

# Medical Visual Question Answering via Conditional Reasoning

Li-Ming Zhan\*  
lmzhan.zhan@connect.polyu.hk  
The Hong Kong Polytechnic  
University

Bo Liu\*  
boliu.kelvin@gmail.com  
The Hong Kong Polytechnic  
University

Lu Fan  
complu.fan@connect.polyu.hk  
The Hong Kong Polytechnic  
University

Jiaxin Chen  
jiach.chen@connect.polyu.hk  
The Hong Kong Polytechnic  
University

Xiao-Ming Wu†  
xiao-ming.wu@polyu.edu.hk  
The Hong Kong Polytechnic  
University

## 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 industry and services, the technology is still in its infancy and is far from practical use. Med-VQA tasks are highly challenging due to the massive diversity of clinical questions and the disparity of required visual reasoning skills for different types of questions. In this paper, we propose a novel conditional reasoning framework for Med-VQA, aiming to automatically learn effective reasoning skills for various Med-VQA tasks. Particularly, we develop a question-conditioned reasoning module to guide the importance selection over multimodal fusion features. Considering the different nature of closed-ended and open-ended Med-VQA tasks, we further propose a type-conditioned reasoning module to learn a different set of reasoning skills for the two types of tasks separately. Our conditional reasoning framework can be easily applied to existing Med-VQA systems to bring performance gains. In the experiments, we build our system on top of a recent state-of-the-art Med-VQA model and evaluate it on the VQA-RAD benchmark [23]. Remarkably, our system achieves significantly increased accuracy in predicting answers to both closed-ended and open-ended questions, especially for open-ended questions, where a 10.8% increase in absolute accuracy is obtained. The source code can be downloaded from <https://github.com/awenbocc/med-vqa>.

## CCS CONCEPTS

• Computing methodologies → Computer vision tasks.

## KEYWORDS

medical visual question answering; attention mechanism; conditional reasoning

\*Equal contribution.

†Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

MM '20, October 12–16, 2020, Seattle, WA, USA

© 2020 Association for Computing Machinery.


ACM ISBN 978-1-4503-7988-5/20/10...\$15.00

<https://doi.org/10.1145/3394171.3413761>


## ACM Reference Format:

Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. 2020. Medical Visual Question Answering via Conditional Reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, October 12–16, 2020, Seattle, WA, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3394171.3413761>

## 1 INTRODUCTION

	Closed-ended		Open-ended
	Question	Are the clavicles broken?	Where is the lesion located?
	Answer	No	Anterior mediastinum
	Question Type	Object condition presence	Position
Organ		Chest	

	Closed-ended		Open-ended
	Question	Is the appendix normal in size?	What cut of the body is this image?
	Answer	Yes	Axial
	Question Type	Size	Plane
Organ		Abdomen	

**Figure 1: Examples of medical visual question answering. There may be multiple question-answer pairs associated with one image. There are two types of questions: closed-ended questions, where the answers are limited to “yes” or “no”, and open-ended questions, where the answers can be free-form text. Questions can also be categorized based on what they are about, such as position, size, or whether there is any abnormal condition.**

Medical visual question answering (Med-VQA) is a domain-specific visual question answering (VQA) problem that requires to interpret medical-related visual concepts by taking into account both image and language information. Specifically, a Med-VQA system aims to take a medical image and a clinical question about the image as input and output a correct answer in natural language. Med-VQA can help patients to get prompt feedback to their inquiries and make more informed decisions. It can reduce the stress on medical facilities so valuable medical resources could be saved for people with urgent needs. It can also help doctors to get a second opinion in diagnosis and reduce the high cost of training medical professionals. Despite the enormous potential of the technology, the research of Med-VQA is still in its embryonic stage, where the

literature is rather limited. Therefore, we will start with introducing general VQA, which, in recent years, has attracted a great deal of attention from both the computer vision and natural language processing research communities.

General VQA focuses on visual perceptual tasks that require common perceptual abilities shared by humans. For instance, given an image with several eggs in a basket, a child and a doctor can both easily answer the question “how many eggs are in the basket?”. General visual perceptual tasks include simple tasks such as “does it appear to be rainy?” and difficult tasks such as “what size is the cylinder that is left of the brown metal thing that is left of the big sphere?”. Hence, it needs multilevel reasoning skills to solve VQA tasks. Simple perceptual tasks require skills such as recognizing specific objects and scene understanding, whereas difficult tasks require higher-level reasoning skills such as performing comparisons, counting, or logical inferring. However, most of existing VQA methods are designed for either simple tasks or difficult tasks. Solving the two kinds of tasks in a single model is challenging and only considered in the high-data regime [32, 39].

Compared with VQA in the general domain, Med-VQA is a much more challenging problem. On one hand, clinical questions are more difficult but need to be answered with higher accuracy since they relate to health and safety. In general, Med-VQA systems need to deal with the above-mentioned two kinds of tasks simultaneously. There are basic perceptual tasks such as recognizing the modality of an image (e.g., CT or MRI) and tasks that require higher-level reasoning skills such as locating specific lesions or evaluating if the size of an organ is normal given prior knowledge. As such, domain-specific expert knowledge and multilevel reasoning skills are indispensable to infer correct clinical answers. On the other hand, well-annotated datasets for training Med-VQA systems are extremely lacking, but it is very laborious and expensive to obtain high-quality annotations by medical experts. To our knowledge, there is only one manually annotated dataset available - VQA-RAD [2], which contains various types of clinical questions but only includes 315 medical images. Therefore, state-of-the-art VQA models like [3, 17] cannot be directly applied to solve Med-VQA, since they commonly use sophisticated objection detection algorithms like Faster R-CNN with ResNet-101 [36], which contain many parameters and need to be trained on large-scale annotated datasets. Directly applying these models will lead to severe overfitting.

Previous attempts in designing Med-VQA systems [1, 2, 48] tried to exploit existing VQA models by using deep architectures pre-trained on general datasets and fine-tuned with limited medical data. However, both the image patterns and language styles of medical data are very different from those in the general domain [24]. Due to the large gap between the general domain and medical domain, knowledge transfer from the general domain by simply fine-tuning on limited medical data can only offer little benefit [35]. To overcome this problem, [30] proposed to enhance the quality of multimodal representations. In particular, they used a visual feature extractor pre-trained on external medical datasets with an unsupervised auto-encoder (CDAE) [29] and a meta-learning algorithm MAML [9] to learn a domain-specific weight initialization for the Med-VQA system. While these pioneering works pushed forward the research front on Med-VQA, they only considered improving the feature extraction module, whereas the reasoning module that

is critical for solving high-level reasoning tasks remains underexplored.

In this paper, we focus on improving the reasoning module of Med-VQA, where the extracted multimodal features are processed and analysed to infer a correct answer to the question. To this end, we devise a novel reasoning framework that endows a Med-VQA system with task-adaptive reasoning ability. Specifically, we propose a question-conditioned reasoning (QCR) module to guide the modulation of multimodal fusion features. In essence, our QCR module allows the Med-VQA system to learn and apply different reasoning skills to find the correct answer according to the question. This is achieved by not only considering the combination of multimodal features but also enabling additional transformation to the fused representations to identify question-specific reasoning information. Moreover, we observe that there are two major types of tasks in Med-VQA, as shown in Figure 1, closed-ended tasks, where answers are limited multiple-choice options, and open-ended tasks where answers may be free-form texts. Open-ended tasks are generally much harder to solve than closed-ended tasks. Though existing Med-VQA systems can achieve a reasonable performance on closed-ended tasks, they typically perform poorly on open-ended tasks. Therefore, to further model the disparity of the required reasoning skills for open-ended and closed-ended tasks, we propose a task-conditioned reasoning (TCR) strategy to consider both types of tasks accordingly.

In summary, our contributions are two-fold:

- **Methodology.** We propose a novel reasoning framework for Med-VQA by learning reasoning skills conditioned on both question information and task type information, which is realized by a question-conditioned reasoning module and a task-conditioned reasoning strategy.
- **Empirical study.** We conduct comprehensive evaluations on the benchmark dataset VQA-RAD [23] to demonstrate the effectiveness of our framework. The experimental results show our Med-VQA system consistently and significantly outperforms state-of-the-art baselines on both closed-ended and open-ended questions.

The rest of the paper is organized as follows. In Section 2, we review related works on VQA and Med-VQA. In Section 3, we discuss the proposed mechanisms of QCR and TCR in detail. In Section 4, we present experimental evaluation on our framework, including comparisons with baselines and an ablation study of the proposed mechanisms. Finally, we conclude this paper in Section 5.

## 2 RELATED WORK

In this section, we briefly review the most relevant works in general VQA and introduce recent development of Med-VQA.

### 2.1 Visual Question Answering

A typical VQA system consists of a 3-stage pipeline: (1) Features from the two modalities are extracted by large-scale pre-trained models such as ResNet or BERT; (2) The extracted features are aggregated to learn a joint representation; (3) The joint representation is then sent to the inference module, which can be a classifier to predict the answer. The VQA benchmarks in the general domain stress on different levels of reasoning skills and can be categorized

into two groups accordingly: the first group with human-annotated questions [5, 11] and the second group based on sophisticated synthetic questions [15, 18, 44].

A great many VQA systems differ in how they aggregate multimodal features. Early works employed simple mechanisms such as concatenation and pooling [5]. Later on, inter-modality relation, which models the connection between visual objects and different parts of the question, was studied extensively [3, 20, 25, 26, 43]. [3] proposed a combined bottom-up and top-down visual attention mechanism that extracts a series of regions of interest at the object level and aggregates region features using an attention mechanism. [20] proposed an efficient bilinear attention networks (BAN) to jointly model the interaction between the modalities. In this paper, BAN is also employed for cross-modal feature fusion in our proposed framework.

Some other VQA systems stress on higher-level reasoning skills [4, 8, 13, 14, 28, 34]. [4] proposed neural module networks (NMN) that decompose questions by a semantic parser and trigger different pre-defined modules accordingly. NMN is fancy but heavily relies on complex annotations and pre-defined structures. [8] proposed fast parameter adaptation for image-text modeling (FPAIT) that controls the image feature extraction layers by generating dynamic normalization parameters from textual features. Neural-symbolic (NS) approaches [27, 45] exploit executable symbolic programs to mimic the human reasoning process.

To solve image understanding tasks and complex reasoning tasks in a unified framework, [39] employed recurrent aggregation to capture interactions among bi-modal embeddings and validated the algorithm across different tasks. Further, to enable the system to perform sophisticated reasoning, [32] proposed to add a trinary relation attention layer to learn trinary relations among objects.

## 2.2 Medical Visual Question Answering

Existing studies attempted to adapt advanced general VQA methods based on large-scale pre-trained models for Med-VQA [1, 2, 38, 42, 48]. These studies are mostly reports of the ImageCLEF VQA-Med Challenge [2, 16]. Typically, visual features are extracted by deep pre-trained architectures such as ResNet or VGGNet, and textual features are extracted by stacked RNN-based layers. In [48], the authors concatenated features from both modalities and employed Bi-LSTM to decode answers. [1] exploited stacked attention network (SAN) [43] and compact bilinear pooling (MCB) [10]. [42] and [38] both employed multimodal factorized bilinear pooling [46] for feature fusion. In addition, [38] extracted question topic representations by using embedding-based topic model and predicted question categories with SVM.

However, due to the large discrepancy between medical data and data in the general domain, such straightforward adaptations severely suffer from data scarcity and lack of multilevel reasoning ability. To overcome data limitation, [30] proposed mixture of enhanced visual features (MEVF) that utilizes external medical datasets to learn a domain-specific weight initialization for the Med-VQA model, which is realized by using the unsupervised convolutional denoising auto-encoder (CDAE) [29] and the meta-learning method MAML [9]. Nevertheless, it only employs a bilinear attention mechanism as the reasoning module for multimodal feature

fusion. In this work, we follow MEVF [30] to initialize the visual feature extractor of our Med-VQA system but focus on improving the reasoning module.

## 3 METHODOLOGY

### 3.1 Problem Formulation

In this paper, we consider VQA as a classification problem with  $C$  candidate answers where  $C$  is usually a large number. Denote by  $\mathcal{D} = \{(v_i, q_i, a_i)\}_{i=1}^n$  the training dataset for a 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 VQA model aims to learn a function  $f$  that maps each  $(v_i, q_i)$  pair to a score vector  $\mathbf{s} \in \mathbb{R}^C$  where the  $i$ -th element  $s_i$  is the score for the  $i$ -th class. The probability for the  $i$ -th class is obtained by the softmax function, i.e.,  $p(i) = \frac{e^{s_i}}{\sum_{j=1}^C e^{s_j}}$ ,  $1 \leq i \leq C$ . The function  $f$  is usually instantiated by a neural network with parameters  $\theta$ . The training is conducted by maximizing the log likelihood of the correct answer  $a_i$  across all training tasks:

$$\theta = \arg \max_{\theta} \sum_{i=1}^n \log p(a_i; f_{\theta}(v_i, q_i)). \quad (1)$$

### 3.2 General VQA Model

A VQA model usually consists of three components: a multimodal feature extraction module, an attention-based feature fusion module, and a classifier. From the architecture perspective, there is no conspicuous difference between general VQA and Med-VQA. We elaborate on each component below.

**Multimodal feature extraction.** On one hand, visual features of images are usually extracted by an intermediate feature map of convolution neural networks (CNNs) such as simple CNNs [47] or faster RCNN [36]. On the other hand, semantic features of questions are commonly extracted by a recurrent neural network (e.g., LSTM [12] or GRU [7]) which takes a padded word embedding sequence as input.

**Attention-based feature fusion.** The attention mechanism constitutes an essential part in existing state-of-the-art VQA models [6, 20, 41, 43]. It models the interactive relationship between visual and textual representations to produce multimodal fusion features, which can be readily used for predicting answers to the questions. Hence, the attention mechanism can be considered as the reasoning module of VQA.

**Making predictions with a classifier.** VQA is usually formulated as a classification problem, where distinct answers are considered as different categories. A common choice of the classifier is multilayer perceptron (MLP).

### 3.3 Our Proposed Conditional Reasoning Framework

Figure 2 provides an overview of our proposed conditional reasoning framework. Our central idea is to learn task-adaptive reasoning skills for different types of Med-VQA tasks. Current approaches commonly adopt neural attention as a simple reasoning module to

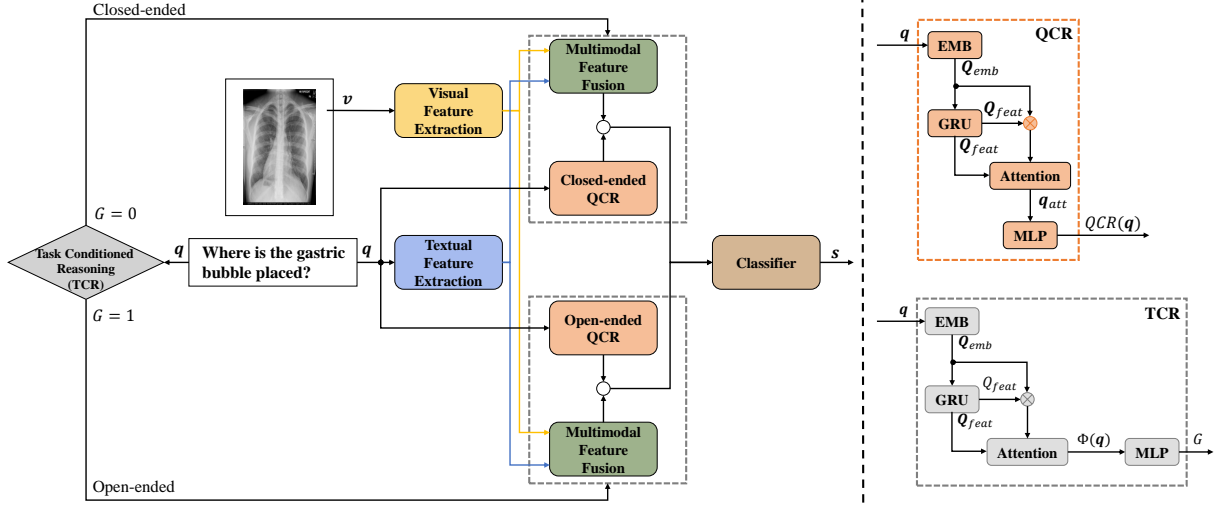


Figure 2: Our proposed question-conditioned reasoning model for Med-VQA.

combine multimodal information for visual reasoning. In comparison, our proposed reasoning module learns an additional question-conditioned modulation for the fused multimodal representations. Moreover, to further model multilevel reasoning skills required by complex Med-VQA tasks, we propose to learn a separate reasoning module for closed-ended and open-ended tasks respectively based on a question type classifier.

**3.3.1 QCR: Question-Conditioned Reasoning Module.** Recent studies in Med-VQA have shown that multimodal feature fusion as a simple reasoning module could perform relatively well for closed-ended questions, but the performance for open-ended questions is considerably poorer [23, 30]. To devise a more powerful reasoning module for Med-VQA, we propose to improve the standard reasoning module by adding a question-conditioned modulation component. The motivations are two-fold. First, similar to human reasoning processes, dealing with different tasks requires corresponding task-specific skills. Second, it has been shown that the question itself contains abundant task information for VQA tasks [19]. As such, our reasoning module aims to extract task information from the question sentence 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 2 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 [33]. Let  $\mathbf{w}_i \in \mathbb{R}^{d_w}$  denote the corresponding word vector for the  $i$ -th word:

$$\mathbf{Q}_{emb} = \text{WordEmbedding}(q) = [\mathbf{w}_1, \dots, \mathbf{w}_l]. \quad (2)$$

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 embedding of the question:

$$\mathbf{Q}_{feat} = \text{GRU}([\mathbf{w}_1, \dots, \mathbf{w}_l]) = [\eta_1, \dots, \eta_l], \quad (3)$$

where  $\mathbf{Q}_{feat} \in \mathbb{R}^{d_G \times l}$ , and  $\eta_i$  denotes the embedding at the  $i$ -th position.

Now, the question embedding  $\mathbf{Q}_{feat}$  does not have any emphasis on different words. To further highlight the important parts of each question, e.g., “how many lesions are in the spleen?”, we design an attention mechanism to impose importance weights on different words:

$$\tilde{\mathbf{Q}} = \mathbf{Q}_{emb} \otimes \mathbf{Q}_{feat}, \quad (4a)$$

$$\mathbf{Y} = \tanh(\mathbf{W}_1 \tilde{\mathbf{Q}}), \quad (4b)$$

$$\tilde{\mathbf{Y}} = \sigma(\mathbf{W}_2 \tilde{\mathbf{Q}}), \quad (4c)$$

$$\mathbf{G} = \mathbf{Y} \circ \tilde{\mathbf{Y}}. \quad (4d)$$

Here,  $\otimes$  denotes feature concatenation in the feature dimension,  $\tilde{\mathbf{Q}} \in \mathbb{R}^{(d_w+d_G) \times l}$ ,  $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d_G \times (d_w+d_G)}$  are trainable weights,  $\sigma$  and  $\tanh$  are the sigmoid activation function and the gated hyperbolic tangent activation respectively, and  $\circ$  is the Hadamard product. In this step, we take the advantages of both context-free (Glove) and contextual embeddings (GRU) to form  $\tilde{\mathbf{Q}}$ , which has been shown useful for many tasks in natural language processing [22, 37].  $\tilde{\mathbf{Y}}$  acts as a gate on the intermediate activation  $\mathbf{Y}$  to control the output  $\mathbf{G} \in \mathbb{R}^{d_G \times l}$  [40].

Then, the attention vector  $\boldsymbol{\alpha} \in \mathbb{R}^{l \times 1}$  for the question embedding  $\mathbf{Q}_{feat}$  can be obtained by

$$\boldsymbol{\alpha} = \text{softmax}((\mathbf{W}_a \mathbf{G})^\top), \quad (5)$$

where  $\mathbf{W}_a \in \mathbb{R}^{1 \times d_G}$  are trainable weights.

Finally, with the attention vector  $\boldsymbol{\alpha}$ , we obtain the final output of QCR as:

$$\mathbf{q}_{att} = \mathbf{Q}_{feat} \boldsymbol{\alpha}, \quad (6)$$

$$\text{QCR}(q) = \text{MLP}(\mathbf{q}_{att}), \quad (7)$$

where  $\mathbf{q}_{att}$  is the aggregated question representation, and MLP is a multilayer perceptron network that provides additional non-linear transformation for importance selection.

On the whole, the mapping function  $f_\theta$  of a general VQA model is composed of two modules, a multimodal feature fusion module  $A_{\theta_m}$  and a classification module  $D_{\theta_c}$ , i.e.,

$$f_\theta(v, q) = D_{\theta_c}(A_{\theta_m}(v, q)).$$

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

$$\mathbf{s} = D_{\theta_c}(A_{\theta_m}(v, q) \circ QCR(q)), \quad (8)$$

where  $\circ$  denotes element-wise product.

**3.3.2 TCR: Type-Conditioned Reasoning.** It has been observed that the closed-ended questions are usually easier to answer than open-ended questions. For example, the question “is the aorta visible in this section?” can be answered by basic image understanding, whereas the open-ended question “what vascular problem is seen above?” usually requires multi-step reasoning. As such, Med-VQA systems need to be endowed with multilevel reasoning abilities, which are lacking in current VQA models [39].

To this end, we propose to use a separate reasoning module for closed-ended questions and open-ended questions respectively, where each module is instantiated by the QCR module proposed in Section 3.3.1. Particularly, we want to train a task type classifier  $G$  that takes a question as input and output 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 “Is\Dose\Are\etc.”, and open-ended questions often start with “What\How many\Where\etc”. The differences of the two types of questions can be captured by question embeddings, which makes it possible to train a reliable classifier that divides Med-VQA tasks into two subbranches as shown by the rhombus module in Figure 2.

Similar to Section 3.3.1, we use Equations (2) - (6) to compute the question embedding and denote the mapping as  $\Phi$ . We then employ a multilayer perceptron  $MLP_T$  to map question embeddings 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  $G$  is then formulated as:

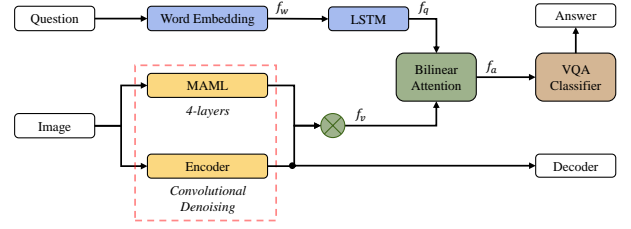
$$G(q) = \begin{cases} 0, & \text{if } p_0^t > p_1^t, \\ 1, & \text{else.} \end{cases} \quad (9)$$

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

$$\mathbf{s} = \begin{cases} D_{\theta_c}(A_{\theta_m}^{cl}(v, q) \circ QCR^{cl}(q)), & \text{if } G(q) = 0, \\ D_{\theta_c}(A_{\theta_m}^{op}(v, q) \circ QCR^{op}(q)), & \text{if } G(q) = 1, \end{cases} \quad (10)$$

where  $cl$  and  $op$  stand for closed-ended and open-ended respectively.

In summary, as depicted in Figure 2, our proposed framework with QCR and TCR reasoning modules solves a Med-VQA task by the following steps. First, visual features and the textual features are extracted by a shared feature extracting module for both closed-ended and open-ended tasks as indicated by the yellow and blue rectangles respectively. Second, these multimodal features are fed into respective reasoning modules (indicated by the two grey dash rectangles) based on the judgement of the question type classifier (TCR module). Third, in each reasoning module, the output is



**Figure 3: The backbone framework [30] used in our experiments.**

produced by element-wise multiplication between the fusion representation and the modulation vector obtained by the QCR module. Finally, the predicted scores for candidate answers are produced by an MLP classifier.

### 3.4 A Case Study

Our proposed conditional reasoning framework including QCR and TCR can be applied to most of the attention-based Med-VQA models that use multimodal fusion features to find answers.

In this paper, we conduct a case study on a state-of-the-art Med-VQA system [30] and implement our proposed framework upon it. As shown in Figure 3, the model in [30] extracts visual features by a mixture of enhanced visual features (MEVF) module that consists of two parallel sub-modules: MAML [9] and convolutional unsupervised auto-encoder CDAE [29], whose parameters are initialized by pre-training on some external medical datasets respectively. On the other hand, the question is encoded by an LSTM encoder.

The model uses bilinear attention (BAN) [20] for multimodal feature fusion, and the fusion features  $\mathbf{f}$  are given by

$$\mathbf{f} = \mathbf{V}\mathcal{A}\mathbf{Q}, \quad (11)$$

where  $\mathbf{V}$  and  $\mathbf{Q}$  are the extracted visual and textual feature matrices respectively, and  $\mathcal{A}$  is the bilinear attention matrix derived from a low-rank bilinear pooling function that takes  $\mathbf{V}$  and  $\mathbf{Q}$  as inputs.

We then implement our framework by applying the proposed QCR modulation on the fusion features  $\mathbf{f}$  and train a pair of parallel reasoning modules for closed-ended and open-ended tasks respectively as illustrated in Figure 2. The experimental results are presented in Section 4.

## 4 EXPERIMENTS

In this section, we conduct comprehensive experiments to evaluate the performance of our proposed conditional reasoning framework including the QCR and TCR components proposed in Section 3.3 on the recently published benchmark VQA-RAD [23]. We compare our approach with current state-of-the-art baselines [23, 30] for Med-VQA. Moreover, we validate the rationality of our model by an ablation study and provide a qualitative evaluation by visualizing some Med-VQA tasks and answers. All the experimental results reported in this paper can be reproduced by the provided code (can be downloaded from <https://github.com/awenbocc/med-vqa>).

**Table 1: Test accuracies of our proposed question-conditioned reasoning model and baselines on VQA-RAD [23]. \* indicates our re-implementation of MEVF+BAN [30].**

	SAN [23, 43] (baseline)	MCB [10, 23] (baseline)	BAN [20, 30] (baseline)	MEVF+SAN [30] (baseline)	MEVF+BAN [30] (baseline)	MEVF+BAN (*) (re-implemented)	Ours
Open-ended (%)	24.2	25.4	27.6	40.7	43.9	49.2	<b>60.0</b>
Closed-ended (%)	57.2	60.6	66.5	74.1	75.1	77.2	<b>79.3</b>

**Table 2: Statistics of the VQA-RAD dataset [23].**

	Training Set (Total: 3064)		Test Set (Total: 451)	
	Open-ended	Closed-ended	Open-ended	Closed-ended
MODALITY	77	77	16	17
PLANE	47	47	14	12
ORGAN	34	15	8	2
ABNORMAL	126	191	18	38
PRESENCE	267	965	45	122
POSITION	439	67	52	8
ATTRIBUTE	70	79	4	14
COLOR	17	67	0	3
COUNT	22	17	2	4
SIZE	25	244	5	41
OTHER	119	52	15	11
Total	<b>1243</b>	<b>1821</b>	<b>179</b>	<b>272</b>

#### 4.1 Dataset

We experiment on the well-composed benchmark VQA-RAD [23], which is specially designed to train and evaluate Med-VQA systems. VQA-RAD contains 315 radiology images and 3515 question - answer pairs generated by clinicians. There may be multiple questions associated with one image. Some examples of the dataset are shown in Figure 1. Specifically, the clinicians could ask different kinds of questions about one image, e.g., questions about “organ system”, “abnormality”, or “object/condition presence”. More generally, the questions can also be categorized into two types: “closed-ended” questions whose answers are “yes/no” or other limited choices, and “open-ended” questions whose answers can be free-form texts. Table 2 shows a detailed summary of the VQA-RAD dataset. In our experiments, we use the processed version of VQA-RAD as in [30], where there are 3064 tasks for training and 451 tasks for testing.

#### 4.2 Training Setup

All our experiments are conducted on a Ubuntu 16.04 server with Titan XP GPU using PyTorch [31]. The training details are provided below.

As shown in Figure 3, the backbone of our Med-VQA system is our re-implementation of [30] with bilinear attention networks (BAN) [20]. To achieve a fair comparison, we follow the training settings in [30]. Specifically, for visual representations, we use the pre-trained initialization from [30] for the MAML [9] and CDAE [29] modules. For semantic textual features, we use Glove [33] to initialize word embeddings and employ a 1024D - LSTM to extract semantic information from questions.

All the GRUs used in QCR and TCR are set to have 1024 - dimensional hidden states. The  $MLP$  in Equation (7) and  $MLP_T$  in Equation (9) are instantiated with hidden units [2048, 1024] and [256, 64] respectively. We pre-train the task type classifier  $G$  with the label “answer\_type” in the training set of VQA-RAD [23], using

Adam [21] optimizer with learning rate 0.0001 for 150 epochs. The trained classifier reaches 99.33% classification accuracy on the test set of VQA-RAD, which demonstrates the effectiveness of our proposed pipeline (Equations (2) - (6)) for extracting semantic textual features as well as the usefulness of the question type classifier  $G$ . In training our reasoning modules, we use Adam optimizer with learning rate 0.005.

Lastly, the model accuracy is computed by:

$$Accuracy = \frac{S_c}{S_{all}} \times 100, \quad (12)$$

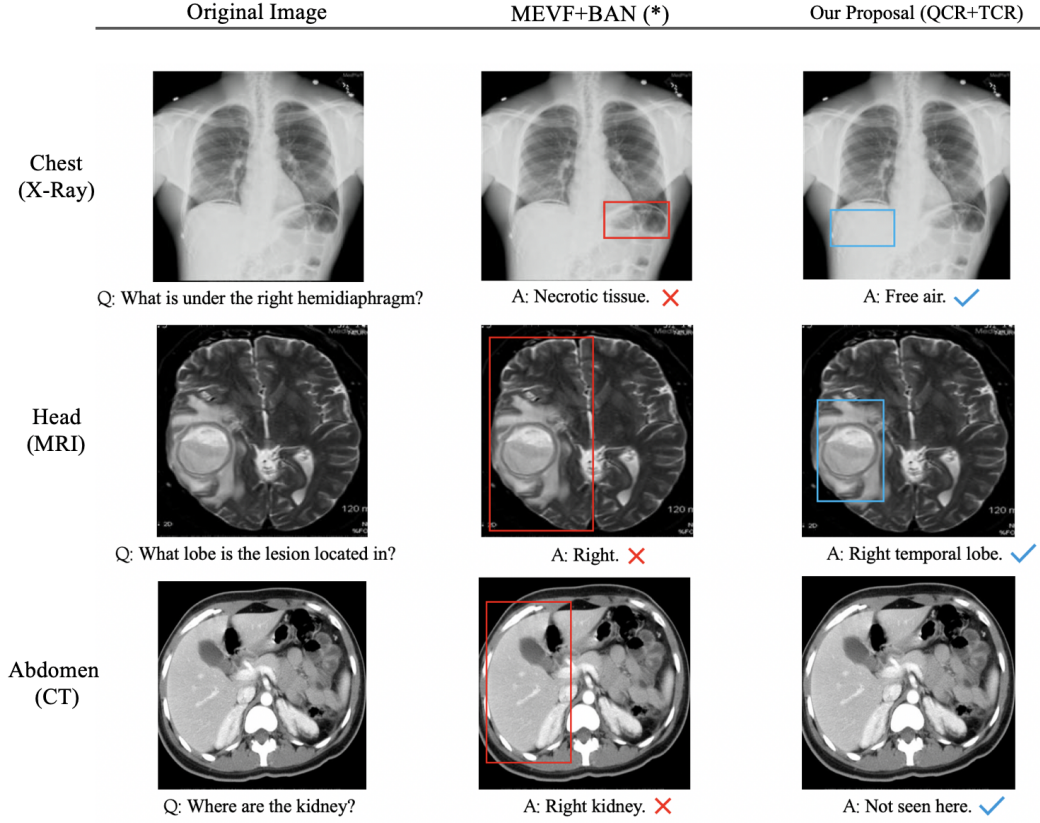
where  $S_c$  and  $S_{all}$  denote the number of correctly answered questions and the total number of questions respectively.

#### 4.3 Comparison with the State-of-the-arts

We compare our approach with state-of-the-art methods used in [23] and [30]. Specifically, [23] directly employed existing VQA models in the general domain to solve Med-VQA, e.g., the stacked attention networks (SAN) [43] and the multimodal compact bilinear pooling (MCB) [10]. To develop a specialized Med-VQA system, [30] proposed a mixture of enhanced visual features (MEVF) framework and combined it with different attention mechanisms such as bilinear attention networks (BAN) [20] and SAN. Below we briefly introduce each baseline.

- **SAN** [43] tried to model the multi-step reasoning process of question answering. It proposed a stacked attention model that takes the question as a query to progressively search for corresponding regions in an image.
- **MCB** [10] argued that using outer product to fuse visual and textual feature vectors is computationally expensive. It proposed a multimodal compact bilinear pooling (MCB) mechanism that projects outer product to a lower dimensional space to reduce the computational cost of feature fusion.
- **BAN** [20] was motivated by the same concern of [10], and it also used bilinear multimodal fusion. Differently, it proposed to utilize low-rank bilinear pooling to reduce the rank of weight matrix so as to reduce the computational overhead. A detailed description of this method is provided in Section 3.4.
- **MEVF+SAN** [30] used the proposed MEVF framework to extract image features that can generalize better on medical images. The details of the MEVF framework is illustrated in Figure 3. In addition, it used the attention mechanism SAN proposed in [43] to fuse the features of images and questions.
- **MEVF+BAN** [30] used the attention mechanism BAN proposed in [20] to fuse the visual features extracted by the MEVF framework and the semantic textual features.





**Figure 4: Visualization of results of our proposal and the baseline MEVF+BAN (\*).** The examples include three radiology images of different modalities on different body parts with open-ended questions. Red and blue boxes locate the corresponding region of each answer. ✓ and ✗ indicate the correctness of the answer given by each model.

Table 1 shows the results of our method and the baselines. For SAN and MCB, we cite the results from [23]. For BAN, MEVF+SAN, and MEVF+BAN, we cite the results from [30]. Our re-implemented version MEVF+BAN (\*) shows slightly higher results than the original implementation in [30]. It can be seen that our method achieves the best performance and significantly outperforms all the baselines. Compared with the best baseline MEVF+BAN (\*), our method achieves 10.8% and 2.1% increases in absolute accuracy for the “open-ended” and “closed-ended” questions respectively. This shows the effectiveness of our proposed conditional reasoning framework for solving Med-VQA tasks.

#### 4.4 Ablation Study

In this subsection, we conduct an ablation study to verify the effectiveness of the QCR and TCR components (described in Section 3.3) in our proposal. The results are summarized in Table 3. To ensure a fair comparison, we use MEVF+BAN (\*), the re-implemented version of [30] as reported in Section 4.3, as the base model.

We first study the impact of QCR and TCR independently. As shown in Table 3, QCR improves the overall accuracy from 66.1% to 67.8%; for TCR, there is a large improvement, i.e., the overall

**Table 3: Ablation study of our proposed QCR and TCR modules on VQA-RAD [23]. \* indicates our re-implemented version of [30].**

Base Model	QCR	TCR	VQA Accuracy (%)		
			Overall	Open-ended	Closed-ended
MEVF+BAN (*)			66.1	49.2	77.2
MEVF+BAN (*)	✓		67.8	51.4	78.7
MEVF+BAN (*)		✓	70.1	56.7	79.0
MEVF+BAN (*)	✓	✓	<b>71.6</b>	<b>60.0</b>	<b>79.3</b>

accuracy increased from 66.1% to 70.1% (the “open-ended” accuracy increased from 49.2% to 56.7% and “closed-ended” accuracy increased from 77.2% to 79.0%). These results show that: (i) Both QCR and TCR can improve the performance of a Med-VQA system, which validates that it is reasonable and effective to exploit question information to learn task-adaptive reasoning skills for different tasks; (ii) The fact that TCR improves the performance by a large margin validates that it is necessary to learn multi-level reasoning skills for different types of Med-VQA tasks.

Finally, we incorporate both QCR and TCR in the base model, and observe further improvements. The overall accuracy is increased from 66.1% to 71.6%, the open-ended accuracy is increased from 49.2% to 60.0%, and the closed-ended accuracy is increased from 77.2% to 79.3%. Remarkably, combining QCR and TCR leads to a huge improvement on the open-ended tasks. Since the open-ended tasks are generally much harder than the closed-ended tasks, the large performance gain suggests that our method is capable of learning higher-level reasoning skills to deal with difficult tasks.

#### 4.5 Qualitative Evaluation

In this subsection, we provide a qualitative comparison of our method and the best baseline MEVF+BAN (\*). Figure 4 shows the visualization of the results of our method and the baseline on three open-ended tasks from VQA-RAD, which cover three modalities of radiology images on different parts of human body.

The first task (top row) is about a chest X-Ray image. While the baseline model wrongly predict the relevant region and answer to the question, our method correctly locates the relevant region and gives the right answer. Note that an radiology image should be read in a laterally inverted fashion. Therefore, “right” in the question means the right side of the body which corresponds to the left region of the image. For the second task (middle row), a head MRI image is presented. Although the baseline method can locate the relevant region in the image, it fails to provide a specific and concrete answer. In contrast, our model gives a concrete and accurate answer. The third task (bottom row) is about an abdomen CT image. The baseline model gives a wrong answer that mistakes the liver as the right kidney. In comparison, our model correctly figures out that there is actually no kidney in the image.

The examples show that our model is capable of providing reasonable answers to complex Med-VQA tasks, by learning advanced reasoning skills such as locating and recognizing specific objects or detecting the mismatch between an image and a question.

#### 4.6 Efficiency Evaluation

In Figure 5, we evaluate the time efficiency of our reasoning modules QCR and TCR, and compare with the base model MEVF+BAN (\*) (our re-implementation of [30]). The results are obtained by averaging the training time and test time of 10 epochs respectively. Compared with the base model, it can be seen that our reasoning modules only slightly increase the computational time for training, while the overhead for testing is negligible. These results show that our method can be efficiently applied to existing Med-VQA systems in practice.

### 5 CONCLUSION

In this paper, we have proposed an effective conditional reasoning framework for Med-VQA, which endows a VQA system with task-specific reasoning ability. This is realized by the use of an attention mechanism conditioned by task information to guide the importance weighting of multimodal fusion features. Our framework is lightweight and can be applied to existing Med-VQA systems in a plug-and-play manner. Empirical evaluation on the recently published benchmark dataset VQA-RAD shows that our approach

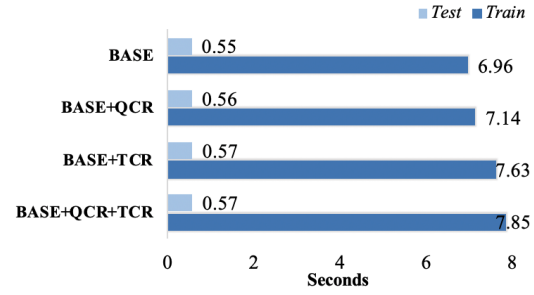


Figure 5: Time efficiency of our method. BASE represents the base model MEVF+BAN (\*).

achieves superior performance compared with state-of-the-art Med-VQA models. Particularly on open-ended tasks where high-level visual reasoning skills are needed, our approach improves the accuracy of answers by a wide margin, demonstrating the effectiveness of the proposed conditional reasoning modules. In future work, we plan to further improve the reasoning ability of our model and conduct more evaluation on Med-VQA tasks.

### ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers for their helpful comments. This research was supported by the grants of P0001175 (ZVJJ) and P0030935 (ZVPY) funded by PolyU (UGC).

### REFERENCES

- [1] Asma Ben Abacha, Soumya Gayen, Jason J. Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman. 2018. NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, Vol. 2125)*. CEUR-WS.org, Avignon, France.
- [2] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, Vol. 2380)*. CEUR-WS.org, Lugano, Switzerland.
- [3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, Salt Lake City, UT, USA, 6077–6086.
- [4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural Module Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, Las Vegas, NV, USA, 39–48.
- [5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision, ICCV*. IEEE Computer Society, Santiago, Chile, 2425–2433.
- [6] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. 2015. ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering. *arXiv e-prints* (Nov. 2015), arXiv:1511.05960.
- [7] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics, Doha, Qatar, 103–111.
- [8] Xuanyi Dong, Linchao Zhu, De Zhang, Yi Yang, and Fei Wu. 2018. Fast Parameter Adaptation for Few-shot Image Captioning and Visual Question Answering. In *2018 ACM Multimedia Conference on Multimedia Conference, MM*. ACM, Seoul, Republic of Korea, 54–62.



- [9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, Sydney, NSW, Australia, 1126–1135.
- [10] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP*. The Association for Computational Linguistics, Austin, Texas, USA, 457–468.
- [11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, Honolulu, HI, USA, 6325–6334.
- [12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.
- [13] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to Reason: End-to-End Module Networks for Visual Question Answering. In *IEEE International Conference on Computer Vision, ICCV*. IEEE Computer Society, Venice, Italy, 804–813.
- [14] Drew A. Hudson and Christopher D. Manning. 2018. Compositional Attention Networks for Machine Reasoning. In *6th International Conference on Learning Representations, ICLR*. OpenReview.net, Vancouver, BC, Canada.
- [15] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE, Long Beach, CA, USA, 6700–6709.
- [16] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba Garcia Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalov, Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Matthew P. Lungren, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux, and Cathal Gurrin. 2018. Overview of ImageCLEF 2018: Challenges, Datasets and Evaluation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF (Lecture Notes in Computer Science, Vol. 11018)*. Springer, Avignon, France, 309–334.
- [17] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0.1: the Winning Entry to the VQA Challenge 2018. *arXiv e-prints* (July 2018), arXiv:1807.09956.
- [18] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, Honolulu, HI, USA, 1988–1997.
- [19] Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Comput. Vis. Image Underst.* 163 (2017), 3–20.
- [20] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS*. NeurIPS, Montréal, Canada, 1571–1581.
- [21] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR, Conference Track Proceedings*. OpenReview.net, San Diego, CA, USA.
- [22] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent Convolutional Neural Networks for Text Classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Press, Austin, Texas, USA, 2267–2273.
- [23] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data* 5, 1 (2018), 1–10.
- [24] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyo Setio, Francesco Ciompi, Mohsen Ghahfoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical Image Anal.* 42 (2017), 60–88.
- [25] Fei Liu, Jing Liu, Richang Hong, and Hanqing Lu. 2019. Erasing-based Attention Learning for Visual Question Answering. In *Proceedings of the 27th ACM International Conference on Multimedia, MM*. ACM, Nice, France, 1175–1183.
- [26] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, NeurIPS*. Barcelona, Spain, 289–297.
- [27] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *7th International Conference on Learning Representations, ICLR*. OpenReview.net, New Orleans, LA, USA.
- [28] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. 2018. Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, Salt Lake City, UT, USA, 4942–4950.
- [29] Jonathan Masci, Ueli Meier, Dan C. Ciresan, and Jürgen Schmidhuber. 2011. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In *Artificial Neural Networks and Machine Learning - ICANN 2011 - 21st International Conference on Artificial Neural Networks, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 6791)*. Springer, Espoo, Finland, 52–59.
- [30] Binh D. Nguyen, Thanh-Toan Do, Binh X. Nguyen, Tuong Do, Erman Tjiputra, and Quang D. Tran. 2019. Overcoming Data Limitation in Medical Visual Question Answering. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Part IV (Lecture Notes in Computer Science, Vol. 11767)*. Springer, Shenzhen, China, 522–530.
- [31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*. Vancouver, BC, Canada, 8024–8035.
- [32] Liang Peng, Yang Yang, Zheng Wang, Xiao Wu, and Zi Huang. 2019. CRA-Net: Composed Relation Attention Network for Visual Question Answering. In *Proceedings of the 27th ACM International Conference on Multimedia, MM*. ACM, Nice, France, 1202–1210.
- [33] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, Doha, Qatar, 1532–1543.
- [34] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press, New Orleans, Louisiana, USA, 3942–3951.
- [35] Maithra Raghu, Chiyuan Zhang, Jon M. Kleinberg, and Samy Bengio. 2019. Transfusion: Understanding Transfer Learning for Medical Imaging. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*. Vancouver, BC, Canada, 3342–3352.
- [36] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, NeurIPS*. Montreal, Quebec, Canada, 91–99.
- [37] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL, Volume 1: Long Papers*. Association for Computational Linguistics, Melbourne, Australia, 440–450.
- [38] Lei Shi, Feifan Liu, and Max P. Rosen. 2019. Deep Multimodal Learning for Medical Visual Question Answering. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, Vol. 2380)*. CEUR-WS.org, Lugano, Switzerland.
- [39] Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2019. Answer Them All! Toward Universal Visual Question Answering Models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE, Long Beach, CA, USA, 10472–10481.
- [40] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway Networks. *arXiv e-prints* (May 2015), arXiv:1505.00387.
- [41] Huijuan Xu and Kate Saenko. 2016. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. In *Computer Vision - ECCV 2016 - 14th European Conference, Proceedings, Part VII (Lecture Notes in Computer Science, Vol. 9911)*. Springer, Amsterdam, The Netherlands, 451–466.
- [42] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu. 2019. Zhejiang University at ImageCLEF 2019 Visual Question Answering in the Medical Domain. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, Vol. 2380)*. CEUR-WS.org, Lugano, Switzerland.
- [43] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked Attention Networks for Image Question Answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Las Vegas, NV, USA, 21–29.
- [44] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. CLEVRER: Collision Events for Video Representation and Reasoning. In *8th International Conference on Learning Representations, ICLR*. OpenReview.net, Addis Ababa, Ethiopia.
- [45] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS*. Montréal, Canada, 1039–1050.
- [46] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering. In

*IEEE International Conference on Computer Vision, ICCV*. IEEE Computer Society, Venice, Italy, 1839–1848.

- [47] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2015. Simple Baseline for Visual Question Answering. *arXiv e-prints* (Dec. 2015), arXiv:1512.02167.
- [48] Yangyang Zhou, Xin Kang, and Fuji Ren. 2018. Employing Inception-Resnet-v2 and Bi-LSTM for Medical Domain Visual Question Answering. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, Vol. 2125)*. CEUR-WS.org, Avignon, France.