用于短文本分类的 DC-BiGRU_CNN 模型

郑 诚 薛满意 洪彤彤 宋飞豹
(安徽大学计算机科学与技术学院 合肥 230601)
(计算智能与信号处理教育部重点实验室 合肥 230601)

摘 要 文本分类是自然语言处理中一项比较基础的任务,如今深度学习技术被广泛用于处理文本分类任务。在处理文本序列时,卷积神经网络可以提取局部特征,循环神经网络可以提取全局特征,它们都表现出了不错的效果。但是,卷积神经网络不能很好地捕获文本的上下文相关语义信息,循环神经网路对语义的关键信息不敏感。另外,利用更深层次的网络虽然可以更好地提取特征,但是容易产生梯度消失或梯度爆炸问题。针对以上问题,文中提出了一种基于密集连接循环门控单元卷积网络的混合模型(DC-BiGRU_CNN)。该模型首先用一个标准的卷积神经网络训练出字符级词向量,然后将其与词级词向量进行拼接并作为网络输入层。受密集连接卷积网络的启发,在对文本进行高级语义建模阶段时,采用文中提出的密集连接双向门控循环单元,其可以弥补梯度消失或梯度爆炸的缺陷,并且加强了每一层特征之间的传递,实现了特征复用;对前面提取的深层高级语义表示进行卷积和池化操作以获得最终的语义特征表示,再将其输入到 softmax 层,实现对文本的分类。在多个公开数据集上的研究结果表明,DC-BiGRU_CNN 模型在执行文本分类任务时准确率有显著提升。此外,通过实验分析了模型的不同部件对性能提升的作用,研究了句子的最大长度值、网络的层数、卷积核的大小等参数对模型效果的影响。

关键词 字符级词向量,双向门控循环单元,密集连接,卷积神经网络,文本分类
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/jsjkx.180901702

DC-BiGRU_CNN Model for Short-text Classification

ZHENG Cheng XUE Man-yi HONG Tong-tong SONG Fei-bao
(School of Computer Science and Technology, Anhui University, Hefei 230601, China)
(Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, Hefei 230601, China)

Abstract Text classification is a basic task in natural language processing. Nowadays, it is more and more popular to use deep learning technology to deal with text classification tasks. When processing text sequences, convolutional neural networks can extract local features, and recurrent neural networks can extract global features, all of which show good effect. However, convolutional neural networks can not capture the context-related semantic information of text very well, and recurrent networks are not sensitive to the key semantic information. In addition, although deeper networks can better extract features, they are prone to gradient disappearance or gradient explosion. To solve these problems, this paper proposed a hybrid model based on densely connected gated recurrent unit convolutional networks (DC-BiGRU_CNN). Firstly, a standard convolutional neural network is used to train the character-level word vector, and then the character-level word vector is spliced with the word-level word vector to form the network input layer. Inspired by the densely connected convolutional network, a proposed densely connected bidirectional gated recurrent unit is used in the stage of high-level semantic modeling of text, which can alleviate the defect of gradient disappearance or gradient explosion and enhance the transfer between features of each layer, thus achieving feature reuse. Next, the convolution and pooling operation are conducted for the deep high-level semantic representation to obtain the final semantic feature representation, which is then input to the softmax layer to complete text classification task. The experimental results on several public datasets show that DC-BiGRU_CNN has a significant performance improvement in terms of the accuracy for text classification tasks. In addition, this paper analyzed the effect of different components of the model on performance improvement, and studied the effect of parameters such as the maximum length of sentence, the number of layers of

到稿日期:2018-09-11 返修日期:2018-12-01

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;**薛满意**(1995-),男,硕士,主要研究方向为自然语言处理、数据挖掘,E-mail:1270405074@qq.com(通信作者);**洪彤彤**(1994-),女,硕士,主要研究方向为自然语言处理、数据挖掘;**宋飞豹**(1994-),男,硕士,主要研究方向为智能优化计算、数据挖掘。

the network and the size of the convolution kernel on the model.

Keywords　Character-level word vector, Bi-directional gated recurrent unit, Dense connection, Convolutional neural network, Text classification

1　引言

随着移动互联网和社交网络的蓬勃发展,日常生活中会产生越来越多非结构化的短文本数据。这些数据通常具有潜在的科研价值和商业价值,如何对其进行准确分类,引起了学术界和工业界的广泛关注。文本分类是自然语言处理中一项比较基础的任务,在垃圾邮件过滤、情感分析、问答系统、信息检索等领域有着很重要的作用。如果对文本分类进行进一步划分,则其包括主题分类、问题分类、实体分类和情感分类等。目前文本分类主要有两种方法:1)基于传统的机器学习方法;2)现阶段最流行的深度学习方法。

1)基于机器学习的方法。传统的文本分类方法专注于特征工程和使用不同类型的机器学习算法作为分类器。虽然词袋法是一种行之有效且比较简单的方法,但也有很大的局限性。这种表示方法没有考虑到已被证明有用的单词顺序信息,而且还会造成维数灾难,并具有高度的稀疏性。此外,研究者还添加了手工制作的 n-gram 或短语,以利用文本数据中单词顺序的信息。对于机器学习算法,线性分类器被广泛使用,诸如 NaiveBayes[1]、支持向量机 SVM[2]、最近邻 KNN[3] 等。虽然更复杂的特征被设计用来捕获更多的上下文语义和词序信息,但是它们仍然存在数据稀疏问题,这严重影响了分类的效果。

2)基于深度学习的方法。近年来,深度学习技术被广泛用于解决自然语言处理的相关问题。将深层神经网络应用于自然语言处理的基础是词嵌入(Word Embedding)技术,即将每个词映射到一个固定维度的稠密向量。词嵌入的优势在于可以避免使用手工设计和提取特征,并可捕捉词汇隐藏的语义和语法特征,它通常由一个大型文本语料库以无监督的方式训练得到。词嵌入主要有两种方法:word2vec[4] 和 Glove[5]。word2vec 是一种基于预测的词向量模型,有 CBOW 和 Skip-Gram 两种训练模式。GloVe 方法是通过双线性回归模型将全局矩阵分解和局部上下文窗口方法相结合。依据文本粒度的粗细,词嵌入分为字符级词嵌入和词语词嵌入两种。深度学习技术虽然取得了不错的效果,但是容易导致梯度消失和梯度爆炸的问题。

本文将字符级词嵌入和单词级词嵌入拼接作为神经网络的输入层,然后通过密集连接循环门控单元获取更深层的高级别全局语义特征,再通过卷积提取局部语义特征,最后将最终语义表示送入 sofmax 层以实现分类任务。本文所提 DC-BiGRU-CNN 模型在多个数据集上取得了很好的效果。

2　相关工作

随着深度学习在计算机视觉、语音识别等领域取得了显著的效果,许多研究者已将深度网络结构迁移到自然语言处理中,它们在绝大多数任务方面的表现都优于传统方法,极大地促进了其发展。被广泛地应用于文本分类的深度网络有:卷积神经网络和循环神经网络。

1)卷积神经网络。最近 CNN 受到了极大的关注,因为它在 NLP 和计算机视觉的各个领域都表现出了先进的性能。Kim 等[6]针对文本分类任务提出了一种基础的且有效的多通道卷积模型,该模型在情感分类方面的表现较优。根据他们的研究,多个卷积层可以提取高层次的抽象特征。Kalchbrenner 等[7]介绍了一个动态卷积神经网络(DCNN)架构,它使用动态 k-max 池化操作进行句子语义建模,模型在问题和情感分类方面达到了很高的性能。Zhang 等[8]使用字符级卷积网络(ConvNets)进行文本分类。他们使用字符代替文字作为输入,模型包含 3 个完全连接层和 6 个卷积层,用于大型文本分类数据集。刘龙飞等[9]验证了在中文文本微博语料库的情感分类任务中,字符级词嵌入比词语级词嵌入作为卷积神经网络原始特征的输入时效果更好。Santos 等[10]将英文单词字符序列作为基本单元,分别训练得到文本的单词级和句子级特征,提高了短文分类的准确性。虽然卷积神经网络可以捕获文本的局部关键信息,但是不能很好地捕获文本的上下文相关信息。

2)循环神经网络。RNN 因其在一段时间内保存序列信息的卓越能力而备受关注。Zhang 等[11]通过引入相邻层的记忆细胞之间的门控直接连接来扩展深层 LSTM,使信息在不同的层之间畅通无阻地流动,从而解决了在建立更深层的 LSTM 时出现梯度消失的问题,实验在语音识别领域取得了很好的效果。Yang 等[12]将 GRU 网络与注意机制相结合,在词级别、句子级别两个粒度上查找关键信息,在文档分类任务中取得了很好的效果。Nie 等[13]通过引用残差网络的思想,将三层双向 LSTM 进行堆叠,即对原始的句子表示以及前面 LSTM 层的输出做拼接,并将其作为下一层 LSTM 的输入。这种句子编码方式能够最大化多层 LSTM 的学习能力,防止网络在层数增加到一定程度时,无法提升性能。该模型在原始 SNLI 数据集上实现了新的先进的编码方式。Qian 等[14]在没有增大模型复杂度的情况下,通过改变损失函数的策略,把语言学规则(如情感词典、否定词、程度副词)结合到句子级 LSTM 模型中,在情感分类数据集上取得了很好的效果,该文还验证了应用双向的 LSTM 模型对文本序列建模比应用单向的 LSTM 模型的效果更好。Zhou 等[15]提出了 BLSTM-2DCNN 模型,首先利用双向 LSTM 对文本的语义进行建模,然后将文本当作图片进行二维卷积和池化处理,取得了不错的文本分类效果。虽然循环神经网络可以捕获文本的长距离依赖信息,但是不具有位置不变性,对关键信息不敏感。

在深度学习任务中,为了获得更高的分类准确性,一些研究者使用了非常深的卷积神经网络。例如,Johnson 等[16]研

究了如何加深词粒度 CNN 对文本进行全局表达，并找到了一种简单的网络结构，通过增加网络深度来提升准确度，但没有过多地增加计算量。Conneau 等[17]提出了非常深层的 CNN 用于文本分类，实验结果表明他们提出的模型的性能随着深度的增加而增加。然而，非常深的神经网络是耗时的，并且容易出现消失和梯度爆炸问题。针对此缺陷，在计算机视觉领域，Huang 等[18]提出了一种深度密集卷积神经网络 Densenet 用于对象识别任务的架构，在该网络结构设计中，每个层使用所有前面层的特征映射作为输入。

由于单词级词嵌入和字符级词嵌入都可以表示出文本的特征信息，因此本文提出将两者相结合，并将其作为网络输入层，以实现优劣互补。循环卷积网络可以提取全局语义特征，卷积神经网络可以更好地提取局部语义特征。研究表明，无论是卷积神经网络还是循环神经网络，更深层次的深度模型可以提取到更多的特征信息，但是堆叠的深度模型容易导致梯度消失和梯度爆炸的问题。基于此，受 Densenet 的启发，本文提出了密集连接双向 GRU 网络模块用来提取更深层的上下文语义信息，在具有传统 L 层的双向 GRU 网络中，以前馈的方式将每一层连接到后面每一层，这种深层网络可以解决梯度消失问题，加强各层之间的特征传递，支持特征复用，并且极大地减少了参数量；然后再通过卷积操作提取局部特征信息。实验结果表明，利用该模型进行文本分类的效果较优。

3 DC-BiGRU_CNN 模型

DC-BiGRU_CNN 模型的总体结构如图 1 所示，其由字符级词嵌入层、词级嵌入层、密集连接双向门控循环单元层、卷积层、池化层、输出层等组成。

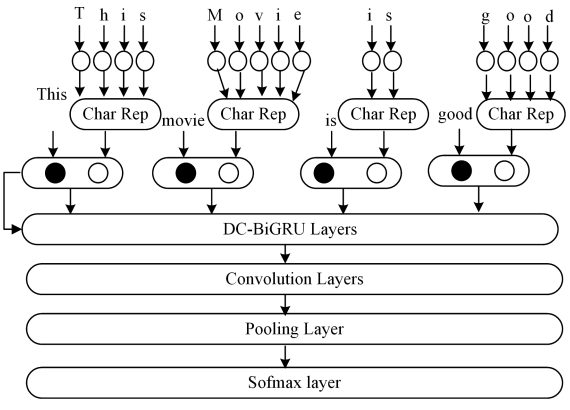

图 1 DC-BiGRU_CNN 模型

Fig. 1 DC-BiGRU_CNN model

3.1 词级嵌入层

3.1.1 字符级词嵌入

已有的研究表明，卷积神经网络是一种可以从单词的字符中提取形态学信息（如单词的前缀与后缀），并将其编码为一个固定维度的向量的有效方法。如给定一个句子 $S=(w_1, w_2, \cdots, w_n)$，其中 n 表示句子的长度，w_i 表示第 i 个词，w_i 包含的字符串长度为 l，w_i 中每一个字符的嵌入向量为 c_j，即每一个字符代表一个特征。如图 2 所示，可以使用一个标准的卷积网络处理每一个词的字符序列得到单词的字符级向量 $e_i^{w_c}$。其计算式如式(1)所示：

$$e_i^{w_c}=\max_{1\leqslant j\leqslant l}\left(W_{\mathrm{CNN}}^{\mathrm{T}}\begin{bmatrix}e^c\left(c_{j-\frac{ke-1}{2}}\right)\\ \cdots\\ e^c\left(c_{j-\frac{ke-1}{2}}\right)\end{bmatrix}+b_{\mathrm{CNN}}\right) \tag{1}$$

其中，W_{CNN} 和 b_{CNN} 表示训练的参数，ke 表示卷积核的大小，max 表示最大池化操作。

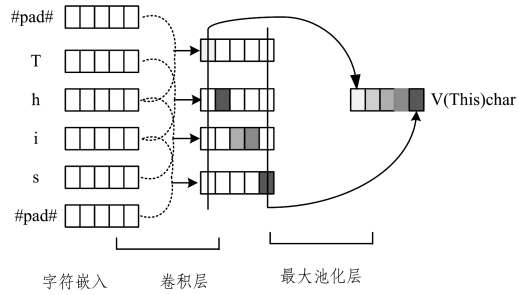

图 2 字符级卷积神经网络的词嵌入结构

Fig. 2 Word-embedding structure of character-level convolution neural network

3.1.2 词级嵌入

一般地，首先加载预训练的词嵌入向量（例如 Glove，word2ve），然后通过 embedding_lookup 查表操作将每个词表示为固定维度的向量 e^w：

$$e_i^w=E(w_i) \tag{2}$$

本文将字符词向量与词向量进行拼接作为网络的输入层，这可看作是对词嵌入的特征拓展，因为字符特性可以为一些非词汇表（OOV）单词提供额外的信息：

$$e_i=[e_i^w;e_i^{w_c}] \tag{3}$$

3.2 密集连接 GRU 网络

3.2.1 GRU 网络

为了解决 RNN 在处理长时间依赖时出现梯度消失和梯度爆炸的问题，相关研究人员提出了中长期循环神经网络（LSTM）和门控循环单元 GRU。在网络结构设计上，GRU 比 LSTM 少了一个门，因此在训练过程中，GRU 的参数比 LSTM 要少，训练速度更快。通过现有研究的比较，GRU 和 LSTM 在大多数任务中的表现相当，但是 GRU 的收敛速度更快。基于此特性，本文使用 GRU 网络对句子进行文本语义特征表示。

GRU 使用门控机制来跟踪序列的状态，而不是使用单独的存储器单元。它有两种类型的门：重置门 r 和更新门 z，两者共同控制信息的更新。在 t 时刻 GRU 计算新状态，如式(4)－式(7)所示：

$$h_t=(1-z_t)\odot h_{t-1}+z_t\odot\tilde{h}_t \tag{4}$$

$$r_t=\sigma(W_r x_t+U_r h_{t-1}+b_r) \tag{5}$$

$$z_t=\sigma(W_z x_t+U_z h_{t-1}+b_z) \tag{6}$$

$$\tilde{h}_t=\tanh(W_h x_t+r_t\odot(U_h h_{t-1})+b_h) \tag{7}$$

其中，W_h，W_z，U_z 表示权重矩阵；b_z，b_h，b_h 为偏置项；σ，tanh 表示激活函数；$\tilde{h}_t$ 表示候选状态；x_t 表示当前 t 时刻的输入；

重置门 r_t 控制过去作用于候选状态的信息量,如果 r_t 的值为 0,则表示忘记之前所有的状态;更新门 z_t 用于控制保留的过去的信息量,以及被添加的新信息量。

在 GRU 对文本序列建模时,每个位置 t 的隐藏状态 h_t 只能对前面的上下文进行正向编码,而不考虑反向上下文。双向 GRU 利用了两个并行通道,一个 GRU 从句首到句末进行文本语义建模,另一个 GRU 从句末向句首进行文本表示,然后将两个 GRU 的隐藏状态进行连接作为每个位置 t 的表示。通过这种方式,当前时刻的输出就不仅仅与之前的状态有关,还与未来的状态有关,因此前后上下文可以同时考虑。其具体的表示如下:

$$\overrightarrow{h}_t = gru(\overrightarrow{h}_t, e(w_t)) \tag{8}$$

$$\overleftarrow{h}_t = gru(\overleftarrow{h}_t, e(w_t)) \tag{9}$$

$$h_t = [\overrightarrow{h}_t \oplus \overleftarrow{h}_t] \tag{10}$$

其中,$\oplus$ 表示连接操作,$\overrightarrow{h}$ 表示前向隐藏状态的输出,$\overleftarrow{h}$ 表示后向隐藏状态的输出。

3.2.2　密集连接模块

顺序堆叠的 GRU 由 L 个相互叠加的 GRU 层组成,即前一层的输出序列形成下一层的输入序列,这种网路具有 L 个连接。虽然这种体系结构能够建立更高层次的表示,但是由于会导致梯度爆炸或消失的问题,训练更深层次的网络通常比较困难。受 Densenet 网络设计的启发,本文使用从任何层到所有后续层的串联操作来进行直接连接,则一个 L 层前馈网络就有 $L(L+1)/2$ 个这样的连接方式,这样处理的优点在于前一层的特征不会被修改,可以实现特征复用。对于每个双向门控循环单元层来说,它都可以直接读取原始的输入特征序列,这样就不需要传递所有的有用信息,只需要向网络添加信息,因此本文每层隐藏层的单元个数较少,其具体结构如图 3 所示。

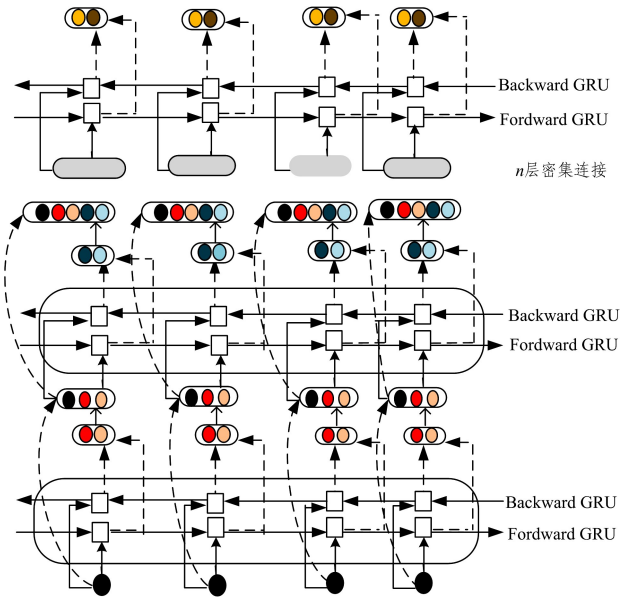

图 3　密集连接双向门控循环单元

Fig. 3　Densely connected bidirectional gated recurrent unit

如第一层输入为 $\{e_{w1}, e_{w2}, \cdots, e_{wn}\}$,第一层输出为 $h^1 = \{h_1^1, h_2^1, \cdots, h_n^1\}$;第二层输入为 $\{[h_1^1; e_{w1}], [h_2^1; e_{w2}], \cdots, [h_1^1; e_{wn}]\}$,第二层输出为 $h^2 = \{h_1^2, h_2^2, \cdots, h_n^2\}$;第三层输入为 $\{[h_1^1; h_1^2; e_{w1}], [h_2^1; h_2^2; e_{w2}], \cdots, [h_n^1; h_n^2; e_{wn}]\}$。

采用这种连接模式,虽然前 $L-1$ 层的 BiGRU 的输出维度都是相同的,但该网络仍然可以不断增加输入特征,然后将网络第 L 层的输出 $H = \{h_1, h_2, \cdots, h_n\}$ 当作语义表示。

3.3　卷积层

本文在密集连接层之后,加入了一个卷积层来捕捉语义的局部特征信息。卷积运算用大小为 k 的卷积核 $m \in R^{k \times d^w}$ 作用于特征图矩阵 H 中的每个窗口以产生新特征。例如,从特征图窗口 $H_{i:i+k-1}$ 产生一个特征 c_i:

$$c_i = f(m \cdot H_{i:i+k-1} + b) \tag{11}$$

其中,b 是一个偏置项;i 的取值范围为 $1 \sim n-k+1$;$\cdot$ 表示矩阵按照元素点乘;f 表示非线性激活函数,如双曲正切 tanh,修正线性单元 relu,sigmod 等。将卷积核 m 应用于特征矩阵 $\{H_{1:k}, H_{1:k+1}, \cdots, H_{n-k+1:n}\}$ 的每个可能的窗口,分别产生一个特征图:

$$C = [c_1, c_2, \cdots, c_{n-k+1}] \tag{12}$$

在实际工作中,可以使用多个卷积核(具有不同的窗口大小)来获得多个不同特征,然后对其结果进行拼接。

3.4　池化层

通过池化操作不仅可以降低文本语义特征的维度,保留主要特征,还可以防止出现过拟合现象。常见的池化操作有两种,即平均池层和最大池化。这里采用最大池化策略获得一个固定长度向量:

$$\hat{C} = \max(c_i) \tag{13}$$

将池化结果进行组合后的结果如下:

$$h^* = [\hat{C}_1, \hat{C}_2, \cdots, \hat{C}_n] \tag{14}$$

3.5　输出层

对于文本分类任务,本文把池化层的输出 h^* 作为输入文本 S 的最终语义表示,然后将其传递到 softmax 分类器层进行归一化,预测文本标签 $\hat{y}$,其计算公式如下:

$$\hat{p}(y|s) = sofmax(W^{(s)} h^* + b^{(s)}) \tag{15}$$

$$\hat{y} = \arg\max_y \hat{p}(y|s) \tag{16}$$

训练目标的最小化损失函数的定义为:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \log(y_i) + \lambda \| \theta \|^2 \tag{17}$$

其中,$t \in R^m$ 是文档的真实标签,用 One-hot 进行编码表示,$y \in R^m$ 是由 softmax 得到每个类的估计概率,m 表示目标分类的个数,λ 是 L2 正则化超参数。

4　实验

4.1　实验数据集

实验数据集的统计结果如表 1 所列。其中,c 为数据集分类的类别数,l 表示平均句子长度,m 为最大的句子长度,train/dev/test 为训练集/验证集/测试集的样本数量,CV 意

味着没有划分出标准的训练集/测试集，本文实验中采用 10 折交叉验证评估算法的准确性。

表 1 语料库统计

Table 1 Corpora statistics

Data	c	m	l	train	dev	test
MR	2	59	21	10662	—	CV
Subj	2	65	23	10000	—	CV
TREC	6	33	10	5452	—	500
SST-1	5	51	18	8544	1101	2210
SST-2	2	51	19	6920	872	1821

实验中的相关数据集如下。

MR(Movie Review)：电影评论数据集，任务是检测正面/负面评论。

Subj(Subjectivity)：主观性数据集，任务是将句子归类为主观或客观。

TREC：问题分类数据集。这项任务涉及判别一个问题的类型，将其划分为 6 种问题类型，即缩写、描述、实体、人、位置、数值。

SST-1(Stanford Sentiment Treebank)：斯坦福大学情感数据库，是 Socher 等对 MR 数据集的延伸。其目的是将评论分类为细粒度标签，即非常消极、消极、中立、积极、非常积极。

SST-2：与 SST-1 数据集相同，但删除了中性评论，只含有二分类标签，即负面、正面。

4.2 实验设置

1)数据集划分。本文将所使用的数据集预先划分为训练集和数据集，其中 SST 数据集还包含划分好的验证集。对于其他不包含标准验证集的数据集，本文随机选取 10%的训练数据作为验证集。

2)训练权重初始化。将本文模型中出现的权重随机初始化为标准差是 0.1 的正态分布随机数。

3)训练超参数。词向量的维度设为 300，字符级词向量的维度设为 50，最小批次 mini_batch 为 200。密集连接层的隐藏单元为 13，最后一层的隐藏层单元为 100，卷积核个数为 100，卷积窗口大小为 4。我们使用 Adam 优化方法加快模型训练速度，学习率初始为 0.01，学习率的下降率为 0.05。为了防止模型出现过拟合，在词嵌入层、池化层设置 *dropout* 值为 0.5。

5 实验结果与分析

5.1 对比实验

1)CNN 模型。CNN-static 模型是由 Kim 等于 2014 年提出的，其在卷积神经网络中直接加载预先训练词向量，且训练时词向量固定不变；CNN-non-static 模型是网络在训练过程中对预训练词嵌入进行微调；CNN-multichannel 模型是以上两种模型的混合模型；MVCNN 模型是由 Yin 等[19]于 2016 年提出的，其分别使用了不同版本的预训练词嵌入，并设置了可变大小的卷积核来提取特征的多粒度短语特征。

2)RNN 模型。LSTM 模型使用标准的单向长短期记忆网络；BLSTM 模型使用标准的双向长短期记忆网络 BLSTM；Tree-LSTM 模型是 Tai 等[20]于 2015 年首次将标准的 LSTM 体系结构推广到树结构网络拓扑中得到的，展示了它在表示句子意义方面的优势；Multi-Task 是 Liu 等[21]提出的，其将 RNN 集成到多任务学习框架中，该框架会将任意文本映射为具有任务特定层和共享层的语义向量表示。

3)混合模型。C-LSTM 模型是由 Zhou 等[22]于 2015 年提出的，首先利用 CNN 提取高维词向量语义表示，然后利用 LSTM 获取文档特征，最后利用 softmax 进行文本分类；DSCNN 模型是由 Zhang 等[23]于 2016 年提出的，其依赖敏感卷积神经网络，首先通过 LSTM 网络层处理词向量，获取句子间和句子内的长期依赖关系，然后通过卷积运算提取特征进行分类；BLSTM-2DCNN 模型是由 Zhou 等[15]于 2016 年提出的，其对文本采用计算机视觉领域中的图片处理模式进行二维卷积和池化操作；Conv-RNN 模型是由 Wang 等[24]于 2017 年提出的，其先通过双向 GRU 对文本进行语义建模，然后对其进行卷积池化操作。

5.2 整体表现

实验结果如表 2 所列。在测试的 5 个数据集中，与 CNN 模型、RNN 模型，以及其他先进的混合模型相比，本文提出的 DC-BiGRU_CNN 模型取得了不错的分类效果，特别是在 MR，Subj，TREC 3 个数据集上分别取得了 83.4%，94.9%，96.2%的精度。相比 Conv-RNN 模型，本文模型在 MR，Subj，SST-1，SST-2 4 个数据集上的分类精度都有提高，分别提高了 1.41%，0.77%，0.23%，0.19%。

表 2 数据集上的测试结果

Table 2 Test results on datasets

(单位：%)

	Model	MR	Subj	TREC	SST-1	SST-2
CNN	CNN-non-static	81.5	93.4	93.6	48.0	87.2
	CNN-static	81.0	93.0	92.8	45.5	86.8
	CNN-multichannel	81.1	93.2	92.2	47.4	88.1
	MVCNN	—	93.9	—	49.6	89.4
RNN	LSTM	—	—	—	46.4	84.9
	BLSTM	—	—	—	49.1	87.5
	Tree-LSTM	—	—	—	51.0	88.0
	Multi-Task	—	94.1	—	49.6	87.9
Others	C-LSTM	—	—	94.6	49.2	87.8
	DSCNN	81.5	93.2	95.4	49.7	89.1
	BLSTM-2DCNN	82.3	94.0	96.1	52.4	89.5
	Conv-RNN	81.99	94.13	—	51.67	88.91
Ours	BiGRU_CNN	81.9	93.8	95.1	50.5	87.9
	DC-BiGRU	82.9	94.5	95.8	51.5	88.5
	No-char-DC-BiGRU_CNN	82.6	94.1	95.0	51.2	88.3
	DC-BiGRU_CNN	**83.4**	**94.9**	**96.2**	**51.9**	**89.1**

为了分析不同部件对性能提升的作用，本文设计了 3 个对比模型：1)仅使用词级别的词嵌入作为网络输入层且其他组件不变的 No-char-DC-BiGRU_CNN 模型；2)在密集连接模块未使用卷积操作且其他组件不变的 DC-BiGRU 模型；3)未使用密集连接组件的 BiGRU_CNN 模型。由表 2 可知，以上 3 种模型相比 DC-BiGRU_CNN 模型在测试时的准确率都有所下降，这证实了我们整合的字/词级嵌入、CNN、密集连接

BiGRU 等组件是十分有效的。

5.3　句子长度对模型的影响

一般地,在对模型进行训练时,需要对其指定一个最大句子长度值。为了研究最大句子长度的选取对分类任务准确率的影响,我们选取 MR 数据集进行实验。本文统计的句子长度分布如图 4 所示。

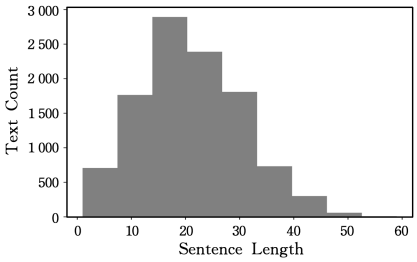

图 4　MR 数据集的句子长度分布

Fig. 4　Sentence length distribution of MR dataset

句子的最大长度值对准确率的影响如图 5 所示。对于 MR 数据集,句子最大长度取值为 35 时的效果最好。如果最大长度值较小,对文本进行截取时会造成语义特征信息丢失;如果最大长度值较大,则 padding 填充对模型准确率的影响也会随之增加。

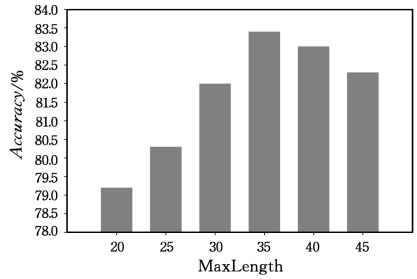

图 5　句子的最大长度值对准确率的影响

Fig. 5　Impact of maximum length of sentence on accuracy

5.4　网络深度对模型的影响

本文选取 MR 数据集研究了在模型中密集连接层的层数对模型性能的影响。实验结果如图 6 所示。结果表明,在一定条件下,模型的性能随着网络层数的加深而提高,但是如果层数太深,模型的准确率则并没有提升,反而有所下降。由实验结果可知,模型在层数为 15 左右时效果最好。

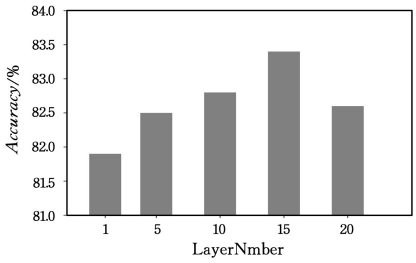

图 6　密集网络的层数对模型准确率的影响

Fig. 6　Impact of layer number of dense network on model's accuracy

5.5　卷积核大小对模型的影响

为了获得最优性能时卷积窗口的尺寸,本文使用 DC-BiGRU_CNN 对 MR 数据集进行实验,并将卷积核的通道数量设置为 100,卷积核大小分别设置为 2×350,3×350,4×350,5×350,6×350。图 7 中,横坐标 C_n 表示 D 卷积窗口的大小为 $n\times350$,纵坐标表示准确率。实验表明,在 MR 数据集上,卷积窗口的大小为 4 时效果最好,预测精度达到了 83.4%。通过分析可知:如果使用更大的滤波器,卷积可以检测到更多的特征,并且性能也可以得到改善。但是,在训练网络时使用太大的滤波器,不仅会占用更多的存储空间,还会消耗更多的时间,且准确率反而有所下降。

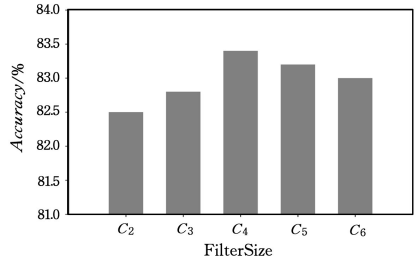

图 7　不同尺寸滤波器的预测精度

Fig. 7　Prediction accuracy of filters with different sizes

结束语　对句子进行语义建模是自然语言处理任务中的核心步骤之一。本文中,首先将字符级特性与词语级特征融合的向量作为网络的输入层;然后通过多层密集连接的双向门控循环单元对语义进行全局建模,充分利用从所有较低层次的层学习的信息,建立底层特征和高层特征之间的跨层连接以丰富语义特征;最后对语义进行卷积,提取局部语义特征,从而实现分类任务。经过多个数据集的验证表明,本文提出的 DC-BiGRU_CNN 模型是十分有效的。近年来,Attention 机制在自然语言处理领域取得了很好的效果,未来工作将结合本文模型对其进行探讨。

参 考 文 献

[1] JOACHIMS T. Text categorization with Support Vector Machines: Learning with many relevant features[C]// European Conference on Machine Learning. Berlin: Springer, 1998: 137-142.

[2] CHEN Z, SHI G, WANG X. Text Classification Based on Naive Bayes Algorithm with Feature Selection[J]. International Journal on Information, 2012, 15(10): 4255-4260.

[3] VRIES A D, MAMOULIS N, NES N, et al. Efficient KNN search on vertically decomposed data[C]// Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data. Madiso: ACM Press, 2002: 322-333.

[4] TOMAS M, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// In Advances in Neural Information Processing Systems. 2013: 3111-3119.

[5] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

[6] KIM Y. Convolutional Neural Networks for Sentence Classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1746-1751.

[7] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A Convolutional Neural Network for Modelling Sentences[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, 2014: 655-665.

[8] ZHANG X, ZHAO J B, LECUN Y. Character-level convolutional networks for text classification[C]// Proceedings of the International Conference on Neural Information Processing Systems. Montreal, 2015: 649-657.

[9] LIU L F, YANG L, ZHANG S W, et al. Convolutional Neural Networks for Chinese Micro-blog Sentiment Analysis[J]. Journal of Chinese Information Processing, 2015, 29(6): 141-149. (in Chinese)
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): 141-149.

[10] SANTOS C, GATTI M. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts[C]// Proc of International Conference on Computational Linguistics. 2014: 69-78.

[11] ZHANG Y, CHEN G G, YU D, et al. Highway long short-term memory RNNS for distant speech recognition[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2016: 5755-5759.

[12] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]// Proceedings of the Conference on the North American Chapter of the Association for Computational Linguistics. Human Language Technologies, 2016: 1480-1489.

[13] NIE Y, BANSAL M. Shortcut-Stacked Sentence Encoders for Multi-Domain Inference[C]// The Workshop on Evaluating Vector Space Representations for Nlp. 2017: 41-45.

[14] QIAN Q, HUANG M, LEI J H, et al. Linguistically regularized LSTMs for Sentiment Classification[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Canada: ACL, 2017: 1679-1689.

[15] ZHOU P, QI Z, ZHENG S, et al. Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling[C]// Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016: 3485-3495.

[16] JOHNSON R, ZHANG T. Deep Pyramid Convolutional Neural Networks for Text Categorization[C]// Meeting of the Association for Computational Linguistics. 2017: 562-570.

[17] CONNEAU A, SCHWENK H, BARRAULT L, et al. Very Deep Convolutional Networks for Text Classification[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2016: 1107-1116.

[18] HUANG G, LIU Z, MAATEN V D L, et al. Densely connected convolutional networks[C]// In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017: 2261-2269.

[19] YIN W, SCHUTZE H. Multichannel variable-size convolution for sentence classification[C]// Proceedings of the Conference on Natural Language Learning (CoNLL). 2015: 204-214.

[20] TAI K S, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[C]// Annual Meeting of the Association for Computational Linguistics (ACL 2015). Beijing, China, 2015: 1556-1566.

[21] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[C]// International Joint Conference on Artificial Intelligence. AAAI Press, 2016: 2873-2879.

[22] ZHOU C, SUN C, LIU Z, et al. A C-LSTM Neural Network for Text Classification [J]. Computer Science, 2015, 1(4): 39-44.

[23] ZHANG R, LEE H, RADEV D. Dependency sensitive convolutional neural networks for modeling sentences and documents [C]// Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016: 1512-1521.

[24] WANG C L, JIANG F J, YANG H X. A hybrid framework for text modeling with convolutional rnn[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 2061-2069.