文章编号：1003-0077（2017）00-0000-00

基于知识增强的多视野表征学习辅助诊断方法

王好天[1] 李鑫[1] 关毅[1] 杨洋[1] 李雪[1] 姜京池[2]

（1.哈尔滨工业大学 语言技术研究中心，黑龙江 哈尔滨 150001；
2.哈尔滨工业大学 物联网与泛在智能中心，黑龙江 哈尔滨 150001）

摘要：针对辅助诊断过程中病人所患疾病不单一，多种疾病之间存在内在关联，及长病历文本特征提取较为困难等问题，该文提出一种基于知识增强的多视野表征学习方法。该方法首先使用 Bi-LSTM 和注意力网络、医疗知识图融合、预训练模型分别从字符视野、实体视野、文档视野提取疾病表征，并通过融合多视野信息从长病历文本中准确抽取疾病诊断相关特征。而后建模疾病间内在关联关系，基于图神经网络方法进行知识融合以增强疾病表征，并实现疾病预测。该模型利用多视野表征学习与知识增强方法，提升了疾病预测的性能，通过结果可视化为模型提供了可解释性。在华为云杯评测数据上的实验表明该方法优于其他基线方法，消融实验证明了该方法各模块的有效性。

关键词：知识增强；多视野表征学习；辅助诊断；多标签分类

中图分类号：TP391 **文献标识码：**A

Multi-view Representation Learning Network Based on Knowledge Augmentation for Auxiliary Diagnosis

WANG Haotian[1], LI Xin[1], GUAN Yi[1], YANG Yang[1], LI Xue[1], and JIANG Jingchi[2]

(1.Language Technology Research Center, Harbin Institute of Technology, Harbin 150001, China ;
2. AIoT Research Center, Harbin Institute of Technology, Harbin 150001, China)

Abstract : To model internal correlations between diseases and extract features from long medical records, we propose a multi-view representation learning network based on knowledge augmentation for auxiliary diagnosis. Firstly, the method uses Bi-LSTM and attention network, medical knowledge graph fusion, and pre-trained models to extract disease representations from character view, entity view, and document view, respectively. Then, the features related to disease diagnosis are accurately extracted from the long medical record text by the fusion of multi-view information. Secondly, the internal correlation between diseases is modeled by knowledge fusion based on the graph neural network to enhance disease representation. Finally, the model uses multi-view representation learning and knowledge enhancement methods to improve the performance of predicting disease. Besides, we provide interpretability by result visualization. Experiments on Huawei Cloud evaluation dataset show that the model is superior to other baseline methods, and ablation studies prove the effectiveness of each module in this method.

Key words: knowledge augmentation; multi-view representation learning; auxiliary diagnosis; multi-label classification

0 引言

随着人们生活水平的不断提高，医疗健康问题也逐渐成为大众日益关注的重要民生问题。国

收稿收稿日期：20XX- XX-XX；**定稿日期**：20XX- XX-XX
基金项目：科技创新 2030——“新一代人工智能”重大项目（2021ZD0113302）

家卫健委在国卫基层函〔2018〕195 号[①]中启动了“优质服务基层行”活动，该文件指出乡镇卫生院及社区卫生服务中心需要至少具备识别和初步诊治 50 种常见病、多发病的能力。基层医院直接对接群众，是医疗卫生服务过程中重要的一环，但由于就诊压力大,医疗设施不完善,医生能力有限等问题，基层医院的误诊情况时有发生。

近年来，人工智能技术取得了重大突破，相关技术已经在医疗、金融、法律等领域实现落地应用[1,2]，这也使利用人工智能技术为医生提供辅助诊断成为可能。医生在问诊时会主动收集患者的性别、年龄、主诉、现病史和既往史等信息，并综合各项信息完成初步诊断。因此利用人工智能技术构建一个基于电子病历（Electronic Medical Record，EMR）文本信息的辅助诊断系统可以为医生提供诊断建议，有效减少误诊的发生。

早在 20 世纪 60 年代，Lealey 等人[3]提出了机器疾病诊断的概念，吸引了大批学者投身于该领域的研究。传统疾病诊断方法需要医学专家人工制定特征，面临海量数据时无法展现出很好的处理能力。随着深度学习技术的快速发展与计算机硬件水平的提高，深度学习模型在海量数据面前表现出了优异的性能，也为辅助诊断提供了良好的思路。目前研究者主要使用序列模型和注意力机制（Attention Mechanism）抽取电子病历中的序列特征以实现疾病的辅助诊断。然而，电子病历文本中含有丰富的专业术语与医学关系，其中蕴含丰富的医学知识，目前深度学习模型难以建模这种结构化知识。因此研究者们开始尝试利用图神经网络抽取电子病历文本中结构化知识表示帮助辅助诊断模型提升性能。

虽然现有模型在辅助诊断任务上已经取得了较大进展，但是仍然存在以下不足：（1）电子病历文本较长，传统的序列模型无法准确地从长文本中抽取诊断需要的特征；（2）病人所患疾病不单一，且疾病之间一般会存在关联，如“糖尿病”与“视网膜病变”之间存在着“并发症”关系，现有模型并未考虑疾病之间的内在关联。

针对现有方法的不足，本文提出了一个基于知识增强的多视野表征学习辅助诊断模型（Multi-View Representation Learning Network，MVRLN）。该模型使用双向长短时记忆网络（Bidirectional Long-Short Term Memory Network，Bi-LSTM）提取语义信息，并利用注意力机制提取疾病的字符视野表征；基于医疗知识图抽取疾病的实体视野表征；基于预训练语言模型（Pretrain Language Model，PLM）抽取疾病的文档视野表征。将不同视野疾病表征融合后，基于疾病关系图利用知识增强方式融合疾病标签之间的内在关系，进而实现疾病辅助诊断。

本文的主要贡献有：

（1）提出一种多视野疾病表征学习方法，分别从字符、实体、文档三个视野抽取疾病表征，有效缓解难以从长文本准确抽取疾病表征的问题；

（2）提出一种基于疾病关系图的知识增强方法，利用图神经网络通过知识融合增强疾病表征，更好地建模疾病之间的内在关系；

（3）在“华为云杯”测评数据集上的实验表明，该模型相比于基线方法有显著提升，验证了该模型的有效性，项目代码已开源[②]。

1 相关工作

随着人们对自身健康关注度的不断提高，基于电子病历的辅助诊断研究也成为了人工智能应用技术的热点之一。现有方法主要是基于序列特征、结构化知识和与预训练语言模型的疾病诊断方法。

1.1 基于序列特征的辅助诊断方法

随着数据量的不断增加，深度学习技术凭借自动提取特征的能力及强大的学习能力受到研究学者们的青睐，开始被应用于疾病辅助诊断任务中，如卷积神经网络（Convolutional Neural Network，CNN）和循环神经网络（Recurrent Neural Network，RNN）。Yang 等人[4]提出使用 CNN 从电子病历文本中抽取高阶语义信息进行疾病诊断。Mullenbach 等人[5]提出了 CAML 模型,通过 CNN 聚合电子病历文本中的特征信息，并利用 Attention 机制得到与每个疾病最相关的特征信息用于疾病诊断。Li 等人[6]提出了 MultiResCNN 模

① http://www.nhc.gov.cn/jws/new_index.shtml

② https://github.com/FutureForMe/MVRLN

型，该模型使用多尺度 CNN 提取文本中的多元特征，通过残差网络扩大感受野，并利用注意力网络得到每种疾病对应的表示用于疾病诊断。Vu 等人[7]使用 Bi-LSTM 提取文本中的序列特征，并通过注意力网络得到每种疾病表示进行疾病诊断。

1.2 基于结构化知识的辅助诊断方法

自从图卷积神经网络[8]（Graph Convolutional Neural Networks，GCN）提出以来，就受到人们的广泛关注。由于医疗文本中包含丰富的专业术语和医学关系，研究学者开始尝试利用电子病历中的结构化知识帮助模型提升性能。刘勘等人[9]使用知识表示模型学习医疗知识图谱结构化知识表示，并与文本特征融合进行并发症的疾病诊断。Zhao 等人[10]基于电子病历数据构造了医疗知识图谱，并利用知识图谱嵌入方法进行概率推理，进行疾病的辅助诊断。Wang 等人[11]使用多尺度标签注意力抽取疾病标签的对应特征，提出溯因因果图融合结构化特征进行辅助诊断。Xie 等人[12]提出了 MSATT-KG 模型，该模型融入了疾病之间树状的层次信息，并结合 CNN 抽取的多元特征信息进行疾病诊断。Yuan 等人[13]从电子病历中抽取结构化因果知识并与文本特征通过注意力网络融合后用于疾病诊断。

1.3 预训练语言模型

预训练语言模型已经在自然语言处理任务上取得了优异的表现，如 Bert[14]、ERNIE[15]等。研究者开始尝试训练医学领域的预训练模型完成疾病诊断任务。Gu 等人[16]使用医疗文献语料库训练 PubMedBERT 模型，并在医疗命名实体识别、文本分类等任务上取得较好效果。Huang 等人[17]提出 PLM-ICD 模型，使用医学领域知识对 PLM 进行预训练，并将长病历文本划分为不同片段微调 PLM，进而完成疾病分类编码任务。

以上方法利用医疗文本中的序列信息建模病人表征，忽视了文本中的多粒度信息，难以从长医疗文本中准确抽取疾病相关特征。此外，疾病诊断任务中的标签并不相互独立，疾病间存在内在关联。为了解决以上问题，本文提出一种基于知识增强的多视野表征学习方法，同时考虑了文本中的多粒度特征以及疾病标签之间的内在关联，有效提升了模型性能。

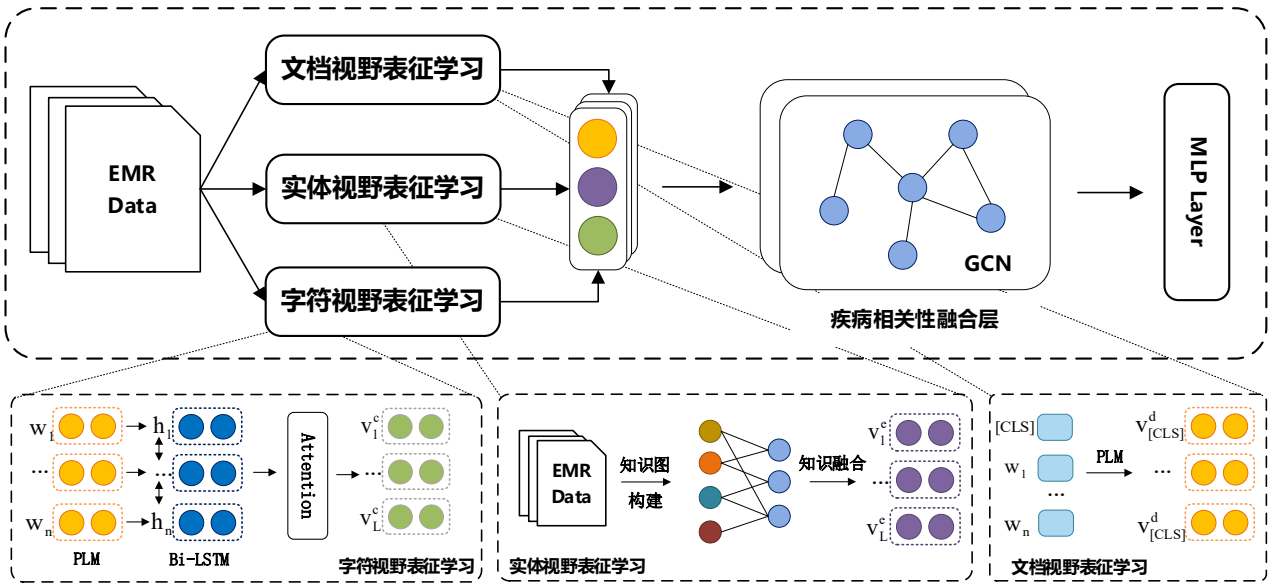


图 1 模型的整体框架

2 辅助诊断方法

本文提出了一种基于知识增强的多视野表征学习辅助诊断方法，框架如图 1 所示。模型主要分为两个模块：疾病的多视野表征学习模块和基于知识增强的标签关系融合模块。

疾病的多视野表征学习模块主要考虑三个视野的特征信息，即字符视野表征学习、实体视野表征学习和文档视野表征学习。三个视野分别对长医疗文本进行不同粒度的疾病表征学习，相比于单一视野可以为每种疾病抽取更丰富的特征。

为了融合疾病标签之间的内在关联信息，利用训练集数据构造标签关系知识图，并使用 GCN 进行知识融合以增强疾病表示，用于最终的疾病辅助诊断。

2.1 字符视野表征学习

2.1.1 Embedding 层

给定医疗电子病历文本 X，对 X 切分可得到对应的token序列$[\text{token}_1, \text{token}_2, \dots, \text{token}_n]$, n表示文本中token的个数。鉴于 PLM 已经在众多自然语言处理任务中取得优异表现，本文使用 PLM 例如 Bert[14]，得到每个token的表示向量$\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_n] \in R^{n\times d}$, d表示词向量的维度。

2.1.2 Bi-LSTM 层

为了提取文本中的语义特征，得到更加契合当前语境的向量表示，我们将文本中token的表示向量作为输入，使用 Bi-LSTM 提取字符级语义信息，获得字符表征，如式（1）至（2）所示。

$$\overrightarrow{\boldsymbol{h}_i} = \overrightarrow{\text{LSTM}}(\boldsymbol{w}_i) \quad (1)$$

$$\overleftarrow{\boldsymbol{h}_i} = \overleftarrow{\text{LSTM}}(\boldsymbol{w}_i) \quad (2)$$

$$\boldsymbol{h}_i = \left[\overrightarrow{\boldsymbol{h}_i}, \overleftarrow{\boldsymbol{h}_i}\right] \quad (3)$$

其中，$\overrightarrow{LSTM}$和$\overleftarrow{LSTM}$分别表示前向编码 LSTM 和后向编码 LSTM，token_i通过拼接得到其最终的向量表示$\boldsymbol{h}_i$，如式（3）所示。Bi-LSTM 可以获得每个token的向量表示$\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \dots, \boldsymbol{h}_n] \in R^{n\times 2u}$, u表示 LSTM 的隐状态向量的维度。

2.1.3 Attention 层

通过 Bi-LSTM 层可以得到每个字符的表示$\boldsymbol{H} \in R^{n\times 2u}$，由于每个标签对于token的关注程度不同，因此模型使用 Attention 机制为每个标签提取对应的表征，如式（4）至（5）所示。

$$\mathbf{A} = softmax(\boldsymbol{WH}) \quad (4)$$

$$\boldsymbol{V}^{char} = \boldsymbol{AH} \quad (5)$$

其中，$\boldsymbol{W} \in \boldsymbol{R}^{2u\times|\boldsymbol{L}|}$为可学习的参数矩阵，用来计算疾病标签与不同字符之间的注意力权重$\boldsymbol{A} \in \boldsymbol{R}^{n\times|\boldsymbol{L}|}$。$\boldsymbol{V}^{char} \in \boldsymbol{R}^{|\boldsymbol{L}|\times 2u}$为使用注意力权重$\boldsymbol{A}$和字符表示$\boldsymbol{H}$计算得到的字符视野疾病表征。其中，$|\boldsymbol{L}|$表示疾病标签的数量。

2.2 实体视野表征学习

2.2.1 医疗知识图构建

在进行实体视野表征学习之前需要根据现有训练集中的电子病历数据构造医疗知识图。首先利用命名实体识别（Named Entity Recognition，NER）技术得到病历文本中的医疗实体；而后根据限定词去掉否认实体（例如“否认高血压”），保留肯定实体。

在得到医疗实体之后，本文借鉴条件概率公式计算医疗实体与疾病标签之间的相关程度，如式（6）所示：

$$\boldsymbol{A}_{i,j} = \frac{n(d_i, f_j)}{n(d_i)+\lambda} \quad (6)$$

其中，d_i表示第i个疾病标签，f_j表示第j个医疗实体，$n(d_i, f_j)$表示在出现疾病标签d_i与医疗实体f_j共同出现的次数，$n(d_i)$表示疾病标签d_i出现的次数，λ是一个超参数，用以消除噪声实体带来的影响。$\boldsymbol{A}_{i,j}$为疾病标签d_i与医疗实体f_j之间的相关性得分。最终，可通过阈值θ对其进行限定以得到标签关系的邻接矩阵$\widehat{\boldsymbol{A}}_{i,j}$，如式（7）所示。

$$\widehat{\boldsymbol{A}}_{i,j} = \begin{cases} \boldsymbol{A}_{i,j}, & if\ \boldsymbol{A}_{i,j} > \theta \\ 0, & else \end{cases} \quad (7)$$

2.2.2 实体视野表征学习

医疗电子病历文本中往往会含有较多的专业术语和医学关系，且文本较长，传统的基于序列模型的辅助诊断模型无法很好地捕捉长距离依赖信息以及文本中蕴含的结构化信息。我们首先使用预训练模型得到实体的初始化向量$\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_{2,\dots,}\boldsymbol{e}_m] \in \boldsymbol{R}^{m\times d}$, m为实体数量，d表示实体向量维度。而后根据实体是否在病人病历中出现构造患者个性化 one-hot 表示$\boldsymbol{p} = [p_1, p_2, \dots, p_m]$, $p_i \in \{0,1\}$。为了融合患者个性化实体特征，我们根据病人的个性化表示对实体向量进行掩码，并基于医疗知识图进行知识融合，如式（8）所示：

$$\boldsymbol{V}^{G} = \widehat{\boldsymbol{A}}_{E}(\boldsymbol{p} * \boldsymbol{E}) \quad (8)$$

其中，$\widehat{\boldsymbol{A}}_{E} \in R^{|\boldsymbol{L}|\times m}$是医疗知识图的邻接矩阵。通过上述过程可以融合病历中实体邻居节点的信息，得到实体视野的疾病表征$\boldsymbol{V}^{ent} \in \boldsymbol{R}^{|L|\times d}$。

2.3 文档视野表征学习

由于预训练模型中[CLS]位置的表示可以融合整段文本的特征信息，因此本文将$\{[\text{CLS}], \text{token}_1, \text{token}_2, \dots, \text{token}_n, [\text{SEP}]\}$作为预训练模型的输入，在预训练模型学习过程中[CLS]关注整段上下文特征，因此本文直接将[CLS]的表示作为文档视野疾病的表征$\boldsymbol{V}^{doc} \in \boldsymbol{R}^{|L|\times d}$。

模型在得到字符视野、实体视野、文档视野的疾病表征之后，将其拼接得到多视野表征学习的疾病表征，如式（9）所示。

$$\boldsymbol{V}^{m}=[\boldsymbol{V}^{char};\boldsymbol{V}^{ent};\boldsymbol{V}^{doc}]\in\boldsymbol{R}^{|L|\times(2u+2d)} \quad (9)$$

2.4 基于知识增强的标签关系融合

在实际诊断过程中，病人患有的疾病往往不单一，且疾病之间不相互独立，会存在一些内在联系，例如“糖尿病”与“视网膜病变”之间存在着“并发症”关系，即患有糖尿病的病人患有视网膜病变的概率要比未患有糖尿病的病人患有视网膜病变的概率高。因此在辅助诊断的过程中考虑疾病标签之间的内在关系会对疾病诊断结果产生积极影响。

为捕捉疾病标签之间的内在联系，本文使用训练集电子病历，利用统计概率方法计算疾病标签之间协同关系的概率，并构建标签相关图（Disease Correlation Graph，DCG），计算过程同2.2.1节。将多视野表征学习得到的疾病表示向量作为疾病标签知识图中疾病节点的初始化向量$\boldsymbol{H}^{0}=\boldsymbol{V}^{m}$，使用GCN进行知识融合，融合不同疾病标签之间关系，如式（10）所示。

$$\boldsymbol{H}^{(l)}=\sigma(\widetilde{\boldsymbol{D}}_{C}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}_{C}\widetilde{\boldsymbol{D}}_{C}^{-\frac{1}{2}}(\boldsymbol{W}\boldsymbol{H}^{(l-1)})) \quad (10)$$

其中，$\widetilde{\boldsymbol{A}}_{C}=\widehat{\boldsymbol{A}}_{C}+\boldsymbol{I}_{C}$，$\widehat{\boldsymbol{A}}_{C}$是标签相关图的邻接矩阵，$\boldsymbol{I}_{C}$是单位矩阵，$\widetilde{\boldsymbol{D}}_{C}$是度矩阵，$l$是GCN的层数。为避免GCN的过平滑问题，我们借鉴残差思想得到最终疾病标签的表示$\boldsymbol{V}^{final}=(\boldsymbol{H}^{0}+\boldsymbol{H}^{1}+\boldsymbol{H}^{2})\in\boldsymbol{R}^{|L|\times(2u+2d)}$（以2层GCN为例），将其输入多层感知机（Multi-Layer Perceptron，MLP）进行疾病预测，如式（11）所示。

$$\hat{y}=\mathrm{sigmoid}(\mathrm{MLP}(\boldsymbol{V}^{final})) \quad (11)$$

通过预定义的阈值计算二元预测结果$\hat{y}_{j}\in\{0,1\}$以得到疾病的最终预测概率。训练的目标为最小化二元交叉熵损失，如式（12）所示。

$$L(E,G,y,\Phi)=-\sum_{j=1}^{L}y_{j}log\hat{y}_{j}+(1-y_{j})\log(1-\hat{y}_{j}) \quad (12)$$

其中，Φ表示所有可训练参数，E表示输入的文本序列，G表示实体视野的知识图和标签相关图。

3 实验设置

3.1 数据集

本文实验用到的所有数据均来自于“华为云杯”2022人工智能创新应用大赛[③]中的真实电子病历数据，对数据筛选后选取诊断标签较多的电子病历构成本文数据集。该数据已经过严格去隐私化处理，病历包含病人的性别、年龄、主诉、现病史、既往史和诊断疾病字段，具体统计信息如表1所示。

表1 数据集信息统计表

名称	数量
训练集病历数	5 422
验证集病历数	721
测试集病历数	720
疾病种类	51
电子病历平均长度	421.61
平均诊断数	2.45

本文将性别、年龄、主诉拼接作为一个字段，并与现病史、既往史共同作为模型输入，字段间使用[SEP]分隔。在构造医疗知识图过程中使用医生人工标注的实体数据集训练Bert+Bi-LSTM+CRF模型[18]，该数据由6511份电子病历构成，包括5种实体类型：疾病、症状、治疗、检查和检查结果，具体统计结果见表2。训练集和验证集划分比例为8:2，最优模型在验证集上的Macro F_1为95.07%，并利用最优模型对本文数据集的电子病历文本进行实体识别。

表2 实体数据信息统计表

实体类型	实体数量
疾病	37 378
症状	69 865
检查	70 507
检查结果	96 696
治疗	36 368

3.2 基线方法

为了全面分析本文提出的MVRLN方法的效果，我们选取了以下基线方法进行实验对比：

- **TextCNN**[19]：使用卷积神经网络抽取文本特征进行疾病辅助诊断。
- **TextRNN**[20]：使用循环神经网络（Bi-LSTM）抽取文本特征进行疾病辅助诊断。
- **TextRCNN**[21]：首先使用循环神经网络提取文本语义特征，而后利用最大池化进行疾病辅助诊断。

③ https://competition.huaweicloud.com

- **CAML**[5]：使用多尺度卷积神经网络进行疾病辅助诊断。
- **MultiResCNN**[6]：利用多尺度卷积神经网络与残差神经网络提取文本 n-gram 特征，并使用 Attention 得到每个标签的表示用于辅助疾病诊断。
- **LAAT**[7]：利用 Bi-LSTM 提取文本时序特征，且通过 Label Attention 得到标签表示并进行疾病辅助诊断。
- **KenMeSH**[22]：利用 Bi-LSTM 分别提取不同字段表示，并结合动态知识增强技术进行疾病辅助诊断。
- **Bert**[12]：使用 Bert 预训练模型④在病历文本数据上微调实现辅助疾病诊断。
- **ERNIE**[13]：使用 ERNIE Health 模型⑤在病历数据上微调实现疾病辅助诊断。
- **Longformer**[23]：长文本预训练模型⑥，在病历数据上微调进行辅助疾病诊断。

3.3 实验设置

模型代码基于 PyTorch 框架实现，使用 RTX 3090 24G 显卡训练测试。预训练模型获取的词向量维度统一为 768；Dropout 参数设置为 0.5，预训练模型学习率为 0.00001，非预训练模型学习率为 0.0005，优化器为 AdamW，GCN 层数为 2；batch size 设置为 12，epochs 设置为 30；阈值θ为 0.5，d_a为 128，LSTM 隐变量u为 384。所有实验结果均取随机种子训练 5 次的平均值。

3.4 评价指标

本方法将辅助诊断任务作为一个多标签分类任务，评价指标为Macro P，Macro R，Macro F_1和Micro P, Micro R, Micro F_1，计算公式如下所示：

$$\text{Macro P}_i = \frac{\overline{\text{TP}_i}}{\overline{\text{TP}_i} + \overline{\text{FP}_i}}$$

$$\text{Macro R}_i = \frac{\overline{\text{TP}_i}}{\overline{\text{TP}_i} + \overline{\text{FN}_i}}$$

$$\text{Macro F}_1 = \sum_{i=1}^{|L|} \frac{2 \times \text{Macro P}_i \times \text{Macro R}_i}{\text{Macro P}_i + \text{Macro R}_i}$$

$$\text{Micro P} = \frac{\overline{\text{TP}}}{\overline{\text{TP}} + \overline{\text{FP}}}$$

$$\text{Micro R} = \frac{\overline{\text{TP}}}{\overline{\text{TP}} + \overline{\text{FN}}}$$

$$\text{Micro F}_1 = \frac{2 \times \text{Micro P} \times \text{Micro R}}{\text{Micro P} + \text{Micro R}}$$

其中，$|L|$表示疾病标签的数量，在计算Macro F_1时，首先计算所有疾病类别的F_1，然后取所有类别的平均值作为结果。在计算Micro F_1时，将每一个样本数据作为独立个体进行预测，并计算F_1。

表 3 对比实验结果

Model Type	Model	Macro P	Macro R	Macro F_1	Micro P	Micro R	Micro F_1
Deep Learning	TextRNN	40.12%	26.57%	31.96%	76.59%	59.38%	66.86%
	TextCNN	51.96%	46.36%	48.99%	76.78%	73.20%	74.94%
	TextRCNN	42.03%	44.20%	43.03%	70.17%	**75.44%**	72.69%
	CAML	42.04%	33.21%	37.07%	79.12%	66.89%	72.48%
	MultiResCNN	50.73%	45.18%	47.76%	78.07%	72.84%	75.36%
	LAAT	48.48%	38.54%	42.93%	79.85%	71.15%	75.25%
	KenMeSH	40.78%	36.92%	37.17%	76.75%	72.66%	74.65%
Pretrain Language Model	Bert	51.87%	42.60%	46.73%	78.50%	70.94%	74.51%
	ERNIE	54.95%	44.33%	49.06%	80.72%	71.86%	76.03%
	Longformer	53.87%	47.21%	50.32%	80.13%	73.77%	76.81%
Our Model	MVRLN_B	56.10%	44.65%	49.72%	82.61%	71.62%	76.72%
	MVRLN_E	58.53%	48.01%	52.75%	81.33%	73.39%	77.15%
	MVRLN_L	**60.14%**	**48.16%**	**53.49%**	**83.28%**	73.16%	**77.89%**

4 实验结果及分析

4.1 实验结果分析

为验证方法的有效性，我们分别结合不同的预训练模型进行向量初始化，并在数据集上与不同的基线方法进行实验对比，实验结果如表 3 所示。其中，MVRLN_B、MVRLN_E和MVRLN_E分别表示

④ https://huggingface.co/bert-base-chinese
⑤ https://huggingface.co/nghuyong/ernie-health-zh
⑥ https://github.com/IDEA-CCNL/Fengshenbang-LM

使用 Bert、ERNIE 和 Longformer 预训练模型初始化向量的模型。通过实验结果可以看出：

（1）本文提出的方法效果超过了所有基线方法，获得了最佳效果。其中，$MVRLN_L$在F_1指标上取得所有基线方法中的最佳结果，其Macro F_1和Micro F_1分别为 53.49%和 77.89%。实验结果表明，多视野表征学习模块能抽取出疾病更准确的表示，通过知识增强技术融合疾病标签内在关联后可以有效提升模型性能，验证了本文方法的有效性。

（2）在所有深度学习基线方法中 TextRNN 模型效果最差，主要原因是电子病历文本较长，而 RNN 模型无法很好地提取长距离依赖特征。LAAT 模型在 Bi-LSTM 基础上增加了 Label Attention 机制，可以为每个疾病提取最相关的特征，有效缓解长距离依赖问题，使得性能有所提升。

TextCNN 相比 TextRNN 效果更好，这是因为电子病历中含有较多的医学名词，使用 CNN 可以抽取文本中的多元特征，取得更好的模型效果。MultiResCNN 在 CNN 基础上进一步改进，通过残差网络增加感受野范围，设置不同卷积核提取特征，取得了深度学习基线方法中的最优效果。

相比于 MultiResCNN 模型，本文提出的方法在 Macro F_1 上提升 5.73%，在 Micro F_1 上提升 2.53%，这表明疾病内在关联在疾病诊断过程中具有优越性，通过知识增强技术可以有效建模疾病间的结构化特征。相比于知识引导的学习方法 KenMeSH，本文提出的方法在Macro F_1上提升 16.32%，在Micro F_1上提升 3.42%，这表明相比于仅使用 Bi-LSTM 提取字符级疾病特征，增加实体视野的结构特征以及文档级的整体特征可以更准确地捕捉疾病相关的语义信息，获得更全面的疾病特征表示。

（3）在所有预训练模型基线方法中，$MVRLN_B$相比于 Bert 在Macro F_1和Micro F_1上提升 2.99%和 2.21%，$MVRLN_E$相比于 ERNIE 在Macro F_1和Micro F_1上提升了 3.69%和 1.12%，$MVRLN_L$相比于 Longformer 在Macro F_1和Micro F_1上提升了 3.17%和 1.08%。这说明本文提出的方法在不同的预训练模型基础上均可以提升模型性能，取得更好的实验结果。

由于医疗文本长度较长，受预训练模型输入文本长度限制，$MVRLN_B$和$MVRLN_E$无法有效建模长文本中的特征，因此$MVRLN_L$相比$MVRLN_B$和$MVRLN_E$凭借着 Longformer 可以建模长文本特征的优势取得更优异的结果。

4.2 消融实验

我们对本文方法的不同模块进行了消融实验，实验结果如表 4 所示。其中，w/o char 表示去除字符视野表征，w/o ent 表示去除实体视野表征，w/o doc 表示去除文档视野表征，w/o DCG 表示去除标签关系融合模块。通过实验结果可以看出：

表 4 $MVRLN_L$消融实验结果

模型	Macro F1	Micro F1
$MVRLN_L$	**53.49%**	**77.89%**
w/o char	52.37%	77.12%
w/o ent	53.12%	77.72%
w/o doc	49.42%	77.21%
w/o DCG	51.48%	77.44%

（1）通过对多视野表征学习模块的消融实验可以发现，相比于 w/o char、w/o ent、w/o doc 方法，本文提出的$MVRLN_L$在Macro F_1上分别提升了 1.12%，0.37%，3.70%；在Micro F_1上分别提升了 0.77%，0.17%，0.51%。其中，在Macro F_1指标文档视野特征贡献度最大，在Micro F_1指标字符视野特征贡献度最大，由此可以看出：文档视野从粗粒度学习疾病特征，可以更全面地提取每类疾病特征，因此对少样本数据的关注度更高；而字符视野使用 Bi-LSTM 和 Label Attention 学习医疗文本的细粒度信息，使得学习到的疾病特征更关注于样本数量较多的疾病。

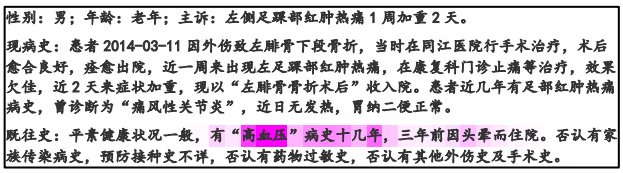
此外，实体视野在Micro F_1和Micro F_1贡献度最小，其主要原因是本文构建实体视野的知识图时仅使用电子病历中的主诉、现病史、既往史字段，文本较少，包含的实体信息并不丰富，后续会融入更多的病历文本，构建蕴含更丰富知识的图。通过以上实验可以验证多视野表征学习模块中每个视野均能在疾病诊断过程中起到作用，这也表明该模块可以帮助模型取得更好的结果。

（2）通过对比基于知识增强的标签关系融合模块的消融实验可以发现，相比 w/o DCG 方法，本文提出的$MVRLN_L$在Macro F_1和Micro F_1上提升 1.64%和 0.28%，证明该模块在疾病诊断过程中起到了积极作用。该模块在Macro F_1指标上的贡献度高于Micro F_1指标上的贡献度，这表明通过知识

增强技术融合标签之间的内在联系，可以增强少样本疾病的表示，保证在辅助诊断任务中更多地关注样本较少的疾病。

4.3 注意力可视化

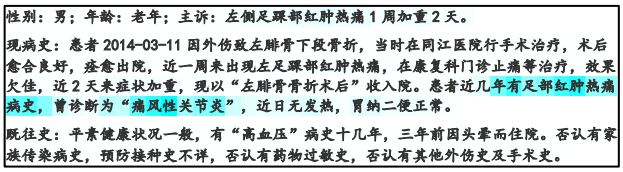
为了探究模型是否关注了医疗电子病历中与疾病最相关的文本片段，我们对字符视野表征学习中 Label Attention 的 score 进行了可视化。随机从测试集中选取一个病人，该病人的入院诊断为“高血压”、“关节炎”和“骨折”，可视化结果如图 2 至图 4 所示。图 2 表示“高血压”疾病注意力可视化图，图 3 表示“关节炎”疾病注意力可视化图，图 4 表示“骨折”疾病注意力可视化图，颜色越深说明该片段对该疾病越重要。


性别：男；年龄：老年；主诉：左侧足踝部红肿热痛 1 周加重 2 天。

现病史：患者 2014-03-11 因外伤致左腓骨下段骨折，当时在同江医院行手术治疗，术后愈合良好，痊愈出院，近一周来出现左足踝部红肿热痛，在康复科门诊止痛等治疗，效果欠佳，近 2 天来症状加重，现以“左腓骨骨折术后”收入院。患者近几年有足部红肿热痛病史，曾诊断为“痛风性关节炎”，近日无发热，胃纳二便正常。

既往史：平素健康状况一般，有“高血压”病史十几年，三年前因头晕而住院。否认有家族传染病史，预防接种史不详，否认有药物过敏史，否认有其他外伤史及手术史。

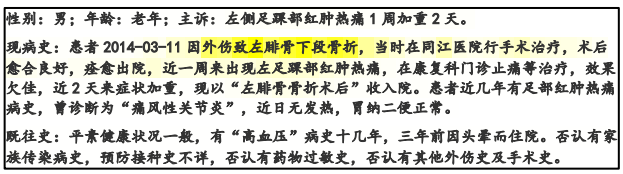
图 2 高血压注意力可视化图


性别：男；年龄：老年；主诉：左侧足踝部红肿热痛 1 周加重 2 天。

现病史：患者 2014-03-11 因外伤致左腓骨下段骨折，当时在同江医院行手术治疗，术后愈合良好，痊愈出院，近一周来出现左足踝部红肿热痛，在康复科门诊止痛等治疗，效果欠佳，近 2 天来症状加重，现以“左腓骨骨折术后”收入院。患者近几年有足部红肿热痛病史，曾诊断为“痛风性关节炎”，近日无发热，胃纳二便正常。

既往史：平素健康状况一般，有“高血压”病史十几年，三年前因头晕而住院。否认有家族传染病史，预防接种史不详，否认有药物过敏史，否认有其他外伤史及手术史。

图 3 关节炎注意力可视化图


性别：男；年龄：老年；主诉：左侧足踝部红肿热痛 1 周加重 2 天。

现病史：患者 2014-03-11 因外伤致左腓骨下段骨折，当时在同江医院行手术治疗，术后愈合良好，痊愈出院，近一周来出现左足踝部红肿热痛，在康复科门诊止痛等治疗，效果欠佳，近 2 天来症状加重，现以“左腓骨骨折术后”收入院。患者近几年有足部红肿热痛病史，曾诊断为“痛风性关节炎”，近日无发热，胃纳二便正常。

既往史：平素健康状况一般，有“高血压”病史十几年，三年前因头晕而住院。否认有家族传染病史，预防接种史不详，否认有药物过敏史，否认有其他外伤史及手术史。

图 4 骨折注意力可视化图

从图 2 可以看出，该病人有“高血压病史十几年”，曾因“头晕”住院，通过以上片段可以判断出该病人患有高血压。

从图 3 可以看出，该病人有“左侧足踝部红肿热痛”、“足部红肿热痛病史”等症状，并且曾被诊断为“痛风性关节炎”，因此可以通过以上信息可推理出该病人患有“关节炎”疾病。

从图 4 可以看出，该病人曾因“外伤致左腓骨下段骨折”行手术治疗，近两天又因“左腓骨骨折术后”入院，因此可以推理出该病人患有“骨折”疾病。

通过对该病人所患三种疾病的可视化结果进行分析可以说明，注意力机制能很好地为每种疾病提取最相关的片段特征，可以帮助模型进行准确地疾病诊断。

5 总结

本文提出了一种基于知识增强的多视野表征学习辅助诊断方法。为解决医疗长文本数据中难以准确抽取疾病相关特征的问题，我们提出了多个视野特征表示模块，分别从字符视野、实体视野和文档视野抽取每种疾病不同粒度的表征。为了解决病人患病不单一，疾病间存在内在关联的问题，我们利用知识增强的方式融合疾病标签之间的内在关系，进一步提升了模型的性能。在真实数据上的实验结果表明，本文提出基于知识增强的多视野表征学习辅助诊断方法可以有效提升疾病诊断的准确率。未来的研究可以尝试引入医疗知识图谱和医疗本体等结构化知识解决罕见疾病的辅助诊断问题，并考虑如何利用逻辑知识增强模型的可解释性问题。

参考文献

[1] Yang Y, Huo H, Jiang J, et al. Clinical decision-making framework against over-testing based on modeling implicit evaluation criteria[J]. Journal of Biomedical Informatics, 2021, 119: 103823.

[2] 安震威,来雨轩,冯岩松. 面向法律文书的自然语言理解[J]. 中文信息学报, 2022, 36(8): 1-11.

[3] Ledley R S, Lusted L B. Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason[J]. Science, 1959, 130(3366): 9-21.

[4] Yang Z, Huang Y, Jiang Y, et al. Clinical assistant diagnosis for electronic medical record based on convolutional neural network[J]. Scientific reports, 2018, 8(1): 1-9.

[5] Mullenbach J, Wiegreffe S, Duke J, et al. Explainable Prediction of Medical Codes from Clinical Text[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018: 1101-1111.

[6] Li F, Yu H. ICD coding from clinical text using multi-filter residual convolutional neural network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 8180-8187.

[7] Vu T, Nguyen D Q, Nguyen A. A label attention model for icd coding from clinical text[J]. arXiv preprint arXiv:2007.06351, 2020.

[8] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.
[9] 刘勘,张雅荃. 基于医疗知识图谱的并发症辅助诊断[J]. 中文信息学报, 2020, 34(10): 85-93,104.
[10] Zhao C, Jiang J, Guan Y, et al. EMR-based medical knowledge representation and inference via Markov random fields and distributed representation learning[J]. Artificial intelligence in medicine, 2018, 87: 4
[11] Wang H, Guan Y, Ma L, et al. Multi-scale Label Attention Network based on Abductive Causal Graph for Disease Diagnosis[C]//2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE Computer Society, 2022: 2542-2549.
[12] Xie X, Xiong Y, Yu P S, et al. Ehr coding with multi-scale feature attention and structured knowledge graph propagation[C]//Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 649-658.
[13] Yuan Q, Chen J, Lu C, et al. The graph-based mutual attentive network for automatic diagnosis[C]//Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. 2021: 3393-3399.
[14] 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.
[15] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.
[16] Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing[J]. ACM Transactions on Computing for Healthcare (HEALTH), 2021, 3(1): 1-23.
[17] Huang C W, Tsai S C, Chen Y N. PLM-ICD: automatic ICD coding with pretrained language models[J]. arXiv preprint arXiv:2207.05289, 2022.
[18] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 260-270.
[19] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2016: 1746－1751.
[20] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning[J]. arXiv preprint arXiv:1605.05101, 2016.
[21] Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]//Twenty-ninth AAAI conference on artificial intelligence. 2015.
[22] Wang X, Mercer R E, Rudzicz F. KenMeSH: Knowledge-enhanced End-to-end Biomedical Text Labelling[J]. arXiv preprint arXiv:2203.06835, 2022.
[23] Beltagy I, Peters M E, Cohan A. Longformer: The long-document transformer[J]. arXiv preprint arXiv:2004.05150, 2020.

王好天（1998—），博士研究生，主要研究领域为自然语言处理、临床推理。
E-mail：wanght1998@hit.edu.cn

李鑫（2000—），硕士研究生，主要研究领域为自然语言处理。
E-mail：22S103169@stu.hit.edu.cn

关毅（1970—），通讯作者，博士，教授，主要研究领域为医疗信息学，知识工程，自然语言处理。
E-mail：guanyi@hit.edu.cn