基于 Skip-gram 的 CNNs 文本邮件分类模型

黄　鹤[1],荆晓远[2],董西伟[2],吴　飞[2]
(1. 南京邮电大学 计算机学院,江苏 南京 210023;
2. 南京邮电大学 自动化学院,江苏 南京 210023)

摘　要:随着互联网广告技术的发展和电子邮件的普及,越来越多的垃圾广告邮件充斥生活,而对如何高效区分垃圾邮件的研究也逐渐成为了热门课题。自然语言在结构上具有很强的前后相关性,而且对于中文邮件直接转化成向量会有过高的维度产生,影响最后分类的准确性。对此,首先对邮件文本进行分词,再利用 skip-gram 模型训练出数据集中每个词的 word embedding,引入的词嵌入(word embedding)是为了将邮件文本转化成低维度特征向量;然后将每个词的 word embedding 组合为二维特征矩阵作为网络的输入,此外在每一次的迭代过程中,输入特征也作为参数进行更新;最后送入提出的 CNN-HIGHWAY 混合模型中进行邮件分类。将该混合模型在 CCERT 中文邮件样本集上进行实验,并与传统的机器学习方法和标准的卷积神经网络模型进行对比,结果表明该模型不仅解决了维度过高的问题,而且提高了邮件分类的准确率。

关键词:自然语言处理;词嵌入;邮件分类;卷积神经网络;深度学习
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2019)06-0143-05
doi:10.3969/j.issn.1673-629X.2019.06.030

CNNs-Highway Text Message Classification Model Based on Skip-gram

HUANG He[1],JING Xiao-yuan[2],DONG Xi-wei[2],WU Fei[2]
(1. School of Computer,Nanjing University of Posts and Telecommunications,Nanjing 210023,China;
2. School of Automation,Nanjing University of Posts and Telecommunications,Nanjing 210023,China)

Abstract:With the development of Internet advertising technology and the popularity of e-mail,more and more spam advertisements are flooding the lives. The research on how to effectively distinguish spam has gradually become a hot topic. The natural language has a strong front-to-back correlation in structure and also too high dimensions for the direct translation of Chinese emails into vectors,which adversely affects the accuracy of the final classification. Therefore,we propose a model which firstly segments e-mail texts and uses the skip-gram model to train the word embedding of each word in the data set. The introduced word embedding is to convert the message text into a low-dimensional feature vector. Then the word embedding of each word is combined into a two-dimensional feature matrix as the input of the network. In addition,during each iteration,the input features are also updated as parameters. Finally,the feature vectors are sent to the proposed CNN-HIGHWAY hybrid model for classification. The hybrid model is tested on the CCERT Chinese mail sample set. Compared with the traditional machine learning methods and the standard convolutional neural network models,this model not only solves the problem of high dimensionality,but also improves the accuracy of mail classification.

Key words:natural language processing;word embedding;mail classification;convolutional neural network;deep learning

0　引　言

随着互联网广告技术的发展和电子邮件的普及,越来越多的垃圾广告邮件充斥生活,垃圾邮件可以说是因特网带给人类最具争议性的副产品之一。占用网络带宽,造成邮件服务器拥塞,进而降低整个网络的运行效率;侵犯收件人的隐私权,侵占收件人信箱空间,耗费收件人的时间、精力和金钱。有的垃圾邮件还盗用他人的电子邮件地址做发信地址,严重损害了他人的信誉;因此,如何快速有效地对这类垃圾邮件进行过滤成为了热门的研究课题,同时也成为了自然语言处

收稿日期:2018-07-11　　**修回日期**:2018-11-15　　**网络出版时间**:2019-03-06
基金项目:国家自然科学基金(61702280)
作者简介:黄　鹤(1989-),男,硕士研究生,研究方向为自然语言处理、模式识别;荆晓远,教授,博导,研究方向为模式识别、图像与信号处理、信息安全、机器学习与数据挖掘等。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190306.0938.046.html

理中一个非常重要的研究方向。近年来,深度学习在计算机视觉领域的惊人表现有目共睹,与此同时,在自然语言处理(natural language processing,NLP)的应用也越来越广泛。对于垃圾邮件分类,国内外的很多学者已经做出了很多探索,使用的方法主要包括传统的机器学习方法和现在流行的深度学习方法。

李婷婷等[1]尝试从文本数据中进行人工特征构建,然后用传统的机器学习方法进行分类,这种方法实质上属于机器学习范畴,其分类效果严重依赖于特征的构建质量并且整个过程非常耗时耗力。陈翠平等[2]利用深度置信网络从高维的原始特征中抽取高度可区分的低维特征,最后用深度学习的思想来完成分类任务。这相比于人工构建特征的方式,更加高效地完成了特征提取任务,但只有网络足够深时才能够提取出较好地反映出文本语义信息的特征,使得模型参数数量和训练时间大大增加。如果用深度学习对文本进行预处理就需要将文本进行数字化表示。以前的词表示方法主要是独热编码 One-hot,但是这样做的缺点是维度过高且数据稀疏,对于自然语言处理来说,也不能很好地保留词语前后的语义信息。目前词嵌入(Word embedding)可以有效地保留词汇语法、语义信息的词向量转换方式。借助于词向量的方法,从而使用深度学习提取有效的邮件文本特征成为了可能。

其次对于垃圾邮件分类问题,有机器学习方法,如 Shen 等[3]用决策树构造三步法进行邮件过滤;Feng 等[4]提出的 SVM-NB 方法取得了较高的垃圾邮件检测精度。2014 年,Kim 等[5]将 Word embedding 与 CNN 相结合应用于情感分析和文本分类等若干自然语言处理任务中,取得了非常好的效果。

在上述研究的基础上,文中利用 Word embedding 将词分布式映射到一个低维空间,解决传统的 One-hot 编码词向量维度过高问题,然后结合 CNN 进行特征提取,其中还加了一层 Highway 网络,它是对提取出的特征进行优化。最后在 CCERT 数据集上进行实验,通过与其他方法的对比,证明该模型提高了准确率。

1 相关工作

1.1 Word embedding

在传统的自然语言处理问题中,首先要把词转换为词向量的形式,这样计算机才可以用各种算法处理自然语言问题。One-hot 是一种很经典的词向量表示方法,这个向量的维度是词表的大小,其中绝大多数是 0,只有一个维度的值为 1,这个维度就代表当前的词,如下所示:

"北京"表示为[0 0 0 1 0 0 0 0 0 0 0…]

"首都"表示为[0 0 0 0 0 0 0 0 1 0 0…]

这里任意两个词之间是独立的,且维度过高,这样构成的词向量非常稀疏,也很难反映出词语间的语义关系。为了解决这个问题,Hinto 等[6]提出了 Word embedding 的词向量表示方法,主要是将词分布式地映射到低维空间,这样就解决了向量稀疏问题。并且,该低维空间中词向量的位置关系很好地反映了它们在语义上的联系,这样能很好地反映了文本的特征。Mikolov 等[7]在 Bengio 等[8]研究的基础上提出了 CBOW 和 Skip-gram 模型,如图 1 所示。

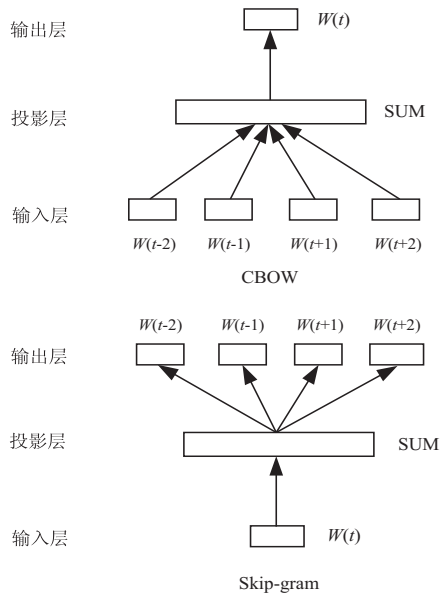

图 1 CBOW 和 Skip-gram 模型

CBOW 模型是根据上下文的词来预测目标词,而 Skip-gram 模型是根据当前词来预测上下文的词。相比其他 Word embedding 模型,Skip-gram 模型训练时间短且效果较好。所以文中使用 Skip-gram 模型来构造 Word embedding 模型[9]:假定有一组词序列 w_1,w_2,…,w_n,Skip-gram 的目标是使式 1 最大化。

$$L=\frac{1}{N}\sum_{n=1}^{N}\sum_{-c\leqslant i\leqslant c,i\neq 0}\log p(w_{n+i}\mid w_n) \tag{1}$$

其中,C 是前后文的词数,C 越大,最后训练出来的 Word embedding 越好,但是会增加训练的时间。

1.2 卷积神经网络与自然语言处理

当谈到卷积神经网络(convolutional neural network,CNNs)时,往往会联想到计算机视觉。CNNs 在图像分类领域做出了巨大贡献,也是当今绝大多数计算机视觉系统的核心技术。Krichevsky 等[10]设计的卷积神经网络在 2012 年的 ImageNet 挑战赛中获得冠军。LeCun 等[11]用卷积神经网络成功解决了手写体数字识别问题。由于卷积神经网络有着诸多优点,现在越来越多的研究者将其应用到各种领域。Honglak 等[12]通过卷积深度置信网络进行音频特征分析。

Collobert 等[13]将卷积神经网络应用于自然语言处理,证明其提出的模型在各项任务中都有出色表现。Shen 等[14]利用卷积神经网络解决了信息检索中的语义分析问题。Kalchbrenner 等[15]利用卷积神经网络对句子进行建模,而且还给出了 pooling 的一个新方式。而文中利用卷积神经网络提取文本邮件的特征,再利用 Highway 网络对卷积特征进行优化,从而提高分类的准确率。

2　模型框架

2.1　数据预处理

2.1.1　分　词

众所周知,图像是由一个个的像素点组成,计算机在处理图片时,输入数据是二维矩阵。因此在处理文本邮件时,首先要把文本数据做成矩阵形式,但在这之前,还要对文本进行分词。对中文文本分词,不像对英文文本那样,只要根据空格和标点符号将词语分割成数组即可,因为英语的句子基本上就是由标点符号、空格和词构成。但是,中文文本是由连续的字序列构成,词与词之间是没有天然的分隔符,所以中文分词相对来说困难很多。中文分词目前来说基本上都还是基于分词用的词典来进行分词的,将字和字组成不同的词然后放入词典中查找。中文分词面临着相当多的挑战,首当其冲的就是歧义问题,不同的分割方式会导致不同的意思。

如下面两句话中词和词序完全一样,但是不同的分词,意思完全不一样:

结婚/的/和/尚未/结婚/的人

结婚/的/和尚/未/结婚/的人

还有个重要的挑战就是未收录的词,人名就是最简单的例子,还有就是网友发明的词,诸如:“草泥马”、“不明觉厉”之类的。所以一个好的分词词典是决定中文分词质量的关键,还有就是做中文分词的话必须经常更新、与时俱进。文中用的是 jieba 分词器,它是基于 Python 的一个中文分词模块,内置词典,词典有 50 万个词条。另外,对于未登录词,jieba 采用了基于汉字成词能力的 HMM(hidden markov model)模型,使用了维特比(Viterbi)算法的四种状态的模型。

2.1.2　停用词处理

邮件文本一般都是短文本,包含的词语少,无关词语带来的影响也会更明显,其中也包括一些无关的表情符号,因此去停用词也是一个重要环节,这里对“哈工大停用词词库”、“四川大学机器学习智能实验室停用词库”、“百度停用词表”进行整理,去重,在提取中文词(而不是大量英文词和中文标点符号)出了一个比较全面的词表,用于文中邮件数据处理。

2.1.3　Word embedding 训练

这里用 Skip-gram 模型进行 Word embedding 训练,它是根据当前词来预测上下文的词。它主要是将词分布式地映射到低维空间,并且,该低维空间中词向量的位置关系很好地反映了它们在语义上的联系,这样能很好地反映文本的特征,如图 2 所示。然后再把做好后的词向量做成二维矩阵,作为卷积神经网络的输入数据。

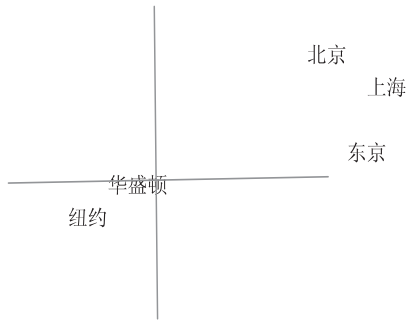

图 2　词向量低维空间

从图 2 的例子可以发现,华盛顿和纽约聚集在一起,北京和上海聚集在一起,且北京到上海的距离与华盛顿到纽约的距离相近。也就是说模型学习到了城市的地理位置,也学习到了城市地位的关系。

在这一步中,首先要找到最大长度的句子,再把其他所有句子都统一填充到这个长度。填充句子到同样的长度是必须的,因为批处理的每个样本都必须有相同的长度,所以这样可以高效地把数据划分成批。这里用零填充的方法。

假设邮件文本中,经过数据预处理后,长度最长的邮件包含 n 个词,该邮件中的第 i 个词所对应的词向量是 $\boldsymbol{v}_i \in R^d$,那么卷积神经网络的输入就是由 n 个 d 维向量组成的 $n \times d$ 的二维矩阵,如图 3 所示。

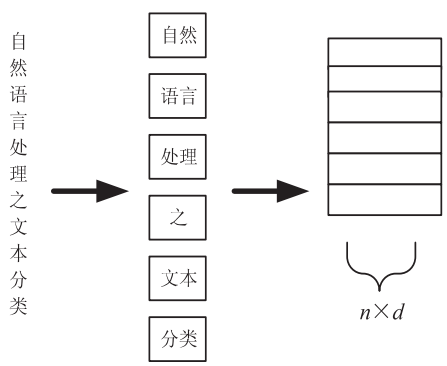

图 3　输入的二维矩阵

图 3 是文本数据经过预处理后,一个句子的矩阵表示,它是由句子中的所有词汇的词向量纵向拼接在一起,可以表示为:

$$\boldsymbol{v} = \boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \otimes \cdots \otimes \boldsymbol{v}_n \tag{2}$$

其中,$\otimes$ 是纵向拼接操作符;$\boldsymbol{v}$ 是一个样本邮件的矩阵表示。

2.2 CNNs-Highway 网络优化和分类

2.2.1 CNNs

文中进行多层卷积操作,卷积核是有三种不同的大小对文本进行卷积,分别提取多组特征向量,若卷积核的高是 h ,维度是 d ,卷积后的特征向量是 t_i ,有:

$$t_i = f(w \cdot x_{i:i+h-1} + b) \tag{3}$$

其中, w 是卷积核的权重参数; b 是偏置值; f 是激活函数,一般常用的是 sigmid 函数或 tanh 函数,文中为了加快收敛速度用 ReLu 函数:

$$f(x) = \max(0, x) \tag{4}$$

最后对所有的邮件文本进行卷积之后得到总特征向量 ($T \in R^{n-h+1}$) :

$$\boldsymbol{t} = [t_1, t_2, \cdots, t_{n-h+1}] \tag{5}$$

得到卷积后的特征向量之后,再进行 max-pooling 操作,进一步提取特征 $T = \max\{t\}$,然后把所有的 T 进行拼接,从而获得最具代表性的特征。在连接全连接层之后,为了防止出现过拟合现象,增加了 Dropout 操作,禁止一部分的神经元参加更新过程,这样就使得权重的更新不依赖于固定节点的作用。

2.2.2 Highway 网络优化层

随着神经网络的发展,网络的深度逐渐加深,网络的训练也变得越来越困难。Highway Networks:一种可学习的门限机制,在此机制下,一些信息流没有衰减地通过一些网络层。一般一个有 L 层的传统前向神经网络,每层网络对输入进行非线性映射变换,表达为:

$$y = H(x, W_H) \tag{6}$$

其中, H 为非线性函数; W_H 为权重; x 为输入; y 为输出。

对于 Highway CNN 网络,在上述基础上增加两个非线性激活函数 T 与 C ,则:

$$y = H(x, W_H) \cdot T(x, W_T) + x \cdot C(x, W_C) \tag{7}$$

一般情况设置 $C = 1 - T$,则式 7 可以改写为:

$$y = H(x, W_H) \cdot T(x, W_T) + x \cdot (1 - T(x, W_C)) \tag{8}$$

其中,参数 x, y, H, T 的维数须一致,不够补零。 T 被称作转换门,可以对输入的信息流进行处理,这种方法已经被证实可以解决训练收敛问题,提高模型性能。

3 实验与分析

3.1 实验环境

实验环境如表 1 所示。

表 1 实验环境及配置

实验环境	环境配置
操作系统	Windows10
编程语言	Python 3.6
分词工具	jieba 0.39
深度学习框架	Tensorflow 1.4
Skip-grm 训练工具	word2vec

3.2 数据集的选择与处理

实验数据来自 CCERT 中文邮件数据集。该数据集中包含 5 000 条正常邮件,5 000 条垃圾邮件。垃圾邮件里包含一些广告,广告商的电话号码,地址,等等。数据的划分是打乱所有数据,90% 作为训练集,10% 作为测试集;数据划分完成后,就进行一系列的数据预处理,处理过程见 2.1 节。

3.3 实验设计

文中尝试用 Skip-gram 的词向量模型,对文本邮件进行词向量化作为卷积神经网络的输入,用 CCERT 中文邮件样本集来完成 CNNs-Highway 混合邮件分类模型的性能测试,具体实验设计如下:

(1)CNNs+Skip-gram。在 Skip-gram 模型训练好的词向量表中查找每条样本中出现的每个词的 word embedding,并组合成 $m \times k$ 的二维数据矩阵作为 CNN 的输入。其中 m 为数据集中最长评论所包含的词数,对于长度小于 m 的样本需要补零; k 为 Word embedding 长度。

(2)CNNs+rand。CNNs 模型部分保持不变,按高斯分布随机初始化 Word embedding。实验目的是通过与 CNNs+Skip-gram 模型的结果相比较,从而验证 Word embedding 在描述原始数据特征分布方面的性能。

(3)传统机器学习模型。在相同数据集上,利用几种常用的机器学习模型作为对比来证明基于 Skip-gram 的 CNNs-Highway 模型在邮件分类任务上的性能优势。为了排除由于特征构建方式的不同而导致实验结果无法比较,传统模型的特征构建方式同样基于 Word embedding,每条样本的特征为该样本中所有 Word embedding 的均值。

3.4 实验结果与分析

表 2 实验结果表明了不同的特征词向量初始化对最后分类的影响。其中 rand 与 non-static 初始化的分类结果相近,这是因为在网络数据中存在大量的新词、表情、链接地址,从而导致该数据集的词典中将近一半的词都没有在词向量中出现过,这也使得随机赋值的词向量在词典中比重较大,使用随机模型 CNN-rand 也就与词向量模型 CNN-non-static 相差甚微。

表 2 不同词向量初始化分类结果对比

模型	准确率/%
rand+CNNs-Highway	91.82
static+CNNs-Highway	90.56
non-static+CNNs-Highway	91.88
Skip-gram+CNNs-Highway	94.48

表 3 的实验结果证明了文中的邮件分类模型相比

传统的机器学习方法获得了出色的性能提升。结合表 2 和表 3 发现,在随机初始化的 rand+CNNs-Highway 基础上准确率就已经超过了传统的机器学习算法 0.71%,结合 Highway 网络对卷积神经网络提取的特征进行优化后,最终的分类结果提高到了 94.48%。

表 3　算法对比

模型	准确率/%
KNeighbors(k =4)	69.31
DecisionTree(entropy)	84.15
DecisionTree(gini)	86.34
RandomForest	90.12
SVM(linear)	91.11
LDA	91.78
Skip-gram+CNNs-Highway	94.48

由表 4 显示,在同样的词向量模型下,加入 Highway 网络优化层比没有加入 Highway 网络的结果提高了 0.47%,提高的不是很明显。从理论上分析,是因为 Highway 网络对深层神经网络优化比较明显,对浅层的网络优化一般。所以这也是下一步需要改进的地方,在卷积网络层加大深度,从而优化文本邮件分类模型。

表 4　有无 Highway 网络层的结果对比

模型	准确率/%
Skip-gram+CNNs	94.01
Skip-gram+CNNs-Highway	94.48

4　结束语

针对如何高效准确地过滤出垃圾邮件的问题,提出了一种基于 Skip-gram 的 CNNs-Highway 文本邮件分类模型。首先对邮件数据集进行中文分词,去停用词,然后用 Skip-gram 模型训练词向量作为 CNNs 的输入,其中经过 CNNs 提取特征以后,还加入了 Highway 网络优化层。实验结果表明,该模型在文本邮件分类准确率上得到了明显提高。未来的研究工作包括以下几方面:改进 CNNs,增加 CNNs 的深度,使之更适应 Highway 网络的优化;在邮件分类问题中,考虑垃圾图片邮件的过滤。

参考文献:

[1] 李婷婷,姬东鸿. 基于 SVM 和 CRF 多特征组合的微博情感分析[J]. 计算机应用研究,2015,32(4):978-981.

[2] 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.

[3] SHEN J J,CHEN Y K,CHU K T,et al. An intelligent three-phase spam filtering method based on decision tree data mining[J]. Security & Communication Networks,2016,9(17):4013-4026.

[4] FENG Weimiao,SUN Jianguo,ZHANG Liguo,et al. A support vector machine based naive Bayes algorithm for spam filtering[C]//2016 IEEE 35th international performance computing and communications conference. Las Vegas,NV,USA:IEEE,2017:1-8.

[5] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of conference on empirical methods in natural language processing. Doha:[s. n.],2014:1746-1751.

[6] HINTON G E. Learning distributed representations of concepts[C]//Proceedings of the 8th annual conference of cognitive science society. Amherst,Mass:[s. n.],1986:12-23.

[7] MIKOLOV T,CHEN Kai,CORRADO G,et al. Efficient estimation of word representations in vector space[J]. Computer Science,2013,2(12):27-35.

[8] BENGIO Y,DUCHARME R,VINCENT P,et al. A neural probabilistic language model[J]. Journal of Machine Learning Research,2003,3:1137-1155.

[9] MIKOLOV T,SUTSKEVER I,CHEN Kai,et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th international conference on neural information processing systems. Lake Tahoe,Nevada:Curran Associates Inc.,2013:3111-3119.

[10] KRICHEVSKY A,SUTSKEVER I,HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. [s. l.]:MIT Press,2012:1097-1105.

[11] LECUN Y,BOTTOU L,BENGIO Y,et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.

[12] LEE H,LARGMAN Y,PHAM P,et al. Unsupervised feature learning for audio classification using convolutional deep belief networks[C]//Proceedings of the 22nd international conference on neural information processing systems. Vancouver,British Columbia,Canada:Curran Associates Inc.,2009:1096-1104.

[13] COLLOBERT R,WESTON J,BOTTOU L,et al. Natural language processing(almost) from scratch[J]. Journal of Machine Learning Research,2011,12:2493-2537.

[14] SHEN Yelong,HE Xiaodong,GAO Jianfeng,et al. Learning semantic representations using convolutional neural networks for Web search[C]//Proceedings of the 23rd international conference on world wide web. New York:ACM Press,2014:373-374.

[15] 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. Baltimore,USA:[s. n.],2014:655-665.