2025 年 第 25 卷 第 2 期
2025, 25(2): 00649-08

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2307782
引用格式:李莉, 张之欣, 王小龙. 基于轻量化改进 ERNIE-RCNN 的中文新闻标题分类[J]. 科学技术与工程, 2025, 25(2): 649-656.
Li Li, Zhang Zhixin, Wang Xiaolong, et al. Chinese news title classification based on lightweight improved ERNIE-RCNN[J]. Science Technology and Engineering, 2025, 25(2): 649-656.

自动化技术、计算机技术

基于轻量化改进 ERNIE-RCNN 的中文新闻标题分类

李莉[1,2], 张之欣[1*], 王小龙[1]
(1. 华北电力大学控制与计算机工程学院, 保定 071003; 2. 河北省能源电力知识计算重点实验室, 保定 071003)

摘 要 针对大型预训练语言模型在处理新闻标题时,面临参数规模庞大、无法高效利用上下文语意特征以及循环卷积神经网络对初始输入元素重要性忽视的问题,提出了一种融合混合专家模型(mixture-of-expert, MoE)的 ERNIE 与注意力机制的循环卷积神经网络(recurrent convolutional neural networks, RCNN)的新闻标题分类方法。首先,借助 MoE 改进 ERNIE 技术进行文本编码,随后利用注意力 RCNN 在保留文本词序和特征的基础上进行分类。为提高分类能力,通过计算输入的融合上下文权重对 RCNN 进行改进。在计算 MoE 中各个专家权重的过程中,选择 Gumbel_Softmax 作为新型的门控函数以改进传统的 Softmax 函数,从而更好地控制平滑程度。根据实验结果,发现相较于传统的分类方法,本文研究提出的分类方法展现出显著优势,极大地减少了参数数量。在此基础上,F_1 相较于传统模型提升了 0.51%。经过消融实验的验证,该分类方法在分类任务上的可行性得到了证实。

关键词 混合专家系统; 知识增强语义表示模型; 注意力机制; 循环卷积神经网络; 文本分类
中图法分类号 TP391.1; **文献标志码** A

Chinese News Title Classification Based on Lightweight Improved ERNIE-RCNN

LI Li[1,2], ZHANG Zhi-xin[1*], WANG Xiao-long[1]
(1. School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China;
2. Hebei Key Laboratory of Knowledge Computing for Energy & Power, Baoding 071003, China)

[**Abstract**] Aiming at the problems that the large-scale pre-training language model faces when dealing with news headlines, such as huge parameters, inefficient use of contextual semantic features and circular convolution neural network's neglect of the importance of initial input elements, a news headline classification method that combines ERNIE (enhanced representation through knowledge integration) of mixture-of-expert model and recurrent convolution neural network with attention mechanism were proposed. Firstly, the text was encoded with the help of MoE's improved ERNIE technology, and then the text was classified with attention RCNN (recurrent convolutional neural networks) on the basis of preserving the word order and characteristics of the text. In order to improve the classification ability, RCNN was improved by calculating the input fusion context weight. In the process of calculating the weights of experts in MoE, Gumbel-Softmax was selected as a new gating function to improve the traditional Softmax function, so as to better control the smoothness. According to the experimental results, it is found that compared with the traditional classification methods, the classification method proposed in this study shows significant advantages and greatly reduces the number of parameters. On this basis, the F_1 value is increased by 0.51% compared with the traditional model. After the ablation experiment, the feasibility of this classification method in the classification task has been confirmed.

[**Keywords**] MoE (mixture of experts); ERNIE (enhanced representation through knowledge integration); attention mechanism; RCNN (recurrent convolutional neural network); text classification

在自然语言处理领域,文本分类是一项至关重要的任务,广泛应用于情感分析[1]、智能客服[2]、新闻推荐系统[3]和舆情分析[4]等信息挖掘领域。新闻文本分类是文本分类领域中的一个关键子任务,它具有广泛的实际应用价值。随着社交媒体的普及,新闻传播速度更快、传播范围更广。一旦发生突发事件并在网络中传播,舆情扩散速度极快。如果传播中的突发事件属于负面新闻,就会造成巨大

收稿日期: 2023-10-07; 修订日期: 2024-10-17
第一作者: 李莉(1980—),女,汉族,重庆人,博士,副教授。研究方向:大数据分析、深度学习。E-mail:haolily12@163.com。
* 通信作者: 张之欣(1998—),男,汉族,河南新乡人,硕士研究生。研究方向:自然语言处理。E-mail:2973916737@qq.com。

的网络舆论,监管机构难以有效控制新闻舆论,也不利于社会稳定。网络舆情治理需要提前识别突发事件,而突发事件主要是以新闻文本为载体在互联网中传播。因此,新闻标题分类在网络舆情前期的监督管理工作中尤为重要,迫切需要研究新闻标题分类技术。

ERNIE(enhanced representation through knowledge integration)通过整合知识图谱和语义网络等语义信息,增强了文本的语义表示能力,从而更好地捕捉文本中的语义信息。然而,这也导致了模型在计算资源需求上的增加。为了解决这个问题,研究者们对模型进行了轻量化的改进,以减少硬件资源的需求,从而可以降低部署、运营和维护成本[5]。同时,轻量化的模型能够更好地运行在移动端设备上,快速且准确的分类模型有助于提高用户体验,随时随地满足用户获取感兴趣类别的新闻需求。

由于预训练语言模型 ERNIE 的参数量庞大且在提取上下文语义特征方面效率不高,导致将模型部署到移动设备上时面对着庞大的参数计算量问题。为了解决这个问题,现提出一种基于轻量化改进 ERNIE-RCNN 的中文新闻标题分类方法,旨在进一步提高中文新闻标题分类效率。通过局部替换 ERNIE 模型中的 encoder 的全连接层为并行处理的混合专家系统层,在保证不失精度的情况下,大幅降低计算资源需求,并且可以高效地进行词嵌入操作。注意力机制通过为每个单词或字符赋予不同的权重,RCNN(recurrent convolutional neural network)具有捕捉文本序列信息和空间特征的能力,注意力机制下的 RCNN 模型可以高效地解决上下文语义的学习。

1 文本分类相关研究

1.1 基于深度学习的文本分类研究

Kim[6]首先将卷积神经网络(convolutional neural network,CNN)应用于文本分类提出了 TextCNN,该模型在文本分类方面表现出了出色的性能。接着,Liu 等[7]通过将循环神经网络(recurrent neural network,RNN)引入文本分类中,设计出了能够有效捕捉更长的序列信息的 TextRNN。然而,TextCNN 在处理含有复杂上下文语境的文本数据方面表现不佳,且 TextRNN 在处理长序列数据时出现梯度爆炸问题。为了克服这些缺陷,TextRCNN[8]应运而生,它融合了 TextRNN 和 TextCNN 的优点,具备了更快的训练速度、更强大的上下文信息捕捉能力,以及适应稍显复杂的语义特性。

尽管如此,CNN 和 RNN 依然面临训练时间过长的挑战。为解决这一问题,Facebook 推出了 FastText[9-10]模型,它在处理简单文本分类任务上表现得更为迅速和高效。然而,FastText 模型在捕捉深层次语义关系方面尚有不足。针对这一问题,Johnson 等[11]提出了 DPCNN 模型,解决了 TextCNN 在获取文本长距离依赖方面的不足,并通过不断加深网络结构以减轻其缺点。总之,众多文本分类算法各具优缺点,在实际应用中,需要根据具体情况进行权衡和调整。

1.2 基于预训练语言模型的文本分类研究

OpenAI 公司提出了生成式预训练模型,模型引入了一种新的自然语言处理范式,即预训练与微调相结合的方式。通过预训练,模型可以直接根据下游任务的需求进行微调,避免了从头开始训练的繁琐过程[12]。这一新范式的出现对自然语言处理领域产生了显著影响,为其发展带来了巨大推进。BERT 模型是由 Devlin 等[13]提出的一种基于深层 Transformer 的预训练语言模型,该模型不仅可以充分利用大规模无标注文本的语义信息,而且还可以加深自然语言处理中各个任务所使用的模型深度。

ERNIE[14]模型在 BERT 模型的基础上进行了优化,能够更准确地理解句子中的实体关系,从而更准确地提取语义信息。相比 BERT,ERNIE 的掩码机制有所不同。它不仅可以对单个字符进行屏蔽,还可以对整个实体进行屏蔽。通过预测被屏蔽的实体来训练模型,进而能够更好地捕捉实体之间的联系。因此,在自然语言处理任务中,ERNIE 模型表现出更高的性能、准确性和可靠性,更适用于中文新闻标题分类任务。

1.3 多模型融合进行文本分类研究

基于词向量的模型、基于上下文机制的模型、基于注意力机制的模型和基于语言的模型都有各自的优缺点,基于此很多学者将各种模型组合在一起进行文本分类研究。杨秀璋等[15]提出一种融合情感词典的改进 BiLSTM-CNN + Attention 模型的情感分类模型,用多通道注意力机制提取 CNN 和 LSTM 输出信息并进行融合,最后结合注意力机制对情感特征进行加成。翟学明等[16]提出一种混合神经网络和条件随机场相结合的文本情感分析,巧妙地运用 CNN 和 BiGRU 两种神经网络来捕获文本的深层语义信息和结构特征,最后采用条件随机场模型作为分类器从而能够准确地判断文本的情感类别。陆晓蕾等[17]选取国家信息中信公布的全国专利信息为实验数据,提出了一种基于预训练语言模型的 BERT-CNN 多层级专利分类模型,并探讨了全局与局部策略在专利多层文本分类上的差异。

1.4　混合专家系统的相关发展研究

混合专家系统[18]是一种新型的监督学习方法,该方法将多层网络进行模块化转换,使用门控网络来决定每个数据应该由哪个模型进行训练。随着深度学习技术的不断进步,计算成本的制约限制了模型规模的进一步扩大。为了解决这一问题,Shazeer 等[19]提出了一种基于稀疏门控的混合专家系统,并将其应用于 RNN 结构中。该方法在确保高效计算的同时,将模型规模提升了 1 000 多倍。

随后,Lepikhin 等[20]将混合专家系统的思想拓展到 Transformer 模型上,并表现出不错的效果。Fedus 等[21]提出了一个高效的预训练大模型 Switch Transformer,主要亮点在于简化了混合专家系统的路由算法,从而显著提高了计算效率。Google 在 2021 年推出了一个超大型模型 GLaM[22],其规模比 GPT-3[23]大 3 倍,但由于采用了稀疏门控的混合专家系统设计,其训练成本仅为 GPT-3 的 1/3。此外,GLaM 在 29 个 NLP 任务上超越了 GPT-3。Xue 等[24]提出了一种名为 WideNet 的结构,旨在解决在压缩模型参数量的情况下如何获得更好效果的问题。该方法首先通过层之间的参数共享来压缩模型大小,然后采用混合专家系统的设计来扩大模型容量。Zuo 等[25]将混合专家系统和知识蒸馏[26]相结合,旨在提高推理速度的同时提高模型效果。

2　模型框架及相关技术

2.1　模型框架

针对 ERNIE 模型具有较大的参数量计算问题、无法高效提取上下文语义特征以及 RCNN 中的输入向量中的元素具有不同的重要性级别被忽略的问题,本文模型在 ERNIE 的编码器层引入 SGMoE(sparsely-gated mixture-of-expert),并在 RCNN 中引入注意力机制,旨在减少模型参数量的同时准确提取特征并高效利用所提取到的特征。

本文模型专为中文文本分类任务设计,其整体架构如图 1 所示,共包含 4 个主要组成部分:①利用 token 进行划分句子,得到一个个分词 $W_1, W_2, \cdots, W_n$作为模型的输入层;②采用改进的 ERNIE 对输入的中文分词进行预训练,获取含有上下文语义的词向量 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$;③将词向量输入注意力 RCNN 网络中,结合上下文语境进行权重分配,得到最终的全局语义信息;④经过 Softmax 激活层处理后,得到最终的输出结果。

2.2　改进 ERNIE 模型

针对 ERNIE 模型处理任务存在海量参数,且通过现有方法进行知识蒸馏来训练小的压缩模型的

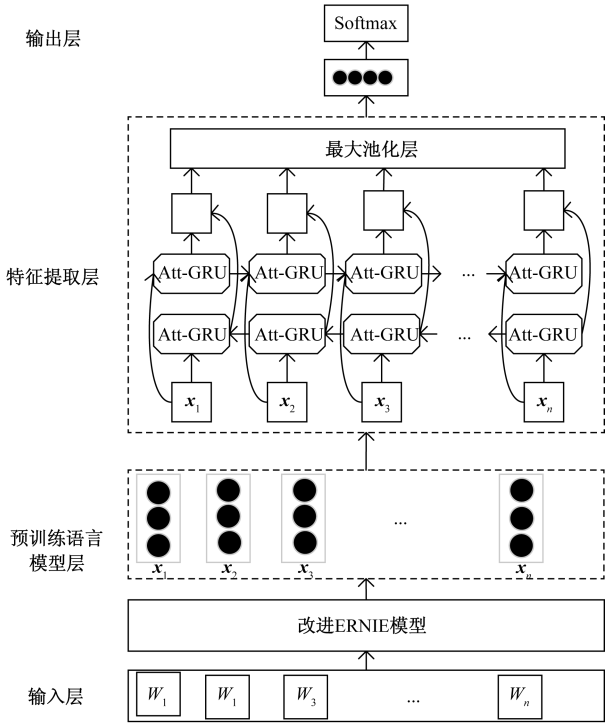


图 1　本文的模型框架

Fig. 1　The model framework of this paper

性能显著下降。采用混合专家结构来增加模型容量和推理速度,通过将预训练模型中的前馈神经网络提供给多个专家网络进行适配,这样预训练模型的表示能力在很大程度上得以保留。

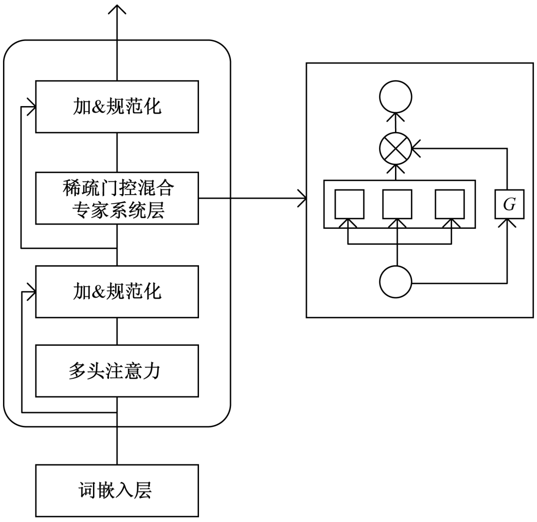
受到 Lepikhin 等[20]提出的 Gshard 模型的启发,改进后的 ERNIE 的 encoder 如图 2 所示,将每隔一个 encoder 的 FFN(feed forward networks)层,替换成 SGMoE 层,将计算稀疏门控值的函数由 Softmax 更换为 Gumbel_Softmax。在推理过程中,自适应地从众多专家网络中选择合适的专家网络以达到负载均衡,从而可以提高效率。

2.3　稀疏门控混合专家系统层

混合专家系统通过集成多个基础模型,旨在提高分类精度。由于不同数据来源的分布存在一定差异,单一模型通常只能处理部分数据,而在其他数据方面表现不佳。针对这一问题,采用多个专家模型处理来自不同来源的数据。每个专家网络在数据分类方面都有其专精的领域,在这些区域中的分类结果优于其他专家网络。通过门控网络进行筛选,决定将输入分配给哪一个专家网络进行处理。结构如图 3 所示。

对于在当前位置的输入 $\boldsymbol{x}$,输出就是所有专家的加权和,即

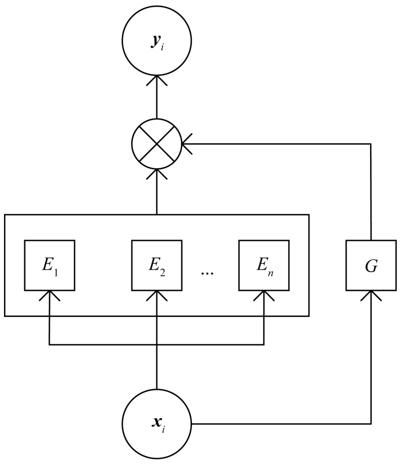
$$\boldsymbol{y} = \sum_{i=1}^{n} G(\boldsymbol{x}_i) E(\boldsymbol{x}_i) \tag{1}$$



G 为门控单元

图2 改进 ERNIE 的 encoder 层

Fig. 2 Improved encoder layer for ERNIE



E_i 为第 i 个词向量的加权，$i=1,2,\cdots,n$

图3 稀疏门控混合专家系统层

Fig. 3 Sparsely-gated mixture-of-expert layer

式(1)中：$\boldsymbol{x}_i$ 为第 i 个分词的词向量；$G(\boldsymbol{x}_i)$ 和 $E(\boldsymbol{x}_i)$ 分别为门控网络的输出和第 i 个专家的输出；$\boldsymbol{y}_i$ 为了经过加权之后的词向量。

其中门控单元 G 为 Softmax 门控，即对输入 $\boldsymbol{x}$ 映射到 n 维后，使用 Softmax 来获取门控值。即

$$G(\boldsymbol{x}) = \text{Softmax}(\boldsymbol{x} * \boldsymbol{W}_g) \tag{2}$$

式(2)中：$\boldsymbol{W}_g$ 为 n 维权重矩阵。

考虑到不同专家之间的差异性以及负载均衡问题，通过 TopK 采样的方式实现稀疏性和将门控机制引入噪声的方式实现负载均衡问题。TopK 采样方法无法进行梯度计算，因此无法更新网络。采用 Gumbel_Softmax[27] 函数代替普通的 Softmax，其优点包括可以近似 TopK 采样的方式、提供采样所需的随机性以及不破坏计算的梯度传播。计算公式为

$$G(\boldsymbol{x}) = \text{Gumbel\_Softmax}(\boldsymbol{x} * \boldsymbol{W}_g) \tag{3}$$

$$\boldsymbol{y}_i = \frac{\exp\left[\dfrac{(\lg\boldsymbol{\pi}_i + g_i)}{\boldsymbol{\tau}}\right]}{\sum_{j=1}^{k}\exp\left[\dfrac{(\lg\boldsymbol{\pi}_j + g_j)}{\boldsymbol{\tau}}\right]} \tag{4}$$

$$\boldsymbol{\pi}_i = \boldsymbol{x}_i * \boldsymbol{W}_{g_i} \tag{5}$$

式(4)中：g_i 和 τ 为在 Softmax 的基础上引入的两个额外变量；g_i 为增加模型的灵活性而引入的服从 Gumbel 分布的噪声；$\boldsymbol{\tau}$ 为温度系数，是一个控制平滑程度的系数；$\boldsymbol{\pi}_i$ 为第 i 个类别的概率；$\boldsymbol{W}_{g_i}$ 为权重矩阵。

2.4 RCNN 模型

与传统的基于窗口的神经网络相比较，RCNN 能够改善文本窗口大小不足的缺陷，在文本分类任务上展现出优越的分类性能。因此，针对新闻文本分类的特点，加入 RCNN 模型作为深度特征提取模块。RCNN 结构如图 4 所示。

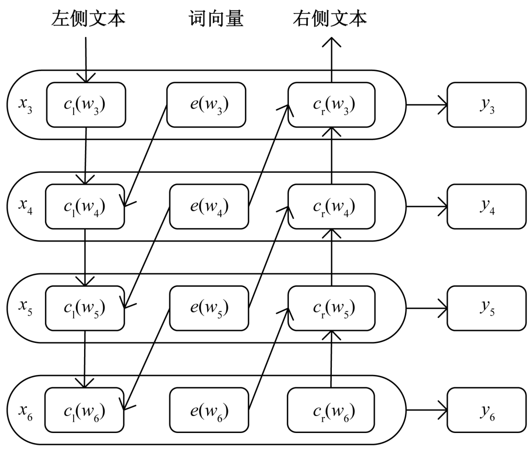


图4 RCNN 结构

Fig. 4 RCNN structure

考虑到参数量计算，利用 BiGRU 提取文本的上下文信息，并将 BiGRU 获得的隐层输出与词向量拼接，组合为新的词表示，即

$$c_l(w_i) = f[\boldsymbol{W}^l c_l(w_{i-1}) + \boldsymbol{W}^{sl}\boldsymbol{e}(w_{i-1})] \tag{6}$$

$$c_r(w_i) = f[\boldsymbol{W}^r c_r(w_{i+1}) + \boldsymbol{W}^{sr}\boldsymbol{e}(w_{i+1})] \tag{7}$$

式中：$c_l(w_i)$ 与 $c_r(w_i)$ 分别为词 w_i 的前向上文表示和后向下文表示；w_i 为输入的第 i 个词；$\boldsymbol{e}(w_{i-1})$ 为单词 w_{i-1} 的词向量；$c_l(w_{i-1})$ 为当前计算词的上一个词的表示形式；$\boldsymbol{W}^l$ 为隐含层的转移矩阵；$\boldsymbol{W}^{sl}$ 为另一个矩阵，用于将当前词的语义与下一个单词的前向上文表示相结合；f 为一个非线性的激活函数。

由式(6)和式(7)可以计算出每个词的前文表示与后文表示。随后，通过式(8)定义出每个词在神经网络中的表示，即

$$\boldsymbol{x}_i = [c_l(w_i), \boldsymbol{e}(w_i), c_r(w_i)] \tag{8}$$

式(8)中:$\boldsymbol{x}_i$ 为将词 w_i 的前文表示、词向量、后文表示拼接得到的结果,再对该结果使用一次 Sigmoid 激活函数,得到的句子表示经过最大池化层,得到特征向量并送入分类器进行分类。

2.5　元素级注意力门控机制

在循环神经网络中,输入向量的元素具有不同的重要性,然而这一特点往往被低估。为了应对这个问题,Zhang 等[28]提出了一种简单且有效的 EleAtt-RNN 结构,使得循环神经网络(RNN)神经元能够具备注意力机制,如图 5 所示。因此,该模型在处理输入元素时,更加注重权重分配,从而提高了整体性能。

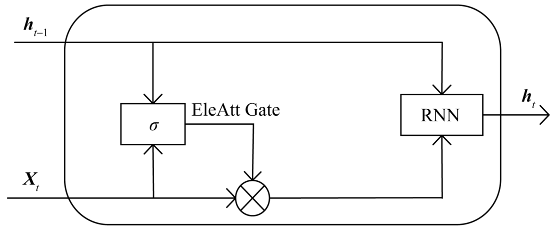


图 5　EleAtt-RNN 结构

Fig. 5　EleAtt-RNN structure

结构单元的计算公式为

$$a_t = \mathrm{Sigmoid}(\boldsymbol{w}_{\mathrm{xa}}\boldsymbol{X}_t + \boldsymbol{w}_{\mathrm{ha}}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{\mathrm{a}}) \tag{9}$$

$$\tilde{\boldsymbol{x}}_t = a_t\boldsymbol{X}_t \tag{10}$$

式中:$\boldsymbol{X}_t$ 为第 t 个词的词向量;$\boldsymbol{h}_{t-1}$ 为前 $t-1$ 个词的前向语境对应的词向量;$\boldsymbol{w}_{\mathrm{xa}}$ 和 $\boldsymbol{w}_{\mathrm{ha}}$ 为两个随机权重矩阵;$\boldsymbol{b}_{\mathrm{a}}$ 为偏置;a_t 为第 t 个词对上文的贡献权重;$\tilde{\boldsymbol{x}}_t$ 为结合上文加权之后的词向量。

在中文文本中,特征分布往往不均匀,不同的字词对上下文环境的贡献程度存在较大差异。为了解决这一问题,在循环神经网络的基础上引入了一种元素级注意力门控 EleAttG。该方法通过逐个字元素地自适应强化重要信息的贡献并抑制不重要信息的影响,从而提高了模型的语义理解能力。

3　实验与结果

3.1　数据集

选用清华大学 THUCNews 新闻语料库中的一部分数据集,该数据集包含了共计 10×10^4 条数据。为了训练、验证和测试模型,从这 10×10^4 条数据中随机抽取了 8×10^4 条作为训练集,1×10^4 条作为验证集,以及剩余的 1×10^4 条作为测试集。每条数据的平均长度为 24 个字符。这个数据集涵盖了 10 个类别:财经、房产、股票、教育、科技、社会、时政、体育、游戏和娱乐。在后续的实验中,将针对这些类别进行分类任务。具体数据示例如表 1 所示。

3.2　实验配置

本研究网络模型的实验配置如表 2 所示。

表 1　数据示例

Table 1　Data sample

输入	类别
皇马剥开胜利有个不得不说问题 走了的人是否会被怀念	7
某知名女星边工作边犒劳员工 率 12 人齐赴戛纳玩	9
常州一小区现楼晃晃 10 栋居民楼晃动半年多	1
多地学校探索多元化评价体系 学生全面发展受重视	3
顺义香悦四季 95～115 m² 3 居新房源在售享 98 折	1

表 2　实验配置

Table 2　Experimental configuration

环境	配置参数
处理器	Intel(R) Core (TM) i7－7700K CPU @4.20 GHz
显卡	NVIDIA GEFORCE RTX 3080Ti
内存	32 GB
框架	PyTorch 1.10

3.3　评价指标

使用精确率、召回率和 F_1 作为评估模型性能的指标。精确率衡量了预测为正实例的样本中真正为正实例的比例,召回率则衡量了所有真实正实例中被正确分类的比例。而 F_1 作为精确率和召回率的加权平均,综合考虑了两者的结果,提供了更全面的模型性能评估。

TP 表示预测为正类且实际也为正类的样本数量,也被称为真正例;TN 则表示预测为反类且实际也为反类的样本数量,被称为真反例。与之相对,FP 表示预测为正类但实际为反类的样本数量,即假正例;而 FN 则表示预测为反类但实际为正类的样本数量,也被称为假反例。

精确率(p_{recision})和召回率(r_{ecall})的计算公式为

$$p_{\mathrm{recision}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\times 100\% \tag{11}$$

$$r_{\mathrm{ecall}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times 100\% \tag{12}$$

加权计算得到 F_1,即

$$F_1 = \frac{2p_{\mathrm{recision}}r_{\mathrm{ecall}}}{p_{\mathrm{recision}}+r_{\mathrm{ecall}}}\times 100\% \tag{13}$$

此外,模型的损失函数采用的是交叉熵损失函数,交叉熵表示为真实概率分布与预测概率分布之间的差异,并且交叉熵的值越小,说明模型分类的结果越好。其公式为

$$L = \frac{1}{N}\sum_{i=1}^{n}\sum_{c=1}^{M} y_{ic}\lg p_{ic} \tag{14}$$

式(14)中:L 为模型的损失值;M 为新闻类别的数量;y_{ic} 为符号函数(0 或 1),即若样本 i 的真实类别等于 c 取 1,否则取 0;p_{ic} 为观测样本 i 属于类别 c 的预测概率。

3.4　实验结果分析

本文模型设计的批尺寸大小(batch_size)设置

为 64,训练迭代次数 epoch 设置为 10。TextCNN 中卷积核的尺寸大小分别取 3、4、5,BiGRU 和 RCNN 的隐藏层的数量为 256,BERT 和 ERNIE 隐藏层的数量为 768,网络模型优化器使用的是 Adam,设置学习率为 0.001,设 Gumbel_Softmax 中温度系数 τ 初始值为 1 并设置衰减率 0.01 来调整 τ 值。若超过1 000 个 batch_size 效果还没有提升就提前结束训练。

在实验中,RCNN、CNN 与 BiGRU 均采用相同的 Word2Vec 模型进行词向量表示,所用 RNN 和 RCNN 网络都是采用的是双向门控单元循环神经网络 BiGRU。通过对这些模型进行实验对比,综合分析了它们的优缺点,并在此基础上提出了本文方法。综合实验对比结果如表 3 所示。

由表 3 可以看出来,RCNN、TextCNN 与 BiGRU 在采用相同词向量表示方法下,得到的标题分类结果的 F_1 达到了 82.95%,相较于 TextCNN 和 BiGRU 分别提升了 1.68% 和 3.56%。通过对比 BERT 和 ERNIE 模型,ERNIE 模型的精确率提升了 0.44%,说明在处理标题分类问题时 ERNIE 模型能够得到更为准确完整的词向量语义表示。最后,通过将本文模型分别与 ERNIE-CNN、ERNIE-BiGRU 和 ERNIE-RCNN,相较于这 3 个模型在 F_1 上分别提升了 1.94%、1.7% 和 0.51%,从而可以看出本章所提模型能够更充分地捕捉上下文语境从而提高模型的分类精度。

为探究混合专家网络模型中专家数的重要性,将专家数设置为 4、6、8、10、12,共计 5 组实验,以 ERNIE-base 为例,各组实验结果如表 4 所示。

表 3 综合实验结果

Table 3 Comprehensive experimental results

模型	精确率/%	召回率/%	F_1/%
TextCNN	81.83	80.71	81.27
BiGRU	79.76	79.02	79.39
RCNN	83.77	82.15	82.95
BERT	90.68	90.55	90.61
ERNIE	91.04	90.99	91.01
ERNIE + CNN	92.32	92.28	92.29
ERNIE + BiGRU	92.56	92.51	92.53
ERNIE + RCNN	93.79	93.66	93.72
本文方法	94.26	94.21	94.23

表 4 专家数对比实验结果

Table 4 The experts compared the experimental results

专家数	精确率/%	召回率/%	F_1/%	参数量/10^6
4	90.84	90.85	90.84	26.64
6	91.21	91.19	91.20	27.83
8	91.37	91.32	91.34	29.01
10	91.51	91.45	91.48	30.19
12	91.38	91.36	91.37	31.38

根据表 4 中的结果分析,当专家数取 10 时,实验表现出最好的效果以及具有相对较小的参数量。

Gumbel_Softmax 可以更好地进行采样,通过自适应学习温度系数控制平滑程度,为验证有效性通过对比 Softmax 进行实验,同时取专家数为 10,实验结果如表 5 所示。从表 5 可以看出,替换门控函数之后,精确率值提升了 1.09%,有着更好的分类效果。

表 5 门控函数对比实验

Table 5 Gating function comparison experiment

门控函数	精确率/%	召回率/%	F_1/%
Softmax	90.42	90.38	90.40
Gumbel_Softmax	91.51	91.49	91.48

3.5 分类错误样本分析

通过分析预测错误的样本发现,预测错误的样本很多都是语境相比较复杂以及可以被标注为多类别的。其中部分分类错误的样本如表 6 所示。

表 6 分类错误样本

Table 6 Classification error sample

输入	类别	预测
体育赛事与明星演唱会联动 为观众带来双重激情盛宴	7	9
广东高考满分作文 17 篇 一道语文题 13 万人吃鸭蛋	3	5
社会助力教育公平 偏远地区儿童获新知	4	3
糖价长期高位运行 果葡糖等替代品需求强劲	0	4

3.5 消融实验

为了验证模型中关键模块设计的合理性和有效性,同时确保模型在保持高精度的同时具有更低的参数计算量,本文进行了消融实验,相关结果如表 7 所示。

通过实验对比与分析,与原始的 ERNIE-RCNN 相比,本文模型在精确率上提升了 0.47% 且参数量约降为原来的 1/3,能够在保持较高分类效率的同时显著降低参数量。改进后的 Att-RCNN 能够更加有效地让输入元素关注到其上下文语境,并在识别性能方面表现得更为优异。

表 7 消融实验

Table 7 Ablation experiment

模型	精确率/%	召回率/%	F_1/%	参数量/10^6
ERNIE-RCNN	93.79	93.66	93.72	101.45
ERNIE-Att-RCNN	94.01	93.99	93.99	103.02
改进 ERNIE-RCNN	93.96	93.98	93.97	31.77
本文模型	94.26	94.21	94.23	33.34

4 结论

随着新媒体平台的不断演进,新闻传播速度正日益加快,但这也可能带来潜在的舆情问题。为了迅速遏制不良舆论蔓延,高效的新闻文本分类变得尤为关键。然而,当前新闻标题特征稀疏、信息处理难度大,并且大规模模型受限于延时需求等多重挑战。因此,提出了一种融合 ERNIE 和改进 RCNN 的混合专家网络新闻标题文本分类模型。实验结果表明,模型在保持较低参数量计算的同时,实现了对短文本的高效分类。这得益于采用了基于 ERNIE 预训练语言模型的向量表示提取方法,确保在初步特征提取阶段保留丰富的语义信息;引入稀疏门控混合专家网络策略显著减少了模型参数量计算;并通过元素级注意力门控机制实现了字词与上下文的紧密结合。

参 考 文 献

[1] 诸林云, 范菁, 曲金帅, 等. 基于 BERT 与多通道卷积神经网络的细粒度情感分类[J]. 科学技术与工程, 2023, 23(33): 14264-14270.
Zhu Linyun, Fan Jing, Qu Jinshuai, et al. Fine-grained sentiment classification based on BERT and multi-channel convolutional neural networks[J]. Science Technology and Engineering, 2023, 23(33): 14264-14270.

[2] 俞学豪, 赵子岩, 马应龙, 等. 基于 BR 和 GBDT 的电力信息通信客服系统多标签文本分类[J]. 电力系统自动化, 2021, 45(11): 144-151.
Yu Xuehao, Zhao Ziyan, Ma Yinglong, et al. Multi-label text classification of power information communication customer service system based on BR and GBDT [J]. Automation of Electric Power Systems, 2015, 45(11): 144-151.

[3] 孟祥福, 霍红锦, 张霄雁, 等. 个性化新闻推荐方法研究综述[J]. 计算机科学与探索, 2023, 17(12): 2840-2860.
Meng Xiangfu, Huo Hongjin, Zhang Xiaoyan, et al. Research review on personalized news recommendation methods [J]. Exploration of Computer Science and Technology, 2023, 17(12): 2840-2860.

[4] 华玮, 吴思洋, 俞超, 等. 面向网络舆情事件的多层次情感分歧度分析方法[J]. 数据分析与知识发现, 2023, 7(4): 16-31.
Hua Wei, Wu Siyang, Yu Chao, et al. Multi-level emotion divergence analysis method for network public opinion events [J]. Data Analysis and Knowledge Discovery, 2023, 7(4): 16-31.

[5] 王军, 冯孙铖, 程勇. 深度学习的轻量化神经网络结构研究综述[J]. 计算机工程, 2021, 47(8): 1-13.
Wang Jun, Feng Suncheng, Cheng Yong. A review of lightweight neural network structures for deep learning [J]. Computer Engineering, 201, 47(8): 1-13.

[6] Kim Y. Convolutional neural networks for sentence classification [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: EMNLP, 2014: 1746-1751.

[7] Liu P, Qiu X, Huang X. Recurrentneural network for text classification with multi-task learning [J]. arXiv preprint arXiv: 1605.05101, 2016.

[8] Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification [J]. AAAI Press, 2015. DOI: 10.1609/aaai.v29i1.9513.

[9] Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification[J]. arXiv preprint arXiv: 1607.01759, 2016.

[10] Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

[11] 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: DPCNN, 2017: 562-570.

[12] 余同瑞, 金冉, 韩晓臻, 等. 自然语言处理预训练模型的研究综述[J]. 计算机工程与应用, 2020, 56(23): 12-22.
Yu Tongrui, Jin Ran, Han Xiaozhen, et al. A review of research on pre-training models for natural language processing[J]. Computer Engineering and Applications, 2019, 56(23): 12-22.

[13] Devlin J, Chang M W, Lee K, et al. Bert: pretraining of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv: 1810.04805, 2018.

[14] Sun Y, Wang S, Li Y, et al. Ernie: enhanced representation through knowledge integration [J]. arXiv preprint arXiv: 1904.09223, 2019.

[15] 杨秀璋, 郭明镇, 候红涛, 等. 融合情感词典的改进 BiLSTM-CNN + Attention 情感分类算法[J]. 科学技术与工程, 2022, 22(20): 8761-8770.
Yang Xiuzhang, Guo Mingzhen, Hou Hongtao, et al. Improved BiLSTM-CNN + Attention emotion classification algorithm based onIntegrated emotion dictionary[J]. Science Technology and Engineering, 2019, 22(20): 8671-8770.

[16] 翟学明, 魏巍. 混合神经网络和条件随机场相结合的文本情感分析[J]. 智能系统学报, 2021, 16(2): 202-209.
Zhai Xueming, Wei Wei. Text sentiment analysis by combining hybrid neural networks and conditional random fields[J]. Journal of Intelligent Systems, 201, 16(2): 202-209.

[17] 陆晓蕾, 倪斌. 基于预训练语言模型的 BERT-CNN 多层级专利分类研究[J]. 中文信息学报, 2021, 35(11): 70-79.
Lu Xiaolei, Ni Bin. Research on BERT-CNN multi-level patent classification based onpre-trained language model[J]. Journal of Chinese Information Technology, 2019, 35(11): 70-79.

[18] Jacobs R A, Jordan M I, Nowlan S J, et al. Adaptive mixtures of local experts[J]. Neural Computation, 1991, 3(1): 79-87.

[19] Shazeer N, Mirhoseini A, Maziarz K, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer[J]. arXiv preprint arXiv: 1701.06538, 2017.

[20] Lepikhin D, Lee H J, Xu Y, et al. Gshard: scaling giant models with conditional computation and automatic sharding[J]. arXiv preprint arXiv: 2006.16668, 2020.

[21] Fedus W, Zoph B, Shazeer N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity[J]. The Journal of Machine Learning Research, 2022, 23(1): 5232-5270.

[22] Du N, Huang Y, Dai A M, et al. Glam: efficient scaling of language models with mixture-of-experts[J]. arXiv preprint arXiv:

2112.06905, 2021.
[23] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.
[24] Xue F, Shi Z, Wei F, et al. Go wider instead of deeper[J]. arXiv preprint arXiv: 2107.11817, 2021.
[25] Zuo S, Zhang Q, Liang C, et al. Moebert: from bert to mixture-of-experts *via* importance-guided adaptation[J]. arXiv preprint arXiv: 2204.07675, 2022.
[26] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7): 38-39.
[27] Jang E, Gu S, Poole B. Categorical reparameterization with gumbel-softmax[J]. arXiv preprint arXiv: 1611.01144, 2016.
[28] Zhang P, Xue J, Lan C, et al. EleAtt-RNN: adding attentiveness to neurons in recurrent neural networks[J]. IEEE Transactions on Image Processing, 2019, 29: 1061-1073.