第 44 卷 第 11 期 计 算 机 工 程 2018 年 11 月
Vol. 44 No. 11 Computer Engineering November 2018

·人工智能及识别技术· 文章编号:1000-3428(2018)11-0209-06 文献标志码:A 中图分类号:TP391.1

结合词性特征与卷积神经网络的文本情感分析

何鸿业,郑 瑾,张祖平

(中南大学 信息科学与工程学院,长沙 410083)

摘 要: 在卷积神经网络模型中,如果输入文本表示不准确,网络训练容易因输入噪音导致过拟合。为改善文本卷积神经网络中输入文本表示的质量,构建一种结合词性特征的文本卷积神经网络模型。利用词性特征捕捉传统词向量无法识别的文本一词多义现象,并与输入文本原始表示方法相结合构造卷积神经网络的双通道输入。基于中文酒店评论和英文影评数据集的实验结果表明,相比于传统文本卷积神经网络,该模型在情感分类准确率、召回率和 F1 值等指标上均有明显提升。

关键词: 自然语言处理; 情感分析; 深度学习; 卷积神经网络; 文本表示

中文引用格式: 何鸿业,郑 瑾,张祖平. 结合词性特征与卷积神经网络的文本情感分析[J]. 计算机工程,2018,44(11):209-214,221.

英文引用格式: HE Hongye,ZHENG Jin,ZHANG Zuping. Text sentiment analysis combined with part of speech features and convolutional neural network[J]. Computer Engineering,2018,44(11):209-214,221.

Text Sentiment Analysis Combined with Part of Speech Features and Convolutional Neural Network

HE Hongye,ZHENG Jin,ZHANG Zuping

(School of Information Science and Engineering,Central South University,Changsha 410083,China)

【Abstract】 In the Convolutional Neural Network (CNN) model, if the input text representation is not accurate, the network training is easy to lead to over-fitted due to the input noises inaccurate text. In order to improve the quality of text representation, Part of Speech (POS) features are utilized in this paper to capture polysemy phenomena of words which typical word embedding models are not sensitive to. Then, a dual-channel CNN model named Word-POS CNN (WP-CNN) is proposed in which the original text representation is enhanced by appending the POS features. According to the experimental results on Chinese hotel reviews and English movie reviews corpus, the proposed model can obviously get better precision, recall rate as well as F1-score in comparison with traditional text CNN models.

【Key words】 Natural Language Processing (NLP); sentiment analysis; deep learning; Convolutional Neural Network (CNN); text representation

DOI: 10.19678/j.issn.1000-3428.0048955

0 概述

在互联网信息中,文本信息占很大的比重,如何对大量文本进行规整分析一直是自然语言处理的研究热点,而文本情感分析是其中一个重要任务。随着社交网络的发展,网站评论区、微博等平台成为互联网用户信息的重要载体,这类平台上的文本通常具有篇幅短、表达不规范等特点,这也给文本情感分析带来了巨大的挑战。如何从这类文本中捕获到用户的情感倾向信息,对于舆情监控有着重要的研究意义。

传统的情感分析任务通常使用基于词典的方法或基于机器学习方法来完成[1]。前者主要依赖于词典数据集,通过点互信息量(Pointwise Mutual Information,PMI)[2]等方法来判断新词的情感倾向,进而对文本整体进行情感分析。后者通常使用词袋(Bag of Words,BOW)模型等将文本表示成定长向量,并使用监督学习的方法对文本情感进行分类,基于机器学习的情感分析一直是研究的主流。近年来,随着词向量工具 Word2Vec 的公布,相关研究增

基金项目: 国家自然科学基金(61379109)。

作者简介: 何鸿业(1993—),男,硕士研究生,主研方向为自然语言处理、深度学习;郑 瑾,副教授;张祖平,教授。

收稿日期: 2017-10-16 **修回日期:** 2017-12-10 **E-mail:** hongyehe@csu.edu.cn

多,文本可以有效地以低维且连续的形式进行表示[3-4],这也成为在自然语言处理(Natural Language Processing,NLP)领域引入深度学习技术的基础,其中卷积神经网络(Convolutional Neural Network,CNN)在情感分析和文本分类领域有着很好的应用[5-6]。在目前文本卷积神经网络中,单词通常使用词向量来表示,文本则根据其中单词对应的词向量来构建矩阵表达,最终模型的分类能力与使用的词向量的质量有直接联系,如果作为输入的原始词向量表示存在语义上的噪音,网络拟合后的分类效果会变差。

为提高输入文本表示的质量,本文构建结合词性特征的文本卷积神经网络模型 WP-CNN。首先在词向量表示基础上引入词性(Part of Speech,POS)特征组合的方法,对原始文本进行词性标注与拼接,通过词性特征进行词义消歧以改进词向量的训练;然后引入原始文本词向量表示和词性拼接表示双通道输入策略,丰富网络的输入特征,避免网络训练由于输入噪音造成的过拟合现象。

1 相关工作

情感分析的目标在于挖掘文本中观点的倾向与态度,文献[7]提出使用机器学习的方法来处理篇章级别的情感分析任务,将文本的倾向判断视为一种文本情感分类情形,并使用支持向量机(Support Vector Machine,SVM)、逻辑斯蒂回归(Logistic Regression,LR)等机器学习分类算法来进行情感分析。早期一般通过计算文本中词的 TF-IDF 权重,并使用词袋模型来表示文本,这是一种高维稀疏的文本表示方式,当文本集过大时,会存在维度灾难的问题,并且无法捕获深层语义。文献[8]通过使用隐含语义分析(Latent Semantic Indexing,LSI)对原始文本表示进行特征降维以应对维度灾难。针对深层语义捕获的问题,文献[9]使用隐含 Dirichlet 分布(Latent Dirichlet Allocation,LDA)对文本进行主题建模以更好地判断文本情感极性。文献[10]则利用二元文法(Bigram)来捕获更多的上下文信息来改善情感分类的效果。

词向量研究的深入为情感分析以及其他 NLP 问题提供了新的思路。文献[11]提出的神经网络语言模型为词向量的研究奠定了基础,而之后由文献[3-4]提出的 CBOW 和 SkipGram 模型,极大地提高了词向量模型训练的效率,使单词能够被高效地映射到低维连续的向量空间上,从根本上解决文本表示的维度灾难等问题。文献[12]基于词向量提出了句向量模型 DM 与 DBOW,在情感分类上获得了很好的效果。另外,词向量低维连续的特性使其成为在 NLP 任务中进入深度学习的基础。文献[5]建立了经典的文本卷积神经网络模型,在此基础上,文本卷积神经网络的诸多改进模型被广泛应用到情感分析任务中:文献[13]对卷积神经网络应用于情感分析任务的调参进行了详细分析;文献[14]利用结合注意力模型(Attention Model)的文本卷积神经网络来完成对文本中特定目标的情感分析。

尽管文本卷积神经网络在情感分析任务上获得了很好的效果,但是文献[15]指出,词向量的训练通常以词为单位,无法捕获一词多义等现象,且训练对文本噪音十分敏感。文本卷积神经网络分类效果与文本词向量直接联系,如果作为输入的词向量语义表达存在语义上的噪音,网络训练容易造成过拟合。

2 结合词性特征的文本卷积神经网络

WP-CNN 模型的介绍分为 3 个部分:第 1 部分为结合词性拼接改进的文本表示方法;第 2 部分为双通道卷积神经网络结构;第 3 部分为模型训练的介绍。

2.1 文本改进表示

本节将详细介绍如何使用词性特征来改进词向量训练,并展示基于词性拼接的文本矩阵表达方式。

2.1.1 基于词性特征的词义消歧

词向量训练能将文本中的词映射为低维连续的向量,很好地解决了传统词袋模型对单词进行独热编码带来的数据稀疏性与维度灾难。但是,传统的词向量模型并不能捕捉上下文关系,因此,对于一词多义不敏感。一词多义在中英文语境下皆有存在,如以下例句所示,英文单词“works”和中文词语“制服”在各自 2 种语境下分别呈现出不同的词义,然而,词向量模型在训练时无法区分单词不同语境下的歧义,因此,会给后续模型的输入带来噪音。

The works(名词,意为艺术作品) of this artist are very amazing.

He works(动词,意为工作劳动) hard in the factory.

那位民警制服(动词,意为使屈服)了一个狡猾的罪犯。

在这儿上班的人必须穿上规定的制服(名词,意为统一的服饰)。

可以利用单词的词性来区分某些一词多义的情形,本文利用词性标注技术获取单词的词性并将其与单词拼接,构成“单词-词性对”(Word-POS),例如(works,verb)、(works,noun)、(制服,动词)和(制服,名词),进而将文本转化为 Word-POS 序列作为 SkipGram 等词向量训练模型的输入。拼接词性后的单词被训练为 Word-POS 向量,相对于原始词向量能达到词义消歧的作用。

2.1.2　文本矩阵表示

对于卷积神经网络,文本需要以矩阵的形式作为输入。给定一篇长度为 l 的文本 $\boldsymbol{W}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l\}$,首先通过词性标注获取每个词的词性 $\boldsymbol{P}=\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_l\}$,然后将文本拼接为 Word-POS 序列 $\boldsymbol{WP}=\{(\boldsymbol{x}_1,p_1),(\boldsymbol{x}_2,p_2),\cdots,(\boldsymbol{x}_l,\boldsymbol{p}_l)\}$。通过预训练的 Word-POS 向量模型映射后,每个 Word-POS 单位都会被转化为 n 维的 Word-POS 向量 $\boldsymbol{wp}_k$,文本被转化为 $l\times n$ 维的矩阵表示。结合词性特征的文本表示的具体流程如图 1 所示。

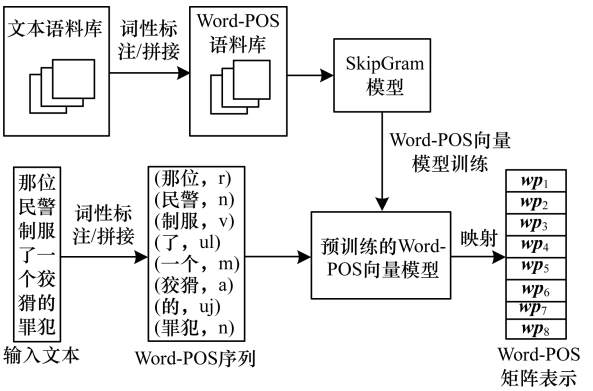

图 1　结合词性特征的文本表示方法流程

2.2　神经网络结构

本文对文献[5]提出的经典文本卷积神经网络模型进行改进,提出一种结合词性特征的文本卷积神经网络 WP-CNN,将基于原始词向量构造的文本矩阵和上节展示的 Word-POS 表达矩阵相结合,形成网络的双通道输入。整体神经网络结构如图 2 所示。

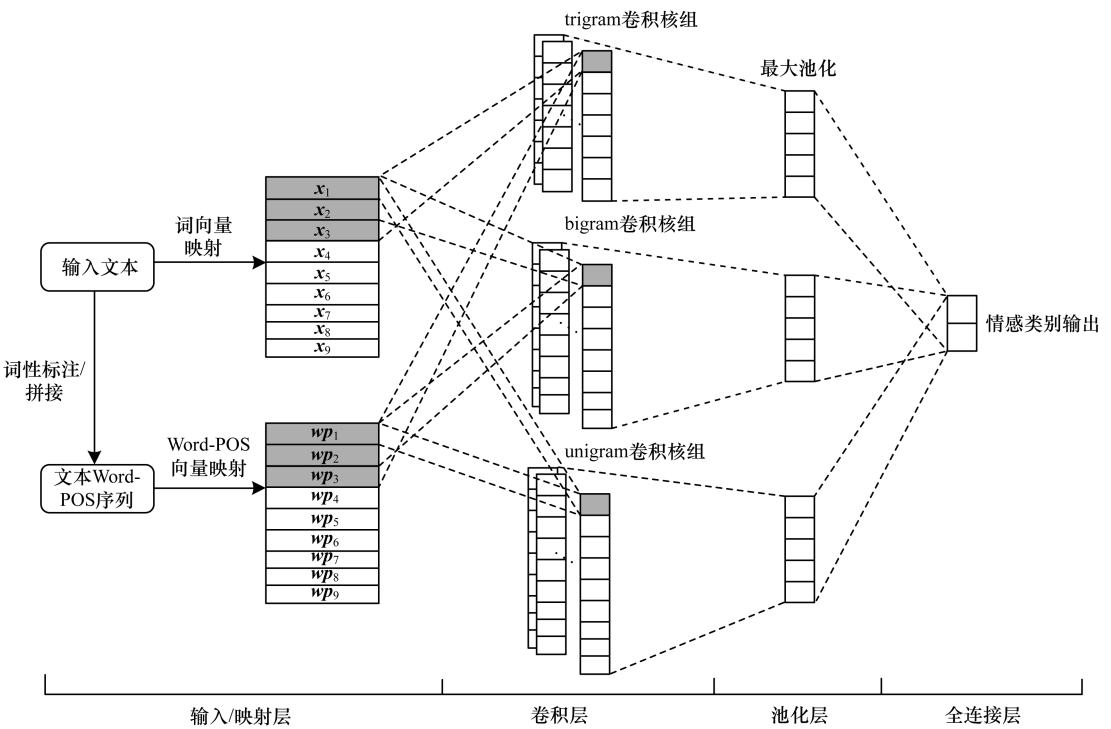

图 2　结合词性特征的双通道文本卷积神经网络

2.2.1　双通道输入

在文本集合数据不充足或者由于原始文本不规范导致词性标注不准确时,训练出 Word-POS 向量也会不精准。使用存在噪音的 Word-POS 矩阵输入会给网络训练带来过拟合现象。为了避免这种情况,本文模型采取双通道输入策略,保留原始文本词向量矩阵的输入,将文本 Word-POS 矩阵作为卷积神经网络第 2 个通道的输入,输入分别表示为 $\boldsymbol{S}_1=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l]$ 和 $\boldsymbol{S}_2=[\boldsymbol{wp}_1,\boldsymbol{wp}_2,\cdots,\boldsymbol{wp}_l]$,其中,$l$ 为文本中单词数量,$\boldsymbol{S}_1$ 为词向量矩阵,$\boldsymbol{x}_k$ 为文本中第 k 个词的词向量,$\boldsymbol{S}_2$ 为词性拼接向量矩阵,$\boldsymbol{wp}_k$ 为文本中第 k 个 Word-POS 单位的向量表示。

2.2.2　卷积操作

在卷积神经网络中,卷积核的作用在于通过窗

口滑动来提取输入数据上的局部特征,而在文本卷积神经网络中,通常选取 $h \times n$ 维大小的卷积核在文本矩阵相邻的词上进行滑动以获取卷积特征,其中,n 为词向量维度,h 代表窗口在多少个相邻词上滑动。卷积核设置为这种大小的目的是在文本的多元文法上提取上下文特征。而为了尽可能地捕获更多的上下文信息,文本卷积神经网络通常会使用多组高度设定不同的卷积核来在文本矩阵上滑动。但是,卷积核组的增加相应地会给网络训练带来更多的消耗,并且过多地引入卷积核组,网络分类效果后续的提升并不明显。考虑到网络整体的训练效率,本文选择 3 组卷积核,分别为 unigram 卷积核($h=1$)、bigram 卷积核($h=2$)与 trigram 卷积核($h=3$)。为展示卷积核在数个邻近单词上进行特征提取的操作过程,本文将邻近位置上的单词以及 Word-POS 的向量拼接分别定义为 $\boldsymbol{S}_{1,i:j}=[\boldsymbol{x}_i,\boldsymbol{x}_{i+1},\cdots,\boldsymbol{x}_j]$ 和 $\boldsymbol{S}_{2,i:j}=[\boldsymbol{wp}_i,\boldsymbol{wp}_{i+1},\cdots,\boldsymbol{wp}_j]$,当卷积核高度指定为 h 时,每次卷积滑动提取的特征值表示为:

$$c_i=f(\boldsymbol{W}_1\cdot S_{1,i:(i+h)}+\boldsymbol{W}_2\cdot S_{2,i:(i+h)}+b) \tag{1}$$

其中,$\boldsymbol{W}_1$、$\boldsymbol{W}_2$分别代表卷积核在 2 个通道上的权重,b 是偏置项,$f(\cdot)$为网络中的非线性激活函数,此处选取了 Relu 函数,如式(2)所示。

$$f(x)=\max(0,x) \tag{2}$$

当卷积核在长度为 l 的文本上滑动完毕后,根据卷积核高度 h 的大小,共可提取 $l-h+1$ 个卷积特征,原始文本将被映射到一个特征向量上,表示为:

$$\boldsymbol{c}=[c_1,c_2,\cdots,c_{l-h+1}] \tag{3}$$

2.2.3 池化操作

池化层负责对卷积层获得的特征进行二次筛选,在提取重要特征的同时提高整体网络的训练效率。如上节所述,卷积核在输入数据上滑动会提取一组局部特征,池化层的作用就是筛选出这其中最重要的特征。常见的池化方法有最大池化(max pooling)与平均池化(average pooling)。本文模型使用最大池化进行特征筛选,当确定卷积核高度 h 时,在经过卷积核操作得到的特征向量 $\boldsymbol{c}$ 上,池化层筛选出的特征值 c'可以表示为:

$$c'=\max(c_1,c_2,\cdots,c_{l-h+1}) \tag{4}$$

假设各组卷积核的数量都为 m,$c'_{i,j}$表示当卷积核组窗口高度 $h=i$ 时,池化层从此组第 j 个卷积核获取的特征向量上筛选出的特征值,最后池化层的输出可以表示为:

$$\hat{\boldsymbol{c}}=[c'_{1,1},\cdots,c'_{1,m},c'_{2,1},\cdots,c'_{2,m},c'_{3,1},\cdots,c'_{3,m}] \tag{5}$$

2.3 模型训练

池化层的输出将作为最后全连接层的输入,通过 softmax 完成情感倾向的预测。$p(y_k)$为文本在第 k 种情感倾向上的输出,代表了文本归为第 k 种情感倾向的概率,$p(y_k)$通过 softmax 归一化后表示为:

$$p(y_k)=\frac{\exp(\boldsymbol{s}_k\cdot\hat{\boldsymbol{c}}+b_k)}{\sum_{i=1}^{n}\exp(\boldsymbol{s}_i\cdot\hat{\boldsymbol{c}}+b_i)} \tag{6}$$

其中,$\boldsymbol{s}_i$与 b_i分别为全连接层的对应输出为 y_i的参数与偏置,n 为输出类别的总数。指定输入文本表示为 $\boldsymbol{X}_i$,其真实情感标签为 y_i,网络参数的集合为 θ。为方便目标函数的表示,本文将整体网络前向传播后输出类别为 y_i的概率简写为 $P(y_i|\boldsymbol{X}_i,\theta)=p(y_i)$,那么网络训练的目标函数可以表示为:

$$L=-\sum_{i=1}^{D}\ln P(y_i|\boldsymbol{X}_i,\theta) \tag{7}$$

其中,D 为训练文本集的大小。模型训练采用随机梯度下降(Stochastic Gradient Descent,SGD)算法来最小化目标函数,每轮训练迭代通过反向传播来更新网络中的各个参数,直到模型达到拟合。

3 情感分析实验

3.1 实验数据

本文使用谭松波整理的中文酒店评论情感分析语料 ChnSentiCorp-Htl-unba-10000,这是一个不平衡语料集合,含有 10 000 篇从携程网等网站上采集整理的酒店评论数据,其中正向评价文本 7 000 篇,负向评价文本 3 000 篇。此外,为验证 WP-CNN 在中英文上的泛用性,本文还在文献[16]整理的英文影评语料上进行了实验,语料含义正负向英文影评各 5 331篇,是一个平衡数据集。对于 2 组语料集的实验,实验各自抽取了 90% 的文本用作网络的训练集,10% 的文本用作测试集。

对于 2 个数据集,本文分别去掉了文本中的特殊符号,针对中文酒店评论语料,使用中文分词工具 jieba 进行了分词处理。此外,为构造用于训练 Word-POS 向量的词性拼接语料,需要对文本进行词性标注,其中,本文使用 jieba 工具中的词性标注功能来获取中文语料词性特征,同时使用自然语言处理工具 nltk 来完成英文词性标注。

3.2 词向量/Word-POS 向量训练

为避免外部数据引入对实验的影响,本文只使用原始语料来进行相关向量训练。其中:词向量的训练使用未拼接的语料库;Word-POS 向量的训练使用经过词性特征拼接后的语料库。本文词向量的训练使用了 Word2Vec 工具,此工具包含 SkipGram 和 CBOW 方法的实现。针对文本的原始表示,实验同

时使用了 SkipGram 和 CBOW 方法来训练词向量来初始化后续用于对比的基准文本卷积神经网络。而对于 Word-POS 向量,本文统一使用 SkipGram 方法进行训练。词向量/Word-POS 向量训练相关参数如下:向量的维度 100 维;上下文窗口大小为 10;向量训练迭代次数为 10 次。

3.3　神经网络参数设置

本文的网络结构使用了 tensorflow 库来进行搭建,相关参数描述如下:输入的通道数为 2,分别对应文本词向量矩阵和 Word-POS 矩阵,每轮输入的样本数量(batch-size)为 64;选取了 3 种窗口大小的卷积核,分别为 1 × 100(unigram 卷积核)、2 × 100(bigram 卷积核)和 3 × 100(trigram 卷积核),其中 100 为词/Word-POS 向量的维度,每组卷积核的数量各为 64;为防止过拟合,网络在 softmax 层使用了 dropout 机制,全连接层的节点会进行随机失活,失活率设置为 0.5。

3.4　评估标准

文本情感分类可以视为一种文本分类任务,本文使用准确率 *Precision*、召回率 *Recall* 以及 F1 值 *F1-Score*作为分类评估标准。准确率定义为正确分类为指定类别的文本数量与分类到了指定类别的全部文本数量的比值;召回率定义为正确分类为指定类别的文本数量与全部指定类别文本数量的比值;F1 值是准确率和召回率的调和均值,能对准确率和召回率进行综合评估。设 *TP* 为正确分类到正类的文档数,*TN* 为正确分类到负类的文档数,*FP* 为误分类到正类的负类文本数量,*FN* 为误分类到负类的正类文本数量,则对于正类,各评估标准可被表示为:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

本文用来评估的准确率和召回率分别为正类与负类的准确率均值和召回率均值,而 F1 值则使用平均准确率和平均召回率通过式(10)计算获得。

3.5　结果对比与分析

3.5.1　对比模型介绍

为比较各种模型在 2 组数据集上的表现以验证 WP-CNN 在情感分析任务中的有效性,本文选取了若干传统机器学习和基准文本卷积神经网络模型与本文提出的模型进行了对比。对于传统机器学习,使用词袋模型对文本进行表示,并选用 SVM 和 LR 2 种算法进行情感分类。根据词向量初始化方式的不同,选择多种用于对比基准文本卷积神经网络,分别为词向量随机初始化并在网络训练过程中通过反向传播动态调整的 Rand-CNN、词向量分别通过 CBOW 模型和 SkipGram 模型初始化的 CBOW-CNN 和 SkipGram-CNN。此外,为分析双通道输入对实验结果的影响,本文使用了 2 种 WP-CNN 模型进行对比,分别为去除了原始词向量矩阵输入,文本只以 Word-POS 矩阵作为单通道输入的 WP-CNN1;文本同时使用词向量矩阵和 Word-POS 矩阵作为双通道输入的 WP-CNN2。

3.5.2　实验结果分析

表 1 和表 2 展示了使用各个模型在数据集上拟合后获得的分类效果。从中可以看出:使用随机初始化输入的 Rand-CNN 相对于传统机器学习方法并没有显著的提高,在英文数据集上效果甚至差于 SVM,F1 值仅为 0.759 7;Rand-CNN 在学习分类网络参数的同时还要对原始词向量进行调整,在训练数据不充足的情况下,训练词向量质量较差;相比之下,使用预训练词向量初始化输入的 CBOW-CNN 和 SkipGram-CNN 模型相对于 Rand-CNN 和有明显的提升,这证实了文本卷积神经网络的分类效果与输入表达方式有直接关系,网络对于输入的噪音相当敏感。

表 1　中文酒店数据情感分析结果

模型	准确率	召回率	F1 值
LR	0.898 4	0.893 8	0.896 1
SVM	0.905 8	0.901 5	0.903 6
Rand-CNN	0.908 8	0.906 7	0.907 7
CBOW-CNN	0.915 5	0.913 7	0.914 6
SkipGram-CNN	0.921 7	0.919 6	0.920 6
WP-CNN1	0.941 2	0.939 5	0.940 3
WP-CNN2	0.946 7	0.946 4	0.946 5

表 2　英文影评数据情感分析结果

模型	准确率	召回率	F1 值
LR	0.760 7	0.753 1	0.756 9
SVM	0.766 1	0.765 9	0.766 0
Rand-CNN	0.760 0	0.759 4	0.759 7
CBOW-CNN	0.781 2	0.780 9	0.781 0
SkipGram-CNN	0.786 6	0.784 6	0.785 6
WP-CNN1	0.808 0	0.807 1	0.807 5
WP-CNN2	0.818 5	0.818 4	0.818 4

通过对比发现,结合词性特征的 WP-CNN 在各项评估指标上要明显优于其他模型。在中文酒店评

论集上,相对于效果最佳的基准模型 SkipGram-CNN,单通道的 WP-CNN1 与双通道的 WP-CNN2 在 F1 值上分别提升了 0.019 7 与 0.025 9;而在英文影评数据上 F1 值分别提升了 0.021 9 与 0.032 8。实验结果证实了通过结合词性特征来对词向量训练进行词义消歧可以改善文本卷积神经网络的输入表达,进而得到更佳的分类能力。同时,基于双通道的 WP-CNN2 在情感分析效果上要好于单通道 WP-CNN1,在 2 个数据集上,前者 F1 值分别高于后者 0.006 2和0.010 9,这也验证结合原始词向量的双通道输入能提高网络分类能力,这点在文本相对更不规范的英文影评情感数据集上表现的更明显。

3.5.3 迭代结果分析

图 3 和图 4 分别展示了在 2 组情感分析实验中各网络前 50 轮迭代的分类准确率。从中可以看出:2 种 WP-CNN 每轮迭代效果都要优于各基准文本卷积神经网络,从第 5 轮迭代开始其分类准确率就明显高于其他网络模型;2 组实验中 Rand-CNN 由于需要动态调整词向量,每轮迭代的分类准确率都是最低的。

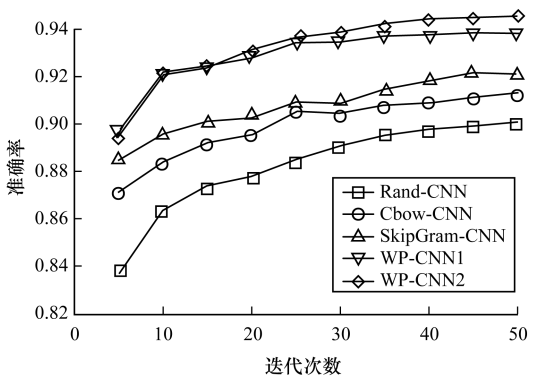

图 3 中文酒店数据集实验前 50 轮迭代结果

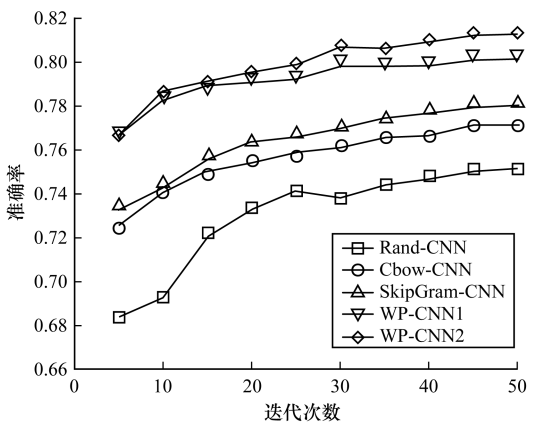

图 4 英文影评数据集实验前 50 轮迭代结果

通过观察可发现,WP-CNN1 收敛需要的迭代次数基本和 CBOW-CNN 和 SkipGram-CNN 保持一致。相对地,由于 WP-CNN2 拥有更多的输入特征,因此随着迭代次数的增加,其能体现出比 WP-CNN1 更好的分类效果,这种趋势也证实了双通道输入策略能够为网络训练带来更大的提升空间,网络不会由于单一通道的输入噪音过早地陷入过拟合。

4 结束语

本文在经典文本卷积神经网络的基础上引入了词性拼接强化策略,对网络的输入文本进行词义消歧,进而构建一种结合词性特征的文本卷积神经网络模型 WP-CNN,并将其应用到文本情感分析任务中。通过词性拼接捕获一词多义现象,WP-CNN 可以改善文本词向量的训练,为文本卷积神经网络提供质量更优的输入,并且其使用结合原始词向量和 Word-POS 向量的双通道输入,能够有效地解决因输入噪音过多造成的网络训练过拟合问题。在基于中文酒店评论与英文影评数据的情感分析实验中,WP-CNN 模型的分类准确率、召回率和 F1 值明显高于传统机器学习方法和基准文本卷积神经网络模型,其有效性得到了验证。在后续工作中,将把 WP-CNN 应用到其他 NLP 任务中评估其效果,并做进一步优化。

参考文献

[1] GIACHANOU A,CRESTANI F. Like it or not:a survey of Twitter sentiment analysis methods[J]. ACM Computing Surveys,2016,49(2):1-41.

[2] TURNEY P D. Thumbs up or thumbs down?:semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of Meeting on Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2002: 417-424.

[3] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of International Conference on Neural Information Processing Systems. New York,USA:ACM Press,2013:3111-3119.

[4] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[C]//Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics. Berlin, Germany:Springer,2013:430-443.

[5] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. (2014-08-25)[2017-05-24]. https://arxiv.org/abs/1408.5882.

[6] KALCHBRENNER N,GREFENSTETTE E,BLUNSOM P. A convolutional neural network for modelling sentences [EB/OL]. (2014-04-08) [2017-05-24]. https://arxiv.org/abs/1404.2188.

(下转第 221 页)

classification[J]. Data Mining and Knowledge Discovery, 2011,23(3):447-478.
[7] YAN S,XU D,ZHANG B,et al. Graph embedding and extensions: a general framework for dimensionality reduction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2007,29(1):40-48.
[8] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering[J]. Advances in Neural Information Processing Systems, 2002,14(6):585-591.
[9] GALLAGHER B, ELIASSIRAD T. Leveraging label-independent features for classification in sparsely labeled networks: an empirical study [C]//Proceedings of International Conference on Advances in Social Network Mining and Analysis. Berlin,Germany:Springer,2008:1-19.
[10] PEROZZI B, ALRFOU R, SKIENA S. DeepWalk: online learning of social representations[EB/OL]. [2017-09-10]. http://www. perozzi. net/publications/14_kdd_deepwalk-slides. pdf.
[11] TANG J, QU M, WANG M, et al. LINE: large-scale information network embedding [C]//Proceedings of International World Wide Web Conferences Steering Committee. New York, USA: ACM Press, 2015: 1067-1077.
[12] GROVER A, LESKOVEC J. node2vec: scalable feature learning for networks[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York,USA:ACM Press,2016:855-864.
[13] WANG D, CUI P, ZHU W. Structural deep network embedding[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York,USA:ACM Press,2016:1225-1234.
[14] 蔡波斯,陈　翔. 基于行为相似度的微博社区发现研究[J]. 计算机工程,2013,39(8):55-59.
[15] CAO S S, LU W, XU Q K. GreRep: learning graph representations with global structural information[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management. New York, USA:ACM Press,2015:891-906.
[16] 何　静,潘善亮,韩　露. 基于双边兴趣的社交网好友推荐方法研究[J]. 计算机工程与应用,2015,51(6):108-113.
[17] 周芝民,龙　华,杜庆志,等. 基于连通性和随机游走的好友推荐算法[J]. 信息技术,2016(8):67-70.
[18] 张中军,张文娟,于来行,等. 基于网络距离和内容相似度的微博社交网络社区划分方法[J]. 山东大学学报(理学版),2017,52(7):97-103.
[19] YANG C,ZHAO D,ZHAO D,et al. Network representation learning with rich text information [C]//Proceedings of International Conference on Artificial Intelligence. [S.l.]:AAAI Press,2015:2111-2117.

编辑　吴云芳

(上接第 214 页)

[7] BO P,LEE L,VAITHYANATHAN S. Thumbs up?:sentiment classification using machine learning techniques [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics,2002:79-86.
[8] BADER B W, KEGELMEYER W P, CHEW P A. Multilingual sentiment analysis using latent semantic indexing and machine learning [C]//Proceedings of IEEE International Conference on Data Mining Workshops. Washington D. C., USA: IEEE Computer Society,2011:45-52.
[9] LIN C,HE Y. Joint sentiment/topic model for sentiment analysis [C]//Proceedings of ACM Conference on Information and Knowledge Management. New York, USA:ACM Press,2009:375-384.
[10] WANG S, MANNING C D. Baselines and bigrams: simple, good sentiment and topic classification [C]//Proceedings of Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics,2012:90-94.
[11] BENGIO Y,SCHWENK H,SENÉCAL J S,et al. Neural probabilistic language models [M]//HOLMES D E, JAIN L C. Innovations in Machine Learning. Berlin, Germany:Springer,2006:1137-1155.
[12] LE Q V, MIKOLOV T. Distributed representations of sentences and documents [EB/OL]. (2014-05-22) [2017-05-24]. https://arxiv. org/abs/1405. 4053.
[13] 王盛玉,曾碧卿,胡翩翩. 基于卷积神经网络参数优化的中文情感分析[J]. 计算机工程,2017,43(8):200-207,214.
[14] 梁　斌,刘　全,徐　进,等. 基于多注意力卷积神经网络的特定目标情感分析[J]. 计算机研究与发展,2017,54(8):1724-1735.
[15] TRASK A, MICHALAK P, LIU J. Sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings[EB/OL]. (2015-12-19) [2017-05-24]. https://arxiv. org/abs/1511. 06388.
[16] BO P,LEE L. Seeing stars:exploiting class relationships for sentiment categorization with respect to rating scales[C]//Proceedings of the Meeting of Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2005: 115-124.

编辑　金胡考