计算机科学
COMPUTER SCIENCE

基于大批次对抗策略和强化特征提取的文本情感分类方法

陈嘉昊, 段利国, 常轩伟, 李爱萍, 崔娟娟, 郝渊斌

引用本文

陈嘉昊, 段利国, 常轩伟, 李爱萍, 崔娟娟, 郝渊斌. 基于大批次对抗策略和强化特征提取的文本情感分类方法[J]. 计算机科学, 2025, 52(10): 247-257.

CHEN Jiahao, DUAN Liguo, CHANG Xuanwei, LI Aiping, CUI Juanjuan, HAO Yuanbin. Text Sentiment Classification Method Based on Large-batch Adversarial Strategy and EnhancedFeature Extraction [J]. Computer Science, 2025, 52(10): 247-257.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240800061

基于大批次对抗策略和强化特征提取的文本情感分类方法

陈嘉昊[1] 段利国[1,2] 常轩伟[1] 李爱萍[1] 崔娟娟[1] 郝渊斌[1]
1 太原理工大学计算机科学与技术学院 太原 030024
2 山西电子科技学院 山西 临汾 041000
(1931668813@qq.com)

摘 要 文本情感分类任务旨在对短文本语句进行分析并判断其对应的情感类别。为解决现有模型在情感分类方面缺乏大规模高质量语料数据集、文本特征非均匀重要性提取不足等问题,提出了一种基于大批次对抗策略和强化特征提取的文本情感分类方法。首先将文本数据集输入预训练语言模型BERT中,得到相应的词嵌入向量表示;再利用BiLSTM进一步学习序列中的上下文依赖关系;之后将局部注意力机制与TextCNN的局部感受野加权结合,实现强化特征提取能力;最后将BiLSTM的输出与TextCNN的输出进行拼接,得到两个空间的深层特征融合,再交由分类器进行情感分类的判断。整个训练过程采取大批次对抗策略,在词嵌入空间中加入对抗性扰动并进行多次迭代,进而提高模型的鲁棒性。在多个数据集上的实验结果验证了该模型的有效性。

关键词:短文本;情感分类;对抗策略;特征提取;词嵌入
中图分类号 TP391

Text Sentiment Classification Method Based on Large-batch Adversarial Strategy and Enhanced Feature Extraction

CHEN Jiahao[1],DUAN Liguo[1,2],CHANG Xuanwei[1],LI Aiping[1],CUI Juanjuan[1] and HAO Yuanbin[1]
1 College of Computer Science and Technology,Taiyuan University of Technology,Taiyuan 030024,China
2 Shanxi Electronic Science and Technology Institute,Linfen,Shanxi 041000,China

Abstract The text sentiment classification task aims to analyze short text sentences and determine their corresponding sentiment categories. In order to solve the problems of lack of large-scale high-quality corpus dataset and insufficient non-uniform importance extraction of text features in the existing models in sentiment classification,this paper proposes a text sentiment classification method based on large-batch adversarial strategy and enhanced feature extraction. Firstly,the text dataset is input into the pre-trained language model BERT to obtain the corresponding word embedding vector representation,and then the BiLSTM is used to further learn the context dependencies in the sequence. Then,the local attention mechanism is combined with the local receptive field weighting of TextCNN to enhance the feature extraction ability. Finally,the output of BiLSTM and the output of TextCNN are spliced to obtain the deep feature fusion of the two spaces,which are handed over to the classifier for the judgment of sentiment classification. In the whole training process,a large-batch adversarial strategy is adopted,and adversarial perturbations are added to the word embedding space and multiple iterations are carried out to improve the robustness of the model. Experimental results on multiple datasets verify the effectiveness of the proposed model.

Keywords Short text,Sentiment classification,Adversarial strategy,Feature extraction,Word embeddings

1 引言

社交媒体平台在国内的快速发展,已经成为网民分享日常生活、交流独特思想的重要渠道。然而,随着信息爆发式的增长,如何从海量的社交媒体平台文本中提取有价值的信息,并且协助政府理解和预测公众情绪的变化,具有重要的研究意义和社会价值。自然语言处理(Natural Language Processing,NLP)技术的发展,尤其是深度学习在文本分类领域的应用,为解决这一问题提供了可能。

在情感分析的研究领域,传统的基于机器学习的情感分类方法虽然提高了准确率,但是仍然需要手动进行标注来提取文本特征,并且还需要训练分类器进行情感分类,这些方法存在泛化能力有限以及所需处理成本较高、处理效率低下等问题。现如今大部分的研究集中于深度学习模型,BERT(Bi-

到稿日期:2024-08-12 返修日期:2024-11-30
基金项目:山西省自然科学基金(202203021221234,202303021211052)
This work was supported by the Natural Science Foundation of Shanxi Province,China(202203021221234,202303021211052).
通信作者:段利国(zhaixing202202@163.com)

directional Encoder Representations from Transformer)[1]模型作为一种预训练语言模型,具有强大的语义理解和表达能力,但是当缺乏大规模高质量训练数据时会限制其泛化能力,可能会影响模型在下游任务中的表现。扩展和丰富训练过程则可以避免这种情况,通过使用自对抗训练(Self-Adversarial Training)[2]模拟潜在的攻击场景促进模型学习到更加强大的特征表示,可以弥补高质量训练数据稀疏的缺陷,有利于促进模型对语义和语法知识的理解,提高情感分类的准确性。长短期记忆神经网络 LSTM(Long Short Term Memory)[3]的使用也大大增强了情感分类成功的结果,但 LSTM 是单向的,只能捕捉序列的单侧上下文,无法完全理解某些需要前后文信息才能准确预测的情况。因此,在大多数情况下使用 LSTM 的变体双向长短期记忆神经网络 BiLSTM(Bidirectional Long Short-Term Memory)[4],利用双向处理信息,能够同时获取到序列的前续和后继上下文信息,可 BiLSTM 设计的初衷是用于捕捉序列数据中的长期依赖关系,不擅长捕捉局部特征和短距离依赖。实验发现,在某些情况下相同情感词语在不同语句和语境下所代表的情感并不相同。总的来说,现有的模型在情感分析方面有着以下不足之处:

1)在缺乏大规模数据集语料和训练过程不够丰富的情况下,模型对信息捕捉不够充分,鲁棒性和泛化性能不足。

2)模型对局部语义信息和文本非均匀重要性的理解不足,对文本中某些对于分类任务比其他部分更重要的信息的认识有缺陷。

3)当相同的词语出现在不同的语句和语境中时,模型的理解不够充分,容易出现情感判断失误的情形。

鉴于以上缺陷,本文引入大批次对抗策略进行改进。自对抗训练属于数据增强技术的一种,通过生成对抗性样本,利用神经网络自身进行对抗性攻击产生的对抗性扰动来提高模型的鲁棒性和泛化能力。大批次对抗策略属于自对抗训练中 Projected Gradient Descent(PGD)[5]的一种优化。通过整合现有的资源,本文提出了基于大批次对抗策略和强化特征提取的情感分类方法。首先将文本数据集输入预训练语言模型 BERT 中,得到相应的词嵌入向量表示,再利用 BiLSTM 进一步学习序列中的时序特征,强化模型表达。为了解决情感分析中局部语义单元(情感词/修饰词/标点)的细粒度表征问题,以及词汇情感极性随语境动态变化的问题,将局部注意力机制(Local Attention Mechanism)[6]与 TextCNN(Text Convolutional Neural Network)[7]的局部感受野加权结合,提供更全面的序列表达,生成更加丰富和信息密集的特征表示。最后,将 BiLSTM 的输出与 TextCNN 的输出进行拼接,得到两个空间的深层特征融合,再交由分类器进行情感分类的判断。在整个训练过程中采取大批次对抗策略,在每次迭代中同步更新模型参数和输入扰动,通过减少梯度计算次数来加速对抗训练过程,进而有效地提高模型的鲁棒性。实验结果表明,基于大批次对抗策略和强化特征提取的情感分类方法能改善现有模型在情感分析方面的不足,在多个分类任务中取得了最好的结果。

本文有以下 3 点贡献:

1)为了解决模型缺乏大规模高质量语料的问题,本文提出了在训练过程中采取大批次对抗策略的方法。

2)将局部注意力机制与 TextCNN 的局部特征提取进行加权结合,实现强化特征提取的能力。

3)在中文微博数据集、酒店和外卖评论数据集上进行实验,与众多模型对比,本文模型均取得了最好效果,足以证明该模型的优越性。

2 相关工作

近年来,随着深度学习的不断发展,各种优秀的模型相继被提出,大大提高了情感分类成功的准确率。

早期的情感分析主要依赖于预先定义好的包含具有情感倾向的词汇及其情感极性标签的情感词典,然后根据这些词汇的情感倾向来预测整个文本的情感。2016 年,Kreutz 等提出了结合领域特定的词嵌入和标签传播算法的 SENTPROP 框架,使用少量种子词来准确地引导领域特定情感词典的生成[8]。同年,Teng 等[9]提出了基于上下文敏感的情感词典方法,使用加权和模型与 BiLSTM 来预测每个情感词的权重和句子级情感偏差分数。这些方法可以准确地反映文本的非结构化特征,易于分析和理解。但是,基于情感词典的方法存在很多的局限性,例如对新词和网络用语的识别能力有限,对词典的依赖性较强,以及在跨领域和跨语言的应用中效果不佳。

随着机器学习技术的发展,深度学习的出现为情感分析带来了革命性的变化。2018 年,Devlin 等提出了跨时代的预训练语言模型 BERT,旨在通过联合考虑所有层中的上下文信息来预训练深度双向表示,证明了语言模型预训练对改进多种自然语言处理任务的有效性。之后,多种以 BERT 为基础的模型相继被提出,例如 RoBERTa 通过动态掩码、更大的训练数据集和更大的批处理大小来优化 BERT 的预训练过程[10];ALBERT 通过跨层参数共享和嵌入层参数分解来减少模型的参数量,同时保持了 BERT 的双向上下文理解能力[11];ELECTRA 使用一个生成器和一个判别器,通过替换标记检测任务进行预训练[12];SpanBERT 通过改进的预训练任务,专注于问答和关系提取任务,提高了模型对句子结构的理解[13];DistilBERT 通过知识蒸馏技术来减小模型规模,同时保持了 BERT 的性能[14]。尽管 BERT 及其变体在自然语言处理领域取得了显著的成就,但这些模型的决策过程不够透明,难以解释模型的预测,其性能在很大程度上依赖于预训练阶段使用的数据质量,限制了模型在特定领域上的应用。2021 年,Feng 等[15]提出了 MCNN-MA 模型,该模型将单词特征与词性、位置和依赖句法特征结合,形成 3 个新组合特征后输入多通道 CNN 中,并集成了多头注意力机制。Peng 等[16]提出了 MSCNN-SPU 模型,利用多尺度卷积层(MSCNN)获取音频和文本的隐藏表示,双向循环神经网络(BiRNN)和注意力机制作为语音情感识别的标准方法,将它们融合用于情感分类任务。为了解决模型在中文情感分类任务中的不足,以及提高多任务学习在模型性能上的潜力,Yang 等[17]提出了多任务学习模型 LCF-ATEPC,将多头注意力机制、BERT 模型以及局部上下文聚焦机制整合在一起来强化特征提取。

2022 年,Zhao 等[18]提出了一种名为 CISS 的新框架,利

用因果推断来识别机器学习模型中的对抗性和脆弱性来源，并提出通过随机化分类器模拟因果效应来实现鲁棒性。Xu等[19]在解决深度神经网络对不可见扰动的脆弱性时，提出了A^2，它是一个高效的自动化攻击者，用于在对抗训练过程中实时生成最优扰动，以增强模型的鲁棒性。Ma等[20]提出了一种新的对抗训练方法FAT，通过实验展示了对抗训练中不同扰动半径对鲁棒公平性的影响，有效减少了由于不同类别之间标准准确性与鲁棒准确性不一致所引发的鲁棒公平性问题。Li等[21]提出了一种非自治神经常微分方程（ASODE），并对其相应的线性时变系统施加约束，确保所有清洁实例成为其渐近稳定平衡点，以防御对抗性攻击。2023年，为了解决依赖树作为句法结构，与情感分类任务的语义特性之间不匹配的问题，Ma等提出了APARN模型，通过路径聚合和关系增强的自注意力机制来提取语义特征[22]。Zhang等[23]提出的关键信息提取算法为本文的文本强化特征提取提供了一定的思路。Liu等[24]提出的注意力改进方法也给予了本文一定参考性。

上述模型虽然一定程度地提高了情感文本分类任务的效果，但归根结底，对相同词语在不同场景的理解和在提取文本特征以及大规模语料方面仍然具有很大的局限性。因此，本文将诸多模型和方法整合在一起，并在某些方面进行改进，使所提方法的性能得到了提升。

3 文本情感分类模型

在情感分析任务中，通常给定一个代表情感极性的标签和需要分类的句子，利用模型将句子分类到对应的情感极性。给定一个数据集样本(T,S)，其中，$T=\{t_1,t_2,\cdots,t_n\}$代表具有情感极性的标签，n代表标签的个数；$S=\{s_1,s_2,\cdots,s_i,\cdots,s_m\}$代表需要分类的句子，$s_i$表示该句话的第$i$个字符，$m$表示句子的长度。

模型结构如图1所示，共分为3层，分别为嵌入学习层、强化提取层以及分类层。嵌入学习层利用BERT和BiLSTM获取文本的词嵌入向量，强化提取层将局部注意力机制与TextCNN模型加权结合来强化特征提取，分类层得到最终的结果。整个模型的训练过程利用大批次对抗策略进行数据增强。

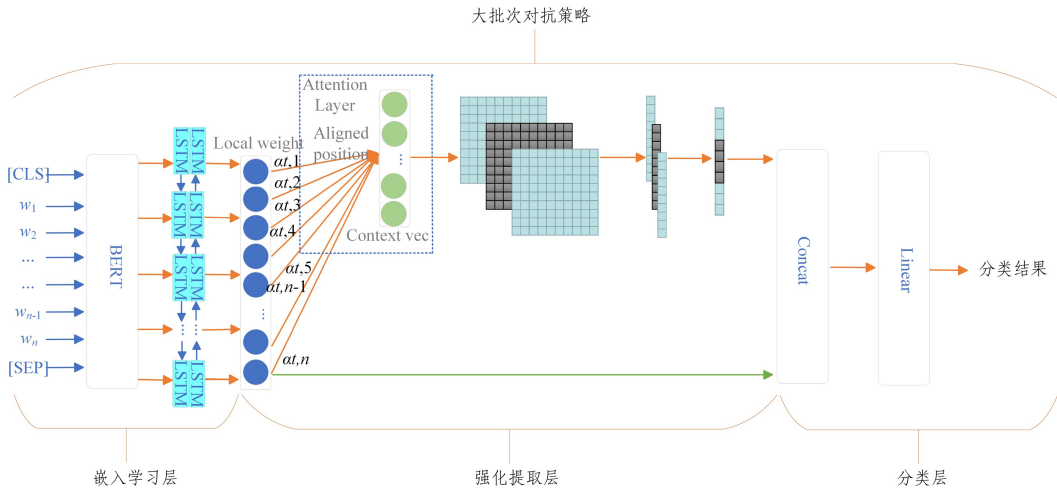

图1 模型结构图

Fig. 1 Structure diagram of model

3.1 嵌入学习层

嵌入层将词汇表中的每个词或字符映射到一个高维空间中的连续向量，同时捕捉词与词之间的语义关系。在文本预处理阶段，采用BertTokenizer分词器，它基于WordPiece算法[25]而创建，基础版本的词汇表大小为21128。WordPiece是一种子词粒度的tokenize算法，它在做合并时，需要找到能够最大化训练集数据似然的集合。在BertTokenizer中，标点符号、生僻字等未出现的token被[UNK]代替，中文被拆分成了字的形式。如：

Text：你知道“鋆”字怎么念吗？

Tokens：你，知，道，[UNK]，[UNK]，[UNK]，字，怎，么，念，吗，[UNK]

具体的分词流程如图2所示。经过预处理切词分词等操作后，得到输入文本的序列化表示$T=\{X_1,X_2,\cdots,X_n\}$，其中X_i表示文本中的第i个分词。在把数据输入BERT之前，首先将其输入标记嵌入层、片段嵌入层以及位置嵌入层，得到对应的嵌入表示$E=\{E_1^a,E_2^a,\cdots,E_n^a\}$，其中，$E_i$表示文本中的第$i$个词语的嵌入向量，$a$表示具体的某一层。最后，累加所有的嵌入表示作为$BERT$的输入表示，即：

$$BERT_{\text{in}}=\{[\text{CLS}],W_1,W_2,\cdots,W_n,[\text{SEP}]\} \tag{1}$$

其中，W_i表示第i个词语的累加嵌入向量。输入经过Encoder编码得到词语级别的输出，表示为$O=\{O_1,O_2,\cdots,O_n\}$，其中O_i代表第i个词语的表示向量。

输入文本经过预训练模型BERT编码后，生成深度的双向语言表征，得到词与词之间的复杂关系。学习层采用BiLSTM处理序列数据，通过其门控机制捕捉长距离依赖关系。LSTM在某时刻的计算式如下：

$$f_t=\sigma(W_f\cdot[h_{t-1},o_t]+b_f) \tag{2}$$

$$i_t=\sigma(W_i\cdot[h_{t-1},o_t]+b_i) \tag{3}$$

$$\widetilde{C}_t=\tanh(W_c\cdot[h_{t-1},o_t]+b_c) \tag{4}$$

$$C_t=f_t*C_{t-1}+i_t*\widetilde{C}_t \tag{5}$$

$$u_t=\sigma(W_o\cdot[h_{t-1},o_t]+b_o) \tag{6}$$

$$h_t = u_t * \tanh(C_t) \tag{7}$$

其中，h_{t-1}是前一个时间步的隐藏状态，o_t是当前时间步的输入(第一次输入代表 BERT 输出的隐藏层向量)，i_t是输入门的激活值，$\tilde{C}_t$是候选记忆单元，C_t是当前时间步的记忆单元，u_t是输出门的激活值。然后，将双向机制加入 LSTM 中，即每个时间步的输入都会受两个 LSTM 的处理：一个处理正向序列，另一个处理反向序列。对此，每个时间步 t 会得到两个隐藏状态：h_t^f(正向)和 h_t^b(反向)。最后，将 BiLSTM 的正向和反向隐藏状态拼接起来得到该模型的最终隐藏层向量表示。

$$h_t = [h_t^f, h_t^b] \tag{8}$$

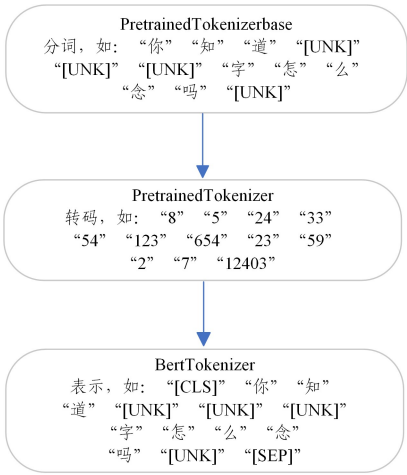

图 2 分词流程

Fig. 2 Process of word segmentation

3.2 强化提取层

为了更有效地捕捉局部特征，将局部注意力机制(Local Attention)与 TextCNN 结合使用。现有的大部分注意力机制通过在解码层的每个时间步，从输入序列所有时刻的隐藏状态中搜索出与当前输出目标最相关的一些隐藏状态，然后将这些隐藏状态与前面时刻的输出一起作为解码层当前时刻的输入，以此来保留输入序列的大部分信息。但是，考虑所有隐藏状态的计算成本非常昂贵，特别是对一些长句子，会导致计算不切实际。局部注意力机制在此基础上做出了改进，即每次解码时不再考虑编码层的全部隐藏状态，只考虑局部的隐藏状态。这种做法的计算简洁，而且更有利于强化特征提取。具体的分析和计算式如下。

在编码层，输入序列为 BERT 的输出。

$$O = \{O_1, O_2, \cdots, O_n\}, O_i \in R^{n \times BAT_0}$$

其中，BAT_o表示输入序列的维度大小，n 表示序列长度。本文输出采用 BiLSTM 前向和后向拼接的隐藏状态 $h_t = [h_t^f, h_t^b]$，即 $h_t = (h_1, h_2, \cdots, h_t)$，$h$ 代表某个词的隐藏向量，l 代表序列的长度。

在解码层，每个时间步的隐藏状态 S_i的计算式如下：

$$z_i = \sigma(W_a s_{i-1} + Y_a k_i) \tag{9}$$

$$r_i = \sigma(W_b s_{i-1} + Y_b k_i) \tag{10}$$

$$\tilde{s}_i = \tanh(W[r_i \circ s_{i-1}] + Y k_i) \tag{11}$$

$$s_i = (1 - z_i) \circ s_{i-1} + z_i \circ \tilde{s}_i \tag{12}$$

其中，$W_a, W_b, W \in R^{m \times q}$，$Y_a, Y_b, Y \in R^{q \times 2q}$均为权重矩阵，$m$ 和 q 分别对应词向量和隐藏层的维度。对于初始的隐藏层向量 s_0，采用编码层中第一个时间步的反向过程的隐藏层向量进行计算，计算式如下：

$$s_0 = \tanh(W_s \overleftarrow{h}_1) \tag{13}$$

其中，$W_s \in R^{n \times n}$。k_i表示随着 i 和权重 α_{ij}而变化的上下文向量，因此其计算式为：

$$k_i = \sum_{j=1}^{n} \alpha_{ij} h_j \tag{14}$$

在计算解码层的每个时间步 t 时刻的权重 α_{ij}时，需要先确定输入序列中与该时刻对齐的一个位置 p_t，然后以该位置为中心，设定一个窗口大小，即$[p_t - D, p_t + D]$，D 为一个整数，由于数据集为短文本，因此设定 D 为 4，对齐方式采用预测对齐(local-p)方法，即在每个时间步都会对 p_t进行预测。其计算式如下：

$$p_t = n \cdot sigmoid(v_p^{\mathrm{T}} \tanh(W_p h_t)) \tag{15}$$

其中，W_p和 v_p均为权重矩阵，$p_t \in [0, n]$。接着，在计算权重向量时，只考虑编码层中在该窗口内的隐藏状态，当窗口的范围超过输入序列的范围时，则直接舍弃超出的部分。同时，采用高斯分布进行修正，其计算式如下：

$$\alpha_t(s) = align(h_t, \bar{h}_s) \exp\left(-\frac{(s - p_t)^2}{2\sigma^2}\right) \tag{16}$$

其中，$\bar{h}_s$表示编码层第 s 时间步的隐藏状态，s 表示输入序列的位置，$\sigma = \frac{D}{2}$。

局部注意力机制能够识别文本中的关键部分，并给予这些部分更大的权重，忽略不相关的信息。TextCNN 使用卷积核，通过卷积层捕捉包含相同词的局部词组的相似特征表示，其多层结构可以学习从原始文本数据中抽象出的高级特征，有助于相同词语在不同语境下的理解。结合局部注意力机制所识别的关键部分和 TextCNN 加权结合进行局部特征提取，可以进一步强化这些部分的特征，并生成一个综合表示，这个表示同时考虑了文本的局部特征和全局上下文。

对于序列 $S_i = (s_1, s_2, \cdots, s_n)$，使用一个窗口大小为 d 的卷积核，卷积核的宽度与上一模型输出的隐藏层向量维度 q 相同，即卷积核 $\omega \in R^{d \times q}$。其计算过程如下：

$$h_f(l) = f(W_f * x_{l:l+d-1} + b_f) \tag{17}$$

其中，$h_f(l)$是第 f 个卷积核在位置 l 的输出，$x_{l:l+d-1}$是输入序列中从位置 l 开始的 d 个词的向量，f 是卷积核的大小，$f(\cdot)$是激活函数，通常使用 ReLU 函数引入非线性，池化层使用最大池化来选择每个卷积特征图中的最大值，这样可以减少参数数量，提取最重要的特征，其计算式如下：

$$h_f(l) = \max(0, h_f(l)) \tag{18}$$

$$h_f = \max_l h_f(l) \tag{19}$$

其中，h_f是经过最大池化后的第 f 个卷积核输出。

一般情况下，使用多个卷积核进行特征提取，$C = \{C_1, C_2, \cdots, C_m\}$用于表示 m 个卷积核的组合，C_m表示第 m 个卷积

核的大小。之后再计算 m 个特征向量并将这些向量展平成一个长向量，得到最终的结果。

$$H_t=[H_1,H_2,\cdots,H_F] \tag{20}$$

3.3 分类层

分类层首先将 BiLSTM 的输出和 TextCNN 的输出进行拼接，利用特征提取的多样性进行多尺度的特征融合，提供更全面的文本表示，减小在特定数据集上过拟合的风险，然后将其输入全连接层中，其计算式如下：

$$H_{\text{out}}=[h_t:H_t] \tag{21}$$

$$H_{\text{final}}=Linear(H_{\text{out}}) \tag{22}$$

最后，获取情感分类任务中对应的标记，本文使用的损失函数为交叉熵损失函数，其表达式如下：

$$L=-\sum_{i=1}^{N}\sum_{j=1}^{4}y_{i,j}\log(\hat{y}_{i,j}) \tag{23}$$

每个样本的标签表示为长度为 4(或 2)的向量，正确的类别对应位置为 1，其他位置为 0。$y_{i,j}$ 是标签向量中的第 j 个元素，如果样本 i 属于类别 j，则 $y_{i,j}=1$，否则 $y_{i,j}=0$。$\hat{y}_{i,j}$ 表示模型预测样本 i 属于类别 j 的概率。

3.4 大批次对抗策略

对抗训练的基本思想是通过在训练过程中引入对抗性扰动来提高模型的鲁棒性。

传统的对抗训练最核心的概念是 Min-Max 公式，用于描述模型在面对对抗性扰动时的优化问题，通常以生成对抗性样本来训练模型并提高其鲁棒性。其具体表达式如下：

$$\min_{\theta}\max_{\|\delta\|\leqslant\varepsilon}L(f_\theta(X+\delta),y) \tag{24}$$

式(24)指，在原始输入样本 X 上添加一个扰动幅度较小且易导致错误分类的对抗性扰动 δ，再通过对网络参数的更新来让模型抵抗这种扰动。其中，θ 表示模型的参数，$\|\delta\|$ 是扰动的范数，ε 是扰动的上限，用于控制对抗样本与原始样本的差异程度，y 是原始输入 X 的真实标签，L 是损失函数，f_θ 是模型的预测函数。

现有的绝大部分对抗算法均是基于梯度一步到位，同时在多步的对抗训练中向词嵌入层添加扰动，然后对扰动后的词嵌入空间进行调整改善，将优化后的词向量作为输入特征重新注入模型进行迭代训练。投影梯度下降算法(PGD)[5]对其进行了创新，设置一个扰动半径，当梯度上升超过这个半径时，就映射回原梯度，以保证扰动不会过大，即 PGD 进行多次迭代，每次走一小步，每次迭代都会将扰动投射到规定范围内，模型最终的梯度就是最开始的梯度加上最后一次扰动产生的梯度。该方法虽然简单有效，但是计算效率不高，而计算成本却很高，每一次的梯度下降都对应 K 步的梯度提升。本文使用一种对 PGD 算法进行改进的方法，由于其在梯度上升时，通过对输入梯度进行计算可以无成本地得到另一个参数梯度，因此在梯度下降的过程中，利用输入梯度和参数梯度，在一次的计算中利用更多信息加速对抗学习的训练，即大批次对抗策略。具体分析如下。

首先，该策略并不会使用全局标量，而是计算每个词向量的梯度信息，将对抗性扰动嵌入在词上。这种以每个词向量的梯度信息为基准的处理将会在模型整个学习的过程中考虑到词汇层面的变化，从而捕捉模型学习文本数据时的细微差异。对抗性扰动在确定哪些词是对抗性攻击的目标时，可以通过计算词的重要性来实现，一般情况下均是删除词，并观察模型输出的变化来评估其重要性。如："这家餐厅的服务非常好，食物也很美味。"大批次对抗策略会逐一删除句子中的词，判断哪些词对模型的正面分类起到了关键性的作用，同时观察模型的预测结果是否发生变化。当删除"好"或"美味"后，模型的预测从正面情感分类变为负面情感或者变得不确定，那么这些词就是潜在的攻击词，因为它们对模型的预测结果具有显著的影响。在这个基础之上添加微小的扰动，以此来构建对抗性样本。这种对抗性样本在人类看来与原句的区别并不大，但是却能够欺骗模型，使其做出错误的预测。具体操作是，在模型的输入 X 上添加一个通过梯度上升和模型参数 θ 的梯度得到的扰动 δ，用于最大化模型的损失函数。对抗性扰动 δ 具体的实现表达式为：

$$\delta_{\text{new}}=\delta_{\text{old}}+\alpha\cdot sign(\nabla_\delta L(f_\theta(X+\delta_{\text{old}}),y) \tag{25}$$

其中，α 为学习率。接着在 PGD 多次进行迭代产生对抗性样本时，在每一步都将对抗训练样本计算得到的参数梯度保留下来，对多轮计算的梯度值进行累加，而不是在每一步后立即更新模型参数。这样就可以将得到的信息虚拟性扩大 K 倍，从而达到改善模型性能的目的。另外，大批次对抗策略利用多达 K 个不同的范数约束，将不变性强加给 K 个对抗性扰动，相比之下，PGD 算法只有单一的范数约束，这样带来的泛化误差会更小。实现表达式如下：

$$g_t=g_{t-1}+\frac{1}{K}\nabla_\theta L(f_\theta(X+\delta_t),y) \tag{26}$$

最后在 K 次迭代后，使用累积的梯度的平均值来一次性更新模型参数 θ，其表达式如下：

$$\theta=\theta-\tau\cdot\frac{1}{K}\sum_{t=1}^{K}g_t \tag{27}$$

其中，τ 是学习率。使用该方法可以在额外开销很小的情况下将原始数据扩大 K 倍，减少了模型更新的次数，提高了训练中梯度的利用率，有助于提高模型的鲁棒性。如果在某一步中扰动的范数超过了预设的限制，则需要将扰动投影回允许的范数范围内，具体表达式如下：

$$\theta proj_\varepsilon(\delta)=\delta\cdot\min\left(1,\frac{\varepsilon}{\|\delta\|_2}\right) \tag{28}$$

其中，δ 是计算得到的扰动，ε 是允许的最大范数。

一般情况下，对抗训练不会与 Dropout 同时使用。Dropout 作为一种正则化手段，会因每次前向传播随机丢弃网络连接而改变网络结构，这可能导致对抗扰动的计算不稳定。但是由于 BERT 预训练模型在每次微调时都会使用 Dropout，因此需要采取一系列措施使对抗策略与 Dropout 相适应。

首先在对抗训练开始时进行初始化，并且在迭代开始前备份当前的 Dropout mask。在迭代过程中生成对抗样本和更新模型参数时，每次前向传播的 Dropout mask 必须一致。然后在 K 步迭代中，当每次前向和反向传播时，使用备份的 Dropout mask，确保扰动的计算和参数更新都是在相同的网络结构下进行。在对抗训练结束后，对各超参数进行调整，确

保模型能够获得最佳性能。大批次对抗策略的整体结构流程如图 3 所示。

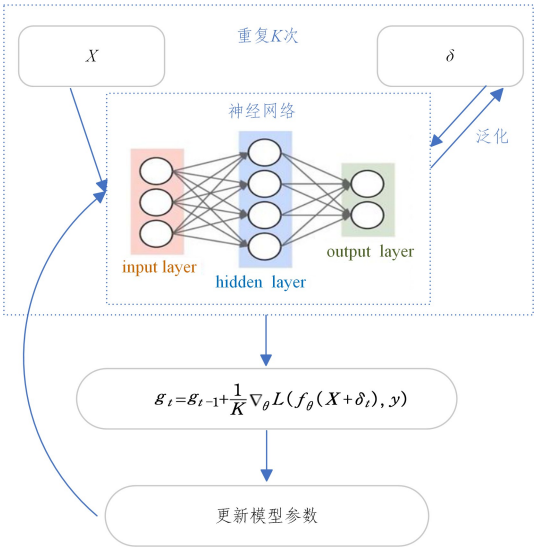

图 3 结构流程图

Fig. 3 Structural flow diagram

4 实验

4.1 数据集

本文所使用的数据集为中文微博数据集(四分类和二分类)、酒店评论数据集(二分类)和外卖评论数据集(二分类),微博四分类数据集包含 364 751 条数据,涉及娱乐、体育、资讯、经济等多个领域,每条数据格式均为 label(标签)和 review(内容),label 包含 4 类,分别为 0,1,2,3,对应喜悦、愤怒、厌恶和低落。其中,喜悦为 199 705 条,愤怒为 53 094 条,厌恶为 56 263 条,低落为 55 689 条。按照标准,本文将训练集以 8∶1∶1 比例划分为训练集、测试集和验证集,划分后的数据集统计如表 1 所列,具体示例如表 2 所列。

表 1 中文微博数据集(四分类)

Table 1 Chinese Weibo dataset(four categories)

数据集	文本数量	情感数量			
		喜悦	愤怒	厌恶	低落
训练集	291 801	160 491	43 770	43 770	43 770
测试集	36 475	20 062	5 471	5 471	5 471
验证集	36 475	20 062	5 471	5 471	5 471

表 2 微博数据集四分类示例

Table 2 Four classification example of Weibo dataset

内容	标签	标签含义
风格不一样嘛,都喜欢! 最喜欢哪张?	0	喜悦
最痛恨别人骗我!	1	愤怒
还有理可言吗? 无语	2	厌恶
唉,每次看球都赶不上精彩的进球	3	低落

微博二分类数据集包含 119 989 条数据,涉及的具体内容同上,每条数据格式均为 label(标签)和 review(内容),label 包含 2 类,分别为 0,1,对应积极和消极。其中,积极占 59 981 条,消极占 60 007 条。同理,本文将训练集以 8∶1∶1 比例划分为训练集、测试集和验证集,划分后的数据集统计如表 3 所列,具体示例如表 4 所列。

表 3 中文微博数据集(二分类)

Table 3 Chinese Weibo dataset(two categories)

数据集	文本数量	情感数量	
		积极	消极
训练集	95 991	47 985	48 006
测试集	11 999	6 000	5 999
验证集	11 999	6 000	5 999

表 4 微博数据集二分类示例

Table 4 Two classification example of Weibo dataset

内容	标签	标签含义
梦想有多大,舞台就有多大! [鼓掌]	1	积极
在地铁站居然被一块砖绊倒了! [抓狂]	0	消极

酒店评论二分类数据集包含 7 766 条数据,每条数据格式均为 label(标签)和 review(内容),label 包含 2 类,分别为 1 和 0,对应积极和消极。其中,积极占 5 322 条,消极占 2 444 条。同理,本文将训练集以 8∶1∶1 比例划分为训练集、测试集和验证集,划分后的数据集统计如表 5 所列,具体示例如表 6 所列。

表 5 酒店评论数据集(二分类)

Table 5 Hotel reviews dataset(two categories)

数据集	文本数量	情感数量	
		积极	消极
训练集	6 213	4 258	1 955
测试集	777	532	245
验证集	776	532	244

表 6 酒店评论数据集二分类示例

Table 6 Two classification example of hotel reviews dataset

内容	标签	标签含义
位置便利,价格便宜,服务周到!	1	积极
房间地毯都是黑的,自助餐难吃!	0	消极

外卖评论二分类数据集包含 11 988 条数据,每条数据格式均为 label(标签)和 review(内容),label 包含 2 类,分别为 1 和 0,对应积极和消极。其中,积极占 7 987 条,消极占 4 000 条。同理,本文将训练集以 8∶1∶1 比例划分为训练集、测试集和验证集,划分后的数据集统计如表 7 所列,具体示例如表 8 所列。

表 7 外卖评论数据集(二分类)

Table 7 Takeaway review dataset(two categories)

数据集	文本数量	情感数量	
		积极	消极
训练集	9 590	6 390	3 200
测试集	1 199	799	400
验证集	1 199	798	400

表 8 外卖评论数据集二分类示例

Table 8 Two classification example of takeaway review dataset

内容	标签	标签含义
不错! 吃一个就很饱	1	积极
“凉皮太辣,吃不下都”	0	消极

4.2 实验参数设置

实验参数如表 9 所列。

表 9 实验参数

Table 9 Experimental parameters

参数	值
操作系统	Linux 3.10
GPU	V100 16GB
框架	PyTorch
Batch Size	32
Learn Rate	1×10^{-6}
Epoch	10
词嵌入维度	768
隐藏层维度	768
Dropout	0.5
最大文本长度	200
损失函数	交叉熵损失函数
优化器	Adam
卷积核	(3,5,7)
SAT 迭代次数	4
扰动(四分类)	0.5
扰动(二分类)	0.4
激活函数	ReLU

本文部分实验参数按照基准模型 BBASM[27] 进行确定，如 Epoch、词嵌入维度、隐藏层维度、激活函数等。但是对最大文本长度进行了改进，原因是中文微博数据集中某些文本较长，需要更大的文本长度。

4.3 实验评估方法

本文采用准确率(ACC)和 F1 值(F1)作为情感分类成功的评估指标。当每句话和其对应的标签精确匹配时，则说明模型的预测结果正确。

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \tag{29}$$

$$P=\frac{TP}{TP+FP} \tag{30}$$

$$R=\frac{TP}{TP+FN} \tag{31}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{32}$$

其中，TP 是实际为正类且预测为正类的样本数，TN 是实际为负类且预测为负类的样本数，FP 是实际为负类但预测为正类的样本数，FN 是实际为正类但预测为负类的样本数。

4.4 实验结果与分析

为了验证所提模型的有效性，本文将该模型与其他几个模型进行对比，具体如下。

ChineseBERT[26]：将汉字的字形和拼音信息合并到语言模型预训练中，处理了汉语中非常普遍的异音(即同一个字不同的读音和不同的意思)现象。

BBASM[27]：该模型是一个结合了 BERT 和 BiLSTM 的注意力机制模型，用于增强文本相似度的计算。

BERT-BiGRU[28]：该模型结合了 BERT 和 BiGRU，用于获得更全面的文本表示。

MCNN-MA：是一种基于多通道卷积神经网络(CNN)和多头注意力机制的短文本情感分析模型，在多个短文本情感分析任务上表现出色。

MCNN-LSTM[29]：将 CNN 和 LSTM 结合，有利于获得更丰富的特征表示。

BERT-CNN-BiLSTM[30]：BERT 将文本转换为高维的语义向量，CNN 提取文本局部特征，BiLSTM 捕捉长期依赖关系，三者结合以提高对文本数据的理解能力。

BBC-PGD：该模型在本文模型的基础上将对抗策略更改为投影梯度下降法，具体参数与本文模型一致。

上述模型与本文模型的对比实验结果如表 10、表 11 所列。

表 10 对比实验(微博)

Table 10 Comparative experiments(Weibo)

(%)

Model	4 Categories		2 Categories	
	ACC	*F*1	*ACC*	*F*1
ChineseBERT	56.83	43.36	74.97	74.84
BBASM	57.04	44.76	75.26	75.03
BERT-BiGRU	56.96	44.13	75.07	74.85
MCNN-MA	57.26	44.62	75.13	75.92
MCNN-LSTM	57.37	44.63	75.22	75.33
BERT-CNN-BiLSTM	58.19	45.73	75.54	75.84
BBC-PGD	59.95	46.16	75.89	76.33
Ours	59.98	46.53	76.17	76.67

表 11 对比实验(酒店和外卖)

Table 11 Comparative experiments(hotels and takeaways)

(%)

Model	Hotels		Takeaways	
	ACC	*F*1	*ACC*	*F*1
ChineseBERT	80.60	85.31	89.12	86.62
BBASM	81.53	86.52	90.35	87.43
BERT-BiGRU	81.24	86.33	89.24	86.91
MCNN-MA	81.11	85.38	89.56	86.84
MCNN-LSTM	82.58	85.63	89.95	87.38
BERT-CNN-BiLSTM	83.37	86.96	90.57	87.77
BBC-PGD	84.48	87.87	90.84	87.96
Ours	85.26	88.54	91.16	88.08

由实验结果可知，本文模型的 ACC 和 F1 值在中文微博数据集上均达到最优。BBASM 模型和 BERT-BiGRU 模型分别在 ChineseBERT 模型的基础上添加了 BiLSTM 和 BiGRU，但是结果提升程度却不高。在短文本、长文本和跨域任务中，对具有特殊含义的标点符号的句子识别错误率仍然较高，如“多么‘棒’的服务，我等了 3 个小时”。由于 BERT 中的自注意力机制已经在某种程度上处理一部分长距离依赖关系，BiLSTM 更擅长处理严格时序信息，相比于 ChineseBERT 模型并未有特别大的进步；而且 BERT，BiLSTM，BiGRU 三者更倾向于理解全局上下文的含义，局部特征提取的能力不足，因此仍然具有很大的局限性。MCNN-MA 模型和 MCNN-LSTM 模型则正好相反，没有 BERT 精确捕捉词义细微差别和长距离依赖关系的加持，两个模型虽然在短文本任务中有了一些进步，但是对于长文本任务和跨域分析的学习能力不够。MCNN-MA 虽然加入了 TextCNN 强化特征提取，但是多头注意力机制在增强对长文本的上下文理解时，可能需要过多的头，这样易导致模型过拟合；过少的头则不利于学习足够的知识。而且其内部工作机制较为复杂，每个头的具体功能难以直观理解，这降低了模型的可解释性。MCNN-LSTM 同样如此，LSTM 的参数较多，梯度需要通过许多层的循环连接进行反向传播，因此在处理长序列文本时计算复杂度很高，处理长时间跨度问题时难度较大，甚至出现梯度爆炸

的问题。BERT-CNN-BiLSTM 结合了上述几个模型的优点，利用 BERT 和 BiLSTM 加强对短文本和长文本的学习，Text-CNN 加强对情感词、标点符号及其周围情感信息局部特征的提取，因此在整个数据集上的准确率有所提高。但是由于缺乏对抗训练这种数据增强技术，并不利于跨域任务的实现。本文认为在没有对抗训练时，数据的差异性较大，模型会对源领域数据过拟合，导致泛化性能下降。而加入对抗训练后，模型在训练时学会抵抗这种扰动以及从扰动中恢复并保持其决策能力，从而提高模型在不同领域的鲁棒性。BBC-PGD 模型在上述几个模型的基础上加入了对抗训练这种数据增强技术；PGD 利用多次迭代和逐步增加的扰动，可以更好地适应非线性模型，有利于处理复杂的神经网络。可以看出，该模型相比于前几个模型，在几个任务上都有了质的飞跃。但是 PGD 在重新计算扰动时，可能会忽略之前迭代中获得的信息，导致实验结果下降。另一方面，PGD 的计算成本较高，且对样本的泛化能力提升有限。

综上所述，相比上述模型，本文模型在中文微博数据集的指标上均取得了最好的效果。

4.5 对抗训练迭代和扰动实验

本节介绍模型中对抗训练超参数对实验结果的影响，本文在中文微博数据集上进行多步对抗训练，寻找到效果最好的迭代次数和扰动范围。迭代次数和扰动的初始值按照 PGD[5] 对抗训练策略中的实验确定，即 $K=3$ 以及 $\delta=0.3$，然后进行逐步增加或减少。首先进行迭代次数的实验（默认将扰动设置为最好的状态，且保持不变），将对抗训练迭代次数分别设置为 3，4，5，6，7 等。具体结果如表 12 所列。

表 12 迭代实验

Table 12 Iterative experiments

(%)

K	4 Categories		2 Categories	
	ACC	*F1*	*ACC*	*F1*
2	57.43	44.36	75.27	74.94
3	58.84	45.66	75.66	75.33
4	59.98	46.53	76.17	76.67
5	59.45	46.27	75.94	76.23
6	58.78	45.53	75.67	75.48

从表 12 中可以得出结论，逐步增加迭代次数可以提高 ACC 和 F1 值，当 K 的值设置为 4 时，在四分类和二分类数据集上的效果均为最佳。但是继续增加迭代次数不仅会造成模型的效果变差，而且会大大提高计算成本，降低模型的训练速度。

表 13 扰动实验

Table 13 Perturbation experiments

(%)

δ	4 Categories		2 Categories	
	ACC	*F1*	*ACC*	*F1*
0.2	57.03	42.36	74.67	74.64
0.3	58.78	44.56	75.68	75.32
0.4	59.06	45.93	76.17	76.67
0.5	59.98	46.53	75.36	75.63
0.6	59.52	46.07	74.86	75.12

同理，在表 13 中可以看到，相比于迭代次数的变化，扰动的设置反而对模型的实验结果影响更大；当扰动的值设置过小或者过大，ACC 和 F1 值反而会比基线模型更小。另外，在四分类和二分类数据集上并没有出现扰动一致 ACC 和 F1 值同时最大的情况，当扰动取值为 0.5 时，四分类数据集 ACC 和 F1 值得到最好结果；而当扰动取值为 0.4 时，二分类数据集 ACC 和 F1 值得到最好结果。究其原因，本文认为是四分类数据集比二分类数据集要大许多，大数据集通常包含更多变化和噪声，通过引入更大扰动，模型可以学习到更广泛的特征表示，这有助于提高模型的泛化能力。

4.6 不同卷积核对比实验

在文本分类中，卷积层应用于词嵌入序列以提取局部特征。使用不同大小的卷积核可以捕捉不同范围的上下文信息，但是过大的卷积核会忽略一些重要的局部信息，而过小的卷积核则会不足以捕获更广泛的上下文信息，甚至造成过拟合。因此，选取合适的卷积核成为重中之重。由于本文的目的侧重于提取局部特征，因此选取较小的卷积核用于提取细节特征。在实际应用中，多个小卷积核的组合通常比单个大卷积核更有效，因为它们可以减少参数数量并保持网络的深度以提高网络的学习能力和泛化能力，故本节选取了(1，3，5)，(2，4，6)，(3，4，5)，(3，5，7)，(5，7，9)5 组卷积核进行四分类对比实验，实验结果如表 14 所列（其他超参数默认设置为最好的状态，且保持不变）。

表 14 卷积核实验

Table 14 Convolutional kernel experiments

(%)

卷积核大小	4 Categories		2 Categories	
	ACC	*F1*	*ACC*	*F1*
(1,3,5)	58.39	44.76	74.98	75.04
(2,4,6)	58.63	45.38	75.68	75.62
(3,4,5)	59.14	46.02	75.93	76.15
(3,5,7)	59.98	46.53	76.17	76.67
(5,7,9)	58.62	45.53	75.36	75.63

通过观察实验的结果可以看到，当卷积核取到(3，5，7)时，四分类和二分类数据集中的 ACC 和 F1 值可以达到最好的效果。(3，4，5)中的 4×4 卷积核的效果要弱于(3，5，7)中的 7×7 卷积核，这是因为 7×7 卷积核有更强的网络感受野，可以捕获更广泛的局部信息。相比之下，(5，7，9)中的 9×9 卷积核较大，会减小网络的深度，所捕获的信息很大程度上已经被之前的模型所捕获，故选择(3，5，7)卷积核的效果最好。同时可以看到，将卷积核的取值(2，4，6)和(3，4，5)进行对比，发现 2×2 和 6×6 的结果相比于 3×3 和 5×5 要差一些。本文认为使用奇数尺寸的卷积核便于中心定位，可以在不使用填充的情况下保持输入和输出的空间维度不变，在绝大多数的情况下是有利的。合理的调整卷积核大小，可以提高模型在不同尺度上的特征提取能力，并且有利于分析相同词在不同句子场景中的情感。

4.7 消融实验

本节消融实验仅考虑模型中的各个模块对实验的影响，各超参数取最优值且保持不变。由于四分类任务相对更有挑战性，因此本文仅在微博四分类数据集上进行实验。具体实验结果如表 15 所列。

表 15 消融实验

Table 15 Ablation experiments

ACC/%	58.26	59.57	58.54	58.37	59.98
F1/%	45.32	46.06	45.49	45.33	46.53
Local-Attention	√		√		√
TextCNN	√	√			√
SAT		√	√	√	√

观察表 15 中的数据可以发现，当模型不使用大批次对抗策略时，ACC 下降了 1.72%，F1 下降了 1.21%，表明自对抗训练有助于模型学习到更一般的特征表示，通过在训练过程中引入词嵌入扰动，使模型对输入数据的微小变化更加不敏感，可提高模型的鲁棒性以及在新数据集上的泛化能力。而当模型不使用局部注意力机制时，最后的实验结果中 ACC 下降了 0.41%，F1 值下降了 0.47%。本文推测局部注意力机制能够增强模型对局部上下文的理解，而且在特征融合上通过加权的方式与 TextCNN 进行结合进而取得了一些进步。但是因为 TextCNN 本身在任务上已经表现得足够好，所以局部注意力机制只能带来一些微小的进步。观察表中的第三列，在模型不使用 TextCNN 时，结果下降明显，ACC 和 F1 值分别下降了 1.44%，0.94%。TextCNN 利用不同的卷积核提取不同长度的局部特征，并且利用最大池化操作从特征图中提取最重要的特征，因此在文本分类中可以帮助分析情感词、修饰词以及标点符号等与周围上下文的关系，同时捕捉从单个词语到短语的多尺度情感信息，这足以说明 TextCNN 的有效性。观察表中第四列，当模型不使用局部注意力机制和 TextCNN 时，结果下降明显，ACC 和 F1 值分别下降了 1.61%和 1.20%。局部注意力通过 TextCNN 的卷积层从输入数据中提取特征表示，然后其 SE 模块利用 TextCNN 的全局池化操作将每个特征图压缩为一个标量值，从而获得一个具有全局视野的特征向量。接着，使用全连接层对压缩后的特征进行处理，并通过激活函数引入非线性来学习每个通道的重要性权重。最后，将注意力权重与 TextCNN 最开始卷积的输出特征图进行加权结合，增强重要特征并抑制不重要的特征，如能够加强对修饰词和标点符号等的学习，大大提高了局部特征提取的能力。以“多么‘棒’的服务，我等了三个小时。”为例，当模型同时去掉局部注意力和 TextCNN 时，无法成功识别；但是加上后，可以成功识别，这足以说明局部注意力和 TextCNN 加权结合的重要性。

4.8 对抗训练的有效性分析

对抗训练在处理数据不平衡问题时可以提高模型的公平性和准确性。为了验证对抗训练的有效性，我们考虑将四分类数据集划分为二分类数据集，即喜悦占 1 类，其他 3 类随机取 1 类。具体数据集划分如表 16 所列。

表 16 改进后的数据集

Table 16 Improved dataset

	喜悦	愤怒	厌恶	低落	总数
文本数量	160491	43770	0	0	204261
	160491	0	43770	0	204261
	160491	0	0	43770	204261

观察表 16 中的数据可以看到，喜悦这一类别的数量远超其他类别，理论上，当数据不平衡时，会降低性能指标的有效性，使模型泛化能力下降。但是通过使用对抗训练这种数据增强技术，可以在一定程度上避免这种情况，使实验的结果得到提高。具体的实验结果如图 4 所示。

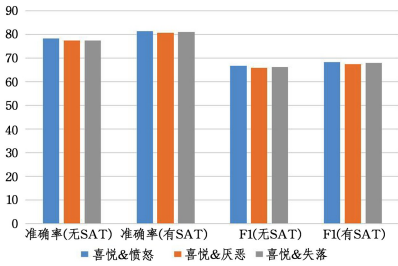

图 4 对抗训练有效性实验的结果

Fig. 4 Results of SAT validity test

本文认为，对抗训练可以尝试作为一种正则化手段来提高模型的泛化能力，通过在词嵌入层添加扰动进而生成数据量较小的类别的对抗样本，提高生成样本的多样性，有利于解决过拟合问题；同时增强模型对较小扰动的稳健性，从而提升模型的性能。

4.9 案例分析

本节将在中文微博四分类数据集上选取几条数据进行案例分析，通过具体案例对比，证明局部注意力机制与 TextCNN的联合特征提取方式，结合大批次对抗策略，能显著提升分类效果，如表 17 所列。

表 17 案例分析(BBASM 和 Ours 对比)

Table 17 Case study(comparison of BBASM and Ours)

模型	句子	情感(括号内为真实情感)
BBASM	多么‘棒’的服务，我等了三个小时。	喜悦(愤怒)
	终于等到这一刻了，不容易。	低落(喜悦)
	这‘美食’真让人大开眼界，不如和垃圾桶亲密接触。	喜悦(厌恶)
	今天真是太‘美好’了，因为我的运气不知道在哪里。	喜悦(低落)
Ours	多么‘棒’的服务，我等了三个小时。	愤怒(愤怒)
	终于等到这一刻了，不容易。	喜悦(喜悦)
	这‘美食’真让人大开眼界，不如和垃圾桶亲密接触。	厌恶(厌恶)
	今天真是太‘美好’了，因为我的运气不知道在哪里。	低落(低落)

将本文模型与 BBASM 模型进行对比。表 17 列出的 4 个案例中，BBASM 模型无法正确预测，本文认为是该模型对包含模糊、讽刺和幽默的表达难以准确判断其情感倾向，对语气词和标点符号的学习也不足。另外，如果关键的情感线索在文本之外，比如前一条语句，该模型则无法准确预测。而本文模型通过使用大批次对抗策略进行数据增强，以及强化

特征提取的方式加强了模型对情感词、修饰词和标点符号及其周围信息的理解，提高了模型的鲁棒性和泛化性。

结束语 本文提出了一种基于大批次对抗策略和强化特征提取的文本情感分类方法。首先，将文本数据集输入预训练语言模型 BERT 中，得到相应的词嵌入向量表示，再利用 BiLSTM 进一步学习序列中的时序特征，强化模型对文本上下文依赖关系的理解。之后，将局部注意力机制与 TextCNN 的输出进行加权结合，强化特征提取能力，提供更全面的序列表达，生成更加丰富和信息密集的特征表示。最后，再将 BiLSTM 的输出与 TextCNN 的输出进行拼接，得到两个空间的深层特征融合，交由分类器进行情感分类判断。在整个训练过程中，本文采取大批次对抗策略，在每次迭代中同步更新模型参数和输入扰动，通过累积梯度的方式来加速对抗训练过程，进而有效提高模型的鲁棒性。

显而易见，本文模型仍然具有很强的局限性。在处理具有隐喻和比喻的情况时模型效果并不好，如句子“他是个狮子般的领导者”，在缺乏外部知识的情况下，无法精确判断“狮子”在这里是象征勇气和力量（正面情感），还是残暴和凶猛（负面情感）。同样地，在区分具有复杂句法结构的长文本，尤其是文章所想要表达的情感词靠后时，往往很难成功分类，如微博数据集中的句子“虽然我对这次长途飞行感到非常焦虑，因为我怕自己无法适应狭小的座位和长时间的不适，但事实上，一旦飞机起飞，我很快就被窗外那令人惊叹的云海景色所吸引，这让我几乎忘记了旅途的疲倦，最终，当飞机平稳降落在目的地时，我感到了一种前所未有的成就感和满足，因为我克服了自己的恐惧，享受了这次旅行。”模型会因为首先识别到“焦虑”“狭小”“不适”等词而错误地将文本分类为低落，但实际上该句话所想要表达的意思是喜悦。另一方面，因为模型具有对抗训练这种数据增强技术，在词嵌入上添加的扰动可以加强对跨域数据的识别，所以当模型在中文微博数据集上进行训练时，在其他数据集上进行测试的识别精度仍然不错。

未来的工作中，我们将考虑在模型中融入包含隐喻和比喻的外部知识，以更好地对有特殊结构的句子进行分类，如上面提到的句子“他是个狮子般的领导者”。针对含有成语的句子，提取其内部语法结构特征（如联合结构、偏正结构等），以增强模型对成语特殊表达方式的理解能力；以及使用更加细致和深入的编码方式对特征融合方式进行改进，如在 BERT 中使用深度残差网络或密集连接网络的思想让网络层次之间可以更有效地交互和进行特征融合；或者在 BERT 的每个 Transformer 层后都进行一次特征融合，确保信息在模型的每一层都能得到有效的整合。以此来更好地实现特征提取，提高情感分类的准确性。

参 考 文 献

[1] 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. ACL, 2019: 4171-4186.

[2] ZHAO W M, ALWIDIAN S A, MAHMOUD Q H. Adversarial Training Methods for Deep Learning: A Systematic Review[J]. Algorithms, 2022, 15(8): 283.

[3] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[4] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. arXiv: 1508.01991, 2015.

[5] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[J]. arXiv: 1706.06083, 2017.

[6] LUONG T, PHAM H, MANNING C D. Effective Approaches to Attention-based Neural Machine Translation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. ACL, 2015: 1412-1421.

[7] KIM Y. Convolutional Neural Networks for Sentence Classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2014: 1746-1751.

[8] KREUTZ T, DAELEMANS W. Enhancing General Sentiment Lexicons for Domain-Specific Use [C]// Proceedings of the 27th International Conference on Computational Linguistics. ACL, 2018: 1056-1064.

[9] TENG Z Y, VO D T, ZHANG Y. Context-Sensitive Lexicon Features for Neural Sentiment Analysis[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. ACL, 2016: 1629-1638.

[10] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. arXiv: 1907.11692, 2019.

[11] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations[J]. arXiv: 1909.11942, 2019.

[12] CLARK K, LUONG M T, LE Q V, et al. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators [J]. arXiv: 2003.10555, 2020.

[13] 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.

[14] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter[J]. arXiv: 1910.01108, 2019.

[15] FENG Y, CHENG Y. Short Text Sentiment Analysis Based on Multi-Channel CNN With Multi-Head Attention Mechanism [J]. IEEE Access, 2021, 9: 19854-19863.

[16] PENG Z, LU Y, PAN S, et al. Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021). 2021: 3020-3024.

[17] YANG H, ZENG B Q, YANG J H, et al. A Multi-task Learning Model for Chinese-oriented Aspect Polarity Classification and Aspect Term Extraction[J]. arXiv: 1912.07976, 2019.

[18] ZHAO H T, MA C, DONG X S, et al. Certified Robustness Against Natural Language Attacks by Causal Intervention[C]//

International Conference on Machine Learning. 2022.
[19] XU Z E,ZHU G H,MENG C H,et al. A2:Efficient Automated Attacker for Boosting Adversarial Training[J]. arXiv: 2210.03543,2022.
[20] MA X S,WANG Z K,LIU W W. On the Tradeoff Between Robustness and Fairness[C]// Advances in Neural Information Processing Systems. 2022.
[21] LI X Y,XIN Z,LIU W W. Defending Against Adversarial Attacks via Neural Dynamic System[C]// Advances in Neural Information Processing Systems. 2022.
[22] MA F K,HU X M,LIU A W,et al. AMR-based Network for Aspect-based Sentiment Analysis[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,2023:322-337.
[23] ZHANG H Y,DUAN L G,WANG Q C,et al. Multi-entity sentiment analysis of long text based on multi-task joint training [J]. Computer Science,2024,51(6):309-316.
[24] LIU D X,DUAN L G,CUI J J,et al. Short text semantic matching strategy based on dual channels of semantic similarity matrix and word vector[J]. Computer Science,2024,51(12): 250-258.
[25] WU Y H,SCHUSTER M,CHEN Z F,et al. Google's Neural Machine Translation System:Bridging the Gap between Human and Machine Translation[J]. arXiv:1609.08144,2016.
[26] SUN Z J,LI X Y,SUN X F,et al. ChineseBERT:Chinese Pretraining Enhanced by Glyph and Pinyin Information[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. ACL, 2021: 2065-2075.
[27] LI X L,LEI Y Y,JI S W. BERT- and BiLSTM-Based Sentiment Analysis of Online Chinese Buzzwords[J]. Future Internet, 2022,14(11):332.
[28] DAI J,CHEN C. Text classification system of academic papers based on hybrid Bert-BiGRU model[C]// 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics(IHMSC). 2020:40-44.
[29] HASIBK M,AZAM S,KARIM A,et al. ,MCNN-LSTM:Combining CNN and LSTM to Classify Multi-Class Text in Imbalanced News Data[J]. IEEE Access,2023,11:93048-93063.
[30] HE Y S. BERT-CNN-BiLSTM:A Hybrid Deep Learning Model for Accurate Sentiment Analysis[C]// IEEE 5th International Conference on Power, Intelligent Computing and Systems (ICPICS 2023). 2023:921-926.

CHEN Jiahao, born in 2001. His main research interest is sentiment analysis.

DUAN Liguo, born in 1970, is a member of CCF (No. 15823S). His main research interest is natural language processing.

(责任编辑:喻藜)