Computer Science and Information Systems 18(3):1023–1040 https://doi.org/10.2298/CSIS201120032L

Multimodal Encoders and Decoders with Gate Attention for Visual Question Answering

Haiyan Li[1] and Dezhi Han[2]

[1] School of Information Engineering, Shanghai Maritime University
Shanghai,201306, China
1977115781@qq.com
[2] School of Information Engineering, Shanghai Maritime University
Shanghai,201306, China
dezhihan88@sina.com

Abstract. Visual Question Answering (VQA) is a multimodal research related to Computer Vision (CV) and Natural Language Processing (NLP). How to better obtain useful information from images and questions and give an accurate answer to the question is the core of the VQA task. This paper presents a VQA model based on multimodal encoders and decoders with gate attention (MEDGA). Each encoder and decoder block in the MEDGA applies not only self-attention and cross-modal attention but also gate attention, so that the new model can better focus on inter-modal and intra-modal interactions simultaneously within visual and language modality. Besides, MEDGA further filters out noise information irrelevant to the results via gate attention and finally outputs attention results that are closely related to visual features and language features, which makes the answer prediction result more accurate. Experimental evaluations on the VQA 2.0 dataset and the ablation experiments under different conditions prove the effectiveness of MEDGA. In addition, the MEDGA accuracy on the test-std dataset has reached 70.11%, which exceeds many existing methods.

Keywords: Deep Learning, Artificial Intelligence, Visual Question Answering, Gate Attention, Multimodal Learning.

1. Introduction

Deep learning has been extensively applied in the domains of Computer Vision (CV) and Natural Language Processing (NLP), such as object detection, image segmentation, machine translation, and has shown excellent performance in these domains. Tasks based on language and vision are attracting more and more researchers' attention. Inspired by the multimodal task of image captioning, people began to study Visual Question Answering(VQA). VQA [3] is a complete artificial intelligence task which takes images and questions as input and combines their information to output an answer using human language, but some questions cannot be answered directly from the picture, which requires certain knowledge reasoning, so it requires not only a detailed understanding of questions but also the analysis of the visual elements of images [39][28]. How to predict suitable answer is one of the most challenging tasks in VQA. The VQA has the practically applied value in helping the blind [18][19] and image retrieval [16][27], etc. Blind people can input photos they photo and questions into the VQA system to solove their questions, which

can help them ”see” the world. VQA has been applied to medical images recently, CGMVQA can better assist doctors in clinical analysis and diagnosis [35]. VQA can also be combined with wireless sensors [4][42] [34][1]. Wireless sensors can be used in military, agriculture, ecological environment, medical treatment [20] [8][9], etc. The data collected by wireless sensors in these scenarios can be processed into picture data as input to the VQA system. We ask questions about the sensors for related questions, the VQA system will give corresponding answers. The development of CV and NLP produces endless VQA models which base on deep learning. Many previous models use VGG-NET[37], ResNet [21] for the extraction of global features information, and then VQA models learn from Faster RCNN [2] in object detection to obtain the region of interest of image which applies an object detector to obtain image categories accurately; from BoW to Long-Short Term Memory(LSTM),Gate Recurrent Unit(GRU), GloVe, Bert [10], these technologies have significantly improved the VQA accuracy.

On the other hand, we utilize the attention mechanism[44][40][7]to improve the accuracy of the VQA model, which proved to be one of the most effective methods. Attention in the human vision refers to obtain an object region by quickly scanning the entire picture. The global features extracted by the early VQA models contain a lot of irrelevant information or noise information. To circumvent this problem, people apply attention mechanism to the VQA. Yang et al. [46] apply it in their model which makes the VQA more conducive to fine-grained visual understanding. But early attention models ignore interactions in different modal and the links between image areas and the words in the question. In order to avoid this defect, recent studies have also proposed the co-attention model [31][45][32], which can learn image attention and text attention simultaneously.

Although the results of experiment indicate that their models have a good improvement on accuracy, their results still contain some irrelevant information. We guess that the VQA model can further filter out irrelevant information and perform better based on the following conjectures: 1) The VQA model can analyze the correlation between image feature informaion or question feature information and the attention results. 2) The model can model the relationship between different visual objects in the image. Experiments verify our ideas. Specifically, inspired by the AoA network [22] in the image captioning task, we apply gate attention to achieve this. We have designed MEDGA, which can acquire the information in the images and questions more effectively to make more accurate reasoning and give more accurate answers. The entire framework is shown in Figure 1. From Figure 1, we can see that MEDGA consists of several encoder and decoder blocks.

The contributions of this article can be summarized as follows:

(1) A VQA model based on multimodal encoders and decoders with gate attention is designed. Self-attention is employed to describe the inter-modal interactions and use cross-attention to better describe the intra-modal interactions of multi-modal data. The proposed MEDGA makes multi-modal reasoning more accurate by stacking multiple encoders and decoders.

(2) We design a new encoder block and decoder block. Gate attention is introduced in the new blocks, that is, make use of self-attention, cross-modal attention results, and queries to better model the contextual relationship between different objects in the picture so that it is conducive to give fine-grained answers to relational reasoning questions.

(3) This paper has proved the effectiveness of MEDGA based on a great deal of experiments and ablation studies. The accuracy on the VQA v2 dataset outperforms many advanced methods.

The rest structure of this paper is arranged as follows: Chapter 2 introduces the related work of Visual Question Answering, Chapter 3 introduces the overview framework of MEDGA; Chapter 4 introduces related experiments and the comparison of MEDGA with other advanced methods; The conclusion of this paper is shown in Chapter 5.

2. Related Work

2.1. Multimodal Features Fusion

The visual question answering task needs to input images and questions at the same time. The image exists in the form of pixels and contains a lot of rich information while the question exists in the form of text and contains limited information. Therefore, the VQA task requires a complex interaction between visual features and language features to obtain fused features that contain abundant information. The visual features and language features are processed into a form of vector, and the two types of features are merged for gaining a joint representation. In vector fusion, the traditional ways are dot produce, dot addition, and full connection. Akira Fukui et al. believed that the outer product of vectors is more expressive, so they raise a Multimodal Compact Bilinear pooling (MCB) model [14], but it may cause a sharp increase in dimensions. J.-H. Kim et al. proposed a Multimodal Low-rank Bilinear Attention Networks(MLB) model [24]. In this method, the tensor of three used for bilinear combination is decomposed into three 2-dimensional weight matrices and applies matrix decomposition to reduce the rank based on the two-dimensional tensor. The method reduces the dimensionality of the tensor and can make the output feature dimensions low to a certain extent. The parameters of experiment are small. However, MLB is sensitive to hyper-parameter and converges slowly. MUTAN [6] proposed by Ben-younes et al. promotes MCB and MLB, which has stronger expressiveness. This method is based on the decomposition of Tucker tensor, including decomposition into three matrices and core tensor, which effectively parameterizes vision and textual representation.

2.2. Attention Mechanism

In recent years, a wide variety of tasks take advantage of the attention model. The attention mechanism is introduced to the image field [44] by Xu et al. and they calculate the probability distribution of attention to highlight the impact of a key input on the output. Their model can identify salient regions in the image and generate subtitles based on these regions. This idea is applied to visual question answering, making the model focus on the image area related to the question. The visual question answering task not only needs to understand the image feature information, but also the question features. Therefore, understanding the image attention features guided by the question and understanding the question attention features guided by the image features is very important in the task. The introduction of attention mechanisms in VQA tasks is of great help in improving accuracy. Lu et al. proposed to focus on images and problems through parallel co-attention

and alternative attention [30], and proposed that in addition to visual attention which is "where to look," the "what words to listen to" is also important in question attention. The co- attention model was used to infer image and text attention jointly. Peng Gao et al. proposed a Dynamic Fusion with Intra- and Inter- modality Attention Flow (DFAF) model [15]. This method can be used to pass information dynamically between visual and linguistic modalities and can well capture the high-level interaction between language and visual area, and the performance of VQA tasks has been significantly improve. Yu et al. proposed a multi-level attention network (MLAN) [11]. The attention in this system includes semantic attention,attribution attention, and visual attention while paying attention to the semantic attributes and image regions related to question. MLAN reduces the semantic gap between vision and language.

2.3. Other Works for Computer Vision and Language

In addition to the methods mentioned above, there are many other methods for VQA. The method proposed by Qi Wu et al. [43] combines external knowledge to make inferences. Shi Yang et al. proposed using question type feature for solving VQA task [36]. The model is able to predict the type of question in advance before answering the question, reducing the search space for the answer, and achieving a good result in the TDIUC dataset. In [48], they propose a neural network component, allowing count objects from the object proposals. In the cases of not affecting other types of questions, this method improves the accuracy of the number category on the VQA v2 dataset.

3. Multimodal Encoders and Decoders with Gate Attention

This section introduces the MEDGA architectures. The overall framework of MEDGA is shown in Figure 1. MEDGA includes the following four components: 1) Basic visual and language features extraction 2) The Encoder to model intra-modal interactions within language modality and the Decoder to capture intra-modal interactions within visual modality and inter-modal interactions across two modality simultaneously. 3) Feature fusion. 4) An answer prediction layer with multi-label classification. The details of these sections will be described next.

3.1. Image and Question Representations

To obtain visual features, this paper uses the top-down and bottom-up model proposed by Peter Anderson et al. [2]. Faster R-CNN uses ResNet-101 for initialization, and then perform fine-tuning on the Visual Genome dataset[26].The object detection in the pictures by Faster R-CNN includes the following two steps: First, object proposals in the image are predicted through the RPN and select the proposals with the highest score as input to the next step. Second, use RoI pooling to select smaller feature maps for each boxing proposal. In this article, we take the top 100 detected objects with the highest probability. Given image I, the obtained vision feature can be expressed as $V \in R^{\mu \times 2048}$,where μ represents the total number of object regions.The i^{th} region feature can be expressed as $r_i \in R^{2048}$.Input a question Q of length L, Q is first tokenized into a sequence of words. These words are represented by one-hot vector and then they are transformed

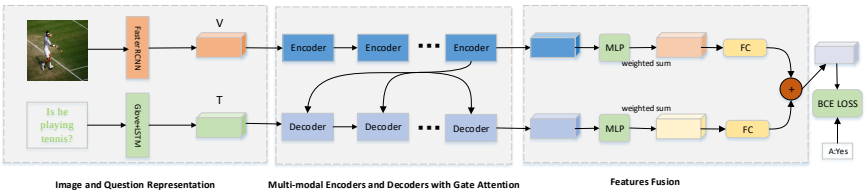


Fig. 1. The overall framework of MEDGA. MEDGA stacks multiple encoders and decoders with self-attention, cross-modal attention, and gate attention. Through MEDGA, we obtain visual features and question features. Input the fused features to answer prediction layer to get the answer to the question.

into a 300-dimensional word embeddings by GLoVe. The resulting word sequence size is $n \times 300$,where n is the number of words in the question. It is then sent to the LSTM. The word vector is encoded into 1024-dimensional features. The process of obtaining visual features V and language features T can be expressed by Equations (1) and (2),where θ_{lstm} and $\theta_{faster\,rcnn}$ re the parameters of the visual and language features.

$$V = Faster\,RCNN(I; \theta_{faster\,rcnn}) \tag{1}$$

$$T = LSTM(Q; \theta_{lstm}) \tag{2}$$

3.2. Encoder and Decoder

This article designs encoders and decoders block. MEDGA takes question representation V and image representation T as input, and then outputs their features with attention learning. Every encoder and decoder includes self-attention, cross-modal attention and gate attention modules.

In short, the attention mechanism is the process of mapping a query and a set of key-value pairs to output [41]. Both self-attention and cross-modal attention use multi-head attention [41]. Multi-head attention is calculated by scaled dot-Product attention h times respectively and it can make the model learn relevant information in different representation subspace. The scaled dot-Product attention mechanism is depicted in the left of Fig 2. The input to it is the query matrix (Q) with d_q dimension, the key matrix (K) with d_k dimension, and the value matrix (V) with d_v dimension. The dot product of the query and all keys are calculated. To prevent the dot product get too large, we adjust by $\frac{1}{\sqrt{d_k}}$ which called scaling factor, and then apply the softmax function to obtain the weight on the values. The formula for the scaling dot-Product attention is shown in Equation (3).

$$f_d(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

Multi-head attention consists of h parallel heads. Each head is an independent scaled dot-Product attention. The parameters attention structure is shown in the right of Fig

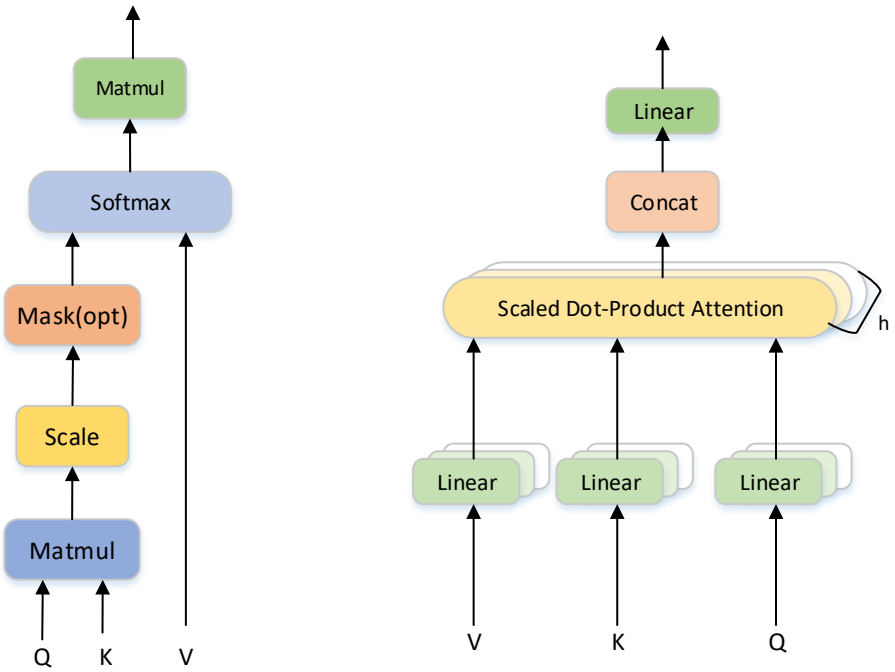


Fig. 2. Scaled dot-Product and Multi-Head Attention

2. For the sake of simplicity,$d_k = d_v = \frac{d_{model}}{h}$ in each attention layer. The specific calculation process is described by Equation (4), where $W_i^Q, W_i^K, W_i^V \in R^{d \times d_h}$ and $W^o \in R^{hd_v \times d_{model}}$ are parameter matrices.

$$\begin{aligned} f_m(Q,K,V) &= Concat(head_1, head_2, ..., head_h)W^o \\ head_i &= f_d(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{4}$$

The gate attention learns from the AoA model proposed in [22], which take the results of multi-head attention calculations and queries as input. We can get a more accurate attended information A from the following Equation (5).

$$A = \sigma(FC_q^g(Q) + FC_f^g(f_m)) \bigodot (FC_q^g(Q) + FC_f^g(f_m)) \tag{5}$$

Encoder Fig 3 illustrates the details of the encoder. The encoder is concerned with the single modal features of the question. When encoding each word, it also pays attention to other words in the input sentence. The encoder can capture some syntactic or semantic

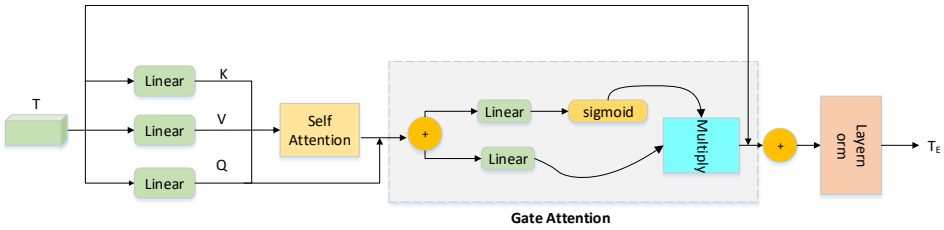


Fig. 3. The Encoder Architecture

features between words in the same sentence. It is easier to capture long-distance interdependent features in sentences, ensuring that important words in the question will be given greater weight. For example, for the question, "Where is the child sitting?", more attention should be attended to the words "child" and "sitting" and therefore the model has the ability to focus on the relevant regions and infer the correct answer. The input of the encoder is language feature $T = [t_1, t_2, t_3, t_4, ...t_m] \in R^{n \times d_t}$ and output a group of attended features $T_s a \in R^{n \times d}$.First, T is transformed into keys, values, and queries which are of the same shape through three independent fully connected layers. They are represented by $T_K, T_V, T_Q \in R^{n \times d}$. The multi-head attention in Equation (2) is used in the self-attention in the encoder.

$$T_K = Linear_k(T) \tag{6}$$

$$T_Q = Linear_q(T) \tag{7}$$

$$T_V = Linear_v(T) \tag{8}$$

Where Linear represents a fully connected layer, and d represents the same dimension of the transformed features in the language modality. The process of obtaining T_{sa} is as Equation (9):

$$\begin{aligned} T_{sa} &= Concat(head_1, head_2, ..., head_h)W^o \\ head_i &= f_d(T_Q W_i^Q, T_K W_i^K, T_V W_i^V) \\ f_d(Q, K, V) &= softmax(\frac{T_Q T_K^T}{\sqrt{d_k}})T_V \end{aligned} \tag{9}$$

Then T_{sa} is fused with the query Q in the original feature, and the fused feature is T'_{sa}.

$$T'_{sa} = Concat(T_{sa}, T_Q), T'_{sa} \in R^{n \times d} \tag{10}$$

The result T'_{sa} reflects the relationship between the words in the question, but in the process of learning self-attention, even if there is no related vector in the query, self-attention will generate a weighted vector, which makes the model confusing. Therefore, this article inputs the results of self-attention into the gate attention layer to get expected useful attended information.

The fused feature T'_{sa} is used as the input of the gate attention layers, which is input to the two linear layers to calculate the attended information vector A_E, which can filter out irrelevant results from attention. The calculation of it is as following, where σ is a non-linear sigmoid function and $FC_q^g, FC_f^g \in R^{d\times d}$.

$$A_E = \sigma(FC_q^g(T_Q) + FC_f^g(T_{sa})) \bigodot (FC_q^g(T_Q) + FC_f^g(T_{sa})) \tag{11}$$

After the gate attention layer, it is then fused with the original features, and Layernorm [5] is used to obtain the question features $T_E \in R^{n\times d}$ that after the encoder module. T_E is obtained by Equation (12). Layernorm plays a role of regularization and the model applying Layernorm is more stable.

$$T_E = Encoder(T) = Layernorm(A_E + T_k) \tag{12}$$

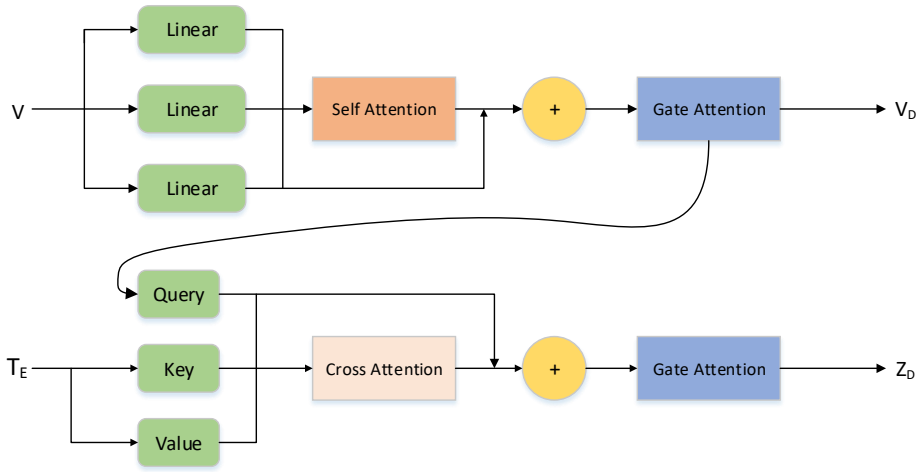


Fig. 4. The Decoder Architecture

Decoder The details of the decoder are shown in Fig 4. The decoder not only pays attention to the relationship between each visual area of each image, but also the connection between each image area and the words in the question, such as the question "What is the man holding with the left hand?", the decoder should focus on the visual area of the individual's left hand, and the relationship between the visual region of the left hand and the object.

The input of self-attention in the decoder is the visual feature $V = [v_1, v_2, v_3, v_4, ..., v_m] \in R^{m\times d_v}$. For an image, the relationship between different visual regions is different, so the weights should be different. Through stacking several

decoders, we can model the relationship of objects in images. Similar to the process of calculating the language feature T_E, we can get V_D which gets through the self-attention layer and gate attention and use it as part of the cross-modal attention input.

Cross-modal attention uses a question's semantic features to guide the attention distribution of each region of the image while using gate attention to filter out unrelated attention results. Cross-modal also uses multi-head attention. Different from self-attention is that its input is the text feature T_E obtained by the encoder and the visual feature V_D obtained by the self-attention and gate attention calculations. The feature Z_{Ca} obtained by the cross-modal attention is calculated by Equation (16). The matrix Z_Ca captures the importance between each object region and the word. The computational procedure the gate attention resembles that of the encoder. After the decoder module, the multi-modal feature Z_D can be obtained.

$$Z_K = Linear_k(T_E) \tag{13}$$

$$Z_Q = Linear_q(T_E) \tag{14}$$

$$Z_V = Linear_v(T_E) \tag{15}$$

$$\begin{aligned} Z_{ca} &= Concat(head_1, head_2, ..., head_h)W^o \\ head_i &= f_d(QW_i^Q, KW_i^K, VW_i^V) \\ f_d(Q,K,V) &= softmax(\frac{Z_Q Z_K^T}{\sqrt{d}})Z_V \end{aligned} \tag{16}$$

3.3. Feature Fusion and Answer Generation

After stacking several encoders and decoders, the visual features $V' = [v_1, v_2, v_3, ..., v_x]$ and language features $T' = [t_1, t_2, t_3, ..., t_y]$ contains rich image and text information. For the two features, first apply a multi-layer perceptron (MLP) with ReLu nonlinear activation function, then the softmax function is devoted to obtain the attention weights which is relevant to the image and the question and finally weight the image and question features from all regions through these attention weights. The Relu activation function makes the output of some neurons zero, which makes the neural network sparse, reduces the interdependence of parameters, and relieves the occurrence of the over-fitting problem. The final weighted sum as the final visual and text features V_{attd}, T_{attd} can be described by the following formulas. V_{attd}, T_{attd} are projected to the same dimension by linear layer. Fusing such features adopts concatenation, or element-wise product, or addition. We use feature addition to obtain the final fused feature H, which can get best performance.

$$\tau_v = softmax(MLP(V')) \tag{17}$$

$$\tau_t = softmax(MLP(T')) \tag{18}$$

$$V_{attd} = \sum_{i=1}^{x} \tau_{v_i} v_i \tag{19}$$

$$T_{attd} = \sum_{i=1}^{x} \tau_{t_i} t_i \tag{20}$$

$$H = Layernorm(W_v^T V_{attd} + W_t^T T_{attd}) \tag{21}$$

In this paper, like other existing methods of visual question answering, we treat the VQA task as a multi-label classification task. The fused multi-modal feature H is input into the answer classifier in the answer prediction layer, and the score is standardized using the sigmoid function. It is between 0 and 1, which is used as the probability of the candidate answer. The final answer is the first 5 answers that appear most often and are used as classification labels.

The loss function in this paper refers to the strategy proposed in [38]. A BCE calculation function is used. The binary cross-entropy calculation function is described in Equation (20), where M is the number of training samples, s represents the probability of answer prediction and N is the number of candidate answers.

$$L = -\sum_{i}^{M} \sum_{j}^{M} s_{ij} \log(s'_{ij}) - (1 - s_{ij}) \log(1 - s'_{ij}) \tag{22}$$

4. Experiments

In this part, specific experiments for evaluating the effectiveness of MEDGA are presented.

4.1. Dataset

The experiments in this paper are performed on the VQA v2 dataset [17]. VQA v2 is a human annotated dataset for open-ended VQA. Each image from Microsoft COCO dataset [29] contains question-answer pairs. Compared with the VQA v1 dataset, it contains more examples for training, verifying, and testing. To prevent the improvement of model accuracy caused by over-fitting, each question corresponds to two images in VQA v2, so each question has two different answers. VQA v2 also minimizes some language biases in the VQA v1 dataset. VQA v2 contains 204721 images from the MSCOCO dataset an abstract scene dataset including 50,000 clipart. Each image in the dataset corresponds to three questions, and every question has 10 answers. This article evaluates MEDGA on 200,000 real images, including about 80,000 pictures in the training set, about 40,000 pictures in the validation set, and about 80,000 pictures in the test. 25% of the data in the test set is called test-dev. Examples of questions and answers on VQA v2 are shown in Figure 5. All questions fall into three categories: Yes / No, Number, Other. As in previous studies, this paper trains on the training and validation sets, and the results are verified on test-dev and test-standard. The experimental results include the accuracy of the three categories and overall accuracy.

Fig. 5. Typical example in VQA v2. The question types are Number, Yes / No, Other

4.2. Experimental Setup

Faster R-CNN is used to extract 2048-dimensional visual features. LSTM encode the language features into a 512-dimensional vector , then the two features are embedded into a 512-dimensional vector through a fully connected layer. The dimension of the fused feature is 1024. The self-attention in the encoder and the cross-modal attention in the decoder have 2 multi-head attention with 256 dimensions for each head. The batch size in every epoch is 64. Each training is 13 epochs. During the process of training, Adam optimization algorithm [25] is used with the parameters $\alpha = 0.001$, $\beta_1 = 0.9$ and $\beta_2 = 0.98$. Gradient clipping and dropout are also used in the experiment. The number of encoders and decoders is from 1 to 10. The candidate choices for the answer only retain the answers that appear more than 8 times in the training set, and the size of the answer vocabulary is 3129. All ablation experiments are performed on the validation set. The experiments in this article are implemented using Pytorch. All initialization is the default initialization in Pytorch.

4.3. Evaluation Metric

The answers to questions of the dataset are given by 10 different people which causes every question have different answers, such as “cat” has the same meaning as “kitty”. It cannot be determined which answer is correct. Therefore, to solve the inconsistency of answers, this paper uses the evaluation method proposed by Antol et al[3]. That means a prediction is right if and only if at least three persons have the same answer. The method can be described as Equation (21) where are the answers given by different annotators, a is the predicted answer, C = 10.

$$Accuracy(accuracy) = \frac{1}{C}\sum_{c=1}^{C}\min(\frac{\sum_{1\leq j\leq c, j\neq c}\prod a = a_j}{3}, 1) \tag{23}$$

4.4. Ablation Studies

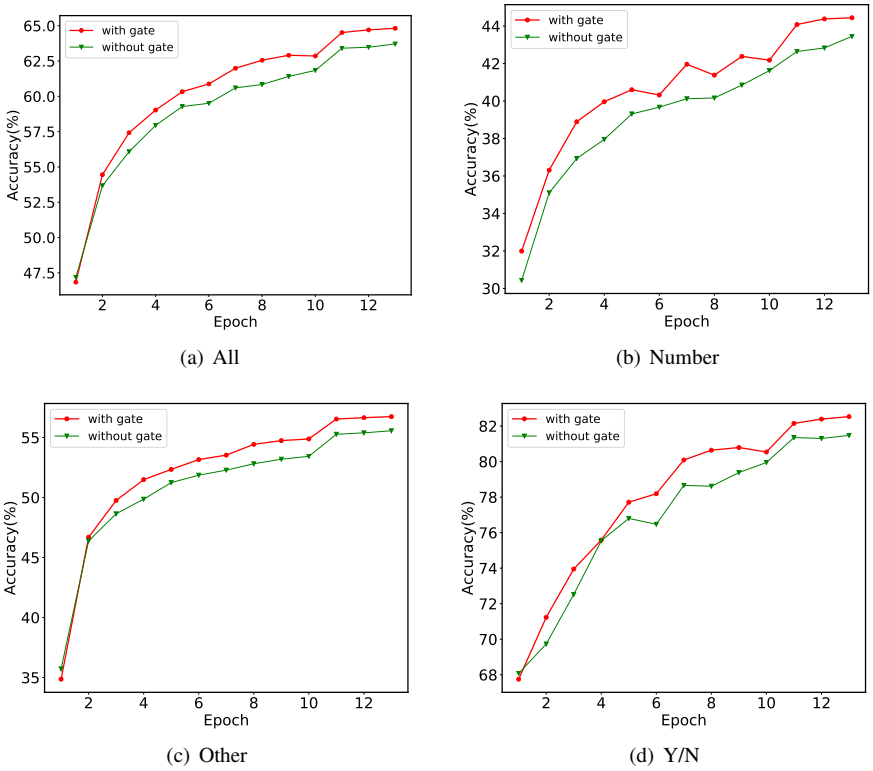
The MEDGA consists of several modules. For testing the impact of each component on the accuracy, this paper conducts several ablation studies under different conditions. With different parameters and settings, different versions of MEDGA are trained on the training

Table 1. Effect of Attention Head

head	All	Y/N	Num	Other
1	64.59	82.22	44.29	56.56
2	**64.82**	82.53	**44.44**	**56.74**
4	64.71	82.55	43.86	56.67
8	64.73	82.44	44.37	56.65

Table 2. The effect of the number of encoders and decoders on the experimental accuracy

Number	All	Y/N	Num	Other
1	64.82	82.53	44.44	56.74
2	65.58	83.13	46.82	57.39
4	66.10	83.41	47.08	57.97
6	66.46	83.94	47.86	58.10
8	**66.63**	**84.10**	**48.42**	**58.16**
10	66.63	84.09	48.83	58.05



(a) All (b) Number

(c) Other (d) Y/N

Fig. 6. The overall and per-type accuracies of the MEDGA along with the variants with gate attention and without gate attention

set of VQA v2, and the effects are displayed on the verification set to verify the effectiveness of MEDGA.

First, we conduct the ablation studies with different number of attention head. To save the run time, the default encoder and decoder are 1. Table 1 shows the influence of multi-head attention. It is found in the experimental results that when h = 2, the overall accuracy is the highest. Too fewer or too many heads reduces the overall accuracy.

Next, we explore the effect of the number of encoders and decoders. Table 2 shows the effect on the overall accuracy when the number of encoders and decoders is 1,2, 4, 6, 8,10. In this experiment, we use 2 head with the best performance As clearly appears from the table 2, when the number is 8, the overall accuracy is the highest. As the number of encoders and decoders continues to increase, the accuracy has not changed much, probably due to over-fitting. Considering the running time, we choose 8 in the final model.

Finally, the effectiveness of gate attention is verified. The default number of encoder and decoder is 1 and head is 2. Figure 6 shows the accuracy of each type in the model with and without gate attention. The model without gate attention only use self-attention and cross-attention. From the results, we can see that MEDGA with gate attention has a better performance than the model without gate attention on overall accuracy. Furthermore, for three categories questions, MEDGA has higher accuracy than without gate attention.

4.5. Visualization of Attention

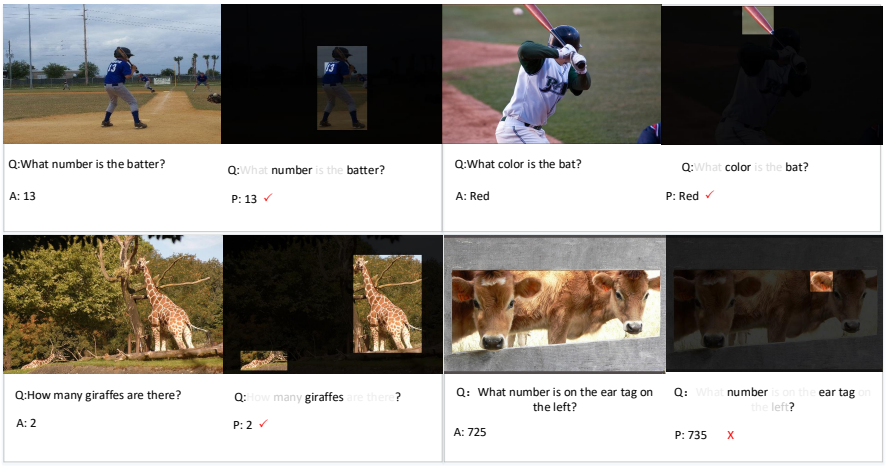


Fig. 7. Typical example in VQA v2. The question types are Number, Yes / No, Other

In Figure 7, four examples of attention visualization are shown, involving types of questions such as color and counting. The left side of each diagram is the input image of the model, and the right side is the learned attention visualization. We observe from Figure 6 that MEDGA highlights the most relevant regions to the question and the most

important words in the sentence. Q is the question asked for the picture, A represents the correct answer, while P denotes the predicted answer by the MEDGA. Among the four visualization examples given, there is an error example. The reason why the model predicts wrong is the model did not give the word "left" enough weight when assigning weights, which caused the model to mislocate in the image. Instead of focusing on the left object, it focused on the right object so that the model gives the wrong answer.

4.6. Comparison with Existing Advanced Methods

Table 3. Compared with some advanced methods on test-dev and test-std. all models are tested on vqa v2 dataset

Model	test-dev				test-std
	Y/N	Num	Other	All	All
BUTD	81.82	44.21	56.05	65.32	65.67
MFH	/	/	/	66.12	/
Graph	82.91	47.3	56.22	/	66.18
ODA-GCN	83.73	47.02	56.57	66.67	66.87
BU+QAA	/	/	/	66.70	67.0
DCN	83.51	46.61	57.26	66.87	66.97
Counter	83.14	51.62	58.97	68.09	68.41
MFH+BUTD	84.27	49.56	59.89	68.76	/
BAN	85.31	50.93	60.26	69.52	/
MDAnet	/	/	/	/	69.74
MEDGA(ours)	**85.97**	51.56	60.09	**69.78**	**70.11**

Table 3 shows the performance of the proposed MEDGA and existing advanced methods on VQA v2, where / indicates that the model has not tested the accuracy of the question type or the dataset. In Table 3, BUTD [2] won the champion in the VQA challenge 2017. It puts forward to use Faster R-CNN to extract features not using ResNet [21]. MFH [47] is a state-of-the-art bilinear pooling method. Graph [33] builds a graph in all the regional propose boxes and conditions this graph on the question.ODA-GCN [49] is also a graph-based visual question answering method, and they introduce a soft attention layer. QAA [12] proposes a question-agnostic attention mechanism that complements existing attention mechanisms. The Dense Symmetric Co-Attention Model (DCN) [32] stacks multiple co-attention modules. Although it does not use Faster RCNN but uses ResNet to extract image features, the experimental results are better than previous advanced methods. Counter [48] makes full use of bounding box information to make it highly accurate in counting type questions. Bilinear Attention Network [23] has 12 stacked bilinear attention modules.MDAnet [13] is a method that uses a multi-modal encoder to replace the RNN in the traditional method, so that the position can be reserved. MEDGA outperforms the

advanced models above on both test-dev and test-std datasets, and it achieves 70.11% accuracy on test-std. On Y/N type, MEDGA perform the best. Although our model is a little worse than Counter on num type, MEDGA outperforms other methods on this type due to the advantages of our model in modeling The overall accuracy of MEDGA on test-dev exceeds the current advanced method BAN by 0.26 percentage points and exceeds DCN by 2.91 percentage points, verifying the performance of cross-modal attention and gate attention. Through the comparison with other methods, we can prove the effectiveness of MEDGA.

5. Conclusion and Future Work

This paper raises and designs a VQA model on the basis of a multi-modal encoder and decoder with gate attention. This model solves the visual question answering task by stacking multiple encoders and decoders. The core of the model is to use self-attention, cross-modal attention, and gate attention. MEDGA pay close attention to major words of the questions and model the relationship between various visual regions in the images. The attention maps obtained only by self-attention and cross-modal attention may have a lot of irrelevant information. The introduced gate attention can solve this problem and make the final attention result more accurate and more conducive to the final answer prediction. The MEDGA presented in this paper is simple and efficacious. The experimental results on the VQA v2 dataset prove the performance of the model. But our model has certain defects in counting and other types. In the future, we will focus on how to make the model more accurate in object detection so that the accuracy of counting is higher. On the other hand, we will study the performance of our model on other datasets, as well as its application in medical images, satellite image recognition, etc. Besides, we will conduct more in-depth research on wireless sensors and combine their wide range of applications with VQA, so that VQA will have a broader application prospect and benefit mankind.

References

1. Ahutu, O.R., El, H.: Centralized routing protocol for detecting wormhole attacks in wireless sensor networks. IEEE Access 8, 63270–63282 (2020)
2. Anderson, P., He, X., Buehler, C., Teney, D., Mark Johnson, S.G., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6077–6086. IEEE Computer Society, Salt Lake City, UT, USA (2018)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433. IEEE Computer Society, Santiago, Chile (2015)
4. B, H., H., Z.: Obstacle-aware fuzzy-based localization of wireless chargers in wireless sensor networks. Electrical and Computer Engineering 43(1), 17–24 (2019)
5. Ba, L.J., Kiros, J.R., Hinton, G.E.: Layer normalization. CoRR (2016)
6. Ben-younes, H., Cadène, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2631–2639. IEEE Computer Society, Venice,Italy (2017)
7. Chen, C., Han, D., Wang, J.: Multimodal encoder-decoder attention networks for visual question answering. IEEE Access 8, 35662–35671 (2020)

8. Cui, M., Han, D., Wang, J.: An efficient and safe road condition monitoring authentication scheme based on fog computing. IEEE Internet Things J. 6(5), 9076–9084 (2019)
9. Cui, M., Han, D., Wang, J., Li, K.C., Chan, C.C.: Arfv: An efficient shared data auditing scheme supporting revocation for fog-assisted vehicular ad-hoc networks. IEEE Transactions on Vehicular Technology PP(99), 1–1 (2020)
10. Devlin, Jacob, M.W.C.K.L., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language, NAACL-HLT. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, MN, USA (2019)
11. Dongfei, Y.: Attention mechanism and high-level semantics for visual question answering. University of Science and Technology of China (2019)
12. Farazi, M.R., Khan, S.H., Barnes, N.: Question-agnostic attention for visual question answering. CoRR (2019)
13. Feng, J., Gong, P., Qiu, G.: Mdanet: Multiple fusion network with double attention for visual question answering. In: Proceedings of The 3rd International Conference on Video and Image Processing. pp. 143–147. ACM, Shanghai, China (2019)
14. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 457–468. The Association for Computational Linguistics, Austin, Texas, USA (2016)
15. Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S.C.H., Wang, X., Li, H.: Dynamic fusion with intra- and inter- modality attention flow for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6639–6648. Computer Vision Foundation / IEEE, Long Beach, CA, USA (2019)
16. Gordo, A., Larlus, D.: Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5272–5281. IEEE Computer Society, Honolulu, HI, USA (2017)
17. Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6325–6334. IEEE Computer Society, Honolulu, HI, USA (2017)
18. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3608–3617. IEEE Computer Society, Salt Lake City, UT, USA (2018)
19. Han, D., Pan, N., Li, K.C.: A traceable and revocable ciphertext-policy attribute-based encryption scheme based on privacy protection. IEEE Transactions on Dependable and Secure Computing PP(99), 1–1
20. Han, D., Yu, Y., Li, K., de Mello, R.F.: Enhancing the sensor node localization algorithm based on improved dv-hop and DE algorithms in wireless sensor networks. Sensors 20(2), 343 (2020)
21. He, Kaiming, X.Z.S.R., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. IEEE Computer Society, Las Vegas, NV, USA (2016)
22. Huang, L., Wang, W., Chen, J., Wei, X.: Attention on attention for image captioning. In: Proceedings of the International Conference on Computer Vision. pp. 4633–4642. IEEE, Seoul, Korea (South) (2019)
23. Kim, J., Jun, J., Zhang, B.: Bilinear attention networks. In: Proceedings of the Conference and Workshop on Neural Information Processing Systems. pp. 1571–1581. Montréal, Canada (2018)

24. Kim, J., On, K.W., Lim, W., Kim, J., Ha, J., Zhang, B.: Hadamard product for low-rank bilinear pooling. In: Proceedings of the International Conference on Learning Representations. OpenReview.net, Toulon, France (2017)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations. San Diego, CA, USA (2015)
26. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. 123(1), 32–73 (2017)
27. Li, H., Han, D.: A novel time-aware hybrid recommendation scheme combining user feedback and collaborative filtering. Mob. Inf. Syst. 2020, 8896694:1–8896694:16 (2020)
28. Li, H., Han, D., Tang, M.: A privacy-preserving charging scheme for electric vehicles using blockchain and fog computing. IEEE Systems Journal pp. 1–12 (2020)
29. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755. Springer, Zurich,Switzerland (2014)
30. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Proceedings of the Conference and Workshop on Neural Information Processing Systems. pp. 289–297. Barcelona, Spain (2016)
31. Nam, H., Ha, J., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2156–2164. IEEE Computer Society, Honolulu, HI, USA (2017)
32. Nguyen, D., Okatani, T.: Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6087–6096. IEEE Computer Society, Salt Lake City, UT, USA (2018)
33. Norcliffe-Brown, W., Vafeias, S., Parisot, S.: Learning conditioned graph structures for interpretable visual question answering. In: Proceedings of the Conference and Workshop on Neural Information Processing Systems. pp. 8344–8353. Montréal, Canada (2018)
34. Qadir, J., Ullah, U., de Abajo, B.S., Bego: Energy-aware and reliability-based localization-free cooperative acoustic wireless sensor networks. IEEE Access 8, 121366–121384 (2020)
35. Ren, F., Zhou, Y.: Cgmvqa: A new classification and generative model for medical visual question answering. IEEE Access 8, 50626–50636 (2020)
36. Shi, Y., Furlanello, T., Zha, S., Anandkumar, A.: Question type guided attention in visual question answering. In: Proceedings of the European Conference on Computer Vision. pp. 158–175. Springer, Munich,Germany (2018)
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations. San Diego, CA, USA (2015)
38. Teney, D., Anderson, P., He, X., van den Hengel, A.: Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4223–4232. IEEE Computer Society, Salt Lake City, UT, USA (2018)
39. Teney, D., Wu, Q., van den Hengel, A.: Visual question answering: A tutorial. IEEE Signal Processing Magazine 34(6), 63–75 (2017)
40. Tian, Q., Han, D., Li, K.C., Liu, X., Castiglione, A.: An intrusion detection approach based on improved deep belief network. Applied Intelligence (3) (2020)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the Conference and Workshop on Neural Information Processing System. pp. 5998–6008. Long Beach, CA, USA (2017)
42. Venugopal, K.R., T., S.P., Kumaraswamy, M.: Qos routing algorithms for wireless sensor networks. Springer (2020)

43. Wu, Q., Shen, C., Wang, P., Dick, A.R., van den Hengel, A.: Image captioning and visual question answering based on attributes and external knowledge. IEEE Trans. Pattern Anal. Mach. Intell. 40(6), 1367–1381 (2018)
44. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning. pp. 2048–2057. JMLR.org, Lille, France (2015)
45. Yang, C., Jiang, M., Jiang, B., Zhou, W., Li, K.: Co-attention network with question type for visual question answering. IEEE Access 7, 40771–40781 (2019)
46. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 21–29. IEEE Computer Society, Las Vegas, NV, USA (2016)
47. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. Neural Networks Learn. Syst. 29(12), 5947–5959 (2018)
48. Zhang, Y., Hare, J.S., Prügel-Bennett, A.: Learning to count objects in natural images for visual question answering. In: Proceedings of the International Conference on Learning Representations. OpenReview.net, Vancouver, BC, Canada (2018)
49. Zhu, X., Mao, Z., Chen, Z., Li, Y., Wang, B.: Object-difference drived graph convolutional networks for visual question answering. Multimedia Tools and Applications (2020)

Haiyan Li is currently pursuing the M.S degree in Shanghai Maritime University, China. Her research interests include visual question answering and deep learning.

Dezhi Han received the B.S. degree in applied physics from Hefei University of technology, China, in 1990, the Ph.D. degree in computing science from the Huazhong University of Science and Technology, China, in 2005. He is currently a Professor with the Department of Computer, Shanghai Maritime University, China. in 2010. His research interests include reinforcement learning and deep learning, wireless communication security, network and information security. He is a member of IEEE.

Received: November 20, 2020; Accepted: February 10, 2021.