基于改进 CNN 的文本情感分析

何　野,杨会成,潘　玥,徐姝琪
(安徽工程大学 电气工程学院,安徽 芜湖 241004)

摘　　要:传统的机器学习技术,包括支持向量机(SVM)等技术,已经被应用到文本情感分析的各种任务中,这使得复杂分类问题的泛化能力很差. 近年来,机器学习在自然语言处理研究方面取得了突破. 卷积神经网络(CNN)和递归神经网络(RNN)是文本分析的两种主流方法. 通过对神经网络模型的研究,提出了一种使用卷积神经网络(CNN)的多个分支与长短时记忆神经网络(LSTM)层的组合内核来进行情感分析的方法,并通过实验验证了其性能优于现有的 CNN 模型和 LSTM 模型.

关　键　词:情感分析;LSTM-CNN;机器学习;自然语言处理
中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1673 - 1670(2021)05 - 0059 - 04

0　引言

随着电子产品及互联网技术的不断进步与革新,包含人们情感评论的大量文本信息已经出现在网络平台上. 在自然语言处理(NLP)领域中,寻找一种有效的数据挖掘和分析方法是一项非常重要的研究,这被称为文本情感分析[1]. 文本情感分析主要包括文本分类、信息提取和文本生成技术,情感分析是一个过程,用于识别和分类意见、观点(从文本到特定主题或产品). 这些评估可以是正面、负面或中立的. 分析可以在文档级别、句子级别或单词级别执行. 目前,基于统计机器学习的情感分析在各种应用中都取得了良好的效果. 然而,机器学习方法中使用的函数非常简单,这可能导致它们在处理复杂分类问题时对新的模型的适应能力较差,并且在样本和计算单位比较少的情况下,表达复杂函数的能力在一定程度上受到限制.

历年来,情感分类方法发生了许多变化,最早是基于初始情感词典的方法, Bengio 等[2]最早使用神经网络构建语言模型. Mikolov 等[3-4]于 2013 年提出了 Word2Vec 技术,推动了词向量的快速发展. 然后再到机器学习方法,例如支持向量机(SVM)[5]、朴素贝叶斯(NB)、决策树、逻辑回归等. 尽管某些机器学习方法可以在某些任务上取得良好的结果,但是由于特征工程的复杂性,这些方法的效果非常依赖于特征表示,并且难以获得可接受的分类结果. 随着人们对深度学习算法的进一步研究与认识,许多深度学习方法被应用于情感分类任务. 与机器学习方法相比,深度学习则更加侧重于取代人工使用自动提取语料特征的方式,这样可以使文本的特征更加广泛和丰富. 但是,深度学习模型的鲁棒性和泛化能力在很大程度上取决于训练阶段可用的数据量,而基于机器学习的分类系统的性能则主要取决于语料库中的标记训练及其有效特征的选择.

卷积神经网络[6](CNN)和递归神经网络是两个广泛使用的用于情感表达的深度学习模型. Yann LeCun 在 1998 年提出的 CNN 具有很强的适应性,并且非常善于从文本中提取局部特征. 由于其独特的权重共享结构,它可以显著降低计算复杂度以及训练参数的数量. 对于句子建模, Kalchbrenner 等人提出了动态卷积神经网络,它能够获得短时和长时关系. 作为另一个流行的网络,循环神经网络[7](RNN)可以处理序列数据并了解长期依赖性. RNN 将当前输出和网络的上一级输出联系在一起,这意味着当前隐藏层的输入不仅包括输入层

收稿日期:2021 - 03 - 31
作者简介:何　野(1996—),男,安徽省安庆市人,安徽工程大学电气工程学院硕士研究生,主要从事机器学习神经网络及其图像处理研究.
通信作者:杨会成(1970—) ,男,安徽省滁州市人,工学硕士,安徽工程大学电气工程学院教授,硕士生导师,主要从事图像信息处理、疲劳驾驶检测研究.

的输出,而且还包括先前隐藏层的输出. 当 RNN 学习到对信息的长期依赖时,它将产生梯度衰减或爆炸. 为了解决这个问题,研究人员提出了一种 LSTM[8] 单元,其中包含一个可以长时间保持状态的存储单元,这样可以确保结构更准确地提取情感信息. 笔者在研究 CNN 和 LSTM 的基础上,将这两种框架的部分结构结合在一起,CNN 仅提取本地特征,而 LSTM 是一种网络类型,其网络内存可以记住输入中的先前数据,并根据该知识做出决策. 因此,LSTM 更适合直接输入文本,因为句子中的每个单词都具有基于周围单词的含义. 充分利用了它们各自的优势,弥补了单个网络框架的缺陷,并通过实验证明了该网络模型对提高文本情感分析具有较为高效的作用.

1 相关技术与概念

1.1 CNN 模型

CNN 是卷积神经网络(Convolutional Neural Network)的简称,它本质上是一个多层感知机. 该模型主要由 3 个部分构成:卷积层、池化层和全连接层(输出层). 结构如图 1 所示.

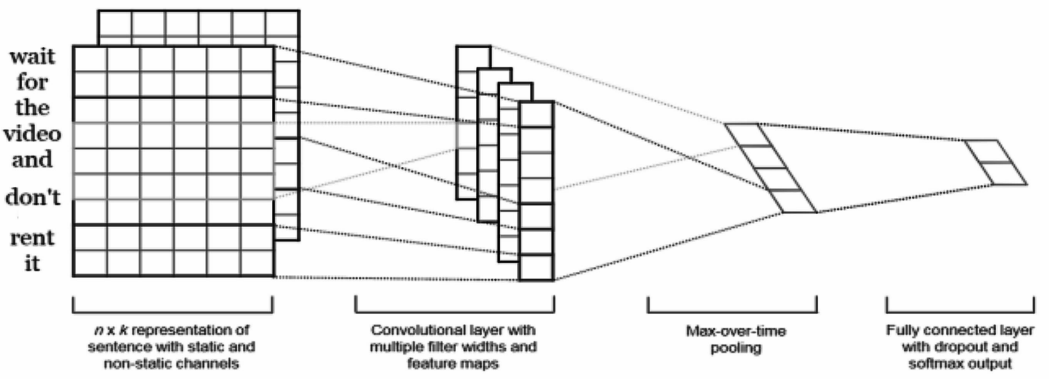

图 1 CNN 模型结构

它是一个具有多个隐层的人工神经网络. 卷积和池化是网络中的最关键的操作,也是网络进行局部和全局特征提取的主要方式. CNN 采用梯度下降法,通常都能够得到最优解,经过多次的训练调整提高网络的参数的合理性.

1.2 长短时记忆 LSTM 模型

在处理时间序列问题上,无论是分类还是预测的问题,循环神经网络(RNN)都有着很好的优势. 它的神经单元经过运算输出结果后,继续将其作为下一个单元的输入并循环往复,这样可以有效利用前面的信息. 在文本处理以及情感分析的问题上,循环神经网络能够贯穿全文,利用上下文的信息,从而使分类问题变得更加精准. 然而,传统循环神经网络难以对长文本进行处理,因其容易造成梯度爆炸和消失的问题.

长短时记忆网络(Long Short Term Memory Network,LSTM),是一种改进之后的循环神经网络,可以解决 RNN 无法处理长距离的依赖的问题,在涉及长时间滞后的任务上,其性能将优于 RNN.

LSTM 网络结构由 4 个主要部分组成:输入门、自循环链接、遗忘门和输出门.

对于输入门 i、遗忘门 f 和输出门 o,在 t 时刻分别有如下操作:

$$i_t = \sigma(\boldsymbol{W}_i x_t + \boldsymbol{U}_i h_{t-1}), \tag{1}$$

$$f_t = \sigma(\boldsymbol{W}_f x_t + \boldsymbol{U}_f h_{t-1}), \tag{2}$$

$$o_t = \sigma(\boldsymbol{W}_o x_t + \boldsymbol{U}_o h_{t-1}), \tag{3}$$

$$c'_t = \tanh(\boldsymbol{W}_c x_t + \boldsymbol{U}_c h_{t-1}), \tag{4}$$

$$c_t = i_t \cdot c'_t + f_t \cdot c'_{t-1}, \tag{5}$$

$$h_t = o_t \cdot \tanh(c_t). \tag{6}$$

其中 $\boldsymbol{W}_i$、$\boldsymbol{W}_f$、$\boldsymbol{W}_o$、$\boldsymbol{W}_c$、$\boldsymbol{U}_i$、$\boldsymbol{U}_f$、$\boldsymbol{U}_o$、$\boldsymbol{U}_c$ 均为权重矩阵,网络结构如图 2 所示.

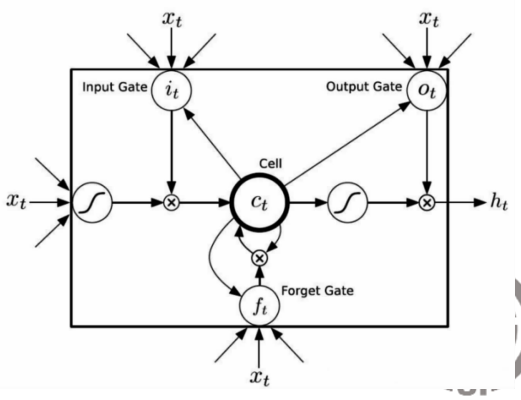

图 2 LSTM 模型结构

2　基于 LSTM-CNN 的模型

LSTM-CNN 模型由一个初始 LSTM 层构成,它将接收词向量输入矩阵作为输入,LSTM 层为原始输入生成一个新的编码. 然后将 LSTM 层的输出紧接着输入到期望可以提取局部特征的卷积层中. 最后卷积层的输出将被汇集到一个较小的纬度,最终输出为正或负标签. 它的结构如图 3 所示.

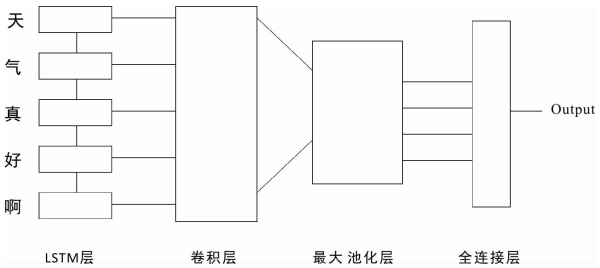

图 3　LSTM-CNN 组合网络

2.1　文本的词向量表示

对于情感分析来说,英文和中文存在着差别,即英文一般都是以单词来表达意思,而中文则以词组来表达情感. 所以在进行中文情感分析之前,首先得对数据集进行预处理,即对句子进行分词,去除无用符号和停用词等,其次将处理好的文本用词向量表示.

2.2　Word2Vec

Word2Vec 是一个非常有效的工具,可以在无须人工干预的情况下提取特定域的功能. 另外,对于太小的文本或单个词语,它都可以很好地工作. 通过提供庞大的语料库上下文并使用 Word2Vec,可以创建具有正确意义的词语并在大型数据集上更快地运行.

单词含义是深度学习的最终视角,使用 Word2Vec 对较大的实体进行分类可以完全满足单词的含义. 在提出的方法中,数据集是在向量上训练的. 具有相同情感标签的单词具有相同的向量,因此可以轻松指定单词相似度.

2.3　LSTM-CNN 网络搭建

该网络由以下 3 个部分组成:

1)卷积神经网络的输入矩阵. 所有的词向量都被连接成二维矩阵,作为卷积神经网络的输入矩阵.

2)卷积神经网络. 在本文模型中,由 4 层卷积层构建卷积神经网络模型,以提取句子中的重要特征信息. 卷积由卷积内核执行. 对于长度为 l 的内核,有:

$$c_i = f(\boldsymbol{\omega} \cdot x_{i:i+l-1} + b)\ . \tag{7}$$

式中,$\boldsymbol{\omega} \in R^{l \times d}$ 是内核的权重矩阵,$x_{i:i+l-1}$ 被用于内核嵌入基质的字. 而对于长度为 n 的句子,则得到特征向量 $c = [c_1, c_2, \cdots, c_i, c_n]$.

3)最大池化层和全连接层. 对于具有 m 个内核的窗口,利用最大池化层来增加这些特征的重要信息,并降低参数的复杂度,得到 $\hat{c} = [\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_m]$,式中 $\hat{c} \in R^m$ 是提取的特征向量. 此外利用激活函数预测情绪倾向:

$$\hat{y}(x) = g(\hat{c})\ . \tag{8}$$

式中,$g(\hat{c})$ 是激活函数,$\hat{y}(x)$ 是输入语句的预测标签(正数为 1,负数为 0).

2.4　模型的训练

该模型训练的损失函数是通过反向传播算法更新参数的形式得到的,损失函数如式(9)所示:

$$F_{\text{loss}} = \frac{1}{2n}\sum \left(\| y(x) - \hat{y}(x) \|^2 + \alpha \cdot \text{sim\_cos}(x, m) \right)\ . \tag{9}$$

式中,α 是控制解卷积效果的参数,$y(x)$ 是输入句子的真实标签,而 $\hat{y}(x)$ 是预测标签,sim_cos( ) 是估计的输入之间的余弦相似度函数 ,并通过反向传播算法更新参数.

3　实验结果与分析

笔者使用电子商务网站的产品评论为数据集,共 20 065 条数据,使用数字 1 代表积极情绪,数字 0 代表消极情绪. 其中,积极评价数据 10 212 条,消极评价数据 9 853 条. 模型中,文本长度为 29,词向量维度为 128,Adam 的学习率为 0.002,dropout 为 0.5,批次大小设置为 100.

为了验证模型的性能,本次实验另外训练了两种其他模型作为对比数据,分别是 CNN 和 LSTM 模型,通过对比各个模型的准确率(accuracy)、精确率(precision)、召回率(recall)和 F-measure 的值来判断优化程度.

实验结果如表 1 所示.

表 1　模型测试结果

模型	准确率	精确率	召回率	F-measure
CNN	0.73	0.6	0.6	0.47
LSTM	0.58	0.37	0.36	0.33
LSTM-CNN	0.76	0.71	0.71	0.72

通过模型 1 和模型 2 的实验对比,发现 CNN 模型在处理文本的各方面均优于 LSTM 模型.

对以上 3 种实验结果进行分析,LSTM-CNN 模型相比其他两种单个模型,在对评论文本情感分析上有着更好的表现,它的 F-measure 值均高于其他两种模型.

4 结论

综上所述,针对文本情感分析问题,在研究了 CNN 和 LSTM 模型的基础上,笔者提出的基于 LSTM-CNN 算法的文本情感分析模型在各方面数据显示其具有较为优异的处理能力,实验结果验证了该模型的可行性和有效性. 将来,可以尝试将模型与其他自然语言处理技术串联起来,以期在 NLP 问题中获得更好的结果.

参考文献:

[1]孙艳,周学广,付伟. 基于主题情感混合模型的无监督文本情感分析[J]. 北京大学学报(自然科学版),2013,49(1):102 – 108.

[2] BENGIO Y, SCHWENK H, SENÉCAL J S, et al. Neural Probabilistic Language Models[M] // Innovations in Machine Learning. Berlin: Springer, 2006: 137 – 186.

[3] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]. Lake Tahoe: 27th Annual Conference on Neural Information Processing Systems, 2013: 3111 – 3119.

[4]蔡慧苹,王丽丹,段书凯. 基于 word embedding 和 CNN 的情感分类模型[J]. 计算机应用研究,2016,33(10):2902 – 2905,2909.

[5]陈海红. 多核 SVM 文本分类研究[J]. 软件,2015,36(5):7 – 10.

[6] KIM Y. Convolutional neural networks for sentence classification[C]. Doha: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1746 – 1751.

[7]胡荣磊,芮璐,齐筱,等. 基于循环神经网络和注意力模型的文本情感分析[J]. 计算机应用研究,2019,36(11):3282 – 3285.

[8] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735 – 1780.

(责任编辑:王彦江)

Study on the Text Sentiment Analysis Based on Improved CNN

HE Ye, YANG Huicheng, PAN Yue, XU Shuqi

(School of Electrical Engineering, Anhui Polytechnic University, Wuhu, Anhui 241004, China)

Abstract: Traditional machine learning techniques, including support vector machines (SVM), have been applied to various tasks of text sentiment analysis, which makes the generalization ability of complex classification problems poor. In recent years, machine learning has made breakthroughs in natural language processing research. Convolutional neural network (CNN) and recurrent neural network (RNN) are two mainstream methods of text analysis. Through the research on the neural network model, a method for sentiment analysis is proposed using a combined kernel of multiple branches of a convolutional neural network (CNN) and a long and short-term memory neural network (LSTM) layer. And it is verified through experiments that the performance is better than the existing CNN models and LSTM models.

Key words: sentiment analysis; LSTM-CNN; machine learning; natural language processing