文章编号: 1006-2475(2021) 04-0122-05

基于 BiGRU-Attention-CNN 模型的垃圾邮件检测方法

赵宇轩,胡怀湘

(华北计算技术研究所基础一部,北京 100083)

摘要: 电子邮件是一种重要的通信工具,但是垃圾邮件问题一直影响着人们日常的工作生活。不断改进垃圾邮件的检测技术、提高垃圾邮件的检测速度和准确率有着重要的研究意义和现实意义。双向门控循环单元(BiGRU) 和卷积神经网络(CNN) 广泛应用于文本分类领域,二者的结合可以充分发挥 BiGRU 上下文依赖关系提取能力以及 CNN 特征提取能力,但是针对垃圾邮件检测问题,还需要考虑邮件中一些特定的词语,因此本文提出一种基于 BiGRU-Attention-CNN 模型的垃圾邮件检测方法来提高垃圾邮件的检测准确率。模型首先将邮件文本转换成特征向量并进行 BiGRU 序列化学习,随后引入注意力机制(Attention) 对特定词语赋予更大的权重,再将注意力层输入 CNN 模型,经过卷积、池化、全连接,最终得到分类结果。本文将模型在 Trec06c 邮件数据集上进行实验,与其他模型进行对比取得了更好的效果,最终模型的准确率达到 91.62%。

关键词: 双向门控循环单元; 注意力机制; 卷积神经网络; 垃圾邮件识别

中图分类号: TP391 **文献标志码**: A **DOI**: 10.3969/j.issn.1006-2475.2021.04.021

Spam Recognition Method Based on BiGRU-Attention-CNN Model

ZHAO Yu-xuan, HU Huai-xiang

(North China Institute of Computing Technology, Beijing 100083, China)

Abstract: E-mail is an important communication tool, but the problem of spam has been affecting people's daily work and life. Continuously improving spam detection technology and increasing the speed and accuracy of spam detection has important research and practical significance. Bi-directional gated recurrent unit (BiGRU) and convolutional neural network (CNN) are widely used in the field of text classification. The combination of them could give full play to BiGRU context dependency extraction capabilities and CNN feature extraction capabilities. But for the problem of spam recognition, it is also necessary to consider some specific words in the email. So this article proposes a spam recognition method based on the BiGRU-Attention-CNN model to improve the accuracy of spam detection. The model first converts the email text into feature vectors and performs BiGRU serialization learning, and then introduces the attention mechanism (Attention) to give greater weight to specific words. After the attention layer is input to the CNN model, through convolution, pooling, and full connection, the classification result is finally obtained. The model is tested on the Trec06c mail data set and compared with other models, better results are achieved. The final accuracy of the model is 91.62%.

Key words: BiGRU; attention; CNN; spam recognition

0 引 言

电子邮件自从 20 世纪 70 年代诞生以来距今已有 50 年,它已成为人们日常生活、工作中最常用的通信工具之一。然而随着邮件技术的普及,垃圾邮件也变得越发猖獗。根据奇安信和 Coremail 联合编撰的 2019 中国企业邮箱安全性研究报告[1] 显示,2019 年全国企业邮箱用户收到的各类垃圾邮件约占企业级用户邮件收发总量的 47.2%,是企业级用户正常邮件数量的 1.2 倍。垃圾邮件已经成为困扰企业邮件安全的重大问题,因此如何快速准确地识别垃圾邮件是一个重要的研究课题。

目前识别垃圾邮件主要有 2 类方法: 一类是基于邮件特征的识别,比如发件人的发信频率,发件地址黑名单 RBL[2]、DBL 等; 另一类是基于邮件内容的识别,传统方法包括字符匹配、词频统计(TF-IDF、

收稿日期: 2021-01-05; **修回日期**: 2021-02-05

作者简介: 赵宇轩(1997—) ,男,吉林长春人,硕士研究生,研究方向: 网络安全,计算机体系结构,E-mail: nsczyx@gmail.com; 胡怀湘(1965—) ,男,研究员,研究方向: 计算机体系结构,网络存储,E-mail: huaixianghu@163.com。

LDA[3]、朴素贝叶斯等算法)。近年来随着深度学习网络的不断发展,卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN) 等技术也逐渐应用在垃圾邮件识别领域。CNN 模型通过"端到端"的学习,能够有效地学习并提取数据样本特征,在图像处理、人脸识别、自动驾驶等领域有着广泛的应用。RNN 模型是一种序列模型,研究人员在 RNN 的基础上提出了变体模型: 双向门控循环单元(BiGRU)。BiGRU 更适合对文本建模、获取文本全局的结构信息。对于邮件文本的分类,邮件中的关键词提取也非常重要。注意力机制(Attention) 可以对邮件中重要的词赋予更高的权重,可以更好提取关键信息。本文综合 3 种模型的优点,提出基于 BiGRU-Attention-CNN 模型的垃圾邮件检测方法。

1　相关工作

研究人员在垃圾邮件识别领域已经做了很多改进、研究和探索。这其中包括对传统文本分类算法的改进,也包括对 CNN、RNN 模型的优化。

王鹿等人[4]应用树结构的思想,对特征词的条件概率进行开方处理,改进了朴素贝叶斯算法的分类效果,并减少了训练需要消耗的资源。吴小晴等人[5]在 TF-IDF 算法里加入卡方统计量 CHI 以及位置影响因子,并且结合逆向最大匹配算法的邮件文本分词和类中心向量算法的特征选择,解决了 TF-IDF 算法未能很好分配词的权重的问题。黄鹤等人[6]在 CNN 的基础上,引用 Skip-gram 以及 Highway,将邮件文本转换成更低维度的特征向量,提高了邮件分类模型的准确率。周枝凝等人[7]针对垃圾邮件分类中词向量学习不充分的问题,引入了 ALBERT 动态词向量生成模型,并提出一种将 ALBERT 动态词向量与循环神经网络相结合的 ALBERT-RNN 模型来进行垃圾邮件的识别。

虽然目前已经有了很多垃圾邮件识别方法,但是在识别速度、识别准确率等方面依然有很大的改进空间。本文研究基于 BiGRU-Attention-CNN 模型的垃圾邮件检测方法,并在 Trec06c 邮件数据集上进行实验,与其他模型进行对比,取得了更好的效果。

2　模型架构

Kim[8]在 2014 年提出了将 CNN 应用在文本分类领域的方法且取得了一定的效果。本文在其研究基础上,针对下述 3 个问题,引入双向门控循环单元 BiGRU 和注意力机制 Attention。BiGRU 可以更好地获得上下文依赖关系和文本特征,CNN 可以提取文本的局部特征,而 Attention 机制可以凸显文本的重要特征,提高模型的准确率和效率。上述 3 个问题是: 1) 传统文本分类算法大都存在数据稀疏、矩阵维度高而性能并不高等情形; 2) CNN 模型在卷积、池化操作时会丢失文本序列中词汇的位置以及顺序信息,不能很好地捕捉文本全局结构信息; 3) RNN 模型存在无法解决长时依赖、文本重要特征无法凸显作用等情形。

如图 1 所示,模型包括文本预处理、BiGRU、Attention、CNN 共 4 大部分。文本预处理将邮件转换为格式化的特征向量,BiGRU 提取上下文信息,Attention 对特定词语增加权重,CNN 最终对邮件进行分类。

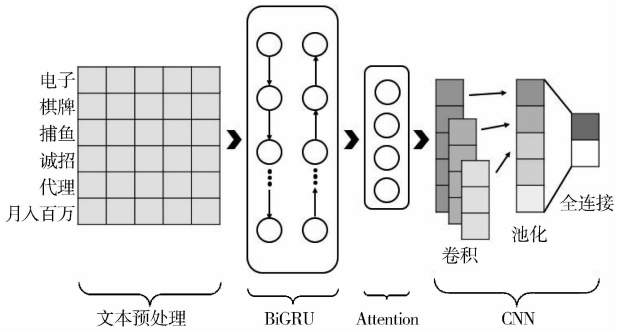

图 1　模型架构

2.1　邮件文本预处理

邮件文本并非格式化数据,要想让计算机能够"读懂"邮件文本内容,需要将邮件文本进行数据清理与处理,将其转换为格式化信息。

2.1.1　去除非文本部分

许多邮件文本为了渲染显示效果会使用 HTML 等前端技术。毫无疑问,HTML 标签对于训练模型来说并不是有效信息,它与邮件文本想要表达的信息没有任何关系。除了 HTML 标签,无用的非文本部分还包括: CSS、JavaScript 代码、URL 地址、标点符号、特殊字符、表情符号等。

为了更有效地挖掘邮件文本的重要信息,去除非文本部分带来的影响,本文采用 Python 的 Beautiful-Soup[9]库以及正则表达式、黑白名单等技术对邮件中的非文本部分进行清洗去除。

2.1.2　去除停用词

停用词属于可以忽略的词,它们对句子的分类并无帮助,去除它们可以节省存储空间并且提高模型效率。停用词包括虚词、语气助词、形容词、副词、连接词、介词等,它们自身并没有特定的含义,只有在完整的语句中才有作用,如"乃""此外""的""在"等词。本文整合了"哈尔滨工业大学停用词表""百度停用

词表”“四川大学机器学习实验室停用词表”[10]，并对它们进行去重操作，形成一个比较完善且准确的停用词表用来对邮件文本进行去除停用词的操作。

2.1.3 中文文本分词

在去除邮件的非文本部分以及停用词后，还需要对邮件文本进行分词处理，这样才能更容易获得邮件文本的特征。英文文本分词非常简单，只需按照英文单词之间的空格分词就可以了，但是中文文本是连续的，并没有天然的分隔符，所以中文文本分词更加困难。

本文模型采用词典与统计相结合的方式对中文文本进行分词。具体使用了“结巴[11](Jieba)”中文分词库，对 Jieba 库不能处理的分词语句，使用自定义的字典来提高邮件文本分词的准确率。

2.1.4 词向量转换

在经过上述步骤处理后，邮件文本已经变得“规整”。如图 2 的数据已经转换成了图 3 的结构。

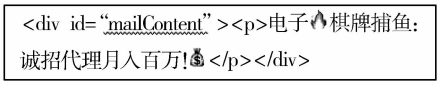

图 2 原始邮件文本

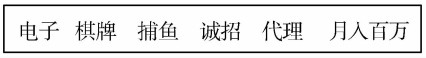

图 3 邮件文本初步处理结果

然而，为了让计算机能够“读懂”，仍需将邮件文本继续转换为实数向量[12]。传统的转换方式是采用词袋模型，词袋模型通过对文本进行分词，分词后统计词典中每个词出现的频率，并将频率作为该词的特征值，最后将词和词频一一对应，形成词向量的转换。如果将词的权重替换为是否出现该词，则是 one-hot[13] 的表达形式。但是 one-hot 有明显的缺点，它无法判断出词与词之间的“相似度”，也称为“语义鸿沟”，one-hot 的基本假设是词与词之间的语义和语法关系是相互独立的，仅仅靠 2 个向量是无法看出 2 个词之间的关系的，而且所产生的向量是高维稀疏向量。

Word2vec[14] 模型可以改进 one-hot 的缺点。它将每个词表示成一个定长的向量，并使得这些向量能较好地表达不同词之间的相似和类比关系。Word2vec 模型训练过程包含 2 个模型：跳字模型(Skip-gram) 和连续词袋模型(Continuous Bag Of Words，CBOW)[15]。跳字模型假设基于某个词来生成它在文本序列周围的词。连续词袋模型假设基于某中心词在文本序列前后的背景词来生成该中心词。

如图 4 所示，本文采用 Word2vec 中的 Skip-gram 模型，Skip-gram 模型训练时间短且训练效果好[6]。Skip-gram 模型的具体目标是：给定一个训练词序列 $w_1, w_2, \cdots, w_t$，使公式(1) 中 J_θ 的值最大。其中 c 是训练上下文的大小，T 是样本个数，训练样本越多，Skip-gram 模型准确率越高。

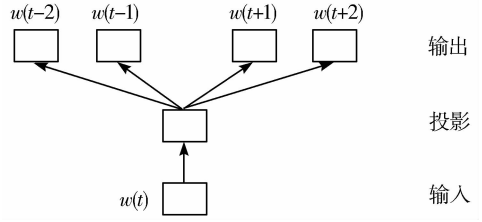

图 4 Skip-gram 模型

$$J_\theta = \left| \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leqslant j \leqslant c; j \neq 0} \log p(w_{t+j} \mid w_j) \right| \tag{1}$$

2.2 BiGRU

RNN 是一种用来处理序列数据的神经网络，它能够提取上下文的关系，但是 RNN 在训练长序列时存在梯度消失以及梯度爆炸的问题。为了解决 RNN 的这些问题，Hochreiter 等人[16] 提出了一种特殊的 RNN-长短期记忆(Long Short-Term Memory，LSTM) 网络，后来 Cho 等人[17] 在 LSTM 的基础上又提出了门控循环单元(Gated Recurrent Unit，GRU) 网络，GRU 与 LSTM 相比模型效果相似但是 GRU 训练所需的资源更少[18]。BiGRU 模型包含前向传播、后向传播 2 个 GRU 模型，相比 GRU 具有更高的分类精度。综上考虑，本文采用 BiGRU 模型。

GRU 输入输出结构如图 5 所示，它会将当前输入 x_t 与包含前面节点信息的 h_{t-1} 结合，得出当前输出 y_t 以及传递给后续节点的隐状态 h_t。

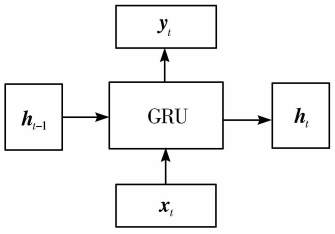

图 5 GRU 输入输出结构

2.3 注意力机制

邮件的特征向量经过 BiGRU 层之后，邮件的前后文信息已经得到了充分的提取，但是邮件中的关键信息、关键词并没有被突出表示。

注意力机制 Attention Mechanism 的原理和人眼看图片的逻辑很相似，无需看清图片的全部细节，而是将注意力聚集在了图片上的焦点区域。本文引入注意力机制，为邮件文本中的关键词赋予更大的权重，以达到突出关键信息的目的。注意力机制实现过程如图 6 所示。

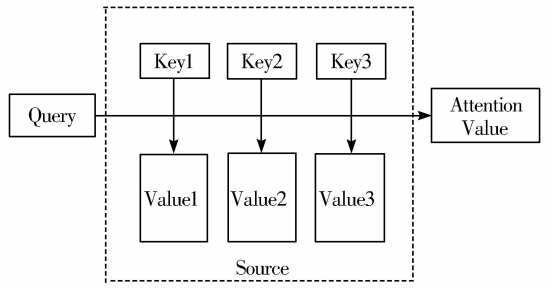

图 6　Attention 实现过程

Attention 机制将 Query 和 Key 进行相似度计算,得到权值,之后将得到的权值进行归一化操作,得到权重,最后将权重和 Value 进行加权求和得到 Attention Value。

2.4　卷积神经网络

卷积神经网络 CNN,是一种前馈神经网络,它广泛应用于图像识别领域,且由于它优秀的局部特征提取能力,也可以应用在文本分类领域。

CNN 包括卷积、池化、全连接等操作。在卷积层中,通过不同的卷积核可以得到语句中不同特征的列向量,卷积核的大小等于词向量的维数与卷积核纵向词个数的乘积; 在池化层中,通过 pooling 操作一方面可以将从卷积层中获得列向量的最大值提取出来,另一方面可以消除由于句子之间长度不同带来的差异; 在全连接层中,通过 softmax 操作可以整合池化数据并获得最终的分类结果。本文利用 CNN 模型进一步处理包含权重信息的邮件文本特征向量,并得到最终的垃圾邮件分类结果。

3　实验与结果

3.1　实验环境与超参数设置

实验环境与模型的超参数如表 1、表 2 所示,模型超参数的选取方法借鉴了 Zhang 等人[19] 的调参实验,首先找到最佳的 region_size,然后通过对不同参数进行调整与实验对比,找到最优的模型超参数。

表 1　实验环境

实验环境	配置数据
操作系统	Ubuntu 18.04.1 LTS
CPU	Xeon(R) E5-2620 v4 @2.10 GHz
内存	32 GB
显卡	GTX 1080Ti
编程语言	Python 3.7.5
模型框架	Tensorflow 2.3.0

表 2　模型超参数

模型超参数	参数数值
Epoches	10
EmbeddingSize	200
HiddenSize	256
DropoutKeepProb	0.2
SequenceLength	200
BatchSize	128
Optimizer	Adam

3.2　实验数据

实验数据来源于公开的垃圾邮件语料库 Trec06c(数据集链接: https://plg.uwaterloo.ca/~gvcormac/treccorpus06/) ,总共包含 64620 封邮件,其中正常邮件 21766 封,垃圾邮件 42854 封。本文对实验数据进行十折交叉验证,即选出 10% 的数据作为验证数据集,90% 的数据作为训练数据集。

3.3　评价指标

本文对模型进行了准确率(Accuracy) 、精确率(Precision) 、召回率(Recall) 、F1 值 4 个维度[20] 的评价,具体的评价指标如下:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

其中: TP 为数据集标记为垃圾邮件且模型检测结果也为垃圾邮件的邮件数量; TN 为数据集标记为正常邮件且模型检测结果也为正常邮件的邮件数量; FP 为数据集标记为正常邮件但模型检测结果为垃圾邮件的邮件数量; FN 为数据集标记为垃圾邮件但模型检测结果为正常邮件的邮件数量。

3.4　对比实验

为了验证基于 BiGRU-Attention-CNN 模型在垃圾邮件分类领域的有效性,本文总共设置了 5 组对比实验,分别是 SVM 模型[21] 、CNN 模型、BiGRU-CNN 模型[22] 、BiLSTM-Attention-CNN 模型[23] 、BiGRU-Attention-CNN 模型。其中 SVM 模型使用 TF-IDF 算法提取邮件文本中关键词的统计特征对邮件进行分类。为了保证实验的公平,5 组对比实验采用了一致的数据集和实验环境。

3.5　实验结果

5 组对比实验结果如表 3 所示。

表 3 实验结果

模型	准确率/%	精确率/%	召回率/%	F1 值/%
SVM	70.20	80.31	72.96	76.46
CNN	82.33	88.51	84.20	86.36
BiGRU-CNN	88.88	93.25	89.74	91.46
BiLSTM-Attention-CNN	91.32	95.13	91.60	93.33
BiGRU-Attention-CNN	91.62	95.36	91.83	93.57

实验结果表明,传统的文本分类模型如 SVM 等准确率最低只有 70.20%,卷积神经网络模型对比传统模型准确率有了很大的提高,准确率数值提高到了 82.33%。在引入 BiGRU 以及注意力机制后,准确率又有了进一步的提高,准确率分别提高到 88.88% 和 91.62%。BiLSTM 和 BiGRU 在实验中虽然准确率等数值区别不大(BiGRU 比 BiLSTM 准确率提高了 0.30 个百分点),但是在对邮件数据集的十折交叉验证过程中,BiLSTM-Attention-CNN 模型的训练时间为 3206 s, BiGRU-Attention-CNN 模型的训练时间为 2897 s,实验表明 BiGRU 的训练速度比 BiLSTM 快。

4 结束语

本文采用基于 BiGRU-Attention-CNN 模型进行垃圾邮件的检测,实验结果表明,该模型方法对比传统的文本分类模型方法以及其他垃圾邮件检测方法有了很大的提升,训练速度与准确率等指标也表现出色。

本次实验在选取邮件数据集进行模型训练与检测时,只选取了邮件的正文(Content)部分,并没有考虑邮件的其他部分(Mail-From、MIME-From、Subject、Sender 等字段),然而这些非正文部分也对垃圾邮件的识别有很大的帮助,所以下一步的研究内容是针对邮件头部的相关检测。

参考文献:

[1] 林延中,裴智勇,刘川琦,等. 2019 年中国企业邮箱安全性研究报告[R]. 北京: 奇安信创新团队, 2020.

[2] 申超. 反垃圾邮件新技术在新华网电子邮局中的应用研究[J]. 中国传媒科技, 2013(15): 58-61.

[3] 林建洪,翟建桐,徐菁. 融合 LDA 与 Word2vector 的垃圾邮件过滤方法研究[J]. 网络安全技术与应用, 2017(3): 73-75.

[4] 王鹿,李志伟,朱成德,等. 基于朴素贝叶斯算法的垃圾邮件过滤研究[J]. 传感器与微系统, 2020,39(9): 46-48.

[5] 吴小晴,万国金,李程文,等. 一种改进 TF-IDF 的中文邮件识别算法研究[J]. 现代电子技术, 2020,43(12): 83-86.

[6] 黄鹤,荆晓远,董西伟,等. 基于 Skip-gram 的 CNNs 文本邮件分类模型[J]. 计算机技术与发展, 2019,29(6): 143-147.

[7] 周枝凝,王斌君,翟一鸣,等. 基于 ALBERT 动态词向量的垃圾邮件过滤模型[J]. 信息网络安全, 2020,20(9): 107-111.

[8] KIM Y. Convolutional neural networks for sentence classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1746-1751.

[9] 迟殿委. 基于 Python 的网页图片爬取[J]. 电脑编程技巧与维护, 2019(5): 129-130.

[10] 官琴,邓三鸿,王昊. 中文文本聚类常用停用词表对比研究[J]. 数据分析与知识发现, 2017(3): 72-80.

[11] 徐博龙. 应用 Jieba 和 Wordcloud 库的词云设计与优化[J]. 福建电脑, 2019,35(6): 25-28.

[12] 景栋盛,薛劲松,冯仁君. 基于深度 Q 网络的垃圾邮件文本分类方法[J]. 计算机与现代化, 2020(6): 89-94.

[13] HARRIS D, HARRIS S. Digital Design and Computer Architecture[M]. Morgan Kaufmann, 2010.

[14] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv: 1301.3781, 2013.

[15] LUO Q, XU W R, GUO J. A study on the CBOW model's overfitting and stability[C]// Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval & Reasoning. 2014: 9-12.

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

[17] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1724-1734.

[18] 胡玉琦,李婧,常艳鹏,等. 引入注意力机制的 BiGRU-CNN 情感分类模型[J]. 小型微型计算机系统, 2020,41(8): 1602-1607.

[19] ZHANG Y, WALLACE B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[C]// Proceedings of the 8th International Joint Conference on Natural Language Processing. 2017: 253-263.

[20] 季威志,薛雷. 基于 BiGRU-CNN-Attention 模型的股市评论情感分析[J]. 工业控制计算机, 2020,33(4): 70-72.

[21] 徐娟,卞良. 基于 SVM 的中文垃圾邮件预测系统研究[J]. 数字技术与应用, 2020,38(1): 38-39.

[22] 郑诚,薛满意,洪彤彤,等. 用于短文本分类的 DC-BiGRU_CNN 模型[J]. 计算机科学, 2019,46(11): 186-192.

[23] 吴小晴. 基于 CNN 的双向 LSTM 注意力机制垃圾邮件分类的研究与分析[D]. 南昌: 南昌大学, 2020.