UNIVERSITY OF CALIFORNIA SAN DIEGO

Leveraging Human Reasoning to Understand and Improve Visual Question Answering

A thesis submitted in partial satisfaction of the
requirements for the degree of Master of Science

in

Computer Science

by

Hammad Abdullah Ayyubi

Committee in charge:

Professor Garrison W. Cottrell, Chair
Professor Manmohan Chandraker
Professor David J. Kriegman

2020

The Thesis of Hammad Abdullah Ayyubi is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am grateful to Prof. Gary Cottrell, who supported me not only in my thesis, but throughout my Master's studies. His unwavering support, especially when things weren't working out, was a critical resource to bank on during my research.

I would also like to thank Prof. David Kriegman and Prof. Manmohan Chandraker. It was a privilege and honor to get the opportunity to work with them. Their insightful suggestions and advice were always important in steering the course of my research.

I owe special thanks to my collaborator and friend Mehrab Tanjim, with whom I have worked on many projects and papers. It's hard to imagine getting through the uncertainties of research without his company. In the same spirit, I want to thank my labmates.

I need to make a special mention of my collaborators from SRI as well - Yi Yao and Ajay Divakaran. I learned important aspects of research and writing academic papers from them. Their guidance and mentorship can't be overstated.

I am also grateful to a number of professors at UC San Diego - Prof. Julian McAuley, Prof. Taylor Kirkpatrick, Prof. Ndapa Nakashole, Prof. Kamalika Chaudhuri, Prof. Lawrence Saul and Prof. Chris Gopal. I have been extremely fortunate to get the opportunity to work with so many amazing mentors and researchers. Their guidance has played a critical role in shaping my perspective and attitude.

Most importantly, I would like to thank my parents and my family, without whose immense sacrifices and support, nothing was possible. Lastly, I would also like to thank my friends: Iftekhar Ahmad, Ahmad Naguib, Ahmed Taha, Ahmed Ismail, Mohammad Yaseen, Mohammad Salah El-Hadri, Daniel Quinnel and Anmol Popli, who provided me with the all important social structure in a foreign country, which allowed me to thrive academically.

Chapter 2, in full, is a reprint of the material as it appears in arXiv:2004.02032. Hammad A. Ayyubi*, Md. Mehrab Tanjim*, Julian J. McAuley and Garrison W. Cottrell. "Generating Rationales in Visual Question Answering". The thesis author was the first co-author of this paper.

VITA

2012-2016 B. Tech in Electrical Engineering, Indian Institute of Technology, BHU

2016-2017 Citicorp Services India Pvt. Ltd.

2017-2018 Soroco India Pvt. Ltd.

2018-2020 M.S. in Computer Science, University of California San Diego

PUBLICATIONS

Hammad A Ayyubi, Yi Yao, Ajay Divakaran, "Progressive Growing of Neural ODEs", ICLR Workshop on Integration of Deep Neural Models and Differential Equations, 2020.

Hammad A. Ayyubi*, Md. Mehrab Tanjim*, Julian McAuley and Garrison W. Cottrell, "Generating Rationales in Visual Question Answering", arXiv:2004.02032.

ABSTRACT OF THE THESIS

Leveraging Human Reasoning to Understand and Improve Visual Question Answering

by

Hammad Abdullah Ayyubi

Master of Science in Computer Science

University of California San Diego, 2020

Professor Garrison W. Cottrell, Chair

Visual Question Answering (VQA) is the task of answering questions based on an image. The field has seen significant advances recently, with systems achieving high accuracy even on open-ended questions. However, a number of recent studies have shown that many of these advanced systems exploit biases in datasets, text of the question or similarity of images in the dataset.

To study these reported biases, proposed approaches seek to identify areas of images or words of the questions as evidence that the model focuses on while answering questions. These mechanisms often tend to be limited as the model can answer incorrectly while focusing on the correct region of the image or vice versa.

In this thesis, we seek to incorporate and leverage human reasoning to improve interpretability of these VQA models. Essentially, we train models to generate human-like language as evidence or reasons/rationales for the answers that they predict. Further, we show that this type of system has the potential to improve the accuracy on VQA task itself as well.

Chapter 1

Introduction

The task of Visual Question Answering (VQA) is formulated as follows: given a question (Q), an image (I), the model must predict the correct answer (A). This answer, A, generally needs to be chosen from apriori defined set of possible choices. For example, in fig. 1.1, given the top left image and the question, *"What is the moustache made of?"*, the model must answer "Banana".

This task is challenging on many levels. First of all, the model needs to understand the text of the question and the visual signals from image. Secondly, it should correctly correlate text with the visual signals. On top of understanding text and visual signals, the model also needs to use commonsense reasoning, knowledge base reasoning, identify context and setting of the image.

Recently proposed systems have started to achieve remarkable success in these multi-modal settings (Chen et al. [2019], Lu et al. [2019]). However, multiple studies have shown that these models could be exploiting biaes in questions, images or datasets. For example, in fig. 1.2 left, Cadene et al. [2019] show that the model memorizes from the training dataset that bananas are associated with yellow color, without actually understanding its color. Hence, when provided with a green banana during test time, it still predicts its color to be yellow. Similarly, Agrawal et al. [2016b] show, in fig. 1.2 right, that the model predicts the car has a license plate, even without looking at the image.

Figure 1.1. Visual Question Answering (VQA), image from Agrawal et al. [2016b]

Existing methods propose to identify such biases by providing evidence of parts of question or image that the model attends to while predicting the answer (Shah et al. [2019], Agrawal et al. [2016a], Das et al. [2017], Goyal et al. [2016]). Such approaches are limiting because the model can still predict the wrong answer despite attending to the correct parts and vice versa.

We seek to address these issues by proposing the task of generating reasons or rationales, as a means to provide evidence for what the model bases its decisions on. The ability to generate correct reasons/rationales requires first and foremost correct understanding and comprehension of text (question) and visual (image) modalities and their correct correlation. Besides, it calls

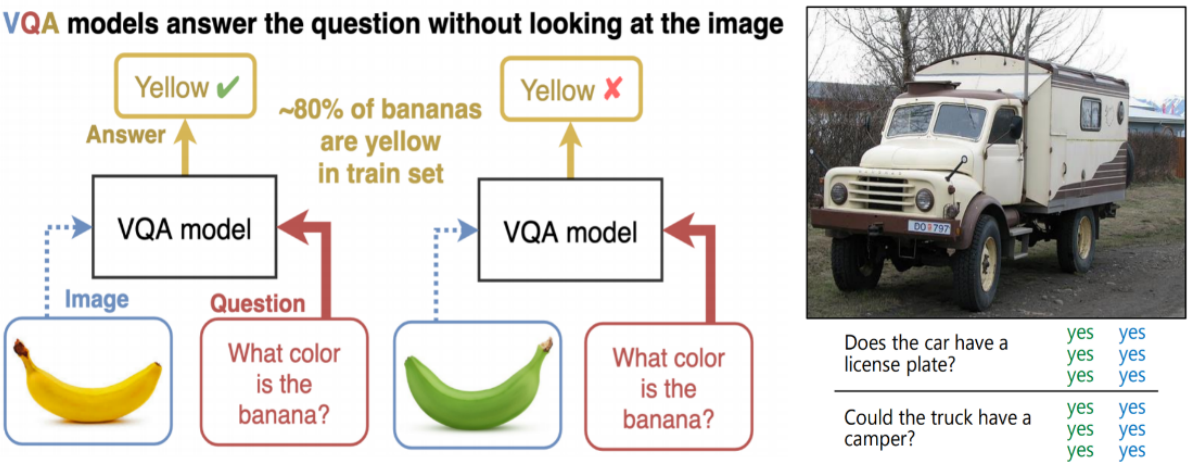


Figure 1.2. VQA bias. Left: figure from Cadene et al. [2019] show that the model learns biases from train dataset and hence can't generalize to a green banana during test. Right: figure from Agrawal et al. [2016b] show that model can answer question without even looking at the image.

into effect the reasoning abilities of the model as well. Another advantage with such a sytem is that the evidence is directly human readable. We don't need to engineer another system to extract evidences.

We seek to study such human readable evidence generation in two settings:

- Rationale Generation: In this setting, the model first predicts the answer from question and image, and then justifies its answer through a rationale as a post hoc measure.
- Reason Generation: The model, in this case, generates a reason from question and the image. Then, it uses this generated reason along with question and image to arrive at the answer.

Further, we show through experiments that such a reasoning and rationalization ability has the potential to improve upon the main task of VQA itself.

Chapter 2

Generating Rationales in Visual Question Answering

2.1 Introduction

Visual Question Answering (VQA) (Agrawal et al. [2016b]) tasks are an important assessment of joint language-vision understanding. To perform well on VQA, a model must understand the given question and then find a relevant answer from the image. A great deal of success has been achieved in this task with state-of-the-art models (Chen et al. [2019]) achieving high accuracy on challenging VQA datasets (Goyal et al. [2017], Ren et al. [2015]).

However, a critical question is how well these models "understand" the image, questions, and the answers that they are predicting. Are they just exploiting biases in the questions (Ramakrishnan et al. [2018], Johnson et al. [2017a], Cadene et al. [2019]), images (Agrawal et al. [2018], Goyal et al. [2017]) or the data (Jabri et al. [2016], Manjunatha et al. [2018])? Answering these questions can help shed light on the limitations of existing VQA approaches, and could also lead to more interpretable/explainable VQA systems.

It is a non-trivial task to evaluate a model's simultaneous understanding of the three components (questions, images, and answers). Previous work has analyzed models' understanding of questions (Shah et al. [2019], Agrawal et al. [2016a]), images (Das et al. [2017], Goyal et al. [2016]) and answers (Chandrasekaran et al. [2018]) individually. They have done so by perturbing words (language modality) or investigating heat-maps of images (vision modality).

A joint measure of question, image and answer understanding requires an approach which can simultaneously understand and test both linguistic and visual modalities.

To address the above challenges, we propose the novel task of generating rationales in VQA as a measure of model's comprehensive understanding. This task not only requires the model to understand the questions (linguistic modality) and the images (visual modality), but it also requires the model to rationalize the predicted answer in relation to the question and image.

As we need gold standard rationales to compare to the generated rationales, we use the dataset provided by the Visual Commonsense Reasoning (VCR) task (Zellers et al. [2018]). This dataset contains questions, images, multiple-choice answers (four choices) and four options for rationales, out of which one is correct. We train models for this particular VQA task in this dataset: choose a correct answer from four options, given a question and image. Then, we task the model with generating rationales for the answers they predict. We compare the generated rationale against the ground-truth rationale from the dataset, as a measure of the model's comprehensive ability.

We employ this approach to investigate one of the leading models on the VCR task—ViLBERT (Lu et al. [2019]). Further, we propose a way to explicitly inject commonsense understanding into the model by jointly training ViLBERT and the language model, GPT-2 (Radford et al. [2019]) on the multi-task objective of predicting the answers and generating rationales. The idea is that by backpropagating the loss of rationale generation into the answer prediction model, sound reasoning can be injected to improve model's comprehensive understanding.

2.2 Approach

The proposed task is illustrated in Figure 2.1. We approach our task of generating rationale by breaking it into two essential components: first calculating a predicted answer embedding, E_{A_p}, from the VQA model (pretrained ViLBERT) by providing it with the given image I and the question Q; and second, feeding the predicted answer embedding E_{A_p} to a

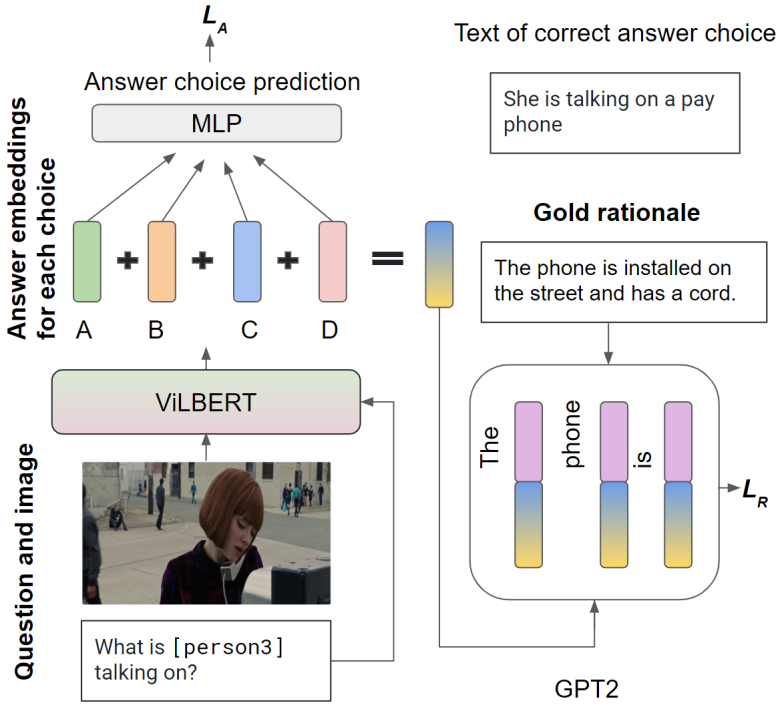


Figure 2.1. High-level overview of the proposed task. We ask the VQA model to generate rationale for the answer it is predicting.

language model (pretrained GPT2).

2.2.1 Predicted Answer Embedding

Our VQA task is represented in terms of a question Q, an image I, and four answer choices $A_1 \ldots A_4$, among which the model must choose the correct option.

The models approach this task by outputting an embedding for each answer options, E_{A_i}. The four answer embeddings are later passed through a linear-softmax layer to predict answer scores. So, if f is a VQA model (e.g., ViLBERT) parameterized by θ, then the embeddings and softmax scores are calculated as follows:

$$E_{A_i} = f(Q, I, A_i; \theta) \quad \forall i \in \{1, \ldots, 4\} \tag{2.1}$$

$$s_i = Softmax(Linear(E_{A_i}; \theta_l)) \quad \forall i \in \{1, , \ldots, 4\} \tag{2.2}$$

The act of choosing the most probable answer embedding out of four options will make the network non-differentiable. To address this issue, we calculate the predicted answer embedding by taking the average of each answer option embedding E_{A_i} weighted by their softmax scores:

$$E_{A_p} = \sum_{i=1}^{4} E_{A_i} \times s_i \tag{2.3}$$

2.2.2 Generating Rationales

We formulate the task of generating rationale as conditional language generation, conditioned on the predicted answer embedding and previously generated rationale tokens. Specifically, if $r = r_1, \ldots, r_n$ is the rationale, and E_{A_p} is the predicted answer embedding, we maximize the following log likelihood:

$$\log(P(r)) = \sum_{i=1}^{n} \log(P(r_i | E_{A_p}, r_1, \ldots, r_{i-1}))$$

Here, r_i is the i^{th} token of the rationale. The language model is then fine-tuned using the gold standard rationale for the corresponding visual question-answer from the VCR dataset.

2.3 Experiments

We use the Visual Commonsense Reasoning (VCR) dataset (Zellers et al. [2018]) since (in addition to visual questions and answers) this dataset includes rationales. The dataset has 290,000 multiple choice questions derived from 110,000 movie scenes. We report all results on the validation set as the test set labels are not available while the VCR challenge is ongoing.

We use ViLBERT (Lu et al. [2019]) as the reference VQA model. For the language model, we use the 124 million parameter pretrained GPT-2 (small) (Radford et al. [2019]). We use a batch size of 32, initial learning rate of 2e-5 and train the models for 20 epochs for all our experiments.

2.3.1 Evaluating VQA Model Understanding

Since we want to investigate how well the VQA reference model (ViLBERT) already "understands" the image, the question and the answer, we freeze the pretrained weights of the model. We extract predicted answer embeddings using eq. (2.3), and generate rationales using GPT-2, conditioned on this answer embedding. We fine-tune GPT-2 using the ground-truth rationale from the dataset. We call this model ViLBERT-Fr (ViLBERT-Frozen).

2.3.2 Injecting Commonsense into VQA

In this setting, we want to explicitly enforce commonsense understanding in the VQA model *while* predicting the answer. We follow the same procedure as in Section 2.3.1, except the weights of ViLBERT are fine-tuned as well. We train ViLBERT with GPT-2 in an end-to-end fashion with the dual loss of answer prediction, $\mathscr{L}_{\mathscr{A}}$ (via a cross-entropy loss) and rationale generation, $\mathscr{L}_{\mathscr{R}}$ (via a causal language modeling loss). The final loss is:

$$\mathscr{L} = \lambda \mathscr{L}_{\mathscr{A}} + \mathscr{L}_{\mathscr{R}} \tag{2.4}$$

where λ is the weight fine-tuned during our experiments.

2.3.3 Evaluation Metric

We use BLEU (Papineni et al. [2002]) and ROUGE (Lin [2004]) to compare generated and gold standard rationales. In addition to these n-gram metrics, we are also interested in comparing the semantic similarity of rationales. We follow Huang [2018] and calculate sentence embeddings using the InferSent model proposed by Conneau et al. [2017], followed by cosine similarity measurement to compare generated rationales with the gold standard.

2.3.4 Results

Multi-task objective

Since we are dealing with multiple losses (and objectives) of answer prediction and rationale generation (eq. (2.4)), we report a comparative study on different losses we explored in

Table 2.1. Comparison of Losses. λ is from eq. (2.4), var = Homoscedastic Uncertainty loss and kldiv = KL-Divergence regularizer added.

Loss	VQA Accuracy	BLEU-1
$\lambda = 1$	70.18	11.24
$\lambda = 3$	69.96	10.78
$\lambda = 10$	70.24	10.55
$\lambda = 1000$	69.84	**11.27**
var	70.19	11.15
kldiv (ViLBERT-Ra)	**70.45**	**11.27**

Table 2.1.

Weighted Losses: We vary the λ in eq. (2.4) as 1, 3, 10 and 1000.

Uncertainty Loss (var): We weight the losses $\mathscr{L}_{\mathscr{A}}$ and $\mathscr{L}_{\mathscr{R}}$ by considering the homoscedastic uncertainty of each task as in Cipolla et al. [2018].

KL-Divergence (kldiv): We add Kullback–Leibler divergence (Kullback and Leibler [1951]) loss between predicted answer scores eq. (2.2) and answer scores from pretrained ViLBERT as an added regularizer. This was done to prevent our model from diverging too much from the trained ViLBERT on the original VQA task.

We see from Table 2.1 that the model trained with KL-Divergence loss performs best on both the VQA task and rationale generation task. As such, we do all further comparison with this model and name it ViLBERT-Ra (ViLBERT-Rationale).

Quantitative Results

Rationale generation: We compare the performance of ViLBERT-Fr and our model, ViLBERT-Ra in Table 2.2. We see that our model consistently out performs ViLBERT-Fr over both n-gram metrics – BLEU and ROUGE and semantic similarity measurement metric – cosine similarity. This demonstrates how we can leverage rationale generation task to improve model comprehension abilities of existing VQA models.

VQA task: We also compare the performance of the two models on the original VQA task in Table 2.1. Since we trained all our models with a batch size of 32 (due to limited compute

Table 2.2. Comparison of generated rationale vs gold standard rationale on the validation set of VCR dataset.

Metrics	Models	
	ViLBERT-Fr	ViLBERT-Ra
BLEU-1	8.92	**11.27**
BLEU-4	0.56	**0.68**
ROUGE-1	13.52	**17.08**
ROUGE-L	11.28	**14.15**
Cosine Similarity	0.57	**0.60**
VQA Accuracy	69.58	**70.45**

resources), we ran a control experiment to train ViLBERT with the same batch size instead of original 64 for fair comparison.

We find that ViLBERT-Ra gives superior performance on the rationale generation task without compromising accuracy on the original VQA task (Table 2.2). In fact, VQA performance is slightly improved. This suggests that training the model to generate rationales can improve model's comprehension which in turn can lead to better answer prediction judgement.

Qualitative Results

Human Evaluation: We presented 100 randomly selected samples from the VCR validation set containing an image, question, correct answer and rationales generated by ViLBERT-Fr and ViLBERT-Ra to human evaluators. The generated rationales were shuffled randomly to hide which rationale came from which model. We then asked them to choose which of the two candidate rationales better explains the answer in the given sample. The results are summarized in Table 2.3. Humans consistently rate the rationales generated by ViLBERT-Ra as better explanations for the answers.

We show an illustrative example of a rationale generated by the two models in Figure 2.2. During training, we replaced tags like [person1], [person2] with random names, so in both rationales we can observe random names being generated. However, we note that our model was able to generate key words of the gold rationale and convey the relevant meaning. We have

Type	Text
Question	What are the occupations of [person1], [person2], [person3], and [person4] ?
Answer	[person1], [person2] and [person3] , and [person4] are among the others military officers.
Gold Rationale	They are all wearing decorated military uniforms.
ViLBERT-Fr	Avon is about to cross the range to save Osiris
ViLBERT-Ra	Myrl Tommie are wearing military uniforms.

Figure 2.2. Generated rationales samples. Highlighted portions show key points of the rationale for the given answer.

Table 2.3. Percent of rationales preferred by three human judges. Here, H stands for Human.

Models	H-1	H-2	H-3	Majority Voting
ViLBERT-Fr	31	29	38	20
ViLBERT-Ra	**69**	**71**	**62**	**80**

provided more such examples in the attached appendix.

2.4 Related Work

Evaluating VQA model comprehension.

VQA models (Malinowski and Fritz [2014], Malinowski et al. [2015], Jiang et al. [2018], Su et al. [2019], Li et al. [2019], Alberti et al. [2019], Lu et al. [2019]) have achieved remarkable results on multiple challenging datasets (Ren et al. [2015], Agrawal et al. [2016b], Goyal et al. [2017]). This naturally raises the question of whether the models are exploiting biases (Ramakrishnan et al. [2018], Johnson et al. [2017a], Cadene et al. [2019], Jabri et al. [2016],

Manjunatha et al. [2018]) or genuinely answering the questions. Researchers have employed various kinds of attention mechanisms over words and images to point out sections of images and words that the model attends to while answering the question (Das et al. [2017], Goyal et al. [2016], Agrawal et al. [2016a]). Another set of approaches use various kinds of 'selection' tasks as a means to interpret models. Goyal et al. [2017] propose picking another image for the same question that has a different answer. Berg and Belhumeur [2013] propose selecting visual facts (image regions) from the image while Zellers et al. [2018] propose picking one rationale from a set of four options.

An orthogonal set of methods Andreas et al. [2016], Hu et al. [2017], Mascharka et al. [2018], Vedantam et al. [2019] approach this task by generating symbolic programs to reason about the question. We note that such an approach would quickly become intractable given the size of the symbol vocabulary required to cover free-form rationale generation, as we are considering.

In contrast to all discussed approaches, we propose generating rationales as a means of interpreting models.

Generating rationales.

The task of generating explanations has previously been employed by Hendricks et al. [2016] to explain fine-grained bird recognition decisions. In our case, we explain answers to visual questions with rationales. Rajani et al. [2019a] generate reasons and rationales to explain question answering tasks, but only in a purely textual mode.

To the best of our knowledge no prior work has proposed the task of generating rationales as a measure of evaluating comprehensive understanding of images, questions and answers in VQA models.

2.5 Conclusion

In this paper, we proposed the novel task of generating rationales as a measure of model understanding for Visual Question Answering tasks. A well-reasoned explanation implies a thorough understanding of all components of the task: the image, the question and the answer. We further proposed an end-to-end training method to improve the model's commonsense understanding. We demonstrated the effectiveness of our proposed method through quantitative and qualitative results.

Chapter 2, in full, is a reprint of the material as it appears in arXiv:2004.02032. Hammad A. Ayyubi*, Md. Mehrab Tanjim*, Julian J. McAuley and Garrison W. Cottrell. "Generating Rationales in Visual Question Answering". The thesis author was the first co-author of this paper.

Chapter 3

Generating Reasons for Understanding and Improving VQA

3.1 Introduction

Visual Question Answering (VQA) (Agrawal et al. [2016b]) tasks require complex reasoning over both text and vision modalities. For example, in fig. 3.1, the task entails understanding the context and dynamics of interaction between [person2] and [person3], before being able to predict that they do indeed know each other. Deep Learning model often learn to predict answers directly from raw data (image and text) in a black box, end-to-end fashion. This obscures the reasoning path taken by the model. The resulting abstraction raises a number of question about their interpretability, trustworthiness and biases. We propose an approach which incorporates the reasoning path into the model prediction, thus making this process explicit and interpretable.

Existing methods which seek to explicit this reasoning process in their models can be broadly categorized into three approaches: (a) Using attention maps, (b) Using neuro-symbolic methods, (c) Using human explanations. Method that use attention maps in their reasoning process, either point to region or patches of image or words in question that the model attends to while answering the question (Das et al. [2017], Goyal et al. [2016], Agrawal et al. [2016a]), or seek to align the image attention of model to a ground truth image region provided by an external source such as humans (Wang et al. [2020], Wu and Mooney [2019]).

Neuro-symbolic methods extract executable programs from question and execute it on

Figure 3.1. Visual Question Answering (VQA) task from Zellers et al. [2018] requires model to predict the correct answer (A) from four given choices, given question (Q) and image (I). This task is abbreviated as $Q \rightarrow A$.

images (Vedantam et al. [2019]) or on a visual scene graph extracted from the image (Hudson and Manning [2019]). Such approaches explicit the reasoning process in the form of executable programs which trace out the reasoning path taken by them.

The last category of reasoning approaches employed is using explanations in the form of human language, thus making the explanations grounded in real world knowledge and directly interpretable. These approaches justify the predicted answer with an explanation in the form of human language (Patro et al. [2020], Li et al. [2018b]). All these methods present their justification in a post-hoc manner, i.e. they first predict the answer and then justify their answer using explanations. In contrast, our proposed method first generates a reason and then uses this generated reason to predict the answer.

In particular, we use the question and image to generate the reason. In the next step, we use these generated reasons along with image and question to predict the answer. Such an approach clearly explicit the reasoning process in the form of generated reason. We hypothesize that this can be a critical approach to address the trust and bias issues of VQA systems. This has the added advantage of being directly interpretable since it's in the form of human language, thus eliminating the need for any third party tool for model investigation. Further, we show that such reasoning architectures has the potential to improve upon the original VQA task itself as well.

3.2 Related Work

Reasoning using Attention Maps

Attention maps have been widely used to justify and provide evidence for the answer predicted by the model. Das et al. [2017], Goyal et al. [2016] point to patches in the image region where the VQA model attends to while predicting the answer. Agrawal et al. [2016a] also add attention over words in question. Further, other approaches use these attention maps more actively. They utilize ground truth attention maps provided by humans or obtained heuristically to improve and refine answer prediction further (Wang et al. [2020], Wu and Mooney [2019]).

Reasoning using Neuro-Symbolic Methods

Symbolic artificial intelligence represent real world objects as symbols and map their relations using rules. They use this as a knowledge base to answer question about the world they have built. As such, they are naturally suited for reasoning tasks. However, mapping symbols from real world objects is a manual process, hence these symbols are not directly co-related with real world objects. On the other hand, neural methods are adept at perceiving world and identifying objects. As such, neuro-symbolic methods seek to combine the best of both worlds by using neural methods to identify/perceive objects from real world, mapping it to symbols, and using symbolic reasoning methods to answer questions on top of the learned world. Johnson et al. [2017b], Vedantam et al. [2019], Hudson and Manning [2019] represent a typical way of employing such an approach.

Reasoning using Explanations in Human Language form

Due to the black box nature of deep learning models, investigating the reasons or decisions taken by it can be difficult at times. As such, certain methods seek to simplify this process by training the models to explain themselves in human readable form. Li et al. [2018b], Park et al. [2018] train VQA models to justify the answers that they predict with an explanation in human language form. However, these approaches use explanation in a post-hoc manner, i.e. explanations are generated after predicting the answer. Our proposed approach, on the other

hand, first generates a reason and then uses the generated reason to predict the answer.

The closest work to our approach is Rajani et al. [2019b], who generate a reason to arrive at the answer as well, but they do it in text only mode. Also, they use a language model to generate reason while we use an encoder-decoder framework. We note that Li et al. [2018a] do use explanation for predicting answer. However, they generate the explanations from image only. As such, it's a description or caption of the image. In contrast, we generate reason conditioned on image and the question, and as such it's relative to the particular question. In addition, we use the image during answer prediction as well unlike their approach.

3.3 Approach

3.3.1 Preliminaries

Objective

The underlying task, that we are addressing, is Visual Question Answering (VQA). Given a question (Q) and an image (I), the model must predict the correct answer (A_p) from four given answer choices ($A_i \forall i \in \{1,\ldots,4\}$). We propose to generate reason first using Q and I, and then use the generated reason along with Q and I to predict the answer. As such, for training, we need question, image and answer triplets along with the associated reasons.

Annotated Reason

Since we need a large dataset of question, image and answer pairs along with reasons, we turn to the Visual Commonsense Reasoning (VCR) dataset (Lu et al. [2019]). Each sample in the dataset consists of a question (Q), image (I), four answer choices ($A_i \forall i \in \{1,\ldots,4\}$), out of which one is correct (A_c), and four rationale choices ($R_i \forall i \in \{1,\ldots,4\}$), out of which one is correct (R_c). There are two sub-tasks associated with the dataset. The first requires models to predict answer (A) from question and image, i.e. ($Q \rightarrow A$). The second asks the models to predict rationale (R) for the correct answer, i.e. ($QA_c \rightarrow R$). We work on the first $Q \rightarrow A$ task and use the provided correct rationales (R_c) as reasons (Re).

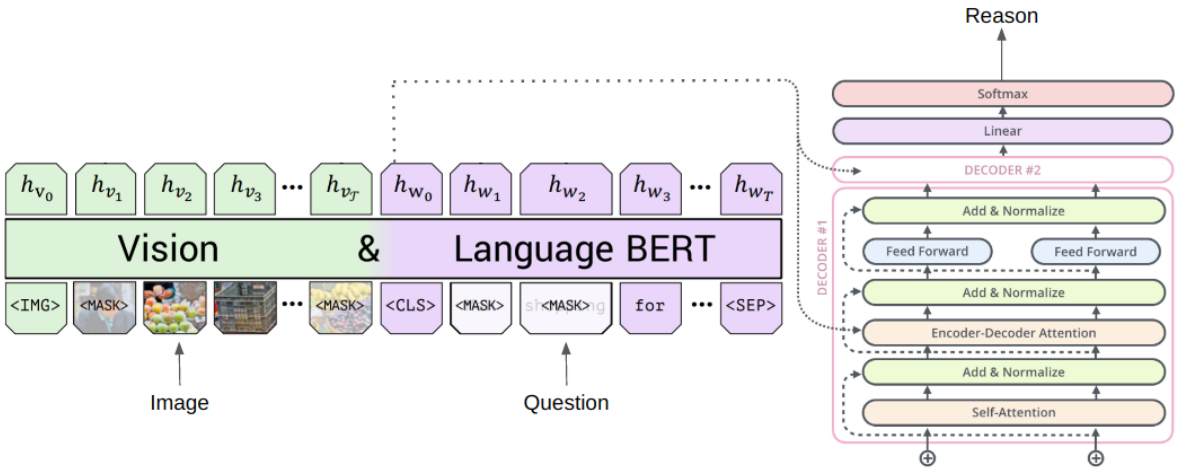


Figure 3.2. Our encoder decoder framework to generate reason from question and image. Encoder network is pretrained ViLBERT (Lu et al. [2019]) model and decoder network is a pretrained transformer.

3.3.2 Proposed Method

We propose to employ a two stage training method. First, we use question (Q) and image (I) to generate reason (Re), i.e. $QI \rightarrow Re$. Then, in the second phase, we use the generated reasons along with question and image to predict the answer our of four given choices ($A_i \forall i \in \{1,\ldots,4\}$), i.e. $QIRe \rightarrow A$.

Generating Reasons

In the first phase of training, we use an encoder decoder framework, as illustrated in fig. 3.2, to generate Re from Q and I. For encoder, we use the recently proposed ViLBERT model (Lu et al. [2019]). ViLBERT model consists of two BERT (Devlin et al. [2018]) transformer blocks, one each for text modality and vision modality. The two transformer blocks interact with each other using co-attention blocks. In co-attention blocks, the key and value associated with text query comes from vision block and vice versa for vision query. This kind of co-attention produces extremely useful and meaningful vision-linguistic representations which have shown to perform exceedingly well on downstream tasks.

For decoder, we experiment with two different kind of transformer architectures, namely

BERT and GPT2 (Radford et al. [2019]). We share the visual-linguistic features from encoder at all your layers of decoder through cross-attention as done by Vaswani et al. [2017].

We input image and question to the ViLBERT encoder and generate reason through the decoder. We train the whole network using cross entropy loss between each generated and ground truth token. Additionally, we use teacher forcing in the decoder network. For both encoder and decoder network, we use pretrained models provided by the original authors and fine tune all layers during training. Once trained, we generate reasons for all examples in training and validation set and store them to be used by the answer prediction network, as described in next section.

Predicting Answer from generated reason

We use another pretrained ViLBERT model for predicting answers from the given question and image, and the generated reason. The question is appended by the generated reason using the '[SEP]' token, followed by an answer option A_i, again separated by '[SEP]' token. This forms four different pairs, one for each answer option A_i. These four pairs are fed to text transformer stream of ViLBERT and the image is fed to the vision transformer stream. The final text and vision representation is combined, followed by a linear layer to predict final probability scores for each of the four pairs.

3.4 Experiments

3.4.1 Dataset

We use the Visual Commonsense Reasoning (VCR) dataset (Lu et al. [2019]) for all our datasets. The dataset has 290,000 multiple choice questions from 110,000 images. All the images are scenes from movies. We report all our results on validation dataset as the test is not available, since the VCR challenge is ongoing. Further, since we are not following the exact protocol of sub-tasks listed in Zellers et al. [2018], we don't submit to that leaderboard.

3.4.2 Generating Reasons

The encoder in our encoder decoder framework, for all experiments, is pretrained ViLBERT model (Lu et al. [2019]). The model has been pretrained on Conceptual Captions dataset Sharma et al. [2018], using BERT like masking and entailment pre-training strategy. For decoder, we experiment with pretrained BERT (Devlin et al. [2018]) and pretrained GPT-2 (Radford et al. [2019]).

For experiments with both GPT2 and BERT, we form our text input using question (Q) only. Additionally, we also train a network where the text input is formed by appending all answer choices ($A_i \forall i \in \{1 \dots 4\}$ to the question separated by '[SEP]' token. We train this model using GPT-2 decoder and call it $GPT2_{ans\_appended}$. We do this following Rajani et al. [2019b]. The image input in all experiments is the ground truth objects from the image, as provided in the dataset.

We use automatic quantitative evaluation metric BLEU (Papineni et al. [2002]) to judge how well the generated reason aligns with gold reason. Further, we also use perplexity metric to measure the language quality of generated sentences.

Training Hyperparameters

We train our models for 20 epochs using BertAdam optimizer. We use a batch size of 384, initial learning rate of 2e-5 and a linear scheduler.

Quantitative Results

The results are summarized in section 3.4.2. We make the following observations:

- **Low BLEU scores.** We see that all the decoders record quite low scores in the range 1.75-2.0. This is not very surprising considering Rajani et al. [2019b] also reported BLEU scores of 4.0 for generated reasons in text only mode. We do further analysis qualitatively in the next section.

- **Low perplexity numbers.** All the decoder models post quite low perplexity values. This

Table 3.1. Comparison of the quality of generated reason using different decoders. All numbers are reported on validation set of VCR dataset.

Decoder	BLEU	Perplexity
BERT	2.088	14.938
GPT2	1.817	14.88
$GPT2_{ans\_appended}$	1.824	13.79

could potentially mean that advanced sampling strategy while decoding has the possibility of improving the quality of generated samples. We leave that for future explorations.

Qualitative Results

We provide samples of correctly generated reason in fig. 3.3 and incorrectly generated reason in fig. 3.4. We mention here that denoted objects ([person1], [person2] etc) are replaced by random unisex names as done by Zellers et al. [2018]. We observe that in fig. 3.3, although the generated reason has the same meaning as the gold reason, there is very less syntactic word overlap. This has a major impact on BLEU metric, resulting in very low scores. Further, the incorrectly generated samples in fig. 3.4 are way off the gold reason. In general, we found that many more samples had poorly generated samples than correctly generated ones. We provide more analysis in the discussion section, detailing the possible reasons which could be the cause behind it.

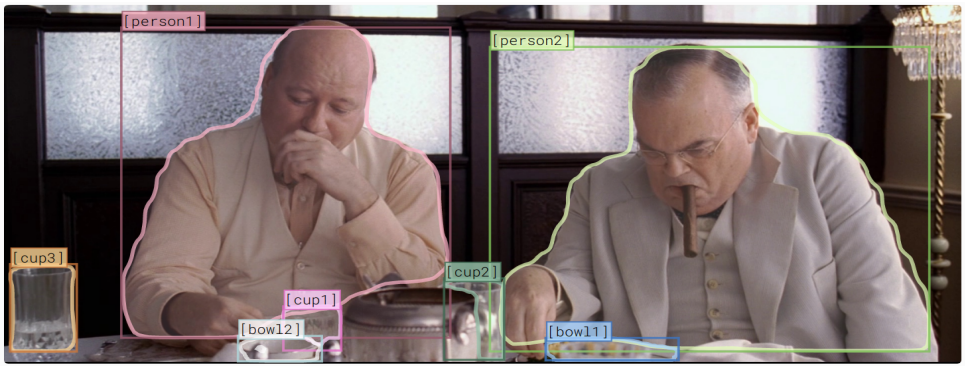
3.4.3 VQA

The text input in these experiments are formed by concatenating Q, Re and answer choice A_i, all separated by '[SEP]' token. The vision input is the ground truth objects from the image, provided in the dataset. The text and image input is fed to ViLBERT model to obtain probability scores for each, A_i, per question.

Type	Text
Question	Are these religious women?
Answer	Yes they are.
Gold Reason	You can tell by the crosses around each of their necks.
Generated Reason	They have religious garb they are wearing wigs which indicates they are in the worship of such material.They are also wearing traditional church garments.

Type	Text
Question	What is going on here?
Answer	This is a shopping center where a group of businessmen are coming.
Gold Reason	Kelin is pushing a shopping cart through the store.
Generated Reason	Kelin is carrying shopping bags.Kelin has an adult supervision box in his hands.

Figure 3.3. Samples of correctly generated reasons.



Type	Text
Question	What are [person1] and [person2] feeling?
Answer	Everyone seems to be trying to avoiding each other.
Gold Reason	Everyone is looking down and avoiding eye contact.
Generated Reason	Mandeep Mica are looking at Mandeep Chandlar with expressions of appreciation, which is common when people talk.Mandeep

Type	Text
Question	Why does [person3] have her hands behind her back?
Answer	[person3] wants to courteous but does not know how.
Gold Reason	Gaylin is in a royal palace but looks like she is dressed in peasant's clothes.
Generated Reason	It's polite to hold a camera to your face when meeting a stranger.Jaz.

Figure 3.4. Samples of wrongly generated reasons.

Training Hyperparameters

We train our models for 20 epochs using BertAdam optimizer. We use a batch size of 512, initial learning rate of 2e-5 and a linear scheduler.

Quantitative Results

We use two baselines in our experiments:

- Q_{only}**:** Only Q appended by A_i is fed as the text input to ViLBERT model. This is similar to what the authors of ViLBERT did, except they trained this model along with the second sub-task of VCR dataset $QA \rightarrow R$ by sharing the same parameters. As such, the accuracy reported in our experiments is lower than theirs.

- QR_c**:** Q appended by gold reason, and followed by A_i is fed as text input. This forms the upper bound of the accuracy models can achieve using reasons as part of input.

The quantitative results are summarized in section 3.4.3. All reported number are averaged across four different seeds. We make the following observations from the results:

- **Advantages of using reasons.** We observe that appending reason to question (QR_c) acheives superhuman accuracy of 91.935. This provides strong evidence of the advantages of using reasoning as part of input to improve VQA accuracy.

- **GPT2$_{\text{ans_appended}}$ performs best.** The reasons generated by appending the answer choices to questions, performs best on the downstream task of VQA. This suggests that appending answer choices to question is a useful strategy and provides more contextual information to generate reason.

- **Input with reasons doesn't have significant effect on VQA accuracy.** As can be seen by the performance of our three models: QR_{BERT}, QR_{GPT2} and $QR_{GPT2_{ans\_appended}}$. The performance improvement to question only input (Q_{only}) is minor to negligible. However, we did see superhuman accuracy of 91.935 while appending gold reason (QR_c). It can be

Table 3.2. VQA results with reason appended to question in the text input. In $QR_{\{.\}}$, $\{.\}$ denotes the decoder used to generate reason. All numbers are reported on validation set of VCR dataset.

Model	VQA Accuracy
Human accuracy	91.0
Q_{only}	69.605
QR_c	91.935
QR_{BERT}	68.509
QR_{GPT2}	67.983
$QR_{GPT2_{ans\_appended}}$	**69.928**

concluded from this that our generated reason were not good enough. We provide more analysis in the discussion section.

3.5 Discussion and Analysis

Model performs well on descriptive reasons

On investigation, we found that reasons which entailed description of objects in the scene or description of the image, were more correctly generated by our model. For example, in fig. 3.3 top, the reason describes that the women are religious because they are wearing cross. The generated reason was also able to pick up that they were religious since it could describe their religious attire. Similarly, for fig. 3.3 bottom, where the reason required description of the shopping center scene, the model generated relevant reasons.

In contrast, samples in fig. 3.4 require generating information which are latent and require sophisticated reasoning abilities. It's not very surprising that the model fails in those cases. This shows that models currently are good at describing images and objects, while it will require more modifications before they can start applying sophisticated reasoning about the world.

Latent/hidden information are too difficult for the model to generate

A number of reasons in this VCR dataset deal with reason which call upon latent or hidden information, which need to be inferred. For expample, in fig. 3.4 bottom, the model needs

world knowledge about what royal places look like and how peasants are dressed. Further, it needs to infer from the context of royal place and the person dressed in peasant clothes, that the person is uncomfortable and doesn't know how to behave. To break it down, first external knowledge base is required to inform how royal places and peasants look like. Secondly, explicit reasoning mechanism is required to connect these sources of information. All this is too much to ask from the model, which looks at just a single image frame.

Similarly, in fig. 3.4 top, reasoning that the two persons are avoiding eye contact requires the model to understand a lot of hidden dynamics. One needs to understand how people interact when they are socializing and when they are trying to avoid each other. Then, those clues need to be identified in the image before generating that kind of reason.

We argue that for generating reasons of the quality provided in this dataset, we need to incorporate (a) external knowledge base (b) explicit reasoning mechanisms to connect different sources of information. It's hardly surprising then that our model, which inputs only the image, generates poor quality reasons.

Using reasoning in VQA has strong potential

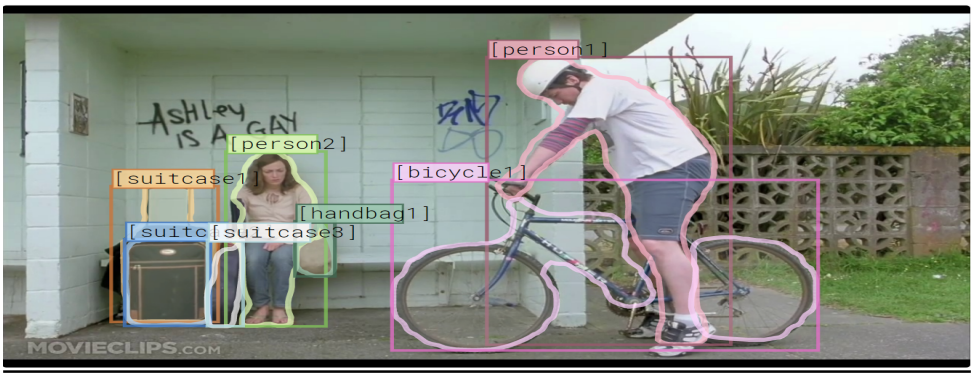
Despite the poor quality of generated reasons, we did see from section 3.4.3 that incorporating correct reasons in modeling VQA systems leads to superhuman accuracy. This suggests that there is strong advantages to be derived out of using such a reasoning mechanism in VQA systems. We hypothesize that using external knowledge base, explicit reasoning mechanism or perhaps advanced sampling strategy while decoding reasons could lead to significant improvements. Further, such reasoning methods also lead to explicti and interpretable models, contributing to its trustworthiness and bias elimination.

Appendix A

Generated Rationales Samples

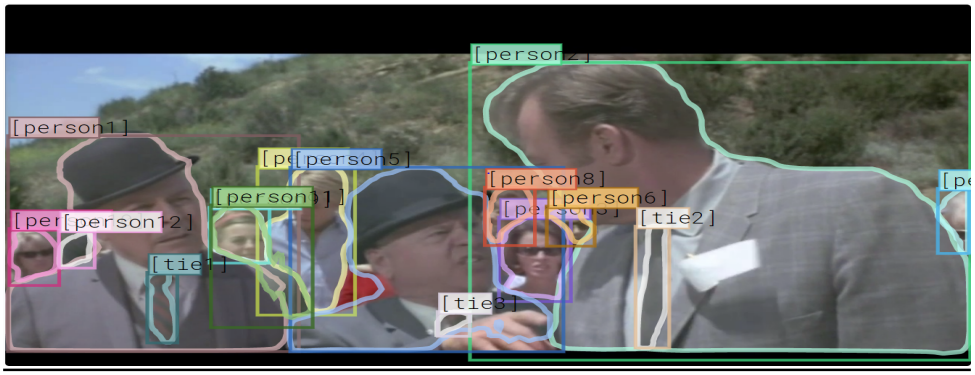
A.1 Comparison of ViLBERT-Fr and ViLBERT-Ra

Here we have provided additional examples of rationales generated by the two models: ViLBERT-Fr and ViLBERT-Ra.



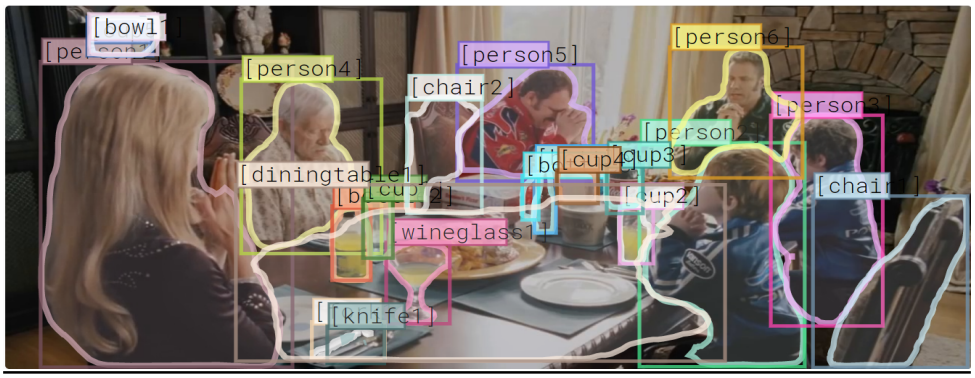
Type	Text
Question	Why is [person2] sitting on the bench?
Answer	[person2] is waiting for the bus.
Gold Rationale	She is sitting and waiting outside with a suitcase and other bags, indicating she is traveling somewhere.
ViLBERT-Fr	Lavern is walking away from the car and looking down at the ground she seems unsure of what to do next.
ViLBERT-Ra	There is a bus stop in front of Aarin

Type	Text
Question	Where are [person9] and [person14] going?
Answer	They are going to school.
Gold Rationale	They are dressed in a school uniform and walking towards a building that resembles a school.
ViLBERT-Fr	She is holding a notebook , one in line with a trip map.
ViLBERT-Ra	They are all carrying backpacks and are walking towards bus.



Type	Text
Question	What is the [person2] doing here?
Answer	He is the official responsible for managing this even.t
Gold Rationale	He has a ribbon on his chest that probably shows he is the official.
ViLBERT-Fr	Eastyn Remy are wearing elaborate discipline officers uniforms there is a government building on the other side of the room Storm is behind Ekam
ViLBERT-Ra	He is wearing the usk badge he has a serious look on his face.

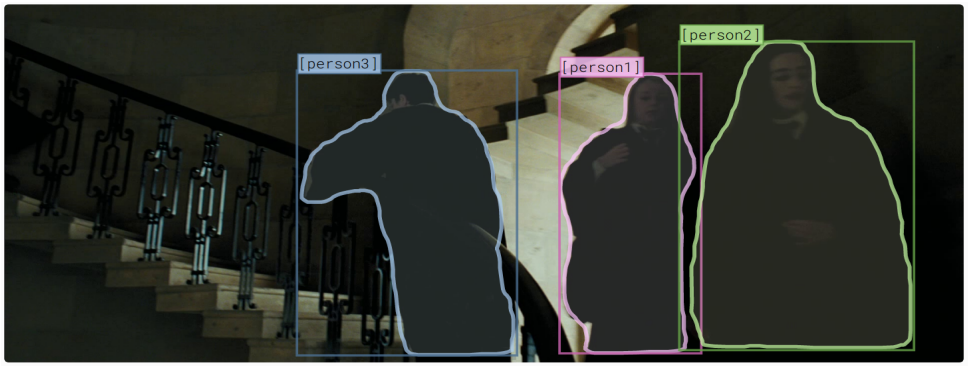
Type	Text
Question	What does the [person2] work?
Answer	[person2] works in a medical facility.
Gold Rationale	[person2] is wearing medical scrubs and is standing next to a hospital bed.
ViLBERT-Fr	Nieves is in Drue ' s office and looking to see what is going On.
ViLBERT-Ra	Kary is wearing a hospital gown and a nurses outfit , and her hair is messy , indicating that she has been there long.



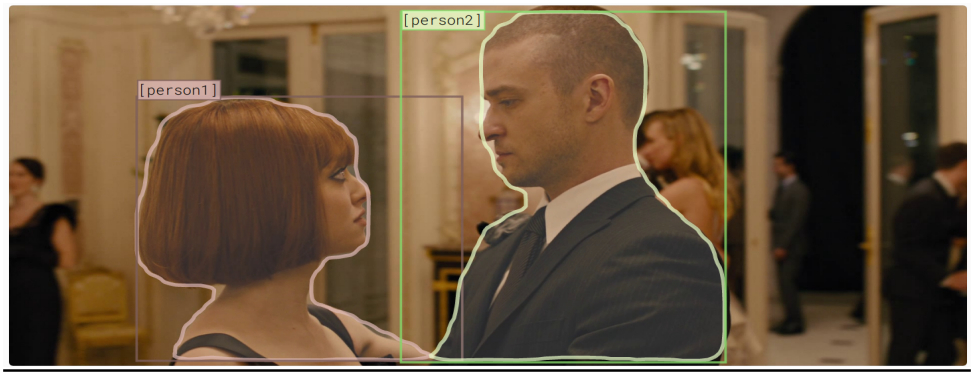
Type	Text
Question	Why does [person5] look upset?
Answer	[person5] is thankful because he hadn't had a good meal in a while and is overcome with emotion.
Gold Rationale	Sometimes when people haven't ate properly or do not come from much money meals may come infrequently and this is why he looks upset.
ViLBERT-Fr	It is breakfast time , and Zaelyn is about to eat.
ViLBERT-Ra	Blind people often eat their cake at the birthday party.

Appendix B

Generated Reason Samples



Type	Text
Question	Is everyone at school?
Answer	Yes, a school or library.
Gold Reason	Amari Ryley are carrying book bags.
Generated Reason	All are wearing uniforms which indicates they are at school.Everyone is wearing school uniforms.They are also wearing uniforms.It looks like



Type	Text
Question	What is [person1] thinking?
Answer	[person1] is wondering if [person2] is going to kiss her.
Gold Reason	They are dancing and they are looking intently at each other.
Generated Reason	Karsen is looking at Karsen.Her eyes are on Karsen.He is looking at Karsen.

Type	Text
Question	What era are [person1] and [person2] from?
Answer	They are from eighteen hundreds.
Gold Reason	Their dress indicates that this is the time period.
Generated Reason	They are wearing the same clothes as the others.The background is that of Shiloh.The clothing of Shiloh, but

Type	Text
Question	What position does [person2] hold?
Answer	[person2] is a butler.
Gold Reason	Butler's wear formal close and greet people as they enter.
Generated Reason	Devlyn's wide, white hair suggests he is of Eastyn's age and is

Bibliography

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016a. doi: 10.18653/v1/d16-1203. URL http://dx.doi.org/10.18653/v1/D16-1203.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31, Nov 2016b. ISSN 1573-1405. doi: 10.1007/s11263-016-0966-6. URL http://dx.doi.org/10.1007/s11263-016-0966-6.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. doi: 10.1109/cvpr.2018.00522. URL http://dx.doi.org/10.1109/CVPR.2018.00522.

Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. doi: 10.18653/v1/d19-1219. URL http://dx.doi.org/10.18653/v1/d19-1219.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. doi: 10.1109/cvpr.2016.12. URL http://dx.doi.org/10.1109/CVPR.2016.12.

Thomas Berg and Peter N. Belhumeur. How do you tell a blackbird from a crow? In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ICCV '13, pages 9–16, Washington, DC, USA, 2013. IEEE Computer Society. ISBN 978-1-4799-2840-8. doi: 10.1109/ICCV.2013.9. URL https://doi.org/10.1109/ICCV.2013.9.

Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering, 2019.

Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh.

Do explanations make vqa models more predictable to a human? *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. doi: 10.18653/v1/d18-1128. URL http://dx.doi.org/10.18653/v1/d18-1128.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations, 2019.

Roberto Cipolla, Yarin Gal, and Alex Kendall. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. doi: 10.1109/cvpr.2018.00781. URL http://dx.doi.org/10.1109/CVPR.2018.00781.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017. doi: 10.18653/v1/d17-1070. URL http://dx.doi.org/10.18653/v1/D17-1070.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, Oct 2017. ISSN 1077-3142. doi: 10.1016/j.cviu.2017.10.001. URL http://dx.doi.org/10.1016/j.cviu.2017.10.001.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. Towards transparent ai systems: Interpreting visual question answering models, 2016.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. doi: 10.1109/cvpr.2017.670. URL http://dx.doi.org/10.1109/CVPR.2017.670.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. *Lecture Notes in Computer Science*, page 3–19, 2016. ISSN 1611-3349. doi: 10.1007/978-3-319-46493-0_1. URL http://dx.doi.org/10.1007/978-3-319-46493-0_1.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. doi: 10.1109/iccv.2017.93. URL http://dx.doi.org/10.1109/ICCV.2017.93.

Kexin Huang. Content-based image retrieval using generated textual meta-data. In *ICAAI 2018*,

2018.

Drew A. Hudson and Christopher D. Manning. Learning by abstraction: The neural state machine, 2019.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. *Lecture Notes in Computer Science*, page 727–739, 2016. ISSN 1611-3349. doi: 10.1007/978-3-319-46484-8_44. URL http://dx.doi.org/10.1007/978-3-319-46484-8_44.

Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the winning entry to the vqa challenge 2018, 2018.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017a. doi: 10.1109/cvpr.2017.215. URL http://dx.doi.org/10.1109/CVPR.2017.215.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning, 2017b.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214/aoms/1177729694.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training, 2019.

Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo. Tell-and-answer: Towards explainable visual question answering using attributes and captions, 2018a.

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions, 2018b.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input, 2014.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based

approach to answering questions about images. *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015. doi: 10.1109/iccv.2015.9. URL http://dx.doi.org/10.1109/ICCV.2015.9.

Varun Manjunatha, Nirat Saini, and Larry S. Davis. Explicit bias discovery in visual question answering models, 2018.

David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. doi: 10.1109/cvpr.2018.00519. URL http://dx.doi.org/10.1109/CVPR.2018.00519.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence, 2018.

Badri N. Patro, Shivansh Pate, and Vinay P. Namboodiri. Robust explanations for visual question answering, 2020.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019a. doi: 10.18653/v1/p19-1487. URL http://dx.doi.org/10.18653/v1/p19-1487.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning, 2019b.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization, 2018.

Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering, 2015.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering, 2019.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://www.aclweb.org/anthology/P18-1238.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Probabilistic neural-symbolic models for interpretable visual question answering, 2019.

Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering, 2020.

Jialin Wu and Raymond J. Mooney. Self-critical reasoning for robust visual question answering, 2019.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning, 2018.