MoE 模型与思维链推理协同优化的医学视觉问答研究

崔芸山[1] 孙希选[1] 陈育德[1] 李殿奎[1*]
CUI Yunshan SUN Xixuan Chen Yude LI Diankui

摘 要 针对目前医学视觉问答（medical visual question answering, Med-VQA）任务中多模态大语言模型存在的领域适应性差、计算成本高、可解释性弱等问题，文章提出了一种面向医疗领域的残差增强型混合专家模型（residual-enhanced mixture of experts for medical domain model, ResMoE-Med），并利用大语言模型强大的思维链推理能力指导本模型处理复杂的医学视觉问答任务。将混合专家模型与思维链推理引入医学视觉问答模型以实现高效的医学诊断，并将模型训练过程分为三个阶段。在模型的推理阶段，提出一种专用于医学领域的模块化思维链推理框架，通过“病灶定位 - 大语言模型指导 - 视觉语言模型鉴别诊断”三步链式推理引导模型生成诊断建议，从而提高模型可解释性。在 VQA-RAD 数据集上的实验表明，配合思维链推理框架的 ResMoE-Med 仅需激活部分参数，就达到了 84.6% 的准确率，比同类型的基线模型提高 0.4%，并且平均推理时间也较同类型基线模型缩短了近 3 s。在 PATH-VQA 数据集的整体性能也优于或基本持平基线模型。消融实验也验证了 MoE 模型与思维链推理的协同优化效果，显著优于传统全量参数激活模型，为资源受限场景下的医学多模态推理提供了一种高效率、低成本的解决方案。

关键词 医学视觉问答；混合专家模型；思维链推理；大语言模型

doi：10.3969/j.issn.1672-9528.2025.11.019

0 引言

医学视觉问答（medical visual question answering, Med-VQA）是一种结合自然语言处理与计算机视觉技术的多模态人工智能任务，其目的是根据输入的医疗图像及相关问题，生成符合医学逻辑的回答。尽管 LLM 和 MLLM 在医学领域取得了一定的学术突破，但医院在训练并将其部署到实际的临床场景仍然面临一些挑战 [1]。首先，不同医学领域具有不同的成像原理和特征，导致模型在一个领域训练后，难以适应其他领域的图像和问题，从而降低模型性能；其次，有关医学视觉问答的多数现有模型在生成回答时，无法提供清晰的推理依据，导致结果缺乏透明性和可解释性；最后，现有医学多模态大模型通常依赖于参数庞大的大语言模型，例如 LLaVA-Med[2] 中使用了 LLaMA3.1(7B)[3] 作为其语言基础模型。这导致模型的训练与部署需要大量的计算资源，使得大多数缺乏足够计算资源的医院对其望而却步，限制了其在临床场景的应用。

针对现有方法的局限性，本文将混合专家模型 [4] 引入到医学视觉问答领域中，通过动态路由机制选择适应专家模块，模拟医院不同领域医疗专家诊断的场景，实现了处理不同领域医学图像的专家模块，混合专家模型 [3] 如图 1 所示。

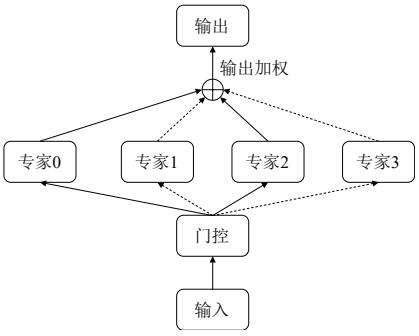

图 1 MoE 模型基础架构

并且，为了更好地提取图像的全局医学特征，受到 ResNet 网络 [5] 残差设计的启发，ResMoE-Med 在专家模块上并联一个全局专家以提取全局特征。这样在模型推理时，只有选定的领域专家和全局专家被激活，进而增强模型稳定性，防止梯度消失。最后，为了进一步提高模型的解决医学复杂任务的能力，增强模型可解释性，受到此前关于大语言模型驱动的医学智能系统研究的启发 [6]，本文提出一个“病灶定位 - 大语言模型指导 - 视觉语言模型鉴别诊断”三步链式推理思维链框架，利用 LLM 的丰富的知识库和强大的思维链推理

1. 佳木斯大学信息电子技术学院 黑龙江佳木斯 154000
[基金项目]2024 年度黑龙江省高等学校基本科研业务费项目（2024-KYYWF-0565）

能力指导 ResMoE-Med 的诊断。这样既可以增强模型的可解释性，还能提升准确性。

本文的贡献主要有以下两点：

（1）提出一个面向医疗领域的残差增强型混合专家模型（ResMoE-Med），在专家模型中并联了全局专家模块，这个全局专家参与所有输入数据的运算，降低了模型在训练过程中出现梯度爆炸的概率，并增强模型的稳定性。为了更好地提升专家模型推理能力，将模型的学习过程分为三阶段。

（2）提出了一个三步链式推理思维链框架，将 LLM 整合到医学问题的解决过程中，利用 LLM 强大的思维链推理能力指导 ResMoE-Med 生成。

1 研究方法

本文所提出的 MoE 和思维链推理协同优化的医学视觉问答模型推理框架如图 2 所示。

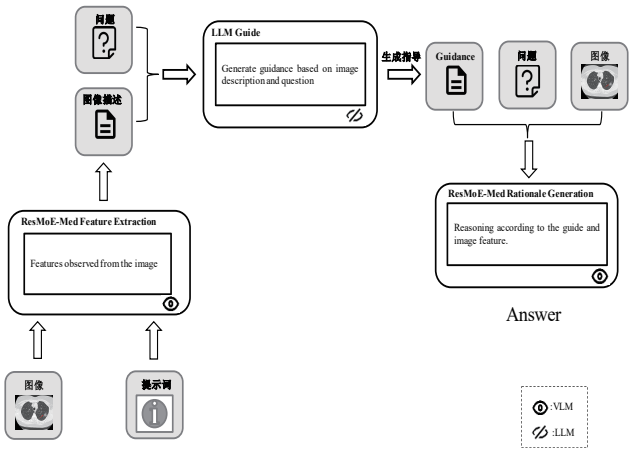

图 2 思维链推理框架

1.1 用于医学领域的模块化思维链推理框架概述

根据 DDCoT[7] 和 Zhang 等人 [8] 的研究表明，大语言模型通过思维链提示生成中间推理链作为推导答案的依据，已经在复杂的推理任务中展示的令人印象深刻的性能。为了更好地配合医学视觉语言模型的推理，本文将思维链提示的方式运用到医学多模态领域，使用思维链提示的方式逐步引导多模态视觉语言模型的生成。尽管 VLM 能够胜任部分简单推理任务，但它们在解读医学图像方面存在推理能力较弱等挑战，因此本文使用 GPT-3.5 为 VLM 提供推理指导，以弥补 VLM 在推理能力上的不足。针对医学视觉问答任务，本文将任务分解为三步： 病灶定位、大语言模型指导、视觉语言模型鉴别诊断。

1.1.1 病灶定位

在思维链推理的第一阶段，由于大语言模型不具备多模态理解能力，本文利用 ResMoE-Med 从输入的医学图像中提取病灶信息。该步骤负责人主要负责对医学图像进行初步解读，提取医学图像的主要特征并转换为文本描述。模型通过预训练的视觉编码器提取图像的全局特征，包括病灶区域、组织结构异常及其他潜在异常信号。为确保生成的描述全面且精准，系统设计了以下提示词引导模型生成图像特征描述：“现在你是一名影像科医生，请仔细检查以下医学图像所描述的特征，包括形状、大小、颜色、纹理和任何不寻常的图案。注意重点识别任何潜在病变。异常和非典型结构，并准确详细地描述出来。”

1.1.2 大语言模型引导 VLM 诊断

接下来，思维链框架引入 GPT-3.5 来生成详细的推理指导。在此步骤中，LLM 使用第一阶段的图像描述 C 和输入问题 Q，构建一条清晰完整的医学思维链，这条医学思维链可用于为医学影像的解读提供详细的指导。首先系统会向 LLM 提示一段提示词，引导其生推理思路，并确保 GPT 在推理过程中保持客观中立，生成的指导具备医学逻辑的可解释性。具体提示设计如下：

“现在你是一名老道的医学专家，任务是指导如何分析医学图像。请你根据给定图像描述和临床问题，分析图像描述和问题之间的关联，并识别潜在的诊断结果，注意重点将图像描述的特征与已知的病理模式和临床证据联系起来。请运用你的医学知识，为解决这个任务提供指导。请不要对问题做任何假设或结论。”这段提示词可以确保 LLM 生成客观中立的指导，LLM 会根据任务要求，利用医学知识分析图像描述所给出的医学特征。图 3 展示了回答病理图像问题的一个示例。

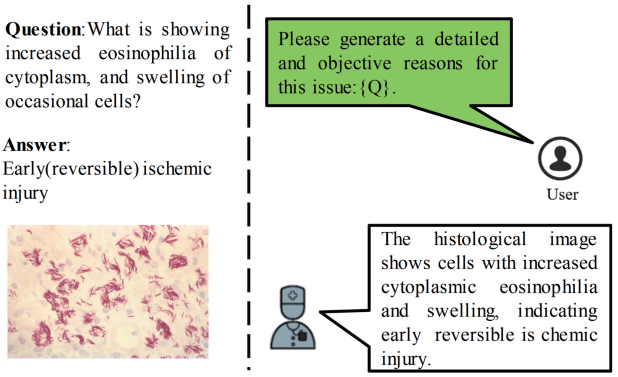

图 3 推理示例

1.2 ResMoE-Med 模型结构概述

1.2.1 模型结构概述

ResMoE-Med 模型由视觉编码器、线性投影层、文本嵌入层、LLM 解码器和 MoE 层组成，具体结构如图 4 所示。模型采用 clip-vit-large 预训练模型初始化视觉编码器和文本编码器。

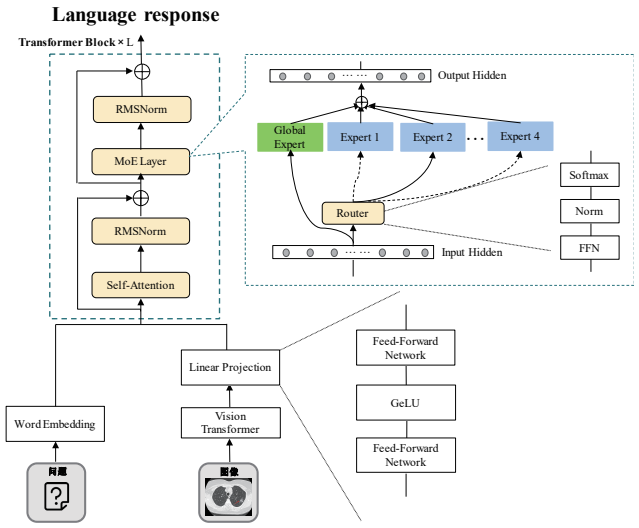

图 4 ResMoE-Med 模型结构

ViT 模型对医学图像进行特征提取，并其输出通过线性投影层得到图像对齐向量 $\boldsymbol{V}=[v_1, v_2, \cdots, v_P]$，该层是一个由两层前馈神经网络及中间的 GeLU 激活函数组成线性投影层，用于将视觉特征映射至与 LLM 输入格式兼容的向量空间中。在文本编码器部分，问题经过分词器后，输入至 CLIP 的文本编码器，生成文本特征向量序列 $\boldsymbol{T}=[t_1, t_2, \cdots, t_N]$。随后，图像对齐向量和文本特征向量拼接后组成联合输入序列 x_0 输入至 LLM 解码器。接下来，联合输入序列 x_0 会通过 L 层 Transformer 块进行计算，每一层 Transformer 块主要进行两个操作：多头注意力机制和混合专家计算。具体公式为：

$$x_0 = \left[v_1, v_2, \cdots, v_P, t_1, t_2, \cdots, t_N\right] \tag{1}$$

$$x_l' = \mathrm{MSA}\left(\mathrm{RMSNorm}(x_{l-1})\right) + x_{l-1}, l = 1, 2, \ldots, L \tag{2}$$

$$x_l = \mathrm{MoE}\left(\mathrm{RMSNorm}(x_l')\right) + x_l' \tag{3}$$

1.2.2 混合专家动态路由机制

本文使用的混合专家模型是通过动态路由机制选择适合的专家来处理输入数据，路由器的任务是为每个输入序列计算得到每个专家的权重，根据这些权重来决定哪个专家被激活。专家模块的设计采用两种类型的专家：全局专家和领域专家。全局专家并不专注于特定医学领域，而是负责处理所有的输入数据并提供一个通用的输出。领域专家则针对特定医学领域进行专门训练学习，这样分工训练的方式既可以提升模型的领域适应性，还可以防止模型不断学习过程中出现灾难性遗忘。具体来说，假设有 E 个领域专家，表示为 $E=\{e_1, e_2, \cdots, e_E\}$。路由器首先会通过一个线性层来计算输入 x 对每个专家的权重，并对结果进行归一化处理，得到权重逻辑值 $f(x)$。

$$f(x) = \mathrm{RMSNorm}(W \cdot x), \quad \boldsymbol{W} \in \mathbf{R}^{D \times E} \tag{4}$$

式中：D 是输入特征的维度；E 是领域专家的数量。路由器接着会对计算得到的转换为概率，$f(x)$ 会通过 Softmax 转换为一个概率分布：

$$P(x)_i = \frac{\mathrm{e}^{f(x)_i}}{\sum_{j=1}^{E} \mathrm{e}^{f(x)_j}}, i = 1, 2, \ldots, E \tag{5}$$

式中：$P(x)_i$ 表示输入 x 对专家 e_i 的激活概率，即输入 x 选择专家 e_i 的可能性。根据计算出的概率，模型会选择概率最大的专家来处理当前输入。每个标记只会激活概率最高的领域专家，构成最终的输出。本文使用指示函数 I 来确保只有概率最高的专家被激活。

$$k = \arg\max_i P(x)_i \tag{6}$$

$$I(i=k) = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases} \tag{7}$$

$$\mathrm{MoE}(x) = E_{\mathrm{global}}(x) + \sum_{i=1}^{E} I(i=k) \cdot P(x)_i \cdot \mathrm{e}_i(x) \tag{8}$$

1.3 ResMoE-Med 模型训练过程概述

由于将 MoE 模型引入 VLM 后导致模型结构复杂，如果直接引入 MoE 后的模型进行微调，会导致训练不稳定、准确率降低的情发生。因此，受到 Lin 等人 [9] 的启发，为了更好地配合思维链框架，本模型采用分阶段训练的方式，具体训练过程如图 5 所示。

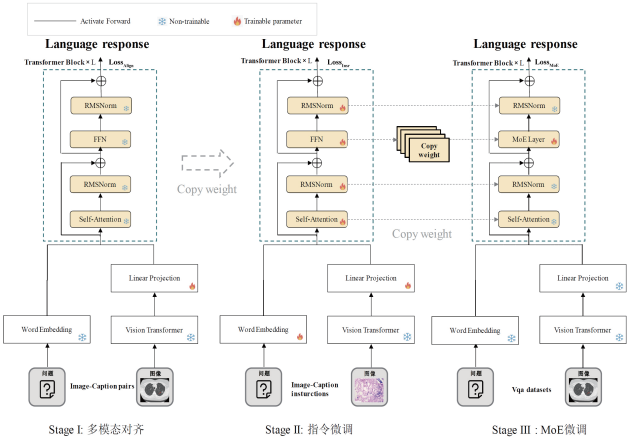

图 5 ResMoE-Med 训练过程示意图

1.3.1 多模态医学对齐

为了使模型能够理解医学图像，不仅需要引入视觉编码器，还需要一个线性投影层来转换视觉编码器的输出，使其转换为 LLM 能够理解的语义序列。在训练的第一阶段需要冻结模型的其他层参数，使用公开的 LLaVA-Med-Alignment 数据集仅训练视觉编码器后的线性投影层以实现多模态对齐。本文将医学图像 I_i 输入到视觉编码器 V_e，生成图像编码向量 $\boldsymbol{T}_i$。同时，将文本描述 C_i 经过分词器转化为文本编码向量 $\boldsymbol{T}_{\mathrm{text}}$，最后将图像编码向量和文本编码向量拼接，并将其输入到 LLM 解码器进行处理。具体公式分别为：

$$\boldsymbol{T}_i = E_v(I_i) \tag{9}$$

$$\boldsymbol{T}_{\text{text}} = \text{Tokenizer}(C_i) \tag{10}$$

$$L_{\text{Align}} = -\sum_{i=1}^{N} \log p\left(\boldsymbol{T}_{\text{text}}^{[P+i]} \mid \boldsymbol{T}_{\text{comb}}, \boldsymbol{T}_{\text{text}}^{[:i-1]}\right) \tag{11}$$

1.3.2 指令微调

第二阶段的目标是提升模型理解和执行多模态复杂指令的能力，使其能够应对复杂多样的医学多模态任务。在这一阶段，输入的医学信息不仅包含图像信息还包含文本信息。为此，本次训练冻结了视觉编码器的所有参数，将指令向量 $\boldsymbol{T}_{\text{instr}}$ 和图像向量 $\boldsymbol{T}_i$ 进行拼接得到 $\boldsymbol{T}_{\text{comb}}$，并将其输入到 LLM 解码器中。目标是最小化以下损失函数：

$$L_{\text{Instr}} = -\sum_{i=1}^{N} \log p\left(\boldsymbol{T}_{\text{resp}}^{[P+i]} \mid \boldsymbol{T}_{\text{comb}}, \boldsymbol{T}_{\text{resp}}^{[:i-1]}\right) \tag{12}$$

式中：$\boldsymbol{T}_{\text{resp}}^{[P+i]}$表示模型在第 $P+i$ 个位置生成的预测向量；$\boldsymbol{T}_{\text{resp}}^{[:i-1]}$表示模型之前生成的所有向量组成的张量。

为了进一步优化模型的推理能力，并提高模型在实际运用中的效率，预训练了一个动态路由器，用于根据输入图像的模态类型选择适当的专家进行推理。该路由器通过在一个已经标注图像类型的小规模数据集上进行预训练，学习如何识别不同类型的医学图像（如 CT、MRI、病理图像等），并根据识别结果来激活相应的专家模块。路由器的训练目标是最小化以下损失函数：

$$L_{\text{Router}} = -\sum_{i=1}^{M} y_i \log p(y_i \mid \boldsymbol{T}_{\text{comb}}) \tag{13}$$

式中：y_i 是输入图像的真实模态标签。

1.3.3 MoE 微调

在模型最终训练阶段，本文将 LLM 中的前馈神经网络（FFN）替换成混合专家模型。在第二阶段训练得到的路由器会处理输入数据并分配与之对应的领域专家模块，同时全局专家将始终保持激活，确保模型能够捕获完整医学信息。由于在 MoE 结构中，多个专家共同处理输入数据，需要保证各个专家均衡利用，如果专家利用不均衡，过于依赖于某个专家，模型可能会退化为普通的 dense 模型，丧失了使用 MoE 的优势。为了防止训练过程中部分专家过载，而其他专家长期闲置而无法有效参与学习情况发生，在损失函数中加上一个辅助损失，辅助损失可以缓解在各个领域专家在训练时出现的负载不均衡的问题，故本阶段的训练目标是最小化损失函数：

$$L_{\text{MoE}} = L_{\text{regressive}} + \alpha L_{\text{aux}} \tag{14}$$

其中，自回归损失优化 LLM 生成医学回答的能力，使其能够逐步生成医学文本序列，其定义为：

$$L_{\text{regressive}} = -\sum_{i=1}^{N} \log p(\boldsymbol{T}_{\text{resp}}[P+i] \mid O_{\text{MoE}}, \boldsymbol{T}_{\text{resp}}[:i-1]) \tag{15}$$

式中：O_{MoE} 是经过 MoE 层的输出。辅助损失定义为：

$$L_{\text{aux}} = E\sum_{i=1}^{E}(F_i - \frac{1}{E})^2 \tag{16}$$

$$F = \frac{1}{K}\sum_{i=1}^{E} 1\{\arg\max P(x) = i\} \tag{17}$$

式中：F 是负载均衡项，如果某个专家实际处理的 token 过多，F 的值就会较大；如果某个专家很少被使用 F 的值就会趋于 0。

通过这一阶段的训练，模型就具备了更强的医学理解能力和推理能力，能够针对不同医学场景自适应地调整激活的领域专家。

2 实验

2.1 数据集和实验参数

本文使用 VQA-RAD[10]、PATH-VQA[11] 两个具有代表性且广泛运用的医学 VQA 数据集来对本文所提出的模型进行微调和评估。采用 Phi-2 模型作为基准模型，clip-vit-large 初始化视觉编码器和文本编码器，领域专家数量 E 设置为 4，设计目标是以较小的参数规模实现高效推理能力。据此基础模型构建了总参数为 5.4 B 的 ResMoE-Med。模型代码基于 PyTorch 框架构建，在 4 张 16 GB NVIDIA Tesla V100 GPU 上进行训练和评估。使用 Adam 优化器来训练模型，采用余弦退火算法来控制学习率。

2.2 对比实验

为了验证基于 MoE 网络和思维链推理的医学视觉问答模型的有效性，将实验结果和一系列基线方法进行了比较。表 1 展示了 ResMoE-Med 在 VQA-RAD 和 PATH-VQA 两个数据集上对比实验的结果，“—”表示原论文中没有报道该项数据。

表 1 本文方法与基线模型在 VQA-RAD 和 PATHVQA 的对比结果

单位：%

方法	VQA Accuracy/Recall			
	VQA-RAD		PATHVQA	
	Open	Closed	Open	Closed
BiomedGPT-S	13.40	57.80	10.70	84.20
BiomedGPT-M	53.60	65.07	12.50	85.70
BiomedCLIP	—	79.80	—	—
PubMedCLIP	60.10	80.00	—	—
VL Encoder–Decoder	—	82.47	—	85.61
M2I2	—	83.50	—	88.00
LLaVA-Med（Llama7B）	**61.52**	84.19	**37.95**	91.21
本文方法	60.74	**84.60**	35.19	**91.80**

实验结果表明，ResMoE-Med 在两个公开数据集上都表现出较好的效果，该模型医学视觉问答任务中相比于传统医学视觉问答模型展现出良好的参数利用率和性能的优势。并且 ResMoE-Med 通过混合专家网络和动态路由机制，在模型大小为 5.4 B 总参数量下仅激活两个专家模块，同时再结合思维链推理，即在 VQA-RAD 数据集封闭式问题上取得 84.60% 的准确率，在开放式问题的召回率也接近于 LLaVA-Med 的 61.52%。在 PATHVQA 数据集上，ResMoE-Med 在封闭问答对准确率高达 91.80%。这个结论验证了 MoE 结构结合思维链推理优化策略的有效性。尽管在 PathVQA 开放式问题上的召回率略低于 LLaVA-Med，未来可通过构建专用思维链数据集进一步优化 ResMoE-Med 的推理能力。

相较于 LLaVA-Med 需要在 8 张 A100 GPU 进行训练，ResMoE-Med 仅需 4 张 V100 GPU，显著地降低了计算资源需求。并且，相较于 LLaVA-Med 推理所需显存也降低了近 45%，同时推理速度得到大幅提升。

2.3 消融实验

为了验证本文方法的有效性，本文在 VQA-RAD 和 PATH-VQA 数据集上针对不同的模块和方法进行一系列的消融实验，实验设置保持一致，其他参数保持不变，以确保不同模块和方法对模型性能的独立影响，具体结果如表 2 所示。

表 2 消融实验结果对比

单位：%

Ablation Setting	VQA-RAD（Open）	VQA-RAD（Closed）	PATH-VQA（Open）	PATH-VQA（Closed）
本文方法	**60.74**	**84.60**	**35.19**	**91.80**
将 MoE 结构替换为 FFN	58.67	83.69	34.24	90.11
将全局专家替换为领域专家	59.93	83.87	—	—
不使用思维链推理策略	59.75	84.37	34.13	91.32

为验证混合专家网络（MoE）对医学多模态推理的贡献，本文对比了使用前馈神经网络与 MoE 结构在不同任务场景下的性能差异。在 VQA-RAD 数据集上，使用 MoE 结构的 ResMoE-Med 相比于传统 FFN 层的版本，在所有任务上均取得了更好的性能提升。其中，在 VQA-RAD 数据集开放式问答任务中，使用 MoE 结构相比使用传统的 FFN 层提升了近 2 个百分点，而在 VQA-RAD 封闭式任务中提升了 0.91%。同样，在 PATH-VQA（Open）和 PATH-VQA（Closed）任务上，使用 MoE 结构的 ResMoE-Med 分别比 FFN 层版本高出 0.95% 和 0.69%。

为了进一步探究本文提出的全局专家对模型性能的影响，本文将全局专家替换为领域专家并将路由策略修改为 Top-2 路由策略，并重新在 VQA-RAD 数据集进行第三阶段训练，该实验表明了全局专家依旧对模型起到了性能提升的作用，在模型中起到了医学知识整合、提高模型稳定性等关键作用，可以有效防止模型再训练过程中出现灾难性遗忘的问题出现，如果去掉全局专家，在 VQA-RAD 数据集开放式和封闭式任务准确性均有所下降。这表明全局专家的存在增强了模型的稳定性，提升模型在复杂的多模态任务上的性能表现。

为了验证思维链推理框架对医学多模态推理任务的贡献，对比了使用模块化思维链推理策略和采用直接推理的模型在不同任务场景下的表现。从实验结果可以看出，模块化思维链推理能够有效提升模型的推理能力，特别是在开放式任务上表现尤为显著。这表明医学视觉问答任务受益于分步思维链推理的方式，采用模块化思维链推理的方式可以通过引导模型逐步分解问题，提高了推理的连贯性。

结合消融实验的结果，可以看出 MoE 结构与思维链推理策略在医学多模态推理任务中具有互补的作用。MoE 通过专家动态路由机制增强模型的领域适应力，防止模型在训练过程中出现灾难性遗忘，并增强模型推理能力。而思维链推理则通过分步引导模型推理的方式，提高模型的可解释性。两者协同优化使得本文方法在资源受限的场景下，依然保持或超越基线模型的性能。

3 结论

本文针对现有医学视觉问答多模态大语言模型普遍存在的计算成本高、领域适应性弱等问题提出了一种面向医疗领域的残差增强型混合专家模型，并将模型结合模块化思维链推理实现了一种高效、可解释的医学视觉问答系统。本文方法在两个医学 VQA 数据集与其他基准模型比较后得出，相比于其他基线模型，本文方法在降低计算成本的同时，其性能不逊于基线模型。同时，ResMoE-Med 的轻量化特性使其更适用于资源受限环境，为现实计算资源受限场景下训练和部署医疗 AI 系统提供了新的技术路径。

参考文献：

[1]LIU B, ZHAN L M, XU L, et al. SLAKE: a semantically-labeled knowledge-enhanced dataset for medical visual question answering[C]//2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). Piscataway:IEEE, 2021: 1650-1654.

[2]LI C Y, WONG C, ZHANG S, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day[EB/OL]. (2023-06-01)[2025-09-18].https://doi.

org/10.48550/arXiv.2306.00890.

[3]GRATTAFIORI A, DUBEY A, JAUHRI A, et al. The Llama 3 herd of models[EB/OL]. (2024-11-23)[2025-09-25]. https://doi.org/10.48550/arXiv.2407.217.

[4]SHEN S, YAO Z W, LI C Y, et al. Scaling vision-language models with sparse mixture of experts[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Brussels: ACL, 2023:11329–11344.

[5]HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[6]FAN Z H, TANG J L, CHEN W, et al. AI hospital: interactive evaluation and collaboration of LLMs as intern doctors for clinical diagnosis[EB/OL].(2024-02-21)[2025-09-12]. https://arxiv.org/html/2402.09742v2.

[7]YOU H X, SUN R, WANG Z C, et al. IdealGPT: iteratively decomposing vision and language reasoning via large language models[EB/OL].(2025-04-11)[2025-09-12]. https://doi.org/10.48550/arXiv.2305.14985.

[8]ZHANG Z S, ZHANG A, LI M, et al. Multimodal chain-of-thought reasoning in language models[EB/OL]. (2024-05-20)[2025-09-25].https://doi.org/10.48550/arXiv.2302.00923.

[9]LIN B, TANG Z Y, YE Y, et al. MoE-LLaVA: mixture of experts for large vision-language models[EB/OL].(2024-12-23)[2025-09-25].https://doi.org/10.48550/arXiv.2401.15947.

[10]LAU J J, GAYEN S, ABACHA A B, et al. A dataset of clinically generated visual questions and answers about radiology images[J/OL]. Scientific data, 2018[2025-09-11].https://www.nature.com/articles/sdata2018251.

[11]HE X H, ZHANG Y C, MOU L T, et al. PathVQA: 30000+ questions for medical visual question answering[EB/OL].(2020-03-07)[2025-09-25].https://doi.org/10.48550/arXiv.2003.10286.

【作者简介】

崔芸山（1999—），男，四川绵阳人，硕士研究生，研究方向：多模态、计算机视觉等。

孙希选（1999—），男，山东聊城人，硕士研究生，研究方向：自然语言处理、深度学习等。

陈育德（1981—），男，黑龙江佳木斯人，硕士研究生，讲师，研究方向：医学信息学、深度学习。

李殿奎（1964—），通信作者（email:datutu2002@sina.com），男，黑龙江齐齐哈尔人，硕士研究生，教授、硕士研究生指导教师，研究方向：图像处理、深度学习。

（收稿日期：2025-05-30 修回日期：2025-11-10）

（上接第 78 页）

[5] 翟晓岩，高刚，李勇根，等．融合注意力机制的二维卷积神经网络测井曲线重构方法 [J]. 石油地球物理勘探，2023, 58(5): 1031-1041.

[6] 周欣，曹俊兴，王兴建，等．基于双向门控循环单元神经网络的声波测井曲线重构技术 [J]. 地球物理学进展，2022, 37(1): 357-366.

[7] 杨满玉，李晶，王锦鹏，等．基于循环神经网络的测井曲线重构技术研究 [J]. 当代化工研究，2023(8): 158-160.

[8] 张海涛，杨小明，陈阵，等．基于增强双向长短时记忆神经网络的测井数据重构 [J]. 地球物理学进展，2022,37(3): 1214-1222.

[9] 尚福华，卢玉莹，曹茂俊．基于改进 LSTM 神经网络的测井曲线重构方法 [J]. 计算机技术与发展，2022,32(6): 198-202.

[10] 王俊，曹俊兴，尤加春．基于 GRU 神经网络的测井曲线重构 [J]. 石油地球物理勘探，2020,55(3): 510-520.

[11] 赵眸光．深度学习与神经网络 [M]. 北京：电子工业出版社，2023.

[12]AURELIEN G. 机器学习实战：基于 Scikit-Learn 和 TensorFlow[M]. 北京：机械工业出版社，2018.

[13] 齐春生，丁磊，焦祥燕，等．基于双向长短时循环神经网络的沉积微相自动识别方法：以莺歌海盆地东方 B 气田为例 [J]. 科技和产业，2023,23(5): 217-221.

[14] 付强，王华伟．基于多层 LSTM 的复杂系统剩余寿命智能预测 [J]. 兵器装备工程学报，2022,43(1): 161-169.

【作者简介】

张茜（1987—），女，山东潍坊人，硕士，工程师，研究方向：油气勘探开发数字化和智能化。

（收稿日期：2025-05-28 修回日期：2025-11-14）