

SUBMITTED VERSION

Hui Li, Peng Wang, Chunhua Shen, Anton van den Hengel

Visual question answering as reading comprehension

Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019 / vol.2019-June, pp.6312-6321

© 2019 IEEE

Published version at: <http://dx.doi.org/10.1109/CVPR.2019.00648>

PERMISSIONS

<https://www.ieee.org/publications/rights/author-posting-policy.html>

Author Posting of IEEE Copyrighted Papers Online

The IEEE Publication Services & Products Board (PSPB) last revised its Operations Manual Section 8.1.9 on Electronic Information Dissemination (known familiarly as "author posting policy") on 7 December 2012.

PSPB accepted the recommendations of an ad hoc committee, which reviewed the policy that had previously been revised in November 2010. The highlights of the current policy are as follows:

- The policy reaffirms the principle that authors are free to post their own version of their IEEE periodical or conference articles on their personal Web sites, those of their employers, or their funding agencies for the purpose of meeting public availability requirements prescribed by their funding agencies. Authors may post their version of an article as accepted for publication in an IEEE periodical or conference proceedings. Posting of the final PDF, as published by IEEE *Xplore*®, continues to be prohibited, except for open-access journal articles supported by payment of an article processing charge (APC), whose authors may freely post the final version.
- The policy provides that IEEE periodicals will make available to each author a preprint version of that person's article that includes the Digital Object Identifier, IEEE's copyright notice, and a notice showing the article has been accepted for publication.
- The policy states that authors are allowed to post versions of their articles on approved third-party servers that are operated by not-for-profit organizations. Because IEEE policy provides that authors are free to follow public access mandates of government funding agencies, IEEE authors may follow requirements to deposit their accepted manuscripts in those government repositories.

IEEE distributes accepted versions of journal articles for author posting through the Author Gateway, now used by all journals produced by IEEE Publishing Operations. (Some journals use services from external vendors, and these journals are encouraged to adopt similar services for the convenience of authors.) Authors' versions distributed through the Author Gateway include a live link to articles in IEEE *Xplore*. Most conferences do not use the Author Gateway; authors of conference articles should feel free to post their own version of their articles as accepted for publication by an IEEE conference, with the addition of a copyright notice and a Digital Object Identifier to the version of record in IEEE *Xplore*.

15 October, 2020

<http://hdl.handle.net/2440/127237>

Visual Question Answering as Reading Comprehension

Hui Li¹, Peng Wang², Chunhua Shen¹ and Anton van den Hengel¹

¹Australian Centre for Robotic Vision, The University of Adelaide, Australia

²School of Computer Science, Northwestern Polytechnical University, China

Abstract

Visual question answering (VQA) demands simultaneous comprehension of both the image visual content and natural language questions. In some cases, the reasoning needs the help of common sense or general knowledge which usually appear in the form of text. Current methods jointly embed both the visual information and the textual feature into the same space. However, how to model the complex interactions between the two different modalities is not an easy task. In contrast to struggling on multimodal feature fusion, in this paper, we propose to unify all the input information by natural language so as to convert VQA into a machine reading comprehension problem. With this transformation, our method not only can tackle VQA datasets that focus on observation based questions, but can also be naturally extended to handle knowledge-based VQA which requires to explore large-scale external knowledge base. It is a step towards being able to exploit large volumes of text and natural language processing techniques to address VQA problem. Two types of models are proposed to deal with open-ended VQA and multiple-choice VQA respectively. We evaluate our models on three VQA benchmarks. The comparable performance with the state-of-the-art demonstrates the effectiveness of the proposed method.

4. Experiments	5
4.1. Datasets	6
4.2. Implementation Details	6
4.2.1 Results Analysis on FVQA	6
4.2.2 Results Analysis on Visual Genome QA	7
4.2.3 Results Analysis on Visual7W	8
5. Conclusion	11

Contents

1. Introduction	2
2. Related Work	2
2.1. Joint embedding	2
2.2. Knowledge-based VQA	3
2.3. Textual Question Answering	3
3. VQA Models	3
3.1. QANet	3
3.2. Open-ended VQA model	4
3.3. Multiple-choice VQA model	5

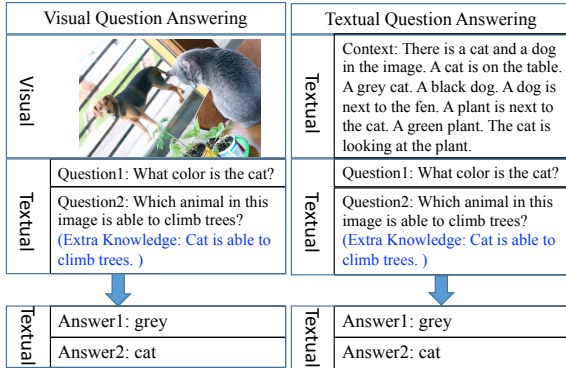


Figure 1. Comparison between VQA and TQA. Question1 is observation based, which can be inferred from the image itself. Question2 is knowledge based, which has to refer knowledge beyond the image. Extra knowledge commonly appears in text, which is easier to be combined to the context paragraph in TQA.

1. Introduction

Visual Question Answering (VQA) is an emerging problem which requires the algorithm to answer arbitrary natural language questions about a given image. It attracts a large amount of interests in both computer vision and Natural Language Processing (NLP) communities, because of its numerous potential applications in autonomous agents and virtual assistants.

To some extent, VQA is closely related to the task of Textual Question Answering (TQA, also known as machine reading comprehension), which asks the machine to answer questions based on a given paragraph of text. However, VQA seems to be more challenging because of the additional visual supporting information. As compared in Figure 1, the inputs in TQA are both pure text, while VQA has to integrate the visual information from image with the textual content from questions. On one hand, image has a higher dimension than text and lacks the structure and grammatical rules of language, which increase the difficulty in semantic analysis. On the other hand, the algorithm has to jointly embed the visual and textual information that come from two distinct modalities.

Most approaches in VQA adopt deep Convolutional Neural Networks (CNNs) to represent images and Recurrent Neural Networks (RNNs) to represent sentences or phrases. The extracted visual and textual feature vectors are then jointly embedded by concatenation, element-wise sum or product to infer the answer. Fukui *et al.* [8] argued that such simple kinds of merging might not be expressive enough to fully capture the complex associations between the two different modalities and they proposed a Multimodal Compact Bilinear pooling method (MCB) for VQA. It would be even complex if extra knowledge is required to be combined for reasoning. Li *et al.* [18] proposed to embed knowledge in memory slots and incorporated ex-

ternal knowledge with image, question and answer features by Dynamic Memory Networks (DMN).

In this work, different from exploring the high-dimensional and noisy image feature vectors to infer the answer, we express the image explicitly by natural language. Compared to image feature, natural language represents a higher level of abstraction and is full of semantic information. For example, the phrase “a red hat” will represent various styles of “red hats” captured in the image. Thus the VQA problem is converted to TQA. With this transformation, we can easily incorporate external knowledge as they are all in the form of natural language. In addition, the complex multimodal feature fusion problem can be avoided. There are works on TQA and image description [11, 15], in this work we move step-forward to connect those methods in answering image based questions.

The main contributions of this work is three-fold:

- 1) We propose a new thought of solving VQA problem. Instead of integrating feature vectors from different modalities, we represent image content explicitly by natural language and solve VQA as a reading comprehension problem. Thus we can resort to the abundant research results in NLP community to handle VQA problem. Using text and NLP techniques allows very convenient access to higher-level information in identifying referred objects, and makes the inferring more interpretable. Moreover, text data is more easier to be collected than images. Our method makes it possible to exploit large volumes of text in understanding images, actions, and commands.
- 2) Two types of VQA models are proposed to address the open-end VQA and the multiple-choice VQA respectively, considering their own characteristics. Based on the converted text description and the attention mechanism used in the models, it becomes more accurate to retrieve related information from the context. The answer inferring process is human-readable. The proposed models show comparable performance with the state-of-the-art on three different types of VQA datasets, which demonstrates their feasibility and effectiveness.
- 3) Most VQA methods cannot handle the knowledge based VQA or have poor performance because of the complicated knowledge embedding. In contrast, our method can be easily extended to address knowledge based VQA as they have the same modality.

2. Related Work

2.1. Joint embedding

Current approaches need to integrate features from both image and text, which is a multimodal feature fusion problem. Most existing approaches use simple manners such as vector concatenation [21, 26, 30], element-wise product or sum [1, 9, 34] to jointly embed the visual feature and textual

feature. Fukui *et al.* [8] argued that these simple manners are not expressive enough and proposed MCB which allows a multiplicative interaction between all elements of image and text vectors. Nevertheless, it needs to project the image and text features to a higher dimensional space firstly (*e.g.*, 16000D for good performance), and then convolves both vectors by element-wise product in Fast Fourier Transform space. Multimodal Low-rank Bilinear pooling (MLB) [13] and Multimodal Factorized Bilinear pooling (MFB) [37] were proposed later. MLB uses Hadamard product to integrate the multimodal features, while MFB expands the multimodal features to a high-dimensional space firstly and then integrates them with Hadamard product. Kim *et al.* [12] also presented Multimodal Residual Networks (MRN) to learn the multimodality from vision and language information, which inherently adopts shortcuts and joint residual mappings to learn the multimodal interactions, inspired by the outstanding performance of deep residual learning.

It can be observed that how to integrate multimodal features plays a critical role in VQA. In contrast to considering the multimodal feature fusion manner, in this work, we convert the visual information directly to text so that all features are from textual information, which escapes the feature jointly embedding issue immediately.

2.2. Knowledge-based VQA

There are some researches in the NLP community about answering questions incorporating external knowledge using either semantic parsing [3, 33] or information retrieval [4, 5]. They are all based on textual features. It is non-trivial to extend these methods to knowledge based VQA because of the unstructured visual input.

In [32], a method was proposed for VQA that combines image representation with extra information extracted from a general knowledge base according to predicted image attributes. The method makes it possible to answer questions beyond the image, but the extracted knowledge is discrete pieces of text, without structural representations. Ahab [28] used explicit reasoning over an resource description framework knowledge base to derive the answer. But the method largely depends on the pre-defined templates, which restricts its application. Wang *et al.* [29] introduced the ‘‘Fact-based VQA (FVQA)’’ problem and proposed a semantic-parsing based method for supporting facts retrieval. A matching score is computed to obtain the most relevant support fact and the final answer. This method is vulnerable to misconceptions caused by synonyms and homographs. A learning based approach was then developed in [23] for FVQA, which learns a parametric mapping of facts and question-image pairs to an embedding space that permits to assess their compatibility. Features are concatenated over the image-question-answer-facts tuples. The work in [39] and [18] exploited DMN to incorporate external knowledge.

Our method is more straightforward to deal with the knowledge-based VQA. By representing the image visual information as text, we unify the image-question-answer-facts tuples into the natural language space, and tackle it using reading comprehension techniques in NLP.

2.3. Textual Question Answering

Textual Question Answering (also known as reading comprehension) aims to answer questions based on given paragraphs. It is a typical cornerstone in the NLP domain, which assesses the ability of algorithms in understanding human language. Significant progress has been made over the past years due to the using of end-to-end neural network models and attention mechanism, such as DMN [17], r-net [31], DrQA [6], QANet [36], and most recently BERT [7]. Many techniques in QA have been inherited in solving VQA problem, such as the attention mechanism, DMN, *etc.* In this work, we try to solve the VQA problem built upon QANet.

3. VQA Models

Our method is build upon the newly proposed QANet [36] for TQA problem. In this section, we firstly outline QANet and its modules that will be used in our VQA models. Then we propose two types of models to tackle the open-ended VQA and the multiple-choice VQA separately.

3.1. QANet

QANet is a fast and accurate end-to-end model for TQA. It consists of embedding block, embedding encoder, context-query attention block, model encoder and output layer. Instead of using RNNs to process sequential text, its encoder consists exclusively of convolution and self-attention. A context-question attention layer is followed to learn the interactions between them. The resulting features are encoded again, and finally decoded to the position of answer in the context. The details can refer [36].

Input Embedding Block: This module is used to embed each word in the context and question into a vector. For each word, the representation is the concatenation of word embedding and character embedding, *i.e.*, $\mathbf{x} = [\mathbf{x}_w, \mathbf{x}_c]$, where \mathbf{x}_w is the word embedding obtained from pre-trained GloVe [24], \mathbf{x}_c is from character embedding, which is the maximum value of each row in the concatenated character representation matrix. A two-layer highway network is applied on x to obtain the embedding features.

Embedding Encoder Block: It is a stack of convolutional layers, self-attention layers, feed forward layers and normalization layers, as illustrated in Figure 2. Depth-wise separable convolutions are adopted here for better memory and generalization ability. Multi-head attention mechanism is applied which models global interactions.

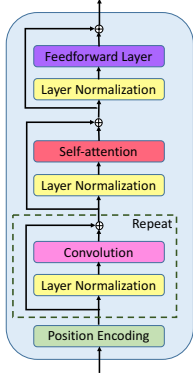


Figure 2. The structure of encoder block used in QANet, which is shared by embedding encoder and model encoder. The number of convolutional layers varies according to design. Layer normalization and residual connection are adopted between every layer for better performance.

Context-question Attention Block: It is designed to extract the most related features between the context and the question words. There are context-to-question attention and question-to-context attention constructed in the model. Denote \mathbf{C} and \mathbf{Q} as the encoded context and question features respectively, where $\mathbf{C} = \{c_1, c_2, \dots, c_n\}$ with n words, and $\mathbf{Q} = \{q_1, q_2, \dots, q_m\}$ with m words. The context-to-question attention is defined as $\mathbf{A} = \bar{\mathbf{S}} \cdot \mathbf{Q}^T$, where $\bar{\mathbf{S}} \in \mathcal{R}^{n \times m}$ is the similarity matrix between each pair of context and question words, and $\bar{\mathbf{S}}$ is the normalization of \mathbf{S} by applying *softmax* on each row. “ \cdot ” is matrix product. The question-to-context attention is defined as $\mathbf{B} = \bar{\mathbf{S}} \cdot \bar{\mathbf{S}}^T \cdot \mathbf{C}^T$, where $\bar{\mathbf{S}}$ is the normalization of \mathbf{S} by applying *softmax* on each column. The similarity function is defined as $f(\mathbf{q}, \mathbf{c}) = \mathbf{W}_0[\mathbf{q}, \mathbf{c}, \mathbf{q} \odot \mathbf{c}]$, where \odot is the element-wise multiplication of each \mathbf{q} and \mathbf{c} , \mathbf{W}_0 is the weight to be learned.

Model Encoder Block: This block takes $[c, a, c \odot a, c \odot b]$ as input, where a and b are a row of the attention matrix \mathbf{A} and \mathbf{B} respectively. It shares parameters with the embedding encoder block.

Output Layer: The output layer predicts the probability of each position in the context being the start or end locations of the answer, based on the outputs of 3 repetitions of model encoder.

3.2. Open-ended VQA model

Instead of merging the visual and textual features into the same space, we convert the image wholly into a descriptive paragraph, so that all the input information is unified as text. It avoids the challenge task of multimodal feature fusion, and can extend to deal with the knowledge-based VQA straightforwardly. The answer inference is more obvious from the semantically high level text description, in contrast to the unstructured image feature. The architecture of our proposed model is presented in Figure 3. Besides the modules such as embedding block, embedding encoder, context-question attention block and model encoder used in QANet, we add another input pre-processing block and modify the output block for the open-ended VQA problem.

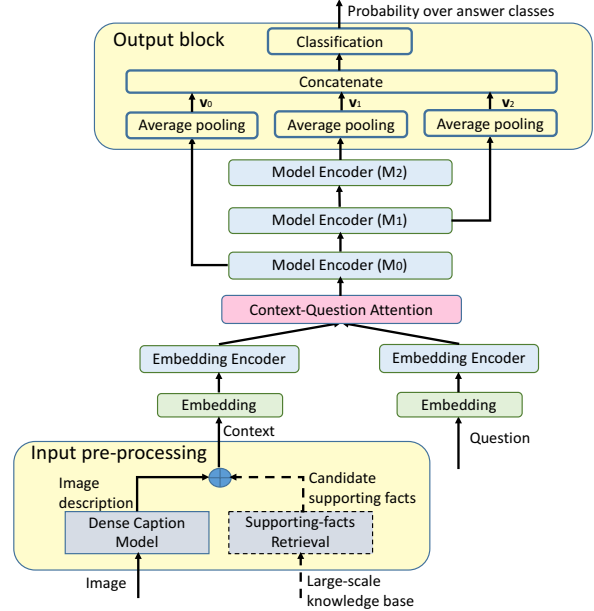


Figure 3. Open-ended VQA model. By representing image with neural language, we convert VQA as a reading comprehension problem. Extra knowledge can be added naturally into the model because of the same modality.

The input pre-processing block may include an image description module or/and external knowledge retrieval module, depending on the task. The image description module aims to represent the image information by a text paragraph. As the question to be asked is undetermined and can be about any part of the image, a simple summary sentence or paragraph is insufficient to cover all the details. It is prefer to collect image information at different levels, from single object, concept, sub-region to the whole image. The Visual Genome dataset [16] provides various human-generated region descriptions for each image, as presented in Figure 4. Regions may overlap with each other but have different focus of interest. The descriptions range from the states of a single object (color, shape, action, *etc.*) to the relationships between objects (spatial positions, *etc.*). Based on this dataset, Johnson *et al.* [11] proposed the dense caption task, which aims to generate sophisticated lever of regions of interest in an image, and describe each region by a sentence. The generated region captions provide a detailed descriptions about the image. Here we combine them as the image description for QANet.

For VQA that requires auxiliary knowledge beyond the image, a supporting-facts retrieval module is needed. It is demanded to extract related supporting facts from a general large-scale knowledge base but ignore the irrelevant ones. Wang *et al.* [29] proposed to query the knowledge bases according to the estimated query types and visual concepts detected from the image. A keyword matching technique is used to retrieve the ultimate supporting fact as well as

the answer. Rather than apply the heuristic matching approach which is vulnerable to homographs and synonyms, here we make use of all the retrieved candidate supporting facts as context. Since both image description and supporting facts are expressed by natural language, they can merge together easily by concatenation. The QANet will then encode the textual information, seek the correlation between context and question, and predict the answer.

The output layer is also task-specific. If the answer is definitely included in the text paragraph, we can continue using the output layer in QANet by predicting the start and end positions of answer in the context. However, in some cases, the answer may not explicitly show up in the context. For example, region descriptions generally do not include the answer to the question “When the image is taken?” proposed for the image shown in Figure 4. Some reasoning is required in this circumstances. It is hoped that the model can learn some clues from region descriptions such as the bright colors presented in the text so as to predict the answer “Day time”. To address this situation, we built the output layer as a multi-class classification layer, and predict the probabilities over pre-defined answer classes based on the output features of three model encoders M_0, M_1, M_2 , as shown in Figure 3. An average pooling layer is adopted firstly. The resulted feature vectors are then concatenated and projected to an output space with the number of answer classes. The probability of being each class is calculated as $\mathbf{p} = \text{softmax}(\mathbf{W}[\mathbf{v}_0; \mathbf{v}_1; \mathbf{v}_2])$, where \mathbf{W} is the parameter to be learned. Cross entropy loss is employed here as the object function to train the model.

3.3. Multiple-choice VQA model

Multiple-choice VQA provides several pre-specified choices, besides the image and question. The algorithm is asked to pick the most possible answer from these multiple choices. It can be solved directly by the aforementioned open-ended VQA model by predicting the answer and matching with the provided multiple choices. However, this approach does not take full advantage of the provided information. Inspired by [8, 10], which receive the answer as input as well and show substantial improvement in per-



Figure 4. 10 region description examples for an image in Visual Genome dataset [16], where each region description corresponds to a bounding box with the same color in the image. The descriptions range from the states of a single object (color, trait, action, etc.) to object relationships.

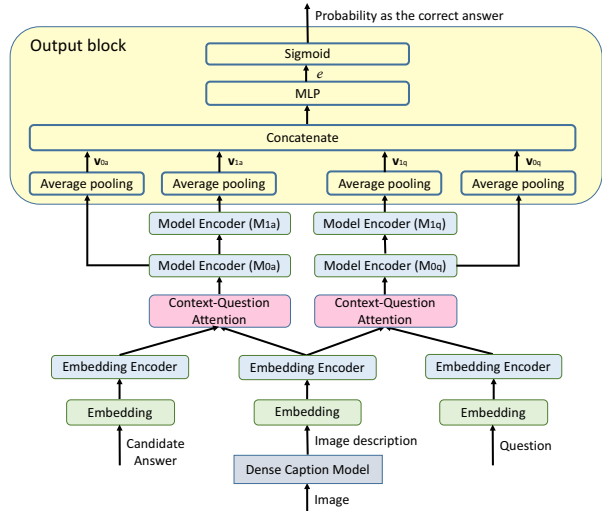


Figure 5. Multiple-choice VQA model. It takes image-question-answer triplet as input and encodes both interactions of question and answer with the context.

formance, we propose another model for multiple-choice VQA problem.

As presented in Figure 5, aside from the question and the converted image description, our model also takes a candidate answer choice as input, and calculates the interaction between the candidate answer and context. If the answer is true, the encoded features of \mathbf{v}_{0a} and \mathbf{v}_{1a} are strong correlated with \mathbf{v}_{0q} and \mathbf{v}_{1q} . Otherwise, the features may be independent. A multilayer perceptrons (MLP) is trained on the concatenated features, i.e., $e = \mathbf{W}_2 \max(0, \mathbf{W}_1[\mathbf{v}_{0a}; \mathbf{v}_{1a}; \mathbf{v}_{0q}; \mathbf{v}_{1q}])$. Dropout with a probability of 0.5 is used after the first layer. The objective is to predict whether the image-question-answer triplet is correct or not. Hence a *sigmoid* function is followed to transform the feature into probability. A binary logistic loss is employed to train the model.

Compared to the open-ended VQA model which selects the top answers as class labels and excludes the rare answers, multiple-choice VQA model encodes the candidate answers directly. Thus It will cover more answer choices. For similar answer expressions, such as “During the day time”, “During daytime”, “In the daytime”, the model can learn the similarity itself by embedding and encoder, rather than use the heuristic answer normalization. Hence, it avoids the chance of regarding them as different classes and learning to distinguish them from the training data.

4. Experiments

In this section, we perform extensive experiments to assess the effectiveness of the proposed approach. All the experiments are conducted on an NVIDIA Titan X GPU with 12 GB memory. The models are implemented in PyTorch.

4.1. Datasets

We evaluate the models on three public available datasets. Each dataset has its own peculiarity.

FVQA [29] (Fact-based VQA) is a dataset that not only provides image-question-answer triplets, but also collects extra knowledge for each visual concept. A large-scale knowledge base (with about 193, 449 fact sentences) is constructed by extracting the top visual concepts from all the images and querying those concepts from three knowledge bases, including DBPedia [2], ConceptNet [19] and WebChild [27]. FVQA collects 2190 images and 5826 questions. The dataset has 5 train/test splits. Each split has 1100 training images and 1090 test images, providing roughly 2927 and 2899 questions for training and test respectively. The questions are categorized into 32 classes.

Visual Genome [16] is a dataset that has abundant information about image and language. It contains 108, 077 images and 1, 445, 233 Question and Answer (QA) pairs. It also supplies 5.4 Million region descriptions as we introduced before. These descriptions give a finer level of details about the image and are used as the ground-truth text representation in our experiments. As there is no official training and test split, we random split 54, 039/4038/50, 000 images for training/validation/test as done by [30], which results in 723, 917/53, 494/667, 911 training/validation/test QA pairs. There are 6 types of questions including *what*, *where*, *how*, *when*, *who*, and *why* (“6W”).

Visual7W [38] is a subset of Visual Genome, which aims exclusively for VQA. It contains 47, 300 images with 139, 868 QA pairs. Answers in Visual7W are in a multiple choice format, where each question has four answer candidates, with only one correct. Here we evaluate our model on the Telling QA subtask, which also consists of the “6W” questions. The QA pairs have been split into 69, 817/28, 020/42, 031 for training/validation/test.

4.2. Implementation Details

FVQA dataset needs to access external knowledge to answer the given question. We follow the question-to-query(QQ) mapping method proposed in FVQA [29] and use the top-3-QQmapping results to extract candidate supporting facts from the whole knowledge base. The extracted supporting facts contain not only the image information, but also demanded knowledge beyond the image. All the facts are combined together into a paragraph. QANet [36] is followed directly to predict the answer position in the paragraph. We use the default parameters in QANet, and fine-tune the model from the one that well-trained on general reading comprehension dataset SQuAD [25]. The model is finetuned with a learning rate of 0.001 for 10 epochs and 0.0001 for another 10 epochs on each training split separately, and tested on the corresponding test split.

Visual Genome provides ground-truth region descrip-

tions. Based on this labeling, Justin *et al.* [11] proposed a fully convolutional localization network to jointly generate finer level of regions and captions. Yang *et al.* [35] proposed a model pipeline based on joint inference and visual context fusion, which achieves much better dense caption results. We re-train these models using our training split, and predict dense captions for test images. The top-5000 frequently appeared answers are selected as class labels to train the open-ended VQA model. Considering the average paragraph length, we use a paragraph limit of 500 words and 4 attention heads in encoder blocks for fast training. The model is trained from scratch using ADAM optimizer [14] for 30 epochs. The learning rate is set to 0.001 initially, with a decay rate of 0.8 every 3 epochs until 0.0001.

As to Visual7W dataset which has multiple-choice answers provided for each question, we train the multiple-choice VQA model. we randomly sample two negative answers from the multiple choices for each positive example, and shuffle all the image-question-answer triplets to train the model.

4.2.1 Results Analysis on FVQA

We use answer accuracy to evaluate the model, following [29]. The predicted answer is determined to be correct if the string matches the corresponding ground-truth answer. (All the answers have been normalized to eliminate the the differences caused by singular-plurals, cases, punctuations, articles, *etc.*) The top-1 and top-3 accuracies are calculated for each evaluated methods. The averaged answer accuracy across 5 test splits is reported here as the overall accuracy.

Table 1. Experimental Results on FVQA. Our method with finetuned QANet achieves the highest top-1 accuracy.

Method	Overall Accuracy (%)	
	top-1	top-3
LSTM-Question		
+Image+Pre-VQA [29]	24.98	40.40
Hie-Question		
+Image+Pre-VQA [29]	43.14	59.44
FVQA (top-3-QQmapping) [29]	56.91	64.65
FVQA (Ensemble) [29]	58.76	-
Question+Visual Concepts [23]	62.20	75.60
Ours-pretrained QANet	55.14	63.34
Ours-QANet-train-from-scratch	47.87	54.24
Ours-finetuned QANet	62.94	70.08

Table 1 shows the overall accuracy of our method based on supporting facts retrieved by using the top-3-QQmapping results in [29]. Our method with finetuned QANet achieves the highest top-1 accuracy, which is 0.7% higher than the state-of-the-art result. It should be note that [23] has the top-3-QQmapping accuracy of 91.97%, which is 9% higher than what we used. The QQmapping results have a direct influence on retrieving the related supporting facts. With the same top-3-QQmapping results, our approach outperforms the method in [29] about 6% on top-

1 and top-3 answer accuracies respectively, and even performs better than the ensemble method in [29]. As this work aims to propose an alternative approach for VQA problem by representing all the input information with natural language and solving VQA as reading comprehension, we leave the improvement of QQmapping as a future work.

In addition, we test the QANet model without finetuned by FVQA training data, *i.e.*, the one trained only by general reading comprehension dataset SQuAD [25]. Experimental results show that the pre-trained QANet model is also feasible on FVQA dataset. The model gives even better results than that trained from scratch solely by FVQA training data, because of the small amount of available data. This phenomenon illustrates that with our framework, we can draw on the experience of well-trained TQA models and make use of the large volumes of general text to improve the VQA performance.

In Figure 6, we show some cases of our method on the FVQA data. Compared to [29] which fails to answer questions in the first two columns because of the wrong supporting fact retrieved, our method leave the exact supporting fact extraction by the context-question attention block in QANet, which is more reliable than the keyword matching approach used in [29]. Method in [23] fails on the third question because of the wrong supporting facts retrieved either. Our method predicts a wrong answer for the last question even if the text representation includes the answer. This may be caused by the similar expressions of “*sth.* belongs to the category of Food” in the paragraph, which confuses the model.

4.2.2 Results Anlysis on Visual Genome QA

We use the top-1 answer accuracy to measure the performance on Visual Genome QA dataset, following [30] for fair comparison. All answers are normalized as well. Answer accuracy for each question type is also reported.

Table 2 lists the evaluation results on Visual Genome QA test split. It can be observed that our method achieves the best performance with the use of ground-truth region descriptions. The overall accuracy is about 5% higher than the result based on ground-truth facts used in [30]. When the predicted region descriptions are applied, our method still has higher accuracies on “5W” questions except “What”, which demonstrates the effectiveness of our method. The superiority is even obvious for “Who” questions, which is almost 10% higher. Nevertheless, since “What” questions account for 60.5% of all questions, its performance has a large effect on the overall accuracy. Answering “What” questions largely depends on the image description, as they mainly concern the states of objects. Using the dense caption model in [35] results in 1% higher overall accuracy than using the model in [11], because of the better dense caption results. As stated in [11], using the ground-truth

region boxes produces the mAP (mean Average Precision) of 27.03, while using the model in [11] only has mAP of 5.39 and the model in [35] obtains mAP of 9.31. The great gap between the predicted and the ground-truth region descriptions causes the VQA performance degradation. However, based on our method, the VQA problem is solved by two subtasks: image description and TQA, which avoids the multimodal feature fusion problem. We believe that as better image description methods become available, the results will improve further. Here we leave the improvement of generating more detailed and correct region descriptions as a future work.

We show some qualitative results produced by our open-ended VQA model tested on Visual Genome QA dataset in Figures 7 8 9.

In Figure 7, all the questions are proposed based on the image shown on the top left. The corresponding text descriptions are presented in the red and blue rectangular boxes on the right, where the red one shows the human-labeled description and the blue one shows the predicted dense captions. Predicted answers based on both descriptions are presented in the table. The results show that 1) our open-ended VQA model can tackle different types of questions; 2) the VQA model works better if the text description is more detailed. Even if the predicted answer is not exactly the same as the ground-truth answer, it is more reasonable based on better description. For example, when asking “What are the man’s hands doing?”, the predicted answer according to human-labeled region description shows “rope”, which is more relevant to the ground-truth answer “holding rein”.

In Figure 8, we present more examples from different input images and questions. According to the weights calculated by the context-question attention block, the sentences containing the higher weighted words in the converted text description are also presented. The results demonstrate that the question can be well answered if there is corresponding description about the question. For questions such as “Why” and “When” which need reasoning, the answer can be learned from the training data.

Figure 9 shows some interesting failure cases, where the predicted answers are very closer to the ground-truth answers. The predicted answer may have the same meaning as the ground-truth or in a general term. But they are not exactly the same and are regarded as incorrect during evaluation. These results expose a drawback of the open-ended VQA model in which multi-class classification is adopted in the output block. It is difficult to deal with synonyms by heuristically normalizing the answers. In addition, as they are divided into different classes, the model will learn to distinguish them from the training data, which is not reasonable.

Table 2. Experimental Results on Visual Genome QA based on the open-ended VQA model. The top-1 accuracies for different question types are also reported. Our method achieves higher accuracies on “5W” question types except “What”. The percentage of each question type is shown in parentheses. “GtDescp” means using the human-labeled region descriptions which is refer to the “GtFact” used in [30]. “PredDescp” means applying the predicted dense caption results in our VQA model.

Method	Accuracy (%)						Overall
	What (60.5%)	Where (17.0%)	When (3.5%)	Who (5.5%)	Why (2.7%)	How (10.9%)	
VGG+LSTM [1]	35.12	16.33	52.71	30.03	11.55	42.69	32.46
HieCoAtt-VGG [20]	39.72	17.53	52.53	33.80	12.62	45.14	35.94
VQA-Machine [30]							
GtFact(Obj+Att+Rel)+VGG	44.28	18.87	52.06	38.87	12.93	46.08	39.30
VQA-Machine [30]							
PredFact(Obj+Att+Rel)+VGG	40.34	17.80	52.12	34.98	12.78	45.37	36.44
Ours-GtDescp	49.6	23.8	56.9	57.2	16.7	59.3	44.8
Ours-PredDescp-by-[11]	36.4	17.9	56.5	48.6	14.7	45.1	33.7
Ours-PredDescp-by-[35]	37.4	18.6	56.6	49.0	14.8	45.8	34.5

Image				
Question	Which object in this image is utilized to chill food?	What animal in the image can rest standing up?	Which object in this image is round?	What sort of food can you see in this image?
Corresponding paragraph	A kitchen with a white refrigerator and a white stove. Brown wooden cabinets. White refrigerator with freezer. Refrigerator belongs to the category of Food. A refrigerator is used for chilling food. Oven belongs to the category of Food preparation appliances. Oven belongs to the category of Food and drink preparation. An oven is a device to heat Food. Stove belongs to the category of Food preparation appliances.	A horse drawn carriage on a city street. A man riding a horse drawn carriage down a street. Horses can rest standing up. Person is related to animate being. Person is related to standing. People is related to animal. People can stand up for themselves. Tree is related to up.	A man playing tennis. Tennis racket in the mans hand. A man holding a tennis racquet on a tennis court. A man swinging a tennis racket at a ball. A tennis ball is round. A tennis ball is often yellow. Tennis balls are spherical in shape. Tennis balls are hollow.	A bunch of fruits and vegetables on a table. A bunch of yellow bananas. Red apples in a bowl. Fruits belongs to the category of Food. Apple belongs to the category of Food. Pear belongs to the category of Foods. Banana belongs to the category of Food. Orange belongs to the category of Food.
Answer (pred)	Refrigerator	Horses	Tennis ball	Banana
Answer (gt)	Refrigerator	Horses	Tennis ball	Fruits

Figure 6. Successful and failure cases of our method on FVQA dataset. Our method correctly predicts answers for the questions in the first three columns, but fails for the last one. In addition, the reason for the answer is obvious from the converted paragraph, which is more semantic and structured than image.

Table 3. Answer accuracies on Visual7W [38] Telling dataset using the multiple-choice VQA model. “GtDescp” means using the human-labeled region descriptions, while “PredDescp” means applying the predicted dense caption results.

Method	Accuracy (%)						Overall
	What (47.8%)	Where (16.5%)	When (4.5%)	Who (10.0%)	Why (6.3%)	How (14.9%)	
LSTM+CNN [1]	48.9	54.4	71.3	58.1	51.3	50.3	52.1
Visual7W [38]	51.5	57.0	75.0	59.5	55.5	49.8	55.6
MCB [8]	60.3	70.4	79.5	69.2	58.2	51.1	62.2
MLP [10]	64.5	75.9	82.1	72.9	68.0	56.4	67.1
MAN [22]	59.0	63.2	75.7	60.3	56.2	52.0	59.4
KDMN-NoKG [18]	59.7	69.6	79.9	68.0	61.6	51.3	62.0
Ours-GtDescp	70.5	74.5	77.0	80.3	63.8	55.7	69.8
Ours-PredDescp-by[11]	58.4	64.9	75.1	70.2	56.3	50.8	60.2
Ours-PredDescp-by[35]	59.7	66.2	75.1	70.8	58.0	51.5	61.2

4.2.3 Results Analysis on Visual7W

We test the multiple-choice VQA model on Visual7W dataset. The results are presented in Table 3. Our method achieves the best performance when applying the ground-

truth region descriptions. It also performs well when we use the predicted dense captions from [35], compared with the results by recently proposed dynamic memory network based methods of [22] and [18] without extra information added. To be specific, our model shows better perfor-



Red flowers climbing the wall. man driving buggy and horse. a shrubbery with red flowers. A man is sitting on a cart being pulled by a horse. A beautiful flowering bush. a chestnut colored horse. a man in blue jeans and ball cap. flowering foliage. a concrete sidewalk path. part of a banana tree. the horse's reins. a green hedge with white blooms. man driving a pony cart. wheels on pony cart. man's hands holding the reins. lush flowering shrub. trimmed hedges in front of a house. air conditioner protruding from house. harness around a pony's neck. A man steering a horse-drawn carriage. A pink entry way. A patch of short grass. A pale blue and red striped shirt. Horse pulling a cart. Dark green leaves with red flowers. Cart with two wheels with rubber tires. Black horse with white facial markings. Green shrub in the middle of the sidewalk. Man with a white hat. Grey brick wall. A horse carrying a man in a carriage. A man riding a horse in a carriage. Green bushes beneath a house. A brown horse with silver hooves. A white house with some pink on it. A street next to pavement. A tree with red on it. worn out cart tire. stone exterior of building. exterior wall painted pink. patch of lawn and weeds. overhanging palm tree fronds. a carriage pulled by a horse. a green palm tree. a tall stone wall. an white apartment building. window with red bars. tan bricks on distant building. a horse's hoofs on pavement. white baseball hat on man's head. shrubs outside a residence. a hole in the man's pants. A sidewalk has a white curb. Man is holding a horse's reins. Two black wheels.

Question	Answer-pred (GtDisp)	Answer-pred (PredDisp)	Answer-gt
What pulls the cart?	horse	horse	horse
Who drives the cart?	man	engineer	man
Where is the man?	on cart	on street	on cart
What color is the horse?	brown	brown	brown
What color is the man's hat?	white	white	white
How many wheels are on the pony cart?	2	2	2
What are the man's hands doing?	rope	riding bike	holding rein
Why is the horse attached to the cart?	riding	taking off	pulling it
What kind of tires does the wagon have?	rubber	tire	rubber
What is the pony doing?	riding	parked	pulling cart

a man riding a horse. white hat on the man. a blue and white shirt. brown horse with black mane. two wheels on a cart. red flowers in a tree. a paved road. a black tire. a palm tree. the man is wearing blue jeans. green bushes in front of the building. a house with a red roof. a horse on the road. a bush in the background. a wooden cart. a chain link fence. a concrete sidewalk. a white door on the wall. a brick wall. a red brick building. a man wearing a blue jeans. a green tree in the background. shadow of the person on the ground. grass growing on the sidewalk. red and white building. the horse is wearing a harness. white line on the road. a black and white sign. the road is grey. a tree with green leaves. a house in the background. a white building with a window. a white light on the wall. a black and white bench. a brown horse with a white stripe.

Figure 7. Success and failure cases of our open-ended VQA method on Visual Genome QA dataset. The model is feasible for different types of questions. “GtDisp” means using the human-labeled region descriptions which are presented in the red box, while “PredDisp” means applying the dense caption results by model in [11] which are shown in the blue box.



Q: Whose head is poking out of the side?
A: Dog
C: Dog sticking head through railing.



Q: What type of bus is turning?
A: Double decker
C: A red double decker bus.



Q: Where was this photo taken?
A: Train station
C: Passenger train stopped at the station.



Q: When is the picture taken?
A: Daytime
C: flower box with green leaves and purple and red flowers.



Q: What color is the floor?
A: White
C: The white tile floor.



Q: How many boys are playing?
A: 2
C: Two boys holding tennis rackets.



Q: Who is standing front of the green wall?
A: Woman
C: The woman is standing.



Q: Why are there shadows?
A: Sunny
C: Elephant shadow on pavement.

Figure 8. Correctly answered examples from Visual Genome QA dataset. “Q”, “A”, “C” denote the question, the properly predicted answer, and the supporting sentence from the predicted image description by model in [11].

mance on “Who” questions and comparable accuracies on “What” and “How” questions. Because the region descriptions contain abundant semantic information about the image. They are helpful to answer questions such as “What color”, “What shape”, “What is the man doing”, “Who

is doing ...”. However, it performs poorly on “Why” and “When” questions even if we use the ground-truth region descriptions. We infer that is because the candidate answers for “Why” and “When” questions are generally longer than others, and are usually not included by the converted text



Q: Where was this photo taken?
A(pred): Street corner
A(gt): Asian city street
C: A Chinese business advertisement sign.



Q: What has lights on?
A(pred): Bicycle
A(gt): Bike
C: Lights and reflectors on bicycle.



Q: Who has their mouths open?
A(pred): Boy
A(gt): Two boys
C: A boy with his mouth open.



Q: How is the triangle logo situated?
A(pred): Upside down
A(gt): Base up
C: Large upside down triangle inside of a red circle.



Q: Why would you ride the bus?
A(pred): Transportation
A(gt): Commute
C: Large bus that says Crosstown on the front.



Q: When was this taken?
A(pred): Day time
A(gt): During day
C: Brown tee-shirt and denim skirt on little girl.

Figure 9. Some failure cases in which the predicted answers are very closer to the ground-truth answers. “Q”, “A(pred)”, “A(gt)”, “C” denote the question, the predicted answer, the ground-truth answer, and the supporting sentence from the predicted image description by model in [11].

Questions	What time of day is it?	When will the children leave the field?	Why are the children running?	How many children are there?	Who is standing in this photo?	Where is this photo taken?
Multi-choices provided and probability predicted	Night time (0.04) Afternoon (0.15) Morning (0.04) Daytime (0.96)	When the game is over (0.45) When they are done playing (0.23) When it is time to eat (0.10) When their parents get ready to take them home (0.08)	They are playing tag (0.26) They are exercising (0.09) They are having fun together (0.66) They are trying to kick the balls (0.18)	Three (0.13) Four (0.04) None (0.18) Two (0.63)	A woman (0.28) A couple (0.01) An old man (0.01) A girl and boy (0.97)	At a park (0.60) In a swimming pool (0.01) At the museum (0.02) On a grassy field (0.72)
Gt_answer	Daytime	When their parents get ready to take them home	They are trying to kick the balls	Two	A girl and boy	On a grassy field



Small child in grass. Small child wearing yellow shirt. Green patch of grass. girl's capris are pink. girl's shirt is yellow. lady bug on girl's shirt. black spots on lady bug. girl's hair is blonde. boy is kicking soccer ball. boy's shorts are red. boy's shirt is red and white. soccer ball is white orange and black. blonde girl soccer with ball. apple on the ground with green. hand with five fingers on it. red shirt with white on the clock. green ball with soccer basket. sun with bank and money coin. lady bug shirt with yellow. boy hand weapon gun knife black. White and black ball. Small patch of green grass. Yellow shirt with red and black design. Small child in the grass. a ball on the grass. a shadow on the grass . pink and blue pants . a white ball. girl wearing shoes. a yellow shirt . a soccer ball .

Figure 10. Qualitative results of our multiple-choice VQA model on Visual7W dataset. Given the image, the predicted dense caption result by [11] is presented in the blue box. We report the probability to each candidate answer choice in brackets. The predicted answer is the one with the largest probability for each question, which is shown in red color. The VQA model will attend the most related words by the context-question and context-answer attentions (as shown in the red words in the text paragraph), which helps the answer inferring.

description. In that case, it becomes difficult for the model to co-attention between question/answer and context. The encoded features of \mathbf{v}_a and \mathbf{v}_q are not strong correlated.

In addition, it should be note that the work in [10] re-

ports the accuracies of 64.5% and 54.9% for “Why” and “How” questions even based the inputs of question and answer, without image, which means their model can infer the correct answer without using image information. It seems

the model is overfit on this dataset. It merely learns the biases from the dataset, which is not accepted from the point of solving VQA problem.

We present some qualitative results produced by our multiple-choice VQA method on different kinds of questions in Figure 10, based on the same input image. The results illustrate that the VQA model performs well if the related information is contained by the text description. Even if the answer is not exactly expressed in the paragraph, the model can infer it according to some related words. The correctly inferred answers to the “How many” and the “Who” questions in Figure 10 prove this point. The “When” and “Why” questions are wrongly answered in this example, because they are totally not mentioned in the text description.

Furthermore, after converting to text which is full of semantic information, the reasoning process is readable from the context-question attention. Other examples show that when the question asks about “color”, all words about color in the context will be higher weighted by the context-question attention. The corrected answer can then be inferred by considering the focused object additionally.

A few more examples are shown in Table 4 which achieves correct answers and in Table 5 which shows failure cases. Our method achieves better results if the answer is included in the converted text description. In Table 4, the predicted results for “What” question shows higher probabilities for both “tree” and “train”, which is understandable, and “train” has higher probability than “tree”. For “Why” and “When” questions, the corrected answers may be learned from the training data. From Table 5, we can see that the failure reasons are mainly caused by the undescribed information in dense captions. The candidate answers usually cannot get higher probabilities, no matter correct or incorrect ones. Furthermore, some improvement directions are observed. For example, for the first image in Table 5, the description includes “the glasses of water”, but does not mention “Food” or “Drink”. Hence external knowledge would be helpful here which explains that “Water belongs to the category of drink”. The second question is a kind of text recognition problem. Therefore, an additional text detection and recognition module is useful, which can extract all the text in the image. Actually, text appeared in the image usually contains lots of semantic information. It can help the image understanding. Last but not least, it is found that summary or analysis about the image would be very useful in answering questions such as “the total number of objects/person show in the image”.

5. Conclusion

In this work, we attempt to solve VQA from the viewpoint of machine reading comprehension. In contrast to explore the obscure information from image feature vector,

we propose to explicitly represent image contents by natural language and convert VQA to textual question answering. With this transformation, we avoid the cumbersome issue on multimodal feature fusion. The reasoning process are readable from the context. The framework can be easily extended to handle knowledge based VQA. Moreover, we can exploit the large volume of text and NLP techniques to improve VQA performance.

Our experiments also show that if the context is too long, it becomes hard to infer the correct answer. Hence, how to generate correct and valid image description, and how to extract proper external knowledge are next work.

References

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015. 2, 8
- [2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, 2007. 6
- [3] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *Proc. Conf. Empirical Methods in Natural Language Processing*, 2013. 3
- [4] A. Bordes, S. Chopra, and J. Weston. Question answering with subgraph embeddings. In *Proc. Conf. Empirical Methods in Natural Language Processing*, pages 615–620, 2014. 3
- [5] A. Bordes, N. Usunier, S. Chopra, and J. Weston. Large-scale simple question answering with memory networks. In *arXiv: abs/1506.02075*, 2015. 3
- [6] D. Chen, A. Fisch, J. Weston, and A. Bordes. Reading wikipedia to answer open-domain questions. In *Proc. Conf. the Assoc. Comput. Linguistics*, pages 1870–1879, 2017. 3
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 3
- [8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proc. Conf. Empirical Methods in Natural Language Processing*, 2016. 2, 3, 5, 8
- [9] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 2296–2304, 2015. 2
- [10] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *Proc. Eur. Conf. Comp. Vis.*, 2016. 5, 8, 10
- [11] J. Johnson, A. Karpathy, and L. Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016. 2, 4, 6, 7, 8, 9, 10
- [12] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. In *Proc. Adv. Neural Inf. Process. Syst.*, 2016. 3

Table 4. Correctly answered examples from Visual7W dataset by our multiple-choice VQA model. We report the probability to each candidate answer choice in brackets. The one with the largest probability for each question is regarded as correct.

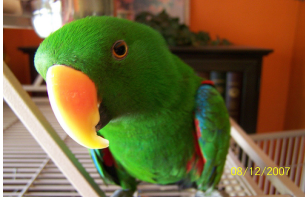








Image			
Question	Who is present?	What is the subject of the photo?	Where was this photographed?
Multi-Choices	Everyone (0.01) All the coworkers (0.02) The boss (0.02) Nobody (0.98)	Clouds (0.15) Trees (0.93) The wilderness (0.05) A train (0.99)	In the country (0.11) At the beach (0.01) In the forest (0.05) City street (0.96)
Answer (pred)	Nobody	A train	City street
Answer (gt)	Nobody	A train	City street
Image			
Question	Why is there a fork?	When was this photo taken?	How many giraffes are in the picture?
Multi-Choices	Because the road divides. (0.03) Because we ran out of spoons. (0.02) Because it is a complete set of silverware. (0.01) To eat the food. (0.77)	At day break (0.31) At noon (0.02) In the afternoon (0.03) During the daytime (0.99)	4 (0.08) 3 (0.11) 6 (0.09) 2 (0.97)
Answer (pred)	To eat the food	During the daytime	2
Answer (gt)	To eat the food	During the daytime	2

Table 5. Failure cases of our multiple-choice VQA model on Visual7W dataset. The failure reasons are mainly caused by the excluded information in the image description.

Image			
Question	What is on the table?	What is written on the plane?	How many signs are pictured?
Multi-Choices	Cups (0.07) Plate (0.08) Food (0.35) Drinks (0.05)	Blastor (0.02) Orbison (0.02) Usa (0.04) Orbast (0.03)	3 (0.15) 0 (0.09) 1 (0.82) 4 (0.05)
Answer (pred)	Food	Usa	1
Answer (gt)	Drinks	Orbast	4

[13] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. In *Proc. Int. Conf. Learn. Representations*, 2017. 3

[14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. Learn. Representations*, 2014.

6

[15] J. Krause, J. Johnson, and L. F.-F. Ranjay Krishna. A hierarchical approach for generating descriptive image paragraphs. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. 2

[16] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz,

- S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comp. Vis.*, 123(1):32–73, 2017. 4, 5, 6
- [17] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. In *Proc. Int. Conf. Mach. Learn.*, pages 1378–1387, 2015. 3
- [18] G. Li, H. Su, and W. Zhu. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018. 2, 3, 8
- [19] H. Liu and P. Singh. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004. 6
- [20] J. Lu, J. Yang, D. Batra, and D. Parikh. hierarchical question-image co-attention for visual question answering. 8
- [21] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 289–297, 2016. 2
- [22] C. Ma, C. Shen, A. Dick, Q. Wu, P. Wang, A. van den Hengel, and I. Reid. Visual question answering with memory augmented networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018. 8
- [23] M. Narasimhan and A. G. Schwing. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Proc. Eur. Conf. Comp. Vis.*, 2018. 3, 6, 7
- [24] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proc. Conf. Empirical Methods in Natural Language Processing*, 2014. 3
- [25] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100, 000+ questions for machine comprehension of text. In *Proc. Conf. Empirical Methods in Natural Language Processing*, 2016. 6, 7
- [26] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016. 2
- [27] N. Tandon, G. de Melo, and G. Weikum. Acquiring comparative commonsense knowledge from the web. In *Proc. National Conf. Artificial Intell.*, 2014. 6
- [28] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel. Explicit knowledge-based reasoning for visual question answering. In *Proc. Int. Joint Conf. Artificial Intell.*, 2017. 3
- [29] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel. Fvqa: Fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–1, 2017. 3, 4, 6, 7
- [30] P. Wang, Q. Wu, C. Shen, and A. van den Hengel. The vqa-machine: learning how to use existing vision algorithms to answer new questions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. 2, 6, 7, 8
- [31] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou. Gated self-matching networks for reading comprehension and question answering. In *Proc. Conf. the Assoc. Comput. Linguistics*, pages 189–198, 2017. 3
- [32] Q. Wu, P. Wang, C. Shen, A. van den Hengel, and A. R. Dick. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016. 3
- [33] C. Xiao, M. Dymetman, and C. Gardent. Sequence-based structured prediction for semantic parsing. In *Proc. Conf. the Assoc. Comput. Linguistics*, 2016. 3
- [34] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *Proc. Int. Conf. Mach. Learn.*, pages 2397–2406, 2016. 2
- [35] L. Yang, K. Tang, J. Yang, and L.-J. Li. Dense captioning with joint inference and visual context. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. 6, 7, 8
- [36] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. In *Proc. Int. Conf. Learn. Representations*, 2018. 3, 6
- [37] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. 2017. 3
- [38] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016. 6, 8
- [39] Y. Zhu, J. J. Lim, and L. Fei-Fei. Knowledge acquisition for visual question answering via iterative querying. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. 3