◆自然语言理解◆

CCM-MF:基于多维度特征融合的中文文本分类模型*

马子晨[1,2],张顺香[1,2**],刘云朵[1,2],王星光[1,2],张友强[1,2]
(1.安徽理工大学计算机科学与工程学院,安徽淮南 232001;2.合肥综合性国家科学中心人工智能研究院,安徽合肥 230088)

摘要:针对中文文本中不同维度特征所携带的语义信息具有差异性的问题,本文提出一种基于多维度特征融合的中文文本分类模型:CCM-MF (Chinese-text Classification Model Based on Fused Multi-dimensional Features)。该模型融合层次维度和空间维度特征,以提高中文文本分类的准确率。首先,在层次维度上,使用预训练模型 ERNIE (Enhanced Representation through Knowledge Integration)获取包含字、词及实体级别特征的词向量;然后,在空间维度上,将包含层次维度特征的词向量分别输入到改进后的深度金字塔卷积神经网络(Deep Pyramid Convolutional Neural Networks,DPCNN)模型及附加注意力机制的双向长短期记忆网络(Attention-Based Bidirectional Long Short-Term Memory Networks,Att-BLSTM)模型中,得到局部语义特征和全局语义特征;最后,将得到的空间维度特征分别作用于 Softmax 分类器,再对计算结果进行融合并输出分类结果。通过在多个公开数据集上进行实验,较现有主流的文本分类方法,本模型在准确率上有更好的表现,证明了该模型的有效性。

关键词:中文文本分类;多维度;ERNIE;DPCNN;Att-BLSTM

中图分类号:TP391 **文献标识码**:A **文章编号**:1005-9164(2023)01-0035-08
DOI:10.13656/j.cnki.gxkx.20230308.004

文本分类是自然语言处理(Natural Language Processing,NLP)中的重要研究方向,其目的是为文本单元(如句子、段落和文章)分配标签,为更深入地进行文本信息挖掘以及后续处理提供技术支持。目前主流的分类方法大多只对单一的模型结构进行深入研究,缺乏同时捕获并利用文本局部语义特征和全局语义特征的能力,且未进一步考虑字、词以及实体级别特征在中文文本分类任务中的不同作用。为了设计一个效果更佳的中文文本分类模型,需要考虑以下方面:①如何充分利用包括字、词及实体级别的文本层次维度特征,使词向量语义信息更加丰富;②如何充分利用包括局部语义和全局语义的文本空间维度特征,使模型更加全面地学习文本中的语义信息;③如何降低特征融合时造成的语义损失,以减少不同

收稿日期:2022-09-16 修回日期:2022-09-23
*国家自然科学基金面上项目(62076006)和安徽省高校协同创新项目(GXXT-2021-008)资助。
【第一作者简介】
马子晨(1998-),男,在读硕士研究生,主要从事自然语言处理研究。
【**通信作者】
张顺香(1970-),男,博士,教授,博士研究生导师,主要从事 Web 挖掘、语义搜索和复杂网络研究,E-mail:sxzhang@aust.edu.cn。
【引用本文】
马子晨,张顺香,刘云朵,等.CCM-MF:基于多维度特征融合的中文文本分类模型[J].广西科学,2023,30(1):35-42.
MA Z C,ZHANG S X,LIU Y D,et al.CCM-MF:Chinese-text Classification Model Based on Fused Multi-dimensional Features [J].Guangxi Sciences,2023,30(1):35-42.

语义特征为模型带来的增益损失。

基于以上 3 点考虑,本文提出一种基于多维度特征融合的中文文本分类模型:CCM-MF(Chinese-text Classification Model Based on Fused Multi-dimensional Features)。该模型首先使用预训练模型 ERNIE(Enhanced Representation through Knowledge Integration)[1],对文本的层次维度特征进行提取,层次维度特征是包括字、词以及实体级别特征的基础语义信息;然后,将得到的基础语义信息分别输入到改进后的深度金字塔卷积神经网络(Deep Pyramid Convolutional Neural Networks,DPCNN)[2]模型和附加注意力机制的双向长短期记忆网络(Attention-Based Bidirectional Long Short-Term Memory Networks,Att-BLSTM)[3]模型中提取空间维度特征,空间维度特征包括局部语义特征和全局语义特征;最后,将得到的局部语义特征和全局语义特征分别作用于 Softmax 分类器,再使用算术平均的方式对结果进行融合得到最终分类结果。CCM-MF 通过 ERNIE 获取包含层次维度特征的词向量,使用 DPCNN 和 Att-BLSTM 来获取空间维度上的特征。在特征融合部分,使用结果融合机制,避免造成语义混乱。

1 相关工作

现有文本分类方法主要包括基于机器学习、神经网络和预训练模型的方法。近期对文本分类任务的研究主要集中在基于神经网络和预训练模型上。

1.1 基于机器学习的分类方法

传统的机器学习方法通过人工方法和浅层分类模型进行特征提取。在文本分类任务中,基于机器学习的算法有支持向量机(Support Vector Machine,SVM)[4]、朴素贝叶斯(Naive Bates,NB)[5]、K 近邻(K-Nearest Neighbor,KNN)[6]等,在当时均取得了一定的成果。但这些方法难以捕捉文本中潜在的深层特征,缺乏泛化和扩展能力。由于深度学习可以在很大程度上解决机器学习中出现的此类问题,因此,在文本分类领域,深度学习更加受到研究人员的青睐。

1.2 基于神经网络的分类方法

2014 年,Kim[7]首次将卷积神经网络(Convolutional Neural Networks,CNN)应用于文本分类领域,并取得较好的分类效果,紧接着便出现了各种 CNN 的变体。其中,腾讯提出的 DPCNN[2]模型既拥有对文本局部信息的提取能力又通过金字塔模型加深了网络深度,减少了长距离文本间的信息丢失,在分类任务中取得了较好的效果。2020 年,汪嘉伟等[8]将注意力(Attention)机制与 CNN 相结合,捕捉文本中的局部语义特征并增强关键词的作用,分类效果有了显著提升。

相较于 CNN,循环神经网络(Recurrent Neural Network,RNN)更适合在时序序列文本上进行学习。长短期记忆网络(Long Short-Term Memory,LSTM)[9]可以提取文本的全局特征,在长文本分类任务上取得了更好的效果。Chen 等[10]使用 LSTM 成功构建了门诊文本分类系统,为用户查询自身健康状态提供了参考。Deepak 等[11]使用双向长短期记忆网络(Bi-directional LSTM,Bi-LSTM)模型来应对 LSTM 无法捕获双向语义的缺陷,在犯罪分类任务上取得较好的效果。Wang 等[12]提出一种结合树结构的区域 CNN 模型,用于检测与任务相关的短语从句,在情感分类任务中取得了较好的效果。Zhou 等[3]第一次将 Bi-LSTM 与 Attention 机制融合,在捕获文本全局语义特征的同时,增强了关键词在分类任务中的作用,取得更好的分类效果。Chen 等[13]将 LSTM 与注意力机制结合并用于预测股价走势,取得比原始 LSTM 更好的效果。

1.3 基于预训练模型的分类方法

2018 年 Devlin 等[14]提出了基于 Transformer 的双向偏码表示模型(Bidirectional Encoder Representations from Transformers,BERT),使用具有强大特征提取能力的双向 Transformer 编码器,通过下一句预测(Next Sentence Prediction,NSP)和掩码语言模型(Masked Language Model,MLM)结合的预训练模型,在多项 NLP 任务中取得了 SOTA(State of the Arts)的成绩。Lan 等[15]使用矩阵分解和参数共享的方法,大大降低了参数量,提高 BERT 的训练速度的同时取得了更好的效果。

2 CCM-MF 模型

本文提出的基于多维度特征融合的中文文本分类模型,通过使用 DPCNN 和 Att-BLSTM 对 ERNIE 中提取的词向量进行学习后,分别得到局部语义特征和全局语义特征,并将其用于预测类别,最后采用算术平均的方式对预测结果进行融合,得到最终的分类结果。模型的框架如图 1 所示。

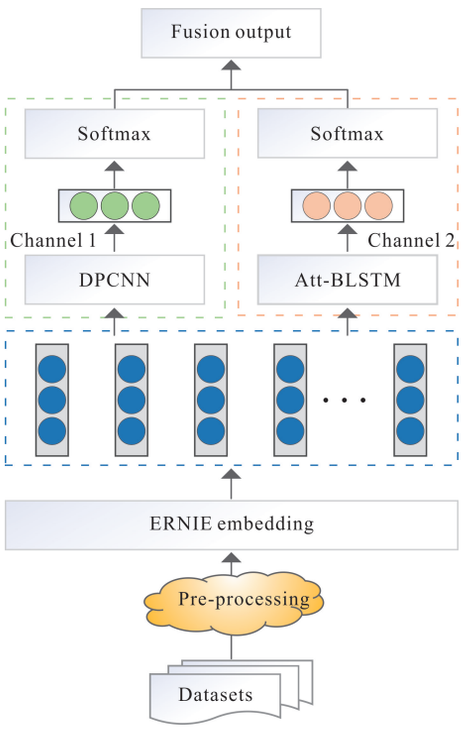

图 1　CCM-MF 框架图

Fig. 1　CCM-MF framework diagram

2.1　嵌入层

ERNIE 模型是对 BERT 模型的进一步优化，在 NLP 领域各个中文任务中取得了 SOTA 的成绩。该模型主要在 BERT 的掩码(mask)机制上做了改进，由 BERT 只针对字级别做 mask，改进为 3 种级别的 mask，分别为字、词以及命名实体级别。通过对 mask 的改进，使得模型可以学习到更高层次的语言表达。BERT 模型和 ERNIE 模型的 mask 机制如图 2 所示。

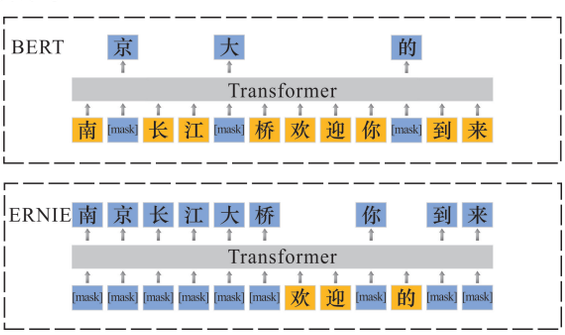

图 2　BERT 模型与 ERNIE 模型的 mask 机制对比

Fig. 2　Comparison of mask mechanism of BERT model and ERNIE model

从图 2 可以清楚地看到，BERT 模型只对文本“南京长江大桥欢迎你的到来”进行字级别的 mask，而 ERNIE 模型除了对字级别进行 mask 外，还对词“到来”以及实体“南京长江大桥”进行了 mask。

ERNIE 模型的 mask 策略分为 3 个阶段学习。第一阶段，采用与 BERT 相同的、基于字级别的 mask；第二阶段加入基于词级别的 mask，mask 掉句中的一部分词组，然后让模型预测这些词组，在这个阶段，词级别的信息被编码到词向量中；第三阶段加入实体级别的 mask，如人名、地名、领域名等，在模型训练完后，将学到的实体信息加入词向量中。本模型将通过 ERNIE 学到的具有层次维度特征的词向量分别作为通道 1 和通道 2 的输入，进行空间维度特征的提取。

2.2　DPCNN 层

对于 CNN 模型来说，解决对文本长距离依赖关系的提取问题，最有效的方法是增加 CNN 的深度，但这同时常会出现模型过大、参数太多而导致梯度爆炸和消失等问题，使得模型效率严重下滑。DPCNN 固定了卷积时的特征映射数量，每经过一个 size 为 3、stride 为 2 的最大池化层(max-pooling)后，文本序列长度会随着卷积块数量的增加呈指数级别减少。这会使得每个卷积层的计算时间减半，形成一个类金字塔形状，从而在加深 CNN 的同时计算效率并未下降。DPCNN 模型结构如图 3 所示。

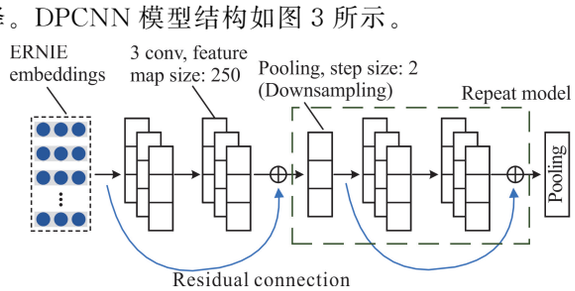

图 3　DPCNN 模型结构图

Fig. 3　DPCNN model structure diagram

在原 DPCNN 论文[2]中，模型使用的是 ReLU 激活函数，如公式(1)所示。ReLU 激活函数是较常用的激活函数之一，具有形式简单、高度非线性等特点，但其将所有负数输入都赋值为 0，在训练过程中极为脆弱，容易导致神经元失活。针对此问题，本文将 DPCNN 中的激活函数从 ReLU 改为 SeLU，如公式(2)所示，并在等长卷积部分的两个卷积后，以及叠加模块中最后一个卷积后各加入一个 SeLU 函数。SeLU 激活函数具有自归一化的特点，即使在输入部分加入噪声，也能使其收敛到均值为 0、方差为 1 或者方差具有上下界的形式，可以有效避免因神经元失活而导致的梯度消失和爆炸问题。

$$\mathrm{ReLU}(x)=\begin{cases}x, x\geqslant 0\\ 0, \text{otherwise}\end{cases}, \tag{1}$$

$$\mathrm{SeLU}(x)=\lambda_{\mathrm{SeLU}}\begin{cases}x, x\geqslant 0\\ \alpha_{\mathrm{SeLU}}(\exp(x)-1), \text{otherwise}\end{cases},\quad (2)$$

其中,x 来自上一层神经网络的输入,λ 和 α 是预先设定的权重,分别为 1.050 7 和 1.673 3。

2.3 Att-BLSTM 层

Att-BLSTM 模型结合了 BLSTM 和 Attention 机制,在 NLP 任务中得到了广泛的应用[16-19]。由于单向 LSTM 只能按照序列顺序从前往后进行学习预测,使得当前时刻的预测只包含之前的信息。而 BLSTM 模型可以同时捕获句子中双向的语义,对文本语义的学习更加丰富。同时,Attention 机制使模型可以自动聚焦于对分类结果有关键影响的部分,用以捕获句子中最重要的语义信息,进一步加强全局语义特征和关键词在中文文本分类任务中产生的作用。Att-BLSTM 模型结构如图 4 所示。

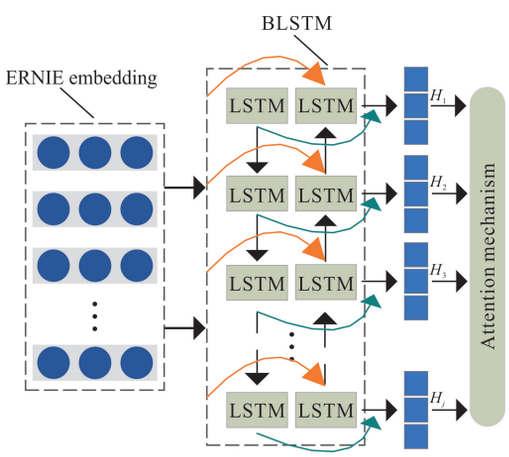

图 4 Att-BLSTM 模型

Fig. 4 Att-BLSTM model

本文将 BLSTM 输出的全局语义特征作为 Attention 机制的输入,赋予特征不同的权重参数,使得模型更加关注对分类结果重要的特征参数。

2.4 融合输出层

通过 DPCNN 层和 Att-BLSTM 层分别获得文本的局部语义特征表示 F_{local} 和全局语义特征表示 F_{global},将 F_{local} 和 F_{global} 分别作为 Softmax 分类器的输入,通过计算分别得到预测概率 p_1 和 p_{g},然后使用算术平均的方式对其融合,得到最终的分类概率 p。

$$p_1=\mathrm{Softmax}(W_1F_{\mathrm{local}}+b_1),\quad (3)$$

$$p_{\mathrm{g}}=\mathrm{Softmax}(W_{\mathrm{g}}F_{\mathrm{global}}+b_{\mathrm{g}}),\quad (4)$$

$$p=\frac{1}{2}(p_1+p_{\mathrm{g}}),\quad (5)$$

其中,W_1 和 W_{g} 为可训练权重,W_1 对应的 size 为(250,num_outputs),W_{g} 对应的 size 为(512,num_outputs),250 和 512 分别为 DPCNN 和 Att-BLSTM 学习到的特征维度,num_outputs 为最终的类别数;b_1 和 b_{g} 为偏置项。

2.5 算法流程介绍

基于多维度特征融合的中文文本分类算法流程描述如下。

算法 1 CCM-MF 中文文本分类算法

输入:中文文本分类领域数据集 S

输出:文本所属类别

①$S=\mathrm{preprocessing}(S)$

②$\mathrm{ERNIE_Embedding}=\mathrm{ERNIE}(S)$

③$F_{\mathrm{local}}=\mathrm{DPCNN}(\mathrm{ERNIE_Embedding})$

④$F_{\mathrm{global}}=\mathrm{Att\text{-}BLSTM}(\mathrm{ERNIE_Embedding})$

⑤$p_1=\mathrm{Softmax}(W_1F_{\mathrm{local}}+b_1)$

⑥$p_{\mathrm{g}}=\mathrm{Softmax}(W_{\mathrm{g}}F_{\mathrm{global}}+b_{\mathrm{g}})$

⑦$p=(p_1+p_{\mathrm{g}})/2$

⑧$\mathrm{Category}=\mathrm{Max}(p)$

⑨End

算法说明:步骤①是对中文文本分类领域数据集进行预处理,包括数据清洗、格式转换等;步骤②采用 ERNIE 模型对文本进行预训练,提取包含字、词以及实体级别基本语义特征的词向量;步骤③和④中将 ERNIE 输出的词向量作为 DPCNN 和 Att-BLSTM 的输入,分别提取局部语义特征和全局语义特征,得到 F_{local} 和 F_{global};步骤⑤和⑥中分别通过局部语义特征向量和全局语义特征向量计算预测的结果;步骤⑦使用算术平均的方法对两个预测概率进行融合;步骤⑧选择概率最大的类别作为最终分类结果。

3 实验与结果分析

3.1 实验环境及数据

实验配置的硬件和软件环境如表 1 所示,主要使用 GPU 为 RTX 3060-12 GB,并使用 Pytorch 深度学习框架进行实验。

在数据处理工作中,首先对原始数据集进行清洗,去除与本文工作无关部分,并对其进行格式转换后得到所需原始实验数据集,如表 2 所示。为了评估该模型在中文文本分类任务上的泛化能力,使用以下 3 个数据集进行实验。①今日头条:属于新闻主题分

表 1　实验环境配置

Table 1　Experimental environment configuration

名称 Name	配置详情 Configuration details
CPU	12th Gen Intel Core i5-12400F six-core
GPU	NVIDIA GeForce RTX 3060-12 GB
Programming language	Python 3.7
Deep learning framework	Pytorch 1.11.0

表 2　数据集详情

Table 2　Details of datasets

数据集 Datasets	类别 Categories	总样本(条) Total sample (item)	训练集(条) Training set (item)	验证集(条) Validation set (item)	测试集(条) Test set (item)
Jinri Toutiao	11	200 000	180 000	10 000	10 000
THUCNews	10	200 000	180 000	10 000	10 000
Online_shopping	2	60 000	50 000	5 000	5 000

3.2　实验设置

对数据集进行清洗并去除和本文工作无关的部分后,利用 Tokenizer 对新闻主题描述文本进行分词处理,然后利用预训练模型 ERNIE 训练词向量。其中,padsize 设置为 64,对于长度不足部分进行 padding 补全,长度超出部分进行剪切;学习率设置为 1e-5,使用批量正则化方式降低过拟合,优化器使用 Adam;在通道 1 中,将 feature map 设置为 250,池化层 size 和 stride 分别设置为 3 和 2;在通道 2 中,将 Dropout 设置为 0.5,隐藏层数量设置为 256;在所有实验中,epoch 均设置为 20,且若超过 1 000 个 batch 后模型学习效果未提升,则提前结束训练。

3.3　评价方法

文本分类任务中常采用准确率(Accuracy,Acc)精确率(Precision,P)、召回率(Recall,R)以及 $F1$ 值这 4 个指标来评估模型。公式如(6)-(9)所示,其中 TP 表示真正例,即预测为正类,实际结果也是正类;TN 表示真反例,即预测为反类,实际结果也是反类;FP 表示假正例,即预测为正类,实际为反类;FN 表示假反例,即预测为反类,实际为正类。

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}, \tag{6}$$

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{7}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{8}$$

$$\mathrm{F1}=\frac{2\times P\times R}{P+R}。 \tag{9}$$

类数据集,包含 11 个新闻主题。②THUCNews:属于新闻主题分类数据集,包含 10 个新闻主题。③Online_shopping:属于情感分析任务数据集,收集了 10 种物品的网购评论,其中,正、负向评论各约 30 000 条。

在本文选用的数据集中,各类别分布均匀。因此,只选用准确率作为模型分类效果的评估标准。

3.4　对比实验

为验证 CCM-MF 在中文文本分类任务上的有效性,分别在 3 个数据集上对不同分类模型进行对比实验,结果如表 3 所示。将 CCM-MF 与以下 5 种模型进行对比评估:

①DPCNN[2]:在 CCM-MF 中,用于提取文本局部语义特征;②Att-BLSTM[3]:在 CCM-MF 中,用于提取文本全局语义特征;③BERT[14]:2018 年由 Google AI 研究院提出的预训练语言模型,基于 Transformer 的双向编码器表示;④B-DLM:将本文所提出模型中 ERNIE 部分替换为 BERT 预训练模型;⑤ERNIE[1]:百度提出的预训练语言模型,作为 CCM-MF 中提取文本层次特征的部分。

表 3　实验结果对比

Table 3　Comparison table of experimental results

模型 Model	准确率(%)Accuracy (%)		
	Online_shopping	今日头条 Jinri Toutiao	THUCNews
DPCNN	85.27	83.74	87.21
Att-BLSTM	85.13	83.71	87.11
BERT	88.31	85.31	90.51
B-DLM	90.16	86.73	90.80
ERNIE	93.28	91.33	94.10
CCM-MF	93.84	91.98	94.64

从表 3 中的实验结果可以看到,本文提出的

CCM-MF 在 Online_shopping、今日头条和 THUCNews 3 个数据集上分别获得了 93.84%、91.98%和 94.64%的准确率,准确率较前 5 个基准中效果最好的 ERNIE 模型分别提升了 0.56%、0.65%、0.54%,表明 CCM-MF 通过融合层次维度以及空间维度的语义特征可以有效提升模型性能。

3.5 消融实验

为了证明 CCM-MF 中各个组件对模型的有效增益,本文进行了相应的消融实验。结果如表 4 所示。其中:

①ERNIE-DPCNN:原模型去掉 Att-BLSTM 及特征融合部分后,直接输出最终分类结果;

②ERNIE-Att-BLSTM:原模型去掉 DPCNN 以及特征融合部分后,直接输出分类结果;

③CCM-MF (ReLU):原模型去掉对通道 1 中 DPCNN 的改进,使用原始 DPCNN 模型。

表 4 消融实验结果对比

Table 4 Comparison of ablation experimental results

模型 Model	准确率(%)Accuracy (%)		
	Online_shopping	今日头条 Jinri Toutiao	THUCNews
ERNIE-DPCNN	93.56	91.41	94.16
ERNIE-Att-BLSTM	93.44	91.47	94.28
CCM-MF (ReLU)	93.69	91.76	94.40
CCM-MF	93.84	91.98	94.64

从表 4 可以看到,实验数据集在 ERNIE-DPCNN 和 ERNIE-Att-BLSTM 上的准确率均高于表 3 中的 ERNIE,说明在 ERNIE 预训练后进一步提取局部或全局语义特征均可以有效地增强模型对文本语义的理解,进而提升模型的效果。此外,由表 4 可知,CCM-MF (ReLU)在各个数据集上的准确率均高于除 CCM-MF 以外的任意模型,说明融合局部和全局语义特征的有效性,这是因为 DPCNN 模型和 Att-BLSTM 模型分别获取文本的局部和全局语义特征,起到互补的作用,但如果使用更多的分类头,会增加模型的冗余,甚至会出现模型准确率降低的情况;同时可以看到 CCM-MF 在 3 个数据集上的准确率均为最优,证明将 ReLU 激活函数改为 SeLU 激活函数的有效性。

3.6 特征融合对比

CCM-MF 通过双通道部分得到局部语义特征向量和全局语义特征向量,为了证明结果融合机制的有效性,本文将 CCM-MF 与使用相加、拼接两种方式融合后的实验结果进行对比,结果如表 5 所示。

表 5 特征融合实验结果对比

Table 5 Comparison of feature fusion experimental results

融合方式 Fusion approach	准确率(%)Accuracy (%)		
	Online_shopping	今日头条 Jinri Toutiao	THUCNews
Add	93.48	91.45	94.19
Concat	93.62	91.71	94.37
Results fusion	93.84	91.98	94.48

从表 5 可以看到,采用结果融合机制与采用相加、拼接的融合方式相比,准确率均有提升。其中,直接相加进行融合的方式准确率最低。这是因为采用直接相加或者拼接的融合方式,可能会引入更多的噪声,从而导致语义混乱,进一步影响模型的性能。

4 结论

本文提出了一种基于多维度特征融合的中文文本分类模型(CCM-MF)。CCM-MF 在层次维度上,通过 ERNIE 获取包含字、词及实体级别特征的词向量;在空间维度上,通过改进的 DPCNN 模型和 Att-BLSTM 模型分别提取包含层次维度特征的局部语义特征和全局语义特征;最后,将提取到的空间维度特征分别作用于 Softmax 分类器,并通过融合机制进行融合,提高分类准确率。分别在 3 个数据集上进行实验验证,评估本文提出方法的性能。实验结果表明:CCM-MF 在各个数据集上的分类准确率均为最高,证明了模型的有效性。

未来工作的重点是在模型中融合更多的中文特征和消减模型的参数量,主要包括以下两个部分:①利用更多的中文特征进行融合,进一步提升模型在中文文本分类任务上的效果;②结合知识蒸馏的方法,消减模型的参数量,提高模型的效率。

参考文献

[1] SUN Y, WANG S H, LI Y K, et al. ERNIE: enhanced representation through knowledge integration [Z/OL]. (2019-04-19)[2022-03-18]. https://doi.org/10.48550/arXiv.1904.09223.

[2] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Vancouver,Canada:Association for Computational Linguistics,2017:562-570.

[3] ZHOU P,SHI W,TIAN J,et al. Attention-based bidirectional long short-term memory networks for relation classification [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2:Short Papers). Berlin,Germany:Association for Computational Linguistics,2016:207-212.

[4] JOACHIMS T. Text categorization with support vector machines:learning with many relevant features [C]// NÉDELLEC C, ROUVEIROL C. Machine Learning: ECML-98. Berlin,Heidelberg,Germany:Springer,1998: 137-142.

[5] MARON M E. Automatic indexing:an experimental inquiry [J]. Journal of the ACM (JACM),1961,8(3): 404-417.

[6] COVER T M,HART P E. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory,1967,13(1):21-27.

[7] KIM Y. Convolutional Neural Networks for Sentence Classification [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha,Qatar:Association for Computational Linguistics,2014:1746-1751.

[8] 汪嘉伟,杨煦晨,琚生根,等.基于卷积神经网络和自注意力机制的文本分类模型[J].四川大学学报(自然科学版),2020,57(3):469-475.

[9] HOCHREITER S,SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[10] CHEN C W,TSENG S P,KUAN T W. Outpatient text classification system using LSTM [J]. Journal of Information Science and Engineering,2021,37(2):365-379.

[11] DEEPAK G,ROOBAN S,SANTHANAVIJAYAN A. A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network [J]. Multimedia Tools and Applications, 2021, 80(18):28061-28085.

[12] WANG J,YU L C,LAI K R,et al. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis [J]. IEEE/ACM Transactions on Audio, Speech,and Language Processing,2020,28:581-591.

[13] CHEN S,GE L. Exploring the attention mechanism in LSTM-based Hong Kong stock price movement prediction [J]. Quantitative Finance,2019,19(9):1507-1515.

[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,Volume 1 (Long and Short Papers). Minneapolis,Minnesota:Association for Computational Linguistics,2019:4171-4186.

[15] LAN Z Z,CHEN M D,GOODMAN S,et al. ALBERT:a lite bert for self-supervised learning of language representations [Z/OL]. (2020-02-09)[2022-03-18]. https://doi. org/10. 48550/arXiv. 1909. 11942.

[16] 万圣贤,兰艳艳,郭嘉丰,等.用于文本分类的局部化双向长短时记忆[J].中文信息学报,2017,31(3):62-68.

[17] 黄培松,黄沛杰,丁健德,等.基于隐含主题协同注意力网络的领域分类方法[J].中文信息学报,2020,34(2): 73-79.

[18] XU G X,ZHANG Z X,ZHANG T,et al. Aspect-level sentiment classification based on attention-BiLSTM model and transfer learning [J]. Knowledge-Based Systems,2022,245:108586.

[19] BASIRI M E,NEMATI S,ABDAR M,et al. ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis [J]. Future Generation Computer Systems,2021,115:279-294.

CCM-MF: Chinese-text Classification Model Based on Fused Multi-dimensional Features

MA Zichen[1,2], ZHANG Shunxiang[1,2**], LIU Yunduo[1,2], WANG Xingguang[1,2], ZHANG Youqiang[1,2]

(1. School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, Anhui, 232001, China;
2. Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, Anhui, 230088, China)

Abstract: In view of the difference of semantic information carried by different dimensional features in Chinese text, a Chinese-text Classification Model based on Fused Multi-dimensional Features (CCM-MF) was proposed. The model combines hierarchical dimension and spatial dimension features to improve the accuracy of Chinese text classification. Firstly, on the hierarchical dimension, the Enhanced Representation through Knowledge Integration (ERNIE) pre-training model is used to obtain word vectors containing features of character, word, and entity levels. Then, on the spatial dimension, the word vectors containing hierarchical dimension features are input into the improved Deep Pyramid Convolutional Neural Networks (DPCNN) model and Attention-Based Bidirectional Long Short-Term Memory Networks (Att-BLSTM) model to obtain local and global semantic features, respectively. Finally, the obtained spatial dimension features are applied to the Softmax classifier, and then the calculation results are fused and the classification results are output. Through experiments on multiple public data sets, this model has better performance in accuracy than the existing mainstream text classification methods, which proves the effectiveness of the model.

Key words: Chinese text categorization; multiple dimensions; ERNIE; DPCNN; Att-BLSTM

责任编辑:梁　晓