Advances in Intelligent Systems Research (AISR), volume 151
2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018)

Multimodal Cross-guided Attention Networks for Visual Question Answering

Haibin Liu[1], Shengrong Gong[2], Yi Ji[1], Jianyu Yang[3] , Tengfei Xing[1] and Chunping Liu[1,*]
[1]School of Computer Science and Technology, Soochow University Suzhou, Jiangsu, China
[2]School of Computer Science and Engineering, Changshu Institute of Technology, Changshu, Jiangsu, China
[3]School of Rail Transportation, Soochow University Suzhou, Jiangsu, China
[*]Corresponding author

***Abstract*—Visual Question Answering (VQA) is an attractive topic combining computer vision with natural language processing. It is more challenging than text-based question answering because of its multimodal nature. The VQA reasoning process requires both effective semantic embedding and fine-grained visual comprehension. Existing approaches predominantly infer answers from visual spatial information, while neglecting important semantic information in questions and the guidance information between images and questions. To remedy this, we imitate the human mechanism of cross-reasoning about visual and textual information and propose a multimodal cross-guided attention network (MCAN) for VQA which employs a cross-guided joint learning strategy with a gated activation learning method, which can simultaneously capture both rich visual spatial information and significant semantic information. We evaluate the proposed model on two public datasets: VQA dataset and COCO-QA dataset. Extensive experiments show state-of-the-art performance on the datasets.**

Keywords—visual question answering; attention; cross-guided; gated activation

I. INTRODUCTION

VQA [1] is a multimodal joint learning task of AI-complete. Affected by the convergence of Computer Vision (CV) and Natural Language Processing (NLP), VQA has received a great deal of attention. A VQA system is designed to automatically answer natural language questions according to the content of a reference image. It is a challenging task which not only requires understanding the image contents and the question semantic information, but also requires exploiting an effective strategy to fuse the low-level image features with the high-level semantic features of question. A VQA system can enhance the human-computer interaction experience and bring convenience to people's work and life. It also has a variety of application prospects, such as visually-impaired assistant devices, blind navigation, image retrieval, and video surveillance.

Previous methods for VQA [1-3] utilize a pre-trained Convolution Neural Networks (CNN) to extract global image features as image representations and encode question via Recurrent Neural Networks (RNN), and finally comply them with a simple joint learning strategy to infer the answer. The results are impressive. However, these methods can't locate the visual fine-grained regions related to the question. Therefore, many approaches [4, 5] picked up spatial representation from CNN and introduced visual attention mechanism to attend to the fine-grained regions relevant to the question. The improvements are significant. However, Das et al. [6] holds that current VQA models using these image representations with visual attention do not seem to focus on the consistent regions as humans do. One possible reason is that current VQA models with attention search for the regions related to the question one by one. As a result, the whole image is separated into several isolated units. Furthermore, the latest VQA models employ a simple one-glimpse attention, only utilizing question-guided visual attention, and ignoring making good use of image guidance issues attention for question.

To address these problems, we propose a novel multimodal cross-guided attention networks (MCAN) that implements a co-attention mechanism and allows multi-step reasoning for VQA. The overview of MCAN is illustrated in FIGURE I. The main contributions of our work are three-fold. First, we propose a multi-layer cross-guided attention networks for VQA task which takes full advantage of multimodal cross-guided information. Second, in order to improve the expressiveness of joint learning features, a gated activation approach is introduced for inferring the answer. Third, we conduct extensive experiments on two public VQA datasets [1, 2], and achieve significant improvements over both one-glimpse and two-glimpse attention models.

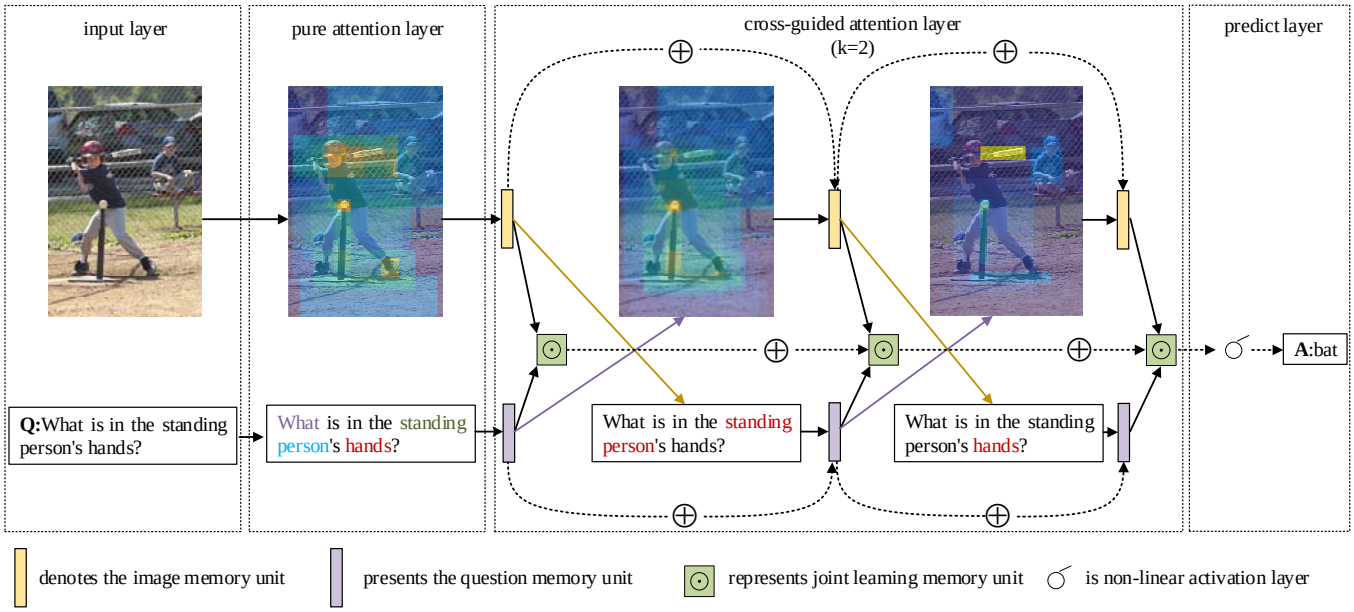

FIGURE I. OVERVIEW OF MCAN IN CASE OF 2 LAYERS . THE DIFFERENT COLORS OF BOUNDING BOXES IN IMAGE AND WORDS IN QUESTION INDICATE THE ATTENTION MAPS PREDICTED BY MCAN

II. RELATED WORK

A. Joint Embedding

Most recent works are based on CNN-RNN architecture. These models utilize CNN to extract semantic representations from images and encode questions via RNN, especially LSTM, and then combine two modalities with an appropriate joint learning method. Many previous methods [1-3, 7] adopt this approach, while some [5, 8] solve VQA task by modifying the basic idea. Besides LSTM, these approaches [3, 9-11] adopted GRU to extract high-level semantic and some [4, 12, 13] utilized CNN to encode question. There are several methods different from above ones, which addressed VQA task as a multi-way classification problem. In [7], the model fed both image and question into LSTM at each time step, and then generated the answer. Wu et al. [14] extracted attributes from image and generated descriptions of image as input of LSTM to generate answer by sequence-to-sequence learning.

B. Attention Mechanisms for VQA

Attention mechanism is widely used in VQA, which allows the model to selectively extract useful visual or textual information. Generally, models with visual attention mechanism pay attention to the significant regions in image and rule out the noise. A number of methods employ question-guided attention to solve VQA task. Yang et al. [4] introduced the soft attention, and proposed a stacked attention model which used question representations to query question-related regions in image via multi-step reasoning. Noh et al. [15] adopted visual attention with joint loss minimization. Xu et al. [16] obtained attention map by calculating the semantic similarity between image regions and the question. Ilievski et al. [17] used an off-the-shelf object detector to catch the important regions, and then fed the regions into LSTM with global image features. Fukui et al. [5] applied a convolutional operation on the concatenated textual representations and image representations to obtain the attention weights all over the regions. However, all of the above attention based methods are one-glimpse attention, just using the question to guide the spatial attention, and ignoring the image information as the important semantic guidance for question.

For better using both visual and textual information, Lu et al. [18] which propose an image-question co-attention mechanism that not only focuses on the relevant image regions but also attends to the important question words. Unlike [18], Nam et al. [19] calculated the textual and visual attention map by a refined multiplication operation. Wang et al. [20] extracted "facts" from image and proposed a novel co-attention approach to address VQA task. Yu et al. [11] trained a concept detector to extract concepts from image and utilized question representations to attend to the relevant regions and related concepts. Impressive performance has been reported by these approaches. However, all of the above proposed co-attention methods were parallel and did not make full use of multimodal interaction information. In this paper, we proposed a multi-layer cross-guided attention networks for VQA, which utilizes the multimodal information by cross-guided strategy adequately.

III. METHOD

In this section, we introduce the image model and the language model firstly and then detail proposed model.

A. Input Representations

Similar to [21], we employ a fine-tuned pre-trained Faster R-CNN in conjunction with the ResNet-101 CNN to encode image. To generate a bounding-box image features set **v** for VQA, we take the final output of the model and perform the non-maximum suppression for each object class with an intersection over union (IoU) threshold, and then select the top-ranked n bounding-boxes features as our image representations. Specially, for each selected bounding-box, we adopt the mean-pooled convolutional feature as the bounding-box feature whose dimension is

2048. Finally, we obtain the features vector $\mathbf{V} = [v_1, \dots, v_n] \in \mathbb{R}^{n \times 2048}$ as our image representations.

For better embedding question, we adopt Bi-GRU to get high-level question semantic. Given an one-hot question representation $q = [q_1, \dots, q_T]$, where q_t is one-hot embedding vector for t-th word, and T is length of question. We first embed the question into a semantic space by $x_t = W q_t$, W is a learned matrix. At each time step, we feed the word embedding feature vector x_t into Bi-GRU:

$$h_t^f = \mathrm{GRU^f}(x_t, h_{t-1}^f) \tag{1}$$

$$h_t^b = \mathrm{GRU^b}(x_t, h_{t+1}^b) \tag{2}$$

where h_t^f and h_t^b are the hidden states at time t for forward GRU and backward GRU respectively. At each time step, we concatenate the two hidden states for the t-th time step question representation:

$$u_t = f([h_t^f, h_t^b]) \tag{3}$$

For better capturing the high-level semantic information in question, proposed method stacks the concatenated hidden states sequentially. Finally, a set of feature vectors $\mathbf{U} = [u_1, \dots, u_T]$ is constructed for representing question, where u_t represents the semantic feature from the first to the t-th words.

B. Gated Activation Method

Inspired by the highway networks [22], we introduce gated activation method (GAM) using a gated hyperbolic tangent activation. It is a gated operations similar to recurrent units such as LSTM and GRU. Given a matrix or vector **x**, the output of GAM for **x** is defined as follows:

$$o_x = f_a(x) \tag{4}$$

expanded as:

$$t = \tanh(Wx + b) \tag{5}$$

$$s = \sigma(W'x + b) \tag{6}$$

$$o_x = t \circ s \tag{7}$$

where W, W' are the learned parameters, b, b' are the bias, σ is the sigmoid activation function, o_x is the output of GMA, and is $\circ$ element-wise multiplication. GMA can effectively enhance the expressive ability of the model by integrating different activation methods.

C. Attention Mechanism

Our proposed model performs a co-attention mechanism including visual and textual attention for extracting significant visual and textual information to infer answer. We perform the same calculation operation to obtain the visual and textual attention map.

Given the input feature matrix $\mathbf{V} = [v_1, \dots, v_n]$ and the attention guiding feature **u**, we first project **V** and **u** to a common space, and then perform GAM on concatenated features of the two modalities. A softmax is adopted to generate the attention distribution over different regions of **V**, and get the weighted feature:

$$\hat{v} = Attend(\mathbf{V}, \mathbf{u}) \tag{8}$$

formulated as:

$$\mathrm{h_p} = f_a([W_v V, W_u u + b_u]) \tag{9}$$

$$a_p = softmax(W_p h_p + b_p) \tag{10}$$

$$\hat{v} = \sum_i^n a_{p,i} v_i \tag{11}$$

where $\hat{v}$ is weighted sum, [.] is concatenate operation, W_v, W_u, b_u, b_p is learned parameters.

D. Multimodal Cross-guided Attention Networks

VQA is a multimodal joint learning problem which requires both the image information and the context semantic information of question. In many cases, answering a question needs a multi-step reasoning. In this work, MCAN set three memory units to maintain the states of image, question and multimodal joint learning features, respectively. The overall architecture of MCAN is illustrated in FIGURE II. Given image feature matrix $\mathbf{V} = [v_1, \dots, v_n]$ and question feature matrix $\mathbf{U} = [u_1, \dots, u_T]$, these units are recursively updated by:

$$m_v^k = m_v^{(k-1)} + \hat{v}^k \tag{12}$$

$$m_u^k = m_u^{(k-1)} + \hat{u}^k \tag{13}$$

$$m_j^k = m_j^{(k-1)} + m_v^k \odot m_u^k \tag{14}$$

where k is the number of cross-guided attention layers which is set to 1 at least, m_v^k,m_u^k, m_j^k are the visual, textual and joint learned memory states of k-th attention layers respectively,$\odot$ is element-wise product, $\hat{v}^k$and $\hat{u}^k$ are the visual and textual attention weighted sum based on Equation(8):

$$\hat{v}^k = Attend(V, m_u^{(k-1)}) \tag{15}$$

$$\hat{u}^k = Attend(U, m_v^{(k-1)}) \tag{16}$$

At each attention layer, the visual and textual weighted sum is generated by a cross-guided strategy that the image features are used to guide to generate textual attention and the question features are used for attending to the image regions. In many cases, question is complicated and a single cross-guided attention layer is not sufficient to extract meaningful information from image and question for predicting answer. Therefore,

MCAN recursively combine the cross-guided attention layer for k steps (see FIGURE II) where each attention layer could effectively locate the question-related image regions and image-related question words. We initial MCAN with a pure learning strategy, followed by:

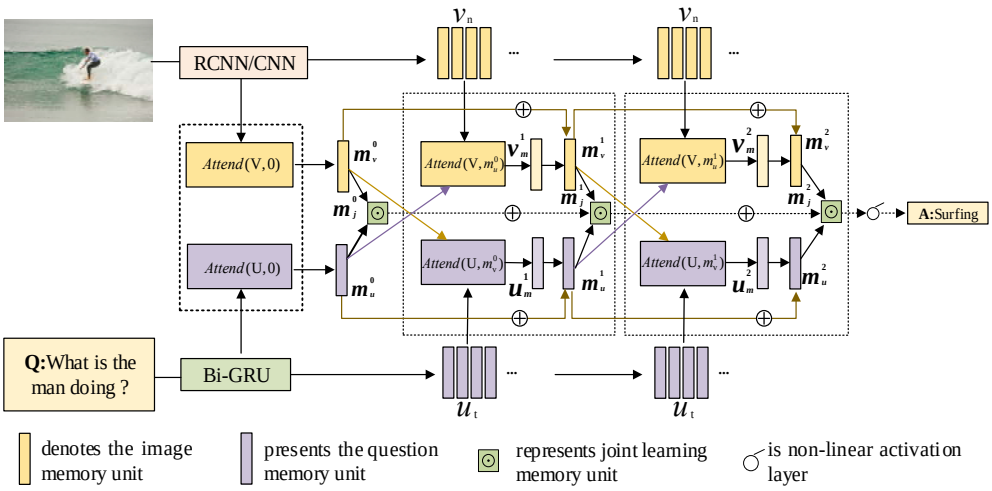

FIGURE II. OVERALL ARCHITECTURE OF MCAN IN CASE OF K=2

$$m_j^0 = m_v^0 \odot m_u^0 \tag{17}$$

$$m_v^0 = Attend(\boldsymbol{V}, 0) \tag{18}$$

$$m_u^0 = Attend(\boldsymbol{U}, 0) \tag{19}$$

where $\mathbf{0}$ represents that generating attention map without any guiding information. In this way, the model could capture the original significant information from image and question.

Similar to most VQA models, MCAN predicts the final answer by a multi-way classifier. We feed the last joint learning memory vector m_j^k into GMA, and then a single-layer softmax classifier with cross-entropy is used to predict the answer:

$$p_{ans} = softmax(W_{ans} f_a(m_j^k) + b_{ans}) \tag{20}$$

where p_{ans} represents the probability over the selected answers.

IV. EXPERIMENTS

A. Dataset and Evaluation Metrics

VQA dataset is one of the most widely used dataset for VQA task. It contains 82,783 images with 248,349 question-answer pairs for training, and 40,504 images with 121,512 question-answer pairs for validation. The size of the test set which involves 81,434 images with 244,302 question-answer pairs is equivalent to the training set. All the questions are divided into three categories: yes/no, number and other. The dataset raises two different tasks which are Multiple-Choice (MC) and Open-Ended (OE).

COCO-QA dataset is also a commonly used dataset. The training set in the dataset contains 78,736 samples and the testing set includes 38,948 samples. There are four types of questions in the dataset: object, number, color and location. The dataset contains 430 single-word answers for training the classifier.

We evaluate proposed model which formulates the VQA task as multi-class classification problem by the accuracy metric. For VQA dataset, the OE task is measured by a voting mechanism to calculate the accuracy and for MC task, the model only selects one of the candidate answers, and then compares with the ground truth answer. For COCO-QA dataset, the measure strategy contains both the accuracy and Wu-Palmer similarity (WUPS) [23]. WUPS is used to measure the correlation between words based on the taxonomy tree. Similar to [18], we use the threshold 0.0 and 0.9 for WUPS.

TABLE I. RESULTS OF OUR PROPOSED APPROACH AND COMPARED METHODS ON VQA DATASET, IN PERCENTAGE AND '-' REPRESENTS THE RESULT IS NOT AVAILABLE

Method	Test-dev					Test-standard				
	Open-Ended				MC	Open-Ended				MC
	y/n	num	other	all	all	y/n	num	other	All	all
VQA-team [1]	80.5	36.8	43.1	57.8	62.7	80.6	36.5	43.7	58.2	63.1
DPPnet[3]	80.7	37.2	41.7	57.2	62.5	80.3	36.9	42.2	57.4	62.7
DualNet[24]	82.0	37.9	49.2	61.5	66.7	81.9	37.8	49.7	61.7	66.7
SAN[4]	79.3	36.6	46.1	58.7	-	79.1	36.4	46.4	58.9	-
MCB(ResNet)[5]	82.2	37.7	54.8	64.2	68.6	-	-	-	-	-
HiecoAtt[18]	79.7	38.7	51.7	61.8	65.8	-	-	-	62.1	66.1
VQA-Machine[20]	81.5	38.4	53.0	63.1	67.7	81.4	38.2	53.2	63.3	67.8
MLAN[11]	82.9	39.2	52.8	63.7	68.9	-	-	-	-	-
MCAN(k=1)	82.7	**39.2**	53.2	63.8	69.9	82.7	38.3	53.3	63.9	69.9
MCAN(k=2)	**82.8**	39.1	**54.5**	**64.5**	**70.1**	82.8	**40.0**	54.1	**64.4**	70.1
MCAN(k=3)	82.6	36.4	54.4	64.0	69.7	**83.0**	36.6	**54.2**	64.2	**70.2**

B. Results and Analysis

TABLE I shows the performance of our approach on VQA dataset and comparison with state-of-the-art methods. We train proposed model on train+val datasets and test on test dataset. We adopt the VQA-team [1] as our baseline method. In the first three rows of TABLE I, we compare MCAN with baseline approaches without attention mechanism and MCAN has made a great improvement. The second two rows of Table 1 exhibit the performance of methods with one-glimpse attention mechanism. SAN [4] employs element-wise addition operation to calculate attention map and gets better performance than the methods in the first part. In MCAN, we replace the addition or the multiplication operation with concatenated operation to compute the joint learning attention map, and gain further improvements. MCB [5] gains state-of-the-art performance in VQA dataset that employs a feedforward CNN to calculate attention weights with high-dimensional joint features. MCAN reduce the dimension of joint features and gains better performance. As we can see in Table1, MCAN (k=2) obtains best performance on the whole. It improves MCB [5] from 68.6% to 70.1% for the Multiple-Choice task on Test-dev set. The third three rows of TABLE I shows the methods with co-attention mechanism which contains several joint attention mechanisms. HiecoAtt [18] uses textual attention cooperated with visual attention which gained better performance than most one-glimpse attention methods. MLAN [11] adopts semantic attention instead of textual attention and slightly im- proved over HiecoAtt [18]. Compared with HiecoAtt [18], and MLAN [11], MCAN (k=2) gains improvements by 4.3%, 1.2% for MC task on Test-dev set. It also improves MLAN [11] from 63.7 to 64.5 for OE task on Test-dev set. These improvements show the advantage of our proposed method with cross-guided attention. We set k=1, 2, 3 and train the different models. As we can see in TABLE I, MCAN (k=2) outperforms than the other models. It shows that the generalization capability of single layer cross-guided attention networks is insufficient, and the performance does not improve as the number of attention layers increased.

We further evaluate our proposed model on COCO-QA dataset and show the performance in TABLE II. We compare the performance by accuracy and WUPS with other methods. MCAN (k=2) improves the accuracy of state-of-the-art QRU [9] from 62.5% to 63.1%. In particular, our model achieves a 5% improvement for the question type of number. However, our model decreases the accuracy in the question types of Color and Location. The possible reason is the unbalanced number of each question type. MCAN also adopts WUPS to measure performance. Comparing with QRU [9], MCAN improves WUPS0.9 by 0.3%. Nevertheless, MCAN drops the performance by 0.3% in WUPS0.0. This may be due to the nature of WUPS evaluation method.

TABLE II. RESULTS OF OUR PROPOSED APPROACH AND COMPARED METHODS ON COCO-QA DATASET, IN PERCENTAGE AND '-' REPRESENTS THE RESULT IS NOT AVAILABLE.

Method	All	WUPS0.0	WUPS0.9
2-VIS-BiLSTM[2]	55.0	65.3	88.6
IMG-CNN[13]	58.4	68.5	89.7
DPPnet[24]	61.2	70.8	90.6
SAN[4]	61.6	71.6	90.9
QRU[9]	62.5	72.6	**91.6**
MCAN(k=2)	**63.1**	**72.9**	91.3

FIGURE III shows the qualitative results of proposed model on VQA test set, the different colors mean the different attention probability generated by MCAN. (a) are the input image and question, (b) are the visualization of $Attend(\mathbf{V}, 0)$, and (c), (d) are the visualization of MCAN(k=2) at different attention layers for $Attend(\mathbf{V}, m_u)$ and $Attend(\mathbf{V}, m_v)$. The first three rows are correct examples, and the last row is the error example. The results of visualization for attention maps indicate proposed model could not only attend to the fine-grained image regions but also significant words in question effectively. For the error example, as we can see, MCAN attends to the correct regions, but the correct answer "watch for children" is not in the classification labels result in predicting error answer.

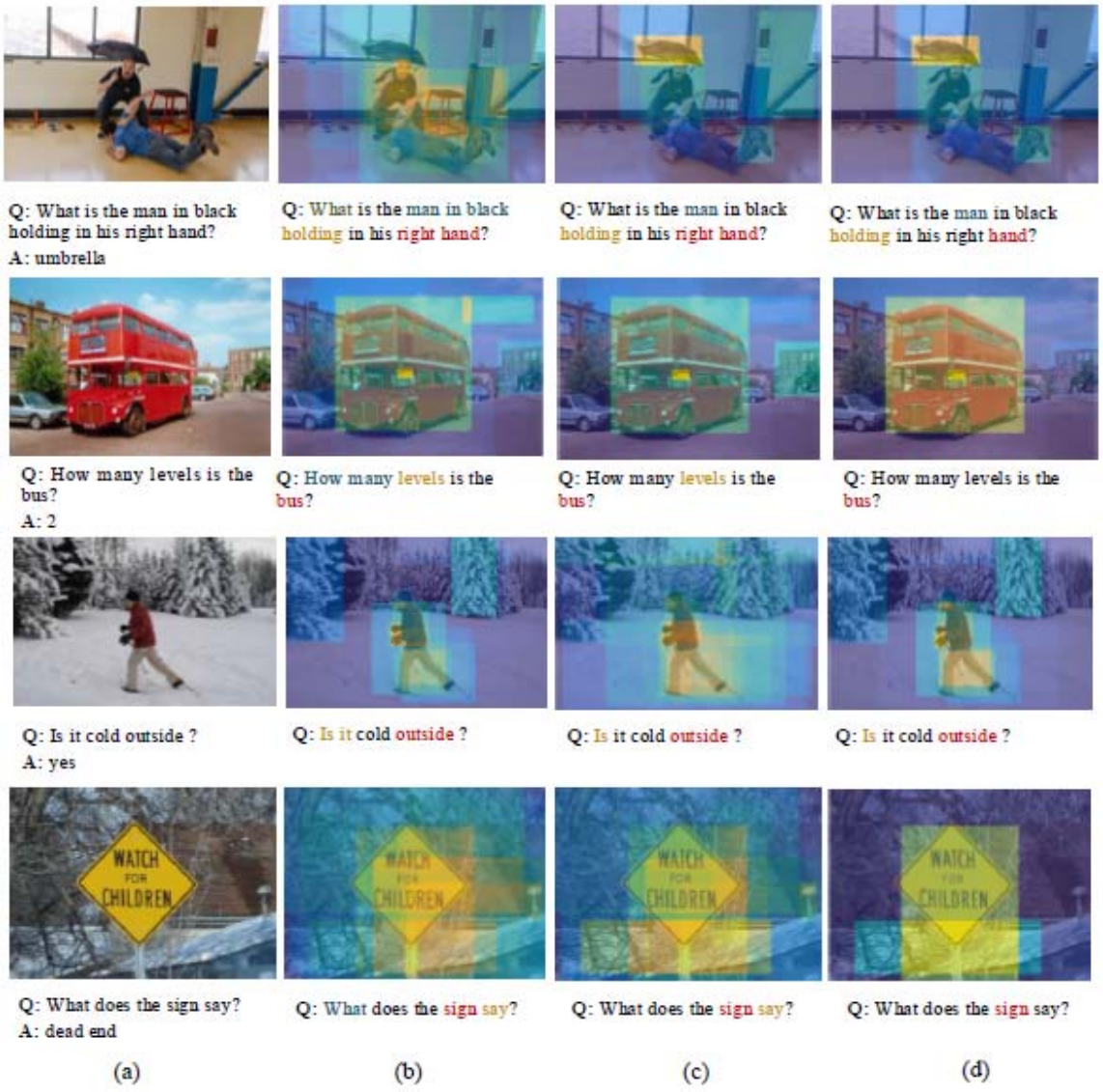

FIGURE III. QUALITATIVE EXAMPLES ON VQA TEST SET

V. CONCLUSION

In this paper, we proposed a novel multimodal cross-guided attention networks to focus on both significant regions in image and meaningful words in question and a novel joint learning strategy is applied for addressing automatic visual question answering. A multi-step reasoning makes it possible to understand fine-grained image regions and high-level semantic representations. Our cross-guided strategy reduces the gap between vision and language effectively for further reasoning. Extensive experiments demonstrate that MCAN outperforms on two public datasets. The visualization of attention layers shows the proposed model attends to the relevant visual clues and textual clues that infer the answer layer by layer. Future works include exploring on extracting image attributes for understanding high-level semantic, and better joint learning methods for these attention mechanisms.

ACKNOWLEDGMENT

This work was partially supported by National Natural Science Foundation of China (NSFC Grant No. 61773272, 61272258, 61301299, 61572085, 61170124, 61272005), Provincial Natural Science Foundation of Jiangsu (Grant No. BK20151254, BK20151260), Science and Education Innovation based Cloud Data fusion Foundation of Science and Technology Development Center of Education Ministry (2017B03112), Six talent peaks Project in Jiangsu Province (DZXX-027), Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University (Grant No. 93K172016K08), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[2] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in Advances in neural information processing systems, 2015, pp. 2953–2961.

[3] H. Noh, P. Hongsuck Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 30–38.

[4] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29.

[5] 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," arXiv preprint arXiv:1606.01847, 2016.

[6] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" arXiv preprint arXiv:1606.03556, 2016.

[7] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1–9.

[8] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual qa," in Advances in Neural Information Processing Systems, 2016, pp. 361–369.

[9] R. Li and J. Jia, "Visual question answering with question representation update (qru)," in Advances in Neural Information Processing Systems, 2016, pp. 4655–4663.

[10] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in International Conference on Machine Learning, 2016, pp. 2397–2406.

[11] D. Yu, J. Fu, T. Mei, and Y. Rui, "Multi-level attention networks for visual question answering," in Conf. on Computer Vision and Pattern Recognition, vol. 1, no. 7, 2017, p. 8.

[12] W. Zhang, C. Zhang, P. Liu, Z. Zhan, and X. Qiu, "Two-step joint attention network for visual question answering," in 2017 3rd International Conference on Big Data Computing and Communications (BIGCOM). IEEE, 2017, pp. 136–143.

[13] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network." in AAAI, vol. 3, no. 7, 2016, p. 16.

[14] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4622–4630.

[15] H. Noh and B. Han, "Training recurrent answering units with joint loss minimization for vqa," arXiv preprint arXiv:1606.03647, 2016.

[16] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in European Conference on Computer Vision. Springer, 2016, pp. 451–466.

[17] I. Ilievski, S. Yan, and J. Feng, "A focused dynamic attention model for visual question answering," arXiv preprint arXiv:1604.01485, 2016.

[18] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in Advances In Neural Information Processing Systems, 2016, pp. 289–297.

[19] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," arXiv preprint arXiv:1611.00471, 2016.

[20] P. Wang, Q. Wu, C. Shen, and A. v. d. Hengel, "The vqa-machine: Learning how to use existing vision algorithms to answer new questions," arXiv preprint arXiv:1612.05386, 2016.

[21] 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 vqa," arXiv preprint arXiv:1707.07998, 2017.

[22] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," arXiv preprint arXiv:1505.00387, 2015.

[23] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994, pp. 133–138.

[24] K. Saito, A. Shin, Y. Ushiku, and T. Harada, "Dualnet: Domain-invariant network for visual question answering," in Multimedia and Expo (ICME), 2017 IEEE International Conference on. IEEE, 2017, pp. 829–834.