

Leveraging Multimodal Perspectives to Learn Common Sense for Vision and Language Tasks

Xiao Lin

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Engineering

Devi Parikh, Co-chair
A. Lynn Abbott, Co-chair
Dhruv Batra
Pratap Tokekar
Harpreet S. Dhillon
Bert Huang

September 6, 2017
Blacksburg, Virginia

Keywords: Common Sense, Multimodal, Visual Question Answering, Image-Caption
Ranking, Vision and Language, Active Learning
Copyright 2017, Xiao Lin

Leveraging Multimodal Perspectives to Learn Common Sense for Vision and Language Tasks

Xiao Lin

ABSTRACT

Learning and reasoning with common sense is a challenging problem in Artificial Intelligence (AI). Humans have the remarkable ability to interpret images and text from different perspectives in multiple modalities, and to use large amounts of commonsense knowledge while performing visual or textual tasks. Inspired by that ability, we approach commonsense learning as leveraging perspectives from multiple modalities for images and text in the context of vision and language tasks.

Given a target task (*e.g.*, textual reasoning, matching images with captions), our system first represents input images and text in multiple modalities (*e.g.*, vision, text, abstract scenes and facts). Those modalities provide different perspectives to interpret the input images and text. And then based on those perspectives, the system performs reasoning to make a joint prediction for the target task. Surprisingly, we show that interpreting textual assertions and scene descriptions in the modality of abstract scenes improves performance on various textual reasoning tasks, and interpreting images in the modality of Visual Question Answering improves performance on caption retrieval, which is a visual reasoning task. With grounding, imagination and question-answering approaches to interpret images and text in different modalities, we show that learning commonsense knowledge from multiple modalities effectively improves the performance of downstream vision and language tasks, improves interpretability of the model and is able to make more efficient use of training data.

Complementary to the model aspect, we also study the data aspect of commonsense learning in vision and language. We study active learning for Visual Question Answering (VQA) where a model iteratively grows its knowledge through querying informative questions about images for answers. Drawing analogies from human learning, we explore cramming (entropy), curiosity-driven (expected model change), and goal-driven (expected error reduction) active learning approaches, and propose a new goal-driven scoring function for deep VQA models under the Bayesian Neural Network framework. Once trained with a large initial training set, a deep VQA model is able to efficiently query informative question-image pairs for answers to improve itself through active learning, saving human effort on commonsense annotations.

Leveraging Multimodal Perspectives to Learn Common Sense for Vision and Language Tasks

Xiao Lin

GENERAL AUDIENCE ABSTRACT

Designing systems that learn and reason with common sense is a challenging problem in Artificial Intelligence (AI). Humans have the remarkable ability to interpret images and text from different perspectives in multiple modalities, and to use large amounts of commonsense knowledge while performing visual or textual tasks. Inspired by that ability, we approach commonsense learning as leveraging perspectives from multiple modalities for images and text in the context of vision and language tasks.

Given a target task, our system first represents the input information (*e.g.*, images and text) in multiple modalities (*e.g.*, vision, text, abstract scenes and facts). Those modalities provide different perspectives to interpret the input information. Based on those perspectives, the system performs reasoning to make a joint prediction to solve the target task. Perhaps surprisingly, we show that imagining (generating) abstract scenes behind input textual scene descriptions improves performance on various textual reasoning tasks such as answering fill-in-the-blank and paraphrasing questions, and answering questions about images improves performance on retrieving image captions. Through the use of perspectives from multiple modalities, our system also makes use of training data more efficiently and has a reasoning process that is easy to understand.

Complementary to the system design aspect, we also study the data aspect of commonsense learning in vision and language. We study active learning for Visual Question Answering (VQA). VQA is the task of answering open-ended natural language questions about images. In active learning for VQA, a model iteratively grows its knowledge through querying informative questions about images for answers. Inspired by human learning, we explore cramming (entropy), curiosity-driven (expected model change), and goal-driven (expected error reduction) active learning approaches, and propose a new goal-driven query selection function. We show that once initialized with a large training set, a VQA model is able to efficiently query informative question-image pairs for answers to improve itself through active learning, saving human effort on commonsense annotations.

Dedication

To my parents, Fuyan Lin and Lianxiang Huang.

Acknowledgments

Special thanks to my advisor, Devi Parikh, for guiding and supporting me through my research. The works in this dissertation would not have happened without her vision, openness, patient guidance and determination.

I would like to thank my committee, Lynn Abbott, Dhurv Batra, Pratap Tokekar, Harpreet Dhillon and Bert Huang. Their insightful comments and suggestions leveraging knowledge from multiple areas pushed me to think more boardly and helped me polish the ideas and the dissertation.

I would like to thank my co-authors and collaborators – Ramakrishna Vedantam, Tanmay Batra, Larry Zitnick, Jiasen Lu, Michael Cogswell, Stanislaw Antol and Qing Sun – as well as everyone in the Computer Vision Lab, the Machine Learning and Perception Lab and the CVML Reading Group. It has been a pleasure brainstorming, polishing and realizing ideas together.

Last but not least, I would like to thank Yuandong Tian and Francis Quek for discussions and inspirations.

Contents

1	Introduction	1
1.1	Overview	2
2	Learning Common Sense Through Visual Abstraction	6
2.1	Introduction	6
2.2	Related Work	8
2.3	Datasets	10
2.3.1	Abstract Scenes Vocabulary	10
2.3.2	Tuple Extraction	11
2.3.3	Tuple Illustration Interface	13
2.4	Approach	15
2.4.1	Model	15
2.4.2	Training	18
2.5	Experimental Setup	18
2.5.1	Visual Features	18
2.5.2	Baselines	19
2.5.3	Evaluation	20
2.6	Results	20
2.6.1	Different Text Models	21

2.6.2	Joint Text + Vision Model	22
2.6.3	Qualitative Results	22
2.6.4	Enriching Knowledge Bases	25
2.7	Discussion	25
3	Leveraging Visual Common Sense for Non-Visual Tasks	27
3.1	Introduction	27
3.2	Related Work	30
3.3	Dataset	32
3.3.1	Fill-in-the-blank (FITB) Dataset	32
3.3.2	Visual Paraphrasing (VP) Dataset	33
3.4	Approach	35
3.4.1	Text Only Model	35
3.4.2	Incorporating Visual Common Sense	36
3.4.3	Scene Generation	40
3.4.4	Answering Questions with Imagined Scenes	41
3.5	Experiments and Results	42
3.5.1	Fill-in-the-blank	42
3.5.2	Visual Paraphrasing	44
3.6	Discussion	46
4	Leveraging Visual Question Answering for Image-Caption Ranking	48
4.1	Introduction	48
4.2	Related Work	50
4.3	Building Blocks: Image-Caption Ranking and VQA	52
4.3.1	Image-Caption Ranking	52

4.3.2	VQA	54
4.4	Approach	55
4.4.1	VQA-Grounded Representations	55
4.4.2	Score-level Fusion	58
4.4.3	Representation-level Fusion	59
4.5	Experiments and Results	60
4.5.1	Image-Caption Ranking Results	60
4.5.2	Ablation Study	63
4.5.3	The Role of VQA and Caption Annotations	64
4.5.4	Number of Question-Answer Pairs	65
4.5.5	Amount of VQA Training Data	66
4.5.6	On the Interpretability of the VQA-Aware Model	67
4.6	Generalization to Flickr8k and Flickr30k	68
4.7	Discussion	70
5	Active Learning for Visual Question Answering: An Empirical Study	71
5.1	Introduction	71
5.2	Related Work	74
5.2.1	Active Learning	74
5.2.2	Visual Conversations	75
5.3	Approach	75
5.3.1	Bayesian LSTM+CNN for VQA	76
5.3.2	Query Strategies and Approximations	77
5.4	Experiment	80
5.4.1	Experiment Setup	80

5.4.2	Active Learning on VQA v1.0 and v2.0	82
5.4.3	Goal-driven Active Learning	82
5.4.4	Quality of Approximations	86
5.5	Discussion	89
6	Conclusion and Future Research Directions	92
	Bibliography	108
A	Learning Common Sense Through Visual Abstraction	109
A.1	Extracting Tuples from Sentences	109
B	Leveraging Visual Common Sense for Non-Visual Tasks	111
B.1	Qualitative Results on Fill-in-the-blanks and Visual Paraphrasing	111
C	Leveraging Visual Question Answering for Image-Caption Ranking	116
C.1	Qualitative Examples	116
C.2	Information of (Q, A) Pairs	116
D	Active Learning for Visual Question Answering: An Empirical Study	121
D.1	Fast Approximation of Goal-driven Scoring Function	121

List of Figures

2.1	We consider the task of assessing how plausible a commonsense assertion is based on how similar it is to known plausible assertions. We argue that this similarity should be computed not just based on the text in the assertion, but also based on the visual grounding of the assertion. While “wants” and “looks at” are semantically different, their visual groundings tend to be similar. We use abstract scenes made from clipart to provide the visual grounding.	7
2.2	A subset of objects from our clipart library.	10
2.3	Snapshot of the interface used to collect human data about plausibility of assertions	12
2.4	Our tuple illustration AMT interface.	13
2.5	We show the original/ background image (last column) to the worker. The worker then illustrates a scene (column 4) containing the relation (column 2). The worker also selects the objects participating in the relation (column 5) and names them (column 1 and column 3). More examples of the data we collected can be found on https://vision.ece.vt.edu/cs/clipart_browser.html	14
2.6	We show some plausible assertions which get a higher score using text + vision than using just text, along with the clipart objects which (visually) support the assertions. More examples can be found on https://vision.ece.vt.edu/cs/assertion_browser.html	22
2.7	Visual and textual similarities are qualitatively different, and capture complementary signals.	24
2.8	Qualitative examples demonstrating visual similarity between tuples.	25

3.1	We introduce two tasks: fill-in-the-blank (FITB) and visual paraphrasing (VP). While they seem like purely textual tasks, they require some imagination – visual common sense – to answer.	28
3.2	Human performance vs. inter-human agreement on the FITB task. Mode of human responses is more accurate when subjects agree with each other. . . .	34
3.3	Scenes generated for an example FITB question.	40
3.4	Scenes generated for an example VP question.	41
3.5	FITB performance on subsets of the test data with varying amounts of human agreement. The margin of improvement of our approach over the baseline increases from 3% on all questions to 6% on questions with high human agreement.	44
4.1	Aligning images and captions requires high-level reasoning <i>e.g.</i> , “a batter up at the plate” would imply that a player is holding a bat, posing to hit the baseball and there might be another player nearby waiting to catch the ball. There is rich knowledge in Visual Question Answering (VQA) corpora containing human-provided answers to a variety of questions one could ask about images. We propose to leverage knowledge in VQA by using VQA models learnt on images and captions as “feature extraction” modules for image-caption ranking.	49
4.2	Our VQA and VQA-Caption network architectures.	55
4.3	Images and captions sorted by $P_I(A Q, I)$ and $P_C(A Q, C)$ assessed by our VQA (top) and VQA-Caption (bottom) models respectively. Indeed, images and captions that are more plausible for the (Q, A) pairs are scored higher. .	56
4.4	We propose score-level fusion (left) and representation-level fusion (right) to utilize VQA for image-caption ranking. They use VQA and VQA-Caption models as “feature extraction” schemes for images and captions and use those features to construct VQA-grounded representations. The score-level fusion approach combines the scoring functions of a VQA-grounded model and a baseline VQA-agnostic model. The representation-level fusion approach combines VQA-grounded representations and VQA-agnostic representations to produce a VQA-aware scoring function.	58

4.5	Qualitative image retrieval results of our score-level fusion VQA-aware model (middle) and the VQA-agnostic model (bottom). The true target image is highlighted (green if VQA-aware found it, red if VQA-agnostic found it but VQA-aware did not).	62
4.6	Left: caption retrieval and image retrieval performances of the VQA-agnostic model compared with our $N = 3000$ score-level fusion VQA-aware model trained using 1 to 5 captions per image. The VQA representations in the VQA-aware model provide consistent performance gains. Right: caption retrieval and image retrieval performances of our score-level fusion and representation-level fusion approaches with varying number of (Q, A) pairs used for feature extraction.	65
4.7	Facts that are most informative for ranking captions for each image in terms of mutual information between the fact and candidate captions. Selected from 3000 (Q, A) pairs using $N = 3000$ representation-level fusion VQA-aware model. 67	
5.1	Performance of two representative VQA models: LSTM+CNN [76] and HieCoAtt [77] on random subsets of the VQA v1.0 dataset. Both models improve by 12% with every order of magnitude of more training data.	72
5.2	Active learning versus passive learning on (top) VQA v1.0 and (bottom) v2.0. All three active learning strategies perform better than passive learning. . . .	81
5.3	Active learning with $N = 20k, 10k, 5k, 2k$ initial training set size. When dataset size is small, active learning is unable to outperform passive learning. The breakpoint when active learning methods start to perform better is around 30k to 50k examples.	83
5.4	Top: Goal-driven active learning of VQA for answering only “yes/no” questions. Our goal-driven active learning approach outperforms passive learning and other active learning approaches. Bottom: Query compositions of active learning approaches, on VQA v2.0 dataset for the task of answering only “yes/no” questions. Our goal-driven active learning approach queries mostly “yes/no” questions.	85
5.5	Goal-driven active learning of VQA for answering only “yes/no” questions, compared to passive learning that “cheats” and queries only “yes/no” questions.	86

5.6	Convergence of Monte Carlo approximation to entropy, curiosity-driven and goa-driven scoring functions in terms of rank correlation. We compute scores using Eq. 5.4 (entropy), 5.5 (curiosity-driven) and 5.7 (goal-driven) for 200 random examples from the pool using $M \in \{1, 2, 5, 10, 20, 50, 100, 200, 500\}$ samples from $q_{\theta}(\omega)$, and compare them with $M = 500$ in terms of rank correlation (Spearman's ρ).	87
5.7	Entropy, curiosity-driven and goal-driven scores of 50 examples under different numbers of model parameter samples.	88
5.8	Our fast approximations using Eq. 5.9 versus the original goal-driven scores computed using Eq. 5.7 under $M = \{2, 5, 10, 20, 50, 100, 200, 500\}$ samples of model parameters. Our approximations have high rank correlation with scores computed using the original method.	90
B.1	Qualitative results of fill-in-the-blanks, sampled based on predictions and ground truth.	112
B.2	Figure B.1 continued. Qualitative results of fill-in-the-blanks, sampled based on predictions and ground truth.	113
B.3	Qualitative results of visual paraphrasing, sampled based on predictions and ground truth.	114
B.4	Figure B.3 continued. Qualitative results of visual paraphrasing, sampled based on predictions and ground truth.	115
C.1	Qualitative results of image retrieval and caption retrieval at rank 1, 2 and 3 using our $N = 3,000$ score-level fusion VQA-aware model and the baseline VQA-agnostic model. The true target images and captions are highlighted.	119
C.2	Figure C.1 continued. Qualitative results of image retrieval and caption retrieval at rank 1, 2 and 3 using our $N = 3,000$ score-level fusion VQA-aware model and the baseline VQA-agnostic model. The true target images and captions are highlighted.	120

List of Tables

2.1	Performance of different text based methods on commonsense assertion assessment.	21
2.2	Text+ vision outperforms text alone on commonsense assertion assessment. .	21
3.1	Fill-in-the-blank performance of different approaches.	43
3.2	Visual paraphrasing performance of different approaches.	45
3.3	Coarse and fine-grained visual paraphrasing. In both coarse- and fine-grained settings, our approach using visual features show improvements on top of the text-only baseline.	46
4.1	Caption retrieval and image retrieval performances of our models compared to baseline models on MSCOCO image-caption ranking test set. Powered by knowledge in VQA corpora, both our score-level fusion and representation-level fusion VQA-aware approaches outperform state-of-the-art VQA-agnostic models by a large margin.	61
4.2	Results on MSCOCO using all 5,000 test images	63
4.3	Ablation study evaluating the gain in performance as more VQA-knowledge is incorporated in the model	64
4.4	Caption retrieval and image retrieval performances of score-level fusion $N = 3000$, when its VQA and VQA-Caption submodules are trained on smaller, randomly sampled subsets of the VQA dataset.	66
4.5	Results on Flickr8k dataset	69
4.6	Results on Flickr30k dataset	69

5.1	On VQA v2.0 for each pair of query strategy, what percentage of (Q, I) pairs are selected by both methods. Active learning (entropy, curiosity-driven, goal-driven) query strategies select $> 80\%$ common (Q, I) pairs and they are very different from passive learning.	84
-----	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----

Chapter 1

Introduction

Learning and reasoning with common sense is a challenging problem in Artificial Intelligence (AI). Humans have the remarkable ability to learn and to use large amounts of commonsense knowledge for reasoning. For example, the commonsense knowledge “before you throw an object you need to be holding it” associates the holding action with the throwing action and can help planning to throw a bowling ball. “People tend to look more often at the things they are interested in” can be useful for inferring people’s intention in photos and novels. Today’s AI agents have read the internet for knowledge, know the location of every coffee shop, and can beat human champions in chess and Go. But they are far from being sapient intelligent entities. Common sense continues to elude them.

Marvin Minsky points out in his book “The Emotion Machine” [87] that learning common sense is not just about building large collections of commonsense knowledge, but also about learning the right representations of commonsense knowledge and using such representations for target tasks. Minsky suggests that it is very unlikely for there to be a single best representation for all commonsense knowledge. Each particular representation has its advantages, as well as limitations. Moreover, attending to all commonsense knowledge at once within the short time of human reaction is intractable and inefficient. Instead, humans have the ability of multimodal thinking. For example when reasoning about the appearance of an object, the visual cortex can be evoked to apply visual knowledge about that object learnt from past experience to help making decisions. In this way we only need to attend to relevant modalities, while at the same time leveraging diverse representations of commonsense knowledge specialized in these modalities.

Taking that perspective, we approach commonsense learning as leveraging perspectives from

multiple modalities for image and text understanding in the context of vision and language tasks. Given a target task (*e.g.*, textual reasoning, matching images with captions), the system first represents input images and text in multiple modalities (*e.g.*, vision, text, abstract scenes and facts). Those modalities provides different perspectives to interpret the input images and text. And then based on those perspectives, the system performs reasoning to make a joint prediction for the target task. In this framework, the system is leveraging not just knowledge directly related to the target task, but also commonsense knowledge from other modalities which can help the performance on the target task. Furthermore, the multimodal perspectives can be used to analyze how final decisions are made and help building learning algorithms that are more transparent. Last but not least, the different perspectives for images and text can be shared across tasks, alleviating the need for large amounts of task-specific training data.

1.1 Overview

We propose three approaches that leverage perspectives learned from multiple modalities for images and text – grounding, imagination and question-answering – for learning and using common sense for vision and language tasks. We show that learning common sense from abstract scenes and Visual Question Answering is able to improve performance, interpretability and data efficiency on a variety of vision and language tasks, *e.g.*, assessing the plausibility of commonsense assertions, solving fill-in-the-blank and paraphrasing questions, and matching images with captions. Specifically:

Leveraging visual common sense in abstract scenes for plausibility assessment of commonsense tuples through alignment (grounding). [122] While some common-sense knowledge is explicitly stated in human-generated text and can be learnt by mining the web, much of it is unwritten. It is often unnecessary and even unnatural to write about commonsense facts. While unwritten, this commonsense knowledge is not unseen. The visual world around us is full of structure modeled by commonsense knowledge. Can machines learn common sense simply by observing our visual world? Unfortunately, this requires automatic and accurate detection of objects, their attributes, poses, and interactions between objects, which remain challenging problems. In this work our key insight is that while visual common sense is depicted in visual content, it is the semantic features that are relevant and not low-level pixel information. In other words, photorealism is not necessary to learn common sense. We explore the use of human-generated abstract scenes made from clipart

for learning common sense.

In particular, we reason about the plausibility of an interaction or, relation, between a pair of nouns using both visual and textual information. A noun-relation-noun tuple is deemed plausible if it has high alignment with the training tuples and visual abstractions. That is grounding the tuple in a collection of abstract scenes. A tuple’s alignment with the visual abstractions provides information on its visual plausibility. We show that by reasoning jointly with visual abstractions with high alignment with text, we can assess the plausibility of commonsense assertions more accurately than by reasoning using text alone.

Leveraging visual common sense in abstract scenes for text reasoning through generation (imagination). [74] When reading novels, humans are able to imagine the scenes behind the words to better understand the story. Can we use the same idea to improve text understanding for machines? In this work we leverage semantic visual common-sense knowledge learnt from abstract images in two textual reasoning tasks: fill-in-the-blank and visual paraphrasing of scene descriptions. Because the space of scene descriptions is exponentially large, given a scene description we may not have seen that exact scenario in images to apply an alignment-based approach. Inspired by human ability of imagining a scenario for reasoning, we propose to “imagine” or generate the scene behind the text as a perspective to understand the scene description, and then leverage visual cues from the “imagined” or generated scenes in addition to textual cues while answering these questions. Because photorealism is not necessary for learning common sense, we imagine the scenes in the space of abstract scenes. Our approach outperforms a strong text-only baseline on these tasks and the proposed tasks can serve as benchmarks to quantitatively evaluate progress in solving tasks that go “beyond recognition”.

Leveraging common sense in Visual Question Answering for image-caption ranking through representation fusion (answering questions). [75] Visual Question Answering (VQA) is the task of taking as input an image and a free-form natural language question about the image, and producing an accurate answer. We view VQA as a “feature extraction” module to extract image and caption representations. Each feature dimension validates (imagines) whether a fact (question-answer pair) could plausibly be true for the image and caption. By validating a large bank of facts, this feature allows the model to interpret images and captions from a wide variety of perspectives and leverage common-sense knowledge in VQA. Our key observation is that these representations are helpful for the task of image-caption ranking – ranking images given a caption, and ranking captions given an image. We propose score-level and representation-level fusion models to incorporate

VQA knowledge in these representations in an existing state-of-the-art VQA-agnostic image-caption ranking model. We find that incorporating the VQA representations and reasoning about consistency between images and captions significantly improves performance.

Complementary to the model aspect, we also study the data aspect of commonsense learning in vision and language. In particular, we study the task of VQA. Today’s state-of-the-art VQA models are deep neural networks. Their performance scales well with the amount of labeled training data, so a naive way to improve performance is to collect larger datasets for training. However, collecting large quantities of annotated data is expensive. Even worse, as a result of long tail distributions, it will likely result in redundant annotations while still having insufficient training data for rare concepts. This is especially important for learning commonsense knowledge, as it is well known that humans tend to talk about unusual circumstances more often than commonsense knowledge which can be boring to talk about.

Active learning helps address these issues. In active learning, a model is first trained on an initial training set. It then iteratively expands its training set by selecting potentially informative examples according to a query strategy, and seeking annotations on these examples.

Active Learning for Visual Question Answering. We present an empirical study of active learning for Visual Question Answering, where a deep VQA model selects informative question-image pairs from a pool and queries an oracle for answers to maximally improve its performance under a limited query budget. Drawing analogies from human learning, we explore cramming (entropy), curiosity-driven (expected model change), and goal-driven (expected error reduction) active learning approaches, and propose a new goal-driven active learning scoring function to pick question-image pairs for deep VQA models under the Bayesian Neural Network framework. We find that deep VQA models need large amounts of training data before they can start asking informative questions. But once they do, all three approaches outperform the random selection baseline and achieve significant query savings. For the scenario where the model is allowed to ask generic questions about images but is evaluated only on specific questions (*e.g.*, questions whose answer is either yes or no), our proposed goal-driven scoring function performs the best.

This dissertation is organized as follows. Chapter 2 describes using an alignment-based approach to learn common sense from abstract scenes for assessing the plausibility of commonsense tuples. Chapter 3 describes using an imagination-based approach to learn common sense from abstract scenes for textual reasoning of scene descriptions. Chapter 4 describes

leveraging common sense in VQA for image-caption ranking by answering questions about images and captions. These works were presented at the International Conference on Computer Vision (ICCV) at 2015, the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR) at 2015 and the 14th European Conference on Computer Vision (ECCV) at 2016. Chapter 5 presents our recent work on using active learning to query informative questions for learning VQA.

Chapter 2

Learning Common Sense Through Visual Abstraction

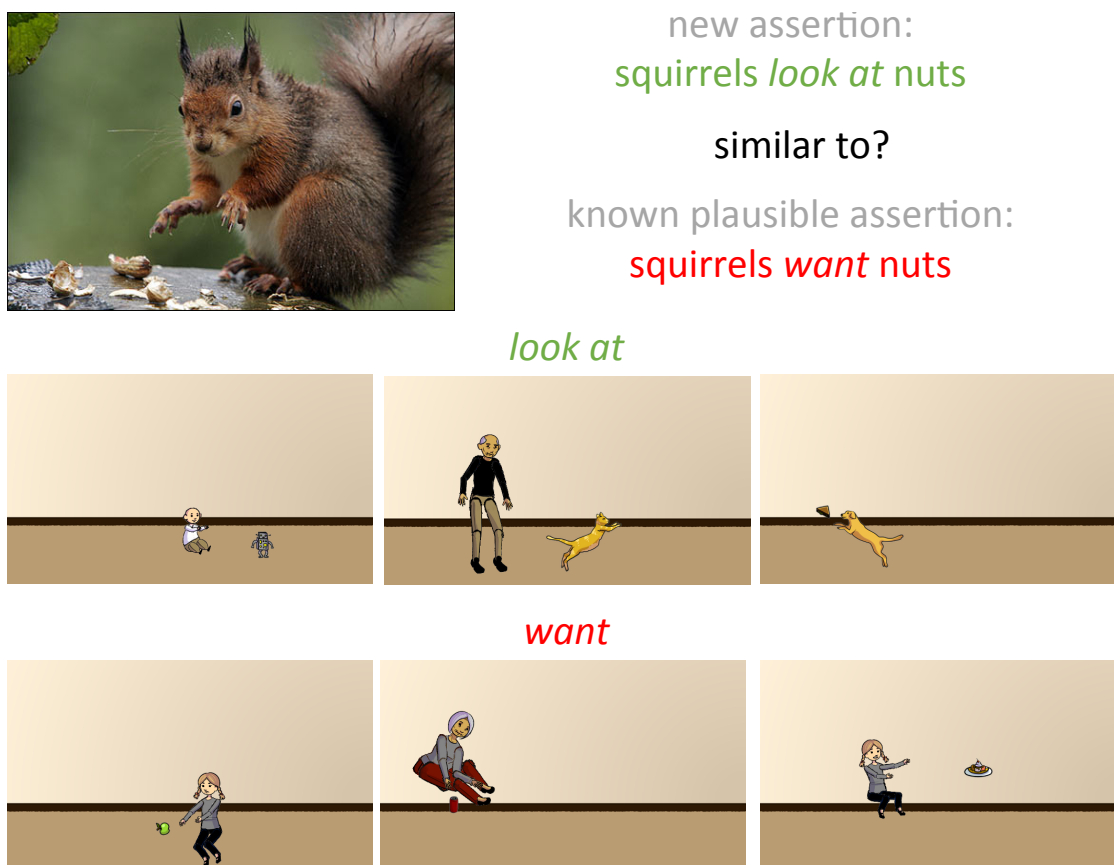
2.1 Introduction

Teaching machines common sense has been a longstanding challenge at the core of Artificial Intelligence (AI) [21]. Consider the task of assessing how plausible it is for a dog to jump over a tree. One approach is to mine text sources to estimate how frequently the concept of dogs jumping over trees is mentioned. A long history of works address the problem in this manner by mining knowledge from the web [11, 53, 70] or by having humans manually specify facts [8, 86, 110, 113] in text. Unfortunately, text is known to suffer from a reporting bias. If the frequency of mention was an indication of occurrence in the real world, people are ~ 3 times more likely to be murdered than they are to inhale, and people inhale ~ 6 times as often as they exhale [43]. This bias is not surprising. After all, people talk about things that are interesting to talk about, and unusual circumstances tend to be more interesting.

While unwritten, commonsense knowledge is not unseen! The visual world around us is full of structure modeled by our commonsense knowledge. By reasoning visually about a concept we may be able to estimate its plausibility more accurately. For instance, while “squirrels wanting nuts” is frequently mentioned in text, “squirrels looking at nuts” is rarely

©2015 IEEE. Reprinted, with permission, from R. Vedantam, X. Lin, T. Batra, C. Lawrence Zitnick, and D. Parikh. Learning common sense through visual abstraction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Figure 2.1: We consider the task of assessing how plausible a commonsense assertion is based on how similar it is to known plausible assertions. We argue that this similarity should be computed not just based on the text in the assertion, but also based on the visual grounding of the assertion. While “wants” and “looks at” are semantically different, their visual groundings tend to be similar. We use abstract scenes made from clipart to provide the visual grounding.



mentioned even though it is equally plausible. However, if we visually imagine a squirrel wanting a nut, we typically imagine a squirrel looking at a nut (Figure 2.1). This is because wanting something and looking at something tend to be visually correlated, even though they have differing underlying meaning. Interestingly, in the word2vec [85] text embedding space that is commonly used to measure word similarity, *look at* is more similar to *feel* than to *want*. Clearly, vision and text provide complementary signals for learning common sense.

Unfortunately, extracting commonsense knowledge from visual content requires automatic and accurate detection of objects, their attributes, poses, and interactions. These remain

challenging problems in computer vision. Our key insight is that commonsense knowledge may be gathered from a high-level semantic understanding of a visual scene, and that low-level pixel information is typically unnecessary. In other words, photorealism is not necessary to learn common sense. In this work, we explore the use of human-generated abstract scenes made from clipart for learning common sense. Note that abstract scenes are inherently *fully* annotated, allowing us to exploit the structure in the visual world, while bypassing the difficult intermediate problem of training visual detectors.

Specifically, we consider the task of assessing the plausibility of an interaction or relation between a pair of nouns, as represented by a tuple (primary noun, relation, secondary noun) e.g., (boy, kicks, ball). As training data, we collect a dataset of tuples and their abstract visual illustrations made from clipart. These illustrations are created by subjects on Amazon Mechanical Turk (AMT). We use this to learn a scoring function that can score how well an abstract visual illustration matches a test tuple.

Given a previously unseen tuple, we assess its plausibility using both visual and textual information. A tuple is deemed plausible if it has high alignment with the training tuples and visual abstractions. When measuring textual similarity between tuples we exploit the significant progress that has been made in learning word similarities from web scale data using neural network embeddings [85, 95]. A tuple’s alignment with the visual abstractions provides information on its visual plausibility. We model a large number of free form relations (213) and nouns (2466), which may form over ≈ 1 billion possible tuples. We show that by jointly reasoning about text and vision, we can assess the plausibility of commonsense assertions more accurately than by reasoning about text alone.

The rest of this paper is organized as follows. We discuss related work in Section 2.2. Our data collection methodology is described in Section 2.3. Our model for classifying novel commonsense assertions (tuples) as plausible or not is presented in Section 2.4. Section 2.5 describes our experimental setup, followed by quantitative and qualitative results in Section 2.6, and a conclusion in Section 2.7.

2.2 Related Work

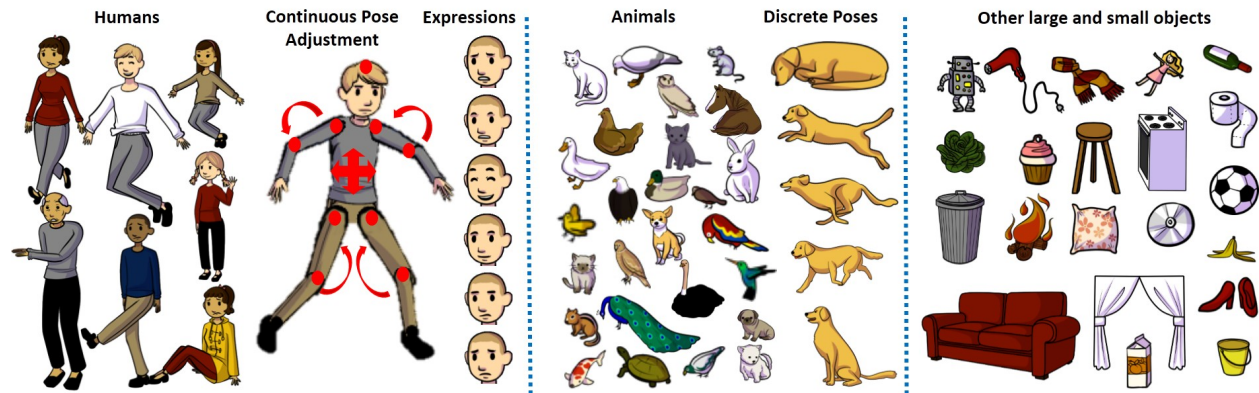
Common sense and text. There is a rich line of works which learn relations between entities to build knowledge bases either using machine reading (e.g., Knowledge Vault [29], NELL [11], ReVerb [32]) or using collaboration within a community of users (e.g., Free-

base [8], Wikipedia¹). We make use of the ReVerb Information Extraction system to create our dataset of tuples (more details in Section 2.3.2). Our goal is to learn common sense from a complementary source: our visual world. A task closely related to learning common sense from text is answering questions. Systems such as IBM Watson [36] combine multiple text-based knowledge bases to answer factual questions. Our work focuses on combining different modalities of information (abstract scenes and text) for the task of assessing the plausibility of commonsense assertions.

Common sense and vision. A popular use of commonsense knowledge in vision has been for modeling context for improved recognition [25, 37, 45, 47, 50]. Recently, there has been a surge in interest in high-level “beyond recognition” tasks which can benefit from external knowledge beyond what is depicted in the image [5, 49, 59, 96, 97]. Zhu *et al.* [136] use attribute and action classification along with information from various textual knowledge bases to perform tasks like zero-shot affordance prediction for human-object interactions. Their dictionary of relations was specified manually and limited to 19 inter-object relations. We explore a larger number of *free-form* relations (213 in total) extracted from text. Johnson *et al.* [56] build a scene graph representation for image retrieval which models attribute and object relations. LEVAN [24] trains detectors for a variety of bigrams (e.g., jumping horse) from google n-grams using web-scale image data. NEIL [16] analyzes images on the web to learn visual models of objects, scenes, attributes, part-of, and other ontology relationships. Our focus is less on appearance models and more on the underlying semantics. Recent work has also looked at mining *semantic* affordances, *i.e.* inferring whether a given action can be performed on an object [12]. In contrast, we are interested in the more general problem of predicting the plausibility of interactions or relations between pairs of objects. Lin and Parikh [74] propose to learn visual common sense and use it to answer textual fill-in-the-blank and visual paraphrasing questions, by imagining a scene behind the text. While they model visual common sense in the context of a scene, our task is at a more atomic level – reasoning about the plausibility of a specific relation or interaction between pairs of objects. Most similar to ours is a concurrent work VisKE [105] which also studies the task of evaluating the plausibility of commonsense assertions using visual cues. Their visual cues are derived from webly-supervised detection models, while we use abstract scenes and text embeddings. A new test tuple can be processed almost instantaneously using our approach, while training their webly-supervised detector takes ~ 30 minutes per tuple. It is conceivable that text, abstract scenes and real images are all complementary sources of information.

¹<http://www.wikipedia.org/>

Figure 2.2: A subset of objects from our clipart library.



Learning from visual abstraction. Visual abstractions have been explored for a variety of high-level scene understanding tasks. Zitnick and Parikh [138] learn the importance of various visual features (occurrence, co-occurrence, expression, gaze, *etc.*) in determining the meaning or semantics of a scene. Zitnick *et al.* also link the semantics of a scene to memorability and saliency of objects [140]. [139] learns the visual interpretation of sentences and generates scenes for a given input sentence. Fouhey and Zitnick [38] learn the dynamics of objects in scenes from temporal sequences of abstract scenes. Antol *et al.* [2] learn models of fine-grained interactions between pairs of people using visual abstractions, and evaluate their models on real images from the web. Lin and Parikh [74] “imagine” abstract scenes corresponding to text, and use the common sense depicted in these imagined scenes to solve textual tasks such as fill-in-the-blanks and paraphrasing. In this work, we are interested in using abstract scenes as a complementary source of commonsense knowledge to text for the task of classifying commonsense assertions as plausible or not.

2.3 Datasets

2.3.1 Abstract Scenes Vocabulary

In order to learn comprehensive commonsense knowledge, it is important for the library of clipart pieces to be expressive enough to model a wide variety of scenarios. Previous works on using visual abstractions depicted a boy and a girl playing in a park [38, 138, 139] with a library of 58 objects, or fine-grained interactions between two people [2] (no

additional objects). Instead, our clipart library allows us to depict a variety of indoor scenes. It contains 20 “paperdoll” human models [2] spanning genders, races, and ages with 8 different expressions. The limbs are adjustable to allow for continuous pose variations. The vocabulary contains over 100 small and large objects and 31 animals in various poses, that can be placed at one of 5 discrete scales or depths in the scene, facing left or right. Our clipart is also more realistic looking than previous work. A snapshot of the library can be viewed in Figure 2.2. Note that while we restrict ourselves to indoor scenes in this work, our idea is general and applicable to other scenes as well. More clipart objects and scenes can be easily added to the clipart library.

2.3.2 Tuple Extraction

Extracting Seed Assertions: To collect a dataset of commonsense assertions, we start by extracting a set of seed tuples from image captions. We use the MS COCO training set [73] containing images annotated with 80 object categories and five captions per image. We pick a subset of 9913 images whose annotated objects all come from a list of manually selected objects from our library of clipart.² Note that MS COCO images are not fully annotated and contain many more objects than those annotated. As a result, captions for these images could contain nouns that may not be part of the annotated object list or our clipart library. Our model can handle this by using word embeddings as described in Section 2.4.1.

We split the images into VAL (4956 images) and TEST (4957 images). We then run the ReVerb [32] information extraction tool on the captions for these images (images are not involved anymore), along with some post-processing (described in Appendix A.1) to obtain a set of (t_P, t_R, t_S) tuples, where t_P is the primary noun, t_R is the relation, and t_S is the secondary noun in the tuple t e.g., (plate, topped with, meat). All tuples containing relations that occur less than four times in the dataset are likely to be noisy extractions, and are removed. Details about ReVerb tuple extraction can be found in Appendix A.1. This gives us a set of 4848 tuples in VAL and 4778 in TEST, 213 unique relations in VAL and 204 in TEST, and 2466 unique nouns in VAL and 2378 in TEST. VAL and TEST have 893 tuples, 814 nouns, and 151 relations in common. These tuples form our seed commonsense assertions.

²List: *person, cat, dog, frisbee, bottle, wine glass, cup, fork, knife, spoon, apple, sandwich, hotdog, pizza, cake, chair, couch, potted plant, bed, dining table, tv, book, scissors, teddy bear* was selected to capture objects in our clipart library that are commonly found in living room scenes.

Figure 2.3: Snapshot of the interface used to collect human data about plausibility of assertions

Below is a list of 20 different scenarios. For each one, please tell us if that scenario typically occurs.

In other words, would you be surprised to encounter this scenario?

Please ignore any minor grammatical errors. But if the scenario doesn't make any sense to you at all, please indicate so.

1. puppy **sit on** leash

☐ Yes, this typically occurs ☐ No, this doesn't occur typically ☐ I don't understand what this scenario is trying to describe.

2. woman **have** cupcake

☐ Yes, this typically occurs ☐ No, this doesn't occur typically ☐ I don't understand what this scenario is trying to describe.

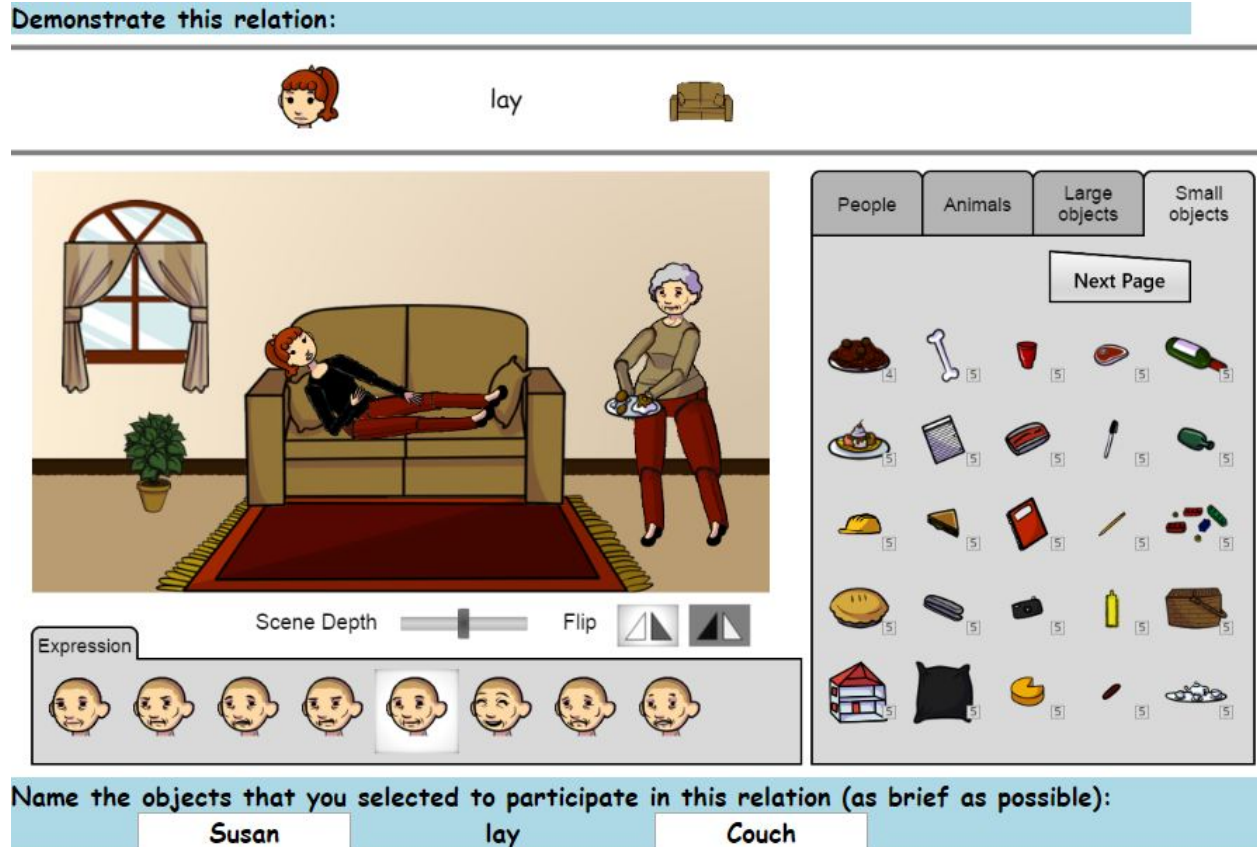
Expanding Seed Assertions: We expand our seed set of assertions by generating random assertions. This is done on both TEST and VAL independently. We iterate through each tuple twice, and pair the corresponding t_R with a random t_P and t_S from all nouns that occur at least 10 times³. So there are twice as many expanded tuples as there are seed tuples. This results in 9700 expanded tuples in VAL and 9554 in TEST. Note that we are sampling from a space of 160 primary nouns (>10 occurrences) \times 204 relations \times 160 nouns i.e., >5 million possible TEST assertions. In total across seed and expanded, our VAL set contains 14548 commonsense assertions spanning 213 relations, and our TEST set contains 14,332 commonsense assertions spanning 204 relations. To the best of our knowledge, ours is the first work that models such a large number of relations and commonsense assertions.

Supervision on Expanded Assertions: We then show our set of assertions (seed + expanded) to subjects on Amazon Mechanical Turk (AMT). Workers on Amazon Mechanical Turk are shown a question and asked to rate if the scenario described by the assertion typically happens or not. We also give workers an option to tell us if the scenario described by the assertion makes no sense. We collect 10 independent human judgments per assertion.

80.1% of annotations on seed tuples were positive. This is not surprising because these tuples were extracted from descriptions of images, and were thus clearly plausible. The creation of random expanded tuples predominantly adds negatives. But we found that some randomly

³This is a coarse proxy for sampling nouns proportional to how often they occur in the seed set.

Figure 2.4: Our tuple illustration AMT interface.

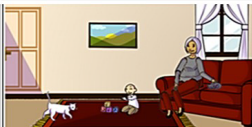






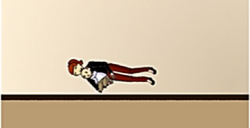





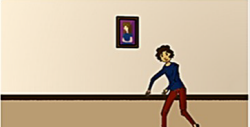



generated assertions such as (puppy, lay next to, chair) and (dogs, lay next to, pepperoni pizza) were rated as plausible (positives). 15.3% of annotations on our expanded tuples were positive. Overall, 36% of the labels in VAL and 37% of the labels in TEST are positives.

2.3.3 Tuple Illustration Interface

We collect abstract illustrations for all 213 relations in VAL. We get each relation illustrated by 20 different workers on AMT using the interface shown in Figure 2.4. Each worker is shown a *background* scene and asked to modify it to contain the relation of interest. We used living room scenes from [1] as background scenes, which were realistic scenes created by AMT workers using the same abstract scenes vocabulary as ours (Section 2.3.1). Priming workers with different background scenes helps increase the diversity in the visual illustrations of relations. For instance, when asked to create a scene depicting ‘holding’, a majority of

Figure 2.5: We show the original/ background image (last column) to the worker. The worker then illustrates a scene (column 4) containing the relation (column 2). The worker also selects the objects participating in the relation (column 5) and names them (column 1 and column 3). More examples of the data we collected can be found on https://vision.ece.vt.edu/cs/clipart_browser.html

Primary	Relation	Secondary	Image	Objects	Original
Cat	Walk	To child			
Puppy	Stand at	Table			
Child	Sleep next to	Woman			
Wine	Served on	Table			
Woman	Shown in	Painting			

workers might default to thinking of a person holding something while standing. But if they are primed with a scene where a woman is already sitting on a couch, then they might place a glass in her hand to make her hold the glass, resulting in a sitting person holding something. Workers are then instructed to indicate which clipart pieces in the scene correspond to the primary and secondary objects participating in the relation, and name them using as few words as possible.

To summarize, we collect 20 scenes depicting each of the 213 relations in VAL (4260 scenes total), along with annotations for the primary and secondary nouns and corresponding clipart objects participating in the relation. These form our set of TRAIN tuples that will be used to train our visual models of what tuples looks like. The VAL tuples will be used to learn

how much visual alignment is weighted relative to the textual alignment. The TEST tuples will be used to evaluate the performance of our approach.

Note that we do not collect illustrations for each VAL *tuple* because tuples may contain nouns that our clipart library does not have. Instead, we collect illustrations for each of the VAL *relations*. Workers choose to depict these relations with plausible primary and second objects of their choice, providing an additional source of commonsense knowledge. Regardless, as will be evident in the next section, our model is capable of dealing with nouns and relations at test time that were not present during training.

Figure 2.5 shows the some sample illustrations created for relations, along with the corresponding tuples (Primary Object, Relation, Secondary Object) phrases. A subset of relations along with all corresponding human illustrations collected to form the TRAIN set can be found on https://vision.ece.vt.edu/cs/clipart_browser.html.

2.4 Approach

We first describe our joint text and vision model, followed by a description of the training procedure.

2.4.1 Model

Let us start by laying out some notation. We are given a commonsense assertion $t' = (t'_P, t'_R, t'_S)$ at test time, whose plausibility is to be evaluated. t'_P is the primary noun, t'_R is the relation, and t'_S is the secondary noun. For each abstract training scene created by AMT workers $i \in I$ we are given the primary and secondary clipart objects c_P^i and c_S^i , as well as a tuple $t^i = (t_P^i, t_R^i, t_S^i)$ containing the names of the primary and secondary objects (nouns), and the relation they participate in. Thus, a training instance i is represented by $\Omega^i = \{c_P^i, c_S^i, t^i\}$.

We score the plausibility of test tuple t' using the following linear scoring function:

$$score(t') = \alpha \cdot f_{text}(t') + \beta \cdot f_{visual}(t') \quad (2.1)$$

Where α and β tradeoff the weights given to the text alignment score f_{text} and the vision alignment score f_{vision} respectively. The text and vision alignment scores estimate how well

the test tuple t' aligns to all training instances – both textual (TRAIN tuples provided by AMT workers) and visual (training abstract scenes provided by AMT workers). Tuples which align well with known (previously seen and/or read) concepts are considered to be more plausible.

Vision and text alignment functions: Both our vision and text alignment functions have the following form:

$$f(t') = \frac{1}{|I|} \sum_{i \in I} \max(h(t', \Omega^i) - \delta, 0) \quad (2.2)$$

Where f can be either f_{text} or f_{vision} . The average goes over all training instances (i.e., abstract scenes with associated annotated tuples) in our training set. The activation of a training instance with respect to a test tuple is determined by h , which has different forms for vision and text. A ReLU (Rectified Linear Unit) function is applied to the activation score offset by δ . We use a threshold of zero for the ReLU because the notion of negative plausibility evidence for a tuple is not intuitive. One can view Equation 2.2 as counting how many times a tuple was observed during training. The parameter δ is used to threshold the activation h to estimate counts. From here on we refer to h as the alignment score (overloaded with f).

Text alignment score: The textual alignment score h_{text} between two tuples is a linear combination of similarities between the corresponding pairs of primary nouns, relations, and secondary nouns. These similarities are computed using dot products in the word2vec embedding space [85]. For nouns or relations containing more than one word (e.g., “gather around” or “chair legs”), we average the word2vec vectors of each word to obtain a single vector.

Let $W(x)$ be the vector space embedding of a noun or relation x . The text alignment score is given as follows:

$$h_{text}(t', \Omega^i) = W(t'_P)^T \cdot W(t_P^i) + W(t'_R)^T \cdot W(t_R^i) + W(t'_S)^T \cdot W(t_S^i) \quad (2.3)$$

Where \cdot denotes the cosine similarity between vectors.

Vision alignment score: The visual alignment score computes the alignment between (i) a given test tuple and (ii) the pair of clipart pieces selected by AMT workers as being the

primary and secondary objects in a training instance i . It measures how well the pair of clipart pieces (c_P^i, c_S^i) depict the test tuple t' . If a test tuple finds support from a large number of visual instances, it is likely to be plausible. Note that we are measuring similarity between words and arrangements of clipart pieces. Consequently, this is a multimodal similarity function.

Given the pair of primary and secondary clipart pieces annotated in training instance Ω^i , we extract features as described in Section 2.5. We denote these extracted features as $u(c_P^i, c_S^i)$. Using these visual features from the training instance Ω^i and text embeddings from test tuple t' , we compute the following vision alignment score:

$$h_{vision}(t', \Omega^i) = u(c_P^i, c_S^i)^T A_P W(t'_P) + u(c_P^i, c_S^i)^T A_R W(t'_R) + u(c_P^i, c_S^i)^T A_S W(t'_S) \quad (2.4)$$

Where A_P , A_R , and A_S are alignment parameters to be learnt. Our vision alignment score measures how well the t'_P , t'_R , and t'_S individually match the visual features $u(c_P^i, c_S^i)$ that describe a pair of clipart objects in training instance Ω_i . One can think of $u(c_P^i, c_S^i)A_P$, $u(c_P^i, c_S^i)A_R$, and $u(c_P^i, c_S^i)A_S$ as embeddings or projections from the vision space to the word2vec text space, such that a high dot product in word2vec space leads to high alignment, and subsequently a high plausibility score for plausible tuples. The embeddings are learnt separately for t'_P , t'_R and t'_S (as parameterized by A_P , A_R and A_S) because different visual features might be useful for aligning to the primary noun, relation, and secondary noun.

The parameters A_P , A_R , and A_S can also be thought of as grounding parameters. That is, given a word2vec vector W , we learn parameters to find the visual instantiation of W . $A_R W(t'_R)$ can be thought of as the visual instantiation of t'_R which captures what the interaction between two objects related by relation t'_R looks like. $A_P W(t'_P)$ and $A_S W(t'_S)$ can be thought of as identifying which clipart pieces and with what attributes correspond to nouns t'_P and t'_S . Our model finds the visual grounding of t'_P , t'_R , and t'_S separately, and then measures similarity of the inferred grounding to the actual visual features observed in training instances. Thus, given a test tuple, we *hallucinate* a grounding for it and measure similarity of the hallucination with the training data. Note that these hallucinations are learnt discriminatively to help us align concepts in vision and text such that plausible tuples are scored highly.

2.4.2 Training

To learn the parameters A_P , A_R , A_S in our vision alignment scoring function (Equation 2.4), we consider the outer product space of the vectors u and W . We learn a linear SVM in this space to separate the training instances (tuples + corresponding abstract scenes, Section 2.3.3), from a set of negatives. Each negative instance is a tuple from our TRAIN set, paired with a random abstract scene from our training data. We sample three times as many negatives as positives. Overall we have 4260 positives and 12780 negatives. Finally, the learnt vectors are reshaped to get A_P , A_R and A_S respectively. We learn the vision vs. text tradeoff parameters α and β (Equation 2.1) on the VAL set of tuples (Section 2.3.2). Recall that these include seed and expanded tuples, along with annotations indicating which tuples are plausible and which are not. We use the vision and text alignment scores as features and train a binary SVM to separate plausible tuples from implausible ones. The weights learnt by the SVM correspond to α and β . Finally, the parameter δ in Equation 2.2 is set using grid search on the VAL set to maximize the average precision (AP) of predicting a tuple as being plausible (positive) or not.

2.5 Experimental Setup

We first describe the features we extract from the abstract scenes. We then list the baselines we compare to.

2.5.1 Visual Features

As explained in Section 2.3.1, we have annotations indicating which pairs of objects (c_P , c_S) in an abstract scene participated in the corresponding annotated tuple. Using these objects and the remaining scene, we extract three kinds of features to describe the pair of objects (c_P , c_S): 1) Object Features 2) Interaction Features 3) Scene Features. These three together form our visual feature set. **Object Features** consist of the type (category, instance) of the object (Section 2.3.1), flip (left facing or right) of the object, absolute location, attributes (for humans), and poses (for humans and animals). The absolute location feature is modeled using a Gaussian Mixture Model (GMM) with 9 components, learnt separately across five discrete depth levels, similar to [139]. The GMM components are common across all objects, and are learnt using all objects present in all abstract scenes. Human attributes are age

(5 discrete values), skin color (3 discrete values) and gender (2 discrete values). Animals have 5 discrete poses. Human pose features are constructed using keypoint locations. These include global, contact, and orientation features [2]. Global features measure the position of joints with respect to three gaussians placed on the head, torso, and feet respectively. Contact features place smaller gaussians at each joint and measure the positions of other joints with respect to each joint. Orientation features measure the joint angles between connected keypoints. **Interaction Features** encode the relative locations of the two objects participating in the relation, normalized for the flip and depth of the first object. This results in the relative location features being asymmetric. We compute the relative location of the primary object relative to the secondary object and vice versa. Relative locations are encoded using a 24 component GMM (similar to [139]). **Scene Features** indicate which types (category, instance) of objects (other than c_P and c_S) are present in the scene. Overall, there are 493 object features each for the primary and secondary objects, 48 interaction features, and 188 global features, resulting in a visual feature vector of dimension 1222.

2.5.2 Baselines

We experiment with a variety of strong baselines that use text information alone. They help evaluate how much complementary information vision adds, and if this additional information can be obtained simply from additional or different kinds of text (e.g., generic vs. visual text).

- **WikiEmbedding:** Our first baseline uses the f_{text} part of our model (Equation 2.1) alone. It uses word2vec trained on generic Wikipedia text.
- **COCOEmbedding:** Our next baseline also uses the f_{text} part of our model (Equation 2.1) alone, but uses word2vec trained on visual text (>400k captions in the MS COCO training dataset).
- **ValText:** Recall that both our TEST and VAL tuples were extracted from captions describing COCO images. Our next baseline computes the plausibility of a test tuple by counting how often that tuple occurred in VAL. This helps assess the overlap between our TEST and VAL tuples (recall: no images are shared between TEST and VAL). Note that the above two baselines, WikiEmbedding and COCOEmbedding, can be thought of as ValText but by using soft similarities (in word2vec space) rather than using counts based on exact matches.

- **LargeVisualText**: Our next baseline is a stronger version of ValText. Instead of using just our VAL tuples to evaluate the plausibility of a test tuple, it extracts tuples from a large corpus of text describing images (>400k captions in the MS COCO training dataset which are not in our test set (Section 2.3.2)). This gives us a set of 91K assertions. At test time, we check how many times the test assertion occurred in this set, and use that count as the plausibility score of the test tuple.
- **BigGenericText (Bing)**: In this baseline, we evaluate the performance of assessing the plausibility of tuple $t' = (t'_P, t'_R, t'_S)$ in the test set using all the text on the web. We query the Bing⁴ search API and compute the log-frequencies of t'_P , t'_R , t'_S as well as t' . We train an SVM on these four features to separate plausible tuples in our VAL set from implausible tuples, and use this SVM at test time to compute the plausibility score of a test tuple.

2.5.3 Evaluation

Recall that we collected 10 human judgements for the plausibility of each test tuple (Section 2.3.2). We count the number of subjects who thought the tuple was plausible ($count_+$). We also count the number of subjects who thought the tuple was not plausible ($count_-$). $count_+ + count_-$ need not be 10 because subjects were allowed to mark tuples as “does not make sense”. These scores are then combined into a single $score = count_+ - count_-$. We threshold these scores at 0 to get our set of positive and negative human (ground truth) labels. That is, a tuple is considered to be plausible if more people thought it is plausible than not. Our method as well as the baselines produce a score for the plausibility of each tuple in the TEST set. These scores are thresholded and compared to the human labels to compute average precision (AP). We also rank tuples based on their predicted plausibility scores and human plausibility scores ($score = count_+ - count_-$). These rankings are compared using a rank correlation, which forms our second evaluation metric.

2.6 Results

We begin by comparing our text-based baseline models. We then demonstrate the advantage of using vision and text jointly, over using text alone or vision alone. We then show qualitative

⁴<http://www.bing.com/>

Table 2.1: Performance of different text based methods on commonsense assertion assessment.

Approach	Test Performance	
	AP	Rank Correlation $\times 100$
WikiEmbedding	68.4	41.7
COCOEmbedding	72.2	49.0
ValText	53.0	31.0
LargeVisualText	58.0	37.6
BigGenericText (Bing)	44.6	20.3

Table 2.2: Text+ vision outperforms text alone on commonsense assertion assessment.




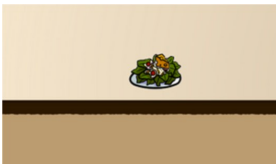


Approach	Test Performance	
	AP	Rank Correlation $\times 100$
Text (COCOEmbedding) + Vision	73.6	50.0
Vision Only	68.7	45.3
Text (COCOEmbedding) Only	72.2	49.0

results. We finally comment on the potential our approach has to enrich existing knowledge bases.

2.6.1 Different Text Models

Of all the text-alone baselines (Table. 2.1), we find that BigGenericText (Bing) does the worst, likely because it suffers heavily from the reporting bias on the web. The LargeVisualText baseline does better than Bing, presumably because the captions in MS COCO describe what is seen in the images which may often be mundane details depicted in the image, and aligns well with the source of our tuples (visual text). ValText performs worse than LargeVisualText because ValText uses less data. But adding soft similarities using word2vec embeddings (WikiEmbedding and COCOEmbedding) significantly improves performance (15.4 and 19.2 in absolute AP). COCOEmbedding performs the best among all text-alone baselines, and is what we will use as our “text only” model moving forward.

Figure 2.6: We show some plausible assertions which get a higher score using text + vision than using just text, along with the clipart objects which (visually) support the assertions. More examples can be found on https://vision.ece.vt.edu/cs/assertion_browser.html

Assertion	Supporting Cliparts		Text Score	Text + Vision Score
dog "stand with" blanket			0.29	0.30
plate "hold" sandwich			0.009	0.011
boy "have" flower			0.01	0.08

2.6.2 Joint Text + Vision Model

We compare the performance of text + vision, vision alone, and text alone in Table. 2.2. We observe that text + vision performs better than text alone and vision alone by 1.4% and 4.9% AP respectively. In terms of rank correlation, text + vision provides an improvement of 1.0 over text alone. Overall, vision and text provide complementary sources of common sense.

2.6.3 Qualitative Results

We first present qualitative examples where using visual cues with text helps (Figure 2.6). The figure shows some assertions which are rated by humans as *plausible*. We see that these

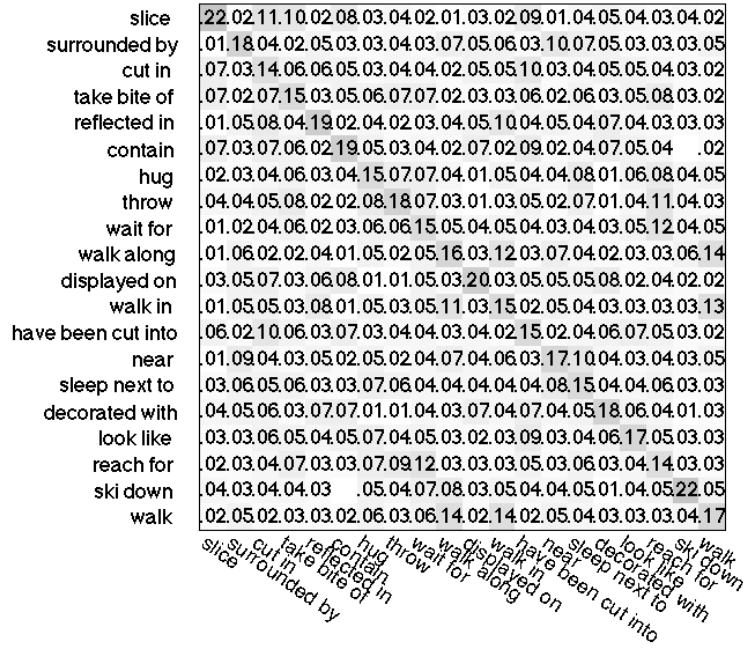
tuples are rated as more plausible when we take the help of vision. For instance, consider the example boy *have* flower. “having” seems to find support from visual instantiations of images one would describe as “beside” (supporting cliparts row) rather than “have”. However, with the visual grounding these lead to a higher score. More examples can be found at https://vision.ece.vt.edu/cs/assertion_browser.html. The predictions of the text+vision model, along with text only and vision only models are given, categorized by relation t_R , at shown. The text tuples and visual illustrations which give most support to the TEST assertion are also shown.

We then visualize relation similarity matrices for text and vision alone (Figure 2.7). Each entry in the text matrix is the word2vec similarity between the relations specified in the corresponding row and columns. Each row is normalized to sum to 1. For vision, each entry in the matrix is the proportion of images depicting a relation (row) whose embeddings – after being transformed by A_R – are most similar to the word2vec representation of another relation (column). This illustrates what our visual alignment function has learnt. We randomly sample a subset of 20 relations for visualization purposes. We can clearly see that the two matrices are qualitatively different and complementary. For instance, visual cues tell us that the relations like “sleep next to” and “surrounded by” are similar. The predictions from the classifier trained on visual features, to predict t_P , t_R , and t_S are shown at https://vision.ece.vt.edu/cs/clipart_browser_w_pred.html. These are qualitative visualizations to see which relations are most similar *visually*. We also show similarity between the predictions and the ground truth tuples using our text model based on word2vec.

In Figure 2.8 we show several scenes created by AMT workers. Note that for clarity we only show the primary and secondary objects as identified by workers, but our approach uses all objects present in the scene. For each scene, we show the “GT” tuple provided by workers, as well as the “Vision only” tuple. This is computed by embedding the scene using our learnt A_P , A_R , and A_S into the word2vec space and identifying the nouns and relations that are most similar. The left most column shows scenes where the visual prediction matches the GT. The next column shows scenes where the visual prediction is incorrect, but reasonable (even desirable) and would not be captured by text. Consider (boy, hold onto, pizza) and (boy, take, pizza) whose similarity would be difficult to capture via text. The next column shows examples where the tuples are visually as well as textually similar. The last column shows failure cases where the visual prediction is unreasonable.

Figure 2.7: Visual and textual similarities are qualitatively different, and capture complementary signals.

(a) Textual similarity between relations



(b) Visual similarity between relations

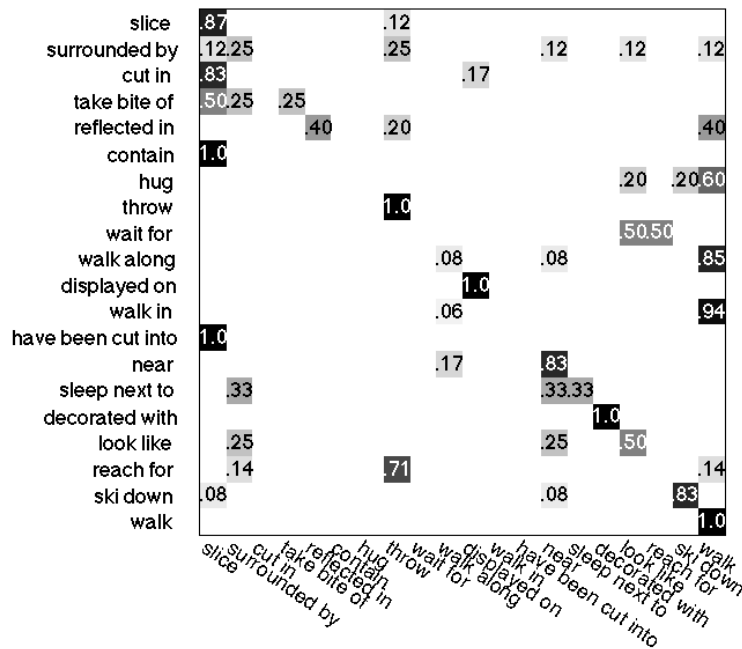
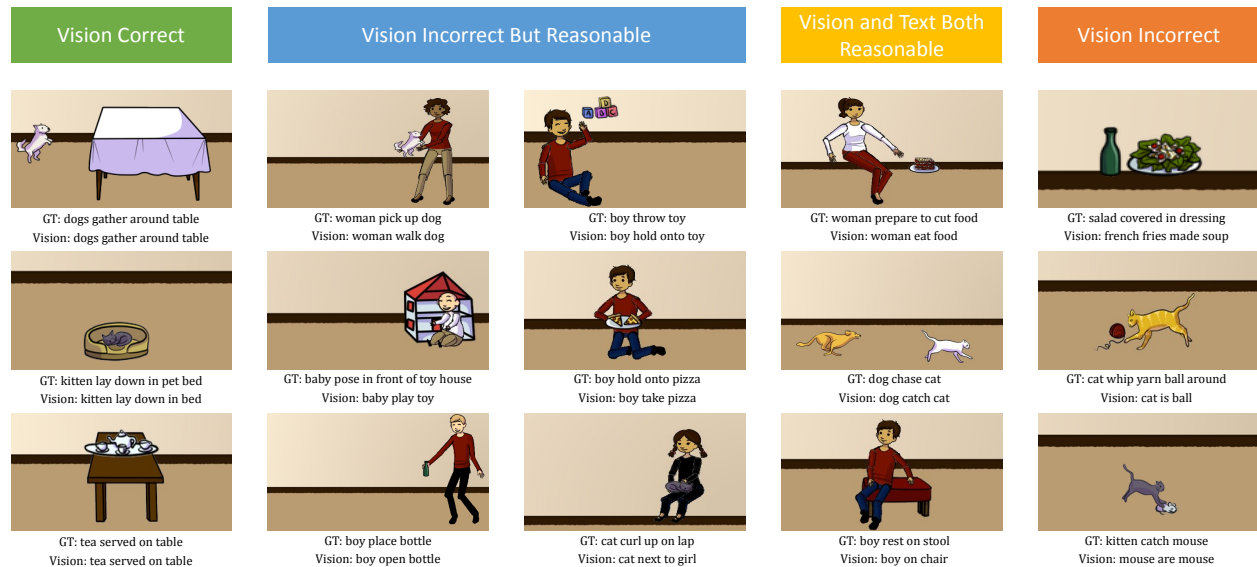


Figure 2.8: Qualitative examples demonstrating visual similarity between tuples.



2.6.4 Enriching Knowledge Bases

ConceptNet [113] contains commonsense knowledge contributed by volunteers. It represents concepts with nodes and relations as edges between them. Out of our 213 VAL relations, only one relation (“made of”) currently exists in ConceptNet. Thus, our approach can add many visual commonsense relations to ConceptNet, and boost its recall.

2.7 Discussion

In this work we considered the task of classifying commonsense assertions as being plausible or not based on how similar they are to assertions that are known to be plausible. We argued that vision provides a complementary source of commonsense knowledge to text. Hence, in addition to reasoning about the similarity between tuples based on text, we propose to ground commonsense assertions in the visual world and evaluate similarity between assertions using visual features. We demonstrate the effectiveness of abstract scenes in providing this grounding. We show that assertions can be classified as being plausible or not more accurately using vision + text, than by using text alone. All our datasets and code are publicly available.

In this work our commonsense assertions dataset is bootstrapped from scene descriptions of scenes that share the same object categories as our clipart library. We try to make sure that plausible commonsense assertions in the dataset can be illustrated using abstract scenes made from the clipart library. For commonsense assertions that are about objects outside of the clipart library, one would need to add new cliparts for those objects in order to use our approach effectively. Our approach does not explicitly handle assertions that are not about visual concepts (where reasoning with vision may not provide a bonus). As a future research direction, it is desirable to tell which assertions are about visual concepts and which are not, so the visual alignment functions can be turned on or turned off accordingly to maximize performance.

In addition, our visual alignment function relies on finding cliparts that correspond to the commonsense assertions in the form of tuples. It is efficient for this type of problems but there exists fine-grained textual input which may involve multiple tuples or more complex structures such as scene descriptions, for which it can be hard to find similar images in a collection. So in Chapter 3, we introduce an approach that imagines the scene behind the text to perform reasoning on scene descriptions.

Acknowledgments

We thank Stanislaw Antol for his help with the tuple illustration interface. This work is supported in part by an Allen Distinguished Investigator Award from the Paul G. Allen Family Foundation and by a Google Faculty Research Award to D. P.

Chapter 3

Leveraging Visual Common Sense for Non-Visual Tasks

3.1 Introduction

Today’s artificially intelligent agents are good at answering factual questions about our world [18, 36, 120]. For instance, Siri¹, Cortana², Google Now³, Wolfram Alpha⁴ *etc.*, when asked “How far is the closest McDonald’s to me?”, can comprehend the question, mine the appropriate database (*e.g.*, maps) and respond with a useful answer. While being good at niche applications or answering factual questions, today’s AI systems are far from being sapient intelligent entities. Common sense continues to elude them.

Consider a simple fill-in-the-blank task shown in Figure 3.1 (left). Answering this question requires the common sense that bears are dangerous animals, people like to stay away from and not be noticed by dangerous animals, and hiding is one way of going unnoticed. Similarly, consider the visual paraphrasing question in Figure 3.1 (right). Answering this question involves common sense that people might throw things when they are angry and in order

©2015 IEEE. Reprinted, with permission, from X. Lin and D. Parikh. Don’t just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

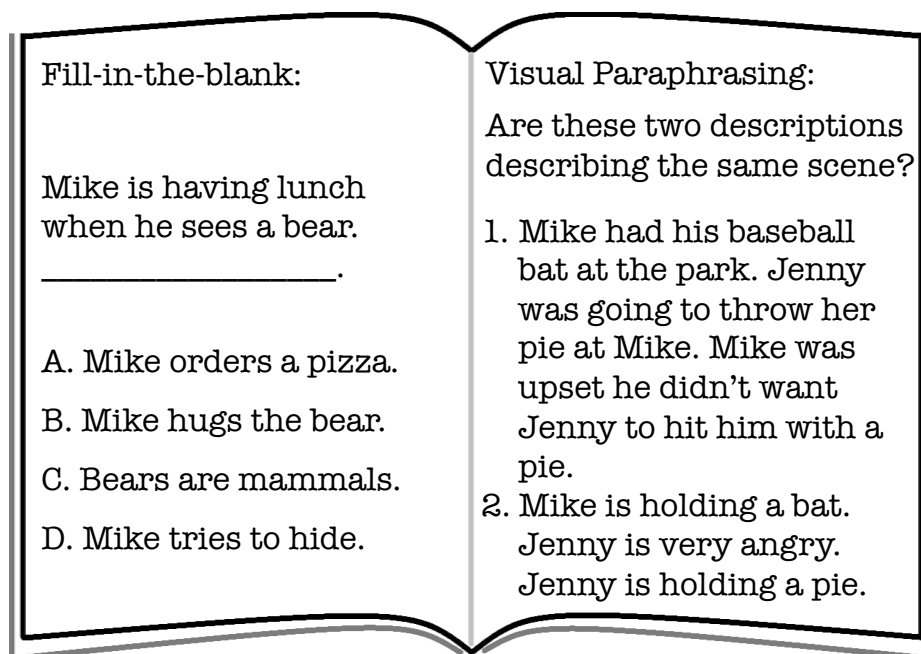
¹<https://www.apple.com/ios/siri/>

²<http://www.windowsphone.com/en-us/how-to/wp8/cortana/meet-cortana>

³<http://www.google.com/landing/now/>

⁴<http://www.wolframalpha.com/>

Figure 3.1: We introduce two tasks: fill-in-the-blank (FITB) and visual paraphrasing (VP). While they seem like purely textual tasks, they require some imagination – visual common sense – to answer.



to throw something, you need to be holding it. Today's systems are unable to answer such questions reliably.

Perhaps this is not surprising. Most existing common sense knowledge bases rely on knowledge described via text – either mined [11, 53, 70] or manually entered [86, 110, 8, 113]. There are a few short-comings of learning common sense from text. First, it has been shown that people tend not to explicitly talk about common sense knowledge in text [43]. Instead, there is a bias to talk about unusual circumstances, because those are worth talking about. Co-occurrence statistics of visual concepts mined from the web has been shown to not generalize to images [84]. Even when describing images, text is likely to talk about the salient “foreground” objects, activities, *etc.*. But common sense reveals itself even in the “background”. Second, much of useful common sense knowledge may be hard to describe in text. For instance, the knowledge that “one person is running after another person” implies that the first person is facing the second person, the second person is looking in the same direction as the first person, and both people are in running poses, is unnatural (and typically

unnecessary) to articulate in text.

Fortunately, much of this common sense knowledge is depicted in our visual world. We call such common sense knowledge that can be learnt from visual data *visual common sense*. By visual common sense we do not mean visual models of commonly occurring interactions between objects [24] or knowledge of visual relationships between objects, parts and attributes [16, 136]. We mean semantic common sense, *e.g.*, the knowledge that if one person is running after another person, and the second person turns around, he will see the first person. It can be learnt from visual data but can help in a variety of visual *and* non-visual AI tasks. Such visual common sense is complementary to common sense learnt from non-visual sources.

We argue that the tasks shown in Figure 3.1 may look like purely text- or language-based tasks on the surface, but they can benefit from visual common sense. In fact, we go further and argue that such tasks can provide exciting new benchmarks to evaluate image understanding “beyond recognition”. Effectively learning and applying visual common sense to such tasks involves challenges such as grounding language in vision and learning common sense from visual data – both steps towards deeper image understanding beyond naming objects, attributes, parts, scenes and other image content depicted in the pixels of an image.

In this work we propose two tasks: fill-in-the-blank (FITB) and visual paraphrasing (VP) – as seen in Figure 3.1 – that can benefit from visual common sense. We propose an approach to address these tasks that first “imagines” the scene behind the text. It then reasons about the generated scenes using visual common sense, as well as the text using textual common sense, to identify the most likely solution to the task. In order to leverage visual common sense, this imagined scene need not be photo-realistic. It only needs to encode the semantic features of a scene (which objects are present, where, what their attributes are, how they are interacting, *etc.*). Hence, we imagine our scenes in an abstract representation of our visual world – in particular using clipart [138, 139, 38, 2].

Specifically, given an FITB task with four options, we generate a scene corresponding to each of the four descriptions that can be formed by pairing the input description with each of the four options. We then apply a learnt model that reasons jointly about text and vision to select the most plausible option. Our model essentially uses the generated scene as an intermediate representation to help solve the task. Similarly, for a VP task, we generate a scene for each of the two descriptions, and apply a learnt joint text and vision model to classify both descriptions as describing the same scene or not. We introduce datasets for

both tasks. We show that our imagination-based approach that leverages both visual and textual common sense outperforms the text-only baseline on both tasks. Our datasets and code are publicly available at <https://filebox.ece.vt.edu/~linxiao/imagine/>.

3.2 Related Work

Beyond recognition: Higher-level image understanding tasks go beyond recognizing and localizing objects, scenes, attributes and other image content depicted in the pixels of the image. Example tasks include reasoning about *what* people talk about in images [5], understanding the flow of time (*when*) [96], identifying *where* the image is taken [49, 59] and judging the intentions of people in images (*why*) [97]. While going beyond recognition, these tasks are fairly niche. Approaches that automatically produce a textual description of images [47, 34, 67] or synthesize scenes corresponding to input textual descriptions [139] can benefit from reasoning about all these different “W” questions and other high-level information. They are semantically more comprehensive variations of beyond recognition tasks that test high-level image understanding abilities. However, these tasks are difficult to evaluate [67, 31] or often evaluate aspects of the problem that are less relevant to image understanding *e.g.*, grammatical correctness of automatically generated descriptions of images. This makes it difficult to use these tasks as benchmarks for evaluating image understanding beyond recognition.

Leveraging visual common sense in our proposed FITB and VP tasks requires qualitatively a similar level of image understanding as in image-to-text and text-to-image tasks. FITB requires reasoning about what else is plausible in a scene given a partial textual description. VP tasks on the other hand require us to reason about how multiple descriptions of the same scene could vary. At the same time, FITB and VP tasks are multiple-choice questions and hence easy to evaluate. This makes them desirable benchmark tasks for evaluating image understanding beyond recognition.

Natural language Q&A: Answering factual queries in natural language is a well studied problem in text retrieval. Given questions like “Through which country does the Yenisei river flow?”, the task is to query useful information sources and give a correct answer for example “Mongolia” or “Russia”. Many systems such as personal assistant applications on phones and IBM Watson [36] which won the Jeopardy! challenge have achieved commercial success. There are also established challenges on answering factual questions posed by humans [18],

natural language knowledge base queries [120] and even university entrance exams [94]. The FITB and VP tasks we study are not about facts, but common sense questions.

[42, 81] have addressed the task of answering questions about visual content. The questions and answers often come from a closed world. [102] introduces self-contained fictional stories and multiple choice reading comprehension questions that test text meaning understanding. [126] models characters, objects and rooms with simple spatial relationships to answer queries and factual questions after reading a story. Our work can be seen as using the entire scene as the “meaning” of text.

Leveraging common sense: Common sense is an important element in solving many beyond recognition tasks, since beyond recognition tasks tend to require information that is outside the boundaries of the image. It has been shown that learning and using *non-visual* common sense (*i.e.* common sense learnt from non-visual sources) benefits physical reasoning [48, 132], reasoning about intentions [97] and object functionality [136]. One instantiation of visual common sense that has been leveraged in the vision community in the past is the use of contextual reasoning for improved recognition [47, 25, 45, 37, 50, 136]. In this work, we explore the use of visual common sense for seemingly non-visual tasks through “imagination”, *i.e.* generating scenes.

Synthetic data: Learning from synthetic data avoids tedious manual labeling of real images. It also provides a platform to study high-level image understanding tasks without having to wait for low-level recognition problems to be solved. Moreover, synthetic data can be collected in large amounts with high density without suffering from a heavy-tailed distribution, allowing us to learn rich models. Previous works have looked at learning recognition models from synthetic data. For instance, computer graphics models were used to synthesize data to learn human pose [108], chair models [3], scene descriptions and generation of 3D scenes [14]. Clipart data has been used to learn models of fine-grained interactions between people [2]. [72] warps images of one category to use them as examples for other categories. [57] uses synthetic images to evaluate low-level image features. Human-created clipart images have been used to learn which semantic features (object presence or co-occurrence, pose, expression, relative location, *etc.*) are relevant to the meaning of a scene [138] and to learn spatio-temporal common sense to model scene dynamics [38]. In this work, we learn our common sense models from human-created clipart scenes and associated descriptions. We also use clipart to “imagine” scenes in order to solve the FITB and VP tasks. Though the abstract scenes [138, 14] are not photo-realistic, they offer a semantically rich world where one can effectively generate scenes and learn semantic variations of sentences and scenes, free

from the bottlenecks of (still) imperfect object recognition and detection. Despite being synthetic, it has been shown that semantic concepts learnt from abstract scenes can generalize to real images [2].

3.3 Dataset

We build our FITB and VP datasets on top of the Abstract Scenes Dataset [138], which has 10,020 human-created abstract scenes of a boy and a girl playing in the park. The dataset contains 58 clipart objects including the boy (Mike), the girl (Jenny), toys, background objects like trees and clouds, animals like dogs and cats, food items like burgers and pizzas, *etc.*. A subset of these objects are placed in the scene at a particular location, scale, and orientation (facing left or right). The boy and the girl can have different poses (7) and expressions (5). Each one of the 10,020 scenes has textual descriptions written by two different people. We use this clipart as the representation within which we will “imagine” our scenes. We also use this dataset to learn visual common sense. While more clipart objects, expressions, poses, *etc.* can enable us to learn more comprehensive visual common sense, this dataset has been shown to contain semantically rich information [138, 139], sufficient to begin exploring our proposed tasks. We now describe our approach to creating our FITB and VP datasets.

3.3.1 Fill-in-the-blank (FITB) Dataset

Every description in the Abstract Scenes Dataset consists of three short sentences, typically describing different aspects of the scene while also forming a coherent description. Since we have two such descriptions for every scene, we arbitrarily place one of the two descriptions (for all scenes) into the source set and the other into the distractor set. For each image, we randomly drop one sentence from its source description to form an FITB question. We group this dropped sentence with 3 random sentences from descriptions of other images in the distractor set. The FITB task is to correctly identify which sentence in the options belongs to the original description in the question.

Removing questions where the NLP parser produced degenerate outputs, our resulting FITB dataset contains 8,959 FITB questions – 7,198 for training and 1,761 for testing. Figure 3.3 shows one example FITB question from our dataset. The scenes corresponding to the ques-

tions in the training set are available for learning visual common sense and text-image correspondence. The scenes corresponding to the test questions are not available at test time.

FITB is a challenging task. Many scenes share the same visual elements such as Mike and Jenny playing football. Sometimes the distractor options may seem just as valid as the ground truth option, even to humans. We conduct studies on human performance on the test set. We had 10 different subjects on Amazon Mechanical Turk (AMT) answer the FITB questions. To mimic the task given to machines, subjects were not shown the corresponding image. We found that the majority vote response (*i.e.* mode of responses) across 10 subjects agreed with the ground truth 52.87% of the time (compared to random guessing at 25%).

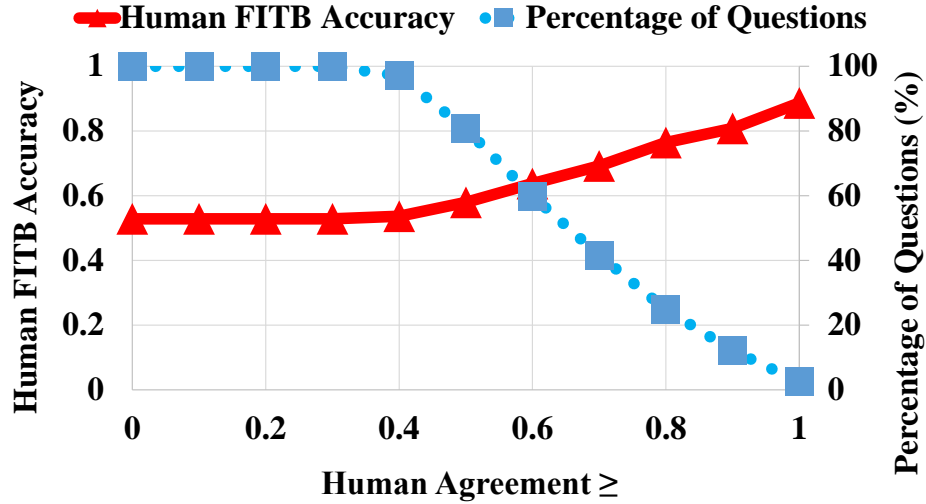
Some questions may be generic and ambiguous and can lead to disagreements among the subjects, while other questions have consistent responses across subjects. We find that 41% of the questions in our dataset have 7 or more subjects agreeing on the response. Of these questions, the mode of the responses across subjects agrees with the ground truth 69% of the time. Interestingly, on the remaining 31% of the questions, 7 out of 10 subjects agree on the *wrong* response. This happens because often the distracting options happen to describe the original image well, or their writing style matches that of the question. In our experiments, we report accuracies relative to the ground truth response, as well as relative to the response that most subjects agree on (the latter might be more relevant from an AI perspective – if the goal is to produce human-like responses).

In Figure 3.2, we consider different subsets of the dataset formed by only considering questions where a certain minimum proportion of subjects agreed on the response (human agreement). For each subset, we can evaluate the accuracy of the mode response. We also look at what percentage of the dataset falls in each subset. Not surprisingly, human accuracy (mode agreeing with ground truth) correlates well with human agreement (percentage of subjects that agree with mode). Note that even if responses were random, on average 43% of subjects would agree on the mode response.

3.3.2 Visual Paraphrasing (VP) Dataset

The VP task is to tell if two descriptions are describing the same scene or two different scenes. The correct answer to a pair of descriptions written by two people describing the same scene is “Yes”, while to randomly drawn descriptions from two different scenes is “No”.

Figure 3.2: Human performance vs. inter-human agreement on the FITB task. Mode of human responses is more accurate when subjects agree with each other.



We build our VP dataset using all 10,020 scenes from the Abstract Scenes Dataset, resulting in a dataset with 10,020 positive pairs. We randomly sample $2 \times 10,020$ pairs as negatives. This leads to a total of 30,060 questions in our dataset. Of these, 24,000 are used for training and the rest 6,060 are used for testing. We choose the negative pairs separately in training and testing sets such that they do not overlap with each other. Figure 3.4 shows one example VP question from our dataset.

We evaluate human performance on our test set. We had 10 different subjects on AMT solve our tasks. We average their responses (0 for No and 1 for Yes) to obtain a score between 0 and 1 for each question. We can use this score to plot a precision-recall curve. Results show that humans can reliably solve this task with 94.78% average precision (AP), compared to chance at 33%.

FITB and VP tasks are ways to evaluate visual common sense. Some applications of FITB tasks may be automatic story telling and automatic Q&A. Some applications of the VP task may be text-based image retrieval and generating multiple diverse descriptions of the same image.

3.4 Approach

We first describe the strong baseline approach of using textual features (common sense) to solve the FITB and VP tasks in Section 3.4.1. We then describe our visual common sense model (Section 3.4.2) and scene generation approach (Section 3.4.3). Finally in Section 3.4.4 we describe our approach to using our model to solve the FITB and VP tasks.

3.4.1 Text Only Model

We first tokenize all words in our dataset and form a vocabulary (1,886 words for the FITB dataset and 2,495 for the VP dataset). We also form a vocabulary of pairs of words by selecting 100 pairs of words which have the highest mutual information in the training data and co-occur more than 100 times.

Both FITB and VP involve reasoning about consistency between two descriptions (question and option for FITB and two input descriptions for VP). Given two descriptions d_1 and d_2 , we extract three kinds of textual features from the pair. The first is term frequency, commonly used for text classification and retrieval, which counts how often each word from our vocabulary occurs in (d_1, d_2) (both descriptions concatenated). The second is a 400D word co-occurrence vector indicating for each (of the 100) pair of words whether: (i) the first word occurred in d_1 and the second word occurred in d_2 or (ii) the first word occurred in d_1 and the second word did not occur in d_2 or (iii) the first word did not occur in d_1 and the second word occurred in d_2 or (iv) the first word did not occur in d_1 and the second word did not occur in d_2 . The third uses a state-of-the-art neural word embedding word2vec [85] trained on questions from our training set to represent each word with a (default) 200D vector. We then average the vector responses of all words in (d_1, d_2) . These features capture common sense knowledge about which words are used interchangeably to describe the same thing, which words tend to co-occur in descriptions, *etc.*.

Fill-in-the-blank. For N fill-in-the-blank questions and M options per question, we denote the question as $q_i, i \in \{1, \dots, N\}$ and the options for q_i as $o_{ij}, j \in \{1, \dots, M\}$. We denote the ground truth option for question q_i as o_i^{gt} , and its index as j_i^{gt} .

The FITB problem is a ranking problem: given q_i , we wish to rank the correct option o_i^{gt} above distractors $o_{ij}, j \neq j_i^{gt}$. For each question-option pair (q_i, o_{ij}) , we extract the three

kinds of textual features as described above using $d_1 = q_i$ and $d_2 = o_{ij}$. Concatenating these three gives us a 2,486D text feature vector $\phi_{fitb}^{text}(q_i, o_{ij})$. We compute scores $s_{ij} = w^T \phi_{fitb}^{text}(q_i, o_{ij})$ for each option that captures how likely o_{ij} is to be the answer to q_i . We then pick the option with the highest score. We learn w using a ranking SVM [13]:

$$\begin{aligned} \min_{w, \xi \geq 0} \quad & \frac{1}{2} \|w\|^2 + C \sum_{(i,j), j \neq j^{gt}} \xi_{ij} \\ \text{s.t.} \quad & w^T \phi_{fitb}^{text}(q_i, o_i^{gt}) - w^T \phi_{fitb}^{text}(q_i, o_{ij}) \geq 1 - \xi_{ij}, \\ & \forall (i, j), j \neq j^{gt} \end{aligned} \tag{3.1}$$

Visual paraphrasing. In visual paraphrasing, for each question i , the goal is to verify if the two given descriptions q_{i1} and q_{i2} describe the same image ($y_i = 1$) or not ($y_i = -1$). We extract all three features described above using $d_1 = q_{i1}$ and $d_2 = q_{i2}$. Let's call this ϕ_{vp1}^{text} . We extract the same features but using $d_1 = q_{i2}$ and $d_2 = q_{i1}$. Let's call this ϕ_{vp2}^{text} . To ensure that the final feature representation is invariant to changing the order of the two descriptions – *i.e.* $\phi_{vp}^{text}(q_{i1}, q_{i2}) = \phi_{vp}^{text}(q_{i2}, q_{i1})$, we use $\phi_{vp}^{text} = [\phi_{vp1}^{text} + \phi_{vp2}^{text}, |\phi_{vp1}^{text} - \phi_{vp2}^{text}|]$ *i.e.* a concatenation of the summation of ϕ_{vp1}^{text} and ϕ_{vp2}^{text} with the absolute difference between the two. This results in a $(2 \times 2,495) + (2 \times 200) + (2 \times 400) = 6,190$ D feature vector ϕ_{vp}^{text} describing (q_{i1}, q_{i2}) . We then train a binary linear SVM to verify whether the two descriptions are describing the same image or not.

3.4.2 Incorporating Visual Common Sense

Our model extends the baseline text-only model (Section 3.4.1) by using an “imagined” scene as an intermediate representation. “Imagining” a scene involves setting values for all of the variables (*e.g.*, presence of objects, their location) that are used to encode scenes. This encoding, along with priors within this abstraction that reason about which scenes are plausible, serve as our representation of visual common sense. This is in contrast with traditional knowledge base representations used to encode common sense via text [136, 97]. Exploring alternative representations of visual common sense is part of future work.

Given a textual description S_i , we generate a scene I_i . We first describe our scoring function that scores the plausibility of the (S_i, I_i) pair. We then (Section 3.4.3) describe our scene generation approach.

Our scoring function

$$\Omega(I_i, S_i) = \Phi(S_i) + \Phi(I_i) + \Psi(I_i, S_i) \quad (3.2)$$

captures textual common sense, visual common sense and text-image correspondence. The textual common sense term $\Phi(S_i) = w^T \phi^{text}(S_i)$ only depends on text and is the same as the text-only baseline model (Section 3.4.1). Of the two new terms, $\Phi(I_i)$ only depends on the scene and captures visual common sense – it evaluates how plausible the scene is (Section 3.4.2). Finally, $\Psi(I_i, S_i)$ depends on both the text description and the scene, and captures how consistent the imagined scene is to the text (Section 3.4.2). We start by describing the representation we use to represent the description and to encode a scene via visual abstractions.

Scene and Description Encoding

The set of clipart in our visual abstraction were described in Section 3.3. More details can be found in [138]. In the generated scenes, we represent an object O_k using its presence $e_k \in \{0, 1\}$, location x_k, y_k , depth z_k (3 discrete scales), horizontal facing direction or orientation $d_k \in \{-1, 1\}$ (left or right) and attributes f_k (poses and expressions for the boy and girl). The sentence descriptions S_i are represented using a set of predicate tuples T_l extracted using semantic roles analysis [99]. A tuple T_l consists of a primary noun A_l , a relation r_l and an optional secondary noun B_l . For example a tuple can be (Jenny, fly, Kite) or (Mike, be angry, N/A). There are 1,133 nouns and 2,379 relations in our datasets. Each primary noun A_l and secondary noun B_l is mapped to 1 of the 58 clipart objects a_l and b_l respectively which have the highest mutual information with it in training data. We found this to work reliably.

Visual Common Sense

We breakdown and introduce the factors in $\Phi(I_i)$ into per-object (unary) factors $\Phi^u(O_k)$ and between-object (pairwise) factors $\Phi^{pw}(O_{k_1}, O_{k_2})$.

$$\Phi(I_i) = \sum_k \Phi^u(O_k) + \sum_{k_1, k_2} \Phi^{pw}(O_{k_1}, O_{k_2}) \quad (3.3)$$

Per-object (unary) factors $\Phi^u(O_k)$ capture presence, location, depth, orientation and attributes. This scoring function will be parameterized by w 's⁵ that are shared across all objects and pairs of objects. Let L be the log probabilities (MLE counts) estimated from training data. For example, $L_e^u(e_k) = \log P(e_k)$, where $P(e_k)$ is the proportion of images in which object O_k exists, and $L_{xyz}^u(x_k, y_k|z_k) = \log P(x_k, y_k|z_k)$, where $P(x_k, y_k|z_k)$ is the proportion of times object O_k is at location (x_k, y_k) given that O_k is at depth z_k .

$$\Phi^u(O_k) = w_e^u L_e^u(e_k) + w_{xyz}^u L_{xyz}^u(x_k, y_k|z_k) + w_z^u L_z^u(z_k) + w_d^u L_d^u(d_k) + w_f^u L_f^u(f_k) \quad (3.4)$$

Between-object (pairwise) factors $\Phi^{pw}(O_{k_1}, O_{k_2})$ capture co-occurrence of objects and their attributes, as well as relative location, depth and orientation.

$$\begin{aligned} \Phi^{pw}(O_{k_1}, O_{k_2}) = & w_e^{pw} L_e^{pw}(e_{k_1}, e_{k_2}) + w_{xyd}^{pw} L_{xyd}^{pw}(dx, dy) + w_z^{pw} L_z^{pw}(z_{k_1}, z_{k_2}) \\ & + w_d^{pw} L_d^{pw}(d_{k_1}, d_{k_2}) + w_f^{pw} L_f^{pw}(f_{k_1}, f_{k_2}) \end{aligned} \quad (3.5)$$

Here the relative x-location is relative to the orientation of the first object *i.e.* $dx = d_{k_1}(x_{k_1} - x_{k_2})$. Relative y-location is $dy = y_{k_1} - y_{k_2}$. These capture where O_{k_2} is from the perspective of O_{k_1} . The space of (x, y, z) is quite large (typical image size is 500 x 400). So to estimate the probabilities reliably, we model the locations with GMMs. In particular, the factor $L_{xyz}^u(x_k, y_k|z_k)$ is over 27 GMM components and $L_{xyd}^{pw}(dx, dy)$ is over 24 GMM components.

Notice that since the parameters are shared across all objects and pairs of objects, so far we have introduced 5 parameters in Equation 3.4 and 5 parameters in Equation 3.5. The corresponding 10 log-likelihood terms can be thought of as features representing visual common sense. The parameters will be learnt to optimize for the FITB (ranking SVM) or VP (binary SVM) tasks similar to the text-only baseline described in Section 3.4.1.

Text-Image Consistency

We now discuss terms in our model that score the consistency between an imagined scene and a textual description. We breakdown and introduce the text-image correspondence factors in $\Psi(I_i, S_i)$ in Equation 3.2 into per-noun factors $\Psi^{n+}(I_i, T_l)$ and per-relation factors

⁵Overloaded notation with parameters learnt for the text-only baseline in Section 3.4.1

$\Psi^{r+}(I_i, T_l)$ for objects that are mentioned in the description, and default per-object factors $\Psi^{u-}(O_k)$ and default between-object factors $\Psi^{pw-}(O_{k_1}, O_{k_2})$ when the respective objects are not mentioned in the description.

$$\Psi(I_i, S_i) = \sum_l \Psi^{n+}(I_i, T_l) + \sum_l \Psi^{r+}(I_i, T_l) + \sum_{k \notin S_i} \Psi^{u-}(O_k) + \sum_{k_1, k_2 \notin S_i} \Psi^{pw-}(O_{k_1}, O_{k_2}) \quad (3.6)$$

The per-noun factors $\Psi^{n+}(I_i, T_l)$ capture object presence conditioned on the nouns (both primary and secondary) in the tuple, and object attributes conditioned on the nouns as well as relations in the tuple. For instance, if the tuple T_l is (Jenny, kicks, ball), these terms reason about the likelihood that cliparts corresponding to Jenny and ball exist in the scene, that Jenny shows a kicking pose, *etc.*. Again, the likelihood of each concept is scored by its log probability in the training data.

$$\Psi^{n+}(I_i, T_l) = w_{abe}^{n+} (L_e^{n+}(e_{a_l}|a_l) + L_e^{n+}(e_{b_l}|b_l)) + w_{arf}^{n+} L_{arf}^{n+}(f_{a_l}|a_l, r_l) + w_{brf}^{n+} L_{brf}^{n+}(f_{b_l}|b_l, r_l) \quad (3.7)$$

The per-relation factors $\Psi^{r+}(I_i, T_l)$ capture relative object location (where is b_l relative to a_l and vice versa), depth and orientation conditioned on the relation. Note that these factors are shared across all objects because “sitting in” in (Jenny, sitting in, sandbox) and (cat, sitting in, sandbox) is expected to have similar visual instantiations.




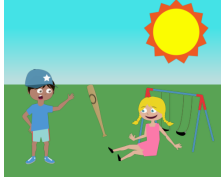

$$\begin{aligned} \Psi^{r+}(I_i, T_l) = & w_{rxyd}^{r+} L_{rxyd}^{r+}(dx, dy|r_l) + w_{rxyd'}^{r+} L_{rxyd'}^{r+}(dx', dy'|r_l) + w_{rz}^{r+} L_{rz}^{r+}(z_{a_l}, z_{b_l}|r_l) \\ & + w_{rd}^{r+} L_{rd}^{r+}(d_{a_l}, d_{b_l}|r_l) \end{aligned} \quad (3.8)$$

Here $dx' = d_{b_l}(x_{b_l} - x_{a_l})$ and $dy' = y_{b_l} - y_{a_l}$ captures where the primary object is relative to the secondary object.

The default per-object factors $\Psi^{u-}(O_k)$ and the default between-object factors $\Psi^{pw-}(O_{k_1}, O_{k_2})$ capture default statistics when an object or a pair of objects is not mentioned in the description. $\Psi^{u-}(O_k)$ captures the default presence and attribute whereas $\Psi^{pw-}(O_{k_1}, O_{k_2})$ captures the default relative location, depth and orientation.

The default factors are object-specific since each object has a different prior depending on its semantic role in scenes. The default factors capture object states conditioned on the object

Figure 3.3: Scenes generated for an example FITB question.

<p style="text-align: center;">Question</p> <div style="border: 1px solid blue; border-radius: 10px; padding: 10px; margin: 10px 0;"> <p>_____. Mike is wearing a blue cap. Mike is telling Jenny to get off the swing</p> </div> <p style="text-align: center;">Answers</p> <div style="border: 1px solid blue; border-radius: 10px; padding: 10px; margin: 10px 0;"> <p>Ground truth: D Vision + text: D Text alone: A</p> </div> <p style="text-align: center;">Original Scene</p> 	<p style="text-align: center;">Options and Generated Scenes</p> <div style="display: flex; justify-content: space-around;"> <div style="width: 45%;"> <p>A. There is a tree near a table.</p>  </div> <div style="width: 45%;"> <p>B. The brown dog is standing next to Mike.</p>  </div> </div> <div style="display: flex; justify-content: space-around;"> <div style="width: 45%;"> <p>C. The sun is in the sky.</p>  </div> <div style="width: 45%;"> <p>D. Jenny is standing dangerously on the swing</p>  </div> </div>
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

not being mentioned in a description. We use notation D instead of L to stress this point. For example $D_e^{u-}(e_k|S_i) = \log P(e_k|k \notin S_i)$, $D_z^{pw-}(z_{k_1}, z_{k_2}|S_i) = \log P(z_{k_1}, z_{k_2}|k_1, k_2 \notin S_i)$.

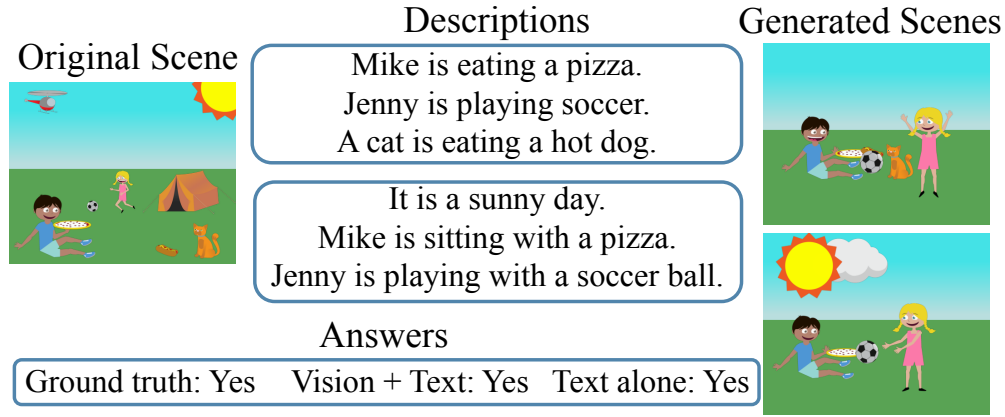
$$\begin{aligned}
 \Psi^{u-}(O_k) &= w_{abe}^{u-} D_{abe}^{u-}(e_k|S_i) + w_{abrf}^{u-} D_{abrf}^{u-}(f_k|S_i) \\
 \Psi^{pw-}(O_{k_1}, O_{k_2}) &= w_{rxyd}^{pw-} D_{rxyd}^{pw-}(dx, dy|S_i) + w_{rz}^{pw-} D_{rz}^{pw-}(z_{k_1}, z_{k_2}|S_i) \\
 &\quad + w_{rd}^{pw-} D_{rd}^{pw-}(d_{k_1}, d_{k_2}|S_i)
 \end{aligned} \tag{3.9}$$

We have now introduced an additional 12 w parameters (total 22) that are to be learnt for the FITB and VP tasks. Notice that this is in stark contrast with the thousands of parameters we learn for the text-only baseline (Section 3.4.1).

3.4.3 Scene Generation

Given an input description, we extract tuples as described earlier in Section 3.4.2. We then use the approach of Zitnick *et al.* [139] trained on our training corpus of clipart images

Figure 3.4: Scenes generated for an example VP question.



and associated descriptions to generate a scene corresponding to the tuples. Briefly, it sets up a Conditional Random Field (CRF) model with a scoring function very similar to $\Phi(I_i) + \Psi(I_i, S_i)$. It samples scenes from this model using Iterative Conditional Modes with different initializations. Details can be found in [139].

3.4.4 Answering Questions with Imagined Scenes

Fill-in-the-blank. For FITB, we generate one scene using each question-answer pair $S_{ij} = (q_i, o_{ij})$. Fig. 3.3 shows qualitative examples of scenes generated for FITB. From the question-answer pair S_{ij} and the generated scenes I_{ij} , we extract features corresponding to our scoring function (Equation 3.2) and use them to learn the ranking SVM (Equation 3.1) to answer FITB questions. We choose the ranking SVM C parameter using 5 fold cross validation.

Visual paraphrasing. For VP we generate one scene for each description $S_{i1} = q_{i1}$ and $S_{i2} = q_{i2}$ in the input pair of descriptions. Fig. 3.4 shows qualitative examples of scenes generated for VP. We capture the difference between the two sentence descriptions by pairing the generated scenes with the *other* description *i.e.* we compute $\Omega(I_{i1}, S_{i2})$ and $\Omega(I_{i2}, S_{i1})$ (Equation 3.2). We extract features for both combinations, concatenate the addition of the features and the absolute difference of the features to make the mapping symmetric. These features are used to train a binary SVM that determines whether the input pair of descriptions are describing the same scene or not. We choose the SVM C parameter using 5

fold cross validation.

3.5 Experiments and Results

3.5.1 Fill-in-the-blank

We present results of our approach on the FITB dataset in Table 3.1. Our approach of “imagining” and joint visual-text reasoning achieves 48.04% accuracy, significantly outperforming the text-only baseline (44.97%) by 3.07% using only 22 extra feature dimensions (compared to 2,486 dimensions of the baseline). This brings the performance closer to human performance at 52.87%. ⁶Leveraging visual common sense does help answering these seemingly purely text-based questions.

By breaking down our 22 parameters (corresponding to visual features) into object presence ($w_e^u, w_e^{pw}, w_{abe}^{n+}, w_{abe}^{u-}, 4D$), attribute ($w_f^u, w_f^{pw}, w_{arf}^{n+}, w_{brf}^{n+}, w_{abrf}^{u-}, 5D$) and spatial configuration ($w_{xyz}^u, w_z^u, w_d^u, w_{xyd}^{pw}, w_z^{pw}, w_d^{pw}, w_{rxyd}^{r+}, w_{rxyd}^{r+}, w_{rz}^{r+}, w_{rd}^{r+}, w_{rxyd}^{pw-}, w_{rz}^{pw-}, w_{rd}^{pw-}, 13D$) categories, we study their individual contribution to FITB performance on top of the text baseline. Object presence contributes the most (47.02%), followed by attribute (46.39%), while spatial information does not help (44.80%). In fact, only using presence and attribute features achieves 48.60%, slightly higher than using all three (including spatial). Visual features alone perform poorly (33.67%), which is expected given the textual nature of the task. But they clearly provide useful complementary information over text. In fact, text-alone (baseline), vision+text (our approach) and humans all seem to make complementary errors. Between text-alone and vision+text, 54.68% of the questions are correctly answered by at least one of them. And between text-alone, vision+text and human, 75.92% of the questions are correctly answered.

Our model is capable of imagining scenes that may contain more objects than the ones mentioned in text. Our model when using only presence does 47.02%, while a visual common sense agnostic model that only infers objects mentioned in the tuples (a_l and b_l) does 46.62%. This further demonstrates the need for visual common sense based imagination, and not treating the text at face value. If the ground truth scenes are available at test time, the performance of our approach reaches 78.04%, while humans are at 94.43%.

⁶Bootstrapping experiments show that the mean bootstrapping (100 rounds) performance of visual+text $46.33\% \pm 0.14\%$ is statistically significantly better than that of text $43.65\% \pm 0.15\%$.

Table 3.1: Fill-in-the-blank performance of different approaches.

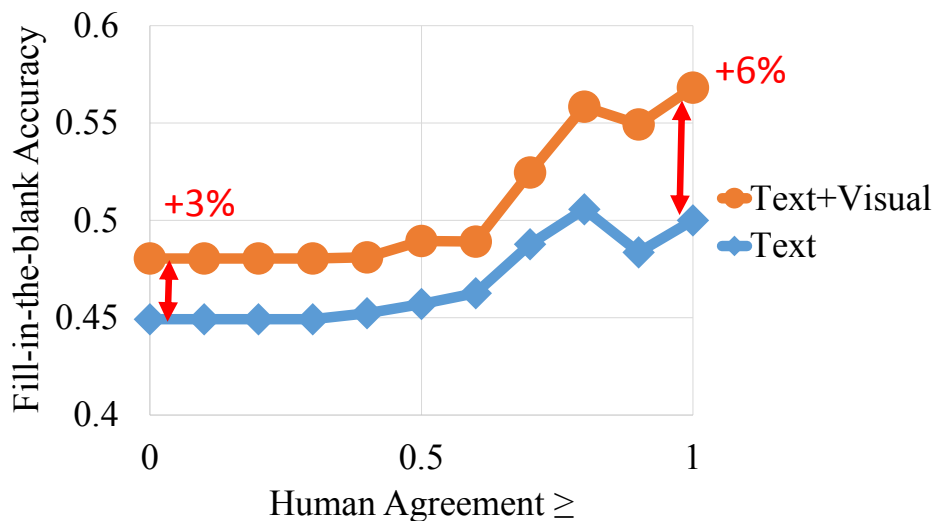
Approach	Fill-in-the-blank Accuracy(%)
Random	25.00
Text baseline	44.97
Visual	33.67
Text + visual (presence)	47.02
Text + visual (attribute)	46.39
Text + visual (spatial)	44.80
Text + visual (presence,attribute)	48.60
Text + visual (all)	48.04
Human Mode	52.87

In addition to predicting ground truth, we also study how well our approach can mimic human responses. Our approach matches the human majority vote (mode) response 39.35% of the times (text alone: 36.40%). When re-trained using the human mode as the labels, the performance increases to 45.43%. The text-only baseline method does 42.25%. These results suggest that mimicking human is a more challenging task (text-only was at 44.97% when training on and predicting ground truth). Note that visual common sense is also useful when mimicking humans.

We also study how the performance of our approach varies based on the difficulty of the questions. We consider questions to be easy if humans agree on the response. We report performance of the text baseline and our model on subsets of the FITB test set where at least K people agreed with the mode. Fig. 3.5 shows performance as we vary K . On questions with higher human agreement, the visual approach outperforms the baseline by a larger margin.

Qualitative results for FITB are presented in Appendix B.1.

Figure 3.5: FITB performance on subsets of the test data with varying amounts of human agreement. The margin of improvement of our approach over the baseline increases from 3% on all questions to 6% on questions with high human agreement.



3.5.2 Visual Paraphrasing

We present results of our approach on the VP dataset in Table 3.2. Our approach of generating and reasoning with scenes does 1.4% better than reasoning only with text⁷. In this task, the performance of the text-based approach is already close to human, while vision pushes it even further to above human performance⁸.

Similar to the FITB task, we break down the contribution of visual features into object presence, attribute and spatial configuration categories. Presence shows the most contribution (0.93%). Spatial configuration features also help (by 0.60%) in contrast to FITB. See Table 3.2.

In VP, a naive scene generation model that only imagines objects that are mentioned in the description does 95.01% which is close to 95.08% where extra objects are inferred. We hypothesize that the VP task is qualitatively different from FITB. In VP, important objects that are relevant to semantic differences between sentences tend to be mentioned in the sentences. What remains is to reason about the attributes and spatial configurations of the

⁷Bootstrapping text+visual $95.11\% \pm 0.02\%$, text $93.62\% \pm 0.02\%$.

⁸Likely due to noise on MTurk.

Table 3.2: Visual paraphrasing performance of different approaches.

Approach	Visual Paraphrasing Average Precision(%)
Random	33.33
Text baseline	94.15
Visual	91.25
Text + visual (presence)	95.08
Text + visual (attribute)	94.54
Text + visual (spatial)	94.75
Text + visual (presence,attribute)	95.47
Text + visual (all)	95.55
Human Average	94.78

objects. In FITB, on the other hand, inferring the unwritten objects is critical to identify the best way to complete the description. Qualitative results are presented in Appendix B.1.

The VP task can be made more challenging by sampling pairs of descriptions that describe semantically similar scenes in the Abstract Scenes dataset [138]. The 10,020 scenes in the Abstract Scenes Dataset are generated from 1,002 sentences. For each of the 1,002 sentences 10 different people drew 10 scenes. And then a new set of workers described each of the 10 scenes (10,020 total). Scenes that are generated from the same sentence belong to the same semantic class, and therefore their sentence descriptions have similar semantic meanings.

We study coarse-grained and fine-grained visual paraphrasing problems. In the coarse-grained visual paraphrasing problem, the objective is to tell sentences describing one semantic class from another. In the fine-grained visual paraphrasing problem, the objective is to tell sentences describing the same semantic class from each other. Results are summarized in Table 3.3. In both coarse-grained and fine-grained visual paraphrasing settings, our approach using both textual features and visual imagination show improvements on top of only using text features.

We would like to stress that FITB and VP are purely textual tasks as far as the input modality is concerned. The visual cues that we incorporate are entirely “imagined”. Our results clearly demonstrate that a machine that imagines and uses visual common sense performs better at these tasks than a machine that does not.

Table 3.3: Coarse and fine-grained visual paraphrasing. In both coarse- and fine-grained settings, our approach using visual features show improvements on top of the text-only baseline.

	Source of positive pairs of sentences	Source of negative pairs of sentences	Random	Text only	Text + Visual	Visual improvement
Original (in main paper)	Same scene	Difference scenes	33.33	94.15	95.50	+1.40
Coarse-grained	Different scenes in the same semantic class	Scenes from different semantic classes	33.33	84.19	86.15	+1.96
Fine-grained	Same scene	Different scenes in the same semantic class	33.33	54.79	56.43	+1.64

3.6 Discussion

Leveraging visual knowledge to solve non-visual tasks may seem counter-intuitive. Indeed, with sufficient training data, one may be able to learn a sufficiently rich text-based model. However in practice, good intermediate representations provide benefits. This is the role that parts and attributes have played in recognition [69, 35, 131]. In this work, the imagined scenes form this intermediate representation that allows us to encode visual common sense.

In this work, we choose clipart scenes as our modality to “imagine” the scene and harness the power of visual common sense. This is analogous to works on physical reasoning that use physics to simulate physical processes [48]. These are both qualitatively different from traditional knowledge bases [16, 136], where relations between instances are explicitly represented and used during inference. Humans cannot always verbalize their reasoning process. Hence, using non-explicit representations of common sense has some appeal. Of course, alternate approaches, including more explicit representations of visual common sense are worth investigating.

Instead of generating one scene per text description, a direction to better capture the uncertainty in imagination might be to generating multiple diverse scenes [4]. Also, our approach learns the scene generation model and textual reasoning models in two separate stages, both as a practical choice and to reduce overfitting. With recent advances in the end-to-end learning technique in deep learning, one could envision a system that learns the scene generation model and the textual reasoning model jointly.

If there will be an oracle providing the ground truth scenes, such scenes would reliably help perform scene description tasks such as FITB and VP. But given a scene description,

different people will imagine different scenes, which may all be different from the ground truth scene. It is desirable to know how much the imagined scenes can realistically improve answering textual questions. Getting humans to draw imagined scenes for the scene descriptions however, is costly. But a high-performance automatic approaches which generate realistic-looking images might be able to provide a good estimate.

Recent deep generative models such as conditional Generative Adversarial Network [88] and conditional Variational Autoencoder [60, 112] have started generating realistic-looking images. It is a promising future direction to use deep learning to improve the performance of the imagination module. While deep generative models still take time to mature, deep classification models have made great significant progress over shallow models on image classification, object detection, scene classification and even answering open-ended questions about images. To make use of the power of such deep classification models, in Chapter 4 we propose an approach that answers questions about images and scene descriptions as features to leverage commonsense knowledge in the Visual Question Answering [1, 41, 42, 44, 82, 101] corpora.

Acknowledgments

We thank Stanislaw Antol for discussions. This work was supported in part by a Google Faculty Research Award and The Paul G. Allen Family Foundation Allen Distinguished Investigator award to Devi Parikh. We thank Larry Zitnick for helpful discussions and his code.

Chapter 4

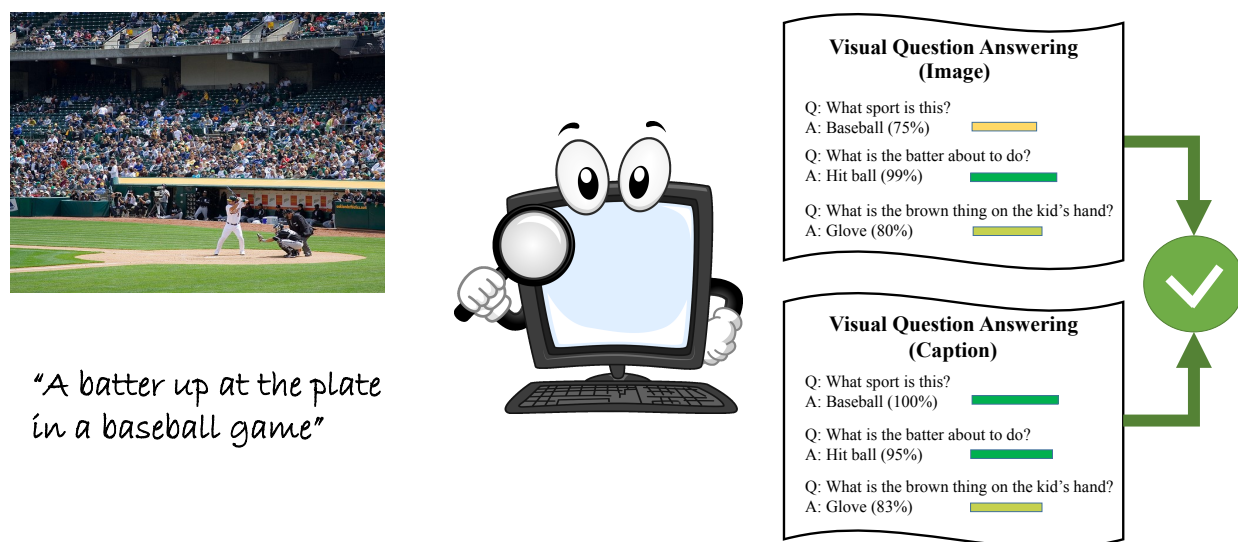
Leveraging Visual Question Answering for Image-Caption Ranking

4.1 Introduction

Visual Question Answering (VQA) is an “AI-complete” problem that requires knowledge from multiple disciplines such as computer vision, natural language processing and knowledge base reasoning. A VQA system takes as input an image and a free-form open-ended question about the image and outputs the natural language answer to the question. A VQA system needs to not only recognize objects and scenes but also reason beyond low-level recognition about aspects such as intention, future, physics, material and commonsense knowledge. For example (*Q*: Who is the person in charge in this picture? *A*: Chef) reveals the most important person and occupation in the image. Moreover, answers to multiple questions about the same image can be correlated and may reveal more complex interactions. For example (*Q*: What is this person riding? *A*: Motorcycle) and (*Q*: What is the man wearing on his head? *A*: Helmet) might reveal correlations observable in the visual world due to safety regulations.

X. Lin and D. Parikh. Leveraging Visual Question Answering for Image-Caption Ranking. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, 2016. ©Springer International Publishing AG 2016, with permission of Springer.

Figure 4.1: Aligning images and captions requires high-level reasoning *e.g.*, “a batter up at the plate” would imply that a player is holding a bat, posing to hit the baseball and there might be another player nearby waiting to catch the ball. There is rich knowledge in Visual Question Answering (VQA) corpora containing human-provided answers to a variety of questions one could ask about images. We propose to leverage knowledge in VQA by using VQA models learnt on images and captions as “feature extraction” modules for image-caption ranking.



Today’s VQA models, while far from perfect, may already be picking up on these semantic correlations of the world. If so, they may serve as an implicit knowledge resource to help other tasks. Just like we do not need to fully understand the theory behind an equation to use it, can we already use VQA knowledge captured by existing VQA models to improve other tasks?

In this work we study the problem of using VQA knowledge to improve image-caption ranking. Consider the image and its caption in Figure 4.1. Aligning them not only requires recognizing the batter and that it is a baseball game (mentioned in the caption), but also realizing that a batter up at the plate would imply that a player is holding a bat, posing to hit the baseball and there might be another player nearby waiting to catch the ball (seen in the image). Image captions tend to be generic. As a result, image captioning corpora may not capture sufficient details for models to infer this knowledge.

Fortunately VQA models try to explicitly learn such knowledge from a corpus of images,

each with associated questions and answers. Questions about images tend to be much more specific and detailed than captions. The VQA dataset of [1] in particular has a collection of free-form open-ended questions and answers provided by humans. These images also have associated captions [73].

We propose to leverage VQA knowledge captured by such corpora for image-caption ranking by using VQA models learnt on images and captions as “feature extraction” schemes to represent images and captions. Given an image and a caption, we choose a set of free-form open-ended questions and use VQA models learnt on images and captions to assess probabilities of their answers. We use these probabilities as image and caption features respectively. In other words, we embed images and captions into the space of VQA questions and answers using VQA models. Such VQA-grounded representations interpret images and captions from a variety of different perspectives and imagine beyond low-level recognition to better understand images and captions.

We propose two approaches that incorporate these VQA-grounded representations into an existing state-of-the-art¹ VQA-agnostic image-caption ranking model [61]: fusing their predictions and fusing their representations. We show that such VQA-aware models significantly outperform the VQA-agnostic model and set state-of-the-art performance on MSCOCO image-caption ranking. Specifically, we improve caption retrieval by 7.1% and image retrieval by 4.4%.

This paper is organized as follows: Section 4.2 introduces related works. We first introduce VQA and image-caption ranking tasks as our building blocks in Section 4.3, then detail our VQA-based image-caption ranking models in Section 4.4. Experiments and results are reported in Section 4.5. We conclude in Section 4.7.

4.2 Related Work

Visual Question Answering. Visual Question Answering (VQA) [1] is the task of taking an image and a free-form open-ended question about the image and automatically predicting the natural language answer to the question. VQA may require fine-grained recognition, object detection, activity recognition, multimodal and commonsense knowledge. Large datasets [81, 101, 129, 41, 1] have been made available to cover the diversity of knowledge

¹To the best of our knowledge on MSCOCO [73], [61] has the state-of-the-art caption retrieval performance. [78] has the state-of-the-art image retrieval performance.

required for VQA. Most notably the VQA dataset [1] contains 614,163 questions and ground truth answers on 204,721 images of the MSCOCO [73] dataset.

Recent VQA models [82, 101, 41, 133, 1, 78] explore state-of-the-art deep learning techniques combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). [1] also explores a slight variant of VQA that answers a question about the image by reading a caption describing the image instead of looking at the image itself. We call this variant VQA-Caption.

VQA is a challenging task in its early stages. In this work we propose to use both VQA and VQA-Caption models as implicit knowledge resources. We show that current VQA models, while far from perfect, can already be used to improve other multimodal AI tasks; specifically image-caption ranking.

Semantic mid-level visual representations. Previous works have explored the use of attributes [33, 10, 125], parts [6, 130], poselets [9, 131], objects [71], actions [104] and contextual information [47, 118, 26] as semantic mid-level representations for visual recognition. Benefits of using such semantic mid-level visual representations include improving fine-grained visual recognition, learning models of visual concepts without example images (zero-shot learning [69, 92]) and improving human-machine communication where a user can explain the target concept during image search [68, 63], or give a classifier an explanation of labels [27, 93]. Recent works also explore using word embeddings [111] and free-form text [30] as representations for zero-shot learning of new object categories. [56] proposes scene graphs for image retrieval. [2] proposes using abstract scenes as an intermediate representation for zero-shot action recognition. Closest to our work is the use of objects, actions, scenes [34], attributes and object interactions [67] for generating and ranking image captions. In this work we propose to use free-form open-ended questions and answers as mid-level representations and we show that they provide rich interpretations of images and captions.

Commonsense knowledge for visual reasoning. Recently there has been a surge of interest in visual reasoning tasks that require high-level reasoning such as physical reasoning [48, 132], future prediction [38, 124, 97], object affordance prediction [136] and textual tasks that require visual knowledge [74, 122, 105]. Such tasks can often benefit from reasoning with external commonsense knowledge resources. [137] uses a knowledge base learnt on object categories, attributes, actions and object affordances for query-based image retrieval. [123] learns to anticipate future scenes from watching videos for action and object forecasting. [74] learns to imagine abstract scenes from text for textual tasks that need

visual understanding. [122, 105] evaluate the plausibility of commonsense assertions by verifying them on collections of abstract scenes and real images, respectively, to leverage the visual common sense in those collections. Our work explores the use of VQA corpora which have both visual (images) and textual (captions) commonsense knowledge for image-caption ranking.

Images and captions. Recent works [58, 15, 61, 127, 83, 79] have made significant progress on automatic image caption generation and ranking by applying deep learning techniques for image recognition [66, 109, 117] and language modeling [17, 116] on large datasets [23, 73]. Algorithms can now often generate accurate, human-like natural-language captions for images. However, evaluating the quality of such automatically generated open-ended image captions is still an open research problem [31, 121].

On the other hand, ranking images given captions and ranking captions given images require a similar level of image and language understanding, but are amenable to automatic evaluation metrics. Recent works on image-caption ranking mainly focus on improving model architectures. [61, 83] study different architectures for projecting CNN image representations and RNN caption representations into a common multimodal space. [79] uses multimodal CNNs for image-caption ranking. [58] aligns image and caption fragments using CNNs and RNNs. Our work takes an orthogonal approach to previous works. We propose to leverage knowledge in VQA corpora containing questions about images and associated answers for image-caption ranking. Our proposed VQA-based image and caption representations provide complementary information to those learnt using previous approaches on a large image-caption ranking dataset.

4.3 Building Blocks: Image-Caption Ranking and VQA

In this section we present image-caption ranking and VQA modules that we build on top of.

4.3.1 Image-Caption Ranking

The image-caption ranking task is to retrieve relevant images given a query caption, and relevant captions given a query image. During training we are given image-caption pairs (I, C) that each corresponds to an image I and its caption C . For each pair we sample $K - 1$ other images in addition to I so the image retrieval task becomes retrieving I from K images

$I_i, i = 1, 2 \dots K$ given caption C . We also sample $K - 1$ random captions in addition to C so the caption retrieval task becomes retrieving C from K captions $C_i, i = 1, 2 \dots K$ given image I .

Our image-caption ranking models learn a ranking scoring function $S(I, C)$ such that the corresponding retrieval probabilities:

$$P_{im}(I|C) = \frac{\exp(S(I, C))}{\sum_{i=1}^K \exp(S(I_i, C))} \quad P_{cap}(C|I) = \frac{\exp(S(I, C))}{\sum_{i=1}^K \exp(S(I, C_i))} \quad (4.1)$$

are maximized. Let $S(I, C)$ be parameterized by θ (to be learnt). We formulate an objective function $L(\theta)$ for $S(I, C)$ as the sum of expected negative log-likelihoods of image and caption retrieval over all image-caption pairs (I, C) :

$$L(\theta) = \mathbb{E}_{(I, C)}[-\log P_{im}(I|C)] + \mathbb{E}_{(I, C)}[-\log P_{cap}(C|I)] \quad (4.2)$$

Recent works on image-caption ranking often construct $S(I, C)$ by combining a vectorized image representation which is usually hidden layer activations in a CNN pretrained for image classification, with a vectorized caption representation which is usually a sentence encoding computed using an RNN in a multimodal space. Such scoring functions rely on large image-caption ranking datasets to learn knowledge necessary for image-caption ranking and do not leverage knowledge in VQA corpora. We call such models VQA-agnostic models.

In this work we use the publicly available state-of-the-art image-caption ranking model of [61] as our baseline VQA-agnostic model. [61] projects a D_{x_I} -dimensional CNN activation x_I for image I and a D_{x_C} -dimensional RNN latent encoding x_C for caption C to the same D_{x_C} -dimensional common multimodal embedding space as unit-norm vectors t_I and t_C :

$$t_I = \frac{W_I x_I}{\|W_I x_I\|_2} \quad t_C = \frac{x_C}{\|x_C\|_2} \quad (4.3)$$

The multimodal scoring function is defined as their dot product $S_t(I, C) = \langle t_I, t_C \rangle$.

The VQA-agnostic model of [61] uses the 19-layer VGGNet [109] ($D_{x_I} = 4096$) for image encoding and an RNN with 1024 Gated Recurrent Units [17] ($D_{x_C} = 1024$) for caption encoding. The RNN and parameters W_I are jointly learnt on the image-caption ranking training set using a margin-based objective function.

4.3.2 VQA

VQA is the task of given an image I and a free-form open-ended question Q about I , generating a natural language answer A to that question. Similarly, VQA-Caption task proposed by [1] takes a caption C of an image and a question Q about the image, then generates an answer A . In [1] the generated answers are evaluated using $\min(\frac{\# \text{ humans that provided } A}{3}, 1)$. That is, A is 100% correct if at least 3 humans (out of 10) provide the answer A .

We closely follow [1] and formulate VQA as a classification task over top $M = 1000$ most frequent answers from the training set. The oracle accuracies of picking the best answer for each question within this set of answers are 89.37% on training and 88.83% on validation. During training, given triplets of image I , question Q and ground truth answer A , we optimize the negative log-likelihood (NLL) loss to maximize the probability of the ground truth answer $P_I(A|Q, I)$ given by the VQA model. Similarly given triplets of caption C , question Q and ground truth answer A , we optimize the NLL loss to maximize the VQA-Caption model probability $P_C(A|Q, C)$.

Following [1], for a VQA question (I, Q) we first encode the input image I using the 19-layer VGGNet [109] as a 4,096-dimensional image encoding x_I , and encode the question Q using a 2-layer RNN with 512 Long Short-Term Memory (LSTM) units [51] per layer as a 2,048-dimensional question encoding x_Q . We then project x_I and x_Q into a common 1,024-dimensional multimodal space as z_I and z_Q :

$$z_I = \text{Tanh}(W_I x_I + b_I) \quad z_Q = \text{Tanh}(W_Q x_Q + b_Q) \quad (4.4)$$

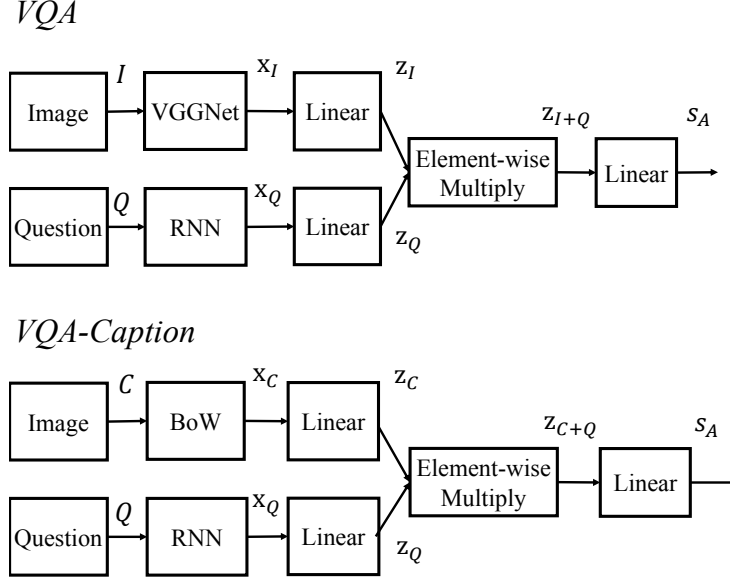
As in [1] we then compute the representation z_{I+Q} for the image-question pair (I, Q) by element-wise multiplying z_I and z_Q : $z_{I+Q} = z_I \odot z_Q$. The scores s_A for 1,000 answers are given by:

$$s_A = W_s z_{I+Q} + b_s \quad (4.5)$$

We jointly learn the question encoding RNN and parameters $\{W_I, b_I, W_Q, b_Q, W_s, b_s\}$ during training.

For the VQA-Caption task given caption C and question Q , we use the same network architecture and learning procedure as above, but using the most frequent 1,000 words in training captions as the dictionary to construct a 1,000 dimensional bag-of-words encoding for caption C as x_C to replace the image feature x_I and compute z_C, z_{C+Q} respectively. Figure 4.2

Figure 4.2: Our VQA and VQA-Caption network architectures.



illustrates the network architectures of our VQA and VQA-Caption models.

The VQA and VQA-Caption models are learnt on the train split of the VQA dataset [1] using 82,783 images, 413,915 captions and 248,349 questions. These models achieve VQA validation set accuracies of 54.42% (VQA) and 56.28% (VQA-Caption), respectively. Next, they are used as sub-modules in our image-caption ranking approach.

4.4 Approach

To leverage knowledge in VQA for image-caption ranking, we propose to represent the images and the captions in the VQA space using VQA and VQA-Caption models. We call such representations VQA-grounded representations.

4.4.1 VQA-Grounded Representations

Let's say we have a VQA model $P_I(A|Q, I)$, a VQA-Caption model $P_C(A|Q, C)$ and a set of N questions Q_i and their plausible answers (one for each question) A_i , $i = 1, 2, \dots, N$.

Figure 4.3: Images and captions sorted by $P_I(A|Q, I)$ and $P_C(A|Q, C)$ assessed by our VQA (top) and VQA-Caption (bottom) models respectively. Indeed, images and captions that are more plausible for the (Q, A) pairs are scored higher.



Then given an image I and a caption C , we first extract the N dimensional VQA-grounded activation vectors u_I for I and u_C for C such that each dimension i of u_I and u_C is the log probability of the ground truth answer A_i given a question Q_i .

$$u_I^{(i)} = \log P_I(A_i|Q_i, I) \quad u_C^{(i)} = \log P_C(A_i|Q_i, C), i = 1, 2, \dots, N \quad (4.6)$$

For example if the (Q_i, A_i) pairs are $(Q_1: \text{What is the person riding?}, A_1: \text{Motorcycle})$ and $(Q_2: \text{What is the man wearing on his head?}, A_2: \text{Helmet})$, $u_I^{(1)}$ and $u_C^{(1)}$ verify if the person in image I and caption C respectively is riding a motorcycle. At the same time $u_I^{(2)}$ and $u_C^{(2)}$ verify whether the man in I and C is wearing a helmet. Figure 4.1 shows another example.

In cases where there is not a man in the image or the caption, *i.e.* the assumption of Q_i

is not met, $P_I(A_i|Q_i, I)$ and $P_C(A_i|Q_i, C)$ may still reflect if there *were* a man or if the assumption of Q_i *were* fulfilled, could he be wearing a helmet. In other words, even if there is no person present in the image or mentioned in the caption, the model may still assess the plausibility of a man wearing a helmet or a motorcycle being present. This imagination beyond what is depicted in the image or caption can be helpful in providing additional information when reasoning about the compatibility between an image and a caption. We show qualitative examples of this imagination or plausibility assessment for selected (Q, A) pairs in Figure 4.3 where we sort images and captions based on $P_I(A|Q, I)$ and $P_C(A|Q, C)$. Indeed, scenes where the corresponding fact (Q, A) (e.g., man is wearing a helmet) is more likely to be plausible are scored higher.²

Based on the activation vectors u_I and u_C , we then compute the VQA-grounded vector representations v_I and v_C for I and C by projecting u_I and u_C to a D_u -dimensional vector embedding space:

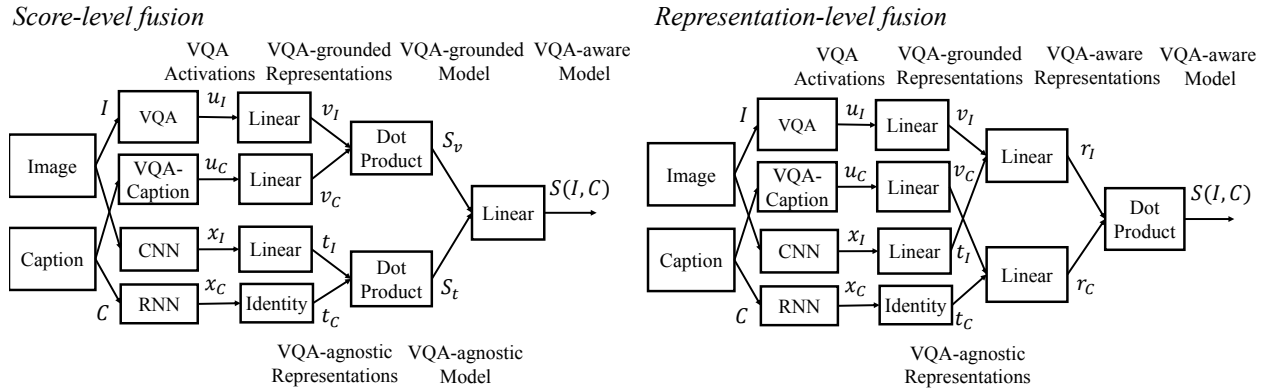
$$v_I = \sigma(W_{u_I}u_I + b_{v_I}) \quad v_C = \sigma(W_{u_C}u_C + b_{v_C}) \quad (4.7)$$

Here σ is a non-linear activation function.

By verifying question-answer pairs on image I and caption C and computing vector representations on top of them, the VQA-grounded representations v_I and v_C explicitly project image and caption into VQA space to utilize knowledge in the VQA corpora. However, that comes at a cost of losing information such as the sentence structure of the caption and image saliency. These information can also be important for image-caption ranking. As a result, We find VQA-grounded representations are most effective when they are combined with baseline VQA-agnostic models, so we propose two strategies for fusing VQA-grounded representations with baseline VQA-agnostic models: combining their prediction scores or score-level fusion (Figure 4.4 left) and combining their representations or representation-level fusion (Figure 4.4 right).

²Nonetheless, checking if a question applies to the target image and caption is also desirable. Contemporary work [100] has looked at modeling $P(Q|I)$, and can be incorporated in our approach as an additional feature.

Figure 4.4: We propose score-level fusion (left) and representation-level fusion (right) to utilize VQA for image-caption ranking. They use VQA and VQA-Caption models as “feature extraction” schemes for images and captions and use those features to construct VQA-grounded representations. The score-level fusion approach combines the scoring functions of a VQA-grounded model and a baseline VQA-agnostic model. The representation-level fusion approach combines VQA-grounded representations and VQA-agnostic representations to produce a VQA-aware scoring function.



4.4.2 Score-level Fusion

A simple strategy to combine our VQA-grounded model with a VQA-agnostic image-ranking model is to combine them at the score level. Given image I and caption C , we first compute the VQA-grounded score as the dot product between the VQA-grounded representations of image and caption $S_v(I, C) = \langle v_I, v_C \rangle$. We then combine it with the VQA-agnostic scoring function $S_t(I, C)$ to get the final scoring function $S(I, C)$:

$$S(I, C) = \alpha S_t(I, C) + \beta S_v(I, C) \quad (4.8)$$

We first learn $\{W_{u_I}, b_{u_I}, W_{u_C}, b_{u_C}\}$ on the image-caption ranking training set, and then learn α and β on a held out validation set to avoid overfitting.

4.4.3 Representation-level Fusion

An alternative to combining the VQA-agnostic and VQA-grounded representations at the score level is to inject the VQA-grounding at the representation level. Given the VQA-agnostic D_t -dimensional image and caption representations t_I and t_C used by the baseline model, we first compute the VQA-grounded representations v_I for image and v_C for caption introduced in Section 4.4.1. And then they are combined with VQA-agnostic representations to produce VQA-aware representations r_I for image I and r_C for caption C by projecting them to a D_r -dimensional multimodal embedding space as follows:

$$r_I = \sigma(W_{t_I}t_I + W_{v_I}v_I + b_{r_I}) \quad r_C = \sigma(W_{t_C}t_C + W_{v_C}v_C + b_{r_C}) \quad (4.9)$$

The final image-caption ranking score is then

$$S(I, C) = \langle r_I, r_C \rangle \quad (4.10)$$

In experiments, we jointly learn $\{W_{u_I}, b_{u_I}, W_{u_C}, b_{u_C}\}$ (for projecting u_I and u_C to the VQA-grounded representations v_I, v_C) with $\{W_{t_I}, W_{v_I}, b_{r_I}, W_{t_C}, W_{v_C}, b_{r_C}\}$ (for computing the combined VQA-aware representations r_I and r_C) on the image-caption ranking training set by optimizing Eq. 4.2.

Score-level fusion and representation-level fusion models are implemented as multi-layer neural networks. All activation functions σ are $ReLU(x) = \max(x, 0)$ (for speed) and dropout layers [114] are inserted after all $ReLU$ layers to avoid overfitting. We set the dimensions of the multimodal embedding spaces D_v and D_r to 4,096 so they are large enough to capture necessary concepts for image-caption ranking. Optimization hyperparameters are selected on the validation set. We optimize both models using RMSProp with batch size 1,000 at learning rate 1e-5 for score-level fusion and 1e-4 for representation-level fusion. Optimization runs for 100,000 iterations with learning rate decay every 50,000 iterations.

Our main results in Section 4.5.1 use $N = 3000$ question-answer pairs, sampled 3 questions per image with their ground truth answers with respect to their original images from 1,000 random VQA training images. We discuss using different numbers of question-answer pairs N and different strategies for selecting the question-answer pairs in Section 4.5.4.

4.5 Experiments and Results

We report results on MSCOCO [73] which is the largest available image-caption ranking dataset. Following the splits of [58, 61] we use all 82,783 MSCOCO train images with 5 captions per image as our train set, 413,915 image-caption pairs in total. Note that this is the same split as the train split in the VQA dataset [1] we used to train our VQA and VQA-Caption models. The validation set consists of 1,000 images sampled from the original MSCOCO validation images. The test set consists of 5,000 images sampled from the original MSCOCO validation images that were not in the image-caption ranking validation set. Same as the train set, there are 5 captions available for each validation and test image.

We follow the evaluation metric of [58] and report caption and image retrieval performances on the first 1,000 test images following [58, 62, 83, 78, 61]. Given a test image, the caption retrieval task is to find any 1 out of its 5 captions from all 5,000 test captions. Given a test caption, the image retrieval task is to find its original image from all 1,000 test images. We report $\text{recall}@ (1, 5, 10)$: the fraction of times a correct item was found among the top (1, 5, 10) predictions.

4.5.1 Image-Caption Ranking Results

Table 4.1 shows our main results on MSCOCO. Our score-level fusion VQA-aware model using $N = 3000$ question-answer pairs (“ $N = 3000$ score-level fusion VQA-aware”) achieves 46.9% caption retrieval $\text{recall}@1$ and 35.8% image retrieval $\text{recall}@1$. This model shows an improvement of 3.5% caption and 4.8% image retrieval $\text{recall}@1$ over the state-of-the-art VQA-agnostic model of [61].

Our representation-level fusion approach adds an additional layer on top of the VQA-agnostic representations, resulting in a deeper model, so we experiment with adding an additional layer to the VQA-agnostic model for a fair comparison. That is equivalent to representation-level fusion using $N = 0$ question-answer pair (“ $N = 0$ representation-level fusion”, *i.e.* deeper VQA-agnostic). Comparing with the VQA-agnostic model of [61], adding this additional layer improves performance by 2.4% caption and 2.6% image retrieval $\text{recall}@1$.

By leveraging VQA knowledge our “ $N = 3000$ representation-level fusion VQA-aware” model achieves 50.5% caption retrieval $\text{recall}@1$ and 37.0% image retrieval $\text{recall}@1$, which further improves 4.7% and 3.4% over the $N = 0$ VQA-agnostic representation-level fusion model.

Table 4.1: Caption retrieval and image retrieval performances of our models compared to baseline models on MSCOCO image-caption ranking test set. Powered by knowledge in VQA corpora, both our score-level fusion and representation-level fusion VQA-aware approaches outperform state-of-the-art VQA-agnostic models by a large margin.

MSCOCO						
Approach	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.1	0.5	1.0	0.1	0.5	1.0
DVSA [58]	38.4	69.9	80.5	27.4	60.2	74.8
FV (GMM+HGLMM) [62]	39.4	67.9	80.9	25.1	59.8	76.6
m -RNN-vgg [83]	41.0	73.0	83.5	29.0	42.2	77.0
m -CNN _{ENS} [78]	42.8	73.1	84.1	32.6	68.6	82.8
Kiros <i>et al.</i> [61] (VQA-agnostic)	43.4	75.7	85.8	31.0	66.7	79.9
N=3000 score-level fusion VQA-grounded only	37.0	67.9	79.4	26.2	60.1	74.3
N=3000 score-level fusion VQA-aware	46.9	78.6	88.9	35.8	70.3	83.6
N=0 representation-level fusion VQA-agnostic	45.8	76.8	86.1	33.6	67.8	81.0
N=3000 representation-level fusion VQA-aware	50.5	80.1	89.7	37.0	70.9	82.9

These improvements are consistent with our score-level fusion approach so this shows that the VQA corpora consistently provide complementary information to image-caption ranking.

To the best of our knowledge, the $N = 3000$ representation-level fusion VQA-aware result is the best result on MSCOCO image-caption ranking and significantly surpasses previous best results by as much as 7.1% in caption retrieval and 4.4% image retrieval recall@1.

Our VQA-grounded model alone (“ $N = 3000$ score-level fusion VQA-grounded only”) achieves 37.0% caption and 26.2% image retrieval recall@1. This indicates that the VQA activations u_I and u_C which evaluate the plausibility of facts (question-answer pairs) in images and captions are informative representations.

Figure 4.5 shows qualitative results on image retrieval comparing our approach ($N = 3000$ score-level fusion) with the VQA-agnostic model. By looking at several top retrieved images from our model for the failure case (last column), we find that our model seems to have picked up on a correlation between bats and helmets. It seems to be looking for helmets in

Figure 4.5: Qualitative image retrieval results of our score-level fusion VQA-aware model (middle) and the VQA-agnostic model (bottom). The true target image is highlighted (green if VQA-aware found it, red if VQA-agnostic found it but VQA-aware did not).



retrieved images, while the ground truth image does not have one. Additional qualitative examples are available in Appendix C.1

We also experiment with using the hidden activations available in the VQA and VQA-Caption models (z_I and z_C in Section 4.3.2) as image and caption encodings in place of the VQA activations (u_I and u_C in Section 4.4.1). Using these hidden activations of the VQA models is conceptually similar to using the hidden activations of CNNs pretrained on ImageNet as features [28]. These features achieve 46.8% caption retrieval recall@1 and 35.2% image retrieval recall@1 for score-level fusion, and 49.3% caption retrieval recall@1 and 37.9% image retrieval recall@1 for representation-level fusion which are as good as our semantic features u_I and u_C . This shows that our semantically meaningful features, u_I and u_C , performs as well as their corresponding non-semantic representations z_I and z_C using both score-level fusion and representation-level fusion. Note that such hidden activations may not always be available in different VQA models and the semantic features have the added benefit of being interpretable (*e.g.*, Figure 4.3).

In addition to using 1,000 images for testing, we also report results on MSCOCO using all 5,000 test images following the protocol of [58] in Table 4.2. Retrieving from 5,000 test

Table 4.2: Results on MSCOCO using all 5,000 test images

MSCOCO 5K test images						
Approach	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.1	0.5	1.0	0.1	0.5	1.0
DVSA [58]	16.5	39.2	52.0	10.7	29.6	42.2
FV (GMM+HGLMM) [62]	17.3	39.0	50.2	10.8	28.3	40.1
Kiros <i>et al.</i> [61] (VQA-agnostic)	18.1	43.5	56.8	12.7	34.0	47.3
N=3000 score-level fusion VQA-grounded only	15.7	37.9	50.3	11.0	29.5	42.0
N=3000 score-level fusion VQA-aware	22.8	49.8	63.0	15.5	39.1	52.6
N=0 representation-level fusion VQA-agnostic	20.6	47.1	60.3	14.9	37.8	50.9
N=3000 representation-level fusion VQA-aware	23.5	50.7	63.6	16.7	40.5	53.8

images is more challenging than retrieving from 1,000 test images so the performances of all models are lower. However, the trends are consistent with results on 1,000 test images reported in the main paper. Our score-level fusion model achieves 22.8% caption retrieval R@1 and 15.5% image retrieval R@1, outperforming the VQA-agnostic model by 4.7% and 2.8%. Our representation-level fusion model achieves 23.5% caption retrieval R@1 and 16.7% image retrieval R@1.

4.5.2 Ablation Study

As an ablation study, we compare the following four models: 1) full representation-level fusion: our full $N = 3000$ representation-level fusion model that includes both image and caption VQA representations; 2) caption-only representation-level fusion: the same representation-level fusion model but using the VQA representation only for the caption, v_C , and not for the image; 3) image-only representation-level fusion: the same model but using the VQA representation only for the image, v_I , and not for the caption; 4) deeper VQA-agnostic: The $N = 0$ representation-level fusion model described earlier that does not use VQA representations for neither the image nor the caption.

Table 4.3 summarizes the results. We see that incrementally adding more VQA-knowledge improves performance. Both caption-only and image-only models outperform the $N = 0$

Table 4.3: Ablation study evaluating the gain in performance as more VQA-knowledge is incorporated in the model

MSCOCO						
Approach	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Deeper VQA-agnostic	45.8	76.8	86.1	33.6	67.8	81.0
Caption-only representation-level fusion	47.3	77.3	86.6	35.5	69.3	81.9
Image-only representation-level fusion	47.0	80.0	89.6	36.4	70.1	82.3
Full representation-level fusion	50.5	80.1	89.7	37.0	70.9	82.9

deeper VQA-agnostic baseline. The full representation-level fusion model which combines both representations yields the best performance.

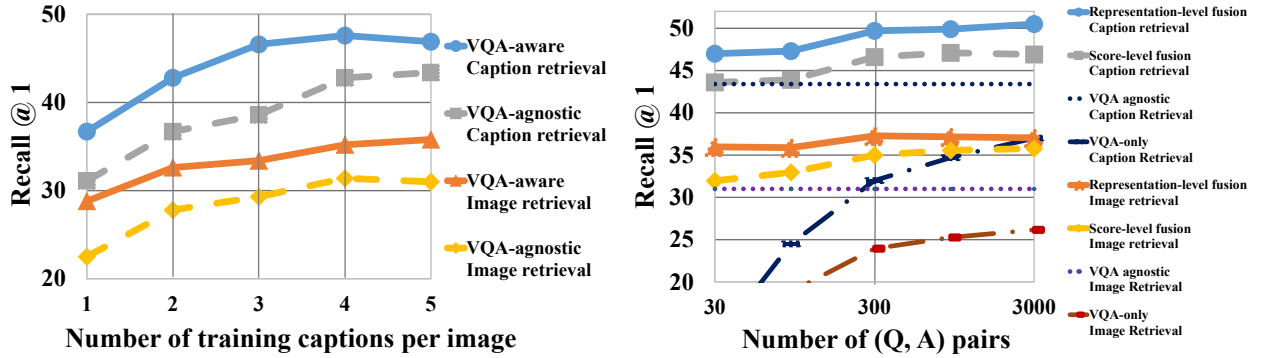
4.5.3 The Role of VQA and Caption Annotations

In this work we transfer knowledge from one vision-language task (*i.e.* VQA) to another (*i.e.* image-caption ranking). However, VQA annotations and caption annotations serve different purposes.

The target language to be retrieved is caption language, and not VQA language. [1] showed qualitatively and quantitatively that the two languages are statistically quite different (in terms of information contained, and in terms of nouns, adjectives, verbs, etc. used). As a result, VQA can not be thought of as providing additional “annotations” for the captioning task. Instead, VQA provides different perspectives/views of the images (and captions). It provides an additional feature representation. To better utilize this representation for an image-caption ranking task, one would still require sufficient ground truth caption annotations for images. In fact, with varying amounts of ground truth (caption) annotations, the VQA-aware representations show improvements in performance across the board. See Figure 4.6 (left).

A better analogy of our VQA representation is hidden activations (*e.g.*, fc7) from a CNN trained on ImageNet. Having additional ImageNet annotations would improve the fc7 feature. But to map this fc7 feature to captions, one would still require sufficient caption annotations. Conceptually, caption annotations and category labels in ImageNet play two

Figure 4.6: **Left**: caption retrieval and image retrieval performances of the VQA-agnostic model compared with our $N = 3000$ score-level fusion VQA-aware model trained using 1 to 5 captions per image. The VQA representations in the VQA-aware model provide consistent performance gains. **Right**: caption retrieval and image retrieval performances of our score-level fusion and representation-level fusion approaches with varying number of (Q, A) pairs used for feature extraction.



different roles. The former provides ground truth for the target task at hand (image-caption ranking), and having additional annotations for the target application typically helps. The latter helps learn a better image representation (which may provide improvements in a variety of tasks).

4.5.4 Number of Question-Answer Pairs

Our VQA-grounded representations extract image and caption features based on question-answer pairs. It is important for there to be enough question-answer pairs to cover necessary aspects for image-caption ranking. We experiment with using $N = 30, 90, 300, 900, 3000$ (Q, A) pairs (or facts) for both score-level and representation-level fusion. Figure 4.6 (right) shows caption and image retrieval performances of our approaches with varying N . Performance of both score-level and representation-level fusion approaches improve quickly from $N = 30$ to $N = 300$, and then starts to level off after $N = 300$.

An alternative to sampling 3 question-answer pairs per image on 1,000 images to get $N = 3000$ questions is to sample 1 question-answer pair per image from 3,000 images. Sampling multiple (Q, A) pairs from the same image provides correlated (Q, A) pairs. For example

Table 4.4: Caption retrieval and image retrieval performances of score-level fusion $N = 3000$, when its VQA and VQA-Caption submodules are trained on smaller, randomly sampled subsets of the VQA dataset.

MSCOCO						
Approach	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Kiros <i>et al.</i> [61] (VQA-agnostic)	43.4	75.7	85.8	31.0	66.7	79.9
Score-level fusion with 80k VQA training data	45.5	78.7	87.1	33.7	68.9	82.8
Score-level fusion with 160k VQA training data	46.7	78.2	87.3	34.8	69.6	82.9
Score-level fusion with 240k VQA training data	46.9	78.6	88.9	35.8	70.3	83.6

(Q : What are these animals? A : Giraffes) and (Q : Would this animal fit in a house? A : No). Using such correlated (Q, A) pairs, the model could potentially better predict if there is a giraffe in the image by jointly reasoning if the animal looks like a giraffe and the if the animal would fit in a house, if the VQA and VQA-Caption models have not already picked up such correlations. In experiments, sampling 3 question-answer pairs per image for correlated (Q, A) pairs does not significantly outperform sampling 1 question-answer pair per image which performs (47.7%, 35.4%) (image, caption) recall@1 using $N = 3000$ score-level fusion, so we hypothesize that our VQA and Caption-QA models have already captured such correlations.

4.5.5 Amount of VQA Training Data

Our model uses the VQA corpora to improve image-caption ranking. Naturally the amount of knowledge the VQA corpora contains will have a significant impact on the performance of our model. We set up an experiment to study that on the $N = 3000$ score-level fusion model. We train the VQA and VQA-Caption submodules in the score-level fusion model with random subsets of 80k, 160k and 240k(entire dataset) training examples in the VQA dataset and evaluate the image-caption ranking accuracy of the final score-level fusion model. Results are summarized in Table. 4.4.

We see that image-caption ranking performance consistantly improves with larger VQA datasets, and the improvements do not seem to be saturating. This suggests that our score-

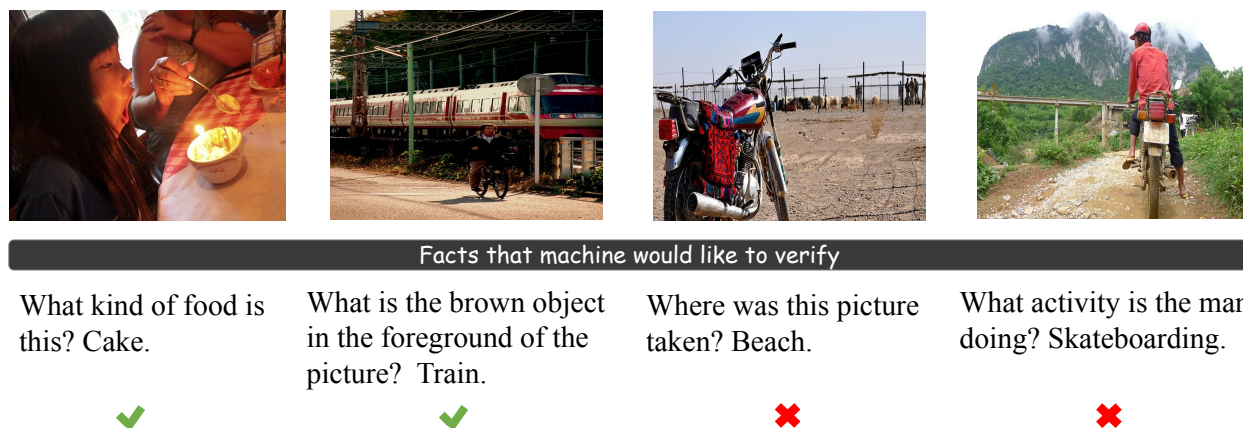


Figure 4.7: Facts that are most informative for ranking captions for each image in terms of mutual information between the fact and candidate captions. Selected from 3000 (Q, A) pairs using $N = 3000$ representation-level fusion VQA-aware model.

level fusion model can still benefit from more VQA training data.

4.5.6 On the Interpretability of the VQA-Aware Model

Deep models are well known to have very low interpretability. Using image-captioning models as an example, it is difficult to tell based on which facts from the image the model generates the caption, or why it fails when it does. Lacking understanding of the model, common practices often resign to adding more training data and using more complex architectures and hoping the performance improves.

By using VQA as a submodule, the VQA-aware model presents opportunities for us to probe the model: “which fact do you want verified for this image for better caption retrieval?”. That could help make the model more transparent and interpretable, allowing us to potentially improve the model more effectively, or strategically (*e.g.*, via active learning).

We ran a proof-of-concept qualitative experiment. Recall that in our VQA-aware model, each (Q, A) pair represents a fact about the image. We compute the mutual information between a fact’s validity for an image and the relevance of a caption for the image. Computing this mutual information requires an estimate of the joint distribution over the fact and the caption. We assume that the fact’s validity and the caption’s relevance are independent conditioned on model parameters (a VQA model for fact validity, and an image-caption model for caption relevance). To marginalize the models out, we use ideas from [39] which

showed that turning on dropout at test time allows unbiased sampling of model architectures. Technical details of this algorithm are described in Appendix C.2. We identify the most “informative” facts or (Q, A) pairs for an image whose validity has the highest mutual information with captions.

Figure 4.7 shows such most informative (Q, A) pairs for caption retrieval selected using our $N = 3000$ representation-level fusion VQA-aware model.

4.6 Generalization to Flickr8k and Flickr30k

So far we have performed image-caption ranking experiments on MSCOCO which the VQA submodule also uses. It is well known that datasets may contain biases [119], so a VQA model trained on MSCOCO may not be as accurate on other datasets. How does our approach generalize across datasets? We test the generalization ability of our approach on Flickr8k [52] and Flickr30k [128] image-caption ranking.

Flickr8k and Flickr30k consist of 8,000 and 30,000 images, respectively, collected from Flickr. Each image in Flickr8k and Flickr30k is annotated with 5 image captions. Following the evaluation protocol of [58] we use 1,000 images for validation, 1,000 images for testing, the rest for training and report $\text{recall}@ (1, 5, 10)$ for caption retrieval and image retrieval on test.

Table 4.5 and Table 4.6 show results on Flickr8k and Flickr30k dataset, respectively. Our VQA-aware model shows consistent improvements over the VQA-agnostic model on both datasets. On Flickr8k our score-level fusion approach achieves 24.3% caption retrieval $R@1$ and 17.2% image retrieval $R@1$, which outperforms the VQA-agnostic model by 2.0% and 2.3%. On Flickr30k our score-level fusion approach achieves 33.9% caption retrieval $R@1$ and 24.9% image retrieval $R@1$, which outperforms the VQA-agnostic model by 4.1% and 2.9%.

Note that the VQA and VQA-Caption models are trained on MSCOCO which is a different dataset. Yet, they consistently improve image-caption ranking on Flickr8k and Flickr30k. It shows that our VQA-grounded image and caption representations generalize across datasets. Fine-tuning on these datasets, and incorporating our approach on top of state-of-the-art captioning approaches on these datasets (Instead of [61] which is state-of-the-art on MSCOCO but not Flickr) may further improve our performance.

Both Flickr8k and Flickr30k are smaller compared with the MSCOCO dataset. Our representation-

Table 4.5: Results on Flickr8k dataset

Flickr8k						
Approach	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.1	0.5	1.0	0.1	0.5	1.0
DVSA [58]	16.5	40.6	54.2	11.8	32.1	43.8
FV (GMM+HGLMM) [62]	31.0	59.3	73.7	21.3	50.0	64.8
<i>m</i> -RNN-AlexNet [83]	14.5	37.2	48.5	11.5	31.0	42.4
<i>m</i> -CNN _{ENS} [78]	24.8	53.7	67.1	20.3	47.6	61.7
Kiros <i>et al.</i> [61] (VQA-agnostic)	22.3	48.7	59.8	14.9	38.3	51.6
N=3000 score-level fusion VQA-grounded only	10.5	31.5	42.7	7.6	22.8	33.5
N=3000 score-level fusion VQA-aware	24.3	52.2	65.2	17.2	42.8	57.2

Table 4.6: Results on Flickr30k dataset

Flickr30k						
Approach	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.1	0.5	1.0	0.1	0.5	1.0
DVSA [58]	22.2	48.2	61.4	15.2	37.7	50.5
FV (GMM+HGLMM) [62]	35.0	62.0	73.8	25.0	52.7	66.0
RTP (weighted distance) [98]	37.4	63.1	74.3	26.0	56.0	69.3
<i>m</i> -RNN-vgg [83]	35.4	63.8	73.7	22.8	50.7	63.1
<i>m</i> -CNN _{ENS} [78]	33.6	64.1	74.9	26.2	56.3	69.6
Kiros <i>et al.</i> [61] (VQA-agnostic)	29.8	58.4	70.5	22.0	47.9	59.3
N=3000 score-level fusion VQA-grounded only	17.6	40.5	51.2	12.7	31.9	42.5
N=3000 score-level fusion VQA-aware	33.9	62.5	74.5	24.9	52.6	64.8

level fusion model overfits to the training sets despite using dropout.

4.7 Discussion

VQA corpora provide rich multimodal information that is complementary to knowledge stored in image captioning corpora. In this work we take the novel perspective of viewing VQA as a “feature extraction” module that captures VQA knowledge. We propose two approaches – score-level and representation-level fusion – to integrate this knowledge into an existing image-caption ranking model. We set new state-of-the-art by improving caption retrieval by 7.1% and image retrieval by 4.4% on MSCOCO.

Improved individual modules, *i.e.*, VQA models and VQA-agnostic image-caption ranking models and end-to-end training may further improve the performance of our approach. In addition, an attention mechanism that selects question-answer pairs (facts) that are useful for ranking captions in an image-specific manner is also a promising direction of future research. From another perspective, that is a machine proposing questions about images that are informative about a target task (*e.g.*, image-caption ranking), which we briefly discussed in Section 4.5.6. Taking forward this idea, in Chapter 5 we study the problem of getting machines to ask questions about images to improve its knowledge quantitatively from an active learning perspective.

Acknowledgments

This work was supported in part by the Allen Distinguished Investigator awards by the Paul G. Allen Family Foundation, a Google Faculty Research Award, a Junior Faculty award by the Institute for Critical Technology and Applied Science (ICTAS) at Virginia Tech, a National Science Foundation CAREER award, an Army Research Office YIP award, and Office of Naval Research YIP award to D. P. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government or any sponsor.

Chapter 5

Active Learning for Visual Question Answering: An Empirical Study

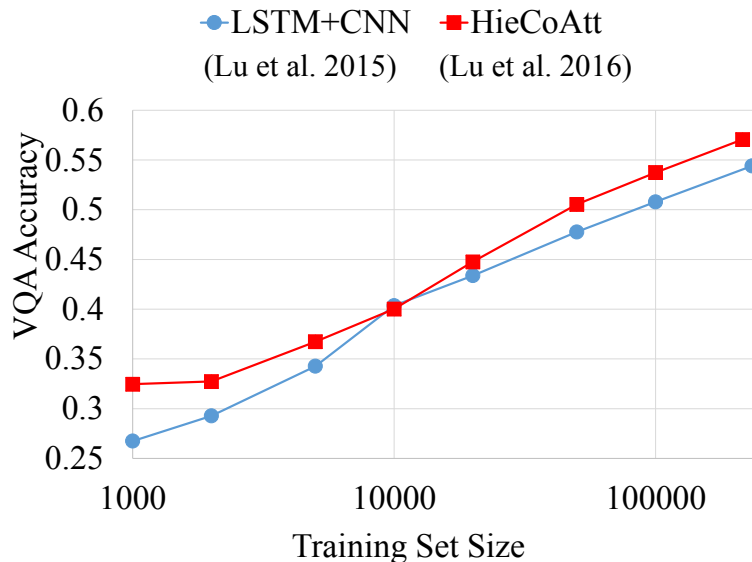
5.1 Introduction

Visual Question Answering (VQA) [1, 41, 42, 44, 82, 101] is the task of taking in an image and a free-form natural language question and automatically answering the question. Correctly answering VQA questions arguably demonstrates machines’ image understanding, language understanding and perhaps even some commonsense reasoning abilities. Previous works have demonstrated that deep models which combine image, question and answer representations, and are trained on large corpora of VQA data are effective at the VQA task.

Although such deep models are often deemed data-hungry, the flip-side is that their performance scales well with more training data. In Fig.5.1 we plot performance versus training set size of two representative deep VQA models: LSTM+CNN [76] and HieCoAtt [77] trained on random subsets of the VQA v1.0 dataset [1]. We see that for both methods, accuracy improves significantly – by 12% – with every order of magnitude of more training data. As performance improvements still seem linear, it is reasonable to expect another 12% increase by collecting a VQA dataset 10 times larger. A similar trend is also seen in ImageNet image classification [89]. Note that improvements brought by additional training data may be orthogonal to improvements in model architecture.

However, collecting large quantities of annotated data is expensive. Even worse, as a result of long tail distributions, it will likely result in redundant questions and answers while still

Figure 5.1: Performance of two representative VQA models: LSTM+CNN [76] and HieCoAtt [77] on random subsets of the VQA v1.0 dataset. Both models improve by 12% with every order of magnitude of more training data.



having insufficient training data for rare concepts. This is especially important for learning commonsense knowledge, as it is well known that humans tend to talk about unusual circumstances more often than commonsense knowledge which can be boring to talk about [43]. Active learning helps address these issues. In active learning, a model is first trained on an initial training set. It then iteratively expands its training set by selecting potentially informative examples according to a query strategy, and seeking annotations on these examples. Previous works have shown that a carefully designed query strategy effectively reduces annotation effort required in a variety of tasks for shallow models. For deep models however, active learning literature is scarce and mainly focuses on classical unimodal tasks such as image and text classification.

In this work we study active learning for deep VQA models. VQA poses several unique challenges and opportunities for active learning.

First, VQA is a multimodal problem. Deep VQA models may combine Multi-Layer Perceptrons (MLPs), Convolutional Neural Nets (CNNs), Recurrent Neural Nets (RNNs) and even attention mechanisms to solve VQA. Such models are much more complex than MLPs or CNNs alone studied in existing active learning literature and need tailored query strategies.

Second, VQA questions are free-form and open-ended. In fact, VQA can play several roles

from answering any generic question about an image, to answering only specific question types (*e.g.*, questions with “yes/no” answers, or counting questions), to being a submodule in some other task (*e.g.*, image captioning as in [75]). Each of these different scenarios may require a different active learning approach.

Finally, VQA can be thought of as a Visual Turing Test [42] for computer vision systems. To answer questions such as “does this person have 20/20 vision” and “will the cat be able to jump onto the shelf”, the computer not only needs to understand the surface meaning of the image and the question, but it also needs to have sufficient commonsense knowledge about our world. One could argue that proposing informative questions about images is also a test of commonsense knowledge and intelligence.

We draw coarse analogies to human learning and explore three types of information-theoretic active learning query strategies:

Cramming – maximizing information gain in the training domain. The objective of this strategy is to efficiently memorize knowledge in an unlabeled pool of examples. This strategy selects unlabeled examples whose label the model is most uncertain about (maximum entropy).

Curiosity-driven learning – maximizing information gain in model space. The objective of this strategy is to select examples that could potentially change the belief on the model’s parameters (also known as expected model change). There might exist examples in the pool whose labels have high uncertainty but the model does not have enough capacity to capture them. In curiosity-driven learning the model will skip these examples. BALD [40, 54] is one such strategy for deep models, where examples are selected to maximize the reduction in entropy over model parameter space.

Goal-driven learning – maximizing information gain in the target domain. The objective of this strategy is to gather knowledge to better achieve a particular goal (also known as expected error reduction). To give an example from image classification, if the goal is to recognize digits *i.e.*, the target domain is digit classification, dog images in the unlabeled pool are not relevant even though their labels might be uncertain or might change model parameters significantly. On the other hand, in addition to digit labels, some other non-digit labels such as the orientation of the image might be useful to the digit classification task. We propose a novel goal-driven query strategy that computes mutual information between pool questions and test questions under the Bayesian Neural Network [7, 39] framework.

We evaluate active learning performance on VQA v1.0 [1] and v2.0 [44] under the pool-

based active learning setting described in Section 5.3. We show that active learning for deep VQA models requires a large amount of initial training data before they can achieve better scaling than random selection. In other words, the model needs to have enough knowledge to ask informative questions. But once it does, all three querying strategies outperform the random selection baseline, saving 27.3% and 19.0% answer annotation effort for VQA v1.0 and v2.0 respectively. Moreover, when the target task is restricted to answering only “yes/no” questions, our proposed goal-driven query strategy beats random selection and achieves the best performance out of the three active query strategies.

5.2 Related Work

5.2.1 Active Learning

Active learning query strategies for shallow models [107, 65] often rely on specific model simplifications and closed-form solutions. Deep neural networks however, are inherently complex non-linear functions. This poses challenges on uncertainty estimation.

In the context of deep active learning for language or image understanding, [134] develops a margin-based query strategy on Restricted Boltzmann Machines for review sentiment classification. [64] queries high-confidence web images for active fine-grained image classification. [106] proposes a query strategy based on feature space covering, applied to deep image classification. Closest to our work, [40] studies BALD [54], an expected model change query strategy computed under the Bayesian Neural Network [7, 39] framework applied to image classification.

In this work we study active learning for VQA. VQA is a challenging multimodal problem. Today’s state-of-the-art VQA models are deep neural networks. We take an information-theoretic perspective and study three active learning objectives: minimizing entropy in training domain (entropy), model space (expected model change) or target domain (expected error reduction). Drawing coarse analogy from human learning, we call them cramming, curiosity-driven and goal-driven learning respectively. We apply the Bayesian Neural Network [7, 39] framework to compute these strategies and derive a novel goal-driven query scoring function that is effective in performance and efficient to compute even for complex multimodal neural networks.

5.2.2 Visual Conversations

Building machines that demonstrate curiosity – machines that improve themselves through conversations with humans – is an important problem in AI.

[91, 90] study generating human-like questions given an image and the context of a conversation about that image. [115] uses reinforcement learning to learn an agent that plays a “Guess What?” game [22]: finding out which object in the image the user is looking at by asking questions. [20] studies grounded visual dialog [19] between two machines in collaborative image retrieval, where one machine as the “answerer” has an image and answers questions about the image while the other as “questioner” asks questions to retrieve the image at the end of the conversation. Both machines are learnt to better perform the task using reinforcement learning.

In this work we study visual “conversations” from an active learning perspective. In each round of the conversation, a VQA model strategically chooses an informative question about an image and queries an oracle to get an answer. Each round of “conversation” provides a new VQA training example which improves the VQA model.

5.3 Approach

We study a pool-based active learning setting for VQA: A VQA model is first trained on an initial training set \mathcal{D}_{train} . It then iteratively grows \mathcal{D}_{train} by greedily selecting batches of high-scoring question-image pairs (Q, I) from a human-curated pool according to a query scoring function $s(Q, I)$. The selected (Q, I) pairs are sent to an oracle for one of J ground truth answers $A \in \{a_1, a_2, \dots, a_J\}$, and (Q, I, A) tuples are added as new examples to \mathcal{D}_{train} .

¹

We take an information-theoretic perspective and explore cramming, curiosity-driven, and goal-driven query strategies as described in Section 5.1. However computing $s(Q, I)$ for those query strategies directly is intractable, as they require taking expectations under the model parameter distribution. So in Section 5.3.1 we first introduce a Bayesian VQA model

¹VQA models require a large training set to be effective. To avoid prohibitive data collection cost and focus on evaluating active learning query strategies, in this work we study pool-based active learning which makes use of existing VQA datasets. Having the model select or even *generate* questions for images it would liked answered, as opposed to picking from a pool of (Q, I) pairs is a direction for future research.

which enables variational approximation of the model parameter distribution. And then Section 5.3.2 introduces the query scoring functions and their approximations.

5.3.1 Bayesian LSTM+CNN for VQA

We start with the LSTM+CNN VQA model [76] introduced in Section 4.3.2. The model encodes an image into a feature vector using the VGG-net [109] CNN, encodes a question into a feature vector by learning a Long Short Term Memory (LSTM) RNN, and then learns a multi-layer perceptron on top that combines the image feature and the question feature to predict a probabilistic distribution over top $J = 1000$ most common answers.

In order to learn a variational approximation of the posterior model distribution, we adopt the Bayesian Neural Network framework [7, 39] and introduce a Bayesian LSTM+CNN model for VQA. Let ω be the parameters of the LSTM and the multi-layer perceptron (we use a frozen pre-trained CNN). We learn a weight-generating model with parameter θ :

$$\begin{aligned}\omega &= \theta \circ \epsilon \\ \epsilon_i &\sim \text{Bernoulli}(0.5)\end{aligned}\tag{5.1}$$

Let $q_\theta(\omega)$ be the probabilistic distribution of weights generated by this model. Following [7, 39], we learn θ by minimizing KL divergence $KL(q_\theta(\omega)||p(\omega|\mathcal{D}_{train}))$ so $q_\theta(\omega)$ serves as a variational approximation to the true model parameter posterior $p(\omega|\mathcal{D}_{train})$. Specifically we minimize

$$KL(q_\theta(\omega)||p(\omega|\mathcal{D}_{train})) = \underbrace{\mathbb{E}_{\omega \sim q_\theta(\omega)}[-\log P(\mathcal{D}_{train}|\omega)]}_{\text{Cross entropy loss}} + \underbrace{KL(q_\theta(\omega)||p(\omega))}_{\text{Deviation from weight prior}}\tag{5.2}$$

using batch Stochastic Gradient Descent (SGD) to learn θ . In practice, $KL(q_\theta(\omega)||p(\omega))$ can be naively approximated with a parametric hybrid $L1 - L2$ norm [39]. Experiments show that such a naive approximation does not have a significant impact on active learning results. So in experiments we set this term to 0. How to come up with a more informative prior term is an open problem for Bayesian Neural Networks.

Let $P(A|Q, I, \omega)$ be the predicted J -dimensional answer distribution of the VQA model for

question-image pair (Q, I) when using ω as model parameters. A Bayesian VQA prediction for (Q, I) using variational approximation $q_\theta(\omega)$ is therefore given by:

$$P(A = a|Q, I) \approx \mathbb{E}_{\omega \sim q_\theta(\omega)} P(A = a|Q, I, \omega) \quad (5.3)$$

5.3.2 Query Strategies and Approximations

We experiment with 3 active learning query strategies: cramming, curiosity-driven learning and goal-driven learning.

Cramming or “uncertainty sampling” [107] minimizes uncertainty (entropy) of answers for questions in the pool. It selects (Q, I) whose answer A ’s distribution has maximum entropy. This is a classical active learning approach commonly used in practice.

$$\begin{aligned} s_{entropy}(Q, I) &= \mathbb{H}(A) \\ &= - \sum_a P(A = a|Q, I) \log P(A = a|Q, I) \end{aligned} \quad (5.4)$$

Curiosity-driven learning or “expected model change” minimizes uncertainty (entropy) of model parameter distribution $p(\omega|\mathcal{D}_{train})$. It selects (Q, I) whose answer A would expectedly bring steepest decrease in model parameter entropy if added to the training set.

$$\begin{aligned} s_{curiosity}(Q, I) &= \mathbb{H}(\omega) - \mathbb{H}(\omega|A) \\ &= \mathbb{I}(\omega; A) \\ &= \mathbb{H}(A) - \mathbb{H}(A|\omega) \end{aligned} \quad (5.5)$$

Intuitively, $\mathbb{H}(A) - \mathbb{H}(A|\omega)$ computes the divergence of answer predictions under different model parameters. If plausible models are making divergent predictions on a question-image pair (Q, I) , the answer to this (Q, I) would rule out many of those models and thereby reduce confusion.

According to BALD [40], the conditional entropy term $\mathbb{H}(A|\omega)$ in Eq. 5.5 can be approximated by:

$$\mathbb{H}(A|\omega) \approx -\mathbb{E}_{\omega \sim q_\theta(\omega)} \sum_a P(A = a|Q, I, \omega) \log P(A = a|Q, I, \omega) \quad (5.6)$$

Goal-driven learning or “expected error reduction” minimizes uncertainty (entropy) on answers A'_t to a given set of unlabeled test question-image pairs $(Q'_t, I'_t), t = 1, 2, \dots, T$, against which the model will be evaluated. The goal-driven query strategy selects the pool question-image pair (Q, I) that has the maximum total mutual information with $(Q'_t, I'_t), t = 1, 2, \dots, T$. That is, it queries (Q, I) pairs which maximize:

$$\begin{aligned}
 s_{goal}(Q, I) &= \sum_t \mathbb{H}(A'_t) - \mathbb{H}(A'_t|A) \\
 &= \sum_t \mathbb{I}(A; A'_t) \\
 &= \sum_t \sum_a \sum_{a'} P(A = a, A'_t = a' | Q, I, Q'_t, I'_t) \log \frac{P(A = a, A'_t = a' | Q, I, Q'_t, I'_t)}{P(A = a | Q, I) P(A'_t = a' | Q'_t, I'_t)}
 \end{aligned} \tag{5.7}$$

For term $P(A = a, A'_t = a' | Q, I, Q'_t, I'_t)$, observe that when the model parameter $\boldsymbol{\omega}$ is given, (Q, I) and (Q'_t, I'_t) are two different VQA questions so their answers – A and A'_t respectively – are predicted independently. In other words, A and A'_t are independent conditioned on $\boldsymbol{\omega}$. Therefore we can take expectation over model parameter $\boldsymbol{\omega}$ to compute this joint probability term:

$$\begin{aligned}
 P(A = a, A'_t = a' | Q, I, Q'_t, I'_t) &= \mathbb{E}_{\boldsymbol{\omega}} P(A = a | Q, I, \boldsymbol{\omega}) P(A'_t = a' | Q'_t, I'_t, \boldsymbol{\omega}) \\
 &\approx \mathbb{E}_{\boldsymbol{\omega} \sim q_{\theta}(\boldsymbol{\omega})} P(A = a | Q, I, \boldsymbol{\omega}) P(A'_t = a' | Q'_t, I'_t, \boldsymbol{\omega})
 \end{aligned} \tag{5.8}$$

Let M be the number of samples of $\boldsymbol{\omega}$, J be the number of possible answers, and U be the number of examples in the pool. Computing $\mathbb{I}(A; A'_t)$ for all U examples in the pool following Eq. 5.8 has a time complexity of $O(UTJ^2M)$. For VQA typically the pool and test corpora each contains hundreds of thousands of examples and there are 1000 possible answers, *e.g.*, $U = 400\text{k}$, $T = 100\text{k}$ and $J = 1,000$. We typically use $M = 50$ samples in our experiments. So computing Eq. 5.8 is still time-consuming and can be prohibitive for even larger VQA datasets. To speed up computation, we approximate $\log(\cdot)$ using first-order Taylor expansion and discover that the following approximation holds empirically (more details can be found in Appendix D.1):

Algorithm 1 Active learning for Visual Question Answering

-
- 1: Initialize \mathcal{D}_{train} with N initial training examples. Use the rest of (Q, I) in VQA TRAIN set as pool.Q
 - 2: Train θ on \mathcal{D}_{train} for K epochs using Eq. 5.2 for initial $q_\theta(\omega)$.
 - 3: **for** $iter = 1, \dots, L$ **do**
 - 4: Sample $\omega \sim q_\theta(\omega)$ M times.
 - 5: Using each ω to make predictions $P(A|Q, I, \omega)$ on all pool and test question-image pairs.
 - 6: Compute $s(Q, I)$ for every (Q, I) in pool using Eq. 5.4, 5.5 or 5.9.
 - 7: Select the top G high-scoring (Q, I) pairs from the pool.²
 - 8: Lookup answers A for (Q, I) pairs in the VQA training set (proxy for querying a human).
 - 9: Add (Q, I, A) tuples to \mathcal{D}_{train} .
 - 10: Update θ on new \mathcal{D}_{train} for K epochs.
 - 11: **end for**
-

$$\begin{aligned}
s_{goal}(Q, I) &\approx \frac{1}{2} \left[\mathbb{E}_\omega \mathbb{E}_{\omega'} \sum_a \frac{P(A = a|Q, I, \omega) P(A = a|Q, I, \omega')}{P(A = a|Q, I)} \right. \\
&\quad \left. \sum_t \sum_a \frac{P(A'_t = a|Q'_t, I'_t, \omega) P(A'_t = a|Q'_t, I'_t, \omega')}{P(A'_t = a|Q'_t, I'_t)} - T \right] \quad (5.9)
\end{aligned}$$

Eq. 5.9 brings drastic improvements to time complexity. It can be computed as a dot-product between two vectors of length M^2 . One only involves pool questions (Q, I) while the other one only involves test questions (Q'_t, I'_t) and can be precomputed for all pool questions. Precomputing the vector for test questions has a time complexity of $O(TJM^2)$. And then computing Eq. 5.9 using the precomputed test vector has a time complexity of $O(UJM^2)$. So the overall time complexity is linear to both dataset size U and T and the number of possible answers J .

Previous works explore this expected error reduction objective for shallow classifiers such as Naive Bayes [103], Support Vector Machines [46] and Gaussian Process [135]. Computing

²Jointly selecting a batch of (Q, I) pairs that optimizes the active learning objectives may further improve active learning performance. Deriving query strategies that can select batches of examples under the Bayesian Neural Network framework is part of future work.

their scoring functions would require learning a new set of model parameters for every possible combinations of (Q, I, A) and then making predictions on (Q'_t, I'_t) using the learnt model. Instead our goal-driven scoring function is designed for Bayesian Neural Networks. The Monte-Carlo approximation of Eq. 5.9 only involves making predictions on (Q, I) and (Q'_t, I'_t) , and avoids training new models for each of $J = 1,000$ answers when computing $s_{goal}(Q, I)$. As a result, our approach has a much lower time complexity and is easy to parallelize.

Our active learning procedure is summarized in Algorithm 1.

5.4 Experiment

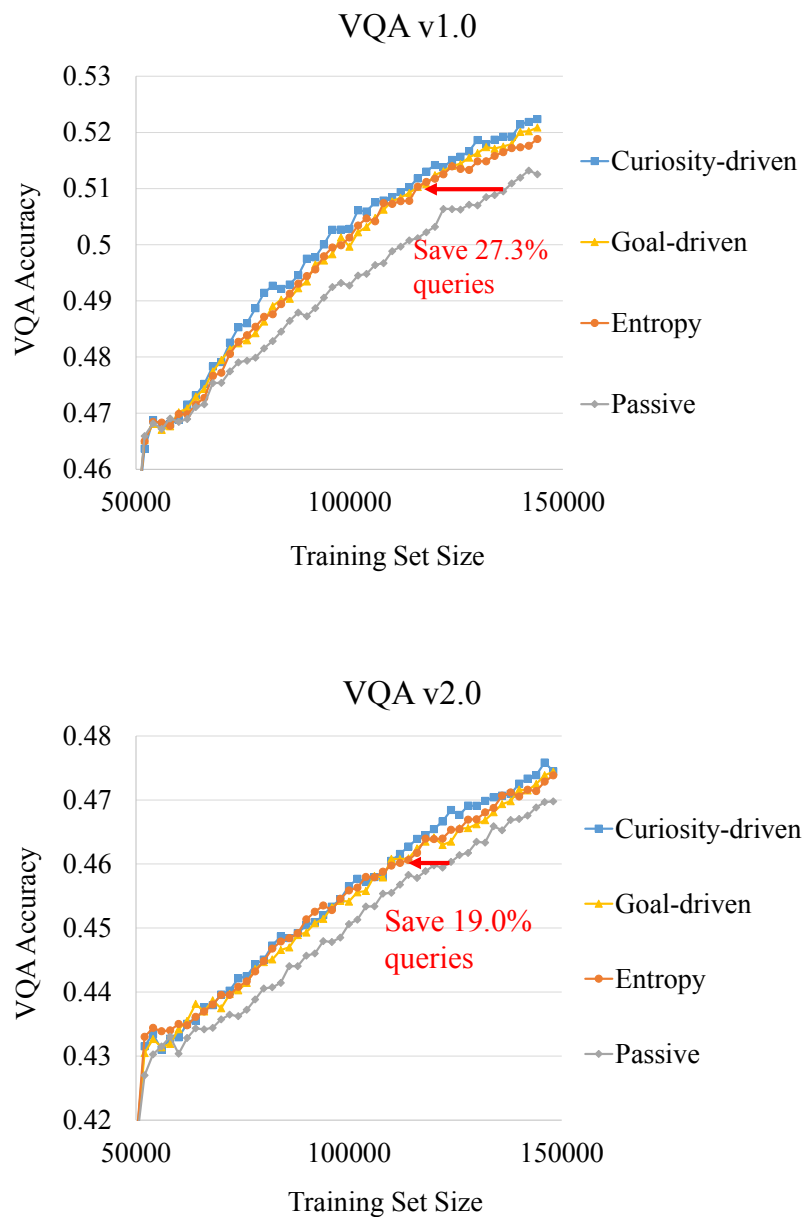
5.4.1 Experiment Setup

We evaluate cramming (entropy), curiosity-driven and goal-driven active learning strategies against passive learning on the VQA v1.0 [1] and v2.0 [44] datasets. The VQA v1.0 dataset consists of 614,163 VQA questions with human answers on 204,721 COCO [73] images. The VQA v2.0 dataset augments the VQA v1.0 dataset and brings dataset balancing: every question in VQA v2.0 is paired with two similar images that have different answers to the question. So VQA v2.0 doubles the amount of data and models need to focus on the image to do well on VQA v2.0.

We choose a random initial training set of $N = 50k$ (Q, I) pairs from the TRAIN split, use the rest of TRAIN as pool and report VQA accuracy [1] on the VAL split. We run the active learning loop for $L = 50$ iterations. We sample model parameter ω $M = 50$ times for query score computation. For passive learning *i.e.* querying (Q, I) pairs randomly, we set $s_{passive}(Q, I) \sim \text{uniform}(0, 1)$. In each iteration $G = 2,000$ (Q, I, A) pairs are added to \mathcal{D}_{train} , resulting in a training set of 150k examples by the end of iteration 50.

For VQA model, we use the Bayesian LSTM+CNN model described in Section 5.3.1. In every active learning iteration we train the model for $K = 50$ epochs with learning rate 3×10^{-4} and batch size 8×128 for learning $q_{\theta}(\omega)$.

Figure 5.2: Active learning versus passive learning on (top) VQA v1.0 and (bottom) v2.0. All three active learning strategies perform better than passive learning.



5.4.2 Active Learning on VQA v1.0 and v2.0

Fig. 5.2 (top), (bottom) show the active learning results on VQA v1.0 and v2.0 respectively. On both datasets, all 3 active learning methods perform similarly and all of them outperform passive learning. On VQA v1.0, passive learning queries 88k answers before reaching 51% accuracy, where as all active learning methods need only 64k queries, achieving a saving of 27.3%. It shows that active learning is able to effectively tell informative VQA questions from redundant questions, even among high-quality questions generated by humans. Similarly at 46% accuracy, active learning on VQA v2.0 achieves a saving of 19.0%. Savings on VQA v2.0 is lower, possibly because dataset balancing in VQA v2.0 improves the informativeness of even a random example.

Table 5.1 shows that for each pair of active learning methods, what percentage of the query (Q, I) pairs are selected by both methods on VQA v2.0 (overlap between their training sets). For the VQA task, active learning methods seem to agree on which (Q, I) pairs are more informative. They have more than 80% of (Q, I) pairs in common, while against passive learning they only share $\sim 27\%$ of (Q, I) pairs.

On VQA v2.0, we also experiment with smaller initial training sets $N \in \{20k, 10k, 5k, 2k\}$ to study the impact of training set size on active learning performance. Fig. 5.3 shows the results. For all initial training set sizes, the breakpoint when active learning methods start to outperform passive learning is around 30k to 50k examples. It shows that active learning methods do require a large training set size to start asking informative questions. Models with smaller initial training set sizes tend to show less and inconsistent data savings compared to $N = 50k$, possibly because such models are less capable of telling informative questions from redundant ones. In addition, entropy shows fluctuating performance while curiosity-driven learning performs consistently better than both entropy and passive learning irrespective of initial training set size.

5.4.3 Goal-driven Active Learning

To evaluate our goal-driven learning approach, we keep the initial training set and the pool unchanged for VQA v2.0 – the model is allowed to ask all kinds of questions from the VQA v2.0 TRAIN split – but will be evaluated on only “yes/no” questions (questions whose answers are “yes” or “no”) in the VAL split. This task tests our proposed goal-driven active learning approach’s ability to focus on achieving the goal of answering “yes/no” questions

Figure 5.3: Active learning with $N = 20k, 10k, 5k, 2k$ initial training set size. When dataset size is small, active learning is unable to outperform passive learning. The breakpoint when active learning methods start to perform better is around 30k to 50k examples.

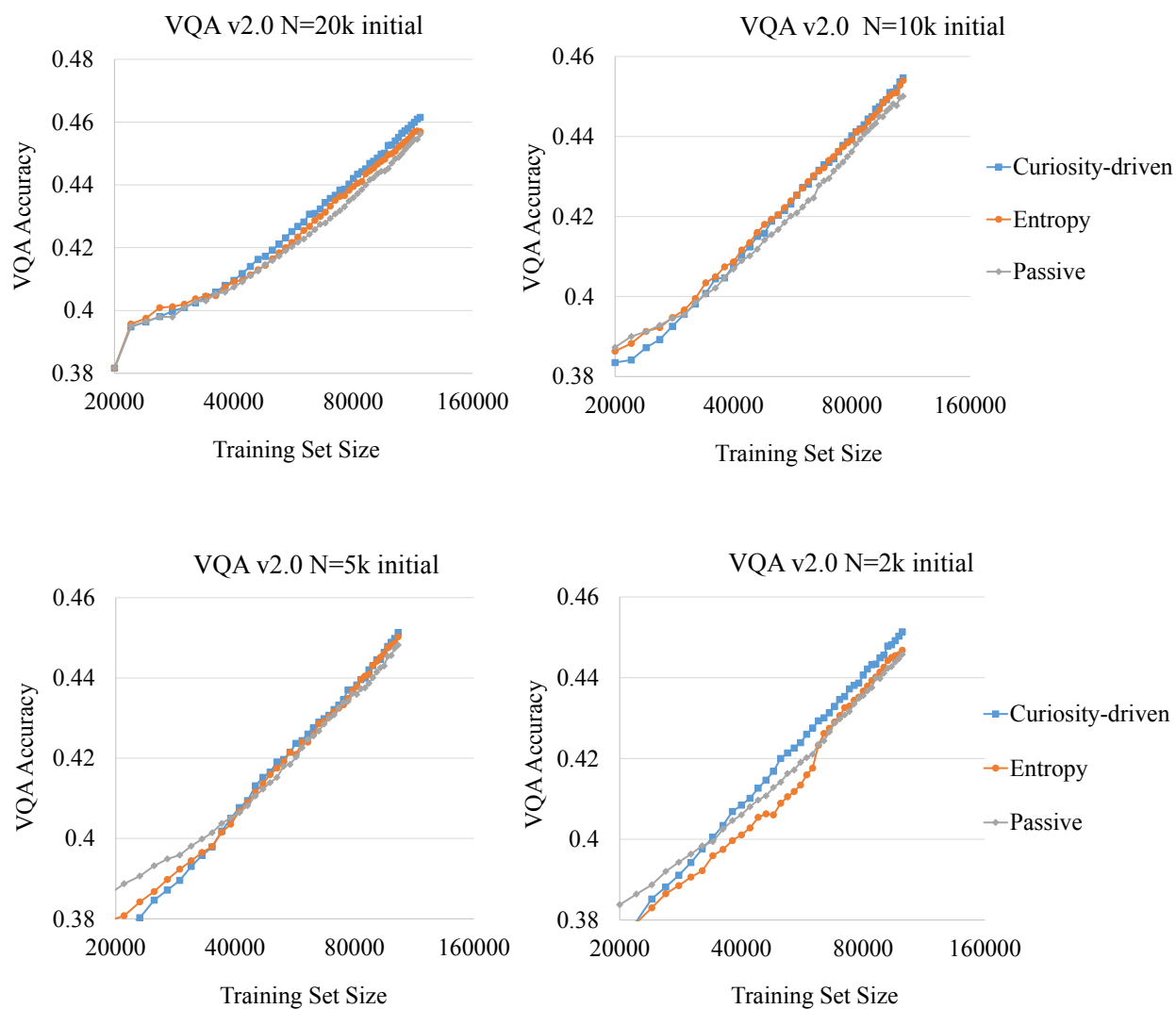


Table 5.1: On VQA v2.0 for each pair of query strategy, what percentage of (Q, I) pairs are selected by both methods. Active learning (entropy, curiosity-driven, goal-driven) query strategies select $> 80\%$ common (Q, I) pairs and they are very different from passive learning.

(Q, I) Overlap (%)	Passive learning	Entropy	Curiosity driven	Goal driven
Passive learning	-	26.70	26.65	26.64
Entropy	26.70	-	83.26	82.52
Curiosity-driven	26.65	83.26	-	85.27
Goal-driven	26.64	82.52	85.27	-

more accurately.

Fig. 5.4 (top) shows the performance of active and passive learning approaches on this task.³ Our goal-driven active learning approach is able to select relevant questions as queries and outperforms passive learning. Curiosity-driven and entropy approaches perform poorly. They are not aware of the task and tend to be attracted to harder, open-ended questions, which are not very relevant to the task.

Fig. 5.4 (bottom) shows a closer examination of the composition of questions queried by the goal-driven learning approach compared to baseline approaches. The goal-driven learning approach queries mostly “yes/no” questions, which are presumably more useful for the task. Note that the approach was not told that the downstream task is to answer “yes/no” questions. The approach figures out which questions will be informative to ask just based on samples from the downstream task. It shows that the goal-driven scoring function in Eq. 5.7, as well as the approximations in Eq. 5.9 are indeed effective for selecting informative questions.

As an “upper bound”, it is reasonable to assume⁴ that “yes/no” questions are more desirable for this task. Imagine a passive learning method that “cheats”: one that is aware that it will be tested only on “yes/no” questions, as well as knowing which questions are “yes/no”

³We also found that updating θ from previous iteration in Algorithm 1 step 10 leads to slight overfitting that affects mutual information approximation. So for this task, θ is initialized from scratch in every iteration.

⁴Note that this is not necessarily the case. Even non-yes/no questions can help a VQA model get better at answering yes/no questions by learning concepts from non-yes/no questions that can later come handy for yes/no questions. For example “Q: What is the man doing? A: Surfing” can be as useful as “Q: Is the man surfing? A: Yes”.

Figure 5.4: **Top:** Goal-driven active learning of VQA for answering only “yes/no” questions. Our goal-driven active learning approach outperforms passive learning and other active learning approaches. **Bottom:** Query compositions of active learning approaches, on VQA v2.0 dataset for the task of answering only “yes/no” questions. Our goal-driven active learning approach queries mostly “yes/no” questions.

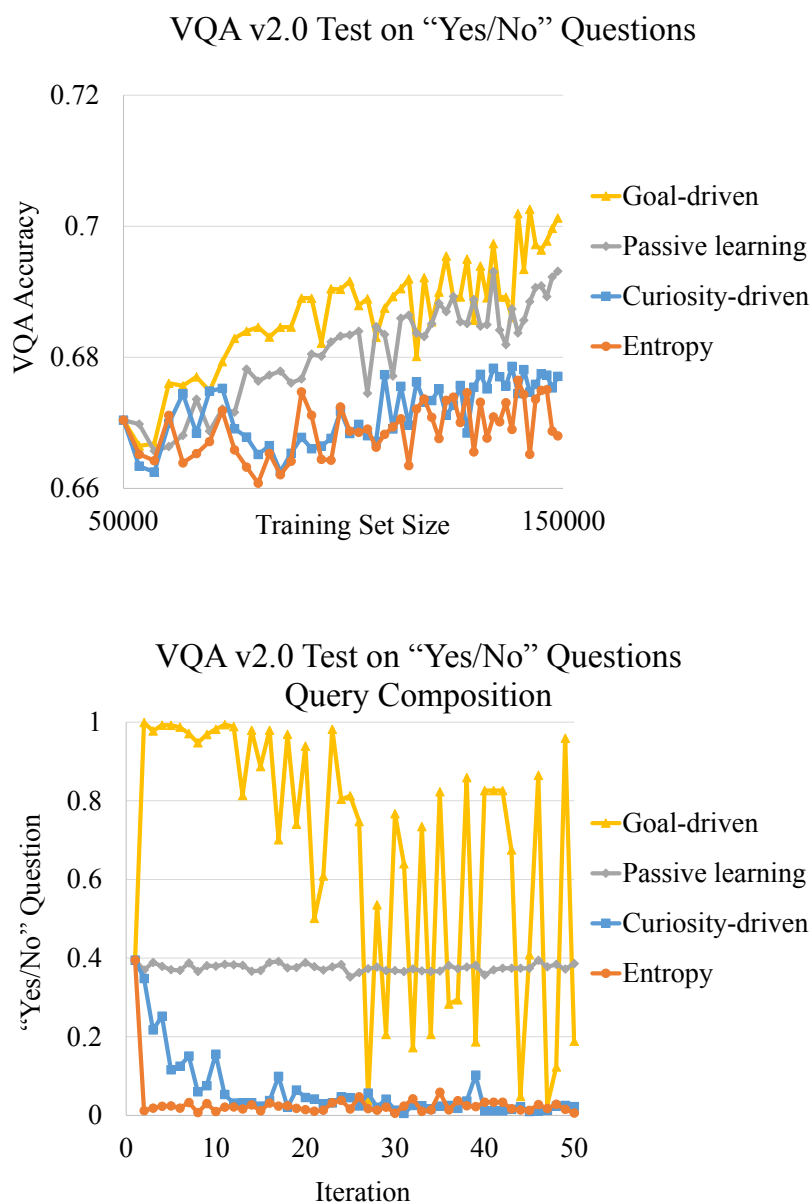
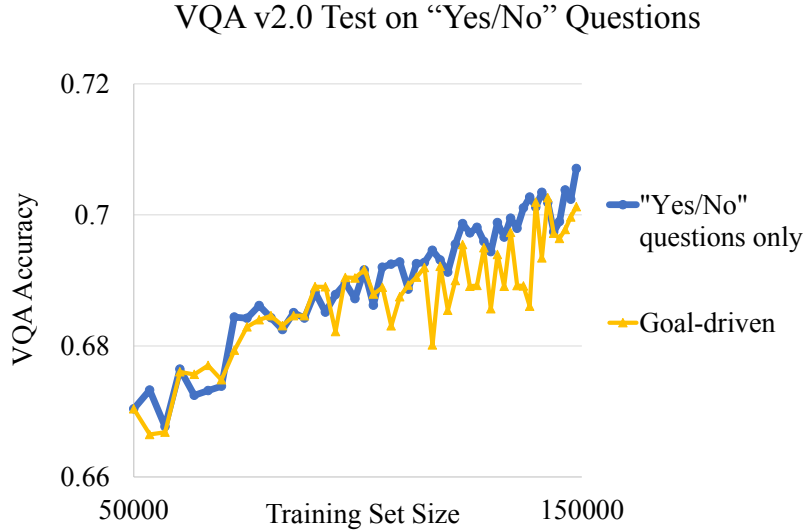


Figure 5.5: Goal-driven active learning of VQA for answering only “yes/no” questions, compared to passive learning that “cheats” and queries only “yes/no” questions.



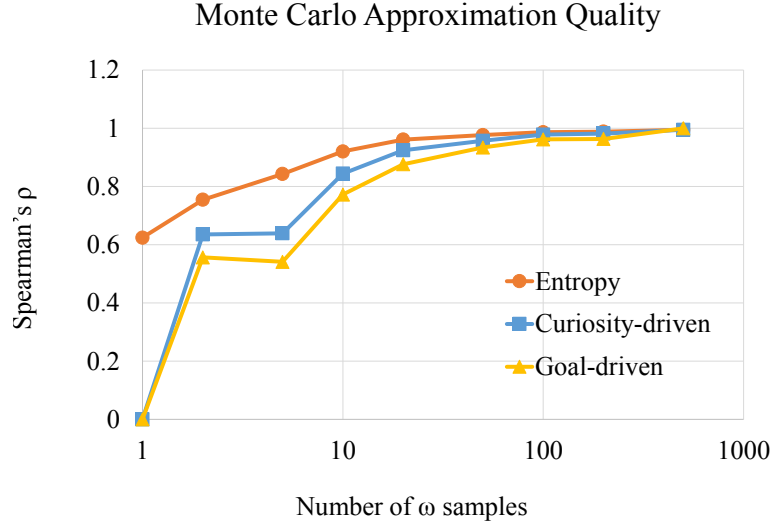
questions in the pool, so it restricts itself to query only “yes/no” questions. How does our goal-driven learning approach compare with such a method that only learns from “yes/no” questions? Fig. 5.5 shows the results. Our goal-driven learning is able to compete with the “cheating” approach. In fact, of all 167,499 “yes/no” questions in the VQA v2.0 TRAIN split, goal-driven learning finds 38% of them by iteration 25, and 50% of them by iteration 50. That might also have made finding the remaining “yes/no” questions more difficult which explains the drop of the rate of “yes/no” question towards later iterations. We expect that a larger pool (*i.e.* a larger VQA dataset) would reduce these issues.

5.4.4 Quality of Approximations

Our entropy, curiosity-driven and goal-driven scoring functions use 3 types of approximations

- Variational distribution $q_{\theta}(\omega)$ as approximation to model parameter distribution $p(\omega|\mathcal{D}_{train})$.
- Monte Carlo sampling over $q_{\theta}(\omega)$ for computing expectation over $p(\omega|\mathcal{D}_{train})$.
- Fast approximation to mutual information in Eq. D.8.

Figure 5.6: Convergence of Monte Carlo approximation to entropy, curiosity-driven and goal-driven scoring functions in terms of rank correlation. We compute scores using Eq. 5.4 (entropy), 5.5 (curiosity-driven) and 5.7 (goal-driven) for 200 random examples from the pool using $M \in \{1, 2, 5, 10, 20, 50, 100, 200, 500\}$ samples from $q_{\theta}(\omega)$, and compare them with $M = 500$ in terms of rank correlation (Spearman’s ρ).

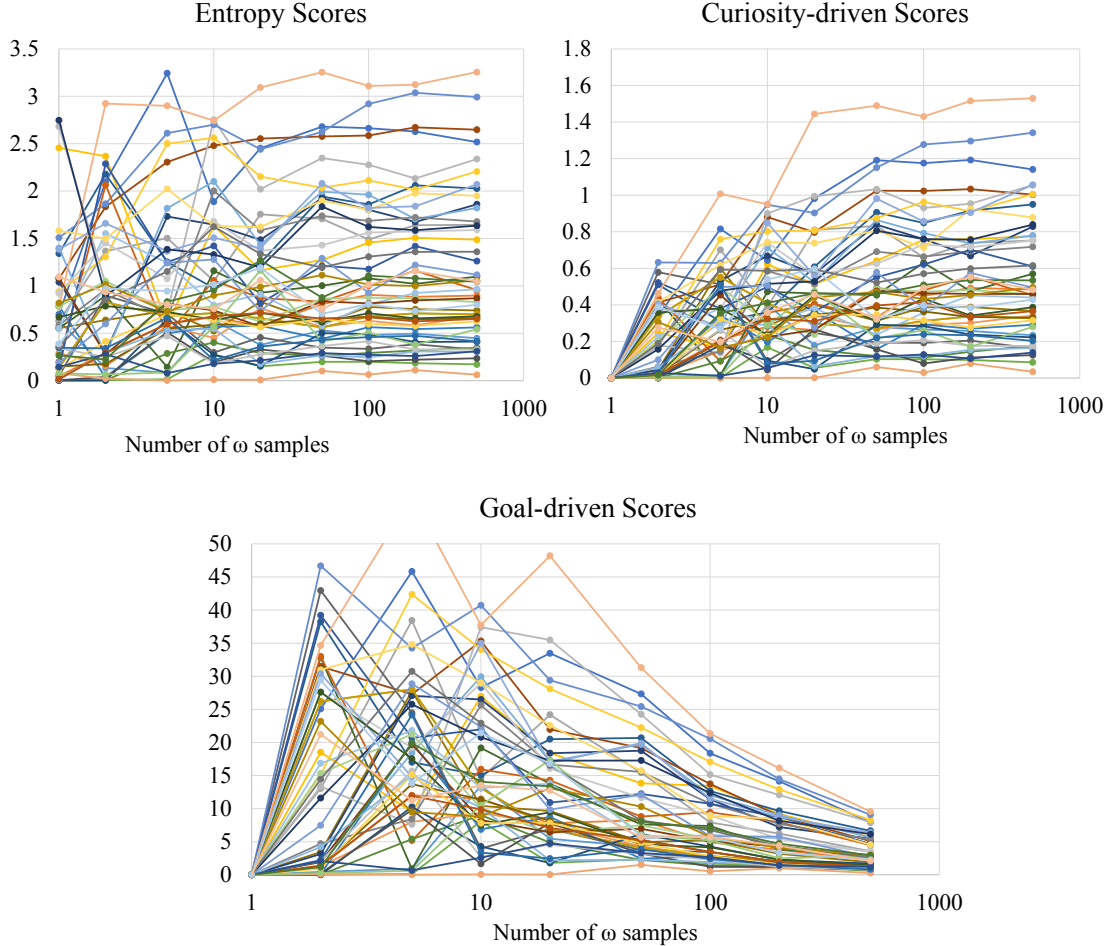


For (a), since the space of model parameters is very large, it is intractable to evaluate how accurate $q_{\theta}(\omega)$ substitutes $p(\omega|\mathcal{D}_{train})$ for expectation computation. But nevertheless our goal-driven learning results in Section 5.4.3 suggest that Eq. 5.9 computed using $q_{\theta}(\omega)$ is indeed useful for selecting relevant examples. It remains as an open problem that how to quantitatively evaluate the quality of $q_{\theta}(\omega)$ for the purpose of uncertainty estimation and expectation computation.

For (b), we study the convergence patterns of Monte Carlo sampling. Specifically, given an arbitrary VQA model⁵, we compute scores using Eq. 5.4 (entropy), 5.5 (curiosity-driven) and 5.7 (goal-driven) for 200 random examples from the pool using $M \in \{1, 2, 5, 10, 20, 50, 100, 200, 500\}$ samples from $q_{\theta}(\omega)$, and compare them with $M = 500$ in terms of rank correlation (Spearman’s ρ). Note that we use different seeds for the different M values, *i.e.* ω samples for $M = 200$ do not overlap with ω samples for $M = 500$. Fig. 5.6 shows the results. Entropy, curiosity-driven and goal-driven scoring functions require increasingly more samples of model parameters to converge in terms of ranking. To reach $\rho = 0.9$, entropy, curiosity-driven

⁵For our experiments we use the model from curiosity-driven learning at iteration 50. This choice is made arbitrarily and does not change conclusions.

Figure 5.7: Entropy, curiosity-driven and goal-driven scores of 50 examples under different numbers of model parameter samples.



and goal-driven scoring functions require 10, 20 and 50 samples from $q_{\theta}(\omega)$ respectively. Fig. 5.8 shows how the actual scores of examples change according to number of samples from $q_{\theta}(\omega)$ for 50 random examples in the pool. The entropy and curiosity-driven scores seem to converge with a large number of samples. The goal-driven scores however, tend to first increase and then decrease with the number of samples and have not yet converged by $M = 500$ samples, which is a limitation of the Monte Carlo sampling approach. Despite that, the relative rankings based on which the queries are selected have mostly converged. Upper- and lower-bounds of Eq. 5.7 that might improve convergence are opportunities for future research.

For (c), we plot goal-driven scores Eq. 5.7 as the x-axis versus our fast approximations Eq. 5.9

as the y-axis for 200 random examples from the pool using $M = \{2, 5, 10, 20, 50, 100, 200, 500\}$ samples from $q_{\theta}(\omega)$. Because Eq. 5.7 does not scale well to large datasets, we use a subset of 200 random (Q'_t, I'_t) pairs from the VAL split as the test domain for both Eq. 5.7 and Eq. 5.9. Fig. 5.8 shows the results. Our fast approximations are mostly linear to the goal-driven scores. The slope changes according to the number of model parameter samples M . That is probably because our approximation $\frac{1}{2}(-x + x^2)$ (see Section D.1 for details) overestimates $x \log x$ for $x > 1$. The rank correlations between goal-driven scores and their fast approximations remain high, *e.g.*, above $\rho > 0.96$ even for $M = 500$, which is sufficient for query selection.

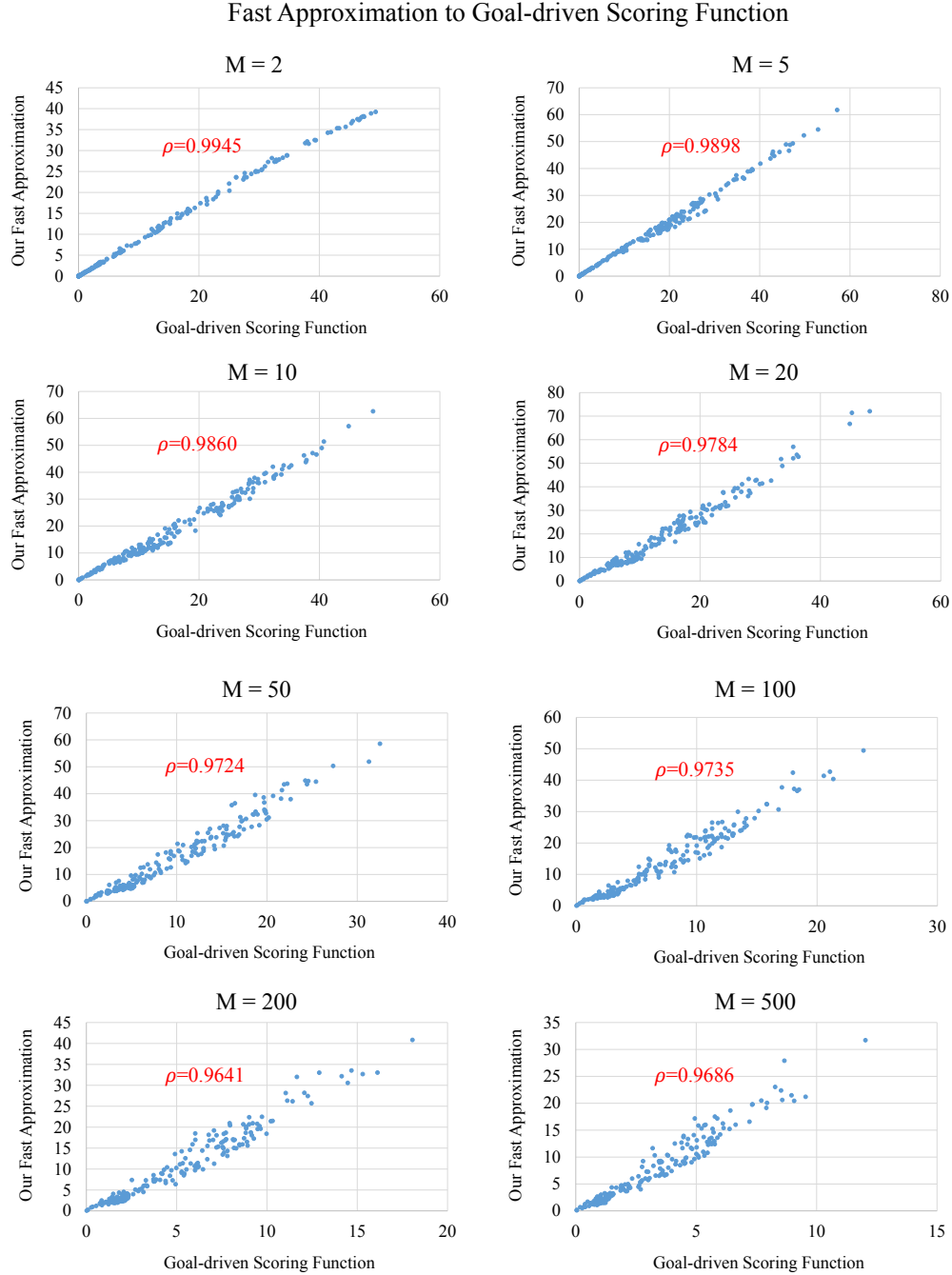
5.5 Discussion

In this work we discussed three active learning strategies – cramming (entropy), curiosity-driven learning and goal-driven learning – for Visual Question Answering using deep multi-modal neural networks. Our results show that deep VQA models require 30k - 50k training questions for active learning before they are able to ask informative questions and achieve better scaling than randomly selecting questions for labeling. Once the training set is large enough, several active learning strategies achieve significant savings in answer annotation cost. Our proposed goal-driven query strategy in particular, shows a significant advantage on improving performance when the downstream task involves answering a specific type of VQA questions.

Jointly selecting batches of examples as queries [106] and formulating active learning as a decision making problem [55] (greedily selecting the batch that reduces entropy by the most for the current iteration may not be the optimal decision) have been shown to improve optimality in active learning query selection. Combining those approaches with deep neural networks under the Bayesian Neural Network framework are promising future directions.

The pool-based active learning setup explored in this work selects unlabeled human generated question-image pairs and asks the oracle for answers. For building VQA datasets however, collecting human-generated questions paired with each image is also a substantial portion of the overall cost. Hence, starting from a bank of questions and an unaligned bank of images, and having the model decide which question it would like to pair with each image to use as a query would result in a further reduction in cost. Note that such a model would need to not only reason about the informativeness of a question-image pair, but also about the relevance

Figure 5.8: Our fast approximations using Eq. 5.9 versus the original goal-driven scores computed using Eq. 5.7 under $M = \{2, 5, 10, 20, 50, 100, 200, 500\}$ samples of model parameters. Our approximations have high rank correlation with scores computed using the original method.



of a question to the image [100, 80]. Evaluating such an approach would require collecting new VQA datasets with humans in the loop to give answers – which we show would require 30k - 50k answers before the model could start selecting informative images and questions. Going one step further, we could also envision a model that generates new questions rather than selecting from a pool of questions. That would require a generative model that can perform inference to optimize for the active learning objectives. We hope that our work serves as a foundation for these future research directions.

Acknowledgements

We thank Michael Cogswell and Qing Sun for discussions about the active learning strategies. This work was funded in part by an NSF CAREER award, ONR YIP award, Allen Distinguished Investigator award from the Paul G. Allen Family Foundation, Google Faculty Research Award, and Amazon Academic Research Award to DP. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

Chapter 6

Conclusion and Future Research Directions

In this dissertation we study leveraging common sense for vision and language tasks from multi-modal perspectives. Input images and text are first represented in multiple modalities (e.g. vision, text, abstract scenes and facts) for a rich set of perspectives. And then the perspectives are used for joint reasoning to make decisions for the target task. We explore grounding, imagination and question answering approaches for leveraging common sense for a variety of vision and language tasks, namely assessing the plausibility of commonsense tuples, solving fill-in-the-blank and paraphrasing questions, and matching images with captions. We show that leveraging common sense learnt from abstract scenes and Visual Question Answering is able to improve performance and interpretability and make more effective use of data.

Complementary to the model aspect, we also study the data aspect for improving common-sense learning, from the perspective of active learning. We study active learning for Visual Question Answering (VQA) where a model iteratively grows its knowledge through querying informative questions about images for answers. Drawing analogies from human learning, we explore cramming (entropy), curiosity-driven (expected model change), and goal-driven (expected error reduction) active learning approaches, and propose a new goal-driven scoring function for deep VQA models under the Bayesian Neural Network framework. We show that once trained with a large initial training set, a deep VQA model is able to effectively query informative question-image pairs for answers to improve itself through active learning, saving human effort on commonsense annotations. For the scenario where the model

is allowed to ask generic questions about images but is evaluated only on specific questions (*e.g.*, questions whose answer is either yes or no), our proposed goal-driven scoring function performs the best.

That also leads to many new research opportunities:

Improving human-AI collaboration with common sense. Gaining trust from users and collaborating with humans is crucial for AI systems for them to be effectively utilized. Building AI systems that can explain their decisions to humans (interpretability), that can improve their decisions using feedbacks from humans (repairability) and that makes predictions which humans are able to predict (predictability) are important aspects of improving human-AI collaboration. Learning and using commonsense knowledge as perspectives from multiple modalities provides unique opportunities to improve human-AI collaboration.

On interpretability, multimodal models provide the opportunity to explain decisions through many relevant modalities. A multimodal image-captioning system with question answering and abstract images as modalities may be able to answer questions about the image or drawing clipart illustrations about the captions to explain the captions that it generates about an image.

On repairability, multimodal models provide flexibility in feedback modalities. Users can provide feedback from multiple modalities. For example, humans may answer questions the model asks about an image, or by annotating cliparts with captions to improve image-captioning. Those feedbacks can be added new training examples for improving commonsense knowledge in modalities that corresponds to the feedbacks.

On predictability, learning common sense reduces the gap in inductive bias between human and AI. Mistakes made by AI systems with sufficient common sense would be closer to those that humans would make. Those mistakes would be easier for humans to predict.

Building machines that learn through question answering. Question answering is a natural interface for humans to interact with as well as to teach AI systems. Initial research above on active learning suggests that a machine with sufficient knowledge is capable of telling informative question-image pairs from non-informative ones in the VQA task, as well as choosing question-image pairs that are informative for a target task in mind. But as AI systems become more intelligent, they have learned enough about generic knowledge, and they will need to ask more fine-grained and more specific questions in order to gain knowledge. Picking questions to ask about images, generating questions to ask about images, asking follow-up questions or even drawing sketches and asking questions about the sketches

are progressively more informative and more challenging stages of learning through asking questions.

On the other hand, mentoring – rather than letting the AI system explore on its own – could also be an efficient form of learning. For the mentor (human or machine), tailoring a curriculum for the student would require combining curiosity from the student which we analyzed in our active learning work, with knowledge from the mentor. How to develop such a curriculum for question-answering is also an interesting open problem.

It is my hope that the framework of leveraging perspectives about images and text from multiple modalities to learn common sense, as well as our exploration on active learning for visual question answering will lay the foundation for many promising new directions in the intersection of vision, machine learning and AI, and build smarter machines that are able to interact with humans and learn multiple ways of thinking.

Bibliography

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433. IEEE, 2015. 13, 47, 50, 51, 54, 55, 60, 64, 71, 73, 80
- [2] S. Antol, C. L. Zitnick, and D. Parikh. Zero-shot learning via visual abstraction. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, pages 401–416. Springer, 2014. 10, 11, 19, 29, 31, 32, 51
- [3] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic. Seeing 3D chairs: Exemplar part-based 2D-3D alignment using a large dataset of CAD models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3762–3769. IEEE, 2014. 31
- [4] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse M-best solutions in Markov random fields. In *Proceedings of the 12th European Conference on Computer Vision (ECCV)*, pages 1–16. Springer, 2012. 46
- [5] A. C. Berg, T. L. Berg, H. Daumé, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and predicting importance in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3562–3569. IEEE, 2012. 9, 30
- [6] T. Berg and P. N. Belhumeur. POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 955–962. IEEE, 2013. 51
- [7] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1613–1622. PMLR, 2015. 73, 74, 76, 118

- [8] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1247–1250. ACM, 2008. 6, 9, 28
- [9] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, pages 168–181. Springer, 2010. 51
- [10] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, pages 438–451. Springer, 2010. 51
- [11] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an architecture for never-ending language learning. AAAI, 2010. 6, 8, 28
- [12] Y.-W. Chao, Z. Wang, R. Mihalcea, and J. Deng. Mining semantic affordances of visual object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4259–4267. IEEE, 2015. 9
- [13] O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with SVMs. *Information Retrieval*, 13(3):201–215, 2010. 36
- [14] D. Chen and C. Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750. ACL, 2014. 31
- [15] X. Chen and C. Lawrence Zitnick. Mind’s eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2422–2431. IEEE, 2015. 52
- [16] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1409–1416. IEEE, 2013. 9, 29, 46
- [17] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST)*, pages 103–111. ACL, 2014. 52, 53
- [18] H. T. Dang, D. Kelly, and J. J. Lin. Overview of the TREC 2007 question answering track. In *Proceedings of the 16th Text REtrieval Conference (TREC 2007)*, pages 105–122. NIST, 2007. 27, 30

- [19] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 75
- [20] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. *arXiv preprint arXiv:1703.06585*, 2017. 75
- [21] R. Davis, H. Shrobe, and P. Szolovits. What is a knowledge representation? *AI Magazine*, 14(1):17, 1993. 6
- [22] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. GuessWhat?! Visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 75
- [23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009. 52
- [24] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3270–3277. IEEE, 2014. 9, 29
- [25] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1271–1278. IEEE, 2009. 9, 31
- [26] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430. IEEE, 2015. 51
- [27] J. Donahue and K. Grauman. Annotator rationales for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1395–1402. IEEE, 2011. 51
- [28] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 647–655. PMLR, 2014. 62

- [29] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 601–610. ACM, 2014. 8
- [30] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2584–2591. IEEE, 2013. 51
- [31] D. Elliott and F. Keller. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 452–457. ACL, 2014. 30, 52
- [32] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. Mausam. Open information extraction: The second generation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, volume 1, pages 3–10. AAAI, 2011. 8, 11
- [33] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1785. IEEE, 2009. 51
- [34] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, pages 15–29. Springer, 2010. 30, 51
- [35] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 46
- [36] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty. Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79, 2010. 9, 27, 30
- [37] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single view geometry. *International Journal of Computer Vision*, 110(3):259–274, 2014. 9, 31
- [38] D. F. Fouhey and C. L. Zitnick. Predicting object dynamics in scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2019–2026. IEEE, 2014. 10, 29, 31, 51

- [39] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1050–1059. PMLR, 2016. 67, 73, 74, 76, 118
- [40] Y. Gal, R. Islam, and Z. Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1183–1192. PMLR, 2017. 73, 74, 77
- [41] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? Dataset and methods for multilingual image question. In *Proceedings of the 28th Advances in Neural Information Processing Systems (NIPS)*, pages 2296–2304, 2015. 47, 50, 51, 71
- [42] D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual Turing test for computer vision systems. volume 112, pages 3618–3623. National Academy of Sciences, 2015. 31, 47, 71, 73
- [43] J. Gordon and B. Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction (AKBC)*, pages 25–30. ACM, 2013. 6, 28, 72
- [44] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 47, 71, 73, 80
- [45] H. Grabner, J. Gall, and L. Van Gool. What makes a chair a chair? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1529–1536. IEEE, 2011. 9, 31
- [46] Y. Guo and R. Greiner. Optimistic active-learning using mutual information. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, volume 7, pages 823–829. AAAI, 2007. 79
- [47] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *Proceedings of the 10th European Conference on Computer Vision (ECCV)*, pages 16–29. Springer, 2008. 9, 30, 31, 51
- [48] J. Hamrick, P. Battaglia, and J. B. Tenenbaum. Internal physics models guide probabilistic judgments about object dynamics. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (CogSci)*. Cognitive Science Society, 2011. 31, 46, 51

- [49] J. Hays and A. A. Efros. IM2GPS: Estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 9, 30
- [50] V. Hedau, D. Hoiem, and D. Forsyth. Recovering free space of indoor scenes from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2807–2814. IEEE, 2012. 9, 31
- [51] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 54
- [52] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 68
- [53] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61, 2013. 6, 28
- [54] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011. 73, 74
- [55] S. Javdani, Y. Chen, A. Karbasi, A. Krause, D. Bagnell, and S. S. Srinivasa. Near optimal Bayesian active learning for decision making. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 430–438. PMLR, 2014. 89
- [56] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678. IEEE, 2015. 9, 51
- [57] B. Kaneva, A. Torralba, and W. T. Freeman. Evaluation of image features using a photorealistic virtual world. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2282–2289. IEEE, 2011. 31
- [58] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137. IEEE, 2015. 52, 60, 61, 62, 63, 68, 69
- [59] A. Khosla, B. An An, J. J. Lim, and A. Torralba. Looking beyond the visible scene. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3710–3717. IEEE, 2014. 9, 30

- [60] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Proceedings of the 27th Advances in Neural Information Processing Systems (NIPS)*, pages 3581–3589, 2014. 47
- [61] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 50, 52, 53, 60, 61, 63, 66, 68, 69
- [62] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using Fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4437–4446. IEEE, 2015. 60, 61, 63, 69
- [63] A. Kovashka, D. Parikh, and K. Grauman. WhittleSearch: Image search with relative attribute feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2973–2980. IEEE, 2012. 51
- [64] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, pages 301–320. Springer, 2016. 74
- [65] A. Krishnakumar. Active learning literature survey. Technical Report, University of California, Santa Cruz, 2007. 74
- [66] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012. 52
- [67] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby Talk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013. 30, 51
- [68] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011. 51
- [69] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 951–958. IEEE, 2009. 46, 51

- [70] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015. 6, 28
- [71] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Proceedings of the 23rd Advances in Neural Information Processing Systems (NIPS)*, pages 1378–1386, 2010. 51
- [72] J. J. Lim, R. R. Salakhutdinov, and A. Torralba. Transfer learning by borrowing examples for multiclass object detection. In *Proceedings of the 24th Advances in Neural Information Processing Systems (NIPS)*, pages 118–126, 2011. 31
- [73] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 11, 50, 51, 52, 60, 80
- [74] X. Lin and D. Parikh. Don’t just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2984–2993. IEEE, 2015. 3, 9, 10, 51
- [75] X. Lin and D. Parikh. Leveraging Visual Question Answering for image-caption ranking. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, pages 261–277. Springer, 2016. 3, 73
- [76] J. Lu, X. Lin, D. Batra, and D. Parikh. Deeper LSTM and normalized CNN Visual Question Answering model. https://github.com/VT-vision-lab/VQA_LSTM_CNN, 2015. xii, 71, 72, 76
- [77] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 29th Advances in Neural Information Processing Systems (NIPS)*, pages 289–297, 2016. xii, 71, 72
- [78] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. *AAAI*, 2016. 50, 51, 60, 61, 69
- [79] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2623–2631. IEEE, 2015. 52

- [80] A. Mahendru, V. Prabhu, A. Mohapatra, D. Batra, and S. Lee. The promise of premise: Harnessing question premises in Visual Question Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 937–946. ACL, 2017. 91
- [81] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the 27th Advances in Neural Information Processing Systems (NIPS)*, pages 1682–1690, 2014. 31, 50
- [82] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–9. IEEE, 2015. 47, 51, 71
- [83] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal Recurrent Neural Networks (m-RNN). In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015. 52, 60, 61, 69
- [84] T. Mensink, E. Gavves, and C. G. Snoek. COSTA: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2441–2448. IEEE, 2014. 28
- [85] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119, 2013. 7, 8, 16, 35
- [86] G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995. 6, 28
- [87] M. Minsky. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*, chapter 6. Simon and Schuster, 2007. 1
- [88] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 47
- [89] D. Mishkin, N. Sergievskiy, and J. Matas. Systematic evaluation of convolution neural network advances on the ImageNet. *Computer Vision and Image Understanding*, 161:11–19, 2017. 71
- [90] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. P. Spithourakis, and L. Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*, 2017. 75

- [91] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1802–1813. ACL, 2016. 75
- [92] D. Parikh and K. Grauman. Relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 503–510. IEEE, 2011. 51
- [93] A. Parkash and D. Parikh. Attributes for classifier feedback. In *Proceedings of the 12th European Conference on Computer Vision (ECCV)*, pages 354–368. Springer, 2012. 51
- [94] A. Peñas, C. Unger, and A.-C. N. Ngomo. Overview of CLEF question answering track 2014. In *Proceedings of the 5th International Conference of the CLEF Initiative*, pages 300–306. Springer, 2014. 31
- [95] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. ACL, 2014. 8
- [96] L. C. Pickup, Z. Pan, D. Wei, Y. Shih, C. Zhang, A. Zisserman, B. Scholkopf, and W. T. Freeman. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2035–2042. IEEE, 2014. 9, 30
- [97] H. Pirsiavash, C. Vondrick, and A. Torralba. Inferring the why in images. *arXiv preprint arXiv:1406.5472*, 2014. 9, 30, 31, 36, 51
- [98] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):74–93, 2017. 69
- [99] C. Quirk, P. Choudhury, J. Gao, H. Suzuki, K. Toutanova, M. Gamon, W.-t. Yih, L. Vanderwende, and C. Cherry. MSR SPLAT, a language analysis toolkit. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 21–24. ACL, 2012. 37
- [100] A. Ray, G. Christie, M. Bansal, D. Batra, and D. Parikh. Question relevance in VQA: Identifying non-visual and false-premise questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 919–924. ACL, 2016. 57, 91

- [101] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Proceedings of the 28th Advances in Neural Information Processing Systems (NIPS)*, pages 2953–2961, 2015. 47, 50, 51, 71
- [102] M. Richardson, C. J. Burges, and E. Renshaw. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 193–203. ACL, 2013. 31
- [103] N. Roy and A. McCallum. Toward optimal active learning through Monte Carlo estimation of error reduction. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 441–448. Morgan Kaufmann Publishers, 2001. 79
- [104] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1234–1241. IEEE, 2012. 51
- [105] F. Sadeghi, S. K. Kumar Divvala, and A. Farhadi. VisKE: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1456–1464. IEEE, 2015. 9, 51, 52
- [106] O. Sener and S. Savarese. A geometric approach to active learning for convolutional neural networks. *arXiv preprint arXiv:1708.00489*, 2017. 74, 89
- [107] B. Settles. Active learning literature survey. Computer Science Technical Report (TR1648), University of Wisconsin, Madison, 2010. 74, 77
- [108] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2821–2840, 2013. 31
- [109] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 52, 53, 54, 76
- [110] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu. Open mind common sense: Knowledge acquisition from the general public. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, pages 1223–1237. Springer, 2002. 6, 28

- [111] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Proceedings of the 26th Advances in Neural Information Processing Systems (NIPS)*, pages 935–943, 2013. 51
- [112] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Proceedings of the 28th Advances in Neural Information Processing Systems (NIPS)*, pages 3483–3491, 2015. 47
- [113] R. Speer and C. Havasi. ConceptNet 5: A large semantic network for relational knowledge. In *The People’s Web Meets NLP: Collaboratively Constructed Language Resource*, pages 161–176. Springer, 2013. 6, 25, 28
- [114] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 59
- [115] F. Strub, H. de Vries, J. Mary, B. Piot, A. Courville, and O. Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2765–2771. AAAI, 2017. 75
- [116] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th Advances in Neural Information Processing Systems (NIPS)*, pages 3104–3112, 2014. 52
- [117] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. IEEE, 2015. 52
- [118] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, and L. Bourdev. Improving image classification with location context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1008–1016. IEEE, 2015. 51
- [119] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528. IEEE, 2011. 68
- [120] C. Unger, C. Forascu, V. Lopez, A. Ngomo, E. Cabrio, P. Cimiano, and S. Walter. Question answering over linked data (QALD-4). In *Working Notes for CLEF 2014 Conference*, 2014. 27, 31

- [121] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575. IEEE, 2015. 52
- [122] R. Vedantam, X. Lin, T. Batra, C. Lawrence Zitnick, and D. Parikh. Learning common sense through visual abstraction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2542–2550. IEEE, 2015. 2, 51, 52
- [123] C. Vondrick, H. Pirsiaavash, and A. Torralba. Anticipating visual representations with unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 98–106. IEEE, 2016. 51
- [124] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3302–3309. IEEE, 2014. 51
- [125] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, pages 155–168. Springer, 2010. 51
- [126] J. Weston, S. Chopra, and A. Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 31
- [127] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 2048–2057. PMLR, 2015. 52
- [128] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 68
- [129] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual Madlibs: Fill in the blank image generation and question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2461–2469. IEEE, 2015. 50
- [130] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 729–736. IEEE, 2013. 51
- [131] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. PANDA: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE Conference on*

- Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1644. IEEE, 2014. 46, 51
- [132] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, and S.-C. Zhu. Beyond point clouds: Scene understanding by reasoning geometry and physics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3127–3134. IEEE, 2013. 31, 51
 - [133] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for Visual Question Answering. *arXiv preprint arXiv:1512.02167*, 2015. 51
 - [134] S. Zhou, Q. Chen, and X. Wang. Active deep networks for semi-supervised sentiment classification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling): Posters*, pages 1515–1523. ACL, 2010. 74
 - [135] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 58–65, 2003. 79
 - [136] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, pages 408–424. Springer, 2014. 9, 29, 31, 36, 46, 51
 - [137] Y. Zhu, C. Zhang, C. Ré, and L. Fei-Fei. Building a large-scale multimodal knowledge base for Visual Question Answering. *arXiv preprint arXiv:1507.05670*, 2015. 51
 - [138] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3009–3016. IEEE, 2013. 10, 29, 31, 32, 37, 45
 - [139] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1681–1688. IEEE, 2013. 10, 18, 19, 29, 30, 32, 40, 41
 - [140] C. L. Zitnick, R. Vedantam, and D. Parikh. Adopting abstract images for semantic scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):627–638, 2016. 10

Appendix A

Learning Common Sense Through Visual Abstraction

A.1 Extracting Tuples from Sentences

As described in Section 2.3.2, we build our VAL and TEST sets using the ReVerb information extraction system to extract our commonsense assertions. The ReVerb system segments the image into (typically) three chunks: primary object clause, relation clause and secondary object clause respectively. We do some post-processing to the ReVerb outputs to map them into our final t_P , t_R , and t_S tuples. We describe this post-processing below.

1. Get the Parts Of Speech (POS) tags for each input sentence.
2. Explore minor clauses in sentences by searching for one of the subordinating words ('because', 'although', 'unless', 'however', 'since') and extracting the shorter (minor) clause. In the minor clause, search for regular expression patterns: "*" is "*" to sample extra sentence chunks.
3. For all relation clauses, remove articles and pronoun instances.
4. For all relation clauses, remove the words "is" and "are".
5. For all primary and secondary clauses, remove pronouns, articles and adjectives.
6. Split to create new relations for each instance of "and". For example "Mike and Jenny *play* baseball" is converted to "Mike *play* baseball" and "Jenny *play* baseball"

7. Drop all relation clauses which contain a noun.
8. Perform lemmatization on all relation words. Lemmatization maps verbs to their root forms. Thus “plays” and “playing” are both mapped to “play”.
9. Convert all plural nouns occurring in primary and secondary clauses to singular form. Also remove all instances of words (‘group’, ‘couple’, ‘pair’, ‘bunch’, ‘crowd’, ‘team’, ‘two’, ‘three’, ‘four’, ‘five’).
10. Remove all clauses with empty primary clause, secondary clause or relation clause to get the tuples.

Appendix B

Leveraging Visual Common Sense for Non-Visual Tasks

B.1 Qualitative Results on Fill-in-the-blanks and Visual Paraphrasing

Figure B.1 to B.4 show qualitative results of our textual+visual approach on fill-in-the-blanks and visual paraphrasing.

Figure B.1: Qualitative results of fill-in-the-blanks, sampled based on predictions and ground truth.

• Scenario 1: human, text baseline and our approach are all correct.

Question

Mike kicked the soccer ball.
The duck is afraid of the soccer ball

Answers

Ground Truth: D
Human: D (8/10)
Text baseline: D
Vision + text: D

Original Scene

A. Jenny and Mike are angry at the dog.
B. The bear has a hamburger and drink.
C. The grill is next to the tree.
D. Jenny wants the soccer ball.

• Scenario 3: human and text baseline are correct while our approach is incorrect

Question

Jenny is petting the cat.
No one is on the riding toy.

Answers

Ground Truth: C
Human: C (8/10)
Text baseline: C
Vision + text: A

Original Scene

A. There is an apple tree behind Mike.
B. There are 3 hot dogs on the grill.
C. Mike is on the slide.
D. Jenny is happy to see Mike.

Question

Jenny is standing on the swing.
Mike is feeling sad.

Answers

Ground Truth: B
Human: B (5/10)
Text baseline: B
Vision + text: B

Original Scene

B. The sun is behind the tree.
C. Jenny is angry because it is raining on her.
D. Jenny is near balloons.

Question

The burger is on the table.
Jenny is standing next to table.

Answers

Ground Truth: D
Human: D (4/10)
Text baseline: D
Vision + text: B

Original Scene

A. Mike is flying a kite.
B. The dog is watching Jenny.
C. Jenny threw the frisbee.
D. Mike is standing next to table.

• Scenario 2: human and our approach are correct while text baseline is incorrect

Question

Jenny is in the sandbox
The cat and Jenny have not left room for Mike

Answers

Ground Truth: B
Human: B (9/10)
Text baseline: C
Vision + text: B

Original Scene

A. Mike sees a pie.
B. The cat is sitting next to Jenny.
C. Mike and Jenny are sitting next a fire.
D. Jenny is playing in the sandbox.

• Scenario 4: human is correct while text baseline and our approach are incorrect

Question

Jenny is holding a pink ball.
Mike threw the beach ball.

Answers

Ground Truth: D
Human: D (7/10)
Text baseline: C
Vision + text: A

Original Scene

A. Mike is sitting next to the tree.
B. There are three hamburgers on the grill.
C. A rocket ship is flying in the sky.
D. Jenny has a pink shovel.

Question

Mike and Jenny are scared of the duck.
Happy duck walks away.

Answers

Ground Truth: B
Human: B (5/10)
Text baseline: A
Vision + text: B

Original Scene

A. Mike was wearing his crown in the sandbox.
B. The ball hits the duck.
C. The sun is shining.
D. Mike is helping Jenny.

Question

Jenny and Mike are fighting.
They are both wearing silly hats

Answers

Ground Truth: A
Human: A (5/10)
Text baseline: D
Vision + text: D

Original Scene

A. Mike is holding a beach ball
B. Mike is wearing the hat.
C. The dog is watching Mike.
D. Jenny kicked the football.

Figure B.2: Figure B.1 continued. Qualitative results of fill-in-the-blanks, sampled based on predictions and ground truth.

• Scenario 5: our approach and text baseline are correct while human is incorrect

Question

The duck is near the soccer ball.
Jenny is sitting near the slide.


Answers

Ground Truth: A
Human: B (8/10)
Text baseline: A
Vision + text: A

Original Scene



A. Mike is standing under the hot air balloon



B. Mike is sitting next to the dog.



C. The snake is sliding behind Mike.



D. Mike is very surprised.



• Scenario 7: text baseline is correct while human and our approach are incorrect

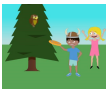
Question

Jenny is jumping up and down.
Mike is holding a frisbee.

Answers

Ground Truth: A
Human: B (7/10)
Text baseline: A
Vision + text: B

Original Scene



A. Mike is wearing his viking hat.



B. Mike and Jenny are camping



C. The rocket is soaring in the sky.



D. Jenny told the bear to leave.



Question

Mike is holding the ball.
Mike is playing with the cat.

Answers

Ground Truth: A
Human: B (4/10)
Text baseline: A
Vision + text: A


Original Scene



A. Mike is wearing sun glasses.



B. Jenny is sitting next to her juice.



C. The bear is roaring angrily.



D. The duck is in the sandbox.




Question

Mike is playing in the sandbox.
Jenny wants to play with Mike.

Answers

Ground Truth: C
Human: D (4/10)
Text baseline: C
Vision + text: D

Original Scene



A. Red apples grow on the tree.



B. Mike is near Jenny.



C. The sun is shining on Mike and Jenny.



D. The pink shovel is on Jenny's lap.



• Scenario 6: our approach is correct while human and text baseline are incorrect


Question

Mike is wearing a hat.
Jenny is holding the pizza.

Answers

Ground Truth: D
Human: C (7/10)
Text baseline: B
Vision + text: D

Original Scene



A. Jenny is trying to catch the soccer ball



B. Mike is holding the shovel.



C. Mike and Jenny are happy.



D. Mike is sitting on the grass.



• Scenario 8: human, text baseline and our approach are all incorrect

Question

Jenny is wearing a crown waving her hand.
The airplane is flying towards a giant cloud.


Answers

Ground Truth: D
Human: A (9/10)
Text baseline: A
Vision + text: A

Original Scene



A. Mike is wearing a pirate hat.



B. Mike is near the swings.



C. Mike has a baseball bat.



D. Mike is happily kicking the soccer ball.



Question

Mike is sitting on the grass.
Jenny is standing by the table.

Answers

Ground Truth: C
Human: D (5/10)
Text baseline: D
Vision + text: C

Original Scene



A. Mike is king for a day



B. Jenny is angry at Mike.



C. Jenny is holding a pizza.



D. Mike is wearing a viking hat.



Question

Jenny is upset she lost her balloons.
Jenny is standing next to the cat.


Answers

Ground Truth: D
Human: C (4/10)
Text baseline: B
Vision + text: B

Original Scene



A. The airplane will not disturb them.



B. Mike is angry that the dog is not listening.



C. The cat is sitting by Jenny.



D. Jenny is afraid the rocket will hit the balloon.



Figure B.3: Qualitative results of visual paraphrasing, sampled based on predictions and ground truth.

<p>• Scenario 1: human, text baseline and our approach are all correct.</p>				<p>• Scenario 3: human and text baseline are correct while our approach is incorrect</p>			
Original Scene(s)	Descriptions	Generated Scenes	Answers	Original Scene(s)	Descriptions	Generated Scenes	Answers
	The bucket is in the sandbox. Mike runs to the ball. Mike is wearing a baseball cap.		Ground truth Yes Human 1.3753 Text baseline 1.221 Vision + Text 2.0805		Mike and Jenny are having a barbecue. Jenny is excited to see a dog. Mike is angry at the dog for begging.		Ground truth Yes Human 1.3753 Text baseline 0.3909 Vision + Text -0.1280
	The bucket is in the sandbox. Mike runs to the ball. Mike is wearing a baseball cap.				Jenny is sitting on the ground. Mike does not like his hamburger. The dog is wearing a blue collar		
	Mike loves throwing the tennis ball. There is a cat looking at Mike. Mike is playing with the cat.		Ground truth Yes Human 4.2825 Text baseline 1.9647 Vision + Text 2.1077		The cool dog is wearing sunglasses. The cat is jealous of the dog. Mike and Jenny play on the slide.		Ground truth Yes Human 1.3753 Text baseline 0.0509 Vision + Text -0.6838
	Mike tries to play catch with the cat. The cat does not want to play catch. Mike threw the tennis ball to the cat.				Mr. Dog is cool in sunglasses. Mike bumps into Jenny. Jenny is surprised by Mr. Dog.		
	Mike is holding a hot dog Jenny is carrying ketchup. Jenny is running.		Ground truth No Human -3.0058 Text baseline -2.2792 Vision + Text -2.5399		It is raining on Jenny. Mike wants Jenny's lunch. Jenny is giving Mike her wet lunch.		Ground truth No Human -1.5452 Text baseline -0.0278 Vision + Text 0.2061
	Mike and Jenny are standing on the picnic table. Mike and Jenny are afraid of the bear. The owl is standing on the beach ball.				Jenny has a blue cap. Mike has a viking helmet. There are 2 trees.		
	The bucket is in the sandbox. Mike runs to the ball. Mike is wearing a baseball cap.		Ground truth No Human -3.0058 Text baseline -1.0911 Vision + Text -1.3115		Jenny wears sunglasses Mike catches the football Jenny is wearing a witch's hat		Ground truth No Human -1.5452 Text baseline -0.6850 Vision + Text 0.1486
	The bucket is in the sandbox. Mike runs to the ball. Mike is wearing a baseball cap.				Mike is kicking the ball. Jenny wants to catch the ball. Jenny is smiling at Mike.		
<p>• Scenario 2: human and our approach are correct while text baseline is incorrect</p>				<p>• Scenario 4: human is correct while text baseline and our approach are incorrect</p>			
	Mike is angry because Jenny won't play. Jenny is crying because Mike is mean. The owl watches the two children argue.		Ground truth Yes Human 1.3753 Text baseline -0.1311 Vision + Text 0.2123		Mike is shooing the dog away. Jenny is waiting for a hamburger. The balloon flies over the playground.		Ground truth Yes Human 4.2825 Text baseline -0.1836 Vision + Text -0.3634
	The helicopter is flying above Jenny. Mike wants Jenny's Frisbee. Jenny is crying because Mike is mad.				Mike is cooking the burger. The dog is standing next to the pit. Jenny is sitting in the grass.		
	It is raining on the tent. Jenny is sitting on the ground. Mike is very mad.		Ground truth Yes Human 2.7909 Text baseline -0.1274 Vision + Text 0.2949		Mike is wearing a beanie cap. The dog wants to eat the hamburger. Jenny is happy to see Mike.		Ground truth Yes Human 2.7909 Text baseline -0.4538 Vision + Text -0.4682
	Jenny is sitting n the grass. Mike is angry with a dog. There is a burger on the grill				Mike is wearing a funny hat Jenny is laughing at Mike's hat Jenny is sitting next to the table		
	A lightning bolt flashes in the sky. Jenny is wearing a crown. Mike is shouting at Jenny.		Ground truth No Human -3.0058 Text baseline 0.2635 Vision + Text -0.2044		Jenny stood next to the fire. The dog watched the hamburgers on the grill. Mike flew into the sky with the mustard on his shirt.		Ground truth No Human -1.5452 Text baseline 1.7038 Vision + Text 1.2092
	Jenny is singing on the swingset. Mike is happy to see Jenny at the park. The hot air balloon is high in the sky.				Mike is near a grill. A dog is near jenny. there are three hot-dogs on the grill.		
	Jenny is running from a snake. Mike is chasing after the snake. It is raining on Jenny.		Ground truth No Human -3.0058 Text baseline 0.1347 Vision + Text -0.5795		Mike is wearing a blue cap. Jenny is wearing a sunglasses. Jenny and Mike are playing catch.		Ground truth No Human -1.5452 Text baseline 0.5427 Vision + Text 0.2067
	Jenny and Mike are afraid of the snake. Jenny is playing with a bat. Mike is jumping up.				Mike is wearing a funny hat. Jenny is jumping off the ground. Mike is scared of something.		

Figure B.4: Figure B.3 continued. Qualitative results of visual paraphrasing, sampled based on predictions and ground truth.

• Scenario 5: our approach and text baseline are correct while human is incorrect					• Scenario 7: text baseline is correct while human and our approach are incorrect				
Original Scene(s)	Descriptions	Generated Scenes	Answers		Original Scene(s)	Descriptions	Generated Scenes	Answers	
	Mike is chasing Jenny. Jenny loves to play on the swings. The big tree is planted in the park.		Ground truth Yes Human -1.5452 Text baseline 0.5894 Vision + Text 0.6304			Mike is holding a hot dog Jenny is carrying ketchup. Jenny is running.		Ground truth Yes Human -1.5452 Text baseline 0.6291 Vision + Text -0.1716	
	The duck is walking towards Mike and Jenny. Mike threw the soccer ball. Jenny is sitting in the grass.		Ground truth Yes Human -1.5452 Text baseline 0.8277 Vision + Text 1.2425			Rain is falling from the cloud. The dog is standing in front of Mike. Mike is wearing sunglasses.		Ground truth Yes Human -1.5452 Text baseline 0.0348 Vision + Text -0.0688	
	Jenny is upset. Jenny doesn't like cats. The dog will cheer Jenny up.		Ground truth No Human 1.3753 Text baseline -0.0449 Vision + Text -0.2418			The dog is on the table. Mike has a hamburger. Jenny has a drink.		Ground truth No Human 1.3753 Text baseline -0.3248 Vision + Text 0.1170	
	Mike is wearing a hat. The bear is roaring at Mike. Mike is in front of a tree.		Ground truth No Human 1.3753 Text baseline -1.1950 Vision + Text -1.1451			Lightning is coming out of the cloud. Mike and Jenny are angry. Mike is playing with a beach ball.		Ground truth No Human 1.3753 Text baseline -0.0142 Vision + Text 0.8637	
	Mike is wearing a pirate hat. Jenny is holding her drink.					Mike is throwing the frisbee. Jenny is throwing the ball. The dog is standing next to the tree.		Ground truth Yes Human -1.5452 Text baseline -0.0217 Vision + Text -0.3078	
• Scenario 6: our approach is correct while human and text baseline are incorrect					• Scenario 8: human, text baseline and our approach are all incorrect				
Original Scene(s)	Descriptions	Generated Scenes	Answers		Original Scene(s)	Descriptions	Generated Scenes	Answers	
	Jenny is upset. Jenny doesn't like cats. The dog will cheer Jenny up.		Ground truth Yes Human -1.5452 Text baseline -0.0771 Vision + Text 0.6696			There is a lightning in the sky. Jenny is running from Mike. Mike is chasing Jenny.		Ground truth Yes Human -3.0058 Text baseline -0.6132 Vision + Text -0.3347	
	Mike and Jenny are sitting on the ground. Two balls are on the ground. Mike is next to the slide.		Ground truth Yes Human -3.0058 Text baseline -0.0863 Vision + Text 0.1524			Mike and Jenny play on the swings. The dog watches Mike on the swing. The tall tree looks pretty.		Ground truth No Human 1.3753 Text baseline 1.1652 Vision + Text 1.0543	
	Jenny is sitting in the grass. Mike is wearing a Vikings hat. Jenny is very surprised.		Ground truth No Human 1.3753 Text baseline 0.2037 Vision + Text -0.0009			Jenny is kicking a ball. Jenny is wearing sunglasses. Mike is smiling.		Ground truth No Human 4.2825 Text baseline 0.0234 Vision + Text 0.1555	
	Mike is wearing a pirate hat. Jenny is wearing a funny hat. A dog is looking for something in the grass.		Ground truth No Human 1.3753 Text baseline 0.3193 Vision + Text -0.1845						

Appendix C

Leveraging Visual Question Answering for Image-Caption Ranking

C.1 Qualitative Examples

Fig. C.1 shows additional qualitative examples of image retrieval and caption retrieval using our $N = 3,000$ score-level fusion model (VQA-aware) and the baseline VQA-agnostic model (VQA-agnostic).

C.2 Information of (Q, A) Pairs

Given an image I , we propose to rank a set of N candidate (Q, A) pairs by how informative their validity $V_1, V_2, \dots, V_N \in \{true, false\}$ is to selecting a caption C for image I from a set of K captions $\{C_1, C_2, \dots, C_K\}$.

We compute information with mutual information $\mathbb{I}(V_i; C)$ between the validity V_i of the i -th (Q, A) pair (Q_i, A_i) and the caption C . By the definition of mutual information

$$\mathbb{I}(V_i; C) = \sum_{v \in \{true, false\}} \sum_k P(V_i = v, C = C_k) \log \frac{P(V_i = v, C = C_k)}{P(V_i = v)P(C = C_k)} \quad (\text{C.1})$$

In order to compute mutual information we need the following three probabilities.

- $P(V_i = v)$: how likely (Q_i, A_i) is true (valid) or false (invalid) given I . We compute that using the prediction from the VQA model.

$$\begin{aligned} P(V_i = \text{true}) &= P(A_i|Q_i, I) \\ P(V_i = \text{false}) &= 1 - P(A_i|Q_i, I) \end{aligned} \tag{C.2}$$

- $P(C = C_k)$: how likely C_k is the chosen caption of I . We compute that using the prediction from our image-caption ranking model.

$$P(C = C_k) = P_{cap}(C_k|I) \tag{C.3}$$

- $P(V_i = v, C = C_k)$: the joint probability of the validity of (Q_i, A_i) and caption C .

Computing the joint probability is an interesting problem. The proposed fusion model is purely feed-forward, so once we have hidden layer activations, the validity of a (Q, A) pair and the caption C are independent. In other words, once we learn a fusion model and feed in the input image/captions/ (Q, A) pairs, the hidden layer activations are already determined, and then V_i and C are independently predicted.

But does that imply V_i and C are independent? No. Because their independence is *conditioned* on the fusion model. From a bayesian perspective, computing the joint probability of V_i and C properly would require marginalizing over the fusion model parameters.

Let Θ be the parameters of the fusion model (includes parameters for both VQA and image captioning). We rewrite the joint probability $P(V_i = v, C = C_k)$ as taking expectation over fusion model parameters Θ to marginalize it out:

$$P(V_i = v, C = C_k) = \mathbb{E}_{\theta \sim P(\Theta)} P(V_i = v, C = C_k | \Theta = \theta) \tag{C.4}$$

Given model parameters, the hidden layer activations are determined and the model would independently predict V_i and C . So we assume that V_i and C are independent given fusion model parameters Θ . Therefore

$$\begin{aligned}
& P(V_i = v, C = C_k) \\
&= \mathbb{E}_{\theta \sim P(\Theta)} P(V_i = v, C = C_k | \Theta = \theta) \\
&= \mathbb{E}_{\theta \sim P(\Theta)} P(V_i = v | \Theta = \theta) P(C = C_k | \Theta = \theta) \\
&= \begin{cases} \mathbb{E}_{\theta \sim P(\Theta)} P(A_i | Q_i, I, \theta) P_{cap}(C_k | I, \theta) & , v = true \\ \mathbb{E}_{\theta \sim P(\Theta)} (1 - P(A_i | Q_i, I, \theta)) P_{cap}(C_k | I, \theta) & , v = false \end{cases}
\end{aligned} \tag{C.5}$$

Marginalizing over all possible model parameters is an intractable task. However recent progress on variational methods for neural networks allows us to compute an approximation. Previous works [39, 7] have established a theoretical foundation that setting dropout layers in neural networks to training mode at test time can be interpreted as sampling from a variational approximation to the posterior model parameter distribution $P(\Theta)$. They also showed that the dropout distribution can be leveraged to approximately compute model uncertainty. In this work we follow the same framework but instead use dropout to approximate joint probability and mutual information.

We compute a monte carlo approximation of Eq. C.5 by sampling θ from the dropout distribution as $P(\Theta)$. Specifically, we

- 1) Sample model parameters θ by setting dropout to training mode;
- 2) Use the same θ to make both VQA and image-caption ranking predictions;
- 3) Average over θ samples to approximate joint probability $P(V_i = v, C = C_k)$.

To our knowledge this is the first study on using dropout to compute joint probability and mutual information between two tasks.

After the three probabilities are computed, they then go into Eq. C.1 to compute mutual information between (Q_i, A_i) and caption C . And then, the (Q_i, A_i) pair that is selected is the most informative one for a given image I in terms of determining which caption is the best match for the image.

In our experiments, image I is randomly selected from the image-caption ranking test set and we use all its $K = 1,000$ candidate captions as $\{C_1, C_2, \dots, C_K\}$. We select the (Q, A) with the highest mutual information from $N = 3,000$ (Q, A) pairs randomly selected from VQA training set. We draw 5,000 dropout samples of θ to approximately compute mutual information. Qualitative examples of selected (Q, I) pairs for examples images is shown in Figure.5 in the main paper.

Figure C.1: Qualitative results of image retrieval and caption retrieval at rank 1, 2 and 3 using our $N = 3,000$ score-level fusion VQA-aware model and the baseline VQA-agnostic model. The true target images and captions are highlighted.


















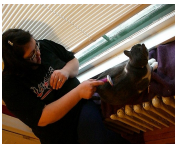































Image Retrieval				Caption Retrieval			
	Rank 1	Rank 2	Rank 3		Rank 1	Rank 2	Rank 3
A man with a red helmet on a small moped on a dirt road.	VQA-agnostic 				VQA-agnostic A woman holding food in a napkin and posing for a bite.	A woman in a bright pink summer shirt smiles and displays a party platter she has made.	A smiling woman standing next to a plate of food she made.
	VQA-aware 				VQA-aware A young girl smiles while enjoying her meal.	A little girl is sitting at a table.	Little girl smiles for the camera as she eats her sandwich.
A zebra standing on the ground with little scattered grass.	VQA-agnostic 				VQA-agnostic A sandwich has lettuce, tomato, as well as other items.	A plate of food containing a sandwich and a salad.	This sandwich has a side of salad on the plate.
	VQA-aware 				VQA-aware A plate of food containing a sandwich and a salad.	A meal at a restaurant of a salad, a toasted sandwich and a pickle.	This sandwich has a side of salad on the plate.
A man and a woman are posing for a photograph.	VQA-agnostic 				VQA-agnostic A couple of people sitting on a bench next to a dog.	A woman is giving her dog a bath.	A man standing next to a dog on the ground.
	VQA-aware 				VQA-aware A woman that is sitting down near a cat.	A man that is laying down underneath a cat.	A woman on a couch with a cat.
A laptop is on a table with a frosty beverage nearby.	VQA-agnostic 				VQA-agnostic Young girl in dress standing on wooden floor in residential home.	An man standing in a kitchen with a small puppy.	A man in the kitchen standing with his dog.
	VQA-aware 				VQA-aware A man in the kitchen standing with his dog.	A woman and a little dog in a very large kitchen.	A man is at a kitchen counter by a dog.

Figure C.2: Figure C.1 continued. Qualitative results of image retrieval and caption retrieval at rank 1, 2 and 3 using our $N = 3,000$ score-level fusion VQA-aware model and the baseline VQA-agnostic model. The true target images and captions are highlighted.

Image Retrieval				Caption Retrieval			
	Rank 1	Rank 2	Rank 3		Rank 1	Rank 2	Rank 3
Two small children standing at a sink brushing their teeth.	VQA-agnostic 				VQA-agnostic A cat laying in front of a bathroom mirror. .	The black cat is alert, lying in front of the bathroom sink.	A large cat stands inside of a clean bathroom sink.
	VQA-aware 				VQA-aware A grey and white cat lays in a sink.	A cat sitting in the sink in the bathroom.	A cute kitty cat in the sink of a bathroom near a brush and other items.
A young boy posing with a baseball bat in hand.	VQA-agnostic 				VQA-agnostic A white plate holding a piece of cheese cake on table. .	A bowl with a piece of cake in it next to a spoon.	A plate holding a grilled cheese sandwich and bowl of soup.
	VQA-aware 				VQA-aware A bowl with a piece of cake in it next to a spoon.	A spoon next to a dessert inside of a bowl.	A green plate sitting on a table with a piece of half eaten food on it.
A couch and ottoman are shown with remotes .	VQA-agnostic 				VQA-agnostic A group of skiers are gathered together as they get ready to ski.	Two people that are standing beside one another while wearing snow skis.	A group of people have backpacks as they stand on snow skis in the snow.
	VQA-aware 				VQA-aware Two people that are standing beside one another while wearing snow skis.	A group of people have backpacks as they stand on snow skis in the snow.	Two people posing on a mountain wearing skis.

Appendix D

Active Learning for Visual Question Answering: An Empirical Study

D.1 Fast Approximation of Goal-driven Scoring Function

In Section 5.3.2, we discuss our proposed goal-driven query strategy that minimizes uncertainty (entropy) on answers A'_t to a given set of unlabeled test question-image pairs $(Q'_t, I'_t), t = 1, 2, \dots, T$, against which the model will be evaluated. It queries (Q, I) pairs which maximize:

$$\begin{aligned} s_{goal}(Q, I) &= \sum_t \mathbb{H}(A'_t) - \mathbb{H}(A'_t|A) \\ &= \sum_t \mathbb{I}(A; A'_t) \\ &= \sum_t \sum_a \sum_{a'} P(A = a, A'_t = a' | Q, I, Q'_t, I'_t) \log \frac{P(A = a, A'_t = a' | Q, I, Q'_t, I'_t)}{P(A = a | Q, I) P(A'_t = a' | Q'_t, I'_t)} \end{aligned} \quad (\text{D.1})$$

Recall that we propose an approximation for term $P(A = a, A'_t = a' | Q, I, Q'_t, I'_t)$ as follows:

$$\begin{aligned}
& P(A = a, A'_t = a' | Q, I, Q'_t, I'_t) \\
&= \mathbb{E}_{\boldsymbol{\omega}} P(A = a | Q, I, \boldsymbol{\omega}) P(A'_t = a' | Q'_t, I'_t, \boldsymbol{\omega}) \\
&\approx \mathbb{E}_{\boldsymbol{\omega} \sim q_{\theta}(\boldsymbol{\omega})} P(A = a | Q, I, \boldsymbol{\omega}) P(A'_t = a' | Q'_t, I'_t, \boldsymbol{\omega})
\end{aligned} \tag{D.2}$$

Let us define four matrices $\mathbf{M}_1, \mathbf{D}_1, \mathbf{M}_2(t), \mathbf{D}_2(t)$ as follows:

$$\mathbf{M}_1 = \begin{bmatrix} P(A = a_1 | Q, I, \boldsymbol{\omega}_1) & P(A = a_2 | Q, I, \boldsymbol{\omega}_1) & \dots & P(A = a_J | Q, I, \boldsymbol{\omega}_1) \\ P(A = a_1 | Q, I, \boldsymbol{\omega}_2) & P(A = a_2 | Q, I, \boldsymbol{\omega}_2) & & P(A = a_J | Q, I, \boldsymbol{\omega}_2) \\ \vdots & & \ddots & \vdots \\ P(A = a_1 | Q, I, \boldsymbol{\omega}_M) & P(A = a_2 | Q, I, \boldsymbol{\omega}_M) & \dots & P(A = a_J | Q, I, \boldsymbol{\omega}_M) \end{bmatrix} \tag{D.3}$$

$$\mathbf{D}_1 = \text{Diag} \left(\begin{bmatrix} P(A = a_1 | Q, I) & P(A = a_2 | Q, I) & \dots & P(A = a_J | Q, I) \end{bmatrix} \right) \tag{D.4}$$

$$\mathbf{M}_2(t) = \begin{bmatrix} P(A'_t = a_1 | Q'_t, I'_t, \boldsymbol{\omega}_1) & P(A'_t = a_2 | Q'_t, I'_t, \boldsymbol{\omega}_1) & \dots & P(A'_t = a_J | Q'_t, I'_t, \boldsymbol{\omega}_1) \\ P(A'_t = a_1 | Q'_t, I'_t, \boldsymbol{\omega}_2) & P(A'_t = a_2 | Q'_t, I'_t, \boldsymbol{\omega}_2) & & P(A'_t = a_J | Q'_t, I'_t, \boldsymbol{\omega}_2) \\ \vdots & & \ddots & \vdots \\ P(A'_t = a_1 | Q'_t, I'_t, \boldsymbol{\omega}_M) & P(A'_t = a_2 | Q'_t, I'_t, \boldsymbol{\omega}_M) & \dots & P(A'_t = a_J | Q'_t, I'_t, \boldsymbol{\omega}_M) \end{bmatrix} \tag{D.5}$$

$$\mathbf{D}_2(t) = \text{Diag} \left(\begin{bmatrix} P(A'_t = a_1 | Q'_t, I'_t) & P(A'_t = a_2 | Q'_t, I'_t) & \dots & P(A'_t = a_J | Q'_t, I'_t) \end{bmatrix} \right) \tag{D.6}$$

Here \mathbf{M}_1 is an $M \times J$ matrix, \mathbf{D}_1 is a $J \times J$ matrix, $\mathbf{M}_2(t)$ is an $M \times J$ matrix and $\mathbf{D}_2(t)$ is a $J \times J$ matrix. With $\mathbf{M}_1, \mathbf{D}_1, \mathbf{M}_2(t), \mathbf{D}_2(t)$ we could rewrite Eq. D.2 in matrix form:

$$\begin{aligned}
& \begin{bmatrix} P(A = a_1, A'_t = a_1 | Q, I, Q'_t, I'_t) & \dots & P(A = a_1, A'_t = a_J | Q, I, Q'_t, I'_t) \\ \vdots & \ddots & \vdots \\ P(A = a_J, A'_t = a_1 | Q, I, Q'_t, I'_t) & \dots & P(A = a_J, A'_t = a_J | Q, I, Q'_t, I'_t) \end{bmatrix} \\
&\approx \frac{1}{M} \mathbf{M}_1^T \mathbf{M}_2(t)
\end{aligned} \tag{D.7}$$

Let $\text{Sum}(\cdot)$ be an operator on a matrix that sums up all elements in that matrix. Combining Eq. D.1 and Eq. D.7, our goal-driven scoring function can be approximately computed as follows:

$$\begin{aligned}
s_{\text{goal}}(Q, I) &= \sum_t \sum_a \sum_{a'} P(A = a, A'_t = a' | Q, I, Q'_t, I'_t) \log \frac{P(A = a, A'_t = a' | Q, I, Q'_t, I'_t)}{P(A = a | Q, I) P(A'_t = a' | Q'_t, I'_t)} \\
&\approx \sum_t \text{Sum} \left\{ \left[\frac{1}{M} \mathbf{M}_1^T \mathbf{M}_2(t) \right] \circ \log \left[\frac{1}{M} \mathbf{D}_1 \mathbf{M}_1^T \mathbf{M}_2(t) \mathbf{D}_2(t) \right] \right\} && \text{Rewriting in matrix form.} \\
&\approx \sum_t \frac{1}{2} \text{Sum} \left\{ -\frac{1}{M} \mathbf{M}_1^T \mathbf{M}_2(t) + \frac{1}{M^2} [\mathbf{M}_1^T \mathbf{M}_2(t)] \circ [\mathbf{D}_1 \mathbf{M}_1^T \mathbf{M}_2(t) \mathbf{D}_2(t)] \right\} && x \log x \approx \frac{1}{2}(-x + x^2). \\
&= \sum_t -\frac{1}{2} + \frac{1}{2M^2} \text{Sum} \left\{ [\mathbf{M}_1^T \mathbf{M}_2(t)] \circ [\mathbf{D}_1 \mathbf{M}_1^T \mathbf{M}_2(t) \mathbf{D}_2(t)] \right\} && \text{Sum of } P(A, A'_t) \text{ reduces to 1.} \\
&= \sum_t -\frac{1}{2} + \frac{1}{2M^2} \text{Tr} \left\{ [\mathbf{M}_1^T \mathbf{M}_2(t)] [\mathbf{D}_1 \mathbf{M}_1^T \mathbf{M}_2(t) \mathbf{D}_2(t)]^T \right\} && \text{Sum}(A \circ B) = \text{Tr}(AB^T). \\
&= \sum_t -\frac{1}{2} + \frac{1}{2M^2} \text{Tr} [\mathbf{M}_1^T \mathbf{M}_2(t) \mathbf{D}_2(t) \mathbf{M}_2^T(t) \mathbf{M}_1 \mathbf{D}_1] \\
&= \sum_t -\frac{1}{2} + \frac{1}{2M^2} \text{Tr} [\mathbf{M}_1 \mathbf{D}_1 \mathbf{M}_1^T \mathbf{M}_2(t) \mathbf{D}_2(t) \mathbf{M}_2^T(t)] && \text{Property of trace.} \\
&= \sum_t -\frac{1}{2} + \frac{1}{2M^2} \text{Sum} \left\{ [\mathbf{M}_1 \mathbf{D}_1 \mathbf{M}_1^T] \circ [\mathbf{M}_2(t) \mathbf{D}_2(t) \mathbf{M}_2^T(t)] \right\} && \text{Tr}(AB^T) = \text{Sum}(A \circ B). \\
&= \sum_t -\frac{1}{2} + \frac{1}{2} \mathbb{E}_{\omega} \mathbb{E}_{\omega'} \left[\sum_a \frac{P(A = a | Q, I, \omega) P(A = a | Q, I, \omega')}{P(A = a | Q, I)} \right. && \text{Rewriting in probability form.} \\
&\quad \left. \sum_a \frac{P(A'_t = a | Q'_t, I'_t, \omega) P(A'_t = a | Q'_t, I'_t, \omega')}{P(A'_t = a | Q'_t, I'_t)} \right] \\
&= \frac{1}{2} \mathbb{E}_{\omega} \mathbb{E}_{\omega'} \left[\sum_a \frac{P(A = a | Q, I, \omega) P(A = a | Q, I, \omega')}{P(A = a | Q, I)} \right. && \text{Rearranging summation.} \\
&\quad \left. \sum_t \sum_a \frac{P(A'_t = a | Q'_t, I'_t, \omega) P(A'_t = a | Q'_t, I'_t, \omega')}{P(A'_t = a | Q'_t, I'_t)} \right] - \sum_t \frac{1}{2}
\end{aligned} \tag{D.8}$$

Which is Eq. 5.9 in Section 5.3.2.

As stated in Section 5.3.2, the above equation can be computed as a dot-product between two vectors of length M^2 . One vector is matrix $\frac{1}{M} \mathbf{M}_1 \mathbf{D}_1 \mathbf{M}_1^T$ expanded into a vector. It only involves pool questions (Q, I) . The other vector is $\frac{1}{M} \sum_t \mathbf{M}_2(t) \mathbf{D}_2(t) \mathbf{M}_2^T(t)$ expanded into a vector. It only involves test questions (Q'_t, I'_t) and it is shared for all pool questions (Q, I) , so it can be precomputed for all (Q, I) . Precomputing $\frac{1}{M} \sum_t \mathbf{M}_2(t) \mathbf{D}_2(t) \mathbf{M}_2^T(t)$ for test questions has a time complexity of $O(TJM^2)$. Note that $\mathbf{D}_2(t)$ is a diagonal matrix, so multiplying $\mathbf{D}_2(t)$ with $\mathbf{M}_2^T(t)$ only takes $O(JM)$ operations. In the same way, comput-

ing $\frac{1}{M}\mathbf{M}_1\mathbf{D}_1\mathbf{M}_1^T$ for all (Q, I) has a time complexity of $O(UJM^2)$. The time complexity of their dot product for all (Q, I) is merely $O(UM^2)$. So the overall time complexity is $O(\max(U, T)JM^2)$. The overall time complexity is linear to both dataset size U and T and the number of possible answers J , so our approach can easily scale to very large datasets and more VQA answers.