文章编号：2096-1472(2023)-06-11-05
DOI:10.19644/j.cnki.issn2096-1472.2023.006.003

基于 BERT-BiLSTM-CNN 模型的新闻文本分类研究

徐建飞，吴跃成

(浙江理工大学机械及其自动控制学院，浙江 杭州 310018)
✉ 992759613@qq.com；wuyuecheng@126.com

摘 要：针对双向长短期记忆网络(BiLSTM)没有考虑到局部关键信息对文本分类的影响，以及卷积神经网络(TextCNN)无法捕获文本的长远距离的特征信息等问题，文章提出了一种基于 BERT-BiLSTM-CNN 混合神经网络模型的新闻文本分类的方法。为了进一步增强文本表示和提高新闻文本分类的效果，首先使用 BERT 预训练模型对文本进行词嵌入映射，其次利用 BiLSTM-CNN 模型进一步提取文本上下文和局部关键特征，最后对新闻文本进行分类；并在 THUCNews 数据上进行对比实验，实验结果表明，BERT-BiLSTM-CNN 模型的文本分类效果优于 Transformer、TextRNN、TextCNN 等深度学习模型。

关键词：BERT；BiLSTM-CNN；深度学习；新闻文本分类
中图分类号：TP391　　文献标识码：A

Research on News Text Classification Based on BERT-BiLSTM-CNN

XU Jianfei, WU Yuecheng

(*Faculty of Mechanical Engineering & Automation, Zhejiang Sci-Tech University, Hangzhou* 310018, *China*)
✉ 992759613@qq.com; wuyuecheng@126.com

Abstract: Aiming at the problems that Bi-directional Long Short-Term Memory (BiLSTM) network does not consider the influence of local key information on text classification, and that TextCNN cannot capture the long-term feature information of text, this paper proposes a method of news text classification based on BERT- BiLSTM-CNN hybrid neural network model. In order to further enhance text representation and improve the effectiveness of news text classification, firstly the BERT (Bidirectional Encoder Representation from Transformers) pre-training model is used to perform word embedding mapping on the text. Secondly, the BiLSTM-CNN model is used to further extract text context and local key features. Finally, the news text is classified and comparative experiments are conducted on THUCNews data. The experimental results shows that the text classification effect of BERT-BiLSTM-CNN model is better than that of Transformer, TextRNN, TextCNN and other deep learning models.

Keywords: BERT; BiLSTM-CNN; deep learning; news text classification

1 引言(Introduction)

在数据爆炸的时代，如何将数据中蕴含的价值提取出来，关键的一个步骤就是对获得的数据文本进行分类。新闻文本分类是自然语言处理领域的一个非常经典的任务，目前常用的文本分类技术[1-2]包含三大类：基于规则的算法、传统机器学习算法及深度学习算法。

基于规则的算法主要是利用人工对文本提取特征进行分类，但由于近年来全球大数据储量呈现爆炸式增长，而且分类方式极易受到标注人的主观认识的影响，所以转而利用机器学习的方法对文本数据特征进行自动标注。基于传统的机器学习算法，如朴素贝叶斯算法(Naive Bayes)[3]、支持向量机(Support Vector Machine，SVM)[4]、K 邻近法(K-Nearest Neighbors，KNN)[5]等，主要通过手工提取特征输入分类器进行训练，而基于传统机器学习算法的文本表示通常存在高纬度、稀疏、特征信息提取不全等问题。目前，深度学习神经网络模型在文本分类上表现出很好的分类效果，常用的深度学习模型有卷积神经网

基金项目：浙江省基础公益研究计划(LGF19E050005).

络(Convolutional Neural Network，CNN)[6]、长短期记忆网络(Long Short-Term Memory，LSTM)[7]、双向长短期记忆网络(Bi-directional Long Short-Term Memory，BiLSTM)[8]，但这些单一的传统神经网络模型在文本分类领域的准确率表现不佳。为提高文本分类的准确率，本文提出了一种基于混合神经网络的文本分类模型(BERT-BiLSTM-CNN)对中文新闻文本进行分类。

2 相关研究(Related research)

深度学习的快速发展使基于深度学习的文本分类模型有了很大的提升。2014年，KIM[6]针对CNN的输入层做了一些变换，提出了用于文本分类的卷积神经网络模型TextCNN，该模型的核心是通过不同的卷积核筛选出文本的局部关键信息，但忽略了上下文语义关联性信息。HOCHREITER等[7]针对RNN(循环神经网络)存在训练时梯度衰减、梯度爆炸，以及只能捕获距离当前位置较近的信息等问题，对RNN网络结构进行了改进和优化，通过控制神经元上的门(gate)开关时间捕获序列长距离和短距离的依赖信息，但LSTM只能沿着一个方向扫描，无法捕获未来的或者下文的语义信息。SCHUSTER等[8]为了更加充分地捕获序列的语义信息，继续对LSTM的网络结构进行优化，从而提出了双向的LSTM模型(BiLSTM)，BiLSTM可以同时用LSTM向前和向后对整个序列进行扫描。在单标签分类中，不管是传统的RNN，还是经过改良后的LSTM模型，在单独进行文本分类时，准确率表现不佳，因此引用Google发布的基于Transformer(自注意力模型)的双向编码器表征量BERT[9-10]模型对中文新闻文本进行分类。其中，CNN无法有效提取文本上下文依赖，而LSTM只能捕获整个句子的语义信息，而无法捕获文本的局部关键性的特征信息。针对以上问题，本文提出了一种基于混合神经网络的新闻文本分类模型(BERT BiLSTM-RCNN)对中文新闻文本进行分类。

3 BERT-BiLSTM-CNN模型设计(BERT-BiLSTM-CNN model design)

3.1 BERT的基础架构

Google提出的BERT(Bidirectional Encoder Representation from Transformers)模型应用在自然语言处理领域的各种任务中都有着非常突出的表现，归结于BERT网络架构能在词向量(Word Embedding)表示中保留更丰富的语义信息。BERT模型利用Transformer架构的编码器部分构造了一个多层双向的网络架构。BERT模型结构图如图1所示。

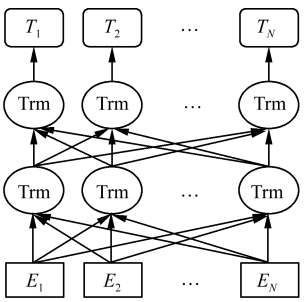

图1 BERT模型结构图

Fig. 1 Structural diagram of BERT model

Transformer模型中的编码端是由6个Encoder组成，解码端是由6个Decoder组成。不仅采用多头自注意机制从多个角度捕获更丰富的语义信息，而且采用了残差网络架构缓解模型梯度消失和爆炸的问题。其中，每一个Encoder的结构都是相同的，都由输入、多头自注意力机制(Multi-Head Self-Attention)及全连接神经网络(Feed Forward Network)构成，如图2所示。

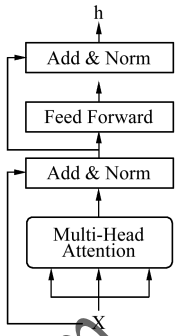

图2 Encoder模块

Fig. 2 Encoder block

图3中清楚地展示了“基金上半年业绩排名”这句话的词嵌入过程，Input为中文文本内容，Token Embedding(*TE*)为字嵌入，Segment Embedding(*SE*)为段嵌入，Position Embedding(*PE*)为位置嵌入。三个部分的嵌入向量相加的总和就是BERT编码器的最终输入词向量*E*。

$$E=TE+SE+PE \tag{1}$$

图3中的BERT输入有2个特殊的标记，[CLS]是文本输入序列的开始标志符，而[SEP]是用于将文本中的句子隔开的标志符，用来区分文本中不同的句子。通过添加“A”或“B”区分句子创建Embedding。词语的语序也非常重要，比如“他欠我100万”和“我欠他100万”，这两句话只是两个字的语序位置不同，但意思天差地别。因此，需要给文本中每个字添加位置编码。图3中，当一个输入序列输入BERT时，它会一直向上移动堆栈。在每一个块中，它会首先通过一个自我注意层再到一个前馈神经网络，之后被传递到下一个编码器，最后每个位置将输出一个大小为隐藏层维数的向量。

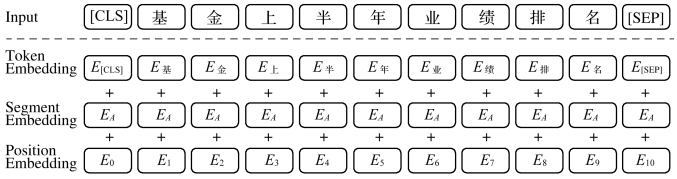

图3 词嵌入过程

Fig. 3 Token embedding process

3.2 BiLSTM

RNN在处理序列输入数据上效果突出，常用于处理文本数据。该模型考虑了时序因素，通过逐词分析文本，并将先前所有文本的语义存储在固定大小的隐藏层中，能够更好地捕捉上下文信息。由于传统RNN在处理长序列信息时会保留较多的冗余信息和存在梯度消失和爆炸的问题，所以本文采用另一种RNN的改进算法LSTM，它可以利用时间序列对输入进行分析，增加了输入门、遗忘门和输出门，可将短期记忆与长期记忆结合起来控制细胞中信息的增加或丢弃，解决

了 RNN 无法记住距离当前时刻更久的信息和梯度消失的问题。输入门控制的是当前时刻序列的输入，遗忘门控制的是上一时刻序列中存储的历史信息，输出门控制的是当前时间序列的输出，其结构如图 4 所示。

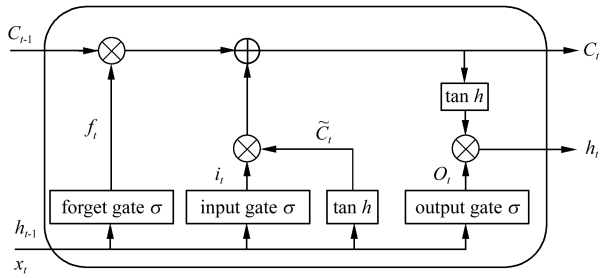

图 4　LSTM 模型

Fig. 4 LSTM model

LSTM 模型各个门和状态的计算公式如公式(2)至公式(7)所示：

$$f_t=\sigma(\boldsymbol{W}_{xf}\cdot x_t+\boldsymbol{W}_{hf}\cdot h_{t-1}+\boldsymbol{b}_f) \tag{2}$$

$$i_t=\sigma(\boldsymbol{W}_{xi}\cdot x_t+\boldsymbol{W}_{hi}\cdot h_{t-1}+\boldsymbol{b}_i) \tag{3}$$

$$O_t=\sigma(\boldsymbol{W}_{xo}\cdot x_t+\boldsymbol{W}_{ho}\cdot h_{t-1}+\boldsymbol{b}_o) \tag{4}$$

$$\widetilde{C}_t=\tan h(\boldsymbol{W}_{xc}\cdot x_t+\boldsymbol{W}_{hc}\cdot h_{t-1}+\boldsymbol{b}_c) \tag{5}$$

$$C_t=f_t*C_{t-1}+i_t*\widetilde{C}_t \tag{6}$$

$$h_t=O_t*\tan h(C_t) \tag{7}$$

其中，f_t 表示遗忘门、i_t 表示输入门、O_t 表示输出门、$\widetilde{C}_t$ 表示候选记忆细胞、C_t 表示新的细胞状态，σ 表示 Sigmoid 函数，上一个阶段的输出表示 h_{t-1}，当前阶段的输入表示 x_t，$\boldsymbol{W}_{xf}$、$\boldsymbol{W}_{hf}$、$\boldsymbol{W}_{xi}$、$\boldsymbol{W}_{hi}$、$\boldsymbol{W}_{xo}$、$\boldsymbol{W}_{ho}$、$\boldsymbol{W}_{xc}$、$\boldsymbol{W}_{hc}$ 分别表示遗忘门、输入门、输出门和输入细胞状态的各个连接权重矩阵，$\boldsymbol{b}_f$、$\boldsymbol{b}_i$、$\boldsymbol{b}_o$、$\boldsymbol{b}_c$ 表示各个阶段的偏移向量。

为了使预测结果更加准确，对于当前时刻的输出可能需要若干之前状态的输入和若干未来状态的输入共同决定，因此提出一种双向循环神经网络 BiLSTM。该模型整体结构由两个方向相反的 LSTM 以叠加的方式组成，正向 LSTM 用来接收从前向后传递的信息状态，反向 LSTM 用来接收从后向前的信息输入，在同一时刻将两个 LSTM 的状态输出进行合并，作为 BiLSTM 当前时刻的输出，从而得到每一时刻的向量特征，弥补 LSTM 无法访问未来的上下文信息的遗憾；BiLSTM 结构如图 5 所示。

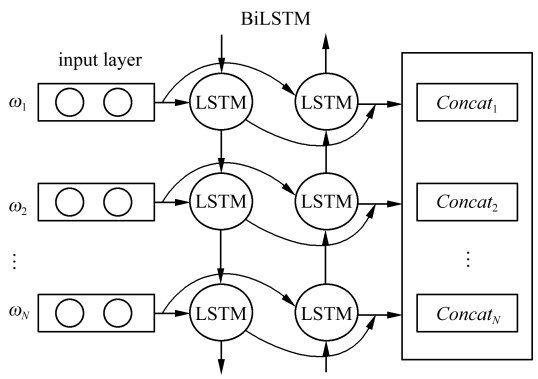

图 5　BiLSTM 模型

Fig. 5 BiLSTM model

使用 BiLSTM 模型提取分别来自 BERT 词向量嵌入层的上下文信息，得到正向词向量特征信息序列 $\overrightarrow{hw_t}$ 和负向词向量特征信息序列 $\overleftarrow{hw_t}$。拼接得到的 hw 为 BiLSTM 提取的文本特征向量，$\oplus$表示向量拼接，计算公式如公式(8)至公式(10)：

$$\overrightarrow{hw_t}=\overrightarrow{LSTM}(w_t) \tag{8}$$

$$\overleftarrow{hw_t}=\overleftarrow{LSTM}(w_t) \tag{9}$$

$$hw=[\overrightarrow{hw_t}\oplus\overleftarrow{hw_t}] \tag{10}$$

3.3 TextCNN

提到卷积神经网络(Convolutional Neural Networks，CNN)时，通常会认为 CNN 属于计算机视觉领域，是用于解决计算机视觉方向问题的模型，而后 KIM[6] 对 CNN 的输入层做了一些的变换，提出了用于文本分类的卷积神经网络模型 TextCNN。如图 6 所示，TextCNN 网络结构比较简单，由若干个卷积层、max pooling、全连接层和 Softmax 层构成，该模型的核心思想是使用不同尺寸的卷积核获取文本相邻的 N-gram 特征表示。首先，在卷积层使用多种尺寸的卷积核提取文本的局部特征，并将得到的特征向量输入池化层，通过下采样方式筛选重要特征信息，降低向量维度，减少计算量。然后，将经过 1-max pooling 处理后的特征向量拼接起来，再经过全连接层和 Softmax 函数进行多分类。TextCNN 对文本近距离浅层特征的抽取能力较强，并且网络架构简单，可以快速提取文本中的局部特征。

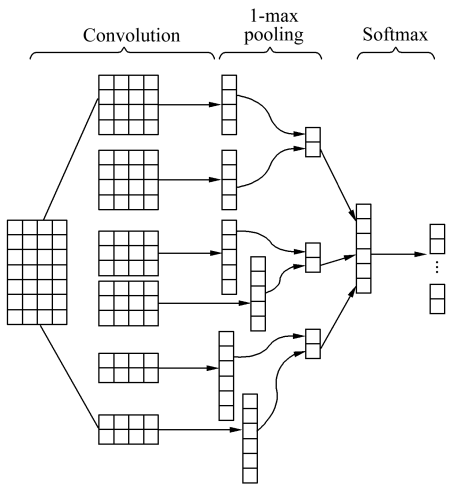

图 6　TextCNN 模型

Fig. 6 TextCNN model

3.4 模型总体架构

如图 7 所示，本文构造了一个基于 BERT-BiLSTM-CNN 的中文新闻文本分类模型，主要由词向量表示模块和深层文本特征提取分类模块构成。词向量转化模块是一个基于 BERT 的中文预训练模型，将文本转化为词向量并进行初步特征处理；深层文本特征提取分类模块利用 BiLSTM-CNN 网络对上下文深层特征和局部关键信息进行提取，最后得到分类结果。

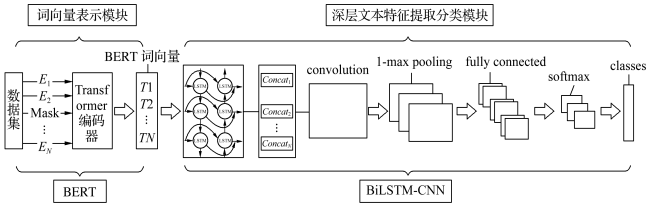

图 7 BERT-BiLSTM-CNN 模型

Fig. 7 BERT-BiLSTM-CNN model

BERT-BiLSTM-CNN 混合网络结构建立的具体步骤如下。

步骤 1：对要进行分类的文本数据集进行预处理，得到输入文本，记为 $E=(E_1,E_2,\cdots,E_i,\cdots,E_n)$，其中 E_i（$i=1,2,\cdots,n$）表示文本的第 i 个字。

步骤 2：将所有的 E_i 输入词向量表示模块，经过语言编码器 Transformer 的编码后，将文本 E 进行序列特征化，输出文本 $w_i=(w_{1i},w_{2i},\cdots,w_{ni})$，其中 w_{ni} 示文本中第 i 句中的第 n 个词的词向量，将 w_1—w_n 词向量拼接$(w_1,w_2,\cdots,w_n)$，得到 BERT 词向量表示矩阵 $\boldsymbol{W}$。

步骤 3：首先，将 BERT 词向量输入深层文本特征提取分类模块，经过双层 BiLSTM 后，文本词向量表示的语义语法信息的依赖性进一步得到增强；其次，经过 2-gram、3-gram 和 4-gram 等多个卷积核的卷积后，进一步提取文本中的关键特征和文本更深层的结构信息；再次，在池化层中对卷积层获得的每个特征图进行 1-max pooling 的特征映射处理得到统一的特征值，经过全连接层把池化后的特征连接起来，并进行 softmax 回归分类，生成最终的分类特征向量；最后输出所属的类别。

4 实验与结果(Experiments and results)

4.1 数据集

为了验证模型性能的优劣，本文在公开的新浪新闻文本分类数据集 THUCNews 上进行了对比实验，它是根据新浪新闻 RSS 订阅频道 2005—2011 年的数据中筛选过滤生成的，包含 74 万篇新闻文档，均为 UTF-8 的纯文本格式，本次实验使用的语料库包含 14 个类别，分别为财经、娱乐、房地产、游戏、股票、体育、教育、政治、科技、社会、彩票、家居、时尚和星座，每个类别包含的数据样本不均等，共 20 万条带类别标签的新闻数据，其中抽取 18 万条数据作为训练集，1 万条作为测试集，1 万条作为验证集。

4.2 实验参数设置

模型参数是模型内部的配置变量，不同的参数设置会影响实验结果，本文实验对比模型包括 TextCNN、BiLSTM、BERT、BERT-CNN 等神经网络模型，超参数设置如表 1 所示。

表1 超参数设置

Tab.1 Hyperparameter setting

参数	数值
word embedding	768
BiLSTM-CNN hidden size	768
Dropout	0.1
learning rate	5e−5
batch size	128
pad size	32
optimizer	Adam
Epoch	5

4.3 实验环境

本次实验环境配置如表 2 所示，预训练模型使用 Hugging Face 发布的 BERT 中文预训练模型。为了避免训练结果产生过拟合现象，本次实验使用提前终止技术(如果每次迭代之后的损失值变化很小，在迭代一定的次数后就不再进行参数的优化)。

表2 实验环境

Tab.2 Experimental environment

实验环境名称	配置
操作系统	Windows 10
GPU	NVIDIA GeForce RTX 3070
CPU	Inter(R)Core(TM)i5-12400F
Python	3.6
Pytorch	1.10 + cu113
编码格式	UTF-8

4.4 分类性能评估指标

评价指标是评价数据表现的标准是数据分析中非常重要的部分。对于此类标签文本分类，常用的评价指标是宏精确率(*Macro-Precision*)、宏召回率(*Macro-Recall*)和宏 $F1$(*Macro-F1*)，这 3 个评价指标通过计算得出的结果可以直观地观察模型在文本分类任务中的性能表现，因此本实验采用以上 3 个评价指标进行分析。其中，宏精确率表示每个类别精确率的算术平均值，如公式(11)所示；宏召回率表示每个类别召回率的算术平均值，如公式(12)所示；宏 $F1$ 表示每个类别 $F1$ 值的算术平均值，如公式(13)所示。

$$Macro\text{-}Precision=\frac{1}{L}\sum_{i=1}^{L}P_i\times 100\% \tag{11}$$

$$Macro\text{-}Recall=\frac{1}{L}\sum_{i=1}^{L}R_i\times 100\% \tag{12}$$

$$Macro\text{-}F1=\frac{2\times Macro\text{-}Precision\times Macro\text{-}Recall}{Macro\text{-}Precision+Macro\text{-}Recall}\times 100\% \tag{13}$$

4.5 实验结果与分析

各个模型的实验结果如表 3 所示，对比 TextCNN 和 BERT-CNN 模型可以看出，分类的宏精确率和宏 $F1$ 分数提高了 0.73%和 0.72%。对比 TextRNN 和 BERT-BiLSTM 模型可以看出，分类的宏精确率和宏 $F1$ 分数提高了 0.49%和 0.45%。由此可以得出，在融合 BERT 预训练模型后，能够更好地提取文本中的特征信息，所以引入 BERT 预训练模型可以提升新闻文本分类的效果。分别对比 BERT-BiLSTM-CNN 模型与 BERT-CNN、BERT-BiLSTM 的宏 $F1$ 分数，分别提高了 0.56%和 0.78%，而比 TextCNN 和 TextRNN 的宏 $F1$ 分数高了 1.28%和 1.23%。与其他一般的深度学习模型如表 3 中的 Transformer、BERT 等模型相比，本文模型的分类效果都是最优的。这是因为 BiLSTM 模型可以提取上下文的全部信息，但是无法提取文本的局部信息；而 TextCNN 模型可以提取文本中的关键信息，但是无法捕获长距离的语义信息。本文构建的 BERT-BiLSTM-CNN 同时弥补了以上两种模型的缺点，并且实验结果也证明了基于 BERT-BiLSTM-CNN 模型的新闻文本分类比一般的深度学习分类模型效果好。

表 3　THUCNews 数据集实验结果

Tab.3 Model experimental results of THUCNews dataset

模型	*Macro-Precision*	*Macro-Recall*	*Macro-F1*
TextCNN	0.908 0	0.908 0	0.908 0
TextRNN	0.908 5	0.908 6	0.908 5
Transformer	0.898 6	0.898 6	0.898 6
BERT	0.917 4	0.917 4	0.917 4
BERT-CNN	0.915 3	0.915 2	0.915 2
BERT-BiLSTM	0.913 4	0.913 1	0.913 0
本文模型	0.920 9	0.920 8	0.920 8

5 结论(Conclusion)

文本分类是 NLP 领域一个非常经典的任务，本文将多种深度学习模型融合并应用在新闻文本分类任务上。首先利用预训练模型 BERT 对文本数据进行高纬度映射，得到嵌入向量，其次通过双层 BiLSTM 对输入文本的特征信息进行提取，可以获取全局上下文信息，再次经过 TextCNN 对其中更深层的关键信息提取，最后经过 softmax 进行文本所属类别的预测。与传统模型相比，BERT-BiLSTM-CNN 算法分类的精确率有所提升，得益于本文对词向量转换过程保留了更多的语义语法信息。在未来的研究中，可以通过文本数据增强和外部知识进一步增强文本特征表示、引入更为优秀的词向量预训练模型、引入对比学习和提示学习等方法，进一步提升新闻文本分类的准确度。

参考文献(References)

[1] 乔嘉琪.一种有效的文本分类方法 MDCC 的实现及应用[D].合肥:安徽大学,2018.

[2] 崔雨萌,王靖亚,刘晓文,等.融合注意力和裁剪机制的通用文本分类模型[J/OL].计算机应用:1-13[2022-10-19].http://kns.cnki.net/kcms/detail/51.1307.tp.20221018.2159.002.html.

[3] PENG F, SCHUURMANS D. Combining naive Bayes and n-gram language models for text classification[C]// SEBASTIANI, FABRIZIO. Proceedings of the 25th European Conference on IR Reserch. Berlin, Heidelberg: Springer, 2002:335-350.

[4] MANEVITZ L M, YOUSEF M. One-class SVMs for document classification[J]. Journal of Machine Learning Research, 2001, 2:139-154.

[5] LI L, WEINBERG C R, DARDEN T A, et al. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method[J]. Bioinformatics, 2001, 17(12):1131-1142.

[6] KIM Y. Convolutional neural networks for sentence classification[C]// MOSCHITTI A, PANG B, DAELEMANS W. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: ACL, 2014:1746-1751.

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

[8] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.

[9] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]// BURSTEIN J, DORAN C, SOLORIO T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. Minneapolis MN USA: ACL, 2019:4171-4186.

[10] 张小为,邵剑飞.基于改进的 BERT-CNN 模型的新闻文本分类研究[J].电视技术,2021,45(7):146-150.

作者简介：

徐建飞(1997-)，男，硕士生. 研究领域：自然语言处理，人机交互.

吴跃成(1966-)，男，博士，副教授. 研究领域：人机交互.