用于短文本分类的BLSTM_MLPCNN模型

郑 诚 洪彤彤 薛满意

(安徽大学计算机科学与技术学院 合肥 230601)

摘 要 文本表示和文本特征提取是自然语言处理的基础工作,直接影响文本分类的性能。文中提出了以字符级向量联合词向量作为输入的BLSTM_MLPCNN神经网络模型。该模型首先将卷积神经网络(CNN)作用于字符以获取字符级向量,并将字符级向量联合词向量作为预训练词嵌入向量,也即双向长短时记忆网(BLSTM)模型的输入;然后联合BLSTM模型的前向输出、词嵌入向量、后向输出构成文档特征图;最后利用多层感知器卷积神经网络(MLPCNN)进行特征提取。在相关数据集上的实验结果表明:相比于CNN,RNN以及CNN与RNN的组合模型,BLSTM_MLPCNN模型具有更优的分类性能。

关键词 字符级向量,词向量,卷积神经网络(CNN),双向长短时记忆神经网络(BLSTM),多层感知器(MLP),多层感知器卷积网络(MLPCNN)

中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.06.031

BLSTM_MLPCNN Model for Short Text Classification

ZHENG Cheng HONG Tong-tong XUE Man-yi

(School of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract Text representation and text feature extraction are essential procedures in natural language processing and directly affect text classification performance. The major output of the present work is the establishment of the BLSTM_MLPCNN neural network model whose inputs are character-level vector integrated with word vector. In this model, firstly the character-level vector is obtained from character via convolutional neural network (CNN), and is integrated with the word vector to compose the pre-training words embedded vectors (also an input to BLSTM model). Then the combination of the BLSTM model's forward output, word embedded vector and backward output forms the document feature map, and finally the MLPCNN model is used to extract feature. The experiments on the pertinent datasets prove the classification performance of BLSTM_MLPCNN model is superior to CNN model, RNN model and CNN/RNN combinatorial model.

Keywords Character-level vector, Word vector, Convolutional neural network (CNN), Bidirectional long short-term memory network (BLSTM), Multi-layer perceptron (MLP), Multi-layer perceptron convolutional neural network (MLPCNN)

1 引言

随着大数据时代的到来,电子文本数量急剧累积,其中短文本数据占比最大,比如问答系统中用户提出的问题、QQ等聊天软件的聊天记录、网店售货后的卖家商品评论、政府直通车平台的网民意见反馈等。面对如此庞大的短文本数据集,对其维护、管理和应用都具有极大的挑战性。文本分类是自然语言处理的基本任务,合理地对杂乱的大数据短文本进行分类,可以方便用户更快、更好地搜索到自己所需的资料,也方便商家及时了解用户的真实需求,推荐更符合用户喜好的商品。

文本表示是文本分类的经典问题。在传统的文本分类中,常用tf-idf计算每个词项的权重,再构建文本向量空间模型(VSM)进行文本特征表示。这种方法虽然简单有效,但是数据维度非常高,即存在所谓的"维度灾难"问题,而且由于文本较短,数据稀疏性严重。为了更好地进行文本特征表示,2013年,Hinton提出了word embedding概念[1]。基于word embedding的表示方法不但能够避免"维度灾难"问题,还能够从更高的语义层面描述词与词之间的关系[2]。词向量作为卷积网络或循环网络的输入,能较大程度地提升文本分类的性能。

卷积神经网络和循环神经网络是两种主流的应用于自然语言处理领域的深度学习算法。循环神经网络(RNN)能够处理任意长度的序列并获取长期依赖。为了避免RNN的梯度消失和梯度爆炸问题,Schmidhuber教授等于1997年提出长短时记忆网络(LSTM),以更好地记忆与记忆存取。

到稿日期:2018-05-18 返修日期:2018-10-29

郑 诚(1964-),男,副教授,硕士生导师,主要研究方向为自然语言处理、数据挖掘,E-mail:csahu@126.com(通信作者);**洪彤彤**(1994-),女,硕士生,主要研究方向为自然语言处理、数据挖掘;**薛满意**(1995-),男,硕士生,主要研究方向为自然语言处理、数据挖掘。

BLSTM 能像访问过去的上下文信息一样访问未来的上下文信息,对文档进行进一步表示。CNN 中的卷积滤波器是广义线性模型(GLM),当潜在概念的样本线性可分时,GLM 可以取得很好的抽象。然而,潜在概念的样本往往是线性不可分的,因此通常需要运用非线性函数来捕获这些概念的表示。本文运用多层感知器取代 GLM 模型。多层感知器是一个通用函数逼近器和一个可通过反向传播训练的神经网络,由具有非线性激活函数的多个完全连接层组成。线性卷积层和多层感知器卷积层都是对上层特征图进行局部感知,然后在更高层将局部信息综合起来得到全局信息的特征图。由于多层感知器具有跨通道聚合效果,因此本文采用全局平均池化层代替最大池化层进行下采样。考虑到实际文本中噪声大、拼写错误的影响,本文运用卷积神经网络训练字符级向量,然后拼接词向量和字符级向量作为当前词的嵌入向量表示。在 5 个公开标准英文数据集上,本文方法取得了不错的分类效果。

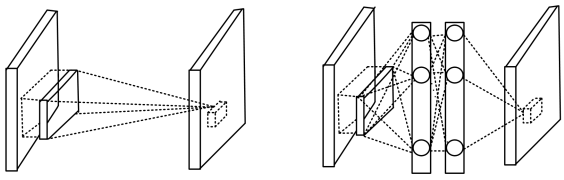

图 1　线性卷积层和多层感知器卷积层的对比

Fig. 1　Comparison linear convolution layer and mlpconv layer

2　相关工作

短文本长度较短,但是信息描述能力强,直接使用传统的文本分类方法无法得到比较好的效果。目前研究中文短文本分类的学者很多,有些学者采用外部知识扩展短文本的方法来提高分类精度,有些学者利用深度学习的方法来解决短文本分类问题,均取得了不错的效果。

1)针对短文本特征扩展的方法。Wang 等[3]在处理短文本分类时,利用一个大的分类知识库为每个类别建立概念模型,并为每个短文本定义一组概念,通过概念相似性对短文本进行分类。宁亚辉等[4]借助知网将短文本的特征词扩展成概念义元,通过对特征进行相似度计算来提高短文本的分类精度。饶高琦等[5]通过 LDA 主题模型获得短文本的主题分布,并把主题中的词作为短文本的特征扩充到原短文本中,以进行文本分类。Sriram 等[6]为微博文本增加作者的配置文件,提出一种文本分类方法。Godin 等[7]和 Mehrotra 等[8]利用 LDA 和微博的标签等特性进行微博文本的分类。

2)基于深度学习的方法。为了捕获不同大小的词关系,Kalchbrenner 等[9]提出了动态的 k-max pooling 机制。Lei 等[10]在标准卷积层使用基于张量的词间操作代替串接词向量的线性运算。Wang 等[11]通过对输入短文本进行语义聚类,构建语义扩展矩阵,进而结合卷积神经网络进行分类。Joulin 等[12]将文本中所有的词向量进行平均,然后将均值化后的词向量直接接入 softmax 层,大大加快了文本分类的效率。Arevian[13]使用循环神经网络对真实世界中的文本进行分类。Tang 等[14]首先用卷积神经网络模拟句子表示,之后利用门控回归神经网络对句子的语义及其关系进行自适应编码,最终实现情感分类。Zhou 等[15]首先用卷积神经网络抽取高维的短语表示,然后将其送入长短时记忆神经网络来获取句子表示。Lai 等[16]在学习单词表示时尽可能使用循环结构来捕捉上下文信息,与传统的基于窗口的神经网络相比,这可能会引入更少的噪声;他们还使用了一个最大池化层,以自动判断在文本分类中起关键作用的词,从而捕获文本中的关键内容。

基于特征扩展的短文本分类方法由于需要借助额外的知识,且待分类文本与知识必须具有语义一致性,因此不仅计算效率较低,而且难以推广。而基于深度学习的方法不需要人工提取特征,同时具有很强的适应性,因此得到了研究者的青睐。

3　BLSTM_MLPCNN 模型

本文提出的 BLSTM_MLPCNN 模型如图 2 所示。该模型主要由三大部分组成:卷积神经网络(CNN)、双向长短时记忆网络(BLSTM)和多层感知器的卷积神经网络(MLPCNN)。其中,CNN 用于字符级向量的训练,BLSTM 获取当前词的上下文信息,MLPCNN 用于局部特征提取和下采样操作。

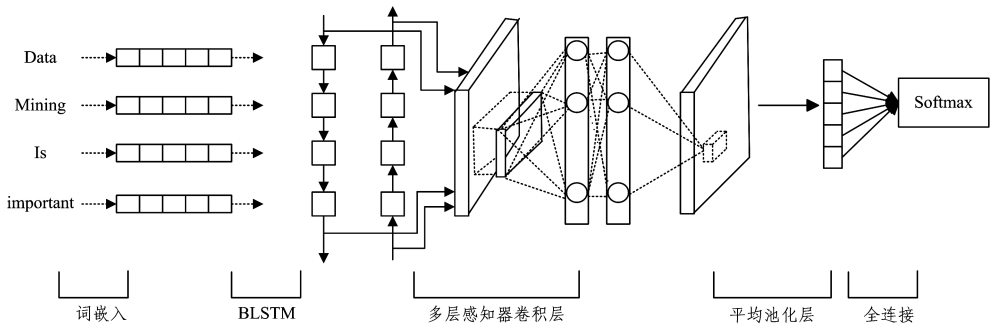

图 2　BLSTM_MLPCNN 模型

Fig. 2　BLSTM_MLPCNN model

3.1　预训练词嵌入向量

BLSTM_MLPCNN 模型输入层的词嵌入向量主要由两部分组成:Glove 词向量和字符级向量。Glove 词向量是已经训练好的英文词向量,每一个词向量的维度为 300。字符级向量是使用卷积网络训练而得的,对于每个字符向量,首先进行随机初始化,然后通过卷积网络使其在模型训练中不断更新。假设单词 w 的 Glove 词向量为($\boldsymbol{v}_1$,$\boldsymbol{v}_2$,…,$\boldsymbol{v}_l$),字符级向量为($\boldsymbol{v}_1'$,$\boldsymbol{v}_2'$,…,$\boldsymbol{v}_{l_0}'$),其中 l 是 Glove 词向量维度,l_0 是字符

级向量维度。输入层单词 w 的嵌入向量为 $e(\boldsymbol{w})=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_l,\boldsymbol{v}_1',\boldsymbol{v}_2',\cdots,\boldsymbol{v}_{l_0}')$。

许多研究人员发现卷积网络(ConvNet)可被用于从原始信号中提取信息。本文将文本作为一种原始信号,利用 ConvNet 进行处理,如图 3 所示。

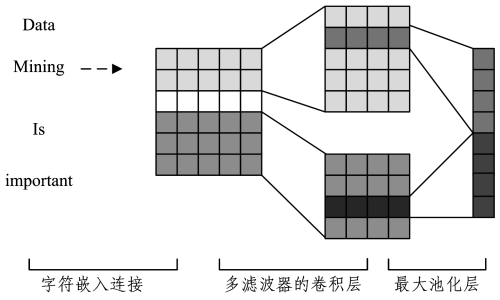

图 3 字符级向量的训练模型

Fig. 3 Training model of character-level vector

设字符词典大小为 m,字符向量嵌入维度为 d,单词 w_t 由一组字符序列$[c_1,c_2,\cdots,c_l]$组成,l 是单词的长度,则单词 w_t 的字符级表示是大小为 $l*d$ 的矩阵,其中第 j 行表示字符 c_j 的嵌入向量。对于大小为 $l*d$ 的矩阵,利用大小为 $w*d$ 的卷积核进行卷积操作,得到特征图 $\boldsymbol{f}^t$,特征图中第 i 行第 j 列元素的计算公式如下:

$$\begin{aligned}\boldsymbol{f}^t[i][j]&=\boldsymbol{f}(x,y)\circ\boldsymbol{w}(x,y)\\&=\sum_{s=1}^{w}\sum_{t=1}^{d}(\boldsymbol{w}(s,t)*\boldsymbol{f}(i-s+w,j-t+d))+b\end{aligned}\tag{1}$$

其中,$\boldsymbol{f}(x,y)$是样本二维矩阵中 x 行 y 列上的取值,$\boldsymbol{w}(x,y)$是滤波器,h'和 d'定义了卷积核大小,b 是偏差。在神经网络模型中,常用的激活函数有许多,如 sigmod 函数、tanh 函数等。为了加快网络的收敛速率,本文选用非线性的 Relu 函数作为激活函数。

获取特征图 $\boldsymbol{f}^t$后,利用最大池化层进行下采样来捕获最重要特征,其计算公式如下:

$$\boldsymbol{y}^t=\max_i \boldsymbol{f}^t[i]\tag{2}$$

假设共有 h 个滤波器,则可以得到 h 个特征图,那么 $\boldsymbol{y}^t=[\boldsymbol{y}_1^t,\boldsymbol{y}_2^t,\cdots,\boldsymbol{y}_h^t]$是单词 w^t的向量表示。

本文所用的数据集包含 70 个字符,包括 26 个英文字母、10 个阿拉伯数字、33 个其他字符。如果数据集中存在某个字符 c,其不在这 70 个字符范围内,则将其标记为 UNK。非空字符如下:

abcdefghijklmnopqrstuvwxyz0123456789

-,;.!?:'''/\|_@#$%^&*~‘+-=<>()[]{}

3.2 文档表示学习

本文先用预训练词的嵌入向量和当前词的上下文信息特征进行词表示,然后构建文档表示。上下文信息有助于获取更为精确的词表示,也可对词消歧。本文模型利用双向长短时记忆神经网络来捕捉当前词的上下文信息。

LSTM 是一种特殊的 RNN,能够学习长期依赖关系。LSTM 网络主要是利用一组门控制来有效地解决 RNN 的梯度消失和梯度爆炸问题。设文档 $\boldsymbol{D}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_t\}$,其中 $\boldsymbol{w}_t$ 表示第 t 个单词。在时间步 t,其输入门 $\boldsymbol{i}_t$、遗忘门 $\boldsymbol{f}_t$、输出门 $\boldsymbol{o}_t$ 被用于控制当前记忆单元 $\boldsymbol{c}_t$ 的更新和当前隐层状态 $\boldsymbol{h}_t$ 的输出。LSTM 转换函数的定义如式(3)—式(8)所示。

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot[h_{t-1},\boldsymbol{w}_t]+\boldsymbol{b}_i)\tag{3}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot[h_{t-1},\boldsymbol{w}_t]+\boldsymbol{b}_f)\tag{4}$$

$$\boldsymbol{q}_t=\tanh(\boldsymbol{W}_q\cdot[h_{t-1},\boldsymbol{w}_t]+\boldsymbol{b}_q)\tag{5}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_o\cdot[h_{t-1},\boldsymbol{w}_t]+\boldsymbol{b}_o)\tag{6}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\otimes\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\otimes\boldsymbol{q}_t\tag{7}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\otimes\tanh(\boldsymbol{c}_t)\tag{8}$$

其中,$\boldsymbol{W}_j$ 是权重矩阵,$\boldsymbol{b}_j$ 是偏差向量,$j\in\{f,i,c,o\}$。$\otimes$表示逐点相乘,σ 表示 S 型函数。$\boldsymbol{f}_t$ 决定哪些信息需要从单元状态中抛弃,$\boldsymbol{i}_t$ 决定哪些值需要更新,$\boldsymbol{o}_t$ 决定模型的输出。

双向 LSTM 模型既可以访问未来的上下文信息,也可访问过去的上下文信息。双向 LSTM 是由两个单向 LSTM 组成的,前向 LSTM 用于访问过去的上下文信息,后向 LSTM 用于访问未来的上下文信息。

设文档 $\boldsymbol{D}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_t,\cdots,\boldsymbol{w}_T\}$,其中 $\boldsymbol{w}_t$ 表示第 t 个单词,且该单词的词表示为 $\boldsymbol{x}_t=[\boldsymbol{h}_t(\boldsymbol{w}_t),\boldsymbol{e}(\boldsymbol{w}_t),\boldsymbol{h}_t'(\boldsymbol{w}_t)]$,其中 $\boldsymbol{h}_t(\boldsymbol{w}_t)$是 BLSTM 的前向输出,$\boldsymbol{e}(\boldsymbol{w}_t)$是当前词的嵌入向量,$\boldsymbol{h}_t'(\boldsymbol{w}_t)$是 BLSTM 的后向输出。因此,文档表示为$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t,\cdots,\boldsymbol{x}_T\}$,其中 $\boldsymbol{x}_t$ 是一维向量表示,则文档表示是二维向量矩阵。

3.3 MLPCNN 模型

经典卷积神经元网络由交替堆叠的卷积层和空间池化层组成。卷积层先通过线性卷积滤波器,随后通过非线性激活函数来生成特征图。非线性激活函数使用 Relu 函数,特征图计算公式如下:

$$f_{i,j,k}=\max(\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_{i,j},0)\tag{9}$$

其中,(i,j)表示特征图下标索引,$\boldsymbol{x}_{i,j}$表示以位置(i,j)为中心的输入区块,k 用于索引特征图的通道。

当潜在概念的实例可线性分离时,这种线性卷积足以进行抽象。但是,实现良好抽象的表示通常是对局部感受野进行更复杂的非线性函数处理。本文提出了多层感知器卷积层(MLPCONV),其可以被看作每个卷积的局部感受野中包含一个多层感知器网络,利用多层感知器对每个局部感受野的神经元进行更加复杂的运算可以提高非线性。本文选择多层感知器是因为:多层感知器本身是一个深层模型,而且多层感知器与卷积神经网络的结构兼容,可以使用反向传播进行训练。多层感知器的卷积层计算如式(10)所示。

$$\begin{aligned}\boldsymbol{f}_{i,j,k_1}^1&=\max(\boldsymbol{w}_{k_1}^{1\mathrm{T}}\boldsymbol{x}_{i,j}+\boldsymbol{b}_{k_1},0)\\\boldsymbol{f}_{i,j,k_n}^n&=\max(\boldsymbol{w}_{k_n}^{n\mathrm{T}}\boldsymbol{f}_{i,j}^{n-1}+\boldsymbol{b}_{k_n},0)\end{aligned}\tag{10}$$

其中,n 表示多层感知器的层数,本文模型中 $n=2$;Relu 函数是多层感知器中的非线性激活函数。

从跨通道汇聚角度看,上述公式等效于在传统卷积层上进行级联跨通道参数汇聚。每个汇聚层在输入特征图上执行加权线性重组,然后通过 Relu 函数进行非线性处理。上层跨通道汇聚特征图是下层跨通道汇聚的输入特征图。跨通道汇聚也等价于具有 $1*1$ 卷积核的卷积层。因此,MLPCONV 实际是由传统 CNN 的卷积层和 n 层卷积核大小为 $1*1$ 的卷积层实现。

由于不同文本内容的上下文信息并不相同,因此本文在卷积层采用了大小不同的滤波器,以提取不同粒度的局部特征。多层感知器的卷积层对局部特征进行特征提取,然后利用全局平均池化层进行下采样操作。

3.4 文档分类

本文分类层采用了 softmax 函数进行分类。卷积神经网络的目标代价函数是交叉熵函数,如式(11)所示。

$$Loss=-\frac{1}{m}\sum_{i=1}^{m}[\boldsymbol{y}^{(i)}\log\boldsymbol{h}_\theta(\boldsymbol{x}^{(i)})+(1-\boldsymbol{y}^{(i)}\log(1-\boldsymbol{h}_\theta(\boldsymbol{x}^{(i)})))]+\lambda\|\boldsymbol{w}\|^2 \tag{11}$$

其中,m 是总样本数,$\boldsymbol{y}^{(i)}$ 是样本 i 的期望输出,$\boldsymbol{h}_\theta(\boldsymbol{x}^{(i)})$ 是样本的网络输出结果。为了防止过拟合,本文引入 L2 正则化来修正损失函数。

4 实验设置

4.1 数据集

在 5 个数据集上对所提出的 BLSTM_MLPCNN 模型进行测试。数据集的汇总统计如表 1 所列。其中,c 为数据集类别数,l 为平均句长,m 为最大句长,$train/dev/test$ 为训练集/验证集/测试集的样本大小,$|V|$ 为词典大小。

表 1　数据集概要

Table 1　Dataset summary

Data	c	l	$train$	dev	$test$	$\|V\|$
MR	2	21	10662	—	CV	20191
Subj	2	23	10000	—	CV	21057
TREC	6	10	5452	—	500	9137
SST1	5	18	8544	1101	2210	17836
SST2	2	19	6920	872	1821	16185

MR:MR 是由 Pang 和 Lee 于 2005 年标注的句子极性数据集,每句话作为一个评论。该数据集分为正、负两类评论,是二分类任务数据集。

Subj:Subj 是主观性数据集,将句子分为主观和客观,包含 5000 个主观句和 5000 个客观句。

TREC:TREC 是问题分类任务数据集。该任务将涉及的问题分为 6 个问题类型(缩写、描述、实体、人员、地点、数值)。

SST1:SST 是由 Socher 等人标注并发布的,是 MR 的扩展。其目的是将评论归类为细粒度标签(非常消极、消极、中立、积极、非常积极)。该数据集中训练集包含 8544 条数据,验证集包含 1101 条数据,测试集包含 2210 条数据。

SST2:该数据集将 SST1 数据集中的中立评论去除,并且把非常消极和消极合并为消极,把积极和非常积极合并为积极。

4.2 对比模型

将本文所提模型与以下模型进行对比分析。

DCNN:Kalchbrenner 等于 2014 年提出的使用 k 最大池化的动态卷积神经网络。

CNN-non-static,CNN-multichannel,CNN-MC:Kim 等于 2014 年首次使用 CNN 进行文本分类;CNN-non-static 模型在训练过程中对词向量进行修改;CNN-multichannel 模型采用多通道进行卷积;CNN-MC 模型先用 CNN 进行特征提取,然后利用全连接进行分类。

CNN-Ana:Zhang 等于 2015 年对一层 CNN 进行敏感度分析,探索了体系结构组件对模型性能的影响。

MVCNN:Yin 等于 2016 年提出的一种新型卷积网络,它结合了不同版本的预训练词嵌入,并利用可变大小的卷积滤波器提取多粒度短语特征。

LSTM,BLSTM 和 Tree-LSTM:LSTM 模型是单向长短时记忆网络,BLSTM 模型是双向长短时记忆网络,Tree-LSTM 模型是 Tai 等于 2015 年提出的将 LSTM 推广到树状网络的拓扑结构。

LSTM-RNN:Le 等于 2015 年提出的 LSTM 的变体。

RCNN:由 Lai 等于 2016 提出,其将循环网络的前向输出、词向量和后向输出进行级联后作为循环网络的输出向量,然后进行最大池化操作,最后利用 softmax 进行分类。

C-LSTM:由 Zhou 等于 2015 年提出,首先利用 CNN 提取高维词表示,然后利用 LSTM 获取文档特征,最后利用 softmax 分类。

DSCNN:Zhang 等于 2016 年提出使用深度敏感卷积神经网络对句子和文档建模,该模型首先利用 LSTM 对预训练词嵌入进行处理,然后利用深度敏感卷积网络进行特征提取,最后利用 softmax 函数进行分类。

4.3 参数设置

1)数据集划分:对于没有标准验证集的数据集,随机选择 10%的训练数据作为验证集。

2)权重初始化:将模型中的所有权重随机初始化为标准差为 0.1 的正态分布随机数。

3)训练参数:字符级向量维度设为 50,最小批次 $mini\_batch$ 为 200,循环层隐藏单元为 100。由 3.3 节的介绍可知,MLPCNN 由一层卷积核大小为 $h*d$ 的卷积层和 n 层卷积核大小为 $1*1$ 的卷积层组成,其中 d 是词表示的维度,h 是卷积核的宽度,本文中 n 为 2。卷积层采用大小不同的滤波器可以提取不同粒度的局部特征,本文中 h 的取值根据不同数据集进行调节,具体取值可参照图 4。卷积核大小为 $h*d$ 的卷积层产生的特征图数为 100,两层卷积核大小为 $1*1$ 的卷积层产生的特征图数均为 150(该值的确定参见 5.2 节)。我们使用 Adam 优化方法训练模型,学习率初始为 0.01,学习率的下降率为 0.05。为了防止过拟合,在全连接层使用 dropout 机制,dropout 取值为 0.5。

5 实验结果及分析

本文在 5 个标准英文数据集上进行实验,以精确率作为评价指标。最终的实验结果如表 2 所列。

表 2　精确率的对比

Table 2　Accuracy comparison

(单位:%)

	Model	MR	Subj	TREC	SST1	SST2
CNN	DCNN	—	—	93.00	48.50	86.80
	CNN-non-static	—	93.40	93.60	48.00	87.20
	CNN-multichannel	81.10	93.20	92.20	47.40	88.10
	CNN-MC	—	93.20	92.00	47.40	88.10
	CNN-Ana	81.02	93.66	91.37	45.98	85.45
	MVCNN	—	93.90	—	49.60	89.40
RNN	LSTM	—	—	—	46.40	84.90
	BLSTM	—	—	—	49.10	87.50
	Tree-LSTM	—	—	—	51.00	88.00
	LSTM-RNN	—	—	—	49.90	88.00
Others	RCNN	—	—	—	47.21	—
	C-LSTM	—	—	94.60	49.20	87.80
	DSCNN	81.50	93.20	95.40	49.70	89.10
Ours	BLSTM-CNN	81.30	93.60	95.20	47.80	87.50
	BLSTM-MLPCNN	83.00	95.00	95.70	49.00	88.20

5.1 整体表现

从表 2 可以看出，本文模型 BLSTM-MLPCNN 在 MR，Subj，TREC 这 3 个数据集上取得了最好的效果，特别是在 MR 和 Subj 数据集上分别取得了 83.0%和 95.0%的精确度。BLSTM-MLPCNN 模型相比于 BLSTM-CNN 模型，在 5 个数据集上的精确度都有明显的提高，分别提高了 1.7%，1.4%，0.5%，1.2%，0.7%，这说明多层感知器的卷积层可以更好地提取文档特征。

本文模型与 CNN，RNN 以及其他模型的对比分析如下。

1）与 CNN 模型对比：CNN 模型直接对词向量输入文本进行局部特征抽取，而 BLSTM-CNN 和 BLSTM-MLPCNN 首先利用 BLSTM 模型获得更具有词表示意义的特征图，然后利用 CNN 进行局部特征提取。从表 2 中可知，相比于 DCNN，CNN-non-sattic，CNN-multichannel，CNN-MC，CNN-Ana，MVCNN 这 6 个 CNN 模型，本文模型在 MR，Subj，TREC 数据集上取得了明显的精度提升。6 种 CNN 模型在 MR，Subj，TREC 数据集上取得的最好的分类精度为81.1%，93.9%，93.6%；而本文模型在这 3 个数据集上的分类精度分别为 83.0%，95.0%，95.7%。

2）与 RNN 模型对比：LSTM 模型将最后一个隐层状态作为输出，然后用 softmax 函数进行分类；而本文是保存 LSTM 模型的每一个隐层状态输出，并将输出与对应词嵌入向量构成一张文档特征图。从表 2 中可知，LSTM，BLSTM，Tree-LSTM 以及 LSTM-RNN 这 4 个 RNN 模型在 SST1 和 SST2 数据集上分别取得了最好的分类精度，为51.0%和 88.0%；本文模型在这两个数据集上的分类精度分别为 49.0%和 88.2%。

3）与其他模型对比：RCNN 和 C-LSTM 是利用 CNN 和 RNN 各自的优点进行模型组合，DSCNN 是深层神经网络。从表 2 可知，DSCNN 的效果远好于 RCNN 与 C-LSTM。在分类精度方面，本文模型相比于 DSCNN 模型，在 MR 数据集上提高了1.5%，在 Subj 数据集上提高了 1.8%，在 TREC 数据集上提高了 0.3%，但是在 SST1 和 SST2 数据集上的表现效果稍弱于 DSCNN 模型。

5.2 多层感知器卷积层的卷积核个数对实验结果的影响

由 3.3 节可知，多层感知器的卷积层实际是由传统 CNN 的卷积层和 n 层卷积核大小为 1 * 1 的卷积层实现。本文在 MR 数据集上探索 n 层卷积核大小为 1 * 1 的卷积核个数 $\{m_1, m_2, \cdots, m_n\}$对实验结果的影响。本文模型中 $n=2$，传统 CNN 卷积层的卷积核个数为 100，实验结果如表 3、表 4 所列。由表 3 可知，当 $m_1=150$ 时，分类效果最佳；由表 4 可知，在 $m_1=150$ 且 $m_2=150$ 时，分类效果最佳。因此，本文模型的两层卷积核大小为 1 * 1 的卷积核个数均为 150。

表 3 m_1 值的确定

Table 3 Determination of m_1 value

m_1	m_2	*Accuracy*/%
50	50	82.1
100	50	81.8
150	50	82.5
200	50	82.2

表 4 m_2 值的确定

Table 4 Determination of m_2 value

m_1	m_2	*Accuracy*/%
150	50	82.5
150	100	81.4
150	150	82.9
150	200	80.4

5.3 多层感知器卷积层的滤波器大小对实验结果的影响

不同大小的卷积窗口可以提取不同粒度的局部特征，因此本文在每个数据集上做了多组实验，选择最好的卷积窗口组合作为该数据集的卷积窗口设置，实验结果表 5 所列；利用柱状图对其进行展示，如图 4 所示。其中，d 表示具有不同滤波器长度的两个并行卷积层，t 表示具有不同滤波器长度的 3 个并行卷积层。

表 5 卷积窗口对实验结果的影响

Table 5 Effect of filter size on experimental results

（单位：%）

	MR	Subj	TREC	SST1	SST2
d:1,2	80.80	94.90	93.25	85.88	49.00
d:2,3	83.00	94.25	95.25	87.65	48.27
d:3,4	82.30	95.00	95.70	86.56	47.96
t:1,2,3	81.70	94.38	95.00	88.20	48.94
t:2,3,4	81.40	94.25	94.25	86.98	48.90
t:3,4,5	82.90	94.13	92.75	86.64	48.82

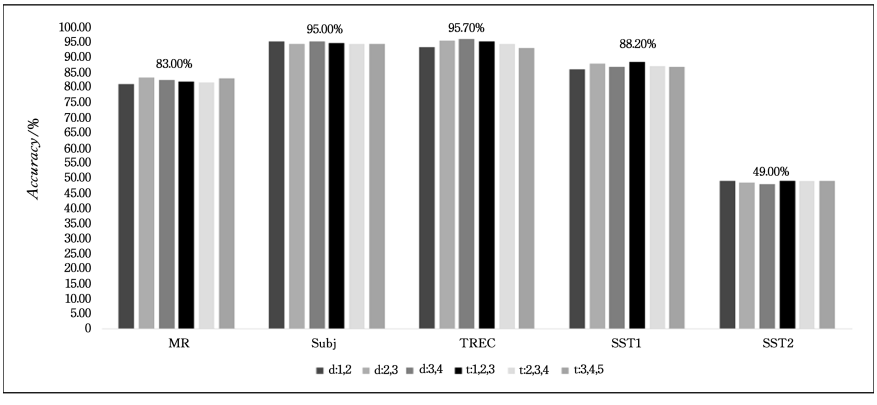

图 4 卷积窗口的确定

Fig. 4 Determination of filter size

结束语 本文提出了 BLSTM-MLPCNN 模型，该模型主要由三大部分组成：卷积网络对字符级向量进行训练构建预

训练词嵌入;BLSTM 模型对预训练词嵌入进行处理,把前向循环输出、预训练词嵌入和后向循环输出级联后作为输出,构成文档特征图;MLPCNN 进行局部特征提取和下采样,以获得文档特征表示向量。在 5 个英文标准数据集上进行实验,结果表明:BLSTM-MLPCNN 模型相比于 CNN,RNN 以及 CNN 与 RNN 的组合模型取得了更好的分类效果,特别是在 MR,Subj 和 TREC 这 3 个数据集上取得了最好的分类精度。更深层的神经网络可以获取更好的特征表示,因此未来的研究方向是构建深层神经网络。

参 考 文 献

[1] HINTON G E. Learning distributed representations of concepts [C]// Proceedings of the Eighth Annual Conference of the Cognitive Science Society. 1986:1-12.

[2] CAI H P. Research on short text classification based on convolutional neural network[D]. Chongqing: Southwest University, 2016. (in Chinese)
蔡慧苹. 基于卷积神经网络的短文本分类方法研究[D]. 重庆:西南大学,2016.

[3] WANG F, WANG Z, LI Z, et al. Concept-based short text classification and ranking[C]// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014:1069-1078.

[4] NING Y H, FAN X H, WU Y. Short text classification based on domain word ontology[J]. Computer Science, 2009, 36(3):142-145. (in Chinese)
宁亚辉,樊兴华,吴渝. 基于领域词语本体的短文本分类[J]. 计算机科学,2009,36(3):142-145.

[5] RAO G Q, YU D, XUN E D. Unsupervised text feature extraction based on natural annotation information and implicit topic model[J]. Journal of Chinese Information Processing, 2015, 29(6):141-149. (in Chinese)
饶高琦,于东,荀恩东. 基于自然标注信息和隐含主题模型的无监督文本特征抽取[J]. 中文信息学报,2015,29(6):141-149.

[6] SRIRAM B, FUHRY D, DEMIR E, et al. Short text classification in Twitter to improve information filtering[C]// Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2010:841-842.

[7] GODIN F, SLAVKOVIKJ V, DE N W, et al. Using topic models for Twitter hashtag recommendation[C]// Proceedings of the 22nd International Conference on World Wide Web. New York: ACM, 2013:593-596.

[8] MEHROTRA R, SANNER S, BUNTINE W, et al. Improving LDA topic models for Microblogs via Tweet pooling and automatic labeling[C]// Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 2013:889-892.

[9] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P A. Convolutional neural network for modelling sentences [J]. arXiv:1404.2188, 2014.

[10] LEI T, BARZILAY R, JAAKKOLA T. Molding CNNs for text: non-linear, non-consecutive convolutions[J]. Indiana University Mathematics Journal, 2015, 58(3):1151-1186.

[11] WANG P. Semantic clustering and convolutional neural network for short text categorization[C]// Proceeding of Meeting of the Association for Computional Linguistics. ACL, 2015:352-357.

[12] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. ACL, 2017:427-431.

[13] AREVIAN G. Recurrent neural networks for robust real-world text classification[C]// IEEE/WIC/ACM International Conference on Web Intelligence. Washington, DC: IEEE, 2007: 326-329.

[14] TANG D Y, QIN B, LIU T. Document modeling with gated recurrent neural network for sentiment classification[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. EMNL, 2015:1422-1432.

[15] ZHOU C H, SUN C, LIU Z, et al. A C-LSTM neural network for text classification[J]. Computer Science, 2015, 1(4):39-44.

[16] LAI S W, XU L H, LIU K, et al. Recurrent convolutional neural networks for text classification[C]// Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI, 2016:2268-2273.