引用格式:张韬政,蒙佳健,李康.基于模型不可知元学习与对抗训练的中文情感分析研究[J].中国传媒大学学报(自然科学版),2023,30(03):31-40.
文章编号:1673-4793(2023)03-0031-10

基于模型不可知元学习与对抗训练的中文情感分析研究

张韬政*,蒙佳健,李康
(中国传媒大学 信息与通信工程学院,北京 100024)

摘要:中文情感分析旨在挖掘出中文文本中的主观情感。目前大多数基于深度学习的中文情感分析模型需要依赖大规模的标注数据去训练,同时深度学习模型在实际应用当中很容易受到对抗性扰动的影响,导致模型的性能下降。针对上述问题,本文提出了基于模型不可知元学习与对抗训练的中文情感分析模型,能够在小规模的数据集下利用元学习加速模型收敛,同时生成对抗样本对模型进行对抗训练,提升模型的抗干扰能力,实验证明模型取得了出色的表现。
关键词:BERT;BiLSTM;模型不可知元学习;对抗训练;情感分析
中图分类号:TP391.1 **文献标识码**:A

Research on Chinese affective analysis based on model-agnostic meta-learning and antagonistic training

ZHANG Taozheng*, MENG Jiajian, LI Kang
(School of Information and Communication Engineering, Communication University of China, Beijing 100024, China)

Abstract: Chinese affective analysis aims to dig out the subjective emotion in Chinese text. At present, most Chinese affective analysis models based on deep learning need to rely on large-scale labeled data for training. Meanwhile, deep learning models are easy to be affected by adversarial disturbance in practical applications, resulting in the degradation of model performance. In response to the above issues, this paper proposes a Chinese affective analysis model based on model-agnostic meta-learning and antagonistic training, which can accelerate the convergence of the model using meta learning under small-scale datasets, and generate confrontation samples to conduct confrontat-ion training on the model, improving the anti-interference ability of the model. Experiments have shown that the model has achieved excellent performance.
Keywords: BERT; BiLSTM; model-agnostic meta-learning; antagonistic training; affective analysis

1 引言

语言,是人类社会繁衍,发展与进步最重要的工具。语言搭建起人们交流互动和思想传递的桥梁。文本情感分析也称为意见挖掘,是对那些带有主观情感色彩的文字进行分析研究[1]。

中文情感分析聚焦以汉字为载体的文本数据,对其蕴含的情感进行研究处理[23]。近些年来随着互联

项目基金:中国传媒大学中央高校基本科研业务费专项资金资助(3132018XNG1829)
作者简介(*为通讯作者):张韬政(1982-),女,副教授,博士,主要从事机器学习和自然语言处理研究。Email:zhangtaozheng@cuc.edu.cn; 蒙佳健(2002-),男,本科生,主要从事机器学习和自然语言处理研究。Email:2679914374@qq.com;李康(2001-),男,本科生,主要从事机器学习和自然语言处理研究。Email:2733089527@qq.com

网与智能手机的普及,网络上的文本数据呈现出爆炸式增长,对这些极具丰富情绪色彩的数据进行研究有着重要的商业价值与社会意义[21]。

自Pang等[2]人首次提出文本情感分析后,该领域受到了广泛的关注。早期的情感分析研究是基于规则开展的[3]。通过对海量文本数据进行处理和人工标记,构建情感词典或语料库,将文本与词典中的词进行匹配并计算每个词的分数,最后通过得分判断文本的情感倾向。但此方法人工干预程度较大,投入成本高,且对不同领域的数据有很大的局限性。基于传统机器学习的研究利用朴素贝叶斯和支持向量机等技术进行文本情感分析[4],其准确率得到了很大的提升,但需要人工提取和构造数据集的特征,当数据集规模较大时,实验难度骤增且准确率变差[22]。

随着深度学习的兴起与高速发展,基于人工神经网络的文本情感分析被广泛研究。Kim等[5]最早将卷积神经网络(Convolutional Neural Networks, CNN)用于文本情感分析,利用不同尺寸的卷积核对词向量进行特征提取,最终的实验效果较传统的机器学习有很大的提升。受限于CNN仅能对文本序列的局部信息进行特征提取,自循环神经网络(Recurrent Neural Network, RNN)提出后[6],基于RNN能记录历史特征信息的特性,人们发现RNN较CNN更适合处理序列数据,以RNN为主体的模型被广泛研究[7]。但随着输入序列的长度增大,RNN出现了梯度弥散和梯度爆炸等问题,模型训练变得异常困难。随着对RNN研究的不断深入,Sak等[8]在RNN的结构基础上提出了长短时记忆网络(Long Short-Term Memory, LSTM),有效缓解了RNN的缺陷。Tang等[9]利用多层LSTM进行多层次的情感分析,模型在关联不同序列上表现出很好的效果。

Devlin等[10]提出全新的预训练语言模型BERT-(Bidirectional Transformers Encoder Representations from Transformers),利用双向的Transformer编码器对序列进行特征提取,在处理有多种语义或者某些上下文关联性强的句子上表现出了比传统深度学习模型更强大的性能。BERT采用微调与预训练的方式对大量无标注的数据进行充分训练,可以学到字符级别、单词级别、句子级别甚至段落级别之间的特征[11],在众多自然语言处理的下游任务上均表现出了很好的效果。

现有的绝大多数中文情感分析模型是以BERT为基础展开的,通过BERT生成动态字向量,可以有效避免分词导致的歧义,同时也能有效解决前后文语境中的一词多义的问题[25]。但深度学习模型大都需要大规模的标记数据去训练[12],该类型数据集的制作过程相当繁琐与漫长。同时在某些特定的研究上,也很难得到大量的数据用于制作数据集。近些年来有研究表明,由于深度学习模型自身存在的可解释性差、高维线性等问题[13],其在实际应用中很容易受到对抗性扰动的影响,导致模型的性能下降。

针对上述问题,本文提出了BERT-BiLSTM(Bidirectional Encoder Representations from Transformers-Bidirectional Long Short-Term Memory)模型用于中文情感分析,该模型利用BERT生成动态的序列向量,然后利用BiLSTM对序列向量进行特征提取与学习,最后通过SoftMax进行情感分类。同时,本文在上述基础上创新性地引入了模型不可知元学习算法以及两种对抗训练算法,快速梯度法(Fast Gradient Method, FGM)与投影梯度下降法(Projected Gradient Descent, PGD)。元学习在样本数量有限的情况下也具备很强的学习能力[14],能加快模型的收敛速度。而对抗训练是在原始序列向量的基础上利用特定算法对其添加微小的扰动生成对抗样本,将对抗样本作为输入进行训练可以有效降低模型对易受扰动特征的敏感程度[15]。本文在BERT-BiLSTM模型的基础上增添模型无关元学习和对抗训练,旨在提升模型的泛化性能。

2 相关技术

2.1 双向长短时记忆网络

长短时记忆网络LSTM在RNN的基础上,引入了多种门结构和记录历史信息的细胞状态C_t,其基本结构如图1所示。

LSTM利用门结构去维护和处理隐藏状态h_t与细胞状态C_t,其共有三种门结构,分别为遗忘门,输入门与输出门。

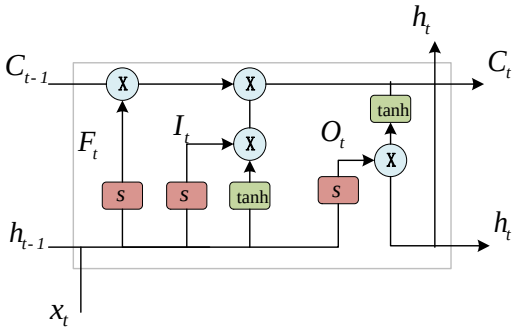

图1 LSTM基本结构

(1)遗忘门F_t利用上一个隐藏状态的信息h_{t-1}与当前时刻的输入x_t来决定从细胞状态中丢弃什么信息：

$$F_t = \sigma\left(w_F \cdot \left[h_{t-1}, x_t\right] + b_F\right) \tag{1}$$

(2)输入门I_t利用上一个隐藏状态的信息h_{t-1}与当前时刻的输入x_t来确定加入细胞状态的新信息：

$$I_t = \sigma\left(w_I \cdot \left[h_{t-1}, x_t\right] + b_I\right) \tag{2}$$

$$\tilde{C}_t = \tanh\left(w_C \cdot \left[h_{t-1}, x_t\right] + b_C\right) \tag{3}$$

(3)通过遗忘门的输出F_t与输入门的输出对细胞状态C_t进行更新：

$$C_t = F_t * C_{t-1} + I_t * \tilde{C}_t \tag{4}$$

(4)输出门O_t确定对输入的x_t的输出信息，并对当前时刻的隐藏状态h_t进行更新与输出：

$$O_t = \sigma\left(w_O \cdot \left[h_{t-1}, x_t\right] + b_O\right) \tag{5}$$

$$h_t = O_t * \tanh\left(C_t\right) \tag{6}$$

双向长短时记忆网络BiLSTM利用两个结构相同，方向相反的LSTM处理序列信息，能够同时利用到序列未来时刻与过去时刻的信息，使得最终得到的特征表达效果更加全面[24]。

2.2 BERT预训练语言模型

BERT的主体由多层Transformer编码器相连接而构成[16]，其结构如图2所示。

图2中T_{rm}即Transformer编码器，输入的字符向量E_i经过多层T_{rm}后输出最终得到的特征表达T_i。Transformer是基于自注意力机制(Self-Attention)实现的[17]，通过注意力机制，可以建立不同序列之间的关联性，使得网络可以长距离地捕获特征信息，提升训练效果。

Transformer编码器的结构如图3所示。编码器将输入的词嵌入向量与位置编码向量相加后送入多头注意力机制层，得到的结果送入前馈网络层，增加特征的非线性变化，然后得到最终的特征表达。另外，Transformer编码器还引入了残差连接与规范化[18]。残差连接能有效避免网络退化导致信息丢失，而规范化可以将参数数值控制在合理范围内，加速模型的收敛。

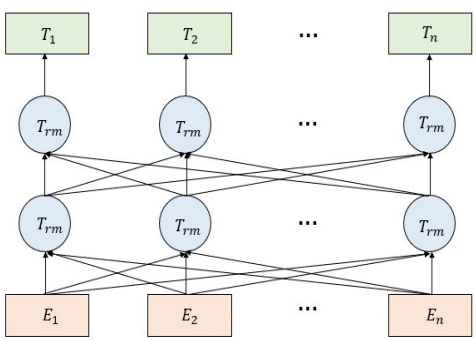

图2 BERT基本结构

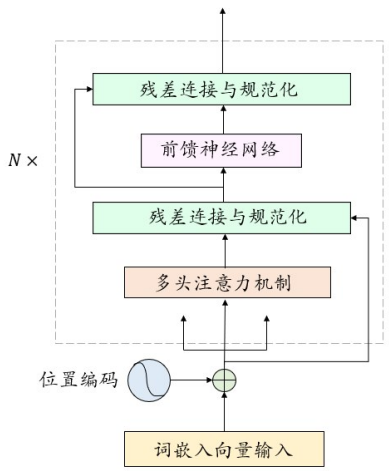

图3 Transformer编码器基本结构

2.3 模型无关元学习

模型无关元学习MAML(Model-Agnostic Meta-Learning)[14]的基本思想是寻求一个更好的网络初始化参数。相较于随机初始化，MAML可以通过更少的梯度下降步骤使得模型收敛，并且对数据样本的需要更少，其基本流程如图4所示。

MAML算法由内循环与外循环构成。内循环通过学习特定的任务并利用梯度下降最小化损失，找出每个任务的最优参数；外循环通过计算每个任务中相对于最优参数的梯度，更新模型的初始参数，这使得

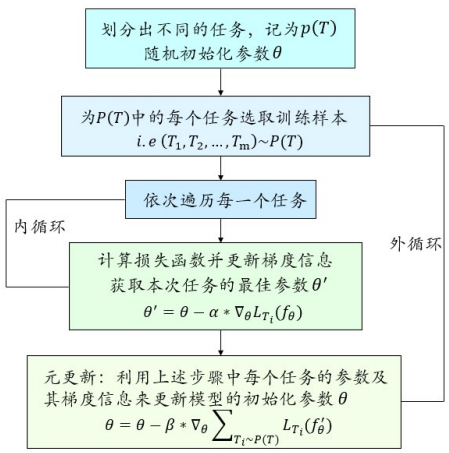

图4 MAML算法基本流程

我们可以将更新后的模型参数作为其初始值进而取代参数的随机初始化。

2.4 对抗样本生成算法

快速梯度法FGM(Fast Gradient Method)[19]是一种有效的对抗样本生成算法,其算法原理如图5所示。

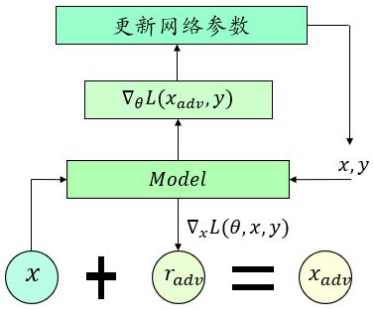

图5 FGM算法基本原理

FGM计算对抗扰动项r_{adv}的方法如下:

$$r_{adv} = \varepsilon * g/\|g\|^2 \tag{7}$$

$$g = \nabla_x L(\theta, x, y) \tag{8}$$

其中,ε为扰动半径;g为梯度信息。

投影梯度下降PGD(Projected Gradient Descent)也是一种有效的对抗样本生成算法[20],其相当于多步的FGM算法。对抗训练本质上是一个非凹的约束优化问题,FGM通过简单的梯度上升可能达不到约束内的最优点,而PGD可以通过多次扰动来接近目标值。PGD的算法原理如图6所示。

PGD计算对抗扰动项r_{adv}的方法如下:

$$x_{t+1} = \prod_{x+S} \frac{x_t + \alpha \cdot g(x_t)}{\|g(x_t)\|^2} \tag{9}$$

$$g(x_t) = \nabla_x L(\theta, x_t, y) \tag{10}$$

其中,$S = r \in R^d, \|r\|^2 \leqslant \epsilon$为扰动的约束空间;$\alpha$为扰动步长。

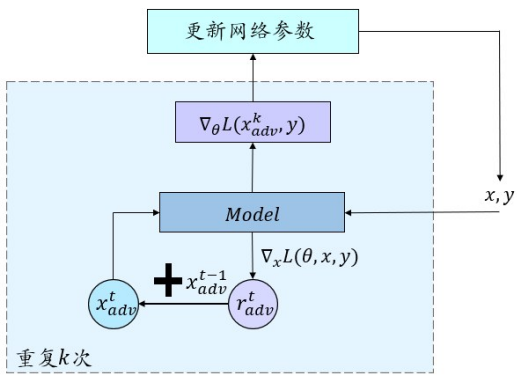

图6 PGD算法基本原理

3 模型结构

本文在上述基础上提出了融合模型不可知元学习与对抗训练的中文情感分析模型,模型的结构如图7所示。

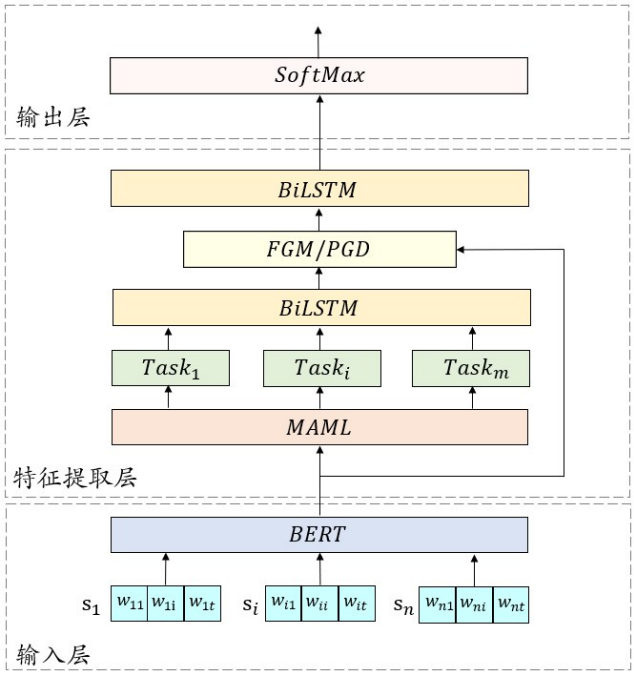

图7 本文模型结构

模型由以下部分组成:

(1)输入层。即文本序列的向量化层,需要将原始中文序列转化为特征提取层当中的BiLSTM网络能够接收并且处理的序列向量。

首先需要对数据集中的原始数据进行清洗,去除其中出现的英文字符,特殊字符以及无任何情感倾向的网页链接和邮箱地址等。假设文本d由n个中文序列组成,即$d = \{s_1, s_2, \cdots, s_n\}$;每个中文序列又由$t$个字符组成,则文本$d$中的第$i$个序列可以表示为$s_i = \{w_{i1}, w_{i2}, \cdots, w_{it}\}$,$w_{ij}$表示文本$d$当中第$i$个序列的第$j$个字,此时文本$d$的形状为$d_Shape = [n, t]$。然后利用BERT预训练语言模型可以得到文本$d$中每个字的向量表示:

$$d' = BertTokenizer(d) \tag{11}$$

$$Input = BERT(d') \tag{12}$$

BERT对w_{ij}生成维度为e的向量表示,$Input$为输入层的最终输出结果,其形状为$Input_Shape = [n, t, e]$。

(2)特征提取层。对$Input$进行特征提取与学习,分为四个步骤完成,如下所示:

①首先将$Input$送入MAML层,MAML将其划分

为m个任务：

$$\{T_1,\cdots,T_m\} = MAML(Input) \tag{13}$$

②然后对BiLSTM层的参数θ(θ主要包括LSTM的隐藏层信息h_0及细胞状态c_0)，依据标准正态分布进行随机初始化，分别求取每个任务的最优参数θ_T，利用每个任务的最优参数θ_T及其对应的梯度信息对BiLSTM层的参数θ进行元更新，计算过程如式(14)-(18)：

$$\theta \sim \{h_0, c_0\} \tag{14}$$

$$\theta = Randn(\theta) \tag{15}$$

$$f_\theta \sim \sum_{i=1}^{m} BiLSTM(T_i) \tag{16}$$

$$\theta_T = \theta - \alpha * \nabla_\theta L_{T_i}(f_\theta) \tag{17}$$

$$\theta' = \theta - \beta * \nabla_{\theta_T} \sum_{i=1}^{m} L_{T_i}(\theta_T) \tag{18}$$

其中，$Randn(x)$为对参数x进行随机标准正态分布赋值；α为MAML内循环的更新步长；β为MAML外循环的更新步长；$\nabla L(x)$为参数x对应的梯度信息。

③将输入层输出的*Input*与BiLSTM层经过元更新后的参数θ'及其梯度信息送入FGM/PGD层，计算对抗扰动，得到对抗样本。

FGM算法先对θ'求L2范数，然后利用θ'的梯度信息计算对抗扰动r_{adv}，利用r_{adv}去干扰BERT网络的词嵌入层参数θ_{em}，再利用干扰后的BERT即可生成对抗样本数据x_{adv}，计算过程如式(19)-(22)：

$$L_2 = Norm(\nabla L(\theta')) \tag{19}$$

$$r_{adv} = \frac{\nabla L(\theta')}{L_2} \tag{20}$$

$$\theta_{em} = r_{adv} + \theta_{em} \tag{21}$$

$$x_{adv} = BERT'(d) \tag{22}$$

PGD算法相当于多步的FGM，其对FGM计算r_{adv}的公式增添了扰动步长α，其余计算与FGM剩余部分相同，但PGD会重复进行K次扰动，计算过程如式(23)-(24)：

$$r_{adv}' = \frac{\alpha \cdot \nabla L(\theta')}{L_2} \tag{23}$$

$$Repeat(K) \sim x_{t+1} = \prod_{x} \frac{x_t + \alpha \cdot \nabla L(\theta')}{L_2} \tag{24}$$

④将得到的对抗样本数据x_{adv}送入BiLSTM层进行对抗训练，BiLSTM由正向LSTM与反向LSTM两部分组成，通过上文与下文更加全面地对数据进行特征提取，计算过程如式(25)-(27)：

$$\overrightarrow{\{h_t, c_t\}} = \overrightarrow{LSTM}(x_{adv}) \tag{25}$$

$$\overleftarrow{\{h_t, c_t\}} = \overleftarrow{LSTM}(x_{adv}) \tag{26}$$

$$Feature = Cat\left(\left[\overrightarrow{h_t}, \overleftarrow{h_t}\right]\right) \tag{27}$$

(3)输出层。将得到的特征表达*Feature*通过SoftMax层进行情感分类，得到最终的结果，计算过程如式(28)-(29)：

$$SoftMax(x) = \frac{e^{h(x,y_i)}}{\sum_{j=1}^{n} e^{h(x,y_i)}} \tag{28}$$

$$Result = SoftMax(Feature) \tag{29}$$

4 实验与分析

4.1 实验环境

实验环境如表1所示。

表1 实验环境

开发环境	参数
代码运行平台	Kaggle Notebook
GPU	GPU P100
内存	20GB
深度学习框架	PyTorch
编程工具	JetBrains Pycharms

4.2 超参数设置与评价标准

在深度学习当中，超参数是在训练开始之前预先设置好的数据，而非通过训练得到的数据，通常情况下，对超参数的优化能够提升模型的性能与效果。本模型超参数的设置如表2所示。

本文选取了模型的分类准确率以及F1-Score作为评价标准。F1-Score同时兼顾了模型的分类精确率和召回率，其计算公式如下：

$$F1\text{-}Score = \frac{2*Precision*Recall}{Precision + Recall} \tag{30}$$

表2 超参数设置

参数名称	参数取值
学习率	0.00002
Batch Size	32
Dropout	0.5
迭代次数	30
最优化算法	Adam
最大序列长度	200
PGD扰动次数	3
PGD扰动步长	0.3
MAML任务数量	5
MAML内循环步长	0.00002
MAML外循环步长	0.00002

4.3 实验数据集

文章共使用三个数据集,前两个数据集来源于GitHub平台上开源的两个中文情感数据集。第一个数据集的主题是关于餐馆评价,共计有2000条数据,记为D_{s1};第二个数据集的主题是微博上关于新冠疫情的评论,共计有8606条数据,记为D_{s2};第三个数据集来自NLPCC 2013会议官网公开的中文微博情绪识别的数据集,共计有4000条数据,记为D_{s3}。

D_{s1}的部分数据如表3所示,数据集总体分布如图8所示。

表3 D_{s1}数据集部分数据

文本内容	情感标签
在自助餐里算是一般的,好在价格便宜种类多	积极
还不错可以再去,希望过节有惊喜	积极
很不错,服务好,环境好,东西也好吃	积极
菜品丰富质量好,服务也不错! 很喜欢!	积极
70+一人,到了里面发现可吃的其实并不多	消极
来上菜的时候,有好几次都弄到我们身上	消极
量太少了,味道也不好,好评都是刷的吧	消极
味道太甜,上菜太慢。。。。。。	消极

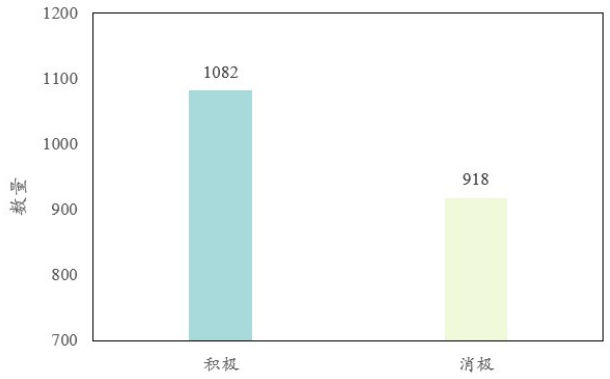

图8 D_{s1}数据集分布

D_{s2}的部分数据如表4所示,数据集总体分布如图9所示。

表4 D_{s2}数据集部分数据

文本内容	情感标签
!!! 一定会好起来的	开心
这次的疫情,哪家旅游软件做得良心,我以后就专门用哪家了。 被扣除了手续费真不爽!	愤怒
#海南封岛#	紧张
江苏挺住啊[泪][泪][泪]	悲伤
数字增长好吓人,希望部分人有点自觉	惊讶
太难了...估计都窝在家里,也没地方敢随便去	悲伤
绍兴! 我慌了啊!	恐惧

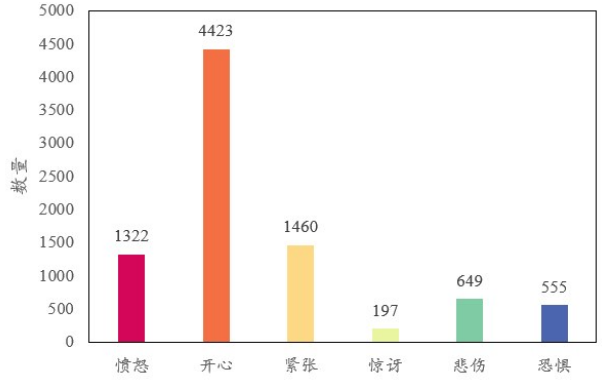

图9 D_{s2}数据集分布

D_{s3}的部分数据如表5所示,数据集总体分布如图10所示。

表5 D_{s3}数据集部分数据

文本内容	情感标签
最讨厌那些所谓的圣人! 装得自己很厉害,认都不认识他却一直对你指手画脚,装的和你很熟的样子。你258,不需要你来指示我! 变态!	厌恶
跟我开玩笑吗? 学校今年不做优秀毕业生评选? 三年的准备,就为这个啊,有没有搞错啊? 玩我?	生气
成年人身上有些孩子气,或者孩子身上有些成人气,都是十分可爱的元素。懂得欣赏这些元素的人,是内心世界丰富的人,也是离幸福体验最近的人。	喜欢
埋头吃着中午带的便当,尝了一块昨晚公公烧的红烧鸭肉,异常好吃啊!!!。	开心
人生不止,寂寞不已。寂寞人生爱无休,寂寞是爱永远的主题、我和我的影子独处、它说它有悄悄话想跟我说、它说它很想念你,原来我和我的影子,都在想你。	悲伤
我就纳了闷了,从渤海湾到雷州半岛,整个的临海被人用岛链加航母封锁的严严实实,这些钢盔加盾牌哪去了?	惊讶
《解梦工具书》上说,蛇在梦中出现是典型的具有性意味的意象,梦中的蛇可能暗示你对性的感觉。	中性
昨晚又看了恐怖片[泪]还是开灯滚床看电影把。	恐惧

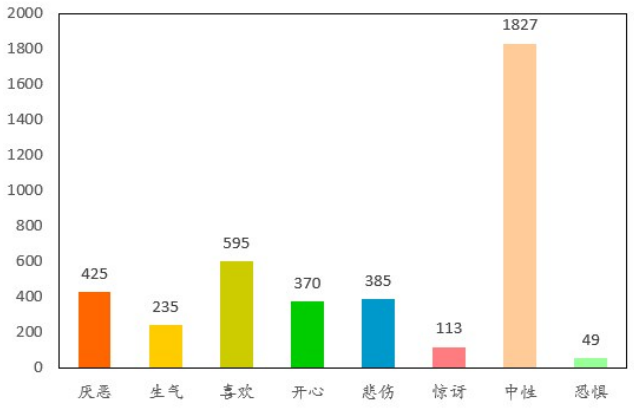

图10 D_{s3}数据集分布

4.4 实验分析与讨论

本文设置了四组对比实验，来验证模型的有效性，分别如下所示：

第一组 不同Batch Size的对比实验。为了验证不同Batch Size对模型特征学习能力的影响，本文选取BERT+BiLSTM模型在D_{s1}数据集下，保持其他超参数相同，分别设置不同的Batch Size进行对比实验，实验结果如表6和图11所示。

表6 第一组对比实验结果

Batch Size	训练总时长(单位/秒)
8	1663.2907
16	1374.2697
32	1043.7001
64	873.3195
100	864.8852

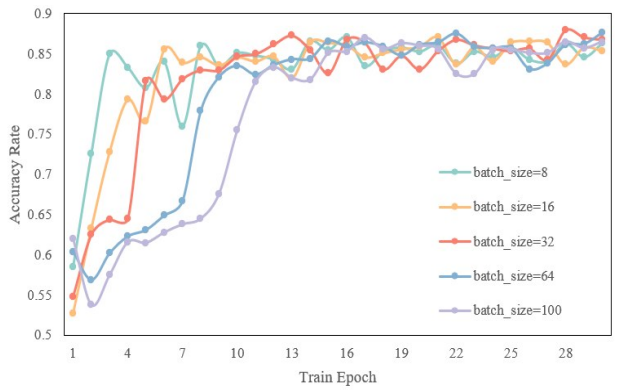

图11 第一组对比实验结果

由表6和图11可以看出，Batch Size影响模型的学习效率和程序运行时间。Batch Size值过小时，模型的学习效率较高，但模型总的训练时间较长；Batch Size值过大时虽然模型的训练时间明显缩短了，但由于数据量的增大，模型的学习效率变得很低。综合二者考虑，当Batch Size值设置为32时是最为合适的，模型的学习效率相较Batch Size为8和16的时候相差不大，并且随着迭代次数的增加，准确率还超过了其余四组；同时程序总体运行时间也在可接受范围内。对此可以看出，设置合适的Batch Size能有效提升模型的训练效率。

第二组 不同PGD扰动次数对比实验。为了验证PGD算法不同扰动次数对模型性能的影响，本文选取BERT+BiLSTM+PGD模型在D_{s1}数据集下，保持其他超参数相同，设置不同的PGD扰动次数进行对比实验，实验结果如图12所示。

由图12可以看出，PGD算法的扰动次数会影响模型的性能。当扰动次数设置为3时，ROC曲线与横轴的面积最大，模型的效果最优。当扰动次数过小时，PGD算法可能会同FGM算法般陷入局部最优，影响模型训练；当扰动次数过多时，扰动范围过大，对模型的干扰程度较为严重。对此可以看出，设置合适的扰动次数才能有效发挥PGD的优势。

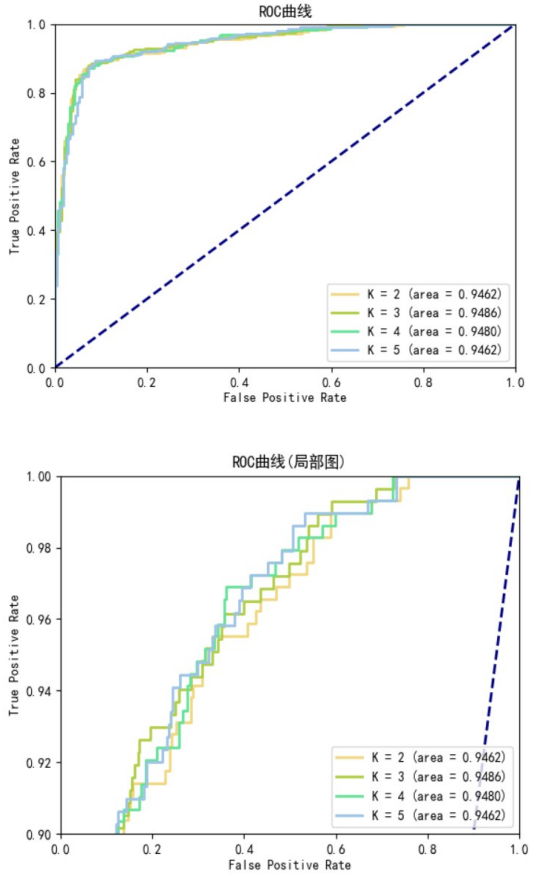

图12 第二组对比实验结果

第三组 MAML有效性对比实验。为了验证MAML能否有效加速模型的收敛，在相同的实验环境下以D_{s3}作为训练数据集，分别对BERT+BiLSTM和BERT+BiLSTM+PGD模型添加或不添加MAML进行训练，得到的实验结果如图13所示。

由图13可以看出，没有添加MAML的模型的前几轮迭代的准确率都比较低，其需要更多次的迭代训练才能逼近最优性能，实现模型收敛；而添加了MAML的模型的初始准确率更高，模型收敛得更加快速。对此，在本模型当中MAML确实能发挥其自身优势，可以利用更少的梯度下降步骤使得模型达到收敛状态。

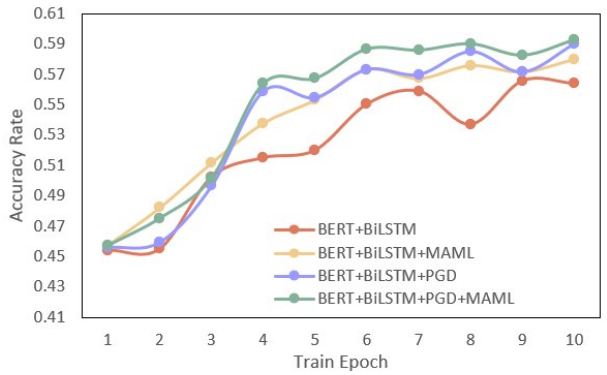

图13 第三组对比实验结果

第四组 不同模型对比实验。为了验证MAML与对抗训练对模型分类准确度的影响,在相同的实验环境下分别以D_{s1},D_{s2}和D_{s3}作为训练数据集,以BERT+BiLSTM模型为基础,对模型分别或同时添加MAML与对抗训练进行实验,得到的实验结果如表7,表8和表9所示。

表7 第四组对比D_{s1}实验结果

模型	准确率	F1-Score
BERT+BiLSTM	88.02%	0.8767
BERT+BiLSTM+MAML	89.58%	0.8946
BERT+BiLSTM+FGM	90.10%	0.9005
BERT+BiLSTM+PGD	**91.15%**	**0.9123**
BERT+BiLSTM+MAML+FGM	89.58%	0.8959
BERT+BiLSTM+MAML+PGD	90.13%	0.8999

表8 第四组对比D_{s2}实验结果

模型	准确率	F1-Score
BERT+BiLSTM	76.29%	0.5923
BERT+BiLSTM+MAML	76.35%	0.5931
BERT+BiLSTM+FGM	**77.50%**	0.6151
BERT+BiLSTM+PGD	76.83%	0.6447
BERT+BiLSTM+MAML+FGM	77.38%	**0.6939**
BERT+BiLSTM+MAML+PGD	77.14%	0.6311

表9 第四组对比D_{s3}实验结果

模型	准确率	F1-Score
BERT+BiLSTM	58.02%	0.3688
BERT+BiLSTM+MAML	58.10%	0.3691
BERT+BiLSTM+FGM	**59.29%**	0.3775
BERT+BiLSTM+PGD	58.61%	0.3751
BERT+BiLSTM+MAML+FGM	59.04%	**0.3906**
BERT+BiLSTM+MAML+PGD	58.45%	0.3746

(1)BERT+BiLSTM:使用BERT生成字向量,送入BiLSTM网络进行特征提取后,利用SoftMax进行情感分类。

(2)BERT+BiLSTM+MAML:使用BERT生成字向量,送入BiLSTM网络进行特征提取,同时在特征提取过程中加入MAML进行模型参数更新,利用SoftMax进行情感分类。

(3)BERT+BiLSTM+FGM:使用BERT生成字向量,送入BiLSTM网络进行特征提取,同时利用FGM算法生成对抗样本对模型进行对抗训练,利用SoftMax进行情感分类。

(4)BERT+BiLSTM+PGD:使用BERT生成字向量,送入BiLSTM网络进行特征提取,同时利用PGD算法生成对抗样本对模型进行对抗训练,利用SoftMax进行情感分类。

(5)BERT+BiLSTM+MAML+FGM:使用BERT生成字向量,送入BiLSTM网络进行特征提取,在特征提取过程中加入MAML进行模型参数更新,同时还利用FGM算法生成对抗样本对模型进行对抗训练,利用SoftMax进行情感分类。

(6)BERT+BiLSTM+MAML+PGD:使用BERT生成字向量,送入BiLSTM网络进行特征提取,在特征提取过程中加入MAML进行模型参数更新,同时还利用PGD算法生成对抗样本对模型进行对抗训练,利用SoftMax进行情感分类。

从表7,表8和表9可以看出,增添了对抗训练的模型相较于基础模型,准确率与F1-Score均有所提升。D_{s1}中仅添加PGD对抗训练的模型的准确率和F1-Score最高,相较于基础模型提升了3.13%和0.0356;D_{s2}中仅添加FGM对抗训练的模型的准确率最高,相较于基础模型提升了1.21%,而同时添加了MAML与FGM对抗训练的模型的F1-Score最高,相较于基础模型提升了0.1002。D_{s3}中仅添加FGM对抗训练的模型的准确率最高,相较于基础模型提升了1.27%,而同时添加了MAML与FGM对抗训练的模型的F1-Score最高,相较于基础模型提升了0.0218。

综合三次实验结果可以看出,对抗训练能有效提升模型的性能,并且相较于PGD算法,FGM算法在特征复杂度更大的数据集上表现出的效果更好。本文引入MAML的目的更多的是研究其在模型训练效率方面的影响,MAML算法并没有对模型的参数组成和网络结构进行修改,而是作为一种训练策略去辅助模型进行训练,可以将其理解为一种新的模型优化方法,对此表格中展示的引入MAML的模型的指标并没有显著提升这一现象,是可以解释的。在本组对比实验中加入MAML的目的是研究MAML在提升模型

训练效率的前提下，是否会干扰模型的最终分类准确率。从表格中展示的数据可以看出，相较于没有添加MAML的模型，增加了MAML后的指标并没有很明显的降低趋势，降低的误差值在可接受的范围内，并且增添了MAML的模型的指标均不差于原始模型，有的甚至还有所提高。证明了MAML在发挥其更好地辅助模型训练这一特性的前提下，也不会对模型的最终性能造成很严重的干扰。

虽然在上述三个表格当中，添加了对抗训练的模型相较于基础模型，性能指标均呈现增长的趋势，但也不难发现，随着数据集规模的增大以及情感类别的增多，增长的幅度越来越低。D_{s1}的主题是餐馆评价，其仅有两种情感类别，对此模型在D_{s1}上的性能表现更好。相比之下，D_{s2}和D_{s3}的情感类别更多，故模型在二者上表现出的效果会差于D_{s1}，同时由于D_{s3}的数据类型更加宽泛，没有像D_{s2}一样聚焦某类主题的数据，对此模型在D_{s3}上的效果要差于D_{s2}。

文章研究的基础模型BERT+BiLSTM主要是依赖BiLSTM网络对文本进行特征学习，BERT更多地是辅助BiLSTM进行情感分类。BiLSTM的学习能力有限，因此随着数据复杂程度的增大，BiLSTM便出现了上述性能饱和的状态。对抗训练是在基础模型上对其进行性能优化，若基础模型的性能已达到饱和状态，对抗训练对其的提升幅度也就相对应的有所降低。

5 结论

本文的实验结果说明，在不同的情感类别上，对抗训练均能有效提升模型的准确率与F1-Score指标；同时模型无关元学习也表现出了出色的模型优化的能力，增添了MAML的模型能够利用更少的迭代次数实现模型收敛。本文的结论为情感分析领域的发展提供了研究方向，如引入对抗训练帮助模型更好地应对真实世界中的挑战性情感表达以及利用元学习帮助模型更快地适应新任务并具备更强的泛化能力等。

针对深度学习模型易受干扰等特性，本文提出的模型在微信和微博等信息量复杂且数量庞大的平台上更能发挥其自身优势。与其他模型相比，本文模型的优势在于，对抗训练能有效提升模型对垃圾信息的处理能力，而模型无关元学习能加快模型对信息的学习效率。

但受限于设备和技术等因素的影响，我们没有对元学习进行更深入的研究，且模型的讨论仅局限在中文领域，没有在不同语言上进行研究。同时由于本文模型是以BiLSTM作为特征学习的核心，BiLSTM学习能力较低也是本文模型的局限。

对于本文研究的后续工作，可以考虑将不同数据集进行融合或者使用多种语言的数据集，研究元学习在多任务问题上的性能；同时利用规模更大的Transformer，GPT和T5等预训练语言模型去替换BiLSTM，在保证模型学习能力足够的前提下，研究对抗训练对模型性能的提升幅度。

当下自然语言处理领域正朝着大规模预训练语言模型的方向发展，模型的训练是通过自监督学习进行的，中文情感分析在未来的发展中可以利用自监督方法来进行情感学习，能有效减少模型对标记数据的依赖；同时随着社交媒体和在线内容的多样性增加，情感分析势必会朝着多模态的方向发展，不仅要考虑文本信息，还要结合图像，音频和视频等其他模态的数据。未来情感分析的发展趋势将会着重于如何有效地融合多个模态的信息，建立更加全面更加准确的情感分析模型。

参考文献(References)：

[1] 罗浩然，杨青. 基于情感词典和堆叠残差的双向长短期记忆网络的情感分析[J]. 计算机应用，2022，42(4)：1099-1107.

[2] Pang B，Lee L.A sentimental education：sentiment analysis using subjectivity summarization based on minimum cuts [C]//Proc of the ACL，2004：271-278.

[3] Alharbi N M，Alghamdi N S，Alkhammash E H，et al. Evaluation of sentiment analysis via word embedding and RNN variants for amazon online reviews[J].Mathematical Problems in Engineering，2021，Special Issue：1-10.

[4] Lee G T，Kim C O，Song M. Semisupervised sentiment analysis method for online text reviews[J]. Journal of Information Science，2021，47(3)：387-403.

[5] Kim Y.Convolutional neural networks for sentence classification[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing，2014：1746-1751.

[6] De Rumelhart，Hinton G E，Williams R J. Learning representations by back propagating errors[J]. Nature，1986，323(6088)：533-536.

[7] Mikolov T，Karafiát M，Burget L. Recurrent neural network based language model[C]//Eleventh Annual Conference of the International Speech Communication Association，2010.

[8] Sak H，Senior A，Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic

modeling[C]//Proceedings of the 15th Annual Conference of the International Speech Communication Association (ISCA), 2014: 338-342.

[9] Tang D, 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, 2015: 1422-1432.

[10] Devlin J, Chang M W, Lee K, et al. BERT: pretraining of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, 1(Long and Short Papers): 4171-4186.

[11] 赵宏, 傅兆阳, 赵凡. 基于BERT和层次化Attention的微博情感分析研究[J]. 计算机工程与应用, 2022, 58(05): 156-162.

[12] 安胜彪, 郭昱岐, 白宇, 等. 小样本图像分类研究综述[J]. 计算机科学与探索, 2023, 17(03): 511-522.

[13] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[DB/OL]. arXiv: 1412.6572, 2014.

[14] Finn C, Abbeel P, Levine S. Model-Agnostic Meta-learning for fast adaptation of deep networks[C]//International Conference on Machine Learning, 2017, 70: 1126-1135.

[15] 金志刚, 周峻毅, 何晓勇. 面向自然语言处理领域的对抗攻击研究与展望[J]. 信息安全研究, 2022, 8(03): 202-211.

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

[17] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in Neural Information Processing Systems, 2014: 3104-3112.

[18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017: 5998-6008.

[19] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations, 2015.

[20] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[C]//International Conference on Learning Representations, 2018.

[21] 杨春霞, 姚思诚, 宋金剑. 一种融合字词信息的中文情感分析模型[J]. 计算机工程与科学, 2023, 45(03): 512-519.

[22] 曾余洋. 基于深度学习的中文文本情感分析研究[D]. 雅安: 四川农业大学, 2022.

[23] 蒋玲玲, 罗娟娟, 朱玉鹏, 周东青. 基于深度学习的对抗攻击技术综述[J]. 航天电子对抗, 2023, 39(01): 10-18+50.

[24] 汪林, 蒙祖强, 杨丽娜. 基于多级多尺度特征提取的CNN-BiLSTM模型的中文情感分析[J]. 计算机科学, 2023, 50(05): 248-254.

[25] 赵宏, 傅兆阳, 王乐. 基于特征融合的中文文本情感分析方法[J]. 兰州理工大学学报, 2022, 48(03): 94-102.

编辑: 王谦