计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(9):171−177 [doi: 10.15888/j.cnki.csa.007495]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结合 TFIDF 的 Self-Attention-Based Bi-LSTM 的垃圾短信识别①

吴思慧[1], 陈世平[2]

[1](上海理工大学 光电信息与计算机工程学院, 上海 200093)
[2](复旦大学 上海市数据科学重点实验室, 上海 201203)
通讯作者: 吴思慧, E-mail: 19916546846@163.com

摘　要: 随着手机短信成为人们日常生活交往的重要手段, 垃圾短信的识别具有重要的现实意义. 针对此提出一种结合 TFIDF 的 self-attention-based Bi-LSTM 的神经网络模型. 该模型首先将短信文本以词向量的方式输入到 Bi-LSTM 层, 经过特征提取并结合 TFIDF 和 self-attention 层的信息聚焦获得最后的特征向量, 最后将特征向量通过 Softmax 分类器进行分类得到短信文本分类结果. 实验结果表明, 结合 TFIDF 的 self-attention-based Bi-LSTM 模型相比于传统分类模型的短信文本识别准确率提高了 2.1%–4.6%, 运行时间减少了 0.6 s–10.2 s.

关键词: 垃圾短信; 文本分类; self-attention; Bi-LSTM; TFIDF

引用格式: 吴思慧,陈世平.结合 TFIDF 的 Self-Attention-Based Bi-LSTM 的垃圾短信识别.计算机系统应用,2020,29(9):171–177. http://www.c-s-a.org.cn/1003-3254/7495.html

Spam Message Recognition Based on TFIDF and Self-Attention-Based Bi-LSTM

WU Si-Hui[1], CHEN Shi-Ping[2]

[1](School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)
[2](Shanghai Key Laboratory of Data Science, Fudan University, Shanghai 201203, China)

Abstract: Mobile phone text messaging has become an increasingly important means of daily communication, so the identification of spam messages has importantly practical significance. A self-attention-based Bi-LSTM neural network model combined with TFIDF is proposed for this purpose. The model first inputs the short message to the Bi-LSTM layer in a vector manner, after feature extraction and combining the information of TFIDF and self-attention layers, the final feature vector is obtained. Finally, the feature vector is classified by the Softmax classifier to obtain the classification result. The experimental results show, compared with the traditional classification model, the self-attention-based Bi-LSTM model combined with TFIDF improves the accuracy of text recognition by 2.1%–4.6%, and the running time is reduced by 0.6 s–10.2 s.

Key words: spam message; text categorization; self-attention; Bi-LSTM; TFIDF

① 基金项目: 国家自然科学基金 (61472256, 61170277, 61003031); 上海重点科技攻关项目 (14511107902); 上海市工程中心建设项目 (GCZXL14014); 上海市一流学科建设项目 (S1201YLXK, XTKX2021.); 上海市数据科学重点实验室开发课题 (201609060003); 沪江基金 (A14006); 沪江基金研究基地专项 (C14001)

Foundation item: National Natural Science Foundation of China (61472256, 61170277, 61003031); Key Science and Technology Project of Shanghai Municipality (14511107902); Engineering Research Center Construction Project of Shanghai Municipality (GCZXL14014); Top Level Discipline Construction Project of Shanghai Municipality (S1201YLXK, XTKX2021); Open Fund of Shanghai Key Laboratory of Data Science (201609060003); Hujiang Foundation (A14006); Special Fund for Research Base of Hujiang Foundation (C14001)

收稿时间: 2019-12-12; 修改时间: 2020-01-03, 2020-01-07; 采用时间: 2020-01-14; csa 在线出版时间: 2020-09-04

21 世纪以来, 手机用户不断增加, 特别是智能手机的使用越来越多, 人们可以通过短信快速高效的获取信息, 但随之而来的是垃圾短信的泛滥, 垃圾短信不仅仅影响到人们正常的手机使用和体验, 更主要的是垃圾短信会带来严重的安全隐患, 很多不法分子通过垃圾短信获取用户的私人信息, 危害到用户隐私安全. 因此, 垃圾短信的识别具有重要的现实意义. 治理垃圾短信不仅需要有关部门的持法监督和相应手机安全厂商的屏蔽, 同时应该利用先进的技术, 直接在源头上消灭垃圾短信.

目前常用的垃圾短信识别的方法主要包括基于黑白名单的方法, 基于规则的方法和基于短信内容的方法这样 3 种[1], 前两种方法要人工手动添加发送垃圾短信号码的名单或者手动添加与垃圾短信对应关键词, 由于手动添加的数据量有限且效率低, 因此目前主要是使用基于短信内容的方法来进行短信识别, 即将文本分类技术用于识别垃圾短信.

文本分类是计算机应用于根据特定的分类系统或者标准自动分类文本[2,3]. 随着深度学习在自然语言预处理领域的应用, 相对于传统的文本分类算法如朴素贝叶斯, 支持向量机等[4–6], 深度学习在文本分类上获得了令人满意的结果. 目前长短时记忆网络 (Long Short-Term Memory, LSTM) 已经广泛应用在文本分类里面, 与循环神经网络 (Recurrent Neural Network, RNN) 相比, LSTM 网络采用了特殊隐式单元, 因此更适合于处理长期依赖关系, 很好的解决了 RNN 的梯度消失或者梯度爆炸的问题, 可以更好的获取文本的全局特征信息. 以 LSTM 网络为基础的双向循环神经网络 (Bidirectional LSTM, Bi-LSTM) 网络是由向前的 LSTM 和向后的 LSTM 组成[7–13], 在处理文本分类上面可以更好的捕捉双向的语义依赖, 对于词向量[14] 的依赖更少, 从而提高文本的分类效率.

专家学者根据基于对人类视觉的研究, 提出注意力 (attention) 机制[15], 目前 Attention 机制已经引入到自然语言预处理领域[16,17], 学习并重点关注目标区域, 使得模型在有效资源的情况下关注重点消息. Attention 机制通常结合编码解码 (encoder-decoder) 模型使用, 应用场景十分广泛, 因此随后出现多种注意力机制的变形, 如自注意力 (self-attention) 机制.

根据以上背景本文提出一种结合 TFIDF 的 self-attention-based Bi-LSTM 神经网络模型. 首先, 使用 Word2Vec[18–20] 将短信文本处理成词向量形式, 随后使用 Bi-LSTM 模型对词向量形式的短信文本的特征信息进行提取, 接着引入自注意机制, 并结合 TFIDF 模型, 对重点词汇进行加权, 最后将输出的特征向量输入 Softmax 分类器得到分类结果. 该模型不仅能够充分利用上下文本信息来进行短信文本特征提取, 还能在对短信文本进行分类时分辨中不同词语的重要程度从而实现对重点词语的提取, 与未使用自注意机制和 TFIDF 模型的 Bi-LSTM 模型相比, 在对垃圾短信和正常短信分类时的分类准确率, 召回率, F1 值, 运行时间等值上有较大的提升, 其中分类准确率达到了 90.1%, 召回率达到了 90.5%, F1 值达到了 90.2%, 说明该模型在对短信文本处理时具有更好的分类能力[21], 同时从实验结果的准确率与训练集大小的性能趋势曲线上来看, 该模型始终优于其他模型, 该模型相较于其他模型需要更少的训练数据就可以到达较高的准确率. 实验结果验证了结合 TFIDF 的 self-attention-based Bi-LSTM 神经网络模型的可行性和有效性.

本文的主要贡献如下:

(1) 将 Bi-LSTM 模型运用到垃圾短信识别中, 既可以利用过去的信息也可以利用将来的信息.

(2) 将自注意机制和 TFIDF 模型相结合, 进一步加强重点词汇的权重, 增强分类效果.

(3) 在 BI-LSTM 模型中引用自注意机制和 TFIDF 模型, 进一步获取对短信文本分类结果产生影响的重点词语的特征.

1 模型构建

本文结合 TFIDF 的 self-attention-based Bi-LSTM 神经网络总体模型如图 1 所示, 模型包含词向量输入层, Bi-LSTM 网络层, 结合 TFIDF 的自注意层, Softmax 层. 模型流程主要是短信文本以词向量的方式输入到 Bi-LSTM 层, 经过特征提取并结合 TFIDF 和自注意层的重点词汇加权从而获得最后的特征向量, 最后通过 Softmax 分类器对最终的特征向量进行分类从而得到短信文本分类结果.

1.1 RNN 模型

RNN 能处理序列问题, 允许信息持久化, 即将上一刻运算结果添加到当前计算的输入中去, 从而实现了“考虑上下文信息”的功能, 可用于一段段连续的语义, 一段段连续的段落等, RNN 包含循环结构, 例如一

个 tanh 层. 具体运行过程是 t 时刻输入当前信息x_t并由神经网络模块 A 接收, 之后由 A 得到 t 时刻的输出h_t, 并且将当前时刻的部分信息传递到下一刻 t+1, RNN 结构如图 2 所示.

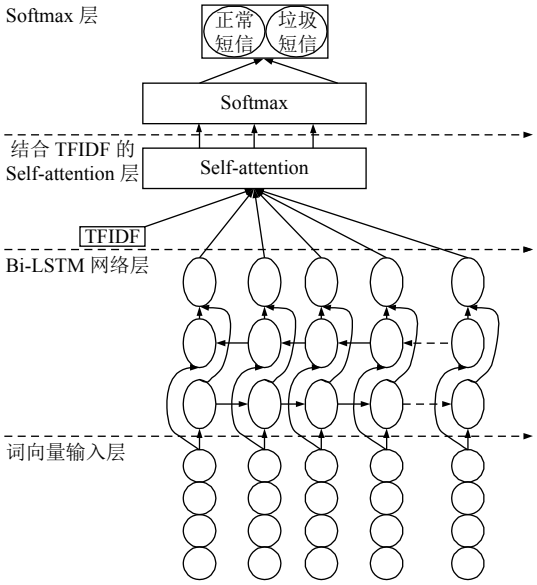

图 1 结合 TFIDF 的 self-attention-based Bi-LSTM 模型

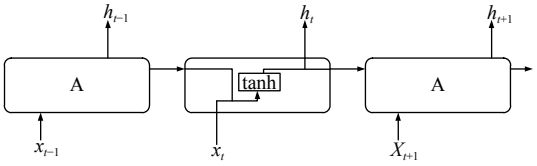

图 2 RNN 结构

1.2 LSTM 模型

在 RNN 模型中仅靠一条线来记录所有的输入信息其工作效果并不是很理想, 很难完美的处理具有长期依赖的信息, 如一段很长的英语句子, RNN 很难记住前面主语的时态形式从而在句子后面选择相应的合适的时态. 因此在 RNN 模型的基础上, 出现了 LSTM. LSTM 是一种特殊的循环神经网络, 可以学习长期依赖信息, 其结构和传统的 RNN 结构相同, 只是重复模块 A 结构更加复杂些, 多了一个单元控制器 Cell, 其能够判断信息是否有用, 从而解决了 RNN 常有的梯度消失或者梯度爆炸的问题. LSTM 结构内部主要包括输入门i_t, 遗忘门f_t, 输出门o_t和 Cell 状态更新向量c_t等部分, LSTM 结构如图 3 所示.

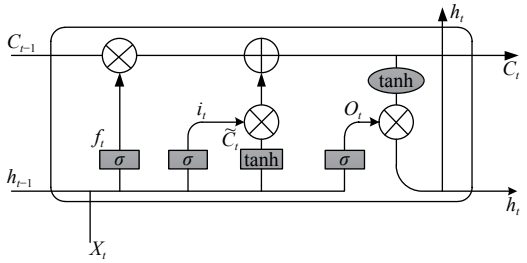

图 3 LSTM 结构

(1) 遗忘门f_t决定需要舍弃的信息部分, 其计算公式如下:

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

其中, W_f和b_f分别表示遗忘门的权重矩阵和遗忘门的偏置矩阵, σ为激活函数, h_{t-1}表示历史信息, x_t表示当前流入 Cell 中新的信息, x_t作用是为了根据当前输入的新的信息来决定要忘记哪些历史信息, 将上一时刻的输出h_{t-1}和本时刻的输入x_t两个向量拼接起来, 通过激活函数输出一个在 0 到 1 之间的数值, 0 表示完全抛弃, 1 表示完全保留, 同时, 绝大部分数值都是接近 0 或者 1 的, 这个数据决定要遗忘多少历史信息, 0 表示完全抛弃, 1 表示完全保留.

(2) 输入门i_t处理当前位置的输入, 确定什么样的新信息被存放在 Cell 中, 此处包含两个部分, 首先, Sigmoid 层的“输入门层”会决定更新哪些值, 接着 tanh 层会建立一个新的候选值向量$\tilde{C}_t$, 在获得了输入门和遗忘门系数之后则更新当前的 Cell 状态, C_{t-1}更新为C_t, 其计算公式如下:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

其中, W_i和b_i分别表示输入门的权重, tanh 为激活函数.

(3) 输出门控制哪些信息用于此刻的输出, 输出门是由历史信息h_{t-1}和新的信息x_t来决定的, 此处包含两个部分, 首先, 运行一个 Sigmoid 层, 主要用于决定 Cell 状态的哪个部分将被输出出去, 将 Cell 状态通过一个 tanh 层进行处理, 得到一个在−1 到 1 之间的值, 将这个值乘以 Sigmoid 门的输出, 最后模型将仅输出确定要输出的部分, 其计算公式如下:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

其中, W_o和b_o分别表示输入门的权重矩阵和输入门的偏置矩阵.

1.3 Bi-LSTM 模型

Bi-LSTM 是对 LSTM 的改进, 因为 LSTM 是序列化处理信息, 所以在信息处理上有先后的顺序, 也就是常常忽略下文信息, 不能综合上下文的信息, 例如: “作业写完了, 我想_手机”, 要在横线中填词, 如果只考虑前面的信息, 那么横线可以填“睡觉”, “玩”, “看电视”等, 但是如果同时结合后面的信息“手机”一词, 那么横线处填“玩”的概率最大, 而 Bi-LSTM 模型包含一个前向的 LSTM 模型和一个后向的 LSTM 模型, 可以获取足够的上下文信息, 并且两个模型都被连接到相同的输出层, Bi-LSTM 结构如图 4 所示.

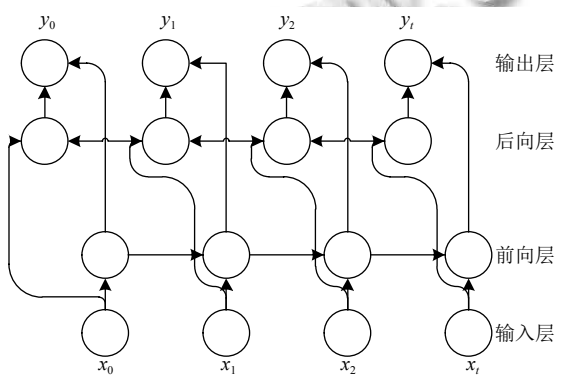

图 4 Bi-LSTM 结构

图 4 中前向的 LSTM 模型捕捉当前时刻的前文特征信息, 后向的 LSTM 模型捕捉当前时刻的后文特征信息, Bi-LSTM 模型 t 时刻的输入的计算公式如下:

$$H_t = \left[\overleftarrow{h_t}, \overrightarrow{h_t}\right] \tag{7}$$

其中, H_t, $\overleftarrow{h_t}$, $\overrightarrow{h_t}$分别表示 BI-LSTM 模型, 前向 LSTM 模型, 后向 LSTM 模型在 t 时刻的输出.

1.4 TFIDF 模型

TFIDF 是一种于咨询检索与咨询勘探的常用加权技术, 主要用以评估一字词对于一个文集的重要程度或者对于一个语料库中的其中一份文件的重要程度, 字词的重要程度与它在文档中出现的次数成正比, 与它在整个语料库出现次数成反比. 该模型主要包括: 词频 (*TF*) 和逆文档频率 (*IDF*) 两个部分, *TF* 表示某个词w_n在文档d_m中的出现频率, IDF 代表该词的类别区分, 计算公式如下:

$$TF(w_n, d_m) = \frac{f_{n,m}}{\sum_{i=1}^{N} f_{i,m}} \tag{8}$$

$$IDF(w_n, d_m) = \log\left(\frac{D}{D_{w_n}+1}\right) \tag{9}$$

其中, d_m为文档集 $D=\{d_1,,d_2,\cdots,d_m,\cdots,d_M\}$中任意一篇, M 为文档集中文档的总数, d_m有词汇集合 $w=\{w_n, w_n,\cdots,w_n,\cdots,w_N\}$, N 为每篇文档的词汇总数, $f_{n,m}$表示词w_n在文档d_m中出现的次数; $\sum_{i=1}^{N} f_{i,m}$表示文档d_m中出现的所有词汇数; D 为文档集合所有的文档数量, D_{w_n}表示出现了词w_n的文档数量, 并且为了不会出现由于语料集不包括词w_n而导致的分母等于零的情况, 在此将分母加上一个常数 1.

TFIDF 权重即为 *TF* 和 *IDF* 的乘积, 计算公式如下:

$$TFIDF_{MN} = \begin{bmatrix} tf_{11}, & tf_{12}, & \cdots & tf_{1N} \\ tf_{21}, & tf_{22}, & \cdots & tf_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ tf_{M1}, & tf_{M2}, & \cdots & tf_{MN} \end{bmatrix} \tag{10}$$

1.5 Self-Attention 机制

短信文本的识别过程中, 文本所包含的词数比较少, 很难获取更多的句子语义信息, 但通过对比语料库可以发现, 在句子中的某些重点词汇可以更快的帮助识别短信类别, 如在“元旦特惠, 原价 xxx 的三星手机现在特惠, 全部八折, 最高直降 xxx”这样的一条垃圾短信中, 就包含了一些重点词汇: “三星”(品牌名称), “特惠”, “八折”, “直降”等. 对于不同的词汇, 其对文本分类起到的作用也不一样, 因此为突出关键词并优化特征词提取过程, 引入 Attention 机制, 结构如图 5 所示.

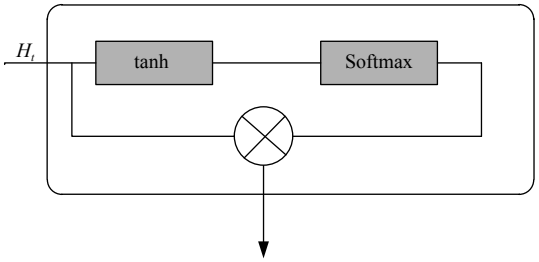

图 5 Attention 机制结构

Attention 机制通过对不同的词汇分配不同的权重从而强化关键信息的比重. 传统的 attention 机制模型需要依赖部分外部信息, 而 self-attention 机制不需要使用其他外部的信息, 它会自动从自身所给的信息训练

来更新参数从而给不同信息分配不同的权重, 因此本文采用 self-attention 机制.

根据 Bi-LSTM 所有的输出向量组成的矩阵 $H=[H_1,H_2,\cdots,H_N]$, 结合 TFIDF 模型所得到了当前输入文本的权重, 更好将注意力集中的重点词汇上, 从而获得更好的分类效果, 基于 TFIDF 的 self-attention 机制的计算公式下:

计算公式如下:

$$Y=\begin{bmatrix} H \\ TFIDF_N \otimes e_N \end{bmatrix} \tag{11}$$

$$M=\tanh(Y) \tag{12}$$

$$\alpha=Softmax\left(W^{\mathrm{T}}M\right) \tag{13}$$

$$\gamma=H\alpha^T \tag{14}$$

H 包含 Bi-LSTM 所输出向量$[H_1,H_2,\cdots,H_N]$, $TFIDF_N$为当前输入文本的权重, e_N为 N 维单位向量, N 为输入文本的长度, W^{T}为随机初始化并在训练中学习的权重矩阵, γ基于 TFIDF 的 self-attention 层的输出值, 将输出值输入激活函数得到分类结果, 短信文本进行“正常短信”和“垃圾短信”的二元分类.

2 实验介绍

2.1 实验数据

本文参与实验的短信数据共有 20 万条, 分为垃圾短信 (negative) 和正常短信 (positive) 两种, 其中正常短信数量为 10 万条, 垃圾短信数量为 10 万条, 这些数据在初始化时已经被分为了垃圾短信或者正常短信.

2.2 实验数据预处理和参数设置

原始短信数据包含了很多非法符号, 例如表情符号这些对于短信分类并没有用, 所以数据不直接使用, 先进行数据的清洗. 经过清洗过的短信数据要进行中文分词处理, 将短信的句子拆成单个中文单词, 本文中使用结巴分词工具对短信进行分词, 分词结束之后去除短信文本中的停用词, 常见的停用词有“的”“得”“在”等, 提高关键词密度, 增强搜索效率.

分词处理之后, 本文使用 Word2Vec 工具初始化词向量, 同时使用 Skip-gram[22,23] 模型训练数据集, 并结合维基中文语料库训练词向量维度, 词向量维度越高可以越好的表达中文单词的语义, 但是随着维度的升高也会增加模型参数的数量, 因此经过实验对比, 将词向量维度设置为 100, 隐藏层设置为 128, 窗口大小设置为 5, 经过预处理之后的短信数据最长的一条为 100 个中文词汇, 因此每条短信的特征矩阵大小均为, 将特征矩阵作为 TFIDF-self-attention-based Bi-LSTM 模型的输入参与到模型的训练中去.

2.3 评价指标

本文以准确率 *Precision*, 召回率 *Recall* 和 *F*1 作为指标来评估模型在垃圾短信识别任务中的有效性, 计算公式如下:

$$Precision=\frac{N_{\text{right}}}{N_{\text{right}}+N_{\text{wrong}}} \tag{15}$$

$$Recall=\frac{TP}{TP+FN} \tag{16}$$

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{17}$$

其中, N_{right}, N_{wrong}, TP, FN 分别表示短信分类准确的数量, 短信分类错误的数量, 正常短信被判断成正常短信的数量和正常短信被判断成垃圾短信的数量.

本文采用十折交叉验证法来评估模型在短信测试集上的准确率.

3 结语

3.1 实验结果

在本文中, 设计了 4 组对比实验, 分别使用了 LSTM, Bi-LSTM, self-attention-based Bi-LSTM (SA Bi-LSTM), 结合 TFIDF 的 self-attention-based Bi-LSTM (TSA Bi-LSTM) 4 组不同的模型, 准确率实验结果和运行时间对比实验结果分别如表 1 和表 2 所示, 对 4 种模型根据不同训练集的大小进行实验结果如图 6所示.

表 1 准确率对比实验结果

模型	*Precision*	*Recall*	*F*1
LSTM	0.855	0.859	0.856
Bi-LSTM	0.861	0.865	0.863
SA Bi-LSTM	0.879	0.883	0.881
TSA Bi-LSTM	0.901	0.905	0.902

表 2 运行时间对比实验结果

模型	时间(s)
LSTM	28.1
Bi-LSTM	20.3
SA Bi-LSTM	18.5
TSA Bi-LSTM	17.9

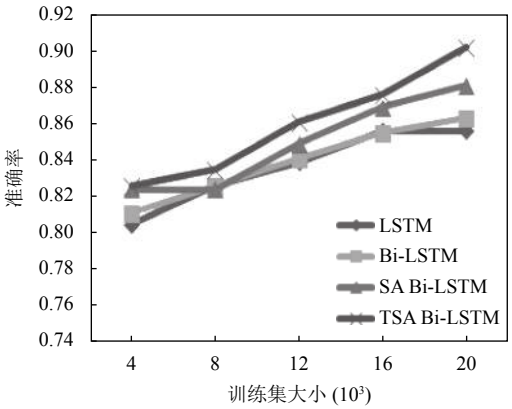

图 6 训练集大小与准确率比较

3.2 模型对比分析

4 种不同模型的准确率可以看出, LSTM 模型的准确率低于 Bi-LSTM 模型的准确率, 在 Bi-LSTM 模型中引入 attention 机制准确率得到了提高, 而 self-attention-based BI-LSTM 模型结合 TFIDF 则又进一步提高了模型准确率.

因此, 通过以上实验结果可得出结论:

(1) 对比 LSTM 模型和 Bi-LSTM 模型, 发现 Bi-LSTM1 模型准确率高于 LSTM 模型, 同时模型用时更短, 所以 Bi-LSTM 模型对文本特征信息提取具有更精确的效果.

(2) 对比 Bi-LSTM 模型和 self-attention-based Bi-LSTM 模型, 可以得出在 Bi-LSTM 模型引入 self-attention 机制之后可以提高模型的准确率并且一定程度上缩短了模型运行时间, 证明了 self-attention 机制的有效性.

(3) 对比 self-attention-based Bi-LSTM 模型和结合 TFIDF 的 self-attention-based Bi-LSTM 模型, 可以得出将注意力机制和 TFIDF 模型相结合, 更能有效提高重点词汇的权重达到更好的分类效果.

4 结语

本文将 self-attention 机制和 TFIDF 模型相结合加入到 Bi-LSTM 模型, 设计出结合 TFIDF 的 self-attention-based Bi-LSTM 模型, 并应用到垃圾短信识别中, 通过 4 组对比实验, 验证了该模型具有良好的使用效果.

由于 self-attention 机制的引用需要消耗一定的计算成本, 因此在未来的工作中, 将考虑如何在减少 self-attention 机制对计算成本消耗的基础上继续优化结合 TFIDF 的 self-attention-based Bi-LSTM 模型, 使得该模型能在未来的应用中达到更好的使用表现.

参考文献

1 Shahi TB, Yadav A. Mobile SMS spam filtering for Nepali text using Naïve Bayesian and support vector machine. International Journal of Intelligence Science, 2014, 4(1): 24–28. [doi: 10.4236/ijis.2014.41004]

2 Kim Y. Convolutional neural networks for sentence classification. arXiv: 1408.5882, 2014.

3 Jiang MY, Liang YC, Feng XY, *et al*. Text classification based on deep belief network and softmax regression. Neural Computing and Applications, 2018, 29(1): 61 –70. [doi: 10.1007/s00521-016-2401-x]

4 Talman A, Yli-Jyrä A, Tiedemann J. Natural language inference with hierarchical BiLSTM max pooling architecture. arXiv: 1808.08762, 2018.

5 雷朔, 刘旭敏, 徐维祥. 基于词向量特征扩展的中文短文本分类研究. 计算机应用与软件, 2018, 35(8): 269–274. [doi: 10.3969/j.issn.1000-386x.2018.08.049]

6 Zhang R, Lee H, Radev D. Dependency sensitive convolutional neural networks for modeling sentences and documents. arXiv: 1611.02361, 2016.

7 Anoop VS, Prem SC, Asharaf S, *et al*. Generating and visualizing topic hierarchies from microblogs: An iterative latent dirichlet allocation approach. Proceedings of 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI). Kochi, India. 2015. 824–828.

8 Wigington C, Stewart S, Davis B, *et al*. Data augmentation for recognition of handwritten words and lines using a CNN-LSTM network. Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto, Japan. 2017. 639–645.

9 Kim Y, Jernite Y, Sontag D, *et al*. Character-aware neural language models. arXiv: 1508.06615, 2015.

10 Gers FA, Schmidhuber E. LSTM recurrent networks learn simple context-free and context-sensitive languages. IEEE Transactions on Neural Networks, 2001, 12(6): 1333–1340. [doi: 10.1109/72.963769]

11 Graves A. Supervised sequence labelling with recurrent neural networks. Berlin, Heidelberg: Springer, 2012.

12 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

13 Kiperwasser E, Goldberg Y. Simple and accurate dependency parsing using bidirectional LSTM feature representations. Transactions of the Association for Computational Linguistics, 2016, 4: 313–327. [doi: 10.1162/tacl_a_00101]

14 唐明, 朱磊, 邹显春. 基于 Word2Vec 的一种文档向量表示. 计算机科学, 2016, 43(6): 214–217, 269. [doi: 10.11896/j.issn.1002-137X.2016.06.043]

15 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, CA, USA. 2017. 5998–6008.

16 Carrasco M, Barbot A. Spatial attention alters visual appearance. Current Opinion in Psychology, 2019, 29: 56–64. [doi: 10.1016/j.copsyc.2018.10.010]

17 郑雄风, 丁立新, 万润泽. 基于用户和产品 Attention 机制的层次 BGRU 模型. 计算机工程与应用, 2018, 54(11): 145–152. [doi: 10.3778/j.issn.1002-8331.1701-0337]

18 Cho K, Van Merrienboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv: 1406.1078, 2014.

19 Mikolov T, Yih WT, Zweig G. Linguistic regularities in continuous space word representations. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, GA, USA. 2013. 746–751.

20 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2013. 3111–3119.

21 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. Proceedings of the 1st International Conference on Learning Representations. Scottsdale, AZ, USA. 2013. 1–12.

22 Zhao MZ, Xu B, Lin HF, *et al*. Discover potential adverse drug reactions using the skip-gram model. Proceedings of 2015 IEEE International Conference on Bioinformatics and Biomedicine. Washington, WA, USA. 2015. 599–698.

23 来斯惟, 徐立恒, 陈玉博, 等. 基于表示学习的中文分词算法探索. 中文信息学报, 2013, 27(5): 8–14. [doi: 10.3969/j.issn.1003-0077.2013.05.002]