DOI:10.19651/j.cnki.emt.2209922

基于改进 Adam 优化算法的中文短文本分类方法*

赵志杰 张艳艳 毛翔宇
(南京信息工程大学 南京 210044)

摘 要：针对 BERT 模型中编码器提取特征信息时因并行计算而缺少文本的时序信息及模型网络复杂度较高易受偏差影响等问题，本文提出一种基于改进 Adam 优化算法的模型 DTSCF-Net。模型采用 BERT 模型提取短文本的语义特征表示，将语义特征输入到 Bi-GRU 中，提取具有上下文时序特征的语义信息，输入 Maxpooling 层筛选最优特征，分类得到该短文本的类别。针对 Adam 算法在拟合中产生的动量偏差添加校正算法来缓解性能下降，对比两个连续时间步上的校正动量值，选取两个时间步中的动量最大值代入梯度计算，并对学习率添加自适应调节因子，利用上一次迭代的梯度值，实现学习率的自适应调节，提高分类精度。实验表明，DTSCF-Net 的分类准确率为 94.86%，相较于同实验环境下的基准模型 BERT、BERT-Bi-GRU 分别提高 2.07%、1.71%。结果证明本文所提方法具有一定的性能提升。

关键词：文本分类；自适应矩估计；BERT；Bi-GRU；短文本
中图分类号：TP391.1 **文献标识码**：A **国家标准学科分类代码**：520.2020

Research on Chinese short text classification method based on improved Adam optimization algorithm

Zhao Zhijie Zhang Yanyan Mao Xiangyu
(Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: The model uses the BERT to extract the semantic feature representation of the short text, inputs the semantic features into the Bi-GRU and extracts the semantic information with contextual timing features. The model feeds the features into the Maxpooling layer to filter the optimal features and classify them to get the category of the short text. A correction algorithm is added to mitigate the performance degradation for the momentum bias generated by the Adam algorithm in the fitting. The Adam algorithm is improved by comparing the corrected momentum values at two consecutive time steps and selecting the maximum value of momentum in the two time steps to substitute into the gradient calculation. The improved Adam algorithm adds an adaptive adjustment factor to the learning rate and uses the gradient value of the previous iteration to achieve adaptive adjustment of the learning rate and improve the classification accuracy. Experiments show that the classification accuracy of DTSCF-Net is 94.86%, which is 2.07% and 1.71% higher than that of the benchmark model BERT and BERT-Bi-GRU respectively in the same experimental environment. The results demonstrate that the proposed method in this paper has certain performance improvement.

Keywords: text classification; Adam; BERT; Bi-GRU; short text

0 引 言

中文短文本分类是自然语言处理领域的重要研究方向之一，其应用领域广泛，如垃圾邮件过滤、个性化推荐[1]、情感分析等。短文本的字符长度一般不超过 160 个字符，文本内容由高度概括的词汇组成[2]。短文本的分类难点在于用词缺乏规范、语义模糊，现有的分类方法表现不佳，且文本长度较短，在有限的长度内提取其完整语义特征的挑战较大[3]。如何提高分词准确性与分类准确度，对短文本分类有重要的研究意义。

文本分类的首要任务是对文本进行表示，即将字词以向量的形式表示。该方法主要分为两类，一类是静态词向量表示方法 Word2Vec、GloVe[4]（global vectors），另一类是 ELMo（embedding from language models）、BERT

收稿日期：2022-05-12
* 基金项目：国家自然科学基金(61705109)、江苏高校优势学科建设工程资助项目、江苏省双创团队人才计划

(bidirectional encoder representations from transformers)等动态词向量方法[5]。Mikolov[6]等提出 Word2Vec 模型,为了让词向量高效地表示上下文信息,它提供了 Sikp-Gram 与 CBOW(continuous bag-of-words)两种训练方法,但它只利用了文本的局部信息,并未高效利用文本的全局信息。为解决此问题,Pennington 等[7]提出了全局词向量模型(global vectors,GloVe),兼顾了文本的局部信息与全局信息。Word2Vec 与 GloVe 等词向量表示方法为文本分类的模型性能带来了有效提升[8],但这类词向量表示方法均为静态词向量表示方法,在不同的上下文中,同一字词在不同语境中的词向量表示相同,无法处理"一词多义"和"一义多词"问题[9],导致文本分类的性能受限。

随着 ELMO、BERT 等动态词向量模型的提出,文本的语境歧义问题得到解决[10],其中最具代表性的 BERT 模型能捕捉词语和句子级别的表示,在多类文本处理任务中表现优异。而 BERT 模型具有 12 层堆叠的多头注意力 Encoder 层[11],其网络结构较为复杂,在受到误差干扰的情况下容易出现性能下降。科大讯飞联合哈工大[12]发表了 BERT-WWM,改变了训练样本的生成策略,由局部词 mask 机制修改为全词遮盖(whole word mask,WWM)。Lan 等[13]提出一种基于 BERT 的轻量级预训练语言模型 ALBERT(a lite BERT),通过嵌入层参数因式分解减少 BERT 参数量,扩展了 BERT 模型的可用性。温超东等[14]结合 ALBER 与门控循环单元(gated recurrentUnit,GRU)模型在专利文本分类任务上取得了不错的效果,但模型分类精度相较于 BERT 有一定程度的下降。

参考上述文献,本文从 BERT 模型网络复杂度较高易受偏差影响与如何有效提取的语义特征两方面开展理论与研究工作,主要贡献如下:1)提出一种改进的 Adam 算法,通过对一阶矩估计与二阶矩估计的校正与对比前后两个迭代阶段的矩估计值,选取最优值更新网络,减小误差对模型的影响,并对学习率添加自适应调节因子,实现网络参数的自适应调整,逼近网络的最优解。2)提出使用 BERT 模型对文本中的字符进行动态词向量表示,引入了 Bi-GRU 最大化获取词向量的上下文时序信息,弥补 BERT 模型在编码时仅使用位置编码而缺乏文本时序信息的不足,并通过 CNN 的最大池化层筛去噪声获取最优特征。

1 相关算法

1.1 BERT

BERT 的特征提取器采用了 Transformer 编码器。Transformer Encoder 是以多头注意力机制作为基础结构,具有并行计算的优点[15]。其自注意力层逐一计算该词与句中其它词的关照程度,通过 Softmax 计算上一层的隐层特征,将该特征转换为相关性的概率分布,并对输入做线性映射,乘以 $\boldsymbol{W}^Q$ 权重矩阵得到每个输入 $\boldsymbol{x}$ 的 Query 向量,同样分别乘以 $\boldsymbol{W}^K$,$\boldsymbol{W}^V$ 计算得出 Key 向量和 Value 向量,利用 Query、Key 和 Value 向量计算句中词语之间的关联度[16]。

上述过程可以用如式(1)所示。

$$\boldsymbol{Z}=Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax(\frac{\boldsymbol{Q}^*\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}})^*\boldsymbol{V} \quad (1)$$

其中,d_k 表示 Key 向量的维度。BERT 中编码器的多头注意力机制为每个注意力头都分配一个权重矩阵。

在训练开始时对权重矩阵进行随机初始化,并将来自较低层的编码器的矢量投影到不同的表示子空间[17]。对于多个注意力头,计算其相应的关照程度向量 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_7$,将多个关照程度向量拼接之后乘以矩阵 W_o 得到注意力矩阵 $\boldsymbol{Z}$[18]。BERT 将每个编码器的输入同多头注意力结果或前馈神经网络运算结果进行拼接,对拼接结果使用归一化函数,对 $\boldsymbol{Z}$ 进行更新,通过注意力矩阵 $\boldsymbol{Z}$ 对文本字符进行动态表示。

1.2 Adam 优化算法

自适应矩估计算法(Adam)可使网络的参数随着模型的训练而对每个参数的学习率独立调整,从而实现自适应学习率[18]。Adam 通过计算梯度得到一阶矩估计 m_t 和二阶矩估计 v_t 来计算不同参数的个体自适应学习率。每一个时间步上的学习率参数值 θ_t 是由当前时间步的一阶矩估计 m_t 和二阶矩估计 v_t 及上一时间步的学习率参数 θ_{t-1} 来更新的。Adam 优化器的基本算法可以描述如下。

设定噪声目标函数 $f_t(\theta)$,表示参数 θ 在第 t 个时间步的随机函数。为减小 $f_t(\theta)$ 的期望,需要使用随机性描述的小样本噪声,计算目标函数关于变量 θ 的梯度如式(2)所示。

$$g_t=\nabla f_t(\theta) \quad (2)$$

计算一阶矩估计、二阶矩估计的公式如式(3)、(4)所示:

$$m_t=\beta_1\cdot m_{t-1}+(1-\beta_1)\cdot g_t \quad (3)$$

$$v_t=\beta_2\cdot v_{t-1}+(1-\beta_2)\cdot g_t^2 \quad (4)$$

其中,参数 β_1、$\beta_2\in[0,1)$,控制 m_t 和 v_t 的衰减速度。在迭代早期,衰减率很小的时候,矩估计值容易偏向 0[19]。

为了解决迭代早期矩估计值偏 0,一般需要对 m_t 和 v_t 分别进行修正,修正后的一阶矩估计 $\hat{m}_t$ 和二阶矩估计 $\hat{v}_t$ 如式(5)、(6)所示。

$$\hat{m}_t=m_t/(1-\beta_1^t) \quad (5)$$

$$\hat{v}_t=v_t/(1-\beta_2^t) \quad (6)$$

$\hat{m}_t/\sqrt{\hat{v}_t}$ 表示信噪比,,当信噪比较小,Δt 趋于无穷小时,目标函数将收敛。令二阶矩估计 v_t 的初始值为 0,则 v_t 在第 t 个迭代周期更新方法如式(7)所示。

$$v_t=(1-\beta_2)\sum_{i-1}^{t}\beta_2^{t-T}g_T^2 \quad (7)$$

每迭代一个周期,都要更新参数 θ_t,其更新方式如式(8)所示。

$$\theta_t=\theta_{t-1}-\alpha\cdot m_t/(\sqrt{v_t}+\varepsilon) \quad (8)$$

其中，α 为学习率；$\varepsilon = 10^{-8}$ 为常数参量。通过参数更新迭代，使目标函数向最优解方向收敛。

Adam 优化算法中，对一阶矩到非中心的二阶矩估计进行了修正，但在面对数据规模较大的中文短文本分类任务时，该算法的迭代曲线会出现震荡，收敛性能变差。

2　基于改进 Adam 算法优化的短文本分类模型

2.1　BERT-BiGRU-Maxpooling 模型

本文提出了 BERT-BiGRU-Maxpooling 模型，联合上下文时序信息提取文本特征信息中的最优特征。BERT 模型可利用双向上下文信息对文本建模，其多头自注意力机制的编码器可并行计算每个字符在当前句中的关联度，根据不同环境对词向量动态表示；Bi-GRU 可最大化获取词向量的上下文时序信息，弥补 BERT 模型在编码时仅使用位置编码而缺乏文本时序信息的不足；CNN 的 Maxpooling 层可筛去当前特征中的干扰噪声，使模型获得益于分类的最优特征。

模型的架构如图 1 所示，分为文本输入、BERT 语义编码层、特征提取层、分类层与输出。BERT 语义编码层将文本输入到 BERT 语义编码层进行语义编码、语义补充，来缓解文本长度较短、语义稀疏的问题；特征提取层用 Bi-GRU 通过时间序列上正序、逆序的特征提取并拼接，得到具有上下文特征的语义特征；分类层对特征提取层的输出进行 Maxpooling 来获取最优特征，最后通过 Softmax 进行分类得到该短文本的类别。

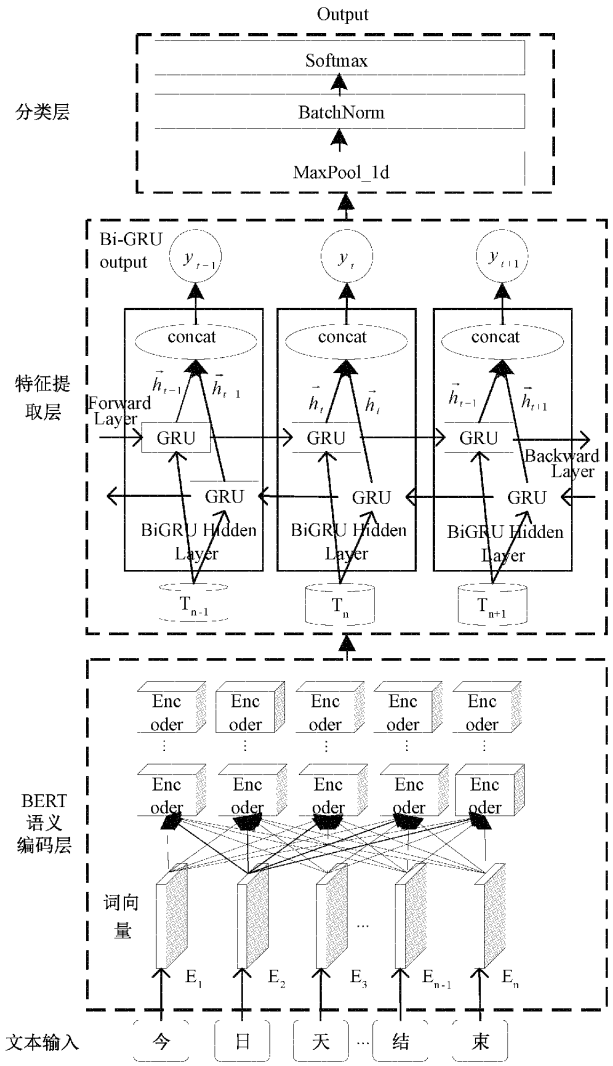

图 1　模型框架

具体步骤如下：文本输入后通过 embedding 层生成 embedding 向量 $\boldsymbol{E}_1,\boldsymbol{E}_2\cdots\boldsymbol{E}_n$，接着 BERT 语义编码层的 12 层 Transformer 编码器并行运行提取 $\boldsymbol{E}_1,\boldsymbol{E}_2\cdots\boldsymbol{E}_n$ 的特征信息，得到输出向量 $\boldsymbol{T}_1,\boldsymbol{T}_2\cdots\boldsymbol{T}_n$，然后将输出向 $\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n$ 传入特征提取层，特征提取层是由 1 个正序的 GRU 和 1 个逆序的 GRU 组成，$\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n$ 表示经过 GRU 计算得到的向量，$\boldsymbol{y}_n$ 是将正序 $\boldsymbol{h}_n$ 与逆序 $\boldsymbol{h}_n$ 向量拼接之后得到的 Bi-GRU 的输出；分类层由 Maxpool、Batch Nromalization 和 Softmax 组成，Maxpool 对 $\boldsymbol{y}_n$ 进行最大池化来筛选最优特征，Softmax 计算上一层的隐层特征，将该特征转换为不同类别的概率，实现短文本的主题分类，如式(9)所示。

$$p(k_j\mid q,w)=\exp(z_i)/\sum_{i=1}^{n}\exp(z_i) \tag{9}$$

其中，j 表示文本分类的标签；q 表示文本，w 表示模型的训练参数；z 表示模型的隐藏层特征；n 表示隐藏层特征的数量。模型采用交叉熵损失函数进行模型的优化。损失函数如式(10)所示。

$$Loss=-\sum_{i=1}^{D}\sum_{j=1}^{N}\hat{y}_i^j\ln(y_i^j) \tag{10}$$

其中，D 表示训练集大小，N 表示文本分类的标签数量，$\hat{y}$ 表示真实的文本类别标签，y 表示模型预测的文本类别标签。

2.2　基于改进 Adam 算法的模型优化

本文提出了改进的 Adam 优化算法，改进内容分为两部分：

1)对 Adam 中存在误差的动量值进行校正，并对比两个连续时间步上的动量值，选取两个时间步中的动量最大值代入计算，减小动量值偏 0 产生的误差。

2)在训练过程中，模型的收敛趋势接近幂指数函数的变化趋势，因此本文对学习率添加调节因子，利用上一次迭代的梯度值进行计算，达到自适应调节学习率的要求，进而改变网络的收敛性能。

1)中的具体内容如下，对存在偏差的 m_t、v_t 进行校正，通过计算得出在时间步 t 上指数移动平均值 $E(v_t)$ 与真实二阶矩 $E(g_t^2)$ 在 $\zeta=0$ 时的比值为 $(1-\beta_2)$ 的关系，如式(11)所示。

$$E(v_t)=E\left[(1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-T}g_T^2\right]=(1-\beta_2)E(g_t^2)+\zeta \tag{11}$$

从而得到经过误差校正后的一阶矩估计 c_t 和二阶矩估计 q_t，如式(12)所示。

$$c_t = \frac{\beta_1 \cdot m_{t-1}}{(1-\beta_1)} + g_t$$

$$q_t = \frac{\beta_2 \cdot v_{t-1}}{(1-\beta_2)} + g_t^2 \tag{12}$$

基于对一阶矩估计 m_t 和二阶矩估计 v_t 的校正，进一步对比两个连续时间步上的动量值，选取两个连续时间步中的动量最大值代入梯度计算，减小训练初期动量值初始化为 0 的误差。

2)中的具体内容如下，设置自适应调节的学习率控制系数，如式(13)所示。

$$\alpha = \alpha_0 m^{-k} \tag{13}$$

其中，α_0 为初始学习率，本文取 α_0 为 2×10^{-5}；m 表示迭代的过程中间量，由当前迭代次数与最大迭代次数决定；参数 k 如式(14)所示。

$$k = \sum_{i=1}^{n}\lambda_i + p \tag{14}$$

其中，λ_i 表示第 i 次迭代过程中步长变化率；p 为常量参数，本文取 0.75。

根据上述计算，在第 t 次迭代学习率的更新过程如式(15)～(17)所示。

$$\alpha_t = \frac{a_{t-1}}{\sqrt{K+L_t}} \tag{15}$$

$$G_t = \varepsilon \cdot g_{t-1}^2 + g_t^2 \tag{16}$$

$$\alpha_{t-1} = \alpha_0\left[1+\frac{t}{R}\right]^{-(k)} \tag{17}$$

其中，R 为最大迭代次数；G_t 为第 t 次迭代的梯度值 g_t 和 $t-1$ 次迭代的梯度值 g_{t-1} 的平方和。ε 为 Adam 算法的衰减因子，默认取值 0.999；K 为常数项，取值为 1。

综合上述两点，改进 Adam 优化算法流程如表 1 所示。

表 1 改进 Adam 算法流程

参数：	α：学习率；$\beta_1,\beta_2\in[0,1)$：矩估计的指数衰减率；$f(\theta)$：θ 的随机标量函数；θ_0：初始参数向量 r；$\lambda\in[0,1)$：解耦权重衰减；$\varepsilon=10^{-8}$
1	$m_0\leftarrow0$(初始化一阶矩估计)
2	$v_0\leftarrow0$(初始化二阶矩估计)
3	$t\leftarrow0$(初始化时间步)
4	while θ_t 不收敛时
5	$t\leftarrow t+1$
6	$g_t\leftarrow\nabla_\theta f_t(\theta_{t-1})$(计算在时间步长上的梯度值)
7	$m_t\leftarrow\beta_1\cdot m_{t-1}+(1-\beta_1)\cdot g_t$(更新一阶矩估计)
8	$v_t\leftarrow\beta_2\cdot v_{t-1}+(1-\beta_2)\cdot g_t^2$(更新二阶矩估计)
9	$c_t=\beta_1\cdot m_{t-1}/(1-\beta_1)+g_t$(校正一阶矩估计)
10	$q_t=\beta_2\cdot v_{t-1}/(1-\beta_2)+g_t^2$(校正二阶矩估计)
11	$\hat{c}_t=\max(c_t,c_{t-1})$(对比与前一时间步、当前时间步的一阶矩估计，以获得最大值)
12	$\hat{q}_t=\max(q_t,q_{t-1})$(对比与前一时间步、当前时间步的二阶矩估计，以获得最大值)
13	$\alpha_t=a_{t-1}/\sqrt{K+L_t}$(利用梯度值计算自适应调节因子控制学习率的变化)
14	$\theta_t\leftarrow\theta_{t-1}-\alpha_t\cdot\hat{c}_t/(\sqrt{\hat{q}_t}+\varepsilon)$(更新参数)
15	End while
16	Return θ_t(获得参数)

其中，$f(\theta)$ 是一个关于 θ 可微的随机标量函数，$g_t=\nabla_\theta f_t(\theta_{t-1})$ 表示梯度即向量在时间步 t 上关于 θ 的偏导数，通过对梯度 g_t 的计算得到一阶矩估计 m_t 和二阶矩估计 v_t 即梯度的均值与梯度的未中心化方差，分别控制模型的更新方向与步长大小。对 m_t 与 v_t 进行校正操作，并为进一步减小训练初期 m_t、v_t 偏零对模型的影响，对进行误差校正后的一阶矩估计和二阶矩估计在当前时间步 t 与上一时间步 $t-1$ 进行对比，选取两个时间步中的最大值作为学习率参数的更新参数代入计算，得出最新的学习率。

3 实验结果与分析

3.1 实验数据集

本文从 THUCNEWS 数据集中选取了 10 类作为训练的数据集，在每类文本中，选取 10 000 个样本作为训练集，另从每类中选取 1 000 个样本作为测试集，总量是 10 000 个文本数据，训练集与测试集相互独立，测试集不参与参数优化训练。

3.2 实验参数设置

BERT 的预训练模型常用的有两种版本分别为

BERT-Base 和 BERT-Large。本实验使用 BERT-Base 的中文预训练模型进行实验，此模型有 12 层，隐藏层的维度为 768，12 个注意力头，包含 110 M 个参数。Pad_size 大小为 32，Learning_rate 为 2e－5，Train_batch_size 为 128，Eval_batch_size 为 128，Num_train_epoch 为 6。其中，Pad_size 表示输入文本的切分长度（长截短填）；Learning_rate 表示模型的学习率；Train_batch_size 表示训练集训练迭代数据的数量；Eval_batch_size 表示验证集训练迭代数据的数量；Num_train_epochs 表示模型训练迭代的次数。

3.3　实验结果与分析

1）对比实验

为了验证本文提出的 DTSCF-Net 的有效性。选取了 DTSCF-Net、BERT-Bi-GRU、BERT、DPCNN、Text-RNN、Text-CNN、FastText 多个模型在相同实验环境下进行对比实验，实验结果如表 2 所示。本文所提方法 DTSCF-Net 取得了较好的表现，其准确率、召回率、F1 值三个指标分别为 94.86%、94.87%、94.85%。与基准算法 BERT 相比，DTSCF-Net 在准确率、召回率、F1 值三个指标上分别提升了 2.08%、2.08%、2.07%。BERT-Bi-GRU 与基准模型 BERT 相比，前者性能也有较大提升，这表明循环神经网络提供的双向时序信息补足了 BERT 编码层仅提供文本位置编码信息的不足，提高了分类的准确率。

表 2　对比实验结果

模型	准确率/%	召回率/%	F1 值/%
DTSCF-Net	**94.86**	**94.87**	**94.85**
BERT-Bi-GRU	93.15	93.14	93.18
BERT	92.79	92.78	92.78
DPCNN	90.86	90.90	90.88
Text-RNN	90.55	90.61	90.56
Text-CNN	91.08	90.99	90.99
FastText	91.92	91.97	91.91

2）改进 Adam 算法对分类精度的影响

对改进 Adam 算法进行分类性能提升判别，分别使用 Adagrad 、Adadelta、RMSProp、Adam 优化算法与改进 Adam 优化算法进行实验，模型网络参数保持一致。训练策略为基于原有的测试集数据逐步增大测试集的数据规模，在 10 000 个文本数据的基础上依次增加 2 000 个文本数据，不同优化算法的分类精度如表 3 所示。

由表 3 可知，本文提出的改进 Adam 优化算法使模型的最大分类精度达到了 94.8%，分类精度均大于其他算法带来的提升。随着文本数据规模的扩大，本文方法的分类精度的波动较小，具有较其他优化算法相比更理想的分类效果。

3）改进 Adam 算法的稳定性

为对本文所提方法进行稳定性判别，在相同实验环境下，对实验 2）中 10 次实验中的每次实验每隔 200 次迭代读取一次数据并计算损失函数的均方差。多种优化算法在同一迭代阶段的损失函数的均方差如表 4 所示。

表 3　多个优化算法的分类精度

次数	Adagrad	Adadelta	RMSProp	Adam	改进 Adam
1	0.906	0.894	0.917	0.915	**0.948**
2	0.909	0.899	0.927	0.918	**0.947**
3	0.913	0.88	0.895	0.921	**0.946**
4	0.91	0.879	0.911	0.918	**0.947**
5	0.911	0.869	0.896	0.933	**0.948**
6	0.899	0.888	0.901	0.918	**0.946**
7	0.894	0.884	0.879	0.926	**0.943**
8	0.912	0.871	0.881	0.922	**0.946**
9	0.92	0.863	0.892	0.939	**0.945**
10	0.895	0.877	0.884	0.94	**0.948**

表 4　同一迭代阶段损失函数均方差（$\times 10^{-2}$）

迭代阶段	Adagrad	Adadelta	RMSProp	Adam	改进 Adam
0	4.45	7.75	6.43	2.69	**2.37**
200	3.24	5.19	3.19	3.17	**1.34**
400	3.67	6.28	5.99	4.09	**2.69**
600	8.19	7.32	9.97	2.25	**1.19**
800	11.8	5.14	10.3	6.17	**2.13**
1 000	13.8	6.11	6.18	12.1	**4.24**
1 200	10.2	9.87	9.24	3.43	**1.27**
1 400	9.24	9.01	8.1	14.5	**6.35**
1 600	8.43	12.2	12.5	12.8	**5.26**
1 800	6.25	8.13	7.13	8.71	**3.15**
2 000	9.01	6.41	6.38	3.94	**1.91**
2 200	11.1	7.72	5.41	16.6	**5.4**
2 400	10.5	11.6	7.05	6.24	**2.16**
2 600	9.8	12.5	9.11	5.48	**1.37**
2 800	8.99	6.61	7.13	6.33	**5.17**

由表 4 可知，在同一迭代阶段，本文所提方法使模型的损失函数的均方差更小，表明改进 Adam 算法可使模型在不同规模的文本数据集上表现更稳定。在不同迭代阶段，由改进 Adam 优化的模型相较于其它模型，均方差的波动范围更小，说明改进 Adam 可使模型对文本数据规模变化的适应性增强。

4）改进 Adam 算法的收敛性

为对本文所提方法进行收敛性判别，基于原有的测试集数据逐步增大测试集的数据规模，在 10 000 个文本数据的基础上依次增加 2 000 个文本数据，改进 Adam 算法的收敛性实验结果如表 5 所示。可以看出，对比其他优化算法，改进的 Adam 算法到达收敛点所花费的迭代次数更

少,收敛速度更快。

表 5 不同算法到达收敛点的迭代次数

次数	Adagrad	Adadelta	RMSProp	Adam	改进 Adam
1	497	610	332	307	249
2	501	597	309	291	217
3	504	603	307	305	231
4	509	602	310	297	219
5	503	616	318	301	221
6	511	604	329	309	219
7	509	611	314	314	221
8	501	605	322	304	216
9	509	612	301	305	210
10	510	622	317	299	217

5)改进 Adam 算法优化下模型的表现

如图 2、3 所示,图 2、3 为基于改进 Adam 算法优化的 DTSCF-Net 与 Adam 算法优化的模型在测试集上的准确率与损失误差的变化趋势。

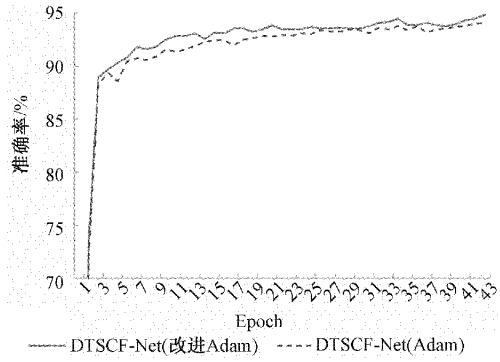

图 2 测试集下的准确率

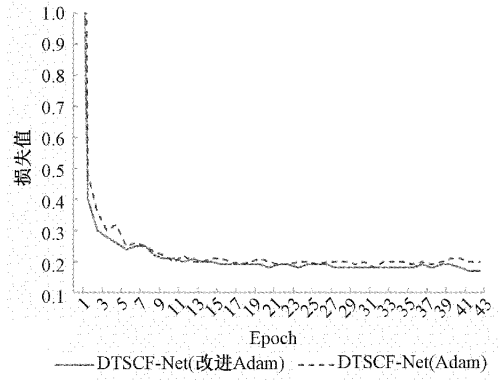

图 3 测试集下的损失值

根据分类准确率的变化趋势,改进 Adam 优化下的模型相较于 Adam 优化下的模型的分类精度更好,前者具有一定的分类性能提升。根据损失误差的变化趋势,训练前期改进 Adam 可使模型在更少的迭代次数下到达收敛点,收敛速度更快。在训练后期改进 Adam 可使模型的损失误差值更小,且在训练的全过程中,改进 Adam 优化下的模型相较于 Adam 优化下的模型,其整体损失误差更低且波动程度更小。综合对比,本文提出的改进 Adam 算法可使模型具有更好的收敛性能与预测准确率。

4 结 论

基于改进 Adam 优化算法的短文本分类方法研究证明,BERT 模型能够丰富短文本语义信息弥补短文本特征分散的不足,Bi-GRU 能够最大化利用上下文时序信息,使模型能够有效的提取短文本特征信息,改进 Adam 优化算法可以加快收敛速度、增强稳定性,从而提高分类精度。本模型在中文短文本分类任务上的准确率达到 94.86%,表现优于同实验环境下的 BERT、BERT-Bi-GRU、DPCNN 等机器学习与深度学习模型。通过对比多种算法在相同实验环境下的实验效果,可以证明本文提出的基于改进 Adam 优化算法的短文本分类方法具有一定的优越性。

参考文献

[1] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2014, 18(7):1527-1554.

[2] 占妮. 基于 Bi-LSTMA-CNNA 的线上评论情感分析模型[J]. 电子测量技术, 2021, 44(3): 83-86, DOI: 10.19651/j.cnki.emt.2005570.

[3] JIA S BA, LI M Z, XIANG Y. Chinese open relation ex-traction and knowledge base establishment[J]. ACM Transactions on Asian Low-Resource Language Information Processin, 2018, 17(3):1-12.

[4] LAI S W, XU L H, LIU K, et al. Recurrent convolutional neural networks for textclassification[C]. Austin, exas: In Proceedings of AAAI, 2015.

[5] 姜鹏. 基于 BERT 的《中图法》文本分类系统及其影响因素分析[J]. 图书馆研究与工作, 2022(5):43-48.

[6] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributedrepresentations of words and phrases and their compositionality[C]. Proceedings of the 26th International Conference on Neuralinformation Processing Systems. Red Hook, NY: CurranAssociates Inc. 2013:3111-3119.

[7] PENNINGTON J, SOCHER R, MANNING C D. GloVe: Globalvectors for word representation[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2014:1532-1543.

[8] CHRISTOPOULOU F, MIWA M, ANANIADOU S. Con-necting the dots: Document-level neural relation extrac-tion with edge-oriented graphs[C]. Proceedings

of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. HongKong:IEEE,2019: 4925-4936.

[9] 王婉,张向先,卢恒,等.融合 FastText 模型和注意力机制的网络新闻文本分类模型[J].现代情报,2022,42(3):40-47.

[10] LU X S, ZHOU M, WU K. A novel fuzzy logic-based text classification method for tracking rare events on twitter[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019, 51(7): 4324-4333.

[11] 杨飘,董文永.基于 BERT 嵌入的中文命名实体识别方法[J].计算机工程,2020,46(4):40-45,52. DOI:10.19678/j.issn.1000-3428.0054272.

[12] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[13] LAN Z Z, CHEN M D, GOODMAN S, et. al. ALBERT:a lite BERT for self-supervised learning of language representations [EB/OL]. (2020-02-09) [2021-05-25]. https://arxiv.org/pdf/1909.11942.pdf.

[14] 温超东,曾诚,任俊伟,等.结合 ALBERT 和双向门控循环单元的专利文本分类[J].计算机应用,2021,41(2):407-412.

[15] WU H, CHENG M, LI D. Chinese text classification based on character-level CNN and SVM [J]. International Journal of Intelligent Information and Database Systems, 2019, 12(3): 212-228.

[16] 杨秀璋,李晓峰,袁杰,等.一种融合语义知识和 BiLSTM-CNN 的短文本分类方法[J].计算机时代,2021(11): 49-54. DOI: 10.16644/j.cnki.cn33-1094/tp.2021.11.013.

[17] XU C, HUANG W, WANG H, et al. Modeling local dependence in natural language with multi-channel recurrent neural networks[C]. Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019, 33: 5525-5532.

[18] 程雅倩,黄玮,金晓祥,等.5G 环境下高校图书馆自媒体平台多标签文本分类方法研究[J].情报科学,2022,40(2):155-161. DOI:10.13833/j.issn.1007-7634.2022.02.021.

[19] HASSAN A, MAHMOOD A. Convolutional recurrent deep learning model for sentence classification [J]. IEEE Access, 2018, 6: 13949-1395.

作者简介

赵志杰,硕士研究生,主要研究方向为语义情感分析方向。

E-mail:zhaozhijie1125@163.com

张艳艳,副教授,主要研究方向为信号处理、图像去噪。

毛翔宇,学士,主要研究方向为信息处理等。