doi:10.3969/j.issn.1001-4616.2025.03.010

基于 NLP 和图像分类模型的中文科技文献双模态分类方法

王　峥[1],丁　熠[2],陈海明[3],陈　盈[4]

(1.台州学院学报编辑部,浙江 台州 318000)
(2.电子科技大学信息与软件工程学院,四川 成都 610054)
(3.宁波大学信息科学与工程学院,浙江 宁波 315211)
(4.台州学院电子与信息工程学院,浙江 台州 318000)

[**摘要**] 随着当前对科技文献管理和组织要求的急剧增加,对于更为可扩展、精确且自动化的文献分类方式的需求也更高. 为了有效应对海量科技文献数据的分析难题,提出了融合 YOLOv7 图像分类模型和自然语言处理(NLP)模型的多模态文献分析引擎. 该架构充分挖掘文档中的自然语言文本、描述性图像以及两者间的内在关联这 3 种关键信息,通过综合训练流程整合不同模态的深度学习网络,达成相较于单模态分类方法更优的分类精准度. 同时,将所提方法应用到中文科技文献数据集,并依据中图分类号对文献进行了分类训练. 结果表明,所提双模态文献分类方法具有更高的分类准确性,有助于企事业单位和研究机构在数据与知识管理方面的效率提升.

[**关键词**] 科技文献分类,图像分类,多模态特征,自然语言处理,深度学习,YOLOv7
[**中图分类号**]TP391.1　[**文献标志码**]A　[**文章编号**]1001-4616(2025)03-0084-09

A Bimodal Classification Method for Chinese Scientific and Technological Literature Based on NLP and Image Classification Models

Wang Zheng[1],Ding Yi[2],Chen Haiming[3],Chen Ying[4]

(1.Journal Editorial Department,Taizhou University,Taizhou 318000,China)
(2.School of Information and Software Engineering,University of Electronic Science and Technology of China,Chengdu 610054,China)
(3.Faculty of Electrical Engineering and Computer Science,Ningbo University,Ningbo 315211,China)
(4.College of Electronics and Information Engineering,Taizhou University,Taizhou 318000,China)

Abstract:Currently,the demand for more scalable,accurate,and automated document classification is increasing due to the sharp increase in the management and organization of technical literature. To solve the problem of effective data analysis from massive scientific literature data,a multi-modal literature analysis engine is proposed,which combines the YOLOv7 image classification model and natural language processing model. This architecture utilizes three types of information,including natural language text in the document,descriptive images,and the relationship between them. By integrating and training deep learning networks of different modals,the multi-modal approach achieves better classification accuracy than the unimodal method. The proposed method is applied to a Chinese scientific literature dataset,and the model is trained to classify documents based on the Chinese Library Classification system. The results show that the proposed method has higher classification accuracy than unimodal methods,which helps promote data and knowledge management for enterprises,institutions,and research organizations.

Key words:classification of scientific and technological literature,document image classification,multi-modal features,natural langliteratureuage processing,deep learning,YOLOv7

收稿日期:2024-12-12.
基金项目:国家自然科学基金面上资助项目(61976149)、浙江省自然科学基金重点资助项目(Z20F020008)、浙江省普通本科高校“十四五”教学改革资助项目(jg20220563)、2025 年度浙江省自然科学基金资助项目(LMS25A010011)、浙江省科技厅软科学研究计划资助项目(2025C35030).
通讯作者:丁熠,博士,教授,研究方向:计算机视觉,人工智能. E-mail:yi.ding@ uestc.edu.cn;陈盈,教授,研究方向:人工智能与大数据. E-mail:ychen222@ tzc.edu.cn

大数据时代带来了数据资源的快速和大规模增长,也显著增加了数据处理的难度. 传统的数据处理技术存在明显的缺点,对基于大数据的数字图书馆资源服务提出了新要求. 因此,数字图书馆必须在新技术应用和相关编辑人员对于新技术的应用能力方面进行持续改进,如数据分析、决策分析、语义分析、数据质量和数据管理及可视化呈现技术等,以实现数字图书馆的更优化建设和管理[1-2].

科技文献分类有助于用户更轻松地检索和获取相关信息. 当前各种文献和专利大部分已实现了数字化并能通过电子方式传输,但许多数字副本是通过扫描仪或相机生成的. 因此,仍然需要编辑人员对文档进行人工校对、文字提取并检查记录的一致性. 此外,科技文献分类的工作量极大,且包含复杂的技术术语,需要掌握特定专业知识的编辑人员来处理. 在对多层次分类系统进行广泛和深入的分类时,编辑人员可能无法准确确定特定文档的标签或类别. 为此,需要通过人工智能方法来减少时间和成本,并确保技术文档的准确分类[3-4].

在计算机视觉领域,数字文件分析得到了广泛研究. 为进行文本搜索和索引,数字化的关键部分是从扫描或拍摄的文档中提取文本内容. 为此,国内外提出了许多光学字符识别(optical character recognition,OCR)算法,将图像转换为字符串,并对复杂文档中的文本进行准确定位和分析. 文献[5]提出了基于卷积循环神经网络的 OCR 算法,从字符、词、文本行不同层面准确获取文本信息. 文献[6]提出了一种 OCR 后处理算法,以提高 OCR 系统对扰动嵌入文本图像的鲁棒性.

文献分类是自然语言处理和文本挖掘中的基本任务. 传统方法使用稀疏词汇特征,例如词频-逆文档频率(term frequency-inverse document frequency,TF-IDF)和 N 元语法(N-gram),并使用线性模型(例如逻辑回归)或基于核的方法(如支持向量机)完成自动分类[7]. 但此类方法扩展性差,不适用于大规模文献的自动管理. 基于深度学习的方法已被证明具有可扩展的分类能力,它可应用各种自然语言处理技术来开发自动分类器,例如卷积神经网络(convolutional neural network,CNN)、循环神经网络(recurrent neural network,RNN)、Transformers 网络和特定的预训练模型等[8]. 文献[9]提出的 FastText 模型将文本视为词袋,然后通过一个或多个多层感知机(multilayer perceptron,MLP)进行分类. 文献[10]利用图卷积网络捕捉远距离语义信息,实现短文本的准确分类. 但这些模型通常都是在不包含专业技术术语的一般文献语料库上进行训练. 针对科技文献分类,文献[11]提出集成了词向量生成器(Word2Vec)和长短时记忆网络(long short-term memory,LSTM)的方法,将数据文献分为 50 个类别. 文献[12]将专利文献表示为固定层次结构向量,并使用基于 LSTM 的架构进行分类. 文献[13]提出了基于 CNN 和词向量的 DeepPatent 算法.

综上所述,虽然深度学习模型取得了优于传统机器学习方法的性能,但如仅使用文本作为模型输入,就忽略了科技文献图表中包含的有价值的信息. 多模态深度学习旨在设计和训练能够利用、处理和关联多种模态信息的模型. 相对于单模态方法,多模态方法在包括检索、匹配和分类等任务上展现出了优势. 文献[14]提出了用于文献分类的融合多模态分类器,通过 MobileNet-v2 模型学习视觉特征,并采用 FastText 处理文本数据. 文献[15]提出了融合知识图谱和多模态信息的文献分类模型,分别使用 Bert 和 ResNet 模型提取文本和图像特征,并通过不同模态的特征加权实现多模态信息互补. 为提高中文科技文献分类的智能化水平和分类精准度,本文提出了结合 CNN 和 YOLOv7 的多模态科技文献分类框架. 此框架首先通过文本特征获取网络提取词、句和文献级别的语义信息,接着借助带协同注意力机制的 YOLOv7 模型提取相应科技文献图表特征,最后利用基于 RNN 的融合网络对文本和图像特征进行整合,以实现准确的中文科技文献分类.

1　图文多模态深度学习网络

所提多模态科技文献分类框架的整个工作流程分为 3 个步骤:(1)数据预处理;(2)图像和文本提升提取;(3)特征融合和文献分类. 在第一步中,使用诸如分词、短语划分、去噪、词形还原和停用词删除等多种文本预处理方法,将文档转换为适合分类模型的表征;采用基于预训练的 CNN 模型,将复合图像分离为单个图像. 在第二步中,通过神经网络分别学习文本和图像特征向量. 在第三步中,将文本和图像特征融合在一起作为输入文档向量,用于网络融合学习和最终的文献分类.

1.1　文本和图像预处理

保留每篇文献的标题和摘要作为文本信息,并对复合图像进行分离后作为视觉信息. 所提方法着重关注标题和摘要文本,因为其与分类任务的相关度更高,可确保分类模型的计算效率和准确性. 相比之

下,技术描述包括更广泛的背景内容,可能会误导深度学习模型的训练.

由于本研究中的科技文献是用中文撰写的,因此需要对通过 OCR 算法得到的文本进行预处理. 文本处理主要包括两个阶段:文本分词和去除停用词. 文本分词非常重要,因为中文字符具有复杂的结构和严格的书写顺序(笔顺). 大多数文本分类算法不考虑笔顺,在中文文本处理中丢失了很多 n 元语法(n-gram)信息. 由于句子中的中文词语之间没有空格,因此需要使用复杂的分词算法. 对于中文文本,需要将其分割成中文字符序列. 使用 jieba 库进行中文分词,jieba 库支持 3 种中文分词模式. 使用 jieba 的标准模式,可得到词语,并通过停用词过滤器对其进行过滤. 停用词可以减少文本噪声,提高文本分类的准确性,模态助词、副词、介词、连词和标点符号不被视为停用词. 为此建立了单词和索引之间的词汇表以及二者的映射关系,当构建文档与索引之间的映射关系时,使用词汇处理器 TensorFlow Vocabulary Processor[16]将文档映射到词语索引序列,获得分类词汇表对象,并将该对象保存为文件. 其后,使用分词文本数据方法生成相应的词语索引值,并将索引值嵌入深度学习模型.

对于科技文献中的图像,须将复合图像分解为可以作为分类模型视觉输入的单个图像,图 1 给出了复合图像分割示例. 使用文献[17]中预训练的 CNN 构建图像分割器,并在实验数据集上对网络进行微调.

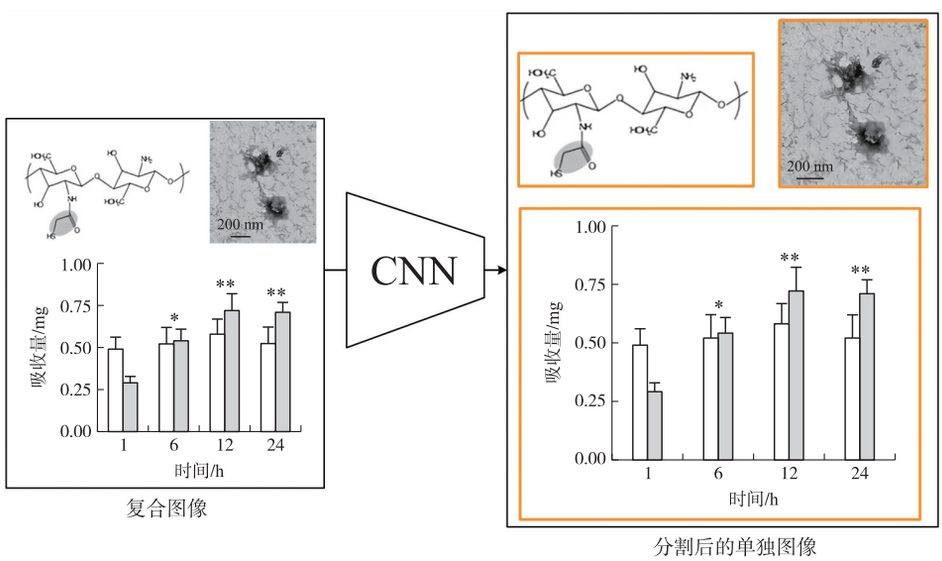


图 1　复合图像分割示例

Fig. 1　Examples of composite image segmentation

1.2　文本特征提取

本文旨在分析科技文献,并基于中图分类号预测文献类别,因此这是一个多标签多分类任务. 针对科技文献内容,提出文本特征提取网络,利用可训练的模型参数将基于词嵌入的句子表征与使用通用模型获得的表征组合起来,网络架构如图 2 所示.

每个文档最多包含 $|S|$ 个句子,每个句子最多包含 $|W|$ 个词语. 每个词(或句子)可以使用多种词嵌入或句嵌入表征. 令 f 和 g 分别为所提选择的词嵌入或句嵌入方法数. d_i^w 和 d_j^s 分别为第 i 和第 j 个词或句嵌入方案的嵌入维度.

每个文献使用两种类型的张量表征:(1)使用不同词嵌入创建的 f 个张量,$f \in \mathbb{R}^{|S|\times|W|\times d_i^w}$;(2)使用不同句编码器构建的 g 个张量,$g \in \mathbb{R}^{|S|\times d_i^w}$. 首先,基于词嵌入,对 f 个张量的子集进行拼接,见图 2 中“词级串联”,生成 p 个张量,$p \in \mathbb{R}^{|S|\times|W|\times D_i^w}$,其中 D_i^w 是第 i 个拼接所得到的维度. 其次,利用 CNN 和 bi-LSTM 网络,为句中词语嵌入在 p 个张量中的句子构建向量表示. 在 CNN 中,沿单词维度应用卷积滤波器,生成基于 bigram、trigram 和 4-gram 的特征. 随后进行最大时间(max over time,MOT)池化[18],为每个滤波器选择最大值,生成句子的表征张量 $p \in \mathbb{R}^{|S|\times c}$,其中 c 是使用的卷积滤波器的总数. bi-LSTM 利用注意机制,将 LSTM 在时间步上的输出聚合成每个句子的向量表示,从而得到张量 $p \in \mathbb{R}^{|S|\times h}$,其中 h 是 bi-LSTM 输出的长度. 通过对所有词嵌入张量应用 CNN 和 bi-LSTM,得到 3 种类型的句级表征张量:(1)使用 CNN 得到的 p 个张量 $p \in \mathbb{R}^{|S|\times c}$;(2)使用 bi-LSTM 得到的 p 个张量 $p \in \mathbb{R}^{|S|\times h}$;(3)使用通用句编码器获得的张量 $g \in \mathbb{R}^{|S|\times d_i^s}$.

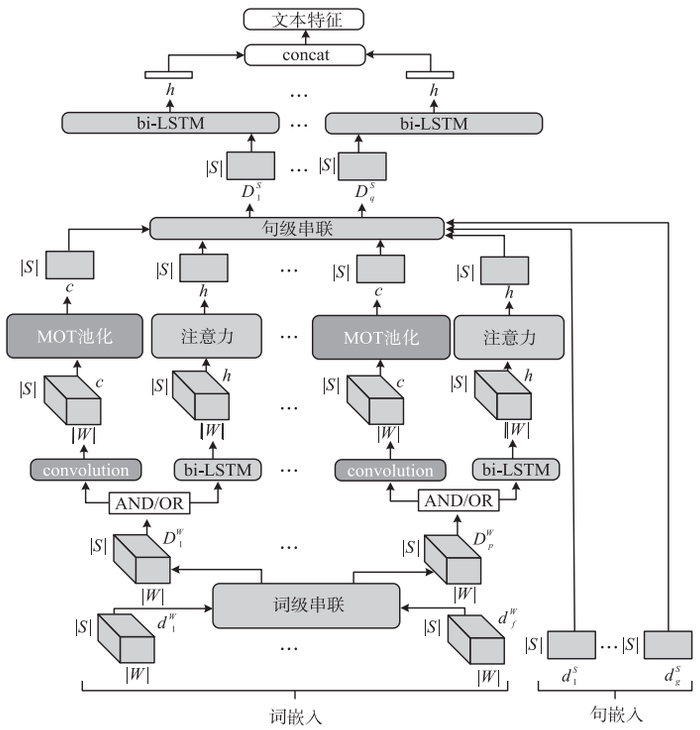


图 2　文本特征提取网络

Fig. 2　Text feature extraction network

基于句级表征张量,对子集进行拼接,得到 q 个张量,$q \in \mathbb{R}^{|S| \times d_j^s}$(图 2 中的句级串联),其中 d_j^s 是第 j 个拼接所得到的维度. 然后,对 q 个张量中的句向量序列中的句子通过 bi-LSTM 和注意力机制进行聚合,共同产生一篇文献的 q 个表征,再将这些向量串联在一起,得到科技文献的整体文本特征.

1.3　图像特征提取

为提高在大规模文献图像特征提取的准确度和效率,使用分类性能优秀的轻量级模型 YOLOv7 进行图像特征提取. YOLOv7[19] 由 YOLOv4 和 YOLOR 的作者提出,与 YOLOv4 一样,仅使用 MS COCO 数据集进行训练,未使用预训练的骨干网络. YOLOv7 通过架构改进提高了分类准确性,而不影响推理速度. 图 3 给出了 YOLOv7 架构.

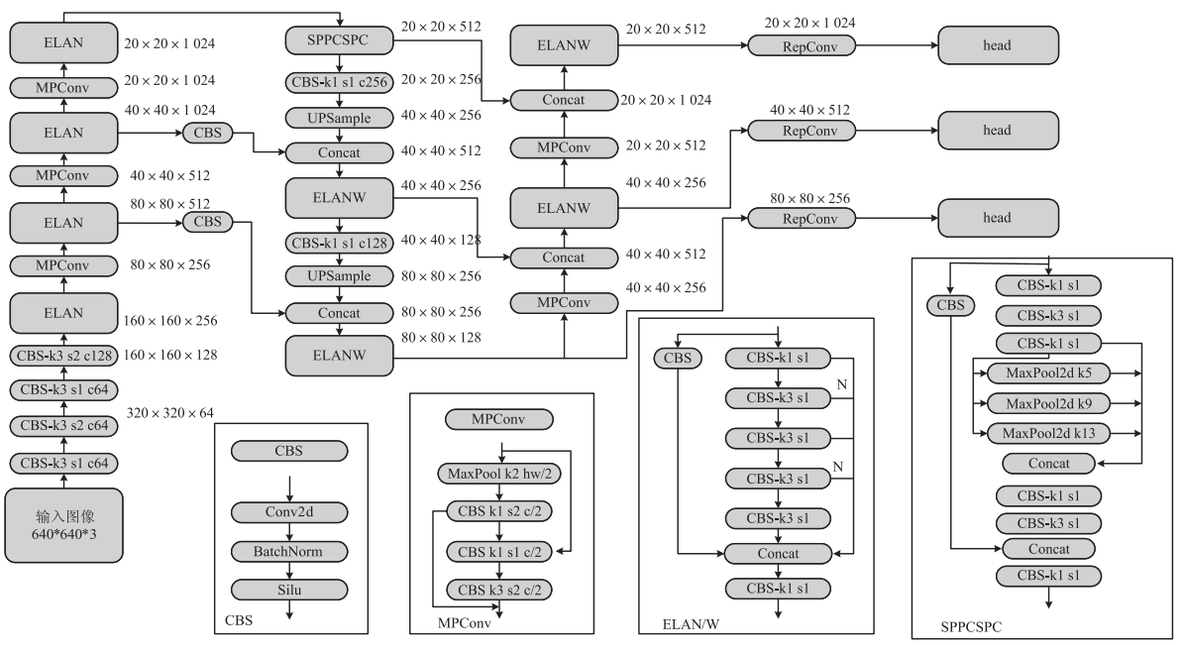


图 3　YOLOv7 架构

Fig. 3　The YOLOv7 architecture

YOLOv7 首先使用扩展高效层聚合网络(extend efficient layer aggregation network,E-ELAN),通过控制最短最长梯度路径,使深度模型更有效地学习和收敛. E-ELAN 支持计算区块的无限制堆叠,通过置乱和合并基数来结合不同组的特征,以增强网络的学习能力,而不破坏原始的梯度路径. 接着,使用基于拼接的模型缩放策略,通过去掉恒等连接并称其为 RepConvN,来改善 ResNet 和 DenseNet 中的残差和拼接. 再次,使用不同的标签分配策略来提高辅助头和主头的训练效果;通过批归一化的均值和方差集成到卷积层的偏置和权重中,提高模型的推理性能. 最后,YOLOv7 将指数移动平均作为最终推理模型,有效地减少了推理时的噪声,提高模型的鲁棒性和稳定性.

为了提高 YOLOv7 对科技文献图像的特征提取性能,所提方法在该网络中结合了协同注意力机制[20]. 令输入 $\boldsymbol{X}$ 的高度为 H,宽度为 W,通道数为 C,使用大小为 $(H,1)$ 和 $(1,W)$ 的两个池化核从空间方向提取特征. 具体来说,宽度 w 和高度 h 的通道 c 的输出分别计算为:

$$z_c^h=\frac{1}{W}\sum_{i=0}^{W-1}x_c(h,i), \tag{1}$$

$$z_c^w=\frac{1}{H}\sum_{j=0}^{H-1}x_c(j,w), \tag{2}$$

式中,$x_c(x,y)$ 为通道 c 在位置 (x,y) 的数值.

$$\boldsymbol{f}=\delta(\mathrm{F1}([\boldsymbol{z}^h,\boldsymbol{z}^w])), \tag{3}$$

式中,$[\boldsymbol{z}^h,\boldsymbol{z}^w]$ 为空间特征串联. F1 表示卷积变换,卷积核大小 1×1,C/r 个过滤器数量为 C/r,r 为约减率. δ 为非线性激活函数. 由此得到大小为 $C/r\times(W+H)$ 的特征.

其后,基于 W 和 H 将特征 $\boldsymbol{f}$ 分割为两部分,得到大小为 $C/r\times H$ 的 $\boldsymbol{f}^h$ 和大小为 $C/r\times W$ 的 $\boldsymbol{f}^w$. 然后通过两个不同的卷积层 F_w 和 F_h 进行转换,生成两个特征向量如下:

$$\boldsymbol{g}^h=\sigma(F_h(\boldsymbol{f}^h)), \tag{4}$$

$$\boldsymbol{g}^w=\sigma(F_w(\boldsymbol{f}^w)), \tag{5}$$

式中,σ 为 sigmoid 函数,则输出 $\boldsymbol{Y}$ 计算为:

$$\boldsymbol{y}_c(i,j)=x_c(i,j)\times\boldsymbol{g}_c^h(i)\times\boldsymbol{g}_c^w(j). \tag{6}$$

1.4　特征融合

在科技文献的图像和文本特征的多模态学习中,需要考虑的一个重要问题是各种模态在特定实例中通常不完全同步. 在这种情况下,图像特征的注意力可能与文本特征相关,但需要通过缩放和线性位移将注意力向量转换,即通过 ReLU 变换完成特征对齐. 例如对文本模态中的注意力向量进行位移或压缩,作为图像特征的先验知识. 图 4 给出了基于注意力 RNN 的特征融合网络架构.

所提方法中,多模态注意力融合 RNN 的更新过程表示为[21]:

$$\begin{cases}\boldsymbol{c}_T=\sum(\boldsymbol{\alpha}_T+F_I(\boldsymbol{a}_I))\cdot\boldsymbol{h}_T\\ \boldsymbol{c}_I=\sum(\boldsymbol{\alpha}_I+F_T(\boldsymbol{a}_T))\cdot\boldsymbol{h}_I\end{cases}, \tag{7}$$

式中,F_I 和 F_T 为线性变换函数,$\boldsymbol{\alpha}_I$ 和 $\boldsymbol{\alpha}_T$ 分别为图像和文本的注意力向量,$\boldsymbol{c}_I$ 和 $\boldsymbol{c}_T$ 分别为图像和文本的上下文向量.

特征融合后,通过逐步串联(concat)操作对融合注意力向量进行联合训练,利用模型末端的 DNN 模块中的多个全连接层构成分层学习架构,并利用 softmax 激活函数完成科技文献分类.

使用扩展二元交叉熵(extended binary cross entropy,EBCE)计算损失[22],通过逐标签二元交叉

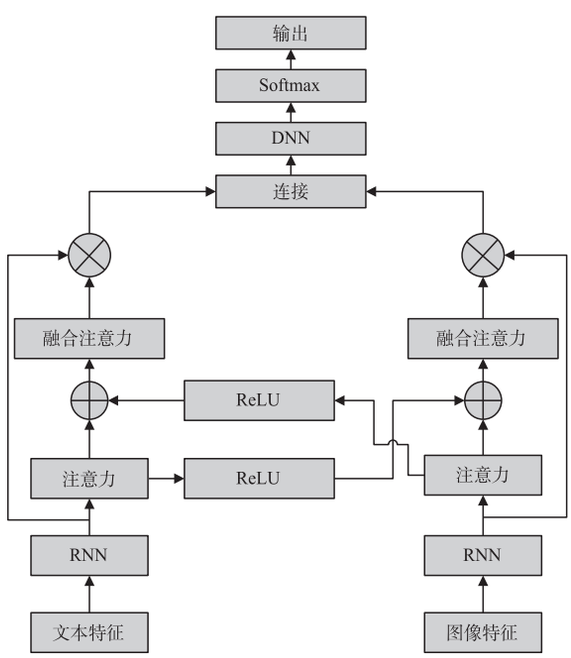


图 4　融合网络架构

Fig. 4　Fusion network architecture

熵的加权均值计算得出最终损失,从而缓解训练数据类别不平衡问题. 损失函数计算如下:

$$L=-\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}w_{jv}\{y_{ij}\log(p_{ij}^{\sigma})\}+(1-y_{ij})\log(1-p_{ij}^{\sigma})\},\tag{8}$$

式中,n 和 m 分别表示文献数量和类别数量. 若文献 x_i 类别为 l_j,则 y_{ij} 为 1;否则,y_{ij} 为 0. p_{ij}^{σ} 为文献 x_i 属于标签 l_j 的估计概率. W_{jv} 为类不平衡权重,计算如下:

$$w_{jv}=\frac{m}{2|\{x_i|y_{ij}=v,1\leqslant i\leqslant m\}|}.\tag{9}$$

2　实验与分析

2.1　实验配置和数据集

实验在一台配置了 CPU 12600KF,GPU nvdia 4060Ti 16G,RAM 32 GB 的台式机上进行,操作系统为 windows 11,采用 Python 编程,深度学习库采用 Pytorch 2.1.0. 在图像方面,所有图像输入的尺寸设置为 640×640,以满足 YOLOv7 模型的输入要求,训练的 epochs 设置为 300,配置文件中 anchor 计算中的 bpr(best possible recall)设置为 0.98. 优化器采用 Adam,学习率为 1e-4,batch size 设置为 64. 在文本特征提取方面,bi-LSTM 层数为 3 层,学习率、batch size 和优化器与图像操作相同,epochs 设置为 30. 在特征融合网络方面,RNN 的层数设置为 3,融合注意力头的数量设置为 8,dropout 设置为 0.5. 其他参数均保持与原模型一致.

科技文献中的专利文件通常包含文本、图像、引用和关于技术、产品系统的大量信息. 此外,文献中的内容必须具有创新性和功能性. 为了分析所提方法在科技文献分类中的性能,从北京大学图书馆与北京各高校图书馆期刊工作研究会联合编辑出版的中文核心期刊要目总览收录的核心期刊中,采集并构建包含 8 个类目的约 5 200 篇高质量科技文献作为实验数据集. 然后,将数据集分为三部分:训练集、验证集和测试集,其中 70%用于训练集,15%用于验证集,15%用于测试集. 验证集主要用于在训练过程中评估模型的性能,并帮助进行超参数调优. 在划分数据集时,使用分层抽样,以保证每个类别在训练集、验证集和测试集中的比例与原始数据集相似. 中文核心期刊中收录的文献都标注了相应的中图分类号. 表 1 列举了实验数据集统计信息,其中,使用科技文献的摘要和关键词作为文本信息,使用图表作为视觉信息进行分类,每篇文献最多对应两个类别.

表 1　数据集统计信息

Table 1　Dataset statistics

类目编号	描述	文献数量	类目编号	描述	文献数量
TD	矿业工程	870	TP	自动化技术、计算技术	1 253
TH	机械、仪表工业	925	TQ	化学工业	370
TK	动力工程	1 025	TU	建筑科学	450
TM	电工技术	630	TV	水利工程	580

2.2　评估指标

由于所提方法针对的是多标签多类分类的问题,即同一篇文献可能被分入一个以上的类别,因此单标签多类分类中使用的度量并不适用. 参照以往研究[23],使用基于实例的度量标准,包括 F1 值(F_1)和准确度(ACC_1),以及基于标签的度量标准,包括宏观 F1(F_{macro})和微观 F1(F_{micro})值.

2.3　实验结果

为了贴近现实中科技文献分类的实际操作场景,本文使用 OCR 生成文本,但 OCR 的输出噪声较大,造成训练数据的清晰度较差. 即使是方向良好的文献,OCR 在处理符号或多语言文本时可能会遇到问题. 此外,特定的用户领域可能使用罕见的词语、代码、缩写或术语,这些在通用词嵌入中不存在. 这样会在文本表征中注入许多非判别特征. 如图 5 所示,在实验数据集的科技文献中,有超过 50 篇文献中包含非词典词语,比例达 1%以上. 为此,本文在文本预处理中使用 TensorFlow Vocabulary Processor 将文档映射到词语索引序

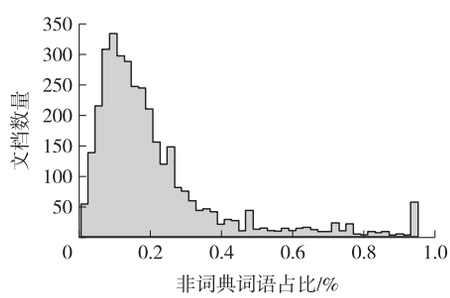


图 5　文章中非词典词语统计

Fig. 5　Statistics of non-dictionary words in the article

列,获得分类词汇表对象,并利用分词方法生成相应的词语索引值,再将索引值嵌入深度学习模型.

表 2 给出了不同方法在实验数据集上的分类结果. 文献[9]、[11]、[13]均为单模态方法,仅考虑文献中的摘要和关键词,而没有纳入图表信息. 其中,文献[9]采用 FastText 算法,能够快速训练和分类大规模文本数据,且能够通过将单词分解为子词,使用子词嵌入来处理未知词语的问题. 但 FastText 的性能取决于语料库的质量和大小,科技文献比较强调创新性,因此其分类性能不佳. 文献[11]使用的 Word2Vec 能够学习词汇的上下文信息,并通过 LSTM 捕捉输入序列的上下文信息,一定程度上提高了分类效果. 文献[13]的方法使用了多个输出层进行多标签分类,可以同时预测多个标签. 此外,使用训练的词向量,可以更好地捕捉词汇的语义信息. 与上述单模态方法相比,文献[14]和本文方法显著提高了分类效果. 其中,文献[14]的方法使用 MobileNetV2 捕捉视觉特征,并利用 FastText 算法提取文本特征,在 CNN 中对两者进行融合. 与之相比,所提方法在文本特征提取过程中考虑到了词级、句级、文档级特征的提取和融合,在视觉特征提取中使用了性能更好的 YOLOv7 模型,并且在该模型中融入了协同注意力机制,在各指标上均取得了最优性能,由此可证明所提方法在科技文献分类中的有效性.

通过消融分析,验证所提多模态方法中不同模块的作用,表 3 给出了不同模块组合的 ACC_1 结果. 其中,TFVP 表示在文本预处理中使用 TensorFlow Vocabulary Processor 生成词语索引值,CIS 表示在图像预处理中使用复合图像分割技术;EBCE 表示在融合模块中使用扩展二元交叉熵损失,BCE 表示使用传统二元交叉熵损失. 从表 3 可发现,所提方法的各模块均能提高科技文献分类的准确度,验证了各模块的有效性.

表 2　不同方法的分类性能比较

Table 2　Comparison of the classification performance of different methods　%

方法	F_1	F_{macro}	ACC_1	F_{micro}
[9]	69.71	61.60	56.33	65.82
[11]	72.88	65.08	60.17	68.80
[13]	75.31	68.56	63.62	71.87
[14]	84.04	76.41	72.55	80.92
所提方法	92.31	84.63	79.37	90.25

表 3　消融分析的准确度结果

Table 3　Accuracy results of ablation analysis　%

模块						ACC_1
文本网络	图像网络	TFVP	CIS	BCE	EBCE	
√				√		60.71
√		√		√		61.18
√		√			√	63.44
√	√	√			√	76.68
√	√	√	√		√	79.37

为了更好地展示所提多模态文献分类方法的性能,图 6 给出了在实验数据集上进行科技文献分类的混淆矩阵结果. 其中:图 6(a)为仅使用图像网络进行分类,从中可发现,分类效果较差. 这是因为科技文献通常包含丰富的文本信息和可能的图像信息,仅依赖图像信息进行分类,可能会忽略文本中提供的关键细节和上下文,从而导致信息的缺失. 此外,并不是所有的科技文献都包含描述性的图像,同时,本文所采用的图像分类方法 YOLOv7 依然受限于图像质量和分辨率等因素,这些因素导致仅使用图像网络进行分类时的效果较差. 对于科技文献,仅使用图像网络分类不能实现准确的分类结果. 图 6(b)为仅使用文本网络进行分类. 图 6(c)给出了结合文本和图像特征的混淆矩阵. 实验结果验证了多模态网络性能的互补性,它能够融合来自不同模态的信息,弥补单一模态信息的不足,增强了方法的鲁棒性,从而也提高了分类的性能.

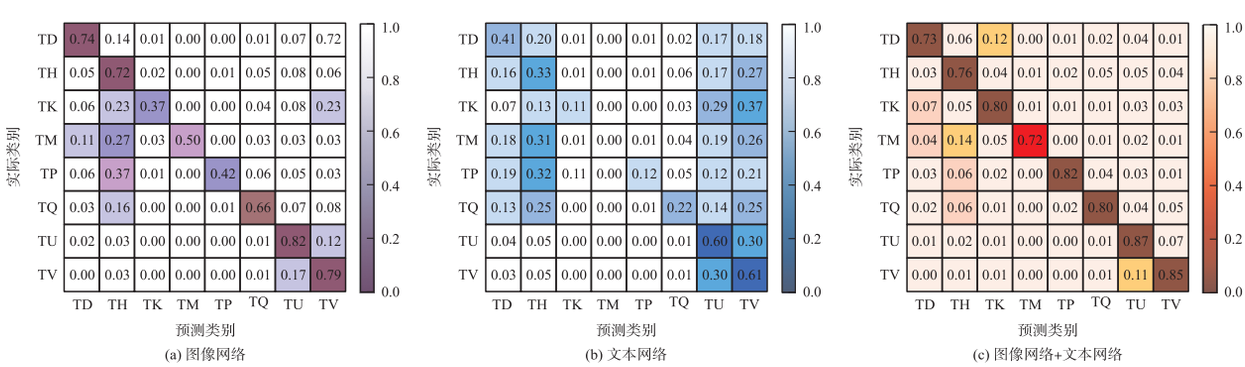


图 6　混淆矩阵结果

Fig. 6　Confusion matrix results

3　结论

本文提出了融合视觉特征和文本特征的多模态科技文献分类方法,联合学习文本和图像特征,并基于文献的融合异构表征进行最终分类. 实验结果表明,所提多模态文献分类方法能够较为准确地自动完成科技文献的分类,性能优于其他基于深度学习的单模态和多模态分类方法. 但所提方法仍有许多改进的空间:首先,本文实验中仅分析了二级类目的文献分类,如何将文献准确分类到成百上千个三级或四级类目,仍需要大量的数据采集和模型训练. 其次,多模态深度学习模型参数量庞大,当前的模型训练需要大量的计算资源. 未来计划通过探索更有效的方法来挖掘多模态信息,特别是视觉信息,从而对分类模型做进一步改进.

[参考文献]

[1] 时莹,王铮. 数字化转型中的高校图书馆数字化服务能力影响因素研究[J]. 图书情报工作,2022,66(23):41-50.

[2] RAHMAN M H,AHMAD A,ZAKARIA S. A literature review on digital content management:trends and future challenges [J]. Digital library perspectives,2023,39(1):97-110.

[3] 柴变芳,李政,赵晓鹏,等. 基于泛化图卷积神经网络的深度文档聚类模型[J]. 南京师大学报(自然科学版),2024,47(1):82-90.

[4] GOYAL A,PREM PRAKASH V. Statistical and deep learning approaches for literary genre classification[C]//Advances in Data and Information Sciences:Proceedings of ICDIS 2021. Singapore:Springer Singapore,2022:297-305.

[5] 熊帆,陈田,卞佰成,等. 基于卷积循环神经网络的芯片表面字符识别[J]. 浙江大学学报(工学版),2023,57(5):948-956.

[6] IMAM N H,VASSILAKIS V G,KOLOVOS D. OCR post-correction for detecting adversarial text images[J]. Journal of information security and applications,2022,66(1):1-15.

[7] 崔磊,徐毅恒,吕腾超,等. 文档智能:数据集,模型和应用[J]. 中文信息学报,2022,36(6):1-19.

[8] MINAEE S,KALCHBRENNER N,CAMBRIA E,et al. Deep learning-based text classification:a comprehensive review[J]. ACM computing surveys(CSUR),2021,54(3):1-40.

[9] UMER M,IMTIAZ Z,AHMAD M,et al. Impact of convolutional neural network and FastText embedding on text classification [J]. Multimedia tools and applications,2023,82(4):5569-5585.

[10] 张虎,柏萍. 融入句子中远距离词语依赖的图卷积短文本分类方法[J]. 计算机科学,2022,49(2):279-284.

[11] KAMATERI E,MICHAIL S,KONSTANTINOS D. An ensemble framework for patent classification[J]. World patent information,2023,75(12):102233-102243.

[12] SHALABY M,STUTZKI J,SCHUBERT M,et al. A LSTM approach to patent classification based on fixed hierarchy vectors [C]//Proceedings of the 2018 SIAM International Conference on Data Mining. San Diego:SIAM,2018:495-503.

[13] HAGHIGHIAN ROUDSARI A,AFSHAR J,LEE W,et al. PatentNet:multi-label classification of patent documents using deep learning-based language understanding[J]. Scientometrics,2022,127(1):207-231.

[14] AUDEBERT N,HEROLD C,SLIMANI K,et al. Multimodal deep networks for text and image-based document classification [C]//Machine Learning and Knowledge Discovery in Databases:International Workshops of ECML PKDD 2019. Würzburg:Springer,2020:427-443.

[15] 景丽,姚克. 融合知识图谱和多模态的文本分类研究[J]. 计算机工程与应用,2023,59(2):102-109.

[16] QIU M,ZHANG Y,MA T,et al. Convolutional-neural-network-based multilabel text classification for automatic discrimination of legal documents[J]. Sensors and materials,2020,32(8):2659-2672.

[17] TSUTSUI S,CRANDALL D J. A data driven approach for compound figure separation using convolutional neural networks [C]//2017 14th IAPR International Conference on Document Analysis and Recognition(ICDAR). Piscataway:IEEE,2018:533-540.

[18] BAI H Y,HUANG Z L,HAO A,et al. Gated character-aware convolutional neural network for effective automated essay scoring[C]//IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. New York:Association for Computing Machinery,2021:351-359.

[19] WANG C Y,BOCHKOVSKIY A,LIAO H Y M. YOLOv7:Trainable bag-of-freebies sets new state-of-the-art for real-time

object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE,2023:7464-7475.

[20] HOU Q,ZHOU D,FENG J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021. Piscataway:IEEE,2021:13713-13722.

[21] ZHANG J,ZONG C. Neural machine translation:challenges,progress and future[J]. Science China technological sciences,2020,63(10):2028-2050.

[22] YAO H S,ZHU D L,JIANG B,et al. Negative log likelihood ratio loss for deep neural network classification[C]//Proceedings of the Future Technologies Conference(FTC)2019. Berlin:Springer,2020:276-282.

[23] RAFI M,ABID F. Learning local and global features for optimized multi-label text classification[C]//2022 International ARAB Conference on Information Technology(ACIT). Piscataway:IEEE,2022:1-9.

[责任编辑:陆炳新]