 doi:10.20944/preprints201804.0313.v1

Cross-Modal Multistep Fusion Network with Co-Attention for Visual Question Answering

MINGRUI LAO[1], YANMING GUO[1], HUI WANG[1], and XIN ZHANG[1]
[1]College of Systems Engineering, National University of Defense Technology, Changsha 410073, China

Corresponding author: Yanming Guo (e-mail: guoyanming@nudt.edu.cn).

ABSTRACT Visual question answering (VQA) is receiving increasing attention from researchers in both the computer vision and natural language processing fields. There are two key components in the VQA task: feature extraction and multi-modal fusion. For feature extraction, we introduce a novel co-attention scheme by combining Sentence-guide Word Attention (SWA) and Question-guide Image Attention (QIA) in a unified framework. To be specific, the textual attention SWA relies on the semantics of the whole question sentence to calculate contributions of different question words for text representation. For the multi-modal fusion, we propose a "Cross-modal Multistep Fusion (CMF)" network to generate multistep features and achieve multiple interactions for two modalities, rather than focusing on modeling complex interactions between two modals like most current feature fusion methods. To avoid the linear increase of the computational cost, we share the parameters for each step in the CMF. Extensive experiments demonstrate that the proposed method can achieve competitive or better performance than the state-of-the-art.

INDEX TERMS visual question answering, cross-modal multistep fusion network, attention mechanism

I. INTRODUCTION

Deep learning has attracted significant attention in recent years [1]-[3] and has brought revolutionary advances in the field of computer vision [4]-[6] and natural language processing [7]-[9]. These advances heightened researchers' confidence for addressing more complex tasks that combine computer vision with language and high-level reasoning. Under such trend, visual question answering (VQA) [10]-[12], as a prime cross-modal example, is receiving increasing attention.

VQA is an "AI-complete" task that needs to answer text-based questions though analyzing the images. It is generally considered to be more challenging than other cross-modal tasks like image captioning [13], [14] or image-text retrieval [15], since it is expected to understand the images/texts logically and reason the answering under different conditions. There are two key components for this task: the feature extraction and the multi-modal fusion.

The feature extraction acts as the base component for the VQA task, and provides the basic elements for the following operations. As a typical cross-modal task, it is essential to have a good alignment of the image and the text, thus most state-of-the-art approaches would contain the attention module. The main idea of the 'attention' is to dynamically adapt the salient features according to the given information, instead of utilizing the static representations throughout [16]-[18].

The early VQA models tend to adopt visual attention to impose regularization on learning models to find the most relevant image regions based on the corresponding quetions. Recently, co-attention methods become popular as it not only consider the visual attention like most existing attention methods, but also explicitly take into account the attention in natural language. By reducing the effect of unimportant textual information, co-attention methods can effectively get richer multimodal representations. The textual attention in common co-attention frameworks [19], [20] is to obtain question attention based on visual features, in the sense that the image representation is used to guide the question attention and the question representation are used to guide image attention in such co-attention frameworks.

However, it is intuitive for human to pay attention to the important textual information by comprehending the overall meaning of the whole question sentence, instead of the corresponding image content. In this paper, we propose a novel textual attention named Sentence-Guide Word Attention (SWA) which relies on the semantic of whole question sentence to calculate the contributions of different question words. And then, a new co-attention framework is proposed by combining SWA and visual attention together for multimodal feature representations in VQA task.

Another important component in VQA is the multi-modal feature fusion. It needs to encode the information from

different modals in a high level, through which to match the correct answer. A common research direction for multi-modal feature fusion is to improve the efficiency of multimodal fusion, and establish sufficient interactions between features from two modals. In contrast to linear pooling methods (such as element-wise addition or concatenation) which just consider simple correlations between multi-modal features, bilinear pooling has become pivot in the research of information fusion due to the second-order interaction between two modals. However, directly using bilinear pooling method would result in sharp increase of the learning parameters and computation resources. Many approaches have been raised to deal with this problem, such as Multimodal Compact Bilinear pooling (MCB) [21], Multimodal Low-rank Bilinear pooling (MLB) [22], Multimodal Factorized Bilinear pooling (MFB) [23] and Multimodal Tucker Fusion (MUTAN) [24].

An alternative solution for the feature fusion is to produce more multi-modal features and establish multiple interactions. This approach has been ignored by the public, possibly owing to the concern that the parameters would increase linearly with the fusion step. In this paper, motivated by the recurrent residual learning architecture in [25], we propose a Cross-modal Multistep Fusion (CMF) network to make up the defects brought by multiple interactions. By introducing fully connected layer with shared residual learning module, CMF can produce multiple multi-modal features without linearly increasing the learning parameters. Compared with bilinear pooling methods, our CMF approach focuses on generating more multi-modal features and fuses two features from different modals in each step , rather than calculating the complex interaction between two features using a large number of computing resources and learning parameters.

The main contributions of our work can be summarized in four aspects:

- We introduce a novel co-attention learning architecture for VQA. For the attention in natural language, a Sentence-guide Word Attention (SWA) is proposed to find important words according to the whole semantic of question sentence. And then, we adopt the attentional word feature to extract visual attention. Our co-attention method can effectively reduce the irrelevant features and obtain richer features for visual and textual representations.
- We describe a new multi-modal fusion architecture CMF which aims to achieve multiple interactions without increasing the learning parameters linearly by the mechanism of shared parameters.. Experimental results demonstrate that it can obtain better performance.
- Detailed visualizations of the attention results validate that our model can effectively focus on important image regions and words for the VQA task.
- We perform the ablation studies to quantify the roles of different components in our model. Experimental results on the widely-used VQA dataset [10] demonstrate that, our proposed method achieve competitive or superior performance over the state-of-the-art.

The remainder of this paper is organized as follows: Section II reviews approaches related to the learning frameworks of visual question answering. Section III describes the details of our proposed co-attention learning method and Cross-modal Multistep Fusion (CMF) network. In section IV, we conduct experiments to evaluate our proposed model on VQA dataset. We also perform ablation studies to quantify the roles of different components in our model. Finally, section V concludes the paper and puts forward future possible directions.

II. RELATED WORKS

Visual question answering is an emerging research task in the crossover of computer vision and natural language processing, whose purpose is to answer a question about an image. It has the potential applications for the visual impaired and the automatic text querying of image collections or surveillance videos.

There are three fundamental components for the general VQA framework: multi-modal feature extraction, multi-modal feature fusion, and answer prediction based on the fusion feature.

A. FEATURE EXTRACTION AND REPRESENTATION

With regard to the process of feature extraction, most approaches use recurrent neural networks [26], typically implemented with LSTM [27] and GRU [28], to extract textual features. As for the visual features, convolutional neural networks [29] are used to obtain region features from image, among which VGG-net [30] and deep residual networks [31] are most common. Recently, Anderson *et al.* [32] used the object features extracted by Faster R-CNN [33] and achieved state-of-the-art results.

In order to get better alignments for image and question sentence, adding the attention mechanisms into learning frameworks is popular. Therefore, many approaches began to introduce visual attention mechanisms [34] into this task to automatically obtain local fine-grained features for the given questions. One classical attention method in this task is "question-guided visual attention" which aims to calculate the contribution of image regions by sentence features. Zhu et al. [35] introduced structure attentions which model the visual attention as a multivariate distribution over a grid-structured Conditional Random Field on image regions, thereby effectively encoding cross-region relations to visual representations. Yang *et al.* [36] regarded visual attention as a process of multi-step reasoning, and proposed a novel stacked attention framework to focus on different regions of the images. Shih *et al.* [37] extracted bounding-box features in picture and scored each box based on the given question. Additionally, there are still some approaches employed co-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 April 2018 doi:10.20944/preprints201804.0313.v1

attention frameworks to improve visual and textual features simultaneously. Lu *et al.* [19] presented a hierarchical co-attention model that jointly reasons about image and question attention. Nam *et al.* [20] proposed Dual Attention Network that attend to special regions in images and words in text through multiple steps and gather essential information from both modalities. Compared with these methods, our co-attention framework combine SWA and Question-Guide Image Attention (QIA) together for multimodal representation. To be specific, the textual attention SWA we proposed aims to guide word attention by the question sentence feature, rather than using image feature.

B. MULTIMODAL FEATURE FUSION

As for fusion strategies, it is intuitive to use linear pooling approach to achieve multi-modal feature fusion. Although these approaches, such as element-wise addition and concatenation, are easy to implement, they may not achieve richer interactions between two modals. With the increasingly widespread use of bilinear pooling method, especially the usages in fine-grained classification [38] and multimodal language modeling [39], this method has been a research that focus on VQA task to get bilinear interactions for deep feature representations. Due to the fact that the out-product operation in bilinear pooling would bring the increase of learning parameters, these are numerous improved bilinear pooling methods proposed to deal with such problem. Fukui et al. [21] introduced Multimodal Compact Bilinear (MCB) pooling based on compression method to combine multimodal features. Motivated by matrix decomposition approach, Kim et al. [22] proposed a Multimodal Low-rank Bilinear (MLB) pooling method and largely decrease learn parameters. Yu et al. [23] refined this MLB method that expand features from two modals to a high dimension space and adapt sum pooling to squeeze high-dimension feature into the compact output features. They introduced this fusion approach named Multimodal Factorized Bilinear pooling (MFB). Recently, a Tucker decomposition framework proposed by Ben-younes *et al.* [24] and achieve state-of-the-art results compare with other fusion methods.

Compared with fusion methods above, our CMF network focuses on performing multiple fusions by generating various multimodal features, instead of achieving a complex interaction between different modals. Liu *et al.* [25] proposed Recurrent Residual Fusion Network (RRF) for multimodal matching. In contrast to their method, there are two main differences: (1) About motivations, RRF aims to use recurrent residual frameworks to recursively improve feature embedding for multimodal matching task, rather than providing richer multimodal features like ours. (2) With regard to learning architecture, our method fuses multimodal features in every step, instead of just using the feature in the final step.

C. ANSWER PREDICTION

For answer prediction, it is noteworthy that different people may use diverse answers or expressions for one exact question, though these answers may have similar meanings. Most visual question answering frameworks adopt answering sampling to randomly pick an answer from a set of candidates, which may ignore sophisticated correlations of the candidate answers. In this paper, we use Kullback-Leibler divergence (KLD) as loss function like [23] to regard answer prediction as a process of label distribution learning.

III. Method

Our overall framework is illustrated in Figure 1. It can be divided into three parts: the first part (in blue color) is the Multimodal Representation. To increase the consistency of the image and text, we propose a novel co-attention architecture, which contains Sentence-guide Word Attention (SWA) and Question-guide Visual Attention (QVA). The section part (in yellow color) denotes the Multimodal Fusion, in which we introduce a multiple fusion module called "Cross-modal Multistep Fusion (CMF)". To be specific, the CMF consists of several cascading CMF units with shared learning parameters. Therefore, it can increase the depth of the network with few storage costs. The third part (in pink color) is the Answer Prediction. In this paper, we employ the Kullback-Leibler Divergence (KLD) loss to transform the multi-class classification problem with sampled answers to the label distribution learning problem with a fixed answer distribution.

In the following, we would describe the three components in the overall framework: Co-Attention module, Cross-modal Multistep Fusion (CMF) network, and answer prediction.

A. CO-ATTENTION MODULE

Attention mechanism in multimodal representation mainly aims to help learning frameworks to understand what image regions or objects are important, and provide better multimodal feature representations to solve question answering problems. Currently, most attention approaches only consider the attention in visual representations, but ignore the fact that there is still some noisy information in textual information, such as the explanatory and colloquial words. A good question attention may result in better image attention, since it provides better textual representations to guide image attention. In order to decrease the influence from unimportant words, we propose a novel Sentence-guide Word Attention (SWA) which aims to score each word by the semantic of question sentence.

The architecture of our co-attention method is shown in Figure 2. We describe two major components in this module: Sentence-guide Word Attention (SWA) and Question-guide Image Attention (QIA).

For the SWA model, we transform the attention score of word into selection expectation, and the expectation p_i of word i can be defined as follow:

$$p_i = softmax\left(conv\left(mlb(w_i, s)\right)\right) \quad (1)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 April 2018 doi:10.20944/preprints201804.0313.v1

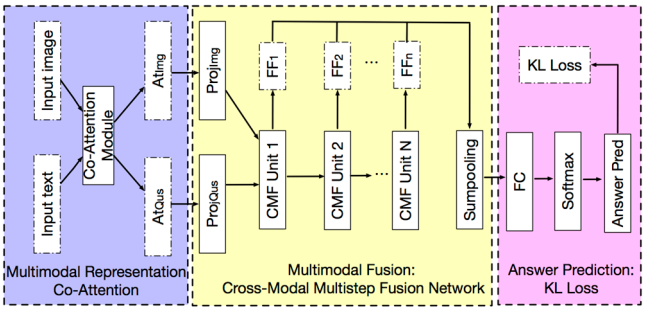


FIGURE 1. **Our learning framework for image question answering. It is comprised of three parts: multimodal representation, multimodal fusion and answer prediction. "At_{Qus}" and "At_{Img}" indicate attentional question feature and attentional image feature from our co-attention framework. "FC" means the fully connected layer. "$Proj_{Qus}$" and "$Proj_{Img}$" imply the fully connected layer which is used to project "At_{Qus}" and "At_{Img}" into the features with same dimension respectively. "FF_N" represents the fusion feature from Cross-Modal Multistep Fusion unit in time step n.**

Where $s \in \mathbb{R}^{n_I}$ and $W = [w_1, \dots, w_N] \in \mathbb{R}^{n_Q \times N}$ are the sentence features from the last hidden state in LSTM and the word features from each time state in LSTM. mlb(·) represents the Multimodal Low-rank Bilinear pooling method and conv(·) is the convolution learning operation consisting of sequential 1×1 convolutional layers and ReLU layers shown in Figure 2.

Then the attentional word representations $\bar{w}$ can be calculated as follows:

$$\bar{w} = \sum_{i=1}^{N} p_i \times w_i \qquad (2)$$

Similar to unstructured visual attention in [32], we adopt the common Question-Guide Image Attention (QIA) measure in this module, and define the attentional image feature $\bar{x}$ as :

$$\bar{x} = \sum_{i=1}^{M} softmax(conv(mlb(x_i, \bar{w})) \times x_i \qquad (3)$$

Where $X = [x_1, \dots x_M] \in \mathbb{R}^{n_I \times M}$ is the image feature extracted by deep residual network (res5c feature maps in ResNet).

To summarize, we propose a novel Sentence-guide Word Attention (SWA) method to obtain word attention by the understanding question sentence, thereby providing richer textual representations to guide image attention. Combining textual and visual attention together, our co-attention framework demonstrates great improvement.

B. CROSS-MODAL MULTISTEP FUSION NETWORK

Multimodal fusion plays an important and fundamental role in image question answering. Most fusion methods focus on achieving complex interactions between two modals, thereby obtaining richer fusion features for answer prediction in single step. Bilinear pooling and related researches are popular, but these methods result in large consumption of learning parameters and computational resources. Motivated by the recurrent residual framework proposed by Liu et al. [39], we introduce a novel fusion framework named Cross-Modal Multistep Fusion (CMF) network for image question answering, thus implementing multistep fusion for feature interactions. The sketch map of CMF network is depicted in the middle part of Figure 1, in which each CMF unit can produce one fusion feature, and all the CMF units share the learning parameters. By obtaining fusion features from all time steps, our method combines them by sum-pooling approach and then gets the ultimate fusion feature for answer prediction.

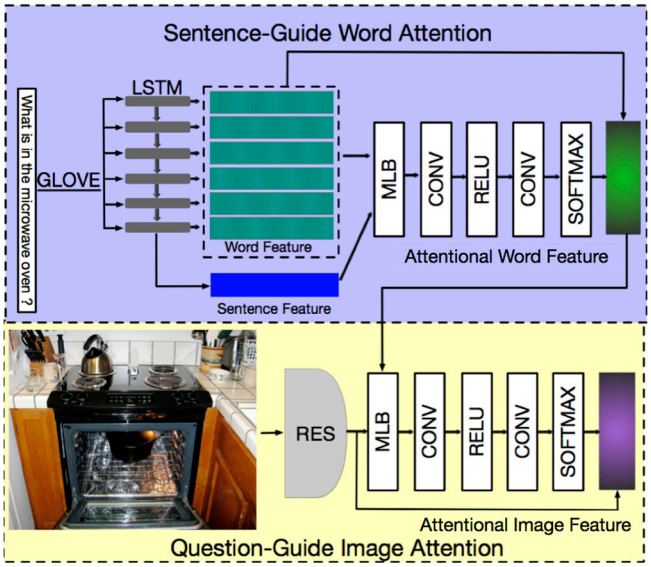


FIGURE 2. **Our co-attention architecture for image question answering which aims to generate attentional image feature and question feature. The image and question are firstly represented as the fine-grained features respectively. GLOVE and RES implies GloVe word embedding model and Residual Learning network.**

Firstly, attentional features from two modals are projected into the two features with the same dimension by the project neural network layers "Proj_{Qus}" and "Proj_{Img}". And then they are fed into the first unit in the CMF network. Finally, the fusion features from all CMF units will be get together by sum-pooling to get the final feature for answer prediction. The detail of CMF unit is illustrated in Figure 3.

There are two inputs in a CMF unit: attentional image feature and attentional question feature from the last CMF unit. Each CMF unit has three output features, among which fusion feature is to provide multistep feature for sum-pooling operation and attentional feature maps from two modals will be regard as inputs for the next CMF unit.

As for the fusion feature, it is formed by Hadamard product between two modal features before being standardized by power normalization and l2 normalization. The fusion feature in step N can be formulated as:

$$FF_n = Norm(IF_{n-1} \circ QF_{n-1}) \qquad (4)$$

Where Norm(·)refers to the power normalization and l2 normalizations. IF_{n-1} and QF_{n-1} imply the attentional image feature and question feature provided by the CMF unit in n-1 step. ∘ is the Hadamard product operation. As for the first unit, the fusion feature is as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 April 2018 doi:10.20944/preprints201804.0313.v1

$$FF_1 = Norm(U^T\bar{x} \circ V^T\bar{q}) \quad (5)$$

Where $\bar{x}$ and $\bar{q}$ refer to the initial attentional feature produced by co-attention module in last part. Two matrix U and V are the fully connected layers ("$Proj_{Qus}$" and "$Proj_{Img}$" shown in Figure 1) to make two modal feature have the same dimension.

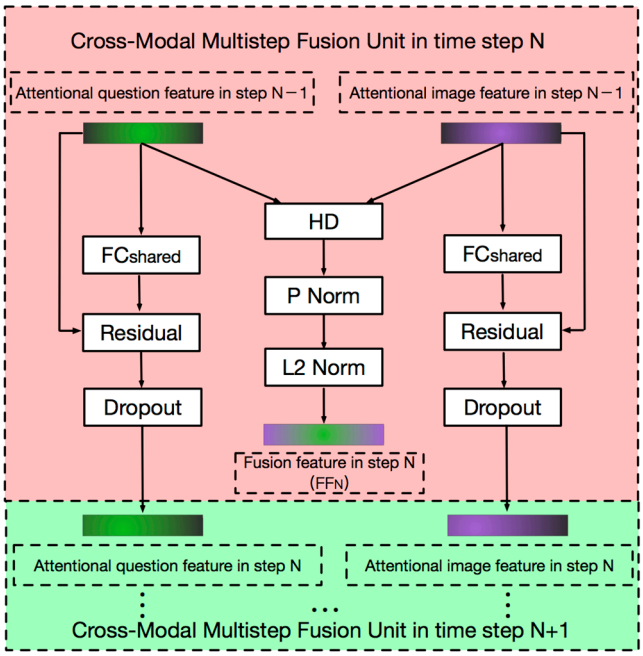


FIGURE 3. Cross-modal Multistep Fusion Unit in CMF network. "FC_{shared}" is the fully connected layer shared by two modal features. "P Norm" and "L2 Norm" imply Power Normalization and L2 Normalization. "HD" is the Hadamard product operation between two features. "Residual" means the residual learning after fully connected layers.

As for the recurrent image feature IF_n and question feature QF_n provided in step n, they can be calculated as follows:

$$IF_n = Drop(Relu(S^T \cdot IF_{n-1} + IF_{n-1})) \quad (6)$$

$$QF_n = Drop(Relu(S^T \cdot QF_{n-1} + QF_{n-1})) \quad (7)$$

Where S^T refers to the shared fully connected layer (shared parameters in each step). $Drop(\cdot)$ and $Relu(\cdot)$ indicate dropout and ReLU activation function.

Finally, the ultimate fusion feature F_{sum} can be computed as a summation across the feature channels:

$$FF_{sum} = \sum_{n=0}^{T} FF_n \quad (8)$$

Where is the number of CMF units in this method.

In summary, compared to single fusion method, our CMF network focuses on achieving multiple interaction between two modal features. Benefited form the sharing mechanism of CMF, CMF network can produce mutlstep fusion feature and the consumption of learning parameters and computation resources will not grow linear along with the time step increasing.

C. ANSWER PREDICTION

Most VQA models use the sampling method to determine the only answer for each question, and transform the answer prediction into the multi-class classification problem. But this method may ignore the complex relations among different answers. In this paper, similar to the MFB framework [21], we use the KL-divergence loss as the loss function, and regard the process of answer prediction as the label distribution learning problem. Accordingly, the loss function $L(p,g)_{KL}$ can be defined as follows:

$$L(p,g)_{KL} = \sum_i p_i \log\left(\frac{p_i}{g_i}\right) \quad (9)$$

Where p and q is the distribution of prediction and ground truth respectively, and i represents the number of answer candidates.

IV. EXPERIMENTS

A. DATASETS AND EVALUATIOB METRICS

We evaluate our proposed model on the commonly used VQA dataset[10]. This dataset is generally considered as a large-scale dataset for the visual question answering task. It contains large amount of training instances and various question types. In VQA dataset, the images could be divided into two types: real images from the MS-COCO dataset and abstract scene images. We select the real images and the corresponding text information as the experimental dataset. The dataset consists of 248.349 training questions, 121,512 validation questions and 244,302 testing questions. Additionally, there is a 25% test subset called *test-dev*. There are three question types including *yes/no*, *number* and *other*. For each question, 10 free-response answers are provided. There are two tasks provided to evaluate performance: Open-Ended (OE) and Multiple-Choices (MC). In this paper, we focus on Open-Ended task where the ground truth answers are given in free natural language phrase.

For the Open-Ended task, we use a voting mechanism to score the accuracy of a predicted answer:

$$Accuracy(a) = \min\left(\frac{Count(a)}{3}, 1\right) \quad (10)$$

Where $Count(a)$ is the count of the answer a voted by different annotators.

B. EXPERIMENT SETTING

We describe our experimental settings here. For the question model, we use the external pre-trained word embedding model with GloVe [40], and the dimension of word vector is 300.A dropout layer with 0.3 ratio is added after each LSTM layer. The number of the answers is 3000, and the maximal length of the questions is 15. For the image model, we extract the visual features by using the 152-layer ResNet model [28] which is pre-trained in the ImageNet dataset [41]. Images are resized to 448×448, and 2048-D $pool5$ features with L2 normalization are used for visual representation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 April 2018 doi:10.20944/preprints201804.0313.v1

For the co-attention model, the number of attention glimpse in both Sentence-guide Word Attention and Question-guide Visual Attention is 2. The joint embedding size of MLB operation is 1200. For the CMF network, the dimension of multimodal feature is 1200. The dropout ratio after the residual leaning is set to 0.1.

With regard to the training parameters, we adopt the Adam solver with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and the initial learning rate is set to 0.007. The batch size is 100, and the number of iterations is fixed to 100K. In the ablation study, we train our model on the training set, and evaluate each component on the validation set. For the comparison with the state-of-the-art, we follow other comparative approaches to train the model on the training set and the validation set.

All experiments are fulfilled with PyTorch toolbox, and performed on the workstations with NVIDIA GTX-1080 Ti and Titan Xp GPUs.

C. ABLASION ANALYSIS

To demonstrate the improvements and effectiveness of our co attention and CMF network, we ablate the full model to analyze each component in our models.

- Baseline: The baseline model employs the spatial image feature from ResNet and extract the last hidden state of LSTM as the question feature. Then two modal features are fused with the MLB method.
- Baseline + Question Sentence-Guide Word Attention (SenAtt): Compared to the baseline model, SenAtt model uses the output of last hidden state of LSTM as the sentence features to guide the visual attentions for image representation compared to the Baseline model.
- Baseline + Question Word-Guide Word Attention (WordAtt): Different with the SenAtt model, the WordAtt model utilizes the features from each timestep in LSTM as the word feature without attention mechanism to guide the visual attentions.
- Baseline + Co-Attention (CoAtt): In CoAtt model, we employ the co-attention framework we proposed to obtain attention weight from two modals, rather than just visual mode in SenAtt and WordAtt
- Baseline + Co-Attention + CMF (CoAtt+CMF): Based on CoAtt model, we replace the simple multimodal fusion MLB with our CMF network with different numbers of CMF units. Additionally, CoAtt+CMF model with one unit (N=1) is same as CoAtt model.

We report the performance of our ablation models on val set and test-dev set of VQA dataset in Table 2 and Table 3.

1) EVALUATION OF CO-ATTENTION FRAMEWORK

In Table 1, we summarized the baseline model and the models with different attention mechanisms. It is obvious that the models with attention outperform the baseline model remarkably, with an average of 3-4 percent increase. This demonstrates the necessity of utilizing attention in this task. In addition, compared with SenAtt and WordAtt, CoAtt could consistently achieve better performance on the Validation and Test-dev datasets, verifying that it is reasonable to jointly consider the textual attention and the visual attention. Owing to the superiority of the CoAtt, we built and evaluated the remaining components on top of the CoAtt model.

TABLE 1. Results of our Co-Attention model compared with other attention model and baseline on val and test-dev datasets.

Model	Val	Test-dev
Baseline	57.52	59.51
SenAtt	60.47	62.78
WordAtt	61.18	63.09
CoAtt	**61.79**	**63.56**

2) EVALUATION OF CMF NETWORK

TABLE 2. Results of our CMF network with different number of CMF units on val and test-dev datasets..

CoAtt+CMF	Val	Test-dev
N=1	61.79	63.56
N=2	62.02	63.97
N=3	**62.69**	**64.79**
N=4	62.55	64.64
N=5	62.29	64.58

From Figure 3, it can be seen that we can build our CMF network with any depth. In Table 2, we show the results of our CMF network with N=1,2,3,4,5. Note that the CoAtt+CMF with N=1 is the same as the CoAtt model, whose fusion method is MLB. Compared with the CoAtt model, all four CoAtt-CMF networks achieved better performance. This verifies the effectiveness of imposing the CMF block in the Co-Att framework. More in detail, the results when N=3 are superior to others. The drop of performance for N=4 and N=5 may be due to the potential overfitting in the model.

E. COMPARISON WITH STATE-OF-THE-ART

In this part, we compare our model with the state-of-the-art on the VQA dataset, and the results are shown in Table 3 and Table 4.Table 3 shows the results of the models without attention mechanism and Table 4 illustrates the counterparts with attention framework. The experiment of our CMF and Coatt+CMF model are performed with N=3.

As can be seen in Table 3, our CMF baseline outperforms all other existing VQA approaches without the attention mechanism. The major improvements are from yes-or-no (Y/N) and number (Num) type answers. Compared with some competitive approaches such as MCB and MFB, which aim to achieve complex interactions between multimodal features, our CMF method focuses on achieving multiple multimodal interaction by generating multistep feature. In addition, our CMF method just employs the fundamental fusion approach

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 April 2018 doi:10.20944/preprints201804.0313.v1

"Hadamard product" to achieve multimodal interaction. It is potential to combine our CMF network with some fusion approaches with complex interaction.

TABLE 3. Results of CMF network and other VQA method without attention mechanism on test-dev set.

Model	Y/N	Num	Other	All
iBOWIMG [42]	76.5	35.0	42.6	55.7
DPPnet [43]	80.7	37.2	41.7	57.2
VQA-team [10]	80.5	36.8	43.1	57.8
AYN [44]	78.4	36.4	46.3	58.4
AMA [45]	81.0	38.4	45.2	59.2
MCB [19]	81.7	36.9	40.0	61.1
MRN [46]	82.3	38.9	49.3	61.7
MFB [21]	81.8	36.7	51.2	62.2
CMF(ours)	**82.8**	**38.4**	**50.4**	**62.4**

Firstly, with the attention mechanism in Table 4, the performances of the models with attention mechanism obtain great improvement in contrast to models without attention. It demotrates that attention mechanism is a major source to boost the performance of VQA. By introducing co-attention learning, the CMF+CoAtt model we proposed delivers an improvement of 4 points compared the CMF model above. Compared to competitive VQA method, our CMF+CoAtt method achieves state-of-art performance in VQA task.

TABLE 4. Results of CMF+ CoAtt method and other VQA method with attention mechanism on test-dev set.

Model	Y/N	Num	Other	All
SMem [12]	80.9	37.3	43.1	58.0
NMN [47]	81.2	38.0	44.0	58.6
SAN [33]	79.3	36.6	46.1	58.7
FDA [48]	81.1	36.2	45.8	59.2
DNMN [49]	81.1	38.6	45.4	59.4
HieCoAtt [35]	79.7	38.7	51.7	61.8
RAU [50]	81.9	39.0	53.0	63.3
MCB+Att [19]	82.2	37.7	54.9	64.2
DAN [36]	83.0	39.1	53.9	64.3
MLB+Att [20]	84.1	38.2	54.9	65.1
MFB+CoAtt [21]	84.0	39.8	56.2	65.9
MF-SIG-T3 [32]	84.3	39.3	56.4	66.0
CMF+CoAtt(ours)	**84.3**	**40.3**	**57.0**	**66.4**

F. QUALITATIVE RESULTS

In this section, we present some examples to visualize the visual/textual attention achieved by our co-attention framework. These examples cover a broad range about the answer types, including the objects, jugement, color and action. For the visual attention, the attention probability distribution is of size 14×14 and the original image is 448×448, we up-sample the attention probability distribution and apply a Gaussian filter to make it the same size as the original image. For the textual attention, the size of attention probabilility is 15×1.

It can be seen that the learned question and image attention are usually closely focus on the key words and the most relevant image regions. For example, consider the question "What is behind the zebra doing ?" which asks the action of the zebra shown in Figure 4(4). In the output of the word attention larer, the model focus on the word "zebra" and "doing" to better understand the intention of the question sentence. In order to find out what the zebra is doing, the visual attention map attends the zebras's legs in the corresponding picture.

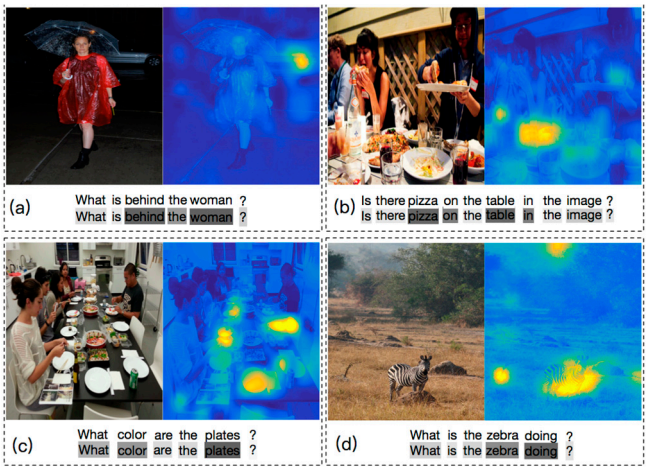


FIGURE 4. Qualitative results of our CMF+CoAtt model on the VQA dataset with attention visualization. For each example, the raw image and the image with visual attention are presented from left to right. The raw question and question with textual attention are shown from top to bottom. The brightness of image and darkness of words represent their attention weights.

V. CONCLUSION

In this work, we propose a co-attention framework and a CMF network for the visual question answering task. For the co-attention framework, we propose a novel SWA textual attention to score each word with the semantic of whole question sentence. For the CMF network, we propose to share the parameters of each CMF unit, and thus we can obtain multiple fusion features without introducing too much computational cost. In the future, we would strive to combine the complex fusion method with multiple fusion network.

REFERENCES

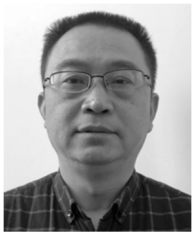
[1] Y. Lecun, Y. Bengio and G. Hinton, "Deep learning.," *Nature*, vol. 521,no. 7553, pp. 436-444, May 2015.
[2] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding," *Neurocomputing*, vol. 187, pp. 27-48, Apr 2016.
[3] A. Hassan, A. Mahmood, "Convolutonal Recurrent Deep Learning Model for Sentence Classification," *IEEE Access*, vol. 6, pp. 13949-13957, Mar 2018.
[4] T. Y. Lin, A. Roychowdhury and S. Maji, "Bilinear Convolutional Neural Networks for Fine-grained Visual Recognition.," *IEEE T Pattern Anal*, Jul 2017, doi: 10.1109/TPAMI.2017.2723400
[5] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," in *Proc. ECCV,* 2016, pp. 499-515.
[6] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Semantic Image Segmentation via Deep Parsing Network," in *Proc. ICCV,* 2016, pp. 1377-1385.

[7] Y. L. Ji and F. Dernoncourt, "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks," in *Proc. NAACL* 2016, pp. 515-520.
[8] W. Yin, S. Ebert and H. Schütze, "Attention-Based Convolutional Neural Network for Machine Comprehension," in *Proc. NAACL* 2016, pp. 15-21.
[9] W. Li and H. Chen, "Identifying Top Sellers In Underground Economy Using Deep Learning-Based Sentiment Analysis," in *Proc. ISI* 2014, pp. 64-67.
[10] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," *Int J Comput Vison*, vol. 123, pp. 4-31, 2017.
[11] D. Teney, Q. Wu and A. V. D. Hengel, "Visual Question Answering: A Tutorial," *IEEE Signal Proc Mag*, vol. 34, no. 6, pp. 63-75, 2017.
[12] 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.
[13] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge," *IEEE T Pattern Anal*, vol. 39, no. 4,pp. 652-663, 2016.
[14] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning," in *Proc. CVPR* 2016, pp. 375-383.
[15] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-Modal Retrieval With CNN Visual Features: A New Baseline," *IEEE T Cybernetics*, vol. 47, no. 2, pp. 449-460, 2017.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," 2017, [online] Available: https://arxiv.org/abs/1706.03762.
[17] L. C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to Scale: Scale-Aware Semantic Image Segmentation," in *Proc. CVPR* 2016, pp. 3640-3649.
[18] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," in *Proc. NAACL* 2016, pp. 1480-1489.
[19] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical Question-Image Co-Attention for Visual Question Answering," in *Proc. NIPS* 2016.
[20] H. Nam, J. W. Ha and J. Kim, "Dual Attention Networks for Multimodal Reasoning and Matching," in *Proc. CVPR* 2016, pp. 299-307.
[21] 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.
[22] J. H. Kim, K. W. On, W. Lim, J. Kim, J. W. Ha, and B. T. Zhang, "Hadamard Product for Low-rank Bilinear Pooling," in *Proc. ICLR* 2017.
[23] 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* 2017, pp. 1821-1830.
[24] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal Tucker Fusion for Visual Question Answering," in *Proc. ICCV* 2017, pp. 2631-2639.
[25] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a Recurrent Residual Fusion Network for Multimodal Matching," in *Proc. ICCV* 2017, pp. 4127-4136.
[26] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký , and S. Khudanpur, "Recurrent neural network based language model," in *Proc. ISCA* 2010, pp. 1045-1048.
[27] R. Dey and F. M. Salem, "Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks," 2017, [online] Available: https://arxiv.org/abs/1701.05923.
[28] M. Sundermeyer, R. Schlüter and H. Ney, "LSTM Neural Networks for Language Modeling," in *Proc. Interspeech* 2012, pp. 601-608.
[29] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional Neural Networks for MATLAB," in *Proc. ACM MM* 2015, pp. 689-692.
[30] L. Wang, S. Guo, W. Huang, and Y. Qiao, "Places205-VGGNet Models for Scene Recognition," 2015, [online] Available: https://arxiv.org/abs/1508.01667v1.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. CVPR* 2016.
[32] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," in *Proc. CVPR* 2018.
[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE T Pattern Anal*, vol. 39, no. 6, pp. 1137-1149, 2017.
[34] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,", 2015, [online] Available: https://arxiv.org/abs/1502.03044v1.
[35] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma, "Structured Attentions for Visual Question Answering," in *Proc. ICCV* 2017, pp. 1291-1300.
[36] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked Attention Networks for Image Question Answering," in *Proc. CVPR* 2016, pp. 21-29.
[37] K. J. Shih, S. Singh and D. Hoiem, "Where to Look: Focus Regions for Visual Question Answering," in *Proc. CVPR* 2016, pp. 4613-4621.
[38] T. Y. Lin, A. Roychowdhury and S. Maji, "Bilinear CNN Models for Fine-Grained Visual Recognition," in *Proc. ICCV* 2016, pp. 1449-1457.
[39] R. Kiros, R. Salakhutdinov and R. Zemel, "Multimodal neural language models," in *Proc. ICML* 2014, pp. II-595.
[40] J. Pennington, R. Socher and C. Manning, "Glove: Global Vectors for Word Representation," in *Proc. EMNLP* 2014, pp. 1532-1543.
[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "ImageNet Large Scale Visual Recognition Challenge," *Int J Comput Vision*, vol. 115, no. 3, pp. 211-252, 2015.
[42] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple Baseline for Visual Question Answering," 2015, [online] Available: https://arxiv.org/abs/1512.02167.
[43] H. Noh, P. H. Seo and B. Han, "Image Question Answering Using Convolutional Neural Network with Dynamic Parameter Prediction," in *Proc. CVPR* 2016, pp. 30-38.
[44] M. Malinowski, M. Rohrbach and M. Fritz, "Ask Your Neurons: A Neural-based Approach to Answering Questions about Images," in *Proc. ICCV* 2015.
[45] Q. Wu, P. Wang, C. Shen, A. Dick, and A. V. D. Hengel, "Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge from External Sources," in *Proc. CVPR* 2016, pp. 4622-4630.
[46] J. H. Kim, S. W. Lee, D. H. Kwak, M. O. Heo, J. Kim, J. W. Ha, and B. T. Zhang, "Multimodal Residual Learning for Visual QA," in *Proc. NIPS* 2016.
[47] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, " Neural Module Networks," 2015, [online] Available: https://arxiv.org/abs/1511.02799.
[48] I. Ilievski, S. Yan and J. Feng, "A Focused Dynamic Attention Model for Visual Question Answering," 2016, [online] Available: https://arxiv.org/abs/1604.01485.
[49] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to Compose Neural Networks for Question Answering," 2016, [online] Available: https://arxiv.org/abs/1601.01705.
[50] H. Noh and B. Han, "Training Recurrent Answering Units with Joint Loss Minimization for VQA," 2016, [online] Available: https://arxiv.org/abs/1606.03647.

 doi:10.20944/preprints201804.0313.v1

MINGRUI LAO received the B.S. degree in information engineering from Xi'an Jiao Tong University, Xi'an, China, in 2017. He is currently persuing the Ph.D degree with the College of Systems engineering in the National University of Defense Technology, Changsha, China. After finishing the B.S. degree, he started to focus on deep learning theories and its applications in multimodal information processing. His research focuses on the multimodal information comprehension, such as visual question answering, and machine reading comprehension.

YANMING GUO is now a lecturer in the College of System Engineering, National University of Defense Technology. He received his B.S. and M.S. degrees from the National University of Defense Technology, in 2011 and 2013, respectively, and the Ph.D. degree in Leiden Institute of Advanced Computer Science (LIACS), Leiden University, in 2017. His current interests include computer vision, natural language processing and deep learning. He has served as reviewers of some journals and conferences, such as TNNLS, TMM, Neurocomputing, MTAP and ICPR.



HUI WANG received his B.S, M.S. and Ph.D. degrees in system engineering from National University of Defense Technology (China), in 1990, 1998 and 2005, respectively. He is currently a professor with the State Key Lab of Information System Engineering, College of Systems Engineering, National University of Defense Technology. His research interests include cross-modal data mining, information extraction, and event analysis.

Xin Zhang received his B.S and Ph.D. degrees in system engineering from National University of Defense Technology (China), in 2000 and 2006, respectively. He is currently a professor with the State Key Lab of Information System Engineering, College of Systems Engineering, National University of Defense Technology. His research interests include cross-modal data mining, information extraction, and event analysis.