文章编号：2095-2163(2022)07-0156-04 中图分类号：TP391.1 文献标志码：A

基于改进 BERT 和多阶段 TCN 的短文本分类

范明炜，张云华
（浙江理工大学 信息学院，杭州 310018）

摘 要： 短文本分类是自然语言处理中一项具有挑战性的任务。目前利用外部知识处理短文本稀疏性和歧义性的传统方法取得了较好的效果，基于 RNN 的方法在并行化方面表现不佳，导致效率较低。基于 CNN 的方法可以捕捉局部特征，但由于忽略上下文相关的特征以及一词多义等问题，准确率还有待提高。针对以上问题，提出基于 CNN 与 TCN 相结合，并加入权重优化与注意力机制的短文本分类模型。使用 Probase 作为外部知识来丰富语义表示，解决特征稀疏和语义不足的问题，通过 BERT 训练词向量，引入词性和词语权重对词向量优化，将优化的词向量作为输入层信息，经过 CNN 和 TCN 相结合的方法提取特征，最后结合注意力机制拼接向量，突显关键信息，获得文本特征表示。实验表明，与几种常用的基于 CNN 和 RNN 的短文本分类方法相比，该方法在短文本分类中更加准确高效。

关键词： 短文本分类；BERT；注意力机制；TCN

Short text classification based on improved BERT and multi-stage TCN

FAN Mingwei, ZHANG Yunhua
(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

【Abstract】 Short text classification is a challenging task in natural language processing. At present, traditional methods using external knowledge to deal with the sparsity and ambiguity of short texts have achieved good results, while RNN-based methods perform poorly in parallelization and results in low efficiency. The CNN-based method can capture local features, but due to ignoring context-related features and polysemy, the accuracy remains to be improved. In response to the above problems, a short text classification model is proposed combining weight optimization and attention based on CNN and TCN. Probase is taken as external knowledge to enrich semantic representation and solve the problem of sparse features and insufficient semantics. BERT is trained with word vectors and part of speech and word weights is introduced to optimize word vectors which is used as input layer information. The combined method extracts features, and finally combines the attention mechanism to splicing vectors to highlight key information and obtain text feature representation. Experiments show that this method is more accurate and efficient in short text classification compared with several commonly used CNN and RNN-based short text classification methods.

【Key words】 short text classification; BERT; attention mechanism; TCN

0 引 言

随着互联网的飞速发展，评论、朋友圈等信息传播中产生了大量的短文本，短文本分类任务已经成为自然语言处理领域的重要研究热点之一[1-2]。

文本中单词往往有多层意思，虽然利用外部知识[3]对文本进行特征扩展，丰富语义关系，在消除文本稀疏性和歧义性有较好的效果，但忽略了单词的位置不同对结果的影响。Google 提出 BERT 模型，采用 MLM 对双向的 Transformers 进行预训练，以生成深层的双向语言表征，但并没有体现出每个词语对整个文本的重要程度。针对短文本分类，传统的机器学习算法，如决策树模型、空间向量模型和支持向量机模型等，主要解决词汇层面的匹配问题，在分类短文本时忽略了词与词之间潜在的语义相关性，导致向量空间稀疏，处理高维数据和泛化能力有所欠缺。

近年来，深度学习在计算机视觉，自然语言处理等领域获得了不错的效果。因此，基于深度学习的短文本分类算法[4]开始受到关注。例如，卷积神经网络(CNN)[5]、Liu 等人[6]提出的循环神经网络(RNN)以及注意力机制等。基于 CNN 的方法虽然可以捕捉局部特征，但忽略了单词之间的顺序和关系，容易丢失之前的信息。而 RNN 是包含循环的网络，允许信息的持久化，但当相关信息和当前预测位置之间的间隔不断增大时，RNN 会丧失学习间隔信

作者简介：范明炜(1996-)，男，硕士研究生，主要研究方向：软件工程技术；张云华(1965-)，男，博士，研究员，主要研究方向：软件工程、系统仿真、智能信息处理。
通讯作者：张云华 Email：605498519@qq.com

收稿日期：2022-01-18

息的能力。LSTM 是 RNN 的一种特殊类型，可以学习长期依赖信息，但由于网络一次只读取解析输入文本中的一个单词或字符，必须等前一个单词处理完才能处理下一个单词，因此无法大规模并行处理，导致效率不高。Bai 等人[7]针对此问题提出时序卷积网络(TCN)，经过与多种 RNN 结构对比，在很多任务上 TCN 都能达到甚至超过 RNN，并且更加高效。

为了提高基于深度学习的短文本分类的有效性和效率，本文提出了基于 CNN 与 TCN 相结合，并加入权重优化[8]与注意力机制[9-12]的短文本分类模型。通过 TF-IDF 计算词的权重，使用 Probase 丰富语义知识，将词和概念通过 BERT-Base 转换词向量，并将词向量与词语权重相乘，将得到优化后的词向量作为输入层。运用 TCN 与 CNN 结合注意力[13]，更加高效且准确的获取最终特征表达。

1　相关知识

1.1　卷积网络(TCN)

TCN 是一种将获取编码时空信息的 CNN 和获取时间信息的 RNN，用一种统一的方法，以层次的方式捕获两个级别所有的信息。因果卷积上一层 t 时刻的值只对下一层 t 之前的值有依赖，是严格时间约束模型。卷积核大小会限制因果卷积对时间的长度，膨胀卷积可以获得更长的历史信息，其允许卷积时的输入存在间隔，用较少的层获得更大的感受野。如图 1 所示，本文选择在 TCN 中加入一个残差块替换一层卷积，使网络可以更好的通过跨层来传递信息。

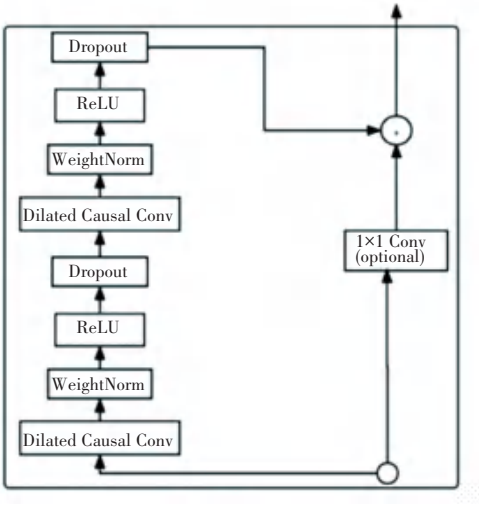

图 1　时间卷积网络

Fig. 1　TCN architecture

1.2　Probase

一个词语可以具有多重意思，如苹果可以指水果，也是一家公司的名称。为了使机器更好理解人类的语言，有学者提出了概念图谱。Probase 可以将短文本进行概念化，其包含了大量如图 2 所示的 is-A 关系，可以很好的解决短文本的稀疏性和歧义性问题。

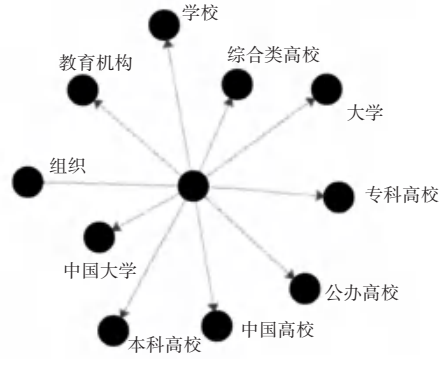

图 2　知识图谱

Fig. 2　CN-Probase

1.3　BERT[14]

BERT 模型通过 MLM 对双向 Transformers 进行预训练，生成双向语言表征，结构如图 3 所示。通过字向量、文本向量和位置向量，将 3 部分的和作为输入层，则可以得到融合了语义信息的表示向量。

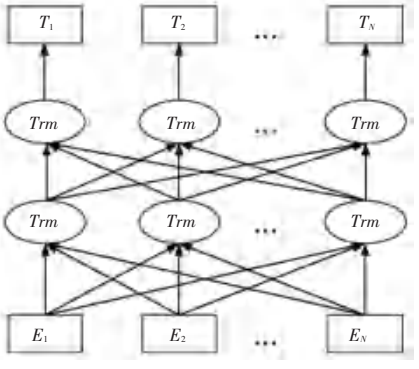

图 3　基于 Transformers 的双向编码器

Fig. 3　Bilateral encoder based on transformer

2　模型构建

图 4 展示了模型的整体设计，通过 CN-Probase 获取短文本概念层[15-16]，使用 BERT 模型将短文本和概念层转换为向量矩阵(矩阵大小为单词数乘词向量维度)，利用 $TF-IDF$ 计算每个词的权重，并与向量矩阵相结合[17]，得到赋予权重矫正的新矩阵。使用 CNN 对新矩阵 K 次卷积，得到 K 个不同阶段的矩阵，以便提取不同的上下文特征表示。将这 K 个矩阵运用 TCN 和注意力[18]获得特征表示，而后拼接向量进入全连接层，经过分类器输入结果。

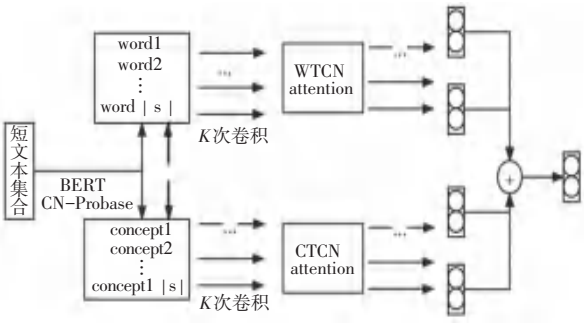

图 4 模型整体设计

Fig. 4 Overall design of the model

2.1 带权重的输入层

2.1.1 文本预处理

由于短文本中含有标点符号和字符等干扰项，因此在处理前需要进行数据清洗。本文使用 Jieba 分词中的全模式，能够快速获取短文本中所有可以成词的词语，并利用四川大学机器智能实验室停用词库去除无用的词和特殊符号。

2.1.2 CN-Probase 扩展

简单依靠分词和停用词并不能很好地消除短文本中的歧义性，因此本文通过 CN-Probase 对分词进行转换，获取到相对应的概念层，可以很好地消除歧义性和稀疏性。

2.1.3 优化的 BERT 获取词向量

字词的重要性与其在文本中出现的次数成正比，但和文本库中出现的频率成反比。利用 $TF-IDF$ 计算出每个单词的权重。计算公式为：

$$TF-IDF = TF \times IDF \tag{1}$$

其中，TF 是词频，IDF 由总文件数目除以包含该词语文件的数目，再将得到的商取以 10 为底的对数得到。

输入向量包含的信息直接决定了后面提取特征能够获得的上限，因此本文采用 BERT-LARGE 模型。该模型有 24 个网络层数、1 024个隐藏维度和 16 个注意力头。BERT 的向量化过程如图 5 所示。

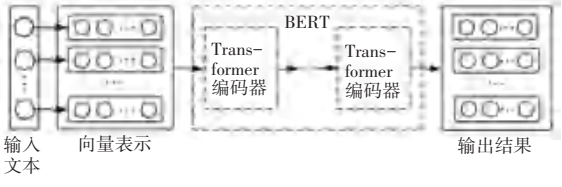

图 5 BERT 文本向量化

Fig. 5 BERT text vectorization

将向量化后的文本矩阵与对应词的 $TF-IDF$ 相乘得到新的词向量矩阵 $\boldsymbol{W}$，即为带权重的输入层。同理，将概念层做相同处理得到概念矩阵 $\boldsymbol{C}$。

2.2 基于注意力机制的 TCN

输入层使用 CNN 对获取的信息进行特征提取，为了更加深层次挖掘词语和概念的局部语义特征，使用多个层次的卷积层进行特征提取。TCN 能够捕获较长的上下文信息，但由于因果卷积的单向性不适合此类任务，因此使用膨胀卷积对 $\boldsymbol{W}$ 和 $\boldsymbol{C}$ 运算，同时加入残差链接，使网络可以更加有效地跨层传递信息，得到 h。为了减少错误词语和概念造成的坏影响，通过注意力机制可以达到该效果。如图 6 所示，将不同阶段获取到的特征向量 $\boldsymbol{O}(\boldsymbol{W})$ 和 $\boldsymbol{O}(\boldsymbol{C})$ 拼接，得到最终的特征表达。

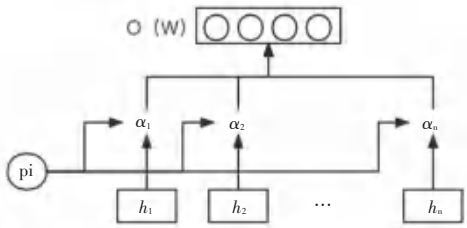

图 6 注意力机制

Fig. 6 Attention mechanism

2.3 输出层

为了充分利用上述处理得到的短文本中词和概念结合的特征表示，采用线性函数训练模型，公式如下：

$$y = \boldsymbol{W} \cdot O + b \tag{2}$$

其中，y 是所有类别的可能性分布；W 是权重；b 是偏移量。使用交叉熵训练 Loss。

3 实验与分析

3.1 数据集与参数设置

为了验证模型的有效性，本文使用了今日头条新闻文本分类数据集。其包含了382 688条数据，并划分为 15 个类别。将这些数据中的新闻标题提取出来，通过十折交叉验证。参数设置见表 1：

表 1 参数设置表

Tab. 1 Parameter set

参数名称	参数值
学习率	0.005
批尺寸	64
隐藏层	128
丢弃率	0.5
卷积核	2
膨胀因子	3
优化器	Adam
训练次数	100

3.2　模型对比

通过在相同数据集与其它模型进行对比，验证本文模型的有效性，主要采用 F_1 值和 Acc 值进行评估。对比模型包括 BERT、LSTM[19]、CNN[20]、Transformer、Seq2seq_Att 等。不同模型在数据集的结果见表 2。

表 2　不同模型在数据集的结果

Tab. 2　Results of different models on the dataset

Model	Value	
	Acc	F_1
BERT	0.938 6	0.930 4
LSTM	0.923 5	0.312 0
CNN	0.893 5	0.887 6
Transformer	0.904 5	0.870 9
Seq2seq_Att	0.896 7	0.889 7
本文模型	0.945 7	0.942 8

从实验结果来看，本文方法相比于基本的深度学习模型，准确率和 F_1 值都有所提高。其主要原因是模型通过外部预料丰富的短文本的语义信息；其次对于 BERT 的权重更改以及多阶段卷积和注意力机制的引入，也使得结果变得更加精确。

4　结束语

针对短文本分类问题，本文提出了利用外部知识丰富短文本语义，优化的 BERT 向量，以及在多阶段卷积中获取不同阶段的特征，利用加入了残差链接的 TCN 和注意力机制，使得模型拥有更加全面的特征获取的能力。从实验结果来看，本文提出的模型能有效提高短文本分类的效果。

参考文献

[1] 刘硕，王庚润，李英乐，等. 中文短文本分类技术研究综述[J]. 信息工程大学学报，2021，22(3)：304-312.

[2] 邓丁朋，周亚建，池俊辉，等. 短文本分类技术研究综述[J]. 软件，2020，41(2)：141-144.

[3] 李珍. 基于语义扩展的短文本分类研究[D]. 西安：西安电子科技大学，2019.

[4] 黄佳佳，李鹏伟，彭敏，等. 基于深度学习的主题模型研究[J]. 计算机学报，2020，43(5)：827-855.

[5] KIM Y. Convolutional Neural Networks for Sentence Classification. arXiv preprint arXiv：1408.5882，2014.

[6] LIU P，QIU X，HUANG X. Recurrent neural network for text classification with multi-task learning[J]. arXiv preprint arXiv：1605.05101，2016.

[7] BAI S，KOLTER J Z，KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. arXiv preprint arXiv：1803.01271，2018.

[8] 朱征宇，孙俊华. 改进的基于《知网》的词汇语义相似度计算[J]. 计算机应用，2013，33(8)：2276-2279，2288.

[9] 彭伟乐，武浩，徐立. 基于注意力机制面向短文本多分类的关键词权重优化[J/OL]. 计算机应用：1-9[2022-01-17]. http://kns.cnki.net/kcms/detail/51.1307.TP.20210727.1453.008.html.

[10] 屈亮亮，侯霞. 一种基于图注意力网络的短文本分类方法[J]. 北京信息科技大学学报(自然科学版)，2021，36(5)：85-90.

[11] 刘媛媛. 融合 CNN-LSTM 和注意力机制的空气质量指数预测[J]. 计算机时代，2022(1)：58-60.

[12] 侯玉兵. 基于注意机制的短文本分类方法[J]. 电脑知识与技术，2020，16(28)：185-186，201.

[13] LIU Y，LI P，HU X. Combining context-relevant features with multi-stage attention network for short text classification[J]. Computer Speech & Language，2022，71：101268.

[14] ASGARI-CHENAGHLU M，FEIZI-DERAKHSHI M R，BALAFAR M A，et al. TopicBERT：A cognitive approach for topic detection from multimodal post stream using BERT and memory-graph[J]. Chaos，Solitons & Fractals，2021，151：111274.

[15] 梁登玉，刘大明. 融合多粒度信息和外部知识的短文本匹配模型[J/OL]. 计算机工程：1-10[2022-01-17].

[16] 贺伟成. 语义一致的实体扩展技术研究[D]. 北京：北京交通大学，2020.

[17] 朱向其，张忠林，李林川，等. 基于改进词性信息和 ACBiLSTM 的短文本分类[J]. 计算机应用与软件，2021，38(12)：179-186.

[18] 陈莉媛，毋涛. 融合主题模型与自注意力机制的短文本情感分析方法[J]. 国外电子测量技术，2021，40(11)：18-23.

[19] 吴岗. 基于有序神经元 LSTM 的短文本相似性检测[J]. 计算机应用与软件，2021，38(12)：314-319，340.

[20] 姜丽婷，古丽拉·阿东别克，马雅静. 基于混合卷积网络的短文本实体消歧[J]. 中文信息学报，2021，35(11)：101-108.

(上接第 155 页)

[12] CAO H，WANG Y，CHEN J，et al. Swin-Unet：Unet-like Pure Transformer for Medical Image Segmentation[J]. arXiv preprint arXiv：2105.05537，2021.

[13] ZHUANG J. LadderNet：Multi-path networks based on U-Net for medical image segmentation[J]. arXiv preprint arXiv：1810.07810，2018.

[14] LI L，VERMA M，NAKASHIMA Y，et al. Iternet：Retinal image segmentation utilizing structural redundancy in vessel networks[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020：3656-3665.