Improving Attention and Multimodal Fusion for Visual Question Answering

Guijie Hou, Jiashuai Xiao

Abstract—**Visual question answer(VQA) is a task that combines computer vision and natural language and plays an important role in visual-text interaction tasks. In the last few years, Transformer has begun to appear in a multimodal set of tasks. The emergence of Transformer has advanced the field of artificial intelligence. However, there are certainly drawbacks. For example, fine-grained relationships within modalities are often ignored, poor representation of features and interference of irrelevant information between different modalities. In the inter-modals fusion module, the useful information between different modalities can not be fused.**

In this paper, in order to solve the above problems, we improve the visual attention mechanism based on the MCAN model and enhance the attention features and design a fusion module to improve the fusion of two different modalities. Firstly , we propose a module to enhance the ability of visual attention, which is used to learn the relationship of fine-grained features within the image, discard irrelevant information and enhance the effective information, so as to obtain a more interesting region of the image. Secondly a cross-modal information fusion module is proposed to enhance the interaction of different modal information. The fusion module is the core of the whole network model, through which different modal information is effectively combined to predict the correct answer. The experiments are evaluated on the VQA2.0 dataset, and compared with the existing methods. The method has significant advantages.

Index Terms—**Visual Question Answer,Attentional Mechanisms, Multimodal Fusion**

I. INTRODUCTION

With the rapid development of Visual-Text tasks and Artificial intelligence, the era of multimodality is soon to be ushered in. Such as video analysis, image classificationand description, and visual questioning.The task of Visual question answer [2]play an important role inthe multimodal era, Visual question answer can processboth visual and linguistic features, fusing multimodal features and thus predicting the best answer. Visual question answer involves complex tasks. For example, "Arethere any animals in the image?" This question not only embodies a object detection task[4,6], but also requires common sense reasoning to detect whether the object is an animal or not, so the visual task is very complex. In the last few years, Transformer have emerged and play a very important role in both vision and language. Attention mechanisms[8,10-13,18,20,23,25,28,29] have been proposed for a wide range of applications in unimodal and the above mentioned multimodal tasks. Attention can take many forms and is mainly divided into two parts. visual attention and textual attention, both of which play a crucial role. However, there are many problems that can be further explored and solved in multimodal tasks. such as,how to extract the key features of each modality and how to make the information of different modalities interact better. This cannot be limited to rough interaction modes. In recent years, researchers have explored and proposed a series of methods basedon this problem. For example, Dense Common AttentionNetwork (BUTD)[1] is proposed to greatly enhance the mutual learning of visual question and answer by modeling the fine-grained features between two modalities. However, it has the limitation of focusing only on modeling the relationship between the two modalities, while ignoring the modeling of the relationship within individual modalities, such as between words within a text and between each region within an image. Therefore, we need to capture the key to focus on both intra-modal and inter-modal feature relationships.

The rise of Transformer has also had a significant impact on the field of visual question answer, with a large number of researchers studying it from the Transformer perspective, making it a starting point. Many recent models have been pr oposed based on Transformer[18] with good performance gains. This is mainly attributed to the great role played by the attentional mechanism, as the presence of the attentional mechanism notonly captures the relational dependencies with in the modality, but also facilitates the joint enhancement of vision and language. To solve the above problem, YU et al. proposed a deep Modular Collaborative Attention Network (MCAN)[22],which consists of six stacked MCA layers, each of which includes Self Attention (SA) and Question-Guided Image Attention (SGA). SA is used to learn the intra-modal feature relations, and SGA is used to learn the interactions between different modalities. To further obtain a dense feature representation between visual texts, Modularized Attention Networks (MCAoAN)[15] are proposed, which is an improved extension of the above Modular Collaborative Attention Networks (MCAN) and adds attentional on attention mechanisms to the SA and CA units. Despite the great improvement in the effectiveness of these models and their rather high degree of flexibility, there are still some limitations. It is not sufficient to use only the traditional Transformer module to capture long-dependent features, which lacks the ability to represent visual features, and thus often misses some of the more interestingregional features and regions more relevant to the text.

In order to solve the visual feature representation ability to pay more attention to the important information and improve

Manuscript received August 29, 2024
Guijie Hou,School of Software,Tiangong University,Tianjin,China.
Jiashuai Xiao,School of Software,Tiangong University,Tianjin,China.

the image text fusion. In this paper, a new visual text interaction model (IAMFN) is proposed based on the MCAN[22] model,which solves the above proposed problem of insufficient interaction between multimodal information and the problem of enhancing the representation ability of visual attention.

Our main contributions:

(1)A visual attention mechanism module (VSA) is proposed that can enhance the visual attention mechanism module, which greatly prevents the loss of valid information and facilitates better learning of more interesting regions of an image. Through a large number of experiments surface this design is superior to the traditional Transformer attention unit.

(2) We designed an image text feature fusion module (VLFM), which enhances the interactive fusion of cross-modal information, facilitates the learning of more relevant information between modalities, improves the model generalization ability, and then predicts more accurate answers.

(3) We conduct a validation of the effectiveness of each module and the performance of the overall network model on the VQA2.0 open-source dataset, and the results of a large number of experiments show that the proposed method achieves state-of-the-art performance.

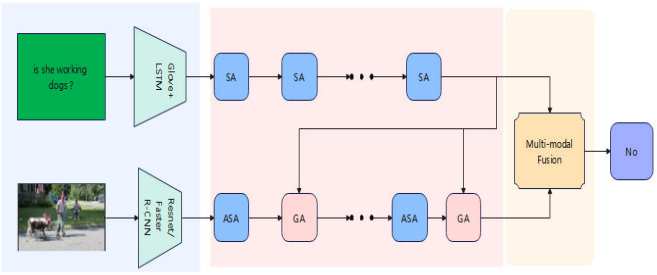

Figure.1, the network architecture (IAMFN) we proposed in our paper, consists of 6 layers of IAMFN layers,each layer includes ASA, SA and GA, ASA represents the augmentedattention module proposed in this thesis, SA and GA represent the traditional self-attention module and guided-attention module, respectively, and finally,we use the VLFM layer to perform the fusion ofthe features between the modalities, and finally output the predicted answer.

II. RELATED WORK

A. Visual question answer Attention

Visual question answer is a task that combines computer vision and natural language understanding, and it is rapidly evolving. Antol et al.[2] first introduced the Visual Question and Answer (VQA) task, which simulates human understanding of a particular visual scene by combining computer vision with natural language processing. The model uses a traditional CNN for image feature extraction and an LSTM for the associated language processing. These features are combined together using simple elemental multiplication and in turn classification is used to predict answers. In the last few years, many visual language models have been proposed to implement visual question and answer tasks. Most of the models[5,17,19] are based on the attention mechanism, which is used to recognize the most relevant information between or within images and text, making the application of the attention mechanism an essential and effective tool. In order to solve the problem of information redundancy and noise due to some global features, Yang et al.[30] designed a multilayer attention network, they learned the image representation related to the question features by executing a visual multilayer attention network, this is the first time that attention is introduced into visual question and answer methods, and good results were obtained. As a result, experiments based on attention were conducted in most of the subsequent studies.Anderson et al.[1] proposed a classical attention model Buttom Up and Top Down (BUTD), which was the first application of Faster R-CNN[16] to image feature extraction. However, it is equally important for the learning of problem features. Later research has considered not only images but also text. Nguyen and Okatani[31] and Kim et al.[32] proposed DCN and BAN methods that focus on inter-modal and intra-modal interactions, respectively, to solve this problem. Enabling model inference is greatly improved. Guo et al.[33] proposed a new multimodal explicit sparse attention network model (MESAN), which focuses the model's attention by explicitly selecting the most relevant text to the image in the input problem in order to reduce the interference caused by irrelevant information and enhance the representation of valid information.

When the Transformer model was first proposed, it was mainly used for natural language processing tasks such as machine translation and text generation. However, with the development of Transformer, some researchers began to experiment with applying Transformer to visual quizzing tasks with some success. Transformer-based models have achieved breakthrough performance on benchmark datasets, which is inextricably linked to the powerful self-attention capability in Transformer.

Currently, Transformer-based methods are categorized into early fusion and late fusion based on the fusion period. The classical model for early fusion is MUAN, which first fuses the features of different modalities and then performs the interaction process, and uses SDPA to model the internal relationships of modalities during the interaction process. The classic models for late fusion are MCAN[22], DFAF[27], etc. The MCAN model applies the scaled dot product (SDPA) attention in Transformer to the textual and visual information streams respectively.The DFAF[27] model textual and visual information streams are fused using self-interaction and cross-interaction respectively.

Many recent researches have been carried out based on the above models, such as the CLVIN[14] model, the MBGAN[33] model, and so on. Their performance is significantly improved compared to the benchmark models. Although good results have been achieved, the ability to express features is still insufficient. In order to solve the above problems, this study also proposes a new model (IAMFN) based on the Transformer architecture, which mainly applies the method of visual attention enhancement to further fuse the output of the attention mechanism with the original input features in two different ways, so that the two

different ways make up for each other to make up for the shortcomings in extracting the features of the important regions, and then obtain the image with the problem of the more relevant regions of the image to the problem. Our improved attention mechanism shows good performance on the VQA2.0 dataset.

B. Visual question answer Fusion

In order to better fuse image and text information, researchers began to explore multimodal representation learning methods,i.e., how to map image and text representations into the same space. A series of models basedon the attention mechanism emerged in this stage, such as the image description generation model with attentionmechanism and the visual question and answer model with attention mechanism.

The multimodal fusion strategy combines the imagefeatures and question features further to improve the generalization ability and robustness of the model, which is helpful to improve the accuracy of answer prediction. Many previous models only use the traditional simple addition or multiplication in the fusion method, which will lose many important information and cannot achieve the desired effect.

Therefore, in this paper, we propose a fusion gating unit that is determined according to the weights, with larger weights representing more important features, which combines image and text fine-grained features more closely and improves the performance of the model, which in turn predicts more accurate answers.

III. MATHOD

We introduce a new model(IAMFN), which optimizes the part of the image attention mechanism that enhances the interaction between any two features of an image. The attention enhancement mechanism compensates for the lack of capturing important information by fusing two fusions with each other, thus discarding invalid information and preventing the excessive loss of valid information. The model fuses the outputs of image features and text features at a finer granularity to improve the generalization ability of the model. Finally, the candidate answers are predicted by the line layer. This paper describes the composition of the modules of IAMFN, which first performs the extraction of visual and textual features, and then inputs the two modal features into the model network for interaction and fusion, respectively, and finally predicts the answer.

A. Visual input features

We use the region image feature as input and the input feature uses ResNext152[36] pretrained on the visual genome[9]. its grid features[7] are first populated to a size of 16×16 and then merged by a kernel size of 2×2 and a step size of 2. The input feature is a transformer with a resolution of 8×2. So the resolution of the transformer is 8×8. using this visual feature as input.

B. Textual input features

First, we label the input questions because the maximum length of each question in the VQA-v2 dataset is 14 words, and only 0.25% of the questions are longer than this length, so the ones larger than 14 can only be discarded, and the ones smaller than 14 make up for the zero. Word embeddings are then performed with a large-scale pretrained Glove[35] word embedding corpus, converting each word into a word vector. We then convert the word vectors into a text feature matrix $X \in \square^{n \times d_x}$ using a single-layer LSTM[34] with dy hidden units, where n denotes the length of the problem.

C. IAMFN Network

Inspired by previous work, this paper proposes an IAMFN network consisting of an IAMF layer.The IAMFN layer consists of three main modular units:visually augmented self-attention (ASA), textual conventional attention (SA) visually-linguistically interacting attention (GA), and fusion unit (VLFM). We applied the proposed augmented attention module to the visual side, the text side and the text-guided image interaction module respectively. The application with the best results was selected as a component of our model. Next, we first in analyzed the computational process of the multi-head attention mechanism[38]. Then, we introduce the IAMFN unit and then the fusion module. Finally, we present our designed IAMFN network.

1)Traditional Attention Module

Attention modules are divided into self-attention modules and guided attention modules, each of which will include a multi-head attention mechanism layer, as shown in Figure.2 Multi-head attention is an extended form of the traditional attention mechanism (Attention), which captures long-distance dependencies and obtains richer and more effective information. Multi-head Attention includes h parallel heads (head) and performs attention computation on each head, and finally stitches together the results of multiple heads to get the final output. Where each head is represented as scaled dot product attention. The input of scaled dot product attention consists of query(Q), key(K) of dimension d_k and value(V) of dimension d_v. Firstly, the similarity score between query, key needs to be calculated, in order to improve the performance of the model, the scaled dot product attention is used here and the dimensions of query, key vectors are equal and both satisfy zero mean and unit variance, the similarity score is calculated and then the scores are normalized using the softmax function. The calculation formula is as follows:

$$f_{att} = f(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d}})V \quad (1)$$

$$f_{att} = MultiHead(Q,K,V) = [head_1, head_2..., head_h]W^o \quad (2)$$

$$head_j = f(QW_j^Q, KW_j^K, VW_j^V) \quad (3)$$

where $,W_j^Q, W_j^K, W_j^V \in \square^{d \times d_h}$ represents the weight matrix of the ith header and $W^o \in \square^{h*d_h \times d}$ is the dimension of the output feature of each header. It is usually set $d_h = d/h$ in order to prevent the dimension from being too large and to improve efficiency.

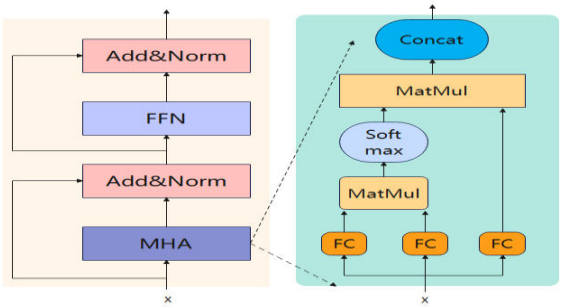

Figure2 Traditional Attention Map

2)Augmented Attention Module (ASA)

In order to solve the problem of frequent loss of effective information in the attention mechanism, this paper proposes an attention enhancement module, which fuses the output of the multi-head attention mechanism with the initial input information in two different ways, and uses the two different fusion methods to make up for each other's deficiencies in capturing information, and to enhance the expression of effective features. We applied the proposed enhanced attention module to visual self-attention module, text self-attention module as in Figure.3 and image-text interaction module as in Figure. 4 for experiments.

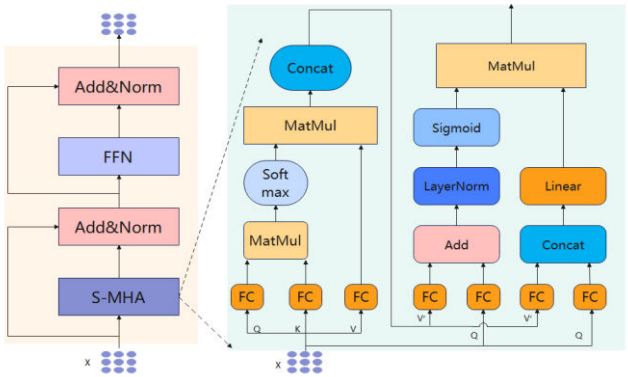

Figure. 3 shows the self-attention augmentation module so that it can be applied to either the visual side or the textual side of the multi-head attention module. S-MHA denotes the self-attention mechanism component.

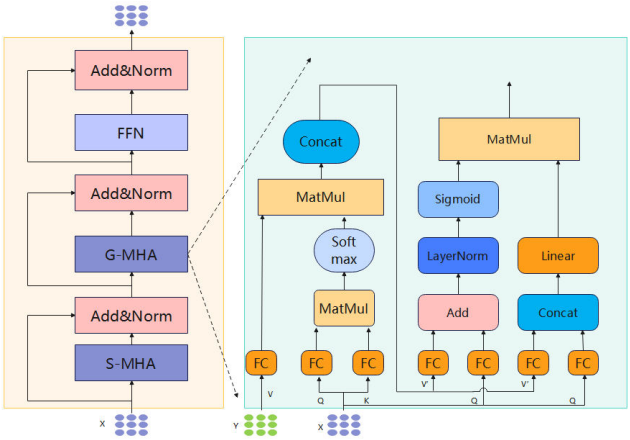

Figure.4 shows the Guided Attention Enhancement Module so that it can be applied to the Image Text Interaction Multi-Headed Attention Module.G-MHA denotes the Guided Attention Mechanism section.

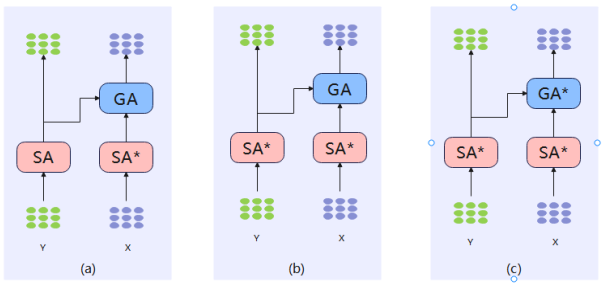

Figure. 5 Attention module(ASA) applied to each of the three cases where the attention module is not used. The ones with * indicate that the traditional attention mechanism is replaced with the enhanced attention proposed in this thesis. In this paper, the validation of the above three cases is given in the ablation experiment section, and the experimental results are shown in Table 3.

The enhanced attention mechanism proposed in this paper extends the multi-head attention mechanism to play a more effective role. We first get the output of the input features through the multi-head attention module proposed above, denoted as V', which V' will be mapped to the same space as Q for the operation, and the two different modal information will be fused in two different ways, namely summing and splicing, respectively, so as to make up for each other's deficiencies. A relationship is established between the two results of the fused output. The specific formula is as follows:

$$V' = MultiHead(Q, K, V) \tag{4}$$

$$I = sigmoid(LayerNorm(W_I^Q Q + W_I^{V'} V')) \tag{5}$$

$$G = Concat(W_G^Q Q, W_G^{V'} V') W' \tag{6}$$

Where, W_I^Q, $W_I^{V'}$, W_G^Q and $W_G^{V'} \in \mathbb{R}^{d \times d}$ represent the weight matrices, in Eq. (5) V' will be summed up after Q and are respectively subjected to linear changes, and the results are normalized by the LayerNorm layer[3], which is used to improve the performance and generalization of the model. In Eq. (6) W' represents the weight matrix after concatenating Q and V' for linear change to speed up the model convergence. Finally, the attention output information is obtained by multiplying I and G through element-by-element multiplication. The formula is as follows:

$$\tilde{I} = I \odot G \tag{7}$$

3) Multimodal Fusion Gate Unit

The model outputs image and text information through the self-attention unit and the modal interaction unit,and fuses[26] the information from the two modalities.In order to further strengthen the close integration between modalities, we design a fusion gating unit, as shown in Figure. 4 Firstly, the fine-grained representations of the two features are respectively enhanced to produce two modal feature representations that contain important information and discard invalid information. Text features are $X_L = [x_1, x_2, ..., x_m] \in \mathbb{R}^{m \times d}$ .Image features are $Y_L = [y_1, y_2, ..., y_m] \in \mathbb{R}^{n \times d}$. The outputs of the six attention

layers X_L and Y_L are output through the MLP layer as follows:

$$X' = \sum_{i=1}^{m} Soft\max(MLP(X_L))x_i \quad (8)$$

$$Y' = \sum_{i=1}^{n} Soft\max(MLP(Y_L))y_i \quad (9)$$

Where m and n denote the length of each image feature and text feature respectively, and d represents the dimension of each image or text feature.

They are then generated by the cross-modal fusion module proposed in this paper. Image features Y' are used X' as the primary modality and Text features are used as the secondary modality. They are fused to generate the predicted answers through a linear layer. The specific formula is as follows:

$$X'' = \mathrm{Tanh}(W_{X'}X') \quad (10)$$

$$Y'' = \mathrm{Tanh}(W_{Y'}Y') \quad (11)$$

$$S = concat(X'', Y'')W_S \quad (12)$$

In the above equation, $W_{X'}$, $W_{Y'}$ and W_S represent the weight matrix, respectively, and Tanh is an activation function used to accelerate model convergence and improve generalization. The output feature S is obtained by splicing and fusion of X'' and Y'' .

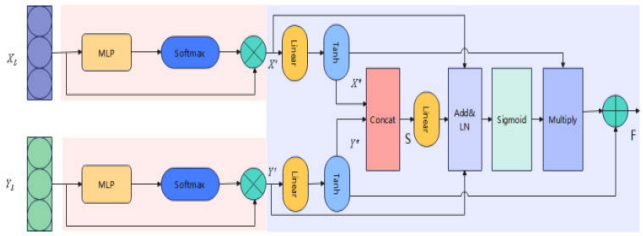

Figure 6 Cross-modal fusion module

Then the output features S and the fine-grained features X', Y' are fused and normalized by the LayerNorm layer to get the output features S' ,the specific formula is as follows:

$$S' = LayerNorm(X' + Y' + S) \quad (13)$$

And S' will be passed to the weight of 0-1 by the sigmoid activation function, and the text output features X'' of Tanh activation function will be multiplied with this weight to get a richer and more effective representation of the image information, and finally the text information will be fused with the image information Y'' to get the final fusion output feature F.

$$F = X'' \odot sigmoid(S') + Y'' \quad (14)$$

D. Network framework

The cross-modal attention network layer is shown in Figure. 7, including the 6-layer IAMF network and the Visual Text Fusion Module (VLFM), which finally predicts the multimodal fusion feature F to predict the answer through the LayerNorm and the linear layer.

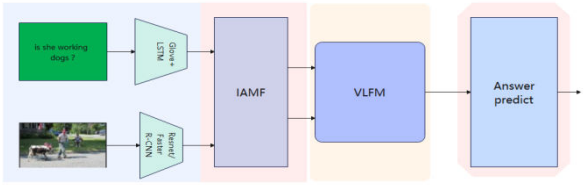

Figure 7 Network Architecture

IV. EXPERIENCE

We conducted a number of experiments on the VQA2.0[37] dataset for validating the effectiveness of our proposed IAMFN model. In this section, we first describe the dataset used in the experiments and the metrics for evaluating the experimental results. Then the experimental details including the specific experimental setup are described. We then perform some ablation studies to compare the performance of different variants of our model. After that, we performed experiments on the test development set and the test standard set to compare our valid results with other classical models. And finally we provide a visualization of the model.

A. Dataset

The VQA dataset is derived from the MSCOCO dataset (Lin et al 2014) and consists of two parts: real i-mages and abstract images. Each image has three questions, and each question contains 10 answers and confid-ence from different annotators. As a detail, the questionsare asked by humans.The VQA 2.0 [37]dataset is divided into training, validation, and testing splits as they ha-ve82783, 40504, and 81434 images and 443757, 214354,and 447793 questions, respectively.There are three typesof answers in VQA 2.0: yes/no, numeric, and other. In order to evaluate the proposed method, we have selectedand used VQA 2.0 for training and testing.

B. Evaluation of indicators

The VQA2.0 dataset can be defined as a multi-categorization problem. Therefore, we follow the same evaluation criteria as the benchmark method in VQA 2.0. In addition, ten different annotators give answers to each question in the VQA 2.0 dataset. Therefore, the predicted answers can be evaluated through a voting mechanism, which is represented by the equation:

$$acc(ans) = min\left\{\frac{\#humans\ that\ said\ ans}{3}, 1\right\} \quad (15)$$

where ans is the answer predicted by the VQA model.

C. Experimental details

The dimension of the visual input features is 2048, the dimension of the question features is 512, the dimension of the fusion features is 1024, the dimensions of the self-attention and FFN are 512 and 2048, respectively, and the number of attention headers is 8, with each header having a dimension of 64. In addition, the number of the set of candidate answers A is N = 3129. let's set the layer k = 6.

During the training of the model, the Adam optimizer was applied with parameters β_1 and β_2 of 0.9 and 0.98, respectively. the basic learning rate was set to $min(2.5te^{-5}, 1e^{-4})$; where t is the current number of iterations from 1. The learning rate was set to; where t is the current number of iterations from 1. After ten iterations, the learning rate decayed by 0.2 every two iterations and the size of each batch was set to 64 and up to 13 periods. The training set used in this paper includes not only the training and validation segmentations, but also a subset of the visual genome as an augmented dataset to facilitate training and improve the generalization of the model.

D. Ablation Experiments

We conducted a large number of experiments to verify the effectiveness of the proposed method, and the results of this experiment are displayed in Tables 1 and 2.

1)Number of Layer (L)

the layer consists of multiple A-units stacked together, as can be seen in Table 1, the number of layers of our model is increasing while the accuracy is increasing, implying that the performance of the model is also increasing up to 6 layers. After that the performance is gradually saturated and we perform a large number of experiments using L = 6 in our final model, here using the validation set for the experiments.

Table 1: Experimental results with different L. Here we use a range of values from 2 to 8 on validation set. Best performance is achieved with L = 6. Therefore, in this paper we choose L = 6 for our work.

	Overall	Y/N	Num	Other
2	67.20	84.52	49.15	58.79
4	67.79	85.36	49.92	59.16
6	67.98	85.46	50.13	59.40
8	67.97	85.56	50.13	59.31

Table 1

2)Testing the effectiveness of each module

In this paper, two important module units are proposed: (1) Attention Enhancement Module (2) Cross-modal Fusion Module, which lifts the interference of irrelevant information between different modal information, prevents the excessive loss of valid information, and tightly combines the image features and the text features to obtain the closest features between different modalities. We verify the effectiveness of each independent module proposed in the paper by conducting experiments using the VQA 2.0 validation set, and the results of this experiment show that the individual modules have a very important impact on the enhancement of the experimental effect of VQA.

Table 2: Effectiveness of each of the proposed modules. We conducted experiments using the VQA2.0 validation set to verify the impact of each of the proposed modules on the performance of the model, and the results in Table 2 show that it is beneficial to improve the performance of the model.

	Overall	Y/N	Num	Other
Baseline	67.45	85.26	49.27	58.72
+ASA	67.84	85.51	49.64	59.21
+VLFM	67.48	85.05	50.02	58.74
+ASA+VLFM	67.98	85.46	50.13	59.40

3)Impact of different attention unit combination methods

the enhanced attention module we propose in this paper can be applied to both images and text, i.e., the traditional part of the multi-head attention mechanism is replaced by using the attention module proposed in this paper. We conduct ablation experiments on the VQA2.0 validation set to verify the effects of the image attention enhancement module, the text attention enhancement module and the interaction attention enhancement module on the model performance. The results in Table 3 indicate that using the self-attention enhancement unit only at the visual end works best.

Table3: In the above table, Base represents the use of the default multi-attention mechanism and the newly proposed image-text fusion module.AVA represents the replacement of the image attention part with the augmented attention unit, ATA represents the replacement of the text attention part with the augmented attention unit, and AGA represents the replacement of the image-text interactive attention part with the augmented attention unit. (Each experiment is performed on the basis of using the newly proposed fusion module). The results of the three experiments except Baseline in the above table correspond to parts (a), (b), and (c) of Figure. 5 species, respectively.

	Overall	Y/N	Num	Other
Baseline	67.45	85.26	49.27	58.72
+VA	67.98	85.46	50.13	59.40
+VA+TA	67.85	85.45	49.96	59.19
+VA+GA	67.69	85.07	49.96	59.15
+VA+TA+GA	67.93	85.50	50.19	59.26

E. Experience Result

We evaluated our model on the VQA2.0 dataset using both the test-dev and test-dev sets and compared it to existing state-of-the-art methods. Table 4 shows the results using online evaluation of test-dev and test-std. The offline evaluation only supports validation splitting. We compare the method proposed in this paper with the state-of-the-art SOTA, including the most classical baseline models and top conference models in the field of visual quizzing such as BUTD[1],MCAN[22],MCAoAN[15]and MBDGAN[40], etc. In addition, our results are not compared with the pre-trained models because they introduce a large number of additional datasets for training the models. This can lead to unfair results.

In Table 4 , our approach has significant improvement over SOTA in this domain, and our model outperforms the test development set by 0.43% and the test standard set by 0.53% compared to BASELINE, where IAMFN-R denotes the use of regional features, and IAMFN-G denotes the use of grid features. So, our model outperforms other models without using large scale pre-training dataset. It shows that the proposed module SAA can get better visual features and the module VLFM can tightly fuse the features between two modalities, image and text.

Model	Test-dev				Test-std
	Overall	Y/N	Num	Other	
BUTD[1]	65.32	81.82	44.21	56.05	65.67
MCAN[22]	70.63	86.82	53.26	60.72	70.90
MCAoAN[15]	70.90	87.05	53.81	60.97	71.16
MMNAS[39]	71.24	87.27	55.68	61.05	71.46
TGAM[41]	71.28	87.47	53.66	61.43	71.60
MBGAN[40]	70.79	85.58	53.52	61.98	71.14
CAT[42]	71.20	87.35	53.17	61.49	71.43
LSAT-G[21]	71.67	87.74	54.51	61.83	71.94
IMCN+IMFN	70.95	87.10	53.74	61.02	71.18
Base	71.39	87.46	53.90	61.62	71.65
IAMFN-R	70.90	87.21	53.04	61.02	71.42
IAMFN-G	71.82	87.79	54.56	62.09	72.18

Table 4

V. Conclusion

In this paper, we propose an enhanced attention network based on MCAN improvement, which consists of an improved visual attention mechanism layer and a proposed inter-modal fusion module to solve the problems in visual question and answer and to improve the performance of the model. Our model enhances the visual attention and inter-modal fusion approach, which has significantly enhanced advantages over the MCAN approach. Our proposed model outperforms existing SOTA methods on the VQA2.0 benchmark dataset.

References

[1] Anderson P, He X, Buehler C, et al. (2018) 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
[2] Antol S, Agrawal A, Lu J, et al. (2015) Vqa: visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2425–2433
[3] Ba JL, Kiros JR, Hinton GE (2016) Layer normalization. arXiv preprint arXiv:1607.06450
[4] Carion N, Massa F, Synnaeve G, et al. (2020) End-to-end object detection with transformers. In: European conference on computer vision, pp 213–229
[5] Gao H, Mao J, Zhou J, et al. (2015) Are you talking to a machine? Dataset and methods for multilingual image question answering. In: Neural Information Processing Systems. MIT Press, pp 2296-2304
[6] Hu H, Gu J, Zhang Z, et al. (2018) Relation networks for objectdetection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3588–3597
[7] Jiang H, Misra I, Rohrbach M, et al. (2020) In defense of grid features for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10,267–10,276
[8] Kim JH, Jun J, Zhang BT (2018) Bilinear attention networks. In:Neural Information Processing Systems, pp 1571–1581
[9] Krishna R, Zhu Y, Groth O et al (2017) Visual genome: Connecting language and vision using crowdsourced dense image annotations Int J Comput Vis 123(1):32–73
[10] Liu Y, Zhang X, Zhang Q et al (2021) Dual self-attention with coattention networks for visual question answering. Pattern Recognit 117(107):956
[11] Liu Y, Wei W, Peng D et al (2022) Depth-aware and semantic guided relational attention network for visual question answering. IEEE Trans Multimed 2022:1–14
[12] Lu J, Yang J, Batra D, et al. (2016) Hierarchical question-image co-attention for visual question answering. In: Neural information processing systems, pp 289–297
[13] Mao A, Yang Z, Lin K et al (2022) Positional attention guided transformer-like architecture for visual question answering. IEEE Trans Multimed 2022:1–13
[14] Chen, Chongqing et al. "CLVIN: Complete language-vision interaction network for visual question answering", Knowledge-based systems 275 (2023)
[15] Rahman T, Chou SH, Sigal L, et al. (2021) An improved attention for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1653–1662
[16] Ren S, He K, Girshick RB, et al. (2015b) Faster r-CNN: towards real-time object detection with region proposal networks. In: Neural information processing systems, pp 91–99
[17] Teney D, Anderson P, He X, et al. (2018) 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
[18] Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. In: Neural information processing systems, pp 6000–6010
[19] Zeng Y, Zhang X, Li H (2022) Multi-grained vision language pre-training: aligning texts with visual concepts. In: International Conference on Machine Learning, PMLR, pp 25,994–26,009
[20] Zhang S, Chen M, Chen J et al (2021) Multimodal feature-wise coattention method for visual question answering. Inf Fusion 73:1–10
[21] X. Shen, D. Han, Z. Guo, C. Chen, J. Hua, G. Luo, Local self-attention in transformer for visual question answering, Appl. Intell. (2022) 1–18,
[22] Yu Z, Yu J, Cui Y, Tao D, Tian Q (2019) Deep modular coattention networks for visual question answering. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 6281–6290
[23] Guo Z, Han D (2022) Sparse co-attention visual question answering networks based on thresholds. Appl Intell :1–15
[24] Chen C, Han D, Chang C-C (2022) CAAN: Context-aware attention network for visual question answering. Pattern Recogn 132:108980
[25] Shen X, Han D, Chang C-C, Zong L (2022) Dual self-guided attention with sparse question networks for visual question answering. IEICE Trans Inf Syst 105(4):785–796
[26] Nguyen D-K, Okatani T (2018) 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 (CVPR), pp 6087–6096
[27] Gao P, Jiang Z, You H, Lu PC, Hoi S, Wang X, Li H (2019) Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 6639–6648
[28] Qin B, Hu H, Zhuang Y (2022) Deep residual weight-sharing attention network with low-rank attention for visual question answering. IEEE Transactions on Multimedia
[29] Cadene, R.; Ben-Younes, H.; Cord, M.; Thome, N. Murel: Multimodal relational reasoning for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 1989–1998.
[30] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 21–29, 2016.
[31] D. Nguyen, T. Okatani, Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, 2018, pp. 6087–6096,
[32] J. Kim, J. Jun, B. Zhang, Bilinear attention networks, in: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal,Canada, 2018, pp. 1571–1581
[33] Guo Z, Han D (2020) Multi-Modal Explicit sparse attention networks for visual question answering. Sensors 20(23):6758
[34] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. Neural Computation, 1997.

[35] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In EMNLP, 2014.
[36] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik LearnedMiller, and Xinlei Chen. In defense of grid features for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10267–10276, 2020.
[37] Yash Goyal, Tejas Khot, Douglas Summersstay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In IEEE Conf. Comput. Vis. Pattern Recog., 2017.
[38] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian.Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6281–6290,2019.
[39] Z. Yu, Y. Cui, J. Yu, M. Wang, D. Tao, Q. Tian, Deep multimodal neural architecture search, in: MM '20: The 28th ACM International Conference on Multimedia, Virtual Event/Seattle, WA, USA, October 12–16, 2020, 2020, pp. 3743–3752,
[40] Linqin C, Nuoying X, Hang T, Kejia C, Haodu F, et al. Multimodal Bi-direction Guided Attention Networks for Visual Question Answering[J], NEURAL PROCESSING LETTERS, 2023, 55(9): 11921-11943.
[41] H. Zhang and W. Wu, "Transformer gate attention model:An improved attention model for visual question answering,"in 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022, pp. 1–7.
[42] Haotian Zhang, Wei Wu. CAT: Re-Conv Attention in Transformer for Visual Question Answering.[C], International Conference on Pattern Recognition, 2022: 1471-1477.