IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 24, 2019, accepted March 17, 2019, date of publication March 28, 2019, date of current version April 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2908035

Co-Attention Network With Question Type for Visual Question Answering

CHAO YANG[1], MENGQI JIANG[1], BIN JIANG[1], WEIXIN ZHOU[1], (Fellow, IEEE), AND KEQIN LI[2]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
[2]Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

Corresponding author: Chao Yang (yangchaoedu@hnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61702176, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2017JJ3038, and in part by the Foundation of Changsha Technological Plan under Grant kq1706020.

ABSTRACT Visual Question Answering (VQA) is a challenging multi-modal learning task since it requires an understanding of both visual and textual modalities simultaneously. Therefore, the approaches used to represent the images and questions in a fine-grained manner play key roles in the performance. In order to obtain the fine-grained image and question representations, we develop a co-attention mechanism using an end-to-end deep network architecture to jointly learn both the image and the question features. Specifically, textual attention implemented by a self-attention model will reduce unrelated information and extract more discriminative features for question-level representations, which is in turn used to guide visual attention. We also note that a lot of finished works use complex models to extract feature representations but neglect to use high-level information summary such as question types in learning. Hence, we introduce the question type in our work by directly concatenating it with the multi-modal joint representation to narrow down the candidate answer space. A new network architecture combining the proposed co-attention mechanism and question type provides a unified model for VQA. The extensive experiments on two public datasets demonstrate the effectiveness of our model as compared with several state-of-the-art approaches.

INDEX TERMS Co-attention, question type, self-attention, visual question answering.

I. INTRODUCTION

Recently, multi-modal learning for computer vision and natural language processing has grown by leaps and bounds, such as visual question answering [1], image captioning [2] and image-text matching [3], etc. The VQA tasks require to provide the correct answer to a question with a corresponding image, as shown in Fig. 1. There are many potential applications for VQA, such as image retrieval [4], aided-navigation for blind individuals [5] and automatic querying of surveillance video [6]. Predicting the best matching answer correctly has always been one of the most challenging tasks for VQA, since it requires a fine-grained understanding of the question text and parsing the visual scene and it may also involve complex reasoning.

Due to in-depth research on computer vision and natural language processing, numerous methods have attacked VQA and have achieved good results. In the early stage, most models directly learn the joint embedding of visual and textual features through linear pooling (such as element-wise addition or multiplication) and then feed it into a classifier to predict the most related answer. Specifically, visual features are obtained with convolutional neural network (CNN) pre-trained on object recognition, and textual features are obtained with recurrent neural network (RNN). However, these visual and textual features are represented at

FIGURE 1. Examples of different questions in VQA. Q=question, A=answer.

The associate editor coordinating the review of this manuscript and approving it for publication was Yanzheng Zhu.

the global-level and may bring irrelevant or noisy information. Hence, the attention-based models have been developed. The idea behind the attention mechanism is to assign different weights to local features instead of considering global features merely. More recently, models based on co-attention have been widely used in VQA tasks, aiming to focus on salient regions of the image and critical words of the question [7]–[9]. For co-attention mechanisms, question-guided visual attention is obtained based on the question information, while image-guided textual attention is obtained based on the image information. The co-attention mechanism can reduce unrelated information and obtain more meaningful features representations for image and question.

In spite that a lot of promising results have been achieved, the capabilities of the methods based on co-attention are still far from satisfaction. We argue that this is mainly because the following reasons. The primary reason is that question-guided visual attention mechanism uses the whole question feature to guide visual attention, which will distract attention and fail to attend to the question-related regions accurately, due to the colloquial words in the question. Another reason is that the image-guided textual attention mechanism uses the image features to guide textual attention, which does not precisely attend to the important words of the question, because there are many interference information in the image that is not related to the question. For the reasons mentioned above, we propose a new co-attention mechanism. More specifically, self-attention is performed on the question without the guidance of image features to obtain discriminative question representation, which is in turn used to guide visual attention. We consider that the advantage of introducing self-attention mechanism is that it can assign greater weights to those important words in the question, thus reducing the negative impact of irrelevant information on the accuracy of the answer prediction. Compared with the traditional co-attention mechanism, our proposed co-attention mechanism: 1) contributes to attend to the image regions which are most relevant to the question; 2) does not exploit the image features when calculating the textual attention, thus reduces the computational overhead.

Moreover, we notice that in the existing models, question/answer type is usually not considered in training. In general, each sample in VQA dataset includes an image, a question/answer pair, and an answer type. Most works use answer type for result analysis but it is not considered during the training process. Compared with the answer type, question type has less variety and is easier to interpret when we only have the question. Meanwhile, question input can be clustered into question types with different semantics. Hence, we divide the questions from VQA datasets into 8 sub-categories, including *color*, *time*, *counting*, *location*, *reason*, *sport*, *judgement* and *other*. Introducing question type will help the model know the type of question before answering, so it can reduce the search space of answers. For this purpose, we fuse the question type by directly concatenating it with the multi-modal joint representation.

In this paper, Co-Attention Network with Question Type (CAQT) is proposed to address the VQA task. CAQT is designed to integrate co-attention mechanism and question type into one unified model for VQA.

The key contributions of this work are three-fold:

- We propose a novel co-attention mechanism for the VQA task. For the given questions, we perform self-attention to assign greater weights to the important words. And then, we use the new representation to guide visual attention of images.
- We concatenate the one-hot encoding of the question type directly to the multi-modal joint representation for later answer generation. Our intuitive motivation is that knowing question type before answering could narrow down the candidate answer space.
- Extensive experiments performed on two benchmark VQA datasets demonstrate the feasibility and effectiveness of CAQT.

The remainder of this paper is organized as follows. Section II reviews the related works and section III introduces preliminary knowledge about the VQA tasks. In section IV, we provide the details of CAQT. We then perform the experimental evaluation in Section V. Finally, we present the conclusion of this paper and provide the future work in Section VI.

II. RELATED WORK

This section is divided into three parts. The first part introduces the related knowledge of the VQA task. The second part introduces the models based on the attention mechanism. The final part introduces the models with question-type.

A. VISUAL QUESTION ANSWERING

VQA lies in the intersection of computer vision and natural language processing, which has attracted increasing interest from multiple research fields. A series of major datasets for VQA have been publicly released, including DAQUAR [10], COCO-QA [11], VQA [1], FM-IQA [12], Visual7W [13], and Visual Genome [14]. A basic framework for the VQA task first encodes question embedding using RNN model and extracts image feature via CNN model, then fuses the question and image features, and finally, uses this feature to predict the answer. Recently, effective bilinear pooling methods such as MCB [7], MLB [15], MFB [16] and MLPB [17] have been proposed, which are superior to linear pooling (concatenation, element-wise addition or multiplication). Moreover, memory-augmented neural networks [18] and attention-based models [8], [19], [23], [25], [31], [32] have also been developed. Therefore, significant progress has been made in the study of VQA.

B. ATTENTION MODELS

Currently, the mainstream VQA models are essentially based on attention mechanisms. Its success mainly relies on the reasonable assumption that humans have the ability to quickly understand the visual scene by attending to selective parts

of the whole image instead of processing the entire scene at once [20]. Attention learns to attend to the most relevant regions of the input space and assigns different weights to different regions. Attention mechanisms are firstly used in machine translation [21] and then are employed to solve the multi-modal tasks, such as image caption [22], VQA [23] and Cross-media Retrieval [9]. In the VQA task, the attention mechanism is used to identify ''where to look'' [24] and ''which word to listen'' before carrying on further computations. For example, for the question ''What color is his hat?'', the image region containing ''hat'' is more informative than other image regions, and the textual information containing ''color'' and ''hat'' are more important than other words in the question. Numerous works have concentrated on using the attention mechanisms to solve the VQA task [8], [9], [16], [23], [25]–[27].

Recently, some works have introduced visual attention to address the VQA task. For example, [23] designed an Attention-based Configurable Convolutional Neural Network (ABC-CNN) to learn question-guided attention. ABC-CNN determined an attention map for each image-question pair by convolving the image feature map with configurable convolutional kernels derived from the question's semantics. And [25] presented a scheme with Stacked Attention Networks (SAN) to obtain the answer, SAN regarded the semantics of the question as a query to find the regions of the image that were associated with the answer. Some recent works integrate visual attention with textual attention to further improve VQA performance. Reference [8] designed a Hierarchical Co-attention Model (HieCoAtt), simultaneously reasoning about image and question attention. Reference [9] proposed a Dual Attention Networks (DANs), which refined specific regions in the images and words in the text through multiple reasoning steps, in order to capture essential information from visual and textual features. In the VQA task, they explored a reasoning way that allowed textual attention and visual attention to steer each other during inference period [9]. In [16] and [26], the networks can find important information in question text without the guidance of the image. Reference [27] designed a high-order attention mechanism for multi-modal input data. These inputs included an image, a question, and 18 candidate answers. They considered that learning high-order correlations could obtain the appropriate information from different data modalities (question, image, and answers), so as to infer a correct answer. In this paper, we compute the textual attention based on the question itself, without having to consider the image feature.

Besides, self-attention [28] was first proposed to solve machine translation problems and achieved good performance. In long-distance dependence, self-attention module calculated the attention at position t in a sequence by attending to all positions. Currently, some methods [29], [30] based on self-attention are used in the field of natural language processing. For example, [29] proposed a model for extracting an interpretable sentence embedding by self-attention, which achieved promising performance on author profiling, sentiment classification, and textual entailment. And [30] proposed self-attention based approach to tackle Semantic Role Labeling, which can directly capture the relationships between two tokens regardless of their distance. Our model is partly motivated by [29], which utilizes self-attention to find the most informative components of the question and uses a matrix to represent the question.

C. MODEL WITH QUESTION-TYPE

When considering the model with question-type for solving the VQA task, [33] introduced Question Type-guided Attention (QTA) that dynamically gated the contribution of ResNet [34] and Faster R-CNN [35] features. QTA utilized the information of question type to guide the visual encoding process, and the experiments over TDIUC [36] dataset showed impressive performance. Motivated by [33], we also consider the question type in our model. The main idea is that if the model knows the question type before answering the question, the search space of answers set can be reduced.

III. PRELIMINARY

In this section, we first formulate the VQA problem addressed in this paper and then illuminate the basic framework for the problem.

A. PROBLEM DESCRIPTION

Given an image I and the related question Q, the VQA model is designed to predict possible answer. The dominant methods typically formalize VQA as a classification problem in the space of candidate answers. This can be formulated as:

$$\hat{a} = \underset{a \in \Omega}{arg\,max}\ p(a|Q, I; \Theta), \tag{1}$$

where Ω is the set of candidate answers, Θ represents the parameters of the method.

It is worth noting that ambiguous or subjective questions might have multiple correct answers. Hence, we use a sigmoid output that allows multiple correct answers for each question, instead of a common single-label softmax. That is, in this paper, we treat VQA as a multi-label regression problem.

B. COMMON FRAMEWORK

The basic framework for VQA always consists of three major components which are image embedding, question embedding, and joint feature learning.

1) IMAGE EMBEDDING

In the literature, CNNs are typically used in extracting image feature. CNNs are pre-trained on ImageNet, with common examples as ResNet-152 [34] and Faster R-CNN [35]. The feature representation extracted from ResNet-152 has a size of $2048 \times 14 \times 14$, where 14×14 indicates the number of image regions and 2048 represents the dimension of each region. While in Faster R-CNN, the input image is passed through it to obtain a vector representation of size $K \times 2048$

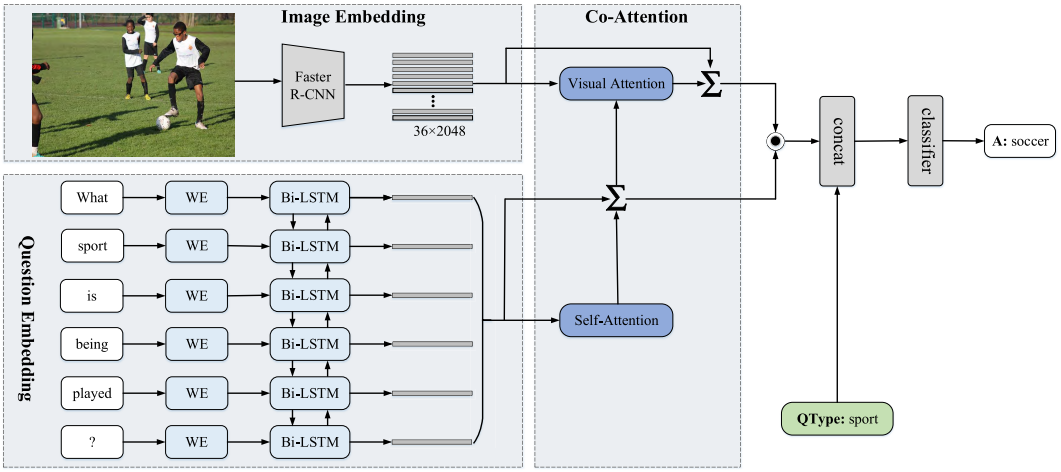

FIGURE 2. Overview of the CAQT model. WE=word embedding, QType=question type, A=answer.

(in [37], $K = 36$), where K is the number of objects in the image and the dimension of each object is 2048. Given I represents the input image, V denotes the output vector, the image feature is obtained by:

$$V = CNN(I). \tag{2}$$

2) QUESTION EMBEDDING

RNNs like Long Short-Term Memory (LSTM) [38] and Gated Recurrent Unit (GRU) [39] are typically used in extracting question feature. Given Q represents the input question, H denotes the output vector, the question embedding is obtained by:

$$H = RNN(Q). \tag{3}$$

Specifically, the number of words in the question is usually limited to 14, the embedding of each word is sequentially fed into the RNN model. The final hidden state of the RNN model is considered as question representation.

3) JOINT FEATURE LEARNING

The image feature V and the question feature H are fused via multi-modal pooling, i.e., concatenation, element-wise addition or multiplication, MCB, MLB, MFB, and etc. The joint feature is:

$$F = f(V, H), \tag{4}$$

where f indicates the multi-modal pooling module. The joint feature F is then fed into the classifier to predict the answer.

Recently, many models incorporate the co-attention mechanism for getting more discriminative visual and textual representations, and experiments have shown that the model based on co-attention has achieved the state-of-the-art results.

IV. MODEL

In this section, we describe our proposed CAQT model and explain it in more details. The overall architecture of our proposed model is given in Fig. 2. CAQT has the following four parts:

- Input representation module which includes image and question features extraction components.
- Co-Attention module which includes self-attention based textual attention and question-guided visual attention.
- Question type module which concatenates the one-hot encoding of the question type directly to the multi-modal joint representation.
- Prediction module is used to infer answers.

A. MODELING

1) INPUT REPRESENTATION

a: IMAGE EMBEDDING

Recently, [40] use bottom-up attention to obtain image features which have achieved promising results in comparison with other image feature extraction methods. In light of this, we employ Faster R-CNN [35] for high-level image feature extraction and obtain the top-K candidate objects of the input image I:

$$V = CNN(I), \tag{5}$$

$$V = [v_1, v_2, \ldots, v_K] \in R^{K \times d_v}, \tag{6}$$

where $v_k \in R^{d_v}$ indicates the k-th object feature, K is the number of objects in the image.

b: QUESTION EMBEDDING

In this paper, we use bi-directional LSTM (Bi-LSTM) to encode questions. The question is first tokenized into words and then transformed to one-hot feature vectors $q = [q^1, q^2, \ldots, q^n]$, where $q^t \in R^D$ is the one-hot encoding at position t, n is the length of the question, D is the number of words in the vocabulary. We first convert q_t into a vector representation x_t by $x_t = W_e q_t$, where W_e is an embedding matrix, x_t is learned along other parameters during training.

At each step, we feed the vector x_t into Bi-LSTM:

$$h_t^f = LSTM^f(x_t, h_{t-1}^f), \tag{7}$$
$$h_t^b = LSTM^b(x_t, h_{t+1}^b), \tag{8}$$
$$h_t = [h_t^f, h_t^b], \tag{9}$$
$$H = [h_1, h_2, \ldots, h_n], \tag{10}$$

where $h_t^f \in R^{d_h}$ and $h_t^b \in R^{d_h}$ represent the hidden states at time t from forward and backward LSTMs, respectively. h_t encodes the semantics of the t-th word in the context of the entire question. Note that we use all the hidden states of Bi-LSTM ($H \in R^{n\times 2d_h}$) instead of the final hidden state as the expression of the question, which helps improve the model performance.

2) CO-ATTENTION

Co-attention mechanism ensures deep cross-domain interactions and obtains discriminative features. Our co-attention mechanism combines self-attention based textual attention and visual attention. We will introduce these two parts separately in the following. We first perform self-attention on the question and receive a new expression of the question, then use this expression to perform visual attention. Self-attention can highlight the focus of the question, question-guided visual attention will give greater weights to regions of the image which are relevant to important semantic of the question.

a: TEXTUAL ATTENTION BASED ON SELF-ATTENTION

Textual attention based on self-attention performs self-attention on the question. In [16], the mechanism for dealing with questions is textual attention without the guidance of image features, but the result is a vector that can only represent a specific component of the question, and this representation is not enough to express the complicated question (especially for reasoning questions with multiple objects). But in our work, we use the self-attention proposed by [29], which can fully express most or even all components of the question. Thus, it is not difficulty in expressing complex questions.

We use a 2-D matrix to represent the sentence embedding, with each row of the matrix attending on a different part of the sentence. The self-attention mechanism takes the whole Bi-LSTM hidden states H as input and outputs a vector of weights a. The question features are weighted by the self-attention weights a and then summed into a single vector m which represents the attended question. The equations are as follows:

$$a = softmax(w_{s2}tanh(W_{s1}H^T)), \tag{11}$$
$$m = aH, \tag{12}$$

where $w_{s2} \in R^{1\times d_a}$ and $W_{s1} \in R^{d_a\times 2d_h}$ are learnable parameters, d_a is a hyperparameter. The self-attention ensures that the more important word in question will be assigned greater weight. $m \in R^{1\times 2d_h}$ usually focuses on a specific component of the question, but questions usually have multiple related words or phrases, so we need to perform multiple hops of attention:

$$A = softmax(W_{s2}tanh(W_{s1}H^T)), \tag{13}$$
$$M = AH, \tag{14}$$

where $W_{s2} \in R^{r\times d_a}$ and $W_{s1} \in R^{da\times 2d_h}$ are learnable parameters, r is a hyperparameter which represents the number of parts extracted from the question. Finally, the question is expressed as $M \in R^{r\times 2d_h}$, which highlights the focus of the question.

b: QUESTION-GUIDED VISUAL ATTENTION

Visual attention uses the question representation M as a guide to attend the objects that are most relevant to the question. M becomes m' by mean function. The question representation m' and the image representation V are firstly projected to the same dimension by non-linear layers. Next, we use element-wise multiplication to fuse these projected representations, and then compute the normalized attention weight α_k of each image object feature v_k through a linear-layer and softmax function. Finally, the image features are weighted by normalized attention weights and summed into a single vector $u \in R^{1\times 2d_h}$ which represents the attended image. The calculation details are listed below:

$$m' = mean(M), \tag{15}$$
$$e = w_e(ReLU(W_q m'^T) \odot ReLU(W_v V^T)), \tag{16}$$
$$\alpha = softmax(e), \tag{17}$$
$$u = \sum_k^K \alpha_k v_k, \tag{18}$$
$$ReLU(x) = max(0, x), \tag{19}$$

where $w_e \in R^{1\times 2d_h}$, $W_q \in R^{d_q\times 2d_h}$ and $W_v \in R^{d_v\times 2d_h}$ are learnable parameters, d_q and d_v are hyperparameters, $2d_h = d_q = d_v$. $\odot$ represents element-wise multiplication.

3) QUESTION TYPE

We believe that adding the question type information before feeding the merged features into the classifier can narrow down the range of answers. For example, questions starting with "how many" will mostly lead to numerical answers. Therefore, we divide the VQA v1.0 and VQA v2.0 datasets into 8 sub-categories, including *color*, *time*, *counting*, *location*, *reason*, *sport*, *judgement* and *other*. The distribution of the percentage of each question type is shown in Tab. 1. For question types that are difficult to distinguish between specific types in the datasets or that have a small proportion, we attribute them to the *other* class. As can be seen from Tab. 1, *judgement* class and *sport* class respectively has the highest proportion and the lowest proportion in training set, validation set and testing set of the VQA datasets. The distributions of question types in the training set and validation set of the two datasets are relatively uniform, but the questions in

TABLE 1. The proportion of question types in the training set, validation set and testing set of the VQA v1.0 and VQA v2.0 datasets, respectively.

Category	VQA v1.0 (%)			VQA v2.0 (%)		
	training set	validation set	testing set	training set	validation set	testing set
color	8.976	9.465	7.925	8.975	9.303	8.002
time	0.673	0.830	0.018	0.657	0.815	0.017
counting	10.640	10.587	9.092	11.347	11.217	9.670
location	2.099	2.511	3.027	2.004	2.346	2.857
reason	1.201	1.051	0.034	1.102	0.911	0.033
sport	0.621	0.547	0.004	0.569	0.507	0.003
judgement	38.396	37.670	78.825	37.745	37.800	78.364
other	37.394	37.340	1.075	37.600	37.102	1.054

the testing set are mainly distributed in *color*, *counting*, *location*, *judgement* and *other* classes, the other three question types account for a small percentage, the sum is less than 1%. Note that *judgement* class holds 78.825% and 78.364% of the VQA v1.0 testing set and VQA v2.0 testing set, respectively.

The fusion of question type is formulated as follows:

$$f = ReLU(W_{m'}m'^{T}) \odot ReLU(W_u u^T), \quad (20)$$

$$f' = concat(f, c), \quad (21)$$

$$concat(f, c) = [f, c]^T, \quad (22)$$

where $c \in R^{e\times 1}$ $(e = 8)$ indicates the one-hot encoding of the corresponding question type, $W_{m'} \in R^{2d_h\times 2d_h}$ and $W_u \in R^{2d_h\times 2d_h}$ are the learnable parameters. We directly concatenate the one-hot encoding of the question type c to the multi-modal joint representation $f \in R^{2d_h\times 1}$ in order to get $f' \in R^{1\times(2d_h+e)}$ for later answer generation. In addition to the one-hot encoding of the question type, we also have tried to map the question type to the same dimension as the question, but the experimental results are not satisfactory.

4) PREDICTION

In this paper, we treat VQA as a multi-label regression task and use a multi-layer perceptron (MLP) to perform this task. A set of candidate answers is pre-determined from all the correct answers in the training set that appear more than 8 times. The joint representation f' is first fed into a non-linear layer and then mapped via a linear layer. Finally, the sigmoid function is employed to predict the score for each candidate answer:

$$\hat{s} = sigmoid(WReLU(W_{f'}f'^{T})), \quad (23)$$

where $W \in R^{d\times 2d_h}$, $W_{f'} \in R^{2d_h\times(2d_h+e)}$ are the classifier parameters, d indicates the number of candidate answers, $\hat{s}$ represents the probability of the answer prediction.

B. LEARNING

1) LOSS FUNCTION

Our loss function is similar to the binary cross-entropy loss while using soft accuracies. The objective function is:

$$L_{vqa} = -\sum_{i=1}^{M}\sum_{j=1}^{N} s_{ij}log\hat{s}_{ij} + (1 - s_{ij})log(1 - \hat{s}_{ij}), \quad (24)$$

where M, N refer to the number of training questions and candidate answers, respectively. The ground-truth scores s are the soft accuracies of ground-truth answers computed in (25).

2) OPTIMIZATION

We choose Adamax as the optimizer. Adamax is a variant of Adam that provides a simpler range of upper learning rates. Compared with Stochastic Gradient Descent (SGD), Adamax does not need to manually adjust the learning rate and has faster convergence. We use weight normalization to accelerate the training. More specifically, gradient clipping technology and dropout ($ratio = 0.5$) are exploited in training.

V. EXPERIMENTS

In order to validate our proposed model, we carry out experiments to answer the following questions:

- **RQ1:** How does our designed approach perform when compared with other benchmark methods?
- **RQ2:** Can the self-attention on the question contribute to the overall effectiveness of CAQT?
- **RQ3:** Is the question type helpful for boosting the performance of CAQT?

A. DATASET

We validate the CAQT model on both balanced and unbalanced version of VQA dataset, namely, the VQA v1.0 dataset [1] and VQA v2.0 dataset [46]. We train our model on the training and validation sets of the VQA datasets and then report the test results on the test-dev and test-standard sets.

1) VQA V1.0

VQA v1.0 dataset [1] consists of 204,721 images from the MSCOCO dataset [48]. There have 248,349 training questions, 121,512 validation questions, 60,864 developing test questions, and 244,302 standard test questions. The questions in VQA1.0 can be divided into three sub-categories: *Yes/No*, *Number* and *Other*. There are three questions for per image and each question exists ten ground-truth answers from ten different annotators. Besides, VQA v1.0 includes two tasks: Open-Ended task and Multiple-Choice task (18 answers choices per question).

2) VQA V2.0

VQA v2.0 dataset [46] consists of 443,757 questions for training, 214,354 questions for validation and

447,793 questions for testing. There has only Open-Ended task. The VQA v2.0 dataset is more balanced as compared with the VQA v1.0 dataset. Specifically, for every question, there are two similar images which have two different answers to the question. We evaluate the CAQT model on the challenging Open-Ended task of both datasets.

B. EVALUATION METRICS

Since each question in the datasets is answered by ten different annotators, the answers sometimes are not the same, especially for ambiguous or subjective questions. In order to explore the inconsistency between answers, we adopt soft accuracies as the regression targets. We report the soft accuracy as:

$$Acc(a) = \frac{1}{K}\sum_{k=1}^{K} min(\frac{\sum_{1\leqslant j\leqslant K, j\neq k}\mathbb{I}(a=a_j)}{3}, 1), \quad (25)$$

where $a_1, a_2, a_3, \ldots, a_K$ are correct answers provided by the different annotators, a is the predicted answer and $K = 10$. $\mathbb{I}$ is an indication function.

C. PARAMETER SETTINGS

For extracting visual object features, we use Faster R-CNN [35] to obtain top 36 ($K = 36$) object regions and each region is represented by 2048 dimensional features. For sentence encoding, a pre-trained GloVe word embedding of dimension (300) is utilized [49]. And, Bi-LSTM is used to encode question, the dimension of word feature vector in each question is 2048 ($d_h = 1024$). For computational efficiency, we limit the length of each question to 14 words ($n = 14$). We set $d_a = 100$, $r = 10$ (according to [29]). The batch size is set to 512, and the epoch is set as 30. All experiments are fulfilled with PyTorch toolbox.

D. COMPARED METHODS

1) COMPARISON ON THE VQA V1.0 DATASET

We compare CAQT with existing state-of-the-art methods on the VQA v1.0 dataset as follows:

- LSTM Q+I [1] uses a two-layer LSTM to get question embedding and uses VGGNet to get image features, then fuses them via element-wise multiplication.
- DPPnet [41] solves the VQA task by learning a CNN with a dynamic parameter layer where the weights are determined adaptively based on questions.
- FDA [42] provides better-aligned image content representation with questions. FDA can find the important words and critical regions in the question and image, respectively. The question feature and image feature are fused via LSTM units.
- DMN+ [43] proposes a model with memory and attention mechanism. The input fusion layer allows interactions between input facts, and a novel attention based GRU allows for logical reasoning over ordered inputs.
- SMem [44] stores neuron activations from different spatial regions of the image in its memory, and uses the question to choose relevant regions for computing the answer.
- SAN [25] uses the semantic representation of question as the query to search for the regions in an image that is related to the answer. It queries an image multiple times to infer the answer progressively.
- MRN [45] uses shortcuts and residual mappings for multimodality, so MRN allows a deeper network structure and can effectively learn the joint representation from vision and language information.
- MCB [7] proposes to utilize Multimodal Compact Bilinear pooling to efficiently and expressively combine multi-modal features.
- MLB [15] proposes low-rank bilinear pooling using Hadamard product for an efficient attention mechanism of multi-modal learning.
- HieCoAtt [8] presents a hierarchical co-attention model for visual question answering. Co-attention model jointly reasons about image attention and question attention.
- MAN [18] exploits memory-augmented neural networks to predict accurate answers for visual questions, the memory network incorporates both internal and external memory blocks and selectively pays attention to each training exemplar.
- DAN [9] attends to specific regions in the image and key words in the question via multiple steps and gathers essential information from both modalities to infer answers.
- MFB [16] develops a Multi-modal Factorized Bilinear pooling approach to combine multi-modal features, which leads to superior performance for the VQA task compared with other bilinear pooling approaches.

2) COMPARISON ON THE VQA V2.0 DATASET

We compare CAQT with existing state-of-the-art methods on the VQA v2.0 dataset as follows:

- VQA team-Prior [46] predicts the most common answer in the training set, for all test questions.
- VQA team-Language only [46] has a similar architecture with LSTM Q+I [1] except that it only accepts the question as input and does not utilize any visual information.
- VQA team-LSTM+CNN [46] uses LSTM to get question embedding and uses CNN to get image features. It combines these two features via element-wise multiplication, and is followed by a MLP classifier to predict a probability distribution over candidate answers.
- MAN [18], MCB [7] and MLB [15] are consistent with the comparison methods in VQA v1.0 dataset.
- Up-Down [37] combines bottom-up and top-down attention mechanisms which enable attention to be calculated at the level of objects and other salient image regions. The Up-Down model employs the visual feature from Faster R-CNN with ResNet-101 and extracts the final

TABLE 2. Results comparison of various models on VQA v1.0 test-dev set and test-standard set. "-" indicates that the result is not available. For the test-dev set and the test-standard set, the best results are bolded.

Model	test-dev				test-standard			
	Overall	Other	Number	Yes/No	Overall	Other	Number	Yes/No
LSTM Q+I [1]	53.74	36.42	35.24	78.94	54.10	36.80	35.60	79.00
DPPnet [41]	57.22	41.69	37.24	80.71	57.36	42.24	36.92	80.28
FDA [42]	59.24	45.77	36.16	81.14	59.54	-	-	-
DMN+ [43]	60.30	48.30	36.80	80.50	60.40	-	-	-
SMem [44]	57.99	43.12	37.32	80.87	58.24	43.48	37.53	80.80
SAN [25]	58.70	46.10	36.60	79.30	58.90	-	-	-
MRN [45]	61.68	49.25	38.82	82.28	61.84	49.41	38.23	82.39
MCB [7]	64.70	55.60	37.60	82.50	-	-	-	-
MLB [15]	65.08	54.87	38.21	**84.14**	65.07	54.77	37.90	**84.02**
HieCoAtt [8]	61.80	51.70	38.70	79.70	62.06	51.95	38.22	79.95
MAN [18]	63.80	54.00	39.00	81.50	64.10	54.70	37.60	81.70
DAN [9]	64.30	53.90	39.10	83.00	64.20	54.00	38.10	82.80
MFB [16]	65.90	56.20	39.80	84.00	65.80	56.30	38.90	83.80
CAQT(ours)	**66.37**	**57.98**	**42.02**	82.63	**66.53**	**58.05**	**41.15**	82.88

state of GRU as the question feature, and then combines these two features via element-wise multiplication.

- MF-SIG-T3* [47] indicates 2-glimpse model by concatenating a Mean Field attention with a sigmoid attention. It is trained on external datasets and involves three reasoning processes.

3) VARIANT MODELS

To demonstrate the effectiveness of self-attention on the question, and question type, we design and compare three variant methods. These variant models are trained and evaluated on the training and validation sets of the two public datasets, respectively. Specifically, the variant models are as follows:

- Baseline: The baseline model first extracts the image feature from Faster R-CNN and uses the output of every hidden state of Bi-LSTM as the question word feature. Then executes question-guided visual attention. Finally, image and text features are fused and fed into the classifier to get the answer.
- Baseline+Self-Attention (Baseline+SelfAtt): Based on the Baseline model, Baseline+SelfAtt performs self-attention on the question, and then uses the new question expression to execute visual attention.
- Baseline+question type (Baseline+QType): Compared with the Baseline model, Baseline+QType concatenates the one-hot encoding of the question type to the multimodal joint representation directly, then feeds these features into the classifier.

E. EXPERIMENTAL RESULTS AND DISCUSSIONS

1) PERFORMANCE COMPARISON (RQ1)

We first list the performance of our method and then compare it with previous published competing approaches. Tab. 2 shows the experimental results on test-dev and test-standard of VQA v1.0 dataset for the Open-Ended task. Tab. 2 is divided into three categories: 1) approaches without using attention mechanism; 2) methods only based on visual attention; 3) approaches integrating both visual attention and textual attention.

In Tab. 2, we have the following observations: CAQT obtains the best *Overall* accuracies on test-dev (66.37%) and test-standard (66.53%). Our method belongs to the third category approach (approach based on visual and textual attention), with significant improvements over the best approach MFB by 0.47% on test-dev and 0.73% on test-standard. We consider that CAQT is superior to MFB probably because we use a matrix to represent questions, but MFB uses a vector, which may lead to information loss. Besides, MFB uses bilinear pooling to fuse image and text features, which actually results in expensive time and space complexity. Comparing with the first category (methods without attention), CAQT outperforms the best approach DMN+, especially with an increase of 6.13% in terms of test-standard *Overall* accuracy. In addition, compared with the second category approaches involving visual attention, our method is still better, with a large margin on the test-dev and test-standard sets.

We further demonstrate the performances of CAQT on the VQA v2.0 dataset with both test-dev and test-standard sets. Tab. 3 summarizes the comparisons of our approach with the results of advanced methods. From the results in Tab. 3, we can see that CAQT achieves 65.46% and 65.80% *Overall* accuracies across all question domains of test-dev set and test-standard set, respectively. CAQT outperforms the state-of-the-art approach (MF-SIG-T3) by 0.73% on *Overall* accuracy of test-dev set, even though the MF-SIG-T3 model was trained via VQA v2.0 dataset and an external dataset (Visual Genome [14]). Furthermore, the improvements of CAQT can be seen in all of the entries (*Other* with 0.79%, *Number* with 0.93%, *Yes/No* with 0.62% on test-dev set, *Other* with 0.24%, *Number* with 0.04%, *Yes/No* with 0.04% on test-standard set), especially for the *Number* accuracy on the test-dev set. The results in Tab. 2 and Tab. 3 explicitly show the advantages of CAQT.

2) EFFECTIVENESS OF SELF-ATTENTION MECHANISM (RQ2)

We summarize the performance of Baseline+SelfAtt on VQA v1.0 and VQA v2.0 validation sets in Tab. 4 and Tab. 5, respectively. When a model does not execute self-attention on

TABLE 3. Results comparison of various models on VQA v2.0 test-dev set and test-standard set. "-" indicates that the result is not available. "*": trained with external datasets. For the test-dev set and the test-standard set, the best results are bolded.

Model	test-dev				test-standard			
	Overall	Other	Number	Yes/No	Overall	Other	Number	Yes/No
VQA-team-Prior [46]	-	-	-	-	25.98	01.17	00.36	61.20
VQA team-Language only [46]	-	-	-	-	44.26	27.37	31.55	67.01
VQA team-LSTM+CNN [46]	-	-	-	-	54.22	41.83	35.18	73.46
MAN [18]	-	-	-	-	62.10	52.60	39.50	79.20
MCB [7] reported in [46]	-	-	-	-	62.27	53.36	38.28	78.82
MLB [15]	-	-	-	-	62.54	52.95	38.64	79.85
Up-Down [37]	-	-	-	-	65.67	56.26	43.90	82.20
MF-SIG-T3* [47]	64.73	55.55	42.99	81.29	-	-	-	-
CAQT(ours)	**65.46**	**56.34**	**43.92**	**81.91**	**65.80**	**56.50**	**43.94**	**82.24**

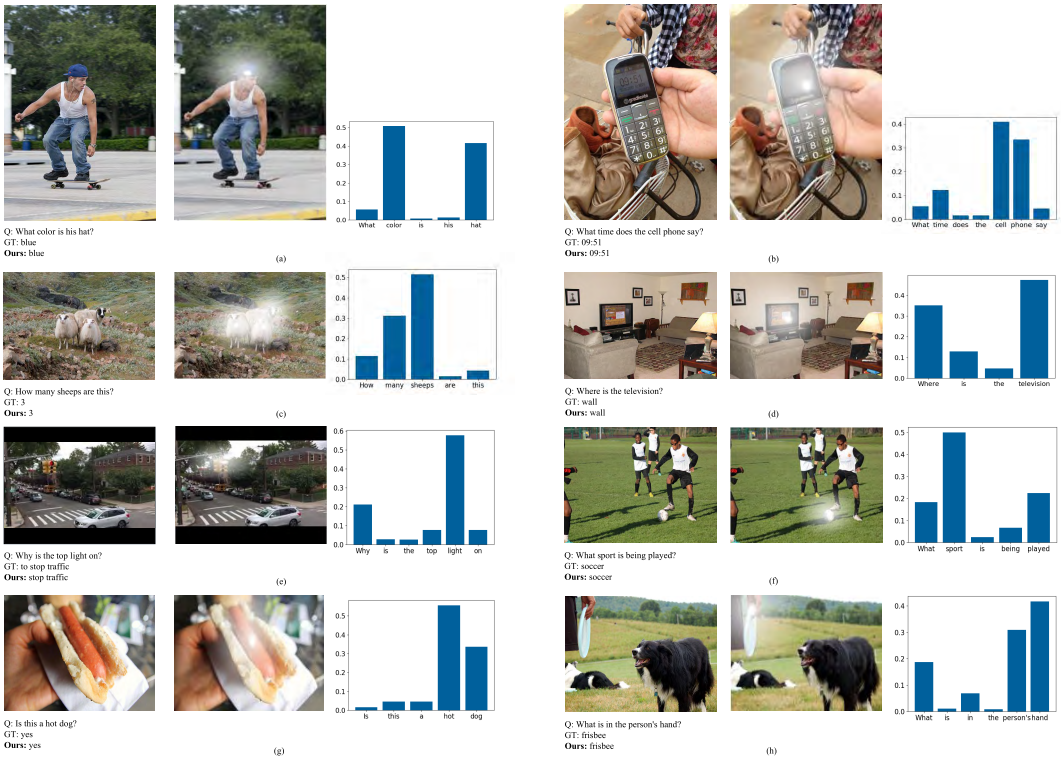

FIGURE 3. Visualization of examples. Original image (left), the output of the visual attention (middle) and the histogram of textual attention weights distribution (right).

TABLE 4. The results of variant models on VQA v1.0 validation set. Self-attention and question type are abbreviated as SelfAtt and QType,respectively. The best results are bolded.

Model	Overall	Other	Number	Yes/No
Baseline	63.40	54.14	44.27	81.49
Baseline+SelfAtt	63.60	54.23	44.19	81.93
Baseline+QType	63.51	54.15	44.15	81.82
CAQT (final)	**63.78**	**54.28**	**44.55**	**82.24**

TABLE 5. The results of variant models on VQA v2.0 validation set. Self-attention and question type are abbreviated as SelfAtt and QType, respectively. The best results are bolded.

Model	Overall	Other	Number	Yes/No
Baseline	63.58	54.63	47.86	80.31
Baseline+SelfAtt	63.95	54.80	**49.00**	80.74
Baseline+QType	63.63	54.68	48.39	80.25
CAQT (final)	**64.25**	**55.39**	48.03	**81.17**

the question, and not consider the question type, the *Overall* accuracy shows the worst performance. In Tab. 4, compared with Baseline, Baseline+SelfAtt has an *Overall* accuracy increased by 0.2%. In Tab. 5, Baseline+SelfAtt outperforms the Baseline model across all of the entries (*Overall* with 0.37%, *Other* with 0.17%, *Number* with 1.14%, *Yes/No* with 0.43%). Compared with Baseline, the improvement of experimental results fully demonstrates that self-attention plays a crucial role in improving the accuracy of the model.

3) VALIDATION OF QUESTION TYPE (RQ3)

Tab. 4 and Tab. 5 also give the performance of Baseline+QType on the validation sets. It can be seen from Tab. 4, Baseline+QType improves Baseline model from 63.40% to 63.51% on the *Overall* accuracy. In Tab. 5, Baseline+QType obtains 63.63% on the *Overall* accuracy, which is better than the Baseline model, especially on the entry *Number*. We believe that the reason for the high accuracy on the two datasets is that question type narrows down the answer space,

hence the model only need to search for the correct answers among the candidate answers of the corresponding type.

In Tab. 4, CAQT obtains the *Overall* accuracy of 63.78% on the VQA v1.0 validation set, outperforming baseline by 0.38%. In Tab. 5, when we add both self-attention and question type to the Baseline, the *Overall* accuracy outperforms Baseline model remarkably, with an increase of 0.67%. Furthermore, compared with Baseline+SelfAtt and Baseline+QType, CAQT could consistently achieve better performance on the two datasets, verifying that it is reasonable to jointly consider self-attention on the question, and question type.

4) CASE STUDY

In this section, we show some examples to demonstrate that visual attention can locate the regions that are relevant to the potential answers, and textual attention highlights the key words. Fig. 3 presents eight examples. They cover question types include *color*, *time*, *counting*, *location*, *reason*, *sport*, *judgement* and *other*. For each example, the exhibited three images from left to right are ordered as the original image, the output of the question-guided visual attention and the histogram of textual attention weights distribution. The white aperture area of the image is the detected attention. Across all those examples, we can see that the visual attention attends to the image regions that are related to the correct answer, and the text attention assigns a relatively large weight to the important words of the question. In Fig. 3(a), the question is ''What color is his hat?'', which asks the color of the hat of the man who stands on the skateboard. In the output of the visual attention layer, the attention is focused on the head of the man. As we can see from the histogram of textual attention weights distribution, ''color'' accounts for the largest proportion, followed by ''hat''. CAQT finally outputs the answer of the question: the hat's color is *blue*.

VI. CONCLUSION AND FUTURE WORK

In this paper, we present a novel deep neural network with the new co-attention mechanism and question type for the VQA task. Our proposed co-attention mechanism combines self-attention based textual attention and question-guided visual attention. We introduce the question type in CAQT model by directly concatenating it with the multi-modal joint representation to reduce the search space of candidate answers. The visualization shows that the CAQT model has the ability to pay more attention to the relevant textual and visual clues that are highly relevant to the answer of the question. Experiments conducted on VQA v1.0 and v2.0 datasets show that CAQT outperforms state-of-the-art approaches.

In the future work, we plan to extend our work in the following two directions: 1) We will strive to explore more complex interactions between textual and visual features; 2) We will design an end-to-end model to generate an explanation for the predicted answer, because an explanation can make the question answering process more understandable and traceable.

REFERENCES

[1] S. Antol *et al.*, ''VQA: Visual question answering,'' in *Proc. ICCV*, Dec. 2015, pp. 2425–2433.
[2] X. Chen *et al.* (Apr. 2015). ''Microsoft COCO captions: Data collection and evaluation server.'' [Online]. Available: https://arxiv.org/abs/1504.00325
[3] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, ''Deep image retrieval: Learning global representations for image search,'' in *Proc. ECCV*, 2016, pp. 241–257.
[4] A. Gordo and D. Larlus, ''Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval,'' in *Proc. CVPR*, Jul. 2017, pp. 5272–5281.
[5] W. S. Lasecki, Y. Zhong, and J. P. Bigham, ''Increasing the bandwidth of crowdsourced visual question answering to better support blind users,'' in *Proc. ASSETS*, 2014, pp. 263–264.
[6] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu, ''Joint video and text parsing for understanding events and answering queries,'' *IEEE Multimedia*, vol. 21, no. 2, pp. 42–70, Apr./Jun. 2014.
[7] 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. EMNLP*, 2016, pp. 457–468.
[8] J. Lu, J. Yang, D. Batra, and D. Parikh, ''Hierarchical question-image co-attention for visual question answering,'' in *Proc. NIPS*, 2016, pp. 289–297.
[9] H. Nam, J.-W. Ha, and J. Kim, ''Dual attention networks for multimodal reasoning and matching,'' in *Proc. CVPR*, Jul. 2017, pp. 2156–2164.
[10] M. Malinowski and M. Fritz, ''A multi-world approach to question answering about real-world scenes based on uncertain input,'' in *Proc. NIPS*, 2014, pp. 1682–1690.
[11] M. Ren, R. Kiros, and R. Zemel, ''Exploring models and data for image question answering,'' in *Proc. NIPS*, 2015, pp. 2953–2961.
[12] 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,'' in *Proc. NIPS*, 2015, pp. 2296–2304.
[13] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, ''Visual7W: Grounded question answering in images,'' in *Proc. CVPR*, Jun. 2016, pp. 4995–5004.
[14] R. Krishna *et al.*, ''Visual genome: Connecting language and vision using crowdsourced dense image annotations,'' *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.
[15] J.-H. Kim, K.-W. On, W. Lim, J.-W. Ha, and B.-T. Zhang, ''Hadamard product for low-rank bilinear pooling,'' in *Proc. ICLR*, Apr. 2017, pp. 1–14.
[16] Z. Yu, J. Yu, J. Fan, and D. Tao, ''Multi-modal factorized bilinear pooling with co-attention learning for visual question answering,'' in *Proc. ICCV*, Oct. 2017, pp. 1839–1848.
[17] M. Lao, Y. Guo, H. Wang, and X. Zhang, ''Multimodal local perception bilinear pooling for visual question answering,'' *IEEE Access*, vol. 6, pp. 57923–57932, 2018.
[18] C. Ma *et al.*, ''Visual question answering with memory-augmented networks,'' in *Proc. CVPR*, Jun. 2017, pp. 6975–6984.
[19] C. Yang, H. Zhang, B. Jiang, and K. Li, ''Aspect-based Sentiment Analysis with Alternating Coattention Networks,'' *Inf. Process. Manage.*, vol. 56, pp. 463–478, May 2019.
[20] M. Corbetta and G. L. Shulman, ''Control of goal-directed and stimulus-driven attention in the brain,'' *Nature Rev. Neurosci.*, vol. 3, no. 3, pp. 201–215, Mar. 2002.
[21] D. Bahdanau, K. Cho, and Y. Bengio. (Sep. 2014). ''Neural machine translation by jointly learning to align and translate.'' [Online]. Available: https://arxiv.org/abs/1409.0473
[22] K. Xu *et al.*, ''Show, attend and tell: Neural image caption generation with visual attention,'' in *Proc. ICML*, 2015, pp. 2048–2057.
[23] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia. (Nov. 2015). ''ABC-CNN: An attention based convolutional neural network for visual question answering.'' [Online]. Available: https://arxiv.org/abs/1511.05960
[24] K. J. Shih, S. Singh, and D. Hoiem, ''Where to look: Focus regions for visual question answering,'' in *Proc. CVPR*, Jun. 2016, pp. 4613–4621.
[25] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, ''Stacked attention networks for image question answering,'' in *Proc. CVPR*, Jun. 2015, pp. 21–29.
[26] M. Lao, Y. Guo, H. Wang, and X. Zhang, ''Cross-modal multistep fusion network with co-attention for visual question answering,'' *IEEE Access*, vol. 6, pp. 31516–31524, 2018.
[27] I. Schwartz, A. Schwing, and T. Hazan, ''High-order attention models for visual question answering,'' in *Proc. NIPS*, 2017, pp. 3664–3674.

[28] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.
[29] Z. Lin *et al.* (Mar. 2017). "A structured self-attentive sentence embedding." [Online]. Available: https://arxiv.org/abs/1703.03130
[30] Z. Tan, M. Wang, J. Xie, Y. Chen, and X. Shi, "Deep semantic role labeling with self-attention," in *Proc. AAAI*, 2017, pp. 4929–4936.
[31] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. CVPR*, Jun. 2018, pp. 7794–7803.
[32] J. Fu *et al.* (Sep. 2018). "Dual attention network for scene segmentation." [Online]. Available: https://arxiv.org/abs/1809.02983
[33] Y. Shi, T. Furlanello, S. Zha, and A. Anandkumar, "Question type guided attention in visual question answering," in *Proc. ECCV*, Sep. 2018, pp. 151–166.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.
[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.
[36] K. Kafle and C. Kanan, "An analysis of visual question answering algorithms," in *Proc. ICCV*, Oct. 2017, pp. 1983–1991.
[37] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. CVPR*, Jun. 2018, pp. 6077–6086.
[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[39] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1724–1734.
[40] D. Teney, P. Anderson, X. He, and A. van den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proc. CVPR*, Jun. 2018, pp. 4223–4232.
[41] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. CVPR*, Jun. 2015, pp. 30–38.
[42] I. Ilievski, S. Yan, and J. Feng. (Apr. 2016). "A focused dynamic attention model for visual question answering." [Online]. Available: https://arxiv.org/abs/1604.01485
[43] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proc. ICML*, Jun. 2016, pp. 2397–2406.
[44] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. ECCV*, 2016, pp. 451–466.
[45] J.-H. Kim *et al.*, "Multimodal residual learning for visual QA," in *Proc. NIPS*, 2016, pp. 361–369.
[46] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. CVPR*, Jul. 2016, pp. 6325–6334.
[47] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma, "Structured attentions for visual question answering," in *Proc. ICCV*, Oct. 2017, pp. 1300–1309.
[48] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.
[49] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

CHAO YANG received the B.E. and M.E. degrees in computer science from Hunan University, Changsha, China, in 1999 and 2005, respectively, and the Ph.D. degree in computational intelligence and systems science from the Tokyo Institute of Technology, Tokyo, Japan, in 2010.

She was a Postdoctoral Fellow with the Tokyo Institute of Technology. Since 2016, she has been an Associate Professor with the College of Computer Science and Electronic Engineering, Hunan University. Her research interests include data mining, machine learning, intelligent systems, recommender systems, and deep learning. She is a member ACM and CCF.

MENGQI JIANG received the bachelor's degree in computer science and technology from the Hunan University of Chinese Medicine, in 2017. She is currently pursuing the master's degree with the College of Computer Science and Electronic Engineering, Hunan University, China. Her research interests include visual question answering and information extraction.

BIN JIANG received the B.A. degree in mathematics and the M.E. degree in soft engineering from Hunan University, Changsha, China, in 1993 and 2006, respectively, and the Ph.D. degree in computational intelligence and systems science from the Tokyo Institute of Technology, Tokyo, Japan, in 2015. He is currently an Associate Professor with the College of Computer Science and Electronic Engineering, Hunan University. His research interests include big data technology, artificial intelligence, machine learning, data mining, intelligent computing, recommender systems, and social computing. He is a member ACM and CCF.

WEIXIN ZHOU received the bachelor's degree in computer science and technology from the China University of Mining and Technology, in 2017. He is currently pursuing the master's degree with the College of Computer Science and Electronic Engineering, Hunan University, China. His research interests include deep learning, recommender systems, and natural language processing.

KEQIN LI is currently a Distinguished Professor of computer science with the State University of New York. He has published over 620 journal articles, book chapters, and refereed conference papers. His current research interests include cloud computing, fog computing and mobile edge computing, energy-efficient computing and communication, embedded systems and cyber-physical systems, heterogeneous computing systems, big data computing, high-performance computing, CPU-GPU hybrid and cooperative computing, computer architectures and systems, computer networking, machine learning, and intelligent and soft computing. He has received several best paper awards. He currently serves or has served on the editorial boards of the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON CLOUD COMPUTING, the IEEE TRANSACTIONS ON SERVICES COMPUTING, and the IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING.