•信息工程•

DOI:10.15961/j.jsuese.201801160

基于串并行卷积门阀循环神经网络的短文本特征提取与分类

唐贤伦[1], 林文星[1], 杜一铭[2], 王　婷[1]

(1.重庆邮电大学 自动化学院, 重庆 400065; 2.重庆邮电大学 计算机学院, 重庆 400065)

摘　要:针对短文本数据特征少、提供信息有限,以及传统卷积神经网络(convolutional neural network, CNN)和循环神经网络(recurrent neural network, RNN)对短文本特征表示不充分的问题,提出基于串并行卷积门阀循环神经网络的文本分类模型,处理句子特征表示与短文本分类。该网络在卷积层中去除池化操作,保留文本数据的时序结构和位置信息,以串并行的卷积结构提取词语的多元特征组合,并提取局部上下文信息作为RNN的输入;以门阀循环单元(gated recurrent unit, GRU)作为RNN的组成结构,利用文本的时序信息生成句子的向量表示,输入带有附加边缘距离的分类器中,引导网络学习出具有区分性的特征,实现短文本的分类。实验中采用TREC、MR、Subj短文本分类数据集进行测试,对网络超参数选择和卷积层结构对分类准确率的影响进行仿真分析,并与常见的文本分类模型进行了对比实验。实验结果表明:去掉池化操作、采用较小的卷积核进行串并行卷积,能够提升文本数据在多元特征表示下的分类准确率。相较于相同参数规模的GRU模型,所提出模型的分类准确率在3个数据集中分别提升了2.00%、1.23%、1.08%;相较于相同参数规模的CNN模型,所提出模型的分类准确率在3个数据集中分别提升了1.60%、1.57%、0.80%。与Text–CNN、G–Dropout、F–Dropout等常见模型相比,所提出模型的分类准确率也保持最优。因此,实验表明所提出模型可改善分类准确率,可实际应用于短文本分类场景。

关键词:特征表示;短文本分类;循环神经网络;门阀循环单元

中图分类号:TP391　　**文献标志码**:A　　**文章编号**:2096-3246(2019)04-0125-08

Short Text Feature Extraction and Classification Based on Serial–Parallel Convolutional Gated Recurrent Neural Network

TANG Xianlun[1], *LIN Wenxing*[1], *DU Yiming*[2], *WANG Ting*[1]

(1.School of Automation, Chongqing Univ. of Posts and Telecommunications, Chongqing 400065, China;
2.School of Computer Sci. and Technol., Chongqing Univ. of Posts and Telecommunications, Chongqing 400065, China)

Abstract: In order to address the problems that the features and information is limited in short text, the short text features are not fully expressed by traditional convolutional neural network (CNN) and recurrent neural network (RNN), a text classification model named convolutional gated recurrent neural network was proposed to represent sentence feature vector and classify short texts. The pooling operation was removed in convolution layerof the model to retain sequential structure and location information in text data. Series–parallel convolution structure was used to extract multi-feature combination of words and local context information as the input of RNN. Then, the gated recurrent unit (GRU) was used as the structure of RNN to represent the sentence features based on the sequential information of text. The features were input to the classifier with additive margin to guide network to learn distinguishing features and realize short text classification. The short text classification data set TREC, MR, and Subj were applied for testing. The influence of network hyper-parameters selection and convolution layer structures on classification accuracy were simulated and analyzed, and common text classification models were compared in experiments. Experimental results demonstrated that the classification accuracy of text data was improved by removing the pooling operation and using smaller convolution kernels for series–parallel convolution in the multi-feature representation. Compared with the GRU with the same number of parameters, the classification accuracy of the proposed model was increased by 2.00%, 1.23% and 1.08% in three datasets respectively. Compared with the CNN with the same number of parameters, the classification accuracy of the proposed model was increased by 1.60%, 1.57% and 0.80% in three datasets respectively. Compared

收稿日期:2018 – 10 – 20

基金项目:国家自然科学基金项目(61673079);重庆市基础科学与前沿技术研究项目(cstc2016jcyjA1919)

作者简介:唐贤伦(1977—),男,教授. 研究方向:模式识别;计算智能. E-mail: tangxl@cqupt.edu.cn

网络出版时间:2019 – 07 – 08 11 : 32 : 57　　**网络出版地址**: http://kns.cnki.net/kcms/detail/51.1773.TB.20190705.1302.001.html

with Text–CNN, G–Dropout, F–Dropout and other common models, the classification results also kept best. Therefore, experiments showed that the classification accuracy was effectively improved by the proposed model, which could be applied to short text classification scenarios.

Key words: feature representation; short text classification; recurrent neural network; gated recurrent unit

对文本数据进行有效的分类是自然语言处理领域的经典问题。当前，针对文本分类的模型已有大量的研究。Diab等[1]使用差分进化算法微调朴素贝叶斯分类器的参数，并成功应用于文本分类领域。Zhang等[2]使用了一种半监督聚类的文本分类模型，该模型将训练集进行聚类并标记簇的标签，测试集归入距离最近的簇，以聚类簇的标签作为分类结果。除此以外，隐马尔科夫模型[3]、决策树[4]也在文本分类领域应用广泛。但传统的文本特征表示方法难以把握文本语义，而语义信息往往在文本分类等领域举足轻重[5]。

深度学习引入了词嵌入（word embeding）的机制，将文本数据映射到一个低维度的词向量，为文本的表示方法引入语义信息。CNN作为特征提取器已被应用于许多领域并获得显著成果。Kim[6]将CNN模型用于处理文本分类，将一段文本当作一幅固定长宽的图像，该模型融合了不同大小的卷积核对词向量进行卷积和降采样操作。孙松涛等[7]使用多通道的CNN模型，有效地利用表情符号提升微博情感分类效果。Meng等[8]为CNN引入注意力机制并成功应用于文档建模中。

CNN模型虽然取得了不错的效果，但都没有考虑文本的词序，将文本中每个词的出现都认为是独立的，不依赖于其他词是否出现。然而，文本是一个长串的序列，文中各个词的出现是与上下文相互关联的。相比上述模型，GRU等RNN模型更适合处理例如文本、语音这样的序列数据。Lu等[9]采用类似CNN滑动窗口的机制，将多个词语混合输入RNN中。Zhou等[10]使用字符级别词嵌入与词语级别词嵌入并行输入RNN，在汉语短文本分类中取得了不错的效果。Yang等[11]采用层次化的注意力RNN模型，对于句子级别和文档级别的文本进行层次化建模。Wang等[12]采用1维卷积网络提取词向量并降采样特征，再以RNN生成时序表达。谢金宝等[13]将注意力神经网络、长短期记忆网络与卷积神经网络的多通道特征融合，建立中文文本分类模型。上述采用卷积网络的模型中均采用了不同程度的池化操作。但短文本数据对卷积层中池化操作比较敏感，LiD等[14]指出在序列建模过程中使用池化会丢失文本局部的位置信息和序列结构。

针对以上问题，作者尝试去除池化操作，以避免对词语时序破坏和短文本信息因采样造成的丢失。采用串并行卷积结构丰富词向量在多尺度卷积下的特征组合。将双层串并行卷积替代等效步幅的一层卷积，起到提高非线性的作用。并将第1层卷积并行到输出GRU中，利用GRU处理序列数据的优势，最终生成句子的向量，辅助特征学习。同时，在Softmax分类器中引入附加边缘余弦距离，引导网络学习出具有区分性的特征表示，使得文本分类的准确率进一步提升。

1 串并行卷积门阀循环神经网络

作者建立了串并行卷积门阀循环神经网络（SPCGRU）处理短文本特征抽取与分类。网络的结构如图1所示，该模型主要由词向量处理层、串并行卷积层、GRU层和分类输出层4个部分组成。

1.1 卷积神经网络

采用的卷积方式是1维宽卷积，通过填充0值，保持卷积后输出的序列长度不发生变化。固定卷积核$\boldsymbol{W}\in\mathbb{R}^{k\times d}$的长度为词向量的维数$d$；$\boldsymbol{W}$的宽度$k$为变量，表示滑动窗口的大小。落入$\boldsymbol{W}$的$k$个词向量依次为$\boldsymbol{x}_j,\boldsymbol{x}_{j+1},\cdots,\boldsymbol{x}_{j+k-1}$，将其表示为矩阵$\boldsymbol{X}_j\in\mathbb{R}^{k\times d}$：

$$\boldsymbol{X}_j=[\boldsymbol{x}_j,\boldsymbol{x}_{j+1},\cdots,\boldsymbol{x}_{j+k-1}] \tag{1}$$

对于卷积核中的词向量进行卷积运算得到当前窗口的特征：

$$c_j=f(\boldsymbol{X}_j*\boldsymbol{W}+b) \tag{2}$$

式中：*为卷积运算；b为偏置；$f(\cdot)$为激活函数，采用的激活函数为修正线性单元(rectified linear units, ReLU)。ReLU激活函数的形式如下：

$$f(x)=\max(0,x) \tag{3}$$

SPCGRU模型采用了多种卷积核形成特征图，对于n个卷积核进行卷积得到的特征可表示为：

$$\boldsymbol{C}_j=[c_{1j},c_{2j},\cdots,c_{nj}] \tag{4}$$

式中，c_{ij}为第i个卷积核对$\boldsymbol{X}_j$进行卷积得到的特征。若不采用最大池化，$\boldsymbol{C}_j$就直接作为后续网络输入。

最大池化是通过最大化词语的特征表示以减少特征参数的方法。但最大池化的非连续的提取特征对文本的时序性也造成了破坏，会造成局部短语内部词语位置信息丢失。

对特征图最大池化可表示为只获取池化窗口内的每一维的最大特征$p_j\in\mathbb{R}$：

$$p_j=\max(\boldsymbol{C}_j,\boldsymbol{C}_{j+1},\cdots,\boldsymbol{C}_{j+l-1}) \tag{5}$$

式中，l为最大池化的窗口大小。

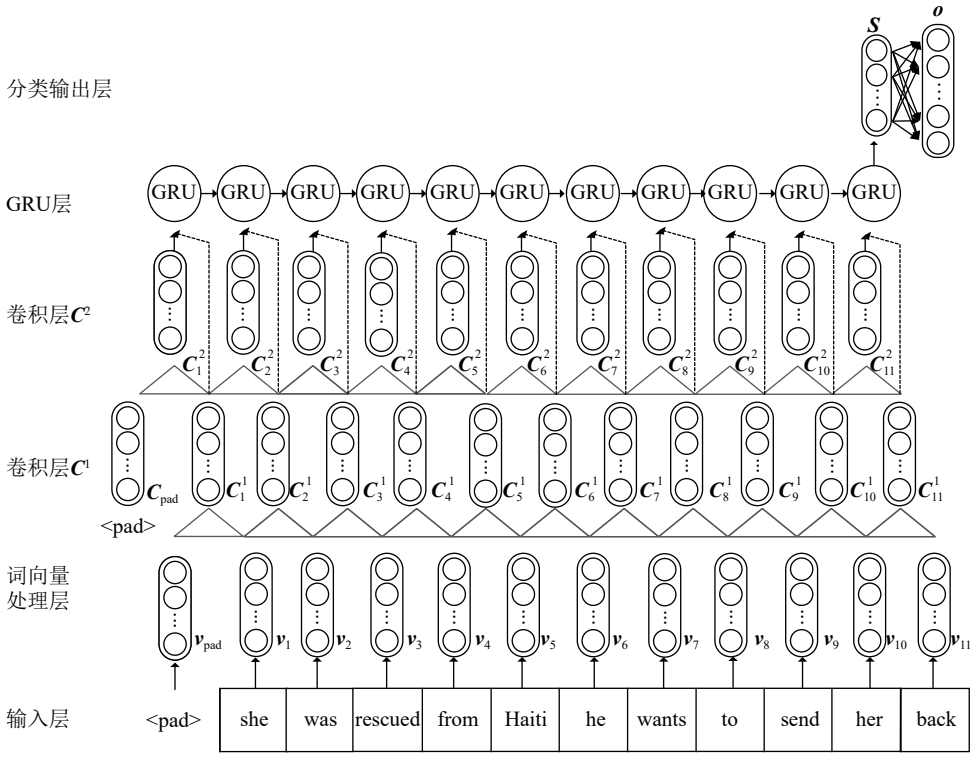

图 1　SPCGRU结构

Fig. 1　SPCGRU architecture

对于n个特征图经过卷积和最大池化得到的结果为：

$$\boldsymbol{P}_j=[p_{1j},p_{2j},\cdots,p_{nj}] \tag{6}$$

若采用最大池化，$\boldsymbol{P}_j$将取代$\boldsymbol{C}_j$作为后续网络输入。

1.2　门阀循环单元

GRU是目前流行的RNN变体之一，Cho等[15]将长短期记忆模型（LSTM）进行部分简化提出了GRU，以处理机器翻译问题。由于GRU相比LSTM存在更少的参数，在收敛时间和迭代次数上都优于LSTM，更适用于样本数较小的文本分类数据。

如图2所示，GRU包含：一个隐含状态$\boldsymbol{h}$、一个候选状态$\tilde{\boldsymbol{h}}$，以及两个门函数，即重置门r（reset gate）与更新门z（update gate）以控制信息通过的比例。

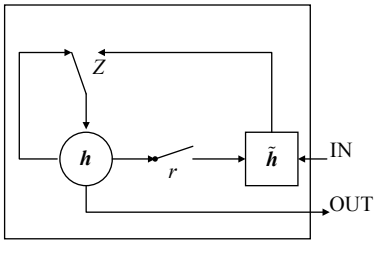

图 2　GRU结构

Fig. 2　GRU architecture

t时刻GRU的计算公式：

$$r_t=\sigma_{\mathrm{g}}(\boldsymbol{W}_{\mathrm{r}}\boldsymbol{x}_t+\boldsymbol{U}_{\mathrm{r}}\boldsymbol{h}_{t-1}) \tag{7}$$

$$z_t=\sigma_{\mathrm{g}}(\boldsymbol{W}_z\boldsymbol{x}_t+\boldsymbol{U}_z\boldsymbol{h}_{t-1}) \tag{8}$$

$$\tilde{\boldsymbol{h}}_t=\tanh(\boldsymbol{W}\boldsymbol{x}_t+\boldsymbol{U}(r_t\odot\boldsymbol{h}_{t-1})) \tag{9}$$

$$\boldsymbol{h}_t=z_t\odot\boldsymbol{h}_{t-1}+(1-z_t)\odot\tilde{\boldsymbol{h}}_t \tag{10}$$

式中：$\boldsymbol{x}_t\in\mathbb{R}^n$为当前时刻GRU的输入；⊙表示对应的元素进行相乘运算；$\boldsymbol{W}_{\mathrm{r}}\in\mathbb{R}^{n\times m}$、$\boldsymbol{W}_z\in\mathbb{R}^{n\times m}$、$\boldsymbol{W}\in\mathbb{R}^{n\times m}$、$\boldsymbol{U}_{\mathrm{r}}\in\mathbb{R}^{n\times m}$、$\boldsymbol{U}_z\in\mathbb{R}^{n\times m}$和$\boldsymbol{U}\in\mathbb{R}^{n\times m}$均是网络的权重参数，$m$为GRU隐含单元个数；$\sigma_{\mathrm{g}}$函数为：

$$\sigma_{\mathrm{g}}(x)=\max\left(0,\min\left(1,\frac{x+1}{2}\right)\right) \tag{11}$$

候选状态$\tilde{\boldsymbol{h}}_t$的计算依赖于上个时刻的隐含状态$\boldsymbol{h}_{t-1}$和当前输入$\boldsymbol{x}_t$。重置门r_t作用于$\boldsymbol{h}_{t-1}$，根据$\boldsymbol{h}_{t-1}$的重要程度，控制过去隐含状态保留程度。r_t越大，$\tilde{\boldsymbol{h}}_t$受$\boldsymbol{h}_{t-1}$的影响程度越大。更新门z_t加权$\tilde{\boldsymbol{h}}_t$与$\boldsymbol{h}_{t-1}$的值，得到当前的隐含状态$\boldsymbol{h}_t$。最后时刻的隐含状态则作为整个句子的特征向量，输入Softmax进行分类。

1.3　附加边缘余弦距离的分类器

Softmax是深度学习中最为常见的分类器之一。尽管Softmax分类器简单易用并且效果出色，但其没能引导网络学习出具有区分性的特征[16]。为了有效学习到类内紧凑、类间离散的特征。Wang等[17]提出了带有附加边缘余弦距离的Softmax（AM–Softmax）分类器。

传统的Softmax函数为：

$$F_{\text{Softmax}}(\boldsymbol{X})=\frac{\mathrm{e}^{\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{X}}}{\sum_{j=1}^{n_{\text{o}}}\mathrm{e}^{\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{X}}}=\frac{\mathrm{e}^{\|\boldsymbol{W}_k\|\|\boldsymbol{X}\|\cos\theta_k}}{\sum_{j=1}^{n_{\text{o}}}\mathrm{e}^{\|\boldsymbol{W}_j\|\|\boldsymbol{X}\|\cos\theta_j}} \tag{12}$$

式中，$\boldsymbol{X}$表示全连接层输入向量，$\boldsymbol{W}_j^{\mathrm{T}}$为关于第$j$个输出节点的全连接层权重，$\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{X}$为分类结果为$k$时对应节点的输出值，$k\in\{1,2,\cdots,n_{\text{o}}\}$。

AM–Softmax函数引入了附加的边缘余弦距离：

$$F_{\text{AM-Softmax}}(\boldsymbol{X})=\frac{\mathrm{e}^{s(\cos\theta_k-d_{\text{m}})}}{\mathrm{e}^{s(\cos\theta_j-d_{\text{m}})}+\sum_{j=1,j\neq k}^{n_{\text{o}}}\mathrm{e}^{s\cos\theta_j}}=\frac{\mathrm{e}^{s(\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{X}-d_{\text{m}})}}{\mathrm{e}^{s(\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{X}-d_{\text{m}})}+\sum_{j=1,j\neq k}^{n_{\text{o}}}\mathrm{e}^{s\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{X}}} \tag{13}$$

式中，s为尺度缩放因子，d_{m}为附加的边缘余弦距离。AM–Softmax函数将$\boldsymbol{W}$与$\boldsymbol{X}$均归一化为1，则$\cos\theta=\frac{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}}{\|\boldsymbol{W}\|\|\boldsymbol{X}\|}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$。

相较于传统的Softmax函数，AM–Softmax函数通过引入附加的边缘余弦距离，获得更大的分类决策边界，使得特征的区分性更强。以二分类问题为例，当$\|\boldsymbol{W}_1\|\|\boldsymbol{X}\|\cos\theta_1>\|\boldsymbol{W}_2\|\|\boldsymbol{X}\|\cos\theta_2$时，Softmax函数将样本分为1类；AM–Softmax要求$\|\boldsymbol{W}_1\|\|\boldsymbol{X}\|(\cos\theta_1-d_{\text{m}})>\|\boldsymbol{W}_2\|\|\boldsymbol{X}\|\cos\theta_2$成立时，才将样本分为1类，加大了对参数学习的约束。

2 短文本特征提取及分类

SPCGRU中短文本特征提取与分类的工作流程如下：

Step 1 初始化参数：句子最大长度$l_{\max}$、词向量维度d、卷积核大小k_{c}、卷积核数目n_1、GRU隐含层节点数目n_2、分类输出层节点数目n_{o}、最大迭代次数$i_{\max}$、每批次数据量m、全局学习率ε、尺度缩放因子s、附加边缘余弦距离d_{m}。

Step 2 词向量层根据输入的文本数据，在预训练的词向量文件中匹配词向量。

Step 3 在第1层卷积中，n_1个卷积核在卷积窗口下对词向量$\boldsymbol{X}_j$进行卷积得到词语的组合特征$\boldsymbol{C}_j^1$：

$$\boldsymbol{C}_j^1=[c_{1j}^1,c_{2j}^1,\cdots,c_{n_1j}^1] \tag{14}$$

式中，c_{ij}为第i个卷积核卷积$\boldsymbol{X}_j$的结果。

Step 4 在第2层卷积中，同样以n_1个卷积核在相同窗口大小下对$\boldsymbol{C}_j^1$进一步卷积得到高层特征$\boldsymbol{C}_j^2$。

Step 5 $\boldsymbol{C}_j^1$与$\boldsymbol{C}_j^2$拼接成新的特征组合，对应$t=j$时刻GRU的输入$\boldsymbol{x}_j$，由式(15)～(16)计算产生GRU隐含状态$\boldsymbol{h}_j$。

$$\boldsymbol{h}_j=f(\boldsymbol{h}_{j-1},\boldsymbol{x}_j;\boldsymbol{\theta}_{\text{g}}) \tag{15}$$

$$\boldsymbol{x}_j=\boldsymbol{C}_j^1\oplus\boldsymbol{C}_j^2 \tag{16}$$

式中，$\boldsymbol{\theta}_{\text{g}}$为GRU的网络参数。

Step 6 将最后时刻GRU的隐含状态作为句子的特征向量$\boldsymbol{S}$，连接一个全连接层使用AM–Softmax函数进行分类，得到样本x分为第k个类别的输出概率：

$$p(y=k|x)=F_{\text{AM-Softmax}}(\boldsymbol{S}) \tag{17}$$

Step 7 计算交叉熵损失函数：

$$J(\boldsymbol{\theta})=-\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{n_{\text{o}}}y\ln p(y=k|x) \tag{18}$$

式中，$\boldsymbol{\theta}$为网络中全部可训练参数向量集，y为样本标签值。

将损失函数作为优化目标，使用RMSprop优化损失函数，基于反向传播调整参数：

$$\boldsymbol{g}\leftarrow\nabla_{\theta}J(\boldsymbol{\theta}) \tag{19}$$

$$G\leftarrow\rho G+(1-\rho)\boldsymbol{g}\odot\boldsymbol{g} \tag{20}$$

$$\Delta\boldsymbol{\theta}=-\frac{\varepsilon}{\sqrt{\delta+G}}\odot\boldsymbol{g} \tag{21}$$

$$\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}+\Delta\boldsymbol{\theta} \tag{22}$$

式中，$\boldsymbol{g}$为梯度项，G代表累计的梯度平方项，ρ为梯度衰减项，δ为防止除0的小常数，ε为全局学习率，$\odot$为按元素相乘。

3 实验结果及分析

3.1 实验数据集

为了验证本文算法的有效性，在TREC、MR以及Subj数据集上进行了大量实验测试。MR、Subj数据集中无标准测试集，为合理评估模型性能，采用十则交叉验证方式进行实验。即将数据集随机分成10等份，每次实验不重复地选取其中1份数据作为测试集评估模型，其余9份数据作为训练集训练模型。数据集详情如表1所示。

表 1 数据集

Tab. 1 Datasets

数据集	分类类别	平均长度	样本总数	测试样本
TREC	6	10	5 952	500
MR	2	20	10 622	CV
Subj	2	23	10 000	CV

注：CV为无标准测试集，采用十则交叉验证评估算法。

TREC数据集：TREC问题类型分类任务，根据问题提问的类型划分为提问与人物相关、地点相关、数目相关等6类。

MR数据集：电影评论情感极性分类任务，将电影评论的情感极性分为正面评价及负面评价两类。

Subj数据集：电影评论情感主观性分类任务，将电影评论按照评价的主观性和客观性分为两类。

3.2 实验环境设置

实验环境为：框架是Tensorflow、Keras；编程语言是Python3.4；处理器是CPU:i5–6500，GPU是1050TI。以Tensorflow作为底层框架，Keras作为顶层框架来实现模型与算法。

3.3 文本数据处理

文本是一种非结构化的数据，要想使文本能够为计算机所识别，就必须将文本数字化表示，即将文本映射到一个d维的实数向量中。这种实数向量又被称为词向量，采用无监督的方式在大规模语料上训练得到。目前，将文本分布式表示为词向量的方法有word2vec[18]、glove[19]。在文本分类任务中，使用预先训练的词向量能够使模型获取一个比较好的初始值，起到改善模型效果的作用。选用由斯坦福大学根据2014年在60亿词汇的维基百科语料库训练得到100维的glove词向量glove.6B.100，作为预训练的词向量[19]。

3.4 超参数设置

卷积核的数目设为512，卷积核的大小设为2，每一批数据为128。模型采用dropout的正则化技巧，以减小训练过拟合的程度。在词向量输入卷积层前和GRU输出到Softmax分类器前，采用比例为0.5的dropout操作。在TREC数据集中，GRU隐含层节点数为100，尺度缩放因子s设置为30，附加边缘余弦距离d_{m}为0.1。在MR数据集中设置GRU隐含层节点数为30，尺度缩放因子s为5，附加边缘余弦距离d_{m}为0.2。Subj数据集中设置GRU隐含层节点数为30，尺度缩放因子s为7，附加边缘余弦距离d_{m}为0.1。尺度缩放因子主要影响模型的收敛快慢，需要根据模型收敛的迭代次数进行调整，尺度缩放因子越大，模型收敛越快。d_{m}、卷积核数、卷积核大小、GRU隐层节点数4个超参数的选取对分类准确率的影响如表2～5所示。

表 2　d_{m}值的选取对分类准确率的影响

Tab. 2　Influence of d_{m} value selection on classification accuracy

d_{m}值	分类准确率/%		
	TREC	MR	Subj
0.1	95.80	82.08	94.85
0.2	95.40	82.33	94.69
0.3	95.60	81.98	94.60
0.4	94.40	82.10	94.62
0.5	95.00	82.10	94.65

表 3　卷积核数对分类准确率的影响

Tab. 3　Influence of the number of convolution filter on classification accuracy

卷积核数	分类准确率/%		
	TREC	MR	Subj
128	95.20	81.90	94.52
256	95.40	81.93	94.60
384	95.40	82.14	94.72
512	95.80	82.33	94.85
640	95.80	82.17	94.72

表 4　卷积核大小对分类准确率的影响

Tab. 4　Influence of the size of convolution filter on classification accuracy

卷积核大小	分类准确率/%		
	TREC	MR	Subj
2	95.80	82.33	94.85
3	95.00	81.80	94.78
4	94.60	81.20	94.55
5	94.60	81.11	94.55

表 5　GRU隐层节点数对分类准确率的影响

Tab. 5　Influence of the number of GRU hidden units on classification accuracy

GRU隐层节点数	分类准确率/%		
	TREC	MR	Subj
30	94.80	82.33	94.85
60	94.80	82.22	94.84
90	95.60	82.23	94.75
100	95.80	82.23	94.80
200	95.60	82.33	94.77
300	95.40	82.25	94.80

d_{m}值的选取对分类准确率的影响如表2所示。附加边缘余弦距离d_{m}是对模型参数学习的约束项，通过实验测定。由表2的实验结果可知，d_{m}的取值不宜太大，在0.1～0.2之间效果较好。

卷积核数对分类准确率的影响如表3所示。由表3可知：卷积核数在一定程度上会影响分类的精确度。卷积核数太小时，分类准确率有所欠缺；而卷积核数超过一定数目后，准确率无明显变化。卷积核数选择512时，分类准确率较为精确。

卷积核大小对分类准确率的影响如表4所示。由表4可知：采用小卷积核模型的分类准确率要高于大卷积核模型的分类准确率。卷积核越大，GRU单次接受的词语个数越多，而对于短文本数据，GRU单次接受的词语个数过多，相邻的GRU输入的相关性就会减弱，不利于序列建模。对于所提出的SPCGRU模型，卷积核的大小为2时，分类准确率最高。

GRU隐层节点数对分类准确率的影响如表5所示。由表5可以看出：对于MR数据集与Subj数据集，隐层节点数的变化对分类准确率的影响不大，其分类准确率的变化维持在0.1%左右。由于MR数据集与Subj数据集为二分类数据集，只需设置隐层节点数为较小的数值，就可获得较高的准确率。而对于TREC多分类数据，隐层节点数太小时，模型表达能力不足，分类准确率较低，将其设为100时分类准确率较高。

3.5 CNN结构设置

图像具有旋转、平移、尺度变换的不变性，因此在图像分类问题中，即使图像经历了一个小的平移、旋转等空间变化之后，依然会产生相同的池化特征，从而最大池化成为CNN在图像分类中常用的降维手段。但文本数据的特征表示不具有图像数据空间上的不变性特点，池化窗口往往对局部的文本空间信息造成破坏，并且在短文本数据中，可利用的文本信息有限，最大池化在降采样的过程中会使文本信息发生丢失。

池化对文本分类准确率的影响如图3所示。

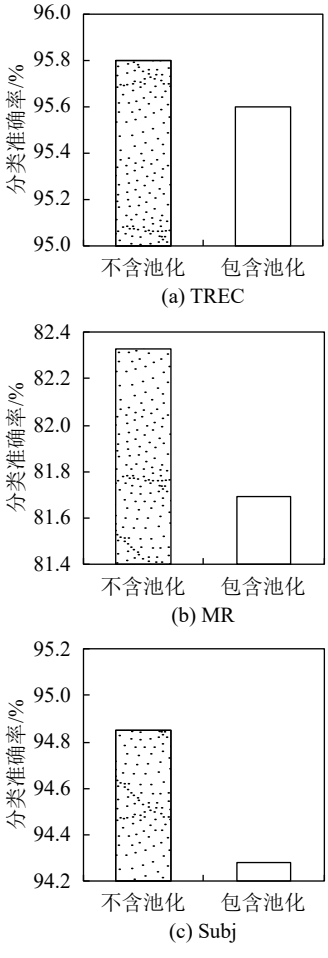

图 3 池化对分类准确率的影响

Fig. 3 Effect of pooling on classification accuracy

由图3可知，在3个短文本数据集中，原有卷积网络中去掉池化后的模型的分类效果均有不同程度的提升。在所提出的SPCGRU网络中，文本全局的时序建模由GRU完成，在卷积层中抛弃池化操作，使文本在输入GRU前保持位置信息和序列结构，有利于提高分类准确率。

串并行卷积层对词语局部的组合特征进行多尺度的融合。卷积层$\boldsymbol{C}^1$、$\boldsymbol{C}^2$均采用小步幅的卷积窗口，串行的两个小步幅的滑动窗口连续卷积所接受到词语范围等同于一个大步幅的滑动窗口卷积的词语范围。例如，作者采用连续两个2步幅的卷积窗口与一个3步幅卷积窗口所接受的词语范围相同，均为3。但串行的双层卷积相比单层卷积增加了更多的非线性。每一层卷积提取到的特征均有助于文本词语组合的特征学习。将第1层卷积$\boldsymbol{C}^1$的结果和第2层卷积$\boldsymbol{C}^2$的结果并行输出，拼接而成新的特征作为GRU的输入。使用串并行的卷积结构在数据集上的分类正确率如图4所示。

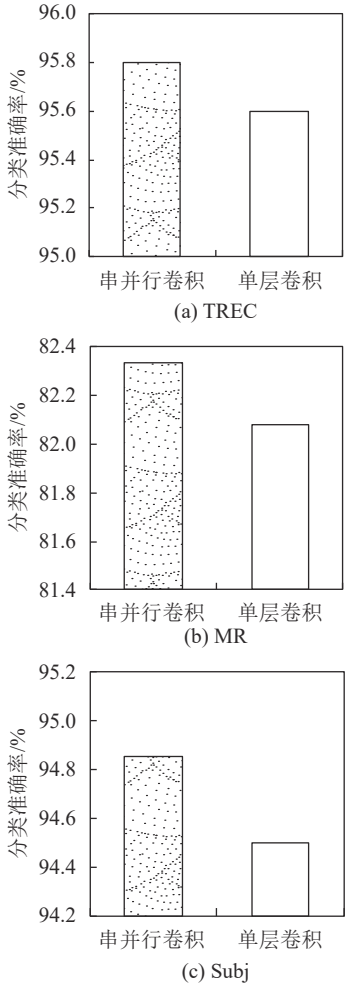

图 4 串并行结构对分类准确率的影响

Fig. 4 Effect of serial–parallel structure on classification accuracy

从图4可知，采用串并行的卷积结构的模型的分类准确率比普通的单层卷积模型的分类准确率更高。关注C^1层低尺度特征的学习能辅助C^2层的特征学习效果。多尺度串并行的卷积结构取低层特征在高层并行化输出，使得提取到的词语特征组合更为丰富，能够更好地表达文本的内容信息，充分融合文本在不同层次上的多元特征，因此有利于提高分类准确率，也体现了所提出方法的有效性。

3.6　与其他方法对比

所提出方法与其他方法的对比结果如表6所示。表6中，SPCGRU–S和SPCGRU–AM–S分别代表采用Softmax和AM–Softmax函数进行分类的SPCGRU模型。G–Dropout[20]、F–Dropout[20]、Text–CNN[6]、CNN、DCNN[21]模型都是非时序的文本建模方法，没有考虑文本中时序信息对特征表达的影响。GRU模型仅仅将单个字符作为模型每个单元的输入，没有计算多个字符组成短语的情形。P–LSTM[9]虽然通过滑动窗口将多个词语混合输入LSTM中，但缺少对词语特征非线性组合。

表 6　与其他方法的对比结果

Tab. 6　Comparison results with other methods

分类方法	分类准确率/%		
	TREC	MR	Subj
G–Dropout[20]	—	79.00	93.40
F–Dropout[20]	—	79.10	93.60
P–LSTM[9]	—	80.17	93.77
DCNN[21]	93.00	—	—
Text–CNN[6]	93.60	81.50	94.40
CNN	94.20	80.76	94.05
GRU	93.80	81.10	93.77
SPCGRU–S	95.40	82.08	94.77
SPCGRU–AM–S	95.80	82.33	94.85

从表6可以看出：采用串并行卷积层的SPCGRU与其他方法相比仍然保持最高的分类准确率。在Softmax分类器中引入边缘余弦距离，加大学习参数的约束，得到的SPCGRU–AM–S的分类效果进一步提升。在3个数据集中，SPCGRU–AM–S模型的分类准确率比相同参数规模的GRU模型的分类准确率分别提升了2.00%、1.23%、1.08%，比相同参数规模的CNN模型的分类准确率分别提升了1.60%、1.57%、0.80%。相较于G–Dropout、F–Dropout、Text–CNN、DCNN模型而言，SPCGRU采用时序结构作为文本表达的方式，更加符合文本建模的特点，具有更好的特征提取能力，能够获取上下文语义的知识。与P–LSTM、GRU模型相比，SPCGRU更能充分提取文本不同层次的非线性特征的多元组合，取得更好的分类效果。

4　结　论

结合CNN与GRU提出基于串并行卷积门阀循环神经网络的短文本分类方法。由于短文本数据不具备空间上的不变性特征导致对最大池化操作比较敏感，最大池化在降采样过程中会导致信息丢失，并对局部文本的空间结构造成破坏，从而不利于提升分类准确率。作者提出的串并行卷积门阀循环神经网络通过引入不含池化结构下的串并行卷积结构提取词语之间的多尺度组合特征。所建的串并行卷积结构中，串行结构双层卷积增加了模型非线性表达能力；并行结构将低层特征并行化输出，辅助高层特征学习，从而丰富了GRU的特征输入。同时，在Softmax分类器中引入边缘余弦距离得到AM–Softmax分类器，使学习到的特征在类内紧凑、类间离散，能够进一步提升分类的效果。在后续研究中，还会继续优化和改进所提出模型的算法、结构与参数设置，以进一步提高模型对文本识别能力。

参考文献：

[1] Diab D M,El Hindi K M.Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification[J].Applied Soft Computing,2017,54:183–199.

[2] Zhang Wen,Tang Xijin,Yoshida T.TESC:An approach to TExt classification using semi-supervised clustering[J].Knowledge-Based Systems,2015,75:152–160.

[3] Vieira A S,Borrajo L,Iglesias E L.Improving the text classification using clustering and a novel HMM to reduce the dimensionality[J].Computer Methods and Programs in Biomedicine,2016,136:119–130.

[4] Wang Yisen,Xia Shutao T,Wu Jia.A less-greedy two-term Tsallis entropy information metric approach for decision tree classification[J].Knowledge-Based Systems,2017,120:34–42.

[5] de Boom C,van Canneyt S,Demeester T,et al.Representation learning for very short texts using weighted word embedding aggregation[J].Pattern Recognition Letters,2016,80:150–156.

[6] Kim Y.Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on empirical Methods In Natural Language Processing.Doha:ACL,2014:1532–1543.

[7] Sun Songtao,He Yanxiang.Multi-label emotion classification for microblog based on CNN feature space[J].Advanced Engineering Sciences,2017,49(3):162–169.[孙松涛,何炎祥.基于CNN特征空间的微博多标签情感分类

[J].工程科学与技术,2017,49(3):162–169.]

[8] Er M J,Zhang Yong,Wang Ning,et al.Attention pooling-based convolutional neural network for sentence modelling[J].Information Sciences,2016,373:388–403.

[9] Lu Chi,Huang Heyan,Jian Ping,et al.A P–LSTM neural network for sentiment classification[M]//Advances in Knowledge Discovery and Data Mining.Cham:Springer,2017:524–533.

[10] Zhou Yujin,Xu Bo,Xu Jiaming,et al.Compositional recurrent neural networks for Chinese short text classification[C]//Proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI).Omaha:IEEE,2016:137–144.

[11] Yang Zizhao,Yang Diyi,Dyer C,et al.Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics.San Diego:NAACL,2016:1480–1489.

[12] Wang Xinyou,Jiang Weijie,Luo Zhiyong.Combination of convolutional and recurrent neural network for sentiment analysis of short texts[C]//Proceedings of the 26th International Conference on Computational Linguistics.Osaka:ACM,2016:2428–2437.

[13] Xie Jinbao,Hou Yongjin,Kang Shouqiang,et al.Multi-feature fusion based on semantic understanding attention neural network for Chinese text categorization[J].Journal of Electronics & Information Technology,2018,40(5):1258–1265.[谢金宝,侯永进,康守强,等.基于语义理解注意力神经网络的多元特征融合中文文本分类[J].电子与信息学报,2018,40(5):1258–1265.]

[14] Li Linchuan,Wu Zhiyong,Xu Mingxing,et al.Combining CNN and BLSTM to extract textual and acoustic features for recognizing stances in mandarin ideological debate competition[C]//Proceedings of the 17th Annual Conference of the International Speech Communication Association.San Francisco:ISCA,2016:1392–1396.

[15] Cho K,van Merriënboer B,Gulcehre C,et al.Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.Doha:ACL,2014:1724–1734.

[16] Liu Weiyang,Wen Yandong,Yu Zhiding,et al.Large-margin softmax loss for convolutional neural networks[C]//Proceedings of the 33th International Conference on Machine Learning.New York:ACM,2016:507–516.

[17] Wang Feng,Cheng Jian,Liu Weiyang,et al.Additive margin softmax for face verification[J].IEEE Signal Processing Letters,2018,25(7):926–930.

[18] Zhang Dongwen,Xu Hua,Su Zengcai,et al.Chinese comments sentiment classification based on word2vec and SVM^{perf}[J].Expert Systems with Applications,2015,42(4):1857–1863.

[19] Pennington J,Socher R,Manning C D.GloVe:Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.Doha:ACL,2014:1532–1543.

[20] Wang S I,Manning C D.Fast dropout training[C]//Proceedings of the 30th International Conference on Machine Learning.Atlanta:ACM,2013:118–126

[21] Kalchbrenner N,Grefenstette E,Blunsom P,et al.A Convolutional neural network for modelling sentences[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.Baltimore:ACL,2014:655–655.

（编辑　赵　婧）

引用格式: Tang Xianlun,Lin Wenxing,Du Yiming,et al.Short text feature extraction and classification based on serial–parallel convolutional gated recurrent neural network[J].Advanced Engineering Sciences,2019,51(4):125–132.[唐贤伦,林文星,杜一铭,等.基于串并行卷积门阀循环神经网络的短文本特征提取与分类[J].工程科学与技术,2019,51(4):125–132.]