doi:10. 16055/j. issn. 1672-058X. 2022. 0003. 003

基于改进的卷积神经网络邮件分类算法研究

宋 丹[1], 陆 奎[1],戴旭凡[2]
(1. 安徽理工大学 计算机科学与工程学院, 安徽 淮南 232001;2. 安徽理工大学 电气与信息工程学院, 安徽 淮南 232001)

摘 要:针对传统文本分类方法中出现的维度过高和数据稀疏问题,通过对卷积神经网络(Convolutional Neural Network,CNN)和 inception V1 模型的深入研究,将两个模型融合起来,提出了一种基于 i-CNN 模型的邮件分类方法;在卷积、池化操作中加入了 1×1 卷积核降低特征向量的厚度,减少了参数,提高了计算性能;通过数据验证,i-CNN 模型对邮件的分类结果高达 92. 18%,在对比实验中,i-CNN 模型相对于几种机器学习分类模型,取得了最高的分类精准率,在有无 inception 结构模型对比中,i-CNN 模型精准率高于 CNN 模型;说明该模型具有较好的分类效果,且 inception V1 模型的融入能提高文本分类的准确率。

关键词:文本分类;卷积神经网络;inception V1;word2vec
中图分类号:TN911 **文献标志码**:A **文章编号**:1672-058X(2022)03-0020-06

0 引 言

随着网络媒体的发展和电子邮件的普及,社交媒体中的广告越来越多地充斥在生活中,给人们带来了很多困扰。垃圾邮件不仅占据空间内存,还侵犯收件人隐私,所以邮件的快速分类就成了现在的热门课题。邮件分类的研究是自然语言处理的一个分支方向,对此,国内外已有很多学者进行了深入研究,主要有传统的机器学习方法,如朴素贝叶斯、k 近邻算法、支持向量机等,深度学习方法,如卷积神经网络、循环神经网络等。

文献[1]在基于词典的基础上,增加了 5 部词典,结合分析文本之间的语义规则,实现文本情感分析;文献[2]在 Skip-Gram 模型中将词向量映射到一个低维度,结合 CNN 提取特征,再加一层 Highway 网络对整体特征优化,提高了模型准确率;文献[3]提出基于 KNN 的方法和一般基于图的分类方法来区分中国评论网站中的串通垃圾邮件发送者,两种方法的分类效果明显优于纯指标分类器;文献[4]用 CNN 进行了一系列分类任务的实验,对该体系结构简单修改,同时使用特定于任务的向量和静态向量,在多个基准上都取得了很好的结果,证明了一个简单的 CNN 与很少的超参数调整和静态向量在多个基准上能取得很好结果;文献[5]提出了一种全局最大池化模型,通过提取多个卷积层和 GMP 层得到更深层的语义特征,在点积层产生情感得分,提高了模型效率;文献[6]在传统卷积神经网络模型基础上去掉了池化层,加上了门阀循环单元 GRU 层,提出来一种串并行卷积门阀循环网络新模型,提高了分类准确率。

传统分类模型结构简单,改进模型或者结合一些其他模型会提高分类的准确率。因此,本文通过对 inception V1 模型特征优化,将稀疏矩阵聚类为密集矩阵,最终到 soft 分类器中分类。实验通过对数据样本多次迭代操作,结果显示:精准率逐渐升高且趋于稳定,通过几种分类模型的对比试验可以得出,i-CNN 模型相对其他机器学习模型有着较好的

收稿日期:2020-12-28;**修回日期**:2021-03-06.
基金项目:国家自然科学基金项目资助(51274011).
作者简介:宋丹(1995—),女,河南驻马店人,硕士研究生,从事文本分类研究.

分类效果。

1　相关工作

1.1　卷积神经网络与文本分类

卷积神经网络是一种深层神经网络模型,本质上就是多层卷积运算,将上一层神经元的输出作为下一层的输入,通过多层卷积计算,对每一层的卷积运算结果进行非线性转换[10]。卷积神经网络相较于传统神经网络具有参数共享、局部连接两大优势,使网络需要训练的参数大大减少,却没有降低准确率。

参数共享也可理解为“平移不变性”,是指一组神经元对应一个权重,而不是每个神经元对应一个权重,这样就减少了很多参数。局部连接是指和上一层神经元中的局部相连,而不是和上层所有神经元相连接,这样又减少了很多参数,减少了后续的计算量。

用传统的机器学习方法做文本分类任务时,通常输入文档的 tf-idf 向量作模型的特征,这样的输入会使 tf-idf 表示丢失文本序列中的顺序。而卷积神经网络对文本数据建模时,输入词向量,然后通过滑动窗口将长短不一的词向量转化为固定长度的向量,这样可以保留原文本中的一些局部特征和顺序上的信息。由于远距离两个词之间的依赖关系很难被捕捉到,所以基于 CNN 的文本分类更适合短文本。

1.2　文本分词

不同于英文文本词与词之间有分隔符,不用进行分词处理,中文文本字词之间没有分隔符,而且单个汉字不能完整的表达意思,所以中文文本要进行分词处理。中文分词会存在很多问题,最主要的就是会出现歧义,对句子进行不同的分词会出现不同的意思。对于这一问题,可以在卷积时使用多种不同的卷积核进行卷积。其次是分词词典要全面、及时更新,中文分词选用 Word2vec 分词器,它是一个基于 gensim 库的中文数据处理工具包,在词典中,每一个词都有唯一的向量对应,文本就可以转化为向量表示。

1.3　停用词处理

停用词就是文本中出现频率很高而且没有实际意义的词,比如“的”、“啊”“了”等,也包括标点符号和网络表情包,它们没有表达任何感情倾向,对文本情感划分没有影响。邮件一般都是比较短的文本,停用词的存在对结果的影响更大。

去停用词常用的方法是选择一个比较全面的词表,过滤去重,将文本中的停用词逐一去掉。去重后的文本不仅降低了特征向量的维度,降低了后续计算的复杂度,在一定程度上还提高了分类的准确率。

1.4　Word Embedding 模型

图像在计算机处理时输入的是二维矩阵,因此处理中文邮件时也要先把文本转变为矩阵形式,通过开源工具 word2vec 将文本中的词转化为词向量。CBOW 和 Skip-Gram 是 word2vec 的两个模型,CBOW 更适合数据样本比较少的数据库,而 Skip-Gram 在大型语料中分类准确率更高,所以实验选用 Skip-Gram 模型构造 Word Embedding 模型[11]。

Skip-Gram 模型通过输入词 w_t 预测上下文 $Sw_t=(w_{t-k},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+k})$,其中 k 为 w_t 卷积窗口大小,即上下文中预测词向量的个数,CBOW 模型则是根据上下文 Sw_t 去预测某个词向量 w_t。Skip-Gram 模型和 CBOW 模型训练目标优化函数分别如式(1)和式(2)所示:

$$L_{\text{Skip-Gram}}=\sum_{w_t\in C}\sum_{-k\leqslant j\leqslant k}\sum_{j\neq 0}\log P(w_{i+j}\mid w_t) \tag{1}$$

$$L_{\text{CBOW}}=\sum_{w_t\in C}\log P(w_t\mid S_{W_t}) \tag{2}$$

其中,C 为数据样本中所有的词向量,k 为 w_t 上下文卷积窗口大小。Word2vec 为得到训练后网络中隐藏层的参数,这些参数就是 Word2vec 学习的词向量。

2　i-CNN 模型

虽然传统的 CNN 模型已经在卷积、池化层通过部分连接,下采样的方式减少了参数,但是随着样本数据的增加,会出现数据稀疏现象,造成空间的浪费,同时也会带来维数过高的问题,使得计算复杂度增高。为了解决这两个问题,将 Inception V1 模型与 CNN 模型结合,提出来 i-CNN 模型,在卷积层后加入 1×1 卷积核,通过降低特征向量的通道维数来减少参数,提高计算效率及分类精准率。

i-CNN 模型的框架流程如图 1 所示:将处理好的词向量输入模型,使用过滤器获取特征向量,改变窗口大小,得到多个特征向量,池化层筛选最强的特征向量,最终到 soft 分类器输出类别的概率,通过比较分类结果和数据集标签计算损失函数,再通过梯度下降算法反向传播,调节模型参数降低损失至网

络收敛,训练过程完成。

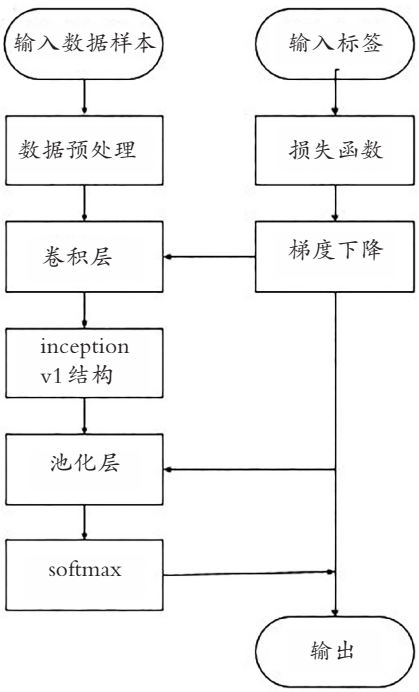

图 1 网络框架流程图

Fig. 1 Flow chart of network framework

2.1 输入层

输入层将中文邮件处理为词向量,一般是 $n\times k$ 的矩阵,其中 n 为文本单词总数,样本数据的词向量维度要保持一致,若长度不够,用 padding 补零,超出则舍弃。用开源工具包 word2vec 将每个词转化 k 维向量,k 维向量作为未知参数由训练得到。

如例句:本公司有部分普通发票、商品销售发票、增值税发票及其他服务行业发票。处理之后变成{本公司,有,部分,普通发票,商品,销售,发票,增值税,发票,及,其他,服务,行业,发票}。

2.2 卷积层的优化

2.2.1 卷积层

在卷积层对样本数据进行特征提取,这是建立模型最重要的一步。在输入的 $n\times k$ 的矩阵上,卷积操作定义如式(3)所示:

$$c_i=f(\boldsymbol{w}\cdot\boldsymbol{x}_{i:i+h-1}+b) \tag{3}$$

其中,$\boldsymbol{x}_{i:i+h-1}$ 代表输入矩阵的第 i 行到第 $i+h-1$ 行所组成的大小为 $h\times k$ 的滑动矩阵窗口,$\boldsymbol{w}$ 和 $\boldsymbol{x}_{i:i+h-1}$ 的维度一样,也是 $h\times k$,b 为偏置参数,f 为非线性激活函数,C_i 为标量。

卷积核在做卷积操作时,假设卷积窗口大小为 2,在 $2\times k$ 的滑动窗口上卷积,就可得到 $(n-1)$ 个结果,然后拼接组成 $(n-1)$ 维的特征向量,如式(4):

$$\boldsymbol{c}=(c_1,c_2,c_3,\cdots,c_{n-1}) \tag{4}$$

卷积层中可以设定不同的卷积窗口 h,提取不同的特征,单一的卷积操作可能会造成对句子意思理解的偏差,设置不同尺寸的卷积核,从每一个卷积核的卷积操作中提取一个特征,就可以提取出不同的特征,减少因语义产生的分类误差。

2.2.2 Inception V1 模型

Inception V1 模型主要改进网络中的卷积层,针对网络参数过多,容易过拟合,以及深层网络梯度下降停滞等问题,1×1 卷积核可以降低特征向量通道维度。在卷积层采用不同尺寸的窗口做特征提取,max pooling 本身也是一种特征提取,也可以作为一个分支,这就是 inception V1 模型。模型结构如图 2 所示,在 3×3,5×5 前,max pooling 后加上 1×1 卷积核,降低特征向量厚度。

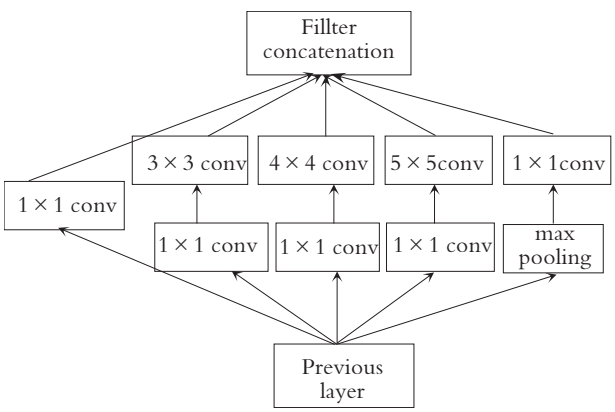

图 2 Inception V1 模型结构图

Fig. 2 Structure diagram of inception V1 model

根据图 2 做对比计算,假设上层输入特征向量的维度为 256,输出特征向量维度也是 256,直接用 3×3×256 维的卷积核输出 256 维特征向量,参数为 256×3×3×256=589 824 个。如果先经过 1×1×64 的卷积核,再经过一个 3×3×64 的卷积核,最后再经过一个 1×1×256 的卷积核,参数为 256×1×1×64+64×3×3×64+64×1×1×256=69 632 个,这样就减少了大部分的参数。

2.3 池化层

池化层对上层特征向量压缩,提取主要的特征,减少网络参数和后续计算复杂度。池化方法有很多种,主要采用 max 池化,即把上一层中的每个卷积核产生的最大特征向量作为特征值,如式(5):

$$\tilde{\boldsymbol{c}}=\max(c_1,c_2,\cdots,c_{n-h+1}) \tag{5}$$

上一层每个卷积核提取的特征生成一个特征

值,然后所有的 $\tilde{c}$ 值构成池化层的特征向量,送给全连接层,并使用 Softmax 函数输出每个文本类别的概率。

2.4　全连接层

全连接层对前面高度抽象化的特征整合,进行维度转换,再加上激活函数的非线性映射,通过 Softmax 函数对分类情况输出,Softmax 表达式如(6)所示:

$$y_i = \frac{e^i}{\sum_{i=1}^{n} e^i} \tag{6}$$

i 表示 n 维向量的第 i 个值,对于输入的 n 维向量,向量的每个值都表示输入对于每个类的概率。

3　实例分析与应用

3.1　实验数据

实验数据是从网络上获取并整理的中文邮件,根据语义将邮件分为正常邮件和垃圾邮件,正常邮件是需要的,垃圾邮件是一些广告、推文等。根据语义将邮件整理放在两个文件夹中,5 000 条正常邮件为正向评论,放在 Positive-examples 文件夹中,5 000 条垃圾邮件为负向评论,放在 Negative-examples 文件夹中。在两个不同数据集上进行实验,对本文所提出模型进行评估。对于每个数据集,随机抽取 80%作为训练集,20%作为测试集,并进行多次迭代实验,直到损失函数收敛。

3.2　实验环境

实验采用 Intel Core i5 处理器、8G 内存、Windows 系统、开发语言是 Anaconda 集成环境中的 Python3.7 编写实验代码,开发工具为 PyCharm Community Edition。主要的几种环境配置如表 1 所示。

表 1　实验环境及配置

Table 1　Experimental environment and configuration

实验环境	环境配置
操作系统	Windows10
编程语言	Python3.7
分词工具	Word2vec
学习框架	tensorflow

3.3　实验设计

本实验利用 gensim 库中的 Word2vec 中文工具包进行分词和词性标注,并将邮件中的词转化为词向量,进而对词向量训练。实验参数主要有词向量维度、卷积窗口大小、池化方法等。其中词向量维度设置为 128,卷积窗口大小在(3,4,5)中取值,卷积核数量设置为 128,比率为 0.5,批量尺寸为 64,评估间隔为 100,采用 max-pooling 方法进行池化。本文网络模型参数设置如表 2 所示。

表 2　网络模型参数

Table 2　Network model parameters

超参数	参数值
词向量维度	128
卷积窗口大小	3,4,5
卷积核数量	128
比率	0.5
批量尺寸	64
评估间隔	50
池化方法	max-pooling
Inception 卷积核	1×1

实验用整理好的中文邮件去验证 i-CNN 模型的分类效果,具体做了以下 3 种实验:

(1) 对训练集数据进行预处理,分词、去停用词、词向量转化输入网络模型,进行训练,每 50 次做一次评估检测保留 accuracy 值,用 accuracy 衡量模型性能。

(2) 为了比较 i-CNN 模型在中文邮件分类中与其他分类方法性能的优劣,用支持向量机(SVM)、朴素贝叶斯(NB)、k 邻近算法(KN)3 种传统的分类模型做对比试验,使用同一组数据集,正向评论和负向评论各 5 000 条。

(3) 为了比较 inception V1 模型对传统 CNN 模型的优化效果,在同一组数据集训练,各种参数保持一致。

3.4　实验结果分析

数据集在模型中的训练结果如图 3 所示,经过 600 次迭代训练,精准率稳定上升达到 90%以上,最高达到 93.7%,验证了 i-CNN 模型对垃圾邮件有较好的分类效果。

由表 3 可以看出,i-CNN 模型的分类精准率达到了 92.18%,除了与支持向量机的分类效果比较接近,其分类性能远高于朴素贝叶斯和 k 近邻算法两种传统的机器学习模型,说明 i-CNN 模型具有较好的分类性能。

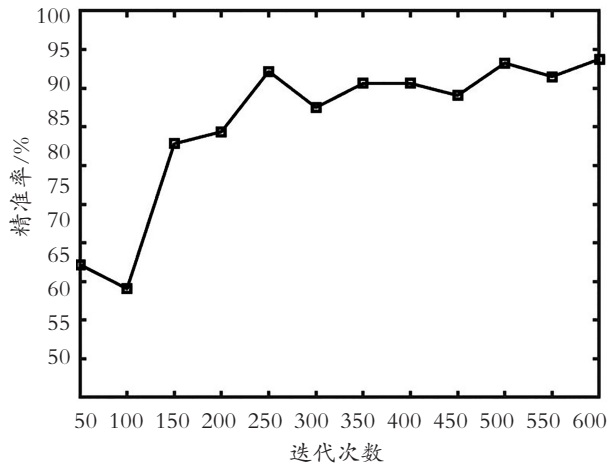

图 3 训练数据精准率

Fig. 3 Accuracy of training data

表 3 数据分类精准率

Table 3 Accuracy of data classification

模 型	精准率/%
SVM	90.11
Naive Bayes	86.5
KNeighbors	68.34
i-CNN	92.18

由表 4 可以看出:相同的数据样本中,在 CNN 模型中加入 inception 模型,精准率提高了 1.45%,效果不是很好,主要是网络的深度不够,inception V1 模型对深层的网络优化效果更好。

表 4 有无 inception 结构的结果对比

Table 4 Comparison of results with and without inception structure

模 型	精准率/%
CNN	90.73
i-CNN	92.18

4 结 论

通过对邮件的分类试验,可以看出 i-CNN 模型在文本分类方面有较好的效果,inception V1 模型确实能优化网络,提高分类效果,但从结果来看,还有很多的提升空间。对本实验来说,数据样本还不够大,种类还不够全面,网络深度不够,目前的网络模型缺乏通用性、灵活性,在提高分类效果方面还需要更多探索,可尝试与其他模型结合或者改进卷积神经网络模型来做进一步研究。

参考文献(References):

[1] 吴杰胜, 陆奎. 基于多部情感词典和规则集的中文微博情感分析研究[J]. 计算机应用与软件, 2019, 36(9): 93—99.
WU Jie-sheng, LU Kui. Research on sentiment analysis of Chinese microblog based on multiple sentiment dictionaries and rule sets[J]. Computer Applications and Software, 2019, 36 (9): 93—99.

[2] 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.
HUANG He, JING Xiao-yuan, DONG Xi-wei, et al. CNNs text mail classification model based on skipgram[J]. Computer Technology and Development, 2019, 29(6): 143—147.

[3] XU C, ZHANG J, CHANG K, et al. Uncovering collusive spammers in Chinese review websites[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. 2013.

[4] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014.

[5] 周锦峰, 叶施仁, 王晖. 基于深度卷积神经网络模型的文本情感分类[J]. 计算机工程, 2019, 45(3): 300—308.
ZHOU Jin-feng, YE Shi-ren, WANG Hui. Text sentiment classification based on deep convolution neural network model[J]. Computer Engineering, 2019, 45(3): 300—308.

[6] 唐贤伦,林文星,杜一铭,等. 基于串并行卷积门阀循环神经网络的短文本特征提取与分类[J]. 工程科学与技术,2019,51(4):93—99.
TANG Xian-lun, LIN Wen-xing, DU Yi-ming, et al. Feature extraction and classification of short text based on series parallel convolution gate valve recurrent neural network[J]. Engineering Science and Technology, 2019, 51 (4): 93—99.

[7] WANG F, CHENG J, LIU W Y, et al. Additive margin softmax for face verification[J]. IEEE Signal Processing Letters, 2018, 25(7): 926—930.

[8] LU C, HUANG H Y, JIAN P, et al. A P-LSTM neural network for sentiment classification[C]//Advances in

Knowledge Discovery and Data Mining. Cham: Springer, 2017.

[9] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016.
ZHOU Zhi-hua. Machine learning[M]. Beijing: Tsinghua University Press, 2016.

[10] 刘龙飞,杨亮,张绍武,等. 基于卷积神经网络的微博情感倾向性分析[J]. 中文信息学报,2015,29(6):159—165.
LIU Long-fei, YANG Liang, ZHANG Shao-wu, et al. Analysis of microblog sentiment tendency based on convolutional neural network[J]. Chinese Journal of Information, 2015,29(6): 159—165.

[11] 陈翠平. 基于深度信念网络的文本分类算法[J]. 计算机系统应用,2015,24(2):121—126.
CHEN Cui-ping. Text classification algorithm based on deep belief network[J]. Computer System Application, 2015, 24 (2): 121—126.

[12] LEE G,JEONG J,SEO S, et al. Sentiment classification with word attention based on weakly supervised learning with a convolutional neural network [J]. Knowledge-Based Systems,2018,152:70-82.

[13] IEIRA J P A,MOURA R S. An analysis of convolutional neural networks for sentence classification[C]// Proceedings of IEEE Computer Conference. Washington D C,USA:IEEE Press,2017.

Research on Mail Classification Algorithm Based on Improved Convolutional Neural Network

SONG Dan[1] , LU Kui [1] , DAI Xu-fan [2]

(1. School of Computer Science and Engineering, Anhui University of Science & Technology, Anhui Huainan 232001, China;2. School of Electrical and Information Engineering, Anhui University of Science & Technology, Anhui Huainan 232001, China)

Abstract: Aiming at the problems of high dimension and sparse data in traditional text classification methods, this paper proposes an e-mail classification method based on i-CNN model by combining convolutional neural network (CNN) and inception V1 model. In the convolution and pooling operation, 1×1 convolution kernel is added to reduce the thickness of eigenvectors, reduce the parameters and improve the computational performance. Through data validation, the result of i-CNN model for e-mail classification reaches as high as 92.18%. In the comparative experiment, compared with several machine learning classification models, i-CNN model achieved the highest classification accuracy. In the comparison with or without the inception structure model, i-CNN model accuracy is higher than CNN model. It shows that the model has a good classification effect, and the integration of inception V1 model can improve the accuracy of text classification.

Key words: text classification; convolution neural network; inception V1; word2vec

责任编辑:李翠薇

引用本文/Cite this paper:

宋丹, 陆奎,戴旭凡. 基于改进的卷积神经网络邮件分类算法研究[J]. 重庆工商大学学报(自然科学版),2022,39(3):20—25.

SONG Dan, LU Kui, DAI Xu-fan. Research on mail classification algorithm based on improved convolutional neural network[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2022,39(3):20—25.