Jestr

Journal of Engineering Science and Technology Review 15 (1) (2022) 8 - 15

JOURNAL OF
Engineering Science and Technology Review

Research Article

www.jestr.org

Modal Feature Contribution Distribution Strategy in Visual Question Answering

Feng Dong[1, 2], Xiaofeng Wang[1,*], Ammar Oad[2] and Masood Nazir Khoso[3]

[1]*College of Information Engineering, Shanghai Maritime University, Shanghai 200120, China*
[2]*College of Information Engineering, Shaoyang University, Shaoyany 422000, China*
[3]*Department of Mechatronic Engineering, Mehran University of Engineering and Technology Jamshoro, Sindh 76020, Pakistan*

Received 25 December 2021; Accepted 2 February 2022

Abstract

The results of the visual question answering (VQA) task are obtained by the joint inference of the information of the image and text modalities, and its performance is affected by multiple factors, such as the modal feature extraction method and the modal feature fusion process. The current popular VQA models do not undertake processing after extracting the modal features from images and texts. Instead, feature fusion is executed straight without considering the modality's feature contribution, which is debatable. To reveal the relationship between the contribution distribution of image and text modal features, and the performance of VQA task, a plug-and-play contribution distribution strategy of modal features was proposed based on nonlinear functions. By globally operating the image and text features on the basis of extracting the features of the two modalities, the feature weight distribution in each mode was processed and analyzed based on the nonlinear function and the global features of the two modes. Extensive experiments were carried out on the basis of the existing model to fully verify the effectiveness of this strategy. Results show that the contribution distribution of the extracted image and text features before the feature fusion stage of the modality harmonizes the relationship between the image and text modalities and strengthens the effective features in the respective modalities. At the same time, the strategy further improves the performance of existing VQA models. This study provides a certain reference on how to reprocess features in VQA tasks.

Keywords: Visual question answering, Modal feature contribution distribution strategy, Modal feature fusion, Plug-and play

1. Introduction

Visual question answering (VQA) [1] is a multimodal processing task that combines visual and linguistic modalities. Answers to open-ended questions are provided in the form of natural language via visual images in the task. The VQA task solution is primarily investigated using a deep learning network framework [2-5], and the network model's performance is continuously enhanced by training on a large number of training sets. As a result, the deep learning network for the VQA task consists of three stages: feature representation, feature fusion, and answer prediction. However, most scholars concentrate on the feature fusion and answer prediction stages, thereby ignoring the impact of the feature representation stage on model performance.

The major goal of the feature representation stage is to extract the features of images and question texts for the upcoming feature fusion stage using the existing target detection pre-training model [6] and natural language pre-training model [7]. The extracted features are immediately passed to the fusion stage, which results in the identical contribution distribution of the two modalities' features to the succeeding stages. Scholars have not conducted extensive research on the distribution of modal contribution, which is quite illogical. For multimodal tasks, distributing the weight of modal features rationally from the perspective of global multimodal features is necessary to identify the effective information within the modes rather than directly fusing the modes without any processing. Meanwhile, VQA task is an artificial intelligence visual processing technology, and it should learn to replicate how humans process VQA tasks. To begin, we should consider the overall image and text, then narrow our focus on the image and text's most critical information, and then integrate this critical information to make logical reasoning for the answer to the question. This should be a top priority.

2. State of the art

At present, scholars have conducted several works on how to effectively improve the performance of VQA tasks. The following solutions were obtained: fusion-based methods, attention-based methods, and graph neural network (GNN) methods. Among these fusion-based methods, linear fusion and bilinear pooling are the most common. Lin [8] proposed a recognition architecture with a bilinear model to solve the fine-grained recognition task of images. The specific solution was to use two convolutional neural networks (CNN) to extract image features and then use a bilinear model to fuse the extracted features so that fine-grained image classification could be performed on the fused vectors. Such method had certain reference for the fusion of image and text features in subsequent VQA tasks, but the disadvantage was that the fusion process was relatively simple and did not capture the key feature information of the image. Based on this, Kim [9] proposed the use of bilinear attention matrix to introduce bilinear fusion into matrix fusion and proposed Bilinear Attention Networks (BAN), which made better use of complete context features and

*E-mail address: xfwang58@163.com
ISSN: 1791-2377
doi:10.25103/jestr.151.02

image features under the guidance of attention mechanism. However, the disadvantage was that the computational workload was relatively large. Ben [10] introduced a novel multimodal fusion method based on block-super diagonal tensor decomposition. This method optimized the expressivity of fusion models and is capable of representing very fine-grained interactions between modalities while maintaining a robust unimodal representation. However, the whole method had a large amount of training, and the accuracy was not very high. At the same time, Fang [11] proposed a bilinear framework called block term decomposition pooling (BTDP). Based on the block-term decomposition theory of tensors, the framework introduced sparsity into bilinear operations to improve the effect of feature fusion between modalities. However, during the fusion process, the BTDP framework did not discriminate important objects in images and texts effectively. Therefore, methods based on attention mechanism were widely used in VQA tasks.

Among attention-based methods, Anderson [12] proposed a combined bottom–up and top–down attention mechanism that computed attention at the level of objects and other salient image regions. The application of this mechanism to the VQA task won the championship of the 2017 VQA challenge, but the attention mechanism only considered the key target of the image and ignored the role of the multimodal fusion process of image features and text features. Li [13] established the positional self-attention with co-attention (PSAC) framework by exploiting the self-attention and co-attention in the transformer structure [14] to replace the attention-based recurrent neural network (RNN). The framework exploited the global dependencies of video questions and temporal information, thereby enabling the process of question and video encoding to execute in parallel. At the same time, the co-attention mechanism was used to fuse the two modal features. Although PSAC achieved good results on the count performance of the VQA task, other performance metrics still needed to be improved. Chen [15] proposed a novel multimodal encoder–decoder attention network (MEDAN). MEDAN was composed of deeply cascaded multimodal encoder–decoder attention (MEDA) layer, each MEDA layer contained an encoder module that modeled the self-attention of the question and a decoder module that modeled the self-attention of the question-guided attention and images. The network achieved an overall accuracy of 71.01% on the test set. However, MEDAN only considered the attention mechanism of the question-guided image and did not consider the image-to-question guidance. Subsequent scholars had extensively applied attention mechanisms to other specialized domains related to VQA tasks, such as remote sensing image (RSI) processing and medical image processing. Zheng [16] introduced attention mechanism and bilinear technique in RSI processing to enhance the features of spatial location and alignment between words and utilized the fully connected layer of SoftMax to output the answer from the perspective of multi-classification task. The advantage was that the proposed method could capture the features between the image and the question text. Sharma D [17] developed an attention-based multimodal deep learning model for the VQA task of medical images, which was referred to as MedFuseNet. MedFuseNet aimed to maximize learning with minimal complexity by decomposing question statements into simpler tasks and predicting answers. Experiments showed that this approach demonstrated the interpretability of model predictions during the visualization of attention. However, the above two VQA methods related to professional fields did not consider the contribution of image and question text modalities in the prediction results of the VQA task.

In addition to the fusion-based and attention-based approaches, the GNN approach opens up a new way to solve the VQA task. Narasimhan [18] applied graph convolution network (GCN) to VQA tasks using common sense reasoning, inferred the correct answer through entity graph and GCN and achieved 7% accuracy improvement in FVQA dataset. However, due to the problem of variance inflation in GCN, performance would decline in the training process. Haurilet [19] proposed a dynamic GNN model based on automatic learning from input. Different from the general VQA method, which relied on the attention mechanism of global embedding of image cell structure, this model automatically built scene graph and focuses on the relationship between nodes to answer a given question. Laugier [20] transformed external knowledge into question-answering system by using GCN and used this external knowledge to understand the contextual knowledge of VQA question text. The methods proposed by Haurilet and Laugier achieve good performance improvements on benchmark datasets, but they all relied on knowledge systems to reason about VQA problems. Therefore, how to construct an effective knowledge base becomes the key of their proposed method to a certain extent. Therefore, Nuthalapati [21] proposed a knowledge base-independent method and designed a new model called conditional enhanced graph attention network (CE-GAT) to encode visual and semantic scene graph pairs. CE-GAT encoded visual and semantic scene graphs with node and edge features, seamlessly integrated with text question encoders, and generated answers through question graph conditioning. Compared with the state-of-the-art, it achieved good results on the GQA dataset. Sharma H [22] proposed a novel VQA model based on GNN and contextual attention. The model fully considered the relationship between regions of interest between modalities and the inference of these regions and achieved high accuracy on VQA 1.0 and VQA 2.0 datasets. Guo [23] improved the BAN and developed a bilinear graph network to simulate the context of the joint embedding of words and objects. The question graph and image graph in this bilinear graph network cooperated with each other to realize multi-step reasoning by establishing relationships and dependencies between objects. Zhang [24] designed an efficient multimodal inference and fusion model to achieve fine-grained multimodal inference and fusion. The method aimed to form a deep multimodal reasoning and fusion network (DMRFNet) by deep stacking of multi-graph reasoning and fusion (MGRF) layers. The network revealed the motivation of the module's decision and enhanced the interpretability of the model. Although the above reasoning models had achieved good results in VQA tasks, due to the huge amount of calculation, complex reasoning process, and long model training time, these methods only stayed in scientific research, and there was still a great distance from practical application.

The above analyses mainly discussed the performance improvement of the VQA task from the aspects of feature fusion between modalities, the application of attention mechanism, and the reasoning of the model, but the contribution of the two modalities of image and question text to the VQA task are not considered. In particular, the existing VQA task models assign the same contribution to both modalities after extracting features from images and

texts, which are then used as inputs for subsequent stages, which is obviously unreasonable. To that end, this study introduces a MFCD module between the two stages of feature representation and feature fusion and employs the nonlinear functions SoftMax [25] and Sigmoid [26] to develop a weight-based MFCD strategy and a gating mechanism-based MFCD strategy that can reasonably distribute the feature weights for image and text modalities to achieve more accurate predictions. From the standpoint of global features, the weight-based method's primary goal is to assign varied weights to the modality's features, whereas the gating mechanism's primary goal is to remove the modality's invalid information and strengthen its effective information. These two distribution strategies are seamlessly integrated into the existing VQA model, thereby providing robust support for subsequent modal feature fusion and response answer prediction.

The remainder of this study is organized as follows. Section 3 presents algorithm for MFCD strategy in detail. Section 4 describes the experimental studies of the proposed strategy, and finally, the conclusions are summarized in Section 5.

3. Methodology

3.1 Contribution distribution of modal features

The VQA model is based on a deep learning framework and is consists of three parts: feature representation, feature fusion, and answer prediction. The idea of MFCD in this study is largely applied between the feature representation stage and the feature fusion stage. Given that no academics have discussed the contribution of modal features, a brief exposition of this idea must be provided.

In the feature representation stage, the general VQA task directly inputs these features into the feature fusion stage after using the CNN to extract the target features of the picture and the RNN to extract the language features of the question text (Figure 1).

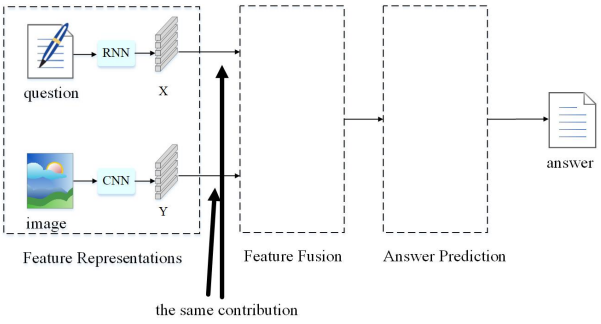

Fig. 1. General VQA model

In Figure 1, the feature information generated in the feature representation stage, such as image and text features, is not processed and directly enters the feature fusion stage, with the same contribution to the results of the VQA task. In using computer simulations for humans to solve the VQA task, when all the modal information is perceived by humans with the same contribution, the intra-modal and inter-modal information have no differences and hence is anomalous. The reason is that people also have choices when dealing with information within and between modalities. Consistent with the attention mechanism, some modal information will receive more attention, whereas other modal information will receive less attention. To sum up, the idea of MFCD in this study is shown in Figure 2. That is, a MFCD module is added between the feature representation stage and the feature fusion stage to simulate the perception of various information modes and to distribute the modal feature contribution adaptively before the modal feature fusion.

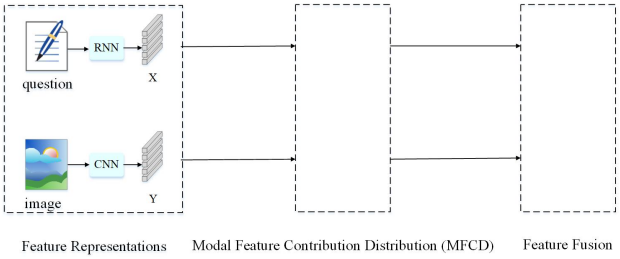

Fig. 2. Idea of MFCD

This study mainly introduces two strategies in MFCD, a weight-based contribution distribution (WCD) strategy and a gating mechanism-based contribution distribution (GMCD) strategy.

3.2 WCD strategy

The primary task of the VQA task feature representation stage is to extract the corresponding features from the input image I and question text Q. Based on the existing model of VQA task, RNN and CNN are generally used to extract text and image features, as expressed in Formulas (1)-(2):

$$X = CNN(I) \tag{1}$$

$$Y = RNN(Q) \tag{2}$$

where X and Y represent the extracted image and text features, respectively. The whole process is also shown in Figures 1 and 2. Based on the idea of the MFCD strategy, this study first introduces WCD strategy.

The WCD strategy uses the element-wise addition of tensors to obtain the global features of images and texts and redistributes the weights of image and text features based on the SoftMax nonlinear function. The multi-modal global feature can be used within each modality to increase the weight of effective features adaptively while reducing the weight of invalid features. The specific process is shown in Figure 3:

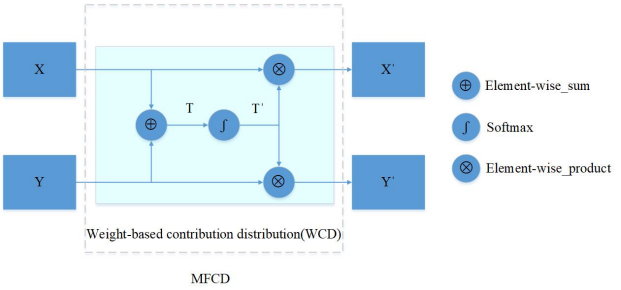

Fig. 3. WCD

First, the dimensions of X and Y are unified through tensor operation, enter MFCD for element-wise addition operation, and obtain T, which is the global feature for images and text. As shown in Formula (3):

$$T = X \oplus Y \tag{3}$$

T aims to form a distribution based on the weight of T through the succeeding SoftMax function, as shown in Formula (4):

$$T' = Soft\ max(T) \tag{4}$$

where T' is the weight distribution coefficient. Then, using the idea of residual structure, the outputs X' and Y' of MFCD are obtained by element-by-element multiplication of T' and X' and Y', respectively. The process is shown in Formulas (5)-(6):

$$X' = X \otimes T' \tag{5}$$

$$Y' = Y \otimes T' \tag{6}$$

where X' and Y' represent the image and text features distributed by WCD, respectively. Finally, these features can be inputted into the subsequent feature fusion stage.

By merging the previous formulas, the contribution distribution of image and text features can be expressed comprehensively by Formulas (7) and (8) as follows.

$$X' = X \otimes Soft\ max(X \oplus Y) \tag{7}$$

$$Y' = Y \otimes Soft\ max(X \oplus Y) \tag{8}$$

3.3 GMCD strategy

In the GMCD strategy, the gating mechanism uses the nonlinear function sigmoid. Consistent with WCD strategy, the GMCD strategy can filter the features in each mode by sigmoid function on the basis of the global features of images and texts. The features that pass sigmoid function continue to enter the feature fusion stage, and those that fail will be eliminated by the gating strategy, as shown in Figure (4):

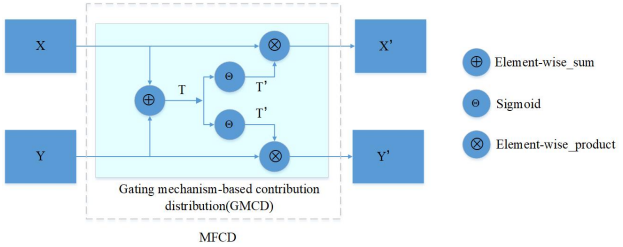

Fig. 4. GMCD

Image feature X and text feature Y are inputted into MFCD, which is the same as the WCD. First, element-wise addition is performed, as represented by Formula (9):

$$T = X \oplus Y \tag{9}$$

Then, using the sigmoid function to generate the gating coefficient T' of T, Formula (10) is expressed as follows:

$$T' = Sigmoid(T) \tag{10}$$

Finally, referring to the residual structure and element-by-element multiplication operation, X and Y are multiplied by the gating coefficient T' to obtain the output X' of the image and Y' of the text respectively. Formulas (11)-(12) are expressed as:

$$X' = X \otimes T' \tag{11}$$

$$Y' = Y \otimes T' \tag{12}$$

The whole process can be expressed by Formulas (13) and (14) as follows:

$$X' = X \otimes Sigmoid(X \oplus Y) \tag{13}$$

3.4 Differences between the two MFCD strategies

The SoftMax and sigmoid functions are used in the two contribution distribution strategies. SoftMax generates a weight distribution with a sum of 1, whereas sigmoid generates a gated coefficient between 0 and 1. In these two strategies introduced in this study, the contribution distribution of intra-modal features based on the global features of images and texts simulates the degree of importance that people place on intra-modal features. Fundamentally, they are both the process of simulating human attention in the application of an attention mechanism in the VQA task. However, these two strategies still have essential differences. GMCD can fundamentally filter out features that are irrelevant to the VQA task, whereas WCD only weakens these features.

4. Result Analysis and Discussion

We apply two mechanisms, namely, WCD and GMCD, to the four models with better performance on the VQA task. Several comparative experiments are performed on the benchmark dataset VQA-v2 to verify the reliability of the MFCD strategy proposed in this study. These four models include MuRel [27], MCAN [28], MMnas [29], and MCAoAN [30], which are all derived from the top conferences of ACM and IEEE in recent years and have certain reference value. All experiments are performed on two RTX 2080Ti GPUs. The questions encountered in the experimental process have three types, namely, yes/no, number, and other. The overall performance is represented by all. The experimental results will be tested on two online test sets, namely, test-dev and test-std, and will appear in the form of percentages.

4.1 VQA-v2 dataset

The VQA-V2 dataset is based on the MS COCO dataset and consists of images, questions, and answers. This dataset contains 82,783 training images, 40,504 validation images, and 1,434 test images in the image section. The questions section contains 443,757 training questions, 214,354 validation questions, and 447,793 testing questions. The final answers section only contains 4,437,570 training answers and 2,143,540 verification answers. The ratio of the number of answers to the number of questions suggests that the training process and verification process provided 10 answers for each question to choose from.

4.2 WCD Experiments

Based on the WCD strategy in Section 3.2, this study analyzes the impact of WCD as a plug-and-play module on model performance in several classical models. As a MFCD module, the WCD strategy is added to the MCAN, MuRel, MMnas, and MCAoAN models. The specific location is after CNN and RNN extract the features of the image and the question text. The MCAN model is taken as an example, as shown in Figure 5:

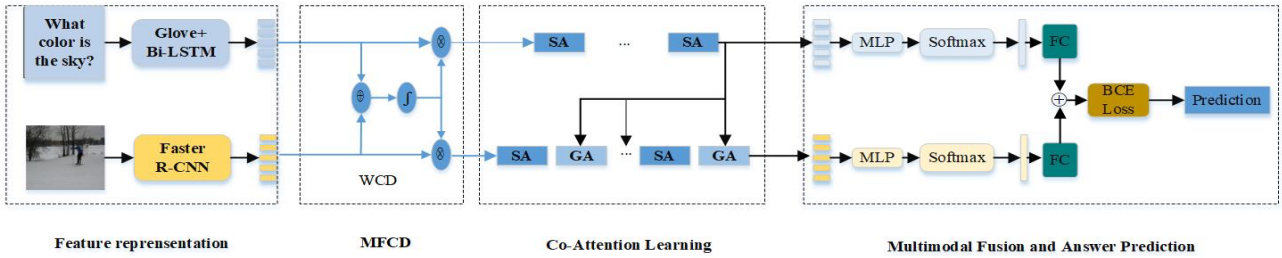

Fig. 5. MCAN model with WCD

The three other model structure modifications are the same as the MCAN model. The experiments were carried out on these four models in turn, and the results are shown in Table 1.

Table 1. WCD comparison experiment

Model	Test-Dev				Test-Std
	All	Other	Yes/no	Number	All
MuRel	68.03	57.85	84.77	49.84	68.41
MuRel+GMCD	**68.13**	**58.05**	**84.87**	**50.09**	**68.59**
MCAN	70.63	60.72	86.82	53.26	70.90
MCAN+GMCD	**70.80**	**60.93**	**86.95**	**53.44**	**71.20**
MMnas	71.24	61.05	87.27	55.68	71.46
MMnas +GMCD	**71.48**	**61.54**	**87.45**	**55.88**	**71.76**
MCAOAN	70.84	61.01	86.96	53.45	71.16
MCAOAN+GMCD	**71.03**	**61.32**	**87.12**	**53.95**	**71.44**

The performance increase part in the table is marked in bold. The experimental results in Table 1 show that the model with the WCD strategy has improved in various performance. For a single model, MCAN+WCD improves the overall performance of the MCAN model by 0.17 and 0.3 on test-dev and Test-Std, respectively. Similarly, MMnas+WCD increased by 0.24 and 0.3, and MCAOAN+WCD increased by 0.19 and 0.28. However, MuRel+WCD only improved by 0.10 and 0.18, which is smaller than the three previous models. From the perspective of model structure, the MuRel model introduces a multimodal relational reasoning unit to gradually improve the interaction between vision and questions, and the interaction between image features and question text features is not very deep in the feature fusion stage. The MCAN and MCAOAN models use the transformer structure. Different from the multi-modal relational reasoning unit of the MuRel model, the self-attention unit and the guided attention unit in the transformer structure fully consider the interaction between the two modalities of image and text and use the cascading method to continuously interact among modalities. The degree of fusion is higher than the MuRel model. In addition, the MMnas model uses a neural structure search mechanism, and the feature fusion is also better than the MuRel model. Therefore, from the perspective of feature fusion, the use of the WCD strategy is more effective in improving model performance for models with better feature fusion effects.

Based on the above point of view, the percentage results for other types of questions for the four models are compared, as shown in Figures 6-8:

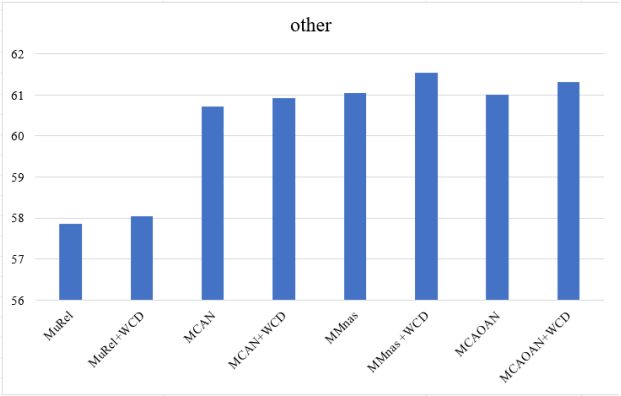

Fig. 6. Performance comparison for other types of questions

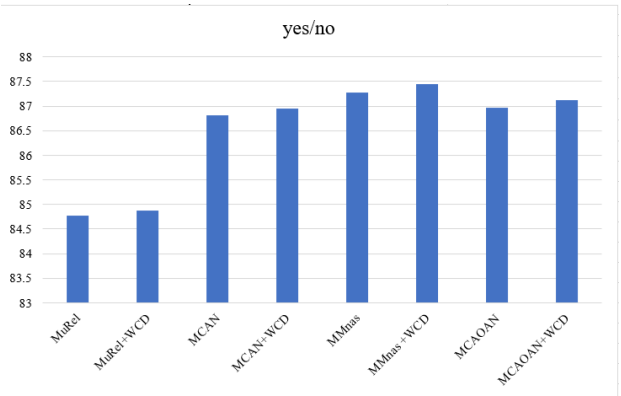

Fig. 7. Performance comparison for yes/no types of questions

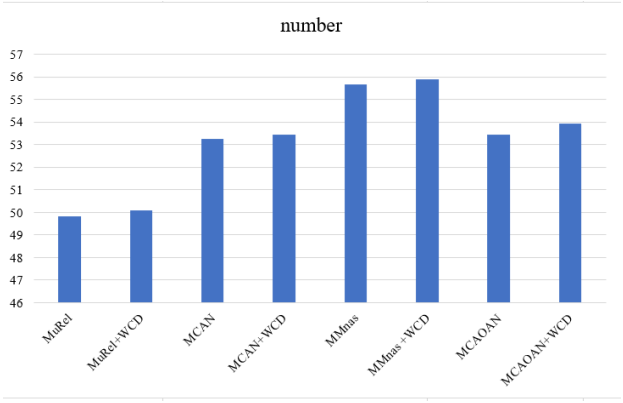

Fig. 8. Performance comparison for number types of questions

Figures 6-8 show that the performance of the MuRel model for the three types of questions with good fusion effect is lower than that of the three other models.

At the same time, according to the characteristics of the model itself, WCD can be added to the model in a targeted manner. Then, it is verified by experiments. If it is helpful to the experimental results, then it will be retained, and if it is not helpful, then it will be removed, thereby reflecting the advantages of WCD's plug and play characteristics.

4.3 GMCD Experiments

Based on the GMCD strategy in Section 3.3, the same as the WCD strategy, this study also analyzes the impact of GMCD as a plug-and-play module on model performance in several classic models. The GMCD strategy is embedded in the same position of the four models. The MCAN model is taken as an example, as shown in Figure 9:

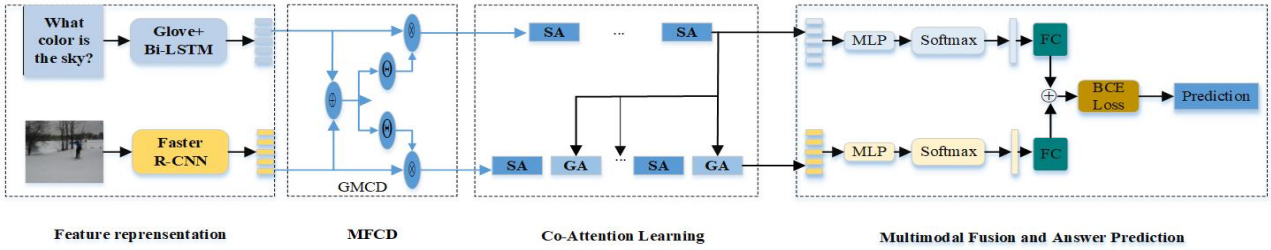

Fig. 9. MCAN model with GMCD

The experimental results are shown in Table 2.

Table 2. GMCD comparison experiment

Model	Test-Dev				Test-Std
	All	Other	Yes/no	Number	All
MuRel	68.03	57.85	84.77	49.84	68.41
MuRel+WCD	**68.21**	**58.07**	**84.92**	**50.19**	**68.73**
MCAN	70.63	60.72	86.82	53.26	70.90
MCAN+WCD	**70.87**	**61.03**	**87.07**	53.22	**71.28**
MMnas	71.24	61.05	87.27	55.68	71.46
MMnas +WCD	**71.44**	**61.51**	**87.08**	**56.17**	**71.71**
MCAOAN	70.84	61.01	86.96	53.45	71.16
MCAOAN+WCD	**71.11**	**61.47**	**87.22**	53.38	**71.48**

The data in Table 2 show that the overall performance of the GMCD strategy is improved on Test-Dev and Test-Std test sets. The MuRel, MCAN, MMnas, and MCAOAN models improved their overall performance by 0.18 and 0.32, 0.24 and 0.38, 0.20 and 0.25, and 0.27 and 0.32, respectively. However, the accuracy of the MCAN and MCAOAN models on the number type question is indeed reduced by 0.04 and 0.07. The main reason is that the GMCD strategy uses the gating function sigmoid to manipulate the features inside the respective modality of the image and text to highlight the important features inside the modality. The problem of the number of targets is considered from the global features of the image. When occlusion exists between the target and the target in the image, the GMCD strategy removes the edge features of the target. At the same time, MCAN and MCAOAN models use the transformer mechanism, so when answering an open question about the number of targets in an image, the weakened features are ignored in the subsequent feature fusion stage, thereby resulting in deviations in target counts. However, for the other and yes/no types of questions, the GMCD strategy, such as the attention mechanism, captures the important information inside the modal to a certain extent.

Notably, under the action of GMCD, the MuRel and MMnas models have improved the number type question because these two models address the inference process based on images and text. The MuRel model infers based on the relationship between images and question text, whereas the MMnas model infers their relationship by finding the best network structure for processing images and text through neural architecture search.

4.4 MFCD strategy analysis

WCD and GMCD are two MFCD strategies. The former distributes the contribution of image and text features based on weights, which simulate human's different emphasis on intra-modal features. The modal feature with high weight will occupy more feature fusion percentage in the subsequent stages, thereby playing a guiding role for another modal feature with low weight. The latter is a feature contribution distribution method based on the gating mechanism, which focuses on the distribution of global features within the modality, which is similar to the attention

mechanism of human vision and achieves the purpose of focusing on important information within the modality.

The mechanisms of WCD and GMCD also have similarities. They both execute their respective operations after performing element-by-element addition operations on the extracted image and text features. The element-by-element addition is the beginning of the fusion of the two modal features, and the result is the global feature of the two modalities. Such an operation is also in line with the requirement of the VQA task to consider image and text feature information.

This study proposes two MFCD strategies, namely, WCD and GMCD, both of which have excellent plug-and-play value. However, while in use, the specific network structure, which is inextricably linked to the model's feature representation and feature fusion processes must be considered. If the model's feature extraction approach fails to collect the global features of the modalities or if these features cannot capture the global information contained in images and texts, then MFCD will be rendered ineffective. When the feature fusion process does not allow the significant interaction between modalities or when the key objects are not aligned, MFCD becomes a superfluous aspect of the model.

Overall, WCD and GMCD are the strategies used for obtaining important parts of features, and each has its own advantages for a specific model.

5. Conclusions

To explore the contribution of image and text modal features to the VQA task and to reveal the impact of the contribution distribution on the performance of the VQA task, this study begins with the weight distribution and gating mechanism and adopts the combination deep learning and experimental research. Inter-modal and intra-modal feature contributions are analyzed. The following conclusions could be drawn:

(1) WCD and GMCD strategies analyze feature information in the image and text modes efficiently. Moreover, they play a critical auxiliary role in the subsequent feature fusion. It is analogous to the second stage of feature preprocessing following the feature representation stage.

(2) The two MFCD strategies begin from the global features of images and texts and adjust the distribution of their respective modal features adaptively with the coefficients generated by the global features.

(3) WCD and GMCD strategies are simple in structure, easy to implement, and have the efficiency of plug and play.

In this study, a new concept of MFCD is proposed by combining laboratory experiment with theoretical research. The established weight-based and gating-based contribution distribution strategies, which have certain reference for improving the performance of VQA task, are very simple and easy to implement. However, in this study, these two strategies are only used as a secondary processing method of modal features before the feature fusion stage, and the use of these strategies in this stage is not fully considered. Therefore, in the future work, we will innovate more modal feature contribution allocation strategies with plug and play characteristics and apply these strategies to other multimodal processing directions and models.

Acknowledgements

This work was supported by National Science Foundation of China (Grant Nos. 61872231 and 61701297).

References

1. 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*, Santiago, Chile: IEEE, 2015, pp.2425-2433.
2. Narayanan, A., Rao, A., Prasad, A., Natarajan, S., "VQA as a factoid question answering problem: A novel approach for knowledge-aware and explainable visual question answering". *Image and Vision Computing*, 116, 2021, pp. 104328.
3. Ouyang, N., Huang, Q., Li, P., Yi, C., Liu, B., Leung, H. F., Li, Q., "Suppressing Biased Samples for Robust VQA". *IEEE Transactions on Multimedia*, 2021, doi: 10.1109/TMM.2021.3097502.
4. Sejnova, G., Tesar, M., Vavrecka, M., "Compositional models for VQA: Can neural module networks really count?". *Procedia Computer Science*, 145, 2018, pp.481-487.
5. Sejnova, G., Vavrecka, M., Tesar, M., Skoviera, R., "Exploring logical consistency and viewport sensitivity in compositional VQA models.". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Macau, China: IEEE, 2019, pp.2108-2113
6. Sun, X., Wu, P., Hoi, S. C., "Face detection using deep learning: An improved faster RCNN approach". *Neurocomputing*, 299, 2018, pp.42-50.
7. Ruder, S., Peters, M. E., Swayamdipta, S., Wolf, T., "Transfer learning in natural language processing". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, USA: ACL, 2019, pp.15-18.
8. Lin, T. Y., RoyChowdhury, A., Maji, S., "Bilinear cnn models for fine-grained visual recognition". In: *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile: IEEE, 2015, pp. 1449-1457.
9. Kim, J. H., Jun, J., Zhang, B. T., "Bilinear Attention Networks". In: *Thirty-second Conference on Neural Information Processing Systems*, Montreal, Canada: Curran Associates Inc., 2018, pp.1571-1581.
10. Ben-Younes, H., Cadene, R., Thome, N., Cord, M., "Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, USA: AAAI, 2019, pp.8102-8109.
11. Fang, Z., Liu, J., Liu, X., Tang, Q., Li, Y., Lu, H., "BTDP: Toward Sparse Fusion with Block Term Decomposition Pooling for Visual Question Answering". *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(2), 2019, pp.1-21.
12. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., 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*, Lake City, USA: IEEE, 2018, pp.6077-6086.
13. Li, X., Song, J., Gao, L., Liu, X., Huang, W., He, X., Gan, C., "Beyond rnns: Positional self-attention with co-attention for video question answering". In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, USA: AAAI, 2019, pp.8658-8665.
14. 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 31st International Conference on Neural Information Processing Systems*, Long Beach, USA: Curran Associates Inc., 2017, pp. 6000-6010.

15. Chen, C., Han, D., Wang, J., "Multimodal encoder-decoder attention networks for visual question answering". *IEEE Access*, 8, 2020, pp.35662-35671.
16. Zheng, X., Wang, B., Du, X., Lu, X., "Mutual attention inception network for remote sensing visual question answering". *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2021, pp.1-14.
17. Sharma, D., Purushotham, S., Reddy, C. K., "MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain". *Scientific Reports*, 11(1), 2021, pp.1-18.
18. Narasimhan, M., Lazebnik, S., Schwing, A. G., "Out of the box: Reasoning with graph convolution nets for factual visual question answering". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montreal, Canada: Curran Associates Inc., 2018, pp. 2659-2670.
19. Haurilet, M., Al-Halah, Z., Stiefelhagen, R., "DynGraph: Visual Question Answering via Dynamic Scene Graphs". In: *German Conference on Pattern Recognition*, Dortmund, Germany: Springer, 2019, pp.428-441.
20. Laugier, L., Wang, A., Foo, C. S., Theo, G., Chandrasekhar, V., "Encoding knowledge graph with graph CNN for question answering". In: *Proceedings of the International Conference on Learning Representations*, New Orleans, USA: ICLR, 2019.
21. Nuthalapati, S. V., Chandradevan, R., Giunchiglia, E., Li, B., Kayser, M., Lukasiewicz, T., Yang, C., "Lightweight Visual Question Answering using Scene Graphs". In: *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, Queensland, Australia: ACM, 2021, pp.3353-3357.
22. Sharma, H., Jalal, A. S., "Visual question answering model based on graph neural network and contextual attention". *Image and Vision Computing*, 110, 2021, pp.104165.
23. Guo, D., Xu, C., Tao, D., "Bilinear graph networks for visual question answering". *IEEE Transactions on Neural Networks and Learning Systems*, 2021, pp.1-12.
24. Zhang, W., Yu, J., Zhao, W., Ran, C., "DMRFNet: deep multimodal reasoning and fusion for visual question answering and explanation generation". *Information Fusion*, 72, 2021, pp.70-79.
25. Marty, M. T., "Memetic algorithm for community detection in networks". *Journal of the American Society for Mass Spectrometry*, 30(10), 2019, pp.2174-2177.
26. Mugunthan, S. R., Vijayakumar, T., "Design of improved version of sigmoidal function with biases for classification task in ELM domain". *Journal of Soft Computing Paradigm (JSCP)*, 3(02), 2021, pp.70-82.27.
27. 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, USA: IEEE, 2019, pp. 19891-1998.
28. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q., "Deep modular co-attention networks for visual question answering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, USA: IEEE, 2019, pp. 6281-6290.
29. Yu, Z., Cui, Y., Yu, J., Wang, M., Tao, D., Tian, Q., "Deep multimodal neural architecture search". In: *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, USA: ACM, 2020, pp. 3743-3752.
30. Rahman, T., Chou, S. H., Sigal, L., Carenini, G., "*An Improved Attention for Visual Question Answering*". [EB/OL]. Retrieved from https://openaccess.thecvf.com/content/CVPR2021W/MULA/papers/Rahman_An_Improved_Attention_for_Visual_Question_Answering_CVPRW_2021_paper.pdf,2021-7/2022-1-5.