A Dual-Attention Learning Network With Word and Sentence Embedding for Medical Visual Question Answering

Xiaofei Huang and Hongfang Gong

Abstract—Research in medical visual question answering (MVQA) can contribute to the development of computer-aided diagnosis. MVQA is a task that aims to predict accurate and convincing answers based on given medical images and associated natural language questions. This task requires extracting medical knowledge-rich feature content and making fine-grained understandings of them. Therefore, constructing an effective feature extraction and understanding scheme are keys to modeling. Existing MVQA question extraction schemes mainly focus on word information, ignoring medical information in the text, such as medical concepts and domain-specific terms. Meanwhile, some visual and textual feature understanding schemes cannot effectively capture the correlation between regions and keywords for reasonable visual reasoning. In this study, a dual-attention learning network with word and sentence embedding (DALNet-WSE) is proposed. We design a module, transformer with sentence embedding (TSE), to extract a double embedding representation of questions containing keywords and medical information. A dual-attention learning (DAL) module consisting of self-attention and guided attention is proposed to model intensive intramodal and intermodal interactions. With multiple DAL modules (DALs), learning visual and textual co-attention can increase the granularity of understanding and improve visual reasoning. Experimental results on the ImageCLEF 2019 VQA-MED (VQA-MED 2019) and VQA-RAD datasets demonstrate that our proposed method outperforms previous state-of-the-art methods. According to the ablation studies and Grad-CAM maps, DALNet-WSE can extract rich textual information and has strong visual reasoning ability.

Index Terms—Medical visual question answering, double embedding, medical information, guided attention, visual reasoning.

I. INTRODUCTION

WING to intensive research in computer vision (CV) and natural language processing (NLP), related multimodal learning tasks, such as automatic image annotation [1], video question answering [2], cross-modal information retrieval [3], and visual question answering (VQA) [4], [5], have attracted great interest. Medical visual question answering (MVQA) is a socially significant application of VQA and one of the current research hotspots in computer-aided diagnosis (CAD) technology. A mature MVQA system can relieve the burden on medical staff, provide them with valuable second opinions on medical images, and reduce the risk of misdiagnosis [6].

MVQA requires us to answer relevant natural language (NL) questions in combination with given medical images. The complete modeling process can be described as follows. First, features are extracted from a given image and question. Second, the content of these features are understood and fused. Finally, the fused vectors are used to formulate the possible answers. Some research work has shown that because of the limitations of specialist medical concepts, challenges remain in understanding clinical texts [7], [8]. Thus, text features that are rich in medical information must be extracted, but most studies have ignored this need. At the same time, previous research on feature fusion is imperfect, which hinders the model’s ability to perform correct visual reasoning and select core regions and keywords related to the answer. Yu et al. proved through experiments that some dense interaction fusion schemes have low scalability [9], and that schemes based on multi-head attention methods are rarely applied to MVQA [6].

Word embedding, a technique for representing words as real-valued vectors, is commonly used in text processing. Most of the text extraction work in VQA and MVQA apply this technique to convert words into question features, thereby obtaining keyword information. The bi-branch model proposed by Liu et al. uses the sum of token, position, and segment embedding as the question feature representation [10]. However, many words have completely different public and medical meanings (e.g., "patient" and "mass"), and the multiplicity of meanings of medical terminology, the incongruity of abbreviations, and other similar factors can hinder the efficacy of word embedding [11]. Therefore, relying on words alone is not enough to gather sufficient medical information. Question-Centric Multimodal Low-rank Bilinear (QCMLB) [12] uses skip-thought vectors [13] to directly extract the sentence semantics of the question, which can obtain medical expertise but lose keyword information. As a result, the correct...
keywords and core areas may not be highlighted in the subsequent feature understanding process.

Multi-Modal Relation Attention Network (MRA-Net) obtains keywords information and semantic relations between words through word attention and self-guided relational attention, respectively. Then, the sum of the outputs of these two attention mechanisms is used as a question feature to answer complex queries [14]. Yang et al. added the question-type information after fusing multimodal features to narrow the candidate answer space [15]. The aforementioned methods can obtain more valuable information from the text. However, because of the single question type and simple relationship between words in clinical NL questions, using these schemes does not sufficiently help in extracting medical information from text. Zhan et al. designed Question-Conditioned Reasoning (QCR) and Type-Conditioned Reasoning (TCR) modules to extract closed-ended and open-ended question features, respectively. Gupta et al. generated word embedding and subword embedding, which were used with the integer sequence of the question to obtain text features [16]. Although these schemes enrich text representation, they are more focused on the use of word information and still fall short in the semantic representation of medical questions.

Attention mechanism [17] is not only successfully applied to unimodal tasks, but also plays an important role in many multimodal tasks like text matching [18] and visual captioning [19]. Recent studies have shown that learning co-attention for visual and textual elements at the same time can lead to a fine-grained understanding of the image and question, thereby improving the visual reasoning ability of the model and enabling more accurate predictions [20]. Most works on co-attention focusing on MVQA are based on VQA, and each of these studies have flaws. Bilinear attention networks (BAN) achieved the purpose of simultaneously learning two modal attention distributions by adding an attention matrix to the bilinear model [21]. Dense co-attention network (DCN) integrated visual and textual features by repeatedly interacting attention weights [22]. Although BAN and DCN can achieve dense interaction of modalities, adding depth provides minimal improvement in performance. Of course, such schemes are beneficial to MVQA at present because the dataset is quite small, but as the amount of data increases, the effect and scalability of the model is affected.

Modular Co-Attention Network (MCAN) proposed by Yu et al. improves the models understanding of image content through the synergy of question-guided attention and self-attention [9]. The Multimodal Encoder-Decoder Attention Networks (MEDAN) used self-attention in the encoder to model question features and employed question-guided attention and self-attention in the decoder to model image information, thus capturing comprehensive question and image features [23]. It is noteworthy that MEDAN and MCAN share similarities in their approach as they both emphasize the influence of the question on the image, while neglecting the impact of the image on the question. Moreover, they have not yet been applied to MVQA. CGMVQA [24] and Multimodal Medical BERT (MBERT) [25], consider both visual and textual inputs as tokens and leverage the BERT structure to enhance the feature information. Nevertheless, they treat all inputs as sequences, which are encoded and modeled using self-attention mechanism. These models may face challenges in capturing the intricate interaction between the modalities.

In this study, we propose a Dual-Attention Learning Network with Word and Sentence Embedding (DALNet-WSE) for MVQA to address the issues of inadequate extraction of medical information and poor granularity of feature understanding. Specifically, we design a module called transformer with sentence embedding (TSE) to extract question features. We used word and sentence embeddings to obtain keywords and medical information for the questions, respectively. A double embedding representation of the question is obtained through TSE. Then, we propose a dual-attention learning (DAL) module to learn self-attention and guided attention. The fusion component consisting of several DAL modules enhances the understanding of features and improves visual reasoning by learning co-attention between vision and text.

The present study’s contributions are summarized as follows:

- A DALNet-WSE is proposed for MVQA, which mainly solves the problems where the extracted question features lack medical information and does not completely understand the features;
- The TSE and DAL modules are designed. TSE extracts question features containing keywords and medical information. DAL models intramodal and intermodal interactions by learning self-attention and guided attention;
- Our DALNet-WSE achieves better performance on the VQA-MED 2019 and VQA-RAD datasets. In addition, ablation experiments are performed to verify the effectiveness of the TSE module, and with the help of Grad-CAM to determine whether DALs can improve feature understanding and enhance visual reasoning.

The rest of this paper is structured as follows. Section II briefly reviews related studies. Section III presents the entire modeling process and the various details of DALNet-WSE. Section IV describes the experiments and presents an analysis of the results. In section V the discussion and conclusion are provided.

II. Related Work

Since the initial hosting of the visual question answering challenge in the medical domain by ImageCLEF in 2018 [26], an increasing number of researchers have participated and explored various methods for MVQA. Overall, the strategies used for MVQA and VQA tasks are similar, and both methods include four core learnable modules: (i) a question encoder for extracting textual features, (ii) an image encoder for extracting visual features, (iii) a fusion algorithm for combining the extracted modal features, and (iv) a classifier that selects the optimal answer among a group of candidates.

A. Question Encoder

The purpose of this device is to encode the input questions, and Recurrent Neural Network (RNN) is often preferred for
VQA question encoder because of the excellent performance of RNN in NLP. However, it cannot obtain the complete context because the auto-regressive structure is limited. The pre-trained Bidirectional Encoder Representations from Transformers (BERT) model proposed by Google adjusts the word vectors according to the context and obtains a context-rich word representation [27]. Lee et al. devised Bidirectional Encoder Representations from Transformers (BioBERT) by adapting the architecture and hyperparameters of BERT and fine-tuning it on biomedical text corpora. BioBERT has achieved exceptional performance on three major biomedical text mining tasks, including named entity recognition, relation extraction, and question answering [28]. Rasmy et al. introduced Med-BERT, a domain-specific language model for electronic health records (EHRs) that incorporates code embeddings, serialization embeddings, and visit embeddings as input features, and demonstrated its effectiveness in predicting diseases with remarkable accuracy [29]. Although differences exist between the corpora in the general and medical domains, many MVQA studies have directly adopted pre-trained BERT to extract words as question features, such as [24] and [25].

In recent years, methods to create sentence embedding have led to a breakthrough in textual data representation. Ryan et al. proposed skip-thought vectors to predict the context of the target sentence through an encoder-decoder composed of RNNs [13]. It can obtain vector representations and learn continuity relations between sentences but is less computationally efficient. Quick-thought vectors proposed by Lajanugen et al. convert the predictive behavior of skip-thought vectors into categorical behavior, thereby improving sentence representation and computational efficiency [30]. Based on BERT, Sentence BERT combines the capabilities of the Siamese and triplet network to create high-quality sentence representations [31]. These methods are rarely used in MVQA.

### B. Image Encoder

The image encoder is suitable for extracting visual features from image inputs, with pre-trained CNNs being the most common in VQA. The widely accepted CNNs are trained and tested on the ImageNet dataset, and researchers select suitable image encoders based on their performance. Visual Geometry Group Networks (VGGNet) [32] and Residual Networks (ResNet) [33] have been successfully applied to VQA and MVQA tasks. The main reason is that the features they extract are more general and effective for datasets other than ImageNet. Mixture of Enhanced Visual Features (MEVF) proposed by Nguyen et al. used model-agnostic meta-learning (MAML) and denoising convolutional autoencoder (DCAE) to overcome the limitations of insufficient medical data and enhance visual features, which are extracted as image coding [34].

It is evident that models such as VGGNet and ResNet, which extract features specific to the entire image, are more suited for image classification. As the research in this field progresses, VQA tasks are becoming increasingly intricate, necessitating richer and more diverse natural image features. Consequently, a more intricate architecture for the VQA image encoder is imperative. Instance segmentation can achieve a more robust representation of natural image features by detecting object masks at the pixel-level. Image detection models are increasingly being preferred as the image encoder in VQA. The Faster Region-based Convolutional Network (Faster R-CNN), for example, uses VGGNet, ResNet or FPN [35] as its backbone architecture and merges it with a head architecture to achieve instance segmentation [36]. However, owing to the lack of a large-scale detection dataset in MVQA at present, most of the existing studies on image encoders are similar to the CNN classification model described above.

### C. Fusion Algorithm

The fusion phase involves modeling the correlation between the extracted visual and textual features. This phase mainly includes the attention mechanism and multi-modal pooling.

1) **Attention Mechanism**: This mechanism has been widely used in the feature fusion phase of VQA and MVQA to help the models understand the visual content of images well. For example, Du et al. used visual attention to understand visual features and find core image regions associated with question words for visual-text correlation learning [37]. However, in VQA, the model also needs to understand NL questions, so the model is required to have both visual and textual learning capabilities.

Lu et al. constructed a co-attention learning framework called HieCoAtt, which alternates learning visual and textual attention in a hierarchical manner [38]. Nam et al. proposed a combined framework for visual and textual attention learning, which draws on multiple steps to learn keywords and core regions [39]. However, these schemes only learn separate attention distributions for each modality, ignoring the intensive interaction between each word and object. Liu et al. proposed Cross-Attentional Spatio-Temporal Semantic Graph Networks (CASSG), a multi-headed, multi-hop attention model with diversity and progressivity, to explore fine-grained interactions among different modalities in an intersectional manner [40]. Learning Cross-Modality Encoder Representations from Transformers (LXMERT) [41] and Vision-and-Language BERT (ViLBert) [42] learned about image and text co-attention with two different Transformer components, but no one has introduced these ideas into MVQA.

2) **Multi-Modal Pooling**: The fusion scheme of multimodal pooling is another common technique in VQA. The fundamental concept is to perform separate pooling operations on the features extracted from the question and the image modalities, and subsequently integrate them to form a global feature vector. The integration operations commonly used are concatenation, summation, and element-wise product.

### III. Method

As with existing visual question answering methods, about the given medical image, MVQA aims to predict the most reasonable and likely answer, \( \hat{a} \), to a question, \( M_Q \). The task can be formulated as:

\[
\hat{a} = \arg \max_{a \in A} P(a|M_V, M_Q, \theta) \tag{1}
\]
where \( \mathcal{A} \) is the set containing all candidate answers, and \( \theta \) represents all the parameters of the model.

The proposed DALNet-WSE model framework is shown in Fig. 1. Specifically, we use the TSE module and pre-trained ResNet-152 to extract question features, \( Q \), and image features, \( V \), respectively. Through the fusion components, the understanding of the features is deepened and the visual reasoning ability is improved. The fusion vectors are then fed into a classifier that outputs the probability distribution of the \( N = |\mathcal{A}| \) answers and obtains the predicted results. Next, we describe each step of the framework in detail.

\[ \theta = \text{Softmax}(\frac{Q^T V}{\sqrt{d}}) \]

Fig. 1. The proposed DALNet-WSE framework. The ResNet-152 is used to extract image features, \( V \). TSE is used to extract the question features, \( Q \). The \( L \) DALs form the fusion module, and the fusion vector is obtained with the help of a weighted combination. The classifier consists of \( \text{MLP} \) and \( \text{Softmax} \).

### A. Image and Question Representations

1) **Image Representation:** Considering the limitations of the dataset and the goal to maximize the role of DAL in feature understanding, we choose the pre-trained ResNet-152 as the image encoder. As the medical images are extremely complex, similar to the CGMVQA approach, our proposed method extracts image features from five convolutional blocks of ResNet-152 to function as a detection model. On the one hand, it can make full use of image information, and on the other hand, it can help the model learn dual-attention better. We begin by using the convolutions with the same number of output channels to unify the number of feature maps output at each block. We then use global average pooling to obtain a vector representation of the features at each level of the image. The features of the given medical image \( M \) can be defined \( \hat{V} = (v_1; v_2; \cdots; v_5) \in \mathbb{R}^{5 \times d} \). Here, \( v_1, v_2, \cdots, v_5 \) denote the features extracted from various blocks of ResNet-152, which embody the semantic features of different levels of image \( M \).

2) **Question Representation:** The word information can ensure that the model selects the correct keywords for visual reasoning to predict the answer during feature understanding. However, using only word representations causes difficulty in understanding the given medical question correctly. Thus, we make up for this deficiency by learning the overall semantics of medical texts. We design the TSE module to obtain a double embedding representation of the question, ensuring that the extracted features are rich in both keywords and medical information.

For the given question \( MQ \), the word representation is formulated as follows. First, we use WordPiece [43] to mark it as several words, then project it to the embedding layer to obtain the word representation, \( \tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_m \), where \( \tilde{q} \in \mathbb{R}^{d_q} \), \( d_q \) denotes the question feature dimension. Here \( n \) denotes the uniform length of the question text after padding. The sentence representation is created by directly extracting the sentence embedding, \( S = \text{SBERT}(MQ) \in \mathbb{R}^{d_s} \), for \( MQ \) using pre-trained Sentence BERT.

The attention mechanism can effectively fuse different levels of information. It is formulated as querying a dictionary with key-value pairs and can be reconstructed based on the similarity of the elements [17]. The input comprises queries and keys, each having a dimension of \( d_{key} \), along with values that have a dimension of \( d_{value} \). For simplicity, we specify \( d_{key} = d_{value} = d \). Calculated as in Equation (2):

\[ \text{Attention}(Q_a, K_a, V_a) = \text{Softmax}(\frac{Q_a K_a^T}{\sqrt{d}}) V_a \]

where \( Q_a, K_a, \) and \( V_a \) are matrices created by packaging queries, keys, and values. Inspired by TUPE (Transformer with Unified Positional Encoding) [44], we add sentence information to the relational modeling of each word pair to reduce the loss of semantic information. Specifically, we combine words and sentences embedding in the following way:

\[ a_{ij} = \frac{1}{\sqrt{2d}} (\tilde{q}_i W^Q) (\tilde{q}_j W^K)^T + \frac{1}{\sqrt{2d}} (SU^Q)(SU^K)^T \]

\[ q_i = \sum_{j=1}^{n} \frac{\exp(a_{ij})}{\sum_{j'=1}^{n} \exp(a_{ij'})} (\tilde{q}_j W^V) \]

where \( W^Q, W^K, W^V \in \mathbb{R}^{d_q \times d_q} \) denote the learnable project matrices of \( \tilde{q} \). \( U^Q, U^K \in \mathbb{R}^{d_s \times d_q} \) represent the learning project matrices of \( S \). Then, we obtain a double embedding representation of the question, \( Q = (q_1; q_2; \cdots; q_n) = \text{TSE}(MQ) \in \mathbb{R}^{n \times d_q} \), as shown in Fig. 2.

### B. Dual-Attention Learning

Before describing feature fusion, we introduce its core component, DAL, which has an encoder-decoder architecture consisting mainly of self-attention and guided attention. DAL can learn simultaneously image- and question-guided attention.
1) Self-Attention and Guided Attention: Multi-head attention based on the Equation (2) can provide information on the encoding of different subspaces and enhance the expressiveness of the model, which can be defined as:

$$MA(Q_a, K_a, V_a) = [\text{head}_1; \text{head}_2; \ldots; \text{head}_d]W^o$$ \hspace{1cm} (5)

$$\text{head}_j = \text{Attention}(Q_aW_j^Q, K_aW_j^K, V_aW_j^V)$$ \hspace{1cm} (6)

where $W_j^Q \in \mathbb{R}^{d \times d_i}$, $W_j^K \in \mathbb{R}^{d \times d_i}$, $W_j^V \in \mathbb{R}^{d \times d_i}$, $W^o \in \mathbb{R}^{hd \times d}$ are learning matrices, $d$ denotes the dimension of the hidden layer of the model.

We build self-attention and guided attention based on the multi-head attention mechanism. Using the image and question encoder, we can extract the corresponding features from the input data and obtain image feature $V$ and question feature $Q$. The self-attention learning for element $v_i$ can be defined as $f_i = MA(v_i, V, V)$, and reconstructed from the normalized similarity of $v_i$ and all samples in $V$. For the feature $V$, the self-attention learning can be expressed as:

$$F_{(V, V)} = MA(V, V, V).$$ \hspace{1cm} (7)

The guided attention learning for element $v_i$ can be expressed as $g_i = MA(v_i, Q, Q)$, which is reconstructed by cross-modal similarity between $v_i$ and all samples in $Q$. The guided attention learning by $Q$ on $V$ is $F_{(V, Q)}$:

$$F_{(V, Q)} = MA(V, Q, Q).$$ \hspace{1cm} (8)

We simulate intensive intramodal and intermodal interactions by learning self-attention and guided attention.

2) DAL for MVOA: The DAL module is shown in Fig. 3, we found that our DAL presents a bi-directional encoder-decoder structure. For example, on the left half of Fig. 3, this part can be considered as an encoder when learning image-guided attention and as a decoder when learning question-guided attention. Both the encoder and decoder in DAL have two layers of multi-head attention that learn self-attention and guided attention. We use the first part of multiple attention for self-attention learning and the second part for guided attention learning. Take visual learning as an example. First, the input image features, $V$, are subjected to self-attention learning of Equation (7), which models the interaction relationship between each pair of hierarchical features $\{v_i, v_j\} \in V$ and reconstructs $V$ based on the similarity between semantic features. Later, after guided attention learning of Equation (8), the interaction relationship between each $v_i \in V$ and $q_j \in Q$ is modeled, and $V$ is reconstructed again based on the similarity between the two modalities.

C. DALs for Feature Understanding and Fusion

Modal information is initially fused during feature interaction and reconstruction. However, individual DAL plays a limited role, so we cascaded the DALs in depth. We learn co-attention between vision and text through the alternation of self-attention and guided attention in DALs, thus enhancing feature understanding while further facilitating the fusion of two forms of modal information. Specifically, the image features, $V$, and question features, $Q$, which are extracted by the image encoder and question encoder, are input to the DALs, denoted as DAL(1), DAL(2), …, DAL(L), where $L$ is the number of DAL; $V^{(l-1)}$ and $Q^{(l-1)}$ are the input of DAL(l), the output features are represented by $V^{(l)}, Q^{(l)}$, which are further fed into the next DAL, namely, DAL(l+1). The recursive process can be described as:

$$[V^{(l)}, Q^{(l)}] = \text{DAL}^{(l)}([V^{(l-1)}, Q^{(l-1)})].$$ \hspace{1cm} (9)

In particular, we take $[V, Q]$ as the input feature of DAL(1), i.e., $[V^{(0)}, Q^{(0)}] = [V, Q]$. After DALs, $V^{(L)} = (v^{(L)}_1; \ldots; v^{(L)}_s) \in \mathbb{R}^{s \times d}$ and $Q^{(L)} = (q^{(L)}_1; \ldots; q^{(L)}_n) \in \mathbb{R}^{n \times d}$ can be represented as the output image and question features. Then, we perform a simple weighted sum of $V^{(L)}$ and $Q^{(L)}$ as in Equation (10):

$$Z = W^V v^{(L)} + W^Q q^{(L)}$$ \hspace{1cm} (10)

where $W^V \in \mathbb{R}^{1 \times s}$, $W^Q \in \mathbb{R}^{1 \times n}$. $Z \in \mathbb{R}^{1 \times d}$ is the ultimate fusion feature, which is used for the prediction of the model.
D. Classifier

Similar to previous studies [6], our proposed model views MVQA as a multi-classification task and does not generate text. Specifically, we pass $Z$ through the classifier, which is composed of a MLP (FC($d$) − GELU − FC($N$)) and Softmax activation, to obtain the probability scores for each potential answer. Here, FC($d$) denotes a fully connected layer with an output dimension of $d$. We then select the answer with the highest probability as the predicted result. The operational procedure can be summarized as follows:

$$\hat{a} = \text{argmax}(\text{Softmax}(\text{MLP}(Z))) \quad (11)$$

where $\hat{a}$ is the predicted result of the model.

IV. EXPERIMENTS AND RESULTS

In this section, we evaluate the performance of our proposed DALNet-WSE on the datasets VQA-MED 2019 and VQA-RAD in detail. In addition, we conduct ablation experiments to validate the effectiveness of the TSE and DALs, and use Gradient-weighted Class Activation Mapping (Grad-CAM) [45] for visual analysis to explore the impact of using different guided attention on visual reasoning.

A. Evaluation Metrics

Accuracy (ACC) and Bilingual Evaluation Understudy (BLEU) score are the most commonly used evaluation metrics in the MVQA task. ACC is employed to assess answer accuracy, while BLEU is used to measure the similarity between predicted and ground truth answers. However, BLEU similarity assessment may not adequately cover biomedical knowledge. To address this, Word-Based Semantic similarity (WBSS) evaluation was added. WBSS is a metric based on the Wu-Palmer similarity [46] and was developed using a semantic similarity algorithm proposed by Gizem et al. specifically for the biomedical field [47]. Combining with Fig. 4, the Wu-Palmer similarity of $C_1$ and $C_2$ is calculated as follows:

$$\text{Sim}_{WP}(C_1, C_2) = \frac{2 \cdot \text{depth}(C_3)}{\text{depth}(C_1) + \text{depth}(C_2)} \quad (12)$$

where depth$(C_1)$, depth$(C_2)$, and depth$(C_3)$ indicate the depth of $C_1$, $C_2$, and $C_3$ nodes, respectively. $C_3$ is the least common superconcept of $C_1$ and $C_2$. The higher the WBSS, the higher the semantic similarity between the predicted result and the standard answer.

It is worth noting that in our study on MVQA, we approach it as a multi-classification task. Hence, BLEU score and WBSS are computed based on class names. Specifically, we consider the predicted and true classes as sentences, and then calculate their BLEU or WBSS to measure the similarity between the predicted and reference answers.

B. Datasets

VQA-RAD [48] is an MVQA dataset on radiology, which has 315 medical images and 3515 question-answer (QA) pairs of 11 question types. They are: Modality, Plane, Organ (Organ System), ABN (Abnormality), Pres (Object/Condition Presence), Pos (Positional Reasoning), Color, Size, Attrib (Attribute Other), Count (Counting), Other. Then, 58% of the questions in VQA-RAD are closed-ended and the rest are open-ended. We train and test these two categories of data separately. VQA-RAD is manually labeled and the trained model has a higher confidence level. In this experiment, we employed the same training sets as previous studies [7], consisting of 3064 data samples for training and 451 samples for testing.

VQA-MED 2019 [49] is presented in the ImageCLEF 2019 challenge. Inspired by VQA-RAD, all questions in VQA-MED 2019 follow the patterns naturally proposed and validated in VQA-RAD. The VQA-MED 2019 dataset covers the four most common categories in medical diagnosis: modality, plane, organ system, and abnormality. Of these, abnormality is open-ended, and the remaining categories are closed-ended questions. During our data analysis, we found an imbalance in class distribution across all categories. For instance, in the training set of the plane category, the single ‘axial’ class accounted for 48.7%. Additionally, we observed some “yes/no” types of questions in the modality and abnormality categories. To ensure comparability of experimental results and accommodate this sort of data, we created a new category called Yes/No. Incidentally, both the VQA-MED 2019 and VQA-RAD test results contain the overall category and are calculated according to the following rules:

$$\text{Overall} = \sum_{i=1}^{M} \frac{C_i \cdot D_i}{T} \quad (13)$$

where $C_i$, $D_i$ represents the result (ACC, BLEU or WBSS) and data volume of the $i$-th category, respectively. $M$ denotes the number of categories, $M = 2$ in VQA-RAD and $M = 5$ in VQA-MED 2019. $T$ represents the amount of data in the entire test set. Notably, DALNet-WSE is considered a classification model in both datasets, regardless of whether the questions are closed- or open-ended.

Pre-training not only speeds up model convergence but also has the potential to achieve better results. To obtain the best performance DALNet-WSE, we pre-train it on the Radiology Objects in Context (ROCO) dataset [50]. In the pre-training phase, we use all images and their corresponding captions and keywords for the Mask Language Model (MLM) [51] based on DALNet-WSE.
C. Implementation Details and Training

In this study, we consider MVQA as a multi-classification task and therefore employ the widely-used cross-entropy loss as the training objective. To further enhance the model’s performance and ensure its strong generalization ability, we also incorporate label smoothing techniques:

\[
Loss = H(q', p) = -\sum_{k=1}^{K} q'(k) \log p(k) = (1 - \epsilon)H(q, p) + \epsilon H(u, p) \tag{14}
\]

where \(p(k)\) means the probability that our model calculates each label \(k \in [1, \ldots, K]\), and \(q'(k) = (1 - \epsilon)q(k) + \epsilon u(k)\), \(u(k)\) is a fixed distribution. During the experiment, we set \(u(k) = 1/K\) as a uniform distribution with respect to \(K\). \(H(q, p)\) and \(H(u, p)\) represent the cross-entropy loss when the true distributions are \(q(k)\) and \(u(k)\), respectively. \(\epsilon\) represents the smoothing coefficient.

In the pre-trained MLM, the task is to predict tokens that are masked. Unlike unimodal interaction, in this task, we not only use the unmasked text but also incorporate visual information. To ensure that the model learns medical knowledge based on medical images and NL questions, we masked only the medical keywords provided in the original dataset. Specifically, we randomly masked 15% of the medical keywords and used the image encoder of DALNet-WSE to extract image features and TSE to extract features from text with masked keywords. The question information, obtained through the co-attention learning of DALs, is then fed into a new classifier that comprises MLP (FC(\(d – GelU – FC(d_{vocab})\)) and Softmax to produce prediction results. Here, \(d_{vocab}\) represents the size of the vocabulary, and the cross-entropy loss is utilized as the objective function during pre-training. This approach ensures that the model can effectively integrate both modalities to predict the masked tokens.

The hyperparameters of DALNet-WSE throughout the experiment are set as follows. The image feature dimension, \(d_i\), question feature dimension, \(d_q\), and hidden layer dimension, \(d\), are equal to 312 (i.e., \(d_i = d_q = d = 312\)), the number of heads in the multi-head attention layer \(h = 12\), the dimension of each head \(d_h = d/h = 26\), the uniform length of the padded text \(n = 20\), and the vocabulary size \(d_{vocab} = 30522\). During the experiment, the number of DAL, \(L = 2\), and the smoothing parameter \(\epsilon = 0.10\) in the loss function Equation (14). In the pre-training stage, \(dropout = 0.0\), and during fine-tuning, \(dropout = 0.1\).

We train our model on a single NVIDIA RTX 3080 GPU. We reshape all images to size \(244 \times 244\) and perform augmentation, including random rotations within \(\pm 10^\circ\), and color jitter, to enrich usable image information. In pre-training, we use the Adam optimizer to optimize the model loss with a learning rate of \(2e-5\), batch size = 16, and iteration of 10 epochs. In fine-tuning, we adopt the Adam optimizer with a learning rate of \(1e-4\), batch size = 16, and iteration of 100 epochs to optimize the model loss. In VQA-MED 2019 (or VQA-RAD), if the loss on the validation set (or training set) does not improve within 10 consecutive epochs, the learning rate is reduced by a factor of 0.1. To avoid the influence of chance, we select the ACC, BLEU score, and WBSS that appear frequently in the test results. Moreover, it is worth noting that VQA-RAD dataset lacks a dedicated validation set [48]. Therefore, to further evaluate the effectiveness of our model, we also performed 10-fold cross-validation on the training data. Specifically, we trained the model on a training subset, fine-tuned it on the validation subset, and eventually evaluated the final model on the test set. The core code is available at https://github.com/Coisini-Glenda/DALNet-WSE.

D. Ablation Studies

We design ablation experiments to verify (a) whether the TSE module can effectively extract medical information from text, and (b) whether the proposed DALs can effectively fuse multimodal information, enhance feature understanding and improve visual reasoning by simultaneously learning image- and question-guided attention. We set the following states: (1) DALNet-SE(NP) denotes that the question feature is solely derived from the sentence information of the text, without pre-training, and uses two types of guided attention. (2) DALNet-WE(NP) has no pre-training and only applies word information as question features using both question- and image-guided attention model. (3) DALNet-BioBERT(NP) refers to the use of pre-trained BioBERT to extract text information, using two types of guided attention without pre-training. (4) DALNet-WSE(NP) has no pre-training and applies word and sentence information as question features using two types of guided attention. (5) Concat-WSE(NP) uses both word and sentence information, replaces DALs with concatenation for modal information fusion, and does not perform pre-training. (6) DALNet-WSE(NP) only with \(F_{(V, Q)}\) (abbreviated as \(F_{(Q, V)}\)) represents a non pre-trained model that uses double embedding of words and sentences, but only uses image-guided attention. (7) DALNet-WSE(NP) only with \(F_{(V, Q)}\) (abbreviated as \(F_{(V, Q)}\)) has no pre-trained model that uses double embedding but only uses question-guided attention. (8) DALNet-WSE(P) is a pre-trained model that uses double embedding and two types of guided attention.

1) Validation of TSE: Table I and Table II depict in detail the performance of DALNet-WSE in different states on the VQA-MED 2019 and VQA-RAD test sets, respectively. DALNet-WSE(NP) is model with sentence semantics added to DALNet-WE(NP) or with word information added to DALNet-SE(NP). Table I shows the following results. Compared with DALNet-WE(NP), DALNet-WSE(NP) outperformed DALNet-WE(NP) significantly in all categories and exceeded it by approximately 6 percentage points in abnormality. In comparison, DALNet-SE(NP) only achieved slightly better results than DALNet-WSE(NP) in the yes/no category, while DALNet-WSE(NP) outperformed DALNet-SE(NP) by a considerable margin in all other categories. When BioBERT was used as the question encoder instead of TSE, the model exhibited performance similar to DALNet-WSE(NP) in the organ and abnormality categories. However, its performance in the plane and yes/no categories was notably weaker than DALNet-WSE(NP).

As reported in Table II, DALNet-WSE(NP) has higher accuracy on the VQA-RAD test set compared to DALNet-WE(NP),
improving to 0.637 on the open-ended and 0.827 on the closed-ended questions. Moreover, DALNet-WSE(NP) outperformed DALNet-SE(NP) in all evaluation metrics, with an approximate 2 percentage point difference in overall performance. Although DALNet-SE(NP) showed slightly better performance than DALNet-WE(NP), it still fell short of DALNet-WSE(NP) in all evaluation metrics. As for DALNet-BioBERT(NP), its overall performance is not substantially different from that of DALNet-WSE(NP), but DALNet-WSE(NP) performs more prominently on open-ended questions. In summary, the performance of the model is significantly improved by effectively integrating the information of words and sentences. This result shows that the TSE module can be used to encode the input questions to obtain medical information-rich question features.

2) Validation of DALs: To assess the effectiveness of the cross attention mechanism in integrating multimodal data, we replaced DALs with concatenation. As shown in Table I and Table II, regardless of the type of cross attention (states (4)(6)(7)) used, the performance of the model surpasses that of Concat-WSE(NP), in all evaluation metrics. As for DALNet-BioBERT(NP), its overall performance is not substantially different from that of DALNet-WSE(NP), but DALNet-WSE(NP) performs more prominently on open-ended questions. In summary, the performance of the model is significantly improved by effectively integrating the information of words and sentences. This result shows that the TSE module can be used to encode the input questions to obtain medical information-rich question features.

E. Qualitative Analysis

We also use an additional evaluation matrix to assess the classification effect of DALNet-WSE(P). We visualized the confusion matrices for the four categories in VQA-MED 2019: modality, plane, organ, and yes/no. Furthermore, we amalgamate similar candidate responses across these categories, such as “ct with gi and iv contrast” with “gi and iv,” and “mr-flair” with “flair”. However, we did not visualize and analyze the confusion matrices for abnormality in VQA-MED 2019 and for the open-ended and closed-ended categories in VQA-RAD, due to the significantly large number of candidate answers for these categories and some classes are severely under-represented. Displaying their confusion matrix can lead to a complex and confusing representation, making the evaluation challenging to interpret.
The confusion matrix visualization of the modality category is shown in Fig. 5. Some types can be accurately predicted, such as “xr-plain film,” “us-ultrasound,” “t2,” and “noncontrast”. However, none of the “sbft-small bowel,” and “noncontrast(mri)” types were correctly predicted. The model’s performance in the remaining types needs improvement, as the accuracy of predictions ranges between 50% and 75%.

A confusion matrix visualization of the plane category is shown in Fig. 6. In this category, no type can be predicted 100% of the time. But the proportion of incorrect predictions for the “axial,” “coronal,” and “sagittal” types is relatively small and can be predicted with greater accuracy. The performance on the other types is poorer.

Fig. 7 shows the confusion matrix visualization of the organ category. Among them, only the “genitourinary,” type can be completely predicted, and most of them are correctly predicted with a high probability, such as “skull and contents,” “lung, mediastinum, pleura,” and “musculoskeletal”. However, the model’s performance seems to fall short in the “heart and great vessels” and “breast” types. We also visualize the confusion matrix for the yes/no category, as shown in Fig. 8. From this, we can conclude that the accuracy, recall, and f1 score of the model on the yes/no category are 0.891.

According to Table I, it is apparent that DALNet-WSE exhibits varying performance across different categories, excelling in some categories while underperforming in others. This indicates that the model’s performance is influenced by the specific characteristics of the category data. Therefore, we hypothesize that the phenomena described in the confusion matrix may be due to class imbalance in the data. We analyze the distribution of these classes in the training set using the example of plane category. As shown in Fig. 9, it is evident that there is a class imbalance issue. Combining the corresponding confusion matrix with the test results, it can be observed that the model can accurately predict some answers with a large proportion, such as “axial”, “coronal”, “sagittal”, etc. However, some answers with a small proportion are difficult to predict correctly or cannot be predicted at all. Moreover, in some incorrect predictions, the predicted results are also mainly based on answers with a larger proportion. This indicates that there may be a positive correlation between the proportion of candidate answers in the training set and the likelihood of predicting these answers in the test results. In other words, the larger the proportion, the higher the likelihood of predicting this answer, partially validating our hypothesis.
Despite the findings from our ablation experiments that DALs can successfully integrate modal information, it is yet uncertain if the implementation of DALs can enhance the model’s reasoning capability. Grad-CAM can provide a localization map that highlights important regions of the image used for prediction and provides a visual interpretation of the model’s decision-making [45]. In order to gain further insight into the model’s decision-making behavior, we used Grad-CAM to aid in verifying if the use of DALs can improve the model’s visual reasoning ability, and to investigate the impact of guided attention techniques on visual reasoning. We randomly selected four examples from two datasets and depicted Grad-CAM maps for different states of DALNet-WSE, see Fig. 10. On the far left are the selected QA pairs, the first column shows the input medical image, and the other columns show the activation feature maps visualized by Grad-CAM to illustrate the focus of DALNet-WSE in the different states for the given examples. Specifically, the second column shows the Grad-CAM maps for DALNet-WSE(NP), the third column shows the Grad-CAM maps for DALNet-WSE(P) only with $F_{(Q, V)}$, and the fourth column shows the Grad-CAM maps for DALNet-WSE(NP) only with $F_{(Q, V)}$. The last column shows the Grad-CAM maps for Concat-WSE(NP).

The top two rows of Fig. 10 illustrate examples from VQA-MED 2019, both with closed-ended questions. In the first example, we can observe an area in the input image with some width and color variations, which are characteristic of doppler ultrasound imaging as it can display the speed and direction of blood flow, usually coded with color. When asked “What imaging method was used?”, all variants of DALNet-WSE can focus on these colored regions for sensible visual reasoning and accurately predict the answer “us-d - doppler ultrasound”. However, DALNet-WSE(NP) can focus on more detailed regions.

As shown in the second line of Fig. 10, the question is “Is this a contrast or noncontrast mri?”. The primary difference between contrast and noncontrast imaging lies in the level of image contrast and anatomical structure visualization (clarity and precision) [58]. A contrast-enhanced image displays higher contrast and sharper edges. To correctly answer such a question, the model may need to attend to all organizational objects. Our findings indicate that DALNet-WSE(NP) and DALNet-WSE(NP) only with $F_{(Q, V)}$ focus on most areas, resulting in a “noncontrast” answer. This could be due to the absence of pronounced bright spots and high contrast in the tissue of the focused area, leading to a “noncontrast” visual inference. While “noncontrast” is an acceptable answer for addressing this question, it is not the correct answer (“noncontrast (mri)”). However, Concat-WSE(NP) and DALNet-WSE(NP) only with $F_{(Q, V)}$ do not entirely focus on the relevant regions, increasing the likelihood of incorrect visual reasoning.

The last two rows of Fig. 10 depict Grad-CAM maps for two instances in VQA-RAD, all open-ended questions. In the example of the third row, we were asked to answer, “In which two ventricles can calcifications be seen on this CT scan?” and the correct answer is “the 3rd ventricle and the lateral ventricles”. To aid our explanation, we highlighted the 3rd ventricle (in red) and lateral ventricle (in blue) in the input image. It is evident that only DALNet-WSE(NP) focuses on these two regions, but unfortunately predicts only “lateral ventricles”. DALNet-WSE(NP) only with $F_{(Q, V)}$ might focus on some regions of the lateral ventricles and provide the corresponding prediction. However, DALNet-WSE(NP) only with $F_{(Q, V)}$ and Concat-WSE(NP) do not attend to the relevant regions, making accurate visualized reasoning a challenging task.

In the final example in the last row of Fig. 10, we are asked the question “What skeletal joint is seen in this image?” and the correct answer is “sacroiliac joint”. To aid in our explanation, we have highlighted the location of the sacroiliac joint in red. It is clear that only the DALNet-WSE(NP) model extensively focuses on the relevant region and accurately predicts the answer. Conversely, in other cases, the model fails to focus on the appropriate area and makes incorrect predictions.

It is important to note that the examples we randomly selected from two test sets may not fully represent the overall visual reasoning capability of the model. However, these examples do offer valuable insights into the effects of guided attention for visual reasoning. From these examples, it can be inferred that learning question- and image-guided attention simultaneously is a more promising approach compared to other feature understanding methods, as it facilitates reasonable visual reasoning and accurate prediction. It should be noted that, although the model that simultaneously learns both types of guided attention can correctly focus on the regions relevant to the answer in the selected examples, it does not always guarantee prediction accuracy, as demonstrated in the third example. This issue deserves further investigation in future research.

F. Comparison With State-of-the-Art

After conducting ablation studies, we compare the best single model DALNet-WSE(P) with several state-of-the-art methods currently on the VQA-MED 2019 and VQA-RAD datasets. The performance of DALNet-WSE(P) on
Fig. 10. Example of medical images and Grad-CAM maps from DALNet-WSE with different states on VQA-MED 2019 and VQA-RAD. The text on the left shows the pairs of question-answer ground truths used in each row and the predictions of DALNet-WSE (NP) (All), DALNet-WSE (NP) only with $F(Q,V)$, DALNet-WSE (NP) only with $F(V,Q)$, and Concat-WSE (NP) (Concat). Green and Red denote the correct and wrong predictions, respectively.

Table III

| Method          | Modality | Plane | Organ | Abnormality | Yes/No | Overall |
|-----------------|----------|-------|-------|-------------|--------|---------|
|                 | ACC      | BLEU  | WBSS  | ACC         | BLEU  | WBSS    |
| TUA1 [52]       | 0.657    | 0.759 | 0.788 | 0.716       | 0.816 | 0.617   |
| Up-Down [53]     | 0.606    | 0.871 | 0.873 | 0.824       | 0.834 | 0.866   |
| COMVQA [24]     | 0.405    | 0.856 | 0.870*| 0.808       | 0.813 | 0.814*  |
| COMVQA(Ens) [24]| 0.819    | 0.880 | 0.866 | 0.844       | 0.844 | 0.844   |
| MMBERT(NP) [23] | 0.806    | 0.856 | 0.865*| 0.816       | 0.816 | 0.818*  |
| MMBERT(Ens) [23]| 0.833    | 0.882 | 0.891*| 0.884       | 0.884 | 0.885*  |
| DALNet-WSE(NP)  | 0.833    | 0.871 | 0.883 | 0.832       | 0.832 | 0.836   |
| DALNet-WSE(P)   | 0.864    | 0.894 | 0.903 | 0.864       | 0.864 | 0.869   |

VQA-MED 2019 and VQA-RAD is shown in Table III, Table IV.

We describe the performance of DALNet-WSE(P) on VQA-MED 2019 in detail. We find that the performance on several categories is significantly better than that of the previous correlation model, especially in abnormality. And the overall category accuracy and BLEU scores both improved by 2.1 percentage points, reaching 0.693 and 0.711, respectively. Combining our re-implemented results, DALNet-WSE(P) achieved a WBSS improvement of at least 2 percentage points, reaching 0.729.

However, DALNet-WSE(P) performed the worst in the organ category of VQA-MED 2019. We attempted to guess the possible reasons, which is mainly an issue with the true label distribution of the dataset. We carefully analyzed the relevant data for the organ category and found that in the training and validation sets, “lung, mediastinum, pleura,” “skull and contents,” “genitourinary,” “musculoskeletal,” and others are the 10 types of answers. However, in the test set, another seven types include “gastrointestinal#lung, mediastinum, pleura,” “heart and great vessels#lung, mediastinum, pleura#spine and contents,” and others, which never appeared during training and accounted for 7.2% of the test set. In other words, the models had never been trained with these types of data, so were not aware of the visual and textual features corresponding to these answers. Thus, all the models performed poorly in the organ category. The result suggests that this dataset needs further refinement.

In VQA-RAD, the accuracy of DALNet-WSE(P) improved to 0.849 on the closed-ended questions and 0.774 in the overall category, outperforming all relevant models. The corresponding WBSS metric also outperformed all compared models. The accuracy on the open-ended question is 0.659, which is about 1 percentage point lower than BiLR’s 0.665.
TABLE IV

| Method               | Open-ended ACC | Open-ended WBSS | Closed-ended ACC | Closed-ended WBSS | Overall ACC | Overall WBSS |
|----------------------|----------------|----------------|-----------------|------------------|-------------|--------------|
| BiAN [54]            | 0.284          | 0.383*         | 0.679           | 0.687*           | 0.522       | 0.566*       |
| MAML(BAN) [55]       | 0.401          | 0.471*         | 0.724           | 0.735*           | 0.596       | 0.630*       |
| MEVF(BAN) [34]       | 0.439          | 0.495*         | 0.751           | 0.767*           | 0.627       | 0.659*       |
| MMQ [56]             | 0.537          | 0.594*         | 0.758           | 0.772*           | 0.670       | 0.701*       |
| CR [7]               | 0.600          | 0.651*         | 0.793           | 0.812*           | 0.716       | 0.748*       |
| MMBERT(P) [25]       | 0.631          | 0.687*         | 0.779           | 0.794*           | 0.720       | 0.752*       |
| CPRD+BAN+CR [57]     | 0.605          | 0.685*         | 0.804           | 0.817*           | 0.725       | 0.765*       |
| BiRL [18]            | **0.665**      | **0.698**      | **0.824**       | **0.837**        | **0.761**   | **0.782**    |
| DALNet-WSE(NP)       | 0.637          | 0.685          | 0.827           | 0.840            | 0.752       | 0.778        |
| DALNet-WSE(P)        | 0.659          | 0.695          | **0.849**       | **0.858**        | **0.774**   | **0.793**    |

The '*' indicates the results were reproduced based on the provided model.

TABLE V

| Method               | Open-ended ACC | Open-ended WBSS | Closed-ended ACC | Closed-ended WBSS | Overall ACC | Overall WBSS |
|----------------------|----------------|----------------|-----------------|------------------|-------------|--------------|
| BiAN [54]            | 0.402*         | 0.471*         | 0.711*          | 0.725*           | 0.588*      | 0.624*       |
| MAML(BAN) [55]       | 0.542*         | 0.606*         | 0.742*          | 0.755*           | 0.663*      | 0.696*       |
| MEVF(BAN) [34]       | 0.525*         | 0.589*         | 0.746*          | 0.761*           | 0.658*      | 0.693*       |
| MMQ [56]             | 0.536*         | 0.594*         | 0.757*          | 0.772*           | 0.669*      | 0.701*       |
| CR [7]               | 0.587*         | 0.658*         | 0.761*          | 0.780*           | 0.692*      | 0.732*       |
| MMBERT(P) [25]       | 0.592*         | 0.668*         | 0.746*          | 0.761*           | 0.685*      | 0.724*       |
| CPRD+BAN+CR [57]     | 0.598*         | 0.651*         | 0.768*          | 0.782*           | 0.701*      | 0.73*        |
| BiRL [18]            | 0.603*         | 0.672*         | 0.79*           | 0.805*           | 0.706*      | 0.752*       |
| DALNet-WSE(P)        | **0.620**      | **0.695**      | **0.805**       | **0.817**        | **0.732**   | **0.769**    |

The 10-fold cross-validation results of DALNet-WSE(P) and related models on the VQA-RAD dataset, including accuracy and WBSS. The '*' indicates the results were reproduced based on the provided model.

Perhaps the advantages of DALNet-WSE on VQA-RAD are not obvious, but while most of the existing work only considers the performance of the respective models on a single dataset, we consider two datasets and achieve better performance on both of them. This result suggests that our proposed DALNet-WSE is more general and effective, allowing for more robust visual inference.

To validate the performance of our model more effectively, we conducted 10-fold cross-validation on DALNet-WSE(P) and the related models using the VQA-RAD dataset. The specific results can be found in Table V. It can be found that DALNet-WSE(P) achieved the best performance in both closed-ended and open-ended tasks. The performance of DALNet-WSE(P) surpassed all comparative models, with an overall accuracy of 0.732 and WBSS of 0.769. Compared to Table IV, cross-validation provided a more comprehensive evaluation of DALNet-WSE and yielded more accurate and reliable assessment results.

V. DISCUSSION AND CONCLUSION

The images and texts given in the MVQA task are more difficult to understand than normal because of the knowledge associated with the medical field. Extracting medical information from the text and making a fine-grained understanding of the image and question features are key to improving model prediction performance. The medical images in the VQA-MED 2019 and VQA-RAD datasets cover almost all organs of the human body, and the corresponding natural language questions are the most common when medical images are read by physicians. In contrast to methods proposed in previous studies, DALNet-WSE is effective on both datasets.

With regard to extracting richer information from medical natural language questions, most previous works only considered information about words or word combinations. In contrast to these schemes, DALNet-WSE considers sentence semantics, aggregate words, and sentence information through TSE to extract text features that are rich in keywords.
and medical information. This idea may be applied to other related fields, such as biomedical text classification, which we believe could be highly interesting.

In current VQA research, both MCAN and MEDAN have demonstrated remarkable performance by employing dual-attention computations, but only question-guided attention was utilized. Specifically, they only considered the influence of text on images, disregarding the impact of images on text, similar to the dual-attention operation in DALNet-WSE only with $F(V, Q)$ in this paper. In contrast, our study examines the interplay between the question and the image simultaneously. We hypothesize that during the VQA task, not only can the question words search the image content, but also the images can filter the question words. We attempted to combine self-attention and different guided attention to fuse and understand modal information. The results demonstrate that using both image- and question-guided attention models performs the best. This confirms our assumption that the model understands both visual and text more effectively by considering the interaction of both modalities simultaneously. However, the classification model-based image encoder restricts the play of DAL, and we anticipate the need for an image detection model to function as an image encoder for MVQA, allowing us to fully leverage the use of guided attention by interacting image objects with question keywords.

In this study, we observed that DALNet-WSE was not effective in tackling open-ended questions. For example, the accuracy in abnormality is only 20%, which merits our attention. Consequently, we conducted a thorough analysis of the abnormality category data distribution and discovered that the training set contained 3082 data with 1483 unique classes. The highest percentage of classes represented only about 1% of the total, and over 85% of the classes account for less than 0.1%. These findings suggest that many classes cannot learn sufficient feature information for model prediction during training. Hence, we plan to explore more effective methods for handling open-ended questions in MVQA in future research.

In this paper, we proposed DALNet-WSE, a new model for MVQA tasks to efficiently extract medical information from natural language questions and explore the role relationship between vision and text. We designed a TSE module to add sentence information to each pair of word relationship models with the help of learnable projection matrices. This module ensures that the extracted question features are rich in keywords and medical information. The proposed DAL module effectively models intramodal and intermodal interactions through self- and guided attention learning. DAIs can improve the model’s fine-grained understanding of features and enhance visual reasoning by learning co-attention between vision and text. We conducted comprehensive experiments on VQA-MED 2019 and VQA-RAD datasets to confirm the effectiveness and generality of DALNet-WSE. We believe that with further research, our approach can perform better and facilitate the development of CAD.

REFERENCES

[1] C. Chaudhary, P. Goyal, D. N. Prasad, and Y. P. Chen, “Enhancing the quality of image tagging using a Visio-Textual knowledge base,” IEEE Trans. Multimedia, vol. 22, no. 4, pp. 897–911, Apr. 2020.
[2] J. Liang, L. Jiang, L. Cao, Y. Kalantidis, L.-J. Li, and A. G. Hauptmann, “Focal visual-text attention for Memex question answering,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 8, pp. 1893–1908, Aug. 2019.
[3] X. Huang, Y. Peng, and M. Yuan, “MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval,” IEEE Trans. Cybern., vol. 50, no. 3, pp. 1047–1059, Mar. 2020.
[4] S. Antol et al., “VQA: Visual question answering,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2425–2433.
[5] S. Mannan and B. C. Koivoo, “Visual question answering: A state-of-the-art review,” Artif. Intell. Rev., vol. 53, no. 8, pp. 5705–5745, Dec. 2020.
[6] Z. Lin et al., “Medical visual question answering: A survey,” 2021, arXiv:2111.10056.
[7] L.-M. Zhan, B. Liu, L. Fan, J. Chen, and X.-M. Wu, “Medical visual question answering via conditional reasoning,” in Proc. 26th ACM Int. Conf. Multimedia (ACM), Oct. 2018, pp. 2385–2393.
[8] H. Pan, S. He, K. Zhang, B. Qu, C. Chen, and K. Shi, “MuVAM: A multi-view attention-based model for medical visual question answering,” 2021, arXiv:2107.03216.
[9] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, “Deep modular co-attention networks for visual question answering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6274–6283.
[10] S. Liu, Y. Yang, and J. Yang, “BIBMVQA: A bi-branch model for medical visual question answering,” BMC Med. Imag., vol. 22, no. 1, pp. 1–19, Dec. 2022.
[11] K. Huang, J. Altosaar, and R. Ranganath, “ClinicalBERT: Clinical notes and predicting hospital readmission,” 2019, arXiv:1904.05342.
[12] M. H. Yu, T. Löffstedt, T. Nyholm, and R. Sznitman, “A question-centric model for visual question answering in medical imaging,” IEEE Trans. Med. Imaging, vol. 39, no. 11, pp. 2856–2868, Sep. 2020.
[13] R. Kiyos et al., “Skip-thought vectors,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2015, pp. 3294–3302.
[14] L. Peng, Y. Yang, Z. Wang, Z. Huang, and H. T. Shen, “MRA-Net: Improving VQA via multi-modal relation attention network,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 1, pp. 318–329, Jan. 2022.
[15] C. Yang, M. Jiang, B. Jiang, W. Zhou, and K. Li, “Co-attention network with question type for visual question answering,” IEEE Access, vol. 7, pp. 40771–40781, 2019.
[16] D. Gupta, S. Suman, and A. Ekbal, “Hierarchical deep multi-modal network for visual question answering,” Expert Syst. Appl., vol. 164, Feb. 2021, Art. no. 113993.
[17] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.
[18] Y. Li et al., “A bi-level representation learning model for medical visual question answering,” J. Biomed. Inform., vol. 134, Oct. 2022, Art. no. 104183.
[19] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, “High-quality image captioning with fine-grained and semantic-guided visual attention,” IEEE Trans. Multimedia, vol. 21, no. 7, pp. 1681–1693, Jul. 2019.
[20] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, “Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 12, pp. 5947–5959, Dec. 2018.
[21] J.-H. Kim, J. Jun, and B.-T. Zhang, “Bilinear attention networks,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2018, pp. 1571–1581.
[22] D.-K. Nguyen and T. Okatani, “Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6087–6096.
[23] C. Chen, D. Han, and J. Wang, “Multimodal encoder-decoder attention networks for visual question answering,” IEEE Access, vol. 8, pp. 35662–35671, 2020.
[24] F. Ren and Y. Zhou, “CGMVQA: A new classification and generative model for medical visual question answering,” IEEE Access, vol. 8, pp. 50626–50636, 2020.
[25] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar, “MBBERT: Multimodal BERT pretraining for improved medical VQA,” in Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI), Apr. 2021, pp. 1033–1036.
[26] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, and M. Lungren, “Overview of ImageCLEF 2018 medical domain visual question answering task,” in Proc. CLEF Work. Notes, Sep. 2018, pp. 10–14.
[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.
