MAMF: A Multi-Level Attention-Based Multimodal Fusion Model for Medical Visual Question Answering

Shaopei Long[1], Zhenguo Yang[2], Yong Li[1], Xiaobo Qian[1], Kun Zeng[3],
and Tianyong Hao[1](✉)

[1] School of Computer Science, South China Normal University, Guangzhou, China
{shaopei-lauv,lycutter,haoty}@m.scnu.edu.cn

[2] School of Computer Science, Guangdong University of Technology, Guangzhou, China
yzg@gdut.edu.cn

[3] School of Computer Science, Sun Yat-Sen University, Guangzhou, China
zengkun2@mail.sysu.edu.cn

Abstract. Medical Visual Question Answering (VQA) targets at accurately answering clinical questions about images. The existing medical VQA models show great potential, but most of them ignore the influence of word-level fine-grained features which benefit filtering out irrelevant regions in medical images more precisely. We present a Multi-level Attention-based Multimodal Fusion model named MAMF, aiming at learning a multi-level multimodal semantic representation for medical VQA. First, we develop a Word-to-Image attention and a Sentence-to-Image attention to obtain the correlations of word embeddings and question feature to image feature. In addition, we propose an attention alignment loss which contributes to adjust the weights of image regions gained from word embeddings and question feature to emphasize relevant regions for improving the quality of predicted answers. Results on VQA-RAD and PathVQA datasets suggest that our MAMF significantly outperforms the related state-of-the-art baselines.

Keywords: Medical Visual Question Answering · Multimodal fusion · Attention mechanism · Deep learning · Medical image

1 Introduction

Visual Question Answering (VQA) has obtained extensive attention from numerous scholars dedicated to research Computer Vision (CV) [1, 2] or Natural Language Processing (NLP) [3, 4] in the past few years. As a specific domain of VQA, the purpose of medical VQA is to answer diagnostically a question asked on a medical image. An outstanding medical VQA model can profit both clinicians and sick person. It can provide subsidiary analysis for clinical diagnoses and therapeutics for doctors. In addition, a Medical-VQA system helps ask for medical consultation whenever patients need. Therefore, developing a medical VQA model helps relieve the burden of healthcare

H. Zhang et al. (Eds.): NCAA 2023, CCIS 1870, pp. 202–214, 2023.
https://doi.org/10.1007/978-981-99-5847-4_15

and make medical diagnoses and treatment more efficient. Although medical VQA has tremendous potential, researches on medical VQA still face many challenges. Compared with general VQA, medical VQA is more challenging. In the foremost, well-annotated medical VQA datasets for training model are extraordinarily rare, since they are time-consuming and strenuous to gain precise annotations by clinicians. For example, the manually annotated dataset VQA-RAD [5] includes varied types of questions but it contains only 315 radioactive pictures. Furthermore, some general VQA models cannot be adopted to develop Medical-VQA systems. The reason is that they always utilize extremely complex visual feature extraction modules such as Faster R-CNN [6] and ResNet-101 [7], which included a great deal of arguments and demanded to be trained with large datasets. The direct employment of these models may result in the overfitting issue. Furthermore, clinical questions are not only harder to be understanded for the VQA system as they are about professional medical knowledge, but also needed to be answered precisely as they are relevant to safety and health.

Some previous works [8, 9] attempted to utilize general VQA models and fine-tuned them on Medical-VQA datasets. Nevertheless, medical images and clinical questions were quite different from those of general VQA. Raghu et al. [10] proposed to transfer knowledge from general VQA, but they gained a subtle improvement. Nguyen et al. [11] employed Model-Agnostic Meta-Learning (MAML) [12] to obtain weights of the visual feature extractor. In addition, they utilized Convolutional Denoising Auto-Encoder (CDAE) [2] to make model more robust. Though these groundbreaking medical VQA works pushed forward the research field, they only focused on making better the feature extractor, while ignored inference module. Zhan et al. [13] concentrated on enhancing the inference ability of models. Specifically, they devised a Question-Conditioned Reasoning (QCR) module to identify the importance of each word. Besides, they proposed a task-conditioned reasoning (TCR) strategy to enlarge the difference of reference abilities for close-ended and open-ended tasks accordingly. Nevertheless, owing to the limitation of medical data, it can only obtain rough fusion features. Li et al. [14] designed two reasoning modules to obtain fine-grained relations between words and image regions. But they ignored the relationships between Word-to-Image attention and Sentence-to-Image attention, which make them unable to gain more fine-grained semantic information.

In order to gain a multi-level multimodal fusion feature, we design a Multi-level Attention-based Multimodal Fusion (MAMF) model by developing a Word-to-Image (W2I) attention and a Sentence-to-Image (S2I) to model the relations of both word embeddings and question feature to the image feature for medical VQA. The W2I attention is adopted to word-level fine-grained reasoning, while the S2I attention is applied to sentence-level coarse-grained reasoning. Besides, we propose an Attention Alignment Loss (AAL) to concentrate on adjusting the weights of the image regions learned from word embeddings and question feature to lay stress on crucial image regions and obtain multi-level multimodal semantic representation to predict the high-quality answer.

To sum up, our contributions are as follows:

1) A novel Multi-level Attention-based Multimodal Fusion (MAMF) model is proposed by developing a Word-to-Image (W2I) attention and a Sentence-to-Image (S2I) attention to capture word-level and sentence-level inter-modality relations of them, as well as to learn a multi-level multimodal semantic representation for medical VQA.

2) An Attention Alignment Loss (AAL) is designed to adjust the importance of the image regions obtained from word embeddings and question feature to identify the relevant and crucial image regions.
3) The evaluations on VQA-Rad and PathVQA datasets show that our proposed MAMF significantly superior to the related state-of-the-art baselines.

2 Related Work

VQA has aroused great research interest among scholars since Antol et al. [15] proposed the first VQA task. VQA models in general domain adopted various methods for extracting image feature and question feature. As for image feature extraction module, researchers commonly utilized object detectors like simple CNNs [16], SSD [17], and Faster-RCNN [6]. As for question feature extractor, they usually adopted models like GTP-3 [12], Bert [3] and RoBerta [18]. After that, the extracted features were aggregated by using bilinear pooling model like Multimodal Compact Bilinear Pooling [19], Multimodal Low-rank Bilinear Pooling [20] or Bilinear Attention Network (BAN) [21] to obtain a fusion feature. The feature was transmitted to the classifier to predict the answer.

However, these models could not be simply adopted to develop a Medical-VQA system, owing to the limitation of medical data. Therefore, Nguyen et al. [11] utilized a meta-learning algorithm MAML [12] and CDAE [2] to obtain weight initialization of visual feature extractor to learn visual features. Do et al. [22] proposed a multiple meta-model quantifying (MMQ) algorithm to learn meta-annotation. Nevertheless, they ignored the reasoning module of the models, which led to limit their performances. Consequently, Zhan et al. proposed a question-conditioned reasoning (QCR) module to adjust the weights of words and task-conditioned reasoning (TCR) method to learn inference abilities for close-ended tasks and open-ended tasks respectively. Gong et al. [23] designed a novel multi-task learning paradigm. However, this needed large-scale medical data. Bo et al. [24] adopted contrastive learning to gain several cumbersome models and train an unsophisticated student model by distilling these models and fine-tuning on VQA-RAD dataset.

Various attention mechanisms were also adopted in the medical VQA field. Vu et al. [25] proposed a multi-glance attention method to obtain the most related image regions. Sharma et al. [26] proposed a MedFuseNet to utilize a co-attention mechanism to improve the quality of fusion feature. However, these previous works neglected to learn multi-level multimodal feature representations which limited their performance. In this paper, we develop a Word-to-Image (W2I) attention and a Sentence-to-Image (S2I) attention to concentrate on learning a multi-level multimodal semantic representation.

3 Methods

3.1 Problem Formulation

Medical VQA is defined as a multiclassification problem. Given an image I and a question q, the output is the predicted answer $\hat{a}$. The both I and q are input into model f to obtain the predicted answer:

$$\hat{a} = \arg\max_{a \in A} f(a|I, q, \theta), \tag{1}$$

where A and a denote candidate answers and one of them, separately, and θ denotes all parameters.

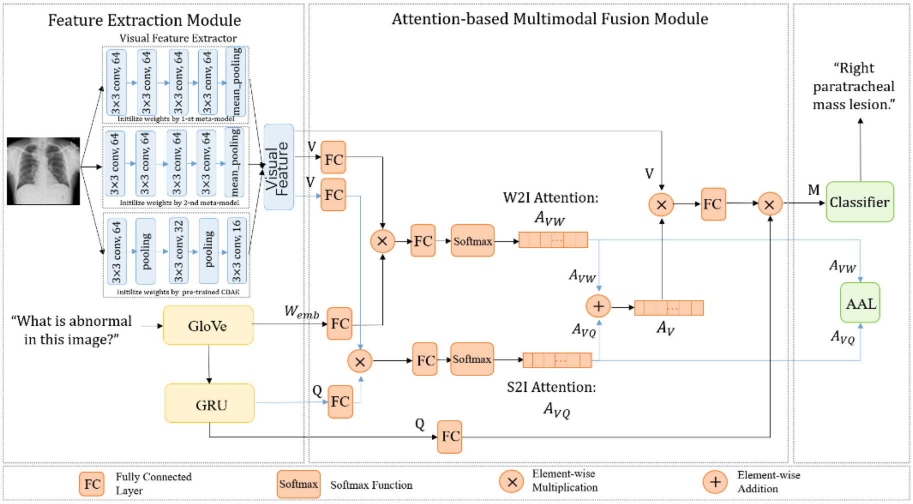

Fig. 1. Overview framework of our proposed MAMF. Each medical image gains three 64-D vector through a CDAE encoder and two meta-model. The vectors are concatenated to generate the visual feature V. GloVe and GRU are adopted to produce the word embedding sequence W_{emb} and the semantic feature Q. A_{VW} and A_{VQ} are W2I attention weight and S2I attention weight respectively, and M is a fusion feature.

3.2 Overview of Our Proposed Model

The structure of MAMF is shown in Fig. 1. Overall, the model includes a visual feature extractor, a word embedding module GloVe [27], a question embedding module GRU [28], an attention-based multimodal fusion module and a classifier. Glove is adopted to convert every word to a 300-dimension word. Then we utilized GRU to generates question feature. The visual feature extractor utilizes the Convolutional Denoising Auto-Encoder [2] and two meta-models obtained from Multiple Meta-model Quantifying (MMQ) [22]. The attention-based multimodal fusion module is adopted to model the relations between visual feature and word embeddings, and between visual feature and question feature, respectively. Finally, the classifier is adopted to classify multimodal semantic representations and then provide predicted answers to the Medical-VQA tasks.

3.3 Word Embedding and Question Representation

In the foremost, given a question q who has l words, GloVe [27] is adopted to generate a word embedding sequence. $w_i \in \mathbb{R}^{d_w}$ express the i-th word vector:

$$W_{\text{emb}} = WordEmbedding(q) = [w_1, ..., w_l]. \quad (2)$$

The word embedding $W_{emb} \in \mathbb{R}^{d_w \times l}$ is then sent to Gated Recurrent Unit (GRU) [28] whose dimension is d_G to gain the semantic feature:

$$Q = GRU(W_{emb}) = [\gamma_1, ..., \gamma_l], \tag{3}$$

where $Q \in \mathbb{R}^{d_G \times l}$, and γ_i is the i-th word embedding.

3.4 Visual Feature Extractor

As for the visual feature, we adopt the two best meta-models obtained from MMQ [22] and a CDAE [2] as visual feature extractor, as shown in Fig. 1. Specifically, each meta-model contains four 3*3 convolutional layers. Each convolutional layer includes 64 filters. Finally, the extractor gains three feature vectors. We concatenated them to obtain the visual feature. It is denoted as $V \in \mathcal{R}^{d_V}$, where $d_V = 192$ represents the dimension of the feature.

3.5 Attention-Based Multimodal Fusion Module

This module calculates the word-based attention A_{VW} and the sentence-based attention A_{VQ} using the following equations respectively.

$$A_{VW} = \text{softmax}(l \times w_1 \times ((w_2 \times V) \circ (w_3 \times W_{\text{emb}})) + b), \tag{4}$$

$$A_{VQ} = \text{softmax}(l\prime \times \text{w}_1' \times ((w_2' \times V) \circ (w_3' \times Q)) + b'), \tag{5}$$

where l and w_x represent the weight matrix and a fully connected layer, respectively, and b denotes a scalar. Besides, $\circ$ indicates element-wise multiplication. The softmax functions in Eq. (4) and Eq. (5) are adopted to normalize the attention weights.

The attention weight of image feature is computed as:

$$A_V = A_{VQ} + A_{VW}. \tag{6}$$

The attention weight A_V and visual feature V are then element-wise multiplied to obtain the visual feature,

$$V' = A_V \circ V. \tag{7}$$

The visual feature and the question feature are both sent to fully connected layers. The vectors from the fully connection layers are element-wise multiplied together to obtain the joint embedding M. M is then sent to a classifier. The predicted answer $\hat{a}$ has the highest probability among the candidate answers. The accuracy is computed as follows:

$$Accuracy = \frac{1}{n_{Test}} \sum^{Test} (Onehot(\arg\max(\hat{a})) \cdot a), \tag{8}$$

where a denotes the correct answer of the task.

3.6 Loss Function

The predicted answers are utilized to obtain binary cross entropy loss during training,

$$L_{CE} = -\frac{1}{n_{Train}} \sum_{i=1}^{n_{Train}} (a \log(\hat{a}) - (1 - a) \log(1 - \hat{a})). \tag{9}$$

In addition, an Attention Alignment Loss (AAL) is proposed to align the word-based attention and the sentence-based attention to emphasize relevant and crucial image regions. The loss function is computed as follows:

$$L_{AAL} = -\frac{1}{\mathrm{n_{Train}}} \sum_{i=1}^{\mathrm{n_{Train}}} \left\| A_{VQ} - A_{VW} \right\|^2. \tag{10}$$

At last, the final loss function is calculated as follows:

$$Loss = \alpha L_{AAL} + L_{CE}, \tag{11}$$

where α is a weighting parameter.

4 Experiments

4.1 Datasets

The prevalent medical VQA datasets are adopted to evaluate our proposed MAMF: (1) VQA-RAD [5]: It contains 3,515 question-answer pairs and 315 radiology images. Some questions are related to the same image. The clinicians or patients ask various questions about position, presence, organ and others. (2) PathVQA [29]: It contains 32,799 question-answer pairs, including “how”, “what”, “where” and other types. There are 3,328 medical images obtained from the PEIR digital library and 1,670 pathological images selected from several medical literatures. The answer types of two datasets are classified as close-ended and open-ended. The close-ended answers are “yes/no” or several options, while the open-ended answers are free-form texts. The question-answer pairs of PathVQA dataset are generated by a semi-automated approach using image captions and then manually reviewed and modified by clinicians.

4.2 Experiment Settings

All experiments are performed on the Ubuntu 20.04.4 server with NVIDIA GTX 1080 GPU based on PyTorch library in version 1.8. We adopt Adam optimizer to optimize our model. The learning rate is set to 1e–4 and batch size is set to 128. For semantic textual features, each question contains 12 words. GloVe [27] is utilized to generate the word embeddings. They are input into GRU [28] to gain question feature. As for visual representations, each 128-dimensional image is input into 2 quantified meta-models obtained from the MMQ [22] and a Convolutional Denoising Auto-Encoder, which generates 3 vectors. The enhanced visual feature is produced by concatenating these vectors. We adopt accuracy, precision, recall and F1-score (denoted as Acc, P, R, F1) as evaluation metrics.

4.3 Baseline Models

The medical VQA baselines including MAML, BiAN, MEVF [11], MMQ [22], CR [13] and CMSA [26] are reimplemented by using the open-source codes. The brief descriptions of baselines are in Table 1.

Table 1. The brief descriptions of baseline models.

Models	Descriptions
MAML	It utilized model-agnostic meta-learning method to obtain semantic representations
BiAN	It utilized ImageNet [30] to initialize the weights of the visual feature extractor
MEVF	It adopted MAML [12] and CDAE [2] to extract visual feature, and then used BAN to fuse them with question features
MMQ	It designed a multiple meta-model quantifying module to utilize meta-annotation
CR	It adopted a QCR module to improve fusion feature and proposed a TCR strategy
CMSA	It introduced a Cross-Modal Self-Attention module to effectively obtain the crucial semantic information

4.4 Results

The results of our proposed MAMF and other baseline models on the VQA-RAD test set are shown in the Table 2. The results of baseline models are re-implemented using available codes. From the table, it suggests that MAMF significantly superior to other state-of-the-art baselines. MAMF gains the best overall accuracy 74.94%, precision 82.39%, recall 74.94% and F1-score 78.02%. As for close-ended tasks and open-ended tasks, we also achieve the best performances except precision of the open-ended. Although we utilize the MMQ methods to enhance our image feature extractor, the reason may be that our model reduces the prediction probability of the true positive samples during fusion stage. The tasks corresponding to open-ended questions are harder for medical VQA models to answer correctly, since their answers can be free-form text. However, our proposed model MAMF still outperforms other baselines benefitting from the W2I attention, S2I attention, and AAL.

We also perform experiments on PathVQA dataset. Compared with VQA-RAD dataset, PathVQA have more diversities. It can verify the robustness of our proposed MAMF. The result is shown in Table 3. Our proposed MAMF gains the best performances reaching the best accuracy 54.28%, precision 65.82%, recall 54.28% and F1-score 52.38% on the entire test set. MAMF obtains dramatic improvement on the open-ended questions compared with other baseline models. The reasons of this improvement are as follows: First, MAMF builds word-level correlation representation of word embeddings and image feature, which filters unrelated regions in the image and retains essential ones for predicting answer. Second, our proposed AAL aligns the attention weights of regions in the image learned from the W2I attention and Q2I attention to recognize essential words and image regions for reasoning.

Table 2. Results on the VQA-RAD.

Models	Overall (%)				Open-ended (%)				Closed-ended (%)			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
MEVF	67.18	71.89	63.19	66.09	49.72	65.14	42.46	43.16	78.68	78.55	76.84	77.45
MMQ	71.80	82.17	72.06	75.71	60.90	**84.45**	61.45	61.71	79.01	81.08	79.04	80.39
CR	71.60	77.67	68.96	72.31	60.10	57.69	56.11	56.18	79.01	77.49	80.95	79.15
CMSA	73.17	79.73	73.17	75.35	61.45	73.17	61.45	60.71	80.88	82.38	80.88	81.46
MAMF	**74.94**	**82.39**	**74.94**	**74.94**	**65.36**	78.23	**65.36**	**65.81**	**81.25**	**83.63**	**81.25**	**82.28**

Table 3. Results on the PathVQA.

Models	Overall (%)				Open-ended (%)				Closed-ended (%)			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
BiAN	35.60	37.32	35.60	37.39	2.90	0.40	2.90	0.06	68.20	82.46	68.20	79.12
MAML	42.90	45.87	42.90	46.32	5.90	7.57	5.90	8.17	79.50	84.57	79.50	84.49
MEVF	44.80	40.28	44.80	40.84	8.10	2.01	8.10	2.50	81.40	83.31	81.40	81.99
MMQ	48.80	45.14	48.80	45.36	13.40	7.51	13.40	7.61	84.00	83.76	84.00	83.51
MAMF	**54.28**	**65.82**	**54.28**	**52.38**	**22.49**	**46.12**	**22.49**	**18.93**	**85.75**	**85.87**	**85.75**	**85.78**

4.5 Ablation Study

Several ablation experiments are conducted to verify the effectiveness of each part of MAMF. The experiment results are shown in Table 4 and Table 5. We remove W2I attention, S2I attention and AAL successively. The performances of MAMF without W2I attention and MAMF without S2I attention datasets dramatically decreased compared with the complete form of MAMF. Without the W2I attention, the model cannot establish word-level correlations between the word embeddings and image feature. Thus, it can only use the coarse sentence-level multimodal semantic representations to roughly reason. Without the S2I attention, the model can neither properly understand the meaning of questions nor predict the high-quality answers. These two ablation instances show the effectiveness of the W2I attention and S2I attention. As for the model MAMF without AAL, it also obtains poor performances on the two datasets. As discussed in Sect. 3.5, AAL is used to align the W2I attention and S2I attention, which helps locate crucial image regions to optimize the model. Furthermore, the complete form of MAMF obtains the best performance. Consequently, our proposed MAMF gains a satisfactory performance that utilizes the W2I attention and S2I attention to obtain the multi-level semantic information of image from word-level feature and sentence-level feature in the question, respectively, and employs the AAL to maximize the similarity of the relevant regions obtained from the W2I and S2I attention respectively.

Table 4. Ablation experiments on the VQA-RAD.

Models	Overall (%)				Open-ended (%)				Closed-ended (%)			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
MAMF w/o W2I	72.72	80.38	72.72	75.52	63.12	77.60	63.12	63.81	79.04	82.24	79.04	80.36
MAMF w/o S2I	72.28	80.45	72.28	75.17	60.33	75.05	60.33	61.10	80.14	82.90	80.14	81.34
MAMF w/o AAL	73.39	79.85	73.39	75.73	63.12	72.51	63.12	63.57	80.14	82.11	80.14	81.00
MAMF	**74.94**	**82.39**	**74.94**	**74.94**	**65.36**	**78.23**	**65.36**	**65.81**	**81.25**	**83.63**	**81.25**	**82.28**

Table 5. Ablation experiments on the PathVQA.

Models	Overall (%)				Open-ended (%)				Closed-ended (%)			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
MAMF w/o W2I	52.17	51.14	52.19	50.88	18.67	17.50	18.67	16.53	85.37	85.44	85.37	85.39
MAMF w/o S2I	51.97	49.48	51.97	49.31	18.55	14.59	18.55	14.00	85.13	85.69	85.13	85.17
MAMF w/o AAL	51.82	49.92	51.82	50.25	18.37	15.16	18.37	15.60	85.04	85.25	85.04	85.08
MAMF	**54.28**	**65.82**	**54.28**	**52.38**	**22.49**	**46.12**	**22.49**	**18.93**	**85.75**	**85.87**	**85.75**	**85.78**

Table 6. α changes from 0 to 2.0 in Eq. (11).

Model	Type	α										
		0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
MAMF	Open-ended (%)	63.12	63.13	63.10	60.33	63.12	63.88	**65.36**	63.70	61.50	61.66	60.10
	Closed-ended (%)	80.14	**81.99**	77.90	80.14	79.04	80.37	81.25	80.10	80.90	80.07	79.01
	Overall (%)	73.39	74.50	72.00	72.28	72.72	73.39	**74.94**	73.60	73.20	72.28	71.60

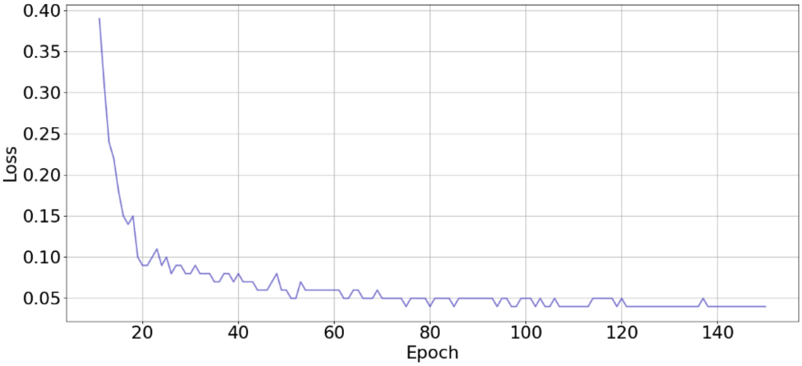

Fig. 2. The loss curve of MAMF.

4.6 Hyperparameter Analysis

We allocate distinct values of the hyperparameter α in the AAL in Eq. (11) and conduct experiments on the VQA-RAD dataset, as shown in Table 6. The overall task and open-ended task can gain the best performances when α is 1.2. Therefore, α is set to 1.2 during training our proposed model.

We train MAMF for 150 epochs. The loss curve and accuracy curve of MAMF are shown in Fig. 2 and Fig. 3, respectively. As shown from the Fig. 2, MAMF gains a relatively stable state after about approximately 150 epochs. From the Fig. 3, we can see that the accuracy curve also slowly becomes stable. Consequently, the value of hyperparameter epochs is set to 150 during training.

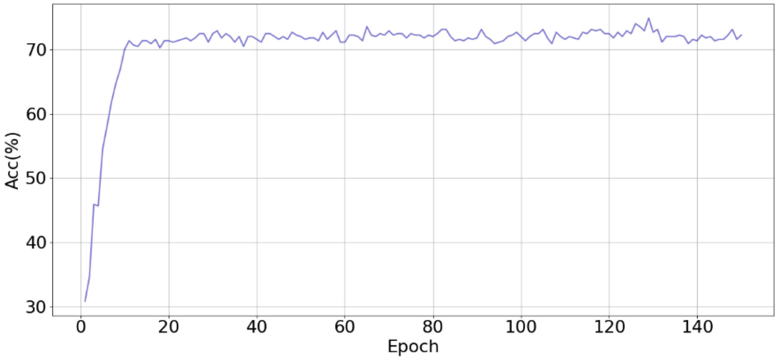

Fig. 3. The accuracy curve of MAMF.

4.7 Qualitative Evaluation

The qualitative evaluation of our proposed MAMF and the best baseline CMSA on the VQA-RAD dataset is shown in Fig. 4. For the first VQA task, while the baseline model

CMSA cannot select all the relevant regions to answer the clinical question, our proposed model locates all the related regions and correctly predicts the answer. The real position of the radiological image is completely opposite to what we see.

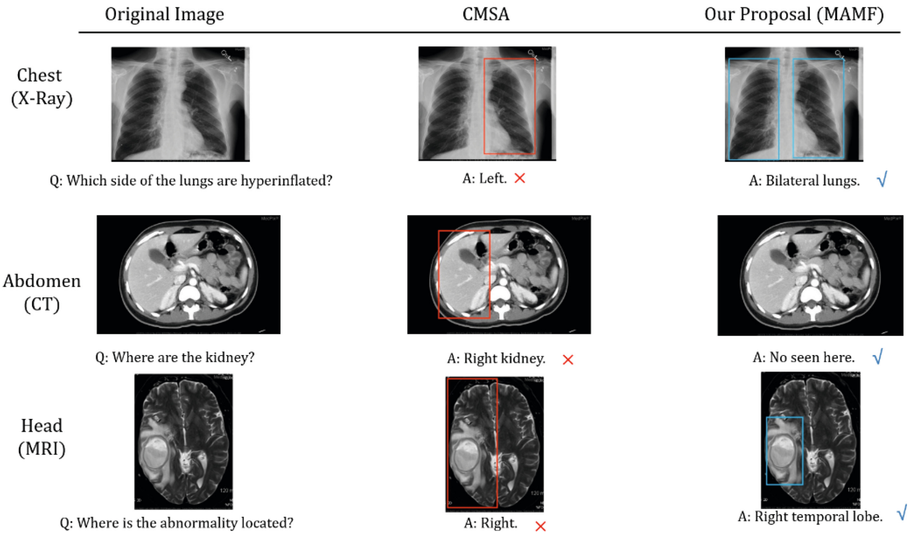

Fig. 4. Visualization of performances of our presented model MAMF and the baseline CMSA.

Therefore, “left” in the answer means the right region of the image. As for the second task, the CMSA identified the liver as the kidney, while our method finds that there is no kidney in the image. For the third task, the baseline can identify the related image region, but it could not recognize the concrete region to answer the question. In contrast, our model identifies the crucial image region and provides an accurate answer.

These instances show that our method has better ability to locate relevant and crucial regions in the medical image and understand well the clinical question. Therefore, it can provide concrete and accurate answer to complex Medical-VQA tasks.

5 Conclusion

This paper presents a Multi-level Attention-based Multimodal Fusion (MAMF) model. MAMF utilizes word embeddings and question features to identify the relevant and key regions of medical image by adopting a W2I attention and a S2I attention. It then contributes to obtain a multi-level multimodal semantic representation. Moreover, we propose an attention alignment loss to align the word-based attention and sentence-based attention to recognize relevant and crucial regions in medical images. This model is beneficial for clinicians in diagnosing different diseases. It also can help patients obtain the answers of health-related questions. Additionally, our model significantly outperforms related state-of-the-art baselines.

Acknowledgements. The work is supported by grants from Humanities and Social Sciences Research Foundation of the Ministry of Education, “Intelligent Analysis and Evaluation of Learning Effection Based on Multi-Modal Data” (No. 21YJAZH072).

References

1. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the International Conference on Machine Learning, pp. 1126–1135 (2017)
2. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: Proceedings of the International Conference on Artificial Neural Networks, pp. 52–59 (2011)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Hao, T., Li, X., He, Y., Wang, F.L., Qu, Y.: Recent progress in leveraging deep learning methods for question answering. Neural Comput. Appl. **34**, 2765–2783 (2022)
5. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Sci. Data **5**, 1–10 (2018)
6. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. **28** (2015)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Abacha, A.B., Gayen, S., Lau, J.J., Rajaraman, S., Demner-Fushman, D.: NLM at ImageCLEF 2018 visual question answering in the medical domain. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, Vol. 2125). CEUR WS.org, Avignon, France (2018)
9. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-Med: overview of the medical visual question answering task at ImageCLEF 2019. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, vol. 2380). CEUR-WS.org, Lugano, Switzerland (2019)
10. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Trans-fusion: understanding transfer learning for medical imaging. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, pp. 3342–3352. NeurIPS, Vancouver, BC, Canada (2019)
11. Nguyen, B.D., Do, T.-T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming data limitation in medical visual question answering. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 522–530. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_57
12. Brown, T., et al.: Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165 (2020)
13. Zhan, L.M., Liu, B., Fan, L., Chen, J., Wu, X.M.: Medical visual question answering via conditional reasoning. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2345–2354 (2020)
14. Li, Y., et al.: A Bi-level representation learning model for medical visual question answering. J. Biomed. Inform. **134**, 104183 (2022)
15. Antol, S., et al.: VQA: visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433 (2015)

16. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2015)
17. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016).https://doi.org/10.1007/978-3-319-46448-0_2
18. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q. V.: XLNet: generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237 (2019)
19. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)
20. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325 (2016)
21. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. arXiv preprint arXiv:1805.07932 (2018)
22. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A.: Multiple Meta-Model Quantifying for Medical Visual Question Answering. arXiv preprint arXiv:2105.08913 (2021)
23. Gong, H., Chen, G., Liu, S., Yu, Y., Li, G.: Cross-Modal Self-Attention with Multi-Task Pre-Training for Medical Visual Question Answering. arXiv preprint arXiv:2105.00136 (2021)
24. Liu, Bo., Zhan, L.-M., Wu, X.-M.: Contrastive Pre-training and representation distillation for medical visual question answering based on radiology images. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 210–220. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_20
25. Vu, M.H., Löfstedt, T., Nyholm, T., Sznitman, R.: A question-centric model for visual question answering in medical imaging. IEEE Trans. Med. Imaging **39**(9), 2856–2868 (2020)
26. Sharma, D., Purushotham, S., Reddy, C.K.: MedFuseNet: an attention-based multimodal deep learning model for visual question answering in the medical domain. Sci. Rep. **11**(1), 1–18 (2021)
27. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543 (2014)
28. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. In: Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pp. 103–111. Association for Computational Linguistics, Doha, Qatar (2014)
29. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: PathVQA: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
30. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)