ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 4,Jul. 2025,pp. 922－933
DOI:10. 16337/j. 1004-9037. 2025. 04. 007

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于多尺度注意力和图神经网络的多模态医学实体识别研究

韩　普[1,2]，刘森岭[1]，陈文祺[1]

(1. 南京邮电大学管理学院,南京 210003; 2. 江苏省数据工程与知识服务重点实验室,南京 210023)

摘　要：随着信息技术的快速发展,医疗健康领域中文文本、图像等多模态数据呈现出了爆发式增长。多模态医学实体识别(Multi-modal medical entity recognition，MMER)是多模态信息抽取的关键环节,近期受到了极大关注。针对多模态医学实体识别任务中存在图像细节信息损失和文本语义理解不足问题,提出一种基于多尺度注意力和图神经网络(Multi-scale attention and dependency parsing graph convolution,MADPG)的MMER模型。该模型一方面基于ResNet引入多尺度注意力机制,协同提取不同空间尺度融合的视觉特征,减少医学图像重要细节信息丢失,进而增强图像特征表示,补充文本语义信息;另一方面利用依存句法结构构建图神经网络,捕捉医学文本中词汇间复杂语法依赖关系,以丰富文本语义表达,促进图像文本特征深层次融合。实验表明,本文提出的模型在多模态中文医学数据集上F_1值达到95.12%,相较于主流的单模态和多模态实体识别模型性能得到了明显提升。

关键词：多模态医学实体识别;多尺度注意力;图卷积神经网络;多模态融合;语义特征
中图分类号：TP391.1;TP183　　**文献标志码**:A

Multi-modal Medical Entity Recognition Based on Multi-scale Attention and Graph Neural Networks

HAN Pu[1,2]，LIU Senling[1]，CHEN Wenqi[1]

(1. School of Management，Nanjing University of Posts and Telecommunications，Nanjing 210003，China; 2. Jiangsu Province Key Lab of Data Engineering and Knowledge Service，Nanjing 210023，China)

Abstract：With the rapid development of information technology，multi-modal data such as Chinese texts and images in the medical and health field has shown explosive growth. Multi-modal medical entity recognition (MMER) is a key step in multi-modal information extraction，and has attracted great attention recently. Aiming at the problems of image detail loss and insufficient text semantic understanding in multi-modal medical entity recognition tasks，this paper proposes a novel MMER model based on multi-scale attention and dependency parsing graph convolution (MADPG). This model introduces a multi-scale attention mechanism based on ResNet to collaborate to extract visual features fused with different spatial scales and to reduce the loss of important details of medical images. Thus the image feature representation and complementing text semantic information are enhanced. Then，the dependency syntactic structure is used to construct the graph neural network to capture the complex grammatical dependencies between

基金项目:国家社会科学基金项目(22BTQ096);南京邮电大学1311人才计划资助。
收稿日期:2025-01-07;**修订日期**:2025-03-03

words in medical texts, so as to enrich the semantic expression of texts and promote the deep integration of image text features. Experiments show that the F_1 value of the proposed model reaches 95.12% on the multi-modal Chinese medical data set, and the performance of the proposed model is significantly improved compared with the mainstream single- and multi-modal entity recognition models.

Key words: multi-modal medical entity recognition (MMER); multi-scale attention; graph convolutional neural network; multi-modal fusion; semantic feature

引　　言

近年来,随着“健康中国”国家战略持续推进以及人工智能的快速发展,医疗健康领域的知识服务正朝着智能化、精准化和个性化方向发展[1]。医疗健康信息抽取和知识挖掘是知识服务智能化的关键,其水平高低很大程度上决定了知识服务的效率和质量[2]。同时,随着医学影像等信息技术和医疗信息系统的迅猛发展,各类平台和系统积累了海量文本、图像等多模态医疗健康数据。而传统的单模态数据处理和分析主要针对单一模态的文本或图像,难以支撑多模态医疗健康数据的语义理解和知识融合。因此,如何增强模态间深层语义理解、实现深度知识融合是当前医疗健康领域信息抽取和知识挖掘的关键问题。

多模态医学实体识别(Multi-modal medical entity recognition, MMER)是多模态医疗健康知识组织的基础和关键任务,旨在通过医学图像、音频等辅助模态补充文本语义进而提升医学实体识别效果[3]。相比于通用领域,医疗健康领域数据更为复杂和专业,MMER任务更具挑战性。现有研究主要利用文本特征和图像视觉信息进行MMER研究。Wang等[4]和Ning等[5]分别利用多模态树和交叉注意力机制将词特征和字符特征进行模态信息融合,进而丰富医学文本语义信息实现多模态实体识别。韩普等[6]将医学图像作为辅助模态以补充文本语义特征,并通过多模态表示学习和双线性注意力机制实现图像文本特征融合。虽然这些方法在MMER任务中取得了一定进展,但仍存在两个关键问题:(1)忽略了医学图像由于对比度差、边界模糊等问题丢失的重要细节信息,造成多模态信息交互不充分;(2)忽略了医学文本中术语和词汇间的复杂依赖关系,未能充分挖掘和理解文本语义信息,限制了多模态特征的有效融合。

鉴于通道和空间注意力机制在多个计算机视觉任务中的显著优势,为了增强图像特征表示,Ouyang等[7]在图像特征提取中提出了多尺度注意力(Efficient multi-scale attention,EMA)方法,该方法将每个特征组中空间语义信息均匀分布,有效减少了图像细节丢失。在文本分类任务中,为了捕捉更丰富的文本特征,Wang等[8]利用图卷积神经网络(Graph convolutional networks, GCN)捕捉词间语义依赖关系进而建模复杂语义表示,增强了模型对文本语义的理解能力。在已有研究基础上,本文提出了一种基于EMA和图神经网络的多模态实体识别模型,一方面在图像特征提取阶段引入EMA捕捉不同尺度信息以减少图像细节丢失;另一方面在多模态特征融合过程中利用句法结构构建GCN以丰富文本语义特征。

本文在中文MMER数据上验证了所提模型的有效性,进一步推进了中文医学领域信息抽取研究进展。本文的主要贡献如下:(1)将EMA用于MMER任务,从多个维度提取图像模态特征,充分利用医学图像中的关键细节信息;(2)引入依存句法结构充分挖掘文本深层语义信息,并利用图神经网络进一步促进图像和文本的深层语义交互。

1 相关研究

1.1 多尺度注意力相关研究

EMA通过引入多个尺度的注意力权重,并结合并行注意力子模块,能够同时关注局部细节和整体结构,可有效提升模型特征表示能力和泛化性能,在图像分类、目标检测和语义分割等任务中表现出色[9-10]。

Jiao等[11]提出了多尺度空洞注意力,通过设置不同空洞率捕捉多尺度语义信息;Wang等[12]结合多尺度和门控机制构建了EMA网络以聚合全局和局部信息,并降低了图像噪声干扰。雷浩等[13]在ResNet-18网络中引入多尺度卷积模块替代单一尺度卷积核,使模型更充分地提取多种局部特征。Ding[14]在语义分割任务中提出了一个新型多尺度卷积注意力模块,通过不同的感受野大小捕获特征,提高了模型语义分割准确性。Shao等[15]利用EMA卷积神经网络(Convolutional neural network, CNN)提取多尺度、低噪声的特征,在仅有少量标记数据的情况下实现高精度的故障检测。Zhong等[16]在医学图像分割任务中引入了一个基于注意力机制的多尺度特征增强模块,促进了多尺度语义信息与全局上下文特征的有效融合,增强了模型特征表示能力。陈龙等[17]在小样本图像识别任务中构建了一种EMA与领域自适应模型,通过突出类别相关特征提升了图像识别准确率。李烨等[18]引入EMA融合模块,利用不同分辨率特征信息之间的差异,提高了多尺度遥感目标检测精度。为提取全面且关键特征信息,刘拥民等[19]结合CNN和EMA融合图像局部特征信息与全局特征信息,提升了小样本图像分类准确率。

1.2 图神经网络相关研究

图神经网络(Graph neural network, GNN)是处理图结构数据的深度学习模型,其核心在于通过多次迭代聚合邻居节点信息更新节点表示,以捕捉数据深层次特征。

为解决传统深度学习仅能处理局部语义关系问题,Kipf等[20]构建了GCN,通过在图结构数据上进行卷积操作,有效捕获了GNN中全局结构特征和节点间复杂关系。GCN已广泛应用于谣言检测、情感分析和语音情感识别等任务,在谣言检测任务中,Cui等[21]采用图卷积网络挖掘文本内以及文本间的深层语义关联,从局部和全局两个层面对文本和知识信息进行建模,从而提升模型检测性能;Wang等[22]提出一种结合语法与语义的双图神经网络,该网络通过GCN捕捉句子中的语义信息,并将其与句法信息相融合,以提升方面级情感分析性能。为了缓解句子信息丢失的影响,充分利用句子间依赖结构,Lu等[23]提出了一种情感交互式图卷积网络,用于捕捉句法中的长距离依赖关系以及语义中的情感交互;Shang等[24]利用GCN模型建模不同单词间的依赖关系以及句子间依存句法关系,提升了情感分析任务性能;常远等[25]在中文医学实体识别任务中引入句法结构捕捉字符间依赖关系,借助关系GCN融合句法依赖信息,以提升模型对口语化和表述不规范文本的实体识别效果;Jain等[26]等将RoBERTa和GCN相结合,利用图卷积网络捕捉依存关系图中的节点特征,增强了模型对实体边界的理解和提取能力。

2 模型设计

为充分挖掘医学图像语义信息和文本深层语义特征,解决医学图像细节丢失和文本语义理解不足问题,本文提出一种基于多尺度注意力和图神经网络(Multi-scale attention and dependency parsing graph convolution, MADPG)的MMER模型。

具体而言,首先利用ResNet和EMA协同提取多尺度图像特征以减少图像细节丢失,接着基于门控注意力机制将文本序列和多尺度视觉序列在RoBERTa中进行多模态特征融合,然后通过依存句法结构构建GCN学习文本深层语义信息以丰富多模态语义特征,最后结合双向长短期记忆网络和条件随机场进行实体识别序列标注预测输出,如图1所示。

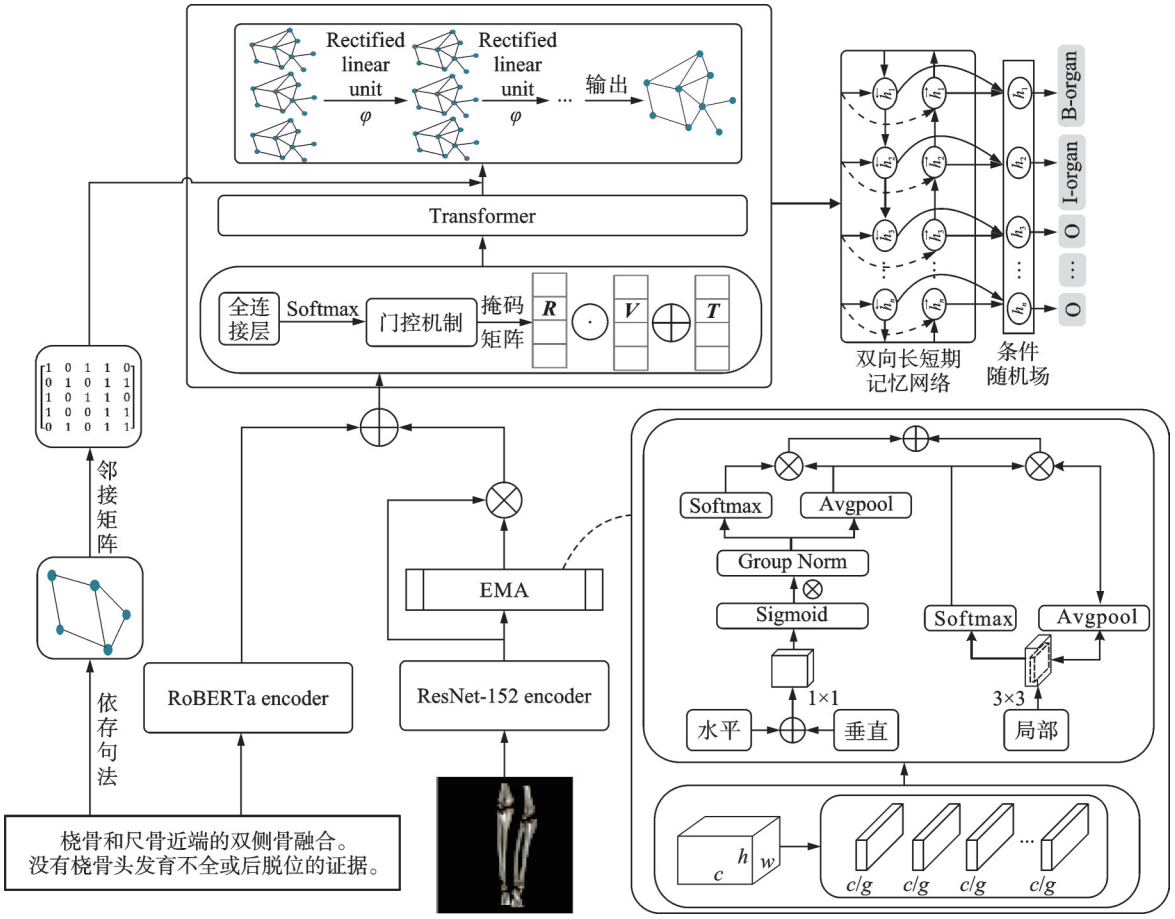

图1　本文模型整体架构

Fig.1　Overall architecture of the proposed model

2.1　多模态特征提取

(1) 文本特征提取

文本特征表示对多模态实体识别具有重要影响[27]。本文利用在中文维基百科大规模语料中进行预训练的语言模型进行文本特征表示。鉴于RoBERTa-WWM[28]能够利用全词掩码技术学习词级先验语义特征,并且能够捕获输入数据中的复杂模式并生成更精细的特征表示,本文选用RoBERTa-WWM作为文本信息编码器,得到文本标记序列的特征表示 $T=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n]$。

(2) 图像特征提取

多维度的图像语义信息能够充分补充文本语义,进而提升多模态实体识别效果。ResNet-152[29]在大规模图像数据集上进行了预训练,学习了丰富的视觉特征,具有较强的泛化能力。但重要病理特征可能存在于图像局部区域,仅使用单尺度卷积难以有效捕捉局部特征。为获得更多维度的视觉上下文信息,本文在ResNet-152基础上引入EMA[30],构建了一个并行层次结构,从不同空间尺度提取图像的局部特征和全局表示。首先将图像输入ResNet提取最后一层卷积层输出作为EMA的输入,然后EMA将输出通道维度按批处理维度重新排列,从全局(水平和垂直)和局部多个维度提取多尺度视觉信息,以补充ResNet卷积层输出,进而增强图像特征表示,最后形成有序的视觉特征序列 $V=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m]$。

2.2 多模态特征融合

传统多模态特征融合方法通常采用简单的特征拼接,无法充分捕捉图像文本模态间的深层语义关联[31],本文提出了一种两阶段融合方法,该方法通过细粒度融合策略来获得图像文本模态间更丰富的交互信息,能够动态调整文本对图像不同区域的关注度,并在多层次上实现模态间深度融合。具体流程如下。

(1) 通过注意力机制实现图像与文本特征的初步融合

在第1阶段特征融合过程中,为引导文本特征和多尺度视觉特征融合,本文首先将医学文本特征和医学图像特征结合,形成初步的多模态序列表示$\boldsymbol{Z}$作为多模态特征融合网络的输入,如式(1)所示,其中[CLS]表示文本与图像之间的关系分类,[SEP]表示文本和图像特征之间的分离。

$$\boldsymbol{Z}=[\mathrm{CLS}]\boldsymbol{w}_1\boldsymbol{w}_2\cdots\boldsymbol{w}_n[\mathrm{SEP}]\boldsymbol{v}_1\boldsymbol{v}_2\cdots\boldsymbol{v}_m \tag{1}$$

接着将多模态序列表示$\boldsymbol{Z}$输入全连接层得到标量值x,然后使用向量归一化函数得到文本-图像关系的相关性得分r。根据相关性得分r,本文在门控注意力单元中引入一个7×7的权重参数矩阵$\boldsymbol{W}$,构建视觉掩码矩阵$\boldsymbol{R}$,即

$$r=\left[\mathrm{Gate}(\mathrm{softmax}(x))\right]_2 \tag{2}$$

$$\boldsymbol{R}=(x_{i,j}=r)\odot\boldsymbol{w}+\boldsymbol{d} \tag{3}$$

式中:$\boldsymbol{d}$表示7×7的可学习偏置矩阵,"⊙"表示矩阵点乘。在视觉掩码矩阵$\boldsymbol{R}$基础上,本文进一步通过Hadamard积对医学图像特征进行加权得到加权后的视觉特征$\boldsymbol{V}'$,并将其与文本特征拼接形成新的多模态特征表示$\boldsymbol{Z}'$后送入Transformer层中进行处理。最后,利用多头自注意力机制和前馈神经网络在不同层次上捕捉文本和图像特征间的全局依赖关系,得到多尺度视觉语义增强的文本特征表示$\boldsymbol{H}^{(0)}$,Concat表示矩阵拼接,具体过程如式(4~6)所示。

$$\boldsymbol{V}'=\boldsymbol{V}\odot\boldsymbol{R}=\begin{pmatrix}\boldsymbol{v}_{11}&\boldsymbol{v}_{12}&\cdots&\boldsymbol{v}_{1d}\\\boldsymbol{v}_{21}&\boldsymbol{v}_{22}&\cdots&\boldsymbol{v}_{2d}\\\vdots&\vdots&\ddots&\vdots\\\boldsymbol{v}_{w1}&\boldsymbol{v}_{w2}&\cdots&\boldsymbol{v}_{wd}\end{pmatrix}\odot\begin{pmatrix}\boldsymbol{r}_{11}&\boldsymbol{r}_{12}&\cdots&\boldsymbol{r}_{1d}\\\boldsymbol{r}_{21}&\boldsymbol{r}_{22}&\cdots&\boldsymbol{r}_{2d}\\\vdots&\vdots&\ddots&\vdots\\\boldsymbol{r}_{w1}&\boldsymbol{r}_{w2}&\cdots&\boldsymbol{r}_{wd}\end{pmatrix} \tag{4}$$

$$\boldsymbol{Z}'=\mathrm{Concat}(\boldsymbol{T},\boldsymbol{V}') \tag{5}$$

$$\boldsymbol{H}^{(0)}=\mathrm{Transformer}(\boldsymbol{Z}') \tag{6}$$

(2) 构建依存句法结构图,利用GCN实现深度融合

在第2阶段特征融合过程中,本文首先使用Spacy(https://spacy.io)的依存句法分析功能从原始输入文本中生成依存句法结构图DpG=(X,E),其中文本中的词向量构成图结构中的节点集合X,词汇间的依存句法关系构成图中的边集合E。接着将依存句法结构图转换为邻接矩阵$\boldsymbol{A}\in\boldsymbol{M}^{n\times n}$,若词与词之间存在依存关系,对应位置赋值为1,否则赋值0,即

$$\boldsymbol{A}_{i,j}^{n}=f_{\mathrm{spa}}(x)=\begin{cases}1 & \boldsymbol{w}_i\in\boldsymbol{M}\text{且}\boldsymbol{w}_j\in\boldsymbol{M},\text{或}i=j\\0 & \text{其他}\end{cases} \tag{7}$$

然后基于邻接矩阵$\boldsymbol{A}$,利用GCN对多尺度视觉语义增强的文本特征表示$\boldsymbol{H}^{(0)}$进行卷积操作,来更新每个节点的特征表示,具体如式(8)所示,$\boldsymbol{H}_i^{(k+1)}$为节点i在第$k+1$层的特征,$\phi(\cdot)$使用了非线性激活函数(Retctifeid linear unit, ReLU),$\boldsymbol{W}^{(l)}$是第k层权重矩阵,$b^{(k)}$为第k层截距。

$$\boldsymbol{H}_i^{(k+1)}=\phi\left(\sum_{j=1}^{n}c^i\boldsymbol{A}_{ij}(\boldsymbol{W}^{(k)}\boldsymbol{H}_j^{(k)}+b^{(k)})\right) \tag{8}$$

鉴于长短期记忆网络(Long short-term memory, LSTM)能够有效解决长期依赖问题,并且缓解了梯度弥散和梯度爆炸影响,利用正反向LSTM构建的双向长短期记忆网络(Bidirectional LSTM, BiLSTM)能够更好地学习文本上下文信息[32]。因此,本文最后将GCN的输出$\boldsymbol{H}_i$送入BiLSTM网络进

行特征抽取,最后得到每个序列标记对应的隐藏状态向量$\boldsymbol{H}$,即

$$\boldsymbol{H}=\left\{\boldsymbol{h}_i=[\overrightarrow{\boldsymbol{h}}_i^{\mathrm{LSTM}};\overleftarrow{\boldsymbol{h}}_i^{\mathrm{LSTM}}]\right\}_{i=1}^{n} \tag{9}$$

2.3　特征解码

由于BiLSTM模型难以学习标签之间的相关性,而条件随机场(Conditional random field, CRF)能够通过全局特征归一化得到求解过程中的全局最优解,有效降低了标记过程中可能出现的偏差问题[33]。本文基于CRF模型给定输入$\boldsymbol{H}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n]$,在训练过程中通过最小化损失函数优化模型参数,最后使用Viterbi算法得到最佳命名实体标签。对于标签序列$\boldsymbol{y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n]$,最大化标签序列$\boldsymbol{y}$的似然概率和损失函数分别定义为

$$s(\boldsymbol{H},\boldsymbol{y})=\sum_{i=0}^{n}T(\boldsymbol{y}_i,\boldsymbol{y}_{i+1})+\sum_{i=0}^{n}U(\boldsymbol{h}_i,\boldsymbol{y}_i) \tag{10}$$

$$\begin{cases}\ln[P(\boldsymbol{y}|\boldsymbol{H})]=s(\boldsymbol{H},\boldsymbol{y})-\ln\left(\sum\limits_{\tilde{\boldsymbol{y}}=\boldsymbol{Y}_{h_1}}^{\boldsymbol{Y}_{h_n}}\mathrm{e}^{s(\boldsymbol{H},\boldsymbol{y})}\right) & \tilde{\boldsymbol{y}}\in\boldsymbol{Y}_H\\ \mathrm{loss}=-\ln[p(\boldsymbol{y}|\boldsymbol{H})]\end{cases} \tag{11}$$

式中:n表示序列长度,$T(\boldsymbol{y}_i,\boldsymbol{y}_{i+1})$表示第$i$步的标签是$\boldsymbol{y}_i$、第$i+1$步标签是$\boldsymbol{y}_{i+1}$的转移分数,$U(\boldsymbol{h}_i,\boldsymbol{y}_i)$表示第$i$步输入$\boldsymbol{h}_i$对应的标签是$\boldsymbol{y}_i$的发射分数,$P(\boldsymbol{y}|\boldsymbol{H})$为条件概率,$\tilde{\boldsymbol{y}}$为真实标注序列,$\boldsymbol{Y}_H$为所有可能的标注序列。

3　实验设计与结果分析

3.1　数据准备

本文实验数据来自爱爱医临床病历数据库(https://www.iiyi.com)、影像园(http://www.xctmr.com)和Radiopaedia(https://radiopaedia.org)等在线医疗健康平台,主要收集文本和图像两种模态数据。文本数据主要由病历记录和诊断报告组成;图像数据是文本对应的患者诊断图像。参照已有研究,本实验定义了疾病(Disease)、症状(Symptom)、器官(Organ)、属性(Attribute)和治疗手段(Treatment)五类医学实体。经过对收集的数据进行脱敏、清洗和归一化处理后得到6 995条高质量的图片文本对,数据集划分细节如表1所示。

同时,在具有专业医学知识的专家指导下采用BIO标注体系(B-X,I-X和O)进行标注,其中X代表实体类别,B-X表示实体X的开始,I-X表示实体X的内部,O表示非实体,具体示例如图2所示。

表1　数据集划分详情

Table 1　Dataset partitioning details

数据集划分	数据数量	实体数量
训练集	5 000	76 070
验证集	1 000	15 236
测试集	995	15 193

图2　医学图像文本对标注示例

Fig.2　Annotation example of medical image and text

3.2　实验环境与参数设置

本文实验均在Ubuntu操作系统进行,服务器CPU型号为Intel(R),频率为2 GHz;GPU型号为24 GB显存的RTX 3090的显卡;服务器内存为43 GB;开发环境为python3.10、cuda12.1和torch2.2.1。实验参数设置如表2所示。

表2　实验参数设置

Table 2　Experimental parameter setting

模型参数	参数值	模型参数	参数值
批尺寸	32	优化器	Adam
迭代次数	16	激活函数	ReLU
学习率	1e−5	注意力头	12
Dropout机制	0.5	最大序列长度	256

3.3 评价指标

参照已有研究[34]，实验采用的评价指标为准确率(Precision)、召回率(Recall)和F_1值(F_1-score)，计算方式如式(12～14)所示。

$$\text{Precision}(P)=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{12}$$

$$\text{Recall}(R)=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{13}$$

$$F_1=\frac{2PR}{P+R} \tag{14}$$

式中：TP是正样本正确预测的数量；FP指负样本预测为正样本的数量；FN是正样本预测为负样本的数量；F_1值是准确率和召回率的调和平均数，能够平衡两者之间的影响，提供一个更为全面的性能评估。

3.4 基准模型对比实验

为验证本文所提模型MADPG-MMER在中文多模态医学实体识别任务中的有效性，本实验将所提模型与经典文本模态模型以及常见多模态模型进行对比。对比实验模型具体如下，实验结果如图3和表3所示。

(1) CNN-CRF[35]：使用CNN提取文本特征表示并使用CRF建模序列标注实体之间依赖关系，输出最优预测结果。

(2) BiLSTM-CRF[36]：使用双向长短期记忆网络BiLSTM学习序列数据的特征表示并利用CRF完成实体识别。

(3) BERT-CRF[37]：使用预训练语言模型BERT学习本文语义表示，编码上下文信息并结合CRF进行实体识别。

(4) RoBERTa-CRF[38]：使用BERT的改进模型RpBERTa提取文本特征并使用CRF完成实体识别任务。

(5) BERT-BiLSTM-CRF[37]：在BERT-CRF模型中引入BiLSTM加强模型捕捉时间序列依赖的能力，提高实体识别效果。

(6) RoBERTa-BiLSTM-CRF[38]：在RoBERTa-CRF模型中引入BiLSTM同时对前后信息进行学习，更全面地捕获上下文信息，提高模型识别实体的准确性。

(7) 统一多模态Transformer(Unified multimodal Transformer, UMT)[39]：使用BERT提取文本特征和ResNet提取视觉对象图像特征，并通过Transformer交互文本和图像信息进行实体识别。

(8) 统一多模态图融合(Unified multi-modal graph fusion, UMGF)[40]：使用BERT提取文本特征、ResNet提取图像特征，并结合多模态图融合方法实现文本和图像语义信息交互完成实体识别任务。

(9) RpBERT[41]：使用BERT和ResNet分别提取文本和视觉对象图像特征，然后基于文本图像关系传播机制进行图像文本特征融合。

(10) EMA-RpRoBERTa：改进RpBERT模型，使用RoBERTa提取文本特征、ResNet和EMA提取多尺度图像特征，然后利用文本图像关系传播机制融合图像文本

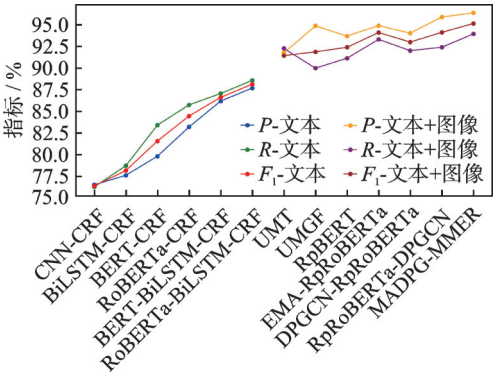

图3 基准模型实验结果

Fig.3 Experimental results of the benchmark models

特征。

(11) DPGCN-RpRoBERTa:改进RpBERT模型,使用RoBERTa和依存句法分析图卷积网络(Dependency parsing graph convolutional networks, DPGCN)提取文本深层语义信息、ResNet提取视觉对象图像特征,然后通过文本图像关系传播机制进行图像文本特征融合。

(12) RpRoBERTa-DPGCN:改进RpBERT模型,使用RoBERTa和ResNet分别提取文本和视觉图像特征,然后使用文本图像关系传播机制融合图像文本特征,最后使用DPGCN挖掘深层语义信息。

表3　基准模型实验结果

Table 3　Experimental results of the benchmark models　%

数据模态	模型名称	P	R	F_1
文本	CNN-CRF	76.49	76.24	76.36
	BiLSTM-CRF	77.58	78.69	78.13
	BERT-CRF	79.79	83.39	81.55
	RoBERTa-CRF	83.17	85.72	84.43
	BERT-BiLSTM-CRF	86.16	87.03	86.59
	RoBERTa-BiLSTM-CRF	87.66	88.56	88.11
文本+图像	UMT	91.73	92.25	91.41
	UMGF	94.85	89.96	91.85
	RpBERT	93.67	91.11	92.37
	EMA-RpRoBERTa	94.88	93.29	94.08
	DPGCN-RpRoBERTa	94.01	91.99	92.98
	RpRoBERTa-DPGCN	95.87	92.38	94.10
	MADPG-MMER	96.35	93.92	95.12

3.5　消融实验

为探究本文所提模型MADPG-MMER中各模块在中文多模态医学实体识别任务中的有效性,本实验分别去掉不同单一模块或组合模块与完整MADPG-MMER模型进行对比,采用同样评价指标,具体结果如表4所示。其中"w/o"代表从MADPG-MMER中去除对应模块,"w/o EMA"代表去除EMA,"w/o DPGCN"代表去除基于依存句法结构的图卷积网络,"w/o Gating mechanism"代表去掉门控注意力机制。

表4　消融实验结果

Table 4　Ablation experiment results　%

模型名称	P	R	F_1
MADPG-MMER	96.35	93.92	95.12
w/o EMA	95.87	92.38	94.10
w/o DPGCN	94.88	93.29	94.08
w/o Gating mechanism	94.78	93.32	94.05
w/o EMA & DPGCN	94.09	92.71	93.39
w/o EMA & Gating mechanism	95.14	91.46	93.26
w/o DPGCN & Gating mechanism	93.53	92.28	92.90
w/o EMA &DPGCN & Gating mechanism	93.23	90.36	91.77

3.6 实验结果分析

(1) 基准模型对比实验结果分析

通过分析表3可知,在单一文本模态模型中,BERT-CRF和RoBERTa-CRF模型F_1值较BiLSTM-CRF模型分别提升3.42%和6.3%,表明通过预训练可以增强模型的语义理解和表征能力,并且动态掩码机制在医学文本特征提取过程中更具优势。进一步分析发现,RoBERTa-BiLSTM-CRF模型相较于RoBERTa-CRF模型,F_1值提升了3.68%,这一结果表明结合预训练模型和LSTM显著增强了模型对医学文本长期依赖关系的捕捉能力,从而提升了实体识别的准确性。

在文本-图像多模态模型中,UMT模型和UMGF模型F_1值相较于单模态基线模型分别提升了3.3%和3.74%,这表明通过引入多模态交互注意力和多模态图融合网络,能够有效利用视觉辅助信息,从而提升文本理解准确性。RpBERT模型F_1值相较于UMGF模型进一步提升了0.52%,这是因为RpBERT模型引入文本-图像关系传播机制,能够降低图像相关性低的区域对文本特征的干扰,使得模型能够更细粒度地利用图像文本间的互补信息,进而提高模型对中文医疗实体识别的准确性。

从图3可以明显看出,文本-图像多模态模型在P、R和F_1这3个评价指标上的折线始终高于所有单一文本模态模型,这表明图像和文本特征融合可有效提升模型实体识别性能。在RpBERT模型基础上,本文所提MADPG-MMER模型在所有指标上均表现最优,其中F_1值为95.12%,相较于最优单模态和多模态基线模型,分别提升7.01%和2.75%。MADPG-MMER模型的优势可以归结于以下两个原因:① 在医学图像特征提取过程中,通过引入EMA模块,模型能够沿水平维度和垂直维度两个方向对图像特征进行加权融合,进而聚合更加全面完整的空间结构信息;而经典的多模态基线模型UMT、UMGF和RpBERT均采用ResNet编码器,这种基于单尺度卷积的方法在医学图像表征过程中会导致细节丢失,因此本文选用EMA模块可为医学实体提供丰富的视觉语义信息,显著提升多模态医学实体识别任务效果。② 在多模态特征融合过程中,通过引入依存句法结构图,模型能够聚合邻居节点信息来理解医学文本中的复杂句法结构,进而捕捉医学实体间的细粒度关系。该方法从更多维度对多尺度视觉语义增强的医学文本特征进行挖掘,不仅丰富了文本特征的多层次语义信息,还进一步增强了图像文本特征间的深层语义融合,显著提高了模型对多模态医学实体识别的准确性。

(2) 消融实验结果分析

分析表4可知,移除MADPG-MMER模型的单一模块或组合模块均会导致模型识别效果下降:①去除EMA模块,模型F_1值降低1.02%,表明EMA能够有效捕捉医学图像多尺度语义特征;②去掉门控注意力机制,F_1值下降1.07%,表明通过计算文本特征和图像特征区域的相关度,有效降低了无关视觉区域的噪声干扰;③移除DPGCN模块的F_1值减少1.04%,表明基于依存句法结构构建图卷积网络能够帮助模型更好地理解文本依存句法信息;④去除EMA和门控注意力组合模块,模型F_1值下降1.86%,说明这两个模块结合能够有效地捕捉图像特征深层语义信息;⑤同时去掉EMA和DPGCN模块,F_1值降低1.73%,表明结合这两种技术能够提高模型对多模态信息的挖掘能力;⑥移除门控注意力和DPGCN模块,F_1值降低2.73%,说明组合这两个模块能够更充分地利用图像文本信息间深层语义关联;⑦同时去掉EMA、门控注意力和DPGCN模块,F_1值降低3.35%,表明这3大模块能够相互协作共同提高模型的多模态实体识别效果。综上所述,实验结果验证了EMA、依存句法结构图卷积网络以及门控注意力机制这3个关键组件在多模态医学实体识别模型MADPG-MMER中的必要性和有效性,它们共同作用能够提升模型性能。

4 结束语

为深入挖掘医学图像细节信息和文本深层语义特征,本文提出了基于EMA和GNN的多模态实体

识别模型。该模型一方面在ResNet中引入EMA模块从不同空间尺度捕捉图像模态特征,能够更充分利用医学图像中的关键细节信息;另一方面通过设计门控注意力机制动态调整文本对图像区域的关注度,减少了无关区域噪声干扰,并利用依存句法结构构建GCN充分挖掘文本深层语义信息,通过多尺度视觉特征和文本特征间深层语义交互,有效提升多模态医学实体识别准确性。与最优单模态、多模态基线模型对比,本文所提模型F_1值分别提升了7.01%、2.75%,表明该模型能够有效降低图像噪声干扰,丰富文本语义特征,促进文本和图像特征有效融合,从而更准确地识别医学实体。尽管如此,本研究还存在一定不足,所构建模型主要基于文本和图像模态,并且未在多个多模态医学实体识别数据集上验证该模型泛化性能。在未来工作中,一方面将继续探究融合声音、视频等其他模态数据的实体识别模型;另一方面将在不同规模多模态医学实体识别数据集上进行实验,以进一步验证本文所提模型的有效性。

参考文献:

[1] 韩普, 叶东宇, 陈文祺, 等. 面向多模态医疗健康数据的知识组织模式研究[J]. 现代情报, 2023, 43(10): 27-34.
HAN Pu, YE Dongyu, CHEN Wenqi, et al. Research on the knowledge organization model of multimodal medical and health data[J]. Journal of Modern Information, 2023, 43(10): 27-34.

[2] 杨善林, 丁帅, 顾东晓, 等. 医疗健康大数据驱动的知识发现与知识服务方法[J]. 管理世界, 2022, 38(1): 219-229.
YANG Shanlin, DING Shuai, GU Dongxiao, et al. Healthcare big data driven knowledge discovery and knowledge service approach[J]. Journal of Management World, 2022, 38(1): 219-229.

[3] 韩普, 李雄. 基于增强异构图融合的多模态医学实体识别研究[J]. 现代情报, 2025, 45(6):34-45.
HAN Pu, LI Xiong. Multimodal medical named entity recognition based on augmented heterogeneous graph fusion[J]. Journal of Modern Information, 2025, 45(6): 34-45.

[4] WANG C, WANG H, ZHUANG H, et al. Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree[J]. Journal of Biomedical Informatics, 2020, 111: 103583.

[5] NING J, LI J, YANG Z, et al. BioNER-CFEM: Biomedical named entity recognition based on character feature enhancement with multimodal method[C]//Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Las Vegas, USA: IEEE, 2022: 787-790.

[6] 韩普, 陈文祺, 顾亮, 等. 融合多模态数据的中文医学实体识别研究[J]. 情报理论与实践, 2024, 47(9): 174-182.
HAN Pu, CHEN Wenqi, GU Liang, et al. Research on Chinese medical named entity recognition with fusion of multimodal data[J]. Information Studies: Theory & Application, 2024, 47(9): 174-182.

[7] OUYANG D, HE S, ZHANG G, et al. Efficient multi-scale attention module with cross-spatial learning[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-5.

[8] WANG Y, WANG Y, HU H, et al. Knowledge-graph- and GCN-based domain Chinese long text classification method[J]. Applied Sciences, 2023, 13(13): 7915.

[9] LI X, FU C, WANG Q, et al. DMSA-UNet: Dual multi-scale attention makes UNet more strong for medical image segmentation[J]. Knowledge-Based Systems, 2024, 299: 112050.

[10] 阎菩提, 邱实, 岳程斐. 结合局部高清图像的遥感集群目标区域超分辨率重建[J]. 南京航空航天大学学报, 2023, 55(6): 956-965.
YAN Puti, QIU Shi, YUE Chengfei. Remote sensing image super-resolution reconstruction with local high-resolution clustered object images[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2023, 55(6): 956-965.

[11] JIAO J, TANG Y M, LIN K Y, et al. DilateFormer: Multi-scale dilated Transformer for visual recognition[J]. IEEE Transactions on Multimedia, 2023, 25: 8906-8919.

[12] WANG Y, LI Y, WANG G, et al. Multi-scale attention network for single image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2024: 5950-5960.

[13] 雷浩, 苑迎春, 何振学. 基于多尺度卷积和注意力机制的枣品种识别[J]. 中国农机化学报, 2024, 45(6): 135-141, 148.
LEI Hao, YUAN Yingchun, HE Zhenxue. Jujube varieties recognition based on multi-scale convolution and attention mechanism[J]. Journal of Chinese Agricultural Mechanization, 2024, 45(6): 135-141, 148.

[14] DING B. LENet: Lightweight and efficient LiDAR semantic segmentation using multi-scale convolution attention[EB/OL]. (2023-01-11). https://doi.org/10.48550/arXiv.2301.04275.

[15] SHAO X, KIM C S. Adaptive multi-scale attention convolution neural network for cross-domain fault diagnosis[J]. Expert Systems with Applications, 2024, 236: 121216.

[16] ZHONG J, TIAN W, XIE Y, et al. PMFSNet: Polarized multi-scale feature self-attention network for lightweight medical image segmentation[J]. Computer Methods and Programs in Biomedicine, 2025, 261: 108611.

[17] 陈龙, 张建林, 彭昊, 等. 多尺度注意力与领域自适应的小样本图像识别[J]. 光电工程, 2023, 50(4): 66-80.
CHEN Long, ZHANG Jianlin, PENG Hao, et al. Few-shot image classification via multi-scale attention and domain adaptation [J]. Opto-Electronic Engineering, 2023, 50(4): 66-80.

[18] 李烨, 周生翠, 张驰. MSDAB-DETR: 一种多尺度遥感目标检测算法[J]. 数据采集与处理, 2024, 39(6): 1455-1469.
LI Ye, ZHOU Shengcui, ZHANG Chi. MSDAB-DETR: A multi-scale remote sensing target detection algorithm[J]. Journal of Data Acquisition and Processing, 2024, 39(6): 1455-1469.

[19] 刘拥民, 肖凤姣, 乔梦媛, 等. CA-MFE: CNN与注意力机制多尺度的图神经小样本图像分类网络[J]. 控制工程, 2025, 32(6): 1008-1015.
LIU Yongmin, XIAO Fengjiao, QIAO Mengyuan, et al. CA-MFE: CNN and attention mechanism multiscale graph neural network for few-shot image classification[J]. Control Engineering of China, 2025, 32(6): 1008-1015.

[20] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2016-09-09). https://doi.org/10.48550/arXiv.1609.02907.

[21] CUI W, SHANG M. KAGN: Knowledge-powered attention and graph convolutional networks for social media rumor detection [J]. Journal of Big Data, 2023, 10(1): 45.

[22] WANG P, TAO L, TANG M, et al. Incorporating syntax and semantics with dual graph neural networks for aspect-level sentiment analysis[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108101.

[23] LU Q, SUN X, GAO Z, et al. Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis[J]. Information Processing & Management, 2024, 61(1): 103538.

[24] SHANG W, CHAI J, CAO J, et al. Aspect-level sentiment analysis based on aspect-sentence graph convolution network[J]. Information Fusion, 2024, 104: 102143.

[25] 常远, 季长伟, 张春玲, 等. 融合多特征嵌入的中文医疗命名实体识别模型MF-MNER[J]. 小型微型计算机系统, 2024, 45(12): 2915-2922.
CHANG Yuan, JI Changwei, ZHANG Chunling, et al. MF-MNER: Chinese medical named entity recognition model fusing multi-feature embedding[J]. Journal of Chinese Computer Systems, 2024, 45(12): 2915-2922.

[26] JAIN A, SHARMA R. Enhancing legal named entity recognition using RoBERTa-GCN with CRF: A nuanced approach for fine-grained entity recognition[M]//Advances in Information Retrieval. Cham: Springer Nature Switzerland, 2024: 261-267.

[27] 范涛, 王昊, 陈玥彤. 基于深度迁移学习的地方志多模态命名实体识别研究[J]. 情报学报, 2022, 41(4): 412-423.
FAN Tao, WANG Hao, CHEN Yuetong. Research on multimodal named entity recognition of local history based on deep transfer learning[J]. Journal of the China Society for Scientific and Technical Information, 2022, 41(4): 412-423.

[28] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[29] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016: 770-778.

[30] SINGH V K, ABDEL-NASSER M, RASHWAN H A, et al. FCA-Net: Adversarial learning for skin lesion segmentation based on multi-scale features and factorized channel attention[J]. IEEE Access, 2019, 7: 130552-130565.

[31] 谢小东, 吴洁, 盛永祥, 等. 基于多模态特征融合的相似专利识别方法研究[J]. 图书情报工作, 2024, 68(18): 112-122.
XIE Xiaodong, WU Jie, SHENG Yongxiang, et al. Research on similar patent identification based on multimodal feature fusion

[J]. Library and Information Service, 2024, 68(18): 112-122.

[32] 刘佳, 边俊伊. 基于混合深度学习的藏医古籍命名实体识别研究[J]. 现代情报, 2023, 43(11): 37-46.
LIU Jia, BIAN Junyi. Research on named entity identification of tibetan medical ancient books based on hybrid deep learning[J]. Journal of Modern Information, 2023, 43(11): 37-46.

[33] 李纲, 潘荣清, 毛进, 等. 整合BiLSTM-CRF网络和词典资源的中文电子病历实体识别[J]. 现代情报, 2020, 40(4): 3-12, 58.
LI Gang, PAN Rongqing, MAO Jin, et al. Entity recognition of Chinese electronic medical records based on BiLSTM-CRF network and dictionary resources[J]. Journal of Modern Information, 2020, 40(4): 3-12, 58.

[34] 王永胜, 李培峰, 王中卿, 等. 多模态信息抽取研究综述[J]. 软件学报, 2025, 36(4): 1665-1691.
WANG Yongsheng, LI Peifeng, WANG Zhongqing, et al. Survey on multimodal information extraction research[J]. Journal of Software, 2025, 36(4): 1665-1691.

[35] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[36] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. (2015-08-09). https://doi.org/10.48550/arXiv.1508.01991.

[37] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2018-10-11). https://doi.org/10.48550/arXiv.1810.04805.

[38] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized BERT pre-training approach[EB/OL]. (2019-07-26). https://doi.org/10.48550/arXiv.1907.11692.

[39] YU J, JIANG J, YANG L, et al. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: [s.n.], 2020: 3342-3352.

[40] ZHANG D, WEI S, LI S, et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14347-14355.

[41] SUN L, WANG J, ZHANG K, et al. RpBERT: A text-image relation propagation-based BERT model for multimodal NER [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(15): 13860-13868.

作者简介:

韩普(1983-),**通信作者**,男,博士,教授,研究方向:智能信息处理、医疗健康数据分析及应用,E-mail:hanpu@njupt.edu.cn。

刘森嶺(2002-),男,硕士研究生,研究方向:数据分析与处理。

陈文祺(2000-),男,硕士研究生,研究方向:数据分析与处理。

(编辑:张蓓,王婕)