An Effective Pre-trained Visual Encoder for Medical Visual Question Answering

Yefan Huang, Xiaoli Wang(✉), and Jinsong Su

School of Informatics, Xiamen University, Xiamen, China
yefanhuang@stu.xmu.edu.cn, {xlwang,jssu}@xmu.edu.cn

Abstract. Medical Visual Question Answering (Med-VQA) is a domain-specific task that answers a given clinical question regarding a radiology image. It requires sufficient prior medical knowledge, resulting in additional challenges compared to general VQA tasks. However, the lack of well-annotated large-scale datasets makes it hard to learn sufficient medical knowledge for Med-VQA. To address the challenge, this paper employs a large-scale medical multi-modal dataset to pre-train and fine-tune an effective model, denoted by ROCOGLoRIA. The model can locate semantic-rich regions implied in medical texts and extract local semantic-focusing visual features from the image. We propose to combine the global visual features with the weighted local visual features, for capturing fine-grained semantics in the image. We further incorporate ROCOGLoRIA as the visual encoder into baselines, to investigate whether it benefits Med-VQA. We conduct extensive experiments on three benchmark datasets and the results show that the method using ROCOGLoRIA as a pre-trained visual encoder outperforms strong baselines in the overall accuracy.

Keywords: Medical visual question answering · pre-trained visual encoder · fine-tuning · transfer learning

1 Introduction

Medical visual question answering (Med-VQA) aims to answer a given question based on a medical image. It has become a hot research topic in computer vision and natural language processing, which processes multi-modal information of visual images and textual language. Med-VQA task has great potential to benefit medical practice. It may aid doctors in interpreting medical images to obtain more accurate diagnoses with responses to closed-ended questions or help patients with urgent needs get timely feedback on open-ended questions raised. Different from general VQA, Med-VQA requires substantial prior domain-specific knowledge to thoroughly understand the contents and semantics of medical visual questions.

Inspired by general VQA, Med-VQA systems have witnessed great successes in recent years while still hindered by challenges [9]. These systems generally need

X. Yang et al. (Eds.): ADMA 2023, LNAI 14180, pp. 466–481, 2023.
https://doi.org/10.1007/978-3-031-46677-9_32

to be trained on well-annotated large-scale datasets to learn enough domain-specific knowledge for understanding medical visual questions. Although benchmark datasets [10,13,19,23] have been published, such as VQA-RAD, SLAKE, and PathVQA, we still suffer from the problem of data limitation on Med-VQA [7]. To tackle the problem, a native solution is data augmentation. VQAMix [9] focused on generating new Med-VQA training samples. However, it may incur noisy samples that affect the performance of models [9]. Recent studies adopted deep encoders to interpret the image and question for predicting an answer. They typically contain four main components: visual feature extraction, textual feature extraction, multi-modal fusion, and answer prediction [22]. Several studies adopted meta-learning to pre-train a visual encoder [6,24]. However, they did not consider the impact of linguistic representation on Med-VQA [9], or pre-trained the visual encoder for specific body regions, limiting the generalization ability to other settings [7]. Current studies adopt transfer learning to pre-train a visual encoder on external medical image-text pairs to capture suitable visual representations for subsequent cross-modal reasoning [4,7,9,20]. These approaches have been particularly successful by performing pre-training using large-scale medical image-text pairs without additional manual annotations. Following such studies, we investigate to what extent using external medical multi-modal datasets without any manual annotation can contribute to Med-VQA.

In this paper, we propose an effective and generic model, denoted by ROCOGLoRIA, to extract useful visual features for Med-VQA. We employ a large-scale medical multi-modal dataset, named by ROCO [26], to train ROCOGLoRIA. ROCO includes a diverse range of body organs and image modalities. Especially, we perform multi-task learning to pre-train an effective visual encoder by fine-tuning the GLoRIA model [12], which has been pre-trained on a medical dataset. We first fine-tune it using ROCO and output semantic-oriented visual features, which are then used as the input of the R2Gen [2] model to generate a medical report about the medical image. We consider both losses of GLoRIA and R2Gen to pre-train the final visual encoder and output both global and local weighted visual features in the medical image. These features are combined and fused with question features to predict a final answer. We further investigate the performance gains when incorporating our ROCOGLoRIA as a visual encoder into state-of-the-art Med-VQA methods. Overall, our contributions are summarized as follows:

- We propose an effective model to perform multi-task learning for pre-training an effective visual encoder for Med-VQA. The model is trained using a large-scale medical multi-modal dataset without additional manual annotations, which can overcome the data limitation problem for Med-VQA.
- Different from previous visual encoders used in Med-VQA, ROCOGLoRIA is pre-trained using medical images from a diverse range of body regions, which can achieve better generalization ability to a wide range of Med-VQA datasets.
- ROCOGLoRIA can locate semantic-rich regions implied in medical texts, and extract local semantic-focusing visual features from the medical image. This

can help understand the contents and semantics of the medical image for predicting a correct answer to the clinical question.

– We conduct extensive experiments on three public datasets. The results show that incorporating ROCOGLoRIA as the visual encoder into baselines can benefit Med-VQA. Our model also shows the new state-of-the-art performance over all three benchmark datasets, which may indicate a better generalization ability.

2 Related Work

2.1 Medical Visual Question Answering

Current Med-VQA systems mainly consist of four components [21]: visual feature extraction, textual feature extraction, multi-modal fusion, and answer prediction.

To address the challenge of lacking well-annotated large-scale datasets for Med-VQA, a native solution is to use data augmentation (e.g. VQAMix [9]). However, such methods may generate noisy training samples which hinder the Med-VQA models from learning effective visual representations. Therefore, recent studies adopted deep encoders to interpret the image and question for predicting an answer. As the textual feature extraction module was reported to have less impact on the Med-VQA task, most existing studies focused on improvements of the visual feature extraction [7].

Medical image features are crucial for Med-VQA. Early studies attempted to use pre-trained models based on large-scale natural scene image datasets (e.g. ImageNet [30]) as the visual extractor. However, they cannot achieve satisfactory performance on Med-VQA due to the data limitation problem [7]. Several studies improved the visual feature extraction module by adopting meta-learning, such as MEVF [24] and MMQ [6]. However, they did not consider the impact of linguistic representation on Med-VQA [9]. Besides, some studies pre-trained the visual encoder for specific body regions. For example, CPRD [22] utilized a large number of unlabeled radiology images to train three teacher models and a student model for the body regions of the brain, chest, and abdomen through contrastive learning, limiting the generalization ability to other settings [7]. Recent studies employed transfer learning by using medical image-text pair datasets to fine-tune the cross-modal pre-trained models [4,7,9,20]. For example, Eslami et al. [7] fine-tuned the CLIP [27] model using the ROCO dataset [26], demonstrating its effectiveness on Med-VQA. These approaches had been particularly successful by performing pre-training using large-scale medical image-text pairs without additional manual annotations. Along this line, we further investigate what extent using external medical multi-modal datasets without any manual annotation can contribute to Med-VQA.

With the effective visual encoder, existing systems employed multi-modal fusion models to generate representations for final answer prediction. SAN [32] and BAN [17] are two representative models, which are effective for Med-VQA.

However, they only support two modalities. MTPT-CMSA [8] introduces a cross-modal self-attention model to capture the long-range contextual relevance from more modalities. While the multi-modal fusion module in M^{3}AE [3] consists of two transformer models.

Several studies are focusing on the improvements of model framework and model reasoning. For example, MMBERT [16] used the transformer framework instead of the typical deep learning framework and employed the image features of Med-VQA for mask language modeling. This work also used ROCO [26] for pre-training. CR [33] proposed a novel conditional reasoning framework, which aimed to automatically learn effective reasoning skills for various Med-VQA tasks. Its research ideas had been adopted in many studies to improve the accuracy of Med-VQA models [7,8,22].

2.2 Transfer Learning

As Med-VQA suffers from the general lack of well-annotated large-scale datasets for training, most studies adopted transfer learning. For multi-modal tasks in the general domain, many studies have proposed multi-modal self-supervised models, such as CLIP [27], simCLR [1] and DALLE [28], to overcome the challenge of requiring additional annotated data. These models can be transferred to most tasks and are competitive with fully supervised methods. Tasks such as VQA, image captioning, and image-text retrieval, can benefit from these models. Inspired by this, current studies have emerged to pre-train effective visual representations. conVIRT [34] contrasts the image representations with the paired descriptive texts via a bidirectional objective between two modalities. GLoRIA [12] is a framework for jointly learning multi-modal global and local representations of medical images by contrasting attention-weighted image regions with words in the paired reports. In this paper, we use ROCO [26] to fine-tune GLoRIA [12] pre-trained on CheXpert [14]. Specifically, we perform the multi-task pre-training based on GLoRIA. The original GLoRIA can learn more image types after fine-tuning, and hence it can be applied to various datasets while ensuring the effectiveness of the extracted visual features.

3 Methodology

This paper proposes an effective model, denoted by ROCOGLoRIA, to extract semantic-oriented visual features for Med-VQA. Figure 1 provides an overview of our model architecture. We first perform multi-task pre-training to extract visual features by fine-tuning the GLoRIA model [12] and then incorporate our ROCOGLoRIA as a visual encoder into state-of-the-art Med-VQA methods for improving the answer prediction accuracy.

3.1 Multi-task Pre-training

The GLoRIA used for fine-tuning is trained based on CheXpert [14] which contains a total of 224,316 chest radiographs from 65,240 patients. We fine-tune it

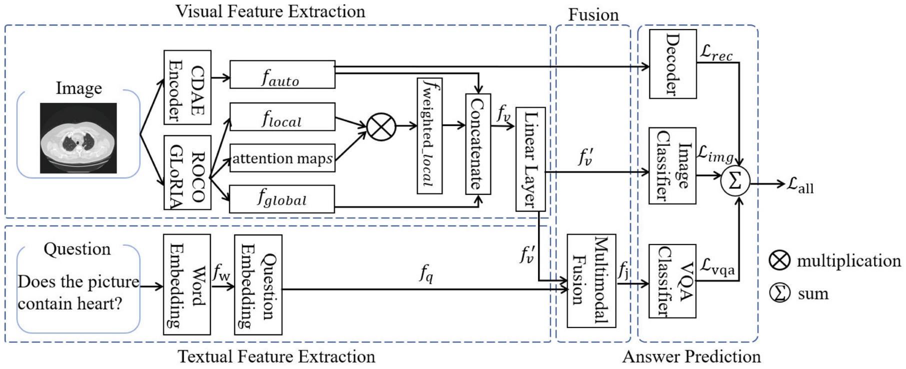

Fig. 1. Model architecture

using the ROCO dataset [26]. As shown in Fig. 2, during multi-task pre-training, we have two pre-training objectives: GLoRIA-based text-image pair similarity calculation [12] and R2Gen-based image report generation [2].

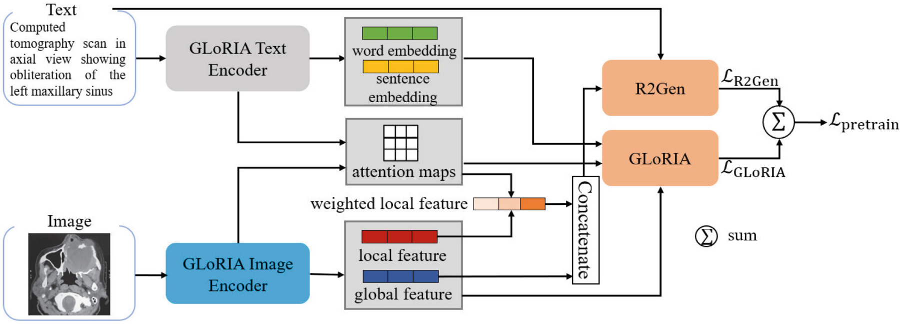

Fig. 2. Multi-task pre-training

GLoRIA-Based Similarity Calculation. Texts and images from ROCO are respectively encoded by the text encoder and the image encoder in GLoRIA. For texts, GLoRIA extracts the word embeddings denoted by f_{word} and the sentence embeddings denoted by f_s. For images, GLoRIA extracts the global features denoted by f_{global} from the final adaptive average pooling layer of ResNet50, and the local features denoted by f_{local} from an intermediate convolution layer. The local features are vectorized to the C-dimensional feature map for each of M image sub-regions. The generated f_{word}, f_s, f_{global}, and f_{local} are used in the GLoRIA model for similarity calculation between text-image pairs. At the same time, in the process of fine-tuning GLoRIA, we output word-based attention maps, which are subsequently multiplied with the image local features of f_{local} to obtain the attention-weighted local features denoted by $f_{weighted\_local}$.

R2Gen-Based Image Report Generation. This task is designed to further ensure the effectiveness of the visual features outputted by fine-tuned GLoRIA. The visual features are obtained by concatenating the global features of f_{global} and the weighted local features of $f_{weighted\_local}$ produced by GLoRIA. We feed them into R2Gen to generate a report about the medical image and calculate the loss denoted by $\mathcal{L}_{R2Gen}$. Specifically, we replace the visual extractor in R2Gen with visual features generated by GLoRIA. Formally, the multi-task loss function is defined as:

$$\mathcal{L}_{pretrain} = \mathcal{L}_{GLoRIA} + \mathcal{L}_{R2Gen}. \tag{1}$$

It is worthwhile to continuously fine-tune the GLoRIA model by using the reports generated from R2Gen. We use the fine-tuned GLoRIA model as a visual feature extractor in our proposed ROCOGLoRIA. The results in Sect. 4.6 also verify the effectiveness of our fine-tuned version, compared with the original GLoRIA model.

3.2 Our Med-VQA Model

In this paper, we evaluate the effectiveness of our proposed ROCOGLoRIA by incorporating it as a visual encoder into three advanced Med-VQA methods: CR [33], VQAMix [9] and M^{3}AE [3]. In Fig. 1, Our Med-VQA model consists of four main components: visual feature extraction for capturing visual features of the medical image, textual feature extraction for generating the embedding of the given question, fusion module for visual-textual feature fusion, and answer prediction.

We follow existing studies (e.g., [6,8,24]) to define the Med-VQA task as a classification task. Given an image denoted by v and an associated question denoted by q, the goal is to select a correct answer from a set of candidate answers denoted by A. F denotes our model, and α denotes the attention maps. Thus, the predicted answer denoted by $\hat{a}$ is formulated: $\hat{a} = \arg\max_{a \in A} F(a|v, q, \alpha)$.

Visual Feature Extraction. Most existing studies obtained the global visual features by averaging the visual feature vectors of different local regions in the medical image [6,24,29]. Existing studies found that the semantic-oriented medical content for general VQA may only lie in a particular area of the image [4]. Thus, the need arises to focus on local semantic-rich features in the medical image. However, it is much more challenging for Med-VQA, as the range of semantic-oriented regions associated with various types of questions can be very different. Single-use of global features or semantic-rich local features cannot achieve a satisfactory result. To tackle this problem, we propose ROCOGLoRIA, which can simultaneously extract global features and weighted local features of the medical image. The extracted features are then combined to capture useful visual information, which can help understand the contents and semantics of the medical image for predicting a correct answer to the clinical question.

Given an image-text pair, the original GLoRIA [12] model considers each word in the text to generate attention maps denoted by $\alpha \in \mathbb{R}^{L\times H\times W}$, L denotes the length of the text. Considering that the lengths of questions in a Med-VQA dataset are different, we need to ensure the dimensions of the weighted local features generated later are consistent. Thus, we convert raw attention maps to average attention maps denoted by $\alpha^{'} \in \mathbb{R}^{1\times H\times W}$, where $H \times W$ is the shape of the input medical image. The weighted local features can be calculated as $f_{weighted\_local} = \alpha^{'} f_{local}^{\mathrm{T}}$, where $f_{weighted\_local} \in \mathbb{R}^{1\times C}$, $f_{local} \in \mathbb{R}^{C\times H\times W}$, and C is the image feature dimension.

In this way, we exploit the attention maps in the question to localize the semantic-oriented regions in the medial image. We concatenate the global features of f_{global}, and the weighted local features of $f_{weighted\_local}$ to generate f_{gloria}, which can be concatenated with the features extracted by CDAE [24] denoted by f_{auto} to generate the final visual features: $f_v = [f_{gloria}; f_{auto}]$, where $f_{auto} \in \mathbb{R}^{1\times C}$ and $f_{global} \in \mathbb{R}^{1\times C}$. Finally, a linear layer is used to convert f_v to $f_v^{'}$ for multi-modal fusion.

Textual Feature Extraction and Multi-modal Fusion. We follow the MEVF [24] to perform textual feature extraction and multi-modal fusion. Specifically, each word of the question is represented as a 600-D vector, and then the word embedding denoted by f_w is fed into a 1024-D LSTM [31] to produce the question embedding denoted by $f_q \in \mathbb{R}^{12\times 1024}$. We do not directly use the text encoder in ROCOGLoRIA to generate question embeddings as the text encoder is trained on lengthy medical reports which may incur noisy information for embedding questions in Med-VQA. Subsequently, question embedding of f_q and visual features of $f_v^{'}$ are fed into the BAN [17] to generate the joint feature denoted by f_j. Noted that for M^{3}AE, we keep other modules in it and only replace its visual encoder with ROCOGLoRIA.

Multi-task Learning. The joint features of f_j are fed into a classifier for Med-VQA answer prediction. Existing studies showed that introducing multi-task learning into Med-VQA can effectively improve the accuracy of the model [4,8]. In this paper, to ensure the effectiveness of the visual features generated by the visual extraction component, we use the generated visual features of $f_v^{'}$ to perform the task of image classification for predicting the image type in parallel. Meanwhile, we also incorporate the CDAE into Med-VQA to adjust the loss function [24], and finally, our multi-task loss function is defined:

$$\mathcal{L}_{all} = \mathcal{L}_{vqa} + \beta\mathcal{L}_{img} + \mathcal{L}_{rec}, \tag{2}$$

where $\mathcal{L}_{vqa}$ is a cross-entropy loss for VQA classification, $\mathcal{L}_{rec}$ stands for the reconstruction loss of CDAE, and $\mathcal{L}_{img}$ represents the loss for image type classification. β is a hyperparameter for balancing the three loss terms. Noted that the $\mathcal{L}_{img}$ in the loss function is not used in our subsequent experiments in the PathVQA dataset [10] as no image type label is provided in this dataset.

4 Experiments

4.1 Datasets

For pre-training, we use a large-scale publicly available medical dataset, named ROCO [26]. It contains image-text pairs that are collected from PubMed articles, which cover a variety of imaging modalities such as X-Ray, MRI, angiography, etc. It also widely covers diverse body regions, such as the head, neck, teeth, etc. We select 87,952 non-compound radiological images with the associated captions, which provide rich semantic information about the content of images. This paper follows the training and validation data splits from the original paper [4].

Three benchmark Med-VQA datasets of VQA-RAD [19], SLAKE [23], and PathVQA [10] are used to train and evaluate our ROCOGLoRIA model. The VQA-RAD dataset contains 315 images and 3,515 corresponding questions posted and answered by clinicians. In this paper, we utilize the English subset of SLAKE dataset, denoted by SLAKE-EN. The SLAKE-EN dataset comprises 642 images and more than 7,000 question-answer pairs. The PathVQA dataset consists of 32,799 question-answer pairs generated from 1,670 pathology images collected from two pathology textbooks and 3,328 pathology images collected from the PEIR digital library[1]. For a fair comparison, we use the same data splits by following previous studies to partition the VQA-RAD dataset [6,24], the SLAKE-EN dataset [7], and the PathVQA dataset [6].

Medical visual questions are usually divided into two types: closed-ended and open-ended questions. Closed-ended questions are typically answered with "yes/no" or other limited choices; while open-ended questions do not have a restrictive structure and can have multiple correct answers.

4.2 Metrics

To quantitatively measure the performance of Med-VQA models, we use accuracy as the evaluation metric by following previous studies [6,8,24]. Let P_i and L_i denote the prediction and the label of sample i in the test set, and T represents the set of samples in the test set. The accuracy is calculated as: $accuracy = \frac{1}{|T|} \sum_{i \in T} l(P_i = L_i)$.

4.3 Competitors

We incorporate ROCOGLoRIA into three strong baselines for Med-VQA: CR [33], VQAMix [9] and M^{3}AE [3]. Our adapted methods are named **ROCOGLoRIA+BAN+CR**, **ROCOGLoRIA+BAN+VQAMix** and **ROCOGLoRIA+M^{3}AE**. We replace the visual encoder in these three methods using our

[1] https://peir.path.uab.edu/library/.

model. Our competitors are as follows:
MEVF [24] leverages the meta-learning MAML and deploys the auto-encoder CDAE for image feature extraction to overcome the data limitation problem.
CR [33] proposes a question-conditioned reasoning module and a type-conditioned reasoning module to automatically learn effective reasoning skills for various Med-VQA tasks.
CPRD [22] leverages large amounts of unannotated radiology images to pre-train and distill a lightweight visual feature extractor via contrastive learning and representation distillation.
MMBERT [16] is pre-trained using the ROCO dataset with masked language modeling using image features for Med-VQA.
MMQ [6] increases metadata by auto-annotation, dealing with noisy labels, and output meta-models which provide robust features for Med-VQA.
PubMedCLIP [7] is a fine-tuned version of CLIP in the medical domain based on PubMed articles.
MTPT-CMSA [8] reformulates image feature pre-training as a multi-task learning paradigm.
VQAMix [9] combines two training samples with a random coefficient to improve the diversity of the training data instead of relying on external data.
MKBN [11] is a medical knowledge-based VQA network that answers questions according to the images and a medical knowledge graph.
M^{3}AE [3] learns cross-modal domain knowledge by reconstructing missing pixels and tokens from randomly masked images and texts.

4.4 Implementation Details

All the models in this paper are implemented based on the PyTorch library [25] and run on a Ubuntu server with NVIDIA GeForce RTX 3090 Ti GPUs. The detailed training process is described below.

Pre-training. The pre-training model in this paper is implemented based on the ViLMedic library [5]. GLoRIA [12] was pre-trained on CheXpert [14] with ResNet50 as the visual feature extractor and R2Gen [2] was pre-trained on MIMIC-CXR [15]. The model is trained by the Adam optimizer [18] with an initial learning rate of 5e-5, and the learning rate is decreased by loss plateau decay.

Med-VQA Model. When adopting CR, the model is trained by the Adam optimizer [18] with an initial learning rate of 0.0009 on both VQA-RAD and SLAKE-EN, and an initial learning rate of 0.002 on PathVQA. When adapting VQAMix, we follow the original experimental settings for VQA-RAD and PathVQA. For SLAKE-EN, we set the epoch to 80 and the learning rate to 0.02. When adapting and reproducing M^{3}AE on PathVQA, the batch size is set to 16.

4.5 Performance

Table 1 summarizes the results of our model compared against the competitors on three benchmark datasets. We provide overall accuracy along with accuracy in answering open-ended and closed-ended questions. Our proposed model can achieve the best overall accuracy on three datasets, and yield the highest accuracy for open-ended and closed-ended questions. Other observations are summarized as follows:

(1) The performance of CR, VQAMix, and M^{3}AE is improved when adopting our ROCOGLoRIA as the pre-trained visual encoder. This verifies the effectiveness of ROCOGLoRIA for Med-VQA. And as shown in Table 1, the improvements in our model are effective for both open-ended and closed-ended questions.

(2) Our ROCOGLoRIA+BAN+CR achieves better overall accuracy than our ROCOGLoRIA+BAN+VQAMix on SLAKE-EN and PathVQA. On VQA-RAD, ROCOGLoRIA+BAN+VQAMix achieves better overall accuracy than ROCOGLoRIA+BAN+CR. This suggests that there may be underlying differences in the question type distribution in these datasets. It also verifies that VQAMix can give full play to its advantages on the small labeled dataset of VQA-RAD, by using data augmentation. While on large datasets like SLAKE-EN and PathVQA, this strategy may introduce noise to degrade the performance. Anyway, our ROCOGLoRIA can further enhance VQAMix by capturing both global and weighted local information, which helps to distinguish meaningless pairs and avoid introducing noise during training.

(3) Existing Med-VQA systems have shown too limited performance on PathVQA, as the pathological images contained in PathVQA are quite different from the clinical images contained in VQA-RAD and SLAKE-EN. The superior advantage of our model on PathVQA also indicates that our model can achieve better generalization ability to a wide range of Med-VQA datasets.

(4) Our model achieves the most significant improvements on open-ended questions in PathVQA. We achieve 289%, 38.8% and 9.5% absolute open-ended questions accuracy gain on CR, VQAMix, and M^{3}AEon PathVQA respectively.

(5) ROCOGLoRIA not only far outperforms classical MEVF, but also compares very well to the multi-stage pre-trained model M^{3}AE [3] and CPRD [22]. CPRD is pre-trained on head, abdomen, and chest images. In contrast, the pre-training images we use target a larger range of body parts. In other words, our model can achieve good results on more datasets. The pre-training of CPRD is more tailor-made for datasets of the three parts of the head, abdomen, and chest. Taking the SLAKE dataset as an example, the head, abdomen, and chest account for 88% of the total.

(6) Compared to advanced models that improve Med-VQA using ROCO, such as PubMedCLIP [7]), ROCOGLoRIA achieves 3.5% and 3% absolute overall accuracy gain on VQA-RAD and SLAKE-EN respectively. This verifies the

Table 1. Comparisons with competitors on three benchmark datasets. The highest and the second-highest accuracy are respectively marked in bold and by an underline. * indicates the reproduced results. The proposed methods yield the highest overall accuracy on every dataset and achieve the highest accuracy for each kind of questions in the datasets. Differences between the highest and second-highest accuracy are shown in brackets.

Dataset	Models	VQA accuracy (%)		
		Open	Closed	Overall
VQA-RAD	MEVF+SAN	49.2	73.9	64.1
	MEVF+BAN	49.2	77.2	66.1
	MMQ+SAN	46.3	75.7	64.0
	MMQ+BAN	53.7	75.8	67.0
	CPRD+BAN	52.5	77.9	67.8
	PubMedCLIP	48.9	76.7	65.5
	MTPT-CMSA	61.5	80.9	73.2
	MMBERT	63.1	77.9	72.0
	PubMedCLIP+CR	58.4	79.5	71.1
	MEVF+BAN+CR	60.0	79.3	71.6
	CPRD+BAN+CR	61.1	80.4	72.7
	MEVF+SAN+VQAMix	60.0	77.2	70.4
	MEVF+BAN+VQAMix	62.4	81.2	73.8
	M^{3}AE	67.2	83.5	77.0
	ROCOGLoRIA+BAN+CR	60.6	82.3	73.6
	ROCOGLoRIA+BAN+VQAMix	62.6	**83.8**(+0.1)	75.4
	ROCOGLoRIA+M^{3}AE	**67.6**(+0.4)	83.7(-0.1)	**77.4**(+0.4)
SLAKE-EN	MEVF+SAN	75.3	78.4	76.5
	MEVF+BAN	77.8	79.8	78.6
	CPRD+BAN	79.5	83.4	81.1
	PubMedCLIP	76.5	80.4	78.0
	MKBN_MGE	77.7	85.1	80.6
	PubMedCLIP+CR	78.4	82.5	80.1
	MEVF+BAN+CR	78.8	82.0	80.0
	CPRD+BAN+CR	81.2	83.4	82.1
	MEVF+SAN+VQAMix*	76.3	79.6	77.6
	MEVF+BAN+VQAMix*	77.4	79.1	78.0
	M^{3}AE	80.3	87.8	83.3
	ROCOGLoRIA+BAN+CR	**81.7**(+0.5)	83.7	82.5
	ROCOGLoRIA+BAN+VQAMix	80	84.9	81.9
	ROCOGLoRIA+M^{3}AE	81.1	**88.5**(+0.7)	**84.0**(+0.7)
PathVQA	MEVF+SAN	6.0	81	43.6
	MEVF+BAN	8.1	81.4	44.8
	MMQ+SAN	9.6	83.7	46.8
	MMQ+BAN	11.8	82.1	47.1
	MEVF+BAN+CR*	7.3	84.0	45.8
	MEVF+SAN+VQAMix	12.1	84.4	48.4
	MEVF+BAN+VQAMix	13.4	83.5	48.6
	M^{3}AE*	26.3	90.3	58.4
	ROCOGLoRIA+BAN+CR	28.4(-0.4)	85.5	57.1
	ROCOGLoRIA+BAN+VQAMix	18.6	83.3	51.1
	ROCOGLoRIA+M^{3}AE	**28.8**(+0.4)	**90.8**(+0.5)	**59.9**(+1.5)

effectiveness of our proposed pre-training and fine-tuning model that may capture more useful semantic information from Med-VQA.

Table 2. Ablation study on ROCOGLoRIA+BAN+CR. "GLoRIA" means using the original Gloria model as the visual encoder. "global" means using only global features. "local" means only using weighted local features. "w/o img classify" means not performing the image type classification.

	VQA-RAD	SLAKE-EN	PathVQA
ROCOGLoRIA+BAN+CR	**73.6**	**82.5**	**57.1**
GLoRIA	72.3	79.7	54.0
global	69.8	81.1	50.5
local	72.3	78.8	54.5
w/o img classify	69.8	81.5	–

4.6 Ablation Study

To verify the effectiveness of each component in our model, we conducted an ablation study based on our adapted method of ROCOGLoRIA+BAN+CR. The results are shown in Table 2. We observe that:

(1) Adopting ROCOGLoRIA as the pre-trained visual encoder in existing Med-VQA systems can further boost the overall accuracy of 1.3%, 2.8%, and 3.1% on VQA-RAD, SLAKE-EN, and PathVQA, respectively, compared to those by adopting the original GLoRIA.

(2) We evaluate the impact of global features and weighted local features on Med-VQA. Using only global features in our model, the performance drops by 3.8%, 1.4%, and 6.6% on VQA-RAD, SLAKE-EN, and PathVQA, respectively; while using only weighted local features, the performance of our model drops by 1.3%, 3.7%, and 2.6%, respectively. The most frequently asked questions in VQA-RAD are about the presence of an abnormality in the images. This requires the visual encoder to detect local features and abnormalities in the image. In this case, our model with better visual localization outperforms that using only the global features. Similarly, the frequently asked questions in PathVQA are about the abnormality in the images. Hence our model using only the local features also performs better in this dataset than that with global ones. However, on SLAKE-EN, the most frequently asked questions ask about the presence of organ type in the image. In this case, the visual encoder needs the overall understanding of the image content, and thus our model using only the global visual features achieves better accuracy than that using only the local ones.

(3) Employing image type classification in our multi-task learning scheme helps to boost the performance of Med-VQA. The results are consistent with previous studies [4,8]. After removing the image type classifier, the performance of our model drops by 3.8% and 1% on VQA-RAD and SLAKE-EN, respectively. No result is shown in PathVQA as no image type label is provided in this dataset.

4.7 Qualitative Analysis

We provide a qualitative comparison of our model with two competitors. Our goal is to illustrate the performance of the original MEVF and VQAMix in comparison with VQAMix when adopting ROCOGLoRIA as the visual encoder for Med-VQA. Examples from three datasets in Fig. 3 show that the MEVF model fails to answer Med-VQA questions. It cannot correctly understand the content of the medical image. For example, in the second left image in Fig. 3, we observe that the given image is from the lung, but the predicted answer from MEVF relates to the heart. In the same image, the original VQAMix model provides answers that located the correct organ to the given image. However, it fails to

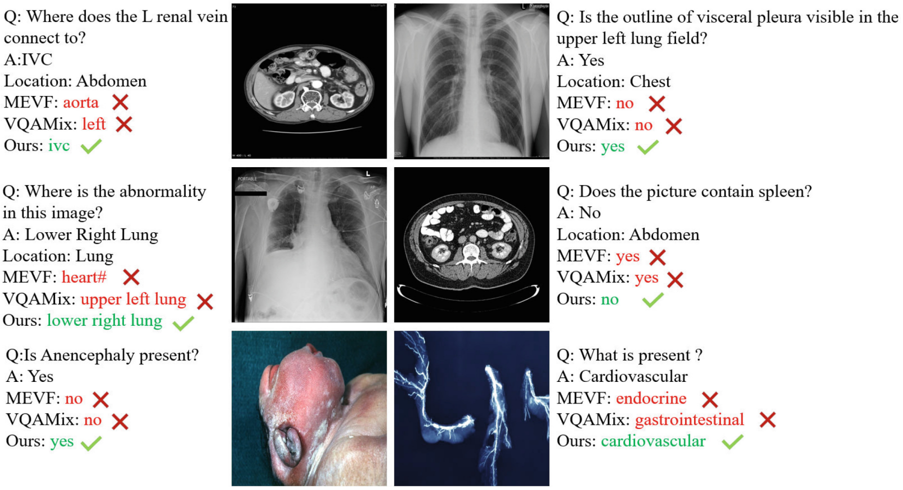

Fig. 3. Examples from VQA-RAD, SLAKE-EN, and PathVQA

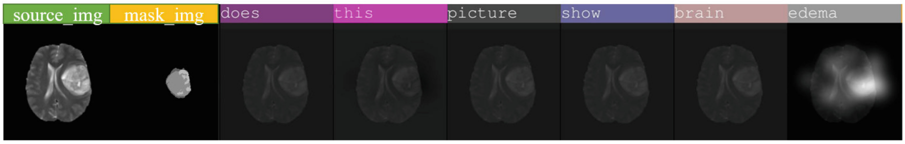

Fig. 4. Visualization of attention maps

predict a correct answer by pointing to a wrong region. In contrast, adopting ROCOGLoRIA as the visual encoder in VQAMix shows an improvement and results in providing answers that are correct throughout all examples. This also indicates that our model not only correctly captures image content holistically but also understands regions of interest as specifically associated with the questions to provide correct answers.

Moreover, we select an image from SLAKE-EN to visualize the attention maps of α generated by ROCOGLoRIA. In Fig. 4, we ask one question "does this picture show brain edema" for the left first medical image namely source_img. ROCOGLoRIA can correctly identify the semantic focusing regions of the image corresponding to each word in the question, and the region identified by "edema" is basically consistent with the annotated region of the image namely mask_img provided by SLAKE-EN. Correct attention maps guarantee the correctness of weighted local features of $f_{weighted\_local}$, which helps to predict a correct answer.

5 Conclusion

To address the problem of data limitation problem on Med-VQA, this paper employs a large-scale medical multi-modal dataset to fine-tune an effective model, denoted by ROCOGLoRIA, which can be used as a pre-trained visual encoder for enhancing existing Med-VQA systems. The model can locate semantic-rich regions implied in medical texts. Experimental results show that our model outperforms the state-of-the-art models. Considering global features and weighted local features at the same time can ensure our model has a better generalization ability.

Acknowledgments. The project was supported by the Natural Science Foundation of Fujian Province of China (No. 2021J01003).

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML, pp. 1597–1607. PMLR (2020)
2. Chen, Z., Song, Y., Chang, T.-H., Wan, X.: Generating radiology reports via memory-driven transformer. ArXiv, abs/2010.16056 (2020)
3. Chen, Z., et al.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022. MICCAI 2022. LNCS, vol. 13435, pp. 679–689. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_65
4. Cong, F., Xu, S., Guo, L., Tian, Y.: Caption-aware medical VQA via semantic focusing and progressive cross-modality comprehension. In: MM, pp. 3569–3577 (2022)
5. Delbrouck, J.-B., et al.: ViLMedic: a framework for research at the intersection of vision and language in medical AI. In: ACL, pp. 23–34 (2022)

6. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A.: Multiple meta-model quantifying for medical visual question answering. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 64–74. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_7
7. Eslami, S., de Melo, G., Meinel, C.: Does clip benefit visual question answering in the medical domain as much as it does in the general domain? ArXiv, abs/2112.13906 (2021)
8. Gong, H., Chen, G., Liu, S., Yu, Y., Li, G.: Cross-modal self-attention with multitask pre-training for medical visual question answering. In: ICMR, pp. 456–460 (2021)
9. Gong, H., Chen, G., Mao, M., Li, Z., Li, G.: VQAMix: Conditional triplet mixup for medical visual question answering. TMI **41**, 3332–3343 (2022)
10. He, X., Zhang, Y., Mou, L., Xing, E.P., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. ArXiv, abs/2003.10286 (2020)
11. Huang, J., et al.: Medical knowledge-based network for patient-oriented visual question answering. IP&M **60**(2), 103241 (2023)
12. Huang, S.-C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: a multimodal global-local representation learning framework for label-efficient medical image recognition. In: ICCV, pp. 3922–3931 (2021)
13. Huang, Y., Wang, X., Liu, F., Huang, G.: OVQA: a clinically generated visual question answering dataset. In: SIGIR, pp. 2924–2938 (2022)
14. Irvin, J., et al.: ChExpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI, vol. 33, pp. 590–597 (2019)
15. Johnson, A.E.W., et al.: Mimic-CXR: a large publicly available database of labeled chest radiographs. ArXiv, abs/1901.07042 (2019)
16. Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D., Jawahar, C.: MMBERT: multimodal bert pretraining for improved medical VQA. In: ISBI, pp. 1033–1036 (2021)
17. Kim, J.-H., Jun, J., Zhang, B.-T.: Bilinear attention networks. NeurIPS **31**, 1571–1581 (2018)
18. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. abs/1412.6980 (2015)
19. 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), 1–10 (2018)
20. Li, P., Liu, G., Tan, L., Liao, J., Zhong, S.: Self-supervised vision-language pretraining for medical visual question answering. ArXiv, abs/2211.13594 (2022)
21. Lin, Z., et al.: Medical visual question answering: a survey. ArXiv, abs/2111.10056 (2021)
22. Liu, B., 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., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 210–220. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_20
23. Liu, B., et al.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: ISBI, pp. 1650–1654 (2021)
24. Nguyen, B.D., Do, T.-T., Nguyen, B.X., Do, T.K.L., Tjiputra, E., Tran, Q.D.: Overcoming data limitation in medical visual question answering. In: MICCAI, pp. 522–530 (2019)
25. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. vol. 32, pp. 8024–8035 (2019)

26. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology objects in COntext (ROCO): a multimodal image dataset. In: Stoyanov, D., et al. (eds.) LABELS/CVII/STENT -2018. LNCS, vol. 11043, pp. 180–189. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01364-6_20
27. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748–8763. PMLR (2021)
28. Ramesh, A., et al.: Zero-shot text-to-image generation. In: ICML, pp. 8821–8831. PMLR (2021)
29. Ren, F., Zhou, Y.: CGMVQA: a new classification and generative model for medical visual question answering. IEEE Access **8**, 50626–50636 (2020)
30. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. IJCV **115**, 211–252 (2014)
31. Sundermeyer, M., Schlüter, R., Ney, H.: LSTM neural networks for language modeling. In: Interspeech (2012)
32. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR, pp. 21–29 (2015)
33. Zhan, L.-M., Liu, B., Fan, L., Chen, J., Wu, X.-M.: Medical visual question answering via conditional reasoning. In: MM, pp. 2345–2354 (2020)
34. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.: Contrastive learning of medical visual representations from paired images and text. ArXiv, abs/2010.00747 (2020)