·人工智能及识别技术· 文章编号：1000-3428(2017)08-0200-08 文献标志码：A 中图分类号：TP391.1

基于卷积神经网络参数优化的中文情感分析

王盛玉[1]，曾碧卿[1,2]，胡翩翩[1]

（1. 华南师范大学 计算机学院，广州 510631；2. 华南师范大学 软件学院，广东 佛山 528225）

摘 要： 卷积神经网络模型的训练需要设计者指定大量模型参数，但因模型对各类参数的敏感度不一，导致实验效果不佳。针对上述问题，研究中文文本情感分析，以词向量维度、词向量训练规模、滑动窗口大小和正则化方法等作为不同模型的影响因素，设计单层卷积神经网络，在不同影响因素下分别进行中文情感分类实验，并根据结果得出卷积神经网络在处理中文情感分析时对各类参数的敏感程度和具体的模型参数优化建议。

关键词： 卷积神经网络；情感分析；参数优化；词向量；深度学习

中文引用格式： 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.

英文引用格式： Wang Shengyu，Zeng Biqing，Hu Pianpian. Chinese Sentiment Analysis Based on Parameter Optimization of Convolutional Neural Network[J]. Computer Engineering，2017，43(8)：200-207，214.

Chinese Sentiment Analysis Based on Parameter Optimization of Convolutional Neural Network

WANG Shengyu[1]，ZENG Biqing[1,2]，HU Pianpian[1]

（1. School of Computer，South China Normal University，Guangzhou 510631，China；
2. School of Software，South China Normal University，Foshan，Guangdong 528225，China）

【Abstract】 The training of Convolutional Neural Network (CNN) model requires designers to set a large number of model parameters. Because the sensitivity of the model to various parameters is different，the experimental results are poor. To address the problem，this paper provides an analysis on Chinese text sentiment analysis and designs a one layer CNN with influence factors of different models，including the dimensionality of word vectors，the training scale of word vectors，slide window size，regularization method，and so on. Chinese sentiment classification experiment is conducted under different influence factors. According to the results，the sensitivity degree of the CNN on various parameters when dealing with Chinese sentiment analysis and the optimization of specific model parameters are proposed.

【Key words】 Convolutional Neural Network (CNN)；sentiment analysis；parameter optimization；word vector；deep learning

DOI： 10.3969/j.issn.1000-3428.2017.08.034

0 概述

近年来，随着互联网的发展，社交网络和电子商务的火热，在互联网上产生大量的个人观点以及商品评论，而基于这些海量评论数据的个人情感分析也成为自然语言处理的关注领域。同时，伴随着深度学习在图像和语音识别领域取得重大突破，国内外学者也在自然语言处理领域对其进行深入研究[1]，在使用神经网络模型建立语言模型、文本特征表示以及文本分类等方面取得进展。

神经网络模型中卷积神经网络(Convolutional Neural Network，CNN)已经在情感分析中取得相关成果。文献[2-3]在这方面做了相关的研究。卷积神经网络首先使用词向量的表示方式，将句子转化成为词向量表示的矩阵，然后将词向量矩阵作为卷积神经网络的输入。通过构建单层的卷积神经网络，在不同的数据集上获得了较支持向量机(Support Vector Machine，SVM)更好的实验结果。与此同时，还有更多的神经网络模型取得了突破，文献[4]在递归自动编码的基础上，结合情感类别的监督信息来训练模型。

但是，基于卷积神经网络的模型方法都面临着

基金项目： 国家自然科学基金(61503143)。

作者简介： 王盛玉(1992—)，男，硕士研究生，主研方向为自然语言处理、情感分析；曾碧卿，教授；胡翩翩，硕士研究生。

收稿日期： 2016-08-04 **修回日期：** 2016-09-05 **E-mail：** wangshengyu_1992@163.com

模型参数优化选择的问题,无论是单层的或多层的卷积神经网络模型都需要实验者根据模型设计可适应的超参,而相较于传统的机器学习方法,如 SVM、逻辑回归,神经网络模型拥有众多的参数,又因为卷积神经网络模型的训练需要使用 GPU 加速,传统的基于 CPU 的训练方式速度较慢,所以基于卷积神经网络模型方法的参数优化选择和调整代价较大。

为此,本文结合卷积神经网络、中文处理方面的特点以及模型训练时的困难,以单层卷积神经网络模型为基础,研究模型训练时训练数据集中评论语句长度、输入词向量维度、训练词向量规模、滑动过滤窗口大小、每个滑动过滤窗口的数量、正则化方法选择以及是否将训练数据集添加到词向量训练等因素对模型的影响,并进行相应实验。

1　相关工作

情感分析主要是面向文本的处理,通过对文本的分析,发掘评论人对于事物或事件的观点和态度。目前,情感分析的研究方法主要集中在基于规则和基于统计的模型。例如通过优选的情感词典,结合机器学习中监督学习、半监督或无监督学习的方法对情感进行分析。机器学习中常用的方法有支持向量机、朴素贝叶斯(Native Bayes,NB)、条件随机场(Condition Random Field,CRF)、最大熵(Maxinum Entropy,ME)等。文献[5]利用机器学习的方法对电影评论进行褒贬分类,其选取了多个训练模型,并使用一元词、二元词、词性标注等若干特征。文献[6]将支持向量机和朴素贝叶斯方法相结合,在多个数据集上取得不错的效果。文献[7]则使用半监督学习方法,同时建议将训练数据分为主观数据与非主观数据(客观)。

在语言模型选择上,主要有词袋模型和 N-Gram 模型。词袋模型虽然在模型假设上存在缺陷,但这并不影响其在自然语言处理领域所取得的成果,已广泛应用于各类机器学习方法中,与贝叶斯模型、文档主题生成模型隐含 Dirichlet 分配(Latent Dirichlet Allocation,LDA)、隐层语义分析(Latent Semantic Analysis,LSA)、SVM 相结合,取得较好的效果,但是传统的词袋模型,无论是向量空间模型(Vector Space Model,VSM)或是 one-hot 模型,都存在着无法获取文本语法语义的缺陷,无法获取文本深层次语义特征。随着文献[8-9]提出深度学习方法,文献[10]使用深度学习的方法建立语言模型,训练生成词向量表示,文献[11-12]通过简化模型,提出 CBOW 和 Skip-gram 模型,使模型在时间上达到实际运用要求,文献[13]则将模型训练的词向量用于分词、词性标注、命名实体识别等任务,并推出了 SENNA 系统。

在解决情感分析问题时,以深度学习的方法建立模型,使用提前训练好的词向量,将文本转化成模型可识别的输入,最终通过模型训练优化参数,以得到最优结果。文献[14]结合 LDA 模型和神经网络模型的特点,混合监督与半监督学习方法获取文本语法和情感信息;文献[15]提出了一种嵌入词情感信息的语言模型,在学习词向量的同时将词情感信息嵌入词向量表示中,取得了更好的情感分类效果。文献[2]设计了单层的卷积神经网络,在不同的数据集上进行情感分类任务的对比实验,而文献[16]则使用 one-hot 模型,结合卷积神经网络进行文本情感分类。文献[3]提出动态 k-max 池化方法,将其应用于自然语言变长的处理,并取得不错的效果。文献[17]将传统情感分析中考虑的因素,如词语的情感极性、人工构建的情感词典融入到 CNN 中,利用情感词典中的词条对文中的词语进行抽象表示,并利用 CNN 抽象词语特征,将其用于情感分析任务。

在目前的相关工作中,利用深度学习卷积神经网络结合词向量表示的方法进行情感分析任务,需要模型设计者设置大量模型参数,而模型参数的选择往往需要大量时间尝试,但又因模型对各类参数的敏感度不一,导致设置结果不佳,故本文通过构建单层的卷积神经网络,充分考虑情感分析任务中对模型效果影响的因素。

2　基于卷积神经网络模型的参数优化选择

2.1　卷积神经网络语言模型架构

利用深度学习卷积神经网络处理文本时,需要将文本转化成卷积神经网络可以识别的输入。首先需要将句子分词后,将词语映射到 d 维实数向量,词向量表示可以使用高斯分布进行初始化,也可以使用 word2vec 或 GloVe 提前训练。令 $\boldsymbol{x}_i \in \mathbb{R}^d$ 代表句子中第 i 个词 d 维的词向量表示,所有的词组成句子矩阵 $\boldsymbol{M}_j \in \mathbb{R}^{l \times d}$,其中,$j$ 代表所有评论集中第 j 条评论句子;l 代表句子中词的数量;d 代表词向量表示的向量维度,矩阵每行代表句中词的词向量表示,将矩阵 $\boldsymbol{M}_j$ 作为卷积神经网络的输入,如图 1 所示。

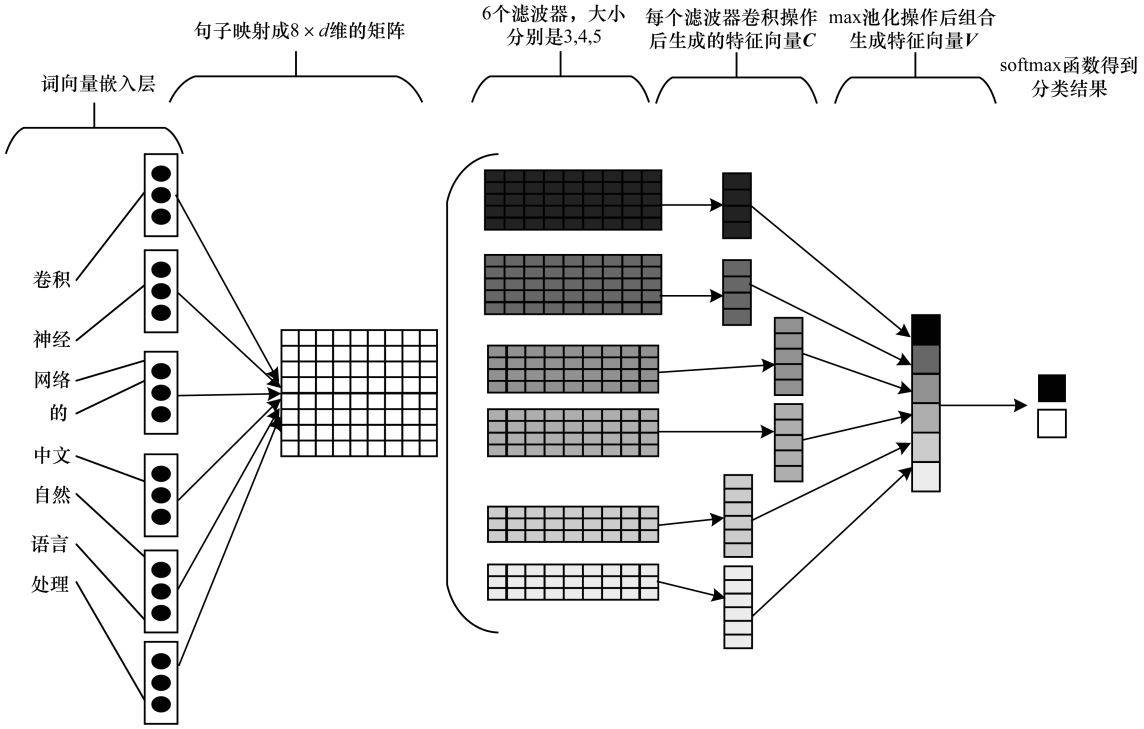

图 1 卷积神经网络自然语言处理架构

对于输入句子矩阵 $\boldsymbol{M}_j$，利用大小为 $h \times d$ 的滤波器（滑动窗口）$w^{h\times d}$ 进行卷积操作，卷积滑块涉及 h 个词，滑块宽度与词向量表示维度相同：

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

其中，b 代表偏置量；$f(\cdot)$ 为非线性卷积核函数；x_{i+h-1} 表示矩阵第 i 行到第 $i+h-1$ 行；c_i 表示由卷积操作所产生的局部特征。故在句子矩阵 $\boldsymbol{M}_j$ 上，卷积滑动窗口将作用于 $\{x_{1:h}, x_{2:h+1}, \cdots, x_{l-h+1:l}\}$ 个局部特征区域，所以：

$$\boldsymbol{C} = [c_1, c_2, \cdots, c_{l-h+1}] \tag{2}$$

其中，$\boldsymbol{C} \in \mathbb{R}^{l-h+1}$。

对于滤波器产生的输出 $C \in \mathbb{R}^{l-h+1}$，采用 Collobert 提出的 max-over-time pooling 方法进行特征映射，得到 $\hat{c}$：

$$\hat{c} = \max(\boldsymbol{C}) \tag{3}$$

故每个特征对应一个滤波器，而使用 max 方法是为了获得滑动窗口获取的局部特征中最重要的特征，因为每个句子的长度不同，且使用的滤波器的大小不同，故卷积操作产生的特征数量不同，但通过 pooling layer 的 max-over-time 方法，可以保证最终得到的特征数量相同。

对于整个卷积神经网络模型，将使用多个滤波器 $w_j^{h\times d}$（h 为不同的值）对输入矩阵 $\boldsymbol{M}_j^{l\times d}$ 进行卷积操作，产生多个特征，将特征组合作为全连接层的输入向量 $\boldsymbol{V}$：

$$\boldsymbol{V} = [\hat{c}_{1,1} \cdots \hat{c}_{1,h} \cdots \hat{c}_{m,h}] \tag{4}$$

其中，$\hat{c}_{m,h}$ 代表大小为 h 的第 m 个滤波器产生的特征。最后将全连接的输出利用 *softmax* 函数生成分类结果，模型利用实际分类标签，使用反向传播算法对参数进行优化。

$$P(y|\boldsymbol{V}, \boldsymbol{W}, b) = softmax(\boldsymbol{W} \cdot \boldsymbol{V} + b) \tag{5}$$

其中，$y \in \{+1, -1\}$，代表情感的消极或积极；$\boldsymbol{W} \in \mathbb{R}^{|V|}$ 代表全连接层的参数；b 为偏置项。

2.2 卷积神经网络模型参数优化

本文以单层卷积神经网络为基础，文献[18]实验参数作为对比参考，结合中文处理特点，选取多组模型参数和可能的数据因素为模型影响因子，通过实验对比影响因子对模型正确率的影响。选取的模型参数和数据因素如表 1 所示。

表 1 模型影响因素

参数	属性
评论数据集语句长度差	紧密度，散密度
词向量维度	50，100，200
词向量训练规模/GB	1.6，3.3
滑动窗口大小	3，5，7，10，20
滑动窗口数量	10，60，120，240
dropout rate	0.1～0.9
L2 正则	1，2，3，5，10，20
评论数据是否作为词向量训练数据	是，否

3 实验与结果分析

3.1 基准实验设置

研究中以单层卷积神经网络[2]为基准实验，不设置传统机器学习方法的对比，如 SVM、逻辑回归。基准实验卷积神经网络模型参数如表 2 所示。

表 2　基准实验参数设置

参数	属性
输入词向量	Word2vec，50 维
词向量训练集	搜狗新闻数据集
滑动窗口大小	3，4，5
滑动窗口数量	120
激励函数	ReLU
Pooling 方法	Max
Dropout rate	0.5
L2	3
迭代次数	50

3.2　实验数据准备与预处理

实验数据分为多个部分，其中词向量表示训练数据分为两大部分，第一部分来自搜狗实验室的全网新闻数据（SogouCA），第二部分来自维基百科中文数据。SogouCA 数据为 2012 年 6 月—7 月期间国内、国际、体育、社会、娱乐等 18 个频道的新闻数据，经过编码、分词等处理后，数据大小为 2.3 GB 左右，维基百科中文数据经过编码、分词等处理后，数据大小为 1.1 GB 左右，两者数据总量在 3.3 GB 左右，共包含超过十亿级中文词汇和少量英文词汇。其语料库规模足够训练出高质量的词向量表示，使每个词语义得以充分表达，模型训练采用 skip-gram 模型。

针对实验模型的训练数据，为了防止单一数据来源导致的偶然性，本次实验选取 2 组实验训练数据。其中一部分来自某酒店的用户评论数据集。该数据集包括 6 000 条评论样本，其中正负样本各3 000 条，分别是用户对酒店的积极评价和消极评价。另一份数据为某电商的家电评论数据，该数据集包括 4 000 条评论样本，其中正负样本各 2 000 条，分别是用户对家电的积极评论和消极评论。每组实验结果以 10 折交叉验证方式对模型准确率进行评估。

词向量表示训练数据在使用 word2vec 工具训练之前，需要将数据编码格式统一，之后统一分词，word2vec 工具使用 gensim，分词工具使用开源中文分词工具 jieba。卷积神经网络模型的构建使用 google 公司开源工具 tensorflow。

3.3　实验结果对比

3.3.1　评论数据集语句长度差的影响

根据前文卷积神经网络语言模型架构的介绍，CNN 在进行自然语言处理时，需要将句子转换成 $m \times d$ 的二维数据矩阵。而为了处理数据的一致性与效率，模型会将二维数据矩阵的大小设置为固定大小，即 m 为最长评论语言包含的词数量，d 为词向量维度，当句子长度不足时，采用补零处理。

通过对评论数据集的分析，从图 2 中可以看到频繁出现的评论语句长度占据了整个评论数据集的 2/3，而过短或过长评论语句出现次数较少。通过对比模型训练数据语句长度密度发现，当数据集中语句长度较为分散时，模型的准确率有所下降，故余下实验均采用占电商评论数据集 80% 以及占酒店评论数据集 2/3 的紧密度数据。实验结果通过 10 折交叉验证得到，如表 3 所示。

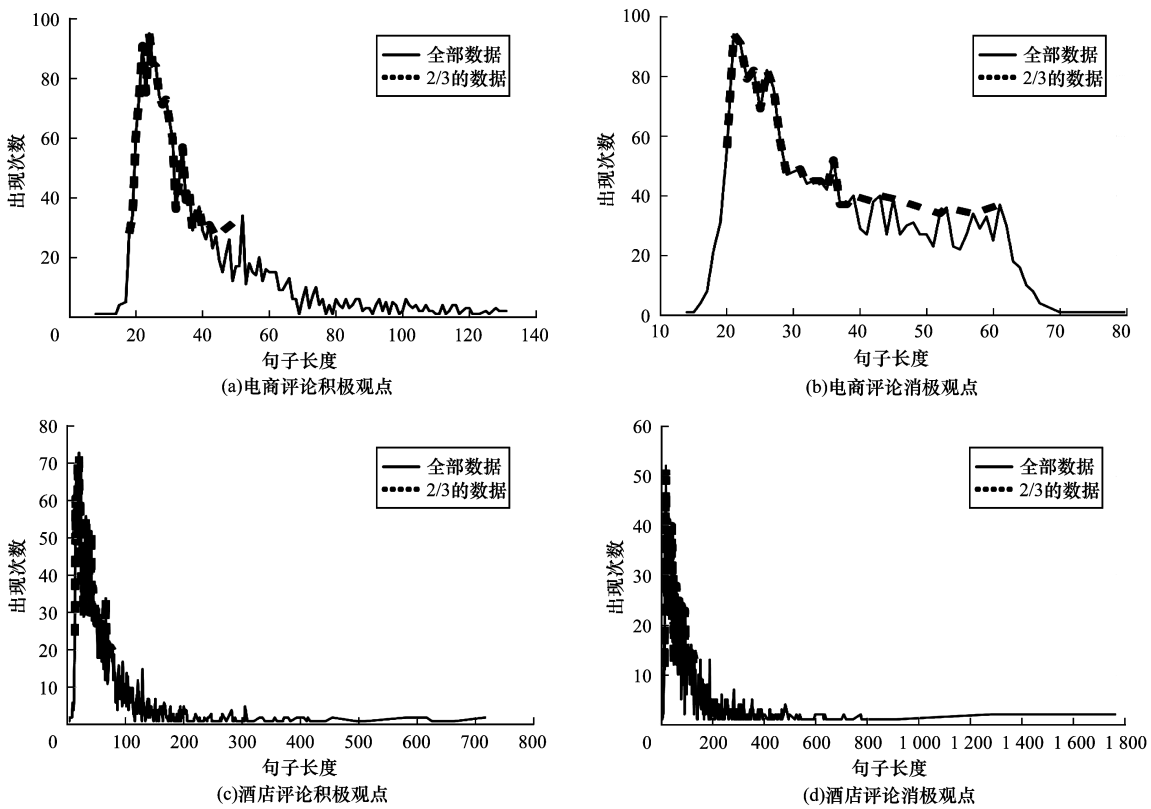

图 2　数据集评论语句长度分析

表 3 数据准确率对比 %

数据集类别	准确率
紧密度电商数据集	90.582 7
散密度电商数据集	89.583 3
紧密度酒店数据集	90.693 3
散密度酒店数据集	89.201 4

3.3.2 词向量表示的影响

卷积神经网络可以在文本上表现出不错的效果,其一部分得益于词向量表示的运用。作为将句子映射为矩阵的关键部分,词向量表示具有非常灵活的表示方式,可以随机初始化,也可以通过提前训练得到。实验通过结合搜狗实验室新闻数据与维基百科中文语料为训练数据集,两者之和包含十亿级规模的中文词汇,将其用于探索训练数据量大小,同时对比词向量维度的影响。

实验固定基准实验的其他参数,仅改变词向量表示的维度或词向量训练数据集的大小。首先验证词向量表示的维度对模型的影响,训练数据集采用搜狗新闻数据,词向量表示维度 $k=\{50,100,200\}$,实验结果通过 10 折交叉验证得到,如图 3 所示。

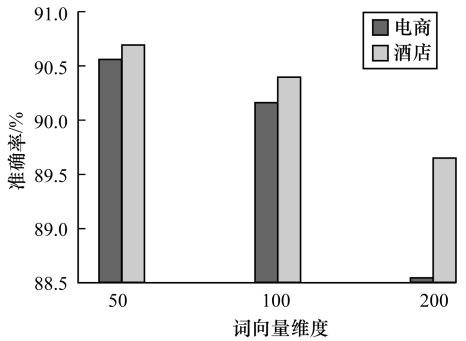

图 3 词向量表示对维度的影响

通过实验对比可以发现,针对搜狗新闻数据集,随着训练词向量维度的增大,模型的准确率随之下降。电商评论数据集随着词向量维度的增大,下降达 2%,酒店的评论数据集也有 1%。故针对不同规模的词向量训练数据集,应使用不同的词维度进行训练。此处的词向量训练数据集包含亿级中文词汇和少量英文词汇,采用 50 维较为合适。

同时,因为词向量维度对训练数据的大小相对敏感,故通过将搜狗新闻数据与维基百科中文数据相结合,一同训练词向量,验证训练数据集大小对词向量质量的影响,继而得出对卷积神经网络模型的影响。通过 10 折交叉验证得到准确率的平均值,对比实验结果如图 4 所示。

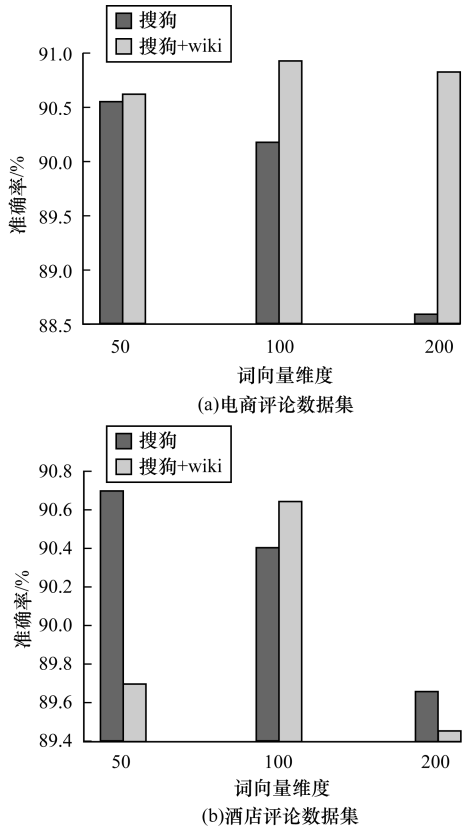

图 4 词向量训练量对比

从实验的对比图表中可以看出,利用搜狗新闻数据与维基百科中文数据共同训练的词向量,当词向量的维度较低时,模型的准确率反而有可能降低。如酒店评论数据集,其下降幅度较小,不到 1%。随着词向量训练数据集的增大,其训练的词向量表示在较高维度上较单一搜狗新闻数据要好,此处两数据集之和约包含十亿级中文词汇和少量英文词汇,词向量维度 100 维较适合。所以若是中文词向量训练数据集的词汇规模在亿级,可将词向量维度控制在 50 维左右,十亿级词汇量规模,将词向量维度增大到 100 维,而对于训练数据集规模达到百亿或千亿级,那么需要将词向量维度扩大到 300 维。针对不同训练数据集规模选择适当的词向量维度,有利于词向量更好地表达语义和句法信息,提高模型的准确率。

针对搜狗新闻数据集与中文维基百科数据集,两数据集的规模大小相似,所包含的中文词汇都在亿级,利用两者单独训练词向量,将其应用于卷积神经网络模型中,观察各自训练的词向量对模型的影响。本次实验以前 2 次实验为基础,设置词向量维度为 50,其余参数与基准实验参数相同,采用 10 折交叉验证,对比结果如图 5 所示。从对比实验的结果中可以看出,搜狗新闻数据相较于维基百科中文数据集取得更好的结果,但是两者之间的准确率之差却并不大,相差不到 1%,因为两数据集的规模和

质量都非常接近,所以其差距并非很大。故两数据集作为 50 维词向量训练数据集,都是非常好的训练资源,若需要更高维度的词向量,可将两数据集合并训练 100 维词向量。

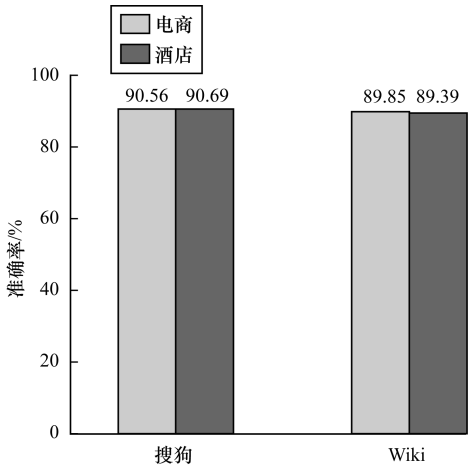

图 5　不同训练集词向量训练对比

3.3.3　滑动窗口大小的影响

对于滑动窗口大小对模型影响的对比,以基准实验为基础,固定每个滑动窗口数量为 120,修改滑动窗口大小。实验仅考虑单一滑动窗口,不考虑滑动窗口组合,令 $h=\{3,5,7,10,20\}$,实验并未设置更大的滑动窗口,因为评论语句分词后长度并不会太长,实验结果通过 10 折交叉验证得到,对比结果如图 6 所示。

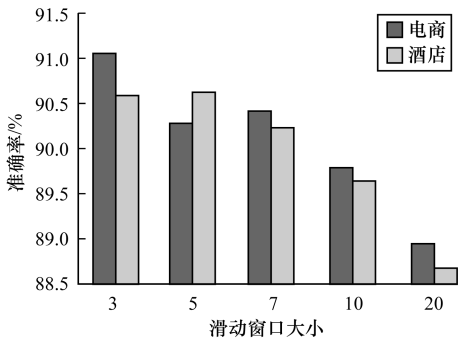

图 6　单一滑动窗口大小对比

通过实验对比可以发现,伴随着滑动窗口大小的增大,卷积神经网络模型在两评论数据集上的准确率都以下降为趋势。同时,滑动窗口大小在 3 ~ 7 之间变化时,模型准确率变化不大,相差最大时为电商评论数据集,前后准确率之差不到 1% 。而随着滑动窗口大小的进一步增加,其准确度下降较明显,又因为两评论数据集的评论语句长度存在差异,酒店的评论语言长度明显要大于电商评论数据集(见实验 1),故模型滑动窗口大小为 5 时,基于酒店评论数据集的准确率出现了上升,达到最优。所以针对不同的评论数据集,根据评论语句的长短,应适当调整滑动窗口的大小,此处酒店评论数据集的模型训练数据长度控制在 150 以下,窗口大小应控制在 3 ~ 10 之间,若评论语句长度达到 300 以上,则需将滑动窗口大小随之扩大。

同时,本文同样考虑多滑动窗口组合对卷积神经网络模型的影响,首先固定每个滑动窗口数量为 120,修改滑动窗口大小的组合。选择窗口大小组合的标准以上一组单一窗口实验结果为参考,选择上一组实验中效果最好的窗口大小,以其周边相邻大小组合窗口,故选择的窗口大小与最好效果单一窗口大小相近,实验通过 10 折交叉验证得到,对比结果如表 4、表 5 所示。

表 4　酒店评论组合对滑动窗口大小的影响　%

滑动窗口大小	准确度
5	90.618 6
3,4,5	90.693 3
4,5,6	90.107 7
5,6,7	89.885 5
5,7,9	90.020 2
10,11,12	90.069 9
12,13,14	90.119 7

表 5　电商评论组合对滑动窗口大小的影响　%

滑动窗口大小	准确度
3	91.061 4
3,4,5	90.562 7
4,5,6	90.189 4
5,6,7	90.096 1
5,7,9	89.940 5
10,11,12	88.819 4
12,13,14	88.072 0

通过 2 组对比实验可以看出,多窗口大小组合的方式并非一定比单一窗口大小取得的模型准确率高。电商评论数据集上单一窗口大小取得最好效果,而酒店评论数据集上组合的效果更好。同时,酒店评论数据集上其组合窗口的大小范围较电商评论数据集要更广,这与其评论语句的长度有关。最后,实验结果肯定了滑动窗口的组合,大小选择相邻于最好单一滑动窗口的大小所取得的模型效果最好。

3.3.4　滑动窗口数量的影响

与其他实验相同,以基准实验为基础,固定其他参数,仅修改滑动窗口数量,设置滑动窗口数量 $n=\{10,60,120,240\}$,实验并未设置比 300 更大的窗口数量,因为伴随着窗口大小的增大,实验的准确率并未有明显增加,且代价越来越高,对比结果如图 7 所示。

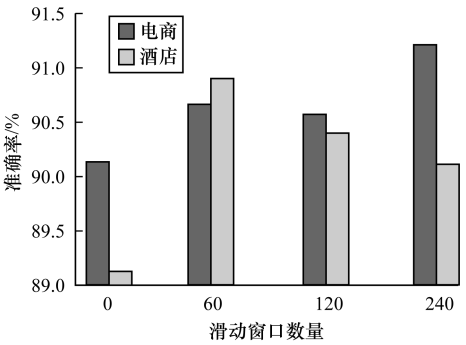

图 7 滑动窗口数量对准确率的影响对比

从对比实验结果可以看出,滑动窗口大小对模型的影响根据不同的训练数据集呈现不同的结果,与模型的准确率并非线性相关,并且其对模型产生的准确率在 1% 左右波动,并非很大。但是在实验过程中发现,随着滑动窗口大小的增大,模型训练的计算量和时间复杂度急剧上升,实验代价随之变大,故应根据情况选择合理的大小,避免产生过高的计算量。

3.3.5 是否包含模型训练集的影响

词向量的初始化有多种方式,可以通过符合高斯分布的函数随机初始化,同样也可以使用提前准备好的优质训练数据集训练得到。但是,对于模型的训练数据集,提前准备的词向量训练数据集往往不能完全覆盖,故当通过模型的嵌入层进行句子映射到矩阵时,未登录词只能根据随机初始化或其他替代方法,而将模型训练数据添加到词向量训练数据集中避免出现这样的情况。实验以基准实验为基础,固定其他模型参数,仅改变词向量数据集,通过 10 折交叉验证得到,对比结果如图 8 所示。

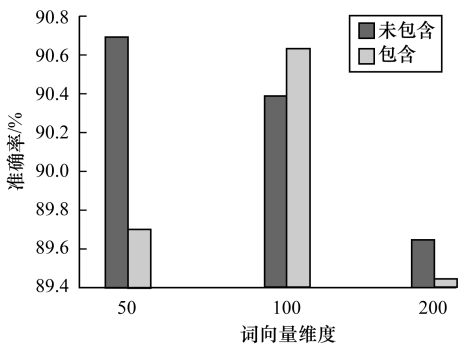

图 8 训练数据词向量维度对准确率的影响对比

通过实验对比可以看出,将模型的训练数据,即酒店评论语料库添加到词向量的训练数据中,并未对模型产生太大的影响。在搜狗数据训练的表现能力最好的 50 维词向量上,反而出现了下降,虽从图中表现的差距较大,但实际两者之差不到 0.7% ,而在更高维度,其影响可以忽略不计。这主要是模型的训练数据规模相较于词向量的训练数据显得过小,所以将其添加后对词向量训练的影响过小,但若模型的训练数据达到一定的规模,可将其作为词向量训练数据的一部分,这样有助于词向量获取语义与句法上的信息。

3.3.6 正则化方法影响

在卷积神经网络模型中,主要有 2 处使用到正则化策略,分别是倒数第 2 层的全连接层,使用 dropout 策略,以及作用于 *softmax* 函数的 L2 正则化方法。dropout 策略按比例接收上层模型训练结果,即放弃一部分上层模型训练的结果,为的是防止模式过拟化,同样防止过拟化,而 L2 正则方法为的是防止 *sotfmax* 函数的参数 W 过拟化。

设置不同的 dropout 层比例,大小从 0.1 ~0.9,同时固定 L2 的参数为 3,与基准实验相同。实验结果如图 9 所示。

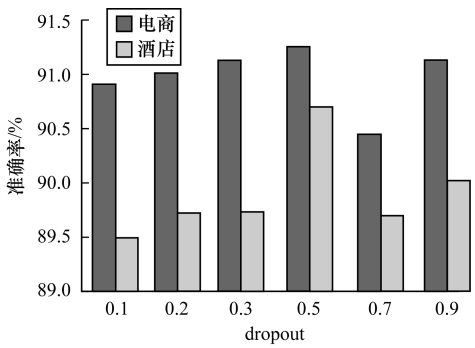

图 9 dropout 参数对准确率的影响对比

同时,对于 L2 参数对模型准确率的影响,固定 dropout 策略参数为 0.5,与基准实验相同。令 L2 参数为 $c = \{1,2,3,5,7,10,20\}$,实验结果如图 10 所示。

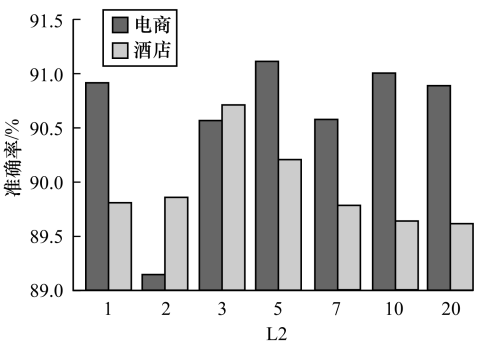

图 10 L2 参数对比

从 dropout 参数影响对比实验结果可以看出,无论是电商评论数据还是酒店评论数据,当 dropout 参数取 0.5 时,模型的准确率达到最高,这说明 0.5 是一个较为合理的参数选择点。但是同时可以看出,dropout 层的参数对模型准确率的影响并不是很大,电商评论数据集上最好与最差准确率之差只有

0.5% ,酒店评论数据集上差距略大,为 1% 左右。

从 L2 参数影响对比实验结果中可以得出,L2 参数在 3 ~ 5 之间,模型准确率波动较小,取得较好的结果。从整个实验结果也可以看出,除个别参数的选择(如电商数据集 L2 = 2 时)波动较大,其余整体变化幅度较小,大概在 1% 以内,对模型准确度的影响较小。

3.4　模型参数敏感度分析

本文通过实验验证了众多参数对卷积神经网络模型准确率的影响,通过实验可以发现,众多参数对模型准确率的影响程度大小不一。

模型的训练数据,即评论数据集的评论语言长度密集度对模型影响较大,应选择尽量紧密的训练数据集,以保证模型的准确率和计算效率。

词向量作为句子到卷积神经网络输入矩阵的映射的核心元素,根据其训练数据规模选择合适的词向量维度对模型影响较大,并且在通常情况下,词向量训练规模越大,其训练的词向量表现能力越好,模型的效果就越好。

卷积神经网络滑动窗口大小的选择及组合对模型的影响较大,而滑动窗口的数量对模型的训练代价影响较大。

正则化方法虽对模型产生影响,但其波动较小。

3.5　参数优化选择建议

通过大量实验结果表明,通常情况下卷积神经网络模型处理中文情感分析时,参数优化选择应注意如下 4 点:

1)对于模型训练的评论数据,尽可能选取评论语言长度差距不大的数据集,保持评论数据的长度在一个合理的范围内。

2)词向量的训练,若训练数据集包含词汇量在亿级规模,则词向量维度选择 50 维左右;十亿级规模,则将词向量维度扩大到 100 维左右;百亿或千亿级规模,则将词向量维度对应扩大到 300 维左右。对于模型训练的评论数据,若数据规模较大,建议将其作为词向量训练的一部分,否则对模型影响并不大。

3)根据实验结果,模型中滑动窗口大小选择 3 ~ 7之间较为合理。但若是整个训练数据集的语句长度普遍偏长,可将大小适量增大。滑动窗口的数量应根据情况,选择适当的不会产生较高实验代价的值,若有 GPU 加速,可将大小选择在 600 以上,若只有 CPU,可将大小降低到 600 以下。

4)对于模型中的 dropout 参数,选择 0.5,模型准确率较高。

4　结束语

本文在充分研究卷积神经网络模型参数的基础上,结合中文情感分析,主要做出了以下改进:

1)对卷积神经网络模型中众多参数对模型的影响做出实验验证。通过实验得出模型对参数的敏感度以及模型参数优化的建议。

2)针对卷积神经网络模型中的词向量训练,利用搜狗新闻数据集与维基百科中文数据集,以十亿数量级词汇量训练出优质的词向量表示,且针对多种维度进行了训练。其他研究者和设计开发人员可将其用于模型实验,不必重复训练,节省了大量研究时间。

3)本文的实验模型由作者实验团队开发,通过实验模型对数据的分析,有利于实验者选择较优的卷积神经网络模型参数,提高实验效果。

4)针对卷积神经网络中文情感分词,提出多个实验中常被忽略的影响因素,较为全面地考虑了卷积神经网络模型中各参数的影响,有效地优化了模型。

本文基于卷积神经网络模型对中文情感分析进行了研究,但是对中文本身所具备的语言特征和结构特征的研究没有涉及,因此下一步计划将中文语言所具有的语言特征融合到 CNN 中。

参考文献

[1] 孙志远,鲁成祥,史忠植,等. 深度学习研究与进展[J]. 计算机科学,2016,43(2):1-8.

[2] Kim Y. Convolutional Neural Networks for Sentence Classification[EB/OL]. (2014-08-25). https://arxiv.org/abs/1408.5882.

[3] Kalchbrenner N,Grefenstette E,Blunsom P. A Convolutional Neural Network for Modelling Sentences[EB/OL]. (2014-04-08). https://arxiv.org/abs/1404.2188.

[4] Socher R, Pennington J, Huang E H, et al. Semi-supervised Recursive Autoencoders for Predicting Sentiment Distributions[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics,2011:151-161.

[5] Pang Bo, 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.

[6] Wang Sida,Manning C D. Baselines and Bigrams:Simple, Good Sentiment and Topic Classification[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics,2012:90-94.

(下转第 214 页)

377-386.
[7] Zhu J. Max-margin Nonparametric Latent Feature Models for Link Prediction[EB/OL]. (2012-07-18). http://techtalks. tv/talks/max-margin-nonparametric-latent-feature-models-for-link-prediction/57429/.
[8] Mohammad H A. Link Prediction Using Supervised Learning[C]//Proceedings of SDM Workshop on Link Analysis Counterterrorism & Security. Washington D. C., USA:IEEE Press,2005:798-805.
[9] 潘 锋,王建东,顾其威,等. 基于图的特征选择算法[J]. 计算机工程,2012,38(9):197-198,201.
[10] Robnik-Sikonja M, Kononenko I. Theoretical and Empirical Analysis of ReliefF and RReliefF[J]. Machine Learning,2003,53(1/2):23-39.
[11] 冯文雅. 半导体生产面积预测不确定性的量化[D]. 成都:电子科技大学,2006.
[12] Newman M E J. Clusteringand Preferential Attachment in Growing Networks[J]. Physical Review E, 2001,64(2).
[13] Lü L,Jin C H,Zhou T. Similarity Index Based on Local Paths for Link Prediction of Complex Networks[J]. Physical Review E,2009,80(4).
[14] Lü L,Zhou T. Link Prediction in Complex Networks:A Survey[J]. Physica A Statistical Mechanics & Its Applications,2010,390(6):1150-1170.
[15] 杨 巧. 基于改进相似度的社会网络链接预测研究[D]. 广州:华南理工大学,2015.
[16] 毛 勇,周晓波,夏 铮,等. 特征选择算法研究综述[J]. 模式识别与人工智能,2007,20(2):211-218.
[17] 徐峻岭,周毓明,陈 林,等. 基于互信息的无监督特征选择[J]. 计算机研究与发展,2012,49(2):372-382.
[18] Boyadzieva D, Gluhchev G. Feature Set Selection for On-line Signatures Using Selection of Regression Variables[C]//Proceedings ofInternational Conference on Pattern Recognition and Machine Intelligence. Washington D. C., USA:IEEE Press,2011:440-445.
[19] Nagaraja V K, Abd-Almageed W. Feature Selection Using Partial Least Squares Regression and Optimal Experiment Design[C]//Proceedings of International Joint Conference on Neural Networks. Washington D. C., USA:IEEE Press, 2015:1-8.
[20] 刘 健,钱 猛,张维明. 基于 Fisher 线性判别模型的文本特征选择算法[J]. 国防科学技术大学学报, 2008,30(5):135-138.
[21] Yi Z, Ding C, Tao L. GeneSelection Algorithm by Combining ReliefF and mRMR[J]. BMC Genomics, 2008,2(1):453-458.
[22] Barabsi A L, Albert R. Emergence of Scaling in Random Networks[J]. Science,1999,286(5439):509-512.
[23] Dorogovtsev S N, Mendes J F. Evolution of Networks[J]. Advances in Physics,2002,51(4):1079-1187.

编辑 陆燕菲

(上接第 207 页)

[7] Li Shoushan, Huang Churen, Zhou Guodong, et al. Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification[C]// Proceedings of Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics,2010:414-423.
[8] Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks[J]. Science, 2006, 313(5786):504-507.
[9] Hinton G E, Osindero S, Teh Y W. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006,18(7):1527-1554.
[10] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research,2003,3(6):1137-1155.
[11] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. New York,USA:ACM Press,2013:3111-3119.
[12] 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.
[13] Collobert R,Weston J,Bottou L,et al. Natural Language Processing (Almost) from Scratch[J]. Journal of Machine Learning Research,2011,12(1):2493-2537.
[14] Maas A L,Daly R E, Pham P T,et al. Learning Word Vectors for Sentiment Analysis[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: Association for Computational Linguistics, 2011: 142-150.
[15] Tang Duyu, Wei Furu, Yang Nan, et al. Learning Sentiment-specific Word Embedding for Twitter Sentiment Classification[EB/OL]. (2015-01-12). http://anthology. aclweb. org/P/P14/P14-1146. xhtml.
[16] Johnson R, Zhang Tong. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks[EB/OL]. (2015-03-26). https://arxiv. org/pdf/1412. 1058v2. pdf.
[17] 陈 钊,徐睿峰,桂 林,等. 结合卷积神经网络和词语情感序列特征的中文情感分析[J]. 中文信息学报, 2015,29(6):172-178.
[18] Zhang Ye, Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification[EB/OL]. (2015-10-19). https://arxiv. org/pdf/1510. 03820v2. pdf.

编辑 顾逸斐