



计算机科学

COMPUTER SCIENCE

数据知识双增强的医学视觉问答网络

闫玉静, 侯霞, 郭玉婷, 张铭梁, 宋文凤

引用本文

闫玉静, 侯霞, 郭玉婷, 张铭梁, 宋文凤. [数据知识双增强的医学视觉问答网络](#)[J]. 计算机科学, 2025, 52(12): 252-259.

YAN Yujing, HOU Xia, GUO Yuting, ZHANG Mingliang, SONG Wenfeng. [Data and Knowledge Enhanced Medical Visual Question Answer Network](#) [J]. Computer Science, 2025, 52(12): 252-259.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[卷积增强自适应分类模型的构造与研究](#)

Construction and Research of Convolution Enhanced Adaptive Classification Model

计算机科学, 2025, 52(11A): 241200069-5. <https://doi.org/10.11896/jsjcx.241200069>

[频域纹理先验与特征增强的医学图像分割模型](#)

Medical Image Segmentation Model Based on Frequency Texture Prior and Frequency Feature Enhancement Fusion

计算机科学, 2025, 52(11A): 241200125-8. <https://doi.org/10.11896/jsjcx.241200125>

[CINN:一种高速且抗JPEG的医学图像水印网络](#)

CINN:A High-speed and JPEG-resistant Medical Image Watermarking Network

计算机科学, 2025, 52(11A): 241100037-7. <https://doi.org/10.11896/jsjcx.241100037>

[基于改进YOLO模型的脑肿瘤病灶区域检测](#)

Detection of Brain Tumor Lesion Areas Based on Improved YOLO Model

计算机科学, 2025, 52(11A): 241000166-8. <https://doi.org/10.11896/jsjcx.241000166>

[融合注意力机制的道路场景三维目标检测算法](#)

Three-dimensional Object Detection Algorithm of Road Scene Based on Attention Mechanism

计算机科学, 2025, 52(11A): 241100112-7. <https://doi.org/10.11896/jsjcx.241100112>

数据知识双增强的医学视觉问答网络

闫玉静¹ 侯霞¹ 郭玉婷² 张铭梁¹ 宋文凤¹

1 北京信息科技大学计算机学院 北京 102206

2 北京航空航天大学虚拟现实技术与系统国家重点实验室 北京 100191

(2022020605@bistu.edu.cn)

摘要 医学视觉问答(Medical Visual Question Answering, Med-VQA)旨在正确回答与给定医学图像相关的临床问题,在临床医学智能化中起着至关重要的作用。虽然该领域研究已获得一定进展,但是在文本和图像多模态输入信息的深度提取,以及小规模数据集上的有效模型训练方面仍然面临挑战。对此,提出一种数据知识双增强的医学视觉问答网络。针对小规模数据集,设计了多模态条件混合模块对输入的图像和文本进行数据增强,利用问题类别作为约束条件对输入样本对进行线性组合,以提高答案生成的合理性。针对多模态特征提取,设计了一个基于卷积神经网络的图像位置识别器,将其捕获的图像位置特征编码到图像特征和问题特征的融合过程中进行知识增强,可在较少的参数下实现有效的特征提取。在 SLAKE 和 VQA-RAD 数据集上的实验结果表明,与基线模型相比,所提模型的性能有明显提升。

关键词: 视觉问答;医学视觉问答;医学图像;数据增强;计算机视觉

中图分类号 TP391

Data and Knowledge Enhanced Medical Visual Question Answer Network

YAN Yujing¹, HOU Xia¹, GUO Yuting², ZHANG Mingliang¹ and SONG Wenfeng¹

1 School of Computer Science, Beijing Information Science and Technology University, Beijing 102206, China

2 State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China

Abstract Med-VQA aims to accurately answer clinical questions based on a given medical image, which is key in advancing clinical medical intelligence. Despite some progress in this field, challenges remain in extracting deep multimodal information from both images and questions and in effectively training models on small-scale datasets. To address these issues, this paper proposes a Med-VQA network that incorporates dual data and knowledge enhancement. Aiming at small-scale datasets, a multimodal conditional mixing module is designed to enhance the input image and question data, and linear combinations of input sample pairs are performed by using the category of questions as constraints to improve the rationality of answer generation. For multimodal feature extraction, an image location recognizer based on convolutional neural networks is designed to encode the captured image location features into the fusion process of image and question features for knowledge enhancement, which can effectively achieve feature extraction under fewer parameters. Experimental results on the SLAKE and VQA-RAD datasets demonstrate that the proposed model significantly outperforms the baseline models.

Keywords Visual question answering, Medical visual question answering, Medical images, Data enhancement, Computer vision

1 引言

视觉问答(Visual Question Answer, VQA)旨在理解图像和相关问题,以提供准确的回答^[1],是重要的视觉语言理解任务之一,被广泛应用于视障人士的辅助技术、教育和文化传承、视觉聊天机器人、商业服务、监控检索和医疗诊断等领域^[2]。医学视觉问答(Medical Visual Question Answer, Med-VQA)则是通过医学图像为相关问题提供准确答案,常被应用于临床决策、远程医疗等特定领域。医疗图像通常是通

CT, MRI, X 射线等多种成像方式获取的灰度图像,而问题则包括研究的医疗细节、使用的成像方法、涉及的器官系统等^[3]。现有的研究表明,Med-VQA 可以辅助放射科医生、病理学家或医疗助理^[3],有助于减少医护人员的工作量,降低误诊率和错误信息的传播,从而提高医疗保健服务的安全性和整体质量。因此,Med-VQA 成为一个广泛研究的课题。与通用 VQA 模型相比,Med-VQA 在医学图像特征提取和数据集两方面面临挑战。

特征提取是 Med-VQA 的基础环节之一,但是由于医学

到稿日期:2024-10-21 返修日期:2025-03-04

基金项目:国家自然科学基金(62572062, 62525204);北京市自然科学基金(L232102)

This work was supported by the National Natural Science Foundation of China(62572062, 62525204) and Beijing Natural Science Foundation(L232102).

通信作者:宋文凤(songwenfenga@163.com)

图像包含比自然图像更复杂的语义信息,专业性更强,因此提升医学图像特征提取质量一直是一个难题。现有工作^[4-6]主要关注图像的语义特征提取,往往忽视了对图像隐含信息的利用,如位置、形态和类别等。

另外,医学图像标注的专业性太强,数据标注困难,导致医学问答数据集规模较小。针对该类问题,许多研究通过改进预训练图像编码器来提升特征提取的效果^[7-9],然而,预训练模型通常依赖额外的资源进行训练。为了进一步提高模型的泛化能力,Liu等^[10]引入数据增强技术,在不增加额外数据的情况下丰富数据源,但也在一定程度上产生了噪声。本文将其作为基线模型,对数据增强方法做出改进。

针对上述两个挑战,本文提出了一种数据知识双增强的医学视觉问答模型(Data and Knowledge Enhanced Network, DKEN)。DKEN包含多模态条件混合(Multimodal Conditional Mixup, MCM)和图像位置识别器(Image Location Recognizer, ILR)。MCM将图像和问题进行线性组合,并通过问题类别对训练样本对进行约束,实现低噪声的数据增强,丰富数据的多样性。ILR通过结合位置识别网络(Location Recognition Network, LRNet)和残差块提取图像的位置特征,作为图像特征和问题特征的引导与补充,实现知识增强。在两个基准数据集 SLAKE 和 VQA-RAD 上的实验结果表明,本文方法优于之前的方法,并实现了先进的性能。

2 相关工作

2.1 Med-VQA 相关工作

Med-VQA 任务由 ImageCLEF Med-VQA 2018 竞赛提出,通常采用联合嵌入方法来处理医学图像和临床问题。一般来说,模型从图像和问题中提取模态特征,然后融合提取的视觉特征与问题嵌入以预测答案或生成答案。大多数模型使用卷积神经网络(CNN)作为视觉编码器,如 VGG Net^[11]和 ResNet^[12],并利用 LSTM^[4]或基于 Transformer^[5]的模型(如 BERT^[6])提取给定问题的特征。特征编码器通常使用预训

练的权重进行初始化,然后训练期间进行冻结或微调。早期的方法常采用矩阵的拼接、张量和、点积等方式进行特征融合^[3]。之后的工作引入了注意力机制提高了模型性能,如堆叠注意网络(SAN)^[13]、双线性注意网络(BAN)^[14]等。答案预测器通常是神经网络分类器或递归神经网络语言生成器。

在最新的研究中,Chen等^[7]使用 VQA-Med-2019 数据集^[15]作为附加训练数据。Eslami等^[16]在外部医学数据的支持下预练了一个 Med-VQA 图像编码器。Allaouzi等^[17]提出的模型以及 Mmbert^[18]借助迁移学习分别在图像数据集 CheXpert^[19]和 ROCO^[20]上训练编码器。Nguyen等^[8]利用元学习思想初始化图像编码器的模型权重,克服了标记数据量的局限性。CPRD^[9]引入对比学习技术,在自监督方案中对未标记的图像进行预训练,降低了模型对医学图像标签的依赖。以上方法对图像编码器进行预训练,通过改进图像特征的提取来提升问答准确率,但是提取到的图像特征仅包含视觉特征,且预训练需花费更多的资源。

2.2 Med-VQA 中的数据增强

为了应对 Med-VQA 领域数据集规模小的挑战,数据增强方法被应用于 Med-VQA 中。Liu等^[10]提出的 M-Mixup 分别对图像和问题进行线性混合,解决了训练样本不足的问题,但未考虑混合过程中产生的噪声。为了提高混合过程的合理性和可解释性,Gong等^[21]提出了非线性混合的 VQAMix 模型来过滤无意义的答案,但是非线性混合的函数更加复杂,混合的比例难以精准把控,因此生成的数据可能失去原始数据的重要特征,难以确保数据的一致性。

3 数据知识双增强的网络结构

本文提出一种基于 MCM 和 ILR 的数据知识双增强的医学视觉问答网络 DKEN。其中,MCM 借助约束条件实现数据增强,ILR 提取图像的位置特征作为多模态特征的知识增强,框架如图 1 所示。

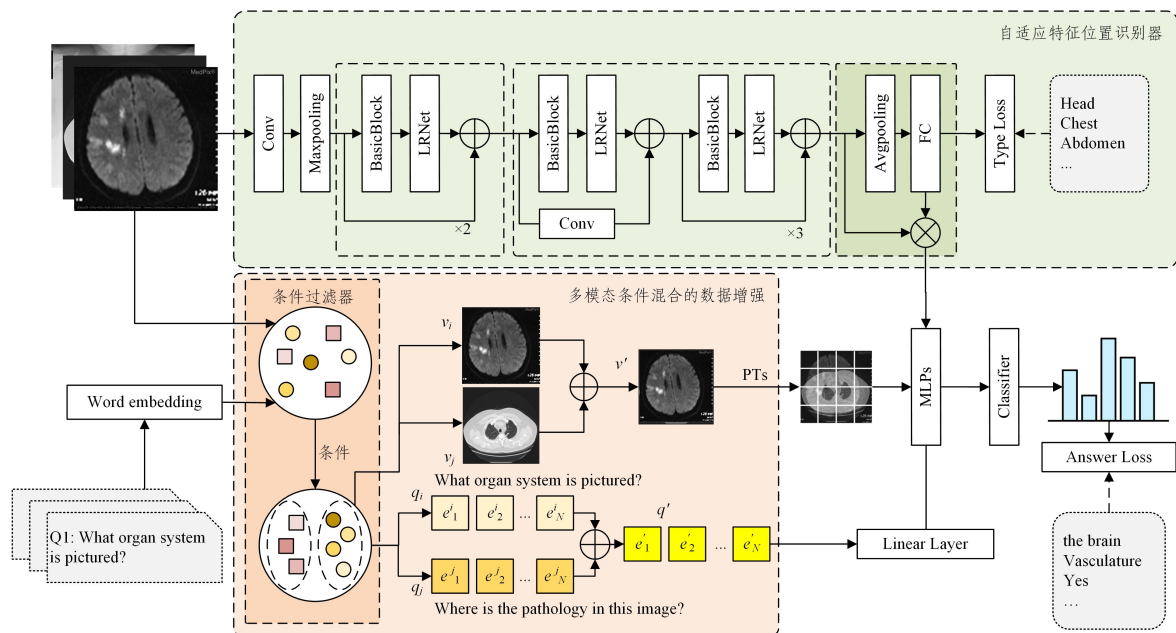


图 1 基于 MCM 和 ILR 的 Med-VQA 网络总体架构

Fig. 1 Overall architecture of Med-VQA network based on MCM and ILR

3.1 网络整体结构

DKEN 包含多模态特征提取、特征融合和答案预测 3 个部分。总流程公式如下：

$$\mathbf{Y}' = \text{Classification}(\text{MLPs}(\mathbf{F}_v, \mathbf{F}_t, \mathbf{F}_q)) \quad (1)$$

其中, \mathbf{Y}' 表示预测答案标签; $\text{Classification}(\cdot)$ 表示线性分类器; $\text{MLPs}(\cdot)$ 表示融合网络; $\mathbf{F}_v, \mathbf{F}_t$ 和 \mathbf{F}_q 分别表示图像特征、图像位置特征和问题特征。

令 $\mathbf{I} \in R^{H \times W \times C}$ 和 $\mathbf{Q} \in R^L$ 分别表示输入的图像和问题, 模型借助 MCM 得到数据增强后的图像 $\mathbf{I}' \in R^{H \times W \times C}$ 和问题序列 $\mathbf{Q}' \in R^L$, 其中 H 和 W 分别表示输入图像的高度和宽度, C 表示图像的通道数, L 表示问题序列的长度。然后使用改进后的池化令牌模块 (Pooling Tokens, PTs)^[10] 提取深层次视觉特征 $\mathbf{F}_v \in R^{M \times D}$, 使用 GloVe^[22] 提取问题特征 $\mathbf{F}_q \in R^{L \times E}$, 其中 M 表示可标记的数量, E 是词向量的维度; 利用 ILR 提取图像位置特征 $\mathbf{F}_t \in R^{N \times D}$ 作为知识增强, 其中 N 表示图像类别数量, D 表示通道数。特征提取的表达式如式 (2)–(4) 所示:

$$\mathbf{F}_v = G^1(\mathbf{I}') \quad (2)$$

$$\mathbf{F}_q = G^2(\mathbf{Q}') \quad (3)$$

$$\mathbf{F}_t = G^3(\mathbf{I}) \quad (4)$$

其中, $\mathbf{F}_v, \mathbf{F}_q$ 和 \mathbf{F}_t 分别表示图像特征、问题特征和图像位置特征; $G^1(\cdot), G^2(\cdot)$ 和 $G^3(\cdot)$ 分别表示提取图像特征、问题特征和图像位置特征的函数; \mathbf{I}' 和 \mathbf{Q}' 分别表示数据增强后的图像和问题特征, \mathbf{I} 表示原图像特征。

为了让模型学习输入信息的模态信息, 在特征融合时添加视觉模态类别嵌入 \mathbf{F}_{vm} 和文本模态类别嵌入 \mathbf{F}_{qm} , 其中 $\mathbf{F}_{vm}, \mathbf{F}_{qm} \in R^D$, 初始化为零向量, 且在训练期间, 借助损失函数反向传播进行更新。融合网络的输入特征表达式如式 (5) 所示, 即将以上特征进行拼接融合得到融合网络初始特征 \mathbf{z}_0 :

$$\mathbf{z}_0 = [\mathbf{F}_v + \mathbf{F}_{vm}; \mathbf{F}_t; \mathbf{F}_q + \mathbf{F}_{qm}] \quad (5)$$

其中, $\mathbf{F}_v, \mathbf{F}_q$ 和 \mathbf{F}_t 表示输入的多模态特征, \mathbf{F}_{vm} 和 \mathbf{F}_{qm} 表示视觉和文本模态嵌入, $[\cdot]$ 表示矩阵级联操作, $[\cdot]$ 表示矩阵加法操作。

特征融合网络由多个多层感知器 (Multilayer Perceptron, MLP) 组成, 并利用残差连接保证模型的稳定性和有效性。特征融合的表达式如式 (6) 和式 (7) 所示:

$$\mathbf{z}_k' = \text{FC}(\text{GELU}(\text{FC}(\text{LN}(\mathbf{z}_{k-1})))) \quad (6)$$

其中, \mathbf{z}_k' 和 \mathbf{z}_{k-1} 分别表示单个 MLP 的输出和输入特征, FC 表示全连接层, GELU 表示激活函数, LN 表示基于 LayerNorm 的归一化层。

$$\mathbf{z}_k = \mathbf{z}_k' + \mathbf{z}_{k-1}, k = 1, \dots, K \quad (7)$$

其中, \mathbf{z}_k 表示特征融合网络的最终特征, \mathbf{z}_k' 和 \mathbf{z}_{k-1} 分别表示单个 MLP 的输出和输入特征, K 表示 MLP 个数, k 的取值为 $1 \sim K$, $+$ 表示向量加法。

在答案预测时, 计算每个通道上的平均特征, 并使用线性分类器预测最终的答案标签 \mathbf{Y}' 。预测过程表达式如式 (8) 和式 (9) 所示:

$$\bar{\mathbf{z}}_k = \frac{1}{M+L} \sum_{k=0}^{M+L-1} (\mathbf{z}_k) \quad (8)$$

其中, $\bar{\mathbf{z}}_k$ 表示融合后特征 \mathbf{z}_k 的平均值, M 表示可标记的

数量, L 表示问题序列的长度, $\sum_{k=0}^{N+L-1} (\cdot)$ 表示从 0 到 $N+L-1$ 对 (\cdot) 进行求和。

$$\mathbf{Y}' = \text{FC}(\text{ReLU}(\text{LN}(\text{FC}(\bar{\mathbf{z}}_k)))) \quad (9)$$

其中, \mathbf{Y}' 表示预测答案标签, $\bar{\mathbf{z}}_k$ 表示分类器的输入, FC 表示全连接层, ReLU 表示激活函数, LN 表示基于 LayerNorm 的归一化层。

3.2 多模态条件混合的数据增强

为了应对医学问答数据集规模小的挑战, 受 Mixup^[23] 的启发, 本文提出了以问题类别作为约束条件的数据增强模块 MCM, 一定程度上丰富了训练数据的多样性。值得注意的是, 不同于视觉信息的直接混合, 本文首先用零将句子填充到相同的长度, 然后使用 GloVe 转换为单词嵌入, 然后在每个单词维度上进行线性插值。具体的流程如算法 1 所示。

将一对样本 (v_i, q_i, y_i, t_i) 和 (v_j, q_j, y_j, t_j) 合成为 (v, q, y, t) 的表达式如下:

$$\begin{aligned} (v, q, y, t) = & (\lambda v_i + (1-\lambda)v_j, \lambda q_i + (1-\lambda)q_j, \lambda y_i + \\ & (1-\lambda)y_j, \lambda t_i + (1-\lambda)t_j) \end{aligned} \quad (10)$$

其中, (v, q, y, t) 表示混合后的输入信息; λ 是混合策略的混合比, $\lambda \sim \text{Beta}(\alpha, \beta)$; v 和 q 分别表示输入的视觉编码和文本编码; y 是答案标签; t 是图像类别; i 和 j 表示样本序号; $+$ 表示向量加法。

在数据混合过程中, 标签 y 中存在噪声分量, 这可能会对深度神经网络的性能产生负面影响。经过分析, 不同领域的答案混合会引起噪声。为了解决这个问题, 本文提出了一种抑制混合数据生成噪声的策略, 称为“条件混合”。

针对 Med-VQA 中两种类型的数据源 (图像和文本), 本文提出 3 种条件混合的方法, 分别是: 1) 仅混合具有相同成像模式的 (v, q, y, t) 元组; 2) 仅混合具有相同问题类别的 (v, q, y, t) 元组; 3) 将具有同一图像模式和问题类别的 (v, q, y, t) 元组进行混合。

通过分析可得出结论: 问题和答案在潜在空间上更为接近, 问题类型可以直接反映答案的类型; 相同模态的图像包含更加相似的器官和疾病类别, 不同模态的图像进行混合可以提高图像的多样性, 使模型学习到更多器官和疾病信息; 数据集中有许多关于图像的模式和器官的问题, 如果以图像类别作为约束条件会减少训练过程中的不确定性, 更容易过拟合。因此, 提出多模态问题条件混合, 即将相同问题类别的两个 (v, q, y, t) 元组进行混合, 增强混合过程的可解释性; 同时, 模型可以从不同模式的图像中学习模式特征, 也可以丰富特征的多样性。

根据问题类别 (封闭式和开放式) 划分类别问题集 Q 作为条件问题约束, 定义为:

$$Q_c = \{q, q' \in Q \mid \text{category}(q) = \text{category}(q')\} \quad (11)$$

其中, $\text{category}(q)$ 和 $\text{category}(q')$ 代表问题集合 Q 中随机问题 q 和 q' 的问题类别, 是通过相应数据集中的问题类型获得的。

算法 1 MCM

输入: 图像 v_1, v_2 ; 问题的 GloVe 嵌入 q_1, q_2 ; 正确答案标签 (one-hot 向量) y_1, y_2 ; 图像类别标签 (one-hot 向量) t_1, t_2 ; 两个数据加载器

loader1,loader2;用于 Beta 分布的超参数 α, β

输出:预测的答案标签 y' ; 预测的图像类别标签 t'

```

1. for( $v_1, q_1, y_1, t_1$ ), ( $v_2, q_2, y_2, t_2$ ) in zip(loader1, loader2) do
2.   if category( $q_1$ ) = category( $q_2$ ) then
3.     //初始化混合权重参数  $\lambda$ 
4.      $\lambda \leftarrow \text{numpy.random.beta}(\alpha, \beta)$ 
5.      $v \leftarrow \lambda * v_1 + (1-\lambda) * v_2$  //混合后的图像
6.      $q \leftarrow \lambda * q_1 + (1-\lambda) * q_2$  //混合后的问题
7.      $y \leftarrow \lambda * y_1 + (1-\lambda) * y_2$  //混合后的答案
8.      $t \leftarrow \lambda * t_1 + (1-\lambda) * t_2$  //混合后的类别标签
9.   //优化器的梯度清零
10.  optimizer.zero_grad()
11.  //通过模型获取预测的答案和类别标签
12.  ( $y', t'$ )  $\leftarrow$  model( $v, q$ )
13.  lossy  $\leftarrow$  criterion1( $y', y$ ) //计算答案损失
14.  losst  $\leftarrow$  criterion2( $t', t$ ) //计算类别标签损失
15.  loss  $\leftarrow$  lossy + losst //总损失为两者的和
16.  loss.backward() //反向传播计算梯度
17.  optimizer.step() //更新模型参数
18. end if
19. end for

```

3.3 自适应特征位置识别器

为了增强模型提取多模态特征的能力,本文设计了用于提取图像位置特征的 ILR 来实现知识增强。ILR 在 ResNet18 的基础上使用 LRNet 放大医学图像的中间特征,在不增加网络深度的情况下,自适应地改进了 ILR 的特征提取能力。具体而言,在挤压和激励网络 (SENet)^[24] 的基础上设计了 LRNet,并将其添加在 ResNet 基础块 (BasicBlock) 之后,与卷积和池化层一起组成新的残差网络。为了进一步增强模型的鲁棒性,将得到的模式类别特征与图像中间特征融合,然后作为外部图像特征输入骨干网络。

图像特征首先完成从 $X \rightarrow U$ 的卷积变换,其中, $X \in R^{h' \times l' \times c'}$, $U \in R^{h \times l \times c}$, $h(h')$ 和 $l(l')$ 分别表示图像的长和宽, c 表示图像特征的通道数。然后特征图 $U = [u_1, u_2, \dots, u_c]$ 沿着空间维度压缩,再通过全局平均池化生成包含通道的空间信息描述符 z 。在这个过程中,每个通道的二维特征 ($h \times l$) 被压缩为 1 个实数,特征图从 (h, l, c) 转换为 ($1, c$), 表示为:

$$z = F_{sq}(u_c) = \frac{1}{h \times l} \sum_{i=1}^h \sum_{j=1}^l u_c(i, j) \quad (12)$$

其中, z 是通道描述符, $F_{sq}(\cdot)$ 表示压缩函数, u_c 是大小为 $h \times l$ 的第 c 层特征图, $u_c(i, j)$ 表示在第 c 个通道上的坐标为 (i, j) 的像素值, $\sum_{i=1}^h \sum_{j=1}^l (\cdot)$ 表示从坐标 ($1, 1$) 到 (h, l) 对 $u_c(\cdot)$ 进行求和, h 和 l 表示图像的长和宽。

为了获得通道级的注意力权重,设置了两个完全连接层和两个激活函数传递通道信息,表示为:

$$w = F_{er}(z, W) = \sigma(W_2 \delta(W_1 z)) \quad (13)$$

其中, w 是通道方向的注意力权重, $F_{er}(\cdot)$ 表示激励函数, z 和 W 表示函数参数, σ 是 ReLU 激活函数, δ 是 Sigmoid 激活函数, W_1 和 W_2 是两层全连接层。

将通道注意力权重应用于从前面的层中导出的特征图。与传统的 SENet 不同,本文加入了残差块来强化通道特征。

缩放操作表示为:

$$u_c' = (u_c \times w) \times u_c \quad (14)$$

其中, u_c' 是权重更新后的特征图, u_c 表示第 c 层通道上的特征图, w 是通道方向的注意力权重, \times 表示两向量的叉乘。

与普通的 SE-ResNet 不同,本文修改了最后一层的输出,使图像位置特征同时受到图像类别和图像中间特征的影响。最终,经过 ILR 网络生成的特征 F_i 可以表示为:

$$F_i = T_c \otimes FC(Avg(T_c)) \quad (15)$$

其中, F_i 表示 ILR 网络的输出, T_c 表示卷积层最终输出的特征, \otimes 表示矩阵拼接操作, Avg 表示平均池化操作, FC 表示全连接层。

4 实验验证

4.1 数据集

为全面评估本文方法的有效性,在两个广泛使用的 Med-VQA 基准数据集 SLAKE^[25] 和 VQA-RAD^[26] 上进行了实验。这两个数据集的详细信息如表 1 所列。

表 1 SLAKE 和 VQA-RAD 数据集详细信息

Table 1 Details of SLAKE and VQA-RAD datasets

| 数据集 | 数据类别 | 训练集 | 验证集 | 测试集 | 模态 |
|---------|------|------|------|------|----|
| SLAKE | 图像 | 450 | 96 | 96 | 3 |
| | 问答对 | 9849 | 2109 | 2070 | |
| VQA-RAD | 图像 | 315 | — | 315 | 3 |
| | 问答对 | 3064 | — | 451 | |

SLAKE 数据集是一个既有语义标签又有结构医学知识库的综合双语数据集,包含 642 张放射学图像,包括 12 种疾病和 39 个全身器官。VQA-RAD 是医学 VQA 领域人工构建的数据集,放射学图像来自 MedPix 开放获取数据库,包含 11 个类别的问题,每个图像都与多个问题相关联。SLAKE 和 VQA-RAD 中的问题分为“开放式”和“封闭式”两类。开放式是指答案没有固定结构,存在多个备选答案;而封闭式的答案形式有限,如“YES/NO”。图 2 给出了一个样本图像及其对应的问答对示例。



图像

问题: 这张图片中的右下器官是什么?
 答案: 脾脏
 问题: 这张照片看起来正常吗?
 答案: 没有
 问题: 这张照片是用什么方式拍摄的?
 答案: CT

图 2 数据集集中的图像和问答对示例

Fig. 2 Examples of image and question and answer pairs in datasets

4.2 实验参数设置

本文中 PTs 模块由 4 层卷积和池化层组成,在每层 PT 中,卷积核数为 3、步长为 1,池化层核数为 2、步长为 2。输入图像大小为 224×224 ,提取的图像特征维度为 $14 \times 14 \times 768$,

输入的问题特征维度为 20×768 。在数据增强部分,选取了混合权重参数 $\lambda \sim \text{Beta}(5, 1)$,并在之后的消融实验中对 λ 进行讨论。

本文使用学习率为 0.005、权重衰减为 0.001 的 AdamW 优化器^[27]进行优化, batch size 设为 128,每次训练 5 000 轮次。所有训练均在单个内存为 24GB 的 NVIDIA GeForce RTX 3090 GPU 上进行。

4.3 实验对比分析

4.3.1 定量评估

为了证明本文模型的有效性,将其与一些现有的 SOTA 方法进行了对比,包括 BAN^[14], MEVF^[8], CPRD^[9], MMixup^[10], VQAMix^[21], PubMedCLIP^[28] 和 MISS^[29]。评价指标采用准确率 (Accuracy, ACC),其中大多数模型将 VQA 视为答案分类或排序任务,而 MISS 模型将 Med-VQA 视为生成任务,因此,使用自动评估方法评估封闭式问题,而对于开放式问题,将生成的回答与基本事实答案进行人工评估。表 2 和表 3 分别列出了在 SLAKE 和 VQA-RAD 数据集上的实验结果。

表 2 在 SLAKE 数据集上不同方法的实验结果

| 方法 | 开放式准确率 | 封闭式准确率 | 总体准确率 (%) |
|----------------------------|--------------|--------------|--------------|
| BAN ^[14] | 74.60 | 79.10 | 76.30 |
| MEVF-SAN ^[8] | 75.30 | 78.40 | 76.50 |
| MEVF-BAN ^[8] | 77.80 | 79.80 | 78.60 |
| CPRD-BAN ^[9] | 79.50 | 83.40 | 80.10 |
| MMixup ^[10] | 79.38 | 85.82 | 81.90 |
| PUBMEDCLIP ^[28] | 78.40 | 82.50 | 80.10 |
| MISS ^[29] | 82.91 | 81.47 | 82.00 |
| Ours | 80.16 | 86.54 | 82.66 |

注:加粗表示最优结果。

表 3 在 VQA-RAD 数据集上不同方法的实验结果

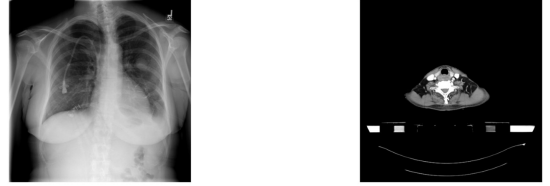
| 方法 | 开放式准确率 | 封闭式准确率 | 总体准确率 (%) |
|----------------------------|--------------|--------------|--------------|
| BAN ^[14] | 37.40 | 72.10 | 58.30 |
| MEVF-SAN ^[8] | 49.20 | 73.90 | 64.10 |
| MEVF-BAN ^[8] | 49.20 | 77.20 | 66.10 |
| CPRD-BAN ^[9] | 52.50 | 77.90 | 67.80 |
| MMixup ^[10] | 53.10 | 81.30 | 70.20 |
| VQAMix ^[21] | 56.60 | 79.60 | 70.40 |
| PUBMEDCLIP ^[28] | 60.10 | 80.00 | 72.10 |
| MISS ^[29] | 71.81 | 80.35 | 76.05 |
| Ours | 53.63 | 81.98 | 70.73 |

注:加粗表示最优结果。

与其他的视觉语言模型相比,本文模型在 SLAKE 数据集上封闭式问题和全类别问题中均达到了最高的准确率,分别为 86.78% 和 82.66%,超过了所有采用答案分类和生成任务的方法。对于 VQA-RAD 数据集,本文模型在封闭式问题上的准确率达到 81.98%,比第二名高出 0.68 个百分点。

本文模型在个别指标上低于 MISS 模型,这主要是由于 MISS 模型使用人工评估方式计算准确率,存在主观性偏差。如图 3 所示,MISS 的预测结果来自其原论文中的数据^[29]。图 3(a)中,与 MISS 模型预测结果相比,本文方法的预测结果

包含更多真实答案中的信息;图 3(b)中,本文方法的预测结果与真实答案完全匹配。除此之外,由于开放式的答案种类较多,语言更加灵活,在较少的数据量上训练无法使模型学习到更多的答案特征,这也是本文模型指标略低的原因之一。



问题:如何预防该图像中左肺叶疾病的发生?
真实答案:戒烟,增强体质
本文模型:注意防寒保暖,增强体质
MISS模型:注意保暖防寒。

(a)

问题:黑色中空组织在人体中扮演什么角色?
真实答案:气体输送
本文模型:气体输送
MISS模型:通风。

(b)

图 3 在 SLAKE 数据集上预测结果的可视化

Fig. 3 Visualization of prediction results on SLAKE dataset

实验数据表明,本文方法在处理封闭式问题上有显著优势,与生成任务方法相比,本文方法在处理开放式问题上仍有提升空间。

4.3.2 定性评估

为了验证本文模型的有效性,对问答结果进行了定性分析。从 SLAKE 数据集和 VQA-RAD 数据集中选取了 6 个样例,在图 4 中对基线模型和本文模型对答案的预测情况进行比较分析,用绿色和红色区分预测答案的正确和错误情况。

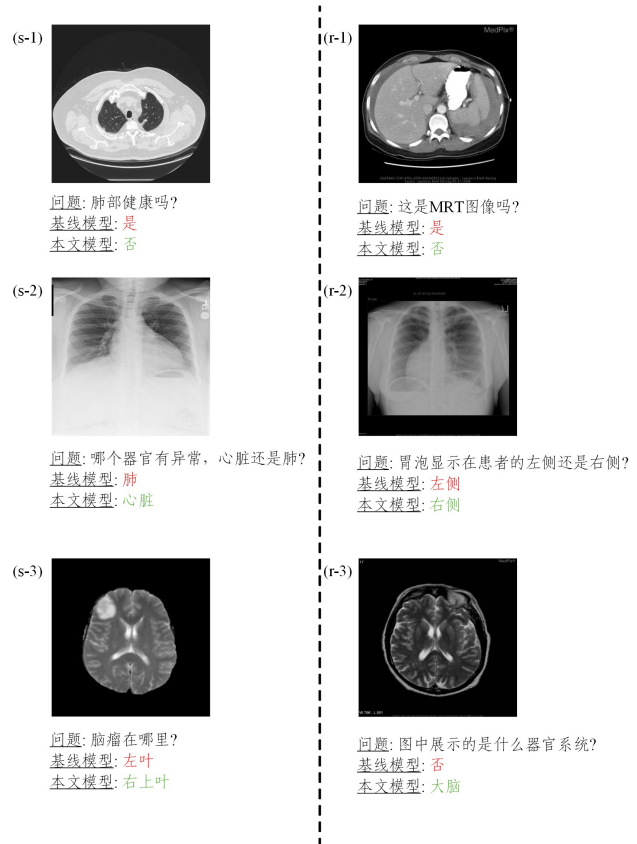


图 4 在 SLAKE 和 VQA-RAD 数据集上的预测结果

(电子版为彩图)

Fig. 4 Prediction result on SLAKE and VQA-RAD datasets

其中,左边的3个示例来自 SLAKE 数据集,右边的3个示例来自 VQA-RAD 数据集。为了保证分析结果的合理性和均衡性,分别选取了封闭式和开放式问题,问题涉及器官、疾病、方位和数量等不同方面。

从图4中的示例可以看出,基线模型^[10]预测出错误答案的原因是基线模型不能准确定位异常信息,如示例(s-1)(s-2)和(r-1)所示。相比之下,本文模型通过 ILR 模块学习不同图像类别的位置特征,增加了模型预测的准确性。对于示例(s-3)和(r-2),基线模型输出与基本事实相反的答案,因为它只关注到了图像中的左侧,而忽略了图像中的左侧代表着人体的右侧这一事实情况。然而,所提模型通过学习到更全面的多模态信息,提供了一个准确的答案。该方法在位置相关问题上表现出色,这是由于 ILR 模块学习到了图像位置特征,增强了模型对图像的隐式特征的理解。在示例 r3 中,基线模型预测的答案与真实答案的类别有很大差别,这是因为基线模型在不合理的数据混合中产生了噪声数据,这也反映了条件约束在有效数据混合中的关键作用。上述示例表明,本文模型通过使用图像位置信息提取和多模态条件数据增强,可以有效地提取丰富而准确的多模态特征,从而提升在 Med-VQA 任务中的正确性。

4.4 消融实验

本节通过消融研究以评估本文方法中每个模块的影响。表4和表5分别列出了两个基准数据集上的实验结果。可以看出,在去掉 MCM 模块(w/o MCM)后,两个数据集的所有类别问题中的准确率都大幅度降低,证明了 MCM 模块能够丰富数据集的多样性,有效提升模型性能。在去掉 ILR 模块(w/o ILR)后,开放式问题的准确率大幅度降低,说明 ILR 模块能够有效提取图像中位置相关的特征,帮助模型更准确地理解图像信息,特别是在处理需要位置感知的开放式问题时效果显著,可以参考图5中的(s-b)。当同时去掉 MCM 和 ILR 模块(w/o MCM & ILR)后,模块的性能进一步下降,且在所有评估指标上表现最差。这表明,MCM 和 ILR 模块是模型性能提升的关键组成部分,两者相辅相成,共同优化模型对多模态和位置特征的处理能力。

表4 在 SLAKE 数据集上的消融实验

Table 4 Ablation studies on SLAKE dataset (%)

| 方法 | 开放式准确率 | 封闭式准确率 | 总体准确率 |
|-------------|--------------|--------------|--------------|
| w/o MCM&ILR | 77.98 | 81.25 | 79.26 |
| w/o MCM | 78.29 | 82.21 | 79.83 |
| w/o ILR | 78.76 | 86.30 | 81.72 |
| Ours | 80.16 | 86.54 | 82.66 |

表5 在 VQA-RAD 数据集上的消融实验

Table 5 Ablation studies on VQA-RAD dataset (%)

| 方法 | 开放式准确率 | 封闭式准确率 | 总体准确率 |
|-------------|--------------|--------------|--------------|
| w/o MCM&ILR | 41.9 | 72.79 | 60.53 |
| w/o MCM | 37.43 | 76.47 | 60.98 |
| w/o ILR | 55.31 | 80.51 | 70.51 |
| Ours | 53.63 | 81.98 | 70.73 |

VQA-RAD 数据集上,本文方法的开放式准确率低于“w/o ILR”,这主要是 VQA-RAD 数据集的开放式问答对数

量过少(1242 个问答对),ILR 模块引入的额外位置特征提取未能完全发挥作用,导致在开放式问题上的表现略差。

但从总体性能来看,ILR 模块依然显著提升了模型的封闭式问题表现和总体准确率,使模型达到平衡的状态。其价值依然得到了验证。

为了进一步阐明 MCM 和 ILR 模块的有效性,图5中展示了在6个 Med-VQA 实例中消融实验的可视化结果,用绿色和红色区分预测答案的正确和错误情况。图5中左侧3个案例来自 SLAKE 数据集,右侧案例来自 VQA-RAD 数据集。在示例(s-b)和(r-c)中,两个模块都能在不同程度上修正模型的响应。从示例(r-a)中可以看出,所提模块有助于模型准确地响应两个基准的封闭式和开放式查询。结合示例(s-a)(s-c)和(r-b)的结果可以看出,ILR 模块在 SLAKE 数据集上的提升效果更好,而 MCM 模块在 VQA-RAD 数据集上能发挥更大的优势,分析其原因可能是 VQA-RAD 数据集的问题种类更为单一,有较多的重复,这也体现出 MCM 模块在丰富数据多样性方面的优势。总体而言,图5中的案例共同证明了 ILR 和 MCM 模块显著提高了模型在应对 Med-VQA 挑战方面的性能。

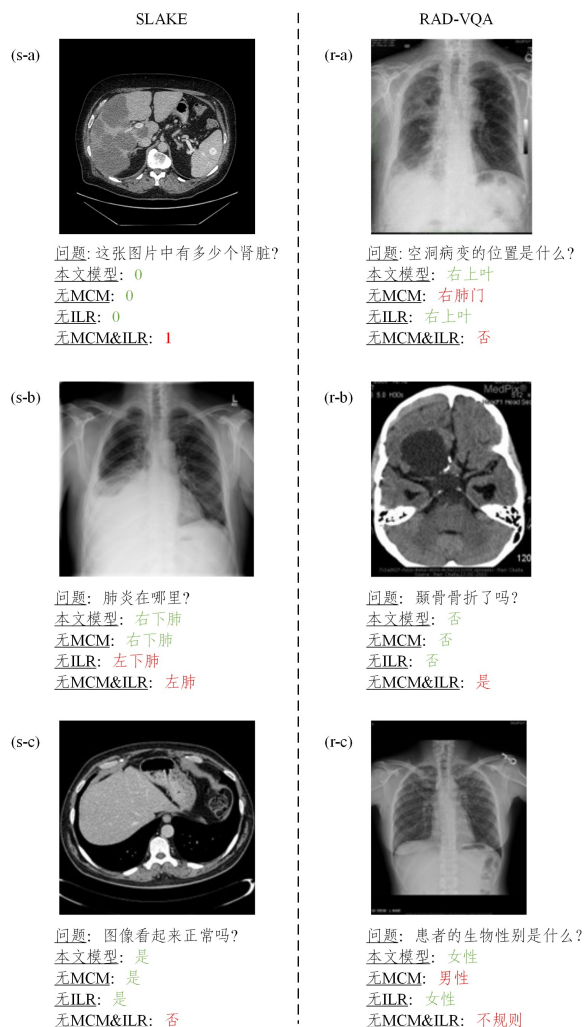


图5 在 SLAKE 和 VQA-RAD 数据集上的消融实验结果 (电子版为彩图)

Fig. 5 Ablation results on SLAKE and VQA-RAD datasets

4.4.1 MCM 中 λ 取值对模型的影响

为了确定 MCM 中最好的 $\lambda \sim \text{Beta}(\alpha, \beta)$ 取值, 比较了不同 Beta 分布对模型性能的影响。表 6 的结果表明, 适当选择 Beta 分布能够积极提升 Med-VQA 任务的性能。Med-VQA 任务要求模型理解两种模态的信息并进行推断。由于新样本是由两个训练样本混合产生的, 因此 Beta 分布的选择直接影响着这些合成样本的质量。当 $\lambda \sim \text{Beta}(0.5, 0.5)$ 时, 两个高比例样本的插值生成了复杂的合成样本, 这些样本中嵌入了过多冗余信息, 导致模型难以学习到有效特征并影响了收敛性。实验结果显示, 随着插值样本的比例差增大, 模型准确率在一定范围内有所提升。经过实验验证, 当 $\lambda \sim \text{Beta}(5, 1)$ 时, 模型表现最佳, 这归因于平衡的插值样本的比例对模型的泛化能力的正向作用。然而, 当 α 参数增大至 $\lambda \sim \text{Beta}(10, 1)$ 时, 模型的效果反而下降, 过高的插值样本比例导致某一样本的影响减弱, 容易变为噪声数据。因此, 选择合适的 Beta 分布有助于提高模型的泛化能力, 从而提高其性能。

表 6 在 SLAKE 数据集上 MCM 模块参数的消融实验

Table 6 Ablation studies of MCM parameters on SLAKE dataset (%)

| $\lambda = \text{Beta}(\alpha, \beta)$ | 开放式准确率 | 封闭式准确率 | 总体准确率 |
|--|--------------|--------------|--------------|
| (0.5, 0.5) | 78.76 | 85.82 | 81.53 |
| (1, 1) | 79.84 | 86.54 | 82.47 |
| (5, 1) | 80.16 | 86.54 | 82.66 |
| (10, 1) | 78.76 | 84.13 | 80.87 |

4.4.2 对数据增强方法的消融实验

为了证明本文所提出的数据增强方法的优越性, 在 SLAKE 数据集上对不同的数据增强方法进行了评估, 实验结果如表 7 所列。其中, “线性混合”方法使用 MMixup 中的数据增强方法, 将两个训练样本按权重参数 λ 进行线性加权相加, 而“非线性混合”方法则采用 VQAMix 中的数据增强方法, 在计算过程中引入权重参数 λ^2 , 通过非线性函数进行样本的混合。在传统线性混合方法的基础上, 本文引入了约束条件, 提出了一种新的数据增强方法, 以进一步优化混合过程并提升模型性能。从表 7 中的实验结果可以看出, 本文提出的数据增强方法在各项指标上均取得了最佳效果。

表 7 在 SLAKE 数据集上数据增强方法的消融实验

Table 7 Ablation studies of data enhancement methods on SLAKE dataset (%)

| 数据增强方法 | 开放式准确率 | 封闭式准确率 | 总体准确率 |
|--------|--------------|--------------|--------------|
| 线性混合 | 80.00 | 85.58 | 82.19 |
| 非线性混合 | 79.69 | 86.30 | 82.28 |
| 本文方法 | 80.16 | 86.54 | 82.66 |

结束语 针对 Med-VQA 中多模态特征提取质量低和数据集规模小的问题, 本文提出了一种新颖的数据知识双增强的 Med-VQA 模型 DKEN。该模型使用数据增强模块 MCM 将图像问题对在问题类别约束下进行线性组合, 丰富了数据集的多样性, 避免了混合过程中产生额外的噪声。通过将 LRNet 和残差块结合提出 ILR, 有效提取图像位置特征, 并将其编码到图像特征和问题特征的融合过程中, 利用知识增强使得提取到的多模态特征更详细、丰富。在 SLAKE 和

VQA-RAD 数据集上进行的大量实验表明, 本文模型在 Med-VQA 任务中具有优越性。

Med-VQA 的可解释性可以为医生的诊断提供更加可靠的回答依据和更加清晰的推理思路, 在一定程度上保证了深度学习模型的安全性, 这也是当前视觉问答领域的研究热点。为了进一步提升模型的可解释性, 计划集成一个推理模块, 这对模型的实际部署和临床应用具有重要意义。此外, 本文使用的公开医学问答数据集 SLAKE 和 VQA-RAD 在数据集规模、病种和问题类型的多样性方面仍有提升空间。因此, 提升基于少量数据集训练的模型准确率, 将是未来工作的一个重要方向。

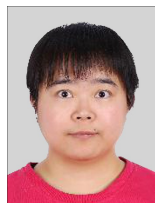
参考文献

- [1] ANTOL S, AGRAWAL A, LU J S, et al. VQA: Visual Question Answering[C]// 2015 IEEE International Conference on Computer Vision. 2015:2425-2433.
- [2] ISHMAM M F, SHOYON M S H, MRIDHA M F, et al. From Image to Language: A Critical Analysis of Visual Question Answering(VQA) Approaches, Challenges, and Opportunities[J]. Information Fusion, 2024, 106:102270.
- [3] LIN Z, ZHANG D, TAO Q, et al. Medical Visual Question Answering: A Survey[J]. Artificial Intelligence in Medicine, 2023, 143:102611.
- [4] SCHMIDHUBER J, HOCHREITER S. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.
- [5] 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. 2017:6000-6010.
- [6] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019:4171-4186.
- [7] CHEN G, GONG H, AND LI G. HCP-MIC at VQA-Med 2020: Effective Visual Representation for Medical Visual Question Answering[C]// CLEF(Working Notes). 2020.
- [8] NGUYEN B D, DO T T, NGUYEN B X, et al. Overcoming Data Limitation in Medical Visual Question Answering[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. 2019:522-530.
- [9] LIU B, ZHAN L M, WU X M. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images[C]// Medical Image Computing and Computer Assisted Intervention. 2021:210-220.
- [10] LIU L, SU X. How well apply multimodal mixup and simple mlps backbone to medical visual question answering? [C]// International Conference on Bioinformatics and Biomedicine. 2022:2648-2655.
- [11] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]// Proceedings of the 3rd International Conference on Learning Representations. 2015.

- [12] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.
- [13] YANG Z,HE X,GAO J,et al. Stacked attention networks for image question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 21-29.
- [14] KIM J H,JUN J,ZHANG B T. Bilinear attention networks [C]// Conference on Neural Information Processing Systems. 2018.
- [15] BEN ABACHA A,HASAN S A,DATLA V V,et al. VQA-Med: Overview of the medical visual question answering task at image CLEF 2019 [C]// Proceedings of CLEF 2019 Working Notes. 2019:9-12.
- [16] ESLAMI S,DE MELO G,MEINEL C. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? [J]. arXiv:2112.13906,2021.
- [17] ALLAOUZI I,AHMED M B,BENAMROU B. An Encoder-Decoder Model for Visual Question Answering in the Medical Domain[C]// CLEF. 2019.
- [18] KHARE Y,BAGAL V,MATHEW M,et al. Mmbert: Multimodal bert pretraining for improved medical vqa[C]// International Symposium on Biomedical Imaging. 2021:1033-1036.
- [19] WANG X,PENG Y,LU L,et al. ChestX-Ray8: Hospital-scale chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017:3462-3471.
- [20] PELKA O,KOITKA S,RÜCKERT J,et al. Radiology objects in Context(ROCO): a multimodal image dataset[C]// International Conference on Medical Imaging Computing and Computer-Assisted Intervention. 2018:180-189.
- [21] GONG H,CHEN G,MAO M,et al. VQAMIX: Conditional triplet mixup for medical visual question answering [J]. IEEE Transactions on Medical Imaging,2022,41(11):3332-3343.
- [22] PENNINGTON J,SOCHER R,MANNING C D. GloVe: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014:1532-1543.
- [23] ZHANG H,CISSE M,DAUPHIN Y N,et al. mixup: Beyond empirical risk minimization[J]. arXiv:1710.09412,2017.
- [24] HU J,SHEN L,SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7132-7141.
- [25] LIU B,ZHAN L M,XU L,et al. SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering[J]. arXiv:2102.09542,2021.
- [26] LAU J J,GAYEN S,ABACHA A B,et al. A dataset of clinically generated visual questions and answers about radiology images[J]. Scientific Data,2018,5(1):1-10.
- [27] LOSHCHILOV I,HUTTER F. Fixing weight decay regularization in adam[J]. arXiv:1711.05101,2017.
- [28] ESAI S,MEINEL C,DE MELO G. Pubmedclip: How much does clip benefit visual question answering in the medical domain? [C]// Findings of the Association for Computational Linguistics. 2023:1181-1193.
- [29] CHEN J,YANG D,JIANG Y,et al. MISS: A Generative Pre-training and Fine-Tuning Approach for Med-VQA[C]// International Conference on Artificial Neural Networks. 2024:299-313.



YAN Yujing, born in 1999, postgraduate. Her main research interests include computer vision and visual question answering.



SONG Wenfeng, born in 1987, Ph.D., associate professor, is a member of CCF (No. 71334S). Her main research interests include pattern recognition, computer vision and machine learning.

(责任编辑:何杨)