

文章编号: 1003-0077(2021)02-0078-11

融合通道特征的混合神经网络文本分类模型

韩永鹏, 陈彩, 苏航, 梁毅

(北京工业大学 信息学部, 北京 100124)

摘要: 基于卷积神经网络与循环神经网络的混合文本分类模型通常使用单通道词嵌入。单通道词嵌入空间维度低, 特征表示单一, 导致一维卷积神经网络不能充分学习文本的空间特征, 影响了模型的性能。因此, 该文提出一种融合通道特征的混合神经网络文本分类模型。该模型使用了双通道词嵌入丰富文本表示, 增加了空间维度, 在卷积的过程中融合了通道特征, 优化了空间特征与时序特征的结合方式, 最终提高了混合模型的分类性能。在 IMDB、20NewsGroups、复旦中文数据集、THUC 数据集上进行实验, 该模型的分类准确率相比于传统卷积神经网络平均提升了 1%, 在 THUC 数据集上准确率最高提升了 1.3%。

关键词: 通道特征; 神经网络; 文本分类

中图分类号: TP391

文献标识码: A

Hybrid Neural Network Text Classification Model with Channel Features

HAN Yongpeng, CHEN Cai, SU Hang, LIANG Yi

(Faculty of Information, Beijing University of Technology, Beijing 100124, China)

Abstract: The hybrid text classification model based on convolutional neural network and recurrent neural network usually uses single-channel word embedding. Single-channel word embedding has low spatial dimension, leading that one-dimensional convolutional neural network fail to fully capture text features. This paper proposes a hybrid neural network text classification model combined with the channel features. The model uses two-channel word embedding to enrich text representation, fuses channel feature in the process of convolution, and optimizes the combination of spatial and temporal features. Tested on IMDB, 20NewsGroups, Fudan Chinese dataset and THUC dataset, the proposed model improves the classification accuracy by an average of 1% compared with the traditional methods, with a top increase of 1.3% on the THUC dataset.

Keywords: channel feature; neural network; text classification

0 引言

文本分类是跨越信息检索、机器学习和自然语言处理的多领域技术, 是信息处理和数据挖掘的重要研究方向, 主要目标是在事先定义好类别的情况下, 根据文本的内容特征或者属性特征, 将要分类的文本自动分配到所属的类别^[1]。根据文本的长度, 文本分类分为短文本分类与长文本分类, 短文本字符数通常不超过 200^[2]。

随着科学技术的发展, 深度学习被广泛应用于文本分类, 常用于文本分类的神经网络模型主要有

循环神经网络(recurrent neural network, RNN)与卷积神经网络(convolutional neural network, CNN)。循环神经网络是一种对序列数据建模的网络, 由于有梯度消失和梯度爆炸等问题, 通常使用其变体长短期记忆网络^[3](long short-term memory, LSTM)。由于 LSTM 只能学习文本的全局时序特征, 不能学习文本中的局部空间特征, 所以一般先使用 CNN 学习局部特征, 再结合 LSTM 学习时序特征。然而现有混合模型使用的单通道词嵌入空间维度低, 特征表示单一, 导致一维卷积神经网络不能充分发挥空间特征学习能力, 影响了模型的性能。

收稿日期: 2019-11-29 定稿日期: 2020-02-29

基金项目: 国家自然科学基金(61672505, 91546111)

为了弥补现有混合模型的不足,本文提出了一种融合通道特征的混合神经网络文本分类模型,该模型使用基于预测与基于统计的方式构建文本的双通道词嵌入。相比于单通道词嵌入,双通道词嵌入能提供更为丰富的特征,并增加文本表示的空间维度。为了充分利用增加的空间维度,本文在卷积的过程中进行了通道特征融合,提高了卷积层的空间特征学习能力。为了更好地将空间特征与时序特征结合,模型在每路卷积后使用双向 LSTM 学习各路时序特征,避免了过早进行卷积特征融合对融合后的时序特征造成破坏。在四个数据集上进行对比实验,结果表明,该模型取得了良好的分类效果,分类准确率相较于传统卷积神经网络平均提升了 1%。

1 相关工作

传统机器学习算法^[4]在文本分类时往往需要进行特征选择,而深度学习算法因可以自动进行特征学习而被广泛使用,常用结构包括卷积神经网络 CNN 与循环神经网络 RNN。RNN 适合处理时间序列数据,被广泛应用于文本分类当中。Liu 等人^[5]提出了基于 LSTM 的三种模型用于处理多任务学习下的文本分类问题。Xu 等人^[6]使用双向 LSTM 结合前馈型神经网络进行情感分析。由于 LSTM 只能输出最后时刻的特征,不能充分利用各时刻的特征,部分学者尝试使用注意力机制优化 LSTM 的特征表示。Wang 等人^[7]使用了注意力机制对 LSTM 的各个时刻的特征进行加权,在情感分类任务中取得良好效果。Long 等人^[8]在双向 LSTM 中引入了 Multi-head Attention 进行情感分类,取得了优于双向 LSTM 的效果。由于 RNN 不能学习空间特征且训练时间长,CNN 在文本领域开始使用。Kim^[9]首次将 CNN 用于文本分类,采用多路卷积提取空间特征,使用全局最大池化保留最重要的特征,通过实验验证了 CNN 在文本分类领域的实用性。由于全局最大池化容易造成特征大量丢失,Kalchbrenner 等人^[10]提出了一种动态池化的思想,在不同池化层采取不同的 K 值,保留了前 K 个最大特征,有效解决了全局最大池化特征丢失严重的问题。Yang 等人^[11]首次将胶囊神经网络用于文本分类,在部分数据集上取得了超过经典 CNN 的效果。王盛玉等人^[12]尝试在 CNN 中结合注意力机制,有效提升了 CNN 学习局部特征的能力。

由于 CNN 与 RNN 各有侧重,许多学者结合两者优点提出混合模型。Lai 等人^[13]提出了循环卷积神经网络 RCNN,使用双向循环结构对特征的上下文进行建模,实现了卷积的核心思想。Zhou 等人^[14]提出了混合模型 C-LSTM,给出了 CNN 与 RNN 结合使用的模式。Hassan 等人^[15]提出的模型使用多路卷积学习空间特征,融合后经由 LSTM 学习时序特征。Chen 等人^[16]提出的模型在每一路通过堆叠卷积池化层提取更抽象的空间特征,融合后结合 LSTM 进行时序特征学习。Zhang 等人^[17]提出的 LSTM-CNN 探索了先时序后空间的特征学习方式。在此基础上,Zheng 等人^[18]提出的 BRCAN 模型使用双向 LSTM 学习上下文信息,然后结合 CNN 与注意力机制对关键的特征进行加权,在多个数据集上取得良好分类效果。江伟等人^[19]探索了多种注意力机制,进行了全面的对比评估。程艳等人^[20]提出的 C-HAN 模型将文本表示分为词—句子、句子—文档两个阶段,并对比了词向量、字向量对模型性能的影响。车蕾等人^[21]提出的 TSOHHAN 模型结合了标题在话题分类中的作用,取得了优于传统层级注意力网络的分类准确率。不同于以上学者的小规模浅层神经网络模型,Google 团队提出了预训练语言模型 BERT^[22],在多项 NLP 任务中取得了卓越的效果。

尽管学者们提出了多种混合模型,但现有混合模型仍存在以下问题:①普遍使用单通道词嵌入,空间维度低,文本的特征表示单一,只能在单通道上使用一维卷积算法,不能充分发挥卷积的空间特征学习能力;②现有的 CNN-RNN 混合模型在融合多路卷积特征时,往往对融合后的特征时序性造成破坏,影响了后续 LSTM 层对时序特征的学习过程。为此,本文分别使用基于预测与基于统计的方法构建双通道词嵌入,丰富文本表示,增加嵌入层空间维度。在此基础上,为了充分利用双通道特征,本文的模型先在两个通道独立学习空间特征,然后使用逐点卷积融合通道特征,增强了卷积层的空间特征学习能力。在融合多路卷积特征时,在每一路均使用结合注意力机制的双向 LSTM 进行时序特征学习,将每路的特征进行拼接表示文本,有效避免了在进入 LSTM 之前,多路卷积特征融合的过程对融合后的时序特征造成破坏的问题。实验表明,本文提出的混合模型在多个数据集上取得了良好的分类性能。

2 模型描述

融合通道特征的混合神经网络文本分类模型结构如图 1 所示。模型的输入为双通道词嵌入,分别由基于预测与基于统计的词向量生成模型在海量语料中预训练得到,使用预训练词嵌入将大大提高模型的泛化能力。相比于单通道词嵌入,双通道词嵌入增加了文本表示的空间维度,增加了特征的多样

性,丰富了特征的表达。之后,模型使用多路卷积提取空间特征,每一路使用不同大小的卷积核提取不同感受野的局部空间特征,在提取空间特征的过程中融合了通道间特征。为了避免在进入 LSTM 之前,多路卷积特征融合的过程对融合后的时序特征造成破坏,在每一路均使用结合注意力机制的双向 LSTM 网络进行时序特征学习,最终将各路特征进行拼接,形成文档的最终表示,然后经过全连接层与 Softmax 层进行文本分类。

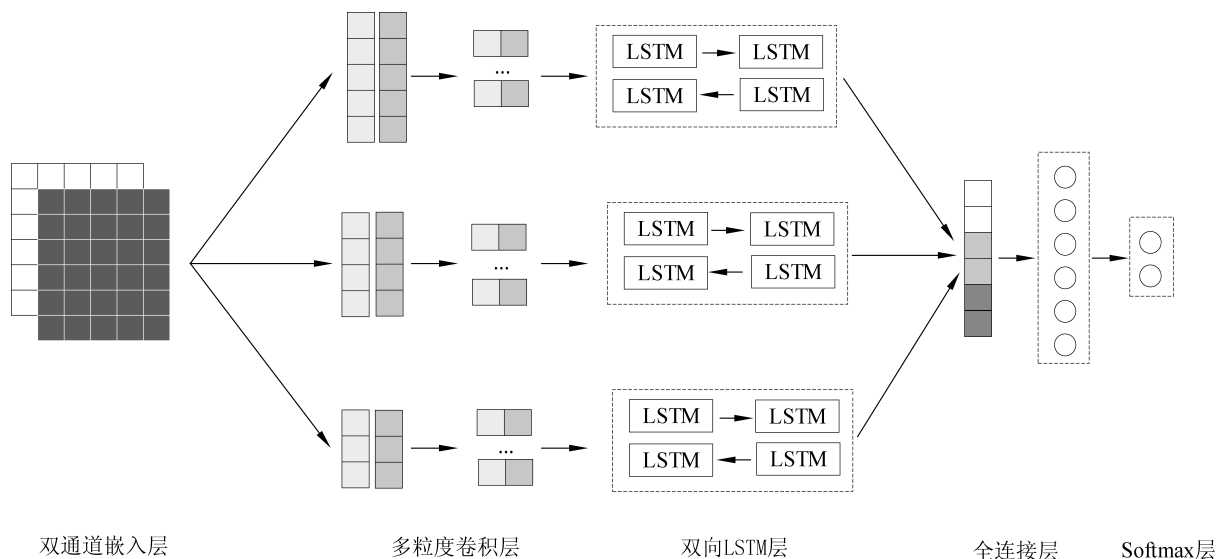


图 1 融合通道特征的混合神经网络文本分类模型

2.1 融合通道特征的多粒度卷积层

以图 1 中的一路卷积为例,详细的卷积层设计原理如图 2 所示。

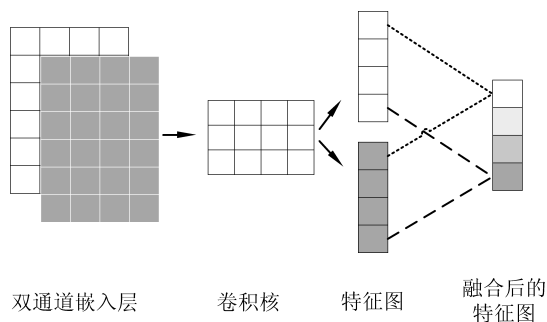


图 2 卷积层设计原理

卷积层的输入为双通道预训练词嵌入矩阵,不受特定分类任务影响,能增加模型的泛化能力。随着神经网络的训练,对双通道嵌入层的权值进行动态调整,使得原本与分类任务无关的词向量变成与特定分类任务相关的词向量,加速整个神经网络模

型收敛的过程。令词嵌入矩阵最多包含 n 个单词,超出 n 个单词的文本被截断,不足 n 个单词的文本用 0 填充。 x_i 表示当前文本中第 i 个单词的预训练词向量,则词嵌入矩阵 $\mathbf{X}_{1:n}$ 可以表示如式(1)所示。

$$\mathbf{X}_{1:n} = x_1 \otimes x_2 \otimes \cdots \otimes x_n \quad (1)$$

其中, \otimes 代表词向量的拼接,卷积操作在词嵌入矩阵 $\mathbf{X}_{1:n}$ 上进行。定义卷积核 \mathbf{W}_c , \mathbf{W}_c 为 $h \times k$ 的二维矩阵, h 代表当前卷积核的感受野大小,而 k 固定为词嵌入的维度,让卷积操作只能沿着时间轴自上而下进行滑动,令 c_i 表示滑动过程中提取到的当前位置的局部特征, f 代表非线性激活函数, b_c 为偏置项,则卷积核形成的特征图 \mathbf{C} 可以由式(2)、式(3)所示。

$$c_i = f(\mathbf{W}_c \cdot \mathbf{X}_{i:i+h-1} + b_c) \quad (2)$$

$$\mathbf{C} = [c_1, c_2, \dots, c_{n-h+1}] \quad (3)$$

由于嵌入层有两个通道,所以在每一个通道上使用同一个卷积核,卷积将形成两张不同的特征图,分别记为 \mathbf{C}_1 与 \mathbf{C}_2 。此时进行逐点卷积,使用 1×1 , 深度为 2 的卷积核 \mathbf{W}_f 对来自两个通道的两张特征

图 C_1 与 C_2 进行通道特征融合,形成融合通道特征之后的特征图 V ,计算如式(4)所示。

$$V = f(W_f \cdot [C_1, C_2] + b_f) \quad (4)$$

其中, f 为非线性激活函数, b_f 为偏置项。

至此可以得到使用一个卷积核 W_c 在双通道嵌入层实施卷积后所形成的一张特征图 V 。由于卷积神经网络通常使用多个卷积核进行空间特征学习,令 N 表示卷积核个数,则使用 N 个相同尺寸的卷积核在双通道嵌入层实施卷积后可以形成 N 张特征图组成特征矩阵 M_o ,如式(5)所示。

$$M_o = [V_1, V_2, \dots, V_N] \quad (5)$$

由于 M_o 的行维度往往较大,如果使用池化降维将导致时序特征丢失,所以模型使用步幅为 K 的卷积核 W_p 对特征矩阵进行卷积降维,形成降维之后的特征图矩阵 M_k ,计算如式(6)所示。

$$M_k = f(W_p \cdot M_o + b_p) \quad (6)$$

其中, f 为激活函数, b_p 为偏置项。由式(6)形成特征矩阵 M_k 保留了时序特征,可以按行的顺序依次输入到 LSTM 当中,完成时序特征的学习。

2.2 融合多路特征的双向 LSTM 层

对于长文本而言,单词的上下文信息充足,往往存在长距离的语义关联,相比于特征少、时序信息不足的短文本,长文本对特征的时序性有着更高的要求。在特征输入 LSTM 之前,多路卷积先进行特征融合,并不能保证融合后特征的时序性,大大影响了 LSTM 对长文本的时序特征学习过程。令 M_1, M_2 分别表示不同路卷积所形成的特征图矩阵,若将 M_1 与 M_2 横向拼接,由于卷积核大小不同造成 M_1 与 M_2 在行维度上不同,只能使用 0 填充,让卷积后的特征图尺寸保持不变,这将导致 M_1 与 M_2 的时序特征不能完全保持对齐,造成整体时序特征质量下降的问题。若将 M_1 与 M_2 纵向拼接,则不能保证拼接后整体特征保持全局有序性。

为了避免上述融合方式的不足,本文的模型在每一路均使用双向 LSTM 学习时序特征,将每一路的双向时序特征进行拼接表示最终文本,避免了各路特征在进入 LSTM 之前就进行融合所导致的时序特征质量下降的问题。由于传统的正向 LSTM 只能学习特征的上文信息,忽视了特征的下文信息,本文使用了双向 LSTM 同时学习特征的上下文信息,极大地提高了模型的时序特征学习能力。为了充分利用 LSTM 所有时刻的输出特征,模型通过注意力机制对 LSTM 每个时刻的特征进行加权求和,

提高 LSTM 的输出质量,本文的双向 LSTM 层如图 3 所示。

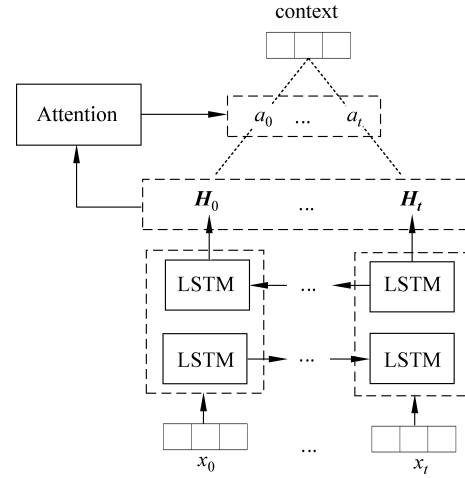


图 3 双向 LSTM 层

令 i 代表第 i 个时刻且 $i \in [0, t]$, x_i 表示第 i 个时刻的输入向量。 M_k 代表一路卷积所形成的特征图矩阵,则 M_k 可以表示成多个行向量的拼接,如式(7)所示。

$$M_k = x_0 \oplus x_1 \oplus \dots \oplus x_t \quad (7)$$

LSTM 按时间顺序接收 x_i 作为输入向量, c_i 表示 LSTM 单元状态, h_i 表示 LSTM 单元最终输出。 f_i, i_i, o_i 分别表示遗忘门、输入门与输出门, σ 表示 Sigmoid 激活函数, $W_f, W_i, W_o, W_c, b_f, b_i, b_o, b_c$ 为网络需要学习的参数, LSTM 的最终输出计算如式(8)~式(13)所示。由于模型使用了双向 LSTM 学习时序特征,所以双向 LSTM 的最终输出由正向 LSTM 输出与反向 LSTM 输出拼接得到。

$$f_i = \sigma(W_f \cdot [h_{i-1}, x_i] + b_f) \quad (8)$$

$$i_i = \sigma(W_i \cdot [h_{i-1}, x_i] + b_i) \quad (9)$$

$$o_i = \sigma(W_o \cdot [h_{i-1}, x_i] + b_o) \quad (10)$$

$$q_i = \tanh(W_c \cdot [h_{i-1}, x_i] + b_c) \quad (11)$$

$$c_i = f_i \odot c_{i-1} + i_i \odot q_i \quad (12)$$

$$h_i = o_i \odot \tanh(c_i) \quad (13)$$

由于 LSTM 只能学习得到最后一个时刻的输出向量,不能对每个时刻的输出充分利用,本文使用注意力机制完成各个时刻输出特征的加权融合。令 H_i 表示第 i 个时刻的双向 LSTM 层的输出向量, e_i 表示 H_i 对整个文本语义表示的重要程度, a_i 表示 H_i 对整个文本语义表示贡献的权重。根据上述定义,双向 LSTM 层的注意力权重计算如式(14)、式(15)所示。

$$e_i = \mathbf{u}^T \cdot \tanh(\mathbf{W}_a \cdot \mathbf{H}_i + b_a) \quad (14)$$

$$a_i = \frac{\exp(e_i)}{\sum_{j=0}^t \exp(e_j)} \quad (15)$$

其中, \mathbf{u}^T 、 \mathbf{W}_a 、 b_a 是网络需要学习的参数, \tanh 为非线性激活函数。在得到双向 LSTM 层的各个时刻的注意力权重后, 使用式(16)对双向 LSTM 层的所有时刻的输出向量进行加权求和, 最终得到的向量 \mathbf{v} 就是整个双向 LSTM 层最终输出的特征向量。

$$\mathbf{v} = \sum_{i=0}^t a_i \mathbf{H}_i \quad (16)$$

令 \mathbf{v}_i 表示第 i 路卷积特征经由双向 LSTM 层之后学习得到的文档表示向量, 则模型最终形成的文档表示向量 \mathbf{v}_d 可表示为 n 路卷积文档表示向量的拼接, 如式(17)所示。

$$\mathbf{v}_d = \mathbf{v}_1 \oplus \mathbf{v}_2 \oplus \cdots \oplus \mathbf{v}_n \quad (17)$$

在得到文本的最终表示向量 \mathbf{v}_d 后, 将 \mathbf{v}_d 经由全连接层与 Softmax 层进行最终的类别输出。令 c 表示某个分类, n 表示分类数, \mathbf{d} 表示文档向量 \mathbf{v}_d 经由全连接后的输出向量, d_c 表示向量 \mathbf{d} 中属于类别 c 的分量值, p_c 表示文本为分类 c 的概率, \mathbf{W}_c 与 b_c 为全连接层网络需要学习的参数, f 为非线性激活函数, 则 p_c 计算如式(18)、式(19)所示。

$$\mathbf{d} = f(\mathbf{W}_c \cdot \mathbf{v}_d + b_c) \quad (18)$$

$$p_c = \frac{\exp(d_c)}{\sum_{k=1}^n \exp(d_k)} \quad (19)$$

3 实验设置

3.1 实验环境与数据集

实验环境如表 1 所示, 所有实验均使用科研机构或学者公开的预训练词向量, 包括: Word2Vec^[23-24] ①②与 GloVe^[25] ③。所有数据集均为公开数据集, 详细信息如表 2 所示。

表 1 实验环境

实验环境	配置参数
操作系统	Windows 10
CPU/内存	Intel Core I7-9750H 6 核 2.6GHz/16 GB
GPU/显存	Nvidia GeForce RTX2070Max-Q/8 GB
运行环境	CUDA9.0 CUDNN7.6.0 Keras2.2.4 Tensorflow1.10.0

表 2 数据集详细信息

数据集	分类	训练样本	测试样本	文本平均单词数
IMDB	2	25 000	25 000	223
20NG	20	11 314	7 532	270
THUC	14	21 000	21 000	656
Fudan	20	9 804	9 833	1 410

各数据集均进行了预处理, 去除了标点符号、特殊字符, 并进行了分词, 对于传统机器学习方法去除了停止词, 对深度学习方法没有去除停止词, 数据集基本介绍如下:

(1) **IMDB**^④: 英文电影评论情感二分类数据集, 分为积极评论与消极评论, 情感极性较为明显, 分类难度较低。

(2) **20NewsGroups(20NG)**^⑤: 英文文本分类数据集, 数据集复杂, 部分分类之间相似度较高, 分类难度大。

(3) **复旦大学中文数据集(Fudan)**^⑥: 由复旦大学自然语言处理小组公开, 文本多为文献内容, 噪声特征较多, 文本篇幅长。

(4) **THUCNews 新闻数据集(THUC)**^⑦: 清华大学公开的中文新闻数据集, 噪声特征少, 由于数据全集样本数过多, 本文从中随机抽取了 42 000 条样本供实验使用。

3.2 基线方法

本文对比了如下方法:

(1) **SVM、NBSVM**: 使用了文献[4]中结合 bi-gram 特征的 SVM 算法与 NBSVM 算法。

(2) **AT-LSTM**: 使用全局信息指导局部注意力机制对 LSTM 各时刻的输出加权进行情感分类, 出自文献[7]。

(3) **BiLSTM-MHAT**: 结合 Multi-head Attention 的双向 LSTM, 出自文献[8]。

(4) **CNN-non-static、CNN-multichannel**: CNN 首次用于文本分类的经典模型, 前者为单通道, 后者为

① <https://github.com/Embedding/Chinese-Word-Vectors>

② drive.google.com/file/d/0B7XkCwpI5KDYNINUTTISS21-pQmM/view

③ <https://nlp.stanford.edu/projects/glove/>

④ <http://ai.stanford.edu/~amaas/data/sentiment/>

⑤ <http://qwone.com/~jason/20Newsgroups/>

⑥ https://download.csdn.net/download/lee0_king/10601701

⑦ <http://thuct.thunlp.org/>

双通道,出自文献[9]。

(5) **Capsule**: 胶囊神经网络在文本分类中的首次探索,使用了文献[11]中的 Capsule-B 模型。

(6) **RCNN**: 对每个特征使用双向 RNN 计算特征的前后文信息,出自文献[13]。

(7) **C-LSTM**: 使用的是文献[14]中三路卷积,不使用池化方式的模型。

(8) **CNN-LSTM-1**: 方法为文献[15]中使用两路卷积,不使用任何池化方式的模型。

(9) **CNN-LSTM-2**: 使用两路卷积,每一路连续使用卷积池化堆叠提取特征,出自文献[16]。

(10) **BRCAN**: 先用双向结构学习时序特征,再使用 CNN 结合注意力机制学习空间特征,出自文献[18]。

(11) **NN-PA**: 短语注意力机制的模型,使用了文献[19]中的 NN-PA2 方法。

(12) **C-HAN**: 结合卷积与层次注意力网络的模型,使用的是文献[20]中基于单词特征的模型。

(13) **CFC-LSTM-single**、**CFC-LSTM-multi**: 本文的混合模型,全称为 Channel Fusion CNN-LSTM, single 代表单路卷积, multi 代表多路卷积。

3.3 参数设置

实验对所有模型的超参数进行了调参范围限定,在有限的范围内搜索出当前最优的超参数组合,中英文预训练词嵌入的维度均为 300 维,模型结构、卷积核大小与原论文的设定保持相同,卷积核个数范围为 16~512, LSTM 隐藏层神经元个数范围为 16~256,全连接层神经元个数范围为 16~256,取值为 2 的整数幂。为了防止模型过拟合,在 LSTM 层与全连接层均使用了 Dropout 正则化, Dropout 取值范围为 0.2~0.5,模型的初始学习率为 0.001,优化算法使用 Adam。模型最大训练轮数为 100,数据的批尺寸大小为 64,在训练样本中,80%用于训练集,20%用于验证集。

3.4 评价指标

在分类问题中通常使用精度(P)、召回率(R)、 F_1 值、准确率(ACC)等评价模型的性能,令 TP 表示预测为正的样本, FP 表示预测为正的负样本, FN 表示预测为负的正样本, TN 表示预测为负的负样本,混淆矩阵如表 3 所示,指标计算如式(20)~式(23)所示。本文使用准确率 ACC 与综合反映分类器性能的宏平均 F_1 值评估分类效

果,宏平均 F_1 值可以看作多个二分类 F_1 指标值的算术平均值。

表 3 混淆矩阵

	预测为正	预测为负
正样本	TP	FN
负样本	FP	TN

$$P = \frac{TP}{TP + FP} \quad (20)$$

$$R = \frac{TP}{TP + FN} \quad (21)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (22)$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (23)$$

4 结果与分析

4.1 模型在公开数据集上的性能对比

表 4 是各种分类方法在公开数据集上的分类准确率与宏平均 F_1 值,第一栏是传统机器学习模型,第二栏是只学习空间或时序特征的单一模型,第三栏是混合模型,第四栏是本文的模型, single 代表单路卷积, multi 代表多路卷积。通过实验结果可以发现,本文的混合模型相比于传统机器学习模型及单一神经网络模型而言,在各个数据集上的分类性能取得了显著的提升,比传统 SVM 的准确率平均提升了 4.3%,比经典 CNN 模型 CNN-non-static 的准确率平均提升了 1%。双通道模型 CNN-multi-channel 相比于单通道模型 CNN-non-static 并没有取得稳定的性能提升,甚至出现下降,这与文献[9]实验结果相同,说明通道数的简单增加,引入更多的特征并不一定有利于分类任务,而本文的混合模型即使在一路卷积的情况下,在各数据集上的分类性能明显超过了 CNN-multichannel 使用三路卷积的模型,原因一方面是本文的混合模型结合了 LSTM 层进行时序特征学习,另一个关键的原因是本文模型使用了更为合理的双通道构建方式以及更为有效的在双通道上执行卷积的方法。由于本文的混合模型使用了双通道丰富文本表示,在卷积过程中融合了跨通道的特征,并优化了空间特征与时序特征结合的方式,在 IMDB、20NG、THUC 三个数据集上相比于其他混合模型均取得了更好的分类性能。在

Fudan 数据集上所有模型的宏平均 F_1 值明显低于准确率,这是由于 Fudan 数据集属于不平衡数据集,宏平均 F_1 值受到了少数类错分的影响。在 Fudan 数据集上,本文的混合模型分类性能不如 RCNN,主要是由于 Fudan 数据集噪声特征较多。

因本文的混合模型没有使用池化,容易受到噪声特征的干扰,而 RCNN 模型模拟了卷积的核心思想,最大池化可以充分过滤噪声特征,因此分类性能更好,所以本文的混合模型在噪声特征较少的数据集上性能表现更好,更为适用。

表 4 各种分类方法在公开数据集上的准确率与宏平均 F_1 值(%)

	IMDB		20NG		Fudan		THUC	
	ACC	F_1	ACC	F_1	ACC	F_1	ACC	F_1
SVM	88.42	88.41	77.87	76.90	88.93	72.22	92.49	92.43
NBSVM	88.97	88.96	78.49	77.51	89.67	72.83	93.09	93.03
AT-LSTM	90.26	90.26	81.74	80.98	94.17	76.48	94.87	94.86
BiLSTM-MHAT	90.38	90.38	81.87	81.05	94.26	76.55	95.05	95.05
CNN-non-static	90.14	90.14	81.70	80.94	94.10	76.42	94.90	94.89
CNN-multichannel	89.84	89.83	81.58	80.87	94.20	76.51	95.01	95.00
Capsule	90.25	90.25	81.82	81.11	94.24	76.54	95.17	95.17
CNN-LSTM-1	89.54	89.53	80.89	80.14	93.56	75.97	93.87	93.86
CNN-LSTM-2	90.28	90.28	81.91	81.09	94.23	76.53	95.12	95.11
C-LSTM	89.79	89.78	81.47	80.70	93.88	76.25	94.60	94.59
NN-PA	90.51	90.51	82.23	81.51	94.52	76.77	95.50	95.50
C-HAN	90.20	90.18	81.68	80.90	94.77	76.96	95.27	95.27
BRCAN	90.36	90.35	81.90	81.14	94.47	76.72	95.22	95.22
RCNN	90.62	90.62	82.09	81.27	95.20	77.32	95.69	95.69
CFC-LSTM-single	90.80	90.80	82.50	81.72	94.70	76.91	95.93	95.93
CFC-LSTM-multi	91.01	91.01	82.73	81.95	95.03	77.19	96.21	96.21

4.2 双通道嵌入层的有效性验证

本节以 CNN-multichannel 验证本文双通道构建方法的有效性。CNN-multichannel 使用了同种预训练词嵌入构建双通道,在训练开始时通道间的差异最小,由于权重只在一个通道更新,随着训练过程通道差异会变大,不变的通道代表了通用特征,更新的通道代表向特定任务调整;本文的双通道使用不同的词嵌入,在训练开始时通道差异最大,权重的更新经由双通道,随着训练过程通道间差异变小,均向特定任务调整。将 CNN-multichannel 的构建方法命名为 Multi-1,本文的方法命名为 Multi-2,以单通道作为基准,图 4 以 Word2Vec 构建 Multi-1,图 5 以 GloVe 构建 Multi-1,Multi-2 则由 Word2Vec 与 GloVe 分别构成。结果表明,相比于单通道,Multi-2 可以取得更为稳定的提升效果,而 Multi-1

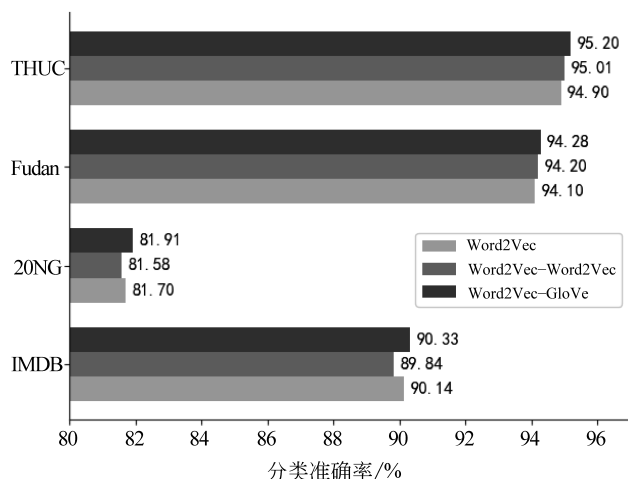


图 4 使用 Word2Vec 作为单通道的对比结果

并没有因为双通道而带来性能上的稳定提升,甚至出现下降,这是由于始终保持权重静止的通道既有

可能为特定任务带来通用特征从而提升分类效果,也有可能因为通用特征的存在导致特定任务特征的重要程度被平均化,反而不如单通道特征。Multi-2除了引入更丰富的特征以外,在双通道上同时向特定任务调整,保证了效果提升更加稳定。图4中,相比 Word2Vec,单通道最大提升 0.3 个百分点;图5中,相比 GloVe,单通道最大提升 0.53 个百分点。

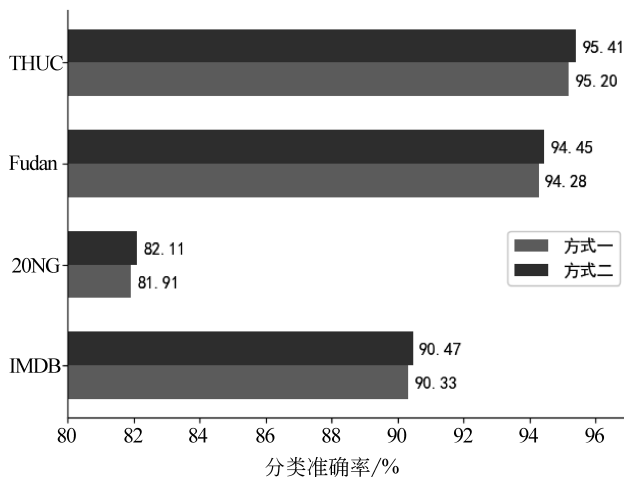


图5 使用 GloVe 作为单通道的对比结果

4.3 卷积模式对模型性能的影响

本节验证卷积模式对模型性能的影响,对以下两种卷积过程进行了对比:①使用 CNN-multi-channel 进行双通道特征学习,权重的更新在双通道同时进行;②本文的卷积方式,在每个通道进行空间特征学习,然后进行跨通道特征融合。在实验中将 CFC-LSTM-multi 中的 LSTM 层取消,保证模型处于同一规模,实验结果如图6所示。

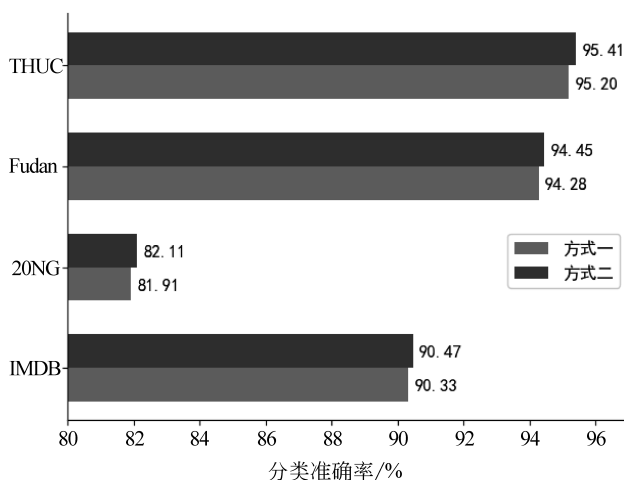


图6 不同卷积模式下的性能对比

可以发现方式二的卷积方式相比于方式一在各

数据集上取得了稳定的提升,说明了将空间特征学习过程与通道特征学习过程进行分离,相比于混合学习空间特征与通道特征更加有效,这种设计思路借鉴了谷歌的图像模型 Xception^[26],说明了在多通道表示下的文本数据,将空间特征学习过程与跨通道特征融合过程分离学习是更为有效的卷积模式,在 THUC 数据集上准确率最大提升 0.21 个百分点。

4.4 时序特征结合方式对模型性能的影响

在公开数据集的对比实验中,C-LSTM 并没有因为 LSTM 的加入,取得超越 CNN-non-static 的效果,关键的原因是由于多路卷积在拼接时,对特征时序性产生了不良影响,无法保证后续 LSTM 层的输入特征质量。本节探究了多路卷积与 LSTM 结合方式对混合模型最终性能的影响。方式一先进行多路卷积特征融合,融合后通过 LSTM 学习时序特征;方式二在每一路卷积之后直接使用 LSTM 学习时序特征。为了减少模型规模造成的干扰,通过堆叠方式一的 LSTM 以增加模型规模,然后在参数设置中指定的超参数范围内进行搜索,实验结果如图7所示。可以发现,方式一的效果在各数据集均不如方式二,最差情况下,准确率比方式二要落后 0.33 个百分点。

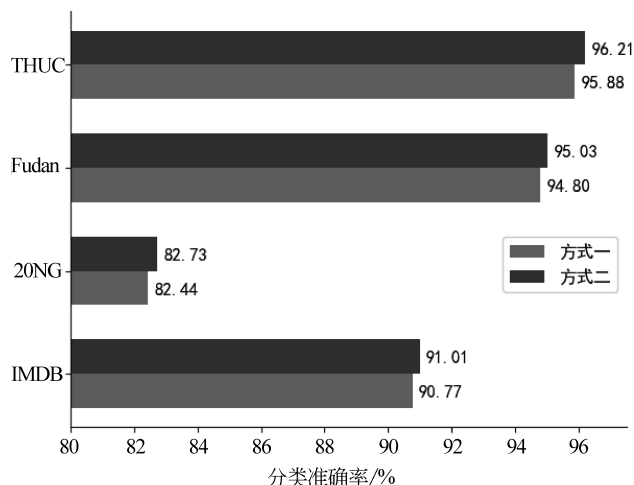


图7 两种 LSTM 结合方式对比

4.5 注意力机制对模型性能的影响

由于 LSTM 在不同时刻所形成的文本表示对最终分类任务的重要程度不同,仅利用最后时刻的输出表示最终文本并不能充分体现文本不同部分的重要程度。本节以 CFC-LSTM-single 为例,探索平

均池化、最大池化与注意力机制对分类性能造成的影响,实验结果如图 8 所示。可以发现平均池化的效果甚至不如直接使用 LSTM 最后时刻作为输出。在主题分类任务中更能突出全局关键特征的最大池化可以取得接近,甚至超过注意力机制的效果,但是在情感分类 IMDB 数据集上,注意力机制优势明显,更容易捕获对全文情感极性造成重要影响的部分。总体而言,相较于原始 LSTM,注意力机制的使用对模型的性能有着稳定的提升效果,平均提升了 0.5 个百分点。

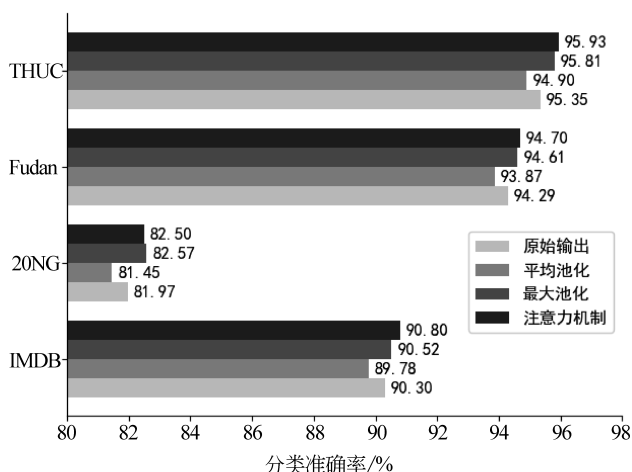


图 8 不同池化方式对模型性能的影响

4.6 模型的训练代价分析

本节对模型的训练代价进行分析,以平均特征数最多的 Fudan 数据集为例,各模型的每轮训练时间如图 9 所示。实验结果表明,传统 SVM 相关模型与单一卷积模型的训练代价明显低于使用了 RNN 结构的模型,说明了 RNN 在进行长文本建模时具有训练效率较低的缺点。在与其他混合模型的

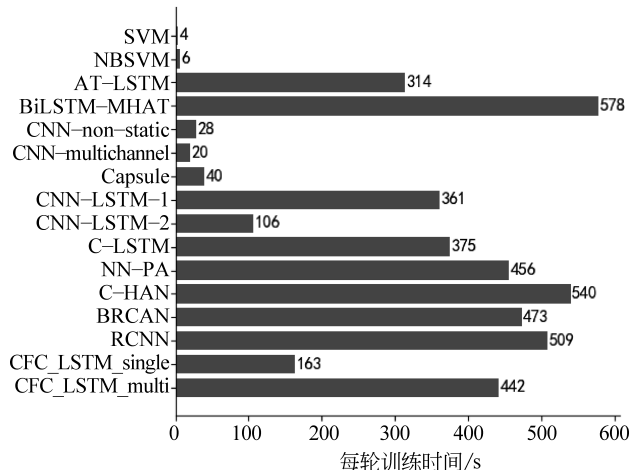


图 9 各模型在 Fudan 数据集的每轮训练时间

对比中,可以发现本文的单路模型 CFC-LSTM-single 的训练代价相对较低,但是多路模型 CFC-LSTM-multi 的训练代价较大。

为了探索造成 CFC-LSTM-multi 模型训练代价较大的具体原因,在图 10 的实验中,将 CFC-LSTM-single 的卷积部分 CFC 单独分离作为对比基准,与 CFC-LSTM-single、CFC-LSTM-multi 进行了各数据集上每轮训练时间的对比。通过图 10 的实验结果可以发现,相比于单路卷积模型 CFC 而言,混合模型的训练时间开销主要有两个方面:一是双向 LSTM 层的引入,二是卷积路数的增加。由于文本数据的特征通常较多,LSTM 的时间步往往上百甚至上千,如果使用双向 LSTM 学习文本的上下文信息,所花费的时间将更长,这也是 LSTM 作为 RNN 系列之一在处理长文本时的固有缺点。相比于单路模型,适当增加模型并联的路数,混合模型的拟合能力更强,有助于提高模型最终的分类型性能,但是模型由于并联路数的增加也带来了参数数量上的明显增多,所以需要耗费更大的时间代价去训练。

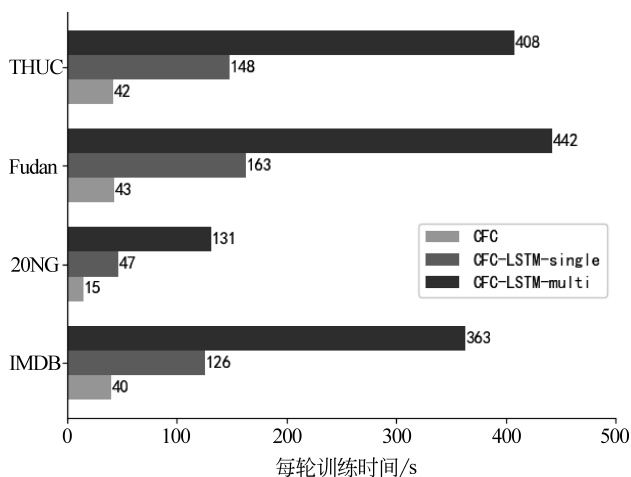


图 10 CFC-LSTM 相关模型每轮训练时间对比

4.7 长短文本数量比例对模型性能的影响

由于 THUC 数据集的样本数与分类数较多,样本中的噪声特征少,故本文选取了 THUC 数据集 10 000 条样本作为训练集,10 000 条样本作为测试集,训练集与测试集均为平衡数据集,在各分类下样本数量基本相同,避免不平衡因素带来的干扰。在此基础上,通过改变样本中长文本与短文本所占的数量比例,验证混合模型在不同长短文本数量比例之下的分类性能表现。在构建数据集时,短文本的最大特征数不超过 100,长文本的最少特征数不低

于 300,实验结果如图 11 所示。通过实验结果可以发现,本文提出的混合模型随长文本数量的增加分类性能越来越好,说明了特征少、时序性不足的短文本分类难度要高于长文本。在完全由长文本组成的数据集中,本文的混合模型性能达到最优,因此本文的模型更偏向于长文本分类任务。

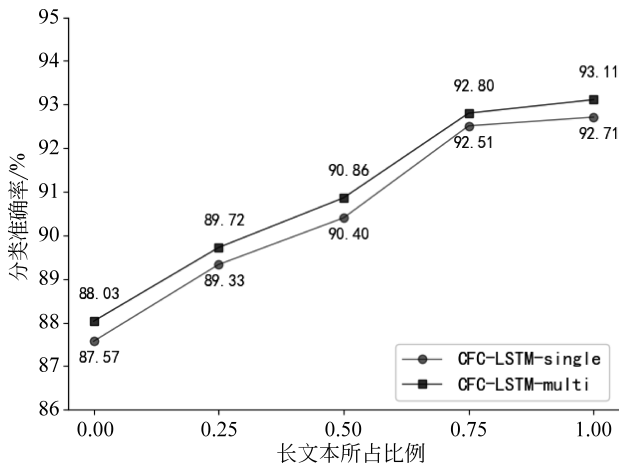


图 11 长文本数量比例对模型性能的影响

5 结束语

本文提出了一种融合通道特征的混合神经网络文本分类模型,使用基于预测与基于统计的方式构建了双通道词嵌入,在卷积中进行了通道特征融合,增强了卷积层空间特征学习能力,为了更好地与时序特征结合,模型在每路卷积后使用双向 LSTM 学习时序特征,避免了过早进行卷积特征融合对融合后的特征时序性造成破坏。实验表明,本文的混合模型在各数据集准确率相较于传统 CNN 模型平均提升了 1%。由于长文本特征多,时序信息足,本文模型更适用于长文本分类任务。未来工作中,我们将对各路卷积的重要程度进行研究,选择最为合适的卷积路数与感受野大小,降低模型的训练时间开销,并尝试用其他注意力机制进一步优化模型性能。

参考文献

[1] 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):1848-1859.
 [2] Song G, Ye Y, Du X, et al. Short text classification: A survey[J]. Journal of Multimedia, 2014, 9(5): 635-644.
 [3] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[4] Wang S, Manning C D. Baselines and bigrams: Simple, good sentiment and topic classification[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2012: 90-94.
 [5] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning[C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016: 2873-2879.
 [6] Xu G, Meng Y, Qiu X, et al. Sentiment analysis of comment texts based on BiLSTM[J]. IEEE Access, 2019, 7: 51522-51532.
 [7] Wang Y, Huang M, Zhao L. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 606-615.
 [8] Long F, Zhou K, Ou W. Sentiment analysis of text based on bidirectional LSTM with multi-head attention[J]. IEEE Access, 2019, 7: 141960-141969.
 [9] Kim Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1746-1751.
 [10] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014, 1: 655-665.
 [11] Yang M, Zhao W, Ye J, et al. Investigating capsule networks with dynamic routing for text classification[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 3110-3119.
 [12] 王盛玉,曾碧卿,商齐,等.基于词注意力卷积神经网络模型的情感分析研究[J].中文信息学报,2018,32(9):123-131.
 [13] Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015: 2267-2273.
 [14] Zhou C, Sun C, Liu Z, et al. A C-LSTM neural network for text classification[J]. Computer Science, 2015, 1(4): 39-44.
 [15] Hassan A, Mahmood A. Convolutional recurrent deep learning model for sentence classification[J]. IEEE Access, 2018, 6: 13949-13957.
 [16] Chen B, Huang Q, Chen Y, et al. Deep neural networks for multi-class sentiment classification[C]//Proceedings of IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City;

- IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2018: 854-859.
- [17] Zhang J, Li Y, Tian J, et al. LSTM-CNN hybrid model for text classification[C]//Proceedings of 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2018: 1675-1680.
- [18] Zheng J, Zheng L. A hybrid bidirectional recurrent convolutional neural network attention-based model for text classification[J]. IEEE Access, 2019, 7: 106673-106685.
- [19] 江伟, 金忠. 基于短语注意机制的文本分类[J]. 中文信息学报, 2018, 32(2): 102-109, 119.
- [20] 程艳, 叶子铭, 王明文, 等. 融合卷积神经网络与层次化注意力网络的中文文本情感倾向性分析[J]. 中文信息学报, 2019, 33(1): 133-142.
- [21] 车蕾, 杨小平, 王良, 等. 面向文本结构的混合分层注意力网络的话题归类[J]. 中文信息学报, 2019, 33(5): 93-102, 112.
- [22] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.
- [23] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013, 247(8): 1045-1050.
- [24] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013: 3111-3119.
- [25] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1532-1543.
- [26] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1251-1258.



韩永鹏(1993—),通信作者,硕士研究生,主要研究领域为深度学习。

E-mail: hanyp0314@163.com



苏航(1978—),博士,讲师,主要研究领域为机器学习与软件自动化。

E-mail: suhang@bjut.edu.cn



陈彩(1963—),副教授,主要研究领域为机器学习与软件自动化。

E-mail: chencai@bjut.edu.cn