations of answers with pre-trained word embeddings. We perform an extensive and rigorous analysis of the trained model. It demonstrates the benefits of the approach and provides us with insights in the ways language semantics are useful for the task of VQA. Moreover, we show that learned answer representations can be used for out-of-vocabulary answer prediction which is an important, yet understudied problem in VQA field (Noh et al., 2019).

The contributions of this work are as follows.

- We formulate VQA as a multitask problem, where we train the model, not only to assign scores to answer candidates but also to perform a regression in a vector space that represents answer semantics.
- We use this multitask formulation to incorporate additional information into the model with a particular loss and initialisation of the semantic answer space. We also show that it allows the model to predict novel answers that were not seen during training.
- We perform an extensive analysis of the model and various ablations. We demonstrate clear advantages on the GQA dataset (Hudson and Manning, 2019a), and obtain insights on the ways in which answer semantics are useful for the task of VQA.

4.2 Related Work

The overarching motivation for research on VQA is that of tackling a complex, open-world and multi-modal task. These aspects are among the foundations required in general artificial intelligence (AI) systems. While the task has attracted considerable attention over the past few years (K. Kafle and Kanan, 2017b; Teney, Wu, et al., 2017), its open-set and open-domain aspects have largely been overlooked. The overarching motivation for research on VQA is that of tackling a complex, open-world and multi-modal task. The common practice of training a model with end-to-end supervision using a fixed dataset is inherently limited. Our discussion focuses on the incorporation of external knowledge and training signals into VQA models.

4.2.1 Answer Embeddings for VQA

Most techniques to incorporate additional information into VQA models are based on representations of language, both of questions and of candidate answers. In (Teney and van den Hengel, 2016) pre-trained word embeddings are used as bag-of-words representations of candidate answers, which are passed to the network as additional inputs, along with question and image features. In (Teney, Anderson, et al., 2017) authors proposed to initialise the weights of the output classifier with pre-trained answer embeddings. They used both a textual branch, initialised with GloVe vectors, and a visual one, initialised with visual features from images representing the candidate answers. In (H. Hu et al., 2018), the authors propose to learn two sets of embeddings, image-question and answer vectors. They optimise a projection of these two embeddings into a joint space where the distances between compatible pairs are minimised. Their experiments showed interestingly that the learned projections was transferable, to some extent, across datasets with different sets of possible answers.

Different from the methods cited above, our model forgoes the notion of a fixed answer set, and the output of the network is a location in a space representing answer semantics. The final prediction is still obtained by searching for the closest representation among answer candidates in this same space, but the formulation offers improved flexibility. This allows us to explore different distance measures in this semantic space. It also allows control over the contribution made by prior and task-specific data. Finally, it easily accommodates multiple representations of a same answer, thereby accounting for polysemy and context-dependent meaning of certain words and expressions.

4.2.2 Class Embeddings for Image Classification

A related line of works use non-visual data to improve image classifiers. Techniques have been proposed to use unannotated text (Frome et al., 2013), knowledge graphs (H. Xu et al., 2018) or hierarchical word databases (Akata et al., 2015) to obtain meaningful class embeddings, which proved beneficial for fine-grained image classification. Our work applies similar ideas to the task of VQA, where the key challenge is to find embeddings semantically connecting both visual and textual modalities.

4.3 Methodology

The main idea is to extend VQA with a regression objective, where the model outputs a high-dimensional vector that represents the semantics of the answer. This is a shift from the traditional classification objective over predefined candidate answers. Our formulation will open the door to compositional and unbounded sets of answers, and the possibility of truly open-set prediction. Technically, our method concerns only the latter stage of a VQA model and is thus applicable to most existing "joint embedding" models, such as (Antol et al., 2015; Saito et al., 2017; B. Zhou et al., 2015). In these models, the network produces a vector $\boldsymbol{x}$ from the fusion of the image and question representations (see Figure 4.2). The traditional approach then feeds this to a classifier and obtains $\boldsymbol{y} = f_\theta(\boldsymbol{x})$, with $\boldsymbol{y} \in \mathbb{R}^A$ being a vector of scores of length A, the cardinality of a predefined set of candidate answers.

4.3.1 VQA as a Regression Task

Our contribution is to learn a supplementary branch from $\boldsymbol{x}$, which produces a projection $\boldsymbol{p} = g_\psi(\boldsymbol{x})$, where ψ are the parameters of the projection. The vector $\boldsymbol{p} \in \mathbb{R}^P$ is interpreted as a representation of the semantics of the predicted answer. The key to this simple approach is both in the objective used to train this branch, and in its use to select an actual textual answer, which we both describe below.

Note that the traditional classifier over $\boldsymbol{x}$ can be interpreted as a special case of our formulation. The classifier $f_\theta(\cdot)$ typically includes a non-linear layer followed by a linear one. They can be interpreted as a non-linear projection followed by the computation of distances (dot products) with representations of answers. These representations then correspond to the rows of the weight matrix of the

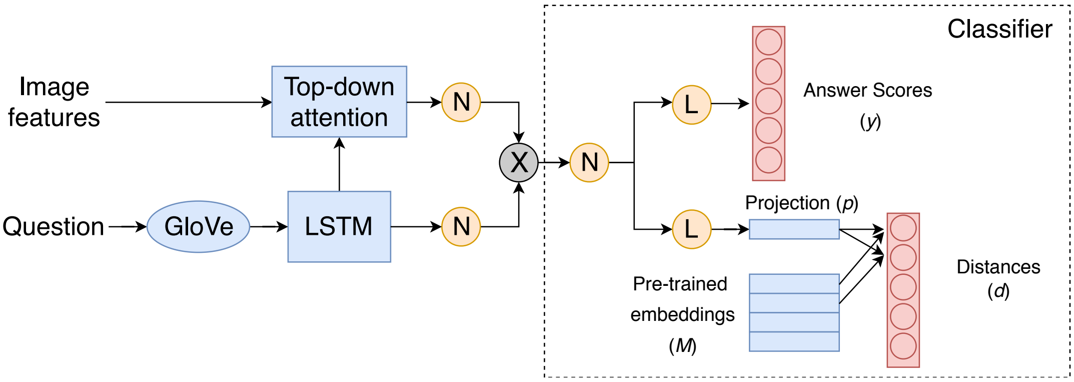


FIGURE 4.2. Our contributions apply to the classifier stage (dashed box) of a VQA model. We feed the fused image/question representation into two separate branches. (1) In the upper branch, a traditional scoring model over predefined candidate answers. (2) In the lower branch, a novel, learned projection to a semantic answer space. The resulting vector $\boldsymbol{p}$ serves to measure pairwise distances ($\boldsymbol{d}$) with pre-trained representations of candidate answers (M). Nodes marked N denote non-linear layers, L linear layers, and X an element-wise product.

linear layer. In this view, our model is a generalisation of the classical approach, with benefits of increased flexibility in the choice of the distance measure, of the optimisation loss, and of the representations of candidate answers including their initial and/or frozen values.

4.3.2 Training

To evaluate the possibility of mutual benefits of the classification and regression objectives, our full model includes both branches on top of the fused representation $\boldsymbol{x}$. Each of their respective outputs $\boldsymbol{y}$ and $\boldsymbol{p}$ is fed into a specific loss. The whole network is trained by backpropagation of the gradient of the two losses through all the layers leading to $\boldsymbol{x}$. The model is therefore trained to minimise the classification error and simultaneously learns the projection into the shared answer embedding space.

Classification Loss

The output of the classification branch $\boldsymbol{y}$ goes through a standard logistic function $\sigma(\cdot)$ and binary cross entropy loss L_c. Denoting with $\hat{\boldsymbol{a}} \in \{0,1\}^A$ the one-hot (multi-hot) vector of the ground truth answer(s) of a specific training instance, we have

$$L_c = \sum_{i=1}^{A} -[\, \hat{a}_i \cdot \log \sigma(y_i) + (1 - \hat{a}_i) \cdot \log(1 - \sigma(y_i))\,] \ , \tag{4.1}$$

where i indexes vector elements. The sum allows for multiple ground truth answers to a single training question.

Regression Loss

The output of the additional regression branch produces the vector $\boldsymbol{p} \in \mathbb{R}^P$. It is interpreted as a location in a high-dimensional space that captures the semantics of the predicted answer. We store in a matrix $M_{A \times P}$ representations of A candidate answers in this space (P-dimensional row vectors). These representations can be learned or initialised using prior knowledge, as described below. The objective of the regression branch is to produce a vector $\boldsymbol{p}$ close to the representation of the ground truth answer, and distinct from those of incorrect ones. Using a metric $\mathrm{dist}(\cdot, \cdot)$, we compute all distances between $\boldsymbol{p}$ and the rows of M, noted as M_i. We have

$$\boldsymbol{d} = [d_1, d_2, ..., d_A] \quad \text{with} \quad d_i = \mathrm{dist}(\boldsymbol{p}, M_i) \ . \tag{4.2}$$

We then define a hinge loss on these distances:

$$L_p = \sum_{i=1}^{A} l_i \quad \text{with} \quad l_i = \begin{cases} d_i & \text{if } \hat{a}_i = 1 \ , \\ \max\{0, \delta - d_i\} & \text{if } \hat{a}_i = 0 \ . \end{cases} \tag{4.3}$$

where δ is a scalar margin hyperparameter.

Total Loss

Our overall optimisation objective is the convex combination of the classification (Equation 4.1) and regression losses (Equation 4.3):

$$L \ = \ \lambda \, L_c \ + \ (1 - \lambda) \, L_p \ , \tag{4.4}$$

where the scalar hyperparameter λ balances the two objectives. By setting $\lambda = 1$, the loss falls back to a unique traditional classification objective, which serves as our baseline.

4.3.3 Predictions

Due to the nature of existing datasets, answer prediction during test time does not differ from the training, since both train and test splits typically share a common answer set. Our major experiments thus simply use the answers predicted by the network with the same combination of the classification and regression branches as the training objective. That is, the final predicted answer $a^\star \in [1...A]$ is the one from the set of candidates with the combination of the highest score and the lowest distance. Formally:

$$a^\star = \arg\max_i \big(\lambda \ \mathrm{softmax}(\boldsymbol{y}) \ + \ (1{-}\lambda) \ \mathrm{softmax}(-\boldsymbol{d})\big) \ . \tag{4.5}$$

To explore the full potential of the proposed task formulation, we conduct an additional set of experiments where train and test answer splits do not intersect. The experimental setting will be described in detail in Section 4.4.9.

4.3.4 Incorporating Prior Knowledge about Answers

The matrix M of the regression branch contains, in each of its rows, the representation of a candidate answer. M can be treated and optimised as any other parameter of the network, but it can also be initialised with values that contain prior knowledge about answers. In particular, we experiment with GloVe embeddings (Pennington et al., 2014) for single-word answers, and averaged (*i.e.* as a bag-of-words) in the case of multi-word ones. The values of M are further fine-tuned during training. In novel answers prediction setting (Section 4.4.9) we use ConceptNet embeddings (Speer et al., 2017) that are frozen during training.

As ablations of our model, we consider two other initialisation schemes of M. They will serve to probe for the source of the gains of our model.

- *Random.* We initialise M with normally distributed random values, as would be any other weight matrix of the network.
- *Shuffled GloVe.* We initialise M with GloVe embeddings as described above but subsequently shuffle its rows randomly, as in (Teney, Anderson, et al., 2017). The rows of M are thus mismatched from their corresponding answers. This allows us to disentangle the anticipated benefits of using the semantic information carried in GloVe vectors, from the mere numerical effects of using them as initial values.

4.4 Experiments

We performed an extensive evaluation to thoroughly validate the benefits of the proposed method, and understand the exact source of improvement. The overall conclusion is that the improvements indeed stem from the information brought in by the use of external data, rather than numerical artefacts or structural modifications to the network architecture.

4.4.1 Datasets

GQA dataset (Hudson and Manning, 2019a) is used for the evaluation of the approach as it provides the most comprehensive suite of metrics and cleanest data of current VQA datasets. We use the validation split for hyperparameter tuning and the test-dev split for model evaluation. The test set is used for comparison with other existing methods. We do not aim to build a data-specific solution, so our model does not utilise scene graphs and functional programs included in the dataset. We do, however, report the model's performance for new metrics proposed by the authors:

- *Validity* measures whether the predicted answer fits the scope of the question (*e.g.* a number for a counting question).
- *Plausibility* checks that the answer is semantically reasonable, defined as occurring at least once with the given question in the whole dataset.
- *Distribution* is the χ^2 distance between the distributions of predicted and ground-truth answers over groups of questions. A lower value means a better ability to predict less frequent answers.
- *Consistency* measures the agreement between answers to pairs of questions about the same image where one entails the other.
- *Grounding* is used for the evaluation of attention-based models and is not tested in our study since attention is not the focus of this research.

The dataset also assigns test questions to categories (Table 4.1), across which the accuracy can be measured separately (as done in Table 4.5).

VQA v2 dataset (Goyal et al., 2017) is used for additional set of experiments. To test the out-of-vocabulary answer prediction, we created a subset of VQA v2 that we call VQA v2 with novel answers. We used the original training and validation splits as our new training and test splits respectively. In each of them, we filtered the questions according to the following rules:

Type	Example
Choose	Is it an indoors or outdoors scene ?
Compare	Are all these animals of the same type ?
Logical	Are there nuts or vegetables ?
Query	What is this bird called ?
Verify	Is there a cat that is not white ?
Attribute	What is the colour of the fence made of metal ?
Category	What piece of furniture is not small ?
Global	Which place is it ?
Object	Is there a train in the picture ?
Relation	What is the vegetable on top of the pizza ?

TABLE 4.1. Examples of each question type of the GQA dataset.

- Every ground truth answer has a corresponding ConceptNet embedding (exact match).
- Every ground truth answer consists of one word only (*e.g.* discarding *black and white* or *don't know*).
- Every ground truth answer must occur in the original dataset between 5 and 500 times (thus discarding very rare and extremely frequent answers such as *yes* and *no*).
- The sets of ground-truth answers in the training and test splits do not intersect.

With this procedure, we obtain 91,255 training questions with 6,928 possible answers and 13,367 test questions with another 1,187 answers.

4.4.2 Experimental Setting

Our contributions are implemented on top of the open-source Pythia framework (Y. Jiang et al., 2018), the winning entry of the 2018 VQA Challenge[1]. The technique is however applicable to a wide range of current and future models. Pythia thus serves as the main baseline. We also evaluate the Pythia model where the weights of the output classifier are initialised with pre-trained answer embeddings (noted 'Pythia+GloVe'). We also compare our method to existing methods designed to inject prior knowledge into the model in the form of answer embeddings. Precisely, we consider the two variants of the factorised

[1] https://visualqa.org/roe_2018.html

Probabilistic Model of Compatibility (fPMC) proposed by H. Hu et al., (2018), using the code provided by the authors. All tested models use the same image features (those provided with the GQA dataset) and representations of question words (300-dimensional pre-trained GloVe embeddings).

Pythia. The baseline Pythia model is the implementation of the classical joint embedding architecture. It uses object image features extracted with the pre-trained Faster R-CNN (S. Ren et al., 2015) model provided with the GQA dataset. On the language side, words are represented with word embeddings initialised with pre-trained GloVe vectors, followed by an LSTM to produce a vector representation of the whole question. A question-guided top-down attention is applied on image features to identify relevant image regions. The image and question features are passed through non-linear layers and finally combined with an element-wise multiplication. The final classifier comprises a non-linear layer and a linear one, which produces a score for each candidate answer. All non-linear layers throughout the network use weight normalisation (Salimans and Kingma, 2016) and ReLU activations.

Pythia serves as a reference for evaluation, and as the base model on which to build our contributions. This choice is justified by a few reasons. It is a high-performing open-source implementation that still outperforms many others on the VQA v2 dataset. This provides us with a strong – and thus challenging – starting point to demonstrate the proposed method. Moreover, the implementation of Pythia is modular and easily allows one to separate, replace, and compare the various blocks of the model. In our case, this enables us to focus specifically on the classification part of the model, leaving the rest unchanged.

Pythia with pre-trained classifier. We compare our method to the Pythia model, in which the output classifier is initialised with pre-trained answer embeddings. As discussed in Section 4.2.1, this is a reasonable approach to embed semantic information about candidate answers within the model. Following a procedure similar to (Teney, Anderson, et al., 2017), we collect 300-dimensional GloVe embeddings for all words in the answer vocabulary (substituting unknown words with zero vectors). We represent each answer directly by its matching word embedding, or, in the case of multi-word answers, by the average embedding of the constituent words. Next, we design the classifier block of the model as follows: one non-linear layer with output dimension equal to the dimensionality of used GloVe embeddings followed by a linear layer with a weight matrix $w \in \mathbb{R}^{300\times A}$. Each row of w thus contains the vector corresponding to one specific answer. Besides the non-random initialisation of w, the only distinction with the original Pythia model is that the output dimension of the non-linear layer is reduced from 5000 to 300 to match the dimensionality of GloVe vectors.

Factorised Probabilistic Model of Compatibility. We also compare the proposed approach to fPMC. In this architecture, a joint image-question embedding is learned alongside the answer embedding, and the model is trained to increase the likelihood of the correct answer. We performed all experiments with the following two variants of the architecture:

- fPMC (SAN$\star$) model, described in the original paper, that utilises stacked attention network (Z. Yang et al., 2016) together with bidirectional LSTM and spatial image features extracted with ResNet-152 (K. He et al., 2016). For obtaining answer embeddings, the model exploits two-layer bidirectional LSTM over GloVe vectors. We used the code provided by the authors of the paper and made the adjustments only required to make it compatible with GQA dataset.
- fPMC (BUTD$\star$) model is our modification of fPMC (SAN$\star$) where we used the "bottom-up and top-down attention" (Anderson, He, et al., 2018) model with object image features for parameterising the joint embedding in the same way as all the other models used in our experiments. We were thus able to explicitly evaluate the approach of learning aligned answer embeddings independently from the impact of different feature initialisations.

4.4.3 Implementation Details

The proposed method builds directly on the open-source Pythia implementation[2], which uses PyTorch (Paszke et al., 2019). Our model is trained for 20,000 iterations with a batch size of 512 and AdaMax optimiser (Kingma and Ba, 2014). We adopted a warm-up learning schedule strategy from the original paper and tuned it to the current setup. Specifically, the starting learning rate of 0.002 is linearly growing up to 0.1 during the first 1000 iterations and then decreased by a factor of 0.1 at 11,000, 13,000 and 15,000 iterations. Importantly, these hyperparameters were selected for the best performance of the **baseline** model on the validation set of the GQA dataset, thus avoiding any unfair advantage for our contributions. The distance function $\text{dist}(\cdot, \cdot)$ (Equation 4.2) is implemented as the Euclidean distance. This choice proved empirically superior, on the GQA validation set, to a dot product or a cosine similarity. The

[2] https://github.com/facebookresearch/pythia/tree/0.1

values of the regression loss margin ($\delta = 1$ in Equation 4.3) and of the loss weight ($\lambda = 0.5$ in Equation 4.4) were determined by grid search for best overall accuracy on the GQA validation set. Every experiment was repeated with five different random seeds, and we report the average over the five runs. The ensembles use the average of the predicted scores/distances of several models trained with different random seeds, before taking the arg max of Equation 4.5.

For VQA v2 dataset the only difference in hyperparameters is the learning schedule. The model is trained for 12000 iterations with a learning rate decreasing at 5000, 7000, 9000 and 11,000 iterations, following the original Pythia implementation. We also found it beneficial to apply L2 normalisation after the projection layer. In the out-of-vocabulary experimental setting, we used a subset of VQA v2 data, so parameters were adjusted to fit the smaller dataset. Specifically, we reduced the batch size to 128 and increased the number of iterations to 30,000 with decreasing steps at 12,000, 17,000, 22,000, and 25,000 iterations.

4.4.4 Quantitative Results

Our main results on the GQA dataset are provided in Table 4.2 (see Appendix A for additional results). Looking at the overall accuracy, our model clearly outperforms all baselines and ablations. The same observations can be drawn on both the binary and open-ended questions. The trend is also confirmed when evaluating an ensemble of our model, versus a similar ensemble of the Pythia baseline. The fPMC model obtains the lowest results, including our modified version fPMC (BUTD$\star$), which indicates its lack of adaptivity to complex feature representation methods. The fPMC model was initially tested only on the very noisy VQA v2 dataset, and a possible reason for its weak performance on GQA is the narrower answer set. A surprising outcome is that Pythia with the pre-trained classifier ('Pythia+GloVe') shows worse accuracy results than the baseline. This occurs mostly due to the overfitting of the pre-initialised classifier to the most common answers in the training set, as observed by the reduced accuracy on both the validation and test-dev sets. Unlike the other described architectures, our model exploits the additional information contained in the representations of answers in an effective way, increasing performance without overfitting.

We present experiments on VQA v2 dataset in Table 4.3. Contrary to our results on GQA, we observe no significant difference compared to the baseline. We attribute this to the nature of the dataset. In VQA v2, a large fraction of the questions (over 37%) are to be answered with *yes* or *no*, and another 13% with a number. Our approach, which focuses on the representation of answer semantics, is already expected to have no influence on this large part of the dataset. Moreover, numbers in VQA v2 are used not only for counting questions, but also to refer to abstract concepts, as in questions like *How old is animal?*, *What time is the clock showing?*, or *What is the size of the TV?*. It would certainly be difficult to infer a single representation of numbers that would encompass such a variety of concepts.

An additional challenge with VQA v2 is that most questions have multiple ground-truth answers that are actually synonyms. Other times, annotation noise means that multiple answers with contradictory meanings are marked as correct. For example, a question *Is the dog male or female?* has both *male* and *female* answers in the annotation. In our model, all ground-truth answers contribute equally to the projection loss, meaning that noisy or incorrect answer labels can push the learned projection in wrong directions. This issue could be mitigated by introducing instance-specific weights in the projection loss. This is an interesting avenue for future work.

Overall, our approach still has a positive impact on VQA v2 for out-of-vocabulary prediction (see Section 4.4.9). And importantly, the above issues did not incur a decrease in performance compared to the baseline model.

4.4.5 Comparison with Existing Models

We compare our model with existing methods reported in (Hudson and Manning, 2019a) and several contemporaneous state-of-the-art models (see Table 4.2). We report the performance of the blind LSTM, the bottom-up top-down attention model (Anderson, He, et al., 2018), MAC (Hudson and Manning, 2018), LXMERT (Tan and Bansal, 2019) and Neural State Machine (NSM) (Hudson and Manning, 2019b). Our model shows better results than all the baselines, and in spite of a much simpler architecture, it notably surpasses the MAC model. However, the newest methods LXMERT and NSM show higher performance which is not surprising. LXMERT model explores a more sophisticated technique of image and language representation and is pre-trained on a significantly larger amount of data. NSM implements a compositional approach and performs explicit multi-step reasoning. Differently, our approach focuses on the output stage of the VQA model, thus the contributions of this work are expected to be applicable to these models.

	GQA validation			GQA test-dev			GQA test		
	Binary	Open	All	Binary	Open	All	Binary	Open	All
Blind LSTM	–	–	–	–	–	–	61.90	22.69	41.07
BUTD	–	–	–	–	–	–	66.64	34.83	49.74
MAC	–	–	–	–	–	–	71.23	38.91	54.06
LXMERT	–	–	–	–	–	–	77.80	45.00	60.30
NSM	–	–	–	–	–	–	**78.94**	**49.25**	**63.17**
Pythia	75.45	45.76	60.13	71.51	38.15	53.46	–	–	–
Pythia + GloVe	74.91	45.77	59.87	71.36	37.94	53.28	–	–	–
fPMC (BUTD⋆)	69.85	42.28	55.62	64.80	35.40	48.90	–	–	–
fPMC (SAN⋆)	71.94	41.78	56.37	67.02	35.83	50.14	–	–	–
Ours + random	75.15	46.33	60.27	70.67	38.14	53.08	–	–	–
Ours + shuffled GloVe	76.17	46.53	60.87	71.80	38.48	53.78	–	–	–
Ours + GloVe	**76.93**	**46.99**	**61.48**	**72.19**	**39.31**	**54.40**	71.35	40.07	54.73
Ensemble: 5× Pythia	77.24	48.41	62.36	73.43	39.85	55.26	–	–	–
Ensemble: 5× Ours + GloVe	**79.32**	**49.48**	**63.92**	**74.35**	**41.40**	**56.52**	–	–	–

TABLE 4.2. Overall accuracy (%) on GQA. Our method shows clear improvements on both binary and open-ended questions.

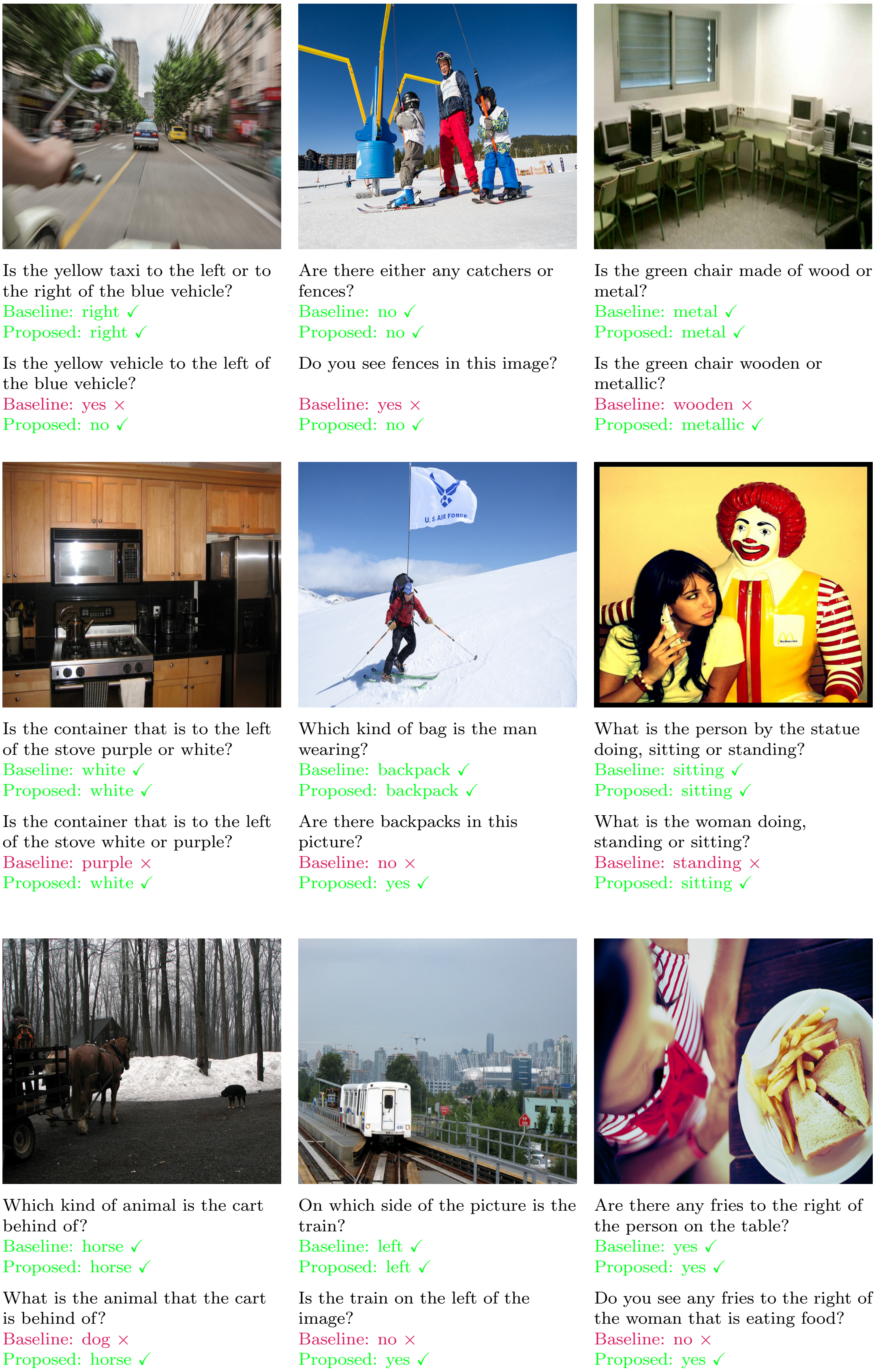
	VQA v2 validation				VQA v2 test-dev			
	Yes/No	Number	Other	All	Yes/No	Number	Other	All
Pythia	**83.11**	44.50	56.86	65.11	**83.42**	45.53	57.13	66.64
Ours + GloVe	82.90	**44.68**	56.93	65.08	83.33	45.46	57.18	66.62
Ours + GloVe (w/o fine-tuning)	82.87	44.55	**57.19**	**65.18**	83.20	**45.54**	**57.53**	**66.74**

TABLE 4.3. Overall accuracy (%) on VQA v2. Our model with fixed (not fine-tuned) GloVe embeddings shows the highest results on the category *other* and on all questions overall for both splits.

4.4.6 In-Depth Analysis

We report the detailed metrics of the GQA dataset in Table 4.4. The first observation is that a similar ranking of methods and ablations can be drawn from most of the metrics. This stability further confirms the benefits of the proposed method. The improvements on these advanced metrics also indicate benefits beyond the sole increase in accuracy. The *validity* and *plausibility* scores, in particular, which are noticeably higher, indicate a generally more robust model. The higher *consistency* score implies that the answers produced over related questions are compatible with one another (see Figure 4.3). The only metric on which our model falls below the baseline is the *answer distribution.* It indicates that the model occasionally favours one answer over most others. We explain this by the fact that some answers are not assigned appropriate initial representations. We also look at the accuracy per question category (Table 4.5). We observe no significant drop in accuracy for any type, and the highest improvements occur on the *choose*, *query*, *attribute*, and *relational* questions.

The ablations of our method ('Ours+random' and 'Ours+shuffled GloVe') are important to determine whether the source of improvements is in the architecture of our model (the additional output branch and loss), in numerical effects from the initialisation of the matrix M with values from GloVe vectors, or in the actual information conveyed in the GloVe vectors. The ablation with random initial values is essentially similar to the Pythia baseline, which shows no significant effect from the architecture alone. Surprisingly, the 'shuffled GloVe' ablation brings some improvement, which we explain by two factors. First, since the values of M are further fine-tuned with the rest of the model, they can still incorporate useful information from the task-specific supervision even if the initial values do not contain relevant semantic information. Second, some answers may actually benefit from the "wrong" initialisation: we have determined that the absolute values of the representations of answers do not play the most significant role, but that their mutual relations are what encodes the critical information. This shows up in particular with pairs of antonym answers such as *yes*/*no* or *left*/*right*. The GloVe embeddings of these pairs are usually similar, whereas the VQA task-specific supervision tends to push their representations apart. This can also be observed on the high accuracy of the 'shuffled' ablation on the *choose* category of questions which do specifically contain this type of antonym answers (see Table 4.1). Despite these effects, the full model still performs clearly better than the ablations, indicating an overall benefit from the information conveyed in the GloVe representations of answers.



Is the yellow taxi to the left or to the right of the blue vehicle?
Baseline: right ✓
Proposed: right ✓

Is the yellow vehicle to the left of the blue vehicle?
Baseline: yes ×
Proposed: no ✓

Are there either any catchers or fences?
Baseline: no ✓
Proposed: no ✓

Do you see fences in this image?
Baseline: yes ×
Proposed: no ✓

Is the green chair made of wood or metal?
Baseline: metal ✓
Proposed: metal ✓

Is the green chair wooden or metallic?
Baseline: wooden ×
Proposed: metallic ✓

Is the container that is to the left of the stove purple or white?
Baseline: white ✓
Proposed: white ✓

Is the container that is to the left of the stove white or purple?
Baseline: purple ×
Proposed: white ✓

Which kind of bag is the man wearing?
Baseline: backpack ✓
Proposed: backpack ✓

Are there backpacks in this picture?
Baseline: no ×
Proposed: yes ✓

What is the person by the statue doing, sitting or standing?
Baseline: sitting ✓
Proposed: sitting ✓

What is the woman doing, standing or sitting?
Baseline: standing ×
Proposed: sitting ✓

Which kind of animal is the cart behind of?
Baseline: horse ✓
Proposed: horse ✓

What is the animal that the cart is behind of?
Baseline: dog ×
Proposed: horse ✓

On which side of the picture is the train?
Baseline: left ✓
Proposed: left ✓

Is the train on the left of the image?
Baseline: no ×
Proposed: yes ✓

Are there any fries to the right of the person on the table?
Baseline: yes ✓
Proposed: yes ✓

Do you see any fries to the right of the woman that is eating food?
Baseline: no ×
Proposed: yes ✓

FIGURE 4.3. Qualitative examples from GQA dataset, with predictions of our model and of the Pythia baseline. We show pairs of questions about a same image where the first entails the second (this information is never provided to the model during training or testing). Our model improves in consistency over the baseline, producing pairs of answers compatible with one another.

	GQA validation			
	Validity	Plausibility	Distribution ↓	Consistency
Pythia	95.07	91.39	**3.93**	83.12
Pythia + GloVe	95.13	91.40	4.07	82.68
fPMC(BUTD⋆)	94.99	90.91	6.20	76.53
fPMC(SAN⋆)	95.11	**91.62**	5.66	78.67
Ours + random	95.07	91.53	4.26	83.00
Ours + shuffled GloVe	95.14	91.48	4.01	83.37
Ours + GloVe	**95.16**	91.55	4.01	**84.57**
Ensemble: 5× Pythia	95.17	91.90	4.78	85.27
Ensemble: 5× Ours + GloVe	**95.25**	**92.06**	**4.56**	**87.33**

TABLE 4.4. Results on additional GQA metrics. Our model noticeably improves in consistency over the baseline. It ranks slightly worse on the distribution metric (see discussion in text).

	GQA test-dev									
	Choose	Compare	Logical	Query	Verify	Attribute	Category	Global	Object	Relation
Pythia	67.93	62.14	**72.11**	38.15	75.28	60.04	45.59	51.85	84.53	44.24
Pythia + GloVe	68.45	**62.72**	71.69	37.94	74.81	59.40	44.72	54.14	84.78	44.51
fPMC (BUTD⋆)	60.39	57.83	63.36	35.40	69.99	51.17	42.90	51.97	81.95	43.04
fPMC (SAN⋆)	64.80	61.53	65.62	35.83	70.69	53.80	43.69	54.01	80.95	43.32
Ours + random	67.42	62.07	70.87	38.14	74.40	58.53	45.01	**55.42**	84.68	44.79
Ours + shuffled GloVe	70.88	62.14	71.38	38.48	75.14	59.65	45.36	54.14	84.60	45.33
Ours + GloVe	**71.12**	62.21	71.14	**39.31**	**76.17**	**60.28**	**46.04**	53.88	**85.01**	**46.01**
Ensemble: 5× Pythia	70.95	63.33	**73.54**	39.85	77.22	61.84	47.87	52.23	**86.50**	45.95
Ensemble: 5× Ours + GloVe	**75.11**	**64.18**	73.04	**41.40**	**77.66**	**63.00**	**48.30**	**53.50**	86.25	**47.70**

TABLE 4.5. Accuracy (%) over question types on GQA test-dev. Our contributions bring clear improvements on most question types, with the highest gain on the *choose* category.

4.4.7 Answer Recall

To obtain deeper insights into the additional knowledge that is actually most beneficial, we examined the improvements of our model over the Pythia baseline on individual answers. We report, in Figure 4.4, the change in answer recall for a random selection of answers. We define the answer recall as, for an answer candidate $\hat{a}$, the ratio of questions with $\hat{a}$ as ground truth that are correctly answered by the model. The recall of most answers improves, but it stays similar or even degrades on some others. We investigated the possible reasons. A degradation is presumably related to less relevant initial representations of the corresponding answer. To assess this, we examined the closest other answers in the space of pre-trained GloVe vectors. Most answers with a negative gain in answer recall have neighbours with no semantic or syntactic connections. For instance, the three closest neighbours to *modern* are $\{$*under*, *rooftop*, *visitor*$\}$. Answers with a high recall improvement, on the contrary, tend to have semantically related neighbours. For example, *basket* has the closest neighbours $\{$*baskets*, *cane*, *sack*$\}$. These observations further support the claim that mutual relations between representations of answers are the major way in which the network stores and uses semantic information.

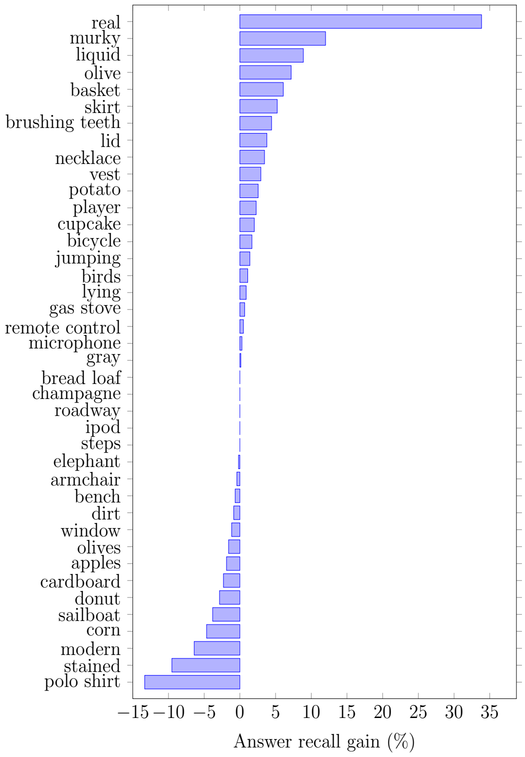


FIGURE 4.4. Absolute gain in answer recall of our model over the Pythia baseline (positive is an improvement). We report an even subset of answers (every 25th one in descending recall gain).

4.4.8 Combination of Losses

Since our architecture is trained to minimise a sum of two losses (classification and regression), we sought to evaluate their possible mutual benefit by varying their relative weight (λ in Equation 4.4). A value of λ=0 corresponds to the regression loss alone, and λ=1 to the baseline using the traditional classification loss alone. Interestingly, a balanced value of 0.5 leads to the highest accuracy (Figure 4.5), demonstrating that they are indeed complementary.

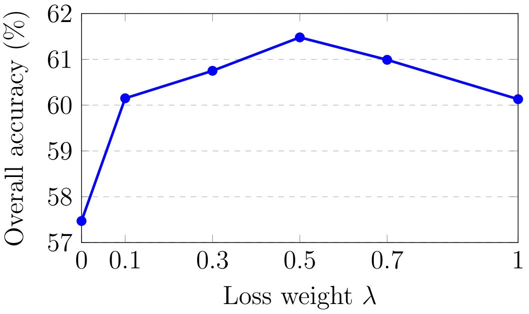
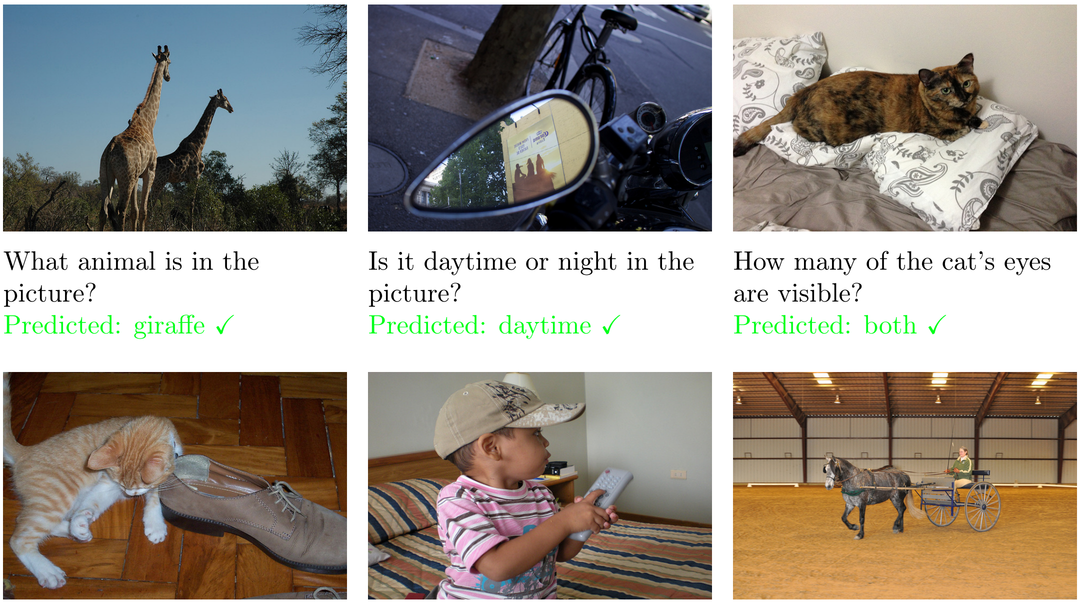


FIGURE 4.5. The performance of our model varies smoothly with the relative weight of the classification and regression losses (Equation 4.4). The value λ=1 corresponds to a traditional classification-only baseline, while the optimal value λ=0.5 corresponds to an even contribution of the two losses.

4.4.9 Prediction of Novel Answers

Our model trained with the regression objective can predict answers at test time that are outside the predefined set of candidates used for training (*i.e.* open-set prediction, or zero-shot VQA Teney and van den Hengel, 2016). This is achieved by replacing the matrix M with new answers and setting λ to 0 at test time. To evaluate this setting, we use ConceptNet embeddings (Speer et al., 2017), which are designed to capture commonsense knowledge.We use the VQA v2 dataset since it features a more diverse set of answers than GQA. We use splits with disjoint sets of answers at training and test time (as discussed in Section 4.4.1). In this setting, our model achieves an accuracy of 27% on the test set, while fPMC model, which also has tools for out-of-vocabulary prediction, obtains about 15% accuracy. Given that test questions feature exclusively answers never seen during training, this clearly demonstrates a capability for predictions beyond the scope of the training set. However, the performance on novel answers is highly dependent on the used answer representations. Embeddings like GloVe and ConceptNet carry only limited, mostly linguistic information, which is insufficient for the full scope of their use in VQA.



What animal is in the picture?
Predicted: giraffe ✓

Is it daytime or night in the picture?
Predicted: daytime ✓

How many of the cat's eyes are visible?
Predicted: both ✓

Where is the cat?
Predicted: floorboard ×
Ground truth: floor

Is this child male or female?
Predicted: boy ×
Ground truth: male

What colour is the ground?
Predicted: ochre ×
Ground truth: gold

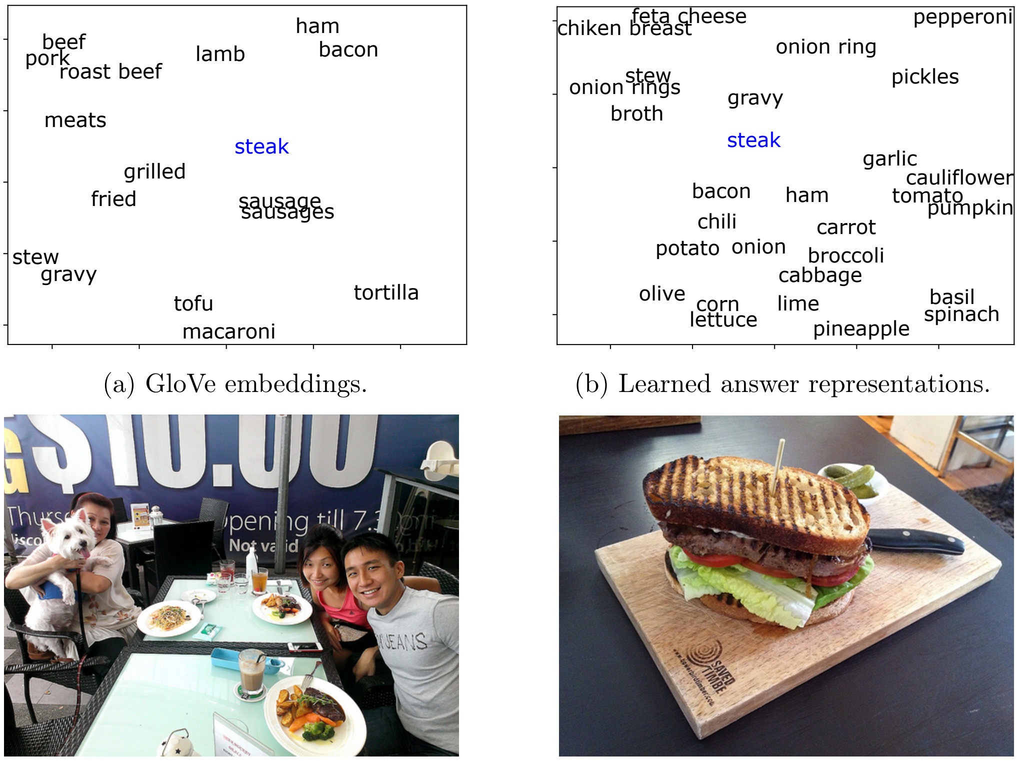
Figure 4.6. Examples of out-of-vocabulary predictions. The model works well when the ground truth answer has a clear and distinct pre-trained embedding (top row), but fails to distinguish between synonymous answers (bottom row).

We analysed the predicted answers in out-of-vocabulary test setting to discover the cause of reduced performance and possible ways for improvement. The reason for many failure cases is due to synonymous and/or related answers (Figure 4.6). When the representations of multiple candidate answers are close in the semantic space, it is difficult for the model to distinguish them, especially when they are both plausible for a given question.

Another important factor in the success of our method is how well the semantic space is covered by answers seen during training. For example, if the training questions all have similar answers, *e.g.* different animal species, the model could generalise well to novel animals, but not as well to anything outside these. In other words, the model is perfectly capable of interpolation, but extrapolation remains a challenge. The VQA v2 dataset was not originally designed to test the out-of-vocabulary prediction, and existing attempts to repurpose it all have notable issues. For our experiments, we created our own splits with novel answers, but we made no particular provision for even coverage of semantic concepts with the training answers. These considerations suggest the need for a specific benchmark to allow a more rigorous evaluation of models designed for out-of-vocabulary and zero-shot VQA.

4.4.10 Learned Representations

We use t-SNE (Van der Maaten and Hinton, 2008) projections to visualise and compare off-the-shelf GloVe embeddings of candidate answers, which we use as prior knowledge to initialise the representations, with these representations after fine-tuning within our VQA model (Figure 4.7). As expected, the GloVe embeddings carry the kind of semantic similarity that emerges from the co-occurrence of words in natural language. In the fine-tuned representations, we rather observe that the proximity of representations captures common co-occurrences of concepts in the same image, such that they are plausible answers to possible questions about this image. For example, the word *steak* is projected close to the words {*potato*, *carrot*, *broccoli*, *tomato*, *pickles*} (Figure 4.7b). We indeed observe co-occurrence of these objects in images from the GQA dataset (Figure 4.7c). This implies that additional knowledge extracted from visual data (*e.g.* as Noh et al., 2019) should be a useful complement to boost out-of-vocabulary performance.



(a) GloVe embeddings.

(b) Learned answer representations.

(c) Example images from the GQA dataset.

FIGURE 4.7. Examples of t-SNE projections in 2D of (a) initial and (b) fine-tuned representations of answers. The proximity of the learned representations better captures typical co-occurrences of the corresponding concepts in images from the dataset (c).

4.4.11 Choice of Answer Embeddings

Alternative Answer Embeddings

	GQA validation					
	Binary	Open	All	V	P	D ↓
GloVe	**76.93**	46.99	**61.48**	**95.16**	91.55	**4.01**
ConceptNet	76.13	46.76	60.97	**95.16**	91.52	**4.01**
Visual	74.66	45.85	59.79	95.05	91.32	5.15
GloVe + shuffled GloVe	76.64	**47.01**	61.34	**95.16**	91.52	4.08
GloVe + ConceptNet	76.18	46.07	60.64	**95.16**	**91.59**	4.06

TABLE 4.6. Accuracy, (V)alidity, (P)lausibility and (D)istribution results for other embeddings on GQA validation set.

Our primary choice for pre-trained answer embeddings is motivated by the popularity and widespread use of GloVe vectors in VQA models. We further investigate the applicability of other word embeddings for the regarded task. Specifically, we evaluate pre-extracted ConceptNet embeddings, visual representations (see description below) and combinations of different embeddings. The results for GQA validation set are given in Table 4.6 and Table 4.7.

ConceptNet is an open-source large-scale knowledge graph that contains commonsense and general knowledge about the world. We use pre-extracted Numberbatch embeddings [3] that were obtained from ConceptNet graph. These embeddings were used in the experiment with novel answers (Section 4.4.9) because they showed superior performance over GloVe. In this setting, answer embeddings are fixed during training and only the regression branch is used for inference. It means that original ConceptNet vectors build semantic space that is more suitable for VQA answers than GloVe. However, when answer embeddings are fine-tuned during training, GloVe vectors (61.48%) outperform ConceptNet (60.97%) in overall accuracy results, while performing on par in other metrics. A possible explanation for such behaviour could be the fact that GloVe vectors are used to initialise word embeddings used for question feature extraction. The model, therefore, can better learn connections between words in questions and answers. The results for *choose* category type, where GloVe achieves 75.36% accuracy and ConceptNet gets 73.15%, support this hypothesis.

[3] https://github.com/commonsense/conceptnet-numberbatch

	GQA validation									
	Choose	Compare	Logical	Query	Verify	Attribute	Category	Global	Object	Relation
GloVe	**75.36**	67.33	81.60	46.99	**76.58**	**66.21**	**57.23**	**65.58**	83.71	**52.94**
ConceptNet	73.15	67.50	**81.68**	46.76	75.94	65.36	57.16	65.57	83.64	52.47
Visual	74.98	65.25	78.08	45.85	73.88	64.38	55.09	63.59	80.61	51.79
GloVe + shuffled GloVe	75.06	67.05	81.63	**47.01**	76.09	65.97	56.87	65.44	**83.76**	52.87
GloVe + ConceptNet	72.96	**67.79**	81.56	46.07	76.19	64.11	56.82	65.28	**83.76**	52.64

Table 4.7. Accuracy over question types for other embeddings on GQA validation set.

	GQA validation									
	Choose	Compare	Logical	Query	Verify	Attribute	Category	Global	Object	Relation
Fixed during training										
GloVe	71.55	67.07	81.53	43.30	76.04	62.91	55.22	63.96	**83.65**	50.24
Shuffled GloVe	72.79	66.57	**81.56**	38.91	**76.90**	63.63	50.00	59.12	83.57	46.65
Trained with 0.1∗*learning rate*										
GloVe	**73.72**	**67.95**	81.52	**46.68**	75.88	**64.64**	**57.17**	**65.83**	83.64	**52.97**

Table 4.8. Accuracy over question types for fixed and slowly trained embeddings on GQA validation set.

Questions in *choose* category always contain the answers in their wordings as one of the options (*e.g. Is it red or blue?*), so the model can learn the link between two possible answers and the question. We tried ConceptNet embeddings as initialisation for both questions and answers, but the overall accuracy fell below the baseline, meaning that GloVe vectors provide better initial representations for the task.

The analysis of learned answer embeddings in Section 4.4.10 revealed that the model tends to learn "visual" representations from the co-occurrence of image concepts. This motivates us to experiment with embeddings that contain visually grounded information. To collect the data, we follow a process similar to the one described in (J. Kiros et al., 2018). Specifically, we use each answer as a search query to retrieve the top ten image results from Google image search. Then we use pre-trained Faster R-CNN model from (Anderson, He, et al., 2018) to extract image features and take the average for ten images as the target visual embedding. We further reduce the dimensionality of the embeddings (J. Kiros et al., 2018) to 300 to match GloVe embeddings. The results for our visual embeddings rank lowest in our experiments (59.79% overall accuracy). The accuracy for binary questions (74.66%) falls below the baseline, which is not surprising because it is hard to find good visual representations for such abstract answers as *yes* and *no*. The low performance of visual embeddings shows that linguistic information is still necessary to build valid answer representations. We also tried multiple variations of the method, including collecting images from Visual Genome (Krishna et al., 2017) instead of Google search, using ResNet (K. He et al., 2016) as a feature extractor, and keeping the original vector size, but all these modifications lead to lower performance.

Our model's architecture allows combining multiple answer embeddings by learning several distinct projections. As different answers may require different types of representations (*e.g.* binary answers benefit when their representations are projected far from each other, while in general, similar answers tend to cluster together in the embedding space), multiple projections can potentially solve the problem. Furthermore, the answers that do not have corresponding pre-trained embeddings of one type, may be present in other embeddings' vocabularies. To combine embeddings, during training we learn two projection branches and projection loss L_p (Equation 4.3) is the sum of individual losses of each branch. During inference we compute softmax over negative distances (Equation 4.5), similarly, for multiple embeddings, we take the average of all softmax results. In our experiments, we combine GloVe and ConceptNet embeddings to investigate whether the information they carry complements each other. We also experiment with combined GloVe and shuffled GloVe embeddings, as we found out that shuffled vectors still benefit some question types (see Section 4.4.6). However, the results show that the overall accuracy for all embeddings combinations is lower than for single GloVe, although it still surpasses baseline performance. 'GloVe + ConceptNet' vectors show a decline in performance for *choose* question category, similarly to single 'ConceptNet' embeddings. 'GloVe + shuffled GloVe' overall achieves comparable results to single GloVe. The information contained in different embeddings is either too similar which does not bring any improvements, or, on the opposite, contradicts each other which harms the overall performance. A proper way to combine embeddings, including weighting or gating mechanisms, is a promising direction for future research.

Fixed Embeddings

	GQA validation					
	Binary	Open	All	V	P	D ↓
Fixed during training						
GloVe	75.70	43.30	58.97	91.97	86.55	58.74
Shuffled GloVe	**76.36**	38.91	57.03	86.75	76.96	106.08
Trained with 0.1∗*learning rate*						
GloVe	76.23	**46.68**	**60.98**	**95.15**	**91.71**	**4.15**

TABLE 4.9. Accuracy, (V)alidity, (P)lausibility and (D)istribution results on GQA validation set for fixed and slowly trained embeddings. GloVe embeddings trained with reduced learning rate outperform fixed embeddings but score lower than normally trained ones.

When pre-trained answer embeddings are further learned as model's parameters, there is a risk that the knowledge encoded in the initial representations will be washed out during training. To examine how original pre-trained embeddings perform in our setting, we conduct two types of experiments: (1) we fix the answer embeddings and do not update them during training, (2) we update the embeddings, but the learning rate used for answer embeddings' parameter group is reduced by a factor of ten. The results for these experiments are shown in Table 4.9 and Table 4.8.

First of all, both fixed and slowly trained embeddings fall behind normally trained GloVe embeddings. That implies that the knowledge contained in GloVe vectors is not fully sufficient for the task, although it provides acceptable initial representations. Furthermore, fixed embeddings not only show the lowest overall accuracy but also lose in validity and plausibility metrics, which indicates the discrepancy between textual (used for GloVe pre-training) and visual (GQA dataset) contexts, where different answers are considered valid or plausible. Interestingly, shuffled GloVe embeddings achieve high accuracy on binary questions, especially for *logical* and *verify* question categories, in which all the questions can be answered with *yes* or *no*. These results further confirm our observation that "wrong" shuffled representations may help to distinguish between antonymous answers. To conclude, fine-tuned GloVe embeddings give the largest overall gain for the task and work best with the current VQA model's setting, thus, GloVe is our primary choice for main experiments.

4.5 Conclusion

In this work, we reformulated VQA as a multitask problem, which allowed us to exploit prior semantic knowledge about answers. We demonstrated that GloVe word embeddings carry information about typical answers that is relevant to the task. In contrast to existing methods for incorporating additional data into VQA models, our technique is both simple and effective, and allows to tune the reliance of the model on general prior knowledge, and learned task-specific information. We evaluated our technique on the GQA dataset and obtained consistent improvement in accuracy in the majority of question categories. The extensive set of metrics also allowed identifying benefits in robustness and consistency of the model across related questions.

The fundamental idea in this work of including a regression task as part of VQA has implications that go beyond what could be demonstrated with existing datasets. This formulation opens the door to the generation of compositional multi-word answers, and to open-set prediction, that is, predicting answers beyond the set of candidate answers predefined at training time.

Statement of Authorship

Title of Paper	Reasoning over Vision and Language: Exploring the Benefits of Supplemental Knowledge
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. "Reasoning over Vision and Language: Exploring the Benefits of Supplemental Knowledge." In Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), pp. 1-18. 2021.

Principal Author

Name of Principal Author (Candidate)	Violetta Shevchenko		
Contribution to the Paper	Design and implementation of experiments, paper writing.		
Overall percentage (%)	70 %		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	18/12/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Damien Teney		
Contribution to the Paper	Discussions, paper writing.		
Signature		Date	21 Dec. 2021

Name of Co-Author	Anthony Dick		
Contribution to the Paper	Discussions, paper revision.		
Signature		Date	22 Dec 2021

Name of Co-Author	Anton van den Hengel
Contribution to the Paper	Discussions, paper revision.
Signature	Date 21/12/21

Chapter 5

Reasoning over Vision and Language: Exploring the Benefits of Supplemental Knowledge

This chapter investigates the injection of supplementary knowledge from general-purpose knowledge bases (KBs) into vision-and-language transformers. The limits of applicability of vision-and-language models are defined by the coverage of their training data. Tasks like visual question answering (VQA) often require commonsense and factual information beyond what can be learned from task-specific datasets. In this work, we use an auxiliary training objective that encourages the learned word representations to align with graph embeddings of matching knowledge entities in a KB. We empirically study the relevance of various KBs to multiple tasks and benchmarks. The technique brings clear benefits to knowledge-demanding question answering tasks by capturing semantic and relational knowledge absent from existing models. More surprisingly, the technique also benefits visual reasoning tasks. We perform probing experiments and show that the injection of additional knowledge regularises the space of embeddings, which improves the representation of lexical and semantic similarities. The technique is model-agnostic and can expand the applicability of any vision-and-language transformer with minimal architectural modifications and computational overhead.

5.1 Introduction

The last few years have seen a surge of interest in vision and language (V&L) tasks. They require processing two modalities and reasoning over textual, visual, and abstract concepts. The current state of the art in V&L are models based on transformers such as BERT (Devlin et al., 2019) that have been extended to handle visual inputs (*e.g.* Tan and Bansal, 2019). These models are usually pre-trained on a collection of datasets of paired textual and visual data. Suitable datasets include image captioning data (X. Chen et al., 2015; Sharma et al., 2018) and visual question answering (VQA) data (Goyal et al., 2017; Hudson and Manning, 2019a). Despite their large scale, these datasets only cover limited domains. Many are based on images from COCO (T.-Y. Lin et al., 2014) and Visual Genome (Krishna et al., 2017) and the linguistic diversity of textual annotation is limited. The V&L tasks that we are ultimately interested in require knowledge beyond current datasets (*e.g.* about specific events, named entities, common sense, and abstract concepts).

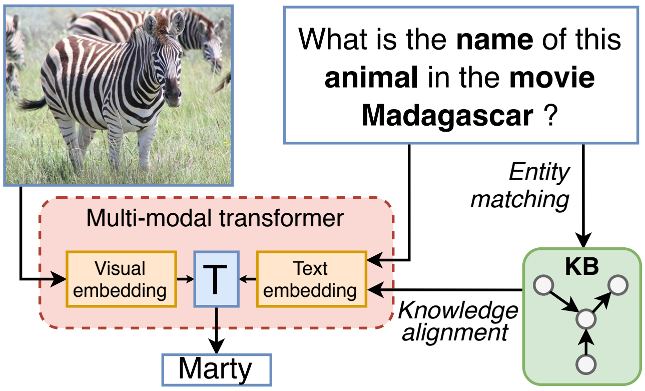


FIGURE 5.1. Additional information from knowledge bases is injected in a vision-and-language transformer. We first preprocess the knowledge base into a set of knowledge embeddings. Then during training, we use an auxiliary objective that aligns its learned word representations with corresponding knowledge embeddings.

This work focuses on the expansion of the applicability of V&L models with additional knowledge (Figure 5.1). During training, we infuse the model with knowledge from an external source, distinct from datasets of paired V&L data.The challenge is that standard V&L data is not annotated or paired with such additional knowledge. Even though techniques have been proposed to exploit additional data in natural language processing (NLP), including text-based question answering (S. Kafle et al., 2019; B. Y. Lin et al., 2019; Lv et al., 2020; Rajani et al., 2019), little work has been done on the extension to V&L. Works in NLP with benchmarks of knowledge-demanding questions (Clark et al., 2018; Mihaylov et al., 2018; Sap et al., 2019; Talmor et al., 2019; Y. Yang et al., 2015) have shown that knowledge bases (KBs) contain information that can benefit large-scale pre-trained models. Motivated by this line of evidence, we aim to evaluate similar mechanisms for V&L tasks.

In the context of VQA, datasets of knowledge-demanding questions (Marino et al., 2019; Shah et al., 2019; P. Wang et al., 2017b) have proved challenging for existing methods. For example, a question like *Is the man eating healthy food?* requires recognising a type of food and relating it to its nutritional quality. Learning this type of knowledge from VQA training examples would be clearly inefficient. Other examples of challenging questions involve references to named entities such as brands, locations, or movie titles. For example, the question *Levi's is a popular brand of what item shown here?* requires knowledge of the brand's specialisation before identifying the correct element in the image. We believe that embedding this type of knowledge in a model is a necessary step to enable progress on complex multi-modal question answering.

This work describes a technique to inject information from KBs into a transformer-based model during its training). We take inspiration from (Goodwin and Demner-Fushman, 2019) and adapt their regulariser from text-based models to V&L transformers. We provide an implementation of our method on top of the popular LXMERT model (Tan and Bansal, 2019) and investigate the suitability of several KBs to different tasks and benchmarks. Our contributions are summarised as follows.

- We describe a method to inject information from knowledge bases during the training of vision-and-language transformers.
- We implement the method on top of the popular LXMERT model with the ConceptNet (Speer et al., 2017) and Wikidata (Vrandečić and Krötzsch, 2014) knowledge bases.
- We perform an extensive empirical evaluation on four downstream tasks. We demonstrate clear improvements on knowledge-demanding VQA and visual reasoning datasets.
- We conduct an in-depth analysis, including ablations and probing experiments. They show that we improve the representation of lexical, semantic, and relational knowledge that is lacking in typical V&L models. This explains the surprising improvements on tasks that do not explicitly depend on external knowledge.

5.2 Related Work

5.2.1 Vision and Language Tasks

V&L tasks require joint processing of visual and textual data *e.g.* for image captioning (Anderson, He, et al., 2018; Hossain et al., 2019) and VQA (Teney, Wu, et al., 2017; Q. Wu, Teney, et al., 2017). They have historically been approached with task-specific models, but recent transformer-based models (Vaswani et al., 2017) were shown to be applicable to a variety of tasks. A transformer can thus be pre-trained on multiple datasets and then fine-tuned for one specific task (Alberti et al., 2019; Y.-C. Chen et al., 2020; G. Li et al., 2020; L. H. Li et al., 2019; Lu et al., 2019; Su et al., 2019; Sun et al., 2019; Tan and Bansal, 2019; L. Zhou et al., 2020). Through the multi-task pre-training, it benefits from a large amount of data from the other tasks and datasets. This work describes a method to embed additional information in a transformer-based model ("additional" to the pre-training and fine-tuning datasets). Our implementation builds on the popular LXMERT model (Tan and Bansal, 2019).

5.2.2 Additional Knowledge in NLP

Inclusion of external knowledge in NLP models can help with tasks requiring commonsense or factual information (Storks et al., 2019). Various techniques have been proposed to improve transformers such as BERT (Devlin et al., 2019). Z. Zhang et al., (2019) proposed ERNIE, which feeds graph embeddings of text entities to the model. Peters et al., (2019) proposed KnowBert, a similar technique suitable to multiple KBs. Levine et al., (2020) used WordNet (Miller, 1998) to aid in the masked-word prediction objective, and improve lexical understanding in downstream tasks. Ye et al., (2019) proposed a multiple-choice question answering pre-training task and improved performance on multiple datasets requiring commonsense reasoning. W. Liu, Zhou, et al., (2020) addressed noise issues by controlling the amount of domain-specific knowledge infused into the model. Goodwin and Demner-Fushman, (2019) proposed OSCAR, a regularisation method to inject ontological knowledge in a pre-trained language model. X. Wang et al., (2019) proposed to simultaneously learn knowledge representations while optimising a masked-language objective, rather than using pre-trained knowledge embeddings. All of these works were applied to NLP tasks. This paper studies the suitability of similar mechanisms to V&L tasks by applying the OSCAR technique (Goodwin and Demner-Fushman, 2019) to a multi-modal transformer.

5.2.3 Knowledge-based VQA

Knowledge-based VQA refers to benchmarks designed to require additional information (Marino et al., 2019; Shah et al., 2019; P. Wang et al., 2017a, 2017b). Models have been proposed that retrieve such information from KBs based on question and image contents (Narasimhan and Schwing, 2018; P. Wang et al., 2017a, 2017b; Q. Wu et al., 2016). Some works use other sources of external information at training (Teney and van den Hengel, 2016) or test time (Teney and van den Hengel, 2018, 2019) but they showed limited improvements on VQA benchmarks. The recently proposed ConceptBert (Gardères et al., 2020) model jointly learns visual, textual and knowledge embeddings to fuse commonsense information into VQA models. This work describes a method applicable to a variety of sources of information and to tasks beyond VQA. We also show that different tasks benefit from different types of information.

5.3 Methodology

We describe a general, simple yet effective technique to embed additional knowledge into a transformer-based model. It is compatible with existing multi-modal transformers and thus suitable for a variety of tasks. The method proceeds in three stages (Figure 5.2):

1. We preprocess the additional knowledge into a set of vector representations that we call knowledge embeddings. For example, with a relational knowledge base, we apply a graph embedding method to obtain a vector representation of every of its entities. Each is associated with a textual expression.

2. We match sentences in the V&L training data with knowledge embeddings. Matches in the training data are referred to as knowledge-rich expressions.

3. During the training of the transformer (pre-training and/or fine-tuning), we optimise an additional objective that aligns its learned representations of knowledge-rich expressions with the matching knowledge embeddings.

We now describe each stage in detail.

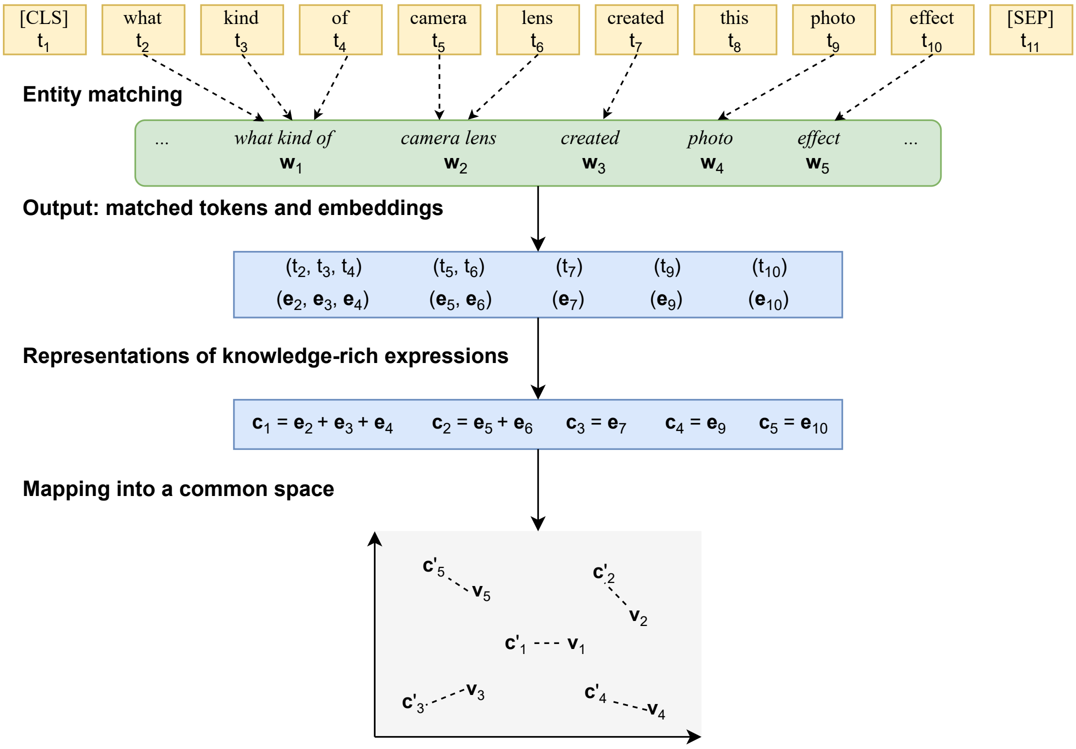


Figure 5.2. Summary of the approach. During training, we first match tokens in the V&L training data (yellow) with entities of the knowledge base (green). We then train the transformer with an additional loss to align the learned representations $\boldsymbol{c}_k$ (sums of word embeddings in knowledge-rich expressions from the transformer, in blue) with the knowledge embeddings $\boldsymbol{v}_k$ derived from the knowledge base.

5.3.1 Representations of Additional Knowledge

The versatility of the approach rests on storing the additional knowledge as a set of knowledge embeddings $\boldsymbol{V} = \{\boldsymbol{v}_i\}$ with $\boldsymbol{v}_i \in \mathbb{R}^{d_v}$. These can be produced by preprocessing sources such as text corpora (with word embedding methods, *e.g.* Mikolov et al., 2013; Pennington et al., 2014) or relational knowledge bases (with a graph embedding method, see Cai et al., 2018). These knowledge embeddings capture semantic information about the relations between entities, which we will incorporate into the V&L model. Each knowledge embedding $\boldsymbol{v}_i$ is associated with an entity $\boldsymbol{w}_i$ in the V&L data. In this work, the $\boldsymbol{w}_i$ are purely textual (single- and multiple-word expressions) but future work could consider visual representations of concepts represented by knowledge embeddings. We denote with $\boldsymbol{W} = \{\boldsymbol{w}_i\}$ the vocabulary of these instantiations.

5.3.2 Matching V&L Training Data with Knowledge Embeddings

We match parts of the training data with entities having additional knowledge. Since our vocabulary $\boldsymbol{W}$ contains textual expressions, we use greedy longest-string matching (Algorithm 1) to identify subsequences in the data that match any $\boldsymbol{w}_i$. Possible improvements left for future work include performing named entity recognition and homonym disambiguation. In practice, entities in our KBs have unique textual representations, making homonyms a non-issue.

We represent text in the training data as a sequence of tokens $T = (t_1, \dots, t_M)$. The transformer internally maps each token t_j to a word embedding $\boldsymbol{e}_j \in \mathbb{R}^{d_e}$. Each correspondence identified by the matching algorithm is of the form of a subsequence $(t_{a_k}, \dots, t_{b_k}) \subset T$ that matches $\boldsymbol{w}_k \in \boldsymbol{W}$. We refer to such a subsequence as a *knowledge-rich expression.* To obtain a fixed-size representation $\boldsymbol{c}_k$ of a knowledge-rich expression, we sum the word embeddings of its constituent tokens *i.e.* $\boldsymbol{c}_k = \boldsymbol{e}_{a_k} + \dots + \boldsymbol{e}_{b_k}$.

Algorithm 1 Entity matching

```
Input: Sentence tokens $(t_1, ..., t_n)$
    Vocabulary of entities $\boldsymbol{W} = \{\boldsymbol{w}_i\}$
Output: Knowledge-rich expression $(t_{a_k}, \dots, t_{b_k}) \subseteq (t_1, \dots, t_n)$
    Corresponding word embeddings $(\boldsymbol{e}_{a_k}, \dots, \boldsymbol{e}_{b_k})$
$i \leftarrow 1$
while $i \leq n$ do
    find the longest series of tokens starting at the $i^{th}$ position
    that match any $\boldsymbol{w}_k \in \boldsymbol{W}$
        $(t_i, ..., t_{i+p}) = \boldsymbol{w}_k$
    if match is found $(p \geq 0)$ then
        Return matched tokens and embeddings
            $a_k \leftarrow i$
            $b_k \leftarrow i + p$
            return $(t_{a_k}, \dots, t_{b_k}), (\boldsymbol{e}_{a_k}, \dots, \boldsymbol{e}_{b_k})$
        Skip tokens that are already matched
            $i \leftarrow i + p + 1$
    else
        $i \leftarrow i + 1$
    end if
end while
```

5.3.3 Aligning Learned Representations with Knowledge Embeddings

The core of the method is an additional training objective that encourages the transformer to produce representations of knowledge-rich expressions ($\boldsymbol{c}_k$) that collectively align with knowledge vectors ($\boldsymbol{v}_k$). A vector $\boldsymbol{c}_k$ as defined above is the representation of a knowledge-rich expression learned by the model. We do not expect these vectors to correspond to their matching knowledge embeddings $\boldsymbol{v}$, but we desire them to capture the information globally represented in the *relations* between the knowledge embeddings in $\boldsymbol{V}$. Therefore, we define a new linear layer that maps a learned representation $\boldsymbol{c}_k \in \mathbb{R}^{d_e}$ to $\boldsymbol{c}'_k \in \mathbb{R}^{d_v}$:

$$\boldsymbol{c}'_k = \boldsymbol{W}_c \boldsymbol{c}_k + \boldsymbol{b}_c \ , \tag{5.1}$$

where $\boldsymbol{W}_c \in \mathbb{R}^{d_e \times d_v}$ and $\boldsymbol{b}_c \in \mathbb{R}^{d_v}$ are learned weights and biases. We then define our alignment loss that encourages each projection $\boldsymbol{c}'_k$ to be close to its corresponding knowledge embedding $\boldsymbol{v}_k$:

$$L_{\text{align}} = \sum_k \|\boldsymbol{c}'_k - \boldsymbol{v}_k\|^2 \ . \tag{5.2}$$

Together, Equation 5.1 and 5.2 encourage the global structure of the learned representations $\boldsymbol{c}$ to align with the set of knowledge embeddings $\boldsymbol{V}$ through the projection $\boldsymbol{W}_c$. The learned representations incorporate information from the knowledge embeddings while allowing the transformer to also represent task-specific information. The model is trained for the combination of its original main loss, and the new alignment loss weighted by a hyperparameter λ:

$$L = L_{\text{main}} + \lambda \, L_{\text{align}} \ . \tag{5.3}$$

At test time, the model is used without any modifications.

5.3.4 Training Strategies

Typically, a transformer-based model is pre-trained on a collection of datasets and then fine-tuned for one specific task. We experimented with the application of the method during pre-training and/or fine-tuning, referred to as FT, PT, and PT+FT below. Enabling the additional objective during pre-training can benefit from the larger overlap between the training data and the additional knowledge. However, most pre-training tasks do not specifically require additional knowledge, and the model may not learn to effectively use it. During fine-tuning on knowledge-demanding tasks, the model is likely to better learn to capture and use relevant additional knowledge.

5.4 Experiments

We performed a suite of experiments with multiple tasks and benchmarks. Our objective is to evaluate the suitability of two popular knowledge bases (ConceptNet and Wikidata) to these tasks. The overall conclusion is that the tasks considered indeed benefit from the inclusion of additional knowledge.

5.4.1 Datasets

	Images	Questions	Answers	Task
OK-VQA	14,031	14,055	14,456	knowledge-based VQA
FVQA	2,190	5,826	954	knowledge-based VQA
SNLI-VE	31,783	565,286	3	entailment
NLVR2	141,480	107,292	2	reasoning

Table 5.1. Summary of datasets used in the experiments.

We evaluated the proposed method on four V&L datasets requiring knowledge-intensive and/or general visual reasoning capabilities (see Table 5.1 for statistics of datasets).

OK-VQA (Marino et al., 2019) is an open-ended VQA dataset where all questions require some sort of outside knowledge. It comprises about 14,000 questions about images from the MS COCO (T.-Y. Lin et al., 2014) dataset. All questions are produced by human annotators based either on information found in Wikipedia, or commonsense knowledge and visual evidence from the images. The questions are divided into ten categories (see Table 5.2) according to the type of knowledge needed to answer them. OK-VQA is one of the most diverse VQA datasets currently available that requires general knowledge. We use it accordingly as a primary benchmark in this study.

FVQA (P. Wang et al., 2017b) is a VQA dataset that contains about 5,000 questions that probe for commonsense knowledge. The questions are produced by annotators in a procedure that forces the question to involve facts found in a reference KB. Each question in the dataset is therefore associated with one specific "supporting" fact that describes a relation between concepts in the question and/or image. We do not use the annotations of these supporting facts. FVQA provides five different training/test splits. We report results that correspond to the average across the five splits. The quality of the questions in FVQA is mediocre in comparison to OK-VQA, and it is also much smaller. We

Alias	Category	Example
VT	Vehicles and Transportation	What is the title of the person driving this vehicle?
BCP	Brands, Companies and Products	Name the laptop model shown in this picture?
OMC	Objects, Material and Clothing	What sort of room is this woman sitting in?
SR	Sports and Recreation	What is this baseball player doing?
CF	Cooking and Food	Which of the foods here have the highest saturated fats?
GHLC	Geography, History, Language and Culture	What city is this meeting taking place?
PEL	People and Everyday Life	What kind of hairstyle does the woman in the black shirt have?
PA	Plants and Animals	What sound does this animal make?
ST	Science and Technology	Which rodent has a similar name to the technical device seen?
WC	Weather and Climate	What kind of clouds are shown?

Table 5.2. Categories of questions in the OK-VQA dataset corresponding to the type of knowledge required.

include it in this study because it previously served to evaluate other models designed to use KBs for VQA.

SNLI-VE (N. Xie et al., 2019) is a dataset for a visual entailment task. The task is an extension of the classical task of natural language inference. The visual version involves an image "premise" and a text "hypothesis". The model must determine whether the hypothesis contradicts the information shown in the image, entails it, or whether there are not enough clues to draw any conclusion. The SNLI-VE dataset contains about 560,000 instances, which were constructed from captions from SNLI (Bowman et al., 2015) paired with images from Flickr30k (Young et al., 2014). Despite similarities in the skills required for this task and for VQA, existing VQA models show relatively poor performance on SNLI-VE. The authors of the dataset attribute it to the need for more fine-grained visual understanding and reasoning, and they suggested the use of external knowledge to improve performance, hence its inclusion in this study.

NLVR2 (Suhr et al., 2019) is a dataset that evaluates visual reasoning over pairs of images. Each of the ∼107,000 instances in the dataset consists of two images with a statement in natural language. The model must predict whether the statement accurately describes the pair of images. The creation of the dataset emphasised the linguistic diversity of the sentences, with the objective for the task to require some compositional reasoning. We use this task in our study to evaluate the suitability of our model to perform compositional reasoning on the knowledge injected into them from KBs.

5.4.2 Knowledge Bases

In this work, we experimented with two popular knowledge graphs that provide well-organised and preprocessed information. To incorporate structural data into an end-to-end model we applied graph embedding techniques and obtained low-dimensional vector representations that we call knowledge embeddings (noted as $\boldsymbol{V}$ in Section 5.3.1). Knowledge graphs and embedding techniques used are described below.

ConceptNet (Speer et al., 2017) is a knowledge graph that encodes the meaning of expressions useful for general language understanding. This decades-old project is built on a number of crowd-sourced and curated sources including dictionaries, encyclopedias, and ontologies. We use the 300-dimensional Numberbatch embeddings distributed by the authors of ConceptNet. They are built using the technique of retrofitting (Faruqui et al., 2015) to combine relational information from the KB with distributional semantics from Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and OpenSubtitles (Tiedemann, 2012). We only use the subset of English expressions (∼500,000 entities).

Wikidata (Vrandečić and Krötzsch, 2014) is a collaboratively edited database of general knowledge. While ConceptNet mostly covers commonsense knowledge, Wikidata spans a larger domain, including historical events, celebrities, locations, science facts, *etc*. We obtain 200-dimensional embeddings with the PyTorch-BigGraph graph embedding method (Lerer et al., 2019). We use a subset of entities that have links to meaningful Wikipedia pages as done in (X. Wang et al., 2019). We also discard entities associated with stop words (*e.g. the*, *are*, *there*) which are common in VQA questions but carry no important semantic information. We retain ∼4.7M entities associated with ∼10M aliases.

5.4.3 Implementation Details

The implementation of our method builds on top of the official implementation of LXMERT[1], the state-of-the-art model on multiple tasks at the onset of this project. This model is pre-trained on five captioning and VQA datasets: COCO (T.-Y. Lin et al., 2014) and Visual Genome (Krishna et al., 2017) captions, VQA v2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019a) and Visual Genome QA (Y. Zhu et al., 2016). We also include experiments with scaled-down pre-training on two datasets (VQA v2 and GQA) which will ease the computational cost of replication.

[1] https://github.com/airsplay/lxmert

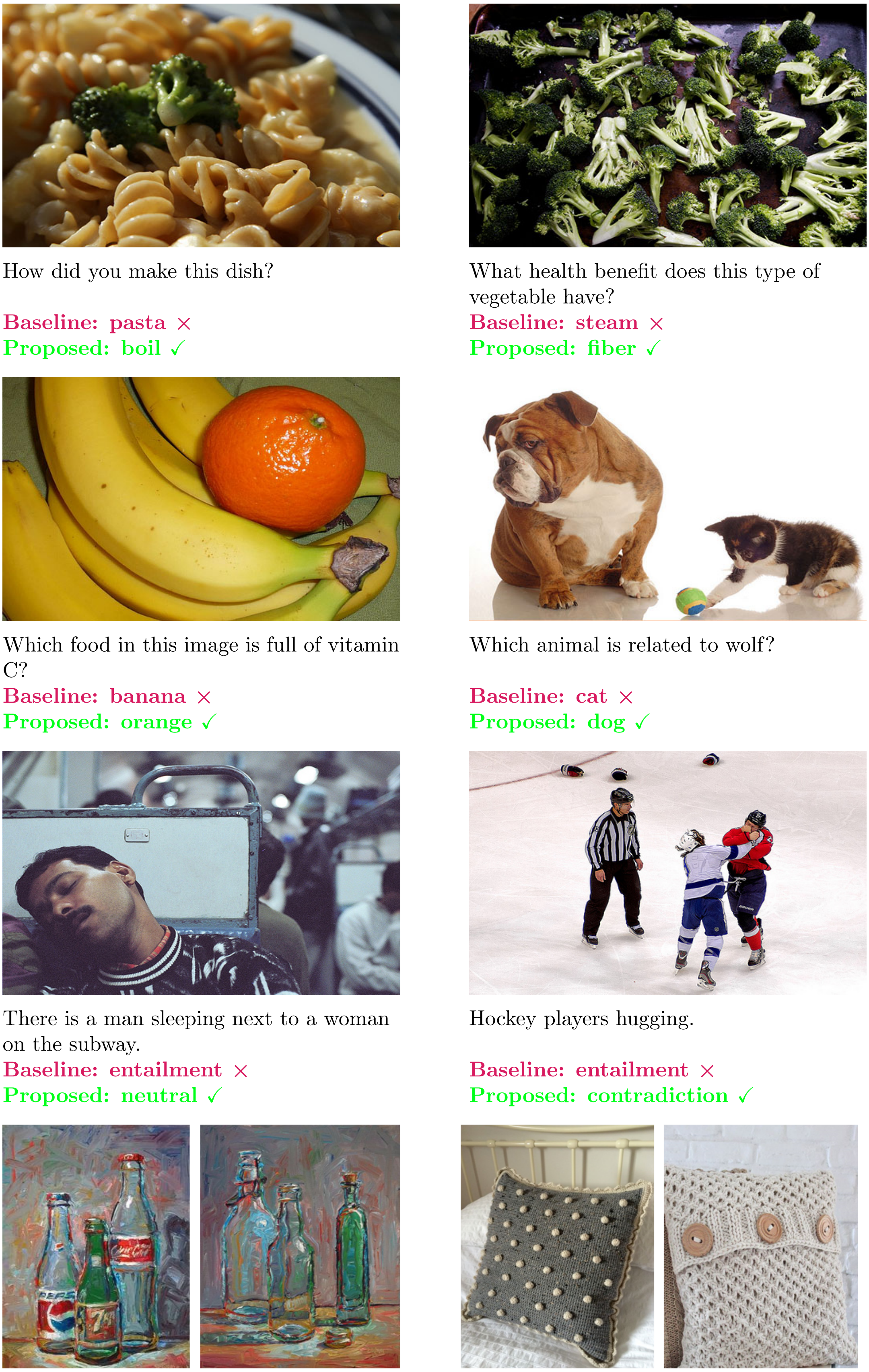
We use values of hyperparameters recommended in the code, including a reduced number of 12 training epochs (compared to 20 mentioned in the paper) and a single-stage training strategy. The loss weight λ (Equation 5.3) was selected by cross-validation for maximum accuracy on the OK-VQA validation set. From the set $\{0.1, 0.5, 1, 10, 100, 200\}$, the optimal values were found to be 0.5 for Wikidata and 100 for ConceptNet. The optimal λ seems to vary in inverse proportion with the size of the KB. The fine-tuning experiments used the same batch size of 32 and the learning rate of 0.00005 as the original LXMERT implementation. The number of fine-tuning epochs was adjusted according to the dataset size: 25 for OK-VQA and FVQA, 8 for NLVR2 and SNLI-VE.

Most transformer-based V&L models map the input text tokens to word, segment, and position embeddings, that are ultimately combined. Our approach is applied to word embeddings. Since the matching of KB entities with the V&L training data proceeds by exact string matching, we processed the KB entities with the same WordPiece tokeniser (Y. Wu et al., 2016) as the LXMERT model does for the V&L training data. To enable fast indexing of the KB, we store the knowledge embeddings $\boldsymbol{V}$ as a hash table indexed by textual expressions $\boldsymbol{W}$.

5.4.4 Quantitative Results

We report the overall accuracy on all datasets in Table 5.3. Our approach clearly outperforms the baseline, with a **higher accuracy** on the knowledge-demanding VQA datasets OK-VQA (+1.78%) and FVQA (+1.97%). We also get clear improvements on the visual reasoning datasets SNLI-VE (+1.19%) and NLVR2 (+1.28%). These do not explicitly require specific knowledge, so we hypothesised that the improvement is due to the richer linguistic representations learned by our model. We verified this hypothesis through probing experiments with the SentEval toolkit (Conneau and Kiela, 2018), which showed that our model better captures multiple semantic and syntactic properties of words (see Section 5.4.11). Example results for all four datasets are given in Figure 5.3 (more results can be found in Appendix B).

The best training strategy is generally to use the additional objective **during both pre-training and fine-tuning** (PT+FT). Only on OK-VQA did the PT strategy perform slightly better. Comparing PT alone with FT alone (the former being superior on all datasets) shows that fine-tuning the representations is not sufficient. Recall that the method relies on the structure of the embedding space to store the additional knowledge. Pre-training the model without



How did you make this dish?

Baseline: pasta ×
Proposed: boil ✓

What health benefit does this type of vegetable have?

Baseline: steam ×
Proposed: fiber ✓

Which food in this image is full of vitamin C?

Baseline: banana ×
Proposed: orange ✓

Which animal is related to wolf?

Baseline: cat ×
Proposed: dog ✓

There is a man sleeping next to a woman on the subway.

Baseline: entailment ×
Proposed: neutral ✓

Hockey players hugging.

Baseline: entailment ×
Proposed: contradiction ✓

All of the bottles in the right image are unlabeled.

Baseline: false ×
Proposed: true ✓

At least one of the pillows has a minimum of 4 different colors.

Baseline: true ×
Proposed: false ✓

FIGURE 5.3. Test cases on which our model (Pre-training with VQA v2, GQA w/ ConceptNet PT+FT) produces better predictions than the baseline.

	OK-VQA	FVQA	SNLI-VE	NLVR2
Pre-training with COCO captions, Visual Genome captions, VQA v2, GQA				
Baseline (LXMERT)	37.26 ± 0.23	52.30 ± 0.09	74.05 ± 0.19	71.31 ± 0.56
w/ ConceptNet PT	**39.04 ± 0.24**	54.08 ± 0.09	75.18 ± 0.21	71.61 ± 0.24
w/ ConceptNet FT	36.99 ± 0.01	52.19 ± 0.15	74.80 ± 0.08	70.82 ± 0.34
w/ ConceptNet PT+FT	38.56 ± 0.31	**54.27 ± 0.28**	**75.24 ± 0.22**	**72.59 ± 0.23**
Pre-training with VQA v2, GQA				
Baseline (LXMERT)	36.71 ± 0.22	50.38 ± 0.24	73.57 ± 0.17	67.88 ± 0.66
w/ ConceptNet PT	38.05 ± 0.41	51.94 ± 0.25	74.07 ± 0.48	69.47 ± 0.20
w/ ConceptNet FT	36.73 ± 0.50	50.54 ± 0.08	74.24 ± 0.15	67.09 ± 0.32
w/ ConceptNet PT+FT	38.12 ± 0.11	51.53 ± 0.17	74.26 ± 0.22	69.69 ± 0.12
w/ Wikidata PT	37.39 ± 0.35	51.00 ± 0.03	73.48 ± 0.17	68.27 ± 0.89
w/ Wikidata FT	36.31 ± 0.28	50.59 ± 0.24	73.60 ± 0.26	67.64 ± 0.19
w/ Wikidata PT+FT	37.43 ± 0.25	50.74 ± 0.29	73.50 ± 0.04	68.25 ± 0.51

Table 5.3. Overall results. Our model with ConceptNet during pretraining and fine-tuning. (ConceptNet PT+FT) generally proves best. We report the average accuracy (%) ± one standard deviation over three random seeds.

the additional objective may cause it to use its capacity in ways not flexible enough to accommodate the additional knowledge during fine-tuning. A second plausible explanation is that the small amount of fine-tuning data does not have enough coverage to capture a beneficial amount of additional knowledge. It is also interesting to note that the knowledge injection during pre-training is effective despite the pre-training tasks not specifically requiring additional knowledge.

5.4.5 ConceptNet vs Wikidata

We now examine the suitability of ConceptNet and Wikidata to the datasets considered. **ConceptNet provides larger improvements than Wikidata on every dataset**. Wikidata shows improvements on knowledge-driven tasks (OK-VQA and FVQA) but fails to improve over the baseline on the visual reasoning ones (SNLI-VE, NLVR2). Wikidata is almost ten times larger than ConceptNet, but it contains more redundant and noisy information due to its open-source nature. ConceptNet, in contrast, is based on a collection of mostly curated sources. Finally, the vector representations of ConceptNet used were obtained through an advanced and proven procedure that involves ConceptNet as well as other pre-trained word representations. The representations of Wikidata used are a more direct representation of the knowledge graph.

	OK-VQA	NLVR2
With less data (no pre-training)		
MLP	20.67	–
BAN	25.17	–
BAN+ArticleNet	25.61	–
MUTAN	26.41	–
MUTAN+ArticleNet	27.84	–
FiLM	–	52.1
ConceptBert (Conceptual Cap.)	33.66	–
VisualBERT (COCO captions)	–	67.00
UNITER (COCO, VG, Conceptual Cap., SBU)	–	**79.50**
LXMERT-Paper (cannot be reproduced, see text)	**42.94**	<u>74.50</u>
This paper (COCO and VG captions, VQA v2, GQA, VG QA)		
LXMERT-GitHub	37.26	71.31
LXMERT-GitHub w/ ConceptNet	<u>39.04</u>	72.59

Table 5.4. Comparison with existing methods. Datasets used for pre-training are given in parentheses.

5.4.6 Comparison with Existing Methods

In Table 5.4 we compare our results with the top entries from the leaderboards of OK-VQA[2] and NLVR2[3]. On OK-VQA, the best accuracy is shown by ConceptBert model pre-trained on a captioning dataset. The other reported results are from the traditional VQA models BAN (J.-H. Kim et al., 2018) and MUTAN (Ben-Younes et al., 2017). The BAN and MUTAN models supplemented with ArticleNet (Marino et al., 2019) obtain each a small improvement (+.44 and +1.43%). This component retrieves Wikipedia articles from which it extracts an answer for each question. These models perform much worse than the LXMERT baseline, which is trained on multiple datasets. We include an LXMERT model provided by its authors (LXMERT–Paper) and one retrained with code they provide (LXMERT–GitHub). The latter uses a simplified training strategy, hence a slight discrepancy (*e.g.* 69.50% on VQA v2 with their model and 68.52% with the retrained one). Our model brings a clear improvement over LXMERT-github, but it does not surpass LXMERT-paper that we could not reproduce.

[2] https://okvqa.allenai.org

[3] http://lil.nlp.cornell.edu/nlvr/

On NLVR2, the classical method FiLM (Perez et al., 2018) expectedly performs worse than transformers pre-trained on multiple datasets. The LXMERT baseline surpasses VisualBERT (L. H. Li et al., 2019), and our knowledge injection brings a small improvement. The state of the art on NLVR2 is obtained by UNITER (Y.-C. Chen et al., 2020) which is pre-trained on a much greater amount of captioning data and uses a significantly larger architecture.

5.4.7 Results on OK-VQA

We examine the accuracy on question categories of OK-VQA in Table 5.5. Each category corresponds to a type of knowledge required. We get **high gains on categories that correspond to a type of knowledge covered in ConceptNet** (objects properties and features, behaviour of people, *etc.*). These include OMC (objects, material and clothing), PEL (people and everyday life), BCP (brands, companies and products) and VT (vehicles and transportation). The only category with a drop in accuracy is ST (science and technology), which is also the smallest (84 questions). The category with the largest gain is GHLC (geography, history, language, and culture), but it contains only 141 questions, and some are distant from these topics (e.g. *What fruit come from these trees?*). These results should not be over interpreted because of the small size of these categories. Some questions also have imprecise labels. For example, the question *What activity are they doing?* is labelled with the correct answer *video game*, and our model's answer *play video game* is considered incorrect.

The **hardest questions for our model** are those referring to exact facts and entities such as place names, famous people, or historical dates. Such precise facts are more difficult to represent and recall than "soft" commonsense knowledge. For example, the question *What year was this picture taken?* requires to recognise a specific event and fetch a precise related fact. Additionally, these exact knowledge entities may lack from ConceptNet making the related questions unanswerable. Other difficult questions refer to precise visual cues, while the text of the question is generic, like *What language is on the sign?*, *Can you guess the place shown in this picture?* or *Which season is it?*. A recent analysis showed that V&L transformers rely primarily on the textual modality (Cao et al., 2020; Singh et al., 2020). Additional mechanisms would be needed to allow the recall of facts solely from visual cues and future improvements on the grounding across modalities could bring benefits here.

	OK-VQA										
	VT	BCP	OMC	SR	CF	GHLC	PEL	PA	ST	WC	Other
Pre-training with COCO captions, Visual Genome captions, VQA v2, GQA											
Baseline (LXMERT)	34.91	29.34	35.22	46.41	40.16	32.58	32.68	35.75	**35.79**	48.68	35.96
w/ ConceptNet PT	**37.40**	**31.94**	37.66	**47.52**	40.82	**38.39**	**35.06**	**37.20**	31.98	48.63	**38.42**
w/ ConceptNet FT	34.46	29.03	35.89	46.02	39.73	33.66	32.57	35.51	35.16	45.84	35.75
w/ ConceptNet PT+FT	37.11	31.47	**39.06**	46.52	**40.89**	36.50	34.08	36.16	33.73	**49.56**	37.05

Table 5.5. Accuracy (%) on the OK-VQA dataset per question category.

	SentLen	WC	TreeDepth	TopConst	BShift	Tense	SubjNum	ObjNum	SOMO	CoordInv
Baseline (LXMERT)	69.10 (6)	56.83 (1)	32.71 (7)	67.04 (4)	66.19 (9)	74.38 (6)	76.38 (8)	75.85 (7)	**51.01 (7)**	59.52 (6)
w/ ConceptNet PT	**70.36 (5)**	**66.19 (1)**	**33.35 (5)**	**71.49 (6)**	**68.23 (8)**	**82.10 (1)**	**80.26 (7)**	**79.02 (7)**	50.70 (5)	**60.24 (7)**

Table 5.6. Results on linguistic probing tasks. For each, we report the maximum score obtained across layers, with the corresponding layer number in parentheses. Our model outperforms the baseline in most probing tasks.

	NLVR2 Test-P			
	All	Consist.	Bal.	Unbal.
Baseline (LXMERT)	71.31	34.65	70.34	72.45
w/ ConceptNet PT	71.61	34.84	71.11	72.60
w/ ConceptNet FT	70.82	34.42	69.86	72.35
w/ ConceptNet PT+FT	**72.59**	**37.56**	**72.12**	**73.71**

Table 5.7. Detailed metrics on NLVR2: consistency (%) and accuracy (%) on balanced and unbalanced subsets. The method brings a clear improvement in consistency.

5.4.8 Results on NLVR2

We examine the extended metrics on NLVR2 in Table 5.7. *Consistency* reflects whether the model answers a given question correctly for all related pairs of images. Our model shows a **clear improvement in consistency** over the baseline (34.65 → 37.56). This suggests that the model can better relate a given textual input to different image contexts. We also report the accuracy on the balanced and unbalanced test sets, designed to evaluate a model's reliance on visual biases. In the balanced set, every image pair appears twice, one with each label (true/false). A drop in performance from the standard test set (*All*) to the balanced set (*Bal.*) would indicate that the method exploits biases. Neither the baseline nor our models show an undesirable reliance on image biases.

5.4.9 Knowledge Ablation

The main source of improvements on knowledge-demanding tasks like OK-VQA is the representation of knowledge relevant to test questions. To illustrate this claim, we create a small knowledge test with OK-VQA to examine how removing certain pieces of knowledge affects model performance. We select, by keyword search, a small subset of 19 test questions that focus on nutrition, on which our model obtains an accuracy of 91.11% *vs* 71.11% for the baseline. We then identify all entities related to nutrition (*e.g. health benefit*, *fiber*, *protein*, *vitamin*, *etc.*) and remove them from the knowledge base. After retraining our model with the pruned KB, the performance on nutrition questions drops markedly to 68.89%. The overall accuracy (on mostly non-nutrition-related questions) is maintained. This confirms that the withheld knowledge was indeed responsible for the high performance on related questions. The automated construction of diagnostic tests of this type could allow a quantitative evaluation and would be an interesting direction for future work.

5.4.10 Nearest Neighbours in Embedding Space

To better understand how the regularisation with additional knowledge changes the structure of the learned embedding space, we examine the nearest neighbours of embeddings of individual words (see Table 5.8). The neighbours are computed using the L2 distance. Using the cosine distance gives qualitatively similar results. Our model clearly captures more lexical and semantic information. For example, the nearest neighbours for *vitamin* are *calcium* and *supplements* with our model, but they are *margarita* and *amphibious* with the baseline. This stark qualitative difference was observed across the board and is not confined to cherry-picked examples. The lack of linguistic information encoded by baseline V&L transformers is unsurprising since their training data has limited linguistic diversity. Initialising a V&L model with a pre-trained BERT has been proposed to address this deficiency, but it was reported to give lower downstream performance by the authors of LXMERT. In comparison, our method brings linguistic information while improving downstream performance.

Word	Nearest neighbors with the baseline	Nearest neighbors with our model
argentina	questioned, [PAD], neutron, dorset	argentine, uruguay, paraguay, mendoza
behaviour	[PAD], absurd, authoritative, mba	behavior, behaviors, demeanor, behavioral
bowling	boxing, smashed, dancing, 75	bowler, bowled, cricket, tennis
cottage	farmhouse, scan, condo, ##tta	cottages, bungalow, farmhouse, ##ode
facebook	jade, brady, institution, utrecht	myspace, twitter, youtube, dit
genes	[PAD], oro, subsistence, ##vah	gene, genetic, genetics, genome
lecturer	greenberg, [unused983], avoidance, ##mour	professor, prof, professors, lectures
playstation	splendid, indo, financial, tapping	xbox, wii, sega, consoles
birth	##gm, dat, sensitive, incorporated	births, childbirth, born, newborn
boxers	dare, indo, briefs, aiding	boxer, briefs, underwear, panties
creeping	goalscorer, fertility, ineffective, [PAD]	crept, creep, crawling, sneaking
dependent	bacterial, [PAD], ##idad, outlaws	depended, depend, depends, dependency
displaced	neptune, roche, peterborough, norway	displacement, refugees, relocated, atletico
down	up, MASK, ##combe, ##ending	up, on, out, ##s
equity	[PAD], implementations, eurasian, newfound	investors, investments, investor, investment
ghosts	[PAD], germain, combustion, ##ignment	ghost, ghostly, haunted, phantom
indication	neptune, musee, converting, legion	indications, indicating, signaled, indicative
limb	stump, branch, limbs, thorn	limbs, branch, leg, ##wara
policemen	cowboys, firefighters, youths, 37	policeman, police, cops, constabulary
quebec	sutton, [PAD], monasteries, frederick	montreal, laval, ontario, sudbury
smells	[PAD], aiding, preston, quentin	smelled, smell, odor, scent
successes	[PAD], kilometres, tina, marne	success, achievements, accomplishments, successful
sugar	yeast, powder, memo, coating	chocolate, butter, candy, celaena
taste	feel, smell, fade, become	tastes, flavor, tasted, flavors
unmarried	[PAD], [unused285], [unused685], ##dium	divorced, childless, widowed, marrying
vintage	retro, antique, victorian, rustic	antique, retro, old, ##60

TABLE 5.8. Selection of nearest neighbours in the space of word embeddings learned by the baseline and by our model. Here, [PAD] is a special token and "##" indicates sub-word tokens.

5.4.11 Linguistic Probing Analysis

The probing tasks aim to identify the linguistic information encoded in the learned representations. Following (Cao et al., 2020), we tested our model on the following ten linguistic probing tasks (Conneau et al., 2018):

- **SentLen:** predict the length of sentences;
- **WC:** word content task requires to predict which words appear in the sentence;
- **TreeDepth:** categorize the sentence according to the depth of its syntax tree;
- **TopConst:** predict the sequence of top constituents;
- **BShift:** check whether two random adjusted words have been inverted in the sentence;
- **Tense:** predict the tense of the main-clause verb;
- **SubjNum:** predict if the main-clause subject is in singular or plural form;
- **ObjNum:** same as SubjNum but for the main-clause object;
- **SOMO:** semantic odd man count task is to determine if a random noun or verb has been replaced.
- **CoordInv**: check whether two coordinate clauses have been inverted.

These tasks are designed to evaluate the quality of sentence embeddings, but LXMERT model learns separate embeddings for each token. To obtain sentence representations, we thus take outputs of 9 intermediate layers in the language encoder and average each of them across all tokens. Every layer output is used as a separate embedding and we report the best result across all layers (Table 5.6). Since the first 9 layers perform attention over textual modality only, no image input is required.

The results show that our model surpasses the baseline in most of the tasks. The highest gain (+9.36%) is seen in WC category meaning that our model better captures the content of words in a sentence. Importantly, the best results for this task are obtained with the first layer outputs. WC accuracy drops for every subsequent layer output till it reaches 16.11% and 18.65% for the baseline and our model respectively. It implies that information about individual words is well encoded in early layers but gets washed off with further self-attention. Surprisingly, we observe a noticeable gain in accuracy for predicting tense and plurality of words (+7.72% in Tense, +3.88% in SubjNum and +3.17% in ObjNum tasks). Slight improvements in BShift and CoordInv tasks may be caused by additional constraints imposed on the word order by our model where the order of tokens that constitute a knowledge-rich expression is important. Finally, SentLen and TreeDepth probes highly rely on the length of sentences and since both tested models limit textual input to 20 tokens, the results may not be reliable.

5.4.12 Discussion and Limitations

Our experiments showed that KBs can expand the domain of applicability of V&L models. The tested approach proved robust in a range of settings with ConceptNet but the larger Wikidata did not fulfil our expectations. Our results suggest that the noise and ambiguities in Wikidata prevent realising its full potential. Methods for better text-to-knowledge matching such as named entity recognition and homonym disambiguation are promising solutions to investigate.

We also identified the grounding of knowledge with visual evidence to be a limitation currently for certain tasks. Questions in OK-VQA about specific places or famous people for example require the model to recall specific facts on the basis of precise visual cues. Although such facts are stored in Wikidata, the tested model did not prove effective at recalling them. Improvements of the visual grounding could also help with visual reasoning tasks like SNLI-VE where the correct interpretation of the text input is heavily dependent on the visual input.

5.4.13 Additional Results

We explored a variety of architectural choices, but some of them did not gain any improvement over the baseline. To measure the distance between learned projections and their embeddings in Equation 5.2 we tried two other options: smooth L1 loss and cosine distance, but they showed worse results than the used mean squared error. To obtain representations for knowledge-rich expressions we first included an additional knowledge embedding layer as the fourth component of textual representations along with word, segment and position embeddings. This technique proved to be inefficient, so eventually, we used word embeddings as a target for knowledge alignment.

Further, we tried to leverage information about objects detected in the image. We used predicted labels for each image object as a knowledge-rich expression and added another learning objective to align visual embeddings with corresponding knowledge. With this objective, the model converged to a smaller loss during pre-training but the accuracy on all fine-tuning tasks was lower than the baseline results indicating possible overfitting. A possible way to better exploit visual information could be to use object labels as additional supervision to enhance entity matching.

5.5 Conclusion

In this work, we described a general-purpose technique to inject external information from knowledge bases into multi-modal transformers for vision-and-language tasks. The current prevailing paradigm is to pre-train large models on collections of datasets. Our experiments demonstrate that some types of commonsense and factual knowledge are not captured within these models. Knowledge bases like ConceptNet and Wikidata can fill in these deficiencies. We showed clear improvements in performance on a variety of tasks and benchmarks requiring visual and multi-modal reasoning, demonstrating the versatility of the procedure.

The value of these results for future research is twofold. On the one hand, they indicate that the combination of heterogeneous sources of information is a promising way to expand the applicability of current machine learning models. On the other hand, by improving the availability of supporting knowledge, the approach opens the door to future advances in reasoning procedures that process this information. Advances on this front would lead to improved capabilities on tasks that require high-level or multi-hop reasoning, for example.

Chapter 6

Conclusion

We have witnessed enormous changes in the visual question answering field over the last few years. What was originally designed as a straight-forward image query task, has grown into a massive benchmark for multi-modal understanding and reasoning evaluation. While the early attempts to solve VQA mostly focused on image modality trying to extract the most relevant information for image understanding, recent studies have shifted their focus towards language inputs as well, in an attempt to perform true multi-modal reasoning. That indicates that VQA is no longer a task that just brings together advances made in computer vision and natural language processing fields, but a separate vision-and-language research area.

In this thesis, we investigated ways of using external information that can not be learned from traditional VQA training datasets but brings real benefit when solving the task. In particular, we explored unsupervised image pre-training that can be applied when VQA annotated data is scarce. Our experiments showed that contrastive training can learn adequate image features that needs little annotated data to fine-tune on a downstream task. We believe that unsupervised pre-training is crucial for small VQA tasks when transfer learning from public datasets is not available, for example, due to image domain mismatch. Further, we proposed to leverage prior knowledge about answers. We showed that information hidden in the semantics of answers can improve the general accuracy and consistency of a VQA model. Moreover, the use of answer embeddings in an additional regression branch opens up the possibility for open-set prediction, which is lacking in the majority of existing models. Lastly, we described a universal technique to inject external knowledge into multi-modal transformers. This method allows the model to capture commonsense and factual knowledge that typically can not be extracted from the training data alone. We found that our method not only boosts the performance on knowledge-demanding and general visual reasoning tasks but also improves the representation of lexical and semantic similarities of textual features. We believe that ability to incorporate additional knowledge in VQA models is vital for expanding their applicability in real-world conditions.

Although we aimed at exploiting varied knowledge sources in our works, most of the information we used in fact came from the language domain. Existing knowledge bases and word embeddings are trained on text corpora, so the information they encode is not grounded in vision. Nonetheless, there are types of commonsense knowledge that can only be extracted from visual modality, as we discussed in Sections 4.4.10 and 5.4.12. A natural future research direction is therefore to collect and exploit visually grounded knowledge that complements existing language knowledge sources. For example, by extracting objects from images and understanding their attributes and relationships, one can build an ontology of visual concepts. The similarity of concepts, in turn, can be decided based on their visual appearance or co-occurrence in images.

Another critical problem we faced during our studies is that current VQA datasets and evaluation metrics are designed for the classification approach where an answer is picked from a closed set of candidates. However, in real-world applications, it is not sensible to expect that a single pre-defined answer set will cover all possible scenarios. Several attempts have been made to tackle the problem of rare or zero-shot answer prediction but with the existing evaluation paradigm, the merits of these methods can not be adequately recognised. The future VQA research should focus on establishing novel evaluation procedures that take account of true open-ended VQA possibility. These evaluation metrics, for example, can measure whether the predicted answer is equivalent, synonymous or contradictive to the ground truth answer and assign a score accordingly. Furthermore, to allow free-form answer generation, additional metrics that measure the quality and similarity of text must be incorporated into the evaluation pipeline. These methods can be adopted from the related natural language understanding field, however, they will require further adjustments to include language-to-image grounding as a part of the text quality evaluation.

Appendix A

Additional Quantitative Results

In the main chapters, we use an average accuracy over multiple runs to report experimental results. This section provides additional tables with minimum and maximum accuracy reported. Table A.1 extends Table 3.3 in Chapter 3, Table A.2 extends Table 4.2 in Chapter 4, and Table A.3 extends Table 5.3 in Chapter 5.

	Test					
	Total		Cube		Sphere	
	min	max	min	max	min	max
ResNet	**98.97**	99.36	98.11	98.72	**99.83**	**100.00**
Baseline	1.78	20.36	0.00	0.11	3.44	40.72
EBM	77.11	91.67	94.33	98.11	59.89	86.06
SimCLR	98.92	**99.44**	**98.56**	**99.06**	99.28	99.83

Table A.1. Minimum and maximum accuracy results of the methods explored in Chapter 3.

	GQA validation					
	Binary		Open		All	
	min	max	min	max	min	max
Pythia	74.87	76.43	44.74	46.63	59.32	61.05
Pythia + GloVe	74.37	76.15	45.35	46.84	59.49	61.02
fPMC (BUTD⋆)	69.57	70.09	41.90	42.62	55.29	55.86
fPMC (SAN⋆)	71.69	72.16	41.51	42.13	56.27	56.48
Ours + random	74.07	76.11	45.26	47.18	59.26	61.02
Ours + shuffled GloVe	75.92	76.51	46.00	47.21	60.49	61.27
Ours + GloVe	**75.99**	**77.46**	**46.09**	**47.67**	**60.55**	**62.09**

Table A.2. Minimum and maximum accuracy results of the methods explored in Chapter 4

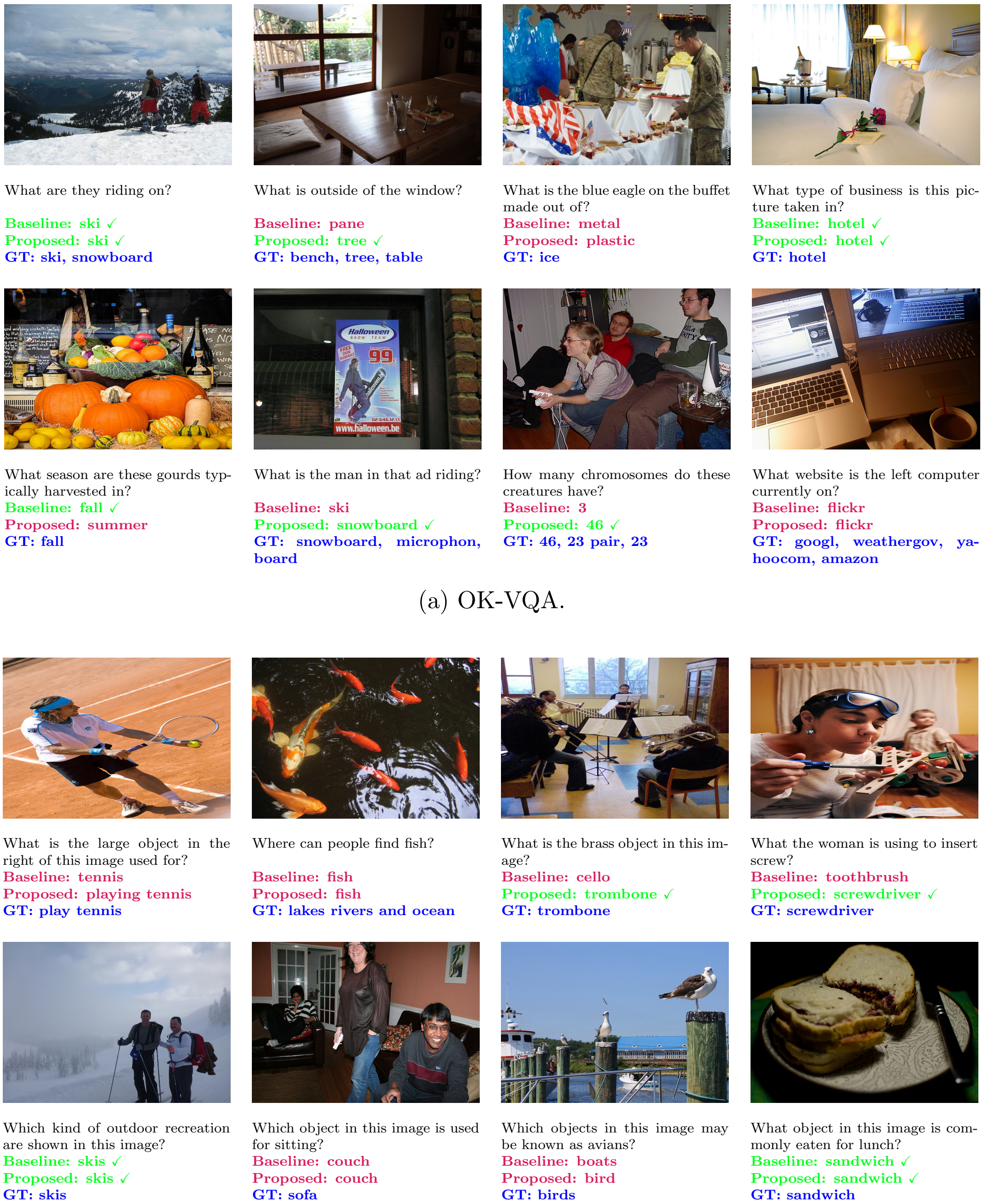
	OK-VQA		FVQA		SNLI-VE		NLVR2	
	min	max	min	max	min	max	min	max
Pre-training with VQA v2, GQA								
Baseline (LXMERT)	36.46	36.86	50.15	50.63	73.41	73.74	67.13	68.39
w/ ConceptNet PT	37.68	**38.49**	**51.79**	**52.22**	73.60	**74.55**	69.27	69.66
w/ ConceptNet FT	36.40	37.30	50.45	50.59	**74.09**	74.39	66.79	67.42
w/ ConceptNet PT+FT	**38.02**	38.24	51.34	51.68	74.01	74.43	**69.57**	**69.80**

Table A.3. Minimum and maximum accuracy results of the methods explored in Chapter 5

Appendix B

Additional Qualitative Results

In this section, we provide additional qualitative results of the knowledge injection method described in Chapter 5. Figure B.1 and Figure B.2 compare test predictions of the baseline (LXMERT) and of our model (pre-trained with VQA v2, GQA w/ ConceptNet PT+FT). See discussion in Section 5.4.4.



What are they riding on?
Baseline: ski ✓
Proposed: ski ✓
GT: ski, snowboard

What is outside of the window?
Baseline: pane
Proposed: tree ✓
GT: bench, tree, table

What is the blue eagle on the buffet made out of?
Baseline: metal
Proposed: plastic
GT: ice

What type of business is this picture taken in?
Baseline: hotel ✓
Proposed: hotel ✓
GT: hotel

What season are these gourds typically harvested in?
Baseline: fall ✓
Proposed: summer
GT: fall

What is the man in that ad riding?
Baseline: ski
Proposed: snowboard ✓
GT: snowboard, microphon, board

How many chromosomes do these creatures have?
Baseline: 3
Proposed: 46 ✓
GT: 46, 23 pair, 23

What website is the left computer currently on?
Baseline: flickr
Proposed: flickr
GT: googl, weathergov, yahoocom, amazon

(a) OK-VQA.

What is the large object in the right of this image used for?
Baseline: tennis
Proposed: playing tennis
GT: play tennis

Where can people find fish?
Baseline: fish
Proposed: fish
GT: lakes rivers and ocean

What is the brass object in this image?
Baseline: cello
Proposed: trombone ✓
GT: trombone

What the woman is using to insert screw?
Baseline: toothbrush
Proposed: screwdriver ✓
GT: screwdriver

Which kind of outdoor recreation are shown in this image?
Baseline: skis ✓
Proposed: skis ✓
GT: skis

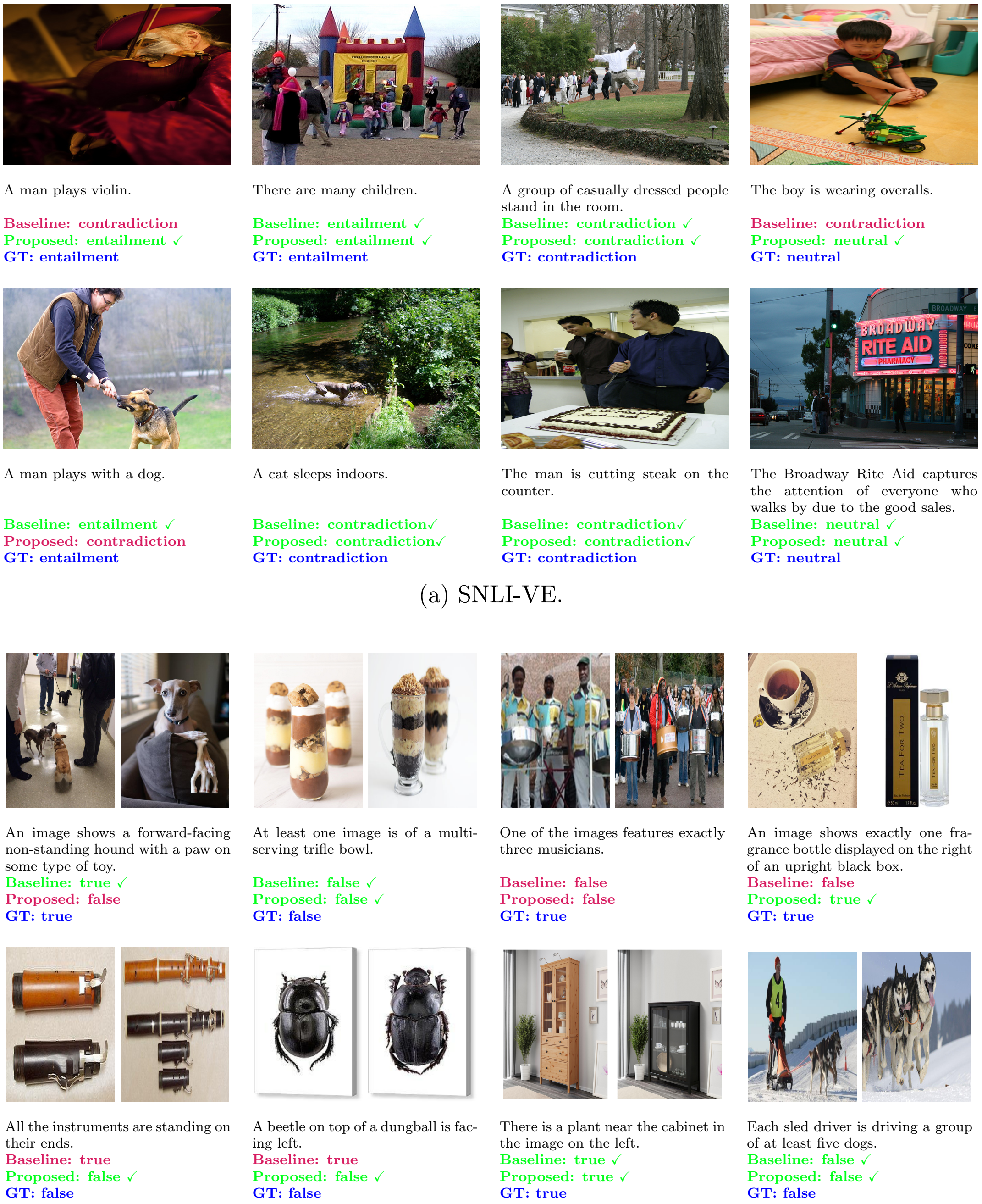
Which object in this image is used for sitting?
Baseline: couch
Proposed: couch
GT: sofa

Which objects in this image may be known as avians?
Baseline: boats
Proposed: bird
GT: birds

What object in this image is commonly eaten for lunch?
Baseline: sandwich ✓
Proposed: sandwich ✓
GT: sandwich

(b) FVQA.

FIGURE B.1. Random selection of test instances from OK-VQA and FVQA datasets, with predictions of the baseline (LXMERT) and of our model (pre-trained with VQA v2, GQA w/ ConceptNet PT+FT).



A man plays violin.
Baseline: contradiction
Proposed: entailment ✓
GT: entailment

There are many children.
Baseline: entailment ✓
Proposed: entailment ✓
GT: entailment

A group of casually dressed people stand in the room.
Baseline: contradiction ✓
Proposed: contradiction ✓
GT: contradiction

The boy is wearing overalls.
Baseline: contradiction
Proposed: neutral ✓
GT: neutral

A man plays with a dog.
Baseline: entailment ✓
Proposed: contradiction
GT: entailment

A cat sleeps indoors.
Baseline: contradiction✓
Proposed: contradiction✓
GT: contradiction

The man is cutting steak on the counter.
Baseline: contradiction✓
Proposed: contradiction✓
GT: contradiction

The Broadway Rite Aid captures the attention of everyone who walks by due to the good sales.
Baseline: neutral ✓
Proposed: neutral ✓
GT: neutral

(a) SNLI-VE.

An image shows a forward-facing non-standing hound with a paw on some type of toy.
Baseline: true ✓
Proposed: false
GT: true

At least one image is of a multi-serving trifle bowl.
Baseline: false ✓
Proposed: false ✓
GT: false

One of the images features exactly three musicians.
Baseline: false
Proposed: false
GT: true

An image shows exactly one fragrance bottle displayed on the right of an upright black box.
Baseline: false
Proposed: true ✓
GT: true

All the instruments are standing on their ends.
Baseline: true
Proposed: false ✓
GT: false

A beetle on top of a dungball is facing left.
Baseline: true
Proposed: false ✓
GT: false

There is a plant near the cabinet in the image on the left.
Baseline: true ✓
Proposed: true ✓
GT: true

Each sled driver is driving a group of at least five dogs.
Baseline: false ✓
Proposed: false ✓
GT: false

(b) NLVR2.

Figure B.2. Random selection of test instances from SNLI-VE and NLVR2 datasets, with predictions of the baseline (LXMERT) and of our model (pre-trained with VQA v2, GQA w/ ConceptNet PT+FT).

Bibliography

Abbasiantaeb, Z., & Momtazi, S. (2021). Text-based question answering from information retrieval and deep neural network perspectives: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1412.

Agrawal, A., Batra, D., Parikh, D., & Kembhavi, A. (2018). Don't just assume; look and answer: Overcoming priors for visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4971–4980.

Akata, Z., Reed, S., Walter, D., Lee, H., & Schiele, B. (2015). Evaluation of output embeddings for fine-grained image classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2927–2936.

Alberti, C., Ling, J., Collins, M., & Reitter, D. (2019). Fusion of detected objects in text for visual question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, 2131–2140.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, *3*, 6.

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., & Van Den Hengel, A. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3674–3683.

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016a). Learning to compose neural networks for question answering. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1545–1554.

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016b). Neural module networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 39–48.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2425–2433.

Arbel, M., Zhou, L., & Gretton, A. (2020). Generalized energy based models. *International Conference on Learning Representations*.

Atzmon, Y., Kreuk, F., Shalit, U., & Chechik, G. (2020). A causal view of compositional zero-shot recognition. *arXiv preprint arXiv:2006.14610*.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The semantic web* (pp. 722–735). Springer.

Ba, J., Mnih, V., & Kavukcuoglu, K. (2014). Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.

Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems*, *32*, 15535–15545.

Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures.

Barra, S., Bisogni, C., De Marsico, M., & Ricciardi, S. (2021). Visual question answering: Which investigated applications? *arXiv preprint arXiv:2103.02937*.

Be my eyes - see the world together. (n.d.). Retrieved November 1, 2021, from https://www.bemyeyes.com/

Ben-Younes, H., Cadene, R., Cord, M., & Thome, N. (2017). Mutan: Multimodal tucker fusion for visual question answering. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2612–2620.

Bhattacharya, N., Li, Q., & Gurari, D. (2019). Why does a visual question have different answers? *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4271–4280.

Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., et al. (2010). Vizwiz: Nearly real-time answers to visual questions. *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 333–342.

Biten, A. F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., & Karatzas, D. (2019). Scene text visual question answering. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4291–4301.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 1247–1250.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Bugliarello, E., Cotterell, R., Okazaki, N., & Elliott, D. (2021). Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts. *Transactions of the Association for Computational Linguistics*, *9*, 978–994.

Cai, H., Zheng, V. W., & Chang, K. C.-C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, *30*(9), 1616–1637.

Cao, J., Gan, Z., Cheng, Y., Yu, L., Chen, Y.-C., & Liu, J. (2020). Behind the scene: Revealing the secrets of pre-trained vision-and-language models. *Proceedings of the European Conference on Computer Vision*, 565–580.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS).*

Chen, K., Wang, J., Chen, L.-C., Gao, H., Xu, W., & Nevatia, R. (2015). Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960.*

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning*, 1597–1607.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325.*

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). Uniter: Universal image-text representation learning. *Proceedings of the European Conference on Computer Vision.*

Chiu, T.-Y., Zhao, Y., & Gurari, D. (2020). Assessing image quality issues for real-world problems. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3646–3656.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cho, S., Park, J.-M., Song, T.-J., & Kim, J.-H. (2020). Human-robot full-sentence vqa interaction system with highway memory network. *Rita 2018* (pp. 131–148). Springer.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). Think you have solved question answering? Try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Community, B. O. (2018). *Blender - a 3d modelling and rendering package*. Blender Foundation. Stichting Blender Foundation, Amsterdam. http://www.blender.org

Conneau, A., & Kiela, D. (2018). Senteval: An evaluation toolkit for universal sentence representations. *Proceedings of the International Conference on Language Resources and Evaluation*.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2126–2136.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., & Batra, D. (2017). Visual dialog. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 326–335.

De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. C. (2017). GuessWhat?! Visual object discovery through multi-modal dialogue. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5503–5512.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, *29*(6), 141–142.

Deng, Y., Bakhtin, A., Ott, M., Szlam, A., & Ranzato, M. (2019). Residual energy-based models for text generation. *International Conference on Learning Representations*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

Du, Y., Li, S., & Mordatch, I. (2020). Compositional visual generation with energy based models. *Advances in Neural Information Processing Systems*, *33*, 6637–6647.

Du, Y., Li, S., Tenenbaum, J., & Mordatch, I. (2020). Improved contrastive divergence training of energy based models. *arXiv preprint arXiv:2012.01316*.

Du, Y., & Mordatch, I. (2019). Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, *32*, 3608–3618.

Dumoulin, V., Shlens, J., & Kudlur, M. (2016). A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*.

Elflein, S., Charpentier, B., Zügner, D., & Günnemann, S. (2021). On out-of-distribution detection with energy-based models. *arXiv preprint arXiv:2107.08785*.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, *88*(2), 303–338.

Fader, A., Zettlemoyer, L., & Etzioni, O. (2014). Open question answering over curated and extracted knowledge bases. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1156–1165.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. *Proceedings of the European Conference on Computer Vision*, 15–29.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1606–1615.

Frolov, S., Hinz, T., Raue, F., Hees, J., & Dengel, A. (2021). Adversarial text-to-image synthesis: A review. *arXiv preprint arXiv:2101.09983*.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, 2121–2129.

Fu, B., Qiu, Y., Tang, C., Li, Y., Yu, H., & Sun, J. (2020). A survey on complex question answering over knowledge base: Recent advances and challenges. *arXiv preprint arXiv:2007.13069*.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., & Xu, W. (2015). Are you talking to a machine? Dataset and methods for multilingual image question. *Advances in Neural Information Processing Systems*, 2296–2304.

Gardères, F., Ziaeefard, M., Abeloos, B., & Lecue, F. (2020). Conceptbert: Concept-aware representation for visual question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 489–498
00002.

Geman, D., Geman, S., Hallonquist, N., & Younes, L. (2015). Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, *112*(12), 3618–3623.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Goodwin, T. R., & Demner-Fushman, D. (2019). Bridging the knowledge gap: Enhancing question answering with world and domain knowledge. *arXiv preprint arXiv:1910.07429*.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6904–6913.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., & Swersky, K. (2019). Your classifier is secretly an energy based model and you should treat it like one. *International Conference on Learning Representations*.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning*, 1321–1330.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., & Bigham, J. P. (2018). Vizwiz grand challenge: Answering visual questions from blind people. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3608–3617.

Gustafsson, F., Danelljan, M., Timofte, R., & Schön, T. B. (2020). How to train your energy-based model for regression. *Proceedings of the British Machine Vision Conference.*

Gustafsson, F. K., Danelljan, M., Bhat, G., & Schön, T. B. (2020). Energy-based models for deep probabilistic regression. *Proceedings of the European Conference on Computer Vision*, 325–343.

Han, T., Nijkamp, E., Fang, X., Hill, M., Zhu, S.-C., & Wu, Y. N. (2019). Divergence triangle for joint training of generator model, energy-based model, and inferential model. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8670–8679.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2961–2969.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, T., McCann, B., Xiong, C., & Hosseini-Asl, E. (2021). Joint energy-based model training for better calibrated natural language understanding models. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 1754–1761.

He, X., Zhang, Y., Mou, L., Xing, E., & Xie, P. (2020). Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286.*

Henaff, O. (2020). Data-efficient image recognition with contrastive predictive coding. *International Conference on Machine Learning*, 4182–4192.

Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136.*

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, *30.*

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, *14*(8), 1771–1800.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, *51*(6), 1–36.

Hu, H., Chao, W.-L., & Sha, F. (2018). Learning answer embeddings for visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5428–5436.

Hu, R., Andreas, J., Rohrbach, M., Darrell, T., & Saenko, K. (2017). Learning to reason: End-to-end module networks for visual question answering. *CoRR, abs/1704.05526*, *3*.

Hu, R., Singh, A., Darrell, T., & Rohrbach, M. (2020). Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9992–10002.

Huang, P., Huang, J., Guo, Y., Qiao, M., & Zhu, Y. (2019). Multi-grained attention with object-level grounding for visual question answering. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 3595–3600.

Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., & Fu, J. (2021). Seeing out of the box: End-to-end pre-training for vision-language representation learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12976–12985.

Huang, Z., Zeng, Z., Liu, B., Fu, D., & Fu, J. (2020). Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Hudson, D. A., & Manning, C. D. (2018). Compositional attention networks for machine reasoning. *International Conference on Learning Representations*.

Hudson, D. A., & Manning, C. D. (2019a). Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6700–6709.

Hudson, D. A., & Manning, C. D. (2019b). Learning by abstraction: The neural state machine. *Advances in Neural Information Processing Systems*, *32*, 5903–5916.

Ilievski, I., Yan, S., & Feng, J. (2016). A focused dynamic attention model for visual question answering. *arXiv preprint arXiv:1604.01485*.

Jabri, A., Joulin, A., & van der Maaten, L. (2016). Revisiting visual question answering baselines. *Proceedings of the European Conference on Computer Vision*, 727–739.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*.

Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E., & Chen, X. (2020). In defense of grid features for visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10267–10276.

Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., & Parikh, D. (2018). Pythia v0.1: The winning entry to the VQA challenge 2018. *arXiv preprint arXiv:1807.09956*.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2017). Inferring and executing programs for visual reasoning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2989–2998.

Kafle, K., & Kanan, C. (2016). Answer-type prediction for visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4976–4984.

Kafle, K., & Kanan, C. (2017a). An analysis of visual question answering algorithms. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1965–1973.

Kafle, K., & Kanan, C. (2017b). Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, *163*, 3–20.

Kafle, K., Shrestha, R., & Kanan, C. (2019). Challenges and prospects in vision and language research. *Frontiers in Artificial Intelligence*, *2*, 28.

Kafle, S., de Silva, N., & Dou, D. (2019). An overview of utilizing knowledge bases in neural networks for question answering. *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, 326–333.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.

Kim, J.-H., Jun, J., & Zhang, B.-T. (2018). Bilinear attention networks. *Advances in Neural Information Processing Systems*, 1564–1574.

Kim, J.-H., On, K.-W., Lim, W., Kim, J., Ha, J.-W., & Zhang, B.-T. (2016). Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

Kim, W., Son, B., & Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kiros, J., Chan, W., & Hinton, G. (2018). Illustrative language understanding: Large-scale visual grounding with image search. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 922–933.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. *Advances in Neural Information Processing Systems*, 3294–3302.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*(1), 32–73.

Krizhevsky, A., Hinton, G. et al. (2009). Learning multiple layers of features from tiny images.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2891–2903.

Lasecki, W. S., Thiha, P., Zhong, Y., Brady, E., & Bigham, J. P. (2013). Answering visual questions with conversational crowd assistants. *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, 1–8.

Lau, J. J., Gayen, S., Abacha, A. B., & Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, *5*(1), 1–10.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., & Huang, F. (2006). A tutorial on energy-based learning. *Predicting structured data*, *1*(0).

Lerer, A., Wu, L., Shen, J., Lacroix, T., Wehrstedt, L., Bose, A., & Peysakhovich, A. (2019). PyTorch-BigGraph: A large-scale graph embedding system. *Proceedings of the 2nd SysML Conference.*

Levine, Y., Lenz, B., Dagan, O., Ram, O., Padnos, D., Sharir, O., Shalev-Shwartz, S., Shashua, A., & Shoham, Y. (2020). Sensebert: Driving some sense into bert. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 4656–4667.

Li, G., Duan, N., Fang, Y., Gong, M., & Jiang, D. (2020). Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 11336–11344.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557.*

Li, S., Du, Y., van de Ven, G. M., & Mordatch, I. (2020). Energy-based models for continual learning. *arXiv preprint arXiv:2011.12216.*

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., & Wang, H. (2020). Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409.*

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. *Proceedings of the European Conference on Computer Vision*, 121–137.

Lin, B. Y., Chen, X., Chen, J., & Ren, X. (2019). Kagnet: Knowledge-aware graph networks for commonsense reasoning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, 2822–2832.

Lin, J., Yang, A., Zhang, Y., Liu, J., Zhou, J., & Yang, H. (2020). Interbert: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint arXiv:2003.13198.*

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *Proceedings of the European Conference on Computer Vision*, 740–755.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., & Wang, P. (2020). K-bert: Enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 2901–2908.

Liu, W., Wang, X., Owens, J. D., & Li, Y. (2020). Energy-based out-of-distribution detection. *arXiv preprint arXiv:2010.03759*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. *Advances in Neural Information Processing Systems*, *29*, 289–297.

Lv, S., Guo, D., Xu, J., Tang, D., Duan, N., Gong, M., Shou, L., Jiang, D., Cao, G., & Hu, S. (2020). Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 8449–8456.

Mahdisoltani, F., Biega, J., & Suchanek, F. (2014). Yago3: A knowledge base from multilingual wikipedias. *7th biennial conference on innovative data systems research*.

Malinowski, M., Doersch, C., Santoro, A., & Battaglia, P. (2018). Learning visual question answering by bootstrapping hard attention. *Proceedings of the European Conference on Computer Vision*, 3–20.

Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in Neural Information Processing Systems*, 1682–1690.

Malinowski, M., Rohrbach, M., & Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1–9.

Marino, K., Chen, X., Parikh, D., Gupta, A., & Rohrbach, M. (2021). Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 14111–14121.

Marino, K., Rastegari, M., Farhadi, A., & Mottaghi, R. (2019). OK-VQA: A visual question answering benchmark requiring external knowledge. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3195–3204.

Mihaylov, T., Clark, P., Khot, T., & Sabharwal, A. (2018). Can a suit of armor conduct electricity? A new dataset for open book question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2381–2391.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A. (1998). *Wordnet: An electronic lexical database*. MIT press.

Mishra, A., Shekhar, S., Singh, A. K., & Chakraborty, A. (2019). Ocr-vqa: Visual question answering by reading text in images. *International Conference on Document Analysis and Recognition*, 947–952.

Nam, H., Ha, J.-W., & Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 299–307.

Narasimhan, M., & Schwing, A. G. (2018). Straight to the facts: Learning knowledge base retrieval for factual visual question answering. *Proceedings of the European Conference on Computer Vision*, 451–468.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., & Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning.

Nguyen, D.-K., & Okatani, T. (2018). Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6087–6096.

Nijkamp, E., Hill, M., Han, T., Zhu, S.-C., & Wu, Y. N. (2020). On the anatomy of mcmc-based maximum likelihood learning of energy-based models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), 5272–5280.

Nijkamp, E., Hill, M., Zhu, S.-C., & Wu, Y. N. (2019). Learning non-convergent non-persistent short-run MCMC toward energy-based model. *arXiv preprint arXiv:1904.09770*.

Noh, H., & Han, B. (2016). Training recurrent answering units with joint loss minimization for vqa. *arXiv preprint arXiv:1606.03647*.

Noh, H., Kim, T., Mun, J., & Han, B. (2019). Transfer learning via unsupervised task discovery for visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8385–8394.

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748.*

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*, 8026–8037.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1532–1543.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018). Film: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence.*

Peters, M. E., Neumann, M., Logan, R., Schwartz, R., Joshi, V., Singh, S., & Smith, N. A. (2019). Knowledge enhanced contextual word representations. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, 43–54.

Qiao, Y., Deng, C., & Wu, Q. (2020). Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia.*

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rajani, N. F., McCann, B., Xiong, C., & Socher, R. (2019). Explain yourself! Leveraging language models for commonsense reasoning. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 4932–4942.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Ren, M., Kiros, R., & Zemel, R. (2015). Exploring models and data for image question answering. *Advances in Neural Information Processing Systems*, *28*, 2953–2961.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*, 91–99.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211–252.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Malaysia; Pearson Education Limited,

Saito, K., Shin, A., Ushiku, Y., & Harada, T. (2017). Dualnet: Domain-invariant network for visual question answering. *Proceedings of the IEEE International Conference on Multimedia and Expo*, 829–834.

Salimans, T., & Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in Neural Information Processing Systems*, 901–909.

Sap, M., Rashkin, H., Chen, D., Le Bras, R., & Choi, Y. (2019). Social iqa: Commonsense reasoning about social interactions. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, 4453–4463.

Shah, S., Mishra, A., Yadati, N., & Talukdar, P. P. (2019). KVQA: Knowledge-aware visual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2556–2565.

Shi, Y., Furlanello, T., Zha, S., & Anandkumar, A. (2018). Question type guided attention in visual question answering. *Proceedings of the European Conference on Computer Vision*, 151–166.

Shih, K. J., Singh, S., & Hoiem, D. (2016). Where to look: Focus regions for visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4613–4621.

Shrestha, R., Kafle, K., & Kanan, C. (2019). Answer them all! Toward universal visual question answering models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10472–10481.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*.

Singh, A., Goswami, V., & Parikh, D. (2020). Are we pretraining it right? Digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744*.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., & Rohrbach, M. (2019). Towards vqa models that can read. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8317–8326.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., & Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, *2*, 207–218.

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence.*

Storks, S., Gao, Q., & Chai, J. Y. (2019). Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172.*

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2019). Vl-bert: Pre-training of generic visual-linguistic representations. *International Conference on Learning Representations.*

Suhail, M., Mittal, A., Siddiquie, B., Broaddus, C., Eledath, J., Medioni, G., & Sigal, L. (2021). Energy-based learning for scene graph generation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 13936–13945.

Suhr, A., Lewis, M., Yeh, J., & Artzi, Y. (2017). A corpus of natural language for visual reasoning. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 217–223.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., & Artzi, Y. (2019). A corpus for reasoning about natural language grounded in photographs. *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). Videobert: A joint model for video and language representation learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7464–7473.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 843–852.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2019). Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4149–4158.

Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, 5103–5114.

Tang, C., Srivastava, N., & Salakhutdinov, R. R. (2014). Learning generative models with visual attention. *Advances in Neural Information Processing Systems*, *27*, 1808–1816.

Teney, D., Anderson, P., He, X., & van den Hengel, A. (2017). Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*.

Teney, D., & van den Hengel, A. (2016). Zero-shot visual question answering. *arXiv preprint arXiv:1611.05546*.

Teney, D., & van den Hengel, A. (2018). Visual question answering as a meta learning task. *Proceedings of the European Conference on Computer Vision*, 219–235.

Teney, D., & van den Hengel, A. (2019). Actively seeking and learning from live data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Teney, D., Wu, Q., & van den Hengel, A. (2017). Visual question answering: A tutorial. *IEEE Signal Processing Magazine*, *34*, 63–75.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., & Li, L.-J. (2016). Yfcc100m: The new data in multimedia research. *Communications of the ACM*, *59*(2), 64–73.

Tian, Y., Krishnan, D., & Isola, P. (2020). Contrastive multiview coding. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 776–794.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. *Proceedings of the International Conference on Language Resources and Evaluation*.

Tu, K., Meng, M., Lee, M. W., Choe, T. E., & Zhu, S.-C. (2014). Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, *21*(2), 42–70.

Tu, L., Pang, R. Y., Wiseman, S., & Gimpel, K. (2020). Engine: Energy-based inference networks for non-autoregressive machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2819–2826.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85.

Wang, P., Wu, Q., Shen, C., Dick, A., & van den Hengel, A. (2017a). Explicit knowledge-based reasoning for visual question answering. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1290–1296.

Wang, P., Wu, Q., Shen, C., Dick, A., & van den Hengel, A. (2017b). FVQA: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Wang, X., Gao, T., Zhu, Z., Liu, Z., Li, J., & Tang, J. (2019). KEPLER: A unified model for knowledge embedding and pre-trained language representation. *arXiv preprint arXiv:1911.06136.*

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., & Cao, Y. (2021). Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904.*

Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. *International Conference on Machine Learning*, 681–688.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., & Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698.*

Whitehead, S., Wu, H., Ji, H., Feris, R., & Saenko, K. (2021). Separating skills and concepts for novel visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5632–5641.

Wikipedia: The free encyclopedia. (2004). Retrieved November 1, 2021, from https://en.wikipedia.org/

Wu, Q., Shen, C., Wang, P., Dick, A., & van den Hengel, A. (2017). Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & van den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*.

Wu, Q., Wang, P., Shen, C., Dick, A., & van den Hengel, A. (2016). Ask me anything: Free-form visual question answering based on knowledge from external sources. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4622–4630.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 133–138.

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3733–3742.

Xiao, Z., Kreis, K., Kautz, J., & Vahdat, A. (2020). Vaebm: A symbiosis between variational autoencoders and energy-based models. *International Conference on Learning Representations*.

Xie, J., Lu, Y., Gao, R., & Wu, Y. N. (2018). Cooperative learning of energy-based model and latent variable model via mcmc teaching. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Xie, N., Lai, F., Doran, D., & Kadav, A. (2019). Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1492–1500.

Xu, H., Qi, G., Li, J., Wang, M., Xu, K., & Gao, H. (2018). Fine-grained image classification by visual-semantic embedding. *IJCAI*, 1043–1049.

Xu, H., & Saenko, K. (2016). Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *Proceedings of the European Conference on Computer Vision*, 451–466.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *International Conference on Machine Learning*, 2048–2057.

Yadav, V., & Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. *Proceedings of the 27th International Conference on Computational Linguistics*, 2145–2158.

Yampolskiy, R. V. (2013). Turing test as a defining feature of ai-completeness. *Artificial intelligence, evolutionary computing and metaheuristics* (pp. 3–17). Springer.

Yang, C., Jiang, M., Jiang, B., Zhou, W., & Li, K. (2019). Co-attention network with question type for visual question answering. *IEEE Access*, *7*, 40771–40781.

Yang, Y., Yih, W.-t., & Meek, C. (2015). Wikiqa: A challenge dataset for open-domain question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2013–2018.

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21–29.

Ye, Z.-X., Chen, Q., Wang, W., & Ling, Z.-H. (2019). Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. *arXiv preprint arXiv:1908.06725*.

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. B. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in Neural Information Processing Systems*, 1039–1050.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). 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.

Yu, D., Fu, J., Mei, T., & Rui, Y. (2017). Multi-level attention networks for visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4709–4717.

Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., & Wang, H. (2021). Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(4), 3208–3216.

Yu, Z., Yu, J., Cui, Y., Tao, D., & Tian, Q. (2019). Deep modular co-attention networks for visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6281–6290.

Yu, Z., Yu, J., Fan, J., & Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1821–1830.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1253.

Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., & Parikh, D. (2016). Yin and yang: Balancing and answering binary visual questions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5014–5022.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). Ernie: Enhanced language representation with informative entities. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1441–1451.

Zhao, S., Jacobsen, J., & Grathwohl, W. (2020). Joint energy-based models for semi-supervised classification. *ICML Workshop on Uncertainty and Robustness in Deep Learning.*

Zheng, W., Yan, L., Gou, C., & Wang, F.-Y. (2020). Webly supervised knowledge embedding model for visual reasoning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12445–12454.

Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., & Fergus, R. (2015). Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167.*

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., & Gao, J. (2020). Unified vision-language pre-training for image captioning and vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 13041–13049.

Zhu, Q., Gao, C., Wang, P., & Wu, Q. (2020). Simple is not easy: A simple strong baseline for textvqa and textcaps. *arXiv preprint arXiv:2012.05153*, *2*.

Zhu, Y., Groth, O., Bernstein, M., & Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4995–5004.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19–27.

Zhu, Z., Yu, J., Wang, Y., Sun, Y., Hu, Y., & Wu, Q. (2020). Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. *arXiv preprint arXiv:2006.09073*.

Zhuang, C., Zhai, A. L., & Yamins, D. (2019). Local aggregation for unsupervised learning of visual embeddings. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6002–6012.

Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. *Proceedings of the European Conference on Computer Vision*, 391–405.