·开发研究与工程应用· 文章编号：1000-3428(2019)03-0300-09 文献标志码：A 中图分类号：TP183

基于深度卷积神经网络模型的文本情感分类

周锦峰，叶施仁，王 晖

（常州大学 信息科学与工程学院，江苏 常州 213164）

摘 要： 为高效提取不同卷积层窗口的文本局部语义特征，提出一种深度卷积神经网络（CNN）模型。通过堆叠多个卷积层，提取不同窗口的局部语义特征。基于全局最大池化层构建分类模块，对每个窗口的局部语义特征计算情感类别得分，综合类别得分完成情感分类标注。实验结果表明，与现有 CNN 模型相比，该模型具有较快的文本情感分类速度。

关键词： 情感分析；情感分类标注；深度学习；卷积神经网络；词向量

中文引用格式： 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.
英文引用格式： ZHOU Jinfeng，YE Shiren，WANG Hui. Text sentiment classification based on deep convolutional neural network model[J]. Computer Engineering，2019，45(3)：300-308.

Text Sentiment Classification Based on Deep Convolutional Neural Network Model

ZHOU Jinfeng，YE Shiren，WANG Hui

（School of Information Science and Engineering，Changzhou University，Changzhou，Jiangsu 213164，China）

【Abstract】 This paper proposes a deep Convolutional Neural Network（CNN） model to efficiently extract the local semantic features of different convolutional layer windows for text. The model avoids manually specifying multiple window sizes and retains local semantic features of different windows by stacking a number of convolutional layers. Classification modules are built based on the Global Max Pooling（GMP） layer to calculate the category score for the local semantic features of each window. The model synthesizes these category scores to complete the sentiment classification annotation. Experimental results show that the model has faster text sentiment classification speed than that of other CNN models.

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

DOI： 10.19678/j.issn.1000-3428.0050043

0 概述

情感分析主要通过人类书写的文本分析和研究人的意见、情感、评价、态度和情绪，是自然语言处理（Natural Language Processing，NLP）中最热门的研究领域之一，并在数据挖掘、Web 挖掘和文本挖掘等应用范畴得到广泛研究[1-3]。例如，分析电商平台上对已购商品的点评，群众对政府新颁布的政策法规的讨论以及消费者对新产品或服务的反馈等。每天数以亿计的用户文本信息包含了丰富的用户观点和情感极性，从中可以挖掘和分析出大量的知识和模式。

深度学习为经典数据挖掘任务提供了新的手段。卷积神经网络（Convolutional Neural Network，CNN）是一种用于处理具有网状拓扑结构数据的深度神经网络（Deep Neural Network，DNN）。CNN 通过卷积操作，组合低层特征形成更加抽象的高层特征，使模型能够针对目标问题，自动学习特征。在文本情感分类应用中，CNN 能够有效避免传统机器学习方法所面临的样本特征表达稀疏、计算复杂等问题[4]。

目前，以 CNN 为基础的文本情感分类方法多数是通过学习文本的一种窗口或多种窗口局部语义信息，然后提取文本最大语义特征进行情感划分。此类方法在文本情感分类标注领域已取得较好的效果。但是目前在文本情感分类标注领域[5-7]，甚至在 NLP 的其他分类问题中[8-10]，使用的 CNN 模型多数采用一个或多个卷积层并行的结构。CNN 模型解决情感分类标注问题时，为了充分捕捉语义的距离

基金项目： 国家自然科学基金（61272367）；江苏省科技厅项目（BY2015027-12）。
作者简介： 周锦峰（1978—），男，硕士，主研方向为机器学习、自然语言处理；叶施仁，副教授、博士；王 晖（通信作者），讲师、博士。
收稿日期： 2018-01-10 **修回日期：** 2018-02-27 **E-mail：** zhouzhou9076@163.com

依赖[11],需要提取不同上下文窗口的局部语义信息,增强情感分类能力。但是,卷积层并行的 CNN 模型使用超参数设定有限种窗口大小,而且随着窗口增加,模型计算量会大幅增加,训练效率和预测速度也随之降低。

为提高模型计算效率,本文提出一种应用于全局最大池化(Global Max Pooling,GMP)层的深度卷积神经网络(GMP-CNN)模型,进行文本情感分类标注。堆叠的卷积层能够逐层深入地提取窗口更大、抽象度更高的局部语义特征。由特殊的卷积层和 GMP 层构成的分类模块为不同窗口的局部语义特征计算情感类别得分,得到文本情感分类标注,并采用斯坦福情感树库(Stanford Sentiment Treebank,SSTb)数据集以验证 GMP-CNN 模型情感分类标注的有效性。

1　相关工作

文献[3]采用朴素贝叶斯模型、最大熵模型和支持向量机模型对文本进行情感分类。此后,以传统机器学习为核心的情感分析模型层出不穷。为提高分类正确率,传统机器学习方法使用大量文本特征。随着特征变多,训练样本在每个特征上的描述会变得稀疏,机器学习的计算复杂性成倍增加。由于文本特征需要人工构造,因此特征越多,人工成本越大。

文献[12]提出分布式表示词向量的概念,从大量未标注的语料库中无监督地学习词向量,通过向量空间上的相似度表示文本语义上的相似度。由词向量序列构成文本的原始表示形式将文本内容的处理简化为 K 维向量空间中的向量运算。分布式表示词向量的出现有效解决了 DNN 输入部分对人工的依赖,并推动 DNN 发展出新模型用于文本情感分类。

文献[13]将 CNN 应用在文本分类任务,并通过实验证明基于 CNN 的文本分类模型能够获得比传统机器学习模型更高的正确率。文本情感分类标注任务也属于文本分类任务,因此,作者使用 CNN 模型完成情感分类标注任务。文献[5]基于单词的构造(以构成单词的字母为单位),提出 CharSCNN 模型。以 CNN 为基础的 CharSCNN 模型,采用 2 个并行的卷积层分别学习单词的构造特征和句子的局部语义特征,充分体现 CNN 对文本局部特征的抽象和提取能力。该模型在短文本情感分类时展示了较好效果,有效论证 CNN 模型在进行句子情感分类标注时的可行性。文献[6]在 CharSCNN 模型基础上,并行多个卷积层,学习多种窗口的文本局部特征。对于中文语料,该模型有效地完成情感二分类标注任务。文献[7]使用 Word2Vec、GloVe 和 FastText 多种词向量形成 CNN 模型的多通道输入,同时使用 avg 池化方法代替 max 池化方法,对于英文和韩文影评语料,均取得较好的标注正确率。目前,多数用于情感分类标注任务的 CNN 模型,在基础结构上类似于文献[13]提出的 CNN 模型,具有以下特点:

1)与计算机视觉领域应用的深度 CNN 不同,一般使用多种卷积层的并行结构,或者只有一个卷积层。

2)CNN 卷积核的大小需要与词向量维度匹配,这使得卷积核至少在一个维度上比较大。

3)通常使用全连接层作为分类器,将卷积层学习到的语义特征表示映射到样本标记空间。

尽管上述 CNN 模型在处理情感分类标注时,特别是情感二分类标注任务,应用效果良好,但是此类模型存在 2 个问题:1)受并行结构的限制,多提取一种窗口类型的局部语义特征需要增加一种并行的卷积层,模型在训练和预测过程中的计算量会大幅增加;2)作为分类器的全连接层参数量过大,特别是以多种窗口的局部语义特征向量作为输入的全连接层,使模型的训练和预测计算量增大,降低了模型速度,还会造成过拟合。针对以上问题,本文提出 GMP-CNN 模型对文本进行情感分类标注。

2　GMP-CNN 模型

如图 1 所示,经典的 CNN 模型解决情感分类标注问题时,通常采用多个池化层并行的结构。将一个句子或一段文本以某种形式(例如词向量序列)输入到并行结构 CNN 模型(parallel-CNN)的多个并行卷积层。经过卷积操作,提取文本的局部抽象语义[13]。最大池化层对该局部语义表示进行降维,同时保留某一个级别的语义特征,通常保留最大语义特征。串接层将这些语义特征向量拼接成一个文本特征向量。全连接层对该特征向量进行进一步抽象,计算出情感分析结果。

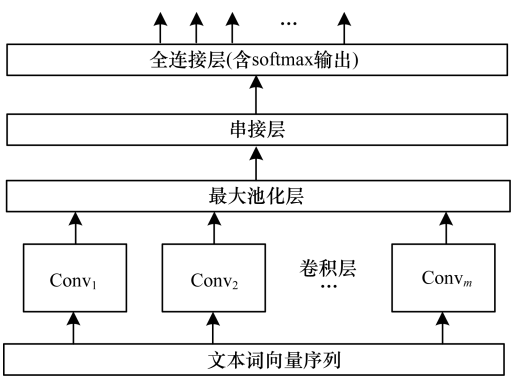

图 1　情感分类标注中并行结构的 CNN 模型

文献[14]指出多个小核卷积层堆叠产生一个大核卷积层的感知野,受此启发,本文提出 GMP-CNN 模型用于文本情感分类标注。如图 2 所示,GMP-CNN 模型通过堆叠多个卷积层,可逐层提取窗口越来越大、抽象度越来越高的文本局部语义特征用于情感多分类标注。在 GMP-CNN 模型中,将卷积层产生的局部语义特征矩阵输入下一个卷积层以及分类模块。在分类模块中,为不同窗口的局部语义特征分别计算情感类别得分。GMP-CNN 模型在点积层综合各分类模块产生的情感类别得分,得到最终的文本情感分类标注。

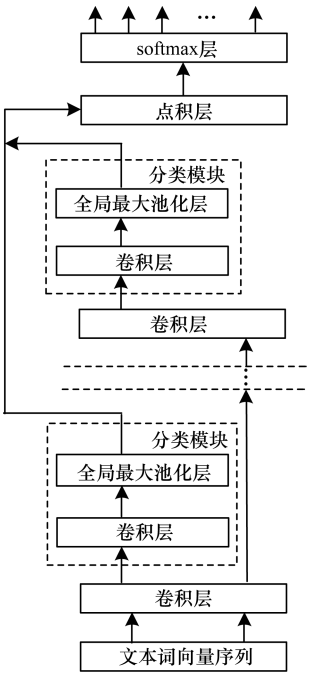

图 2 GMP-CNN 模型

2.1 输入层

词向量是词的分布式表示,将词表示为一个稠密低维度的向量,包含一个词的语法和语义信息。给定由 n 个单词组成的一个文本样本 $\{wrd_1, wrd_2, \cdots, wrd_n\}$,转换每个单词为其对应的 d^{wrd} 维词向量。设该样本中第 i 个单词 wrd_i 对应的词向量为 $\boldsymbol{x}_i$,$\boldsymbol{x}_i \in \mathbb{R}^{d^{wrd}}$。该样本可以初始地表示成一个维度为 $\mathbb{R}^{d^{wrd} \times n}$ 的文本表示矩阵 $\boldsymbol{S} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$。该初始表示矩阵作为 GMP-CNN 模型中第 1 个卷积层的输入。

2.2 深度卷积结构

CNN 模型通常使用不同的窗口对文本的词向量序列进行卷积操作,提取局部语义特征。目前,多数 CNN 模型以超参数方式设定单窗口大小[5]或多窗口大小[6,13]。通常指定的卷积窗口越多,可以提取窗口种类越多的局部语义特征,有助于完成情感分类标注任务。但由于超参数优化、网络规模和计算性能的限制,预先能够指定的窗口种类有限。

基于文献[14]的思想,GMP-CNN 模型可以堆叠多个卷积层形成深度 CNN 模型。设模型有 u 层卷积层,相比第 k 层卷积层,第 $k+1$ 层卷积层在第 k 层卷积层提取的局部语义特征基础上,能提取窗口更大和抽象级别更高的语义特征。由于 $k \in [1, u]$,因此如果 u 值足够大(即堆叠层数足够多),则上下文窗口可以覆盖数据集中最长的文本长度,相当于需要用一个大核的卷积层来捕捉语义的远距离依赖特征。因此,使用小卷积核的卷积层堆叠产生一个大核的卷积层效果,而且多个非线性操作代替一个单一的非线性(或线性)操作能使决策函数更具判别能力[14]。GMP-CNN 模型能产生大窗口的局部语义特征,同时产生多种较小窗口的局部语义特征,这些局部语义特征分别送至对应的分类模块进行分类计算。在超参数设定和调优方面,只需为整个模型设定第 1 个卷积层的窗口大小 w。

2.2.1 same 卷积层

在 GMP-CNN 模型中,每个卷积层均执行窗口大小为 w 的 same 卷积操作。当卷积层输出通道数量与词向量维度相同时,same 卷积操作可以确保每一层卷积层的输入矩阵与输出矩阵为同型矩阵,方便深度堆叠卷积层。

设第 k 层 same 卷积层有输入矩阵 $\boldsymbol{P}_k = [\boldsymbol{p}_{k,1}, \boldsymbol{p}_{k,2}, \cdots, \boldsymbol{p}_{k,n}]$,$\boldsymbol{P}_k \in \mathbb{R}^{d^{wrd} \times n}$,$\boldsymbol{p}_{k,i} \in \mathbb{R}^{d^{wrd}}$。如图 3 所示,$\boldsymbol{P}_k$ 是第 $k-1$ 层 same 卷积层的输出,模型中第 1 层的 same 卷积层输入矩阵 $\boldsymbol{P}_1 = \boldsymbol{S}$。

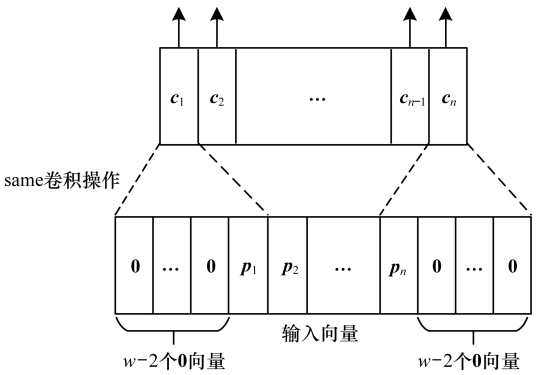

图 3 基于 same 卷积操作的卷积层

在进行 same 卷积计算时,首先在 $\boldsymbol{P}_k$ 左右两端分别填充 $w-2$ 个 $\boldsymbol{0}$ 向量,形成矩阵 $\boldsymbol{Q}_k$。

$$\boldsymbol{Q}_k = [\boldsymbol{0}, \cdots, \boldsymbol{p}_{k,1}, \boldsymbol{p}_{k,2}, \cdots, \boldsymbol{p}_{k,n}, \cdots, \boldsymbol{0}] \tag{1}$$

其中,$\boldsymbol{0}$ 向量维度为 $\mathbb{R}^{d^{wrd}}$,$\boldsymbol{Q}_k \in \mathbb{R}^{d^{wrd} \times [n+2*(w-2)]}$。对 $\boldsymbol{Q}_k$ 进行卷积操作,然后通过激活函数 tanh,计算得到第 k 层卷积层的局部语义特征矩阵 $\boldsymbol{C}_k = [\boldsymbol{c}_{k,1}^{\mathrm{T}}, \boldsymbol{c}_{k,2}^{\mathrm{T}}, \cdots, \boldsymbol{c}_{k,n}^{\mathrm{T}}]$,其中 $\boldsymbol{C}_k \in \mathbb{R}^{d^{wrd} \times n}$,即基于窗口大

小为(k –1)(w –1) + w 的局部语义特征矩阵。

在进行 same 卷积层堆叠时,相邻 same 卷积层相互衔接,并没有加入池化层。这是因为池化层是降采样,虽然保留某种显著特征,但也会过早地丢弃其他特征信息。为使后继 same 卷积层在前层的基础上有效提取局部语义特征,GMP-CNN 模型中相邻 same 卷积层之间没有增加任何形式的池化层。

2.2.2　分类模块

受到 Anytime-Prediction 思想[15]和全局平均池化(Global Average Pooling, GAP)层[16]的启发,本文设计分类模块代替全连接层的功能,如图 4 所示。

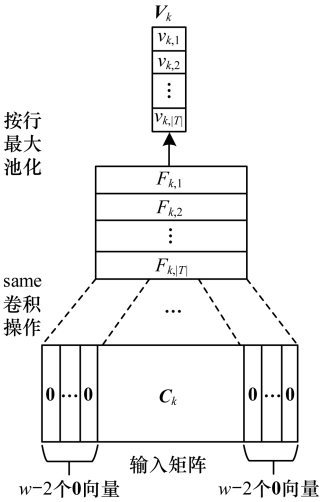

图 4　分类模块

GMP-CNN 模型的分类模块包含 2 个层:输出通道数量与情感类别数量相同的卷积层和 GMP 层。将每层 same 卷积层的输出矩阵送入分类模块的卷积层中,为每个类别生成一个类别特征矩阵,然后 GMP 对类别特征矩阵应用全局最大池化操作,产生一个类别得分向量。分类模块具体操作如下:

1)卷积层。第 k 层的局部语义特征矩阵 $\boldsymbol{C}_k$ 输入第 k 个分类模块的卷积层中,产生类别特征矩阵 $\boldsymbol{F}_k$,$\boldsymbol{F}_k \in \mathbb{R}^{|T| \times n}$,其中 $|T|$ 是情感分类数量。该卷积层采用 same 卷积操作,输出通道与情感分类数量相同。计算过程与 same 卷积计算过程类似。

2)全局最大池化层。将第 k 个分类模块的卷积层生成类别特征向量 $\boldsymbol{F}_k$ 输入全局最大池化模块后,对 $\boldsymbol{F}_k$ 按类别求最大值,即求各行的最大值,产生类别得分向量 $\boldsymbol{v}_k$,$\boldsymbol{v}_k \in \mathbb{R}^{|T|}$,具体操作如下:

$$v_{k,t} = \max(F_{k,t,:}) \tag{2}$$

其中,$v_{k,t}$ 表示基于窗口为(k –1)(w –1) + w 的局部语义特征,得到第 t 类得分,$t \in T$。以 $v_{k,t}$ 为基础,形成 $\boldsymbol{v}_k = [v_{k,1}, v_{k,2}, \cdots, v_{k,|T|}]$。然后以 $\boldsymbol{v}_k$ 为基础,形成文本类别得分矩阵 $\boldsymbol{V} = [\boldsymbol{v}_1^{\mathrm{T}}, \boldsymbol{v}_2^{\mathrm{T}}, \cdots, \boldsymbol{v}_u^{\mathrm{T}}]$,其中 $\boldsymbol{V} \in \mathbb{R}^{|T| \times u}$。

对于传统 CNN 模型,由于全连接层像黑盒一样存在于卷积层和代价函数之间,因此对于分类信息如何回传至卷积层的解释非常困难。GMP 层加强了卷积层和代价函数之间的关联,在理论上具有可解释性[16]。

2.3　点积层与输出层

本文模型综合考虑基于各局部语义特征的类别得分,计算出文本情感分类得分。由于各种局部语义特征的窗口大小和抽象级别不同,因此不同类别得分对文本情感分类贡献不同。点积层对文本类别得分矩阵 $\boldsymbol{V}$ 按列进行加权求和计算,得到文本的情感分类得分向量 $\boldsymbol{scr}$,$\boldsymbol{scr} \in \mathbb{R}^{|T|}$,计算过程如下:

$$scr_t = \sum_{i=0}^{u-1} (\boldsymbol{W} \cdot \boldsymbol{V})_{t,i} \tag{3}$$

其中,scr_t 表示文本对第 t 类的得分,以 scr_t 为基础形成 $\boldsymbol{scr} = [scr_1, scr_2, \cdots, scr_{|T|}]$,$\boldsymbol{W} \in \mathbb{R}^{|T| \times u}$ 是贡献权重矩阵。

输出层对 $\boldsymbol{scr}$ 应用 softmax 函数将句子的情感分类得分转换为情感分类条件概率分布。句子对情感分类 t 的条件概率分布计算如下:

$$p(t|\boldsymbol{S}, \boldsymbol{\theta}) = \mathrm{e}^{scr_t} \Big/ \sum_{i=1}^{|T|} \mathrm{e}^{scr_i} \tag{4}$$

3　GMP-CNN 模型训练

GMP-CNN 模型是通过最小化负对数似然函数进行训练。对式(4)取对数:

$$\ln p(t|\boldsymbol{S}, \boldsymbol{\theta}) = scr_t - \ln\left(\sum_{i=1}^{|T|} \mathrm{e}^{scr_i}\right) \tag{5}$$

采用随机梯度下降(Stochastic Gradient Descent, SGD)算法最小化负对数似然函数,得到:

$$J(\boldsymbol{\theta}) = \sum_{(x_i, y_i \in D)} -\ln p(y_i|\boldsymbol{S}_i; \boldsymbol{\theta}) \tag{6}$$

其中,D 代表训练语料,$\boldsymbol{S}_i$、y_i 表示训练语料的句子及其对应的情感标签,$\boldsymbol{\theta}$ 表示模型所有参数。

过拟合是由训练数据集采样噪声产生,并不是真实地存在于测试数据集[17],会降低模型的泛化能力。此外,SSTb 数据集中长句训练集的样本数量较少,在进行 CNN 模型训练时,过拟合现象较容易发生。在训练过程中,GMP-CNN 模型在输入层使用 Dropout 技术[17],并且模型中各全局池化层对整个网络在结构上做正则化处理[16],因此,本文模型可有效防止过拟合,明显降低泛化误差。

4 实验结果与分析

4.1 情感分析数据集

SSTb 数据集的语料内容来源于在线影评,属于网络短文本[18]。SSTb 不仅有显式的情感实证概率,而且影评相较其他正式类型的文本具有更加主观的表达,因此选用 SSTb 论证 GMP-CNN 模型。SSTb 包含 11 845 个句子和 227 385 个短语,其中短语由句子的语法解析树产生,本文实验只使用句子作为样本数据。数据集有句子和短语的情感实证概率。根据分类标准界限[0.0,0.2]、(0.2,0.4]、(0.4,0.6]、(0.6,0.8]、(0.8,1.0],情感实证概率可映射到五分类中,即表达非常负面、负面、中性、正面、非常正面的情感。在忽略中性类后,分类标准界限为[0,0.4]、[0.6,1.0],将情感实证概率映射到二分类中,即负面和正面情感。

本文按上述标准分别为二分类和五分类划分出 2 套实验数据集。无论在二分类还是五分类实验数据集中,均只包含句子,不包含短语。由于二分类过滤了中性类样本,因此过滤约 20% 的样本,SSTb 数据集划分结果见表 1。

表 1 SSTb 数据集划分结果

数据集	二分类	五分类
训练集	6 920	8 544
验证集	872	1 101
测试集	1 821	2 210

4.2 模型超参数设定

若窗口每次处理范围包含一个词及其上下文,则窗口大小值最小为 3[14],因此 GMP-CNN 模型中卷积层窗口大小 w 设定为 3。考虑到模型中每个卷积层的输出通道数量与词向量维度 d^{wrd} 相同,不宜过低,因此设定为 100。在 GMP-CNN 模型的输入层执行 Dropout 操作,参照文献[17]中的设置,以 $p_{in}=0.5$ 的概率随机保留输入单元。$|D|$ 为每个训练批次包含的样本数,预先设定 $|D|\in\{16,32,64,128\}$,SGD 学习率 λ 为 0.001,通过验证集确定 $|D|$ 为 32。所有超参数设定值见表 2。

表 2 GMP-CNN 超参数设定

参数	参数说明	参数值
w	卷积层窗口大小	3
d^{wrd}	词向量维度	100
p_{in}	Dropout 保留概率	0.5
$\|D\|$	每个批次包含的样本数	32
λ	学习率	0.001

为验证 GMP-CNN 模型的有效性,本文对一系列 parallel-CNN 模型进行实验。除卷积层窗口大小之外,其他参数与表 2 中的设置相同,另外 parallel-CNN 在全连接层的输入也执行 Dropout 操作,随机保留输入单元概率为 p_{in}。parallel-CNN 的卷积层窗口大小设置为相同卷积层数量时 GMP-CNN 模型的等效窗口大小。

4.3 词向量预训练

实验选择 GloVe 算法[19]进行词向量预训练。由于 Twitter 与 SSTb 同属社交网络文本,Twitter 语料库的词语空间分布接近于 SSTb 的词语空间分布,因此本文使用 Twitter 语料库进行词向量预训练。在训练词向量后,得到一个包括一百多万条目的单词表。对于 SSTb 中未出现在单词表中的单词,使用在区间(–0.01,0.01)中的均匀分布随机数进行初始化[20]。

4.4 GMP-CNN 模型结构设置

为实现实验结果的有效对比和论证,在训练过程中对 GMP-CNN 模型的卷积层层数和词向量做不同设定,见表 3。

表 3 实验模型结构设定

模型	说明
GMP-CNN-3-nostatic	堆叠 3 个卷积层,词向量可训练
GMP-CNN-3-static	堆叠 3 个卷积层,词向量不可训练
GMP-CNN-5-nostatic	堆叠 5 个卷积层,词向量可训练
GMP-CNN-5-static	堆叠 5 个卷积层,词向量不可训练
GMP-CNN-7-nostatic	堆叠 7 个卷积层,词向量可训练
GMP-CNN-9-nostatic	堆叠 9 个卷积层,词向量可训练
GMP-CNN-11-nostatic	堆叠 11 个卷积层,词向量可训练
parallel-CNN-3	卷积窗口{3,5,7},词向量可训练
parallel-CNN-5	卷积窗口{3,5,7,9,11},词向量可训练
parallel-CNN-7	卷积窗口{3,5,7,9,11,13,15},词向量可训练
parallel-CNN-9	卷积窗口{3,5,7,9,11,13,15,17,19},词向量可训练

4.5 结果分析

实验选用 Intel I5-4200 的 CPU,8 GB 内存,256 GB的 SSD 硬盘,Linux 操作系统,未使用 GPU。实验开发和运行的操作系统环境是 ubuntu 16.04,在 Anaconda 集成环境中使用 python3.5 语言编写实验代码。实验模型的构建、训练和预测功能模块都是基于深度学习开源软件库 TensorFlow r1.2。

CharSCNN 模型[5]是 CNN 应用在情感分类标注问题的经典模型,采用 2 个相同大小窗口的并行卷积层分别提取单词的构造特征和句子的局部语义特征,并在 SSTb 数据集上验证了该模型的有效性。因此,为验证实验正确率及说明多窗口局部语义特

征的重要性,本文还将给出 CharSCNN 模型在 SSTb 数据集上的实验结果。

4.5.1　GMP-CNN 训练与预测效率分析

从图 5、图 6 可以看出,无论是情感二分类标注还是五分类标注,随着卷积层增加,GMP-CNN 的训练时间和预测时间是近似线性增长,而 parallel-CNN 的训练时间和预测时间增长速率远大于 GMP-CNN,近似为指数增长,主要原因为:1)堆叠结构使得 GMP-CNN 模型计算得到某个大窗口的局部语义特征,同时计算得到一系列较小窗口的局部语义特征,因此 GMP-CNN 在卷积部分的计算量明显少于 parallel-CNN。假设 GMP-CNN 卷积层窗口为 3,词向量维度为 100,卷积层的输出通道数量为 100,文本长度为 20。在堆叠 5 层后,GMP-CNN 取得{3,5,7,9,11}窗口的局部语义特征,这 5 层卷积层共需进行 $5\times[(3\times100)\times20]\times100=3\times10^6$ 次计算。对于 parallel-CNN 模型同样取得{3,5,7,9,11}窗口的局部语义特征,需要进行$[(3+5+7+9+11)\times100\times20]\times100=7\times10^6$ 次计算,GMP-CNN 计算量大幅减少。2)GMP-CNN 使用分类模块代替 parallel-CNN 中的全连接层,分类判别的计算量远小于全连接层的计算量。

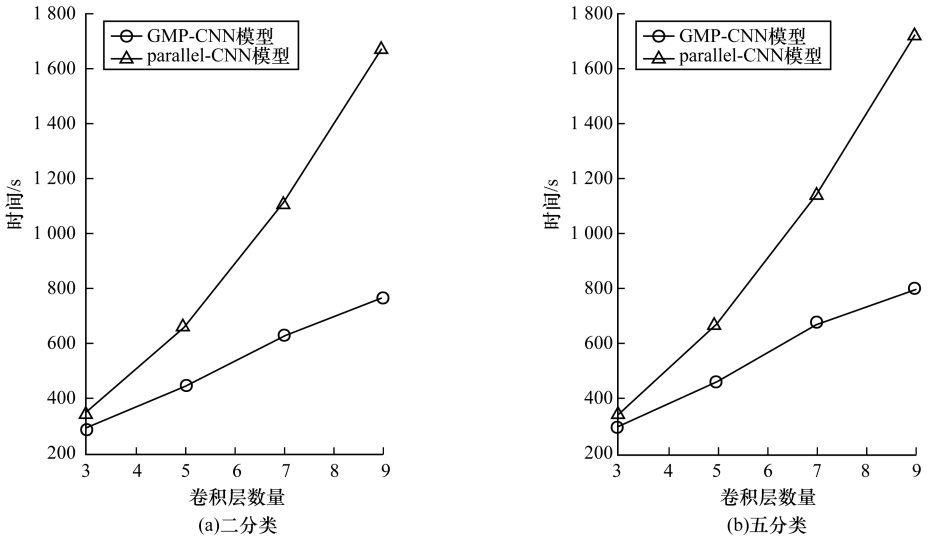

图 5　GMP-CNN 与 parallel-CNN 模型训练时间对比

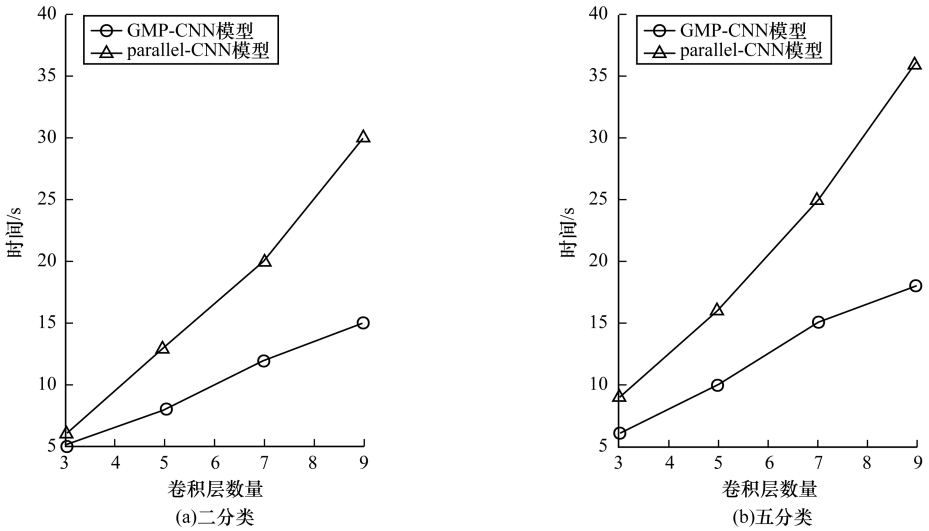

图 6　GMP-CNN 与 parallel-CNN 模型预测时间对比

从图 7、图 8 可以看出,无论在情感二分类还是五分类标注训练过程中,窗口大小相等的 GMP-CNN 模型和 parallel-CNN 模型的训练正确率收敛相似,在相同训练批次,GMP-CNN 模型的训练时间远比 parallel-CNN 模型要少,因此 GMP-CNN 模型在训练效率上比 parallel-CNN 高很多,从而认为其比大多数基于 parallel-CNN 基础结构的 CNN 模型训练效率高。

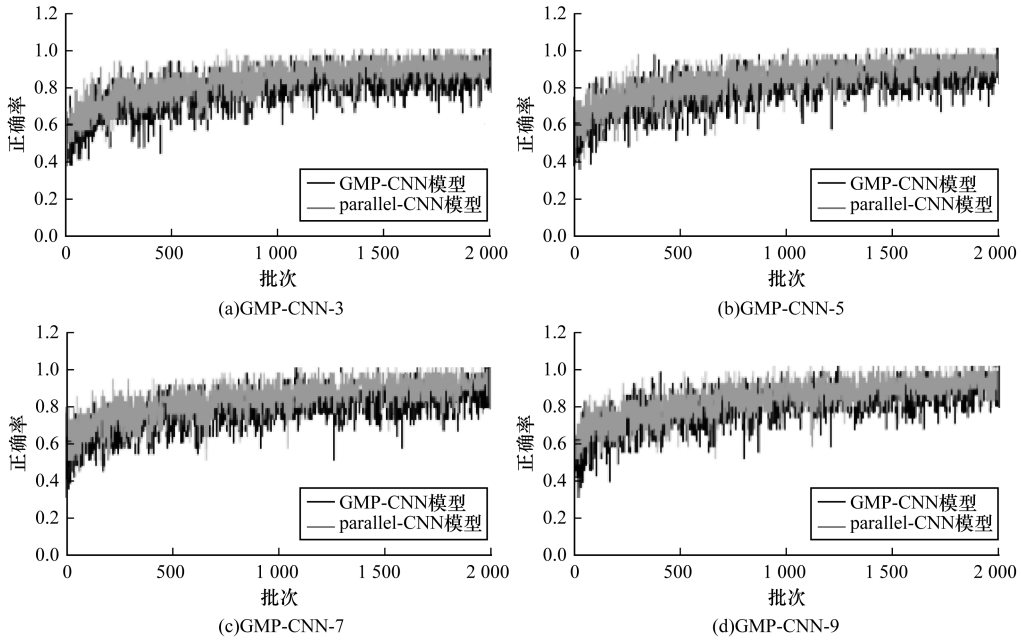

图 7 GMP-CNN 与 parallel-CNN 模型训练正确率对比(二分类)

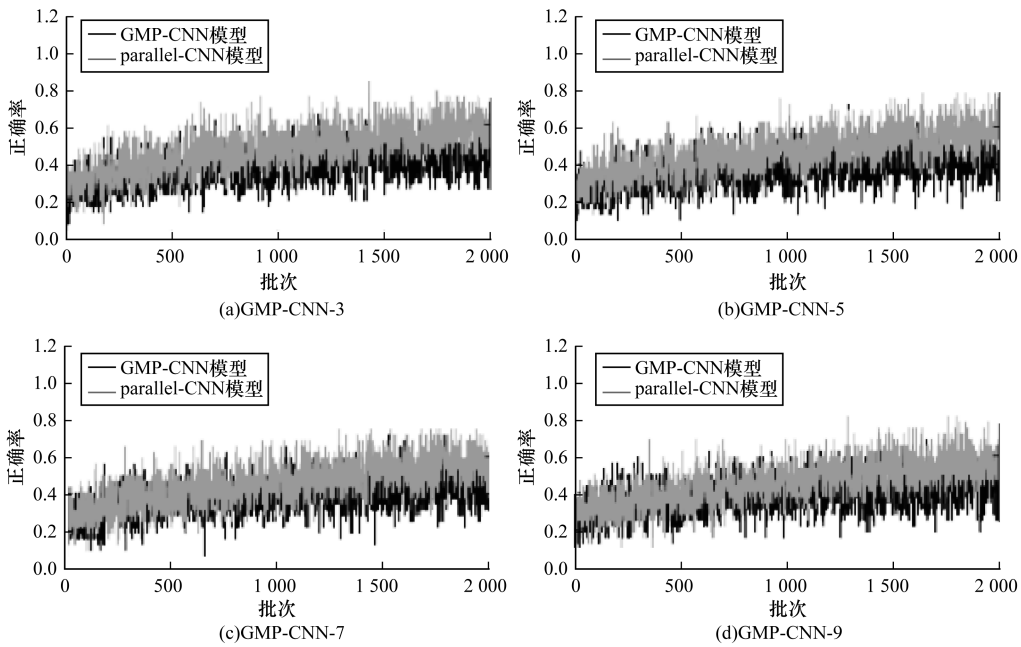

图 8 GMP-CNN 与 parallel-CNN 模型训练正确率对比(五分类)

4.5.2 GMP-CNN 情感分类标注正确率分析

从表 4 可以看出,当进行情感二分类标注任务时,在词向量可以调整的情况下,所有 GMP-CNN 模型正确率均大于 CharSCNN 模型,特别是当卷积层达到 11 层时,正确率比 CharSCNN 模型高 1.8%。在进行情感五分类标注任务时,当卷积层达到 9 层时,GMP-CNN 模型开始优于 CharSCNN 模型,当卷积层达到 11 层时,正确率比 CharSCNN 模型高1.4%,从而验证 GMP-CNN 模型应用于情感分类标注的有效性。

表 4 在 SSTb 数据集上不同模型分类标注正确率 %

模型	二分类	五分类
GMP-CNN-3-nostatic	83.2	42.3
GMP-CNN-5-nostatic	83.5	42.5
GMP-CNN-7-nostatic	83.5	42.7
GMP-CNN-9-nostatic	83.8	43.6
GMP-CNN-11-nostatic	84.1	44.9
GMP-CNN-3-static	79.7	39.8
GMP-CNN-5-static	80.1	40.1
parallel-CNN-3	83.7	44.3
parallel-CNN-5	83.8	43.8
parallel-CNN-7	84.0	45.2
parallel-CNN-9	84.1	44.5
CharSCNN	82.3	43.5

从表 4 还可看到,无论是情感二分类还是五分类标注,parallel-CNN 均优于 CharSCNN。其原因为 CharSCNN 设定一种窗口大小的卷积层,只能提取一种窗口的局部语义特征,而多窗口的局部语义特征可以捕捉更多不同距离上的语义依赖性,这种依赖性对判断文本整体情感分类影响较大,特别是情感多分类标注任务。下文实例说明了在远距离上的语义依赖性对整个句子情感的影响:

实例 1　at all clear what it's trying to say and even if it were -- I doubt it.

实例 2　at all clear what it's trying to say and even if it were -- I doubt it would be all that interesting.

可以看出,实例 2 的负面情感程度比实例 1 弱一些,因为 doubt 后面 4 个词距离上的 all 影响了其强烈程度,从而影响全句负面情感的强烈程度。实例 1 的真实分类是负面,而实例 2 的真实分类是中性。可见,parallel-CNN 正确率虽然有时略高于 GMP-CNN,但总体上基本持平。

4.5.3　GMP-CNN 卷积层层数对标注正确率的影响

GMP-CNN 随着卷积层层数增加,二分类标注和五分类标注正确率总体提高,由此认为正确率的提高主要是因为每增加一层卷积层,就会抽取更大窗口的局部语义特征。虽然每增加一个卷积层,也会增加一个分类模块,使得整个模型规模增加,带来过拟合的可能性,但增加的 GMP 层具有结构上的正则化性[16],从而有效防止模型过拟合的发生。

4.5.4　词向量调整对标注正确率的影响

根据表 4 中 GMP-CNN-3-nostatic、GMP-CNN-3-static 和 GMP-CNN-5-nostatic、GMP-CNN-5-static 实验对比可以看出,对于分类正确率,词向量在训练过程中是否可调整是非常重要的。预训练好的词向量保存词与词之间的通用语法关系,但这种语法关系受限于训练词向量的语料库[7]。同时,SSTb 数据集中有一千多不存在于预训练词向量库中的词,只用随机数代替。因此,将词向量作为 GMP-CNN 训练参数,在训练过程调整词向量。对于预训练好的词向量,这种调整策略可以更好地反映 SSTb 数据集的词与词之间的语法关系。对于随机数代替的词向量,该过程类似针对 SSTb 数据集的情感分类标注任务进行词向量训练。表 5 列举了在二分类标注任务中,GMP-CNN-7-nostatic 训练 2 000 批次后,词向量变化最大的前 10 个词,可以看出这些词有以下特点:1)情感极性强烈的词,如 worst、bad、unfortunately、powerful、problem、unpleasant;2)在文中出现频率较高且能直接影响其他情感词的词,如 too;3)本身有较多词意,但在影评语境下突出某个词意的词,如 works、treat、worth。

表 5　经模型训练后的词向量变化情况

排名	词语	新旧词向量的欧氏距离	出现频次
1	worst	1.95	53
2	too	1.91	449
3	bad	1.83	236
4	unfortunately	1.78	29
5	powerful	1.75	51
6	works	1.75	85
7	treat	1.71	24
8	problem	1.68	53
9	unpleasant	1.61	15
10	worth	1.56	98

5　结束语

本文提出一种多个卷积层堆叠的 GMP-CNN 模型。GMP-CNN 模型能提取出包含多个抽象级别和多种窗口的局部语义特征。实验结果表明,在文本情感分类标注任务中,与其他 CNN 模型相比,GMP-CNN 模型可有效提高训练效率、加快预测速度。下一步将研究更深层次的 CNN 模型在情感分类标注任务中的应用,并综合不同窗口的局部特征,提高 GMP-CNN 模型的情感分类标注正确率。

参考文献

[1] MEDHAT W, HASSAN A, KORASHY H. Sentiment analysis algorithms and applications: a survey[J]. Ain Shams Engineering Journal, 2014, 5(4): 1093-1113.

[2] KUBLER S, MCDONALD R, NIVRE J. Synthesis lectures on human language technologies[EB/OL]. [2018-01-05]. http://www.morganclaypool.com/doi/abs/10.2200/S00416ED1V01Y201204HLT016.

[3] PANG B, LEE L, VAITHYANATHAN S, et al. Sentiment classification using machine learning techniques[C]//Procedings of Empirical Methods in Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 2002: 79-86.

[4] MA M, HUANG L, ZHOU B, et al. Dependency-based convolutional neural networks for sentence embedding[EB/OL]. [2018-01-05]. http://www.oalib.com/paper/4048778.

[5] SANTOS C N D, GATTIT M. Deep convolutional neural networks for sentiment analysis of short texts[C]// Proceeding of the 25th International Conference on Computational Linguistics. Dublin, Ireland: [s.n.], 2014: 69-78.

[6] 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.

[7] LEE G,JEONG J,SEO S,et al. Sentiment classification with word attention based on weakly supervised learning with a convolutional neural network [EB/OL]. [2018-01-05]. https://arxiv.org/abs/1709.09885.

[8] SANTOS C N D, XIANG B, ZHOU B. Classifying relations by ranking with convolutional neural networks[J]. Computer Science,2015,86:132-137.

[9] WANG L, CAO Z, MELO G, et al. Relation classification via multi-level attention CNNs [C]// Proceeding of the 54th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2016: 1298-1307.

[10] LIN Y,SHEN S,LIU Z,et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Philadelphia,USA:Association for Computational Linguistics,2016:2124-2133.

[11] DONG L, WEI F, XU K, et al. Adaptive multi-compositionality for recursive neural network models[J]. IEEE Transactions on Audio, Speech, and Language Processing,2016,24(3):422-431.

[12] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research,2003,3(6):1137-1155.

[13] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of Empirical Methods in Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 2014: 1746-1751.

[14] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [C]// Proceedings of International Conference on Learning Representations. Washington D. C., USA: IEEE Press, 2015:1-7.

[15] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016:770-778.

[16] LIN M,CHEN Q,YAN S. Network in network [EB/OL]. [2018-01-05]. https://arxiv.org/abs/1312.4400.

[17] SRIVASTAVA N, HINTON G E, KRIZHEVSKY A, et al. Dropout:a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research,2014,15(1):1929-1958.

[18] SOCHER R,PERELYGIN A,WU J Y,et al. Recursive deep models for semantic compositionality over a sentiment treebank [C]//Proceedings of Empirical Methods in Natural Language Processing. Philadelphia, USA:Association for Computational Linguistics,2013: 1631-1642.

[19] PENNINGTON J,SOCHER R,CHRISTOPHER D,et al. GloVe: global vectors for word representation [C]// Proceedings of Empirical Methods in Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics,2014:1532-1543.

[20] TANG D,QIN B,LIU T,et al. Aspect level sentiment classification with deep memory network [EB/OL]. [2018-01-05]. https://arxiv.org/abs/1605.08900.

编辑 陆燕菲

(上接第 299 页)

[14] 路荣,项亮,刘明荣,等. 基于隐主题分析和文本聚类的微博客中新闻话题的发现[J]. 模式识别与人工智能,2012,25(3):382-387.

[15] IWATA T, YAMADA T, SAKURAI Y, et al. Online multiscale dynamic topic models[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press,2010:663-672.

[16] CHEN C C, CHEN Y T, SUN Y, et al. Life cycle modeling of news events using aging theory [C]// Proceedings of European Conference on Machine Learning. Berlin,Germany:Springer,2003:47-59.

[17] 蚂蚁软件. 2017 年度社会热点事件传播特点分析[EB/OL]. [2018-01-22]. http://www.eefung.com/hot-report/20180122160439.

[18] 吴平博,陈群秀. 基于时空分析的线索性事件的抽取与集成系统研究[J]. 中文信息学报,2006,20(1):21-28.

[19] ZHANG Y,CHEN M D,LIU L Z. A review on text mining[C]//Proceedings of the 6th IEEE International Conference on Software Engineering and Service Science. Washington D. C. ,USA:IEEE Press,2015:5.

[20] FAHAD A,ALSHATRI N,TARI Z,et al. A survey of clustering algorithms for big data: taxonomy and empirical analysis[J]. IEEE Transactions on Emerging Topics in Computing,2014,2(3):267-279.

编辑 陆燕菲