Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.038598
Article

Improved Blending Attention Mechanism in Visual Question Answering

Siyu Lu[1], Yueming Ding[1], Zhengtong Yin[2], Mingzhe Liu[3,*], Xuan Liu[4], Wenfeng Zheng[1,*] and Lirong Yin[5]

[1]School of Automation, University of Electronic Science and Technology of China, Chengdu, 610054, China
[2]College of Resource and Environment Engineering, Guizhou University, Guiyang, 550025, China
[3]School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou, 325000, China
[4]School of Public Affairs and Administration, University of Electronic Science and Technology of China, Chengdu, 611731, China
[5]Department of Geography and Anthropology, Louisiana State University, Baton Rouge, 70803, LA, USA
*Corresponding Authors: Mingzhe Liu. Email: liumz@cdut.edu.cn; Wenfeng Zheng. Email: winfirms@uestc.edu.cn
Received: 20 December 2022; Accepted: 10 April 2023; Published: 26 May 2023

Abstract: Visual question answering (VQA) has attracted more and more attention in computer vision and natural language processing. Scholars are committed to studying how to better integrate image features and text features to achieve better results in VQA tasks. Analysis of all features may cause information redundancy and heavy computational burden. Attention mechanism is a wise way to solve this problem. However, using single attention mechanism may cause incomplete concern of features. This paper improves the attention mechanism method and proposes a hybrid attention mechanism that combines the spatial attention mechanism method and the channel attention mechanism method. In the case that the attention mechanism will cause the loss of the original features, a small portion of image features were added as compensation. For the attention mechanism of text features, a self-attention mechanism was introduced, and the internal structural features of sentences were strengthened to improve the overall model. The results show that attention mechanism and feature compensation add 6.1% accuracy to multimodal low-rank bilinear pooling network.

Keywords: Visual question answering; spatial attention mechanism; channel attention mechanism; image feature processing; text feature extraction

1 Introduction

A complete visual question answering model typically includes the attention mechanism, multi-modal feature fusion, and answer generation. The model is typically improved through the use of attention mechanisms [1] and feature fusion [2].

The study of attention is inspired by human visual attention. When people observe an image, they do not pay attention to all the pixels of the entire image equally but focus on the part that is most interesting at the moment. The force is also affected by the content seen before, and there is a

state transition relationship between the attention of different content. Inspired by this phenomenon, Mnih et al. [3] proposed an attention mechanism based on a recurrent neural network. The model focus on the image region of interest through the attention mechanism and the current learning state is affected by the state of the previous step. The model no longer processes the entire image equally but only processes part of the area, reducing the computational complexity and parameter storage.

The attention mechanism is a commonly used optimization method in neural network models. Some researchers proposed an attention pooling network [4], an ABCNN [5] network, and other networks in the question-answering system to weigh the features of two sentences so that the corresponding parts of the two sentences stand out. The earliest Xu et al. [6] applied the attention mechanism to image caption generation. Zhu et al. [7] applied the spatial attention mechanism to the LSTM network, using the features of the words extracted from the question to weigh the image with the attention mechanism. However, Chen et al. [8] added the features of the question sentence into the image attention mechanism and weighted the relevant image region information by the input question information. Later, Xu et al. [9] proposed a step-by-step attention mechanism processing method. Later, Shih et al. [10] proposed to perform target detection on the picture to obtain the picture area information and then select the obtained picture area through the question. And Lu et al. [11] proposed a multi-level coordinated attention model (hierarchical co-attention model, HieCoAtt), which first extracts image features and question features and uses the image features in the attention mechanism of the question. In contrast, text features are used in the attention mechanism of pictures, and vice versa. Then the image and text features weighted by attention are processed for visual question answering classification.

The attention mechanism applied to deep learning is equivalent to the weighting of features, and feature regions with different degrees of importance have different weights [12,13]. In the traditional convolutional neural network image classification task, the complete image is used as input for feature extraction and then operation. The problem is that the data is too large and redundant, and many irrelevant data are involved in the operation. The attention mechanism was proposed to reduce the existence of this redundancy [14,15]. There are several different consideration angles for image visual attention mechanism methods: attention mechanism in the spatial domain [16–18], attention mechanism in the channel domain [19,20], and hybrid attention mechanism [21].

Among them, the attention mechanism of the spatial domain is mainly for the work done on the linear change of the image space. The target objects in the image will have different spatial positions, rotation angles, and size differences. The human eye can easily handle these differences but not the neural network. Since the convolutional neural network treats all pixels equally in image processing, the spatial distribution angle of the object will significantly impact the recognition result. Jaderberg et al. [22] proposed an attention mechanism in the spatial domain for irrelevant, redundant spatial features.

The attention mechanism based on the channel domain is mainly inspired by the principle of signal decomposition. A complex signal can be decomposed into the representation of multiple simple signals, and the image in deep learning can be understood as the signal is decomposed after convolution with different convolution kernels. To different channels, and the features on each channel are equivalent to the decomposition amount of the image on the convolution kernel. The channel domain attention mechanism method weighs the entire layer features equally on the channel-based feature layer, and the weighting adopts a soft attention mechanism. The channel domain attention mechanism is intended to solve the diversity problem of human attention, such as people notice not only the shape of objects but also colorful and bright areas when observing pictures. Since

the attention mechanism is more complex due to other non-graphic requirements, a channel-based attention mechanism is proposed to solve this problem.

The typical model of the channel attention mechanism is the SEN network model proposed by Hu et al. [23], which won the championship in the ImageNet competition. The SEN model adds a series of network structures based on feature layers into the classical neural network and weights the features of different channels in the middle. After training, good results are obtained in the experiment. Combined with a classic deep neural network, it only needs to be inserted after the convolution module.

Although the attention mechanism of the channel domain makes up for the neglect of the channel information by the spatial attention mechanism, the attention mechanism of the channel domain also has corresponding problems. It ignores the feature information in space and only adds to the deep neural network. It is simple to destroy deep feature information when the attention weight is used. To combine the advantages of the two and make up for the shortcomings of each other, Wang et al. [24] proposed a mixed-domain attention mechanism. This method draws on the idea of a residual network, adds the original image features to the attention mechanism parameters of the channel domain through additional link channels, supplements the spatial information for the channel attention mechanism, and ensures the transmission performance of deep network parameters. Thus, it makes up for the lack of channel attention mechanism to a certain extent.

Inspired by the visual attention mechanism, scholars have also added the attention mechanism in natural language processing. The addition of the attention mechanism has improved the performance of machine translation and question answering systems in natural language. In natural language processing, the initial processing method is the seq2seq model. That is, the natural language sequence is encoded by a recurrent neural network and then decoded into a language sequence. This method is mostly used in machine translation and question answering systems. In the traditional Recurrent Neural Network (RNN) model, the hidden state at a certain time during decoding is only related to the hidden layer at the current time and the output value of the previous time. For the recurrent network based on the attention mechanism, the decoding time also needs to depend on the context features.

In addition to the basic attention mechanism, there are many variants of attention mechanism in practical problem processing. For example, in operations involving multi-dimensional tensors, the attention mechanism will generate a high-dimensional attention matrix for the hidden layer attention mechanism under multi-dimensional tensors so as to obtain attention features in different spatial dimensions. In addition, attention mechanisms also include multi-level attention mechanism [25], self-attention mechanism [26], and attention mechanism based on memory representation [27].

In this article, we propose a VQA model using attention mechanism. In the visual dimension, spatial and channel attention are combined, a small part of image features is added as compensation. Self-attention is used in the text dimension, and the internal structural features of the sentence are strengthened. The experimental results show that the accuracy of the model has been improved.

2 Materials

2.1 Parameter Settings

To optimize the neural network, we use the Adam algorithm. The parameter Alpha is set to 0.001, the first-order moment parameter is 0.9, and the second-order moment parameter is 0.999.

We use random numbers for the initialization of network parameters. Although the neural network of feature fusion and answer generation is not deep, it has a high dimension and large

data. Therefore, we cannot just use small random numbers to initialize parameters. We refer to the initialization method recommended by the research of He et al. [28] and others and standardize the variance of random numbers to n/2, where n is the number of inputs, and the initialized parameter value is entered into Eq. (1).

$$w_i = \frac{random\,(n)}{\sqrt{\frac{2}{n}}} \tag{1}$$

2.2 Evaluation Method

In the visual question answering task, this paper mainly studies open-ended questions. The answer to open-ended questions is different from yes-no questions. It has no deterministic answer, so the definition of accuracy differs from general classification questions. As described by the VQA data improver documentation, each open-ended question for each picture in the data has answers from ten different answerers. The officially recommended way of judging is if the predicted answer is within ten. If three people mention the manual answer, the answer is judged to be correct, as Eq. (2).

$$accuracy = \min\left(\frac{humans \quad that \quad provided \quad that \quad answer}{3}, 1\right) \tag{2}$$

2.3 Dataset

Model training and testing are extremely important. The basic resources that need to be used are the data sets. The datasets for VQA generally contain pictures, questions, and corresponding answers. Some datasets also include annotation information of pictures, description information of pictures, multiple options of multiple-choice questions, and so on. The earliest widely used datasets are the DAQUAR [29], COCO-QA [30], and VQA-real [31] datasets, all of which contain real images. The difference between the datasets is mainly due to the different forms of questions. For example, some questions are open-ended, and some are multiple-choice. These different questions will have different response results to the model. In addition, the size of the data set, the deviation of the data, the complexity of the problem, and the different evaluation methods are also the differences between the data sets.

VQA 1.0 is used in this paper, which refers to the Microsoft COCO image data set as the image source, including 248,349 training data, 121,512 validation set data, and 244,302 data for testing, while the questions for images are given manually. Questions are divided into various types, such as yes or no, the answer is only yes or no; open-ended questions, the answer is given by a word or phrase; multiple-choice questions, one of 18 alternative answers is correct, etc. The verification methods of the answers are all given by the official, and it is a question of whether it is a non-issue. In the remaining questions, each question has ten answers given independently by humans, and the critique of whether the answer is correct or not depends on whether there are more than three prediction results. People who give the same answer, if so, the accuracy of the answer can be judged to be 100% correct.

3 Methods

3.1 Improved Hybrid Visual Attention Mechanism

The visual attention mechanism is the earliest application in the visual question answering task, which uses the semantic features extracted from the question as guiding features for the attention weight calculation of image features so as to use the key information in the question to improve the accuracy of finding the answer from the picture. Rate and reduce the amount of computation.

3.1.1 Spatial Domain Attention Mechanism Operation

(1) Down-sampling of image feature. The image features v_i are down-sampled, and the maximum pooling method is generally used to reduce the image features to a low-resolution feature map with relatively strong semantic information to obtain the global image semantic information, as shown in Eq. (3).

$$v_I = \max pooling\,(v_i) \tag{3}$$

(2) Generation of attention weight. The commonly used model is based on multi-layer perceptron. The feature encoding v_Q of the problem is obtained through LSTM, and the image features v_I are obtained through a convolutional neural network and maximum pooling. As shown in Eq. (4), the attention weight calculation is firstly performed the linear operation of graphic features h_Z through the shallow neural network, and then the graphic features are summed and then passed through the soft-max layer.

$$\begin{aligned} h_Z &= \tan h\left(W_{I,Z}v_I \oplus \left(W_{Q,A}v_Q + b_A\right)\right) \\ p_I &= soft\max\left(W_P h_Z + b_p\right) \end{aligned} \tag{4}$$

Among them p_I is the attention vector, which contains the attention weight v_Q corresponding to each image region. $W_{I,Z}$, $W_{Q,A}$, W_P are the weight values that can be obtained through learning. b_A, b_p are the thresholds that can be learned.

(3) Layer adjustment of attention weight. The obtained PI attention weight needs to be restored to the image feature size. The methods used are the up-sampling method or interpolation method. This paper adopts the soft attention mechanism, and the continuous numerical attention weight layer in the spatial domain is obtained after up-sampling.

3.1.2 Channel Attention Mechanism Operation

(1) Calculates the multichannel feature of an image. I is the image feature, suppose $I \in R^{H'\times W'\times C'}$, $V \in R^{H\times W\times C}$, the multi-channel feature V is obtained through the convolution kernel filter operation. As shown in Eq. (5), the image I has a total of C' channels, and the convolutional neural network has a total of C convolution kernels. f_k^s is the sth channel of the corresponding image. The convolution kernel in this module is two-dimensional. One layer is the operation for each channel, $*$ is the convolution operation, and the final feature is $V = \{v_1, v_2, \ldots, v_C\}$.

$$v_k = f_k * I = \sum_{s=1}^{C'} f_k^s * i^s \tag{5}$$

(2) Compress multichannel features. The features at this time are independent of each other. In order to further study the relationship between the features of different channels, the author compresses the features into one channel through global pooling to generate new features. The method is to average from the length and width of the feature, as shown in Eq. (6).

$$z_k = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} v_k\,(i,j) \tag{6}$$

(3) Dimensionality reduction and generalization of features. In the subsequent processing, the activation function is added, and the feature is dimensionally reduced and generalized, as

shown in Eq. (7). Take ReLU as the activation function δ, and then use the fully connected layer to reduce the dimension of the model.

$$P_c = \sigma\left(W_2\delta\left(W_1 z\right)\right) \tag{7}$$

Among them W_1 is the fully connected layer, which contains the decay index r, and its function is to reduce the dimension and generalize the original feature. And W_2 is the final output layer, and its role is to rearrange the dimensions to the dimensions required for combining with the features.

3.1.3 Attention Mechanism Mixing and Improvement

After obtaining different attention mechanism weights, we perform mixed attention weighting on image features and apply the obtained spatial attention and channel attention weights to multi-scale image features. For the loss of feature information, we added the original features in a certain proportion, where the image features are H_I, such as Eq. (8).

$$H_{Att} = A\left(P_I + P_c + b\right) * H_I \tag{8}$$

where P_I and P_c are the spatial attention mechanism and the channel attention mechanism, respectively, and b is the tiny weight added to prevent image features from being damaged after being multiplied by the attention mechanism. The final improved attention module is shown in Fig. 1.

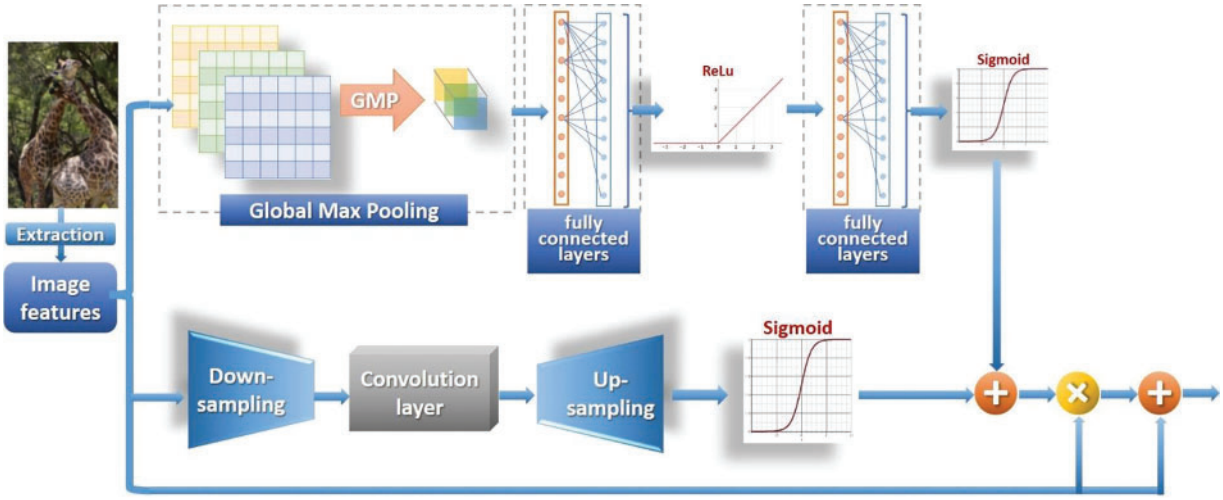

Figure 1: Hybrid attention module

3.2 Text Processing Based on Self-Attention Mechanism

The early attention mechanism of the visual question answering system did not consider the question text. It was not until the study of Lu et al. [11] that the text attention mechanism was added to the visual question answering model. But Lu's research is mainly about the interaction between image attention and text attention and does not consider that the text features can be optimized by the attention mechanism.

The self-attention mechanism is a method for characterizing the internal relationship of sentence sequences. It has been successfully used in text information understanding, article content summarization, and text information focus extraction. This paper will introduce the self-attention mechanism

into text feature processing to extract and emphasize the key features of sentences. The self-attention mechanism method can be divided into the following steps:

(1) Embedding layer, since sentences are unstructured data, the sentences in the visual question answering are first mapped to the vector space before the calculation, as shown in the figure—dotted box. The mapping first maps words to discrete feature vectors via one-hot encoding and then converts them to continuous feature vectors via word2vec.

(2) Encoding layer, in visual question answering, the usual processing method is to encode the sentence through a convolutional neural network and then decode the encoded features in the answer generation stage to obtain the natural language output, as shown in Fig. 2. This paper's visual question answering framework is based on a bilinear feature fusion network, which only needs to encode sentence features and the feature sequence through a recurrent convolutional neural network.

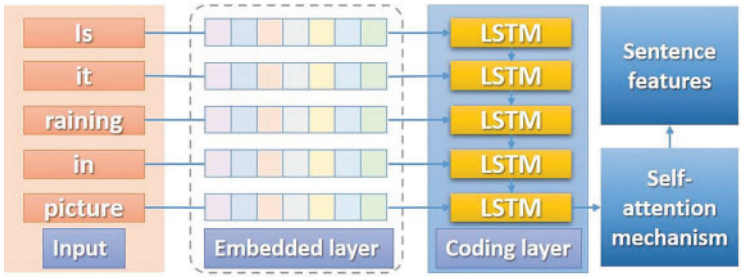

Figure 2: Sentence feature encoding

(3) Attention layer. The text attention mechanism can be expressed as the mapping of query (Q, which is the sentence feature after mapping to the feature space) to the key (K, keyword) and value (V, mapping matrix), as shown in Fig. 3. The value of K is the keyword feature set according to the task, and in the self-attention mechanism, K and Q are the same input, as shown in Eq. (9).

$$Att\,(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_q}}\right)V \tag{9}$$

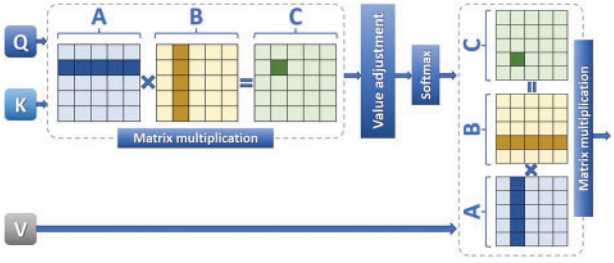

Figure 3: Self-attention mechanism

Since the input Q and K features are both d_q dimensional vectors, to prevent the inner product's maximum value after the dot product operation, a division $\sqrt{d_k}$ is added to the operation to adjust the operation structure.

3.3 Implementation of VQA Model Based on Attention Mechanism

After proposing an improved hybrid attention mechanism and introducing a self-attention mechanism, we combine the spatial and text attention mechanisms in the basic visual question answering model to optimize feature extraction. The overall structure of the visual question answering model is as follows.

(1) Feature extraction

Firstly, we use the pre-trained ResNet-152 model on the ImageNet dataset as the image feature extractor, remove the last fully connected layer of the model, and use the output features of the last convolutional layer as image features. For text features, word embedding is first performed to obtain word features, and then the GRU threshold-based recurrent neural network skip-thought is used to encode sentence features.

(2) Attention mechanism

The attention mechanism includes an improved hybrid attention mechanism for image features and a self-attention mechanism for text features. This paper adopts an improved hybrid attention mechanism to weigh the image features extracted by a deep neural network. F This paper introduces a self-attention mechanism for text features o weigh them individually.

(3) Feature fusion

Finally, we perform feature fusion on the features of different modalities. The fusion method adopts the bilinear network, which extracts the image features using the ResNet152 network pre-trained on the ImageNet dataset. It extracts the question features using the skip-thought cyclic convolutional neural network based on the GRU threshold structure. For the compression of the fused features, we use the Multimodal Low-rank Bilinear Pooling network (MLB) [32] and tucker dimensionality reduction to perform feature dimensionality reduction and compression, respectively. H is the feature after fusion and dimensionality reduction, such as Eq. (10).

$$H = \left\lfloor B\left(H_{Att}\right), B\left(v_Q\right)\right\rfloor \tag{10}$$

(4) Answer generation

In the answer generation network, we use a four-layer multi-layer perceptron to process the fusion feature H and finally use softmax to normalize the probability of the output feature layer. As in Eq. (11) for text generation, we input the obtained feature A into the skip-thought recurrent neural network based on the GRU threshold for feature decoding and obtain the final text output.

$$\begin{aligned} A &= \text{softmax}\left(f\left(H\right)\right) \\ S &= \text{skip-thought}\left(A\right) \end{aligned} \tag{11}$$

The complete visual question answering model structure in this paper is shown in Fig. 4. In the research of this paper, for the attention mechanism of the visual question answering model, we perform a separate self-attention weighting on the text features, while the attention weighting on the image features uses the text features as the guiding vector.

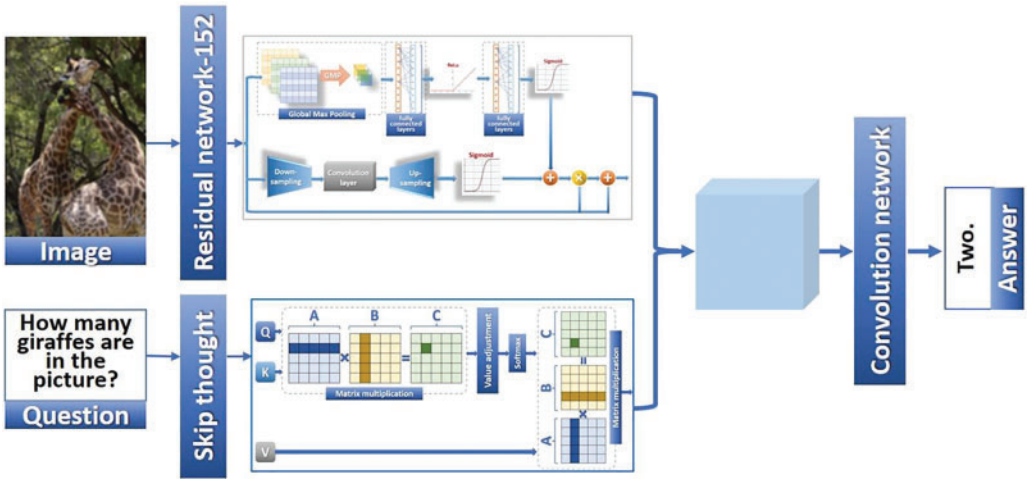

Figure 4: VQA model of this paper

4 Results

4.1 Self-Attention Mechanism VQA Model Experiment

We used the self-attention mechanism to process the text features. The model is tested based on the MLB method and the MUTAN [33] network. We respectively add a self-attention module after the text feature extraction module of the two fusion feature models. The test results are completed by the evaluation method of top 5. Top5 refers to the obtained prediction result scores. The network learning is successful if the five answers with the highest scores contain correct results. The test set experimental results are shown in Fig. 5.

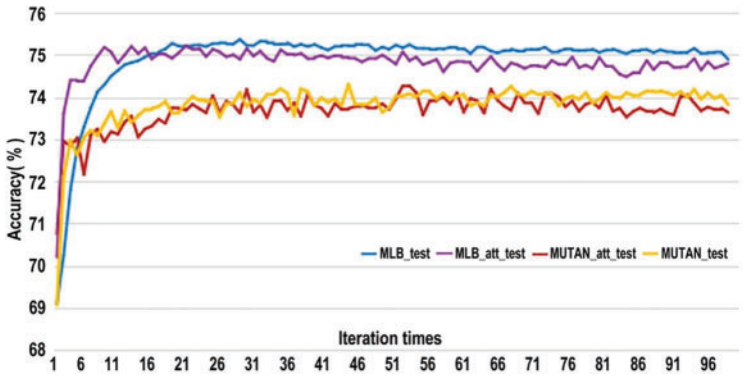

Figure 5: Comparison of MLB and MUTAN adding attention experiment

Combined with the experimental data in Table 1, in the test set, the top 5 accuracy of the MUTAN model without the self-attention mechanism is slightly higher than that of the model without the attention mechanism. The MLB network has slightly improved accuracy after using the self-attention mechanism, but the effect is very weak.

Table 1: Text attention mechanism

	MLB (%)	MUTAN (%)
Without attention mechanism	74.29	73.89
Self-Attention mechanism	74.34	73.32

According to the analysis of the test results, the performance of the MLB model does not improve significantly when the text attention mechanism is added, and the MUTAN network does not improve after adding the self-attention mechanism.

4.2 Improved Hybrid Attention Mechanism VQA Model Experiment

Different attention mechanisms and combinations of attention mechanisms have different effects on the structure of the experiment. In this section, the proposed improved hybrid attention mechanism is used in visual question answering tasks. About the influence of each attention in it, we will separately Experiment with different models carried out to illustrate.

Table 2 shows that when the spatial attention mechanism is added to the MLB network alone, the performance is significantly improved by nearly 5 percent, while the performance improvement is slight when the channel attention mechanism is added alone. When the spatial attention mechanism and the channel attention mechanism are added simultaneously, the performance is improved compared with that without the attention mechanism. Still, the accuracy is lower than the spatial attention mechanism added alone. However, when the spatial attention mechanism is used together with the channel attention mechanism and a certain amount of original data compensation is added, the performance is improved even more, which is 1 percentage higher than that of only using the spatial attention mechanism and nearly 6 percent compared with that without using the attention mechanism.

Table 2: Combination experiments of different attention mechanisms under MLB fusion mode

Type of attention	MLB (%)	MUTAN (%)
Spatial attention	79.36	73.52
Channel attention	75.92	72.17
Spatial attention + Channel attention	76.34	71.20
Spatial attention + Channel attention + Compensation feature	**80.39**	73.71
Without attention mechanism	74.29	73.89

The corresponding MUTAN model decreases the overall accuracy after adding the attention mechanism. When the spatial attention mechanism is used only, it drops by about 0.3 percentage points and drops more after adding the channel attention mechanism. Although the accuracy rate increases after adding the compensation feature, the overall accuracy rate still needs to improve.

In Fig. 6, we compare the accuracy of the MLB network and the MUTAN network with the improved hybrid attention mechanism and the accuracy without the attention mechanism and qualitatively show that the performance of MLB improves after the attention mechanism is used. On the other hand, the performance of the MUTAN network decreases after using the attention mechanism. The final model will use the MLB network as a multimodal feature fusion network.

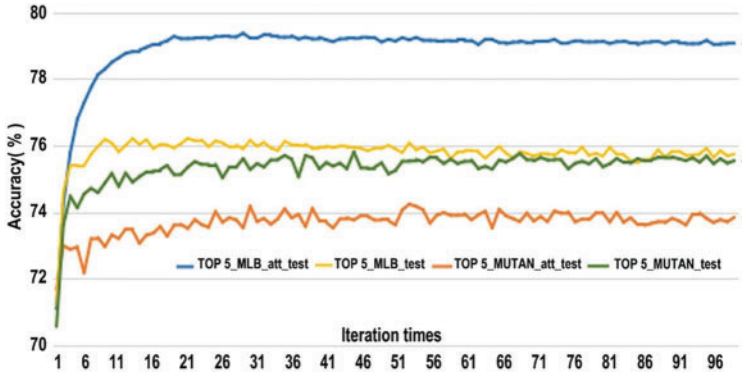

Figure 6: Improved hybrid attention mechanism

5 Discussion and Conclusions

This paper proposes an improved hybrid attention mechanism for image feature processing and introduces the self-attention mechanism for text feature extraction. In this paper, the self-attention mechanism is applied to text features, and its influence on the VQA system is explored through experiments. The experimental results show no improvement after adding the self-attention mechanism. As a result, after considering the operation cost, this paper does not use a text attention mechanism in the follow-up work. In addition, in this paper, the spatial and channel attention mechanisms are applied to the VQA model, and the effects of different attention mechanisms on the VQA system are discussed through experiments. Finally, the accuracy of the VQA system is improved to a certain extent by adding feature compensation and a mixed attention mechanism, the results show that attention mechanism and feature compensation add 6.1% accuracy to Multimodal Low-rank Bilinear Pooling network.

Although this paper has made specific improvements to the visual question answering model in terms of visual feature extraction and attention mechanism and achieved specific results, there is still room for future improvement and optimization. The model's incorrect answer to the question may also be caused by some object relationships' lack of knowledge training. Since the visual question answering database is artificially questioned for the image data set, much human prior knowledge may be included in the question and answer. These issues are difficult to solve with limited data training. Therefore, some scholars try to introduce external knowledge databases for training or use other visual question answering databases to enhance the model to achieve better results. Therefore, combining knowledge base and VQA, enhancing data through other methods, or optimizing the knowledge relationship of VQA databases will be the following potential research directions.

Acknowledgement: We would like to thank the anonymous reviewers for their helpful comments.

Funding Statement: This work was supported by the Sichuan Science and Technology Program (2021YFQ0003).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Pan, S. He, K. Zhang, B. Qu, C. Chen *et al.,* "Amam: An attention-based multimodal alignment model for medical visual question answering," *Knowledge-Based Systems*, vol. 255, pp. 109763, 2022.

[2] J. Cao, X. Qin, S. Zhao, J. Shen and L. Systems, "Bilateral cross-modality graph matching attention for feature fusion in visual question answering," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2022.

[3] V. Mnih, N. Heess and A. Graves, "Recurrent models of visual attention," in *28th Conference on Neural Information Processing Systems (NIPS)*, Montreal, CANADA, 2014.

[4] C. d. Santos, M. Tan, B. Xiang and B. Zhou, "Attentive pooling networks," arXiv preprint arXiv, pp. 03609, 2016.

[5] W. Yin, H. Schütze, B. Xiang and B. Zhou, "Abcnn: Attention-based convolutional neural network for modeling sentence pairs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 259–272, 2016.

[6] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville *et al.,* "Show, attend and tell: Neural image caption generation with visual attention," in *Int. Conf. on Machine Learning*, Lille, France, PMLR, pp. 2048–2057, 2015.

[7] Y. Zhu, O. Groth, M. Bernstein and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, pp. 4995–5004, 2016.

[8] K. Chen, J. Wang, L. -C. Chen, H. Gao, W. Xu *et al.,* "ABC-CNN: An attention based convolutional neural network for visual question answering," arXiv preprint arXiv:1511.05960, 2015.

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

[10] K. J. Shih, S. Singh and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, pp. 4613–4621, 2016.

[11] J. Lu, J. Yang, D. Batra and D. Parikh, "Hierarchical question-image co-attention for visual question answering," *30th Conference on Neural Information Processing Systems (NIPS)*, Barcelona, Spain, 2016.

[12] X. Shen, D. Han, C. Chen, G. Luo and Z. Wu, "An effective spatial relational reasoning networks for visual question answering," *PLoS One*, vol. 17, no. 11, pp. e0277693, 2022. https://doi.org/10.1371/journal.pone.0277693

[13] F. Yan, W. Silamu, Y. Li and Y. Chai, "SPCA-Net: A based on spatial position relationship co-attention network for visual question answering," *Vis. Comput.*, vol. 38, pp. 3097–3108, 2022. https://doi.org/10.1007/s00371-022-02524-z

[14] H. Sharma and S. Srivastava, "Context-aware and co-attention network based image captioning model," *The Imaging Science Journal*, pp. 1–13, 2023.

[15] Z. Guo and D. Han, "Multi-modal co-attention relation networks for visual question answering," *Vis. Comput.*, vol. 37, no. 1, pp. 2048, 2022. https://doi.org/10.1007/s00371-022-02695-9

[16] S. Liu, X. Zhang, X. Zhou and J. Yang, "BPI-MVQA: A bi-branch model for medical visual question answering," *Bmc Medical Imaging*, vol. 22, pp. 79, 2022. https://doi.org/10.1186/s12880-022-00800-x

[17] Y. Miao, W. Cheng, S. He and H. Jiang, "Research on visual question answering based on gat relational reasoning," *Neural Processing Letters*, vol. 54, pp. 1435–1448, 2022. https://doi.org/10.1007/s11063-021-10689-2

[18] L. Gómez, A. F. Biten, R. Tito, A. Mafla, M. Rusiñolet *et al.,* "Multimodal grid features and cell pointers for scene text visual question answering," *Pattern Recognition Letters*, vol. 150, no. 10, pp. 242–249, 2021. https://doi.org/10.1016/j.patrec.2021.06.026

[19] X. Wang, Q. Chen, T. Hu, Q. Sun and Y. Jia, "Visual-semantic dual channel network for visual question answering," in *Int. Joint Conf. on Neural Networks (IJCNN)*, Shenzhen, China, pp. 1–10, 2021.

[20] W. Tian, B. He, N. Wang and Z. Zhao, "Multi-channel co-attention network for visual question answering," in *Int. Joint Conf. on Neural Networks (IJCNN)*, ELECTR NETWORK, 2020.

[21] Y. Miao, S. He, W. Cheng G. Li and M. Tong, "Research on visual question answering based on dynamic memory network model of multiple attention mechanisms," *Sci. Rep.*, vol. 12, pp. 16758, 2022. https://doi.org/10.1038/s41598-022-21149-9

[22] M. Jaderberg, K. Simonyan, A. Zisserman and K. Kavukcuoglu, "Spatial transformer networks," in *29th Annual Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2015.

[23] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[24] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li *et al.,* "Residual attention network for image classification," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 3156–3164, 2017.

[25] P. H. Seo, Z. Lin, S. Cohen, X. Shen and B. Han, "Hierarchical attention networks," *arXiv preprint arXiv*, vol. 2, pp. 2393, 2016.

[26] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan *et al.,* "DISAN: Directional self-attention network for rnn/cnn-free language understanding," in *32nd AAAI Conference on Artificial Intelligence*, New Orleans, LA, 2018.

[27] D. Britz, A. Goldie, M. -T. Luong and Q. Le, "Massive exploration of neural machine translation architectures," in *Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Association for Computational Linguistics, pp. 1442–1451, 2017.

[28] K. He, X. Zhang, S. Ren and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 1026–1034, 2015.

[29] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *28th Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2014.

[30] M. Ren, R. Kiros and R. Zemel, "Image question answering: A visual semantic embedding model and a new dataset," *Proceedings Advances in Neural Inf. Process. Syst.*, vol. 1, no. 2, pp. 5, 2015.

[31] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra *et al.,* "Vqa: Visual question answering," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 2425–2433, 2015.

[32] J. H. Kim, K. W. On, W. Lim, J. Kim, J. W. Ha *et al.,* "Hadamard product for low-rank bilinear pooling," arXiv.1610.04325, 2016.

[33] H. Ben-younes, R. Cadene, M. Cord and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *2017 IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, pp. 2631–2639, 2017.