文章编号:1007－6735(2023)02－0189－09　　DOI: 10.13255/j.cnki.jusst.20211108001

基于预训练模型融合深层特征词向量的中文文本分类

汤英杰，　刘媛华
（上海理工大学 管理学院，上海 200093）

摘要：为解决传统模型表示出的词向量存在序列、上下文、语法、语义以及深层次的信息表示不明的情况，提出一种基于预训练模型(Roberta)融合深层特征词向量的深度神经网络模型，处理中文文本分类的问题。通过 Roberta 模型生成含有上下文语义、语法信息的句子向量和含有句子结构特征的词向量，使用 DPCNN 模型和改进门控模型(RGRU)对词向量进行特征提取和融合，得到含有深层结构和局部信息的特征词向量，将句子向量与特征词向量融合在一起得到新向量。最后，新向量经过 softmax 激活层后，输出结果。在实验结果中，以 *F*1 值、准确率、召回率为评价标准，在 THUCNews 长文本中，这些指标分别达到了 98.41%，98.44%，98.41%。同时，该模型在短文本分类中也取得了很好的成绩。

关键词：预训练模型；Roberta 模型；DPCNN 模型；特征词向量；中文文本分类
中图分类号：TP 391.1　　**文献标志码：**A

Chinese text classification based on pre-training model and deep feature word vector

TANG Yingjie,　LIU Yuanhua
(*Business School, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: A deep neural network model based on pre-training model (Roberta) and deep feature word vector was proposed to deal with the problem of Chinese text classification, in order to solve the problem that the word vector represented by the traditional model has unclear sequence, context, grammar, semantics and deep information representation. The sentence vector containing context semantics and grammar information and the word vector containing sentence structure features were generated by Roberta model. The word vector was extracted and fused by DPCNN model and revised gate recurrent unit(RGRU) to obtain the feature word vector containing deep structure and local information. The sentence vector and feature word vector were fused together to obtain a new vector. Finally, after the new vector passed through the softmax activation layer, the result was output. In the

收稿日期: 2021−11−08
基金项目: 国家自然科学基金资助项目(71771152)
第一作者: 汤英杰(1994−)，男，硕士研究生．研究方向：自然语言处理、文本挖掘. E-mail：1143195254@qq.com
通信作者: 刘媛华(1974−)，女，副教授. 研究方向：系统工程、经济与管理统计、复杂系统理论方法及应用、社会经济复杂系统．E-mail：liuyuanhua@usst.edu.cn

experimental results, $F1$ value, accuracy and recall were chosen as the evaluation criteria, they reached 98.41%, 98.44% and 98.41% in the long text of THUCNews. At the same time, the model had also achieved good results in short text classification.

Keywords: *pre-training model*; *Roberta model*; *DPCNN model*; *feature word vector*; *Chinese text classification*

随着移动互联网技术的飞速发展，以及社交平台、购物平台的不断涌现，人们畅游在网络世界中，在享受高度便利和快捷生活的同时，海量信息也随之充斥网络，让人们难辨真伪和善恶。对网络信息进行正确的文本分类，可以有效降低互联网舆论中负面的影响，如：造谣、诋毁、恶意中伤等事件。同时，正确的文本分类，可以建立起智能信息推荐系统，根据用户的个人兴趣来定位并推荐相关的新闻资料、商品信息等；也可以建立垃圾信息过滤系统，减少生活中琐碎、烦心事件，极大地简便公众的生活。

文本分类的方法包括使用传统的机器学习方法和深度神经网络构建模型的方法。使用机器学习进行文本分类时经常会提取TF-IDF(term frequency-inverse document frequency)或者词袋结构，然后对模型进行训练，如支持向量机[1]、逻辑回归、XGBoost[2]等。利用传统机器学习方法进行文本分类的基本流程是获取数据、数据预处理、特征提取、模型训练、预测。TF-IDF和词袋模型都需要手动去构建词典，统计词汇，进而计算出相关顺序(使用欧式距离或夹角余弦相似度)。这两种方法都存在较大的缺陷，如计算繁琐、可解释性差、语义不明等。

软件、硬件技术的快速发展，使得文本分类问题开始从传统的机器学习转移到深度学习，词向量Word2vec[3]的发展，推动了深度学习模型在自然语言处理中的应用。深度学习模型包括卷积神经网络(convolution neural network, CNN)、循环神经网络(recurrent neural networks, RNN)、图神经网络(graph convolution network, GCN)等方法。

针对当前使用神经网络处理文本分类的问题，本文提出了预训练Roberta[4]模型，对输入数据的随机掩码和双向动态的向量表示方法进行训练，加强了向量表示的灵活性，实现了数据增强。利用DPCNN(deep pyramid convolutional neural networks for text categorization)和改进门控网络提取深层词向量的特征，强化了有效信息，降低了无效信息和梯度消失的影响。运用注意力机制的方法融合句向量与深层词向量，增强了文本向量的语义丰富性，捕捉重要词与句之间的潜在语义关系，有效丰富了特征向量中的结构、语义和语法信息。

1 相关工作

目前的文本分类深度学方法主要包括两种，分别为基于卷积神经网络、循环神经网络和图神经网络进行改进的神经网络，以及基于预训练的神经网络。

1.1 基于CNN、RNN和GNN进行改进的神经网络

文本分类模型使用较多的是TextCNN(text convolutional neural network)，该模型由Kim等[5]提出，第一次将卷积神经网络用于自然语言处理的任务中。TextCNN通过一维卷积来获取句子中n-gram的特征表示，对文本抽取浅层特征的能力很强。在长文本领域，TextCNN主要靠filter窗口抽取特征，但信息抽取能力较差，且对语序不敏感。文献[6]通过采用多个滤波器构建多通道的TextCNN网络结构，从多方面提取数据的特征，捕捉到了更多隐藏的文本信息。文献[7]提出图卷积神经网络对文本内容进行编码，文献[8]使用了异构图注意网络进一步提升了模型的编码能力。

文献[9]提出了循环神经网络的文本分类模型，但RNN结构是一个串行结构，对长距离单词之间的语义学习能力差，同时可能伴随有梯度消失和梯度爆炸的问题。随后，LSTM(long short term memory, LSTM)和GRU(gate recurrent unit, GRU)模型被应用在自然语言处理任务中，LSTM由输入门、输出门和遗忘门控制每个时间点的输入、输出和遗忘的概率，有效缓解了梯度消失和爆炸问题。文献[10]中，提出了ONLSTM(ordered neurons long-short memory)结构，在LSTM结构中引入层级结构，可以提取出文本的层级信息。GRU通过

将输入门和遗忘门组合在一起，命名为更新门，减少了门的数量，在保证记忆的同时，提升了网络的训练效率。文献 [11] 用 BiGRU 模型进行文本情感分类任务，提出了使用 BiGRU 模型对文本进行情感分析。

1.2　预训练神经网络

2018 年，谷歌团队提出了 transformer 模型[12]，并采用了 self-attention 机制[13]。相比于循环神经网络模型，transformer 模型是并行结构，其运算速度得到了大大的提高。Transformer 模型由 encoder 模块和 decoder 模块两部分组成，decoder 模块与 encoder 类似，只是在 encoder 中 self-attention 的 query，key，value 都对应了源端序列，decoder 中 self-attention 的 query，key，value 都对应了目标端序列。注意力机制开始被应用于图像处理上，Bahdanau 等[14] 首次将其应用在了 NLP（自然语言处理）任务中，NLP 领域也迎来了巨大的飞跃。文献 [15] 针对文本分类任务提出了基于词性的自注意力机制网络模型，使用自注意力机制学习出特征向量表示，并融合词性信息完成分类任务。

在 Transformer 模型和注意力机制的基础上，Devlin 等[16] 提出了预训练 Bert 模型（bidirectional encoder representation from transformers），开启了预训练神经网络的时代。Bert 的新语言表示模型代表了 Transformer 的双向编码器，从而生成了文本的双向动态句子向量。孙红等[17] 基于 Bert+GRU 的网络结构对新闻文本进行分类，运用 Bert 得到特征词向量，利用 GRU 网络作为主题网络提取上下文的文本特征。文献 [18] 提出了一种基于 Bert 的构建双通道网络模型的文本分类任务，提升了混合语言文本分类模型的性能。

尽管上述研究证明了对文本进行特征提取和融合之后，可以为分类器提供足够的信息，提高了文本分类问题的准确率。但是，如何对句与词之间的结构、语义和语法等信息进行提取，未作出明确的说明和研究，这也是本文所关注的重点。简而言之，在提取文本信息时，既要提取出文本主要信息，同时也需要注重词与句之间内容关系的提取。

在此基础上，本文利用 Roberta 模型的强大功能，训练出含有上下文语义、语法信息的句子向量和含有句子结构特征的词向量。并分别利用 DPCNN 网络和改进门控网络 (RGRU)，对词向量进行特征提取，使用注意力机制将两部分输出的词向量进行融合，得到深层特征词向量。其中，DPCNN 的主要作用是负责强化局部上下文的关系，RGRU 负责词与词之间的时序关系，注意力机制对局部上下文关系和时序关系进行通盘考虑，使用注意力机制也能够更好地将特征中的重点表现出来。最后，将词向量与句向量相融合来提升模型的性能。

2　模型设计

2.1　模型结构

本文提出的模型应用于中文文本分类任务，模型结构图 1 主要由 3 个部分组成：a. Roberta 模型对输入的中文文本进行预训练，得到含有上下文语义、语法信息的句子向量和词向量；b. 将词向量分别输入至 DPCNN 特征提取层和改进门控神经网络中，然后使用注意力机制将两部分的特征词向量相融合，得到含有深层结构和局部信息的特征词向量；c. 将句子向量与词向量进行融合，得到最终的文本向量表示，最后经过 softmax 激活层后，输出结果。

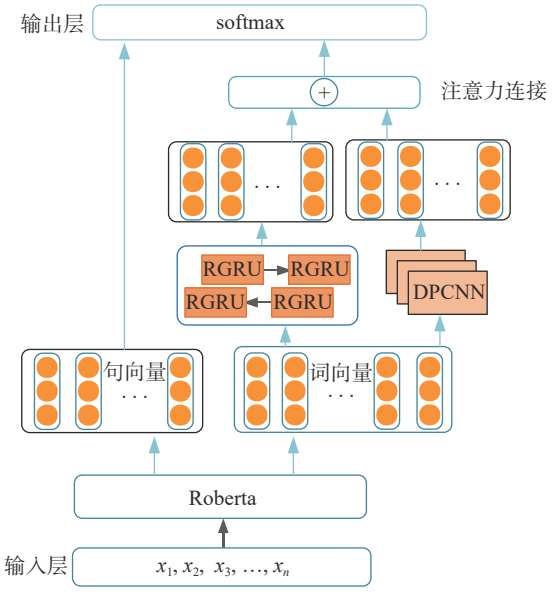

图 1　模型结构图

Fig.1　Model structure diagram

2.2　预训练模型 Roberta

Bert 预训练模型具有以下 3 方面优势：参数规模大、通用能力强、综合性能好。预训练模型中包含着丰富的文本信息知识，因此，近些年的文本分类任务中通常会使用 Bert 行文本特征提取。但是，Bert 的预训练阶段并没有使用全词覆盖的

方式，mask(掩码)字符不利于文本信息的提取，且使用 NSP 任务也会损害 Bert 的特征提取能力。为避免这些问题，本文使用了 Roberta 模型。同时，由于 Roberta 相较于 Bert 使用了更大规模的数据集，使得模型消耗的资源增加，训练时间增长。

Roberta 模型结构不仅继承了 Bert 的双向编码器表示，而且将输入的句子表示为字向量、句向量、位置向量三者之和，经过多层双向 Transformer 编码器(见图 2)得到文本的向量化表示。图中：Add 表示残差连接；Norm 表示层标准化；Feed Forward 表示前向传播；*Nx* 表示 *N* 个堆叠的相同 *x*。

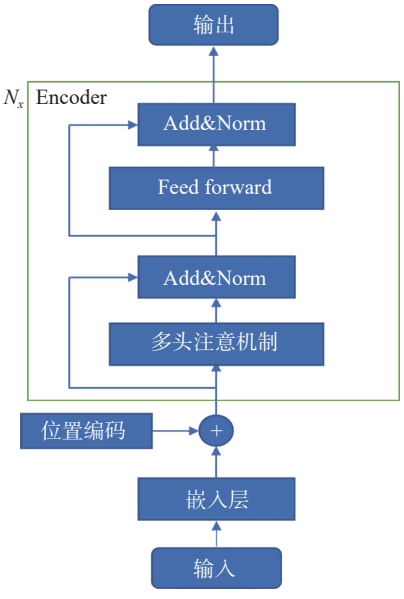

图 2 Transformer 编码器

Fig.2 Transformer encoder

多头注意力机制：假设输入句子为 $\boldsymbol{X}$，$\boldsymbol{X}=[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n]$，$n$表示样本句子中字的个数，对字使用 one-hot 编码表示，其维度为k，则$\boldsymbol{X}$所对应的字嵌入矩阵为$\boldsymbol{Y}=[\boldsymbol{y}_1\ \boldsymbol{y}_2\ \cdots\ \boldsymbol{y}_n]$，$\boldsymbol{x}_i$所对应的向量表示为$\boldsymbol{y}_i$。通过训练模型可得出$\boldsymbol{Q}$(Query)、$\boldsymbol{K}$(Key)、$\boldsymbol{V}$(Value)矩阵，$d_k$表示$\boldsymbol{K}$中列向量的维度大小，从而计算得到注意力值为

$$\boldsymbol{A}_{\text{attention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

多头注意力机制则是通过h个不同的$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$线性变换对进行投影，最后将不同的$\boldsymbol{A}_{\text{attention}}$结果拼接起来得到多头注意力值为

$$\boldsymbol{M}_{\text{multihead}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\boldsymbol{C}_{\text{concat}}(h_1,\ldots,h_h)\boldsymbol{W}^{o} \tag{2}$$

$$h_i=\boldsymbol{A}_{\text{attention}}(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{K}_i^{\boldsymbol{V}}) \tag{3}$$

式中：$\boldsymbol{W}_i^{\boldsymbol{Q}}$，$\boldsymbol{W}_i^{\boldsymbol{K}}$，$\boldsymbol{W}_i^{\boldsymbol{V}}$分别表示第$i$个$h$的$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$权重矩阵；$\boldsymbol{W}^{o}$表示附加权重矩阵；$\boldsymbol{C}_{\text{concat}}(\cdot)$表示连接函数。

Bert 模型中掩码 mask 是静态的，即 Bert 在准备训练数据时，只会对每个样本进行一次随机的 mask(在后续训练中，每个 epoch(训练数据)是相同的)，后续的每个训练步都采用同样的 mask。Roberta 模型相比于 Bert，建立在 Bert 的语言掩蔽策略的基础上，将静态 mask 修改为动态 mask，对数据进行预处理时会对原始数据拷贝 10 份，每一份都随机选择 15% 的 Tokens(字符)进行 mask，图 3 为 Roberta 掩码方式。

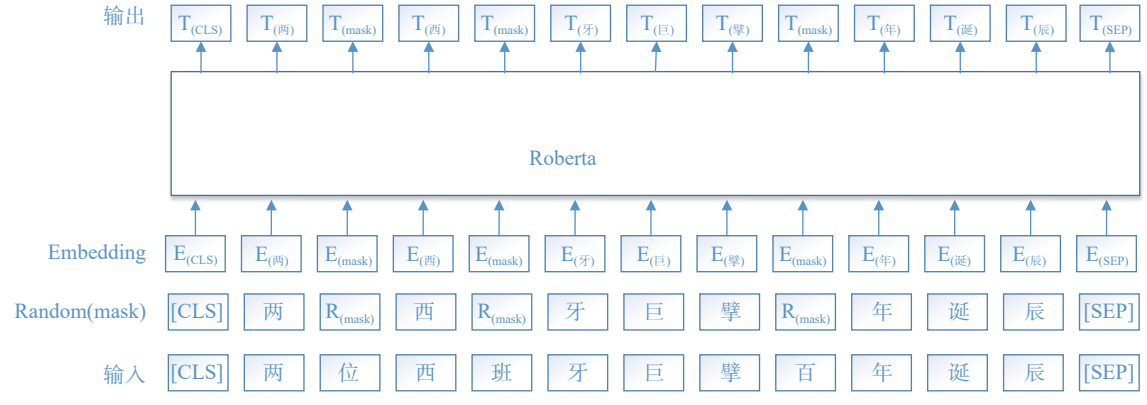

图 3 Roberta 掩码方式

Fig.3 Roberta masking method

同时，Roberta 取消了 Bert 的 NSP(next sentence prediction) 任务，采用了更大规模的数据集进行训练，更好地表现出了词的语义和语法信息，文本向量表示更加完善。Roberta 也修改了 Bert 中的关键超参数，使用更大的 batch 方式和学习率进行训练，增长了训练序列，使得 Roberta 表示能够比

Bert 更好地推广到下游任务中。

2.3　DPCNN 特征提取层

DPCNN[19] 模型相比于 TextCNN模型是更为有效广泛的深层卷积模型，如图 4 所示。图中：$\sigma(\cdot)$为逐分量非线性激活函数；权重 W 和偏差 b(每层唯一)为所要训练的参数。

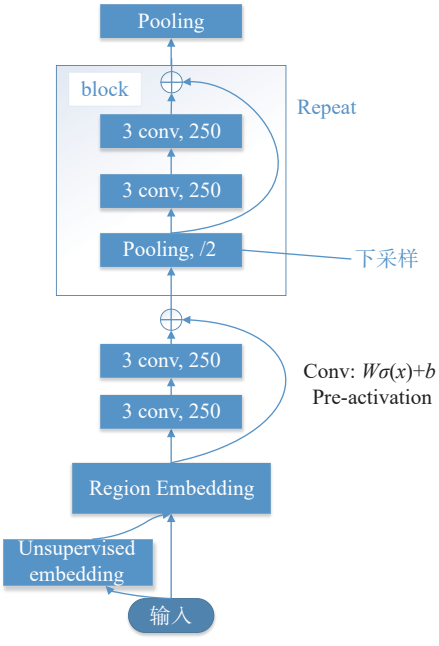

图 4　DPCNN 结构

Fig.4　DPCNN structure

DPCNN 的底层为 Region embedding 层，该层由多个不同大小的卷积核组成，经卷积操作后生成 embedding，作为模型的嵌入层。本文使用两层等长卷积层来捕获长距离模式，提高对词位 embedding 表示的丰富度。

下采样的操作采用固定数量的滤波器，通过最大池化的方法，将原词向量的长度减少一半，计算复杂度也相对减少，但其中包含的文本内容却得到了加长。然后进行两层等长卷积，这两部分组合成 block 模块，重复 block 模块的操作，直至满足任务。随着模型深度的变化，词向量中的深层结构信息和全局语义信息会不断得到加强。

为了解决卷积过程中的梯度消失和爆炸问题，模型在 block 模块进行前与 region embedding 使用 pre-activation 策略进行残差连接，或者直接连接到最后的输出层，有效缓解了梯度问题。模型随着序列长度的加深呈现出深层次的金字塔结构。

2.4　改进门控网络

改进门控模型结构见图 5，对于t时刻而言，输入为$\boldsymbol{q}_t$，隐藏层输入为$\boldsymbol{n}_{t-1}$，隐藏层输出为$\boldsymbol{n}_t$，计算过程如式(4)～(7)所示。

$$\boldsymbol{r}_t = \Phi(\boldsymbol{M}_r[\boldsymbol{n}_{t-1}, \boldsymbol{q}_t] + \boldsymbol{b}_r) \tag{4}$$

$$\boldsymbol{Z}_t = \Phi(\boldsymbol{M}_Z[\boldsymbol{n}_{t-1}, \boldsymbol{q}_t] + \boldsymbol{b}_Z) \tag{5}$$

$$\tilde{\boldsymbol{n}}_t = \Phi(\boldsymbol{M}_n[\boldsymbol{r}_t \otimes \boldsymbol{n}_{t-1}, \boldsymbol{q}_t] + \boldsymbol{b}_n) \tag{6}$$

$$\boldsymbol{n}_t = (1 - \boldsymbol{Z}_t) \otimes \boldsymbol{n}_{t-1} + \boldsymbol{Z}_t \otimes \tilde{\boldsymbol{n}}_t \tag{7}$$

式中：$\boldsymbol{M}_r$，$\boldsymbol{M}_Z$，$\boldsymbol{M}_n$分别表示 3 个门控r，Z，n的参数矩阵；$\boldsymbol{b}_r$，$\boldsymbol{b}_Z$，$\boldsymbol{b}_n$分别表示r，Z，n的偏置项；$\boldsymbol{r}_t$为t时刻重置门输出值；$\boldsymbol{Z}_t$为t时刻更新门输出值；$\tilde{\boldsymbol{n}}_t$为t时刻候选隐藏状态。

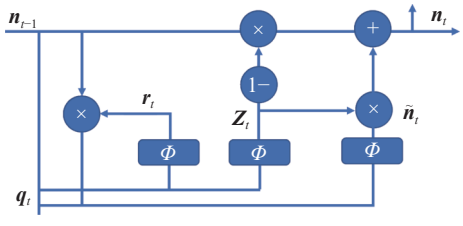

图 5　改进门控网络

Fig.5　Improved GRU

传统的门控神经网络中重置门和更新门都是使用的$\sigma(\sigma = 1/1 + e^{-g})$激活函数，$\sigma$函数存在以下两个缺点：a. 容易出现梯度消失的现象，当激活函数接近饱和区时，变化太缓慢，导数接近 0，从而无法完成深层网络的训练；b. σ的输出不是 0 均值，这会导致后层的神经元的输入是非 0 均值的信号，会对梯度产生影响，导致收敛变慢。本文的Φ（$\Phi = e^{g} - e^{-g}/e^{g} + e^{-g}$）激活函数具有以下 3 个优点：a. 解决了上述$\sigma$函数输出不是 0 均值的问题；b. Φ函数的导数取值范围在 0～1 之间，优于σ函数的 0～0.25，一定程度上缓解了梯度消失的问题；c. Φ函数在原点附近与$y = x$函数形式相近，当输入的激活值较低时，可以直接进行矩阵运算，训练相对容易。

3　实　验

3.1　实验环境

CPU 6× Xeon E5-2678 v3，内存 62 G，显存 11 G，NVIDIA GeForce RTX 2080 Ti，操作系统为 Windows10 64 位，python 版本为 3.8，深度学习框

架为 PyTorch。

3.2 实验数据集

本实验采用网上公开的清华 THUCNews 文本分类数据集中的短、长文本数据集，用于预测模型的性能。选取 THUCNews 数据集中的 10 个类别进行测试，短文本的类别包括：体育、娱乐、房产、教育、时政、游戏、社会、科技、股票、金融；长文本的类别包括：体育、娱乐、家居、房产、教育、时尚、时政、游戏、科学、金融。实验数据集信息如表 1 所示。

表 1 实验数据集信息

Tab.1 Information of experimental data set

数据集	类别	样本数量	训练集	测试集	验证集	平均词长	训练最大长度
THUCNews短文本	10	100000	80000	10000	10000	42	32
THUCNews长文本	10	65000	50000	10000	5000	880	150

3.3 参数设置

文献 [20] 在使用 Bert 作文本分类时给出了 fine-tune 建议。多相关任务的前提下，选择多任务学习进行 Bert fine-tune，目标任务的实现需要考虑文本的预处理、图层选择和学习率。

$$\theta_j^l = \theta_{j-1}^l - \alpha^l \nabla_{\theta^l} J(\theta), j = 1,2,\cdots,k \tag{8}$$

式中：l表示层数；$J(\theta)$表示 softmax 损失函数；j表示迭代次数；k表示迭代的总次数；α为学习率。

$$\alpha^k = \beta\alpha^{k-1} \tag{9}$$

式中，β为学习率衰减系数。

进行学习率衰减，$\beta = 0.95$时模型效果最佳。Roberta 模型只需要一个较小的学习率，同时使用 warm-up 策略，有助于缓解 mini-batch 的提前过拟合现象，保持分布的平稳，同时也有助于保证模型深层的稳定性。以 Adam 算法为基础，采用手动阶梯式衰减、lambda 自定义衰减、三段式衰减和余弦式调整的 4 种方法（见图 6），调整学习率。

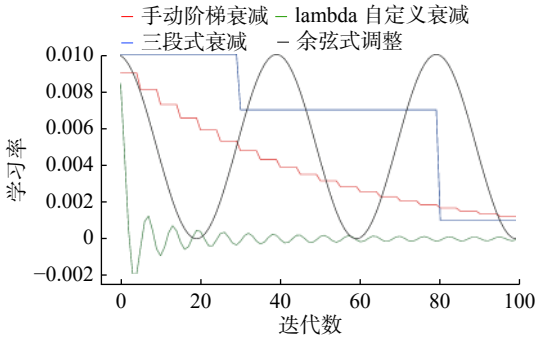

图 6 学习率衰减策略

Fig.6 Learning rate decay strategy

宋明等[21] 在 Bert 作文本分类时，运用 Focal Loss[22] 作为损失函数，提高了模型对困难文本分类的准确率，本文采取 Focal Loss 作为损失函数。

本文中 Roberta 模型的学习率为1.0×10^{-5}，但是在 DPCNN 的结构中需要一个较大的学习率，取 0.001。THUCNews 长文本中句子长度取 150，batch size 取 32；THUCNews 短文本中句子长度取 38，batch size 为 128。DPCNN 结构中，等长卷积 kernel size 为 3。

3.4 评价标准

将准确率 (accuracy)、精确率 (precision)、召回率（recall）和 $F1$ 值作为实验的评价标准，相关的混淆矩阵结构如表 2 所示。

表 2 混淆矩阵

Tab.2 Confusion matrix

预测结果	positive	negative
true	true positive（TP）	true negative（TN）
false	false positive（FP）	false negative（FN）

准确率计算公式为

$$E_{\text{accuracy}} = \frac{H_{\text{TP}} + H_{\text{TN}}}{H_{\text{TP}} + H_{\text{TN}} + H_{\text{FP}} + H_{\text{FN}}} \tag{10}$$

精确率计算公式为

$$P_{\text{precision}} = \frac{H_{\text{TP}}}{H_{\text{TP}} + H_{\text{FP}}} \tag{11}$$

召回率计算公式为

$$R_{\text{recall}} = \frac{H_{\text{TP}}}{H_{\text{TP}} + H_{\text{FN}}} \tag{12}$$

$F1$ 值的计算公式为

$$F1 = \frac{2PR}{P+R} \tag{13}$$

式中，H 表示混淆矩阵各值。

3.5 实验结果

为验证本文所提模型的合理性和有效性，采

用了 8 种模型在两个数据集上进行测试，最后的结果也表现出本文所提出的模型效果优于其他 7 种模型。

a. FastText[23]。Facebook 在 2016 年发布了这种简单快速实现文本分类的方法。FastText 会自己训练词向量，同时采用层次化 softmax 和*n*-gram 让模型学习到局部单词顺序的部分信息。

b. TextCNN。采用多通道 CNN 结构，经过嵌入层后词向量维度为 300，经过卷积核尺寸分别为 2，3，4，通道数为 256 的卷积层后，将输出的 3 个词向量拼接在一起，经过全连接层和 softmax 激活函数后输出结果。

c. LSTM。LSTM 的结构为 2 层全连接层，隐藏层中神经元的个数为 128，方向为双向；LSTM 输出的词向量经过全连接层和 softmax 激活函数后输出结果。

d. DPCNN。深层金字塔卷积结构，采用图 4 中的结构设置。

e. Bert+DPCNN。采用谷歌提供的 Bert 模型作为预训练模型，下游任务连接 DPCNN 网络结构，参数设置与 Roberta+DPCNN 网络结构一样。

f. Roberta+LSTM。将 Roberta 模型中 encoder 层的输出作为 LSTM 模型的输入，得到输出，将此输出与 Roberta 模型中最后一层的输出拼接在一起，经过全连接层和 softmax 激活函数后输出最后的结果。

g. Roberta+TextCNN。将 Roberta 模型中 encoder 层的输出作为 TextCNN 模型的输入，得到输出，将此输出与 Roberta 模型中最后一层的输出拼接在一起，经过全连接层和 softmax 激活函数后输出最后的结果。

所有模型的实验结果对比见表 3 和表 4，其中本文所提出的模型为基于余弦式调整学习率的方法。

THUCNews 短文本分类中，无迁移学习的模型中 FastText 模型的效果最优。而在迁移学习的模型中，本文所采用的模型结合余弦式调整学习率的方法，所得出的结果在所有模型中最优，*F*1 值可以达到 96.98%，比 FastText 模型高出了 2.99%，比使用 Roberta+TextCNN 高出了 1.02%。在 THUCNews 长文本分类中，本文模型相比于无迁移学习的 DPCNN 模型，准确率高出了 5.23%，比 Roberta+LSTM 模型高出了 1.56%。在其他 3 项评价标准上，效果也明显优于其他模型。

THUCNews 短文本分类中，无迁移学习的模型中效果最好的是 FastText。这是因为 FastText 将短文本中的所有词向量进行平均，句子中的序列、语义和结构信息保存都较为完整。

表 3　THUCNews 短文本各模型指标对比

Tab.3　Comparison of various model indicators of THUCNews short text

分类模型	准确率	精确率	召回率	*F*1值
FastText	93.97	94.01	93.97	93.99
TextCNN	92.69	92.7	92.69	92.68
LSTM	93.43	93.45	93.43	93.43
DPCNN	92.28	92.39	92.28	92.33
Bert+DPCNN	94.31	94.31	94.31	94.31
Roberta+LSTM	94.77	94.78	94.77	94.77
Roberta+TextCNN	95.97	95.96	95.96	95.96
本文	96.96	96.99	96.97	96.98

表 4　THUCNews 长文本准确率

Tab.4　Accuracy of THUCNews long text

分类模型	准确率	精确率	召回率	*F*1值
FastText	91.29	91.32	91.29	91.07
TextCNN	93.07	93.17	93.07	93.01
LSTM	93.20	93.43	93.20	93.31
DPCNN	93.18	93.31	93.18	93.24
Bert+DPCNN	94.62	94.62	94.62	94.62
Roberta+LSTM	96.85	96.86	96.85	96.85
Roberta+TextCNN	96.80	96.81	96.80	96.80
本文	98.41	98.44	98.41	98.41

在无迁移学习的模型结构中，长文本分类使用 FastText 模型，效果不如卷积神经网络和循环神经网络。其原因是 *n*-gram 结构所能获取的上下文语义信息不如神经网络模型结构完整。

使用预训练模型 Roberta 连接下游任务，模型的整体性能优于传统模型。这是因为预训练模型中的参数从海量数据中训练得来，相较于传统神经网络的自己从头开始训练，预训练模型的收敛速度更快，泛化效果更好。

学习率作为监督学习以及深度学习中重要的超参，决定着目标函数能否收敛到局部最小以及何时收敛到最小。合适的学习率能使目标函数在合适的时间内收敛到局部最小。表 5 为学习率衰减实验结果，从中可以发现，以 *F*1 值为评价标准，Roberta 模型使用余弦调整的方式分层调整其

表 5 结合不同学习率衰减实验结果（*F*1 值）

Tab.5 Combined with different learning rate attenuation experimental results (*F*1 value)

数据集	手动阶梯式	自定义	三段式	余弦调整
THUCNews短文本	96.83	96.91	96.88	96.96
THUCNews长文本	97.48	97.53	97.41	98.41

学习率，模型效果可以得到小幅度的提升。

从实验结果还可以看出，Roberta 模型在放弃 NSP 任务后，得到的句向量和词向量的内容更为丰富。使用 DPCNN 和 RGRU 模型作为模型的深层特征提取层，能再次提取句子中的有效信息，模型的泛化能力得到了进一步增强。

从图 7 和图 8 各个类别的 *F*1 值中可以看到，短文本分类模型中股票和金融类别的 *F*1 值较低，而长文本分类中只有金融这一个类别的 *F*1 值较高。从短文本中选取一部分相关性较高的数据（表 6），结合图 9，短文本分类里金融类被识别为

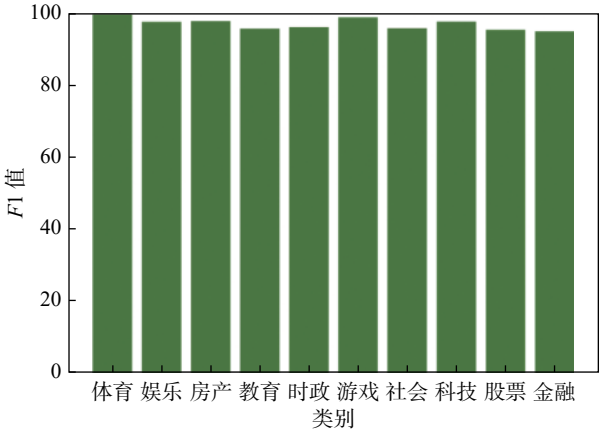

图 7 短文本分类 *F*1 值

Fig. 7 *F*1 value for short text category

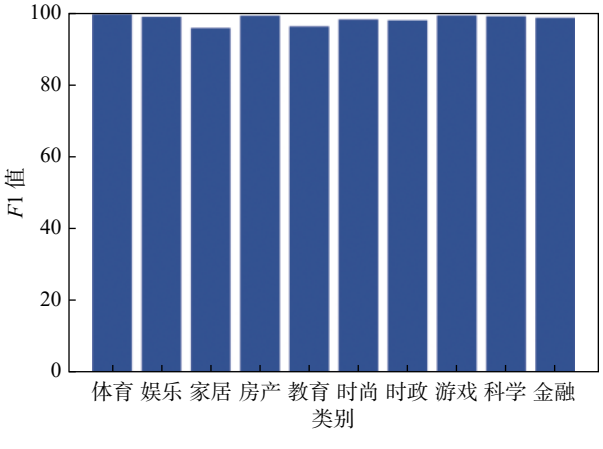

图 8 长文本分类 *F*1 值

Fig. 8 *F*1 value for long text category

表 6 样本数据示例

Tab.6 Sample data example

序号	内容	标签	预测
1	11名小学生课堂遭老师殴打 区教委将介入调查	社会	教育
2	两位西班牙诗界巨擘百年诞辰纪念活动(短文本)	教育	时政
3	警方披露八种高考骗局 高科技作弊器诈骗家长	教育	科技
4	教育部: 将学术道德教育纳入高校课程	时政	教育
5	标普下调斯洛文尼亚评级至AA-	股票	经济
6	收评: 沪基指跌0.36% 近九成封基下挫	经济	股票

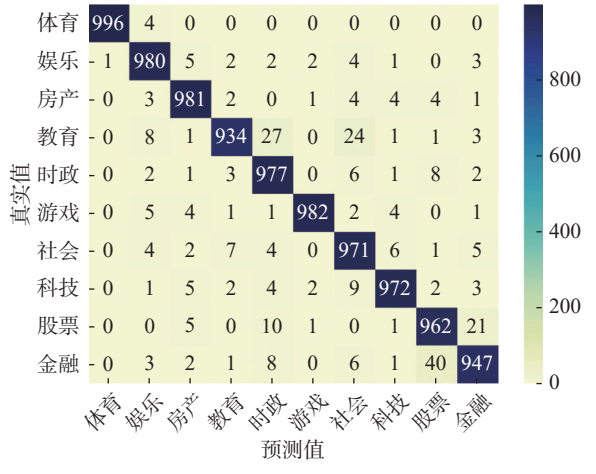

图 9 短文本在测试集上的混淆矩阵热力图

Fig. 9 Confusion matrix heat map of short text on the test set

股票类的有 40 个，股票类被识别为金融类的有 21 个，说明这两个分类在短文本分类模型里相互干扰较为严重。

针对 THUCNews 数据集出现的这种情况，在扩大数据集的同时，需要对数据进行进一步的预处理，同时也需要调整模型，使模型能更好地将不同的数据区分开来。如在序号为 1 的内容中，需要给予殴打、调查等动词更多权重，同时减少小学生、老师、区教委等名词的权重。

4 结束语

以预训练模型结构为基础，连接下游任务的模型结构，其性能优于无迁移学习的网络模型。本文使用了 Roberta 预训练模型连接下游任务的深层特征提取模型，同时针对 Roberta 模型、卷积神经网络和循环神经网络的特点，给予不同的学习率，分层调试其参数，最后得到的文本特征向量信息十分丰富。利用 Roberta 模型中掩码 mask 的策略，使得同一样本在每轮训练的时候，mask

位置不同，提高了模型输入数据的随机性，得到更加符合语义环境的动态词向量，最终提升了模型的学习能力。

通过分析混淆矩阵，得出了当前模型中所存在的不足，下一步将会针对不同类别的数据权重进行研究，尝试将每个词的语义和类型融入到输入层中，进一步增强文本向量的表示信息。同时需要对模型的整体结构进行调整，找出能够提升模型效果的参数，使模型可以更加优秀地处理自然语言处理中的文本分类任务。

参考文献：

[1] CHEN P H, LIN C J, SCHÖLKOPF B. A tutorial on *v*-support vector machines[J]. Applied Stochastic Models in Business and Industry, 2005, 21(2): 111–136.

[2] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016: 785–794.

[3] 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 (EMNLP). Doha: Association for Computational Linguistics, 2014: 1532–1543.

[4] JOSHI M, CHEN D Q, LIU Y H, et al. SpanBERT: improving pre-training by representing and predicting spans[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 64–77.

[5] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics, 2014: 1746–1751.

[6] 陈珂, 梁斌, 柯文德, 等. 基于多通道卷积神经网络的中文微博情感分析 [J]. 计算机研究与发展, 2018, 55(5): 945–957.

[7] YAO L, MAO C S, LUO Y. Graph convolutional networks for text classification[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019: 7370–7377.

[8] HU L M, YANG T C, SHI C, et al. Heterogeneous graph attention networks for semi-supervised short text classification[C]//EMNLP-IJCNLP 2019: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language. Hong Kong, China: Association for Computational Linguistics, 2019: 4821–4830.

[9] LIU P F, QIU X P, HUANG X J. Recurrent neural network for text classification with multi-task learning[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York: AAAI, 2016: 2873–2879.

[10] SHEN Y K, TAN S, SORDONI A, et al. Ordered neurons: integrating tree structures into recurrent neural networks[C]//7th International Conference on Learning Representations. New Orleans: OpenReview. net, 2019.

[11] 王伟, 孙玉霞, 齐庆杰, 等. 基于 BiGRU-attention 神经网络的文本情感分类模型 [J]. 计算机应用研究, 2019, 36(12): 3558–3564.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc. , 2017: 6000–6010.

[13] FENIGSTEIN A. Self-consciousness, self-attention, and social interaction[J]. Journal of Personality and Social Psychology, 1979, 37(1): 75–86.

[14] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by Jointly learning to align and translate[C]//3rd International Conference on Learning Representations. San Diego, 2015.

[15] CHENG K F, YUE Y N, SONG Z W. Sentiment classification based on part-of-speech and self-attention mechanism[J]. IEEE Access, 2020, 8: 16387–16396.

[16] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019: 4171–4186.

[17] 孙红, 陈强越. 融合 BERT 词嵌入和注意力机制的中文文本分类 [J]. 小型微型计算机系统, 2022, 43(1): 22–26.

[18] 张洋, 胡燕. 基于多通道深度学习网络的混合语言短文本情感分类方法 [J]. 计算机应用研究, 2021, 38(1): 69–74.

[19] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017: 562–570.

[20] SUN C, QIU X P, XU Y G, et al. How to fine-tune BERT for text classification?[C]//18th China National Conference on Chinese Computational Linguistics. Kunming: Springer, 2019: 194–206.

[下转第 204 页]

Knowledge Discovery & Data Mining. London: Association for Computing Machinery, 2018: 1840–1849.

[12] ABEBE R, KLEINBERG J, PARKES D, et al. Opinion dynamics with varying susceptibility to persuasion [C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: Association for Computing Machinery, 2018: 1089–1098.

[13] BERMAN A, PLEMMONS R J. Nonnegative matrices in the mathematical sciences[J]. Society for Industrial and Applied Mathematics, 1994, 35(1): 5–17.

[14] FRIEDKIN N E, JOHNSEN E C. Social influence and opinions[J]. The Journal of Mathematical Sociology, 1990, 15(3/4): 193–206.

[15] GABAY D, MERCIER B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation[J]. Computers & Mathematics with Applications, 1976, 2(1): 17–40.

[16] BOYD S, PARIKH S, CHU E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers[J]. Foundations and Trends® in Machine Learning, 2010, 3(1): 1–122.

[17] LI Y H, CHEN W, WANG Y J, et al. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships[C]//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. Rome: Association for Computing Machinery, 2013: 657–666.

[18] MUSCO C, MUSCO C, TSOURAKAKIS C E. Minimizing polarization and disagreement in social networks[C]//Proceedings of the 2018 World Wide Web Conference. Lyon: International World Wide Web Conferences Steering Committee, 2018: 369–378.

（编辑：石 瑛）

[上接第 197 页]

[21] 宋明, 刘彦隆. Bert 在微博短文本情感分类中的应用与优化 [J]. 小型微型计算机系统, 2021, 42(4): 714–718.

[22] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318–327.

[23] 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: Volume 2, Short Papers. Valencia: Association for Computational Linguistics, 2017: 427–431.

（编辑：丁红艺）