Article

Multi-Layer Cross-Guided Attention Networks for Visual Question Answering

Haibin Liu [1], Shengrong Gong [2], Yi Ji [1], Jianyu Yang[3],Tengfei Xing[1], Chunping Liu[1]*

[1] School of Computer Science and Technology, Soochow University Suzhou, Jiangsu, China; liuhaibin210317@hotmail.com; jiyi@suda.edu.cn; tfxing@stu.suda.edu.cn

[2] School of Computer Science and Engineering,Changshu Institute of Technology,Changshu, Jiangsu, China; shrgong@cslg.cn

[3] School of Rail Transportation, Soochow University Suzhou, Jiangsu, China; jyyang@suda.edu.cn

* Correspondence:cpliu@suda.edu.cn

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. The existing approaches predominantly infer answers from visual spatial information, while neglecting sequential information in question and guidance information between image and question. To remedy this, we imitate the human mechanism of cross-reasoning about visual and textual information and propose a multi-layer cross-guided attention network (MCAN) for visual question answering which employs a cross-guided joint learning strategy with a non-linear activation learning method and simultaneously capture both the rich visual spatial information and the sequential semantic information. We evaluate the proposed model on two public datasets: VQA dataset and COCO-QA dataset, and extensive experiments show state-of-the-art performance on the datasets.

Keywords: Visual Question Answering; multimodal; cross-guided; attention

1. 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. Comparing with text-based QA system in NLP, VQA is more abstract as it answers a question based on visual information. The development of VQA can effectively alleviate the gap between CV and NLP, and further improve the capability of machine understanding the real world. 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. In essence, VQA is a complex computational process that predicts the answer by learning multimodal joint features. 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, video surveillance.

Previous methods for VQA [1–4] utilize a pre-trained Convolution Neural Networks(CNN) such as VGG net [5] and ResNet [6] to extract global image features as image representations and encode question via Recurrent Neural Networks(RNN) [7,8], and finally comply them with a 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. Latest state-of-the-art models [9–12] employ visual attention mechanism to focus on the relevant regions to question which further improve the performance and obtain more accurate answers. But these approaches ignore the sequential semantic information in question which means that there is still potential for improving the performance.

Moreover, these approaches select the pool5 (VGG net) layer or res5c (ResNet) layer features from CNN as image representations which retains original image spatial information (e.g. the dimension of res5c layer features in ResNet is 14 × 14 × 2048 where 14 × 14 is the number of regions). However, recent work [13] holds that current VQA models using these image representations with 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 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.

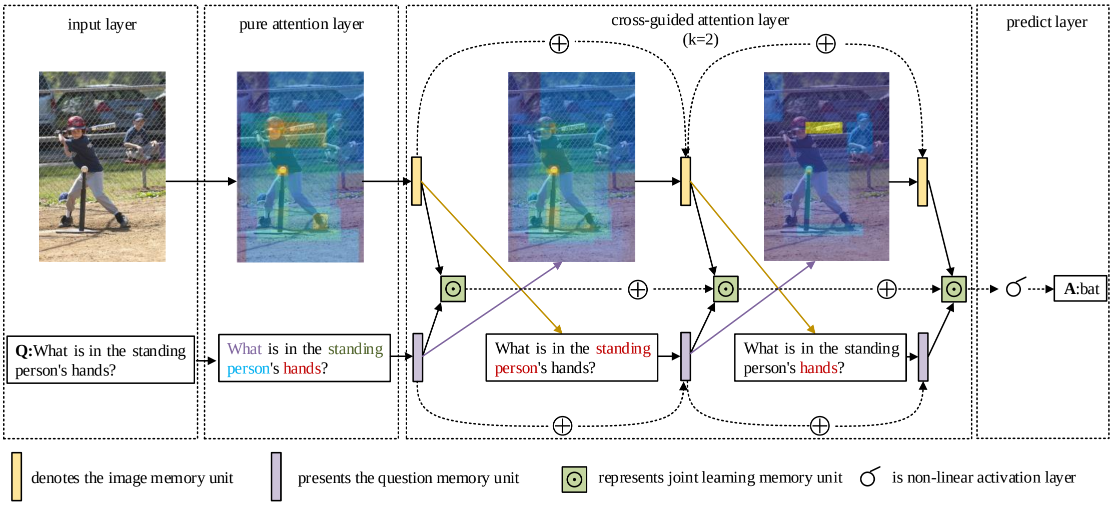

input layer
pure attention layer
cross-guided attention layer
(k=2)
predict layer
Q:What is in the standing person's hands?
What is in the standing person's hands?
What is in the standing person's hands?
What is in the standing person's hands?
A:bat
denotes the image memory unit
presents the question memory unit
represents joint learning memory unit
is non-linear activation layer

Figure 1. Overview of Multi-Layer Cross-Guided Attention Networks (MCAN) which sets the number of layers k to 2. The different colors of bounding boxes in image and words in question indicate the attention maps predicted by MCAN.

To address these problems, we propose a novel multi-layer cross-guided attention networks (MCAN) that implement a co-attention mechanism and allow multi-step reasoning for VQA. The overview of MCAN is illustrated in Figure 1. MCAN is a typical CNN + RNN architecture with joint learning model. It consists of four major modules: (1) the input module which contains a fine-tuned Faster R-CNN [14] in conjunction with ResNet-101 [6] for extracting high-level image representations and a Bidirectional Gated Recurrent Unit(Bi-GRU) [7] for encoding questions. (2) The pure attention module which only focuses on significant regions in image and meaningful words in question. (3) The multi-layer cross-guided attention module where MCAN maintain the states of image, question and joint features via three memory vectors, and previous state is used for guiding to generate attention maps of next step for image and question by a crossing strategy, and then utilize a non-linear activation layer (NLA) to enhance the powerful performance of visual and textual attention weighted sum. (4) The predict module which feeds the joint learning memory vector into NLA and is attached with a classifier for inferring the final answer. We evaluate the proposed model on two public datasets, VQA dataset [1] and COCO-QA dataset [2], and obtain state-of-the-art results.

The main contributions of our work are followed as:

- we propose a multi-layer cross-guided attention networks for VQA task which take full advantage of multimodal cross-guided information.
- we introduce a pure attention mechanism without guiding information to initialize the cross-guided attention networks.
- in order to better improve the expressiveness of joint learning features, a novel non-linear activation approach is introduced for inferring the answer.

- We conduct extensive experiments on two public VQA datasets [1,2], and achieve significant improvements over one-glimpse and two-glimpse attention models.

2. Related Work

Recent works have benefited a lot from deep neural network architectures in both the fields of CV and NLP, such as classification tasks [15–19], detection tasks [20–23] and recognition tasks [24–26]. inspired by this, the deep neural networks are widely-used in VQA task. VQA is a multimodal joint learning task closely related to image captioning. Most early proposed models [2,3,27] are transformed from image captioning [28–30]. The development of deep learning has greatly promoted the development of VQA. With the release of VQA dataset [1] and online evaluation method, more and more VQA approaches have been proposed.

Joint feature embedding learning for VQA. Most existing VQA approaches [1,3,27,31–33] are based on CNN-RNN architecture. These models utilize CNN to extract high-level semantic representations from images and encode questions via RNN, and then combine two modalities with an appropriate joint learning approach. They solve the task as a multi-way classification problem. Zhou et al. [31] proposed a simple baseline for VQA, which is a typically joint learning framework that learns image features with CNN and question representation from Long Short-Term Memory (LSTM) [8], and concatenates these two features to infer answer. Besides LSTM, Li et al. [34], Kafle et al. [35] and Xiong et al. [36] utilized GRU and Yang et al. [9], Zhang et al. [37] and Ma et al. [38] trained CNN for encoding question. In particular, Andreas et al. [39] employed a compositional structure for question. There are several methods different from above ones which addressed VQA task as a multi-way classification problem. Malinowski et al. [40] fed both image and question into LSTM at each time step, and then generated the answer. Wu et al. [41] extracted attributes from image and generated descriptions of image as input of LSTM to generate answer by sequence to sequence learning. Fukui et al. [42] introduced a high-order approach to address VQA task. They designed a multimodal compact bilinear Pooling (MCB) algorithm to address VQA task. However, MCB is prone to dimensional disaster and need more resources for calculation.

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. [9] 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. [43] adopted visual attention with joint loss minimization. Xu et al. [11] proposed a question-guided attention model which obtained the attention map by calculating the semantic similarity between image regions and the question. Shih et al. [10] projected the question representations and representations of image region proposals to a common semantic space and selected question-related regions via spatial attention mechanism. Ilievski et al. [44] 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. [42] 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. [32] proposed an image-question co-attention mechanism that not only focuses on the relevant image regions but also attends to the important question words. Unlike Lu, Nam et al. [45] proposed a dual and parallel attention network, which calculated the textual and visual attention map by a refined multiplication operation. Wang et al. [46] extracted "facts" from image and proposed a novel co-attention approach to address VQA task. Yu et al. [47] trained a concept detector to extract concepts from image and utilized question representations to attend to the relevant regions and related concepts. Impressive performance had 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 crossing adequately.

3. Method

The ultimate goal of VQA is to infer the best answer for a related question. Given an image representations **v** and a related question representations **u**, we have to predict a set of answers and then choose the most likely answer $\hat{a}$ in the set. The process of predicting an answer can be formulated as:

$$\hat{a} = \underset{a \in A}{argmax}\, p(a|\mathbf{v}, \mathbf{u}; \theta) \tag{1}$$

where θ are the model parameters and A is the answers set. For the image embedding **v** and the question embedding **u**, we are aimed to pick up key information from image and question and encode the relationship between **v** and **u** with an efficient joint learning strategy, and it becomes easier to learn a classifier for Equation (1).

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

3.1. Input Representations

Bounding-boxes image encoding. Similar to [48], 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 the non-maximum suppression for each object class with an intersection over union (IoU) threshold is performed, and then select the top-ranked 36 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 $\mathbf{v} = [v_1, v_2, ..., v_n] \in \mathbb{R}^{n \times 2048}$ as our bounding-boxes image representations. This features can effectively eliminate the gap between the regions, while making the various regions communicate with each other closely.

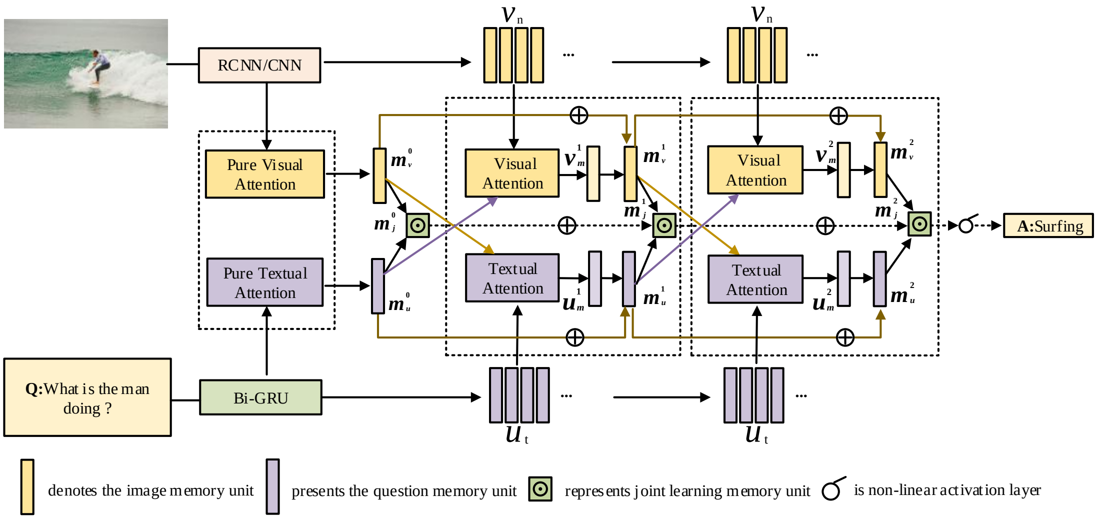


Figure 2. Overall architecture of MCAN in case of the number of cross-guided attention layers k being set to 2. Our framework is a multi-layer attention network with multimodal information joint learning by a cross guiding strategy.

Question encoding. We adopt Bi-GRU to encode question. The Gated Recurrent Unit (GRU) has proven to achieve better performance in recent VQA methods [34–36,49]. It is a variant of LSTM, but it is simpler than LSTM. A GRU cell contains an update gate z and a reset gate r. Given an one-hot question representations $q = [q_1, q_2, ..., q_T]$, 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 $x_t = Wq_t$, where W is embedding weights. At each time step, we feed the word embedding feature vector x_t into Bi-GRU:

$$h_t^f = GRU^f(x_t, h_{t-1}^f) \tag{2}$$

$$h_t^b = GRU^b(x_t, h_{t+1}^b) \tag{3}$$

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:

$$u_t = f([h_t^f, h_t^b]) \tag{4}$$

For better capturing the sequential semantic information in question, our networks stack the concatenated hidden states sequentially. Finally, a set of feature vectors $\mathbf{u} = [u_1, u_2, ..., u_T]$ is constructed for representing question, where u_t represents the semantic feature from the first to the t-th words.

3.2. Attention Mechanism

Our method performs visual attention and textual attention simultaneously through a multiple layer network and three context memory vectors generated by image and question representations maintain states of the representations at each attention layer.

In this section, we detail our attention mechanism employed at each layer, which depends on the context memory vector.

3.2.1. Visual Attention

Our model implements two visual attention mechanisms. The first attention mechanism is similar to image saliency called Pure Visual Attention Mechanism (PVA), which only considers original information of the image. In this work, we take the features extracted from bounding-box image regions as the whole image representations. However, these bounding-boxes maybe contain the same image regions which retains important information in the image. PVA dedicates to attending to these common regions and picks up the key features for the next attention layer. For each location $i = 1, 2, ..., n$ in the image, we attend to the regions by:

$$v_p = PVA(\mathbf{v}) \tag{5}$$

the formula expands as:

$$h_{v,p} = tanh(W_{v,p}\mathbf{v} + b_{v,p}) \tag{6}$$

$$\alpha_{v,p} = softmax(h_{v,p}) \tag{7}$$

$$v_p = \sum_{i=1}^{n} \alpha_{v,p} v_i \tag{8}$$

where $W_{v,p}$ is a learned weights matrix, $b_{v,p}$ is bias, $\alpha_{v,p}$ is the attention map, v_p is the pure visual attention weighted sum. The second attention mechanism is more complicated. It is a classical question-guided visual attention (QVA) mechanism similar to most attended VQA models [9,11]:

$$v_m = QVA(\mathbf{v}, m_u) \tag{9}$$

where m_u is a memory vector of question. For each location $i = 1, 2, ..., n$ in the image, we first tile question memory features m_u to the dimensions the same as the image features $\mathbf{v}$, and then feed the concatenated features into NLA f_a(see section 3.3). Finally, a linear transform is applied for the output of NLA and a scalar attention weight v_m is obtained, formulated as:

$$h_{v,m} = W_{v,m} f_a([\mathbf{v}, m_u]) + b_{v,m} \tag{10}$$

$$\alpha_{v,m} = softmax(h_{v,m}) \tag{11}$$

$$v_m = \sum_{i=1}^{n} \alpha_{v,m} v_i \tag{12}$$

where $W_{v,m}$ is a learned parameters matrix, $\alpha_{v,m}$ is a attention weights vector, $b_{v,m}$ is bias, and v_m is the question-guided attention weighted sum.

3.2.2. Textual Attention

Similar to the visual attention mechanism, we also implement two textual attention mechanisms to attend to the specific words in question which tell the model "what to say".

The first textual attention mechanism called Pure Textual Attention Mechanism (PTA) without image-guided information is aimed to focus on the original meaningful words in the sentence. A 2-layer feed-forward neural network is executed for obtaining the textual attention weights, formulated as:

$$u_p = PTA(\mathbf{u}) \tag{13}$$

expanded as:

$$h_{u,p} = tanh(W_{u,p}\mathbf{u} + b_{u,p}) \tag{14}$$

$$\alpha_{u,p} = softmax(h_{u,p}) \tag{15}$$

$$u_p = \sum_{t=1}^{T} \alpha_{u,p} u_t \tag{16}$$

where $W_{u,p}$ is a weights matrix, $b_{u,p}$ is bias, and u_p is the pure textual attention weighted sum.

Generally, the image contains more information related to the question. Inspired by this, we achieve image-guided textual attention (ITA) mechanism to pay attention to the meaningful words in question:

$$u_m = ITA(\mathbf{u}, m_v) \tag{17}$$

where m_v is image memory state. We repeat the image features by T times and concatenated with question features. Commonly, we perform a linear transform after feeding the features into NLA, and then calculate the probability distribution over the question words, finally, we get a weighed textual feature vector u_m by:

$$h_{u,m} = W_{u,m} f_a([\mathbf{u}, m_v]) + b_{u,m} \tag{18}$$

$$\alpha_{u,m} = softmax(h_{u,m}) \tag{19}$$

$$u_m = \sum_{t=1}^{T} \alpha_{u,m} u_t \tag{20}$$

where $W_{u,m}$ is a weights matrix, $b_{u,m}$ is bias, u_m is the image-guided attention weighted sum.

3.3. Non-linear activation layer

Inspired by the highway networks [50], we introduce NLA using a gated hyperbolic tangent activation. It is a gated operations similar to recurrent units such as LSTM and GRU. Given a vector x, the output of NLA for x is defined as follows:

$$o_x = f_a(x) \tag{21}$$

expanded as:

$$t = tanh(Wx + b) \tag{22}$$

$$s = \sigma(W^{'}x + b^{'}) \tag{23}$$

$$o_x = t \circ s \tag{24}$$

where W, $W^{'}$ are the learned parameters, b, $b^{'}$ are the bias, σ is the sigmoid activation function, o_x is the output of NLA, and $\circ$ is element-wise multiplication.

3.4. MCAN for Visual Question Answering

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 units are recursively updated by:

$$m_v^k = m_v^{(k-1)} + v_m^k \tag{25}$$

$$m_u^k = m_u^{(k-1)} + u_m^k \tag{26}$$

$$m_j^k = m_j^{(k-1)} + m_v^k \bigodot m_u^k \tag{27}$$

where k is the number of cross-guided attention layers which is set to 1 at least, m_v^k and m_u^k are the visual and textual memory states of k-th attention layers, v_m^k and u_m^k are the visual and textual attention weight sums based on Equation (9) and Equation (17), respectively, $\bigodot$ is element-wise product, m_j^k is the context joint learning memory unit which contains the significant contents about image and question. The initial visual memory state m_v^0, textual memory state m_u^0 and context memory state m_j^0 are calculated by Equation (5) and Equation (13),

$$m_j^0 = m_v^0 \bigodot m_u^0 \tag{28}$$

where $m_v^0 = v_p$ and $m_u^0 = u_p$. MCAN is a crossed network which contains question-guided attention mechanism and image-guided attention mechanism where 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 2) where each attention layer could effectively locate the question-related image regions and image-related question words. The cross-guided strategy is followed as:

$$v_m^k = QVA(\mathbf{v}, m_u^{(k-1)}) \tag{29}$$

$$u_m^k = ITA(\mathbf{u}, m_v^{(k-1)}) \tag{30}$$

where $k = 1, 2, ..., n$. After k steps iterating, MCAN could capture more fine-grained visual attention information and more high-level textual semantic information.

Similar to most VQA models, MCAN predicts the final answer by a multi-way classifier. In this paper, MCAN selects top 3000 frequent answers as the final labels. We feed the last joint learning memory vector m_j^k into NLA f_a, and then a single-layer softmax classifier with cross-entropy is used to predict the answer by Equation (1):

$$h^k = f_a(m_j^k) \tag{31}$$

$$p_a = softmax(W_a h^k + b_a) \tag{32}$$

where h^k is the output of f_a, p_a is the probability distribution over the candidate answers, W_a is learned weights matrix, b_a is bias.

4. Experiments

4.1. Dataset and evaluation metrics

We evaluate MCAN on two public datasets, VQA [1] and COCO-QA [2].

VQA dataset is one of the most widely used dataset for VQA task. The images in the dataset are all from the Microsoft COCO dataset [51]. 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 and Open-Ended. For Multiple-Choice task, the answer to each question is chosen from the 18 candidate answers. For Open-Ended task, each question has 10 answers for evaluation. We combine training set and validation set for training, and take the top 3000 frequent answers as labels.

COCO-QA dataset is also based on Microsoft COCO dataset. Images mainly come from the training set and validation set in Microsoft COCO 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 Open-Ended task is measured by a voting mechanism to calculate the accuracy of the generated answer:

$$Accuracy_{OE} = min(\frac{\#\, human\ that\ provided\ that\ answer}{3}, 1) \tag{33}$$

and for the Multiple-Choice 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) [52]. WUPS is used to measure the correlation between words based on the taxonomy tree. Similar to [32], we use the threshold 0.0 and 0.9 for WUPS.

4.2. Implementation Details

We implement our proposed model on the tensorflow platform. For obtaining better image representations, we feed the image into Faster R-CNN and fix the number of bounding-box to 36, and then take the features from pool5 layer. Finally, the image representations are 36 × 2048. For embedding the questions, the length of questions is set to 26 and the word is embedded by a pre-trained word vectors called Glove [53] which represents a word with a 300-dimension vector. The network sets the hidden state of GRU to 512, and the joint learning features are set to 1200. For details, we train our networks by stochastic gradient descent and use the Adam solver with learning rate 0.001, momentum 0.9, weight decay 10e-8, dropout ratio 0.5. The batch size is set to 300, and MCAN gets best performance at epoch 90. For VQA dataset, train+val datasets are used for training and the candidate answers are set to 3000. For COCO-QA dataset, the candidate answers are set to 430.

Table 1. 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
iBOWIMG[31]	76.5	35.0	42.6	55.7	61.7	76.8	35.0	42.6	55.9	62.0
DPPnet[49]	80.7	37.2	41.7	57.2	62.5	80.3	36.9	42.2	57.4	62.7
NMN[39]	81.2	38.0	44.0	58.6	-	81.2	37.7	44.0	58.7	-
SAN[9]	79.3	36.6	46.1	58.7	-	79.1	36.4	46.4	58.9	-
Smem[11]	80.9	37.3	43.1	58.0	-	80.8	37.5	43.5	58.2	-
MRN(ResNet)[4]	82.3	38.8	49.2	61.7	66.2	82.4	38.2	49.4	61.8	66.3
FDA[44]	81.1	36.2	45.8	59.2	-	-	-	-	59.5	-
DMN+[36]	80.5	36.8	48.3	60.3	-	-	-	-	60.4	-
RAU(ResNet)[43]	81.9	39.0	53.0	63.3	67.7	81.7	38.2	52.8	63.2	67.3
MCB(ResNet)[42]	82.2	37.7	54.8	64.2	68.6	-	-	-	-	-
HiecoAtt(ResNet)[32]	79.7	38.7	51.7	61.8	65.8	-	-	-	62.1	66.1
VQA-Machine(ResNet)[46]	81.5	38.4	53.0	63.1	67.7	81.4	38.2	53.2	63.3	67.8
DAN[45]	**83.0**	39.1	53.9	64.3	69.1	82.8	38.1	54.0	64.2	69.0
MLAN[47]	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**

Table 2. Results of our proposed approach and compared methods on COCO-QA dataset, in percentage and '-' represents the result is not available.

Method	All	Object	Number	Color	Location	WUPS0.9	WUPS0.0
2-VIS+BLSTM[2]	55.0	58.2	44.8	49.5	47.3	65.3	88.6
ATT-VGG-SEG[12]	58.1	62.5	45.7	46.8	53.7	68.4	89.9
IMG-CNN[38]	58.4	-	-	-	-	68.5	89.7
DPPnet[49]	61.2	-	-	-	-	70.8	90.6
SAN[9]	61.6	65.4	48.6	57.9	54.0	71.6	90.9
QRU[34]	62.5	65.1	46.9	**60.5**	**57.0**	72.6	**91.6**
MCGN(k=2)	**63.1**	**66.5**	**51.9**	55.8	56.2	**72.9**	91.3

4.3. Results and analysis

Table 1 shows the performance of our approach on VQA dataset and comparison with state-of-the-art methods. We train our networks on train+val datasets and test on test dataset. We use the VQA-team [1] as our baseline method. In the first part of Table 1, we compare MCAN with baseline approaches without attention mechanism and MCAN has made a great improvement. The second part of Table 1 exhibits the performance of methods with one-glimpse attention mechanism. SAN [9] employs element-wise addition operation to calculate attention map and gets better performance than the methods in the first part. DAN [45] adopts element-wise multiplication operation to obtain joint features and compute the attention weights, and then gets further improvements. In MCAN, we replace the addition or the multiplication with concatenating the image features and question features to compute the joint attention map, and gain further improvements. MCB [42] 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 with 2 layers obtains best performance on the whole. MCAN with 2 layers improves MCB [42] from 68.6% to 70.1% for the Multiple-Choice task on Test-dev set. The third part of Table 1 shows the methods with co-attention mechanism which contains several joint attention mechanisms. HiecoAtt [32] and DAN [45] use textual attention cooperated with visual attention which gained better performance than most one-glimpse attention methods. MLAN [47] adopts semantic attention instead of textual attention and slightly improved over HiecoAtt [32]. Compared with these two-glimpse attention methods, MCAN improves DAN [45] from 69.1% to 70.1% on Test-dev set, and from 69.0% to 70.2% on Test-standard set for Multiple-Choice task. MCAN achieves 0.2% improvements of overall accuracy than DAN [45] for Open-Ended task on Test-dev and Test-standard set, respectively. These improvements show the advantage of our proposed method with cross-guided attention. We train MCAN with k=1,2,3, and MCAN with k=2 outperforms others.

We evaluate our model on COCO-QA dataset and show the performance in Table 2. We also compare the performance by accurate and WUPS with other methods. MCAN with k=2 improve state-of-the-art QRU [34] from 62.5% to 63.1%. In particular, our model achieves a 5% improvement for the question type number. However, our model decreases the accuracy in the question types of Color and Location. The possible reason is that MCAN discards words longer than 30 in question. This strategy may have a certain impact on the accuracy. MCAN also adopts WUPS to measure performance. Comparing with QRU [34], 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 3. Ablation study on the VQA Test-dev set.

Method	Test-dev
SAN[9]	58.7
MCAN without NLA	60.6
MCAN without pure attention layer	61.0
MCAN (VGG)61.6Visual-Attention-Only	61.8
MCAN with LSTM	63.1
MCAN full	**64.5**

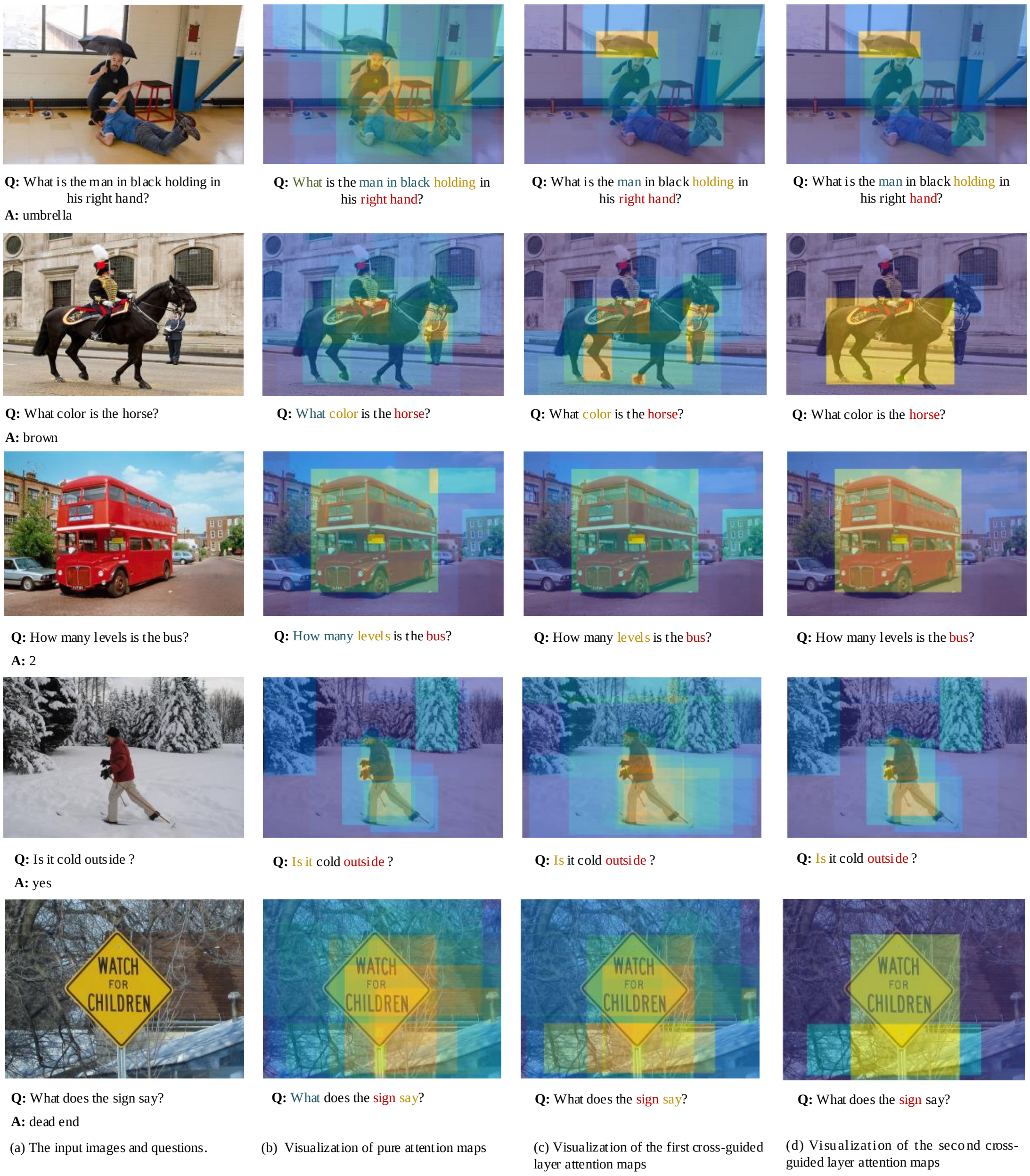
4.4. Ablation study

In this section, we conduct ablation studies to analyze the contribution of each component on our proposed model. We ablate our full model and evaluate on VQA test-dev set. The certain ablation experiments are followed as:

- SAN (VGG net) [9]: a two-layer attention networks without PVA and NLA, multi-layer LSTM is employed for encoding question.
- MCAN without NLA: the full MCAN model discards NLA.
- MCAN with VGG net: the images are encoded by VGG net instead of Faster R-CNN and ResNet-101.
- MCAN without pure attention layer: MCAN removes PVA layer and PTA layer.
- Visual-Attention-Only: the textual attention module is removed from MCAN.
- MCAN with Bi-LSTM: MCAN replaces the Bi-GRU with Bi-LSTM.

Table 3 shows the performance of our ablation experiments on VQA test-dev set. The Visual-Attention-Only model performs better than SAN [9] which does not contain PVA and NLA by 3.1%. This demonstrates that PVA firstly filters out the noise in image and locates important regions all over the image. Furthermore, NLA effectively activates the feature vectors and reduces global loss. To further understand the function of NLA, we re-train MCAN without NLA, and the performance is dropped by 3.9% compared with full MCAN. The pure attention layer is important for full MCAN, and It seems to be a pre-attended locator which retains the useful visual and textual information for the further cross-guided attention layer. We break down full MCAN and ablate pure attention layer, and the model obtains performance almost the same as MCAN without NLA. To measure the impact of image model on performance, we take VGG net to encode image and train MCAN (VGG) for 200 epoches. Experiments show MCAN full get better performance by 2.9%. Our hypothesis is that existing approaches employing attention search for the regions related to question one by one, resulting in each region becoming an isolated unit, while our image encoder solves the problem and strengthens linkages between regions. As a result, MCAN full gets better performance. For demonstrating the contribution of linguistic model for our proposed model, we design MCAN with Bi-LSTM, and the result on Table 3 shows the advantage of Bi-GRU.

Ablation study does demonstrate the rationality of the proposed model and that the linguistic model has the least impact on performance.


Q: What is the man in black holding in his right hand?
A: umbrella
Q: What is the man in black holding in his right hand?
Q: What is the man in black holding in his right hand?
Q: What is the man in black holding in his right hand?
Q: What color is the horse?
A: brown
Q: What color is the horse?
Q: What color is the horse?
Q: What color is the horse?
Q: How many levels is the bus?
A: 2
Q: How many levels is the bus?
Q: How many levels is the bus?
Q: How many levels is the bus?
Q: Is it cold outside ?
A: yes
Q: Is it cold outside ?
Q: Is it cold outside ?
Q: Is it cold outside ?
WATCH FOR CHILDREN
Q: What does the sign say?
A: dead end
Q: What does the sign say?
Q: What does the sign say?
Q: What does the sign say?

(a) The input images and questions.

(b) Visualization of pure attention maps

(c) Visualization of the first cross-guided layer attention maps

(d) Visualization of the second cross-guided layer attention maps

Figure 3. Qualitative examples on VQA test set. The first four rows are the correct examples, and the last row is the error example.

4.5. Qualitative evaluation

We visualize the multi-layer attention maps generated by the proposed model in Figure 3 and present five qualitative examples from VQA test set. Figure 3 (a) are the input images, questions and the generated answers. Figure 3 (b) are the visualization of the attention maps generated by MCAN without any additional information. Figure 3 (c), (d) show that MCAN attention layers with k=2 attend to the image regions related to question and meaningful words corresponding to image layer by layer. The example of the first two rows belong to other question type in VQA test set, and the heat maps are generated for inferring correct answers. The third row shows the example of number question type, even without a clear amount of information in the image. However, MCAN reasons that the bus is two levels. In the fourth example, MCAN infers the correct answer by multi-step reasoning. In some cases, the correct answer not existing in training labels, resulting in the model predicts an incorrect answer, such as the last row example. In the example, "watch for children" is not in the classification labels despite the fact that the model attends to the image regions correctly.

5. Conclusions

In this paper, we proposed a novel multi-layer 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 and the pure attention mechanism captures visual and textual original semantic information adequately for further reasoning. Extensive experiments demonstrate that MCAN outperforms on two public datasets, VQA dataset and COCO-QA dataset. 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 and exploiting concept attention mechanism to attend semantic concepts, and better joint learning methods for these attention mechanisms.

Acknowledgments: 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.

Author Contributions: Haibin Liu and Chunping Liu conceived and designed the experiments; Haibin Liu, Jianyu Yang and Tengfei Xing performed the experiments; Shengrong Gong and Yi Ji analyzed the data; Haibin Liu wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C.L.; Parikh, D.; Batra, D. Vqa: Visual question answering. *International Journal of Computer Vision* **2017**, *123*, 4–31.
2. Ren, M.; Kiros, R.; Zemel, R. Exploring models and data for image question answering. Advances in neural information processing systems, 2015, pp. 2953–2961.
3. Malinowski, M.; Rohrbach, M.; Fritz, M. Ask your neurons: A neural-based approach to answering questions about images. Proceedings of the IEEE international conference on computer vision, 2015, pp. 1–9.
4. Kim, J.H.; Lee, S.W.; Kwak, D.; Heo, M.O.; Kim, J.; Ha, J.W.; Zhang, B.T. Multimodal residual learning for visual qa. Advances in Neural Information Processing Systems, 2016, pp. 361–369.
5. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.

6. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
7. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* **2014**.
8. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural computation* **1997**, *9*, 1735–1780.
9. Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A. Stacked attention networks for image question answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29.
10. Shih, K.J.; Singh, S.; Hoiem, D. Where to look: Focus regions for visual question answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4613–4621.
11. Xu, H.; Saenko, K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. European Conference on Computer Vision. Springer, 2016, pp. 451–466.
12. Chen, K.; Wang, J.; Chen, L.C.; Gao, H.; Xu, W.; Nevatia, R. ABC-CNN: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960* **2015**.
13. Das, A.; Agrawal, H.; Zitnick, C.L.; Parikh, D.; Batra, D. Human attention in visual question answering: Do humans and deep networks look at the same regions? *arXiv preprint arXiv:1606.03556* **2016**.
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence* **2017**, *39*, 1137–1149.
15. Zhou, L.; Li, Q.; Huo, G.; Zhou, Y. Image Classification Using Biomimetic Pattern Recognition with Convolutional Neural Networks Features. *Computational intelligence and neuroscience* **2017**, *2017*.
16. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors* **2015**, *2015*.
17. Zhang, H.; Xiao, L.; Wang, Y.; Jin, Y. A generalized recurrent neural architecture for text classification with multi-task learning. *arXiv preprint arXiv:1707.02892* **2017**.
18. Kowsari, K.; Brown, D.E.; Heidarysafa, M.; Meimandi, K.J.; Gerber, M.S.; Barnes, L.E. Hdltex: Hierarchical deep learning for text classification. *arXiv preprint arXiv:1709.08267* **2017**.
19. Yu, S.; Jia, S.; Xu, C. Convolutional neural networks for hyperspectral image classification. *Neurocomputing* **2017**, *219*, 88–98.
20. Ren, Y.; Zhu, C.; Xiao, S. Object Detection Based on Fast/Faster RCNN Employing Fully Convolutional Architectures. *Mathematical Problems in Engineering* **2018**, *2018*.
21. Yan, S.; Xia, Y.; Smith, J.S.; Lu, W.; Zhang, B. Multiscale Convolutional Neural Networks for Hand Detection. *Applied Computational Intelligence and Soft Computing* **2017**, *2017*.
22. Sladojevic, S.; Arsenovic, M.; Anderla, A.; Culibrk, D.; Stefanovic, D. Deep neural networks based recognition of plant diseases by leaf image classification. *Computational intelligence and neuroscience* **2016**, *2016*.
23. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence* **2016**, *38*, 142–158.
24. Shi, B.; Bai, X.; Yao, C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence* **2017**, *39*, 2298–2304.
25. Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D.L.; Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **2017**, *33*, i37–i48.
26. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* **2013**, *35*, 221–231.
27. Gao, H.; Mao, J.; Zhou, J.; Huang, Z.; Wang, L.; Xu, W. Are you talking to a machine? dataset and methods for multilingual image question. Advances in Neural Information Processing Systems, 2015, pp. 2296–2304.
28. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.
29. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.
30. Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; Yuille, A. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632* **2014**.

31. Zhou, B.; Tian, Y.; Sukhbaatar, S.; Szlam, A.; Fergus, R. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167* **2015**.
32. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical question-image co-attention for visual question answering. Advances In Neural Information Processing Systems, 2016, pp. 289–297.
33. Wu, Q.; Shen, C.; Wang, P.; Dick, A.; van den Hengel, A. Image Captioning and Visual Question Answering Based on Attributes and External Knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**.
34. Li, R.; Jia, J. Visual question answering with question representation update (qru). Advances in Neural Information Processing Systems, 2016, pp. 4655–4663.
35. Kafle, K.; Kanan, C. Answer-type prediction for visual question answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4976–4984.
36. Xiong, C.; Merity, S.; Socher, R. Dynamic memory networks for visual and textual question answering. International Conference on Machine Learning, 2016, pp. 2397–2406.
37. Zhang, W.; Zhang, C.; Liu, P.; Zhan, Z.; Qiu, X. Two-step joint attention network for visual question answering. 2017 3rd International Conference on Big Data Computing and Communications (BIGCOM). IEEE, 2017, pp. 136–143.
38. Ma, L.; Lu, Z.; Li, H. Learning to Answer Questions from Image Using Convolutional Neural Network. AAAI, 2016, Vol. 3, p. 16.
39. Andreas, J.; Rohrbach, M.; Darrell, T.; Klein, D. Neural module networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 39–48.
40. Malinowski, M.; Rohrbach, M.; Fritz, M. Ask your neurons: A deep learning approach to visual question answering. *International Journal of Computer Vision* **2017**, *125*, 110–135.
41. Wu, Q.; Wang, P.; Shen, C.; Dick, A.; van den Hengel, A. Ask me anything: Free-form visual question answering based on knowledge from external sources. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4622–4630.
42. Fukui, A.; Park, D.H.; Yang, D.; Rohrbach, A.; Darrell, T.; Rohrbach, M. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* **2016**.
43. Noh, H.; Han, B. Training recurrent answering units with joint loss minimization for vqa. *arXiv preprint arXiv:1606.03647* **2016**.
44. Ilievski, I.; Yan, S.; Feng, J. A focused dynamic attention model for visual question answering. *arXiv preprint arXiv:1604.01485* **2016**.
45. Nam, H.; Ha, J.W.; Kim, J. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471* **2016**.
46. Wang, P.; Wu, Q.; Shen, C.; Hengel, A.v.d. The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions. *arXiv preprint arXiv:1612.05386* **2016**.
47. Yu, D.; Fu, J.; Mei, T.; Rui, Y. Multi-level attention networks for visual question answering. Conf. on Computer Vision and Pattern Recognition, 2017, Vol. 1, p. 8.
48. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and VQA. *arXiv preprint arXiv:1707.07998* **2017**.
49. Noh, H.; Hongsuck Seo, P.; Han, B. Image question answering using convolutional neural network with dynamic parameter prediction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 30–38.
50. Srivastava, R.K.; Greff, K.; Schmidhuber, J. Highway networks. *arXiv preprint arXiv:1505.00387* **2015**.
51. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. European conference on computer vision. Springer, 2014, pp. 740–755.
52. Wu, Z.; Palmer, M. Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994, pp. 133–138.
53. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.