

Medical Visual Question Answering via Conditional Reasoning and Contrastive Learning

Bo Liu¹, Student Member, IEEE, Li-Ming Zhan¹, Li Xu¹ and Xiao-Ming Wu¹, Member, IEEE

Abstract—Medical visual question answering (Med-VQA) aims to accurately answer a clinical question presented with a medical image. Despite its enormous potential in healthcare services, the development of this technology is still in the initial stage. On the one hand, Med-VQA tasks are highly challenging due to the massive diversity of clinical questions that require different visual reasoning skills for different types of questions. On the other hand, medical images are complex in nature and very different from natural images, while current Med-VQA datasets are small-scale with a few hundred radiology images, making it difficult to train a well-performing visual feature extractor.

This paper addresses above two critical issues. We propose a novel conditional reasoning mechanism with a question-conditioned reasoning component and a type-conditioned reasoning strategy to learn effective reasoning skills for different Med-VQA tasks adaptively. Further, we propose to pre-train a visual feature extractor for Med-VQA via contrastive learning on large amounts of unlabeled radiology images. The effectiveness of our proposals is validated by extensive experiments on existing Med-VQA benchmarks, which show significant improvement of our model in prediction accuracy over state-of-the-art methods. The source code and pre-training dataset are provided at <https://github.com/Awenbocc/CPCR>.

Index Terms—Medical visual question answering, conditional reasoning, contrastive learning.

I. INTRODUCTION

Medical visual question answering (Med-VQA) considers the problem of taking a medical image and a clinical question in natural language related to the image as input and inferring the correct answer also in natural language. Med-VQA has enormous potential in assisted diagnosis and patient education. It can help doctors to get a second opinion on diagnosis and reduce the high cost of training medical professionals. It can also help patients to get prompt feedback for their inquiries and better understand their disease and treatment, hence saving valuable medical resources and reducing the stress on medical facilities. As a domain-specific branch of visual question answering (VQA), the research of Med-VQA is still in an early stage, where the literature is rather limited.

This paper is an extended version of the conference paper [70].

Manuscript received August 26, 2022; revised October 20, 2022; accepted December 15, 2022. (Corresponding author: Xiao-Ming Wu)

Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu are with the Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong (e-mail: bokelvin.liu@connect.polyu.hk; lmzhan.zhan@connect.polyu.hk; cslxu@comp.polyu.edu.hk; xiao-ming.wu@polyu.edu.hk).



	Closed-ended	Open-ended
		
Question	Are there any abnormalities?	Where is the lesion in this image?
Answer	Yes	Left Lower Lung
Chest X-Ray		
		
Question	Is this an MRI image?	What is the organ on the left in the picture?
Answer	No	Liver
Abdomen CT		

Fig. 1. Examples of Med-VQA tasks. For closed-ended questions, the answers are limited, e.g., “yes” or “no”. For open-ended questions, the answers can be free-form text.

Hence, we start with introducing VQA, which has recently attracted a great deal of attention from both the computer vision and natural language processing research communities.

VQA focuses on visual perceptual tasks that require common perceptual abilities shared by humans. For example, given a scenery image with beautiful sunset, either a child or an adult can easily answer the question “what color is the sunset?”. Generally, visual perceptual tasks consist of easy tasks such as “does the man wear glasses?” and difficult tasks such as “which object in the picture has the same color as the pet dog in front of the man?”. It requires multi-level reasoning skills to solve both kinds of tasks. Easy perceptual tasks require basic skills, e.g., basic-level object recognition and scene understanding, while difficult tasks require higher-level reasoning skills such as counting, comparing, or logical inferring. Nevertheless, most of the existing VQA models are designed for coping with either easy tasks or difficult tasks. Simultaneously solving the two kinds of tasks in a single model is challenging and only considered in the high-data regime [46], [54].

Med-VQA tasks are, however, much more challenging than general VQA tasks. On the one hand, accurate answers are imperative for clinical questions, as they are related to health services and education. To this end, a Med-VQA system should be capable of handling multi-level tasks, including basic perceptual tasks such as identifying the body regions in an image, and difficult tasks such as counting the number of nodes, locating lesions, or evaluating the health of an organ by its size. Therefore, to infer correct answers, it is essential for the system to acquire domain-specific knowledge and multi-level reasoning skills. On the other hand, well-annotated

Med-VQA datasets are extremely lacking, since it requires medical expertise to construct high-quality datasets, which is both costly and time-consuming. To our best knowledge, there are only two manually annotated datasets available - VQA-RAD [36] and SLAKE [39]. Both of them only contain hundreds of radiology images but include various types of clinical questions. Therefore, it is not effective to train a typical large VQA model from scratch for Med-VQA with the small-scale training datasets. Moreover, it is impossible to apply popular object-detection-based VQA models such as UpDn [4], Pythia [26], and VL-BERT [57] for Med-VQA, due to the lack of visual object labels and the small size of training data.

Previous research tried to apply existing VQA models for Med-VQA. More specifically, they employed deep architectures pre-trained on general datasets such as ImageNet [34] and then fine-tuned the models on small-scale Med-VQA training data [1], [2], [71]. However, due to the large differences in image patterns and language styles of medical data and non-medical data [38], transfer learning provides little benefit [49]. To overcome this problem, [44] proposed mixture of enhanced visual features (MEVF) to learn an initialization for the visual extractor of a Med-VQA model. In particular, they combined an auto-encoder pre-trained with an image reconstruction task on undisclosed external medical datasets and a 4-layer convolutional neural network pre-trained with an auxiliary classification task on the VQA-RAD dataset [36] using the meta-learning algorithm MAML [14]. While this work alleviates the problem to some extent, it cannot be easily applied on other Med-VQA datasets since the auxiliary classification task is dataset-dependent and requires extra laborious annotations. Besides, it does not explore improving the reasoning module which is of critical importance in solving high-level reasoning tasks. There are some recent attempts [29], [43] to pre-train a multimodal Transformer on large medical vision-language datasets and fine-tune it on Med-VQA tasks. However, large Transformer models tend to overfit on existing small-scale Med-VQA training datasets.

In this paper, we explore lightweight models like MEVF [44] and focus on improving both the reasoning module and the visual feature extractor of a Med-VQA system. First, to make the system possess task-adaptive reasoning ability, we design a novel conditional reasoning mechanism, which includes a question-conditioned reasoning (QCR) module and a type-conditioned reasoning (TCR) strategy. QCR enables the model to gain question-specific reasoning skills by leveraging question attention information to modulate multimodal fusion features. Further, it can be seen that Med-VQA tasks mainly consist of two types, closed-ended questions and open-ended questions, as shown in Figure 1. For closed-ended questions, the answers are limited choices according to the prompt words, e.g., the answer to the question starting with “Does” can only be “Yes” or “No”. For open-ended questions, the answers are free-form, e.g., no fixed choices for questions starting with “What”. Generally, open-ended tasks are harder to solve than closed-ended ones, and current Med-VQA models usually perform poorly on open-ended tasks. Therefore, motivated by the disparity of the needed reasoning skills for open-ended and

closed-ended tasks, we design a TCR strategy to handle the two different types of tasks separately, by learning different sets of reasoning skills.

Second, to address the data scarcity problem, we propose to pre-train a visual feature extractor for Med-VQA in an unsupervised manner without requiring any human annotations. We observe that 1) there involve various types of organs and imaging modalities (e.g., brain MRI, brain CT, chest X-Ray, and abdomen CT) in Med-VQA tasks; and 2) there are many such types of unlabeled radiology images available in open-access sources. Therefore, we propose to leverage these publicly available datasets to pre-train a visual extractor for learning high-level patterns and characteristics of different organs and imaging modalities through contrastive self-supervised learning. Having learned prior knowledge of radiology images, the pre-trained feature extractor can be readily adapted to train Med-VQA systems, even with small-scale training datasets.

This paper makes the following contributions:

- We design a novel conditional reasoning mechanism to empower the reasoning ability of Med-VQA models, which contains a question-conditioned reasoning function and a type-conditioned reasoning strategy, by leveraging both question content and task type.
- We propose to leverage publicly available resources to pre-train a generic visual feature extractor for Med-VQA via contrastive self-supervised learning, which can be easily adapted to existing small-scale training datasets.
- We conduct an extensive evaluation on existing Med-VQA benchmarks to validate the effectiveness of the proposed conditional reasoning mechanism and the pre-trained visual feature extractor and observe significant improvements over state-of-the-art methods.

II. RELATED WORK

In this section, we briefly review the literature of previous studies in VQA and Med-VQA. In addition, we summarize recent advances in contrastive self-supervised learning.

A. Visual Question Answering

A typical VQA system consists of 4 basic components: (I) a visual feature extractor to obtain the visual image features; (II) a textual feature extractor to obtain the textual question features; (III) a multimodal feature fusion module to aggregate both the visual and textual features to produce a joint representation; (IV) a classifier to predict the final answers based on the joint representation. Various VQA systems differ in how they extract and combine multimodal features.

Early studies mainly employed VGGNet [55] and LSTM [20] to extract visual and textual features respectively and combined them by simple mechanisms such as concatenation and pooling [6]. In recent years, a lot of effort has been devoted to studying inter-modality relation by exploring the connection between visual and textual semantics [4], [30], [41], [57], [64]. Stacked attention network (SAN) was proposed in [64] to progressively search for related image regions using question semantic representations. Based on low-rank bilinear

pooling, bilinear attention network (BAN) [30] was proposed to generate bilinear attention maps to fuse multimodal features, which is also employed in this work. UpDn [4] utilized Faster R-CNN [50] to extract regions of interest (ROI) at object level and aggregated region features with weights generated under the guidance of the question. Based on ROI features, some methods such as LXMERT [58] and Pythia [26] achieved promising results. Very recently, Transformer [60] based vision-and-language pre-training (VLP) becomes a popular paradigm. A typical process is to extract question features with BERT [12] and fuse them with visual features via self-attention mechanism. According to the different ways of visual feature extraction, VLP can be divided into object-based methods such as ViLBERT [40], VL-BERT [57], and VisualBERT [37], convolution-based methods such as Pixel-BERT [22], and image-patch-based methods such as ViLT [31]. Since object-based methods rely on visual object labels, which are not available in existing Med-VQA datasets, in this paper, we only explore convolution-based and image-patch-based methods.

Besides, a recent line of research [5], [13], [23], [48] focused on developing VQA systems with higher-level reasoning skills. [5] proposed to split questions into a series of semantic segments, which would accordingly activate pre-specified neural network modules. However, it is difficult to train the network due to complex pre-defined structures and annotations. [13] focused on solving VQA tasks in the few-shot setting by generating additional normalization parameters from questions to control the visual feature extractor's inner layers. To solve the highly challenging compositional questions [24], [27], [65], MAC [23] proposed to use various recurrent cells such as memory, attention, and composition for reasoning, while neural-symbolic (NS) approaches [42], [66] exploited executable symbolic programs to mimic human reasoning process.

B. Medical Visual Question Answering

Existing studies mainly applied popular VQA models on Med-VQA tasks [1], [2], [25], [53], [61], [63], [71]. Specifically, the visual features of medical images are extracted by deep pre-trained networks (e.g., ResNet [19] or VGGNet [55]), and the textual features of clinical questions are obtained through stacked RNN-based layers. For multimodal feature fusion, [71] adopted a simple concatenation operation, [1] exploited stacked attention networks (SAN) [64] and compact bilinear pooling (MCB) [15], [63] and [53] employed multimodal factorized bilinear pooling (MFB) [67], and [61] proposed to query an image by means of a written question based on the multimodal low-rank bilinear (MLB) module.

However, because of the large difference between radiology images and general images and the small scale of training datasets for Med-VQA, such straightforward adaptation suffers from severe overfitting. Moreover, state-of-the-art VQA methods such as UpDn [4], which leverage algorithms like Fast R-CNN [50] for object detection, cannot be applied to Med-VQA due to lack of annotation of existing datasets. To conquer the difficulty of data scarcity, [44] proposed mixture of enhanced visual features (MEVF) that pre-trained a small visual

feature extractor (several convolutional layers) on VQA-RAD dataset [36] and an undisclosed external medical datasets, by using auto-encoder and meta-learning method MAML [14]. While this work achieves good performance on VQA-RAD dataset, the pre-training approach is dataset-dependent and requires extra annotation effort. Besides, it simply employs a bilinear attention mechanism for multimodal feature fusion, which lacks multi-level reasoning ability. Some recent works including MMBERT [29] and MedViLL [43] try to pre-train a multimodal Transformer on medical vision-language datasets and then fine-tune it on Med-VQA tasks. However, due to the small scale of existing Med-VQA training datasets, large models could easily overfit. In this paper, we explore lightweight models like MEVF [44] and aim to improve both the reasoning module and the visual feature extraction module.

C. Contrastive Self-supervised Learning

Lately, there is a growing interest in learning data representations with deep neural networks in an unsupervised or self-supervised manner, to reduce the need for laborious annotation works. Several recent studies have shown promise in learning image representations by designing proper pretext tasks and loss functions. [16] proposed to randomly rotate an image by 0, 90, 180, or 270 degrees and train a neural network to predict the rotation angle. CPC [45] pioneered in using a contrastive objective (InfoNCE loss) to learn data representations with a context auto-encoding task and achieved promising results in various domains including speech, image, and text. Recent development of contrastive self-supervised learning includes MoCo [18] that utilizes a queue to efficiently store a large number of negative samples to remove the restriction of mini-batch size, SimCLR [8] that uses stronger data augmentation and a very large batch to accommodate more negative samples, and MoCo-v2 [9] that combines the design improvements of SimCLR with MoCo. In the medical domain, contrastive self-supervised pre-training methods have also gained much attention recently [33], [72]. In this paper, we utilize MoCo-v2 to pre-train useful representations of radiology images for Med-VQA.

III. METHODOLOGY

In this section, we present our method for training a Med-VQA model, which consists of two stages as shown in Figure 2. In stage I, we propose to learn prior knowledge of radiology images for the visual module tailored for Med-VQA from a collected dataset of publicly available unlabeled radiology images by contrastive self-supervised learning [9]. In stage II, we propose a conditional reasoning mechanism with a question-conditioned reasoning component and a type-conditioned reasoning strategy to adaptively learn effective reasoning skills for different Med-VQA tasks.

A. The paradigm of Med-VQA

The goal of Med-VQA is to automatically answer a clinical question about a radiology image. By convention, it is formulated as a single-label classification task where there are C

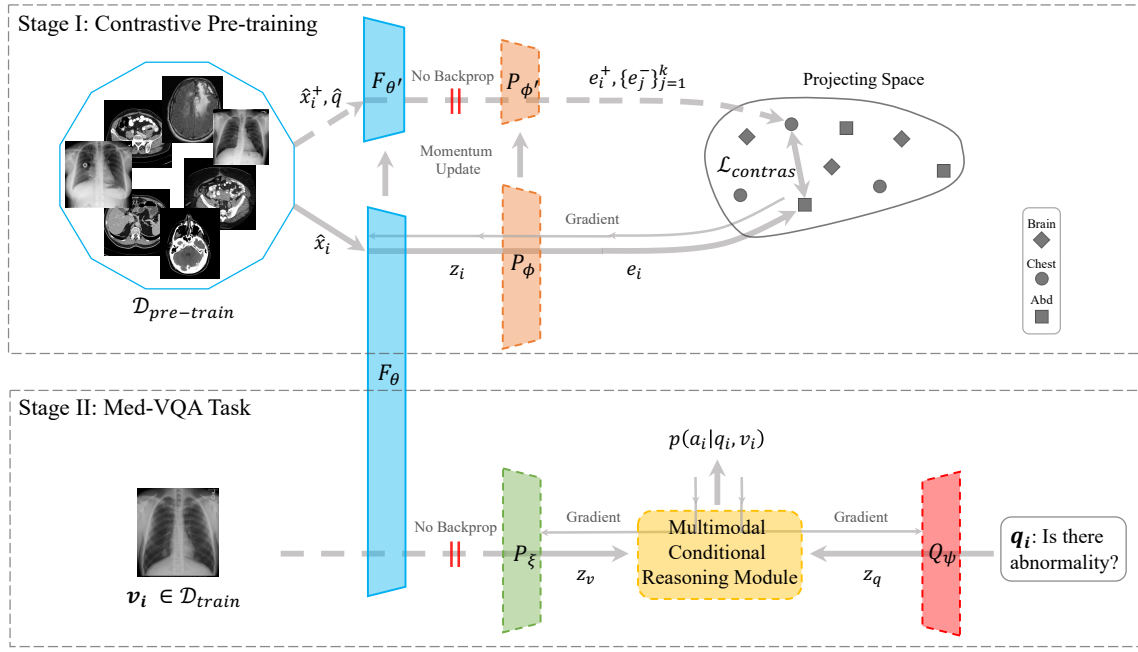


Fig. 2. Our proposed method for training a Med-VQA model. In stage I, we pre-train a visual feature extractor for Med-VQA by contrastive self-supervised learning. In stage II, we solve Med-VQA tasks by introducing a conditional reasoning mechanism.

candidate answers and each answer is a label. Unlike general VQA [4] where a question may have multiple answers, there is only one correct answer for each question in Med-VQA. Denote by $\mathcal{D}_{med-vqa} = \{(v_i, q_i, a_i)\}_{i=1}^N$ the training dataset for a Med-VQA model, where N is the number of training examples, and v , q , and a denote the image, question, and answer of a task respectively. A typical Med-VQA model aims to learn a function f that maps each (v_i, q_i) pair to a score vector $s \in \mathbb{R}^C$ where the j -th element s_i^j is the score for the j -th answer. The probability for the j -th answer is obtained by the softmax function, i.e., $p(s_i^j) = \frac{e^{s_i^j}}{\sum_{j=1}^C e^{s_i^j}}$, $1 \leq j \leq C$. The function f is usually instantiated as a neural network with parameters δ , and optimized by minimizing the cross-entropy loss:

$$\mathcal{L}_{vqa} = -\frac{1}{N} \sum_{i=1}^N a_i \log p(f_\delta(v_i, q_i)). \quad (1)$$

The function f usually consists of an image feature extractor, a question feature extractor, an attention-based feature fusion module, and an answer classifier, which are trained together in an end-to-end manner. In this paper, we focus on designing the visual feature extractor and the feature fusion module. We use long short-term memory network (LSTM) and multi-layer perceptron (MLP) as default choices for the question feature extractor and the answer classifier respectively.

B. Contrastive Pre-training (CP)

Due to the characteristic of radiology images, it is not effective to directly apply deep models (e.g., ResNet [19]) pre-trained on general datasets such as ImageNet to extract visual features of radiology images. Moreover, due to the small

scale of existing Med-VQA datasets (only a few hundred of radiology images available) [36], [39], fine-tuning pre-trained large models on them may lead to severe overfitting [44]. As such, to address the vast diversity of radiology images in terms of different organs and imaging modalities, we propose to pre-train a visual feature extractor by contrastive self-supervised learning on unannotated radiology images. Specifically, we collect a large set of radiology images of different organs and in different modalities, e.g., brain CT, brain MRI, chest X-Ray, and abdomen CT, and train a deep neural network that can pull together similar images and push away the dissimilar ones. Further, to avoid overfitting, after pre-training, we freeze the parameters of the large model and train an additional small network on Med-VQA tasks. We discuss the impact of pre-training datasets and strategies to avoid overfitting in Section IV-D.1.

Particularly, denote by $\mathcal{D}_{pre-train}$ the set of unlabelled radiology images collected for pre-training and \mathcal{D}_{train} the training set of the Med-VQA dataset respectively. As shown in Figure 2 stage I, we randomly sample a radiology image x_i and a queue $q = \{x_j^-\}_{j=1}^K$ of K images disjoint with x_i from $\mathcal{D}_{pre-train}$. Then, a set of data augmentation operations, denoted as Aug , which includes random crop, color distortion, resize with random flip, and Gaussian blur, is applied to all images:

$$\hat{x}_i = Aug(x_i), \hat{x}_i^+ = Aug(x_i), \hat{q} = \{\hat{x}_j^- = Aug(x_j^-)\}_{j=1}^K, \quad (2)$$

where \hat{x}_i and \hat{x}_i^+ are generated by applying Aug on x_i twice and considered as two different views of x_i . A feature extractor (usually a convolutional neural network such as ResNet) is used to obtain the feature representation of the anchor point \hat{x}_i , i.e., $z_i = F_\theta(\hat{x}_i) : \mathcal{X} \rightarrow \mathcal{F}$, where \mathcal{X} and \mathcal{F} are the

input image space and the feature space respectively. Further, a non-linear layer projects the feature representation into the projection space \mathcal{P} , i.e., $e_i = P_\phi(z_i) : \mathcal{F} \rightarrow \mathcal{P}$. Similarly, another two networks $F_{\theta'}$ and $P_{\phi'}$ that share the same structure as F_θ and P_ϕ respectively are used to map \hat{x}_i^+ and \hat{q} to obtain the feature representations $\{z_i^+, z_1^-, z_2^-, \dots, z_K^-\}$ and the projections $\{e_i^+, e_1^-, e_2^-, \dots, e_K^-\}$ respectively. Since e_i and e_i^+ are the projections of different views of x_i , e_i should be similar to e_i^+ (*positive pair*), and dissimilar to $\{e_1^-, e_2^-, \dots, e_K^-\}$ (*negative pairs*).

Following SimCLR [8], we conduct contrastive learning in the projection space using the InfoNCE contrastive loss [45] with dot product similarity:

$$\mathcal{L}_{e_i, e_i^+, \{e_j^-\}} = -\log \frac{\exp(e_i \cdot e_i^+ / \tau)}{\exp(e_i \cdot e_i^+ / \tau) + \sum_{j=1}^K \exp(e_i \cdot e_j^- / \tau)}, \quad (3)$$

where τ is a temperature hyper-parameter [62].

Since the length K of the queue q is much larger than the training mini-batch size, it is costly to update $F_{\theta'}$ and $P_{\phi'}$ by gradient back-propagation. Following MoCo-v2 [9], we update them in a momentum-based way:

$$\theta' \leftarrow m\theta' + (1 - m)\theta, \quad (4)$$

$$\phi' \leftarrow m\phi' + (1 - m)\phi, \quad (5)$$

where m is a momentum coefficient close to 1.

After pre-training, in stage II, we can apply F_θ to extract the visual features of radiology images from \mathcal{D}_{train} . However, we observe in experiments that directly fine-tuning the pre-trained large model easily leads to overfitting on the small-scale training set of Med-VQA datasets. Hence, we propose to keep θ fixed and append a non-linear layer P_ξ after F_θ for adaptation on \mathcal{D}_{train} . We set the input and output dimensions of P_ξ to be the same. As such, the visual features are obtained by: $z_v = P_\xi(F_\theta(v_i))$. Notice that unlike MEVF [44] that designs an auxiliary classification task and requires additional annotation effort for pre-training, our method leverages large amounts of unlabeled images to achieve better generalization.

C. Conditional Reasoning (CR)

Besides improving the feature extraction ability of the Med-VQA model, another key issue is to improve its reasoning ability. Here, we propose a conditional reasoning mechanism, aiming to solve different Med-VQA tasks with task-adaptive reasoning skills, as illustrated in Figure 3. It includes a question-conditioned reasoning module and a type-conditioned reasoning module, building on top of a basic multimodal reasoning module (the multimodal feature fusion module indicated by the yellow block in Figure 3). We first review the multimodal reasoning module and then elaborate on our proposed reasoning modules.

1) Multimodal Reasoning: In this paper, we utilize bilinear attention networks (BAN) [30], a popular model used in general VQA, for multimodal feature fusion and reasoning. Given the extracted visual features $\mathbf{Z}_v \in \mathbb{R}^{d_v \times N}$ (d_v is the dimension of image features and N is the number of channels), textual features $\mathbf{Z}_q \in \mathbb{R}^{d_q \times L}$ (d_q is the dimension of word features and L is the number of words in the question), and the number of reasoning steps – glimpse G , BAN models multimodal feature interaction in the i -th reasoning step via:

$$\mathbf{f}_i = (\mathbf{Z}_v^T \mathbf{W}_v)_j^T \mathbf{M}_i (\mathbf{Z}_q^T \mathbf{W}_q)_j, \quad (6)$$

$$\mathbf{M}_i = \text{softmax}(((\mathbf{1} \cdot \mathbf{p}_i^T) \circ \mathbf{Z}_v^T \mathbf{W}_v) \mathbf{W}_q^T \mathbf{Z}_q), \quad (7)$$

where $\mathbf{f}_i \in \mathbb{R}^J$ with $J \leq \min(d_v, d_q)$, $i \in \{1, \dots, G\}$ is the index of reasoning step, $j \in \{1, \dots, J\}$ is the index of matrix column, $\mathbf{W}_v \in \mathbb{R}^{d_v \times J}$ and $\mathbf{W}_q \in \mathbb{R}^{d_q \times J}$ are trainable weights, $\mathbf{1} \in \mathbb{R}^N$ is an all-one vector, $\mathbf{p}_i \in \mathbb{R}^J$ is a learnable vector, and \circ is the element-wise product. We follow MEVF to set the number of channels of visual features to $N = 1$.

After G reasoning steps, the final fused features \mathbf{f} are obtained by:

$$\mathbf{f} = \text{SumPool}(\sum_{i=1}^G ((\mathbf{W}_i \mathbf{f}_i) \cdot \mathbf{1}^T + \mathbf{f}_{i-1})), \quad (8)$$

where $\mathbf{f} \in \mathbb{R}^{d_q}$, $\mathbf{W}_i \in \mathbb{R}^{d_q \times J}$, $\mathbf{1} \in \mathbb{R}^L$, $\mathbf{f}_0 = \mathbf{Z}_q$, and SumPool is the sum operation over the length dimension L . We discuss the impact of the hyper-parameter G in Section IV-D.2.

2) Question-Conditioned Reasoning (QCR): Recent studies [36], [44] have shown that BAN has limited reasoning ability for Med-VQA, especially for open-ended questions. This is because it can not fully capture the interaction between visual and textual features. For example, BAN merely utilizes bilinear matrix multiplication to fuse multimodal features. To equip the Med-VQA model with more powerful reasoning ability, we improve the standard reasoning module by incorporating an additional question-conditioned modulation function. Our motivations are two-fold. First, similar to human reasoning processes, solving different tasks requires corresponding task-specific reasoning skills. Second, the question itself contains rich task information which could be helpful [28].

Hence, the QCR function is designed to extract task information from the question and use it to guide the modulation over multimodal features. In this process, high-level reasoning skills are learned by imposing importance selection over the fusion features.

The details of QCR are illustrated on the right side of Figure 3 within the orange dashed rectangle. First, a question string q , with L words in it, is converted into a sequence of word embeddings pre-trained by Glove [47]. Let $\mathbf{w}_i \in \mathbb{R}^{d_w}$ denote the corresponding word vector for the i -th word:

$$\mathbf{Q}_{emb} = \text{WordEmbedding}(q) = [\mathbf{w}_1, \dots, \mathbf{w}_L]. \quad (9)$$

The word embedding sequence $\mathbf{Q}_{emb} \in \mathbb{R}^{d_w \times L}$ is further processed by a d_g -dimensional Gated Recurrent Unit (GRU) to obtain the question embedding:

$$\mathbf{Q}_{feat} = \text{GRU}([\mathbf{w}_1, \dots, \mathbf{w}_L]) = [\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_L], \quad (10)$$

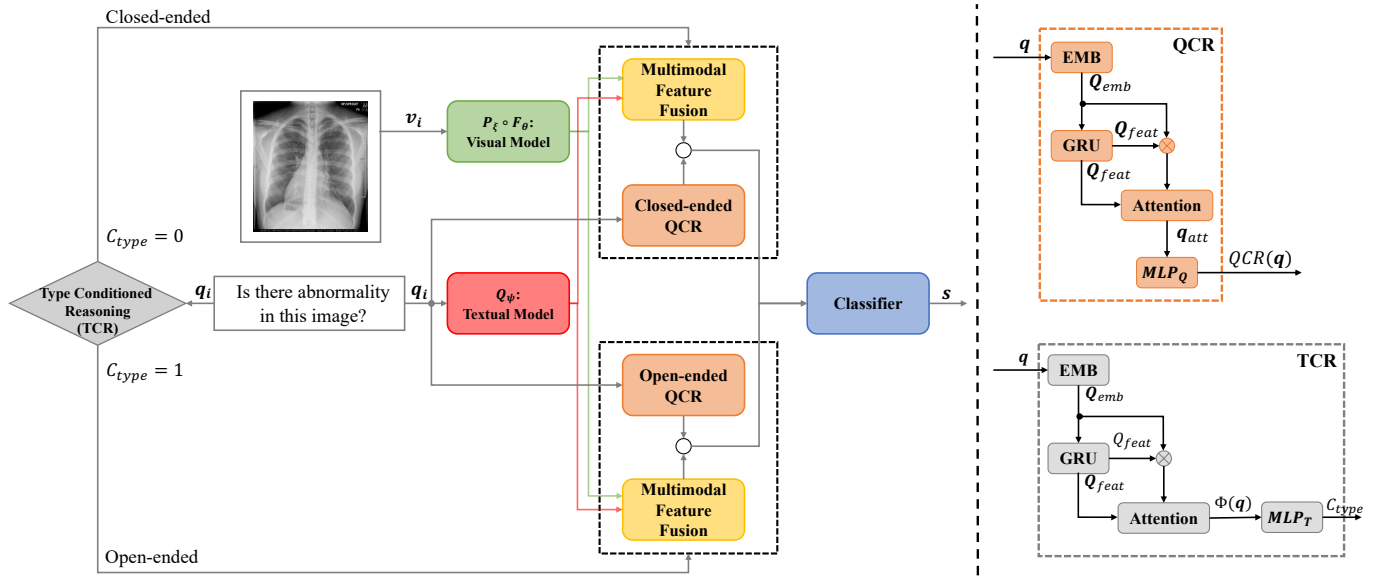


Fig. 3. Our proposed Med-VQA model with conditional reasoning. To prevent overfitting, we freeze the visual model F_θ (pre-trained in stage I) and append a non-linear layer P_ξ for fine-tuning on medical images. First, the TCR module classifies the question as open-ended or closed-ended and chooses the corresponding branch for reasoning. Then, the question features extracted by the textual model Q_ψ will be fused with the image features by the multimodal feature fusion module (e.g., BAN) and our QCR module. Finally, the answer is obtained by an MLP classifier.

where $Q_{feat} \in \mathbb{R}^{d_g \times L}$, and η_i denotes the embedding at the i -th position.

Since the question embedding Q_{feat} is generated by the GRU network word-by-word sequentially, it may put more emphasis on later words. To further highlight the important words, e.g., “where do nodes locate in the lung?”, we design an attention mechanism to re-calculate attention weights on different words:

$$\tilde{Q} = Q_{emb} \otimes Q_{feat}, \quad (11a)$$

$$Y = \tanh(W_1 \tilde{Q}), \quad (11b)$$

$$\tilde{Y} = \sigma(W_2 \tilde{Q}), \quad (11c)$$

$$\mathcal{G} = Y \circ \tilde{Y}. \quad (11d)$$

Here, \otimes denotes feature concatenation in the feature dimension, $\tilde{Q} \in \mathbb{R}^{(d_w+d_g) \times L}$, $W_1, W_2 \in \mathbb{R}^{d_g \times (d_w+d_g)}$ are trainable weights, σ and \tanh are the sigmoid activation function and tanh activation function respectively, and \circ is the Hadamard product. \tilde{Q} can be formed by both context-free embeddings (e.g., Glove) and contextual embeddings (e.g., GRU), which has been demonstrated effective in many NLP tasks [35], [52]. σ and \tanh (Equations 11b – 11c) make up the gated hyperbolic tangent activation [4], [11], which is a special case of highway networks [56] that outperforms traditional ReLU or tanh layers in many scenarios. \tilde{Y} acts as a gate on the intermediate activation Y to control the output $\mathcal{G} \in \mathbb{R}^{d_g \times L}$ [56].

Then, the attention vector $\alpha \in \mathbb{R}^L$ for the question embedding Q_{feat} can be obtained by

$$\alpha = \text{softmax}((W_a \mathcal{G})^T), \quad (12)$$

where $W_a \in \mathbb{R}^{1 \times d_g}$ are trainable weights.

Finally, with the attention vector α , we obtain the final output of QCR as:

$$q_{att} = Q_{feat} \alpha, \quad (13)$$

$$QCR(q) = MLP_Q(q_{att}), \quad (14)$$

where q_{att} is the aggregated question representation, and MLP_Q is a multilayer perceptron network that provides additional non-linear transformation for importance selection.

In this paper, we propose to impose the proposed QCR module on the multimodal feature fusion module A_{δ_m} by an element-wise multiplication between their outputs: $QCR(q)$ and $A_{\delta_m}(Z_v, Z_q)$. The final representations are then fed to the classifier D_{δ_c} , and the prediction scores are given by

$$s = D_{\delta_c}(A_{\delta_m}(Z_v, Z_q) \circ QCR(q)), \quad (15)$$

where \circ denotes element-wise product.

3) Type-Conditioned Reasoning (TCR): It has been observed that closed-ended questions are generally easier than open-ended questions. For example, the closed-ended question “Is this an MRI image?” can be correctly answered with a simple image understanding process, but the open-ended question “What is the abnormality of the patient’s right brain in this radiology image?” needs multi-step reasoning, since the model must locate the abnormality in the right brain first and then diagnose the type of abnormality, e.g., brain tumor. Thus, Med-VQA systems need to be empowered with multi-level reasoning abilities, which are lacking in present VQA models [54].

To this end, we propose to use a separate reasoning component for closed-ended questions and open-ended questions respectively, in which the proposed QCR module is applied on top of the multimodal feature fusion module, as shown in the two black dash rectangles on the left of Figure 3. Particularly,

we want to train a task type classifier C_{type} that takes a question as input and outputs the question type, i.e., closed-ended or open-ended. We observe that different types of questions put emphasis on different words. For example, closed-ended questions usually start with “Do\Are\Is\etc.”, and open-ended questions often start with “What\How many\Where\etc”. The differences between the two types of questions can be captured by question embeddings, which makes it possible to train a reliable and light-weight classifier that divides Med-VQA tasks into two subbranches, as shown by the rhombus module in Figure 3.

Similar to Section III-C.2, we use Equations (9) - (13) to compute the question embedding and denote the mapping as Φ . We then employ a multilayer perceptron MLP_T to map question embedding into classification scores. The binary classification probabilities are computed by $\mathbf{p}^t = \text{softmax}(MLP_T(\Phi(q)))$, and p_0^t and p_1^t are the probabilities for closed-ended and open-ended respectively. The binary question type classifier C_{type} is then formulated as:

$$C_{type}(q) = \begin{cases} 0, & \text{if } p_0^t > p_1^t, \\ 1, & \text{else.} \end{cases} \quad (16)$$

Hence, the predicted scores \mathbf{s} of candidate answers for a task (v, q) can be obtained by

$$\mathbf{s} = \begin{cases} D_{\delta_c}(A_{\delta_m}^{cl}(\mathbf{Z}_v, \mathbf{Z}_q) \circ QCR^{cl}(q)), & \text{if } C_{type}(q) = 0, \\ D_{\delta_c}(A_{\delta_m}^{op}(\mathbf{Z}_v, \mathbf{Z}_q) \circ QCR^{op}(q)), & \text{if } C_{type}(q) = 1, \end{cases} \quad (17)$$

where *cl* and *op* stand for closed-ended and open-ended respectively. For the basic multimodal reasoning module, we use a different number of reasoning steps for the open-ended branch $A_{\delta_m}^{op}$ and the closed-ended branch $A_{\delta_m}^{cl}$ and conduct an ablation study in Section IV-D.2.

D. Proposed Med-VQA Model

The pre-trained visual feature extractor and the conditional reasoning mechanism can be naturally combined to train an end-to-end Med-VQA model. As depicted in Figure 3, our proposed Med-VQA model works as follows. First, the image features \mathbf{Z}_v are obtained by the visual feature extractor $P_\xi \circ F_\theta$ (with F_θ pre-trained and fixed), as indicated by the green rectangle. The question features \mathbf{Z}_q are obtained by the textual feature extractor Q_ψ , as indicated by the red rectangle. Second, the TCR module classifies the question as open-ended or closed-ended and chooses the corresponding reasoning module (black dash rectangle). Note that the visual and textual feature extractors are shared for both branches. Third, the chosen reasoning module will generate fused features, which is the element-wise multiplication between the output of the basic reasoning module (e.g., BAN or SAN) (yellow block) and the modulation vector produced by the QCR module (orange block). Finally, an MLP classifier gives the prediction score \mathbf{s} for candidate answers.

IV. EXPERIMENTS

In this section, we conduct comprehensive experiments to evaluate the performance of our proposed framework on the

TABLE I
MED-VQA DATASET STATISTICS

Dataset	Images	Answers	Questions		
			Overall	Open	Closed
VQA-RAD [36]	315	458	3,515	1,420	2,095
SLAKE-EN [39]	642	219	7,033	4,252	2,781

only two available manually-annotated Med-VQA datasets, VQA-RAD [36] and SLAKE [39]. We compare our approach with current state-of-the-art baselines, evaluate the effectiveness of each component of our framework by ablation studies, and present qualitative results by visualizing the attention maps of both the images and questions of some Med-VQA tasks.

A. Datasets

VQA-RAD [36] and SLAKE [39] are the only two available manually-annotated radiology-based datasets for Med-VQA. The statistics of the two datasets are summarized in Table I.

VQA-RAD [36] contains 315 radiology images (e.g., CT, MRI, and X-Ray) and 3,515 clinical question-answer pairs (tasks), with 3,064 tasks for training and 451 tasks for testing. The number of candidate answers is 458. There may be multiple questions associated with one image. For example, the clinicians may ask different types of questions regarding a radiology image such as “imaging modality”, “abnormality”, or “organ system”.

SLAKE [39] is a newly released bi-lingual Med-VQA dataset. It includes more question types such as “organ shape” and “common fact”. In our experiments, we use the English version of SLAKE, referred to as SLAKE-EN, which contains 642 radiology images, 7,033 question-answer pairs, and 219 candidate answers. We follow the original dataset splitting, where 4,919 tasks about 450 images are used for training, 1,053 tasks about 96 images for validation, and 1,061 tasks about 96 images for testing.

In both datasets, there are closed-ended questions and open-ended questions, as shown in Figure 1.

B. Implementation Details

We conduct all experiments on a Ubuntu 16.04 server with 8 Titan XP GPUs using PyTorch. The implementation details for stages I & II of our method (Figure 2) are provided below.

Stage I. We collect 22,995 unlabeled radiology images from an online open-access resource¹ to form $\mathcal{D}_{pre-train}$, which contains 7,592 brain MRI and CT images, 7,592 abdomen CT images, and 7,811 chest X-Ray images. We use ResNet-50 as the backbone for F_θ and $F_{\theta'}$, and use a non-linear layer with ReLU activation for P_ϕ and $P_{\phi'}$ to project the representations into a 128-dimensional space. Then we train them with the loss $\mathcal{L}_{contras}$ (Equation (3)) for 800 epochs, which takes approximately 23 hours. In each epoch, the mini-batch size is 128, and the model is trained in parallel over 4 GPUs. The length K of the queue q is 4,096, the temperature

¹<http://medicaldecathlon.com/>

TABLE II

TEST ACCURACY OF OUR PROPOSED METHODS AND THE BASELINES ON VQA-RAD [36] AND SLAKE [39]. “Fw.” IS THE ABBREVIATION OF “FRAMEWORK”. * MEANS RESULTS CITED FROM THE ORIGINAL PAPERS.

Models	#Parameters (M)	Accuracy (%) on VQA-RAD [36]			Accuracy (%) on SLAKE-EN [39]		
		Overall	Open-ended	Closed-ended	Overall	Open-ended	Closed-ended
► <i>General VQA Frameworks:</i>							
MFBCoAtt Fw. [68]	58.20	50.6	14.5	74.3	73.3	72.2	75.0
SAN Fw. [36], [64]	36.54	54.3	31.3	69.5	76.0	74.0	79.1
MFH Fw. [69]	72.11	57.9	35.2	72.8	75.9	73.6	79.3
MCB Fw. [15], [36]	36.29	58.1	38.0	71.3	76.1	73.2	80.5
MUTAN Fw. [7]	58.46	58.1	34.1	73.9	76.8	73.6	81.7
BAN Fw. [30], [44]	42.19	58.3	37.4	72.1	76.3	74.6	79.1
► <i>Vision-and-Language Transformers:</i>							
Pixel-BERT-R50 [22]	137.37	61.7	48.2	70.5	77.4	77.1	77.9
ViLT-B/32 (w/o pre-training) [31]	113.12	59.6	38.5	73.5	76.0	75.8	76.2
ViLT-B/32 (w/ pre-training) [31]	113.12	66.5	52.0	76.1	78.1	76.9	80.0
MMBERT* [29]	111.53	72.0	63.1	77.9	-	-	-
MMBERT [29]	111.53	68.5	57.5	75.7	79.0	76.1	83.4
MedViLL* [43]	129.78	70.3	59.5	77.7	-	-	-
MedViLL [43]	129.78	69.6	58.7	76.8	78.4	76.3	81.7
► <i>Vision-Language Contrastive Pre-training:</i>							
GLoRIA+SAN [21]	135.01	67.4	56.4	74.6	76.8	75.2	79.3
GLoRIA+BAN [21]	139.15	69.2	57.5	76.8	79.4	78.1	81.3
► <i>Med-VQA Models:</i>							
MEVF+SAN* [44]	13.99	60.8	40.7	74.1	-	-	-
MEVF+SAN [44]	13.99	64.1	49.2	73.9	76.5	75.3	78.4
MEVF+BAN* [44]	19.64	62.6	43.9	75.1	-	-	-
MEVF+BAN [44]	19.64	66.1	49.2	77.2	78.6	77.8	79.8
CP+BAN (ours)	18.44	68.1	53.1	77.9	80.9	79.1	83.7
MEVF+BAN+CR (ours)	27.21	71.6	60.0	79.3	80.0	78.8	82.0
CP+BAN+CR (ours)	26.01	72.5	60.5	80.4	81.9	80.5	84.1

parameter τ is 0.2, and the momentum coefficient m is 0.999. We utilize SGD optimizer with an initial learning rate of $1.5e^{-2}$ decayed by cosine schedule. After training, we save the weights of ResNet-50 in the last epoch for training in stage II.

Stage II. We use F_θ (ResNet-50) pre-trained in stage I combined with P_ξ (a non-linear layer with ReLU activation) to extract visual features. For a fair comparison, we follow MEVF [44] to set the dimension of visual features to 128, use Glove [47] to initialize word embeddings, and employ a 1024-dimensional LSTM to extract textual features. Moreover, the hidden size of all GRUs in the QCR and TCR modules (Figure 3) is 1024. The MLP_Q in Equation (14) and MLP_T in Equation (16) are instantiated with hidden units 1024 and 64 respectively. For each dataset, we pre-train a task type classifier C_{type} (with about 2.4M parameters) for 150 epochs by using the “answer_type” label in the training set and Adam optimizer [32] with learning rate $1e^{-4}$, and freeze the pre-trained weights during both the training and inference stages. The trained classifiers reach 99.33% and 99.81% classification accuracy on the test set of VQA-RAD and SLAKE respectively. For the training of the Med-VQA model, we use Adamax optimizer with initial learning rate $2e^{-3}$ for 100 epochs. Notice that different from general VQA that formulates open-ended question answering as a multi-label

classification task (e.g., open-ended questions in the VQA v2.0 dataset normally have more than one correct answer [17]) or a text generative task [10], in Med-VQA each question has only one correct answer regardless of question type. Hence, we follow previous works [36], [44] to formulate Med-VQA as a single-label classification task and use accuracy as evaluation metric for both open-ended and closed-ended questions.

C. Comparison with the State-of-the-arts

We compare our method with existing Med-VQA models including general VQA frameworks, vision-and-language Transformers, vision-language contrastive pre-training, and the recently proposed MEVF method [44].

- **General VQA frameworks.** We follow [36] to compare with stacked attention network (SAN) [64] and multi-modal compact bilinear pooling (MCB) [15]. In addition, we also compare with other frameworks including bilinear attention network (BAN) [30], multi-modal factorized bilinear pooling with co-attention (MFBCoAtt) [68], multimodal factorized high-order pooling (MFH) [69], and multi-modal tucker fusion (MUTAN) [7]. These VQA frameworks are usually named after their respective reasoning modules.
- **Vision-and-language Transformers.** Transformer-based

vision-and-language pre-training (VLP) has achieved impressive results in vision-language tasks. In this paper, we compare our method with both general and medical VLP models. Due to the lack of regional object labels in existing Med-VQA datasets, object-based VLP methods cannot be applied. Hence, we compare with a convolution-based method – Pixel-BERT [22] and a patch-based method – general vision-and-language Transformer (ViLT) [31]. For medical VLP models, we compare with multimodal medical BERT (MMBERT) [29] and medical vision language learner (MedViLL) [43].

- **Vision-Language Contrastive Pre-training** is a self-supervised approach that pre-trains a model by pulling the paired image-text instances closer in the embedding space. The representative method in the medical domain is GLORIA [21], which uses ResNet-50 and BioClinicalBERT [3] as the visual encoder and text encoder respectively. Since there is no feature fusion module in this paradigm, we use BAN [30] and SAN [64] instead.
- **MEVF** [44] is a recently proposed lightweight model for Med-VQA, which pre-trains a visual module on medical datasets and combines it with different attention reasoning modules such as BAN [30] and SAN [64].

Table II shows the results of our methods and the baselines. For all general VQA frameworks, ResNet-50 pre-trained on ImageNet and 1024- D LSTM network are used as visual extractor and textual extractor, respectively. Note that these VQA frameworks are usually named after their respective reasoning modules. We re-implement Pixel-BERT [22] (the ResNet-50 version) since both the source code and pre-trained weights are not provided. We use the original implementation of ViLT [22] and fine-tune the model on Med-VQA datasets with or without the pre-trained weights. We re-implement MMBERT, MedViLL, and MEVF+BAN/SAN using the code released by the authors. We cannot reproduce the results of MMBERT as reported in the original paper using the configurations provided by the authors² (probably some key information is missing). Our re-implementation of MEVF+BAN/SAN achieves much better results than the original paper due to longer training epochs (100 epochs for all methods). For MEVF+BAN/SAN, the visual extractor is MEVF, and the reasoning module is BAN/SAN. To compare our proposed visual extractor CP (Section III-B) with MEVF, we combine it with BAN (denoted as CP+BAN). For a fair comparison, we strictly follow MEVF+BAN to use a 1024- D LSTM network to extract textual features with word embeddings pre-trained by GloVe [47], set the dimension of visual features to 128, and set the number of reasoning steps of BAN to 2. When applying the proposed CR mechanism to boost MEVF+BAN and CP+BAN, we set the number of reasoning steps of BAN in open-ended and closed-ended branches to 2 and 1, respectively.

According to the performance of each model, we can make the following observations:

- Our method (CP+BAN+CR) performs much better than general VQA frameworks, vision-and-language Trans-

²<https://github.com/VirajBagal/MMBERT/issues/4>

TABLE III

COMPARISON OF DIFFERENT VISUAL MODULES ON VQA-RAD [36]. † INDICATES PRE-TRAINING ON IMAGENET WITH STANDARD SUPERVISED CLASSIFICATION. ‡ INDICATES PRE-TRAINING ON IMAGENET WITH CONTRASTIVE SELF-SUPERVISED LEARNING (MoCo-v2 [9]).

Visual Modules	#Parameters (M)	Accuracy (%)		
		Overall	Open	Closed
VGG-16† [55]	134.83	56.8	35.2	71.0
ResNet-50† [19]	23.77	58.3	37.4	72.1
ResNet-50‡ [9]	23.77	59.9	34.6	76.5
MEVF [44]	1.22	66.1	49.2	77.2
CP (F_{θ})	23.77	62.3	38.5	77.9
CP ($P_{\xi} \circ F_{\theta}$)	23.79	61.2	38.5	76.1
CP ($P_{\xi} \circ F_{\theta}$, w/ F_{θ} frozen)	0.02	68.1	53.1	77.9

formers, and vision-language contrastive pre-training, with much fewer parameters, showing the benefit of utilizing small models for solving Med-VQA tasks. Also, it can be seen that large models tend to overfit the small-scale Med-VQA training data.

- Our pre-trained feature extractor (indicated by CP) is more effective than MEVF. CP+BAN achieves **2%** and **2.3%** absolute improvement over MEVF+BAN [44] in overall accuracy on VQA-RAD and SLAKE-EN respectively. Also, it can be noted that since MEVF is pre-trained on VQA-RAD, combining it with BAN (i.e., MEVF+BAN) only brings $\sim 2.3\%$ improvement over BAN on SLAKE-EN, much lower than the $\sim 4.6\%$ improvement brought by our CP model. Besides, CP+BAN+CR improves over MEVF+BAN+CR by $\sim 2\%$ in absolute overall accuracy on SLAKE-EN dataset, which is comparable to the improvement of CP+BAN over MEVF+BAN. While the improvement on VQA-RAD dataset is smaller, CP+BAN+CR still consistently achieves better performance than MEVF+BAN+CR for each question type. These results demonstrate the effectiveness and generality of our pre-trained model CP.
- Our conditional reasoning (CR) mechanism can further bring consistent and significant performance gains on top of different visual feature extractors, including MEVF and our CP model. It can be seen that CP+BAN+CR significantly outperforms CP+BAN while MEVF+BAN+CR significantly outperforms MEVF+BAN, on both datasets. The best performance is achieved by our CP+BAN+CR method. Remarkably, for open-ended Med-VQA tasks, incorporating CR leads to very large performance gains on VQA-RAD.

D. Ablation Study and Analysis

In this subsection, we conduct experiments to analyze the effectiveness of our proposed contrastive pre-training (CP) and conditional reasoning (CR) modules. We report results in test accuracy on VQA-RAD [36].

1) Ablation study on contrastive pre-training:

I. Comparison of different visual feature extractors. Table III shows the comparison of different visual feature

TABLE IV
COMPARISON OF PERFORMANCE ON VQA-RAD [36] BY USING DIFFERENT DATASETS FOR CONTRASTIVE PRE-TRAINING (CP IN SECTION III-B).

Dataset	Accuracy (%)
	Overall
Random initialization	60.3
ImageNet (22,995)	65.2
Brain (22,995)	61.4
Chest (22,995)	64.3
Abdomen (22,995)	63.0
Brain (11,500), Chest (11,495)	66.3
Chest (11,500), Abdomen (11,495)	67.2
Abdomen (11,500), Brain (11,495)	64.5
Brain (7,592), Chest (7,811), Abdomen (7,592) (ours)	68.1

extractors for Med-VQA. For all the methods, we use BAN with 2 reasoning steps for multimodal feature fusion and a 1024- D LSTM network as the textual extractor, and set the dimension of extracted visual features to 128. For P_ξ , we keep the input dimension same as the output dimension. It can be seen that our CP module (last row) outperforms MEVF and surpasses general visual backbones VGG-16 and ResNet-50 by large margins (8 ~ 11%), demonstrating its effectiveness.

In addition, the following observations can be made. First, compared with ResNet-50 \ddagger pre-trained on ImageNet also using MoCo-v2³ (row 3), it can be seen that ResNet-50 pre-trained on medical images, i.e., our pre-trained F_θ (row 5) performs better, showing the benefit of domain-specific pre-training. Second, appending a non-linear layer P_ξ to F_θ (row 6) leads to worse performance than F_θ , indicating more overfitting. Finally, freezing the parameters of F_θ and only fine-tuning P_ξ (row 7) leads to significantly better performance. This effective and efficient design helps to avoid overfitting caused by fine-tuning large models on small-scale Med-VQA datasets and greatly reduces the training parameters.

II. Comparison of different pre-training datasets. We observe that existing Med-VQA datasets contain radiology images of different body regions and imaging modalities, e.g., brain MRI, chest X-Ray, and abdomen CT, as shown in Figure 7. Therefore, we collect a dataset $\mathcal{D}_{\text{pre-train}}$ (Section III-B) with a similar composition for contrastive pre-training. Specifically, it contains brain CT and MRI images, chest X-ray images, and abdomen CT images. Since the brain images do not have modality labels, we conduct an ablation study of $\mathcal{D}_{\text{pre-train}}$ w.r.t. different body parts and compare it with other pre-training datasets. For a fair comparison, we fix the size of all datasets as 22, 995, and use the same visual module $P_\xi \circ F_\theta$ (with F_θ frozen), textual model 1024- D LSTM, and reasoning module BAN (with 2 reasoning steps).

The results in Table IV show the importance of pre-training with images of different body parts. Specifically, for the

³The pre-trained weights can be downloaded from <https://github.com/facebookresearch/moco>. We choose the 800-epoch version, which uses the same number of epochs as in our pre-training.

TABLE V
ABLATION STUDY OF OUR PROPOSED CONDITIONAL REASONING (CR) MECHANISM ON VQA-RAD [36].

Base Model	QCR	TCR	#Parameters (M)	Accuracy (%)		
				Overall	Open-ended	Closed-ended
MEVF+BAN	✓	✓	19.64	66.1	49.2	77.2
			23.98	67.8	51.4	78.7
			22.87	70.1	56.7	79.0
	✓	✓	27.21	71.6	60.0	79.3
CP+BAN	✓	✓	18.44	68.1	53.1	77.9
			22.78	69.6	56.4	78.5
			21.67	71.4	58.9	79.7
			26.01	72.5	60.5	80.4

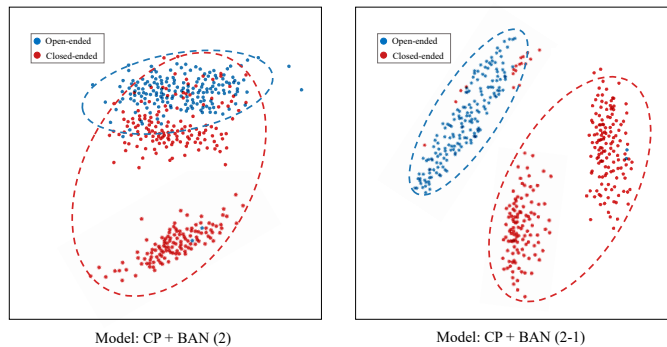


Fig. 4. t-SNE [59] visualization of the multimodal features (input to the classifier layer) of Med-VQA tasks in the test set of VQA-RAD [36] learned by CP+BAN(2) and CP+BAN(2-1) respectively. The TCR module in CP+BAN(2-1) disentangles the representations of open-ended and closed-ended tasks.

datasets containing images of only a single body part (rows 3 ~ 5), the pre-trained models perform worse than the model pre-trained on ImageNet (row 2), though they are better than random initialization (row 1). By increasing the diversity of the pre-training datasets (rows 6 ~ 8), the performance of the pre-trained models is significantly improved. The best performance is achieved with $\mathcal{D}_{\text{pre-train}}$ which contains images of three different body parts (row 9).

2) Ablation study on conditional reasoning:

I. Effect of QCR and TCR. We first evaluate the effectiveness of our proposed conditional reasoning modules QCR and TCR with two base models, MEVF+BAN and our CP+BAN. For both of them, the number of reasoning steps of BAN is set to 2. When incorporating TCR into the base models, we set the number of reasoning steps of BAN as 2 for open-ended questions and 1 for closed-ended questions. When only QCR is included (e.g., MEVF+BAN+QCR), the model does not differentiate the question type, and hence there is only one reasoning module. Our QCR module is then imposed on the basic reasoning module BAN to enhance reasoning ability. When only TCR is included (e.g., MEVF+BAN+TCR), the model differentiates the question type and chooses the corresponding reasoning module. However, only the basic reasoning module BAN is utilized for reasoning, without the QCR enhancement.

As shown in Table V, QCR improves the overall accuracy of MEVF+BAN and CP+BAN by 1.7% and 1.5% respectively,

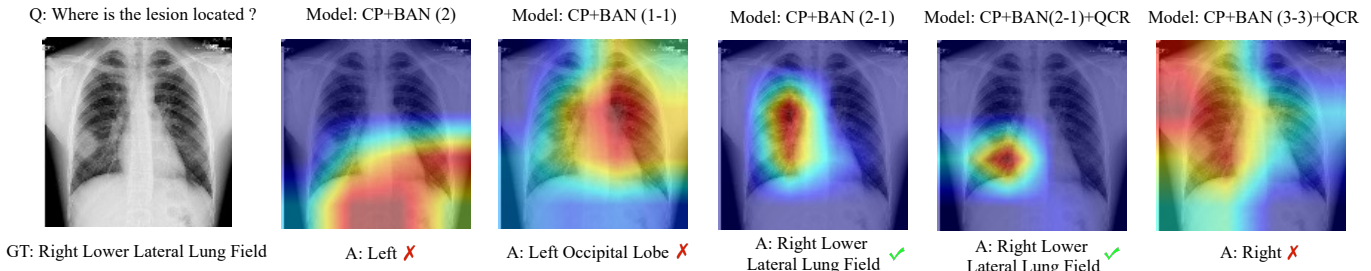


Fig. 5. Visual comparison of the prediction results for an open-ended task in VQA-RAD dataset by variants of CP+BAN. The Grad-CAM maps [51] of the visual model are plotted based on the predicted answers. ✓ and ✗ indicate the correctness of the answer given by each model.

TABLE VI

EFFECT OF THE NUMBER OF REASONING STEPS FOR OPEN-ENDED AND CLOSED-ENDED QUESTIONS IN VQA-RAD [36].

Base Model	G	#Parameters (M)	Accuracy (%)			
			Overall	Open	Closed	
MEVF+BAN	1	17.40	63.6	44.7	76.1	
	2	19.64	66.1	49.2	77.2	
	3	21.87	65.9	52.5	74.6	
	4	24.11	66.1	53.1	74.6	
	5	26.34	65.6	52.5	74.2	
	6	28.57	64.5	52.0	72.8	
CP+BAN	1	16.20	64.5	48.6	75.4	
	2	18.44	68.1	53.1	77.9	
	3	20.67	67.4	53.6	76.5	
	4	22.90	66.3	52.5	75.4	
	5	25.14	66.5	53.1	75.4	
	6	27.37	65.2	52.0	73.9	
► Incorporating TCR:						
CP+BAN	G_{open}	G_{closed}				
	1	1	19.43	68.7	53.6	78.7
	2	1	21.67	71.4	58.9	79.7
	1	2	21.67	70.3	55.3	80.1
	2	2	23.90	71.2	58.7	79.4
	3	2	26.14	69.8	57.5	77.9
	2	3	26.14	70.5	56.4	79.8
	3	3	28.38	69.2	55.9	77.9
	► Further Incorporating QCR:					
CP+BAN	G_{open}	G_{closed}				
	1	1	23.77	70.3	58.1	78.3
	2	1	26.01	72.5	60.5	80.4
	1	2	26.01	71.4	59.2	79.4
	2	2	28.25	71.6	61.4	78.3
	3	2	30.49	72.1	60.3	79.8
	2	3	30.49	71.1	59.2	79.0
	3	3	32.73	69.6	58.1	77.2

while TCR improves them by 4% and 3.3% respectively. The results show that both QCR and TCR are useful, indicating the importance of utilizing question information to learn task-adaptive reasoning skills for different Med-VQA tasks. Further, the large performance improvement brought by TCR demonstrates the necessity of learning multi-level reasoning skills for different types of Med-VQA tasks.

When both QCR and TCR are incorporated, we observe further improvements. The overall accuracy of MEVF+BAN and CP+BAN is increased by 5.5% and 4.4% respectively, where the accuracy for open-ended questions is increased by 10.8% and 7.4% respectively and the accuracy for closed-ended questions is increased by 2.1% and 2.5% respectively. The large improvements on open-ended tasks show the effectiveness of our conditional reasoning mechanism in learning

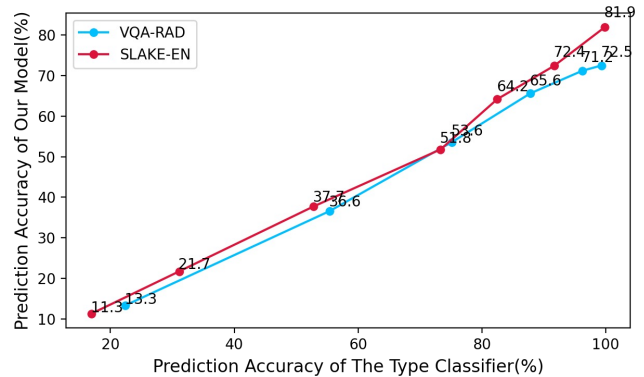


Fig. 6. Impact of the prediction accuracy of the type classifier on our model CP+BAN+CR. Note that the prediction accuracy of our model refers to the overall metric.

higher-level reasoning skills to solve difficult tasks.

II. The reasoning ability required for solving open-ended and closed-ended questions. We further study the reasoning ability required for solving open-ended and closed-ended questions, including the number of reasoning step G (Equation 8) and the proposed QCR module. We use $BAN(G)$ to denote BAN with G reasoning steps and $BAN(G_{open}-G_{closed})$ to denote that in our TCR we use BAN with G_{open} reasoning steps for open-ended questions and BAN with G_{closed} reasoning steps for closed-ended questions.

From the results in Table VI, we can make the following observations. (I) As the number of reasoning steps (G , G_{open} , or G_{closed}) increases, the corresponding model performance first increases and then decreases, indicating a gradual transition from underfitting to overfitting. (II) By incorporating TCR, even BAN(1-1) outperforms BAN with any G reasoning steps, showing the benefit of using a separate reasoning module for different types of questions. As shown in Figure 4, the TCR module disentangles the representations of open-ended and close-ended tasks. (III) The best performance is achieved by CP+BAN(2-1), indicating that solving open-ended questions requires stronger reasoning ability than closed-ended questions. (IV) Incorporating QCR can further improve the reasoning ability on both open-ended and closed-ended questions. These observations can also be reflected in the visual comparison in Figure 5, where CP+BAN(2-1)+QCR can more accurately find the relevant regions and give the correct answer.

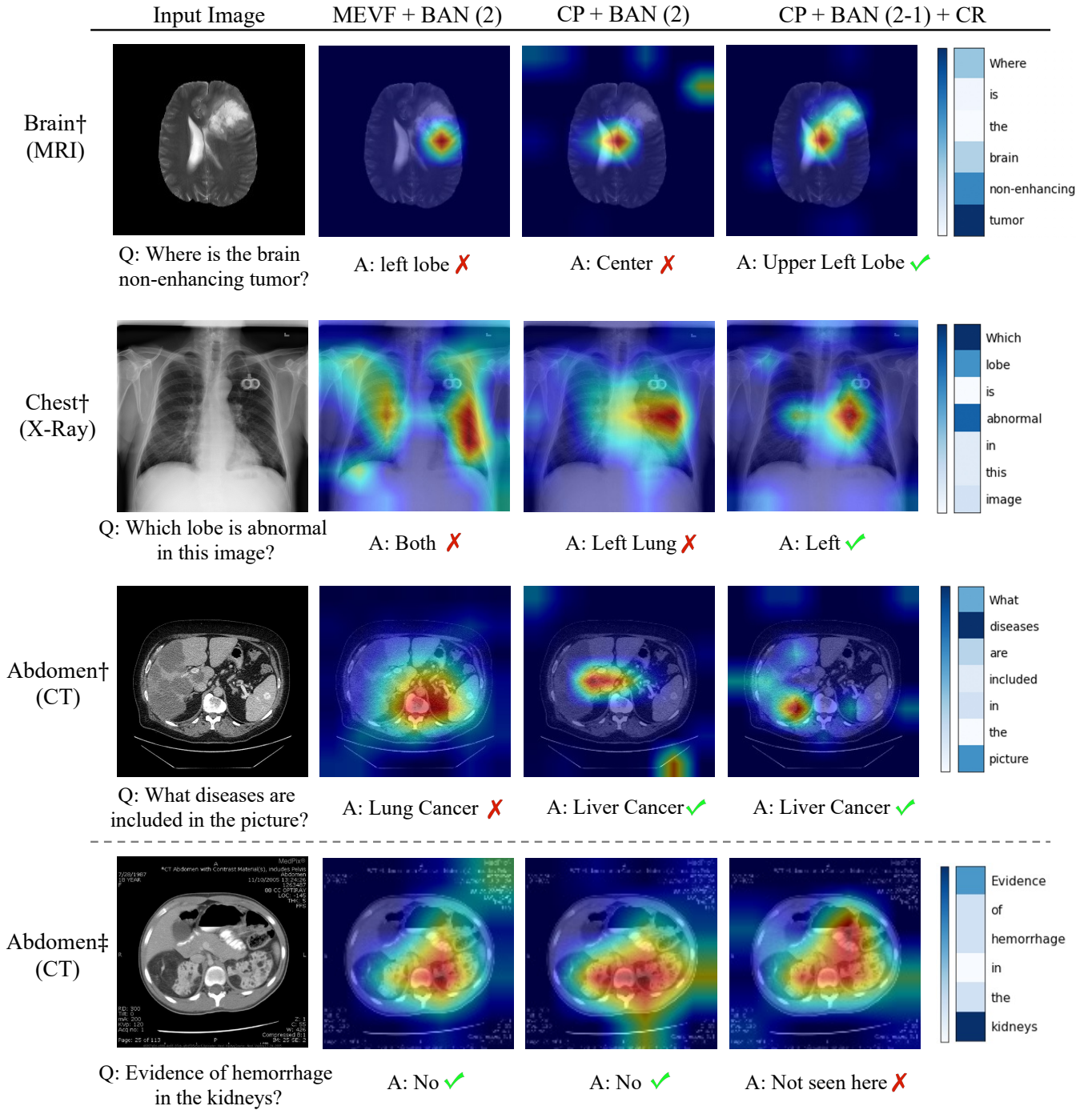


Fig. 7. The Grad-CAM maps of the visual modules of our methods and baseline MEVF+BAN. The attention map of our QCR module is displayed in the right column, and darker color indicates higher weight. ✓ and ✗ indicate the correctness of the answer given by each model. † indicates the test image comes from SLAKE [39], and ‡ indicates it comes from VQA-RAD [36]. The last row shows a failure case of our method with the conditional reasoning module, which is caused by the misclassification of question type.

III. Impact of the prediction accuracy of the type classifier. We analyze how the prediction accuracy of the type classifier C_{type} in the TCR module affects the prediction accuracy of our model CP+BAN+CR on both VQA-RAD and SLAKE-EN datasets. As shown in Figure 6, the more accurate C_{type} is, the better performance our model can achieve. An inaccurate C_{type} will result in extremely poor performance.

Fortunately, with our proposed algorithm (Equations (9) - (13) & Equation (16)), we can easily train a highly accurate C_{type} , achieving 99.33% and 99.81% classification accuracy on the test set of VQA-RAD and SLAKE respectively.

E. Qualitative Evaluation

We provide a qualitative comparison between our proposed methods and baseline MEVF+BAN. Figure 7 shows the Grad-

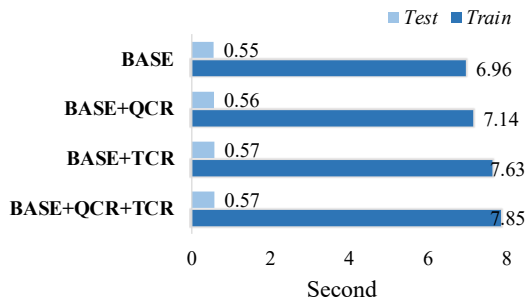


Fig. 8. Time efficiency of the proposed conditional reasoning mechanism (i.e., TCR and QCR modules). BASE represents the base model MEVF+BAN [44]. BASE+QCR does not differentiate the question type, and hence there is only one reasoning module. Our QCR module is imposed on the basic reasoning module BAN to enhance reasoning ability. BASE+TCR differentiates the question type and chooses different reasoning module correspondingly. However, only the basic reasoning module BAN is used for reasoning, without our QCR enhancement. ■ denotes training time (seconds) per epoch. ■ denotes test time (seconds) per epoch.

CAM [51] maps of the visual models based on the predicted answers of four tasks in SLAKE-EN and VQA-RAD, which cover different human body parts and imaging modalities. We also provide a visualization of the attention weights (Equation 12) in QCR.

The first task is about a brain MRI image. While MEVF+BAN and our CP+BAN both give wrong answers, our CP+BAN+CR can find the relevant regions and predict the right answer, demonstrating the effectiveness of conditional reasoning. The second task is about a chest X-Ray image. Compared with MEVF+BAN, our CP+BAN can better find more relevant regions, though it also gives a wrong answer. With conditional reasoning, our CP+BAN+CR can find the right answer. The third task is about an abdomen CT image. Both our CP+BAN and CP+BAN+CR give the right answer, but MEVF+BAN still fails. The last task is also about an abdomen CT image. In this case, both MEVF+BAN and our CP+BAN give the right answer, but our CP+BAN+CR fails. This is because the question type classifier of the TCR module gives a wrong prediction, mistaking a close-ended question as an open-ended one. Notice that even though the prediction accuracy of the question type classifier is as high as 99.33%, it may still fail in some rare cases.

F. Efficiency Evaluation

We progressively evaluate the time efficiency of the proposed QCR and TCR modules on top of the base model MEVF+BAN [44]. The results are provided in Figure 8, where we report the average training time and test time of 10 epochs. Compared with the base model, our reasoning modules only increase the training time by a small factor, and the overhead in test time is negligible. It shows that the proposed conditional reasoning mechanism can be efficiently applied to existing Med-VQA systems.

V. CONCLUSION

Despite the recent booming interest in VQA, there has been little work in Med-VQA. In this paper, we have concerned

with the design of two key modules of a Med-VQA system – the reasoning module and the visual feature extraction module. For the former, we have proposed an effective conditional reasoning mechanism that endows the system with task-specific reasoning ability, which is lightweight and can be applied to existing Med-VQA models in a plug-and-play manner. For the latter, we have proposed to pre-train a visual feature extractor via contrastive learning to tackle the data scarcity problem, which can be readily used by any Med-VQA model on a small-scale dataset. Empirical evaluation on existing benchmarks demonstrates the high effectiveness of our proposals compared with the state-of-the-arts. We hope this work will serve as a solid step to advance the research of Med-VQA.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their helpful comments. This research was supported by the grant of P0038194 (1-ZVVX) funded by PolyU (UGC).

REFERENCES

- [1] A. B. Abacha, S. Gayen, J. J. Lau, S. Rajaraman, and D. Demner-Fushman, “NLM at imageclef 2018 visual question answering in the medical domain,” in *CLEF (Working Notes)*, 2018.
- [2] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, “Vqa-med: Overview of the medical visual question answering task at imageclef 2019,” in *CLEF Workshop*, 2019.
- [3] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, “Publicly available clinical bert embeddings,” *arXiv preprint arXiv:1904.03323*, 2019.
- [4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” in *CVPR*, 2018.
- [5] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Neural module networks,” in *CVPR*, 2016.
- [6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: visual question answering,” in *ICCV*, 2015.
- [7] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, “Mutan: Multimodal tucker fusion for visual question answering,” in *ICCV*, 2017.
- [8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *ICML*, 2020.
- [9] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” *arXiv preprint arXiv:2003.04297*, 2020.
- [10] J. Cho, J. Lei, H. Tan, and M. Bansal, “Unifying vision-and-language tasks via text generation,” in *ICML*. PMLR, 2021.
- [11] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in *ICML*, 2017.
- [12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [13] X. Dong, L. Zhu, D. Zhang, Y. Yang, and F. Wu, “Fast parameter adaptation for few-shot image captioning and visual question answering,” in *ACM MM*, 2018.
- [14] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *ICML*, 2017.
- [15] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” in *EMNLP*, 2016.
- [16] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” *arXiv preprint arXiv:1803.07728*, 2018.
- [17] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V in VQA matter: Elevating the role of image understanding in visual question answering,” in *CVPR*, 2017.
- [18] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *CVPR*, 2020.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016.
- [20] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, 1997.

- [21] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *ICCV*, 2021, pp. 3942–3951.
- [22] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers," *arXiv preprint arXiv:2004.00849*, 2020.
- [23] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," in *ICLR*, 2018.
- [24] —, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *CVPR*, 2019.
- [25] B. Ionescu, H. Müller, M. Villegas, A. G. S. de Herrera, C. Eickhoff, V. Andrearczyk, Y. D. Cid, V. Liauchuk, V. Kovalev, S. A. Hasan, Y. Ling, O. Farri, J. Liu, M. P. Lungren, D. Dang-Nguyen, L. Piras, M. Riegler, L. Zhou, M. Lux, and C. Gurrin, "Overview of imageclef 2018: Challenges, datasets and evaluation," in *CLEF*, 2018.
- [26] Y. Jiang, V. Natarajan, X. Chen, M. Rohrbach, D. Batra, and D. Parikh, "Pythia v0.1: the Winning Entry to the VQA Challenge 2018," *arXiv e-prints*, p. arXiv:1807.09956, 2018.
- [27] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *CVPR*, 2017.
- [28] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Computer Vision and Image Understanding*, 2017.
- [29] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar, "Mmbert: multimodal bert pretraining for improved medical vqa," in *ISBI*. IEEE, 2021, pp. 1033–1036.
- [30] J. Kim, J. Jun, and B. Zhang, "Bilinear attention networks," in *NeurIPS*, 2018.
- [31] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *ICML*. PMLR, 2021.
- [32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
- [33] N. A. Koozbanani, B. Unnikrishnan, S. A. Khurram, P. Krishnaswamy, and N. Rajpoot, "Self-path: Self-supervision for classification of pathology images with limited annotations," *IEEE Transactions on Medical Imaging*, 2021.
- [34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012.
- [35] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *AAAI*, 2015.
- [36] J. J. Lau, S. Gayen, A. B. Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific data*, 2018.
- [37] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.
- [38] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, 2017.
- [39] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," *arXiv preprint arXiv:2102.09542*, 2021.
- [40] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *NIPS*, vol. 32, 2019.
- [41] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NeurIPS*, 2016.
- [42] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," in *ICLR*, 2019.
- [43] J. H. Moon, H. Lee, W. Shin, and E. Choi, "Multi-modal understanding and generation for medical images and text via vision-language pre-training," *arXiv preprint arXiv:2105.11333*, 2021.
- [44] B. D. Nguyen, T. Do, B. X. Nguyen, T. Do, E. Tjiputra, and Q. D. Tran, "Overcoming data limitation in medical visual question answering," in *MICCAI*, 2019.
- [45] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
- [46] L. Peng, Y. Yang, Z. Wang, X. Wu, and Z. Huang, "Cra-net: Composed relation attention network for visual question answering," in *ACM MM*, 2019.
- [47] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.
- [48] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018.
- [49] M. Raghu, C. Zhang, J. M. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *NeurIPS*, 2019.
- [50] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *NeurIPS*, 2015.
- [51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.
- [52] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Hénao, and L. Carin, "Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms," in *ACL*, 2018.
- [53] L. Shi, F. Liu, and M. P. Rosen, "Deep multimodal learning for medical visual question answering," in *CLEF Workshop*, 2019.
- [54] R. Shrestha, K. Kafle, and C. Kanan, "Answer them all! toward universal visual question answering models," in *CVPR*, 2019.
- [55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [56] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway Networks," *arXiv preprint arXiv:1505.00387*, 2015.
- [57] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint*, 2019.
- [58] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.
- [59] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.
- [61] M. H. Vu, T. Löfstedt, T. Nyholm, and R. Sznitman, "A question-centric model for visual question answering in medical imaging," *IEEE transactions on medical imaging*, 2020.
- [62] Z. Wu, Y. Xiong, S. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance-level discrimination," *arXiv preprint arXiv:1805.01978*, 2018.
- [63] X. Yan, L. Li, C. Xie, J. Xiao, and L. Gu, "Zhejiang university at imageclef 2019 visual question answering in the medical domain," in *CLEF Workshop*, 2019.
- [64] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016.
- [65] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "CLEVRER: collision events for video representation and reasoning," in *ICLR*, 2020.
- [66] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, "Neural-symbolic VQA: disentangling reasoning from vision and language understanding," in *NeurIPS*, 2018.
- [67] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *ICCV*, 2017.
- [68] —, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *ICCV*, 2017.
- [69] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE transactions on neural networks and learning systems*, 2018.
- [70] L.-M. Zhan, B. Liu, L. Fan, J. Chen, and X.-M. Wu, "Medical visual question answering via conditional reasoning," in *ACM MM*, 2020.
- [71] Y. Zhou, X. Kang, and F. Ren, "Employing inception-resnet-v2 and bilstm for medical domain visual question answering," in *CLEF Workshop*, 2018.
- [72] Y. Zhou, T. Zhou, T. Zhou, H. Fu, J. Liu, and L. Shao, "Contrast-attentive thoracic disease recognition with dual-weighting graph reasoning," *IEEE Transactions on Medical Imaging*, 2021.