计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200400116

用于文本分类的 CNN_BiLSTM_Attention 混合模型

吴汉瑜[1,2]　严　江[2]　黄少滨[1]　李熔盛[1]　姜梦奇[1]
1 哈尔滨工程大学计算机科学与技术学院　哈尔滨 150001
2 中电科大数据研究院有限公司提升政府治理能力大数据应用技术国家工程实验室　贵阳 550000

摘　要　文本分类是许多自然语言处理任务的基础。卷积神经网络可以提取文本的短语级特征,但是不能很好地捕获文本的结构信息;循环神经网络可以提取文本的全局结构信息,但是对关键模式信息捕获能力不足;而注意力机制能够学习到不同词或短语对文本整体语义的分布,关键的词或短语会被分配较高的权重,但是同样对全局结构信息不敏感。另外,现有模型大多只考虑词级信息,而忽略了短语级信息。针对上述模型中存在的问题,文中提出一种融合 CNN,RNN,Attention 的混合模型,该模型同时考虑不同层次的关键模式信息和全局结构信息,并把它们融合起来得到最终的文本表示,最后把文本表示输入 softmax 层进行分类。在多个文本分类数据集上进行了实验,实验结果表明该模型相较于现有模型可以实现更高的准确率。此外,还通过实验分析了模型的不同组件对模型性能的影响。

关键词:文本分类;关键模式信息;全局结构信息;混合模型;文本表示
中图法分类号　TP391.1

CNN_BiLSTM_Attention Hybrid Model for Text Classification

WU Han-yu[1,2], YAN Jiang[2], HUANG Shao-bin[1], LI Rong-sheng[1] and JIANG Meng-qi[1]
1 College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
2 Big Data Application on Improving Government Governance Capabilities National Engineering Laboratory, CETC Big Data Research Institute Co., Ltd., Guiyang 550000, China

Abstract　Text classification is the basis of many natural language processing tasks. Convolutional neural network (CNN) can be used to extract the phrase level features of text, but it can't capture the structure information of text well; Recurrent neural network (RNN) can extract the global structure information of text, but its ability to capture the key pattern information is insufficient. Attention mechanism can learn the distribution of different words or phrases to the overall semantics of text, key words or phrases will be assigned higher weights, but it is not sensitive to global structure information. In addition, most of the existing models only consider word level information, but ignore phrase level information. In view of the problems in the above models, this paper proposes a hybrid model which integrates CNN, RNN and attention. The model considers the key pattern information and global structure information of different levels at the same time, and fuses them to get the final text representation. Finally, the text representation is input to the softmax layer for classification. Experiments on multiple text classification datasets show that the model can achieve higher accuracy than the existing models. In addition, the effects of different components on the performance of the model are analyzed through experiments.

Keywords　Text classification, Key pattern information, Global structure information, Hybrid model, Text representation

1　引言

文本分类是许多自然语言处理任务的重要组成部分,它可以被应用在情感分类、问题分类、网页检索等任务上,而文本表示在文本分类中扮演着重要的角色。传统的文本分类模型主要是基于机器学习算法,如支持向量机(Support Vector Machine,SVM)[1]、朴素贝叶斯[2]等。这些算法在进行文本分类时大多存在高维、数据稀疏性等问题,性能普遍不高。近年来,随着深度学习的兴起,越来越多的研究者倾向于使用神经网络来解决文本分类的问题。

卷积神经网络(Convolutional Neural Network,CNN)可以通过滑动窗口提取到文本的 n-gram 信息,可以提取文本的短语级特征,并且最大池化技术可以挑选出文本中最具判别力的单词或短语,但是如何选择窗口的大小是一个重要的问题,窗口太小会造成结构信息丢失,窗口太大会造成参数太多,给训练带来麻烦;并且 CNN 不能很好地捕获文本的全局结构信息。循环神经网络(Recurrent Neural Network,RNN)是一种序列模型,天生适合对文本进行建模,它可以捕获文本的全局结构信息,但是对文本中的关键模式信息不敏感。注意力机制被应用在许多自然语言处理任务中,不同于最大池

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:提升政府治理能力大数据应用技术国家工程实验室开放基金
This work was supported by the Big Data Application on Improving Government Governance Capabilities National Engineering Laboratory Open Fund Project.
通信作者:吴汉瑜(wuhanyu@hrbeu.edu.cn)

化只选择最重要的信息，它可以学习到文本中每一部分信息对文本整体语义信息的贡献比例，重要的单词或短语会被分配较高的权重，因此能够更好地提取文本关键模式信息，但是它忽略了词序信息，造成不能很好地利用文本全局结构信息。

由于传统神经网络不能有效利用关键模式信息和全局结构信息，且现有的一些文本分类模型通常只考虑词级信息，而忽略了短语级信息，本文对 CNN，RNN 以及 Attention 进行了探索，提出了一种新颖的混合神经网络模型。该模型使用注意力机制作用于词的隐藏表示，得到词级关键模式信息，使用双向 LSTM 作用于词的隐藏表示，得到词级全局结构信息；然后使用 CNN 作用于词表示，得到短语级特征表示，在短语级特征表示上使用注意力机制得到短语级关键模式信息，在短语级特征表示上使用双向 LSTM 得到短语级全局结构信息。最终，文本的表示由这 4 部分信息融合而成。最后把最终文本表示输入 softmax 函数来进行文本分类。

2 相关工作

近年来，基于深度学习的神经网络模型在自然语言处理领域的许多任务中都展现了强大的性能，比如机器翻译、情感分析、文本分类等。其中应用较为广泛的神经网络模型有卷积神经网络、循环神经网络、注意力机制。

2.1 基于卷积神经网络的模型

一些研究者试图把卷积神经网络从计算机视觉领域引入自然语言处理中。Kim[3]提出用多个不同大小的卷积核提取文本的特征，之后使用最大池化技术来提取出文本中的关键信息来行句子分类。Kalchbrenner 等[4]把一种动态 k-max 池化机制与 CNN 结合，选取特征图中的前 k 大的特征值，而不是像最大池化那样只保留最大的一个特征值，并且在保留前 k 大的特征值时还保持了它们在原始序列中的顺序。该模型被应用在句子建模中并取得了不错的效果。Zhang 等[5]提出一种字符级 CNN 模型并应用在文本分类中，该模型输入的基本单位是字符而不再是单词，因此通用性更强，模型试图从字符序列中学习到文本表示。由于浅层 CNN 不能很好地处理句子中的长距离依赖，一些深层 CNN 模型被提出。例如，Conneau 等[6]提出了 VDCNN，但是随着深度的增加，模型的参数也在迅速增加，这给模型的训练和超参数的调整带来了巨大的麻烦。虽然卷积神经网络加上最大池化技术在捕获文本的短语级特征以及局部关键信息方面是有效的，但是它不能很好地捕获文本的全局结构信息。

2.2 基于循环神经网络的模型

循环神经网络可以捕获文本的结构信息，因此也被许多研究者用来进行文本分类。Tang 等[7]为了在文档中对句子之间的内在(语义或句法)关系进行编码，提出了门控循环神经网络，并用它来进行情感分类。为了把位置不变性引入 RNN 中，Wang[8]将非连续循环神经网络用于文本分类，它断开了 RNN 的信息传输，将最大传输步长限制为固定值 k，使得每一步的表示仅依赖于前 $k-1$ 个单词和当前单词，不仅减轻了文档建模的负担，而且达到了较高的性能。循环神经网络尽管擅长捕获文本的全局结构信息，但是在关键模式信息提取方面存在不足，并且普通 RNN 还存在梯度消失和梯度爆炸的问题。

2.3 基于注意力机制的模型

Yang 等[9]提出用层级注意力网络来对文档进行建模和分类，该模型首先把句子分割成单词序列，使用双向 GRU 对句子进行建模，同时引入单词级注意力机制，使句子学习到重要的单词信息，然后把文档分割成句子序列，使用双向 GRU 对文档进行建模，同时引入句子级注意力机制，使文档能够学习到重要的句子信息。这种层级文档建模的方式符合人们对文档层次结构的直观认识，也实现了最佳的性能。Lin 等[10]通过引入自注意力机制提出一种用来抽取可解释句子嵌入的模型，并把它应用在情感分类和文本蕴涵等领域中，取得了不错的效果。Liu 等[11]结合卷积神经网络和嵌套长短期记忆网络提出一种基于基于注意力的 CNLSTM 模型用于中文新闻分类；Gu 等[12]提出一种基于卷积注意力机制的神经网络模型来提取文本的局部最优情感极性，捕捉文本情感极性转移的语义信息，用在情感分类领域。但是，注意力机制忽略了词序信息，不能很好地利用文本全局结构信息。

3 CNN_BiLSTM_Attention 混合模型

本文提出的 CNN_BiLSTM_Attention 混合模型如图 1 所示。

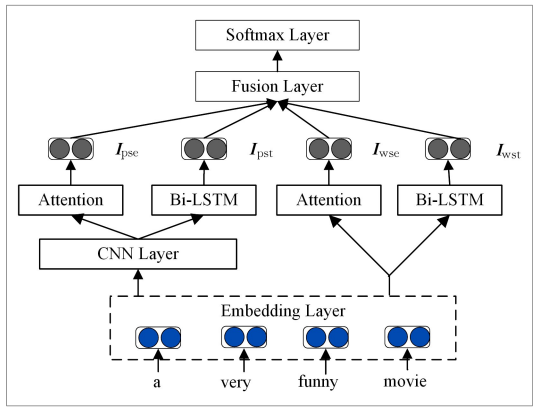

图 1 CNN_BiLSTM_Attention 混合模型

Fig. 1 CNN_BiLSTM_Attention hybrid model

3.1 词嵌入

模型的输入是一段长度为 s 的文本 T，它由一系列单词 $w_1, w_2, w_3, \cdots, w_s$ 组成。为了将文本转换成计算机能够识别的数字，本文使用 GloVe[13]词向量对神经网络的 embedding 层进行初始化。对于文本 T 的每个单词 w_i，通过查表将其映射成词向量 x_i。

$$x_i = E(w_i) \tag{1}$$

其中，$E \in R^{|V| \times d}$，$|V|$ 是词表的大小，d 是词向量的维度。

3.2 词级信息学习

词级信息的学习如图 2 所示。

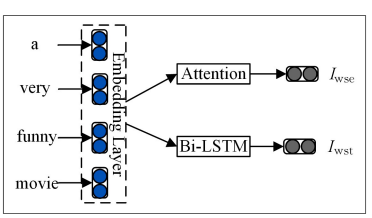

图 2 词级信息学习

Fig. 2 Learning of word level information

3.2.1 词级关键模式信息

在获取到文本中每个词的词嵌入后，使用 Attention 捕获词级关键模式信息：

$$\alpha_i=\frac{\exp(u^{\mathrm{T}}x_i)}{\sum_i\exp(u^{\mathrm{T}}x_i)} \tag{2}$$

$$I_{wse}=\sum_{i=1}^{s}\alpha_i x_i \tag{3}$$

其中，u 是需要训练的注意力参数，u^{T} 是 u 的转置。

3.2.2 词级全局结构信息

为了解决普通 RNN 存在的梯度消失和梯度爆炸的问题，研究者提出了长短期记忆网络(Long Short-Term Memory，LSTM)，其在提取文本的全局结构信息方面表现不错。它引入了输入门、遗忘门和输出门，具有长期记忆功能，它在第 t 个时间步的状态转移公式如式(4)－式(9)所示：

$$i_t=\sigma(W_i[h_{t-1},x_t]+b_i) \tag{4}$$

$$f_t=\sigma(W_f[h_{t-1},x_t]+b_f) \tag{5}$$

$$q_t=\tanh(W_q[h_{t-1},x_t]+b_q) \tag{6}$$

$$o_t=\sigma(W_o[h_{t-1},x_t]+b_o) \tag{7}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot q_t \tag{8}$$

$$h_t=o_t\odot\tanh(c_t) \tag{9}$$

其中，i_t 决定哪些值需要被更新，f_t 决定哪些值需要被遗忘，o_t 表示在 t 时刻模型的输出，h_t 是 t 时刻的隐藏状态。W_i，W_f，W_q，W_o 权重矩阵，b_i，b_f，b_q，b_o 是对应的偏置。σ 表示激活函数，$\odot$ 表示元素对应乘法。

使用双向 LSTM 作用在词向量上，来捕获词级全局结构信息 I_{wst}：

$$\overrightarrow{I_{\mathrm{wst}}}=\overrightarrow{LSTM}(x_1,x_2,x_3,\cdots,x_s) \tag{10}$$

$$\overleftarrow{I_{\mathrm{wst}}}=\overleftarrow{LSTM}(x_1,x_2,x_3,\cdots,x_s) \tag{11}$$

$$I_{\mathrm{wst}}=[\overrightarrow{I_{\mathrm{wst}}};\overleftarrow{I_{\mathrm{wst}}}] \tag{12}$$

词级全局结构信息 I_{wst} 是由前向 LSTM 的最终状态 $\overrightarrow{I_{\mathrm{wst}}}$ 与反向 LSTM 的最终状态 $\overleftarrow{I_{\mathrm{wst}}}$ 拼接而成。

3.3 短语级信息学习

短语级信息的学习如图 3 所示。

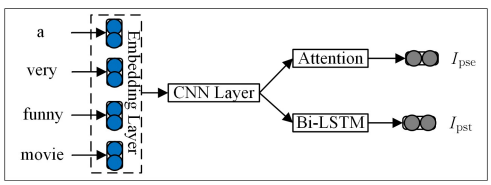

图 3 短语级信息学习

Fig. 3 Learning of phrase level information

由于 CNN 能够捕获文本的短语级特征，因此首先在词向量上使用 CNN 来捕获短语级特征 D。使用卷积核 $F\in\mathcal{R}^{d\times n}$，其中 d 是词向量的维度，n 是窗口大小。输入句子可以看作一个矩阵 $X\in\mathcal{R}^{d\times s}$，卷积核 F 与矩阵 X 进行卷积得到一个特征图 $d\in\mathcal{R}^{s-n+1}$，特征图 d 中的每一个元素 d_k 是通过以下方式计算得到的：

$$d_k=\sigma(\sum_{i,j}(F\odot X_{k:k+n-1})) \tag{13}$$

其中，σ 表示激活函数，$\odot$ 表示元素对应(element-wise)乘法，$X_{k:k+n-1}$ 表示输入词向量矩阵的第 k 列到第 $k+n-1$ 列。

本文使用相同填充，使得卷积层的输入长度与输出长度相同，即 $d\in\mathcal{R}^s$。使用 m 个卷积核，得到 m 个特征图，即短语级特征 $D\in\mathcal{R}^{m\times s}$。

3.3.1 短语级关键模式信息

在得到短语级特征后，使用 Attention 作用于短语级特征 D 来捕获短语级关键模式信息 I_{pse}。

$$\beta_i=\frac{\exp(v^{\mathrm{T}}d_i)}{\sum_i\exp(v^{\mathrm{T}}d_i)} \tag{14}$$

$$I_{\mathrm{pse}}=\sum_{i=1}^{s}\beta_i d_i \tag{15}$$

其中，v 是需要学习的注意力参数，v^{T} 是 v 的转置，d_i 是短语级特征 D 的第 i 个表示向量。

3.3.2 短语级全局结构信息

同样使用双向 LSTM 作用于短语级特征 D，得到短语级全局结构信息 I_{pst}。

$$\overrightarrow{I_{\mathrm{pst}}}=\overrightarrow{LSTM}(d_1,d_2,d_3,\cdots,d_s) \tag{16}$$

$$\overleftarrow{I_{\mathrm{pst}}}=\overleftarrow{LSTM}(d_1,d_2,d_3,\cdots,d_s) \tag{17}$$

$$I_{\mathrm{pst}}=[\overrightarrow{I_{\mathrm{pst}}};\overleftarrow{I_{\mathrm{pst}}}] \tag{18}$$

短语级全局结构信息 I_{pst} 是由前向 LSTM 的最终状态 $\overrightarrow{I_{\mathrm{pst}}}$ 与反向 LSTM 的最终状态 $\overleftarrow{I_{\mathrm{pst}}}$ 拼接而成。

3.4 特征融合

对于词级关键模式信息 I_{wse}、词级全局结构信息 I_{wst}、短语级关键模式信息 I_{pse}、短语级全局结构信息 I_{pst} 的融合，本文探索了两种融合方式来得到最终文本的表示 I_T：静态融合和基于注意力机制的动态融合。

对于静态融合，本文对 4 种信息进行加权平均：

$$I_T=(I_{\mathrm{wse}}+I_{\mathrm{wst}}+I_{\mathrm{pse}}+I_{\mathrm{pst}})/4 \tag{19}$$

对于动态融合，本文使用注意力机制作用于 4 种信息，令 I_{wse}，I_{wst}，I_{pse}，I_{pst} 分别为 I_1，I_2，I_3，I_4，即：

$$\gamma_i=\frac{\exp(z^{\mathrm{T}}I_i)}{\sum_i\exp(z^{\mathrm{T}}I_i)} \tag{20}$$

$$I_T=\sum_{i=1}^{4}\gamma_i I_i \tag{21}$$

其中，z 是注意力机制的参数，z^{T} 是 z 的转置。

3.5 输出

文本 T 的表示向量 I_T 可被认为是文本的高层次表示，因此把文本表示向量 I_T 发送给 softmax 分类器，来得到每个类别对应的概率。

$$p=softmax(W_c I_T+b_c) \tag{22}$$

其中，W_c 是 softmax 分类器的权重，b_c 是偏置。

为了训练得到模型参数，将训练目标的最小化损失函数定义为：

$$L(y,p)=-\sum_{i=1}^{N}\sum_{j=1}^{C}y_{ij}\log(p_{ij}) \tag{23}$$

其中，N 是数据集中样本的数量，C 是类别的数量，y_{ij} 是第 i 个样本在第 j 个类别上的真实值，p_{ij} 是第 i 个样本在第 j 个类别上神经网络的预测概率值。

4 实验

4.1 数据集介绍

(1)MR 数据集(Pang 等[14])是一个二分类的电影评论数据集，它由 5331 个积极样本和 5331 个消极样本组成。

(2)SUBJ 数据集(Pang 等[15])是一个二分类的主题数据集，数据集中所有的句子都被分为主观的(objective)和客观的(subjective)。

(3)TREC 数据集(Li 等[16])是一个六分类的问题分类数据集，数据集中的样本标签为 abbreviation，entity，description，location，numeric，human。

(4)CR 数据集(Hu 等[17])是一个包含 Customer Reviews 的二分类数据集，它的标签分别为 Positive 和 Negative。

(5)斯坦福情感树库(Stanford Sentiment Treebank, SST5)数据集(Socher 等[18])是一个五分类电影评论数据集,它的标签由非常消极的、消极的、中性的、积极的、非常积极的组成。

(6)AGNews 数据集(Zhang 等[5])是一个新闻分类数据集,它的标签分别为 World, Sports, Business, Sci/Tech。

更详细的数据集统计信息如表 1 所列。

表 1 数据集统计信息

Table 1 Statistics of experiment datasets

Dataset	*Class*	*Length*	*Size*	*Test Size*
MR	2	20	10662	CV
SUBJ	2	23	10000	CV
TREC	6	10	5952	500
CR	2	19	3775	CV
SST5	5	18	11855	2210
AGNews	4	45	127600	7600

表 1 中,*Class* 代表数据集中的类别数,*Length* 代表数据集中样本的平均长度,*Size* 代表数据集中样本的总数,*Test size* 代表测试集的数量(其中 CV 表示数据集中没有划分训练集和测试集,所以使用十折交叉验证)。

4.2 实验设置

实验环境:操作系统 Windows10,内存 8 GB;CPU:Intel Core i7;使用深度学习框架 Keras 开发,且其底层支持为 TensorFlow。

超参数:使用 200 维 GloVe(Pennington 等[13])词向量,对于不在 GloVe 中的单词,使用[-1,1]之间的均匀分布对其词向量进行初始化。为防止模型过拟合,分别在词嵌入层之后和卷积层之后增加 dropout 层,dropout 率设置为 0.5。实验中详细参数设置如表 2 所列。

表 2 实验参数设置

Table 2 Experimental parameter setting

参数名称	参数值
词向量维度	200
卷积核大小	3
卷积核个数	200
Dropout_pro(丢弃率)	0.5
BiLSTM 隐藏单元个数	100
Optimizer(优化器)	Adam Optimizer[19]
学习率	5×10^{-5}
batch_size(批次大小)	64

4.3 实验结果与分析

4.3.1 对比实验

本文选择的基线模型主要分为 4 部分。

(1)基于 CNN 的模型,其中包括 kim[3]于 2014 年提出的 CNN-static 和 CNN-non-static,其区别在于是否对预训练词嵌入进行微调;Yin 等[20]于 2016 年提出的 MVCNN,Conneau 等[6]于 2016 年提出的 VDCNN。

(2)基于 RNN 的模型,其中包括 Cho 等[21]于 2014 年提出的 LSTM 和 BiLSTM。

(3)基于 Attention 的模型,包括 Lin 等[10]于 2017 年提出的 Self-attentive。

(4)混合模型,包括 Zhou 等[22]于 2015 年提出的 C-LSTM,Wang 等[23]于 2017 年提出的 Conv-RNN,Zhou 等[24]2016 年提出的 BiLSTM-2DCNN,以及 Gu 等[12]2020 年提出的 CNN_attention_LSTM,Zhou 等[25]2018 年提出的 BGRU-CNN,以及 Zheng 等[26]2019 年提出的 DC-BiGRU_CNN。

实验结果如表 3 所列,评估指标为分类准确率(accuracy)。准确率是文本分类领域中常用的评估指标,它指的是分类正确的样本占总样本的比例,分类准确率越高,则分类器的性能越高。

表 3 准确率对比

Table 3 Comparison of accuracy

(单位:%)

Model	MR	SUBJ	TREC	CR	SST5	AGNews
CNN-non-static	81.5	93.4	93.6	84.3	48.0	92.3
CNN-static	81.0	93.0	92.8	84.7	45.5	—
MVCNN	—	93.69	—	—	49.6	—
VDCNN	—	88.2	85.4	—	49.6	91.3
LSTM	75.9	89.3	86.8	78.4	—	86.1
BiLSTM	79.3	90.5	89.6	82.1	—	86.2
Self-attentive	80.1	92.5	—	—	47.2	—
C-LSTM	—	—	94.6		49.2	—
Conv-RNN	81.9	94.13	—	—	51.67	—
BiLSTM-2DCNN	82.3	94.0	96.1	—	52.4	—
CNN_Attention_LSTM	82.3	—	—	—	48.0	—
BGRU-CNN	82.3	94.4	—	86.0	50.2	—
DC-BiGRU_CNN	83.4	94.9	**96.2**		51.9	—
Ours-static	83.0	94.4	95.3	85.2	50.1	92.2
Ours-dynamic	**83.8**	**95.1**	**96.2**	**87.0**	**52.6**	**93.8**

其中,Ours-static 指的是本文提出的静态加权平均进行特征融合的模型,Ours-dynamic 指的是本文提出的注意力机制进行特征融合的模型。

从结果中可以看出,本文提出的混合模型在 6 个数据集上均取得了最佳的性能。基于 CNN,RNN,Attention 的混合模型大多比单独使用 CNN,RNN,Attention 模型的性能更高,这体现出了混合模型在提取文本特征能力上的优势。而在混合模型中,本文提出的模型能够从不同层次捕获文本的特征,所以达到了最佳的性能。

4.3.2 不同层次的信息对模型性能的影响

为了探究不同层次的信息对模型性能的影响,本文在 MR,SUBJ,TREC 这 3 个数据集上进行了对比实验:仅使用词级信息(词级关键模式信息、词级全局结构信息,见图 2),仅使用短语级信息(短语级关键模式信息、短语级全局结构信息,见图 3),使用词级信息和短语级信息融合(见图 1)。信息的融合方式均选择注意力机制。得到的结果如表 4 所列。

表 4 不同层次的信息对模型性能的影响

Table 4 Influence of different levels of information on performance

(单位:%)

使用的信息	MR	SUBJ	TREC
词级信息	82.7	94.3	95.6
短语级信息	83.2	94.4	95.3
词级信息、短语级信息	**83.8**	**95.1**	**96.2**

从表 4 中可以看出,仅使用词级特征与仅使用短语级特征的性能相差不大,而同时使用词级特征和短语级特征则可以进一步改进模型的性能。因此,同时使用不同层次的关键模式信息和全局结构信息能够改进文本分类的性能。

4.3.3 最大池化与注意力机制提取关键模式信息的比较

为了探究最大池化技术以及注意力机制在关键模式信息方面的提取能力,本文在 MR,SUBJ,TREC 3 个数据集上进行了一组对比实验:在其他参数不变的情况下,把词级关键模式信息和短语级关键模式信息的获取方式从 Attention 更改为 Max-pooling,得到的结果如图 4 所示。

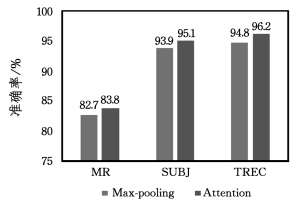

图 4 最大池化与注意力机制的比较

Fig. 4 Comparison of max-pooling and attention mechanism

从图 4 中可以看出，基于注意力机制的模型在 3 个数据集上的准确率全部超越了基于最大池化的模型，这说明注意力机制能够更好地学习到关键模式信息，这可能是因为最大池化技术只选取特征的最大值，而注意力机制能够为不同的信息分配不同的权重，所适用的场景更加广泛，因此拥有更好的性能。

4.3.4 卷积核尺寸对模型性能的影响

为了研究不同卷积核窗口大小对模型性能的影响，本文在 MR 数据集上进行实验，选取的卷积核尺寸分别为 1，3，5，7，得到的结果如图 5 所示。

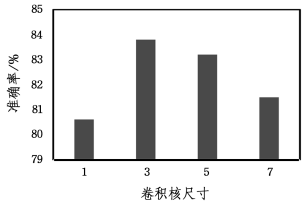

图 5 卷积核尺寸对性能的影响

Fig. 5 Influence of convolution kernel size on performance

实验结果表明，卷积核窗口大小为 3 时模型可以取得最佳的性能。窗口小于 3 时可能由于特征捕获不足导致性能较低，而随着窗口大小的增加，卷积操作的特征捕获能力增强，但是使用更大的卷积核意味着更多的模型参数，消耗更多的时间和空间，导致模型性能开始下降。

4.3.5 不同循环单元对模型性能的影响

本文探究了 3 种不同类型的循环单元(Recurrent Units)对结构信息提取能力的影响，包括 Simple RNN，GRU，LSTM。本文在 MR 数据集上进行实验，实验结果如图 6 所示，评估指标为分类准确率。这里的循环单元均指双向的。实验中除循环单元不同之外，其他设置均相同。

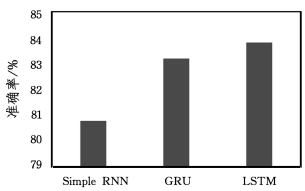

图 6 不同循环单元的性能比较

Fig. 6 Performance comparison of different recurrent units

可以发现，在 MR 数据集上，LSTM 达到了最好的效果，总体上 GRU 的性能与 LSTM 相差不大，而 Simple RNN 的性能远远落后于 LSTM 和 GRU，这可能与 Simple RNN 的梯度消失和梯度爆炸问题有关。

结束语 本文结合 CNN，RNN，Attention 的优点，提出了一种 CNN_BiLSTM_Attention 混合神经网络模型来进行文本分类，该模型能够捕获不同层次的关键模式信息和全局结构信息并对它们进行融合。实验结果显示，本文所提出的模型在几个公开的文本分类数据集上达到了最高的性能。下一步，将把本模型应用在处理中文四险一金领域语料上，为构建四险一金领域知识图谱提供支持。

参 考 文 献

[1] JOACHIMS T. Text categorization with support vector machines: Learning with many relevant features[C]// European Conference on Machine Learning. Berlin: Springer, 1998: 137-142.

[2] CHEN Z, SHI G, WANG X. Text classification based on Naive Bayes algorithm with feature selection[J]. International Information Institute (Tokyo). Information, 2012, 15(10): 4255.

[3] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv: 1408.5882, 2014.

[4] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences[J]. arXiv: 1404.2188, 2014.

[5] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification[C]// Advances in Neural Information Processing Systems. 2015: 649-657.

[6] CONNEAU A, SCHWENK H, BARRAULT L, et al. Very deep convolutional networks for text classification[J]. arXiv: 1606.01781, 2016.

[7] TANG D, QIN B, LIU T. Document modeling with gated recurrent neural network for sentiment classification[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1422-1432.

[8] WANG B. Disconnected recurrent neural networks for text categorization[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 2311-2320.

[9] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 1480-1489.

[10] LIN Z, FENG M, SANTOS C N, et al. A structured self-attentive sentence embedding[J]. arXiv: 1703.03130, 2017.

[11] LIU Y, ZHAI D H, REN Q N. News Text Classification Based on CNLSTM Model with Attention Mechanism[J]. Computer Engineering, 2019, 45(7): 303-308, 314.

[12] GU J H, PENG W T, LI N N, et al. Sentiment classification method based on convolution attention mechanism[J]. Computer Emgineering and Design, 2020, 41(1): 95-99.

[13] 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 (EMNLP). 2014: 1532-1543.

[14] PANG B, LEE L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales[C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005: 115-124.

[15] PANG B, LEE L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 271.

(下转第 34 页)