

# 基于 ERNIE 模型的中文文本分类研究

毕云杉,钱亚冠,张超华,潘俊,徐庆华

(浙江科技学院理学院,杭州 310023)

**摘要:** 针对基于深度学习的中文文本分类任务中词向量表示无法充分利用语义信息的问题,提出一种基于知识增强语义表示(enhanced representation through knowledge integration, ERNIE)模型的中文文本分类方法。首先,通过 ERNIE 模型获得语义表达更充分的分布式文本表示;然后引入深度卷积神经网络对上下文的编码特征进一步提取,以获得更深层次的文本特征表达;最后采用分类器(soft maximum, softmax)实现中文文本分类。在 3 个公开的中文数据集上进行了多组对比试验,发现本模型与传统基于双向编码器表征量(bidirectional encoder representation from transformers, BERT)的分类模型相比,准确率和  $F_1$  值分别平均提升了 6.34%、4.82%,表明基于 ERNIE 模型的文本分类方法能有效提高中文文本分类的性能。本方法在多领域中文文本数据集上能够更准确地实现文本的分类,可为后续自然语言处理领域研究提供参考。

**关键词:** 自然语言处理;文本分类;深度学习;卷积神经网络;ERNIE

**中图分类号:** TP391.1      **文献标志码:** A      **文章编号:** 1671-8798(2021)06-0461-08

## Research on Chinese text classification based on ERNIE model

BI Yunshan, QIAN Yaguan, ZHANG Chaohua, PAN Jun, XU Qinghua

(School of Sciences, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China)

**Abstract:** In response to the problem that word vector representation can not fully utilize semantic information in the Chinese text classification task based on deep learning, a Chinese text classification method was proposed on the basis of ERNIE (enhanced representation through knowledge integration) model. First, a more semantically expressive distributed text representation was obtained through the ERNIE model. Then, the deep convolutional neural network was introduced to further extract the encoding features of the context to obtain a deeper representation of the text features. Finally, a classifier (soft maximum, softmax) was used to realize Chinese text classification. A series of comparative experiments were conducted on three published Chinese data sets, and it was found that compared with the traditional classification model based on BERT (bidirectional encoder representation from transformers), this model has raised the accuracy and  $F_1$  value by 6.34% and 4.82% respectively, indicating that the text classification

**收稿日期:** 2020-12-15

**基金项目:** 科技部重点研发项目(2018YFB2100400);国家自然科学基金项目(61902082)

**通信作者:** 钱亚冠(1976—),男,浙江省嵊州人,副教授,博士,主要从事机器学习、大数据处理、对抗性机器学习研究。E-mail: qianyg@yeah.net。

method based on ERNIE model can effectively improve the performance of Chinese text classification. The proposed method can achieve text classification more accurately on multi-domain Chinese text data sets, and can provide a reference for subsequent research in natural language processing.

**Keywords:** natural language processing; text classification; deep learning; convolutional neural network; ERNIE

近年来,随着移动互联网的迅猛发展,文本数据呈爆炸式增长,而文本分类作为自然语言处理(natural language processing, NLP)中的关键技术之一,受到了越来越广泛的关注。文本分类是一种将文本进行转换并自动分类到指定的某个或某几个类别之中的技术, NLP 领域中的多种应用均可划分为文本分类,如意图识别、领域识别、情感分析等。文本分类的方法包括基于规则的方法<sup>[1]</sup>、基于统计机器学习的方法<sup>[2-3]</sup>及基于深度学习的方法<sup>[4-5]</sup>,其中,由于深度学习强大的表示学习的能力,被广泛应用于文本分类任务中。而随着深度学习预训练语言模型的发展,文本分类任务的性能取得了更进一步的提升。

早期的预训练语言模型有文本向量化(word to vector, word2vec)<sup>[6]</sup>、全局词向量(global vectors for word representation, GloVe)<sup>[7]</sup>等,但这些模型不能解决一词多义问题;之后,从语言模型中获取词向量(embedding from language model, ELMo)<sup>[8]</sup>等联系上下文的词向量预训练方法被提出,解决了一词多义问题,但 ELMo 采用的是长短期记忆网络(long-short-term memory network, LSTM)对特征进行提取,而 LSTM 提取特征的能力相对较弱,不能很好地融合上下文的特征。近几年,生成性预训练(generative pre-training, GPT)模型<sup>[9]</sup>、双向编码器表征量(bidirectional encoder representation from transformers, BERT)模型<sup>[10]</sup>等被相继提出。GPT 是第一个引入了转换器<sup>[11]</sup>的预训练语言模型,转换器特征提取能力强于 LSTM,且转换器并行计算能力强,但 GPT 本质上还是一个单向语言模型,无法获得上下文相关的特征表示,无法捕捉更多的信息。BERT 模型是一种基于双向转换器构建的语言模型,由谷歌公司提出后被广泛应用到自然语言处理任务中<sup>[12-13]</sup>。谌志群等<sup>[14]</sup>将 BERT 模型与双向(bi-directional) LSTM 相结合,构建了 BERT-BiLSTM 模型,用其对微博评论的倾向性进行分析,与传统方法相比提高了分类的准确性。马强等<sup>[15]</sup>提出基于 BERT 与卷积神经网络(convolutional neural network, CNN)的文本分类方法,引入 CNN 对 BERT 模型提取到的句子特征进一步提取,获得了更好的分类效果。由于 LSTM 并行处理的能力弱, CNN 虽拥有较强的并行处理能力,但无法获取长距离的文本依赖关系;而深度卷积神经网络(deep pyramid convolutional neural networks, DPCNN)通过不断加深网络,可以提取出深层次的文本特征。如李颖<sup>[16]</sup>提出基于 BERT 模型与 DPCNN 的垃圾弹幕识别方法,与 LSTM、CNN 方法相比提高了分类的准确性。但是, BERT 预训练模型只关注了自身的文本信息,不能全面地获取潜在的文本特征。

针对上述问题,本研究利用知识增强语义表示(enhanced representation through knowledge integration, ERNIE)模型,它在 BERT 模型基础上进一步优化,加入了大量语料中的词法、句法等先验语义信息,能更全面地捕捉训练语料中的潜在信息,并引入深度卷积神经网络获取长距离的文本依赖关系,对文本特征进行深层次的提取,构建 ERNIE-DPCNN 模型进行中文文本分类,通过提升文本语义表征的能力来增强中文文本分类的效果。

## 1 ERNIE 模型与 DPCNN 网络

### 1.1 ERNIE 模型

ERNIE 模型<sup>[17]</sup>可以同时充分利用词法、句法和语义信息,相比 BERT 模型, ERNIE 模型的改进主要是在掩码机制上。BERT 模型与 ERNIE 模型不同的掩码策略如图 1 所示, BERT 模型建模对象主要聚焦在原始语言信号上,没有充分利用训练数据当中的词法结构、句法结构及语义信息去学习建模,在训练中文文本时, BERT 模型随机掩盖掉 15% 的词进行预测,通过这种强行掩盖的方式,把“杭”和“州”字与字之间的关系给拆散了,模糊了“杭州”和“西湖”词与词之间的联系;而 ERNIE 模型采用的掩码语言模型是一种带有先验知识的掩码机制,通过对词、短语等语义信息进行建模,得出“杭州”与“西湖”的关系,学到“杭州”是“西湖”的“所在地”及“杭州”被称为“人间天堂”,从而推断出此时掩盖掉的词是“杭州”,使得模型学习到完整概念的语义表示。

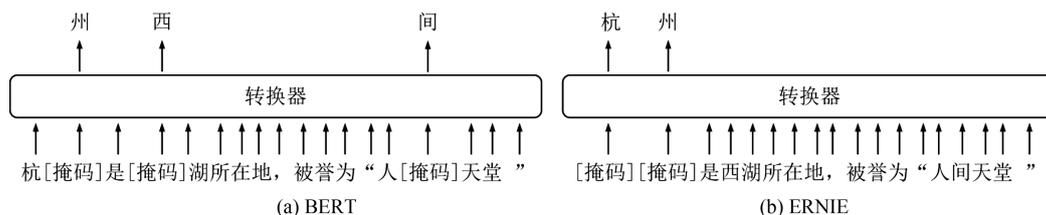


图 1 BERT 模型与 ERNIE 模型不同的掩码策略

Fig. 1 Different mask strategies for BERT model and ERNIE model

### 1.2 DPCNN 网络

深度卷积神经网络<sup>[18]</sup>是由腾讯人工智能实验室提出的,其通过不断加深网络,可以获得长距离的文本依赖关系。在 DPCNN 网络中卷积使用等长卷积,即输出序列的长度等于输入序列的长度;在进行池化操作之前要固定特征图的数量,如果增加特征图的数量会大幅地增加计算时间,而精度并没有提高。在固定了特征图的数量后,每当使用一个卷积核大小为 3、步长为 2 的池化层进行最大池化,每个卷积层的序列长度就被缩短为原始长度的一半,感知到的文本片段比之前长了一倍,计算时间减半,从而形成一个金字塔结构,进一步提取了有效特征。

## 2 ERNIE-DPCNN 模型

我们提出的 ERNIE-DPCNN 模型主要分为输入层、ERNIE 层、DPCNN 层和文本分类输出层,模型结构如图 2 所示。

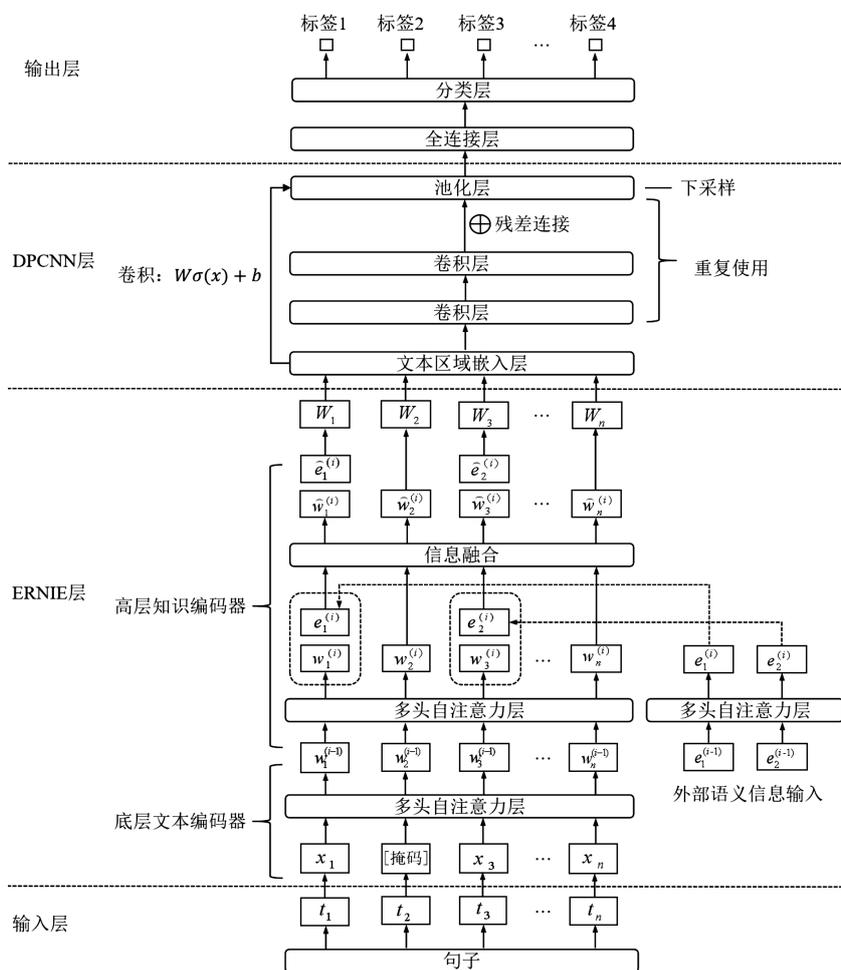


图 2 ERNIE-DPCNN 模型结构

Fig. 2 Architecture of ERNIE-DPCNN model

### 2.1 输入层

将原始的句子转化为大小为(batch\_size, padding\_size)的词嵌入表示  $\{t_1, t_2, t_3, \dots, t_n\}$ , 其中, batch\_size 为每批次训练样本大小, padding\_size 为句子最大序列长度。

### 2.2 ERNIE 层

ERNIE 层利用多层转换器的自注意力双向建模性能, 主要采用了转换器的编码器部分。转换器通过矩阵间的计算来获得人们所关注的信息, 抑制其他无用的信息, 获得每个词新的表征, 从而实现自注意力机制。转换器编码单元如图 3 所示, 首先将输入向量矩阵  $T$  通过位置编码后得到矩阵  $X$ , 与各自权重矩阵  $W^q$ 、 $W^k$ 、 $W^v$  相乘, 用于获得词语之间的相互关联程度, 得到查询矩阵  $Q$ 、键矩阵  $K$  和值矩阵  $V$ :

$$Q = XW^q; \quad (1)$$

$$K = XW^k; \quad (2)$$

$$V = XW^v. \quad (3)$$

然后将查询矩阵与键矩阵相乘, 将乘积的结果除以键矩阵的秩  $d$  的算术平方根  $\sqrt{d}$ , 以保证训练过程具有更稳定的梯度, 从而获得新的关联度。之后经过分类器(soft maximum, softmax)归一化后对值矩阵  $V$  加权求和, 将关联程度与语义相结合, 最终得到自注意力层输出

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \quad (4)$$

ERNIE 层由底层文本编码器和高层知识编码器堆叠组成。底层编码器是一个多头双向的转换器, 用于获取输入的基础词汇和语义信息, 底层采用 BERT 的掩码策略, 对获取的语义信息进行初步的掩码。高层知识编码器负责将外部的知识信息整合融入模型中, 多头注意力层对底层文本编码器的输出  $\{\omega_1^{(i-1)}, \omega_2^{(i-1)}, \omega_3^{(i-1)}, \dots, \omega_n^{(i-1)}\}$  与外部语义信息输入  $e_1^{(i)}, e_2^{(i)}$  分别进行处理, 经信息融合后得到新的语义信息输出  $\{\hat{\omega}_1^{(i-1)}, \hat{\omega}_2^{(i-1)}, \hat{\omega}_3^{(i-1)}, \dots, \hat{\omega}_n^{(i-1)}\}$  与语义信息输出  $\hat{e}_1^{(i)}, \hat{e}_2^{(i)}$ 。ERNIE 层输出的维度为(batch\_size, hidden\_size), 其中 hidden\_size 为 ERNIE 隐藏层的层数。

### 2.3 DPCNN 层

将 ERNIE 层最后一层的输出作为 DPCNN 层的特征输入, DPCNN 层底层是文本区域嵌入层, 对一个文本区域或者片段进行一组卷积操作后嵌入模型中; 采用 2 个等长卷积生成特征, 在每层卷积层中加入线性修正单元(rectified linear unit, ReLU)<sup>[19]</sup> 作为激活函数, 增强神经网络模型的非线性, 减少参数的相互依存关系, 同时加快网络训练速度, 防止梯度消失, 缓解模型过拟合的问题; 卷积层与池化层之间采用残差连接<sup>[20]</sup>, 在每 2 层中增加一个捷径, 构成 1 个残差块, 增加了多尺度信息, 极大地缓解了梯度弥散的问题。2 层等长卷积和 1 个池化下采样组成 1 个模块, DPCNN 层重复使用这种模块就能够抽取文本的长距离依赖关系, 进一步提取文本特征。最后的池化层将整个文本表示为 1 个向量  $V$ 。

### 2.4 输出层

将最终提取的特征向量输入全连接层采用 softmax 分类器进行分类, 从而得到模型最终预测的文本分类结果

$$N = \text{softmax}(w_0V + b_0). \quad (5)$$

式(5)中:  $w_0$  为权重系数矩阵;  $b_0$  为偏置项;  $N$  为最终输出的文本分类标签。

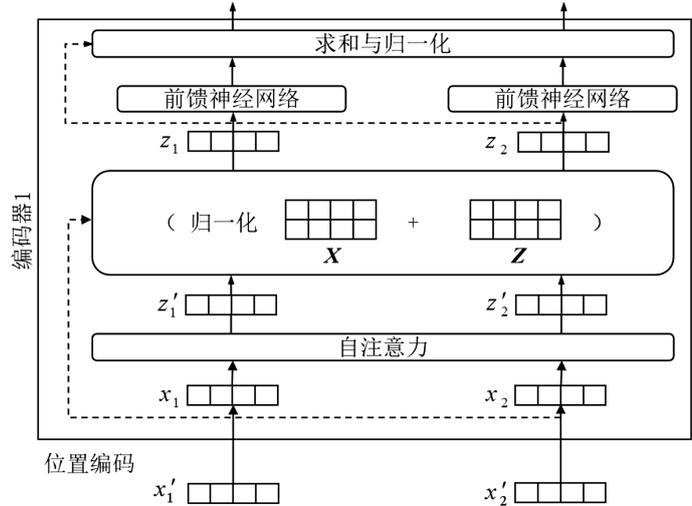


图 3 转换器编码单元

Fig. 3 Encoder unit of transformer

### 3 试验与分析

#### 3.1 试验设置

选取3个中文数据集作为试验语料,数据集详情见表1。

新闻文本数据集<sup>[21]</sup>共有金融、房产、股票、教育等10个分类,包含160 000条训练数据,20 000条验证数据,20 000条测试数据。

网上购物数据集<sup>[22]</sup>共有书籍、平板、手机等10类评论数据,包含50 216条训练数据,6 277条验证数据,6 277条测试数据。

第四届全国应用统计专业学位研究生案例大赛的旅游评论数据集<sup>[23]</sup>中,涉及正面、中性及负面评论,包含3 600条训练数据,400条验证数据,由于数据集较小,测试数据等同于验证数据。

ERNIE-DPCNN模型最大序列长度设置为128,每批次处理的数据量大小设置为64,迭代次数设置为3,学习率设置为0.000 02,使用自适应矩估计优化器(adaptive moment estimation, Adam)更新网络参数。预训练模型使用了百度发布的预训练好的“ERNIE 1.0 Base 中文”模型,以中文维基百科、百度百科、百度新闻、百度贴吧等数据集为训练语料,本模型采用了12层的转换器,隐藏层层数为768,多头注意力机制参数设置为12,模型总参数大小为110 MB。

#### 3.2 评价指标

为了评估ERNIE-DPCNN模型在中文文本分类问题上的可行性,本研究采用准确率( $A$ )、损失值( $L$ )、精度( $P$ )、召回率( $R$ )及 $F_1$ 值作为评价指标。准确率用以衡量分类器预测结果与真实结果之间的差异;损失值为分类器预测错误的概率;精度指在被所有预测为正的样本中实际为正样本的概率;召回率指在实际为正的样本中被预测为正样本的概率。 $F_1$ 值定义如下:

$$F_1 = \frac{2PR}{P+R} \quad (6)$$

如式(6)所示, $F_1$ 值是精度与召回率的调和平均,在给定阈值下, $F_1$ 值越接近1,模型的性能越好。对于多分类的文本分类问题,在计算各项评价指标时,若选定其中一类为正类,则其余的类别作为负类,从而将多分类问题转换为二分类问题进行处理。

#### 3.3 对比试验设置

为了验证ERNIE-DPCNN模型在文本分类任务中的有效性,将我们提出的分类模型与下列性能较高的分类模型进行对比试验,对比模型如下。

1)BERT模型:经中文预训练BERT模型得到文本特征后,通过全连接层输入到分类器中实现文本分类。

2)BERT-CNN模型:在模型1)的基础上,将预训练得到的文本特征表示输入到卷积神经网络中,做进一步的特征提取<sup>[15]</sup>,通过全连接层输入分类器中实现文本分类。

3)BERT-BiLSTM模型:经模型1)中的预训练模型得到文本表示后,利用BiLSTM模型提取句子中各个词的上下文信息<sup>[14]</sup>,通过全连接层输入分类器中实现文本分类。

4)BERT-DPCNN模型:经模型1)中的预训练模型得到文本表示后,利用深度卷积神经网络提取文本特征<sup>[16]</sup>,并在神经网络中引入非线性ReLU激活函数,通过全连接层输入分类器中实现文本分类。

5)ERNIE模型:预训练采用“ERNIE 1.0 Base 中文”模型,直接连接全连接层实现文本分类。

在模型参数设置过程中,尽量保持设置上的一致性;由于模型结构的差异而无法达到一致时,采取最优的参数设置。试验对比模型参数设置见表2。

表1 数据集详情

Table 1 Data set details

数据集	类别	总数据量	训练集	验证集	测试集
新闻文本数据集	10	200 000	160 000	20 000	20 000
网上购物数据集	10	62 770	50 216	6 277	6 277
旅游评论数据集	3	4 000	3 600	400	400

表 2 试验对比模型参数设置

Table 2 Parameter settings of experimental comparison model

模型	参数	数值
BERT	转换器层数	12
	隐藏层层数	768
	多头注意力机制头数	12
	最大序列长度	128
	每批次训练样本个数	64
	迭代次数	3
	学习率	0.000 02
BERT-CNN	总参数大小/MB	110
	最大序列长度	128
	每批次训练样本个数	64
	迭代次数	4
	卷积核个数	256
BERT-BiLSTM	卷积核窗口大小	[2,3,4]
	丢弃率	0.5
	最大序列长度	128
	每批次训练样本个数	64
BERT-DPCNN	迭代次数	4
	卷积核个数	250
	隐藏层层数	768
ERNIE	最大序列长度	128
	每批次训练样本个数	64
	迭代次数	3
	总参数大小/MB	110

### 3.4 试验结果与分析

#### 3.4.1 试验结果

基于 ERNIE 的文本分类模型与基准模型各项评价指标的试验结果见表 3~5。由表 3~5 可知,在 3 个数据集上的各项评价指标的值均有不同程度的提升,ERNIE-DPCNN 模型在新闻文本、网上购物及旅游评论数据集上分类结果的  $F_1$  值分别达到了 94.75%、87.56%、63.61%,表明采用 ERNIE-DPCNN 模型进行文本分类的可行性和有效性。

BERT-CNN、BERT-DPCNN 模型均属于 CNN

类别模型,BERT-BiLSTM 模型属于 RNN 类别模型,CNN 类别模型在中文文本分类任务上的各项评价指标的值都要高于 RNN 类别模型,说明了在中文文本分类中更注重词的提取,而 RNN 类别模型更注重于联系上下文的信息,并没有展现出更好的性能。同时,BERT-DPCNN 模型在新闻文本、网上购物及旅游评论数据集上分类结果的准确率分别为 88.74%、85.02%、82.63%,而 ERNIE-DPCNN 模型分类结果的准确率分别达到了 94.75%、89.79%、85.12%,说明 ERNIE 模型比 BERT 模型特征提取能力更强。

此外,在新闻文本数据集上各模型分类结果的各项指标的值最高,文本分类性能最好,在旅游

表 3 新闻文本分类数据集的试验结果

Table 3 Test results on news-text data set %

模型	A	P	R	$F_1$
BERT	88.30	88.27	88.30	88.28
BERT-CNN	88.61	88.57	88.61	88.52
BERT-BiLSTM	87.94	87.87	87.94	87.90
BERT-DPCNN	88.74	88.77	88.74	88.75
ERNIE	94.68	94.68	94.67	94.67
ERNIE-DPCNN	94.75	94.75	94.75	94.75

评论数据集上各模型分类结果的各项评价指标的值较低,可能是由于数据集数据量小、分类数据不均衡。

表4 网上购物数据集的试验结果

模型	A	P	R	$F_1$
BERT	83.97	84.27	77.95	80.99
BERT-CNN	84.53	84.84	78.05	81.30
BERT-BiLSTM	83.85	82.41	77.74	80.01
BERT-DPCNN	85.02	85.94	78.56	82.08
ERNIE	89.66	89.74	85.44	87.49
ERNIE-DPCNN	89.79	90.05	85.20	87.56

表5 旅游评论数据集的试验结果

模型	A	P	R	$F_1$
BERT	81.25	54.42	53.99	54.20
BERT-CNN	81.38	55.37	54.10	54.73
BERT-BiLSTM	80.62	53.55	53.19	53.37
BERT-DPCNN	82.63	56.12	56.00	56.06
ERNIE	85.00	62.63	61.47	62.05
ERNIE-DPCNN	85.12	63.66	63.51	63.61

### 3.4.2 多头注意力机制的参数对文本分类的影响

为了探究改变多头注意力机制的头数对ERNIE-DPCNN模型进行中文文本分类的影响,选择最具有代表性的新闻文本数据集进行试验,结果见表6。

从表6中可以发现,随着混合模型多头注意力机制头数的增加,文本分类的效果总体上呈逐渐提升的趋势。当多头注意力机制头数为12时,ERNIE-DPCNN模型分类效果最好,此时损失最低,损失率为17%, $F_1$ 值为94.75%;当多头注意力机制头数达到16时,文本分类效果减弱,准确率、 $F_1$ 值降低,损失升高。试验结果表明,适当增加多头注意力机制的头数,使得模型捕获到更多层面的语义特征,能提高ERNIE-DPCNN模型进行文本分类的性能。

表6 多头注意力机制头数对分类结果的影响

头数	A	L	P	R	$F_1$
4	88.77	35	88.88	88.77	88.82
6	89.98	31	90.00	89.98	89.99
8	93.30	21	93.33	93.30	93.31
12	94.75	17	94.75	94.75	94.75
16	91.02	28	91.06	91.02	91.04

### 3.4.3 激活函数对文本分类的影响

为了探究改变激活函数对ERNIE-DPCNN模型进行中文文本分类的影响,选取随机线性修正单元(randomized leaky rectified linear unit, RReLU)<sup>[19]</sup>和参数化线性修正单元(parametric rectified linear unit, PReLU)<sup>[19]</sup>作为激活函数,与ReLU激活函数进行了试验对比,选择最具有代表性的新闻文本数据集进行试验,结果见表7。

表7 激活函数对分类结果的影响

激活函数	A	L	P	R	$F_1$
RReLU	94.12	17	94.13	94.12	94.13
PReLU	94.18	17	94.18	94.18	94.18
ReLU	94.75	17	94.75	94.75	94.75

从表7可以看出,ERNIE-DPCNN模型采用这3个激活函数进行文本分类造成的损失值相同,均为17%,其中,采用ReLU激活函数进行文本分类的结果各项评价指标的值最高,分类试验效果最好,证明了ReLU激活函数对解决深度卷积网络中梯度弥散问题的有效性,在一定程度上提升了ERNIE-DPCNN模型进行文本分类的性能。

### 3.4.4 学习率对文本分类的影响

为了探究学习率对ERNIE-DPCNN模型进行中文文本分类的影响,选择最具有代表性的新闻文本数据集进行试验,结果见表8。

从表8可以看出,在一定范围内,设置小的学习率,文本分类结果的准确率、 $F_1$ 值会适当地提高;当学习率设为0.000 05时,ERNIE-DPCNN模型进行文本分类的效果最好,此时损失率最低,超过了这个范围,降低学习率会使文本分类结果的准确率、 $F_1$ 值

表8 学习率对分类结果的影响

学习率	A	L	P	R	$F_1$
0.000 2	94.37	18	94.38	94.37	94.37
0.000 1	94.49	18	94.52	94.49	94.51
0.000 05	94.75	17	94.75	94.75	94.75
0.000 01	93.68	18	94.30	94.26	94.28

下降,损失率提高。

## 4 结 语

本研究将 ERNIE 和深度卷积神经网络相结合,建立中文文本分类模型。为了证明模型的有效性,我们将提出的模型与其他主流中文文本分类方法进行试验对比,准确率和  $F_1$  值平均分别提升了 6.34%、4.82%,结果表明 ERNIE-DPCNN 模型能有效提高中文文本分类的性能。但是,本研究还存在一些不足之处,模型对分类数据不平衡的数据集,是否能通过加入条件随机场模型等方法来提高这类数据集分类的精度;另外,模型在词向量表示和特征提取等方面还有一定的提升空间,这都是下一阶段工作的研究方向。

## 参考文献:

- [1] 史东辉,蔡庆生,倪志伟,等.基于规则的分类数据离群挖掘方法研究[J].计算机研究与发展,2000,37(9):1094.
- [2] 张孝飞,黄河燕.一种采用聚类技术改进的 KNN 文本分类方法[J].模式识别与人工智能,2009,22(6):936.
- [3] 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展[J].软件学报,2006,17(9):1848.
- [4] 韩众和,夏战国,杨婷.CNN-ELM 混合短文本分类模型[J].计算机应用研究,2019,36(3):663.
- [5] 孟先艳,崔荣一,赵亚慧,等.基于双向长短时记忆单元和卷积神经网络的多语种文本分类方法[J].计算机应用研究,2020,37(9):2669.
- [6] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]//Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics. Cham: Springer,2013:430.
- [7] PENNINGTON J, SOCHER R, MANNING C. GloVe: global vectors for word representation[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press,2014:1532.
- [8] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: NAACL Press,2018:2227.
- [9] RADFORD A, NARASIMHAN K, SALMANS T, et al. Improving language understanding by generative pre-training[EB/OL]. (2018-06-12)[2020-08-23]. <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf>.
- [10] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: NA ACL Press,2019:4171.
- [11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-06-12)[2020-08-23]. <https://arxiv.org/pdf/1706.03762.pdf>.
- [12] 张冬瑜,崔紫娟,李映夏,等.基于 Transformer 和 BERT 的名词隐喻识别[J].数据分析与知识发现,2020(4):104.
- [13] 杨玉亭,冯林,代磊超,等.面向上下文注意力联合学习网络的方面级情感分类模型[J].模式识别与人工智能,2020,33(8):753.
- [14] 湛志群,鞠婷.基于 BERT 和双向 LSTM 的微博评论倾向性分析研究[J].情报理论与实践,2020,43(8):173.
- [15] 马强,赵鸣博,孔维健,等.一种基于 BERT 与 CNN 层级连接的中文文本分类方法:CN111177376A[P]. 2020-05-19.
- [16] 李颖.基于 BERT-DPCNN 的垃圾弹幕识别改进及应用[D].上海:上海师范大学,2020.
- [17] SUN Y, WANG S H, LI Y K, et al. ERNIE: enhanced representation through knowledge integration[EB/OL]. (2019-04-19)[2020-08-23]. <https://arxiv.org/pdf/1904.09223.pdf>.
- [18] 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. Vancouver: ACL Press,2017:562.
- [19] XU B, WANG N, CHEN T, et al. Empirical evaluation of rectified activations in convolutional network[EB/OL]. (2015-05-05)[2020-08-23]. <https://arxiv.org/pdf/1505.00853.pdf>.

(下转第 476 页)