Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 19(10): 10192–10212.
DOI: 10.3934/mbe.2022478
Received: 20 May 2022
Revised: 25 June 2022
Accepted: 05 July 2022
Published: 20 July 2022

Research article

Medical visual question answering via corresponding feature fusion combined with semantic attention

Han Zhu[1], Xiaohai He[1,*], Meiling Wang[1], Mozhi Zhang[2] and Linbo Qing[1]

[1] College of Electronics and Information Engineering, Sichuan University, Chengdu 610064, China
[2] Department of Computer Science, University of Maryland, College Park, MD 20742, USA

* **Correspondence:** Email: hxh@scu.edu.cn.

Abstract: Medical visual question answering (Med-VQA) aims to leverage a pre-trained artificial intelligence model to answer clinical questions raised by doctors or patients regarding radiology images. However, owing to the high professional requirements in the medical field and the difficulty of annotating medical data, Med-VQA lacks sufficient large-scale, well-annotated radiology images for training. Researchers have mainly focused on improving the ability of the model's visual feature extractor to address this problem. However, there are few researches focused on the textual feature extraction, and most of them underestimated the interactions between corresponding visual and textual features. In this study, we propose a corresponding feature fusion (CFF) method to strengthen the interactions of specific features from corresponding radiology images and questions. In addition, we designed a semantic attention (SA) module for textual feature extraction. This helps the model consciously focus on the meaningful words in various questions while reducing the attention spent on insignificant information. Extensive experiments demonstrate that the proposed method can achieve competitive results in two benchmark datasets and outperform existing state-of-the-art methods on answer prediction accuracy. Experimental results also prove that our model is capable of semantic understanding during answer prediction, which has certain advantages in Med-VQA.

Keywords: multimodal learning; pre-training model; residual network; long short-term memory; semantic attention

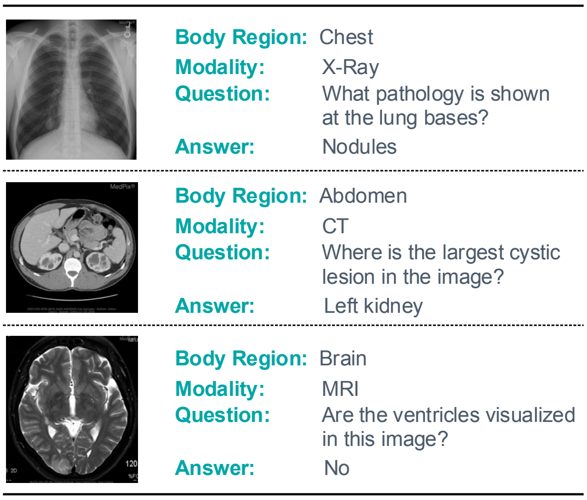
1. Introduction

In the medical field, medical imageology is a mandatory course to be undertaken by every doctor. Different types of imaging techniques, such as computed tomography (CT), magnetic resonance imaging (MRI) and X-ray, play an irreplaceable role in the clinical diagnosis of patients [1−5]. Neural network technology has been gradually introduced into health informatics with the continuous advancement of medical empowerment [6]. Additionally, the effectiveness of this technology has been proven in radiology image analysis [7], and deep learning models have been utilized in the detection and analysis of various diseases, such as lung diseases [8] and chest cancer [9]. However, obtaining visual information exclusively from radiology images has the disadvantages of limited interactive channels and fixed interactive scenes.

In recent years, Visual Question Answering (VQA) [10] have gained ever-increasing attention as a challenging multimodal task. VQA combines the two disciplines of computer vision and natural language processing. A VQA task takes an image and a related question presented with the image as inputs, then it outputs the correct answer to the question through a series of processes. Most methods of VQA [11,12] are based on the framework of supervised learning, which requires large-scale well-annotated multimodal data to train the model. For VQA tasks, Malinowski et al. proposed the DAQUAR dataset [13] in 2014, and Ren et al. constructed the COCO-QA dataset [14] in 2015 based on the MSCOCO image database. Nevertheless, the datasets used in these studies were small in scale, and the question-answer pairs were machine generated, which led to a high repetition rate; besides, the cluttered image contents made questions difficult to be answered. Subsequently, the Visual Genome dataset [15] proposed by Krishna et al. and the Visual7W dataset [16] proposed by Zhu et al. were formulated. These datasets contained a large amount of data, the images and question-answer pairs were manually annotated and screened by volunteers. However, owing to the uneven distribution of answers and biases in the questions, the generalization performance of the models trained on these datasets was mediocre. Goyal et al. proposed the VQA 2.0 dataset [17] in 2017 based on MSCOCO image data. VQA 2.0 contains 240,721 pictures and 1,105,904 question-answer pairs. The scale of VQA 2.0 is sufficiently large, and it overcomes the unbalanced answer distribution. Therefore, VQA 2.0 has been widely used in current studies on general field VQA tasks.

Medical VQA (Med-VQA) aims to improve the quality and efficiency of modern medical diagnosis and alleviates the pressure on the currently strained medical resources. An example of Med-VQA is shown in Figure 1. Different types of radiology images are accompanied by annotations (such as Body Region and Modality) and corresponding clinical question-answer pairs. Each of these radiology images may correspond to several different questions and answers; however, we only list one of them in Figure 1. The Med-VQA task is used to predict the true answer through the provided radiology images and questions. Med-VQA technology can help patients find possible abnormalities in their bodies and—in combination with radiology images—help them easily understand the disease they are suffering from. Additionally, it can assist outpatient doctors with clinical diagnosis and simultaneously indicate abnormal problems that may be overlooked in radiology images.

Unlike in the general field, VQA, in medical domain, is confronted with the lack of large-scale annotated datasets for model training. On the one hand, there are only a few ways to obtain well-labeled radiology images; to annotate a radiology image is difficult and requires the cooperation of experienced doctors. On the other hand, the medical domain requires highly accurate and professional datasets, and different doctors have different ways of generating questions and using words, all of which make it challenging to produce Med-VQA datasets. To the best of our knowledge, ImageCLEF [18] first began to host challenges in Med-VQA early in 2018. VQA-RAD [19] is the earliest benchmark dataset proposed for Med-VQA, which has been representative and well-recognized over the years. It was sampled from MedPix (https://medpix.nlm.nih.gov/), which is a publicly available database of medical radiographic imaging and medical teaching cases. The question-answer pairs in VQA-RAD are generated by the natural-communication manner of professional clinical practitioners, and these questions are closer to the ones communicated between doctors and patients in real life than those generated from a template. SLAKE [20] is a Chinese/English bilingual VQA dataset recently proposed by Liu et al., which contains questions that cover more aspects than the previous datasets. Moreover, a knowledge graph is introduced into SLAKE to expand the scope of questions. Radiology images in the dataset were sampled from three open-source datasets (http://medicaldecathlon.com, https://nihcc.app.box.com/v/ChestXray-NIHCC, https://doi.org/10.5281/zenodo.3431873); furthermore, question-answer pairs were generated by professional doctors based on a pre-set template.



Body Region: Chest
Modality: X-Ray
Question: What pathology is shown at the lung bases?
Answer: Nodules

Body Region: Abdomen
Modality: CT
Question: Where is the largest cystic lesion in the image?
Answer: Left kidney

Body Region: Brain
Modality: MRI
Question: Are the ventricles visualized in this image?
Answer: No

Figure 1. Example of medical visual question answering (Med-VQA) (radiology images with annotations and corresponding question-answer pairs).

For Med-VQA, researchers [21,22] first leveraged transfer learning methods to pre-train the model with a large amount of annotated data from the general VQA domain. Then, they migrated the model to the medical domain for further fine-tuning. However, owing to the significant differences between these two domains, the performance of the model migrated from the general domain was not impressive. Subsequently, many studies [23−25] turned to the unlabeled radiology images. They pre-trained the visual feature extractor through unsupervised learning or self-supervised learning methods and then moved in Med-VQA for fine-tuning. This achieved a better performance in answer prediction. From another point of view, previous researchers paid much attention to improving visual feature extraction through various approaches, while neglecting that the textual feature extraction is equally indispensable in Med-VQA. Furthermore, there were few works emphasized the importance of the interaction between visual features and corresponding textual features as well as the specific semantic information contained in different questions.

Based on the aforementioned factors, we briefly summarize our contributions as follows:

- Considering the interaction between visual features and corresponding semantic features, we propose a novel corresponding feature fusion (CFF) method to integrate multimodal features and build a semantic attention (SA) module to enable our model to focus on important information contained in different clinical questions
- Extensive experimental results illustrate the effectiveness of our proposed method on two benchmark datasets. Compared with previous state-of-the-art methods, our model achieves competitive performance in Med-VQA.

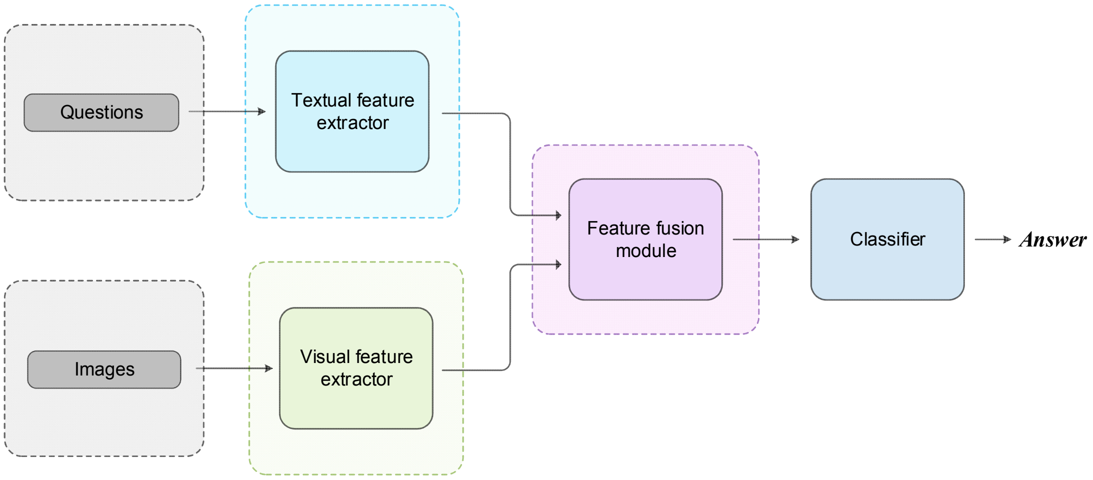


Figure 2. A basic model design for VQA. Visual and textual features are extracted respectively, and sent into the feature fusion module. The joint feature representations are then sent into the classifier for answer prediction.

2. Related works

2.1. Medical visual question answering

The structure of a VQA model in the medical domain is similar to that in the general domain. Generally, in a VQA framework, as shown in Figure 2, the following are required: 1) a visual feature extraction module to obtain the image feature representation, 2) a textual feature extraction module to obtain question feature representation, and 3) a feature fusion module to fuse the multimodal inputs and feed them into a final classifier for answer prediction. Most of the current methods [26−32] choose to use a CNN-based neural network such as ResNet or VGGNet for visual feature extraction. In [33−35], researchers used recurrent neural network (RNN)-based neural networks such as long short-term memory (LSTM) [36], gate recurrent unit (GRU) [37], or transformer-based models such as BERT [38] and BioBERT [39], to extract the textual features. Simultaneously, classical models such as stacked attention networks (SAN) [40], bilinear attention networks (BAN) [41], and multimodal compact bilinear pooling (MCB) [42] are commonly used for multimodal feature fusion to learn visual and textual joint feature representations.

In the past few years, methods such as meta learning and transfer learning have been introduced in modern few-shot tasks. Nguyen et al. designed the mixture of enhanced visual features (MEVF) [43] method from a large number of un-annotated radiology images, using model-agnostic meta-learning (MAML) [44] and convolutional denoising autoencoder (CDAE) [45] to initialize the model weights for the visual feature extraction. Li-Ming Zhan et al. [46] added a conditional reasoning (CR) module on the basis of MEVF; questions were divided into the two categories: "Open" and "Closed", according to the manner in which they were asked, to analyze them further. Khare et al. [47] proposed to pretrain the multimodal medical BERT on a ROCO dataset with a masked language modeling method introduced as a pretext task to learn richer feature representations. Do et al. [48] improved MAML [44] in meta-learning and proposed the multiple meta-model quantifying method without using external datasets for training; this increased the meta-data by auto-annotation and utilized the features output from meta-models for Med-VQA.

2.2. Multimodal learning

Robust feature representation is the condition that a model must fulfill to correctly predict the answer in Med-VQA. Feature extraction during the multimodal learning process is particularly critical. In [49], the author proved through extensive experiments that pre-training can greatly improve the model performance for a domain-specific task. Recently, Allaouzi et al. [23] proposed to use an external chest dataset [50] to pretrain a DenseNet-based neural network for visual feature extraction. Liu et al. [24] noticed that the brain, chest, and abdomen are mainly involved in the current radiological benchmark datasets; they pre-trained three visual feature extraction models targeting these three body regions through a contrastive learning method to obtain better feature representations. Gong et al. [25] used a multitask method to pre-train CNN-based neural networks in three external unlabeled radiology image datasets corresponding to the brain MRI [51], chest X-ray (https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia), and abdomen CT (https://www.synapse.org/#!Synapse:syn3193805/wiki/217753) to extract visual features. The above methods made certain progress in the visual channel process of multimodal input. Whereas, these methods focused extensively on learning better visual feature representations and overlooked the importance of textual features as well as the interactions between textual and visual channels in the multimodal learning processes. Based on the previous studies, we made further progress on the connection between specific visual features and corresponding semantic features while guiding the model to learn the pivotal information contained in various questions in a more targeted manner.

3. Materials and methods

In view of the latest studies on Med-VQA [24,25] and benchmark datasets [19,20], the radiology images mainly focus on three categories of human body regions: abdomen, brain, and chest. Motivated by this observation, as shown in Figure 3, we utilize a type classifier to classify each pair of multimodal inputs (radiology images and clinical questions) into given categories. A Semantic Attention (SA) module was built to help the model focus on semantic features of questions during the feature extraction stage. Thereafter, fusion is performed on the visual and textual features from the same category for the terminal answer prediction. Figure 3 presents an overview of our proposed method, which will be introduced in further detail in this section.

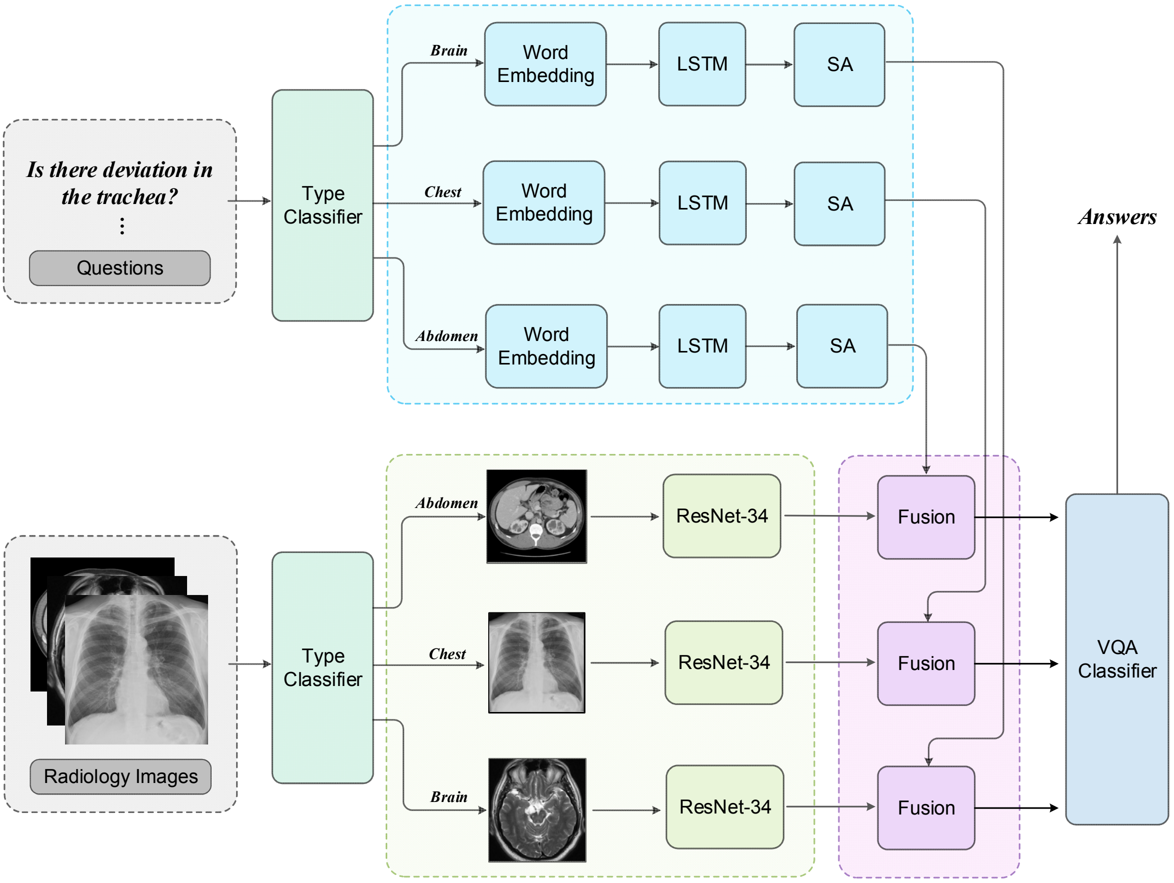


Figure 3. Overview of our proposed corresponding feature fusion (CFF) method. Classified images and questions are proceeded respectively. Corresponding features are fused and then sent into classifier for answer prediction.

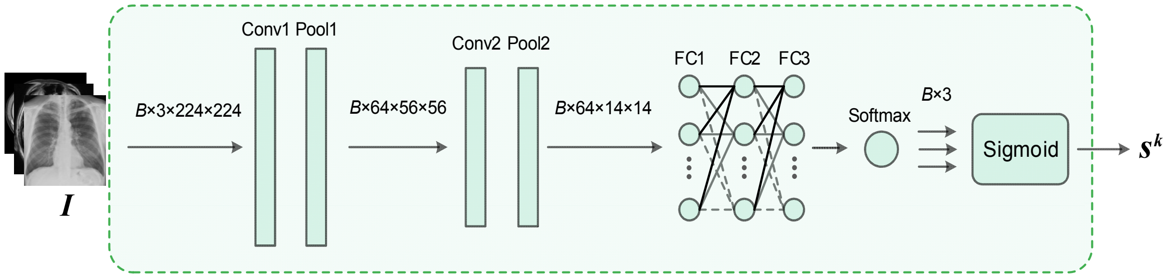


Figure 4. Framework of the type classifier where $k \in \{Abdomen, Brain, Chest\}$.

3.1. Radiology image and question classification

During the production stage of the Med-VQA dataset, doctors prepared questions based on the visual information presented by radiology images. Regarding a chest X-ray, doctors were more likely to ask a question, such as “what abnormalities are observed within the lungs?” rather than “where are the brain lesions located?". Considering this for a chest X-ray, we hope to fuse its visual features with corresponding textual features and then send it to a classifier for answer prediction. It would be confusing for the model if the visual features of a chest radiology image were combined with the textual features from a question that asks about brain diseases.

Based on the above considerations, we propose the CFF method. The preferential step of this method is to classify the input images and questions into specific categories. We first perform some preprocessings on the radiology images. We set the image size and number of channels in the format of $3 \times 224 \times 224$ to be consistent with the scale of the input images used in the pretraining stage of visual feather extractors, which we will elaborate in Section 3.2. Within one batch, we set the batch size as B and images $I \in \mathbb{R}^{B \times 3 \times 224 \times 224}$. For the questions, we set the maximum length of each question to 12; questions with a length less than 12 were zero-padded to ensure that the tensor dimensions are the same in the subsequent operation, and we use $q \in \mathbb{R}^{B \times 12}$ to denote input questions. Considering the scale of the datasets and to prevent overfitting, we designed a lightweight CNN-based type classifier module, as shown in Figure 4, to classify input radiology images with related questions. First, the module extracts visual features of input images I through two convolution and average pooling layers; subsequently, it sends the extracted visual features into a three-layer multilayer perceptron (MLP) for a nonlinear transformation. Ultimately, after the processes of the Softmax and Sigmoid layers, the classifier finally outputs the three-category prediction score S^k of the input images, $k \in \{Abdomen, Brain, Chest\}$ represents the scores of each category. And the highest score S^k corresponds to the category of the input images and related questions. The input images and questions in each batch are divided into three categories:

$$B = B^{Abdomen} + B^{Brain} + B^{Chest} \tag{1}$$

the classified radiology images $I^k \in \mathbb{R}^{B^k \times 3 \times 224 \times 224}$ and the questions $q^k \in \mathbb{R}^{B^k \times 12}$ are to be used for the follow-up works.

3.2. Visual feature extraction

Following the previous work [25], we send I^k from different categories into three ResNet-34 models, which are pre-trained in external radiology image databases that correspond to brain MRIs, chest X-rays, and abdominal CTs separately; pretrained visual feature extractors are utilized to extract the specific visual features contained in the input images from different categories:

$$\mathcal{V}^k = Visual\ \ Feature\ \ Extractor(I^k) \tag{2}$$

where $k \in \{Abdomen, Brain, Chest\}$ and $\mathcal{V}^k \in \mathbb{R}^{B^k \times 512 \times 7 \times 7}$ represents the extracted feature representations of the abdomen, brain and chest.

3.3. Semantic feature extraction

3.3.1. Textual feature representation

We chose to use 200-D BioWordVec [52], which is pre-trained on PubMed and MeSH (two open-source databases in the medical domain), to obtain the word embedding of each word contained in the question:

$$\tilde{q}^k = WordEmbedding(q^k) \tag{3}$$

where $q^k \in \mathbb{R}^{B^k \times 12}$ and $\tilde{q}^k \in \mathbb{R}^{B^k \times 12 \times 200}$.

Right after word embedding, in Eq (4), the 1024-D LSTM network was leveraged to extract textual features from the input questions $\tilde{q}^k$,

$$Q^k = LSTM(\tilde{q}^k) \tag{4}$$

and obtain the preliminary textual feature representations $Q^k \in \mathbb{R}^{B^k \times 12 \times 1024}$ of questions in different categories.

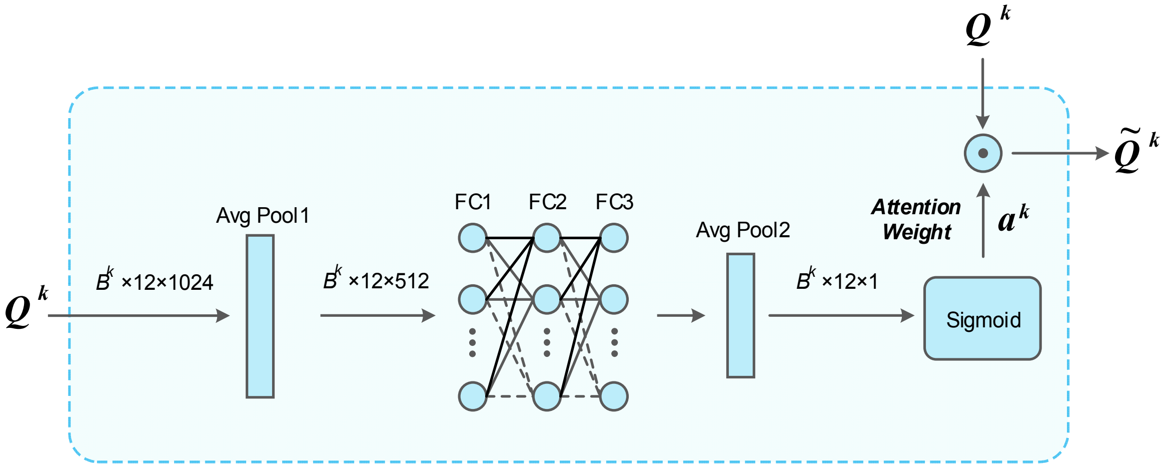


Figure 5. Framework of our proposed SA module where $k \in \{Abdomen, Brain, Chest\}$.

3.3.2. SA module

For each question, we hope that the model is able to distinguish the specific pathological nouns and the questioning methods contained in different question categories during the learning process as humans are able to do. For example, a clinical question such as, “What are the abnormal cranial nerves?”, in which “abnormal” implies that the question may be enquiring about a certain disease. Combining this with the phrase “cranial nerves” indicates that it is a brain-type question which may relates to some brain pathologies; and this instructs the model not to focus on answers with regard to lung or abdominal pathologies. Furthermore, the word “What” suggests that the answer to the question is possibly an open-type answer rather than a limited answer. Considering the aforementioned factors, distinguishing these specific semantic features will not only help the model learn better feature representations but also strengthen its understanding of questions.

To make our model focus more on this type of specific information, we designed an SA module to further process the textual feature representations Q^k of different questions. This step was inspired by [53]. The structure of SA is shown in Figure 5.

$$f^k = AvgPooling1(Q^k) \tag{5}$$

$$f^{k'} = MLP(f^k) \tag{6}$$

$$F^k = AvgPooling2\left(f^{k'}\right) \tag{7}$$

$$a^k = Sigmoid(F^k) \tag{8}$$

To begin with, we utilized an average pooling layer, in Eq (5), to initially obtain the global feature $f^k \in \mathbb{R}^{B^k \times 12 \times 512}$ of questions from different categories. Then, as shown in Eq (6), the global feature f^k was sent into a three-layer MLP for nonlinear transformation; meanwhile, the ReLU activation function was used to connect the layers. Subsequently, another average pooling layer, in Eq (7), was used to compress the global features of the question into $F^k \in \mathbb{R}^{B^k \times 12 \times 1}$. Next, we sent F^k into a Sigmoid layer to get the SA weight $a^k \in \mathbb{R}^{B^k \times 12 \times 1}$ of the whole questio; the value of this attention weight determines which semantic features in the question our model should focus on and what unnecessary information should be ignored. Finally, we multiplied a^k by the previously obtained textual feature representations Q^k:

$$\tilde{Q}^k = a^k \odot Q^k \tag{9}$$

where "$\odot$" indicates the dot product.

After processing of SA, we obtained the final semantic feature representations of different questions $\tilde{Q}^k \in \mathbb{R}^{B^k \times 12 \times 1024}$, which corresponds to the visual features extracted before, where $k \in \{Abdomen, Brain, Chest\}$.

3.4. Feature fusion and loss calculation

We obtained the corresponding features in the above work of the CFF. Next, each pair of the corresponding visual and semantic features from different categories were sent into the fusion module. After fusing $\mathcal{V}^k$ with its corresponding $\tilde{Q}^k$, we sent the joint feature representations into the VQA classifier for answer prediction:

$$\hat{p}^k = Classifier(Fusion(\mathcal{V}^k, \tilde{Q}^k)) \tag{10}$$

Meanwhile, we utilized a cross-entropy method for the loss calculation of answer prediction:

$$\mathcal{L}^k = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}^k \log(\hat{p}_{ij}^k) \tag{11}$$

where $k \in \{Abdomen, Brain, Chest\}$. p^k represents the real answer targets of different categories and $\hat{p}^k$ represents the predicted answer targets; n indicates the quantity of candidate answers the model needs to classify; in other words, it represents the total number of candidate answers; and m is the batch size of different categories. For all multimodal inputs in one batch, the model traverses candidate answers for each input and calculates the cross-entropy between the predicted answer targets and the real answer targets. The sum of the cross-entropy is the loss of answer predicting.

Notably, the total loss of answer prediction contains the sum of three categories $\{Abdomen, Brain, Chest\}$:

$$\mathcal{L}_{pred} = \mathcal{L}^{Abdomen} + \mathcal{L}^{Brain} + \mathcal{L}^{Chest} \tag{12}$$

In addition, the type classifier module participates in gradient backpropagation; therefore, we need to calculate its loss of classification and update the model parameters. We set the real category targets of the input image as y and the predicted category target as $\hat{y}$, which is calculated in Section 3.1.

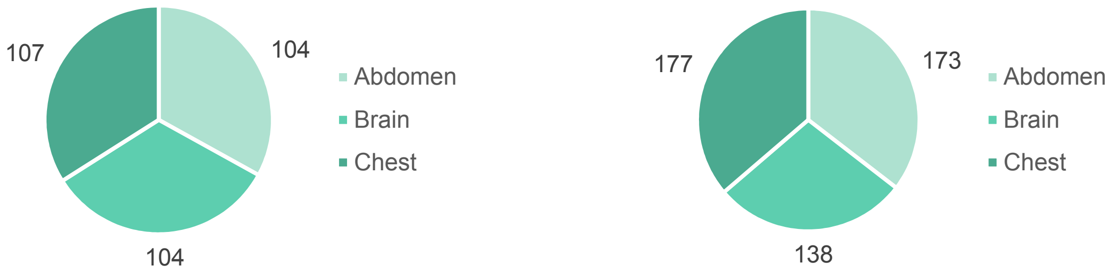
$$\mathcal{L}_{cls} = -\frac{1}{m'}\sum_{i=1}^{m'}\sum_{j=1}^{n'} y_{ij}\log(\hat{y}_{ij}) \tag{13}$$

In Eq (13), m' represents the total batch size that has not yet been classified, and n' represents the number of categories to be classified.

Lastly, we combined the losses of answer prediction and type classification as the final loss for the model evolution through a balancing approach:

$$\mathcal{L}_{final} = \lambda\ \mathcal{L}_{pred} + (1-\lambda)\mathcal{L}_{cls} \tag{14}$$

where λ is leveraged to balance the loss.



(a) VQA-RAD (b) SLAKE*

Figure 6. Radiology image statistics of VQA-RAD and SLAKE*. (a) Radiology image distribution in VQA-RAD. (b) Radiology image distribution in SLAKE* (* indicates the filtered version).

3.5. Datasets

Our model was validated on VQA-RAD [19] and filtered SLAKE [20]. VQA-RAD is a relatively well-recognized dataset in previous benchmarks. As shown in Figure 6(a), there are a total of 315 radiology images in VQA-RAD, including 104 abdomen CTs, 104 brain MRIs, and 107 chest X-rays; SLAKE is a recently proposed bilingual dataset for Med-VQA, which contains 642 radiology images and 7032 question-answer pairs. There are some clinical questions based on knowledge graph and radiology images that are not within the scope of our research; we followed the data distribution of VQA-RAD and screened out 488 radiology images. As shown in Figure 6(b), SLAKE* represents the dataset after filtering. The number of chest X-rays (177) exceeds the remaining image categories, followed by abdomen CTs (173), which are slightly fewer, and the brain MRIs constitute nearly 28% (138) of the radiology images. Figure 7 shows the comparative statistics of questions in different categories from two datasets. We calculated the number of questions corresponding to three categories and the number of “Open/Closed” questions in these two datasets. Here, “Open” and “Closed” refers to whether the question can be answered with limited options such as yes/no or with free-form texts. Thus, the questions are divided into two categories: (1) closed-ended questions, (2) open-ended questions. In particular, SLAKE* uses the original data split with reference to VQA-RAD, and there are a total of 8392 question-answer pairs generated by clinicians in these two datasets, which cover more than 10 aspects such as “Plane”, “Modality” and “Organ System”.

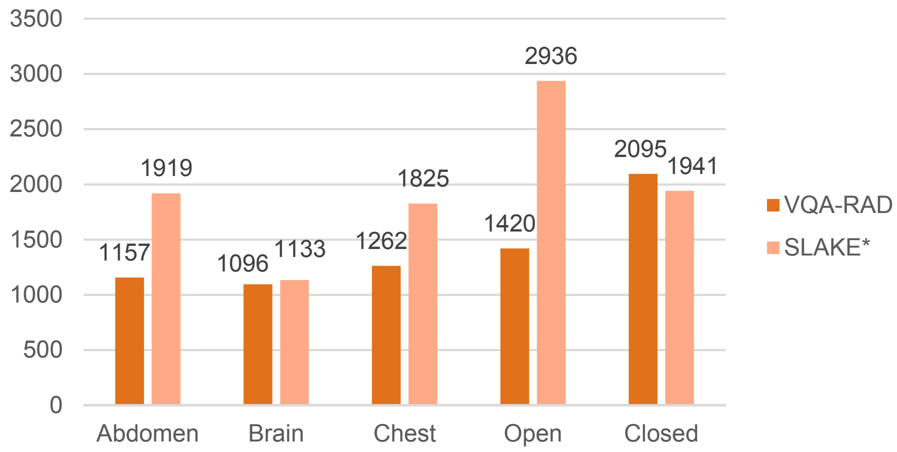


Figure 7. Question statistics of VQA-RAD and SLAKE* (* indicates the filtered version).

3.6. Evaluation metrics

Accuracy is generally used in Med-VQA experiments to evaluate the model performance and is calculated as follows:

$$Accuracy = \frac{N_C}{N_T} \times 100 \tag{15}$$

where N_C represents the number of correctly answered questions and N_T refers to the entire number of questions.

3.7. Evaluation Metrics

All of our experiments were conducted on the Ubuntu 16.04 operating system, and the graphics card used was Nvidia GTX 2080Ti; the deep learning framework was CUDA 10.2 and Pytorch 1.6.0 loaded on the Python programming language 3.7.0; we selected Adamax as the gradient descent optimizer.

Table 1. Experiment for hyperparameter selection.

Accuracy (%)	
Reference model	75.4
Reference: epochs = 200	
Reference: batch size = 64	
Batch size = 32	−0.8
Reference: RNN = LSTM	
RNN = GRU	−1.3
Reference: dropout = 0.5	
No dropout	
Reference: learning rate = 0.005	
Learning rate = 0.010	−1.1
Learning rate = 0.002	−0.4
Reference: gradual warmup steps = 0.0025, 0.0050, 0.0075, 0.0100	
Reference: decay rate = 0.25	
Decay rate = 0.30	−0.5
Decay rate = 0.20	−0.6
Reference: decay step = 38	
Decay step = 30	−0.4
Decay step = 45	−0.7

Before training, as shown in Table 1, we conducted an experiment on the selection of hyperparameters; references were the eventual parameter settings of our proposed model. The left side of the table shows the hyperparameters, and the right side presents their effects on the prediction accuracy of the model. Notably, “RNN” indicates the network we use for textual feature extraction. Furthermore, we utilize the warm-up learning method to speed up model convergence, where gradual warm-up steps indicate the learning rate setting during the warm-up period; the decay rate represents the decay ratio of the learning rate to the previous epoch in each decay step, and the decay step is the number of epochs contained in each decay period. Notably, we calculated the classification accuracy of the proposed type classifier in each epoch, and the current classifier parameters are saved for subsequent training only when the classification accuracy outnumbers the previous best result.

4. Results and discussion

4.1. Comparison with the State-of-the-Art

As shown in Table 2, we validated our model with five other state-of-the-art methods from different periods on VQA-RAD and SLAKE*.

We briefly review the previous methods. Kim et al. [41] proposed a method to extract joint feature representations from multimodal inputs through a low-rank bilinear pooling method while cutting down the consumption of learning attention distributions for each pair of multimodal input channels at the same time. MEVF + BAN [43] used model-agnostic meta-learning (MAML) [44] and a convolutional denoising autoencoder (CDAE) [45] to initialize the visual feature extractor and utilized the proposed MEVF framework to extract image features while combining BAN for feature fusion. MEVF + BAN + CR [46], on the basis of a previous work [43], added the CR module to process the open-ended and closed-ended tasks. Liu et al. [24] proposed a contrastive pre-training and representation distillation (CPRD) method that used contrastive learning to pre-train visual feature extraction networks on an open-source database and filtered the model for adaptability to small-scale datasets. In [25], Gong et al. considered the compatibility and applicability of the pre-trained features and proposed a cross-modal self-attention (CMSA) multimodal feature fusion method combined with a pre-trained visual feature extraction network for answer prediction.

Table 2. Performance comparison on VQA-RAD and SLAKE* datasets (* indicates the filtered version).

Methods	VQA-RAD			SLAKE*		
	Accuracy (%)					
	Overall	Open	Closed	Overall	Open	Closed
BAN [41]	58.3	37.4	72.1	77.1	75.6	78.7
MEVF + BAN [43]	66.1	46.2	77.2	79.2	77.6	80.4
MEVF + BAN + CR [46]	71.6	60.0	79.3	80.7	78.5	83.2
CPRD + BAN + CR [24]	72.7	61.1	80.4	–	–	–
CMSA [25]	73.2	61.5	80.9	81.9	79.9	**84.8**
CFF + SA + CMSA (Ours)	**75.4**	**62.6**	**83.9**	**82.4**	**81.3**	84.5

Experimental results illustrate the progressive results on both datasets after employing the CFF method combined with the SA module, which make certain progress based on former work. As shown in Table 2, our model achieves 2.2, 1.1, and 3.0% increase in accuracy for predicting “Overall”, “Open”, and “Closed” questions, respectively, compared to the current optimal method in VQA-RAD. In SLAKE*, our method achieves 0.5 and 0.4% increase in prediction accuracy of the “Overall” and “Open” questions, respectively, despite a 0.3% decrease in the prediction accuracy of “Closed” questions. Experimental results adequately demonstrate the effectiveness of our proposed model. Furthermore, our model could be further combined with a CR module [46] for an even better performance in Med-VQA.

4.2. Case study

We intuitively compare our model with the current optimal method [25] in more detail to further demonstrate the advantages of our proposed method. As shown in Figure 8, we selected five image-question pairs to calculate the model attention degree on the specific words contained in the questions during the training period. The figure clearly presents the comparison of our method and CMSA in the semantic comprehension of questions.

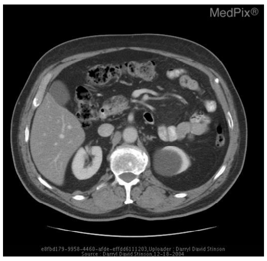


Image A

CMSA:

Q: What is the mass most likely?

A: Exophytic cyst ×

Ours:

Q: What is the mass most likely?

A: Kidney cyst √

	CMSA	Ours	
What	0.6812	0.523	What
is	0.7226	0.5157	is
the	0.7225	0.5162	the
mass	0.6815	0.5209	mass
most	0.7008	0.5169	most
likely	0.7140	0.5211	likely
[padding]	0.6829	0.5119	[padding]
[padding]	0.6446	0.5072	[padding]
[padding]	0.6189	0.5059	[padding]
[padding]	0.5902	0.5053	[padding]
[padding]	0.5787	0.5049	[padding]
[padding]	0.5688	0.5048	[padding]

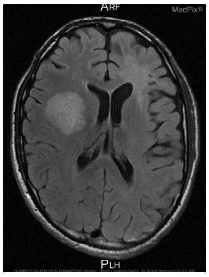


Image B

CMSA:

Q: What abnormality is seen on the left side of the frontal lobe?

A: Abscess ×

Ours:

Q: What abnormality is seen on the left side of the frontal lobe?

A: Regression of left frontal mass √

	CMSA	Ours	
What	0.7192	0.5221	What
abnormality	0.7208	0.5224	abnormality
is	0.7206	0.5171	is
seen	0.7254	0.5177	seen
on	0.7277	0.5176	on
the	0.729	0.5188	the
left	0.7286	0.5196	left
side	0.7234	0.5217	side
of	0.7212	0.5181	of
the	0.7216	0.5184	the
frontal	0.7239	0.5221	frontal
lobe	0.7272	0.5225	lobe

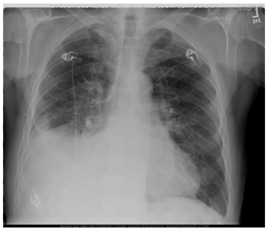


Image C

CMSA:

Q: Is the right diaphragm visualized?

A: Yes ×

Ours:

Q: Is the right diaphragm visualized?

A: No √

	CMSA	Ours	
Is	0.6739	0.5105	Is
the	0.6745	0.5136	the
right	0.6694	0.5243	right
diaphragm	0.7141	0.5234	diaphragm
visualized	0.7004	0.5192	visualized
[padding]	0.692	0.5104	[padding]
[padding]	0.6895	0.5079	[padding]
[padding]	0.6737	0.5062	[padding]
[padding]	0.6568	0.5054	[padding]
[padding]	0.6441	0.5051	[padding]
[padding]	0.6239	0.505	[padding]
[padding]	0.6177	0.5049	[padding]

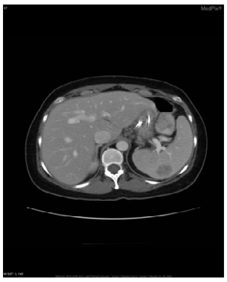


Image D

CMSA:

Q: Are there multiple or just one metastatic focus?

A: Not sure ×

Ours:

Q: Are there multiple or just one metastatic focus?

A: One √

	CMSA	Ours	
Are	0.654	0.5127	Are
there	0.6376	0.5131	there
multiple	0.6876	0.5173	multiple
or	0.6839	0.5234	or
just	0.6774	0.5193	just
one	0.7044	0.5187	one
metastatic	0.7198	0.5229	metastatic
focus	0.7222	0.52	focus
[padding]	0.6816	0.513	[padding]
[padding]	0.7122	0.5085	[padding]
[padding]	0.6852	0.5061	[padding]
[padding]	0.6698	0.5055	[padding]

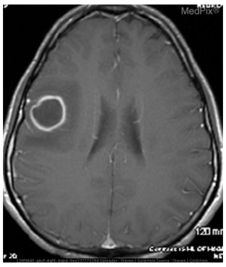


Image E

CMSA:

Q: Is this an MRI or a CT scan?

A: CT with contrast ×

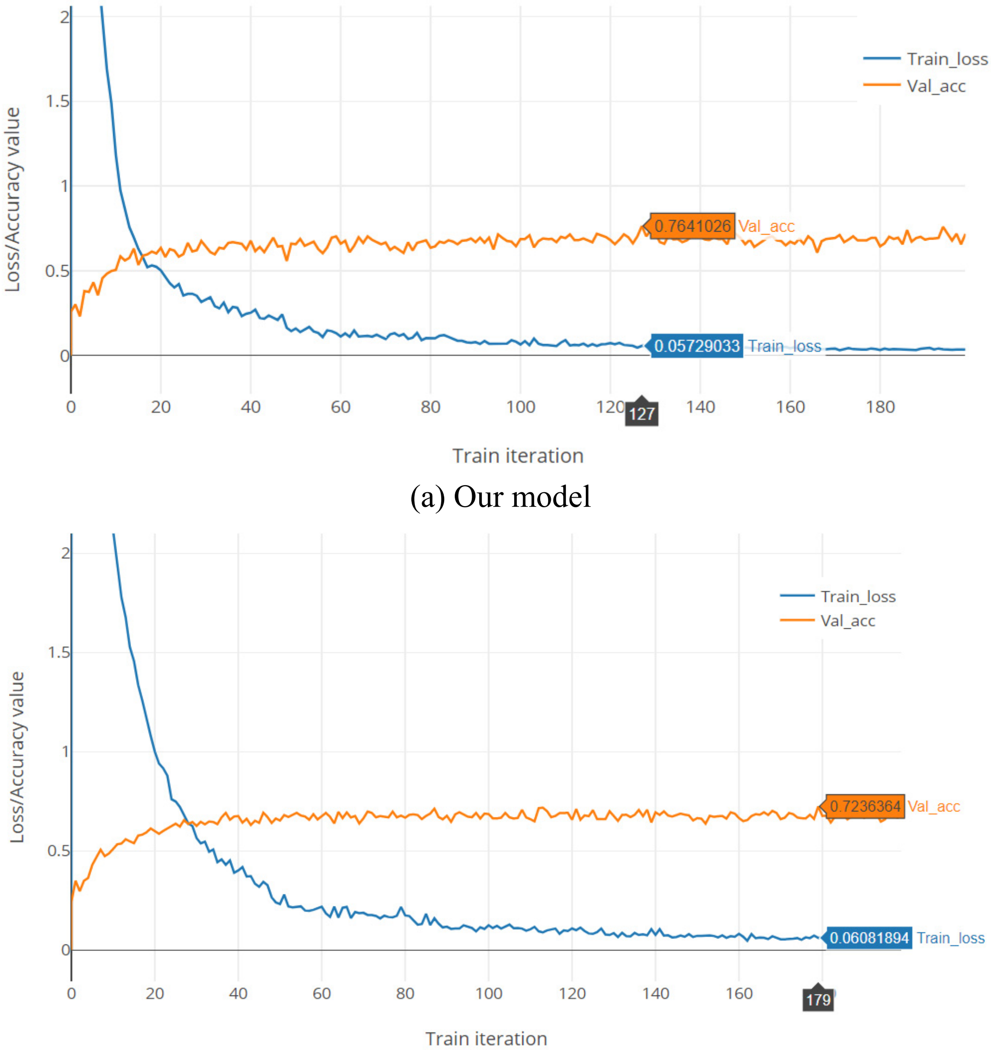
Ours:

Q: Is this an MRI or a CT scan?

A: MRI √

	CMSA	Ours	
Is	0.6739	0.5105	Is
this	0.659	0.515	this
an	0.6153	0.5151	an
MRI	0.7014	0.5206	MRI
or	0.7237	0.5259	or
a	0.7087	0.5179	a
CT	0.7195	0.5172	CT
scan	0.7257	0.5177	scan
[padding]	0.6927	0.5103	[padding]
[padding]	0.7132	0.5078	[padding]
[padding]	0.6765	0.5062	[padding]
[padding]	0.6602	0.5056	[padding]

Figure 8. Semantic attention comparison between our proposed model and CMSA.



(a) Our model

(b) CMSA

Figure 9. Comparison of the Loss/Accuracy curves. (a) The Loss/Accuracy curves of our model. (b) The Loss/Accuracy curves of CMSA.

The attention map in Figure 8 represents the attention paid to each word in the question; CMSA [25] is presented on the left side while our method is on the right. Notably, the darker red shade indicates that more attention is paid on the word and vice versa; “[padding]” means zero padding, and this has no implication in the question. It can be seen from the comparison that our model can focus better on the necessary information contained in a question and is more sensitive to information, such as pathology, interrogative pronouns, and orientation. It neglects “the”, “[padding]”, and other unnecessary information within a question. These concerns are in line with a human understanding of a question and thus can help the model predict answers more accurately and reasonably.

Figure 9(a),(b) shows the comparative training loss and validation accuracy between our proposed model and CMSA, respectively. The training loss decreased with increasing training iterations, and the training loss gradually converged to a stable value, proving the convergence of the model. Comparing these two figures, we can find that the training loss of our proposed method has a faster convergence speed. At approximately 20 epochs, the training loss has dropped to approximately 0.5, and at the 127th epoch, the training loss has dropped to approximately 0.0572; however, more epochs are needed for CMSA. Comparing the two models simultaneously, it can be seen that with increasing training iterations, our proposed model can achieve a higher answer prediction accuracy much earlier than CMSA; this proves the certain advantages that our proposed method has in Med-VQA.

4.3. Ablation study

VQA-RAD has been widely cited and recognized by previous studies. It is also more representative in Med-VQA. Therefore, to verify the effectiveness of our proposed model, we conducted experiments on the VQA-RAD dataset.

First, for fairness, we replaced CMSA with BAN as the multimodal feature fusion module and compared with the current state-of-the-art methods that employed BAN as a feature fusion module as well. As shown in Table 3, our method out-performed all other methods on the answer prediction accuracy of "Open" and "Closed" questions, with a prediction accuracy of the "Overall" questions that is only 0.2% lower compared with CPRD + BAN + CR. The experimental results objectively verify that our proposed method can still achieve competitive performance when combined with BAN for feature fusion.

Table 3. Ablation study of the feature fusion module.

Methods	VQA-RAD		
	Accuracy (%)		
	Overall	Open	Closed
MEVF + BAN [43]	66.1	46.2	77.2
CPRD + BAN [24]	67.8	52.5	77.9
MEVF + BAN + CR [46]	71.6	60.0	79.3
CPRD + BAN + CR [24]	**72.7**	61.1	80.4
CFF + SA + BAN (Ours)	72.5	**62.4**	**82.2**

Table 4. Ablation study of our proposed methods.

Methods	VQA-RAD					
	Accuracy (%)					
	Overall	Open	Closed	Abdomen	Brain	Chest
BAN [41]	58.3	37.4	72.1	–	–	–
BAN + CFF	71.3	57.7	79.8	64.8	73.4	77.6
BAN + CFF + SA	72.5	62.4	82.2	74.8	76.7	73.9
CMSA [25]	73.2	61.5	80.9	–	–	–
CMSA + CFF	74.3	59.3	82.6	68.5	78.3	75.9
CMSA + CFF + SA	75.4	62.6	83.9	71.0	78.2	77.6

Second, as shown in Table 4, ablation experiments were conducted on our model to further analyze the proposed CFF method and the SA module. We calculated the prediction accuracy of the questions from different categories for more intuitive comparison. Meanwhile, BAN and CMSA were combined with our method individually to compare the applicability of our model.

Comparing BAN with BAN + CFF in the table, it can be seen that the CFF method significantly improves the answer prediction accuracy of BAN [41]. The same enhancement also occurs after the combination of CMSA [25], even though the prediction accuracy of "Open" questions demonstrates a slight decline. After introducing SA module, the model shows a certain degree of improvement in the answer prediction performance on almost all types of questions compared to the former method, as observed from the table; the prediction accuracy of questions from different categories ($\{Abdomen, Brain, Chest\}$) is further improved, which affirms that the introduction of SA module can result in a positive impact, which leads to specific features contained in the questions from different categories being exploited. In addition, it can be seen from the comparison of BAN and CMSA that our model can achieve higher prediction accuracy when combined with CMSA; this shows that our method has better adaptability. In conclusion, we proved from experiments that our proposed method is able to exert positive impact on the model to obtain better answer prediction results in Med-VQA.

5. Conclusions

In this study, we propose a CFF method to strengthen the interactions between radiology images and questions from different categories in Med-VQA, utilizing a CNN-based type classifier to classify multimodal inputs and subsequently perform feature fusion for the corresponding image-question pairs. Notably, considering the specific semantic information contained in different questions, we propose an SA module to help our model continuously learn these specific semantic features during the training process and deepen the model's understanding of each question. In addition, extensive experiments were conducted on the benchmark dataset VQA-RAD and a recently proposed bilingual dataset SLAKE to verify the effectiveness of our proposed method. In contrast to previous state-of-the-art methods, our model surpasses several others in answer prediction and achieves better performance in Med-VQA. However, current methods for Med-VQA, including ours, still have certain limitations. The questions that can be answered by the model are only intuitive questions raised according to the content of clinical images. There are some shortcomings, such as limited interaction channels, fixed interaction scenes and narrow description range, which cannot meet the interaction needs of diversified channels in real clinical diagnosis. In order to solve this problem, we intend to introduce knowledge-based question answering (KBQA) [54] into Med-VQA in our future work. On the one hand, for a variety of clinical questions, the model can provide answers in combination with the giant medical information provided by external knowledge bases. On the other hand, the knowledge graph containing a large number of structured triples of medical knowledge [55], which can help the model answer multi-hop questions as well as the reasoning questions. If knowledge-based visual question answering methods [56] could be leveraged in the medical domain, it will better serve the needs of doctors and patients in real life, and help to promote the realization and application of intelligent inquiry in clinical diagnosis.

Acknowledgments

This study was partly supported by the Chengdu Major Technology Application Demonstration Project [No.2019-YF09-00120-SN] and the Sichuan Science and Technology Program [No.2021YFS0239].

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Z. Chen, X. Guo, P. Y. M. Woo, Y. Yuan, Super-resolution enhanced medical image diagnosis with sample affinity interaction, *IEEE Trans. Med. Imaging*, **40** (2021), 1377–1389. https://doi.org/10.1016/j.media.2020.101839
2. W. A. Al, I. D. Yun, Partial policy-based reinforcement learning for anatomical landmark localization in 3d medical images, *IEEE Trans. Med. Imaging*, **39** (2019), 1245–1255. https://doi.org/10.1109/TMI.2019.2946345
3. A. Jungo, R. Meier, E. Ermis, M. Blatti-Moreno, E. Herrmann, R. Wiest, et al., On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2018), 682–690. https://doi.org/10.1007/978-3-030-00928-1_77
4. Y. Tang, Y. Tang, Y. Zhu, J. Xiao, R. M. Summers, A disentangled generative model for disease decomposition in chest x-rays via normal image synthesis, *Med. Image Anal.*, **67** (2021), 101839. https://doi.org/10.1016/j.media.2020.101839
5. H. Abdeltawab, F. Khalifa, F. Taher, N. S. Alghamdi, M. Ghazal, G. Beache, et al., A deep learning-based approach for automatic segmentation and quantification of the left ventricle from cardiac cine MR images, *Comput. Med. Imaging Graphics*, **81** (2020), 101717. https://doi.org/10.1016/j.compmedimag.2020.101717
6. J. Ker, L. Wang, J. Rao, T. Lim, Deep learning applications in medical image analysis, *IEEE Access*, **6** (2017), 9375-9389. https://doi.org/10.1109/ACCESS.2017.2788044
7. X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, S. Yu, A survey on incorporating domain knowledge into deep learning for medical image analysis, *Med. Image Anal.*, **69** (2021), 101985. https://doi.org/10.1016/j.media.2021.101985
8. C. Li, G. Zhu, X. Wu, Y. Wang, False-positive reduction on lung nodules detection in chest radiographs by ensemble of convolutional neural networks, *IEEE Access*, **6** (2018), 16060–16067. https://doi.org/10.1109/ACCESS.2018.2817023
9. D. Bardou, K. Zhang, S. M. Ahmad, Classification of breast cancer based on histology images using convolutional neural networks, *IEEE Access*, **6** (2018), 24680–24693. https://doi.org/10.1109/ACCESS.2018.2831280
10. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, et al., Vqa: Visual question answering, in *IEEE International Conference on Computer Vision*, (2015), 2425–2433. https://doi.org/10.1109/ICCV.2015.279
11. P. Gao, H. You, Z. Zhang, X. Wang, H. Li, Multi-modality latent interaction network for visual question answering, in *IEEE/CVF International Conference on Computer Vision*, (2019), 5825–5835. https://doi.org/10.1109/ICCV.2019.00592
12. Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 6274–6283. https://doi.org/10.1109/CVPR.2019.00644

13. M. Malinowski, M. Fritz, A multi-world approach to question answering about real-world scenes based on uncertain input, *Adv. Neural Inf. Proces. Syst.*, **2014** (2014), 1682–1690.
14. M. Ren, R. Kiros, R. Zemel, Exploring models and data for image question answering, *Adv. Neural Inf. Proces. Syst.*, **2015** (2015), 2953–2961.
15. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, *Int. J. Comput. Vision*, **123** (2017), 32–73. https://doi.org/10.1007/s11263-016-0981-7
16. Y. Zhu, O. Groth, M. Bernstein, F. Li, Visual7w: Grounded question answering in images, in *IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 4995–5004. https://doi.org/10.1109/CVPR.2016.540
17. Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering, in *IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 6904–6913. https://doi.org/10.1007/s11263-018-1116-0
18. B. Ionescu, H. Müller, R. Péteri, A. B. Abacha, M. Sarrouti, D. Demner-Fushman et al., Overview of the ImageCLEF 2021: Multimedia retrieval in medical, nature, internet and social media applications, in *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, Cham, (2021), 345–370. https://doi.org/10.1007/978-3-030-85251-1_23
19. J. J. Lau, S. Gayen, A. B. Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, *Sci. Data*, **5** (2018), 180251. https://doi.org/10.1038/sdata.2018.251
20. B. Liu, L. M. Zhan, L. Xu, L. Ma, Y. Yang, X. Wu, SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, in *IEEE International Symposium on Biomedical Imaging*, (2021), 1650–1654. https://doi.org/10.1109/ISBI48211.2021.9434010
21. A. B. Abacha, S. Gayen, J. J. Lau, S. Rajaraman, D. Demner-Fushman, NLM at ImageCLEF 2018 visual question answering in the medical domain, in *Working Notes of CLEF*, (2018).
22. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
23. I. Allaouzi, M. B. Ahmed, B. Benamrou, An encoder-decoder model for visual question answering in the medical domain, in *Working Notes of CLEF*, (2019).
24. B. Liu, L. Zhan, X. Wu, Contrastive pre-training and representation distillation for medical visual question answering based on radiology images, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2021), 210–220. https://doi.org/10.1007/978-3-030-87196-3_20
25. H. Gong, G. Chen, S. Liu, Y. Yu, G. Li, Cross-modal self-attention with multi-task pre-training for medical visual question answering, in *International Conference on Multimedia*, (2021), 21–24. https://doi.org/10.1145/3460426.3463584
26. S. Liu, X. Zhang, X. Zhou, J. Yang, BPI-MVQA: a bi-branch model for medical visual question answering, *BMC Med. Imaging*, **22** (2022), 79. https://doi.org/10.1186/s12880-022-00800-x
27. U. Naseem, M. Khushi, J. Kim, Vision-language transformer for interpretable pathology visual question answering, *IEEE J. Biomed. Health Inf.*, (2022), forthcoming 2022. https://doi.org/10.1109/JBHI.2022.3163751
28. J. Li, S. Liu, Lijie at imageclefmed vqa-med 2021: Attention model based on efficient interaction between multimodality, in *Working Notes of CLEF*, (2021), 1275–1284.

29. Q. Xiao, X. Zhou, Y. Xiao, K. Zhao, Yunnan university at vqa-med 2021: Pretrained biobert for medical domain visual question answering, in *Working Notes of CLEF*, (2021), 1405–1411.
30. N. M. S. Sitara, K. Srinivasan, SSN MLRG at VQA-MED 2021: An approach for VQA to solve abnormality related queries using improved datasets, in *Working Notes of CLEF*, (2021), 1329–1335.
31. H. Gong, R. Huang, G. Chen, G. Li, et al., Sysu-hcp at vqa-med 2021: A data-centric model with efficient training methodology for medical visual question answering, in *CEUR Workshop Proceedings*, (2021), 1613.
32. Y. Li, Z. Yang, T. Hao, Tam at vqa-med 2021: A hybrid model with feature extraction and fusion for medical visual question answering, in *Working Notes of CLEF*, (2021), 1295–1304.
33. A. Al-Sadi, H. A. Al-Theiabat, M. Al-Ayyoub, The inception team at VQA-Med 2020: Pretrained VGG with data augmentation for medical VQA and VQG, in *Working Notes of CLEF*, (2020).
34. K. Gasmi, Hybrid deep learning model for answering visual medical questions, *Supercomput.*, **2022** (2022), 1–18. https://doi.org/10.1007/s11227-022-04474-8
35. Z. Liao, Q. Wu, C. Shen, A. Van Den Hengel, J. Verjans, AIML at VQA-Med 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering, in *Working Notes of CLEF*, (2020).
36. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.*, **9** (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
37. K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, et al., Learning phrase representations using RNN encoder–decoder for statistical machine translation, preprint, arXiv:1406.1078.
38. J. Devlin, M. V. Chang, K. Lee, K. B. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (2019), 4171–4186. https://doi.org/10.18653/v1/N19-1423
39. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. So, et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, **36** (2020), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682
40. Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in *IEEE conference on computer vision and pattern recognition*, (2016), 21–29. https://doi.org/10.1109/CVPR.2016.10
41. J. H. Kim, J. Jun, B. T. Zhang, Bilinear attention networks, *Adv. Neural Inf. Process. Syst.*, **31** (2018), 1571–1581.
42. A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, preprint, arXiv:1606.01847.
43. B. D. Nguyen, T. T. Do, B. X. Nguyen, T. Do, E. Tjiputra, Q. D. Tran, Overcoming data limitation in medical visual question answering, in *Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, (2019), 522–530. https://doi.org/10.1007/978-3-030-32251-9_57
44. C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in *Proceedings of the 34th International Conference on Machine Learning*, (2017), 1126–1135.
45. J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked convolutional auto-encoders for hierarchical feature extraction, in *International conference on artificial neural networks*, (2011), 52–59. https://doi.org/10.1007/978-3-642-21735-7_7

46. L. Zhan, B. Liu, L. Fan, J. Chen, X. Wu, Medical visual question answering via conditional reasoning, in *The 28th ACM International Conference on Multimedia*, (2020), 2345–2354. https://doi.org/10.1145/3394171.3413761
47. Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, C. V. Jawahar, MMBERT: Multimodal BERT pretraining for improved medical VQA, in *IEEE 18th International Symposium on Biomedical Imaging*, (2021), 1033–1036. https://doi.org/10.1109/ISBI48211.2021.9434063
48. T. Do, B. X. Nguyen, E. Tjiputra, M. Tran, Q. D. Tran, A. Nguyen, Multiple meta-model quantifying for medical visual question answering, in *Medical Image Computing and Computer Assisted Intervention*, (2021), 64–74. https://doi.org/10.1007/978-3-030-87240-3_7
49. S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, et al., Don't stop pretraining: Adapt language models to domains and tasks, preprint, arXiv:2004.10964.
50. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2019), 590–597. https://doi.org/10.1609/aaai.v33i01.3301590
51. J. Cheng, Brain tumor dataset, *Figshare Datasets,* (2017). https://doi.org/10.6084/m9.figshare.1512427.v5
52. Y. Zhang, Q. Chen, Z. Yang, H. Lin, Z. Lu, BioWordVec, improving biomedical word embeddings with subword information and MeSH, *Sci. Data*, **6** (2019), 52. https://doi.org/10.1038/s41597-019-0055-0
53. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 7132–7141. https://doi.org/10.1109/CVPR.2018.00745
54. X. Wang, S. Zhao, B. Cheng, Y. Yin, H. Yang, Explore modeling relation information and direction information in KBQA, *Neurocomputing*, **471** (2022), 139–148. https://doi.org/10.1016/j.neucom.2021.10.094
55. M. Gao, J. Lu, F. Chen, Medical knowledge graph completion based on word embeddings, *Information*, **13** (2022), 205. https://doi.org/10.3390/info13040205
56. L. Liu, M. Wang, X. He, L. Qing, H. Chen, Fact-based visual question answering via dual-process system, *Knowl. Based Syst.*, **237** (2022), 107650. https://doi.org/10.1016/j.knosys.2021.107650