

基于改进BERT集成结构的文本分类模型研究

刘振宗¹, 王超群², 陈乐², 陶永辉³, 王丹³

(1. 上海电力大学 计算机科学与技术学院, 上海 201300;
2. 航天智慧能源研究院; 3. 上海航天能源股份有限公司, 上海 201100)

摘要: 为了改进上下文文本特征提取问题, 捕获文本间语义关系, 并有效融合全局和局部信息, 提出BERT-Transformer-TextCNN并行文本分类模型。首先, 通过BERT模型对输入文本进行预处理, 得到文本特征向量。其次, 采用Transformer编码层提取文本特征向量的全局信息, 并在编码层引入L2正则化、残差连接和余弦相似度来克服过拟合、梯度消失以及向量长度的影响。再次, 利用TextCNN提取文本特征向量的局部信息, 并在此过程中引入残差连接、He初始化以及平均池化层, 以应对梯度消失和信息利用不充分的问题。最后, 将全局与局部信息合并, 通过Softmax分类器进行文本分类, 得到最终的分类结果。实验结果表明, 相比传统模型, 改进模型在THUCNews数据集上的准确率与F1值分别提高了12%和8%, 在IMDB数据集上的准确率与F1值分别提高了13%和8%, 证明了该模型在提取全局和局部信息、融合语义关系方面的有效性。

关键词: 文本分类; 特征提取; Transformer; TextCNN

DOI: 10.11907/rjdk.241141

开放科学(资源服务)标识码(OSID):



中图分类号: TP391.1

文献标识码: A

文章编号: 1672-7800(2025)005-0079-08

Research on Text Classification Model Based on Improved BERT Ensemble Structure

LIU Zhenzong¹, WANG Chaoqun², CHEN Le², TAO Yonghui³, WANG Dan³

(1. School of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 201300, China;
2. Aerospace Smart Energy Research Institute;
3. Shanghai Aerospace Energy Co., Ltd., Shanghai 201100, China)

Abstract: In order to improve the extraction of contextual text features, capture the semantic relationships between texts, and effectively fuse global and local information, the BERT-Transformer-TextCNN parallel text classification model is proposed. This model preprocesses the input text through the BERT model to obtain text feature vectors. The Transformer coding layer is used to extract the global information of text feature vectors, and L2 regularization, residual connection and cosine similarity are introduced in the coding layer to overcome the effects of overfitting, gradient disappearance and vector length. TextCNN is used to extract local information of text feature vectors, and in the process, residual connections, He initialization and average pooling layers are introduced to cope with the disappearance of gradients and insufficient information utilization. Finally, the global and local information are combined, and the text is classified through the Softmax classifier to obtain the final classification result. experiment result shows, Compared with the traditional model in the THUCNews data set, the improved model's accuracy increased by 12%, and its F1 value also increased by 8%. On the IMDB data set, the accuracy and F1 value increased by 13% and 8% respectively, proving the effectiveness of the model in extracting global and local information and integrating semantic relationships.

Key Words: text classification; feature extraction; Transformer; TextCNN

收稿日期: 2024-02-21

扫描二维码阅读全文:



作者简介: 刘振宗(1999-), 男, 上海电力大学计算机科学与技术学院硕士研究生, 研究方向为自然语言处理; 王超群(1979-), 男, 航天智慧能源研究院工程师、硕士生导师, 研究方向为自然语言处理以及机器学习; 陈乐(1996-), 男, 硕士, 航天智慧能源研究院高级工程师, 研究方向为自然语言处理; 陶永辉(1998-), 男, 硕士, 上海航天能源股份有限公司工程师, 研究方向为机器学习; 王丹(1983-), 女, 上海航天能源股份有限公司工程师, 研究方向为机器学习。

0 引言

随着数字化进程的推进,大量文本数据如社交媒体评论、新闻文章等不断涌现。文本分类作为理解和处理这些数据的关键,能够提高工作效率和决策质量。然而,文本分类具有复杂性,并且常常受到语义多义性的影响,因此传统方法的效果较差^[1]。近年来,随着深度学习的发展,基于神经网络的端到端学习成为解决文本分类问题的新方法^[2-3]。

在文本分类中,经典方法如词袋模型^[4]和词频-逆文档频率(Term Frequency-Inverse Document Frequency, TF-IDF)^[5]可捕捉词汇信息,但忽略了语义关系;卷积神经网络(Convolutional Neural Network, CNN)^[6]和文本卷积神经网络(Text Convolutional Neural Network, TextCNN)^[7]能捕获语义关系,但忽略了上下文依赖性,而且存在梯度消失和信息利用不充分的问题;循环神经网络(Recurrent Neural Network, RNN)^[8]和长短期记忆网络(Long Short-Term Memory, LSTM)^[9]能处理上下文依赖性问题,但针对大规模文本速度较慢;Transformer架构^[10]虽然在处理全局依赖性以及上下文关系建模上表现出色,但其存在过拟合、梯度消失以及受向量长度影响的问题;预训练的双向编码器表示(Bidirectional Encoder Representations from Transformers, BERT)^[11]预训练表现出色,但在处理较长的文本时需要进行截断或分段处理,可能会导致上下文信息丢失。

为了改进上下文文本特征提取问题、捕获文本间语义关系,并有效融合全局和局部信息,本文提出BERT-Transformer-TextCNN并行文本分类模型^[12]。该模型综合运用BERT预处理^[13]、Transformer编码层处理全局信息^[14],以及TextCNN提取局部信息^[15]。

为应对Transformer处理全局信息时出现的过拟合和梯度消失问题,在编码层的前馈神经网络中引入正则化^[16]和残差连接^[17],同时在多头注意力机制^[18]中引入余弦相似度^[19]。在TextCNN处理局部信息时,针对梯度消失和信息利用不充分的问题,在卷积层引入残差连接和He初始化^[20],在池化层则采用平均-最大池化替代原有的最大池化^[21]。

1 相关技术

1.1 模型介绍

本文模型将原始文本数据分段后输入到BERT模型中进行文本预处理,并将数据转换为词嵌入矩阵。然后,将词嵌入矩阵分别传入Transformer编码层和TextCNN层,得到全局特征信息与局部特征信息。通过特征融合整合成一个包含全局和局部信息的特征向量,再通过全连接层和Softmax分类器进行文本分类。

本文采用将长文本数据分割成短段的方式,以规避输入长度限制。虽然该方式可能导致信息不充分,但通过并行使用Transformer编码层和TextCNN模型,综合全局和局部信息,能够更全面地理解文本,并降低长文本截断带来的影响。为确保各个模型结构的独立性,Transformer编码层和TextCNN模型的参数相互独立。这有助于每个结构在学习任务相关特征时不受其他结构影响,保持了模型的灵活性。本文结合两种方式,平衡了文本输入和模型性能的需求。

BERT-Transformer-TextCNN结构如图1所示。

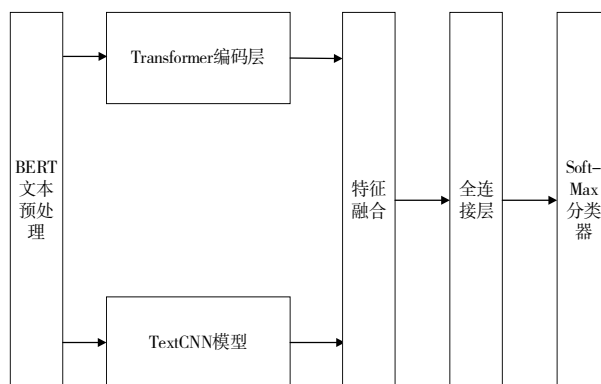


Fig. 1 BERT-Transformer-TextCNN model structure

图1 BERT-Transformer-TextCNN模型结构

1.2 Transformer 编码层

Transformer 编码层是Transformer模型中的一个重要组成部分,在对输入序列进行特征提取的基础上,进一步实现对输入序列的深度表示学习,以期后续模型的信息处理与任务执行奠定基础。其由两个关键组件组成:前馈神经网络与多头注意力机制。Transformer 编码层结构如图2所示。

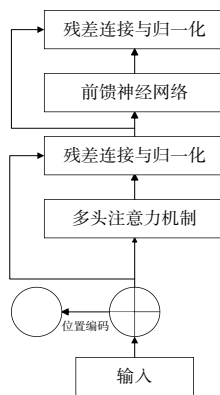


Fig. 2 Transformer coding layer structure

图2 Transformer编码层结构

1.2.1 前馈神经网络

前馈神经网络包括输入层、隐含层和输出层。在每个神经元中,通过线性组合将输入特征与相应权重相乘并求和,得到神经元的输出;通过对输入特征进行线性组合和非线性映射,以学习数据的表示和关系。线性组合公式如下:

$$\text{line}(x_i) = \text{ReLU}(w_i * x_i + b_i) \quad (1)$$

$$z = \text{ReLU} * \sum_{i=1}^{n-1} [w_{i+1} * \text{line}(x_i) + b_{i+1}] \quad (2)$$

其中, z 是神经元的线性组合, w_i 与 w_{i+1} 是权重, $\text{line}(x)$ 是一个全连接的隐含层, x_i 与 x_{i+1} 是输入特征, b_i 与 b_{i+1} 是偏置项, ReLU 是激活函数。

经过线性组合后,通过 ReLU 激活函数进行非线性映射,从而使神经网络能够适应更复杂的数据。 ReLU 激活函数公式如下:

$$\text{ReLU}(x) = \max(0, x) \quad (3)$$

其中, x 是输入值,函数 $\text{ReLU}(x)$ 对其进行运算, $\text{ReLU}(x)$ 的输出值是 x 与 0 之间的极大值。

1.2.2 多头注意力机制

多头注意力机制是一种深度学习模型组件,允许模型同时关注输入序列的不同部分,以学习多种关系和模式。在文本处理中,该机制能够使模型从不同角度理解文本,提升处理效果。多头注意力机制通过多次转换获取查询、键、值向量,计算注意力得分,并将得分与值向量相乘,最后将多个头的结果结合起来。涉及到的公式如下:

$$\begin{aligned} Q &= h * W_q \\ K &= h * W_k \\ V &= h * W_v \end{aligned} \quad (4)$$

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i) = \text{Softmax}\left(\frac{Q_i * K_i^T}{\sqrt{d_k}}\right) V_i \quad (5)$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_i) W_o \quad (6)$$

其中, h 是输入序列, Q 是查询向量, K 是键向量, V 是值向量, W_q 、 W_k 、 W_v 是可学习的权重矩阵, head_i 是每个头注意力计算的结果; Softmax 通过归一化得到注意力权重; MultiHead 是多头注意力计算结果,与 W_o 权重矩阵结合,用于线性变换。

1.3 TextCNN 模型

TextCNN 模型是一种用于文本分类任务的卷积神经网络模型,通过使用一维卷积核在文本数据上进行滑动,自动提取文本的局部特征。

TextCNN 的卷积层是网络的核心组成部分,卷积层使用多个不同大小的卷积核来捕捉不同长度的局部特征。具体公式如下:

$$h_i = f(w * x_{i:i+k-1} + b) \quad (7)$$

其中, h_i 是卷积层输出特征的第 i 个元素; w 是卷积核; $x_{i:i+k-1}$ 是输入特征向量的一部分,包含 k 个连续元素; b 是偏置项; f 是 ReLU 激活函数。

TextCNN 池化层对卷积层的特征向量进行池化操作,以提取出最重要的特征信息。首先采用最大池化的方式进行特征聚合,在每个滑动窗口产生的特征向量中筛选出一个最大特征,然后将这些特征拼接起来构成向量表示。最大池化公式为:

$$\text{Max} = \text{MaxPooling}(X) = \max(X) \quad (8)$$

其中, X 是一维输入数据, $\max(X)$ 是输入中的最大值,最终得到最大的特征。

2 改进的 Transformer 编码层模型

Transformer 编码层是用于处理输入序列特征提取和表示学习的模型,由两个关键组件组成:前馈神经网络与多头自注意力机制。为了降低过拟合风险、改善梯度消失以及受向量长度影响的问题,本文通过改进前馈神经网络与多头自注意力机制来提升模型性能。

2.1 融合 L2 正则化与残差连接的前馈神经网络

前馈神经网络具有一个输入层、一个隐含层和一个输出层。为了更有效地处理数据,本文采用 6 层隐含层。此时面临两个主要问题:过拟合与梯度消失。

因此,本文通过引入正则化方法、限制权重大小来减少过拟合的风险,提高泛化性能;通过引入残差连接,以解决梯度消失问题。

2.1.1 过拟合问题优化方案

本文通过正则化优化公式(式(2))中的 z 添加正则化项(根据表(1)的实验结果,选择 L2 正则化),在训练过程中限制权重大小,并防止过拟合^[17]。L2 正则化公式如下:

$$L2 = \lambda * \sum_{i=1}^n w_i^2 \quad (9)$$

其中, λ 是 L2 正则化参数, w_i 是权重的第 i 个分量, w_i^2 是权重的平方, n 是权重的数量。L2 正则化项受到正则化强度 λ 控制,作用机制为对模型权重的平方和施加惩罚。

L2 正则化的作用是迫使权重保持较小的值以降低模型复杂性,从而降低过拟合风险。结合式(1)、式(8),修改后的公式如下:

$$Z = z + L2 \quad (10)$$

模型训练时,式(10)中的 Z 考虑到 L2 正则化项,鼓励权重保持较小的值,以降低过拟合风险。

2.1.2 梯度消失问题解决方案

在前馈神经网络内部,距离输出层较远的隐含层之间会出现梯度消失问题。因此,本文在较远的隐含层之间引入残差连接来解决此问题。

残差连接的核心思想是通过引入跳跃连接,允许神经网络更轻松地学习恒等映射,使梯度更容易传播,从而降低梯度消失风险^[18]。残差连接架构如图 3 所示。

具体公式如下:

$$y = x + F(x) \quad (11)$$

$$F(x) = W_2 * \text{ReLU}(W_1 * x + b_1) + b_2 \quad (12)$$

其中, x 是输入特征向量; $F(x)$ 是输入特征进行变换得到的残差,包括一系列网络层和操作; y 是经过残差连接后的输出; W_1 和 W_2 是权重参数; b_1 和 b_2 是偏置项; ReLU 是整流线性单元激活函数。

通过式(11)、式(12),原始输入 x 与残差块的输出 F

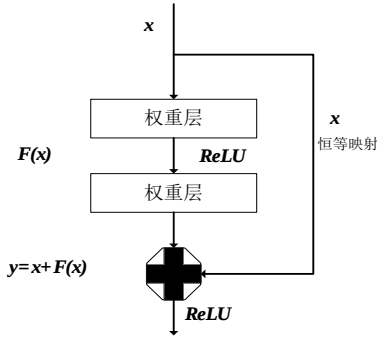


Fig. 3 Residual connection architecture

图3 残差连接架构

(x)相加,得到最终输出。该方式使模型可以同时学习到输入到输出的映射以及输入与输出之间的残差信息,从而缓解梯度消失问题。

改进后的前馈神经网络结构如图4所示。

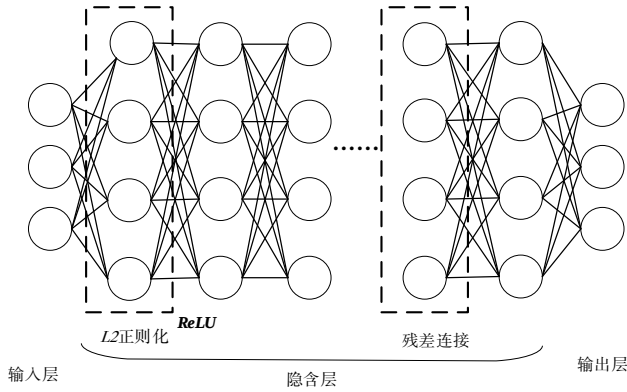


Fig. 4 Improved feedforward neural network architecture

图4 改进后的前馈神经网络结构

2.2 融合余弦相似度的多头注意力机制

多头注意力机制运用点积作为相似度函数,但其对查询向量和键向量的长度比较敏感。因此,点积的值会受到向量长度的影响。

本文采用余弦相似度替代点积作为相似度函数^[20]。由于余弦相似度衡量的是向量夹角而非长度,因此使用余弦相似度可以消除向量长度对相似度的影响。该方式可以对多头注意力机制进行改进,尤其是当输入向量长度差异较大时。余弦相似度计算方式如下:

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \quad (13)$$

其中, A 、 B 分别是查询向量和键向量, $\cdot$ 是向量点积, $\|A\|$ 、 $\|B\|$ 分别是查询向量的长度。

余弦相似度使用键之间的夹角余弦来计算相似度,从而消除向量长度的影响,且仅考虑向量方向,更符合语义空间的特性。将余弦相似度结合公式(14),注意力机制公式变为:

$$\text{Attention}(Q, K, V) = \frac{\text{Softmax}(\text{Cosine Similarity}(Q, K))V}{\text{Softmax}(\text{Cosine Similarity}(Q, K))} \quad (14)$$

其中, Q 是查询向量, K 是键向量, V 是值向量, Soft-

max 通过归一化将原始分数转化为概率分布。

改进后的部分多头注意力机制结构如图5所示。

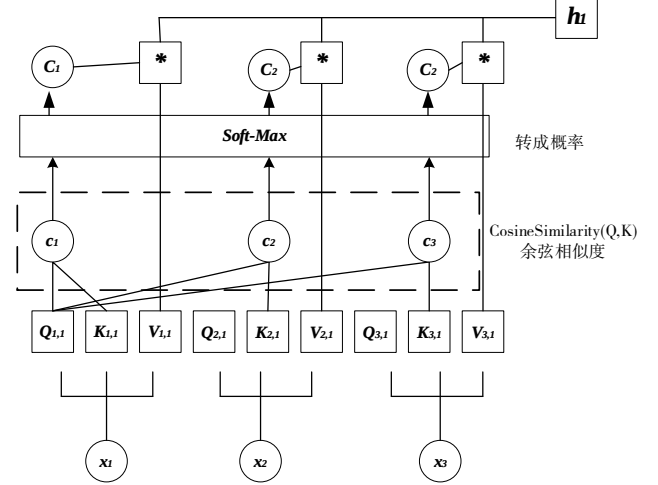


Fig. 5 Improved partial multi-head attention mechanism architecture

图5 改进后的部分多头注意力机制结构

2.3 改进后的Transformer编码层

在改进后的Transformer编码层中,首先,数据经过改进的多头注意力机制,计算每个位置对序列中其他位置的注意力权重。其次,执行加权和操作,将上下文信息与输入相结合,并通过残差连接保持信息传递。再次,通过进行层标准化以稳定梯度,数据通过改进的前馈神经网络进行非线性变换。最后,经过残差连接和层标准化对数据进一步处理^[18]。上述步骤使编码层能够从输入序列中提取特征和上下文信息。改进后的Transformer编码层如图6所示。

3 改进的TextCNN模型

TextCNN模型由卷积层、池化层等多个部分组成,为了解决梯度消失以及信息利用不充分的问题,本文将TextCNN中的卷积层和池化层进行改进。

3.1 融合残差连接与He初始化的卷积层

为了更好地提取局部特征,本文引入4层卷积层(在交叉实验中,具有4层卷积层的模型效果最佳)。为解决可能出现的梯度消失问题,在卷积层后引入残差连接以及He初始化。

在每个卷积层之后加入残差块,残差连接旨在解决梯度消失问题。在式(10)中,残差块采用一种特殊的方式,将卷积输出与输入相加,而不是直接替代输入。通过式(11)产生残差信息,再通过ReLU激活函数进行非线性处理来缓解梯度消失问题。具体过程如图3所示。

在卷积层中,ReLU激活函数在负半轴上的导数为零会导致梯度消失问题,使得网络难以更新权重。

本文引入He初始化,通过调整权重的标准差,使权重更好地匹配ReLU激活函数的性质,有助于减少梯度消失问题^[21]。He初始化后的权重为:

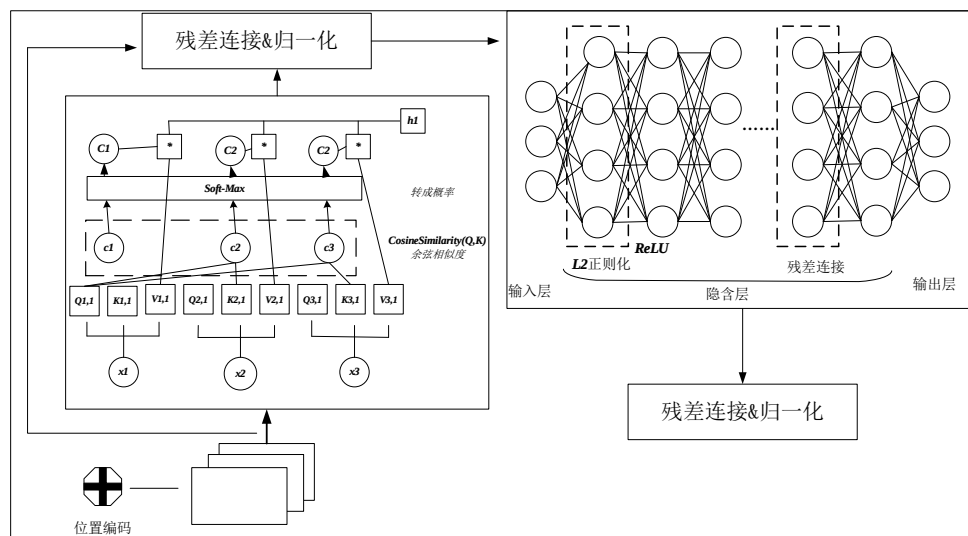


Fig. 6 Improved Transformer coding layer

图6 改进后的Transformer编码层

$$W \sim N(0, \sqrt{2/n}) \quad (15)$$

其中, W 表示网络层的权重矩阵, $N(0, \sqrt{2/n})$ 表示均值为0、方差为 $\sqrt{2/n}$ 的高斯分布。

对于卷积层中的每个权重参数, He 初始化通过保持合适的权重范围, 确保在 ReLU 激活函数的正区域内梯度的平均值保持稳定。残差连接与 He 初始化共同作用, 以缓解卷积层梯度消失问题。

3.2 集成平均—最大池化的池化层

最大池化仅选择池化窗口中的最大值, 而丢弃其他次要的特征值以及有用信息, 导致信息的不完全利用。因此, 本文引入平均池化, 将平均池化与最大池化相结合以解决该问题。

在平均最大池化中, 针对每个特征映射, 先计算所有元素的平均值和最大值, 再将两个值合并以获得最终的特征表示。平均池化公式如下^[22]:

$$Avg = AvgPooling(X) = \frac{1}{N} \sum_{i=1}^N X_i \quad (16)$$

其中, X 是池化窗口内的输入数据, N 是窗口内的元素数量, X_i 是窗口内的第 i 个元素。

结合式(8)、式(16), 平均—最大池化的公式如下:

$$AvgMaxPooling(X) = Concat(Max, Avg) \quad (17)$$

$Concat(Max, Avg)$ 表示将最大池化与平均池化的信息连接起来组成更丰富的语义信息, 得到最终的特征表示。

在平均最大池化中, 首先, 通过最大池化计算池化窗口内元素的最大值, 同时通过平均池化计算元素的平均值。其次, 将最大值与平均值连接, 得到一个综合的特征表示。这一优化旨在克服最大池化可能导致信息不完全利用问题, 在保留最显著特征的同时, 考虑了窗口内所有元素的平均特征, 形成更全面的特征表示。

3.3 改进后的TextCNN架构

本文TextCNN架构包括4层卷积层, 每个卷积层后都引入了残差连接以改进梯度传播和特征学习。在最后一个卷积层之后使用平均—最大池化层, 以进一步减小特征尺寸。该结构允许网络学习复杂的特征表示, 同时保持梯度的有效传播。具体结构如图7所示。

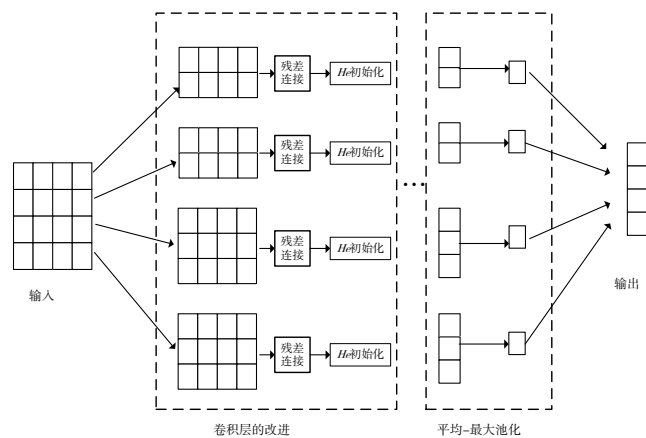


Fig. 7 Improved TextCNN structure

图7 改进后的TextCNN结构

4 实验结果与分析

4.1 实验数据集

THUCNews 数据集是一个中文新闻文本分类数据集, 包含了10种类别的新闻数据, 本文采用了其中的5万条数据。数据集划分为: 训练集4万条(4折交叉验证), 测试集1万条。

IMDB 数据集中有5万条电影评论, 分为正向和负向两种情感倾向。数据集划分为: 训练集4万条(4折交叉验证), 测试集1万条。

4.2 实验评价标准

本文采用准确率、精确率、召回率和 $F1$ 值作为文本分类的评估标准。

TP : 样本为正, 预测为正。 FP : 样本为负, 预测为正 (假阳性)。 TN : 样本为负, 预测为负。 FN : 样本为正, 预测为负 (假阴性)。

准确率 ($Accuracy$): 模型正确分类的样本数量与总样本数量之比。公式如下:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (18)$$

精确率 ($Precision$): 衡量分类器正确预测为正类别的样本在所有预测为正类别样本中的比例。公式如下:

$$Precision = \frac{TP}{TP + FP} \quad (19)$$

召回率 ($Recall$): 衡量分类器正确预测为正类别的样本在所有实际正类别样本中的比例。公式如下:

$$Recall = \frac{TP}{TP + FN} \quad (20)$$

$F1$ 值 ($F1$ Score): 精确率和召回率的调和平均, 综合考虑了两者的表现。公式如下:

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (21)$$

4.3 实验过程与结果分析

4.3.1 模型参数交叉验证

为防止数据泄露, 本文在 4 万条训练集上进行 4 折交叉验证。本文选取了两个重要参数, 分别为模型中 Transformer 编码层内前馈神经网络的维度以及 TextCNN 中卷积层的维度。同时固定其他参数, 并选择准确率作为验证集的评估标准。

验证结果如图 8 所示。随着前馈神经网络隐含层的增加, 模型性能呈上升趋势, 达到 6 层时性能最佳。若继续增加层数会导致模型性能下降, 并可能伴随计算成本增加和过拟合现象出现。前馈神经网络的层数在 1 层与 6 层时, TextCNN 卷积层均在 4 层时达到性能峰值, 若继续增加层数则导致性能下降。固定卷积层并增加前馈神经网络

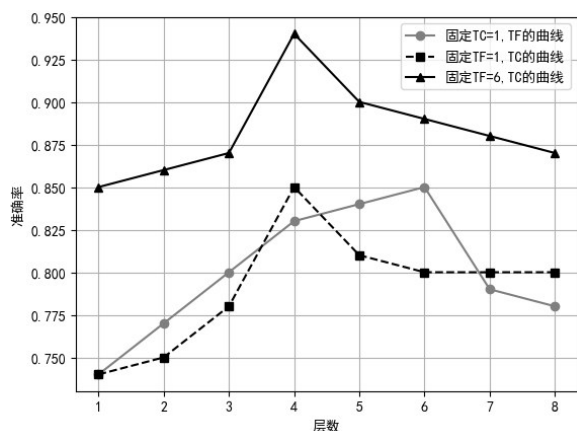


Fig. 8 Cross-validated line chart

图 8 交叉验证折线图

层数, 在 6 层隐含层时性能达到最佳。考虑综合性能和计算成本, 本文最终选择 6 层隐含层和 4 层卷积层的模型配置 (模型在两个数据集上的实验结果一致)。

4.3.2 Transformer 编码层对比实验

在本次对比实验中, 对比模型如下:

Model A: 使用 Transformer 编码层。

Model B: 使用 L1 正则化训练和残差连接改进前馈神经网络的 Transformer 编码层。

Model C: 使用 L2 正则化和残差连接改进前馈神经网络的 Transformer 编码层。

Model D: 使用余弦相似度改进的多头注意力机制的 Transformer 编码层。

Model E: 本文改进的 Transformer 编码层模型。

对比结果如表 1 所示。

Table 1 Comparison results of Transformer models

表 1 Transformer 模型对比结果

数据集	模型	准确率	精确率	召回率	F1
THUCNews 数据集	Model A	0.81	0.82	0.86	0.84
	Model B	0.83	0.84	0.87	0.85
	Model C	0.85	0.87	0.89	0.88
	Model D	0.83	0.84	0.86	0.85
	Model E	0.88	0.89	0.90	0.89
IMDB 数据集	Model A	0.82	0.84	0.86	0.84
	Model B	0.84	0.85	0.87	0.86
	Model C	0.85	0.86	0.87	0.86
	Model D	0.83	0.88	0.86	0.87
	Model E	0.90	0.89	0.91	0.90

在表 1 的两个数据集中, Model A 与其他模型相比, 准确率和 F1 值均低于其他模型, 这可能是由过拟合与梯度消失问题导致的。观察加入了 L1 正则化与残差连接的 Model B, 其在 THUCNews 数据集中的准确率及 F1 值分别有 0.02 和 0.01 的提升, 在 IMDB 数据集中的准确率及 F1 值均有 0.02 的提升, 可确定 Model A 存在过拟合以及梯度消失问题, 并且证明了本文改进方式的有效性。与 Model B 相比, Model C 的准确率和 F1 值再次得到提升。说明 L1 正则化使某些特征权重为 0, 导致某些数据特征消失。但与 L2 正则化相比, 其效果欠佳。此结论在 IMDB 数据集实验结果中也得到了验证。因此, 本文选择了 L2 正则化。对比两个数据集中的 Model A 与 Model D, 可以看出采用余弦相似度代替点积的计算方式, 降低了向量长度对点积带来的影响, 提高了模型性能。本文 Model E 结合了 Model C 和 Model D 两个改进方案, 其在 THUCNews 数据集中的准确率及 F1 值分别有 0.07 与 0.05 的提升, 在 IMDB 数据集中的准确率以及 F1 值分别有 0.08 和 0.06 的提升, 证明了本文改进方式的有效性。

4.3.3 TextCNN 对比实验

实验对比模型如下:

T1: TextCNN 模型。

T1.1: 使用 4 层卷积层的 TextCNN 模型。

T2:使用残差连接以及 He 初始化改进的 TextCNN 模型。

T3:使用平均—最大池化改进的 TextCNN 模型。

T4:本文改进的 TextCNN 模型。

对比结果如表 2 所示。

在表 2 的两个数据集中,T1 模型与其他模型相比,其准确率与 F1 值均是最低的。T1.1 增加了 4 层卷积层来增强其局部处理信息的能力,模型性能在两个数据集中均有小幅提升,可能出现了梯度消失问题。为此,T2 引入残差连接与 He 初始化,其在 THUCNews 数据集中的准确率及 F1 值均有 0.02 的提升,在 IMDB 数据集中的准确率及 F1 值分别有 0.03 和 0.01 的提升,证明了本文改进方法的有效性。T1.1 的最大池化层导致部分信息利用不充分,T3 则运用平均—最大池化层使模型性能得以提升。T4 结合了 T2 与 T3 的两种改进方式,其在 THUCNews 数据集中的准确率及 F1 值分别有 0.07 和 0.05 的提升,在 IMDB 数据集中的准确率及 F1 值分别有 0.10 和 0.06 的提升,证明了本文改进 TextCNN 模型的有效性。

Table 2 Comparison results of TextCNN models

表 2 TextCNN 模型对比结果

数据集	模型	准确率	精确率	召回率	F1
THUCNews 数据集	T1	0.80	0.82	0.84	0.83
	T1.1	0.81	0.83	0.84	0.83
	T2	0.83	0.85	0.86	0.85
	T3	0.84	0.84	0.85	0.84
	T4	0.87	0.89	0.88	0.88
IMDB 数据集	T1	0.81	0.83	0.84	0.83
	T1.1	0.82	0.84	0.86	0.85
	T2	0.85	0.86	0.87	0.86
	T3	0.85	0.87	0.86	0.86
	T4	0.90	0.90	0.89	0.89

4.3.4 本文模型对比实验

实验对比模型如下:

Model 1:BERT 模型。

Model 2:BERT-Transformer 模型。

Model 3:BERT-TextCNN 模型。

Model 4:BERT-Transformer-TextCNN 模型。

Model 5:BERT-改进的 Transformer-TextCNN 模型。

Model 6:BERT-Transformer-改进的 TextCNN 模型。

Model 7:本文的改进模型。

对比结果如表 3 所示。

在两个数据集中,以 BERT 模型作为基准模型进行文本数据预处理。实验结果表明,Model 1、Model 2、Model 3 加入 BERT 文本预处理之后性能均有所提升,因此选择 BERT 模型进行预处理。Model 4 采用 Transformer 编码层与 TextCNN 模型并行处理数据,其在 THUCNews 数据集中的准确率及 F1 值分别有 0.08 和 0.05 的提升,在 IMDB 数据集中的准确率与 F1 值分别有 0.08 和 0.04 的提升。由此证明,本文模型架构具有可行性。Model 5 将 Model 4 原有的

Table 3 Comparison results of the proposed model

表 3 本文模型对比结果

数据集	模型	准确率	精确率	召回率	F1
THUCNews 数据集	Model 1	0.82	0.86	0.85	0.85
	Model 2	0.87	0.89	0.88	0.88
	Model 3	0.86	0.88	0.87	0.87
	Model 4	0.90	0.89	0.89	0.90
	Model 5	0.91	0.92	0.91	0.91
	Model 6	0.92	0.91	0.92	0.91
	Model 7	0.94	0.93	0.93	0.93
IMDB 数据集	Model 1	0.83	0.84	0.88	0.85
	Model 2	0.86	0.88	0.89	0.88
	Model 3	0.85	0.87	0.86	0.86
	Model 4	0.91	0.90	0.89	0.89
	Model 5	0.92	0.91	0.93	0.92
	Model 6	0.93	0.90	0.92	0.91
	Model 7	0.96	0.94	0.93	0.93

Transformer 编码层替换为本文改进的 Transformer 编码层,性能获得了提升。Model 6 替换为改进的 TextCNN 模型,性能同样得以提升。因此,本文模型 Model 7 结合两种改进方式,在 THUCNews 数据集中的准确率及 F1 值分别有 0.12 和 0.08 的提升,在 IMDB 数据集中的准确率及 F1 值分别有 0.13 和 0.08 的提升,证明了该模型在提取全局和局部信息、融合语义关系方面的有效性。

5 结语

本文提出基于 BERT-Transformer-TextCNN 并行文本分类模型的改进方法。通过对 Transformer 编码层底层的前馈神经网络以及多头注意力机制的改进,有效降低了过拟合风险、缓解了梯度消失问题,并消除了向量长度对语义相似度计算的影响。同时,本文对于 TextCNN 卷积层及池化层的改进,有效解决了梯度消失以及信息利用不充分的问题。实验结果表明,本文的并行文本分类模型取得了很好的效果。未来的研究方向将会集中于对 BERT 模型的进一步优化,以适应不同领域和多样化的文本数据,提升其文本预处理能力,进一步增强文本分类性能。

参考文献:

[1] FU W J, YANG D, MA H M, et al. Short text classification method based on BTM and BERT [J]. Computer Engineering and Design, 2022, 43 (12): 3421-3427.
付文杰,杨迪,马红明,等.融合 BTM 和 BERT 的短文本分类方法[J].计算机工程与设计,2022,43(12):3421-3427.

[2] GAN Y T, AN J Y, XU X. Survey of short text classification methods based on deep learning [J]. Computer Engineering and Applications, 2023, 59 (4): 43-53.
涂亚婷,安建业,徐雪.基于深度学习的短文本分类方法研究综述[J].计算机工程与应用,2023,59(4):43-53.

[3] LIANG S P, DOU M M. Text classification method based on hybrid neural network [J]. Computer Engineering and Design, 2022, 43(2): 573-579.
梁顺攀,豆明明.基于混合神经网络的文本分类方法[J].计算机工程

- 与设计,2022,43(2): 573-579.
- [4] HUANG C M, WANG S L. Research on short text classification based on bag of words and TF-IDF[J]. Software Engineering, 2020, 23(3): 1-3.
黄春梅,王松磊. 基于词袋模型和TF-IDF的短文本分类研究[J]. 软件工程, 2020, 23(3): 1-3.
- [5] JING L, HE T T. Chinese text classification model based on improved TF-IDF and ABLCNN[J]. Computer Science, 2021, 48(S2): 170-175, 190.
景丽,何婷婷. 基于改进TF-IDF和ABLCNN的中文文本分类模型[J]. 计算机科学, 2021, 48(S2): 170-175, 190.
- [6] LI C F, MA K. Electronic medical record text classification based on attention mechanism combined with CNN-BiLSTM[J]. Science Technology and Engineering, 2022, 22(6): 2363-2370.
李超凡,马凯. 基于注意力机制结合CNN-BiLSTM模型的电子病历文本分类[J]. 科学技术与工程, 2022, 22(6): 2363-2370.
- [7] BAO T, LUO R, GUO T, et al. Agricultural question classification model based on BERT word vector and TextCNN[J]. Journal of Southern Agriculture, 2022, 53(7): 2068-2076.
鲍彤,罗瑞,郭婷,等. 基于BERT字向量和TextCNN的农业问句分类模型分析[J]. 南方农业学报, 2022, 53(7): 2068-2076.
- [8] LI Q X, LIAO W, MENG J W. Dual-channel DAC-RNN text classification model based on attention mechanism[J]. Computer Engineering and Applications, 2022, 58(16): 157-163.
李启行,廖薇,孟静雯. 基于注意力机制的双通道DAC-RNN文本分类模型[J]. 计算机工程与应用, 2022, 58(16): 157-163.
- [9] SUN L C, HUANG R C. News text classification model fuses LSTM and attention mechanism[J]. Transducer and Microsystem Technologies, 2022, 41(9): 38-41.
孙刘成,黄润才. 融合LSTM和注意力机制的新闻文本分类模型[J]. 传感器与微系统, 2022, 41(9): 38-41.
- [10] ZHI G, ZHONG X Y, WANG X, et al. Transformer-based sequence generation for multi-label text classification[J]. Journal of Shanxi University (Natural Science Edition), 2023, 46(1): 10-19.
支港,钟学燕,王欣,等. 基于Transformer的序列生成多标签文本分类[J]. 山西大学学报(自然科学版), 2023, 46(1): 10-19.
- [11] YANG X R, ZHAO S W, ZHANG R X, et al. BiLSTM-Attention text classification of improved BERT word vectol[J]. Transducer and Microsystem Technologies, 2023, 42(10): 160-164.
杨兴锐,赵寿为,张如学,等. 改进BERT词向量的BiLSTM-Attention文本分类模型[J]. 传感器与微系统, 2023, 42(10): 160-164.
- [12] GE F Y, LEI J S, TANG X L. An integrated model integrating LC-Transformer XL text classification[J]. Computer Applications and Software, 2023, 40(6): 118-123, 132.
葛夫勇,雷景生,唐小岚. 融合LC-Transformer XL文本分类的集成模型[J]. 计算机应用与软件, 2023, 40(6): 118-123, 132.
- [13] LIU B, PU Y F. BERT-based approach for long document classification[J]. Journal of Sichuan University (Natural Science Edition), 2023, 60(2): 81-88.
刘博,蒲亦非. 基于BERT的长文本分类方法[J]. 四川大学学报(自然科学版), 2023, 60(2): 81-88.
- [14] ZHANG D Y, CUI Z J. Identifying noun metaphors with Transformer and BERT[J]. Data Analysis and Knowledge Discovery, 2020, 4(4): 100-108.
张冬瑜,崔紫娟. 基于Transformer和BERT的名词隐喻识别[J]. 数据分析与知识发现, 2020, 4(4): 100-108.
- [15] LI T F, SHENG L, WU D. Study on text classification method of BERT-TECNN model[J]. Computer Engineering and Applications, 2021, 57(18): 186-193.
李铁飞,生龙,吴迪. BERT-TECNN模型的文本分类方法研究[J]. 计算机工程与应用, 2021, 57(18): 186-193.
- [16] HASTIE T, TIBSHIRANI R, FRIEDMAN J. The elements of statistical learning[M]. Berlin: Springer, 2016.
- [17] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
- [18] LI C F, MA K. Text classification model based on multi-channel attention mechanism[J]. Microelectronics & Computer, 2022, 39(4): 33-40.
李超凡,马凯. 基于多通道注意力机制的文本分类模型[J]. 微电子学与计算机, 2022, 39(4): 33-40.
- [19] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.
- [20] HE K M, ZHANG X, REN S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1026-1034.
- [21] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2921-2929.

(责任编辑:黄健 何佳慧)