Incorporating Textual Resources to Improve Visual Question Answering

Jialin Wu
The University of Texas at Austin, 2022

Supervisor: Raymond J. Mooney

Recently, visual question answering (VQA) emerged as a challenge multi-modal task and gained in popularity. The goal is to answer questions that query information associated with the visual content in the given image. Since the required information could be from both inside and outside the image, common types of visual features, such as object and attribute detection, fail to provide enough materials for answering the questions. Textual resources, such as captions, explanations, encyclopedia articles, can help VQA systems comprehensively understand the image, reason following the right path, and access external facts. Specifically, they provide concise descriptions of the image, precise reasons for the correct answer, and factual knowledge beyond the image.

We presented completed work on generating image captions that are targeted to help answer a specific visual question. We introduced an approach that generates textual explanations and used these explanations to determine which answer is mostly supported. We

used explanations to recognize the critical objects for solving the visual question and trained the VQA systems to be influenced by these objects most. We also explored using textual resources to provide external knowledge beyond the visual content that is indispensable for a recent trend towards knowledge-based VQA. We further propose to break down visual questions such that each segment, which carries a single piece of semantic content in the question, can be associated with its specific knowledge. This separation aims to help the VQA system understand the question structure to satisfy the need for linking different aspects of the question to different types of information within and beyond the image.

Contents

Chapter 1

Introduction

Over the past few years, Visual Question Answering (VQA), spanning both the visual and linguistic domains, emerged as a challenging task that attracts tons of attention. The goal is to answer open-ended natural language questions that query information associated with the visual content in the given image. Part of the increasing attraction comes from the belief that VQA offers a step forward to achieving "AI-complete" tasks by stressing the need for various AI capabilities required to access, process, and reason upon multi-modal information for solving the visual questions. For example, fine-grained object, attribute, relation, and activity detection is necessary for understanding the image. Furthermore, commonsense knowledge is indispensable when the implication of the visual content is the main focus. Recently, there is a trend towards knowledge-based VQA, querying the information from other knowledge resources beyond the images.

Due to the necessity of the multi-modal information from various aspects, the input features and supervision choices are critical. We have witnessed a clear shift from using grid features pretrained on image classification tasks to using region-based features pretrained on object and attribute detection. Those features can provide common labels, pair-wise relations, and attributes for the objects in the image; however, they still lack representation power for all the required information, which could be from both inside and outside the image. For

example, the relationship among several objects is hard to characterize; commonsense and specific knowledge beyond the images are missing. Textual resources, such as captions, explanations, encyclopedia articles, can help VQA systems comprehensively understand the image, reason following the right path, and access external facts by providing concise descriptions of the image, precise reasons for the correct answer, and factual statements. On the supervision side, textual explanations help the VQA systems understand why the answer is correct, where the common annotations only provide what is correct.

We generate image captions targeted to help answer a specific visual question. We also introduce an approach that generates textual explanations and used these explanations to determine which answer is most supported. We use explanations to recognize the key objects for solving the visual question and trained the VQA systems to be mostly influenced by these objects. We also explore using textual resources to provide external knowledge beyond the visual content indispensable for knowledge-based VQA.

Finally, we propose some short-term and long-term goals to better utilize textual resources to improve VQA performance. First, we would like to break down visual questions such that each segment, carrying a single piece of semantic content, can be associated with its specific knowledge. This separation aims to help the VQA system understand the question structure to satisfy the need for linking different aspects of the question to different types of information within and beyond the image. Second, since the retrieval process is not jointly trained with the VQA systems, it is possible that the retrieved facts, though they are right from their perspective, sometimes mislead the answer predictor. Therefore, we also want to determine the relevance of the retrieved knowledge to ensure the precision of knowledge and achieve better performance. Finally, as a long-term goal, we would like to step toward building a more interpretable VQA system using textual resources.

Chapter 2

Related Work

We review some of the relevant background knowledge for our work. We first introduce different types of visual question answering (VQA) and common approaches. Then, we discuss the textual content generation, including image captioning and explanations. Finally, we present graphical networks for VQA.

2.1 Visual Question Answering

We review the visual question answering literature in three aspects: question types, visual content representation, and VQA systems.

2.1.1 Visual Questions Types

Visual Question Answering (VQA) is a multimodal task of answering a question that queries information associated with the visual content in the image. The questions cover a vast set of visual information, focusing on different visual aspects, and require different AI capabilities due to their multimodal nature (Antol et al., 2015). General visual questions (Antol et al., 2015; Krishna et al., 2017) mainly focus on asking single visual features present in the image, including colors, shapes, materials, quantity, attributes, relations of objects as shown in 2.1

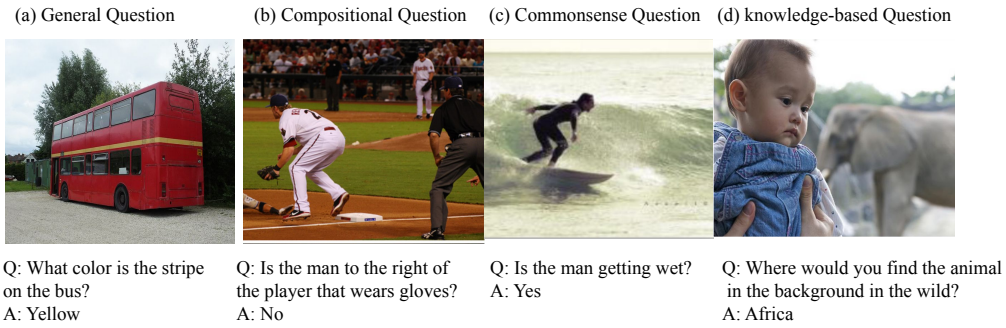

Figure 2.1: Examples of different types of visual questions.

(a). Compositional visual questions (Hudson and Manning, 2019) (in 2.1 (b)) stress the need for a deeper semantic understanding of vision and language. Instead of answering the question in one step, VQA systems need to properly understand every hop in the questions and linked these hops to the corresponding visual content. As a step further, an intelligent should understand both "what is in the image" and the implication of the visual scene that requires a significant amount of commonsense knowledge (Park et al., 2018). For example, the system should understand that when people is surfing, they will get wet as shown in 2.1 (c). Besides, there is also a trend towards knowledge-based VQA (Wang et al., 2017; Marino et al., 2019) where the VQA systems need to retrieve relevant knowledge to correctly answer the questions. For example, the system not only need to recognize the "animal" in the question refers to the elephant, but also need to retrieve the knowledge that this type of elephant lives in Africa.

2.1.2 Visual Representations

The visual and textual representation of the image and question is crucial to provide the VQA systems sufficient yet concise information to predict the answers. An LSTM+CNN baseline system (Antol et al., 2015) uses grid features as inputs produced by CNNs pre-trained on a large-scale image classification dataset (Russakovsky et al., 2015). In order

to enable attention to be calculated at the level of objects and other salient image regions, Up-Down systems introduce object and attribute detection (Ren et al., 2015b) as the visual representation. For better answering commonsense questions where the key clues cannot be visually detected, human justifications (Park et al., 2018) are employed to teach the VQA systems the right reason for the right answer. In order to gather sufficient outside knowledge beyond the image, Wikipedia (Wikipedia contributors, 2004) and concepts from conceptnet (Speer et al., 2017) are commonly used as additional inputs (Marino et al., 2021, 2019; Wu et al., 2021).

2.1.3 Visual Question Answering Systems

VQA systems have witnessed significant progress on the modeling side. LSTM+CNN baseline systems (Antol et al., 2015; Ren et al., 2015a; Fukui et al., 2016; Goyal et al., 2017; Li et al., 2018a) encode the image and question using CNN and RNN, respectively. Up-Down systems (Anderson et al., 2018; Li et al., 2018b; Wu and Mooney, 2019b) use object-level features to determine important ones for the question using Top-Down attention. In order to better present relations between objects, graph neural nets are used to link objects together, enabling the VQA systems to reason among groups of objects according to the visual questions. As VQA requires a vast set of AI skills, recent multimodal transformers (Yu et al., 2019a; Zhou et al., 2020; Lu et al., 2020a, 2019; Tan and Bansal, 2019; Liu et al., 2019; Li et al., 2019, 2020a; Chen et al., 2020) are pretrained on many auxiliary tasks, including VQA (Antol et al., 2015), referring-resolution (Yu et al., 2016), image captioning (Chen et al., 2015; Sharma et al., 2018), etc., using various multimodal datasets. Cross attention modules are built over the textual and visual modalities to learn a joint representation for the entire question and the detected objects. With a large amount of training data and a wide range of pretraining tasks, these models achieve promising performance on various VQA benchmarks(Antol et al., 2015; Hudson and Manning, 2019; Singh et al., 2019; Marino et al., 2019). In order to incorporate knowledge from various external sources, knowledge-based

VQA systems often employ fact graphs (Tompson et al., 2014; Narasimhan et al., 2018; Li et al., 2020b; Marino et al., 2021) as a different modality of inputs.

2.2 Image Captioning

Recent image captioning models have experienced a clear shift from attention-based deep-learning models (Donahue et al., 2015; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Luo et al., 2018; Liu et al., 2018) to multimodal transformers (Cornia et al., 2020; He et al., 2020; Liu et al., 2021). With the help of large image description datasets (Chen et al., 2015), these models have demonstrated remarkable results.

However, deep neural models still tend to generate general captions based on the most significant objects(Vijayakumar et al., 2016). Although previous works (Luo et al., 2018; Liu et al., 2018) build captioning models that are encouraged to generate different captions with discriminability objectives, the captions are usually less informative and fail to describe most objects and their relationships diversely.

We would like to use the captions to provide additional information to enrich the VQA features sets. We develop an approach to generating captions that directly focus on the critical objects in the VQA process and provide information that can help the VQA module predict the answer for a particular question.

2.3 Explanations for Visual Question Answering

2.3.1 Visual Explanation

Several approaches have been proposed to visually explain decisions made by vision systems by highlighting relevant image regions. For example, GradCAM (Selvaraju et al., 2017) analyzes the gradient space to find visual regions that most affect the decision. Attention mechanisms (Singh et al., 2018; Anderson et al., 2018) in VQA models can also be directly utilized to determine highly-attended regions and generate visual explanations.

In order to train a VQA system to be right for the right reason(Ross et al., 2017), recent research has collected human visual attention highlighting image regions that most contribute to the answer. Two popular approaches are to have crowdsourced workers deblur the image (Das et al., 2017) or select segmented objects from the image (Park et al., 2018). Then, the VQA systems try to align either the VQA system's attention (Zhang et al., 2019; Qiao et al., 2018) or the gradient-based visual explanation (Selvaraju et al., 2019) to the human attention. These approaches help the systems focus on the right regions and improve VQA performance when the training and test distributions are very different, such as in the VQA-CP dataset (Agrawal et al., 2018).

2.3.2 Textual and Multi-Modal Explanation

While visual explanations highlight key image regions behind the decision, textual explanations (Park et al., 2018) explain the reasoning process and crucial relationships between the detected objects. As a step further, there has also been some work on multimodal explanations that link textual and visual explanations. A recent extension of this work (Hendricks et al., 2018) first generates multiple textual explanations and then filters out those that could not be grounded in the image. We argue that a good explanation should focus on referencing visual objects that actually influenced the system's decision, therefore generating more faithful explanations.

2.4 Graphical Networks

Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017) generalize Convolutional Networks (CNN) to accommodate graph-structured input. Various types of graph input for VQA have been explored including scene graphs generated by an object and relation detector (Ren et al., 2015b; Yang et al., 2018), and knowledge graphs retrieved from a wide range of sources, such as DB-Pedia (Auer et al., 2007), ConceptNet (Liu and Singh, 2004), VisualGenome (Krishna et al., 2017) and hasPart KB (Bhakthavatsalam et al., 2020). Most

KB-VQA systems (Ramnath and Hasegawa-Johnson, 2021; Narasimhan et al., 2018; Li et al., 2020b; Marino et al., 2021) build their GCNs on top of these knowledge graphs and extract relevant evidence using the entire question representation.

Chapter 3

Completed Work

We present our completed work on utilizing textual resources in the following sections.

3.1 Generating Captions for VQA

Human Captions :
1) A man on a blue surfboard on top of some rough water.
2) A young surfer in a wetsuit surfs a small wave.
3) A young man rides a surf board on a small wave while a man swims in the background.
4) A young man is on his surf board with someone in the background.
5) A boy riding waves on his surf board in the ocean.

Question 1: Does this boy have a full wetsuit on?
Caption: A young man wearing **wetsuit** surfing on a wave.
Question 2: What color is the board?
Caption: A young man riding a wave on a **blue surfboard**.

Figure 3.1: Examples of our generated question-relevant captions. During the training phase, our model selects the most relevant human captions for each question (marked by the same color).

Exploiting textual features from the image, tersely encoding the necessary information to answer the questions, is not sufficiently studied. This information could be richer than the visual features in that the sentences have fewer structural constraints and can easily include the attributes of and relation among multiple objects. In fact, we observe that appropriate captions can be very useful for many VQA questions. We explore a novel approach that generates **question-relevant** image descriptions, which contain information that is directly relevant to a par-

ticular VQA question. Fig. 3.1 shows examples of our generated captions given different questions.

Specifically, our model first extracts image features $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K\}$ and question features $\mathbf{q}$ to produce their joint representation that are further used to generate question-related captions as shown in the left in 3.2. Next, our caption embedding module encodes the generated captions as caption features $\mathbf{c}$ as shown in the bottom of 3.2. After that, a VQA module is built on question, image, and caption features to predict the answer (right part in 3.2). The work is published as a conference paper at ACL 2019.

3.1.1 Generating Question-Relevant Captions

Image Captioning Module. We adopt an image captioning module similar to that of Anderson et al. (2018), which takes the object detection features as inputs and learns attention weights over those objects' features in order to predict the next word at each step. The key difference between our module and theirs lies in the input features and the caption supervision. Specifically, we use the question-attended image features $\mathbf{V}^q$ as inputs and only use the most relevant caption, which is automatically determined in an online fashion (detailed below), for each question-image pair to train the captioning module. This ensures that only question-relevant captions are generated.

Selecting Relevant Captions for Training. Previously, Li et al. (2018b) selected relevant captions for VQA based on word similarities between captions and questions; however, their approach does not take into account the details of the VQA process. In contrast, during training, our approach dynamically determines for each problem the caption that will most improve VQA. We do this by updating with a shared descent direction (Wu et al., 2018) which decreases the loss for *both* captioning and VQA. This ensures a consistent target for both the image captioning module and the VQA module in the optimization process.

During training, we compute the cross-entropy loss for the i-th caption using Eq. 3.1,

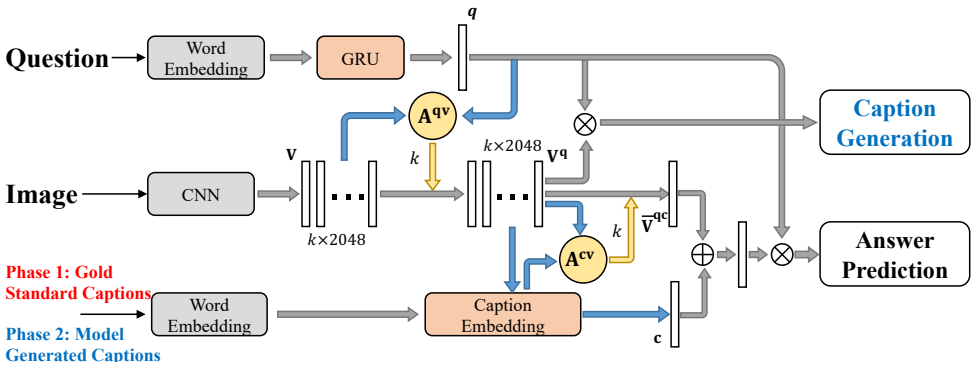

Figure 3.2: Overall structure of our model that generates question-relevant captions to aid VQA. Our model is first trained to generate question-relevant captions as determined in an online fashion in phase 1. Then, the VQA model is fine-tuned with generated captions from the first phase to predict answers. $\otimes$ denotes element-wise multiplication and $\oplus$ denotes element-wise addition. Blue arrows denote fully-connected layers (fc) and yellow arrows denote attention embedding.

and back-propagate the gradients only from the most relevant caption determined by solving Eq. 3.2.

$$\mathcal{L}_i^c = -\sum_{t=1}^{T} \log(p(w_{i,t}^c | w_{i,t-1}^c)) \tag{3.1}$$

In particular, we require the inner product of the current gradient vectors from the predicted answer and the human captions to be greater than a positive constant ξ, and further select the caption that maximizes that inner product.

$$\begin{aligned} &\arg\max_i \sum_{k=0}^{K} \left(\frac{\partial \hat{s}_{\text{pred}}}{\partial \mathbf{v}_k^q}\right)^T \frac{\partial \log(p(\mathbf{W}_i^c))}{\partial \mathbf{v}_k^q} \\ &s.t. \sum_{k=0}^{K} \left(\frac{\partial \hat{s}_{\text{pred}}}{\partial \mathbf{v}_k^q}\right)^T \frac{\partial \log(p(\mathbf{W}_i^c))}{\partial \mathbf{v}_k^q} > \xi \end{aligned} \tag{3.2}$$

where the $\hat{s}_{\text{pred}}$ is the logit[1] for the predicted answer, $\mathbf{W}_i^c$ denotes the i-th human caption for the image and k traverses the K object features.

[1]The input to the softmax function.

Therefore, given the solution to Eq. 3.2, $i^\star$, the final loss of our joint model is the sum of the VQA loss and the captioning loss for the selected captions as shown in Eq. 3.3. If Eq. 3.2 has no feasible solution, we ignore the caption loss.

$$\mathcal{L} = \mathcal{L}^{vqa} + \mathcal{L}^{c}_{i^\star} \tag{3.3}$$

3.1.2 Utilizing Captions for VQA

As illustrated in Fig. 3.2, we use both question features **q** and caption features **c** to generate the visual attention $\mathbf{A}^{cv}$ to weight the images' feature set **V**, producing attended image features $\overline{\mathbf{v}}^{qc}$. Finally, we add $\overline{\mathbf{v}}^{qc}$ to the caption features **c** and further perform element-wise multiplication with the question features **q** (Anderson et al., 2018) to produce the joint representation of the question, image and caption, which is then used to predict the answer.

3.1.3 Experimental Evaluation

We first report the experimental results on the VQA task and compare our results with the state-of-the-art methods in this section. We use the VQA v2.0 dataset (Antol et al., 2015) for the evaluation of our proposed joint model, where the answers are balanced in order to minimize the effectiveness of learning dataset priors. This dataset is used in the VQA 2018 challenge and contains over 1.1M questions from the over 200K images in the MSCOCO 2015 dataset (Chen et al., 2015). After that, we perform ablation studies to verify the contribution of additional knowledge from the generated captions and the effectiveness of using caption representations to adjust the top-down visual attention weights.

As demonstrated in Table 3.1, our single model outperforms other state-of-the-art single models by a clear margin, $i.e.$ 2.06%, which indicates the effectiveness of including caption features as additional inputs. In particular, we observe that our single model outperforms other methods, especially in the 'Num' and 'Other' categories. This is because the generated captions can provide more numerical clues for answering the 'Num' questions

	Test-standard			
	Yes/No	Num	Other	All
Prior (Goyal et al., 2017)	61.20	0.36	1.17	25.98
Language-only (Goyal et al., 2017)	67.01	31.55	27.37	44.26
MCB (Fukui et al., 2016)	78.82	38.28	53.36	62.27
Up-Down (Anderson et al., 2018)	82.20	43.90	56.26	65.32
VQA-E (Li et al., 2018b)	83.22	43.58	56.79	66.31
Ours(single)	**84.69**	**46.75**	**59.30**	**68.37**
Ours(Ensemble-10)	**86.15**	**47.41**	**60.41**	**69.66**

Table 3.1: Comparison of our results on VQA with the state-of-the-art methods on the test-standard data. Accuracies in percentage (%) are reported.

since the captions can describe the number of relevant objects and provide general knowledge for answering the 'Other' questions. Furthermore, an ensemble of 10 models with different initialization seeds results in a score of 69.7% for the test-standard set.

3.2 Self-Critical Reasoning for VQA under Changing Prior

A critical aspect of a VQA system being trustworthy is to answer the question with correct rationale. A number of recent VQA systems (Trott et al., 2018; Zhang et al., 2019; Selvaraju et al., 2019; Qiao et al., 2018) learn to not only predict correct answers but also be "right for the right reasons" (Ross et al., 2017; Selvaraju et al., 2019). These systems are trained to encourage the network to focus on regions in the image that humans have somehow annotated as important (which we will refer to as "important regions."). However, many times, the network also focuses on these important regions even when it produces a wrong answer. Previous approaches do nothing to actively discourage this phenomenon, which we have found occurs quite frequently.

For example, as shown in Figure 3.3, we ask the VQA system, "What is the man eating?". The baseline system predicts "hot dog" but focuses on the banana because hot dog appears much more frequently in the training data. What's worse, this error is hard to detect when only analyzing the correct answer "banana" that has been successfully grounded in the image.

Our self-critical approach prevents the most common answer from dominating the correct answer. We first construct a proposal set of influential objects using textual explanations. Then, we penalize the network for focusing on this region when its predicted answer for this question is *wrong*. Figure 3.4 shows an overview of our approach. Besides the UpDn VQA system (left top block), our approach contains two other components, we first recognize and strengthen the most influential objects (left bottom block), and then we criticize incorrect answers that are more highly ranked than the correct answer and try to make them less sensitive to these key objects (right block). As recent research suggests that gradient-based methods more faithfully represent a model's decision making process (Selvaraju et al., 2019; Zhang et al.; Wu et al., 2018; Jain and Wallace, 2019), we use a modified GradCAM (Selvaraju et al., 2017) to compute the answer a's sensitivity to the i-th

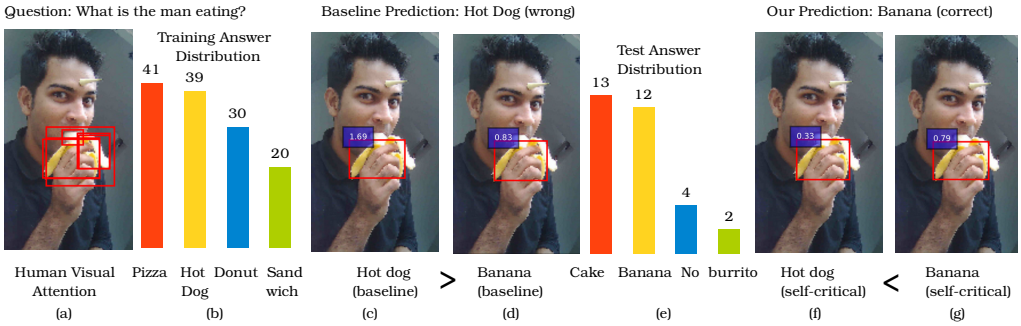

Figure 3.3: Example of a common answer misleading the prediction even though the VQA system has the right reasons for the correct answer. Figure (a) shows the important regions extracted from human visual attention. Figure (b), (e) show the answers' distribution for the question "What is the man eating?" in the training and test dataset. Figure (c), (d) show the most influential region for the prediction "hot dog" and "banana" using the baseline UpDn VQA system and Figure (f), (g) show the influential region for the prediction "hot dog" and "banana" using the VQA system after being trained with our self-critical objective. The number on the bounding box shows the answer's sensitivity to the object.

object features $\mathbf{v}_i$ as shown in Eq. 3.4.[2] The work is published as a conference paper at NeurIPS 2019.

$$\mathcal{S}(a, \mathbf{v}_i) \coloneqq \left(\nabla_{\mathbf{v}_i} P(a|V, q)\right)^T \mathbf{1} \tag{3.4}$$

3.2.1 Constructing Influential Object Set

Our approach ideally requires identifying important regions that a human considers most critical in answering the question. However, directly obtaining such a clear set of influential objects from either visual or textual explanations is hard, as the visual explanations also highlight the neighbor objects around the most influential one, and grounding textual explanations in images is still an active research field. We relax this requirement by identifying a

[2] $\mathbf{1}$ denotes a vector with all 1's.

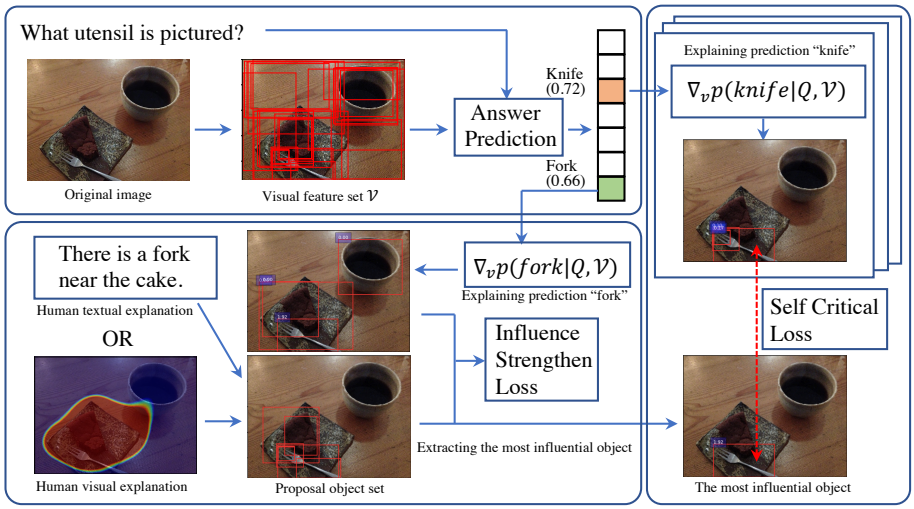

Figure 3.4: Model overview. In the left top block, the base UpDn VQA system first detects a set of objects and predicts an answer. We then analyze the correct answer's sensitivity (Fork) to the detected objects via visual explanation and extract the most influential one in the proposal object set as the most influential object, further strengthened via the influence strengthen loss (left bottom block). Finally, we analyze the competitive incorrect answers' sensitivities (Knife) to the most influential object and criticize the sensitivity until the VQA system answers the question correctly (right block). The number on a bounding box is the answer's sensitivity to the given object.

proposed set of influential objects $\mathcal{I}$ for each QA pair. This set may be noisy and contain some irrelevant objects, but we assume that it includes the most relevant object. We explore three different methods for constructing this proposal set, as described below:

Construction from Visual Explanations. Following HINT (Selvaraju et al., 2019), we use the VQA-HAT dataset (Das et al., 2017) as the visual explanation source. HAT maps contain a total of $59,457$ image-question pairs, corresponding to approximately 9% of the VQA-CP training and test set. We also inherit HINT's object scoring system that is based on the normalized human attention map energy inside the proposal box relative to the normalized energy outside the box. We score each detected object from the bottom-up attention and build the potential object set by selecting the top $|\mathcal{I}|$ objects.

Construction from Textual Explanations. Recently, (Park et al., 2018) introduced a textual

explanation dataset that annotates $32,886$ image-question pairs, corresponding to 5% of the entire VQA-CP dataset. To extract the potential object set, we first assign part-of-speech (POS) tags to each word in the explanation using the spaCy POS tagger (Honnibal and Montani, 2017) and extract the nouns in the sentence. Then, we select the detected objects whose cosine similarity between the Glove embeddings (Pennington et al., 2014) of their category names, and any of the extracted nouns' is greater than 0.6. Finally, we select the $|\mathcal{I}|$ objects with the highest similarity.

Construction from Questions and Answers. Since the above explanations may not be available in other datasets, we also consider a simple way to extract the proposal object set from just the training QA pairs alone. The method is quite similar to the way we construct the potential set from textual explanations. The only difference is that instead of parsing the explanations, we parse the QA pairs and extract nouns from them.

3.2.2 Recognizing and Strengthening Influential Objects

Given a proposal object set $\mathcal{I}$ and the entire detected object set $\mathcal{V}$, we identify the object that the correct answer is most sensitive to and further strengthen its sensitivity. We first introduce a sensitivity violation term $\mathcal{SV}(a, \mathbf{v}_i, \mathbf{v}_j)$ for answer a and the i-th and j-th object features $\mathbf{v}_i$ and $\mathbf{v}_j$ as the amount of sensitivity that $\mathbf{v}_j$ surpasses $\mathbf{v}_i$, as shown in Eq. 3.5.

$$\mathcal{SV}(a, \mathbf{v}_i, \mathbf{v}_j) = \max\big(\mathcal{S}(a, \mathbf{v}_j) - \mathcal{S}(a, \mathbf{v}_i), 0\big) \tag{3.5}$$

Based on the assumption that the proposal set contains at least one influential object that a human would use to infer the answer, we impose the constraint that the most sensitive object in the proposal set should not be less sensitive than any object outside the proposal set. Therefore, we introduce the influence strengthen loss $\mathcal{L}_{infl}$ in Eq. 3.6:

$$\mathcal{L}_{infl} = \min_{\mathbf{v}_i \in \mathcal{I}} \Big(\sum_{\mathbf{v}_j \in \mathcal{V} \setminus \mathcal{I}} \mathcal{SV}(a_{gt}, \mathbf{v}_i, \mathbf{v}_j) \Big) \tag{3.6}$$

where the a_{gt} denotes the ground truth answer. The key differences between our influence strengthen loss and the ranking-based HINT loss are that (1) we relax the unnecessary constraint that the objects should follow the exact human ranking, and (2) it is easier to adapt to different types of explanation ($e.g.$ textual explanations) where such detailed rankings are not available.

3.2.3 Criticizing Incorrect Dominant Answers

Next, for the incorrect answers ranked higher than the correct answer, we attempt to decrease the sensitivity of the influential objects. For example, in VQA-CP, bedrooms are the most common room type. Therefore, during testing, systems frequently incorrectly classify bathrooms (rare in the training data) as bedrooms. Since humans identify a sink as an influential object when identifying bathrooms, we want to decrease the influence of sinks on concluding bedroom.

In order to address this issue, we design a self-critical objective to criticize the VQA systems' incorrect but competitive decisions based on the most influential object $\mathbf{v}^*$ to which the correct answer is most sensitive as defined in Eq. 3.7.

$$\mathbf{v}^* = \arg\min_{\mathbf{v}_i \in \mathcal{I}} \Big(\sum_{\mathbf{v}_j \in \mathcal{V} \setminus \mathcal{I}} \mathcal{SV}(a_{gt}, \mathbf{v}_i, \mathbf{v}_j) \Big) \tag{3.7}$$

Precisely, we extract a bucket of at most B predictions with higher confidence than the correct answer $\mathcal{B} = \{a_1, a_2, ..., a_{|\mathcal{B}|}\}$ and utilize the proposed self-critical loss $\mathcal{L}_{crit}$ to directly minimize the weighted sensitivities of the answers in the bucket $\mathcal{B}$ to the selected most influential object, as shown in Eq. 3.8.

$$\mathcal{L}_{crit} = \sum_{a \in \mathcal{B}} w(a)(\mathcal{S}(a, \mathbf{v}^*) - \mathcal{S}(a_{gt}, \mathbf{v}^*)) \tag{3.8}$$

where a_{gt} denotes the ground truth answer. Because several answer candidates could be similar ($e.g.$ cow and $cattle$), we weight the sensitivity gaps in Eq. 3.8 by the cosine

	Expl.	VQA-CP v2 test			
		All	Yes/No	Num	Other
GVQA(Agrawal et al., 2018)		31.3	58.0	13.7	22.1
UpDn (Anderson et al., 2018)		39.7	42.7	11.9	46.1
UpDn+AttAlign (Selvaraju et al., 2019)		38.5	42.5	11.4	43.8
UpDn+AdvReg. (Ramakrishnan et al., 2018)		41.2	65.5	**15.5**	35.5
UpDn+SCR (ours)	QA	**48.47**	**70.41**	10.42	**47.29**
UpDn+HINT (Selvaraju et al., 2019)	HAT	47.7	70.0	10.7	46.3
UpDn+SCR (ours)	HAT	49.17	71.55	10.72	47.49
UpDn+SCR (ours)	VQA-X	**49.45**	**72.36**	10.93	**48.02**

Table 3.2: Comparison of the results on VQA-CP test dataset with the state-of-the-art systems. The upper part includes VQA systems without human explanations during training, and the VQA systems in the bottom part use either visual or textual human explanations. The "Expl." column shows the source of explanations for training the VQA systems. SCR is the short hand for our self-critical reasoning approach.

distance between the answers' 300-d Glove embeddings (Pennington et al., 2014), $i.e.$ $w(a) = cosine\_dist(Glove(a_{gt}),\ Glove(a))$. In the multi-word answer case, the Glove embeddings of these answers are computed as the sum of the individual word's Glove embeddings.

3.2.4 Experimental Evaluation

Table 3.2 shows results on the VQA-CP dataset, comparing our results with the state-of-the-art methods. VQA-CP (Agrawal et al., 2018) is a diagnostic reconfiguration of the VQA v2 dataset where the distribution of the QA pairs in the training set is significantly different from those in the test set. Most state-of-the-art VQA systems are found to highly rely on language priors and experience a catastrophic performance drop on VQA-CP. We evaluate our approach on VQA-CP in order to demonstrate that it generalizes better and is less sensitive to distribution changes.

Our system significantly outperforms other state-of-the-art systems (e.g., HINT (Selvaraju et al., 2019)) by 1.5% on the overall score for VQA-CP when using the same human

visual explanations (VQA-HAT), which indicates the effectiveness of directly criticizing the competitive answers' sensitivity to the most influential objects. Moreover, using human textual explanations as supervision is even a bit more effective. With only about half the number of explanations compared to VQA-HAT, these textual explanations improve VQA performance by an additional 0.3% on the overall score, achieving a new state-of-the-art of 49.5%.

Without human explanations, our approach that only uses the QA proposal object set as supervision clearly outperforms all of the previous approaches, even those that use human explanations. We further analyzed the quality of the influential object proposal sets extracted from the QA pairs by comparing them to those from the corresponding human explanations. On average, the QA proposal sets contain 57.1% and 54.3% of the objects in the VQA-X and VQA-HAT proposal object sets, respectively, indicating a significant but not perfect overlap.

Note that our self-critical objective remarkably improves VQA performance in the 'Yes/No' and 'Other' question categories; however, it does not do as well in the 'Num' category. This is understandable because counting problems are generally more challenging than the other two types and require the VQA system to consider *all* of the objects jointly. Therefore, criticizing only the most sensitive ones does not improve the performance.

3.3 Competing Explanations for VQA

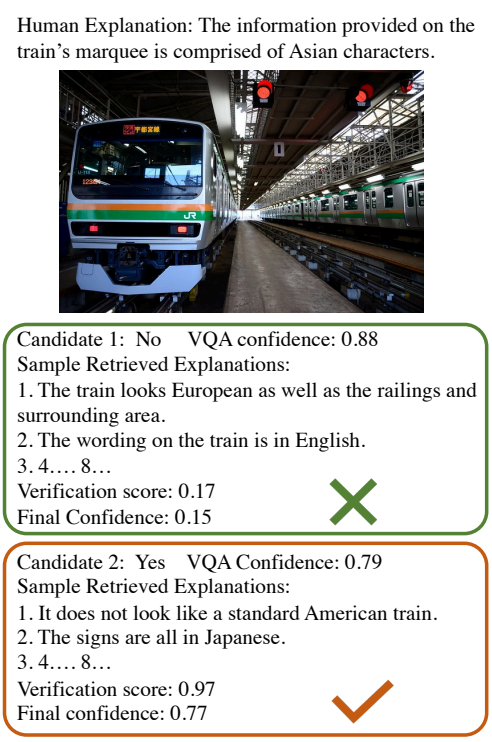

Figure 3.5: An example of utilizing retrieved explanations to correct the original VQA prediction. Though the original VQA confidence of the correct answer "Yes" is lower than that of the incorrect answer "No", the retrieved explanations for "Yes" support their answer better, resulting in a higher verification score and a final correct decision.

Most state-of-the-art VQA systems (Anderson et al., 2018; Kim et al., 2018; Ben-Younes et al., 2017; Jiang et al., 2018; Cadene et al., 2019; Lu et al., 2019; Liu et al., 2019; Tan and Bansal, 2019) are trained to fit the answer distribution using question and visual features and achieve high performance on simple visual questions. However, these systems often exhibit poor explanatory capabilities and take shortcuts by only focusing on simple visual concepts or question priors instead of finding the right answer for the right reasons (Ross et al., 2017; Selvaraju et al., 2019). This problem becomes increasingly severe when the questions require more complex reasoning and commonsense knowledge.

For more complex questions, VQA systems need to be right for the right reasons in order to generalize well to test problems. Textual explanations encode richer information such as detailed attributes, relationships, or commonsense knowledge that is not necessarily directly found in the image. Therefore, we adopt textual explanations to guide VQA systems. In particular, our approach considers explanations for multiple competing answers, comparing these explanations when choosing a final answer, as shown in Figure 3.3.

As shown in Figure 3.6, after the base VQA system computes the top-k answers, our approach retrieves the most supportive explanations for each answer from the training set to construct the set of competing explanations. Then, these explanations are used to help generate explanations for the current question. Next, we learn to predict verification scores that indicate how well the retrieved or generated explanations support the predictions given the input question and visual content. The final answer is determined by jointly considering the original answer probabilities and these verification scores. The work is published as a workshop paper at AAAI 2020.

3.3.1 VQA Module and Candidate Answers Generation

Many recent VQA systems (Fukui et al., 2016; Ben-Younes et al., 2017; Ramakrishnan et al., 2018) utilize a trainable top-down attention mechanism over convolutional features to recognize relevant image regions. These systems first extract a visual feature set $\mathcal{V} = \{\mathbf{v}_i, ..., \mathbf{v}_{|\mathcal{V}|}\}$ for each image whose element $\mathbf{v}_i$ is a feature vector for the i-th detected object. On the language side, UpDn systems sequentially encode each question Q to produce a question vector $\mathbf{q}$. Let f denote the answer prediction operator that takes both visual features and question features as input and predicts the confidence for each answer a in the answer candidate set $\mathcal{A}$, $i.e.$ $P(a|\mathcal{V}, Q) = f(\mathcal{V}, \mathbf{q})$. The VQA task is framed as a multi-label regression problem with the gold-standard soft scores as targets in order to be consistent with the evaluation metric. Finally, binary cross-entropy loss with soft score is used to supervise the sigmoid-normalized outputs.

We briefly introduce two variants of this approach adopted in our experiments:

UpDn. This is the original UpDn system, which uses a single layer GRU to encode questions. The question vector is then used to compute single-stage attention over the detected objects to produce attended visual features. Finally, a two-layer feed-forward network computes answer probabilities given the joint features of the question and visual content.

LXMERT. In order to learn richer representations for both questions and visual content,

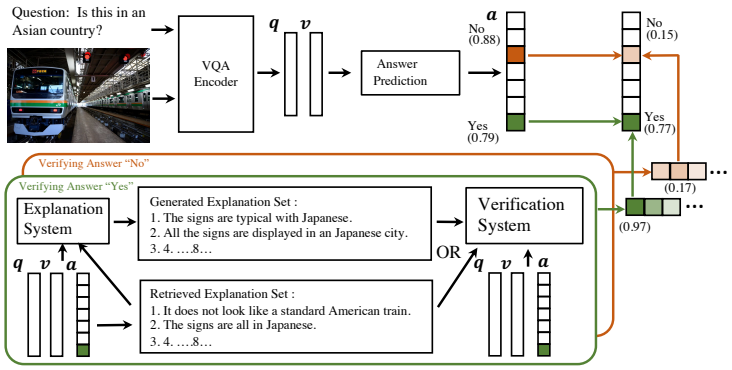

Figure 3.6: Our approach first predicts a set of answer candidates and retrieves explanations for each based on the answer, question, and visual content. These explanations are then used to generate improved explanations. Finally, either retrieved or generated explanations are employed to predict verification scores that are used to reweight the original predictions and compute the final answer.

LXMERT (Tan and Bansal, 2019) uses transformers (Vaswani et al., 2017; Devlin et al., 2019) that learn multiple layers of attention over the input. In particular, it first learns 9 layers over the input question and 5 layers over detected objects, then finally learn another 5 layers of attention across the two modalities to produce the final joint representation.

3.3.2 Collecting Explanations for Candidate Answers

Retrieving Explanations This section presents our approach to retrieving the most supportive human textual explanation from the training set for each answer candidate. Ideally, we would dynamically retrieve explanations for each answer at each iteration. However, this would be very computational costly because the question and visual features have to be computed for each image from the training set. Therefore, we adopt the below relaxation for computational efficiency that only needs to compute the features once.

In particular, we first pretrain the VQA system, and extract the question and visual embeddings, **q** and **v** , for each $\mathcal{QV}$ pair in the training set. For UpDn, we use the attended visual features and the question GRU's last hidden state as the visual and question embeddings. For LXMERT, we use the last cross-modal attention layer's visual and question output as the

embeddings.

Then, for each $\mathcal{QV}$ pair, we only compute the top-10 answer candidates since the top-10 answers together achieve high recall. After that, for each answer candidate a, we extract explanations from the training set that have the same ground truth answer [3] as the current candidate. We then sort these explanations by the L2 distance between the explanations' $\mathcal{QV}$ embeddings, $\mathbf{q} \odot \mathbf{v}$, and the example's and pick the closest 8 explanations as the competing explanations set denoted as $\mathcal{X}_a$.

Generating Explanations Next, the retrieved explanations for similar VQA examples from the training set are used to help generate even better explanations.

We adopt the explainer from (Wu and Mooney, 2019a), a two-layer LSTM network similar to the UpDn captioner (Anderson et al., 2018), as our baseline. Since the current VQA systems are built upon detected objects, we use them as visual inputs instead of segmentations.

The baseline explainer first computes a set of question-attended visual features, $\mathcal{U}$, and an average pooled version, $\bar{\mathbf{u}}$. The explainer then uses $\bar{\mathbf{u}}$ and $\mathcal{U}$ together with question and answer embeddings as inputs to produce explanations.

Our approach simply replaces the average pooled question-attended visual features $\bar{\mathbf{u}}$ with the retrieved explanations' features, $\mathbf{x}$. We use a single-layer GRU to encode all of the retrieved explanations for the correct answer, and then max pool the last hidden states among these explanations to compute $\mathbf{x}$. We sample 8 explanations for each answer candidate to construct the generated explanation set.

3.3.3 Learning and Utilizing Verification Scores

A verification system is trained to score how well a generated or retrieved explanation supports a corresponding answer candidate given the question and visual content. The verification system takes four inputs: the visual, question, answer and its explanation features;

[3] More specifically, the soft score of the answer candidate in the retrieved explanation's example is over 0.6

and outputs the verification score, $i.e.$ $S(Q, \mathcal{V}, a, x) = \sigma(f_2(f(\mathbf{q}), f(\mathbf{v}), f(\mathbf{a}), f(\phi(x)))$. where $\mathbf{a}$ is the one-hot embedding of the answer, and $\phi(x)$ is the feature vector for the explanation, x, encoded using a GRU (Cho et al., 2014), ϕ. We use f_n to denote n consecutive feed-forward layers (for simplicity n is omitted when $n = 1$). We use σ to denote the sigmoid function. The verification system is similar to the answer predictor in architecture except for the number of outputs.

Given the VQA examples with their explanations in the VQA-X dataset, we use binary cross-entropy loss $\mathcal{L}_m$ to maximize the verification score for the matching human explanations, $i.e.$ $\mathcal{L}_m = -\log(S(Q, \mathcal{V}, a, x))$.

Intuitively, we want the verification score S to be high only when the explanation is matched to the VQA example, $i.e.$ replacement of any of the four input sources should lower the score. Therefore, we designed the five kinds of replacements below for constructing negative examples. Specifically, we replace visual, question, answer, explanation and answer-explanation pairs once a time, producing five losses $i.e.$

$$\mathcal{L}_r^q = -\log(1 - S(Q', \mathcal{V}, a, x)) \tag{3.9}$$

$$\mathcal{L}_r^v = -\log(1 - S(Q, \mathcal{V}', a, x)) \tag{3.10}$$

$$\mathcal{L}_r^a = \mathbb{E}_{a' \sim p(a'|QV),\ s(a') < 0.6}[-\log(1 - S(Q, \mathcal{V}, a', x))] \tag{3.11}$$

$$\mathcal{L}_r^a = \max_{x' \in \mathcal{X}_{a'}} [-\log(1 - S(Q, \mathcal{V}, a, x'))] \tag{3.12}$$

$$\mathcal{L}_r^{ax} = -\log(\max_{x' \in \mathcal{X}_{a'}} (1 - S(Q, \mathcal{V}, a', x'))) \tag{3.13}$$

Finally, the total verification loss is the sum of the aforementioned 6 losses as shown in Eq. 3.14:

$$\mathcal{L}_{verification} = \lambda \mathcal{L}_m + \mathcal{L}_r^q + \mathcal{L}_r^v + \mathcal{L}_r^a + \mathcal{L}_r^x + \mathcal{L}_r^{ax} \tag{3.14}$$

Since we have more negative examples (5 ways to form negative examples) and only one

positive example, we assign a larger loss weight ($i.e.$ $\lambda = 10$) for the only positive example.

Using Verification Scores The original VQA system provides the answer probabilities conditioned on the question and visual content, $i.e.$ $P(a|Q, \mathcal{V})$. The verification scores $S(Q, \mathcal{V}, a, x)$ are further used to reweight the original VQA predictions so that the final predictions $\tilde{P}(a|Q, \mathcal{V})$, shown in Eq. 3.15, can take the explanations into account.

$$\tilde{P}(a|Q, \mathcal{V}) = P(a|Q, \mathcal{V}) \max_{x \in \mathcal{X}_a} S(Q, \mathcal{V}, a, x) \tag{3.15}$$

where $\mathcal{X}_a$ denotes the generated or retrieved explanation set for the answer a.

Since we try to select the correct answer with its explanation, the prediction $\tilde{P}(a|Q, \mathcal{V})$ should only be high when the answer a is correct and the explanation x supports a, which is enforced using the loss in Eq. 3.16:

$$\mathcal{L}_{vqae} = -\log(P(a|Q, \mathcal{V})S(Q, \mathcal{V}, a, x_a)) - \log(1 - \tilde{P}(a'|Q, \mathcal{V})) \tag{3.16}$$

where x_a denotes the human explanation for the answer a.

During testing, we first extract the top 10 answer candidates $\mathcal{A}$, and then select the explanation for the answer candidate with the highest verification score. Then, we compute the explanation-reweighted score for each answer candidate to determine the final answer $a^\star = \arg\max_{a \in \mathcal{A}} \tilde{P}(a|Q, \mathcal{V})$.

3.3.4 Experimental Evaluation

Table 3.3 reports the results of our competing explanation approach in VQA v2 and VQA-X dataset. VQA-X datasets contains visual questions that require humans to be of age 9 or higher and are believed to require more commonsense knowledge. Our approach combined with UpDn pretrained on the entire VQA v2 dataset achieves the best results. When training only on the VQA-X training set, we improve the original UpDn and LXMERT by 4.5 % and 1.2 %, respectively. UpDn benefits more from using competing explanations than LXMERT,

	VQA-X Pretrain		VQA v2 Pretrain	
	Gen. Expl.	Ret. Expl.	Gen. Expl.	Ret. Expl.
UpDn (Anderson et al., 2018)	74.2	74.2	83.6	83.6
UpDn+E (ours)	78.0	78.7	85.1	85.4
LXMERT (Tan and Bansal, 2019)	76.8	76.8	83.7	83.7
LXMERT+E (ours)	77.3	78.0	84.1	84.7

Table 3.3: Question answering accuracy on VQA-X using both UpDn and LXMERT as a base system,"+E" denotes using our competing explanations approach. "Gen. Expl." and "Ret. Expl." denote using generated and retrieved explanations, respectively.

but both improve. By using transformers, LXMERT already creates better but less flexible representations, which are harder to improve upon by using explanations. Because we do not use the official LXMERT model parameters pretrained on multiple large datasets (VQA-X test set is used as the training set for the official released model) and only train the LXMERT on the VQA v2 dataset, the performance of LXMERT is not better than UpDn.

3.4 Multi-Modal Answer Validation for Knowledge-Based VQA

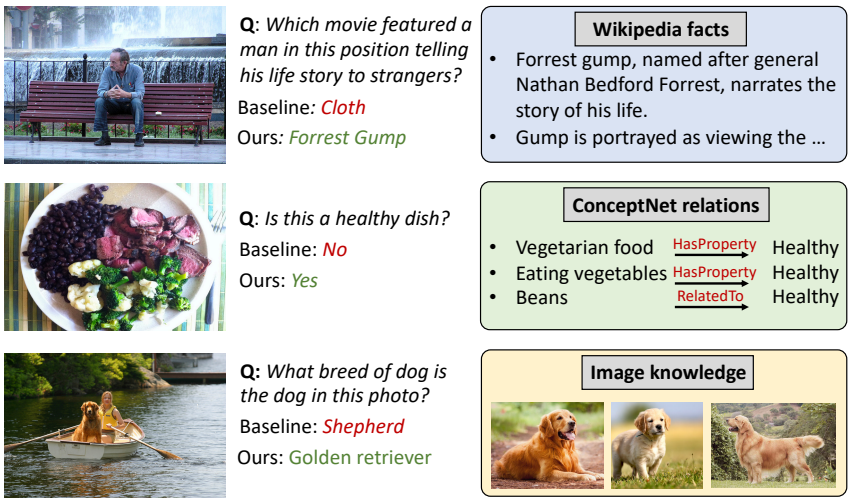

Figure 3.7: We address the problem of knowledge-based question answering. Retrieving relevant knowledge among diverse knowledge sources (visual knowledge, textual facts, concepts, etc.) is quite challenging. The goal in this paper is to learn what knowledge source should be used for a particular question and how to validate a set of potential answer candidates using that source.

Over the past few years, the domain of Visual Question Answering (VQA) has witnessed significant progress (Antol et al., 2015; Zhu et al., 2016; Hudson and Manning, 2019; Singh et al., 2019). There is a recent trend towards knowledge-based VQA (Marino et al., 2019) which requires information beyond the content of the images. To correctly answer those challenging questions, the model requires not only the ability of visual recognition, but also logical reasoning and incorporating external knowledge about the world. These knowledge facts can be obtained from various sources, such as image search engines, encyclopedia articles, and knowledge bases about common concepts and their relations.

Figure 3.7 illustrates a few visual questions and the knowledge from different external sources required to answer them. Each question needs a different type of external knowledge. For example, to identify the movie that featured a man telling his life story to strangers, we need to link the image content and question to some textual facts (blue box in the figure); Vegetarian food and eating vegetables is related to the concept of health (green box); and

What English city is famous for a tournament for the sport this man is playing?

Question + Image

The modern game of tennis originated in Birmingham, England, in the late 19th century as lawn tennis.

Question + Image + Incorrect Answer (Copenhagen)

It is popular for sports fixtures and hosts several annual events including a free opera concert at the opening of the opera season, other open-air concerts, carnival and labour day celebrations, and the Copenhagen historic grand prix, a race for antique cars.

Question + Image + Correct Answer (Wimbledon)

Wimbledon is notable for the longest running sponsorship in sports history due to its association with slazenger who have supplied all tennis balls for the tournament since 1902.

Figure 3.8: Examples of retrieved Wikipedia sentences using different sets of search words. The sentences retrieved using only the words in questions and objects in images (top) and the wrong answer (middle) are hardly helpful to answer the question. However, with the correct answer "Wimbledon" (bottom), the quality of the retrieved fact is significantly improved.

the retrieved images for 'golden retriever' (yellow box) are visually similar to the dog in the question image. *The challenge is to effectively retrieve and correctly incorporate such external knowledge* in an open domain question answering framework.

However, knowledge retrieved directly for the question and image is often noisy and not useful for predicting the correct answer. For example, as shown in Figure 3.8, the sentences retrieved using only the words in questions and objects in images (top) or a wrong answer (middle) are hardly helpful to answer the question. This increases the burden on the answer predictor, leading to only marginal improvements from the use of retrieved knowledge (Marino et al., 2019). Interestingly, with the correct answer "Wimbledon" (bottom), the quality of the retrieved fact is significantly improved, making it useful to answer the question. This observation motivates us to use retrieved knowledge for *answer validation* rather than for producing the answer.

To address this challenge, we propose a new framework called MAVEx or **M**ulti-modal **A**nswer **V**alidation using **Ex**ternal knowledge. The key intuition behind MAVEx is that verifying the validity of an answer candidate using retrieved knowledge is more reliable compared to open knowledge search for finding the answer. Therefore, we learn a model to evaluate the validity of each answer candidate according to the retrieved facts. For this approach to work, we need a small set of answer candidates to start with. We observe that while state-of-the-art VQA models struggle with knowledge-based QA, these models are surprisingly effective at generating a small list of candidates that often contains the correct

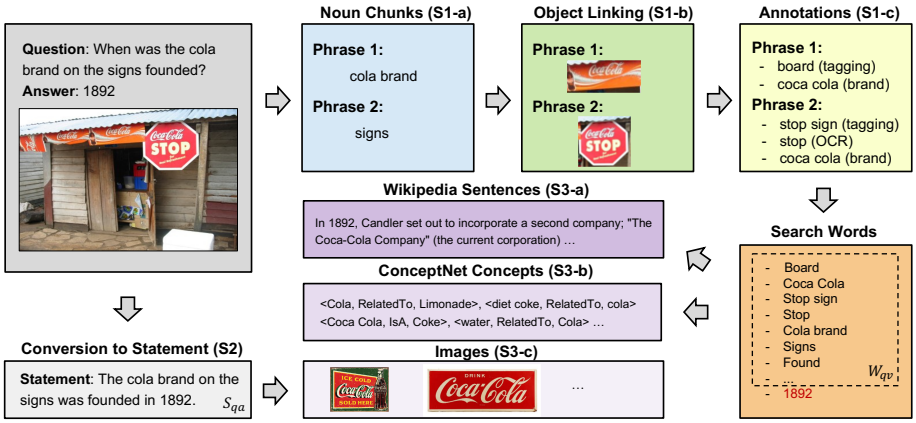

Figure 3.9: An example of the retrieval process for one question-answer pair.

answer. Using these candidates to guide knowledge search makes retrieved facts less noisy and often more pertinent to the question, as shown in Figure 3.8.

3.4.1 Multi-Modal Knowledge Retrieval

Answer Candidate Generation. In order to use answer candidates to inform knowledge retrieval, we use ViLBERT (Lu et al., 2019), a state-of-the-art VQA model, to generate answer candidates. Note that any VQA model can be used for this purpose. As discussed in the experiments section, we found ViLBERT to be particularly effective at generating a small set of promising candidates.

Given a question q about an image I and an answer candidate a from a set of possible answers, we retrieve external knowledge in support of a in three main steps. Figure 3.9 shows the entire process for an example question and a candidate answer.

S1: Answer-Agnostic Search Word Extraction. We first generate short phrases in q and concepts represented in I as a starting point for retrieving external information. This involves the following sub-steps:

Extract Noun Chunks from q: We parse the question using a constituency parser to

compute the parse tree. Then, we extract all the nouns on the leaves of the parse tree together with the words that describe the nouns and belong to one of the types from 'ADJP', 'ADVP', 'PP', 'SBAR', 'DT' or 'JJ'. Those words help us to link the mentioned objects to the images. We use AllenNLP (Gardner et al., 2018) constituency parser. See Figure 3.9 (S1-a).

Link Nouns to Objects: As images usually contain plenty of question-irrelevant contents, making the retrieval process hard to operate, we propose to narrow down the search field to the objects referred to by the question. In particular, we use ViLBERT-multi-task (Lu et al., 2020b) as the object linker, where it outputs scores given the noun phrases from the questions. We approve the linking when the linker's score is higher than 0.5 and extract the linked objects.

Annotate Objects: We automatically provide the category labels, OCR readings and logo information for the linked objects using Google APIs to enrich the retrieved knowledge. See Figure 3.9 (S1-c).

The set of answer-agnostic search words, W_{qv}, consists of all of noun chunks and verbs in q, OCR, tagging (detection), and logo annotation of the referred objects, if any.

S2: Conversion to a Natural Language Statement. In order to use the answer candidate a to inform the retrieval step, we convert q and a into a natural language statement S_{qa} using a rule-based approach (Demszky et al., 2018). Such conversion has been found to be effective as statements occur much more frequently than questions in textual knowledge sources (Khot et al., 2017).

S3: Answer Candidate Guided Retrieval. We now use the search words W_{qv} from step S1, along with the answer candidate a and the statement S_{qa} from step S2, to retrieve relevant information as follows:

Retrieval of textual facts: We query each search word w $\in W_{qv}$ and collect all sentences from the retrieved Wikipedia articles.[4] For each answer candidate a, we first collect answer-specific sentences that contain a (ignoring stop words and yes/no). Then we

[4]We use the python API https://github.com/goldsmith/Wikipedia.

rank those sentences based on the BERTScore (Zhang et al., 2020) between the statement S_{qa} and the sentences. We then encode each of the top k^w_{sp} sentences using a pre-trained BERT (Devlin et al., 2019) model and extract the final layer representation of the [CLS] token. This results in an answer-specific (denoted sp) feature matrix $\mathbf{K}^w_{sp}(a) \in \mathbb{R}^{k^w_{sp} \times 768}$ for each question-answer pair. We also store the retrieved sentences and their corresponding BERTScores for all answer candidates. We then choose the top k^w_{ag} non-repeated sentences according to the stored scores as the answer-agnostic knowledge. Those sentences are also encoded using pre-trained BERT, resulting in an answer-agnostic (denoted ag) feature matrix $\mathbf{K}^w_{ag} \in \mathbb{R}^{k^w_{ag} \times 768}$ for each question.

Retrieval of concepts: While Wikipedia articles provide factual knowledge that people need to look up when they answer a question, ConceptNet offers structured knowledge of concepts. Similar to Wikipedia article retrieval, we also query each search word in W_{qv} and collect all retrieved concepts. For each answer candidate a, we extract the concepts whose subject, relation, or object contains the candidate a, and push all retrieved concepts to the answer-agnostic concept pool. We rank those extracted concepts based on the maximum cosine similarity between the Glove embedding (Pennington et al., 2014) of the words in W_{qv} and those in the concept, and select the top k^c_{sp} concepts as answer-specific knowledge. We also select the top k^c_{ag} concepts similarly from the answer-agnostic concept pool. The subjects, relations, and objects in the selected concepts are first converted into a sentence by handcrafted rules, and then encoded using pre-trained BERT model. Finally, the last layers' representation vectors are concatenated, resulting in a feature matrix $\mathbf{K}^c_{sp}(a) \in \mathbb{R}^{k^c_{sp} \times 768}$ for each question-answer pair, and a feature matrix $\mathbf{K}^c_{ag} \in \mathbb{R}^{k^c_{ag} \times 768}$ for each question.

Retrieval of visual knowledge: Pure textual knowledge is often insufficient due to two main reasons: (1) textual knowledge might be too general and not specific to the question image, (2) it might be hard to describe some concepts using text, and an image might be more informative. Hence, visual knowledge can complement textual information, further enriching the outside knowledge feature space. We use Google image search to retrieve

the top k_i images using the statement S_{qa} as the query. The images are then fed into a MaskRCNN (He et al., 2017) finetuned on the Visual Genome dataset (Zhu et al., 2016) to extract at most 100 object features. We average the object features of visual detection results as the answer-specific visual knowledge representation, resulting in a feature matrix $\mathbf{K}^i_{sp}(a) \in \mathbb{R}^{k^i_{sp} \times 768}$ for each question-answer pair. For answer-agnostic knowledge, we simply use the zero vector.

3.4.2 VQA Module

We use cross-modal attention (Yu et al., 2019b) in the knowledge embedding module, which treats the question-image embedding as a query to mine supportive knowledge from each source.

We first briefly introduce the Self-Attention (SA) and Guided-Attention (GA) units[5] as the building blocks. The SA unit takes as input a group of feature vectors $\mathbf{X} = [x_1; ...; x_m] \in \mathbb{R}^{m \times d}$ and learns the pairwise relationship between each sample pair within $\mathbf{X}$ using a multi-head attention layer by treating all possible combinations of x_i and x_j as queries and keys. Different from SA, the GA unit uses another group of features $\mathbf{Y} = [y_1; ...; y_n] \in \mathbb{R}^{n \times d}$ to guide the attention learning in $\mathbf{X}$. In particular, the GA unit learns the pairwise relationship between each pair across $\mathbf{X}$ and $\mathbf{Y}$ and treats each y_i as query and each x_i as keys. The values of the keys are weighted summed to produce an attended output features $\mathbf{T} \in \mathbb{R}^{m \times d}$ for both SA and GA. Finally, a feed-forward layer with residual links is built upon $\mathbf{T}$ to transform the output features to a new features space.

Given an image and the corresponding question, we first use ViLBERT to extract visual features $\mathbf{v} \in \mathbb{R}^{1024}$ and question features $\mathbf{q} \in \mathbb{R}^{1024}$ from the last layer of ViLBERT's [IMG] and [CLS] tokens, respectively. We then compute a joint feature $\mathbf{U}$ by element-wise multiplication of $\mathbf{q}$ and $\mathbf{v}$. $\mathbf{U}$ is used as a query to mine answer-agnostic features $\mathbf{z}^j_{ag}$. $\mathbf{U}$ and the BERT embeddings of the answer candidates are used to mine answer-specific features

[5]Please refer to (Yu et al., 2019b) for detailed model architectures.

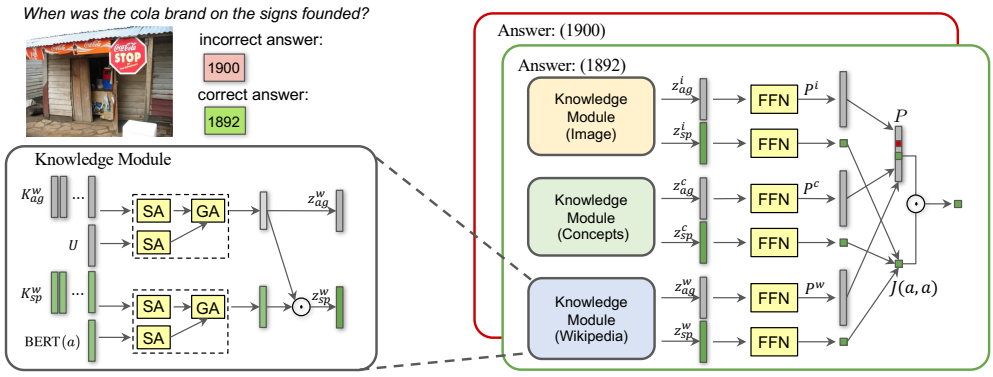

Figure 3.10: Model overview for validating two candidate answers. We explore three sources of external knowledge, *i.e.* Wikipedia, ConceptNet, and Google Images presented by the three parallel knowledge embedding modules. The grey blocks denote answer-agnostic features shared by all answer candidates and the green blocks denote answer-specific features.

$\mathbf{z}^j_{sp}(a, a')$ for the answer candidate a from each one of the three knowledge sources j as described in Eqs. (3.17) and (3.18):

$$\mathbf{z}^j_{ag} = \text{GA}(\text{SA}(\mathbf{U}), \text{SA}(\mathbf{K}^j_{ag})) \tag{3.17}$$

$$\mathbf{z}^j_{sp}(a, a') = \mathbf{z}^j_{ag} \odot \text{GA}(\text{SA}(\text{BERT}(a)), \text{SA}(\mathbf{K}^j_{sp}(a'))) \tag{3.18}$$

where a and a' are two answer candidates and the index j denotes one of the knowledge sources (Wikipedia w, ConceptNet c, or Google images i). Specifically, the answer-specific features $\mathbf{z}^j_{sp}(a, a')$ encode the joint features of a and the knowledge retrieved using a', and are further used to predict how well the knowledge retrieved by a' supports a.

3.4.3 Answer Validation Module

The validation module uses the attended knowledge features $\mathbf{z}^j_{sp}$ and $\mathbf{z}^j_{ag}$ from the three sources to validate the answer candidates. Our approach lets each knowledge source predict its own supportiveness score. The goal of this setting is to prevent misleading knowledge

Method	Knowledge Resources	Performance
ArticleNet (AN) (Marino et al., 2019)	Wikipedia	5.3
Q-only (Marino et al., 2019)	—	14.9
MLP (Marino et al., 2019)	—	20.7
BAN (Kim et al., 2018)	—	25.2
+ AN (Marino et al., 2019)	Wikipedia	25.6
+ KG-AUG (Li et al., 2020b)	Wikipedia + ConceptNet	26.7
MUTAN (Ben-Younes et al., 2017)	—	26.4
+ AN (Marino et al., 2019)	Wikipedia	27.8
Mucko (Zhu et al., 2020)	Dense Caption	29.2
KRISP (Marino et al., 2021)	Wikipedia + ConceptNet	32.3
ConceptBert (Gardères et al., 2020)	ConceptNet	33.7
ViLBERT (Lu et al., 2019)	—	35.2
MAVEx (ours) – w/o answer validation	Wikipedia + ConceptNet + Google Images	37.6
MAVEx (ours)	Wikipedia + ConceptNet + Google Images	**38.7**
MAVEx (ours) (Ensemble 5)	Wikipedia + ConceptNet + Google Images	**39.4**

Table 3.4: MAVEx outperforms current state-of-the-art approaches on the OK-VQA dataset. The middle column lists the external knowledge sources, if any, used in each VQA system.

from contaminating valid knowledge from other sources. In particular, we compute the supportiveness score J^j for each source as $J^j(a, a') = \text{FFN}(\mathbf{z}^j_{sp}(a, a'))$, where FFN denotes a feed-forward layer. Then, the final score is computed by taking the maximum support score across the three sources as $J(a, a') = \max_j \{J^j(a, a')\}$, where $j \in \{w,\ c,\ i\}$ denotes the source index. We use the answer-agnostic features to predict single source VQA scores P^j for all answers in the set as $P^j = \text{FFN}(\mathbf{z}^j_{ag})$, and the final VQA score P is computed as $P = \max_j\{P^j\}$. The overall architecture of the model is shown in Figure 3.10.

3.4.4 Experimental Evaluation

We present results on OK-VQA dataset (Marino et al., 2019) that contains 14055 questions manually selected that require knowledge beyond the image. Table 3.4 shows that MAVEx consistently outperforms prior approaches by a clear margin. For example, MAVEx outperforms recent state-of-the-art models Mucko (Zhu et al., 2020), KRISP (Marino et al., 2021), and ConceptBert (Gardères et al., 2020) by 9.5, 6.4, 5.0 points, respectively. Our approach also outperforms ViLBERT (Lu et al., 2019) base system by 3.5 points. We consider a

MAVEx baseline model that uses the retrieved knowledge ($\mathbf{K}_{ag}^{j}$) as additional inputs without answer validation. This model achieves 37.6 overall scores, 2.4% higher than the ViLBERT model and 1.1% lower than the late fusion model, indicating that using answer-guided retrieved knowledge is helpful and answer validation further improves the performance. An ensemble of 5 MAVEx late fusion models with different initializations improves the results to 39.4.

Chapter 4

Proposed Work

We present the proposed work for both short-term research that will be completed for the final thesis and long-term research that aims at more ambitious goals and may not be included in the dissertation.

4.1 Short Term Proposals

The short-term goals involve breaking down visual questions and generating a customized set of answer candidates for visual questions.

4.1.1 Breaking Down Visual Questions

Solving real-world visual questions that cover a wide range of real-world topics could require multiple steps of reasoning. Therefore, the ability of VQA systems to understand each step in the question and link them to relevant knowledge sources is crucial. Our completed work mainly focuses on the visual and knowledge representation side, aiming to train the VQA systems to focus on the right objects and provide sufficient outside knowledge to answer the questions. However, since VQA systems take as inputs from multiple modalities, the information across more modalities has to be properly utilized

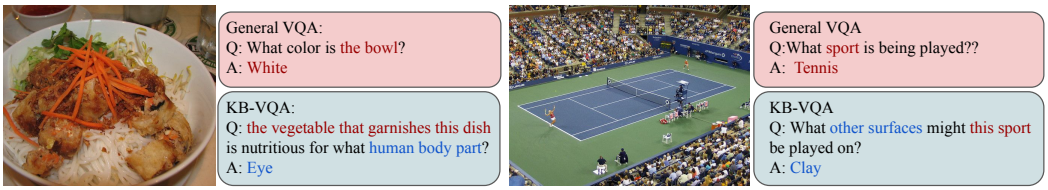

Figure 4.1: Examples of general and knowledge-based (KB) visual questions. The question and answer segments that focus on visual content within the image are highlighted in red, and the segments that requires external knowledge are highlighted in blue.

jointly by VQA systems to achieve good performance. This introduces significant challenges that knowledge representations can vary significantly across different knowledge sources, including factual sentences (Wu et al., 2021; Marino et al., 2019), knowledge triples (Wang et al., 2017), concepts (Gardères et al., 2020) and images (Wu et al., 2021). More importantly, a system needs to understand which knowledge is useful for different semantic segments of the question.

As shown in Fig. 4.1, KB-VQA systems need to link the segment "the vegetable that garnishes this dish" to the carrot on the plate and then query knowledge bases to find out which "human body part" particularly benefits from the nutrients in carrots. Simply encoding the entire question for either retrieving or filtering the knowledge, as most KB-VQA systems (Wang et al., 2017; Marino et al., 2019; Zhu et al., 2020; Li et al., 2020b; Marino et al., 2021; Wu et al., 2021) do, can be confusing since different parts of the question focus on different aspects that can be either outside or inside the image. As depicted in Fig. 4.1, searching for "human body part" and "other surfaces" within the image may cause VQA systems to focus on irrelevant aspects of the image.

We propose to segment visual questions into several semantic chunks to address this issue, assuming that each chunk focuses on a single aspect. Those segments serve as semantic units and are used to retrieve knowledge from various sources. Then, a Graph Neural Network (Veličković et al., 2018) is constructed, which assembles the retrieved knowledge to predict the answer.

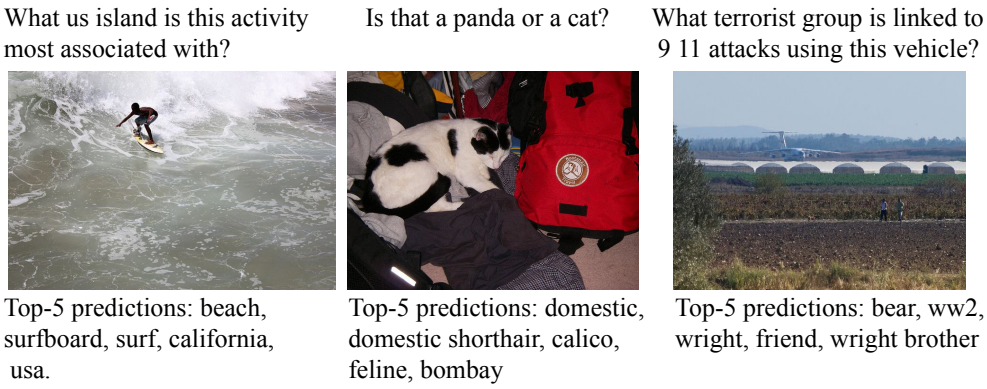

Figure 4.2: Examples where the correct answer is not in the top-5 predictions from a ViLBERT model.

4.1.2 Learning-based Answer Candidate Generator

Most high-performing systems frame the VQA task as an answer classification problem, where the answers are collected from all possible candidates in the training set. For general visual questions, with a tremendous amount of training data, those answers cover most of the cases in the test set as most answers are relatively simple and common.

However, this framework has a few drawbacks. First, using all the potential answers fails to explore the ontology of the key queried objects sufficiently. For example, as shown in the left of Figure 4.2, the questions asked about "US islands"; however, none of the top predictions is an actual US island. Second, using the same set of answer candidates for all visual questions leads to ignorance of the questions' linguistic features. As shown in the middle example in Figure 4.2, if the question is an alternative question, the answer should be in the question. Though the top predictions indicate that the VQA system understands the content of the question and does reasonable visual recognition, the linguistic features are ignored. Third, VQA systems should read and comprehend the retrieved knowledge to infer the answer that might have never been seen before. For example, as shown in the right of Figure 4.2, "Al-Qaeda" never appears in the training set; however, knowledge-based

VQA systems can infer the answer from the retrieved Wikipedia sentence "The September 11 attacks, often referred to as 9/11, were a series of four coordinated terrorist attacks by the Wahhabi Islamist terrorist group al-Qaeda against the United States on the morning of Tuesday, September 11, 2001.". Besides, as the answer candidates are fixed, it is hard for the systems to absorb more training data in an online learning fashion. For each batch of new data, the systems do not know whether the new answers should be kept or not. This prevents VQA systems from serving as continuously improving tools that are essential for lots of daily applications. The issue becomes more problematic for the knowledge-based visual questions that require different domain expertise.

There is a recent trend to verify the potential answer candidates(Wu et al., 2020; Si et al., 2021; Wu et al., 2021). During verification, the model can utilize more answer-specific knowledge, such as explanations and retrieved Wikipedia sentences. As the verification process is computationally costly, it is not possible to verify every answer. Instead, most papers choose the top-K answers from a baseline model for simplification. Building a better answer candidate set would be helpful for all of these types of VQA systems.

We propose to use a learning-based approach to generate a set of answer candidates. We will explore the idea of first converting visual questions to a set of possible textual questions and then using a textual QA or reading comprehension module to generate the answer candidate set according to the converted questions. While collecting multi-modal annotations is hard, there are more data available in textual QA domains (Yang et al., 2015), potentially leading to a better set of generated answer candidates. Technically, we first break down the visual question and extract a set of noun chunks; then, we define a possible set of replacements for each noun chunk, including the nucleus noun and the common object labels with attributes of the top referenced objects. The set of possible textual questions $\mathcal{P}$ are formed by traversing all of the combinations of the replacements for those extracted noun chunks. Then, we train a model to rank those converted questions based on whether the conversion would help generate the correct answer given the document. In particular,

Figure 4.3: Examples of different factual knowledge retrieved for different answers for visually similar objects.

a binary classifier, taking as inputs the converted question and the visual representation of the image, will be used to output the score. We will employ a triplet loss to train the model where the positive examples are the conversion that leads to correct answers, and the negative examples are constructed by randomly replacing the noun chunks with objects in the image. We also finetune the textual QA system that takes the positive conversion as inputs. As the goal is to generate candidates instead of directly predicting the correct answer, we sample multiple conversions and merged the generated answers during inference.

4.2 Long Term Proposals

4.2.1 Verifying Retrieved Knowledge

Our completed work explores various textual resources to improve VQA performance, including image captions, justifications, factual statements, concept sentences, etc. However, most of those resources are not guaranteed to provide supportive evidence for solving the visual questions. Though human justifications reveal the desired underlying reasoning process, they are much more expensive to collect, especially when the questions are domain-specific and require human expertise. Image captioning suffers from object hallucination (Rohrbach et al., 2018) issues that could involve objects that do not appear in the image.

General knowledge retrieved from the Internet may contain fake information. Though factual statements and concepts from certified resources are valid in their respects, they may not be relevant to solving the visual questions due to the two-step retrieval-prediction process. In some worse cases, they can also mislead the VQA systems because the tiny visual difference often ignored in the retrieval process can lead to a significant difference in the factual statement, as shown in Figure 4.3.

While it is important for the VQA systems to access various types of resources for knowledge acquisition, the appropriateness of the retrieved knowledge for the specific visual question plays a crucial role in how likely the knowledge could help improve the performance. Therefore, we propose to ground the knowledge sentences in the image to verify the existence of the prerequisites of utilizing them in the VQA system. We would like first to chunk each sentence to a set of attribute phrases (Hendricks et al., 2018) as a checklist. Then, we compute a score for each item from the checklist to represent how likely the required feature is in the given image. Finally, this score is fed to the VQA system so that the system could decide whether use the knowledge sentence or not.

4.2.2 Explainable VQA systems

VQA systems' ability to explain their reasoning is critical to their utility. The opacity nature of recent high-performing learning-based VQA models hinders the users from trusting the model predictions as a final decision.

Texts not only provide additional information to answer visual questions better, but they also reveal a more interpretable nature than visual features, providing us opportunities to better understand the underlying reasoning process (Rajani et al., 2019). Previous work explored generating single sentence explanations (Wu and Mooney, 2019a) to mimic human justifications for commonsense question answering. Instead, we aim to build an explainable VQA system for knowledge-based visual questions that utilize external textual information and provide explanations using the supporting facts. The goal is more ambitious because

it requires the explanation system to combine each reasoning step in the question that may utilize different textual resources.

Besides, as more modalities of knowledge sources are involved, the format of the explanation can be different for different visual questions. For example, we could provide single sentence textual explanations for simple commonsense questions(Park et al., 2018; Wu and Mooney, 2019a); while referencing a Wikipedia's link is a good solution for explaining some knowledge-based questions. Finding similar visual content from other web images with richer annotations also helps convince users to trust the model predictions.

Chapter 5

Conclusion

This proposal explores utilizing various textual resources to improve visual question answering in terms of performance, robustness to distribution shift, and interpretability.

We presented the approach that generates image captions to serve as textual inputs for better VQA scores to complement visual inputs.

We observe that VQA systems are easy to take short-cuts and focus on superficial statistics when predicting the answer. The problem becomes more severe as the distribution of the training and test set are different. We present a self-critical training approach that first encourages the model to focus on the proper object when predicting the correct answer and then discourage focusing on that object when predicting a wrong answer. The right set of objects is parsed from human textual explanations. Our approach prevents the VQA systems from simply relying on superficial statistics, and therefore helps them be robust to distribution shift. Besides, we also use competing explanations to improve VQA performance. We present two sets of competing explanations, generated and retrieved explanations.

We presented MAVEx, a novel approach for knowledge-based visual question answering. The goal is to retrieve answer-specific textual and visual knowledge from different knowledge sources and learn what sources contain the most relevant information. We formulate the problem as answer validation, where the goal is to learn to verify the validity

of a set of candidate answers according to the retrieved knowledge. MAVEx demonstrates the clear advantages of answer-guided knowledge retrieval, achieving new state-of-the-art performance on the OK-VQA dataset, the largest knowledge-based dataset to date.

Finally, we proposed some short and long-term future extensions to our work. In the short term, we focus on breaking down visual questions into multiple semantic segments such that they can drive the retrieval of relevant knowledge from multiple external sources. This would be especially helpful when the visual questions cover multiple aspects inside and outside the image. Also, as another short-term goal, we would like to generate a customized set of answer candidates based on the ontology and the retrieved knowledge.

Previous works mainly focusing on exploring different types of textual resources; however, the appropriateness of using the resources requires further investigates. Therefore, our first long-term focus is on verifying the properness of the retrieved knowledge for a better trustworthy system. Secondly, we plan on work on building an interpretable VQA system.

Bibliography

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *CVPR*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and VQA. In *CVPR*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A Nucleus for a Web of Open Data. In *The semantic web*. Springer.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *ICCV*.

Sumithra Bhakthavatsalam, Kyle Richardson, Niket Tandon, and Peter Clark. 2020. Do Dogs Have Whiskers? A New Knowledge Base of hasPart Relations. *arXiv preprint arXiv:2006.07510*.

Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. 2019. Murel: Mul-

timodal relational reasoning for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1989–1998.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: UNiversal Image-TExt Representation learning. In *ECCV*. Springer.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *Computer Vision and Image Understanding*.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming Question Answering Datasets into Natural Language Inference datasets. *arXiv preprint arXiv:1809.02922*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ACL*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *CVPR*, pages 2625–2634.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *EMNLP*.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. ConceptBert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *ICCV*.

Sen He, Wentong Liao, Hamed R Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. 2020. Image captioning through image transformer. In *Proceedings of the Asian Conference on Computer Vision*.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. To appear.

Drew A Hudson and Christopher D Manning. 2019. GQA: A New Dataset for Compositional Question Answering over Real-World Images. *CVPR*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not Explanation. In *NAACL*.

Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0. 1: the Winning Entry to the VQA Challenge 2018. *arXiv preprint arXiv:1807.09956*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-semantic Alignments for Generating Image Descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. In *ACL*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *NeurIPS*.

Thomas N Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. *ICLR*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *IJCV*.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A UNiversal encoder for Vision and Language by Cross-Modal Pre-training. In *AAAI*.

Guohao Li, Xin Wang, and Wenwu Zhu. 2020b. Boosting Visual Question Answering with Context-aware Knowledge Aggregation. In *ACM Conference on Multimedia*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557*.

Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo. 2018a. Tell-and-Answer: Towards Explainable Visual Question Answering using Attributes and Captions. *arXiv preprint arXiv:1801.09041*.

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018b. VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions. *ECCV*.

Bei Liu, Zhicheng Huang, Zhaoyang Zeng, Zheyu Chen, and Jianlong Fu. 2019. Learning Rich Image Region Representation for Visual Question Answering. *arXiv preprint arXiv:1910.13077*.

Hugo Liu and Push Singh. 2004. ConceptNet: a Practical Commonsense Reasoning Tool-kit. *BT technology journal*.

Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. 2021. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. *ECCV*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020a. 12-in-1: Multi-Task Vision and Language Representation Learning. In *CVPR*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020b. 12-in-1: Multi-task vision and language representation learning. In *CVPR*.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability Objective for Training Descriptive Captions. In *CVPR*.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA. In *CVPR*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *CVPR*.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out-of-The-Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering. In *NeurIPS*.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *CVPR*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*.

Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2018. Exploring Human-Like Attention Supervision in Visual Question Answering. In *AAAI*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming Language Priors in Visual Question Answering with Adversarial Regularization. In *NeurIPS*.

Kiran Ramnath and Mark Hasegawa-Johnson. 2021. Seeing is Knowing! Fact-based Visual Question Answering using Knowledge Graph Embeddings. *AAAI*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015a. Exploring Models and Data for Image Question Answering. In *NIPS*, pages 2953–2961.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015b. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. In *NIPS*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations. In *IJCAI*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *ICCV*.

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Dhruv Batra, and Devi Parikh. 2019. Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded. In *ICCV*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*.

Qingyi Si, Zheng Lin, Mingyu Zheng, Peng Fu, and Weiping Wang. 2021. Check it again: Progressive visual question answering via visual entailment. *arXiv preprint arXiv:2106.04605*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models that Can Read. In *CVPR*.

Jasdeep Singh, Vincent Ying, and Alex Nutkiewicz. 2018. Attention on Attention: Architectures for Visual Question Answering (VQA). *arXiv preprint arXiv:1803.07724*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*.

Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In *NeurIPS*.

Alexander Trott, Caiming Xiong, and Richard Socher. 2018. Interpretable Counting for Visual Question Answering.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *NeurIPS*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph Attention Networks. *ICLR*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. *arXiv preprint arXiv:1610.02424*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-Based Visual Question Answering. *PAMI*.

Wikipedia contributors. 2004. Plagiarism — Wikipedia, the free encyclopedia. [Online; accessed 22-July-2004].

Jialin Wu, Liyan Chen, and Raymond J Mooney. 2020. Improving vqa and its explanations by comparing competing explanations. *arXiv preprint arXiv:2006.15631*.

Jialin Wu, Dai Li, Yu Yang, Chandrajit Bajaj, and Xiangyang Ji. 2018. Dynamic Filtering with Large Sampling Field for Convnets. *ECCV*.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2021. Multi-Modal Answer Validation for Knowledge-Based VQA. *arXiv preprint arXiv:2103.12248*.

Jialin Wu and Raymond J Mooney. 2019a. Faithful Multimodal Explanation for Visual Question Answering. In *ACL BlackboxNLP Workshop*.

Jialin Wu and Raymond J Mooney. 2019b. Self-critical reasoning for robust visual question answering. *arXiv preprint arXiv:1905.09998*.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph R-CNN for Scene Graph Generation. In *ECCV*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling Context in Referring Expressions. In *ECCV*, pages 69–85. Springer.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019a. Deep Modular Co-Attention Networks for Visual Question Answering. In *CVPR*.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019b. Deep modular co-attention networks for visual question answering. In *CVPR*.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-Down Neural Attention by Excitation Backprop. *IJCV*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *ICLR*.

Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. 2019. Interpretable Visual Question Answering by Visual Grounding from Attention Supervision Mining. In *WACV*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *AAAI*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded Question Answering in Images. In *CVPR*, pages 4995–5004.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering. In *IJCAI*.