基于交叉模态注意力特征增强的医学视觉问答

刘凯，任洪逸，李莹，季怡，刘纯平*

（苏州大学计算机科学与技术学院，江苏 苏州 215006）

摘　要： 医学视觉问答（Med-VQA）需要对医学图像内容与问题文本内容进行理解与结合，因此设计有效的模态表征及跨模态的融合方法对 Med-VQA 任务的表现至关重要。目前，Med-VQA 方法通常只关注医学图像的全局特征以及单一模态内注意力分布，忽略了图像的局部特征所包含的医学信息与跨模态间的交互作用，从而限制了图像内容理解。针对以上问题，提出一种交叉模态注意力特征增强的 Med-VQA 模型（CMAG-MVQA）。基于 U-Net 编码有效增强图像局部特征，从交叉模态协同角度提出选择引导注意力方法，为单模态表征引入其他模态的交互信息，同时利用自注意力机制进一步增强选择引导注意力的图像表征。在 VQA-RAD 医学问答数据集上的消融与对比实验表明，所提方法在 Med-VQA 任务上有良好的表现，相比于现有同类方法，其在特征表征上性能得到较好改善。

关键词： 跨模态交互；注意力机制；医学视觉问答；特征融合；特征增强

中图分类号： TP18　　**文献标志码：** A　　**DOI：** 10.19678/j.issn.1000-3428.0068910

Medical Visual Question Answering Based on Cross-Modal Attention Feature Enhancement

LIU Kai, REN Hongyi, LI Ying, JI Yi, LIU Chunping*

(School of Computer Science and Technology, Soochow University, Suzhou 215006, Jiangsu, China)

【Abstract】 Medical Visual Question Answering (Med-VQA) requires an understanding of content related to both medical images and text-based questions. Therefore, designing effective modal representations and cross-modal fusion methods is crucial for performing well in Med-VQA tasks. Currently, Med-VQA methods focus only on the global features of medical images and the distribution of attention within a single modality, ignoring medical information in the local features of images and cross-modal interactions, thereby limiting the understanding of image content. This study proposes the Cross-Modal Attention-Guided Medical VQA (CMAG-MVQA) model. First, based on U-Net encoding, this method effectively enhances the local features of an image. Second, from the perspective of cross-modal collaboration, a selection guided attention method is proposed to introduce interactive information from other modalities. In addition, a self-attention mechanism is used to further enhance the image representation obtained by selective guided attention acquisition. Ablation and comparative experiments on the VQA-RAD medical question-answering dataset show that the proposed method performs well in Med-VQA tasks and improves feature representation performance compared to similar methods.

【Key words】 cross-modal interaction; attention mechanism; Medical Visual Question Answering (Med-VQA); feature fusion; feature enhancement

0　引言

医学视觉问答（Med-VQA）作为一个新兴研究领域，在医疗诊断中具有巨大的潜力，但尚未得到良好的发展。Med-VQA 以医学图像和医学图像相关的自然语言问题作为输入，以自然语言答案作为输出。对比其他的多模态学习任务，Med-VQA 需要同时对图像和文本内容进行理解，并将图像和文本信息进行融合，因此 Med-VQA 是一项具有挑战性的任务。早在 2018 年，PENG 等[1]受多模态双线性融合与注意力机制的启发，提出了 UMass 模型。该模型将高阶多模态双线性池化与协同注意力机制相结合，增强了图像全局特征与问题文本特征的匹配程度，从而引导模型给出正确回答。然而该模型并没有考虑到模态间的交互作用，从而限制了其在 Med-VQA 上的表现。ABBISHEK 等[2]用 DenseNet[3]对图像与文本特征进行处理，在图像与文本信息损失较小的情况下，将图像与文本特征进

收稿日期： 2023-11-27　　**修回日期：** 2024-03-12

基金项目： 国家自然科学基金（62376041）；江苏省研究生科研与实践创新计划（SJCX21_1341）。

通信作者 E-mail： * cpliu@suda.edu.cn

行融合,从而获得更准确的答案预测。REN 等[4]受基于 Transformer 的双向编码器表示(BERT)[5]模型启发,提出 CGMVQA(Classification and Generative Model for Medical Visual Question Answering)模型。该模型因结合特征金字塔与 Transformer 结构,极大地减少了特征处理与融合过程中的信息损失,同时得益于 Transformer 结构,模型在图像特征与文本特征的模态融合上有良好表现。GONG 等[6]认为,与一般 Med-VQA 解决方案不同,基于 Visual Genome 与 ImageNet 数据集预训练的模型对医学图像处理并不能有效地提取图像中的医学信息,因此提出了基于多个医学图像分类数据集预训练的跨模态自注意力模型(CMSA-MTPT),并将自注意力模块与多模态融合相结合,从而极大地提升了模型在 Med-VQA 任务上的表现。随着预训练的发展,KIM 等[7]提出了基于 Transformer 不带卷积区域监督的多模态预训练方法,从海量的文本-图像对中学习语言和视觉模态间的交互。然而,由于所有预训练的数据都是通用领域的文本-图像对,因此该研究对 Med-VQA 的帮助有限。为了充分利用从大量未标记的医学图像中学习到的图像-文本融合特征,LIU 等[8]提出了对比预训练和表示蒸馏方法(CPRD),利用对比学习在放射学图像未标注数据集中进行预训练和表示蒸馏。

现有 Med-VQA 模型虽然用预训练图像分类模型对图像特征进行提取,但却忽略了图像局部特征信息在答案预测中的重要作用。此外,传统的基于注意力的特征融合方法导致了图像与文本中冗余信息对图像与文本进行匹配时的干扰,在进行图像与文本的特征融合时会出现匹配性能不足的问题。本文针对上述问题提出了 CMAG-MVQA 模型。本文贡献主要在于:1)用基于医学数据集预训练的分割模型进行图像局部特征提取,并使用预训练的 ResNet-34 进行图像全局特征提取,将局部特征与全局特征结合,从而解决了特征信息不足的问题;2)提出选择注意力引导模块,减少了图像与问题文本中冗余信息的干扰,从而增强了模型多模态匹配的能力。

1 相关工作

1.1 视觉问答

在视觉与语言交叉模态研究中,视觉问答(VQA)模型通常以文本问题和自然图像作为输入来获得相关答案。现有研究方法一般都是采用基于 ImageNet 预训练的卷积神经网络、递归神经网络或者基于 Transformer 结构的网络模型作为特征提取器。得益于 ANDERSON 等[9]提出的基于 Faster R-CNN 的 bottom-up 特征以及跨模态特征对齐与融合方法,VQA 受到了广泛的关注并取得了极大的发展。

对于 VQA 的跨模态特征对齐与融合问题,目前主要有基于注意力机制与基于双线性特征融合的两大类方法。KIM 等[10]提出双线性池化注意力方法,用双线性池化代替传统线性注意力计算,然后根据计算所得注意力将视觉特征与文本特征相结合,提高模型在任务中的表现。YU 等[11]在 Top-down 注意力机制基础上,结合 Transformer 网络结构,考虑视觉与文本单一模态内的作用与两种模态间的交互作用,基于模块化设计思想设计了深度模块化协同注意力网络,使问题表征能与其相关的图像区域特征进行融合。邹品荣等[12]提出了多模态协同注意力模型,根据视觉场景中对象间关系的动态交互和文本上下文表示,通过图注意力机制建模不同类型对象间关系。ZHENG 等[13]提出知识引导的元学习框架,并设计了一个新颖的基于三阶张量 Tucker 分解的双线性融合模型来建模图像和文本模态间的精细和丰富的交互。为了充分提取外部医学图像-文本对的特征,CHEN 等[14]遵循预训练和微调范式来尝试对 Med-VQA 这一特定领域进行视觉和语言建模。白亚龙[15]则基于视觉关系推理,利用图像中的高级语义信息使模型更加强大并易于解释。

1.2 融合机制

无论是 VQA 还是 Med-VQA,图像特征与问题特征的有效对齐与融合是决定模型在任务上表现的关键[16]。目前多数图像与问题特征的融合方法都采用双线性池化方式。该方式关键在于对比简单线性模型能有效地建立视觉特征与文本特征更丰富的多模态表征。FUKUI 等[17]认为传统的特征融合方式不如向量外积更具有表达性,因此提出了多模态紧凑双线性方法(MCB)。虽然双线性池化在模态融合上优于传统方法,但是高维度特征以及高复杂度的计算量限制了双线性池化方法的适用范围,因此,YU 等[18]设计了多模态拆分双线性池化模型(MFB),以实现更高效的多模态特征融合。由于简单的双线性池化不足以对复杂推理或高层次的任务进行建模,CADENE 等[19]提出 MuRel 方法,在双线性池化方法的基础上对图像与问题间的交互进行表示并对成对结合区域关系进行建模。与传统卷积层在训练后滤波器保持不变相比,动态生成滤波器是一个新的研究方向。为了学习问题和医学图像间的跨模

态交互，在用动态调制方式来动态融合语义和视觉特征思想的基础上，YU 等[20]提出了一种新颖的问题引导的动态过滤网络(DFN)，利用问题语义信息来动态调节语义特征和视觉特征的特征组合过程。

2 CMAG-MVQA 模型

本文提出的 CMAG-MVQA 模型主要关注医学图像局部信息以及跨模态特征融合时的信息交互。图 1 给出了本文所提出的 Med-VQA 模型的详细结构。该方法在已有的基本框架基础上，重点进行视觉与文本特征表征的描述，体现在以下方面：1)图像局部特征表征学习；2)交叉模态引导的图像、文本特征表征提取；3)交叉引导后图像表征特征的自注意增强。

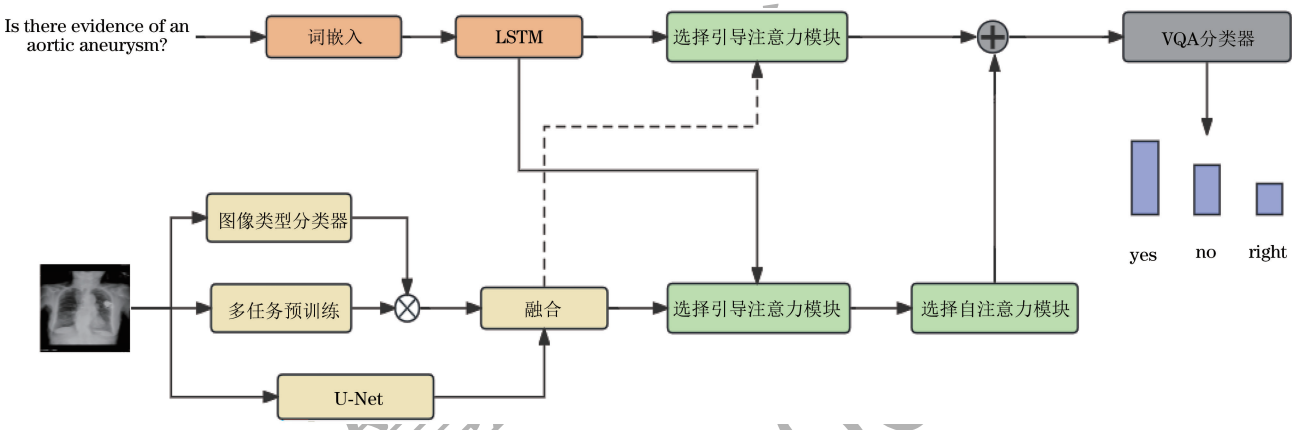

图 1 CMAG-MVQA 模型整体结构

Fig. 1 Overall structure of CMAG-MVQA model

为了训练 Med-VQA 模型，采用由预测答案和医学图像分类交叉熵损失结合的多模态损失函数 L，表示如下：

$$L = L_{\text{vqa}} + \alpha L_{\text{type}} \tag{1}$$

式中：L_{vqa} 表示答案预测的交叉熵损失；L_{type} 表示医学图像分类的交叉熵损失。在实验中，设定 $\alpha = 0.5$，用来平衡 2 种损失。

2.1 视觉表征与视觉特征增强

受限于 Med-VQA 数据集的局限性，现有多数方法都采用迁移学习来获取医学图像的视觉特征表示。早期 Med-VQA 基于 ImageNet 数据集预训练的卷积神经网络(CNN)来提取医学图像的视觉特征，后来也有采用如 CheXpert 胸部射线的大型数据集预训练 DenseNet 作为视觉特征编码器。目前虽然仍采用预训练图像分类模型[如残差网络(ResNet)和视觉几何组网络(VGG)]对图像视觉特征进行提取，但忽视了图像局部特征，导致了图像局部信息缺失。对医学图像而言，图像中的微小结构变化在疾病诊断过程中可能至关重要，在相似度极高的背景组织中的细微变化有可能就代表着某种病变，图像局部特征所包含的信息也至关重要。为了解决图像局部特征不完全导致的信息缺失问题，CMAG-MVQA 模型使用了一种多模态特征增强的方法。多模态特征增强方法主要由多任务预训练的卷积神经网络 ResNet-34 和 U-Net 组成，其中 ResNet-34 用于提取图像的全局特征，而 U-Net 用于提取图像的局部特征，通过将局部特征和全局特征进行拼接融合，从而增强模型的视觉特征。

全局视觉特征提取基于多任务预训练模型获得，具体过程为：采用外部数据集预训练 3 个独立的 ResNet-34 网络获取视觉特征，判断图像是否属于核磁共振成像(MRI)、计算机断层扫描(CT)和 X 射线(X-Ray)3 种类别之一，并通过分类模块获取图像在 3 种类别中的比重。式(2)给出了全局视觉特征表示：

$$v_{\text{global}} = w_1 v_{\text{MRI}} + w_2 v_{\text{CT}} + w_3 v_{\text{X-Ray}} \tag{2}$$

式中：v_{global} 表示全局视觉特征；v_{MRI}、v_{CT}、$v_{\text{X-Ray}}$ 分别表示从 ResNet-34 输出的 MRI、CT 和 X-Ray 的视觉特征；$w_1 \sim w_3$ 为分类模块输出的每个医学图像特征的权重。

局部视觉特征通过 U-Net 视觉特征编码形成，如式(3)所示：

$$v_{\text{local}} = \text{UNet}_{\text{Encoder}}(I) \tag{3}$$

式中：$\text{UNet}_{\text{Encoder}}$ 表示 U-Net 编码器部分；I 表示输入的医学图像。

最后，将全局视觉特征与局部视觉特征通过拼接操作得到最终视觉特征，如式(4)所示：

$$v_{\text{final}} = \text{Contact}(v_{\text{global}}, v_{\text{loacl}}) \tag{4}$$

2.2 文本表征初始描述

单模态文本表征初始描述主要使用词嵌入向量来表达。在 NGUYEN 等[21]工作中，每个输入的问

题被设定在长度最大为 12 个单词，若长度不够 12 个单词则用 0 填充。基于此，CMAG-MVQA 模型中的文本表征初始描述用 BioWordVec 方法[22]将每个单词处理为 200 维词嵌入，并连接来自 VQA-RAD 数据集[23]中的 200 维的增强词嵌入，从而形成 400 维的词表示，然后借助长短期记忆网络(LSTM)形成问题文本的初始描述。

2.3 交叉模态间的交互信息提取

为了充分利用 Med-VQA 任务中的视觉与文本模态的交互信息，CMAG-MVQA 模型中提出了两种基于注意力的模态特征优化的策略，分别是选择自注意力模块(见图 2)与选择引导注意力模块(见图 3)。

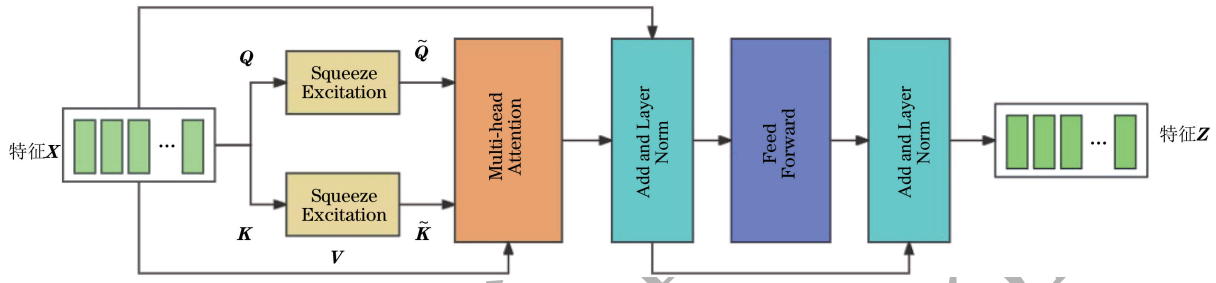

图 2 选择自注意力模块

Fig. 2 Selective self-attention module

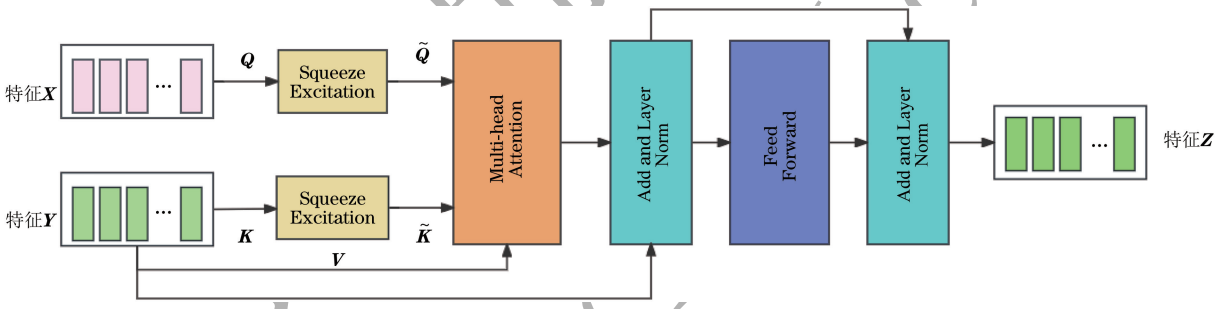

图 3 选择引导注意力模块

Fig. 3 Selective guided attention module

如图 2 所示，选择自注意力模块由通道注意力[24]、多头注意力[25]与前向传播层组成。对于每一个输入特征 $\boldsymbol{X}=[x_1,\cdots,x_m]\in\mathbb{R}^{m\times d_x}$，查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 以及值矩阵 $\boldsymbol{V}$ 都由输入特征 $\boldsymbol{X}$ 映射所得。通道注意力计算每个特征 x 在后续多头注意力计算中的权重，多头注意力学习每个特征对 $<x_i,x_j>$ 的关系，前向传播则使用两侧带有 ReLU 激活函数的全连接层对多头注意力部分的输出进行处理，最终得到处理后的特征 $\boldsymbol{Z}\in\mathbb{R}^{m\times d}$。其中，注意力计算主要包括通道注意力与多头注意力两个基础单元。首先是通道注意力的计算。给出查询矩阵 $\boldsymbol{Q}\in\mathbb{R}^{l\times d}$、键矩阵 $\boldsymbol{K}\in\mathbb{R}^{m\times d}$ 以及值矩阵 $\boldsymbol{V}\in\mathbb{R}^{m\times d}$，其中，$l$ 表示查询值的数量，m 表示键值对的数量。键与查询的全局空间注意力由式(5)、式(6)计算获得：

$$\boldsymbol{K}_c=\frac{1}{n\times d}\sum_{i=1}^{m}\sum_{j=1}^{d}K(i,j) \tag{5}$$

$$\boldsymbol{Q}_c=\frac{1}{l\times d}\sum_{i=1}^{l}\sum_{j=1}^{d}Q(i,j) \tag{6}$$

式中：$\boldsymbol{K}_c$ 表示带有通道注意力分布的键矩阵；$\boldsymbol{Q}_c$ 表示带有通道注意力分布的查询矩阵。同时，为了计算键与查询在点积注意力中的权重，使用带 Sigmoid 激活函数的简单门控机制来完成，如式(7)、式(8)所示：

$$\boldsymbol{K}_s=\sigma(g(\boldsymbol{K}_c,W_k))=\sigma(W_{k2}\delta(W_{k1}\boldsymbol{K}_c)) \tag{7}$$

$$\boldsymbol{Q}_s=\sigma(g(\boldsymbol{Q}_c,W_q))=\sigma(W_{q2}\delta(W_{q1}\boldsymbol{Q}_c)) \tag{8}$$

式中：$\boldsymbol{K}_s$ 表示经过激活函数处理后带有通道注意力分布的键矩阵；$\boldsymbol{Q}_s$ 表示经过激活函数处理后带有通道注意力分布的查询矩阵；δ 表示 ReLU 函数；σ 表示 Sigmoid 函数；$\boldsymbol{W}_{k1},\boldsymbol{W}_{q1}\in\mathbb{R}^{\frac{C}{r}\times C}$；$\boldsymbol{W}_{k2},\boldsymbol{W}_{q2}\in\mathbb{R}^{C\times\frac{C}{r}}$；$r$ 为缩放参数，用于缩小特征向量维度从而减少计算量。因此，通道注意力计算可以通过式(9)、式(10)获得：

$$\tilde{\boldsymbol{K}}=\boldsymbol{K}_s\boldsymbol{K}+\boldsymbol{K} \tag{9}$$

$$\tilde{\boldsymbol{Q}}=\boldsymbol{Q}_s\boldsymbol{Q}+\boldsymbol{Q} \tag{10}$$

式中：$\tilde{\boldsymbol{Q}}$ 为带有通道注意力的查询矩阵；$\tilde{\boldsymbol{K}}$ 为带有通道注意力的键矩阵。如式(11)所示，值矩阵 $\boldsymbol{V}$ 的注意力分布可以用 Softmax 函数对查询 $\tilde{\boldsymbol{Q}}$ 与键 $\tilde{\boldsymbol{K}}$ 的点积与 $\sqrt{d}$ 的商的计算来获取，然后将值 $\boldsymbol{V}$ 与其注意力分布进行点积，得到携带注意力分布的 $\tilde{\boldsymbol{V}}$，

式中，d 为查询 $\widetilde{\boldsymbol{Q}}$ 与键 $\widetilde{\boldsymbol{K}}$ 2 个矩阵的向量维度。

$$\widetilde{\boldsymbol{V}}=\text{Attntion}(\widetilde{\boldsymbol{Q}},\widetilde{\boldsymbol{K}},\widetilde{\boldsymbol{V}})=\text{Softmax}\left(\frac{\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{K}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{11}$$

为了提升特征的表达能力，引入多头注意力机制。多头注意力机制共包含 h 个“头”，每个“头”将查询 $\widetilde{\boldsymbol{Q}}$ 、键 $\widetilde{\boldsymbol{K}}$ 与值 $\boldsymbol{V}$ 映射到了不同且相互独立的特征空间，其计算如式(12)所示。本文将 h 个映射后的特征进行融合，得到最终带有自注意力分布的输出特征 $\boldsymbol{f}_{\text{s}}$，如式(13)所示：

$$H_{\text{head},j}=\text{Attntion}(\widetilde{\boldsymbol{Q}}\boldsymbol{W}_{\widetilde{\boldsymbol{Q}},j},\widetilde{\boldsymbol{Q}}\boldsymbol{W}_{\widetilde{\boldsymbol{K}},j},\boldsymbol{Q}\boldsymbol{W}_{\boldsymbol{V},j}) \tag{12}$$

$$\boldsymbol{f}_{\text{s}}=\text{MHAtt}(\widetilde{\boldsymbol{Q}},\widetilde{\boldsymbol{K}},\boldsymbol{V})=[H_{\text{head},1},H_{\text{head},2},\cdots,H_{\text{head},h}]\boldsymbol{W}_{\text{o}} \tag{13}$$

式中：$\boldsymbol{W}_{\widetilde{\boldsymbol{Q}},j}$，$\boldsymbol{W}_{\widetilde{\boldsymbol{K}},j}$，$\boldsymbol{W}_{\boldsymbol{V},j}\in\mathbb{R}^{d\times d_h}$ 是第 j 个头的映射空间，并且 $\boldsymbol{W}_{\text{o}}\in\mathbb{R}^{h\times d_h\times d}$；$d_h$ 则是多头映射空间的维度。考虑模型大小，实验中设置 $d_h=d/h$。

将带有自注意力分布的特征 $\boldsymbol{f}_{\text{s}}$ 经过残差处理和层归一化处理即可得到选择自注意力模块的输出特征 $\boldsymbol{Z}$，如式(14)所示：

$$\boldsymbol{Z}=\frac{\boldsymbol{X}+\boldsymbol{f}_{\text{s}}-\mu}{\sigma} \tag{14}$$

式中：$\boldsymbol{X}+\boldsymbol{f}_{\text{s}}$ 为残差计算结果；μ 为 $\boldsymbol{X}+\boldsymbol{f}_{\text{s}}$ 的均值；σ 为 $\boldsymbol{X}+\boldsymbol{f}_{\text{s}}$ 标准差。

选择引导注意力结构与选择自注意力结构类似，图 3 展示了选择引导注意力的整体结构。与选择自注意力不同，选择引导注意力需要 2 种不同的模态特征作为输入，由特征 $\boldsymbol{X}$ 引导特征 $\boldsymbol{Y}$ 生成特征 $\boldsymbol{Y}$ 的注意力分布。其中，特征 $\boldsymbol{Y}\in\mathbb{R}^{m\times d_y}$ 经过线性映射得到键矩阵 $\boldsymbol{K}\in\mathbb{R}^{m\times d}$ 和值矩阵 $\boldsymbol{V}\in\mathbb{R}^{m\times d}$，特征 $\boldsymbol{X}\in\mathbb{R}^{n\times d_x}$ 经过线性映射得到查询矩阵 $\boldsymbol{Q}\in\mathbb{R}^{n\times d}$，其中，$d$ 为映射空间特征维度。在得到键矩阵 $\boldsymbol{K}$、值矩阵 $\boldsymbol{V}$ 和查询矩阵 $\boldsymbol{Q}$ 之后，通道注意力计算和多头注意力计算与选择自注意力部分相同，最终将带有引导注意力分布的特征 $\boldsymbol{f}_{\text{g}}$ 经残差和层归一化处理得到输出特征 $\boldsymbol{Z}$，如式(15)所示：

$$\boldsymbol{Z}=\frac{\boldsymbol{Y}+\boldsymbol{f}_{\text{g}}-\mu}{\sigma} \tag{15}$$

式中：$\boldsymbol{Y}+\boldsymbol{f}_{\text{g}}$ 为残差计算；μ 为 $\boldsymbol{Y}+\boldsymbol{f}_{\text{g}}$ 均值；σ 为 $\boldsymbol{Y}+\boldsymbol{f}_{\text{g}}$ 标准差。

3 实验及结果分析

3.1 数据集

实验在 Med-VQA 领域常用基准测试数据集 VQA-RAD 上进行。该数据集包含 315 张放射图像、3 064 个训练问题和 451 个测试问题。实验在训练集上进行模型训练，并在测试集上进行结果展示。训练集和测试集由开放集(Open)和封闭集(Closed)组成。其中：开放集由开放性问题组成，例如提问医学图像对应的疾病名称；而封闭集由封闭性问题组成，问题答案只能从“Yes”和“No”中选择，例如提问对医学图像中的疾病诊断是否正确。

3.2 实验设置

对于 Med-VQA 中的每个问题，首先将问题解析为表计划中的单词，并归一化为 12 个单词长度，然后使用 biowordvec 单词嵌入，把每个单词转换为 400 维的向量，再把词向量输入到 LSTM 网络，生成对应问题的文本特征表示。每个图像则缩放至 224×224 像素大小，然后输入 U-Net 以及预训练的 ResNet 网络进行处理，对应的特征向量维度为 512 维。在模型注意力机制模块中，实验中将“头”个数设定为 8 个，将选择自注意力的维度设定为 51 维，将选择引导注意力部分的维度设定为 1 024 维。

在 Med-VQA 任务中，大多数研究主要采用基于分类的评价指标来衡量方法的性能。基于分类的评价指标是深度学习分类任务中常见的评价指标，该类指标将答案视为分类结果，并计算精确匹配的准确率、精确度和召回率等。在 Med-VQA 任务中，主要采用准确率 A 作为评价指标，该指标能够客观有效地反映方法性能，其计算过程如式(16)所示：

$$A=\frac{n_{\text{c}}}{n_{\text{t}}}\times 100\% \tag{16}$$

式中：n_{c} 表示方法正确回答的问题的个数；n_{t} 表示在 Med-VQA 数据集中问题的总数。

3.3 实验结果

为了证明提出的方法在 Med-VQA 中的有效性，实验对比了目前常见的 7 个解决同类问题的模型，表 1 给出了对比实验结果。其中：加粗数据表示最优值，下同；SAN-RAD 是一个堆叠注意力模型，采用预训练的 VGG-16 和 LSTM 进行特征提取，并将视觉特征和文本特征结合生成注意力分布；MCB-RAD 与 SAN-RAD 类似，针对视觉特征和文本特征融合，该模型采用了紧凑型双线性池化；MEVF-SAN 在 SAN-RAD 的基础上引入未标注的医学图像并采用模型无关元学习和卷积去噪自动编码器对模型特征提取的权重进行初始化；MEVF-BAN 在 MEVF-SAN 的基础上采用双线性池化注意力机制进行注意力计算；MMQ-SAN 和 MMQ-BAN 在 SAN 与 BAN 模型基础上引入了多元模型量化过

程，主要通过自动标注增加元数据，并通过多元学习预测分数的不确定性处理训练阶段的噪声标签；CP-BAN 在 BAN 采用问题条件推理和类型条件推理，将模型通过大量未标记的医学图像进行对比学习，对模型的视觉特征提取器进行初始化；CMSA-MTPT 采用多任务学习初始化视觉特征提取器，并且采用了跨模态自注意力模块对视觉特征和文本特征进行融合处理。

表 1 不同模型性能比较

Table 1 Performance comparison of different models %

对比模型	Open	Closed	All
SAN-RAD[23]	24.2	57.2	44.2
MCB-RAD[23]	25.4	60.6	46.5
MEVF-SAN[21]	40.7	74.1	60.8
MEVF-BAN[21]	4.9	75.1	6.6
MMQ-SAN[26]	46.3	75.7	64.0
MMQ-BAN[26]	5.7	75.8	67.0
CP-BAN[27]	5.1	77.9	68.1
CMSA-MTPT[6]	56.1	77.3	68.8
CMAG-MVQA(ours)	**57.7**	**77.8**	**69.8**

如表 1 所示，在 Open 类型的问题、Closed 类型的问题和问题的总体准确率上，CMAG-MVQA 在 VQA-RAD 数据集上的性能表现明显优于其他模型。其中：SAN-RAD、MCB-RAD 都是通用 VQA 领域的方法，并未对 Med-VQA 领域采取特定的视觉特征方法，因此无法有效提取医学影像中的相关特征信息，模型难以将医学图像和问题所包含的信息相结合并产生正确的答案；而 MEVF-SAN、MEVF-BAN、MMQ-SAN、MMQ-BAN、CP-BAN、CMSA-MTPT 模型引入了额外数据集，针对医学图像训练了特定视觉特征提取器，但模型本身仅提取了医学图像的全局特征而忽视了医学图像的局部特征信息，而局部特征对于病变判断起着重要作用，此外，这 6 种模型采用的基模型仅通过浅层注意力机制高亮了视觉和文本特征中的重要部位，并未对视觉和文本特征的特征交互进行建模从而忽视了跨模态间的特征交互信息，从而影响了模型整体性能表现。本文提出的 CMAG-MVQA 在针对医学图像训练了特定视觉特征提取器的基础上引入预训练的 U-Net 提取了医学图像的局部特征从而进行了特征增强，同时采用了选择自注意力和选择引导注意力模块对视觉特征和文本特征间的交互作用进行建模并通过注意力机制对 2 种模态的表征进行了增强，因此模型在 VQA-RAD 数据集中的表现优于其他模型。

3.4 消融实验

为了进一步验证所提方法中注意力机制改进对 Med-VQA 的性能的改善，根据提出改进的视觉局部特征和交叉模态交互信息表征子模块，分别进行消融实验，实验结果如表 2 所示。

表 2 消融实验结果

Table 2 Results of ablation experiment %

方法模型	Open	Closed	All
CMSA-MTPT	56.1	77.3	68.8
CMSA-MTPT+U-Net	56.9	77.3	69.2
CMSA-MTPT+改进注意力	56.1	77.8	69.2
CMAG-MVQA	**57.7**	**77.8**	**69.8**

从表 2 所示消融实验结果可以看出，在视觉特征表征方面，CMSA-MTPT 方法仅使用多任务预训练的 ResNet-34 图像类别分类模型提取医学全局特征，缺少了视觉的局部特征信息，而本文方法将 U-Net 模型提取的视觉局部特征与 ResNet-34 提取的全局特征相结合，形成全局与局部结合的医学图像特征表征，因此提升了基模型 CMSA-MTPT 方法在 Med-VQA 任务上的表现能力。同时，为了验证选择自注意力机制与选择引导注意力机制的有效性，本文设置了 CMSA-MTPT+改进注意力的消融实验，从实验结果可以看出，使用选择自注意力机制与选择引导注意力机制对视觉特征和文本特征进行跨模态特征，相比于 CMSA-MTPT 使用传统的自注意力机制有较好的性能改善。这是因为单一的自注意力机制仅对视觉特征和文本特征中的重要部分进行高亮并未对视觉特征和文本特征间的交互作用进行建模，缺少了对多模态特征交互作用的信息挖掘，因此使用改进的选择注意力机制结构能有效地改善 CMSA-MTPT 的性能表现。

3.5 实验结果可视化

基于 VQA-RAD 数据集的 CMAG-MVQA 与基模型 CMSA-MTPT 的可视化结果对比如图 4 所示。左侧是 X-Ray、CT 和 MRI 3 种不同采集方式获取的医学图像，分别对应了胸部、腹部和头部 3 种不同的人体部位，每张图像的下方都有该图像对应的问题。中间为 CMSA-MTPT 给出的模型关注图像内容范围的可视化，以及模型给出问题对应的答案。右侧则是提出的 CMAG 模型关注的医学图像内容可视化展示，以及模型给出问题对应的正确答案。在第一个任务中，对于 X-Ray 采集到的胸部医学图像，CMSA-MTPT 针对问题只能给出“Right”区域回答而无法识别器官部位。而在第 2 个问题中，对于 CT 采集到的腹部图像，CMSA-MTPT 完

全定位错误病变部位，CMAG 则能很好地判断具体病变部位。在第 3 个问题中，对于 MRI 采集的头部图像，CMSA-MTPT 则并没有识别出问题是一个二分类问题，答案只能在“axial”和“saggital”中选择，而给出了错误答案“PA”，CMAG-MVQA 模型则识别到了这一点，并且针对问题给出了正确回答。

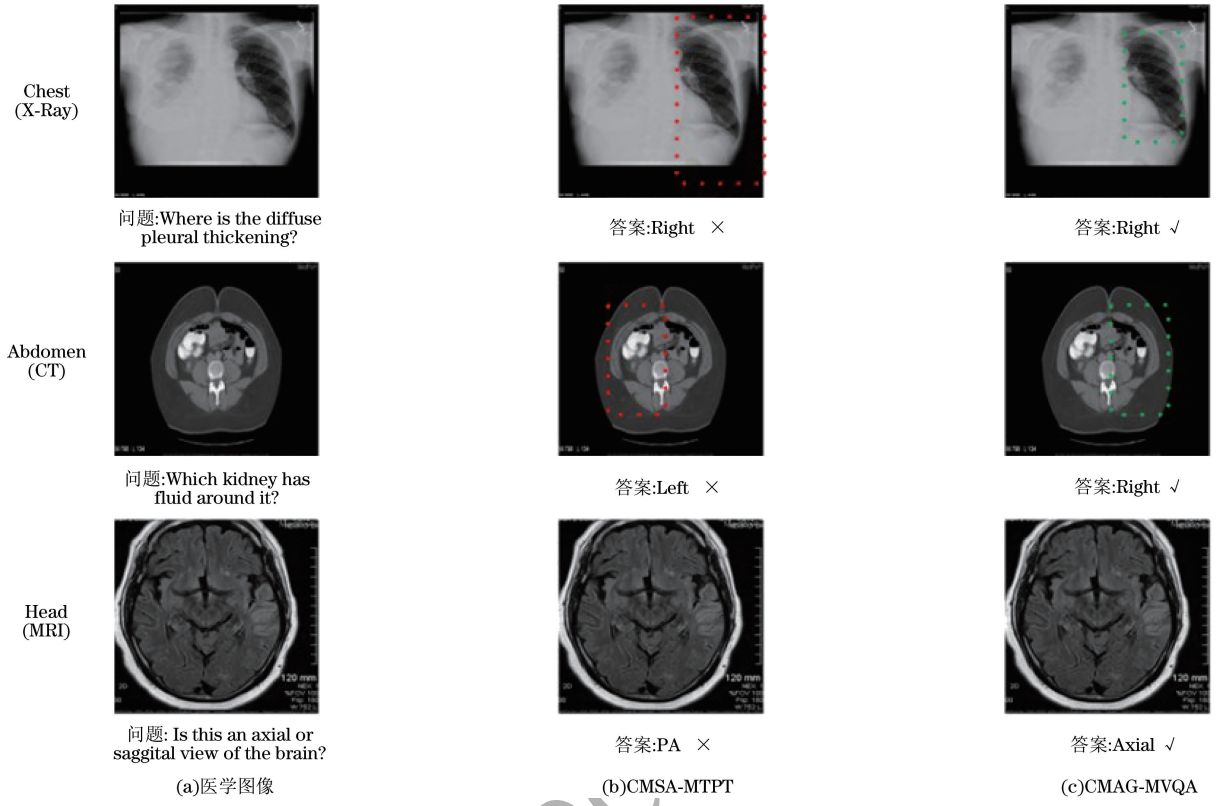

图 4 基模型与本文方法的可视化结果

Fig. 4 Visualization results of the base model and the proposed method

4 结束语

本文提出基于改进注意力机制的特征增强模型，用于解决 Med-VQA 任务。所提方法使用预训练的 U-Net 网络提取图像的局部特征并与多任务预训练分类网络提取的图像全局特征相结合，提升了模型在视觉问答任务的表现能力。此外，针对图像与文本中冗余信息的干扰问题，采用改进的注意力机制，增强了多模态特征的匹配与融合能力，从而提升了模型性能。相比现有方法，本文提出的方法在 VQA-RAD 数据集上带来显著的性能提升。后续将在本文基础上，针对多模态特征存在语义鸿沟的问题进行多模态特征对齐的研究，并改进多模态特征交互的注意力机制结构。

参考文献

[1] PENG Y L, LIU F. UMass at ImageCLEF Medical Visual Question Answering (Med-VQA) 2018 task[C]//Proceedings of ImageCLEF 2018. New York, USA: ACM Press, 2018: 1-9.

[2] ABBISHEK T, KRISHNAMOORTHI M. MIT Manipal at ImageCLEF 2019 visual question answering in medical domain[C]//Proceedings of ImageCLEF 2019. New York, USA: ACM Press, 2019: 1-6.

[3] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 4700-4708.

[4] REN F J, ZHOU Y Y. CGMVQA: a new classification and generative model for medical visual question answering[J]. IEEE Access, 2020, 8: 50626-50636.

[5] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2023-10-11]. https://arxiv.org/abs/1810.04805v2.

[6] GONG H F, CHEN G Q, LIU S S, et al. Cross-modal self-attention with multi-task pre-training for medical visual question answering[C]//Proceedings of the 2021 International Conference on Multimedia Retrieval. New York, USA: ACM Press, 2021: 456-460.

[7] KIM W, SON B, KIM I. Vilt: vision-and-language transformer without convolution or region supervision[C]//Proceedings of International Conference on Machine Learning. [S. l.]: PMLR, 2021: 5583-5594.

[8] LIU B, ZHAN L M, WU X M. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images[C]//Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention. Berlin, Germany: Springer, 2021: 210-220.

[9] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-up

and top-down attention for image captioning and visual question answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 6077-6086.
[10] KIM J H, JUN J, ZHANG B T. Bilinear attention networks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Berlin, Germany: Springer, 2018: 1571-1581.
[11] YU Z, YU J, CUI Y H, et al. Deep modular co-attention networks for visual question answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 6281-6290.
[12] 邹品荣, 肖锋, 张文娟, 等. 面向视觉问答的多模块协同注意模型[J]. 计算机工程, 2022,48(2):250-260.
ZOU P R, XIAO F, ZHANG W J, et al. Multi-module co-attention model for visual question answering[J]. Computer Engineering, 2022, 48(2): 250-260. (in Chinese)
[13] ZHENG W B, YAN L, WANG F Y, et al. Learning from the guidance: knowledge embedded meta-learning for medical visual question answering[C]// Proceedings of the 27th International Conference on Neural Information Processing. Berlin, Germany: Springer, 2020: 194-202.
[14] CHEN Z H, DU Y H, HU J P, et al. Multi-modal masked autoencoders for medical vision-and-language pre-training[C]// Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention. Berlin, Germany: Springer, 2022: 679-689.
[15] 白亚龙. 面向图像与文本的多模态关联学习的研究与应用[D]. 哈尔滨: 哈尔滨工业大学, 2018.
BAI Y L. Research and application of multimodal relevance learning for image and text[D]. Harbin: Harbin Institute of Technology, 2018. (in Chinese)
[16] 何俊, 张彩庆, 李小珍, 等. 面向深度学习的多模态融合技术研究综述[J]. 计算机工程, 2020, 46(5): 1-11.
HE J, ZHANG C Q, LI X Z, et al. Survey of research on multimodal fusion technology for deep learning[J]. Computer Engineering, 2020, 46(5): 1-11. (in Chinese)
[17] FUKUI A, PARK D H, YANG D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding[EB/OL]. [2023-10-11]. https://arxiv.org/abs/1606.01847v3.
[18] YU Z, YU J, FAN J P, et al. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 1821-1830.
[19] CADENE R, BEN-YOUNES H, CORD M, et al. MUREL: multimodal relational reasoning for visual question answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 1989-1998.
[20] YU Y L, LI H F, SHI H R, et al. Question-guided feature pyramid network for medical visual question answering[J]. Expert Systems with Applications, 2023, 214: 119148.
[21] NGUYEN B D, DO T T, NGUYEN B X, et al. Overcoming data limitation in medical visual question answering[C]// Proceedings of the 22nd International Conference on Medical Image Computing and Computer Assisted Intervention. Berlin, Germany: Springer, 2019: 522-530.
[22] ZHANG Y J, CHEN Q Y, YANG Z H, et al. BioWordVec, improving biomedical word embeddings with subword information and MeSH[J]. Scientific Data, 2019, 6: 52.
[23] LAU J J, GAYEN S, BEN A A, et al. A dataset of clinically generated visual questions and answers about radiology images[J]. Scientific Data, 2018, 5: 180251.
[24] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 7132-7141.
[25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2017: 6000-6010.
[26] DO T, NGUYEN B X, TJIPUTRA E, et al. Multiple meta-model quantifying for medical visual question answering[C]// Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention. Berlin, Germany: Springer, 2021: 64-74.
[27] LIU B, ZHAN L M, XU L, et al. Medical visual question answering via conditional reasoning and contrastive learning[J]. IEEE Transactions on Medical Imaging, 2022, 42(5): 1532-1545.

编辑 金胡考