

Visual Question Answering as Reading Comprehension

Hui Li^{*1}, Peng Wang^{*2}, Chunhua Shen¹, 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. Nevertheless, how to model the complex interactions between the two different modalities is not an easy work. 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.

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,

^{*}The first two authors equally contributed to this work. C. Shen is the corresponding author.

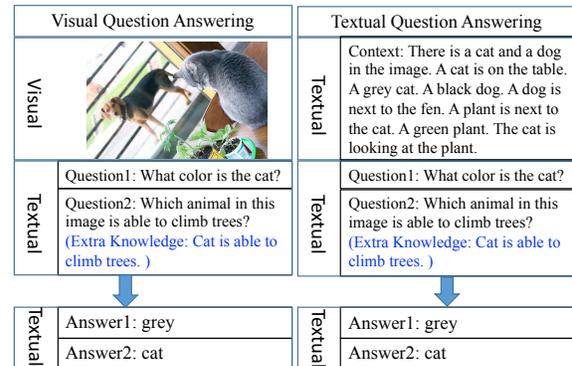


Figure 1 – Comparison between VQA and TQA. Question1 is observation based, which can be inferred from the image itself. Question2 is knowledge based, which has to refer knowledge beyond the image. Extra knowledge commonly appears in text, which is easier to be combined to the context paragraph in TQA.

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.* [17] proposed to embed knowledge in memory slots and incorporated external knowledge with image, question and answer features

by Dynamic Memory Networks (DMN).

In this work, different from exploring the high-dimensional and noisy image features 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 [26]. Through this transformation, all inputs are transferred into text, which avoids the joint embedding of image and text features into a hidden space. Instead, the multimodal fusion is conducted in text domain, which is better for explicitly preserving semantic information that is the core concern of VQA. In addition, external knowledge that is commonly described by text can be easily integrated into the model. With the attention mechanism used in text domain, the proposed model is able to provide semantic-level (*i.e.*, text) supporting facts, and hence make the reasoning process more interpretable.

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 reading comprehension. Thus we can resort to the abundant research results in NLP community to handle VQA problem. Using text and NLP techniques allows convenient access to higher-level information, and enables transfer learning from TQA to VQA models. 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. Based on the converted text description and the attention mechanism used in the models, semantic level supporting facts can be retrieved from the context, which makes the answer inferring process 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 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.

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 [20, 24, 29], element-wise product or sum [1, 9, 34] to jointly embed the visual feature and textual feature. Fukui *et al.* [8] argue that these simple manners are not expressive enough and propose 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] are 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] present 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, which by itself is a challenging problem. In this work, we describe the visual information directly by text which unifies the input information in advance in text domain.

2.2. Knowledge-based VQA

There are some researches in 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.

Wu *et al.* [32] propose to combine image representation with extra information extracted from a general knowledge base according to predicted image attributes for VQA. The method makes it possible to answer questions beyond the image, but the extracted knowledge is discrete pieces of text, without structural representations. Ahab [27] uses explicit reasoning over a 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.* [28] introduce the “Fact-based VQA (FVQA)” problem and propose 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 is then developed in [22] 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 [17] exploit 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-

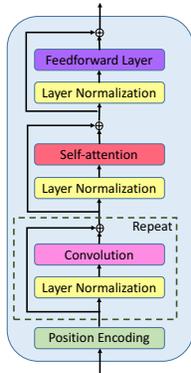


Figure 2 – The structure of encoder block used in QANet, which is shared by embedding encoder and model encoder. The number of convolutional layers varies according to design. Layer normalization and residual connection are adopted between every layer for better performance.

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 [16], r-net [30], 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. A two-layer highway network is applied 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.

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 $\mathbf{S} \in \mathcal{R}^{n \times m}$ 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 $[\mathbf{c}, \mathbf{a}, \mathbf{c} \odot \mathbf{a}, \mathbf{c} \odot \mathbf{b}]$ as input, where \mathbf{a} and \mathbf{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

Questions and answers usually appear in the form of text, which are semantic-level information. It is widely accepted that one of the primary attentions of VQA is to evaluate the semantic-level visual understanding ability of AI systems. Considering that a diverse range of semantic visual information can be described in natural language, in this work, we attempt to convert the image wholly into a descriptive paragraph, so as to preserve as much semantic information as possible for semantic questions. Since all inputs are unified in text domain, our method avoids the challenge task of multimodal feature fusion in hidden space, and can extend to deal with the knowledge-based VQA straightforwardly. The architecture of our proposed model is presented in Figure 3. Besides the basic modules used in QANet, we add another input pre-processing block and modify the output block for the open-ended VQA.

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 para-

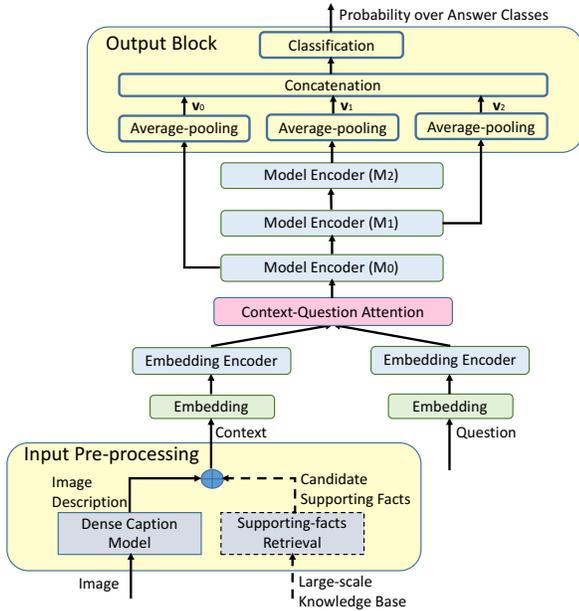


Figure 3 – Open-ended VQA model. By representing image with neural language, we convert VQA as reading comprehension. Extra knowledge can be added naturally into the model because of the same modality.

graph. Dense captions [11] provide a finer level of semantic representation for image content, ranging from the states of a single object (color, shape, action, *etc.*) to the relationships between objects (spatial positions, *etc.*), so we infer that they include most of the supporting visual information required by VQA. It should be note that there is existing work of using semantic labels or concepts for VQA. In [31], the authors use a vocabulary of 256 attributes as image representation, and achieve a significant improvement over CNN image features in VQA. Our work is another manner of using semantic information. Furthermore, the dense caption results are even richer than a few discrete attribute labels, which makes them work well. The generated region captions are combined together as the image description for QANet. Because of the use of self-attention, the model works better for encoding long-term dependency than RNNs that are commonly adopted in VQA.

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.* [28] 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 applying 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 and predict 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 answers to questions like “When” and “Why”. To address this case, 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. It is hoped that the model can learn some clues from region descriptions so as to infer the answer. 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 answer 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 performance, we propose another model for multiple-choice VQA problem.

As presented in Figure 4, 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.

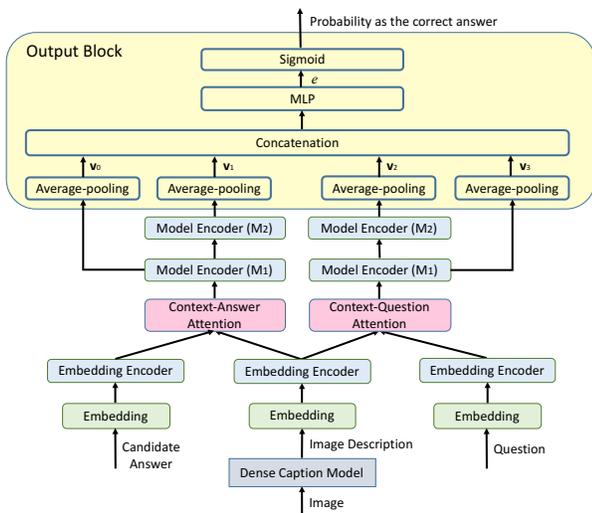


Figure 4 – Multiple-choice VQA model. It takes image-question-answer triplet as input and encodes both interactions of question and answer with the context.

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 using the heuristic answer normalization. Moreover, 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 particularity.

FVQA [28] (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 [18] and WebChild [25]. 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 [15] 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 which give a finer level of semantic information 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 [29], 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 [28] 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 [23]. 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 descriptions. Based on these annotations, 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 both the average paragraph length and training speed, 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.

Method	Overall Accuracy (%)	
	top-1	top-3
LSTM-Question +Image+Pre-VQA [28]	24.98	40.40
Hie-Question +Image+Pre-VQA [28]	43.14	59.44
FVQA (top-3-QQmapping) [28]	56.91	64.65
FVQA (Ensemble) [28]	58.76	-
Question+Visual Concepts [22]	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 – Experimental Results on FVQA. Our method with finetuned QANet achieves the highest top-1 accuracy.

4.2.1 Results Analysis on FVQA

We use answer accuracy to evaluate the model, following [28]. 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 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 shows the overall accuracy of our method based on supporting facts retrieved by using the top-3-QQmapping results in [28]. 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 [22] 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 [28] about 6% on top-1 and top-3 answer accuracies respectively, and even performs better than the ensemble method in [28]. As this work aims to propose an alternative approach to 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 finetuning on FVQA training data, *i.e.*, the one trained only on general reading comprehension dataset SQuAD [23]. Experimental results show that the pre-trained QANet model is also feasible on FVQA dataset. The model gives even better results than the one trained from scratch solely on 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 5, we show some cases of our method on FVQA data. Our method leaves the exact extraction of supporting fact to the context-question attention block in QANet, which is more reliable in comparison to the work

of [28, 22]. Our method gives 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 confuse the model.

4.2.2 Results Analysis on Visual Genome QA (VGQA)

We use the top-1 answer accuracy to measure the performance on VGQA dataset, following [29] for fair comparison. All answers are normalized.

As presented in Table 3, our method achieves the best performance when using the ground-truth region descriptions. The overall accuracy is about 5% higher than that based on ground-truth facts used in [29]. 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 larger 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 the model in [11] only generates 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. 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.

An ablation experiment is also conducted to test the performance with different maximum paragraph length, using the ground-truth dense caption results from [15]. As shown in Table 2, the overall accuracy increases slightly with a longer paragraph, but the computation speed decreases a lot. Trading off between the performance and computational speed, we use the paragraph limits of 500 words.

Maximum length (# Words)	Overall Accuracy (%)	Inference Time (# QA/sec)	Training Time (# QA/sec)
250	43.92	159	49
300	44.35	146	45
500	44.83	90	29
750	44.89	49	17
1000	45.01	37	12

Table 2 – The influence of maximum paragraph length on performance. Generally, a longer text description leads to a relatively higher accuracy. But the computation burden increases largely.

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 5 – Successful and failure cases of our method on FVQA dataset. Our method correctly predicts answers for questions in the first three columns, but fails for the last one. In addition, the reason to the answer is readable from the semantic-level text descriptions, compared to images.

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
HicCoAtt-VGG [19]	39.72	17.53	52.53	33.80	12.62	45.14	35.94
VQA-Machine [29]							
GtFact(Obj+Att+Rel)+VGG	44.28	18.87	52.06	38.87	12.93	46.08	39.30
VQA-Machine [29]							
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

Table 3 – Experimental Results on VGQA 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 [29]. “PredDescp” means applying the predicted dense caption results in our VQA model.

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 [21]	59.0	63.2	75.7	60.3	56.2	52.0	59.4
KDMN-NoKG [17]	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

Table 4 – 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.

4.2.3 Results Analysis on Visual7W

We evaluate the multiple-choice VQA model on Visual7W dataset. The results are presented in Table 4. 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 [21] and [17] without extra information added. To be specific, our model shows better performance on “Who” questions and comparable accuracies on “What” and “How” questions. Because the region descrip-



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 .

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)	When the game is over (0.45)	They are playing tag (0.26)	Three (0.13)	A woman (0.28)	At a park (0.60)
	Afternoon (0.15)	When they are done playing (0.23)	They are exercising (0.09)	Four (0.04)	A couple (0.01)	In a swimming pool (0.01)
Probability predicted	Morning (0.04)	When it is time to eat (0.10)	They are having fun together (0.66)	None (0.18)	An old man (0.01)	At the museum (0.02)
	Daytime (0.96)	When their parents get ready to take them home (0.08)	They are trying to kick the balls (0.18)	Two (0.63)	A girl and boy (0.97)	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

Figure 6 – 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.

tions 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 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] reports the accuracies of 64.5% and 54.9% for “Why” and “How” questions even based solely on 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 overfits 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 6. 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, and shows a kind of multi-hop reasoning ability. As proved by the “How many” and “Who” QA pairs in Figure 6, the predicted answers are based on two separate sentences from the paragraph. 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. 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.

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 joint-embedding of multimodal features in a hidden space. Via conducting multimodal fusion in text domain, semantic information is preserved which is more valuable for VQA. With the adopted attention mechanism in text domain, the reasoning process is more interpretable. The framework can be easily extended to combine external knowledge as it appears in text generally. 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.

Acknowledgements

This work was in part supported by the ARC Centre of Excellence for Robotic Vision. P. Wang’s participation was in part supported by the National Natural Science Foundation of China (#61876152).

References

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015.
- [2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, 2007.
- [3] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proc. Conf. Empirical Methods in Natural Language Processing*, 2013.
- [4] Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embeddings. In *Proc. Conf. Empirical Methods in Natural Language Processing*, pages 615–620, 2014.
- [5] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. Large-scale simple question answering with memory networks. 2015.
- [6] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In *Proc. Conf. the Assoc. Comput. Linguistics*, pages 1870–1879, 2017.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
- [8] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proc. Conf. Empirical Methods in Natural Language Processing*, 2016.
- [9] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei 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.
- [10] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In *Proc. Eur. Conf. Comp. Vis.*, 2016.
- [11] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.
- [12] Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. In *Proc. Adv. Neural Inf. Process. Syst.*, 2016.
- [13] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *Proc. Int. Conf. Learn. Representations*.
- [14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. Learn. Representations*, 2014.
- [15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comp. Vis.*, 123(1):32–73, 2017.
- [16] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *Proc. Int. Conf. Mach. Learn.*, pages 1378–1387, 2015.
- [17] Guohao Li, Hang Su, and Wenwu Zhu. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.
- [18] H. Liu and P. Singh. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004.
- [19] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. hierarchical question-image co-attention for visual question answering.
- [20] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 289–297, 2016.
- [21] Chao Ma, Chunhua Shen, Anthony Dick, Qi Wu, Peng Wang, Anton van den Hengel, and Ian Reid. Visual question answering with memory augmented networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.
- [22] Medhini Narasimhan and Alexander G. Schwing. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Proc. Eur. Conf. Comp. Vis.*, 2018.
- [23] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In *Proc. Conf. Empirical Methods in Natural Language Processing*, 2016.
- [24] Kevin J. Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.
- [25] Niket Tandon, Gerard de Melo, and Gerhard Weikum. Acquiring comparative commonsense knowledge from the web. In *Proc. National Conf. Artificial Intell.*, 2014.
- [26] Damien Teney, Qi Wu, and Anton van den Hengel. Visual question answering: A tutorial. *IEEE Signal Process. Magaz.*, 34(6):63 – 75, 2017.
- [27] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Explicit knowledge-based reasoning for visual question answering. In *Proc. Int. Joint Conf. Artificial Intell.*, 2017.
- [28] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. FVQA: Fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–1, 2017.
- [29] Peng Wang, Qi Wu, Chunhua Shen, and Anton 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.

- [30] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering. In *Proc. Conf. the Assoc. Comput. Linguistics*, pages 189–198, 2017.
- [31] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.
- [32] Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Anthony 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.
- [33] Chunyang Xiao, Marc Dymetman, and Claire Gardent. Sequence-based structured prediction for semantic parsing. In *Proc. Conf. the Assoc. Comput. Linguistics*, 2016.
- [34] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *Proc. Int. Conf. Mach. Learn.*, pages 2397–2406, 2016.
- [35] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.
- [36] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. QANet: Combining local convolution with global self-attention for reading comprehension. In *Proc. Int. Conf. Learn. Representations*, 2018.
- [37] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. 2017.
- [38] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.
- [39] Yuke Zhu, Joseph J. Lim, and Li Fei-Fei. Knowledge acquisition for visual question answering via iterative querying. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.