基于深度学习的医学影像问答模型*

赵晋稷 刘 旻

（1 天津中医药大学第一附属医院 天津 300193 2 国家中医针灸临床医学研究中心 天津 300193）

〔摘要〕 介绍医学影像问答相关研究进展和不足，采用深度学习语义图卷积方法，提出基于语义图的医学影像问答模型，详细阐述该模型构建方法，分析实验结果，指出该模型可提高医学影像问答准确率。

〔关键词〕 医学影像；问答系统；深度学习；语义图；人工智能

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2023.05.013

A Medical Image Question Answering Model Based on Deep Learning *ZHAO Jinji, LIU Min, 1First Teaching Hospital of Tianjin University of Traditional Chinese Medicine, Tianjin 300193, 2National Clinical Research Center for Chinese Medicine Acupuncture and Moxibustion, Tianjin 300193, China*

〔**Abstract**〕 The paper introduces the research progress and shortcomings of medical image question－answering, proposes a medical image question－answering model based on semantic graph by using the deep learning semantic graph convolution method, elaborates the model construction method, analyzes the experimental results, and points out that the model can improve the accuracy of medical image question－answering.

〔**Keywords**〕 medical image; question answering system; deep learning; semantic graph; artificial intelligence (AI)

1 引言

我国卫生资源总体缺乏，优质卫生资源严重不足。医生在临床工作中需要阅读大量医疗检查报告，存在人为错误的可能性，医疗资源不足可能会加剧这一现象。近年来，随着“互联网＋医疗”模式的推广，在线问诊平台发展迅速，患者可以通过平台直接与医生沟通[1－2]。中医问诊平台中，大量舌象图像咨询成为急需解决的问题。伴随人工智能赋能医疗行业，医学领域出现了一系列智能分析系统，如医学影像问答系统，能够依托平台辅助解答并分流大量信息，提高工作效率，减轻医务工作者压力[3]。

2 相关工作

医学影像问答（medical visual question answering, Med－VQA）是医学领域的问答任务。在该过程中，输入医学影像和与之相关的临床问题，将自动输出答案[4]。患者可以在提问后及时得到反馈，医生也可以在诊断疾病时将系统反馈的答案作为参考意见。医学影像问答系统可以节省宝贵的医疗资源，辅助医生诊断。已有医学影像问答系统及相关模型[4]直接将自然场景下的图像问答模型迁移到医

〔修回日期〕 2022－10－26

〔作者简介〕 赵晋稷，硕士研究生；通信作者：刘旻，主任医师，博士生导师，发表论文 50 余篇。

〔基金项目〕 国家重点研发计划“中医药现代化研究”重点专项“基于系统辨证脉学的系列新型智能化脉诊仪研发”（项目编号：2018YFC1707503）。

学场景使用。考虑到自然图像和医学图像中包含的语义有较大差异，有研究者提出注意力堆叠网络（stacked attention networks，SAN）[5]、双线性池化（multimodal compact bilinear，MCB）[6]、增强医学图像中的视觉信息（mixture of enhanced visual features，MEVF）[7]、问题为前提的推理（question - conditioned reasoning，QCR）[8]、双线性注意力网络（bilinear attention networks，BAN）[9]等操作，以缓解数据缺乏问题。已有模型性能较差，主要存在 3 方面问题：一是面对数据分布不相同的自然图片和医学影像，适用于自然图片问答系统的模型在医学影像问答系统并不一定有效；二是医学影像问答任务相关数据集需要人工标注，因此很多数据集包含训练样本较少，限制模型训练效果；三是医学影像问答任务相对于自然图像问答难度更大，因此在完成该任务时需要模型有更强的语义分析和多模态融合处理能力，见表 1。本文在上述工作基础上采用语义图卷积（semantic graph convolution，SGC）进一步获取医学影像和文本之间的关联，从而更好地解决医学影像问答任务。针对问题一，通过元学习和自编码器增强医学影像相关数据，提高模型对噪声的鲁棒性；针对问题二，增加数据可以缓解训练样本较少的问题；针对问题三，通过设计语义图结构，进一步增强提取医学影像和文本之间关联的能力，并通过门控线性单元选择出文本中的重要部分，从而更好地完成医学影像问答任务。

表 1 多种医学影像问答方法的基本原理和缺点

方法	基本原理	缺点
SAN[5]	用堆叠注意力模块提取图像和文本信息	缺少对医学影像数据本身特点的考虑
MCB[6]	用多模态线性压缩融合图像和文本信息	缺少对医学影像数据中噪声的处理
MEVF[7]	用自编码器进行无监督数据增加	没有提取图像和文本信息之间的重要关联
QCR[8]	用条件推理模块选择重要的特征	对文本特征的建模能力不足
BAN[9]	用双线性注意力网络融合图像和文本信息	缺少对医学影像数据不足问题的处理

3 方法

3.1 医学影像问答任务

首先输入一张待诊断图片，如电子计算机断层扫描（computed tomography，CT）图像、舌象图像等；然后输入一个与该图像相关的问题；模型充当医生角色，根据输入图片回答给定问题，从而进行智能问诊。本文提出基于语义图卷积的医学影像问答模型，见图 1。

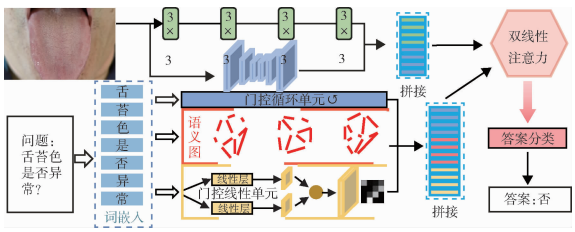

图 1 研究框架

3.2 图像信息处理

输入医学图像 I 之后，对该图像采用如下预处理。一是将图像首先输入 4 层 3×3 的卷积网络，然后进行全局平均池化，并采用元学习（model - agnostic meta - learning，MAML）方法初始化网络权重[10]，最终该操作得到的特征维度为 64。二是采用自编码器对图像进行进一步处理[11]。三是将上述两步操作得到的特征进行拼接，得到图像特征 $V \in R^{1\times 128}$。

3.3 文本信息处理

在实际应用中，当患者将待诊断图像上传后，会提出并输入相应问题 $q \in R^n$，其中 n 为问句长度。对该问句采取如下特征处理过程。

第 1 步：对问句中每个文字进行词嵌入（word embedding，WE），得到语句特征 $Q_e = [w_1, w_2, \cdots, w_n] \in R^{n\times d_e}$。

第 2 步：考虑到语句本身包含上下文序列信息，为了编码整个语句序列特征 $Q_s \in R^{n\times d_s}$，采用门控循环单元（gate recurrent unit，GRU）[12]，该步的输入为上一步得到的词嵌入特征 Q_e，输出为包含上下文序列信息的语句特征 Q_s。

第 3 步：考虑到语句中单个文字之间的关联以及整个语句中的语义结构，将整个语句嵌入到图结构中构成语义图[13]，并采用图卷积网络提取特征，语义图本质上是一种特殊的特征结构方式。首先提取语句之间的关联强度作为语义图的邻接矩阵：

$$A_0 = softmax\left(\frac{(Q_e W_{e1})(Q_e W_{e2})^T}{\sqrt{d_a}}\right) \quad (1)$$

其中 $W_{e1}, W_{e2} \in R^{d_e \times d_a}$，$d_a$ 表示图结构的隐式特征维度，$A_0 \in R^{n \times n}$ 表示该语义图的邻接矩阵。接下来基于邻接矩阵进行图卷积操作：

$$Q_g = A_0 Q_e W_{e1} W_{g0} \quad (2)$$

其中 $W_{g0} \in R^{d_a \times d_e}$ 为可学习的参数，$Q_g \in R^{n \times d_e}$ 为经过语义图嵌入后的语句特征，不仅包含语句文字之间的关联信息，而且融合了语句整体信息。在语义图中，每个文字是 1 个单独的图节点，每个图节点之间边的权重通过关联强度决定，图卷积过程相当于对整个图结构的边不断进行更新，训练完成后得到整个语义图最优结构。

第 4 步：为了进一步提取并突出语句中重要的文字特征，采用门控线性单元（gated linear unit，GLU）[14]：

$$U_1 = \varphi_1(Q_e W_1) \quad (3)$$

$$U_2 = \varphi_2(Q_e W_2) \quad (4)$$

$$Q_l = \varphi_3(U_1 \odot U_2) W_l \quad (5)$$

其中 $W_1, W_2 \in R_e^d \times d_l$ 和 $W_2 \in R_l^d \times d_l$ 为可学习的参数，$\varphi_1, \varphi_2, \varphi_3$ 为非线性激活函数，$\odot$ 表示元素乘法，$Q_l \in R^n \times d_l$ 为得到的重要性相关语句特征。

第 5 步：在得到上下文序列信息的语句特征 Q_s、语句的语义图特征 Q_g、重要性相关语句特征 Q_l 后，将不同层次级别的语句特征进行特征融合，具体实现方式如下：

$$Q_{fea} = [Q_s; Q_g; Q_l] W_q \quad (6)$$

其中，[;] 表示特征拼接操作，$W_q \in R^{(d_s+d_g+d_l)} \times d_q$ 是可学习的参数，$Q_{fea} \in R^{n \times d_q}$ 表示最终提取得到的文本特征。

3.4 问答系统

采用双线性注意力网络对图像特征 V 和文本特征 Q_{fea} 进行融合：

$$y = BAN(V, Q_{fea}) \quad (7)$$

然后通过分类器预测答案的置信分数 s，将概率最大的作为最终结果。

4 实验结果与分析

4.1 实验设置

本文在医学问答数据集（visual questions and answers about radiology images，VQA - RAD）上进行训练和测试[15]，该数据集包含 315 张医学领域相关待诊断图片和 3 515 条医生标注的问诊对话。其中，问诊对话包含两种类型：固定式问答和开放式问答。固定式问答的答案为“是”或“否”两种特定选项，例如问题为“该胸部 CT 图像中是否有异常状况”，答案为“是”；开放式问答的答案没有固定形式，例如问题为“该头颅核磁中的病灶在什么位置”，答案为对应的特定位置。在实验中，采用 3 064 条问答对话作为训练集，451 条对话作为测试集。训练过程中采用自适应动量优化器进行梯度下降优化，学习率为 0.000 1，训练轮数为 150 轮。

4.2 评价指标

$$\text{Accuracy} = \frac{A_c}{A_{all}} \times 100\% \quad (8)$$

其中 A_c 表示正确回答的问题数量，A_{all} 表示整个数据集中的问题数量。为了更准确地分析方法效果和性能，实验结果部分对开放式问答和固定式问答分别进行统计。

4.3 实验结果

本文提出的方法（SGC）与已有方法应用于开放式问答和固定式问答任务的准确率对比结果，见表 2。

表 2 不同方法准确率对比

方法	开放式问答准确率（%）	固定式问答准确率（%）
SAN	24.2	57.2
MCB	25.4	60.6
BAN	27.6	66.5
MEVF + SAN	40.7	74.1
MEVF + BAN	43.9	75.1
QCR	52.8	76.8
SGC	55.6	79.3

可以看到与 QCR 相比，本文方法在开放式问答数据集准确率方面提升 2.8 个百分点，达到 55.6%；在固定式问答数据集准确率方面提升 2.5 个百分点，达到 79.3%，见图 2。其中红色代表正确回答，绿色代表错误回答。为了进一步分析图 2 中的问答结果，将本文方法（SGC）与已有方法（QCR）模型提取的中间层特征进行可视化，见图 3。图 3 针对图 2 中的问题 1 和问题 3 进行分析，高亮部分分别为本文方法（红色）和已有方法（绿色）重点关注区域，结合问题 1（“动脉瘤”）和问题 3（“右侧颞叶”）文本部分可知，本文方法可以更好地捕捉与文本相关的医学影像区域，从而更好地回答问题。通过图 3 所示的可视化结果可以看到，本文方法通过设计语义图结构可以更好地提取医学影像和文本中的关联信息，优于已有方法。

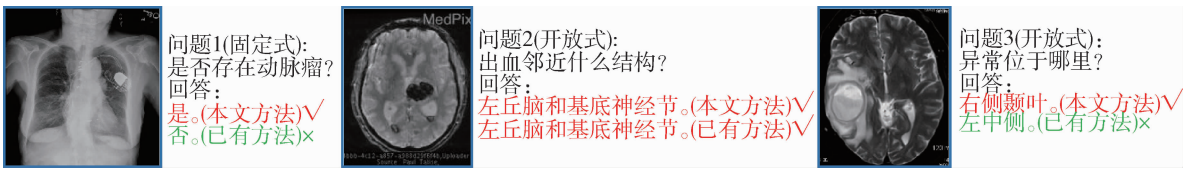

图 2 本文方法与已有方法的问答结果

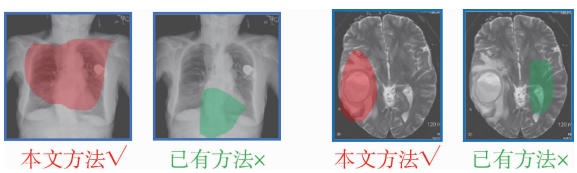

图 3 本文方法与已有方法对医学影像的关注区域可视化

根据前文表 2 中的实验结果和上述分析可知本文提出的基于语义图问答模型的有效性，针对开放式问答和固定式问答，通过语义图模块可以加强文本特征之间关联的表达，同时利用门控线性单元筛选文本特征中的重要信息，提升了整个模型的问答准确率。

4.4 消融分析

为了进一步探讨本文提出模型（SGC）中各模块对于任务的作用和效果，从语义图模块、门控模块以及两模块中包含的激活函数 3 方面进行消融分析。不同模块消融后的基于问答准确率的实验结果，见表 3。

表 3 不同模块消融实验

语义图	门控线性模块	激活函数	开放式问答准确率（%）	固定式问答准确率（%）
√	–	√	48.3	73.4
–	√	√	45.5	74.5
√	√	–	53.3	77.5
√	√	√	55.6	79.3

其中“√”表示使用该模块，“–”表示不使用此模块。从表 3 第 1 行可以看到门控线性模块对提升问答准确率的重要作用，固定式问答任务性能提升约 6%；从第 2 行可以看到语义图模块能够在较难的开放任务上有效捕捉文本内部关联；从第 3 行可以看到语义图模块和门控线性模块中的激活函数对任务准确度也有一定影响。

为了分析不同维度对模型性能的影响，实施相关模块对模型维度敏感度分析，见图 4、图 5。可以看到不同语义图嵌入维度对模型性能影响较小，当维度达到足以表征语义图时，模型性能达到饱和，更高维度的隐式空间是不必要的；门控线性模块对不同维度选择有一定要求，合适的隐藏层维度有利于该模块寻找重要语句文本。

针对语义图维度和门控模块维度，从模型表达能力而言，两者在一定范围内的提高均可以提升相应表达能力，但是更高维度会带来更大的计算复杂度，在神经网络训练过程中有较大开销，延缓整个模型的收敛效率，有一定可能造成性能下降，因而隐藏层维度并不是越大越好，这也与图 4 和图 5 的实验现象一致。从模型过拟合而言，虽然双向编码器表征模型[16]的隐藏层维度可达 768 甚至 1 024，却能同时拥有更强表达能力，对比可知，造成本文模型维度受限的另一个主要原因是训练集数据量较少，较深维度容易导致过拟合性能下降，由于医学领域标注的数据集有限，拥有专业医学知识的标注人员稀缺，标注难度较大，这也是医学问答领域乃至“人工智能 + 医学信息”领域目前的重要挑战。

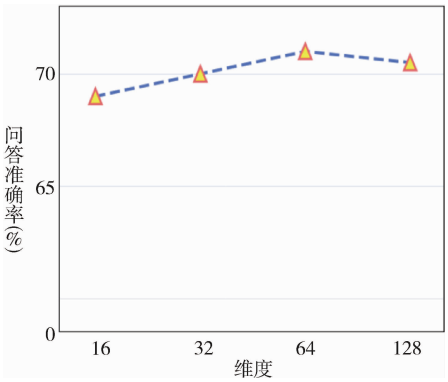

图 4 不同语义图维度模型性能

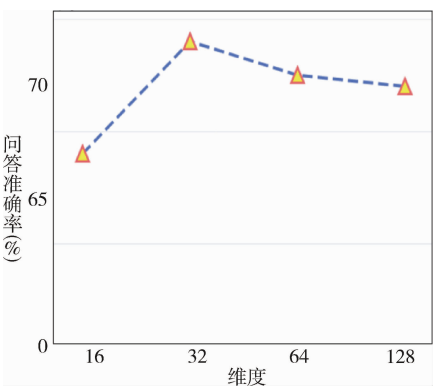

图 5 不同门控线性模块维度模型性能

5 结语

本文使用深度学习方法解决医学影像问答问题，通过元学习和自编码器模块提取医学影像视觉特征，通过语义图卷积提取问题中的文本特征，并获取视觉特征和文本特征之间的重要关联。实验结果表明本文方法在相关医学影像问答数据集上较已有工作有一定提升，开放式问答准确率提升到 55.6%；固定式问答准确率提升到 79.3%。本文方法在开放式问答性能提升方面尚有较大空间，开放式问答任务本身需要生成对应答案，因此需要提升模型生成能力。在未来工作中将从以下两个方面进行改进，首先在语言特征提取时采用预训练的大模型[16]提升特征表达能力，其次是收集并标注更多医学领域语料库用于训练，使模型具备更好的文本生成能力。

参考文献

1 张芳芳，马敬东，王小贤，等. 国外医学领域自动问答系统研究现状及启示［J］. 医学信息学杂志，2017，38（3）：2－6，25.

2 张飞. 基于知识库的医学智能问答系统设计与实现［D］. 北京：北京协和医学院，2021.

3 满坚平，黄国立，赖聪，等. 智能体在医疗健康领域的研究与应用［J］. 医学信息学杂志，2022，43（4）：20－26.

4 ANTOL S，AGRAWAL A，LU J，et al. Vqa：visual question answering［C］. Santiago：IEEE International Conference on Computer Vision（ICCV），2015.

5 YANG Z，HE X，GAO J，et al. Stacked attention networks for image question answering［C］. Las Vegas：IEEE International Conference on Computer Vision and Pattern Recognition（CVPR），2016.

6 FUKUI A，PARK D H，YANG D，et al. Multimodal compact bilinear pooling for visual question answering and visual grounding［C］. Austin：Conference on Empirical Methods in Natural Language Processing（EMNLP），2016.

7 NGUYEN B D，DO T T，NGUYEN B，et al. Overcoming data limitation in medical visual question answering［C］. Shenzhen：Medical Image Computing and Computer－Assisted Intervention（MICCAI），2019.

8 ZHAN L M，LIU B，FAN L，et al. Medical visual question answering via conditional reasoning［C］. Seattle：ACM International Conference on Multimedia，2020.

9 KIM J H，JUN J，ZHANG B，et al. Bilinear attention networks［C］. Montreal：Neural Information Processing Systems（NIPS），2018.

10 FINN C，ABBEEL P，LEVINE S. Model－agnostic meta－learning for fast adaptation of deep networks［C］. Sydney：International Conference on Machine Learning（ICML），2017.

11 MASCI J，MEIER U，CIRESAN D，et al. Stacked convolutional auto－encoders for hierarchical feature extraction［C］. Berlin：International Conference on Artificial Neural Networks（ICANN），2011.

12 DEY R，SALEM F M. Gate－variants of GRUneural networks［C］. Boston：Midwest Symposium on Circuits and Systems（MWSCS），2017.

13 WANG Z，ZHENG L，LI Y，et al. Linkage based face clustering via graph convolution network［C］. Long Beach：IEEE International Conference on Computer Vision and Pattern Recognition（CVPR），2019.

14 VENESS J，LATTIMORE T，BHOOPCHAND A，et al. Gated linear networks［C］. Virtual Event：Association for the Advancement of Artificial Intelligence（AAAI），2021.

15 LAU J J，GAYEN S，ABACHA A，et al. A dataset of clinically generated visual questions and answers about radiology images［J］. Scientific data，2018，5（1）：1－10.

16 JACOB DEVLIN，CHANG M，LEE K，et al. BERT：Pre－training of Deep Bidirectional Transformers for Language Understanding［C］. Minneapolis：The North American Chapter of the Association for Computational Linguistics（NAACL），2019.