The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Re-Attention for Visual Question Answering

Wenya Guo,[1] **Ying Zhang,**[1*] **Xiaoping Wu,**[1] **Jufeng Yang,**[1]
Xiangrui Cai,[2] **Xiaojie Yuan**[1,2]
[1]College of Computer Science, Nankai University, Tianjin, China
[2]College of Cyber Science, Nankai Univeristy, Tianjin, China
guowenya@dbis.nankai.edu.cn, xpwu95@163.com,
{yangjufeng, caixr, yuanxj, yingzhang}@nankai.edu.cn

Abstract

Visual Question Answering (VQA) requires a simultaneous understanding of images and questions. Existing methods achieve well performance by focusing on both key objects in images and key words in questions. However, the answer also contains rich information which can help to better describe the image and generate more accurate attention maps. In this paper, to utilize the information in answer, we propose a re-attention framework for the VQA task. We first associate image and question by calculating the similarity of each object-word pairs in the feature space. Then, based on the answer, the learned model re-attends the corresponding visual objects in images and reconstructs the initial attention map to produce consistent results. Benefiting from the re-attention procedure, the question can be better understood, and the satisfactory answer is generated. Extensive experiments on the benchmark dataset demonstrate the proposed method performs favorably against the state-of-the-art approaches.

Introduction

Visual Question Answering (VQA) is one of the fundamental tasks that involve multiple modalities, *i.e.*, text and images. It can be formulated as a classification problem, which predicts the correct answer for the given question according to an image. Besides answering the given question, VQA also benefits to various applications in practice, such as education and blind person assistance (Gurari et al. 2018).

In the past years, extensive works are proposed to tackle the VQA problem (Anderson et al. 2018; Yu et al. 2019). They attempt to understand the image and question in the fine-grained scenario. Some of the existing methods are designed to obtain critical visual information relevant to the question, in which the visual attention mechanism is widely applied. In these models, the performance is improved by learning meaningful regions or objects according to a unified question representation (Anderson et al. 2018). Some other methods (Gao et al. 2019; Lu et al. 2016) also propose that it is significant to focus on key words in the question. Both the informative visual contents in images and important words are utilized to achieve better performance.

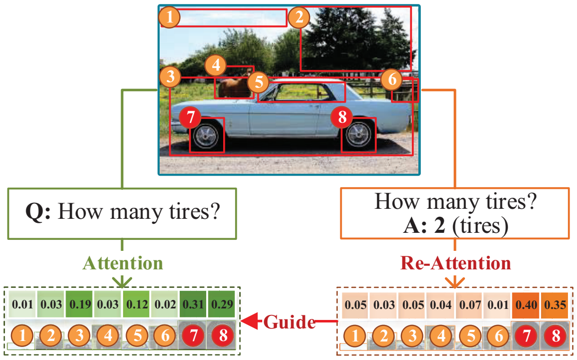

Figure 1: Example of the re-attention VQA scheme. Normally, for the question "Q", the VQA model generates initial attention for visual image and may focus on several relevant components of the car. The proposed re-attention mechanism obtains more accurate attention when taking additional cues from answer "A" into consideration.

Actually, VQA includes three elements: images, questions, and answers (Antol et al. 2015). The goal of VQA is to predict the correct answer to the question about the image. Existing methods treat the answer as a classification label. Yet the answer contains richer information. Both the questions and answers contribute to describing visual contents. Take the question ("How many tires?") in Figure 1 as an example. In order to predict the correct answer "2", VQA methods need to understand the question and figure out the "tires" in the image. As shown in in Figure 1, when only the question is considered, the "car" numbered as "3" and the "tires" are assigned higher weights compared to other objects. Obviously, the weight assigned to the "car" is out of place. After the answer is taken into account, *i.e.*, "How many tires? 2 (tires)", the learned visual attention is more centralized on the tires numbered as "7" and "8". This accurate attention map can guide the visual attention learning as expected. And accurate visual attentions help VQA models predict correct answers.

In this paper, we propose a novel re-attention framework, in which the answer information is used for VQA. In the re-

[*]corresponding author.

attention pattern, answers are exploited to help visual attention learning. Specifically, the answers are used to compute attention weights for the image. Then, an attention consistency loss is defined to measure the distance between the visual attention maps learned with questions and answers. The accurate attention map learned with answers can guide the visual attention learning with questions by minimizing the consistency loss.

Besides, as already mentioned, in most of the existing attention-based methods, questions are used to learn meaningful visual contents in images (Yu et al. 2019). However, the words related to the visual content are also important. It is necessary to consider visual contents when learning the importance of words as well. We propose a new attention module to model the relationship between the image and question in a more fine-grained manner. Specifically, the attention maps for the image and question are derived from a matrix that contains the similarity scores of every object-word pair. Then attended features of image and question are fused into a unified representation. The fused representation is delivered into two branches. In the forward inference branch, it is classified into the correct answer. In the re-attention branch, the representation is used to compute the re-attention maps for the visual contents in images.

Our contributions are summarized as follows: First, we address the VQA problem via a novel re-attention pattern, which sufficiently represents the answer and enables the accurate attention of key question-related contents in visual image. Second, we propose a new attention module which correlates the relationships of each object-word pairs in a fine-grained perspective and generates attention maps for image and question with the guidance of each other. The proposed method is evaluated on the commonly used large dataset VQA v2 (Goyal et al. 2017). The method performs favorably against the state-of-the-art methods for visual question answering.

Related Work

In this section, we review fusion-based approaches and attention-based models for visual question answering that are relevant to our work.

Visual Question Answering (VQA)

VQA is an emerging research area to reason the accurate answer of a given question about the visual content in an image. Fusion-based methods are the most straightforward strategy for this task. The image and question are represented as global features and then fused into a unified representation to predict the correct answer.

For the images, most image features are extracted from a pre-trained CNN (Antol et al. 2015). For the questions, feature extraction methods have developed from the simple bag-of-words model (Zhou et al. 2015) to more complex and useful language models, *e.g.*, LSTM (Antol et al. 2015; Ma et al. 2018) and GRU (Zhang, Hare, and Prügel-Bennett 2018). Fusion of question and image features is usually implemented by existing effective multimodal feature fusion methods. Many multimodal fusion methods are used to learn accurate answer representations, such as residual networks (Kim et al. 2016), multimodal compact bilinear pooling (Fukui et al. 2016), and factorized high-order pooling (Yu et al. 2018). There are some other methods fusing the multimodal features by reasoning the complex interaction between question and image. (Wu et al. 2018) proposes a dynamic multi-step and dynamic model to reason the changed relations between objects. (Haurilet, Roitberg, and Stiefelhagen 2019) propose a model that aimes to capture the interplay among objects guided by the query. (Cadene et al. 2019) utilizes the MuRel cell to progressively refine the interaction between images and questions. More recently, some other methods focus on utilizing the graph to explore the process of understanding images with questions and predicting answers. (Norcliffe-Brown, Vafeais, and Parisot 2018) uses a graph-based methods to learn semantic and spatial representations of image that capture question specific interactions. (Tang et al. 2019) proposes to construct binary trees to encode the relationships among objects.

Attention Models

The aforementioned fusion-based methods may lose critical information to correctly answer the question about local image contents (*e.g.*, "How many people are wearing shorts at the forefront of this photo?"). Consequently, a large amount of attention-based deep neural networks are proposed for VQA (Zhou et al. 2019; Anderson et al. 2018).

On the one hand, many works introduce visual attention mechanisms to adaptively learn the informative image features guided by the given question (Gao et al. 2019; Qiao, Dong, and Xu 2018; Lin et al. 2018). (Yang et al. 2016) propose a stacked attention network to iteratively learn the important visual regions according to questions. In (Shih, Singh, and Hoiem 2016), questions are answered by selecting relevant image regions in line with the text-based query. (Anderson et al. 2018) propose to align questions with relevant object proposals in images generated by Faster R-CNN (Ren et al. 2015), and use a bottom-up and top-down attention method to learn important candidate objects in an image according to the given question. The generated visual features are widely used in later methods. The VQA method in (Lu et al. 2018) performs attention on both free-form regions and detected object proposals based on questions to better utilize complementary information.

On the other hand, in addition to understanding the visual contents, VQA also requires fully understanding the semantic of questions. A number of co-attentin based methods are proposed to learn both the textual attention for questions and visual attention for images (Kim, Jun, and Zhang 2018; Yu et al. 2019; Nguyen and Okatani 2018). Lu *et al.* (Lu et al. 2016) propose a co-attention model to jointly reason for image and question. Our proposed attention module is more accurate than the existing co-attention based methods. The attention weights are derived from a similarity matrix that contains the interaction between each pair of words and objects. When there is more than one word relevant to multiple visual regions, the aggregated visual and textual features in our proposed attention module contains more fine-grained relationship between questions and images.

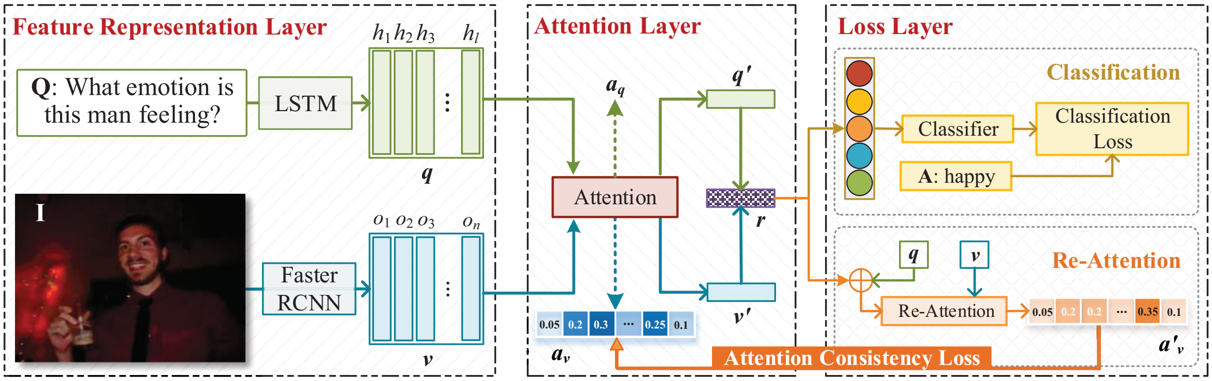

Figure 2: Illustration of the proposed method. We can obtain the feature representation of each candidate objects in image and words in question from LSTM (Hochreiter and Schmidhuber 1997) and the pre-trained Faster RCNN (Ren et al. 2015). The attention layer learns the importance weights for all objects and words by considering the relationship of each object-word pairs. Then we fuse the attended image and question features and deliver them into two branches. One of them is utilized to predict the answer, and the other is used to re-attend the objects in images with the guidance of answer representation from the proposed attention layer.

Methodology

The overall structure of the proposed method is illustrated in Figure 2. It takes the image-question pair as input and predicts the accurate answer by associating the question to relevant image objects. Specifically, the method consists of three parts: the feature representation layer, the attention layer, and the loss layer. Features of the input question and image are extracted from the representation layer. And the informative textual contents in questions and visual contents in image are fused into a unified answer representation through the proposed attention module. The answer representation is delivered into two branches. One is the classification branch, in which the answer representation is used to predict the correct answer about the given question. In the other branch, the fused representation is used to perform re-attention for the image. We use the answer representation to guide the model to attend the objects in the image by minimizing the difference between the re-attention weights and the visual weights learned in the proposed cross attention layer.

Feature Representation Layer

For an image-question pair, we first extract features of the input image I and the question Q in the representation layer. Following (Anderson et al. 2018), we represent the image as a set of visual object features which are extracted from the Faster R-CNN model (Ren et al. 2015) pre-trained on the Visual Genome dataset (Krishna et al. 2017). We obtain a dynamic number of objects for each image by controlling the threshold to the detected object probabilities. The number of detected objects is denoted as n. We use zero-padding to fill n to 100 if there are less than 100 candidates from Faster R-CNN. The i-th object in image I is denoted as $o_i \in \mathbb{R}^{d_v}$. The object features of the image I are represented as $v = [o_1, o_2, \cdots, o_n] \in \mathbb{R}^{n \times d_v}$.

The question Q with l words can be represented as a word sequence $Q =< w_1, w_2, \cdots, w_l >$. Every word is represented as a real-valued vector x_i, where $x_i \in \mathbb{R}^k$ is the k-dimensional word vector corresponding to the word w_i in the question. In our paper, we use fixed-length 300-dimensional word embeddings extracted from GloVe (Pennington, Socher, and Manning 2014). We use random vectors to initialize the words that are not in the dictionary of GloVe. The word vectors are fed into an LSTM network (Hochreiter and Schmidhuber 1997). We use the sequence of the output from all the LSTM cells as the question representation. The question Q is denoted as $q = [h_1, h_2, \cdots, h_l]$, $h_i \in \mathbb{R}^{d_q}$, where h_i is the output of the i-th LSTM cell. The length of q is set to 14 following (Teney et al. 2018) in this paper.

Attention Layer

Based on the representations from the feature representation layer, we can obtain the answer representation by fusing the informative contents of the question and image in this layer. Because the words in question are related to some visual contents in image. It is necessary to learn the textual attention for the question and the visual attention for the image. Different from the self-attention mechanism or the question-guided attention mechanism used in (Anderson et al. 2018; Yu et al. 2019), in which semantic of the whole question is used to guide the visual attention. In this section, we propose a cross attention mechanism. The interactions between every word in the question and every object in the image are considered. The weights for textual and visual contents are computing according to the similarity of every word-object pair. The structure of the proposed cross attention mechanism is shown in Figure 3.

The extracted features of question and image, *i.e.*, q and v, are first transformed into the same dimensional space:

$$\hat{q} = m_q(q) , \quad \hat{v} = m_v(v) , \tag{1}$$

where m_q and m_v are the transformation function for the question and image, and $\hat{q}$ and $\hat{v}$ are the transformed representations with the same dimension of d_c. $\hat{q}$ =

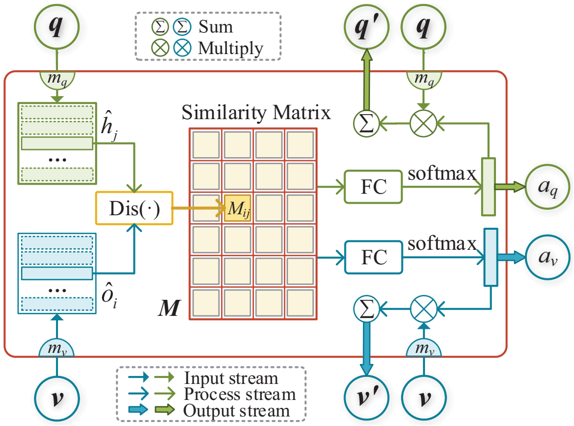

Figure 3: Overview of the proposed cross attention module. First, a similarity matrix M is conducted by computing the distance between the input features of the question $\hat{q}$ and image $\hat{v}$. Every element M_{ij} is computed by evaluating the distance of every pair of words and objects $< \hat{h}_i, \hat{o}_j >$. Then the matrix is mapped into two attention maps for the input question and image, respectively. Finally, the attention module outputs the weight vectors (*i.e.*, a_v and a_q) and attended features of question and image (*i.e.*, q' and v').

$[\hat{h_1}, \hat{h_2}, \cdots, \hat{h_l}], \hat{h_i} \in \mathbb{R}^{d_c}$ and $\hat{v} = [\hat{o_1}, \hat{o_2}, \cdots, \hat{o_n}], \hat{o_i} \in \mathbb{R}^{d_c}$. Then, we define a matrix M to represent the similarity between every pair of object and word. $M \in \mathbb{R}^{n \times l}$ can be calculated as follows:

$$M_{ij} = g(\hat{o}_i, \hat{h}_j)\,, \tag{2}$$

where $g(\hat{o}_i, \hat{h}_j)$ is the function to compute the distance between the i-th object $\hat{o}_i$ in the image and j-th word $\hat{h}_j$ in the question, which is defined as $M_{ij} = \hat{o}_i \times \hat{h}_j^\top$. The importance of words and objects are learned from M:

$$a_v = softmax(f_v(M)), a_v \in \mathbb{R}^{n \times 1}\,, \tag{3}$$

$$a_q = softmax(f_q(M)), a_q \in \mathbb{R}^{l \times 1}\,, \tag{4}$$

where f_v and f_q are two transformation functions for the image and the question. And a_v and a_q are learned attention weights for objects and words, respectively. The attended image and question features are calculated as the weighted summation of the features of the input image and the input question, *i.e.*, $\hat{v}$ and $\hat{q}$:

$$v' = \sum\nolimits_{i=1}^{n} a_{vi} \hat{o}_i\,, \quad q' = \sum\nolimits_{i=1}^{l} a_{q_i} \hat{h}_i\,. \tag{5}$$

Then, the obtained question and image representations, *i.e.*, v' and q', are fused into a unified representation as follows:

$$r = v' + q'\,, \tag{6}$$

where r is used as the answer representation in the following loss layer.

Loss Layer

The goal of this layer is to define an objective function to train the whole model. Specifically, the loss layer consists of two branches. One of them is the classification branch, which is used to determine the accurate answer. We use the binary cross-entropy as the loss function for the branch training following (Teney et al. 2018). The loss function L_c is defined as follows:

$$L_c = -\sum_{i}^{N} \sum_{j}^{C} y_{ij} \log(\hat{y_{ij}}) - (1 - y_{ij}) \log(1 - \hat{y_{ij}})\,, \tag{7}$$

where N and C are the number of training samples and candidate answers, respectively. y is the ground-truth answer and $\hat{y}$ denotes the predicted answer. $\hat{y} = \sigma(f_c(r))$, where $\sigma(\cdot)$ is the sigmoid function and f_c is the function to project the answer representation r to a vector with the dimension of C.

The other branch is the re-attention procedure, which utilizes answer representations to guide visual importance learning. We use the answer representation to learn the importance of visual contents again:

$$a'_{vi} = softmax(\varphi((r + avg(\hat{q})) \odot \hat{o}_i))\,, \tag{8}$$

where $\varphi(\cdot)$ is a transformation function. The model is trained to minimize the difference between the importance of visual contents learned in the proposed cross attention module and that learned in the re-attention branch. Specifically, the concatenation of r and $\hat{q}$ is used to learn the importance of image objects again, and a'_v indicates the learned attention weight. We define the difference between a'_v and a_v as the attention consistency loss, denoted as L_r:

$$L_r = \sum_{i}^{n} (a_{vi} - a'_{vi})^2\,. \tag{9}$$

Therefore, we explicitly train the proposed network to optimize the joint loss of the two branches:

$$Loss = L_c + \lambda_r L_r\,, \tag{10}$$

where λ_r is the trade-off of the strength of answer guidance. The λ_r is also estimated by the large-scale VQA dataset in the experiments. Since all the parameters can be derived, we can conduct an effective end-to-end representation learning using Adam (Kingma and Ba 2015) optimizer to minimize the joint loss function. With this scheme, we can predict the correct answer for the given question about the input image, during which the learned visual attention is guided by the obtained answer representation. It should be noted that we minimize the joint loss only in the model training.

Experiment

In this section, we compare our method against state-of-the-art methods to demonstrate the effectiveness of the proposed method for visual question answering.

Experimental Setup

In the following subsections, we present the setups in our experiments, including the used dataset and implement details of the proposed method.

Table 1: Performance evaluated on the test-dev and test-std splits of the VQA v2 dataset. The proposed method is compared with several state-of-the-art methods. The best results are in bold.

Method	test-dev				test-std			
	Yes/No	Number	Other	Overall	Yes/No	Number	Other	Overall
MCB (Fukui et al. 2016)	-	-	-	-	78.82	38.28	53.36	62.27
HAN (Qiao, Dong, and Xu 2018)	78.54	37.94	53.38	61.99	-	-	-	-
ReNet-3000 (Ma et al. 2018)	-	-	-	-	79.2	39.5	52.6	62.10
PMC (Hu, Chao, and Sha 2018)	-	-	-	63.90	-	-	-	-
UpDn (Anderson et al. 2018)	81.82	44.21	56.05	65.32	-	-	-	65.67
Dual-MFA (Lu et al. 2018)	83.59	40.18	56.84	66.01	83.37	40.39	56.89	66.09
rTV/rTV (Lin et al. 2018)	82.50	45.80	57.34	66.40	82.44	44.93	57.60	66.52
CoR-3 (Wu et al. 2018)	85.22	47.95	59.15	68.62	85.76	48.40	59.43	69.14
Graph (Norcliffe-Brown et al. 2018)	-	-	-	-	82.91	47.13	56.22	66.18
DCN (Nguyen and Okatani 2018)	84.48	41.66	57.44	66.83	84.61	41.27	56.83	66.66
Counting (Zhang et al. 2018)	83.14	51.62	58.97	68.09	83.56	51.39	59.11	68.41
BLOCK (Ben-Younes et al. 2019)	82.86	44.76	57.30	66.41	-	-	-	-
MuRel (Cadene et al. 2019)	84.77	49.84	57.85	68.03	-	-	-	68.41
VCTREE-HL (Tang et al. 2019)	84.28	47.78	59.11	68.19	84.55	47.36	59.34	68.49
Ours	**87.00**	**53.06**	**60.19**	**70.43**	**86.97**	**52.62**	**60.72**	**70.72**

Dataset In our experiments, we use the VQA v2 dataset (Goyal et al. 2017) to evaluate the performance of the proposed method. VQA v2 is the most commonly used balanced benchmark with significantly reduced language biases. The images are from the Microsoft COCO dataset (Lin et al. 2014). Every image corresponds to several questions, and every question corresponds to ten answers collected from human annotators. The complete dataset contains almost 1.1 million image-question pairs. The dataset is typically split into train set, validation set, and test set, which contain 82k, 40k, and 81k images with 443k, 214k, and 447k questions, respectively. Performances of VQA models are evaluated online on the developing (test-dev) and standard (test-std) subsets.

Implementation Details The hyper-parameters used in our model are set as follows. The number of objects in the image and the number of words in the question are padded as 100 and 14, respectively. For the hyper-parameters like the hidden size of the LSTM, d_q, the experimental results are stable when h_q changes. Therefore, it is set as 512 following (Tang et al. 2019) after we comprehensively consider the trade-off between model complexity and performance. The input size of the LSTM is set as 300 in the same way. The dimension of the object features d_v is 2048. The dimension of $\hat{q}$ and $\hat{v}$, *i.e.*, d_c is 512. The size of the answer set is $3,129$ following the strategy in (Teney et al. 2018). All the models are trained with batch size 64. Our framework is implemented using PyTorch and trained with Adam (Kingma and Ba 2015). All of our approaches are trained on an NVIDIA GTX 1080ti with 11GB on-board memory. Only the train split is used during model training for the results evaluated on the validation split. For the performance on the test split, part of samples in the Visual Genome dataset (Krishna et al. 2017) is used as the augmented dataset to facilitate model training following (Yu et al. 2019).

Comparison with State-of-the-Art Methods

In this section, we evaluate the proposed method against the state-of-the-art algorithms for visual question answering, including the fusion-based methods, the attention-based methods, and visual reasoning methods.

We compare our method with fusion-based methods, such as, **MCB** (Fukui et al. 2016) and **ResNet-3000** (Ma et al. 2018). We also compare it with some attention-based methods. **HAN** (Qiao, Dong, and Xu 2018) uses the generated human-like attention maps as supervision to an attention-based VQA model. The **UpDn** (Anderson et al. 2018) proposes to use the object features from Faster RCNN which is used in many later methods. And **DCN** (Nguyen and Okatani 2018) uses the stack of multiple co-attention layers and significantly improves the performance. Besides, beyond the attention mechanism, there are other methods reasoning the relationship among images, question, such as, **rTV/rTV** (Lin et al. 2018), **Counting** (Zhang, Hare, and Prügel-Bennett 2018), **Dual-MFA** (Lu et al. 2018), and **Graph** (Norcliffe-Brown, Vafeais, and Parisot 2018). We also compare with other VQA methods, including **BLOCK** (Ben-Younes et al. 2019), **MuRel** (Cadene et al. 2019), **VCTREE-HL** (Tang et al. 2019), and **PMC** (Hu, Chao, and Sha 2018).

Performances of our method and the compared methods are reported in Table 1. Our method outperforms all of the compared fusion-based methods, attention-based methods, and other reasoning methods. Compared with other attention-based methods, like the classical attention-based method UpDn (Anderson et al. 2018), our method gains significant performance improvement of 5.11% for the overall accuracy on the test-dev split. The UpDn utilizes the feature of the whole question to learn the importance of different objects, while our method models the relationship between each pair of words and objects. Benefiting from it, both the

Table 2: Accuracies of the “base+re-att” and “base+co+re-att (Ours)” when the value of λ_r in Equation (10) varies from 0 to 2.0. All the accuracies are evaluated on the validation split of the VQA v2 dataset. Note that $\lambda_r = 0$ represents that only the forward branch is employed in the model training phase. “base+re-att” denotes the baseline model with the re-attention branch. “base+co+re-att” is the baseline model with the proposed cross attention and re-attention.

Method	Type	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
base+re-att	Yes/No	83.28	84.74	84.65	84.66	**84.74**	84.72	84.74	84.71	84.66	84.67	84.56
	Number	44.71	49.35	49.29	49.04	**49.30**	49.16	49.35	49.46	49.67	49.35	49.16
	Other	57.15	57.86	57.98	**58.05**	57.97	57.95	57.86	57.96	57.90	57.86	57.74
	Overall	65.34	66.84	66.86	66.87	**66.89**	66.86	66.84	66.89	66.88	66.82	66.69
base+co+re-att (Ours)	Yes/No	84.40	84.49	84.35	84.64	**84.85**	84.41	84.64	84.55	84.55	84.65	84.61
	Number	48.62	49.27	**49.55**	49.15	49.30	49.75	49.15	48.85	49.05	49.15	49.05
	Other	57.86	58.10	58.02	58.12	**58.40**	58.15	58.12	58.07	58.06	58.13	58.17
	Overall	66.62	66.86	66.80	66.91	**67.15**	66.92	66.90	66.81	66.83	66.92	66.90

Table 3: Ablation analysis on the validation split. “base” indicates the baseline model with a self-attention for questions and a question guided image attention module. “base+co”, “base+re-att”, and “base+co+re-att” denote the baseline model with the cross attention module, re-attention branch, or both, respectively.

Method	Yes/No	Number	Other	Overall
base	83.28	44.71	57.15	65.34
base+co	84.40	48.62	57.86	66.62
base+re-att	84.74	49.30	57.97	66.89
Ours	**84.85**	**49.30**	**58.40**	**67.15**

question and image can be better understood and the satisfactory answer will be generated. This demonstrates the significance of the interaction between every word and object. In addition, our method performs favorably against the state-of-the-art method, VCTREE-HL (Tang et al. 2019) by 2.24 points for the overall accuracy on the test-dev split. We all consider interactions between images and questions, while in the proposed re-attention strategy, answers are insightfully used to re-attend meaningful visual regions in images. The rich information in answers are fully used to help our model learn more accurate visual attention maps.

Ablation Study

This experiment is conducted to verify the effect of every part in the proposed model. We show the results of the proposed method with different configurations on the validation set. We perform the ablation study to illustrate the effectiveness of the proposed attention module and the answer-guided re-attention module. The “base” model is a basic attention-based VQA model with a self-attention for the question and a question-guided attention module for the input image. The “base+co” method is the “base” model with the proposed cross attention module to further learn the informative textual and visual contents based on features of the “base” model. The “base+re-att” method is the “base” model with a re-attention branch, in which the difference between the attention learned in the question-guided module and that learned in the re-attention branch are minimized. “base+co+re-att” is our proposed method, in which the proposed re-attention branch is build on the top of “base+co” model. The four methods use the same classifier and representation layer.

As reported in Table 3, “base+co+re-att (Ours)” achieves the best performance. Both of the “base+co” method and the “base+re-att” method outperform the “base” model. The performance of “base+re-att” is slightly better than the performance of the “base+co” method. We can draw the following conclusions: First, the proposed cross attention method proposed in this paper achieves better performance by taking into account fine-grained relationships between the questions and images. Second, our proposed re-attention model achieves the best accuracy by utilizing the answer information to help the visual attention learning.

Therefore, it is necessary to consider the fine-grained interaction between questions and images, and the information contained in answers.

Hyperparameter Analysis

Since the objective function consists of the classification loss and attention consistency loss. In this section, we analyze the performance of the proposed method when the trade-off of the re-attention branch, *i.e.*, λ_r in Equation (10), is set as different values. In Table 2, we report detailed accuracies of the “base+re-att” and “base+co+re-att” (Ours) with different λ_r for the types of “Yes/No”, “Number”, and “Other”. As shown in Table 2, the performance varies along with the λ_r. And both of the “base+re-att” method and “base+co+re-att” (Ours) method achieve best performance when λ_r is 0.8. Therefore, we set λ_r as 0.8 in the section of ablation study and the comparison with state-of-the-art methods.

Visualization

The learned image and question attentions and the predicted answers of nine typical examples in the VQA v2 dataset are represented in Figure 4. They cover types as broad as “Yes/No”, “Number”, and “Other”. The highlighted parts in the image are the attended object proposals. Words in bold

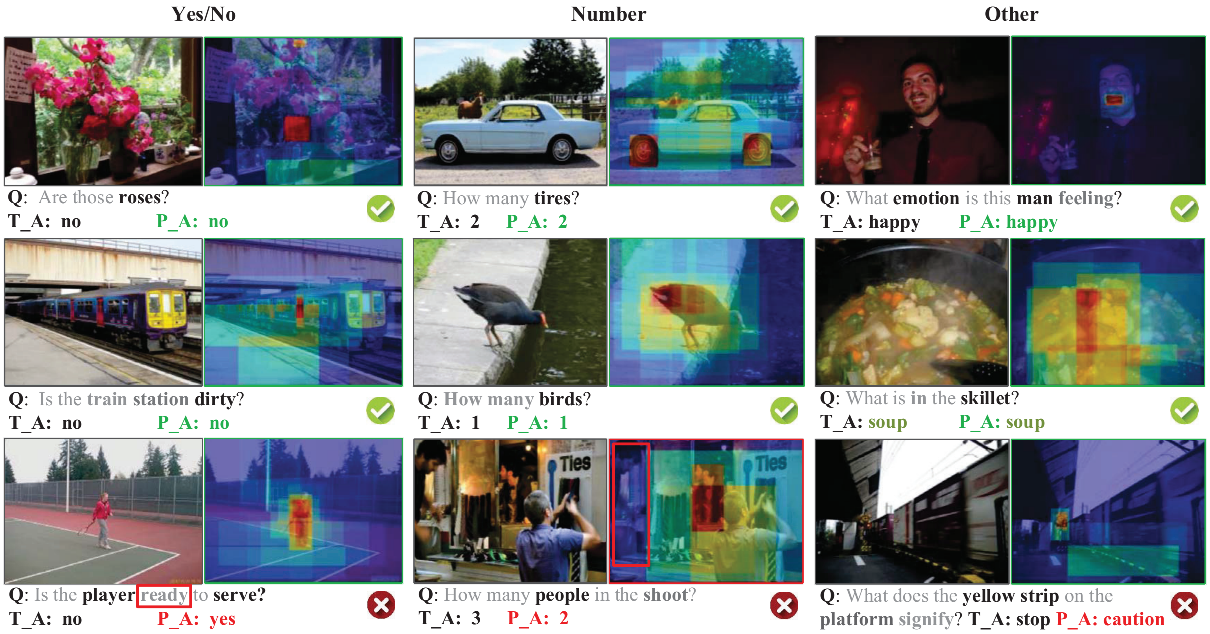
Yes/No
Number
Other
Q: Are those roses?
T_A: no P_A: no
Q: How many tires?
T_A: 2 P_A: 2
Q: What emotion is this man feeling?
T_A: happy P_A: happy
Q: Is the train station dirty?
T_A: no P_A: no
Q: How many birds?
T_A: 1 P_A: 1
Q: What is in the skillet?
T_A: soup P_A: soup
Q: Is the player ready to serve?
T_A: no P_A: yes
Ties
Q: How many people in the shoot?
T_A: 3 P_A: 2
Q: What does the yellow strip on the platform signify? T_A: stop P_A: caution

Figure 4: Visualization of learned attentions for some typical examples in the VQA v2 dataset. For each example, the image on the left indicates the input image; the learned image attention is visualized in the right. We also display the question attention (Q), ground-truth answer (T_A), and predicted answer (P_A) below the image. The highlighted part in the images and words in bold denote the attended visual and textual contents, respectively. The content in red indicates the missed attention or the wrong answer predicted by the method.

are the learned meaningful words. In the last row, the content in red indicates the missed attention or the wrong answer predicted by the method.

For the correctly answered questions, the keywords in questions and objects closely related to the questions are accurately attended. Then, the model gives the correct answers according to the attended visual and textual contents. Take the first sample in the type of “Number” as an example, the proposed model focuses on the keyword “tires” and the question is correctly understood. And then, the tires of the car in the image are attended according to the semantic of the question. The correct answer “2” is predicted based on the accurate understanding of the question and the image. It is the same to the first sample in the type of “other”, our model correctly attended the smiling mouth of the man, thus giving the answer that the emotion of the man in the image is “happy”. For the incorrectly answered questions, there are two main limitation: First, the model fails to distinguish the keywords in the question or key objects in the image, which hinders the model learning the correct answer (*e.g.*, the model fails to focus on the keyword “ready”, which causes the incorrect answer for the first failure instance); Second, the model fails to predict the correct answer even from the closely attended textual and visual contents. Take the last failure instance as an example, because some common sense is involved, which is difficult to understand through the attention-based methods. Our model predicts a wrong answer, although the model focuses on the key words “yellow strip”, “platform”, and the closely relevant objects in the image. These observations can help us to further improve our model.

Conclusion

In this paper, we propose a re-attention model to address the problem of visual question answering. Inspired by the observation that both questions and answers contribute to describing visual contents, rather than only using answers as classification labels, we utilize the answer representation to guide visual attention learning. Specifically, we conduct an attention consistency loss to evaluate the difference between the learned visual attention by only questions and that learned in the re-attention. Besides, a new attention module is proposed to correlate the relationship of each object-word pair in the fine-grained perspective than other co-attention based methods. Our method outperforms the compared methods on a commonly used benchmark dataset, and experimental results show the effectiveness of the proposed cross attention and re-attention module.

Acknowledgments

This work was supported by the Major Project for New Generation of AI Grant (NO. 2018AAA010040003), NSFC (NO. 61876094, U1933114, U1836109, 61772289), Natural Science Foundation of Tianjin, China (NO. 18JCYBJC15400, 18ZXZNGX00110), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. VQA: Visual question answering. In *ICCV*.

Ben-Younes, H.; Cadene, R.; Thome, N.; and Cord, M. 2019. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *AAAI*.

Cadene, R.; Ben-younes, H.; Cord, M.; and Thome, N. 2019. Murel: Multimodal relational reasoning for visual question answering. In *CVPR*.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *CVPR*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*.

Haurilet, M.; Roitberg, A.; and Stiefelhagen, R. 2019. It's not about the journey; it's about the destination: Following soft paths under question-guidance for visual reasoning. In *CVPR*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*.

Hu, H.; Chao, W.-L.; and Sha, F. 2018. Learning answer embeddings for visual question answering. In *CVPR*.

Kim, J. H.; Lee, S. W.; Kwak, D. H.; Heo, M. O.; and Zhang, B. T. 2016. Multimodal residual learning for visual qa. In *NIPS*.

Kim, J.; Jun, J.; and Zhang, B. 2018. Bilinear attention networks. In *NIPS*.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *ECCV*.

Lin, Y.; Pang, Z.; Wang, D.; and Zhuang, Y. 2018. Feature enhancement in attention for visual question answering. In *IJCAI*.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*.

Lu, P.; Li, H.; Zhang, W.; Wang, J.; and Wang, X. 2018. Co-attending free-form regions and detections with multimodal multiplicative feature embedding for visual question answering. In *AAAI*.

Ma, C.; Shen, C.; Dick, A. R.; Wu, Q.; Wang, P.; van den Hengel, A.; and Reid, I. D. 2018. Visual question answering with memory-augmented networks. In *CVPR*.

Nguyen, D.-K., and Okatani, T. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *CVPR*.

Norcliffe-Brown, W.; Vafeais, E.; and Parisot, S. 2018. Learning conditioned graph structures for interpretable visual question answering. In *NIPS*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Qiao, T.; Dong, J.; and Xu, D. 2018. Exploring human-like attention supervision in visual question answering. In *AAAI*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Shih, K. J.; Singh, S.; and Hoiem, D. 2016. Where to look: Focus regions for visual question answering. In *CVPR*.

Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2019. Learning to compose dynamic tree structures for visual contexts. In *CVPR*.

Teney, D.; Anderson, P.; He, X.; and van den Hengel, A. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *CVPR*.

Wu, C.; Liu, J.; Wang, X.; and Dong, X. 2018. Chain of reasoning for visual question answering. In *NIPS*.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *CVPR*.

Yu, Z.; Yu, Z.; Chenchao Xiang, J. F.; and Tao, D. 2018. Beyond bilinear: Generalized multi-modal factorized high-order pooling for visual question answering. *TNNLS*.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *CVPR*.

Zhang, Y.; Hare, J. S.; and Prügel-Bennett, A. 2018. Learning to count objects in natural images for visual question answering. In *ICLR*.

Zhou, B.; Tian, Y.; Sukhbaatar, S.; Szlam, A.; and Fergus, R. 2015. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*.

Zhou, Y.; Ji, R.; Su, J.; Sun, X.; and Chen, W. 2019. Dynamic capsule attention for visual question answering. In *AAAI*.