杨传德, 李海军,王汝旭,等. 基于混合神经网络的多模态抑郁症检测算法[J]. 智能计算机与应用,2025,15(7):48-55.
DOI:10.20169/j.issn.2095-2163.250707

基于混合神经网络的多模态抑郁症检测算法

杨传德, 李海军, 王汝旭, 刘 聪
(德州学院 计算机与信息学院, 山东 德州 253023)

摘 要: 抑郁症检测是心理健康领域的重要分类任务。抑郁症检测中,单一模态方法分类精度低下,本文提出了一种基于混合神经网络的多模态抑郁症检测算法。首先,针对中文文本信息,搭建了一种基于 ESimCSE-BiLSTM-CNN 的混合神经网络的算法模型,在考虑局部关键信息的同时加快收敛速度;其次,针对中文音频信息,将音频特征在进行 GhostVLAD 特征聚合后,送入搭建好的基于多头自注意力的 CNN-GRU 混合神经网络模型,使得算法模型能够更好地捕捉输入序列的结构和规律;最后,通过融合上述混合神经网络模型特征,充分利用模态间的互补关系,提高了算法模型的表达能力。基于 EATD-Corpus 中文多模态抑郁症检测数据集的测试结果表明,本文提出的算法 $F1$ 值达到了 90.90%,在众多主流多模态抑郁症检测算法中具有强大的竞争力。

关键词: 混合神经网络; ESimCSE; 多模态抑郁症检测; 抑郁症检测
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)07-0048-08

Multimodal depression detection algorithm based on hybrid neural network

YANG Chuande, LI Haijun, WANG Ruxv, LIU Cong
(School of Computing and Information, Dezhou University, Dezhou 253023, Shandong, China)

Abstract: Depression detection is an important classification task in the field of mental health. Aiming at the low classification accuracy of single modal algorithm in depression detection, a multi-modal depression detection algorithm based on hybrid neural network was proposed. For Chinese text information, a hybrid neural network algorithm model based on ESimCSE-BiLSTM-CNN is built, which can accelerate the convergence speed while considering the local key information. For Chinese audio information, after GhostVLAD feature aggregation, the audio features are fed into the constructed CNN-GRU hybrid neural network model based on multi-head self-attention, so that the algorithm model can better capture the structure and rule of the input sequence. Finally, by blending the features of the hybrid neural network model and making full use of the complementary relationship between the modes, the expression ability of the algorithm model is improved. The test results based on EATD-Corpus Chinese multimodal depression detection dataset show that the $F1$ value of the proposed algorithm model reaches 90.90%, which has strong competitiveness in many mainstream multimodal depression detection algorithms.

Key words: hybrid neural network; ESimCSE; multimodal depression detection; depression detection

0 引 言

信息技术推动时代飞速发展,社会因素的复杂性与个体的差异性得到放大,在竞争高度激烈的当代社会,人们的心理状态也愈发容易发生变化。科学研究表明:青少年人群由于自身缺乏正确认知及相关知识,患病率逐年提高,并且复发率高[1]。尽管抑郁症会给人们的日常学习工作带来不利影响,但是如果能及时通过客观指标发现心理健康状况,完全可以通过药物、物理和心理等方式缓解甚至治愈病情[2]。因此,寻求一种客观指标来自动诊断识别出抑郁症病患情况是预防治疗抑郁症的关键。而在人工智能向各个产业赋能的今天,利用深度学习来进行抑郁症的检测也逐渐受到研究者的重视。

基金项目: 德州市企业研发计划项目(2022dzkj094); 山东省大学生创新训练项目(S202310448077)。
作者简介: 杨传德(2003—),男,本科生,主要研究方向:深度学习; 王汝旭(2002—),男,本科生,主要研究方向:音频分析; 刘 聪(2002—),男,本科生,主要研究方向:多模态融合。
通信作者: 李海军(1974—),男,硕士,副教授,主要研究方向:计算机视觉。Email:sdlclhj@126.com。
收稿日期: 2023-11-26

当前，国内外融合抑郁症患者的音频和文本两个模态的心理疾病识别研究较少，多是用音频单模态或者音频融合脑电的方法或音视频融合来提升识别准确率。Ozdas A 等[3]说明了抑郁症会改变身体和自主神经系统，从而会潜在地改变发声和发音肌肉的机制；李金鸣等[4]通过自主搭建的多尺度差分归一化音频特征提取算法，提高了抑郁症检测精度；Cai 等[5]使用 KNN（K-Nearest Neighbor）的方法进行多模态脑电数据融合音频数据的抑郁症检测，分类准确率最高达到 86.98%；郭威彤[6]从音视频融合角度出发，建立跨模态共注意力网络，平均分类准确率为 81.2%。上述研究表明音频数据中包含丰富的情感信息，融合其他模态特征可以提高识别率，因此本文将音频与中文文本融合进行抑郁症检测。

尽管也可以看到很少的一些基于音频与文本的多模态融合方法取得了一定的成果，减少了情感数据中可能出现的情感特征质量不佳等问题，弥补了以往研究中忽视人类情感本身由多模态因素组成的缺陷。但多模态抑郁症音频文本数据相比单模态抑郁症数据的时间序列和局部特征具有更高的复杂性，需要同时关注数据的时间依赖关系和局部关键特征。因此需要一种有效方案来学习数据在多模态特征空间中的复杂依赖关系和局部关键线索，使得模型着重关注于与当前特征关联程度更紧密的特征信息，忽略与之相关性较低甚至会影响模型性能的噪声数据。如 Williamson 等[7]只是将特征进行简单的融合，没有挖掘各模态间的互补信息；Hanai 等[8]将文本和声音的特征输入到多层双向 LSTM（Long Short-Term Memory）模型中，虽然能够实现双模态之间的互补性，但忽视了捕捉小样本数据集上珍贵的局部序列特征和全局上下文信息，因此无法适应于复杂环境或有噪声的数据上。上述方法在英文音频和英文文本上取得显著的分类效果，但是由于中文字符的复杂性、词序之间的灵活性、数据的稀疏性以及语义表达的特殊性，使得上述方法提出的网络在中文音频与中文文本融合时会出现有效信息丢失、难以收敛等问题。本文在此基础上，从文本音频数据特征的优缺点出发，适应性的提出基于 ESimCSE-BiLSTM-CNN 进行文本特征学习，利用基于多头自注意力的 CNN-GRU 进行音频特征学习，并使用公开数据集对所提出的算法进行实验和验证，结果证明了本文算法在抑郁症检测中的可行性。

1　基于混合神经网络的多模态抑郁症检测算法

在音频模型的搭建上，本文使用的是 GRU（Gate Recurrent Unit）网络，其结构简单，拥有足够强大的长期依赖的建模能力。音频是一种非线性时间序列的转换信号，可以用 GRU 网络对抑郁音频进行建模。由于 GRU 在处理输入序列时难以有效识别和处理重要性较低的信息，导致其常忽略实质上更重要的信息，与之互补的 CNN（Convolutional Neural Networks）能够捕捉局部重要信息[9]。两者互补结合，模型尽管拥有不错的效果，但是仍然难以处理长距离的依赖关系，而多头自注意力机制不仅可以处理这样的问题，还能够使模型具有更丰富的信息交互能力。因此，搭建基于多头自注意力机制的 CNN-GRU 混合网络，可以学习到对音频信息的更佳表示。

在文本算法模型的搭建上，本文使用了 Bi-LSTM（Bidirectional LSTM）网络，由于该网络可以将前向和后向获取的信息合并在一个隐藏状态里，适用于需要考虑整个输入序列的任务[10]。因文本信息与时间背景密切相关，故文本建模使用 Bi-LSTM 建模。在此基础上，搭建 Bi-LSTM-CNN 混合网络模型，可以增加模型参数共享，权衡感受野的能力，使得模型能够适应于语义复杂性大的中文文本。

基于混合神经网络的多模态抑郁症检测算法的模型架构图如图 1 所示。在文本与音频模态融合时，因每个模态都可以提供独特的信息，故使用特征层融合的方法，以此充分利用多个模态信息，缓解模态间不平衡的问题。

1.1　中文音频特征学习

1.1.1　对数梅尔频谱图表示

为了让算法模型拥有人类听觉系统对声音的感知能力，同时弥补音频信号受噪声影响大的缺点，本文将音频信号进行了对数梅尔频谱图表示，之后将音频信号在梅尔滤波器组中分成 80 个频段，同时每次取对数进行一个小的偏移量的加法操作，让信号转化为对数梅尔频谱图表示时，能够保留关键特征，同时实现一种动态范围的压缩，使得较大振幅的信号相对较小振幅的信号变化更加明显。3 类情绪的部分声音特征的对数梅尔频谱图表示样例如图 2 所示。

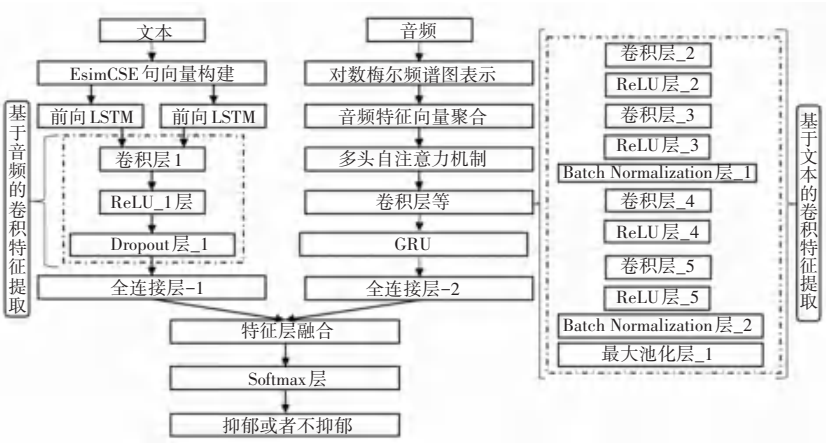
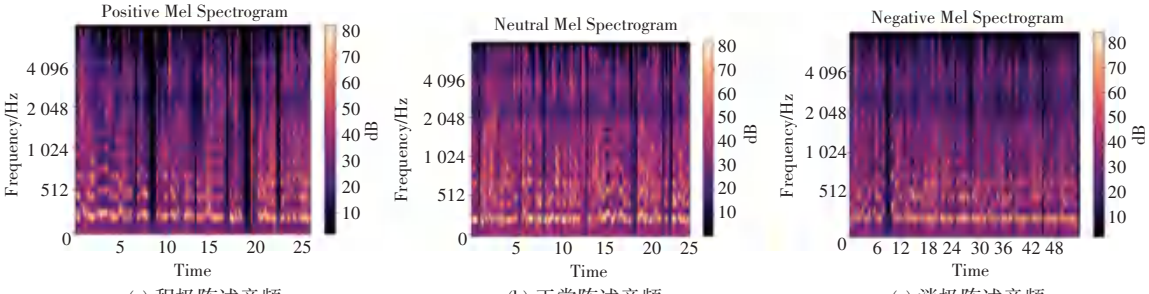


图 1 模型架构

Fig. 1 Model Architecture



(a) 积极陈述音频 (b) 正常陈述音频 (c) 消极陈述音频

图 2 3 类情绪的对数梅尔频谱图表示样例

Fig. 2 Sample representation of the log-Mayer spectrum diagram of three kinds of emotions

1.1.2 音频特征向量聚合

为了将对数梅尔频谱图的特征聚合成一个固定的长度,从而减小数据维度,降低运算成本,本文采用了 GhostVLAD(Ghost Vector of Locally Aggregated Descriptors)这一算法模块以期从梅尔频谱图上获得更好的特征表示。

GhostVLAD 基于 NetVLAD(Network - based Vector of Locally Aggregated Descriptors)层可实现类似于 VLAD(Vector of Locally Aggregated Descriptors)编码的编码[11]。NetVLAD 以 N 个 D 维的局部特征 x_i,K 个聚类中心 c_k 作为 VLAD 的参数,输出的特征是 $K \times D$ 维的[12]。实际上 VLAD 输出矩阵 $\boldsymbol{V}$ 是先会被转化为向量,再经过规范化操作,才能生成图片的全局特征。$\boldsymbol{V}(j,k)$ 计算过程如下式:

$$\boldsymbol{V}(j,k) = \sum_{i=1}^{N} \frac{e^{\omega_k^{\mathrm{T}} x_i + b_k}}{\sum_{k'} e^{\omega_{k'}^{\mathrm{T}} x_i + b_{k'}}} (x_i(j) - c_k(j)) \quad (1)$$

其中,b_k 和 c_k 都是可学习训练的参数,$k \in [1, 2, \cdots, k]$;ω_k^{T} 表示一个指示函数;$x_i(j)$ 和 $c_k(j)$ 分别表示第 i 个局部特征中第 j 维的值和第 k 个聚类中心中第 j 维的值。

GhostVLAD 的创新在于增加聚类中心个数:由 K 到 $K+G$ 个,但是增加的聚类中心在构建聚合特征矩阵时不参与贡献权重,相当于引入了一些噪声以增强网络本身的多样性,但又不会参与到权重的更新中,在端到端的训练中,可以让网络自主选择特征,甄别特征作用大小,GhostVLAD 结构如图 3 所示。通过采用 GhostVLAD 实现音频特征向量聚合,特征表示的判别性及捕捉强度信息的能力得到大幅加强。

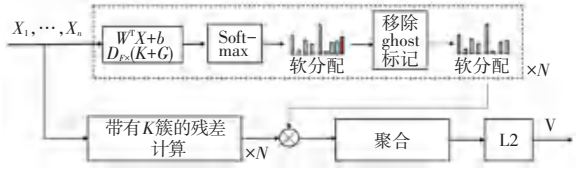


图 3 GhostVLAD 结构

Fig. 3 GhostVLAD structure

1.1.3 多头自注意力机制

多头自注意力机制是一种用于捕捉序列不同位置之间依赖关系的机制,在给定相同的查询、键和值的集合时,实现在相同的注意力机制情况下,学习到不同的表达[13]。将音频向量聚合后的向量通过 3 次不同的映射操作生成查询矩阵 $\boldsymbol{Q}$,键矩阵 $\boldsymbol{K}$ 及值

矩阵 $\boldsymbol{V}$。其注意力输出矩阵如下式：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}) \quad (2)$$

其中，d_k 为每个键的特征维度，表示缩放点积的数量级，平衡点积的数值范围；Softmax 是能够将注意力权重限制在 0~1 之间的归一化函数。

多头自注意力机制将音频时间序列划分为 h 个子空间，每个头对子空间进行自注意力运算，得到丰富的上下文信息，$head_i$ 表示模块中第 i 个头，公式如下：

$$head_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}}, \boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{Q}}, \boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{Q}}) \quad (3)$$

其中，$\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}}$，$\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{Q}}$，$\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{Q}}$ 分别表示查询矩阵、键矩阵以及值矩阵的权重矩阵。

最后，将 h 个头的结果进行拼接集成及线性变换，得到最终结果，$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$ 表示最终的输出结果，公式如下：

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(head_1,\cdots,head_h)\boldsymbol{W}^m \quad (4)$$

其中，$\boldsymbol{W}^m$ 表示线性变换的权重，Concat 表示拼接操作。

1.1.4　基于多头自注意力机制的 CNN-GRU 音频算法模型

借助多头自注意力机制，模型获得了更丰富的表达能力。在此基础上，采用 CNN-GRU 网络架构具有显著优势，主要体现在两者之间具有良好的互补性。具体而言，通过设置 CNN 网络，可以在多头自注意力机制提取的基础上，进一步提取出不同角度（如空间特征、语义关系）和不同子空间（如时序信息、动态特征）的丰富信息。这种设计使得 CNN 和 GRU 在特征提取和时序建模方面实现了协同作用，共同提升了模型的整体性能。

加入 CNN 网络后，模型往往难以高效训练。对于序列化的音频信号处理，采用循环神经网络对序列化特征进行提取，而 GRU 与 LSTM 能力相当的同时，能够大幅度提高训练效率[14]。GRU 的计算步骤可以分为以下 4 个主要步骤：

首先，对于重置门 r_t，通过当前位置输入 x_t 和上一位置隐层的输出 h_{t-1}，进行线性变化后再通过 *Sigmoid* 函数，得到一个 0 ~ 1 之间的值，而这个值决定了有多少信息需要被遗忘，公式如下：

$$r_t = \sigma(\boldsymbol{W}_r \cdot [h_{t-1}, x_t]) \quad (5)$$

其中，$\boldsymbol{W}_r$ 表示权值矩阵，σ 表示为 *Sigmoid* 函数。

其次，对于更新门 z_t，类似于重置门，更新门也通过当前位置输入和上一位置隐层的输出，计算出一个 0 ~ 1 之间的值。这个值决定了哪些信息需要被更新，公式如下：

$$z_t = \sigma(\boldsymbol{W}_z \cdot [h_{t-1}, x_t]) \quad (6)$$

其中，$\boldsymbol{W}_z$ 表示权值矩阵。

对于更新后的值 h'，是重置门 r_t、上一位置输出 h_{t-1} 和当前位置输入 x_t 共同作用的结果，公式如下：

$$h_t' = \tanh(W_h' \times [r_t \times h_{t-1}, x_t]) \quad (7)$$

其中，tanh 表示为双曲正切函数。

最后，对于当前位置的输出 h_t，其状态是由更新门选择更新后的值，与上一位置的输出进行加权相加得到，公式如下：

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times h_t' \quad (8)$$

综合上述多头自注意力机制与 CNN-GRU 的优异特性，基于多头自注意力机制的 CNN-GRU 网络模型能够更加精准地控制输入和记忆之间的交互。音频算法模型层次结构及参数配置见表 1。

表 1　音频算法模型层次结构及参数配置表

Table 1　Audio algorithm model hierarchy and parameter configuration table

层/块名称	输入大小	输出大小	卷积核大小
Multi-Head Attention	1 280	1 280	
Conv1d	1 280	512	3
Conv1d	512	256	3
BatchNorm1d	256		
Conv1d	256	128	3
Conv1d	128	64	3
BatchNorm1d	64		
MaxPool1d			2
GRU	64		
FC	64	2	

1.2　中文文本特征学习

1.2.1　ESimCSE 句向量构建

以 Transformer 为基础的文本预训练模型作为句嵌入映射的方法已经成为当前自然语言处理任务的主流，这些文本大模型通过预先学习大量文本数据，使得下游模型能够学习到更为丰富的语言表示，而 ESimCSE（Enhanced Sample Building Method for Contrastive Learning of Unsupervised Sentence Embedding）算法模型在情感分析方面已被证明具有突出的性能表现，这为无监督句子的向量构建提供了更鲁棒的解决方案[15]。

ESimCSE 模型是对 SimCSE 的一次改进，SimCSE 是通过 Transformer 来实现将位置嵌入编码

一个句子的长度信息[16]。正负对所包含的信息是不同的,由于这种差异性,使得用 SimCSE 训练的语义模型可能存在偏差。为避免这样的问题,ESimCSE 对批处理执行字符重复操作,其算法示意图如图 4 所示。通过对字符的重复操作,正数对的长度变化虽然发生了改变,但句子语义不变,这样便实现了削弱模型在预测正对时等长的提示。除此之外,ESimCSE 还实现了动量对比,即将前面几个小批量的模型输出保持在同一队列,这样损失计算中涉及的负对数量得到了扩充,EsimCSE 的损失函数也进一步变为:

$$\ell_i = -\log \frac{e^{sim(h_i,h_i^+)/\tau}}{\sum_{j=1}^{N} e^{sim(h_i,h_j^+)/\tau} + \sum_{m=1}^{M} e^{sim(h_i,h_m^+)/\tau}} \tag{9}$$

其中,$sim(h_i,h_j^+)$ 是相似度度量;τ 是一个温度超参数;N 是输入批次的批次大小;m 是队列大小;h_m^+ 表示动量更新后队列嵌入的句子。

利用动量对比的方法来增加损失计算中涉及的负对的数量,使得模型可以被激励,向更精细化的学习方向发展。

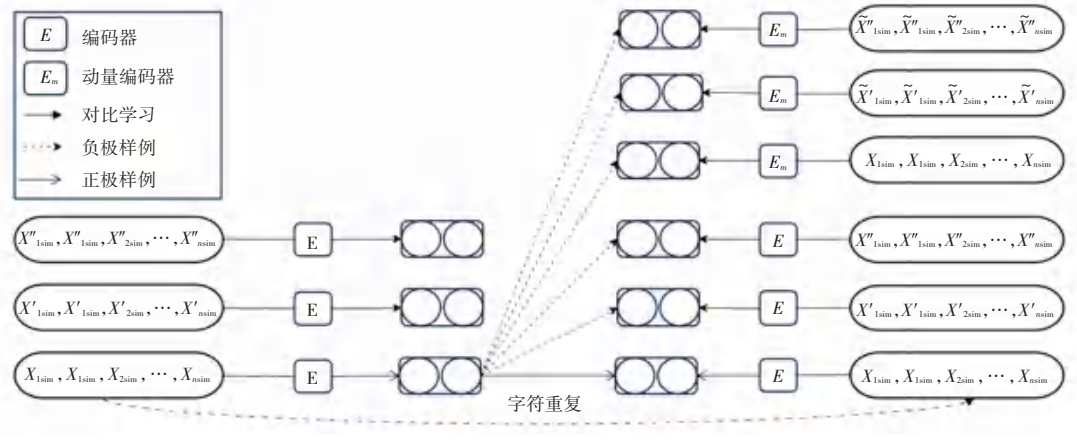


图 4 ESimCSE 算法示意图

Fig. 4 ESimCSE algorithm diagram

1.2.2 基于 EsimCSE-BiLSTM-CNN 文本算法模型

本文搭建了一个基于 EsimCSE-BiLSTM-CNN 的中文抑郁症文本分类算法模型,如图 5 所示,主要由句嵌入映射模块和深层特征提取分类模块构成。句嵌入映射模块通过 EsimCSE 的中文预训练模型来进行编码,将文本转化为句向量并进行对比学习与句嵌入映射;深层文本特征提取分类模块利用 BiLSTM-CNN 混合网络对上下文和局部特征信息进行提取,得到样本类别归属。

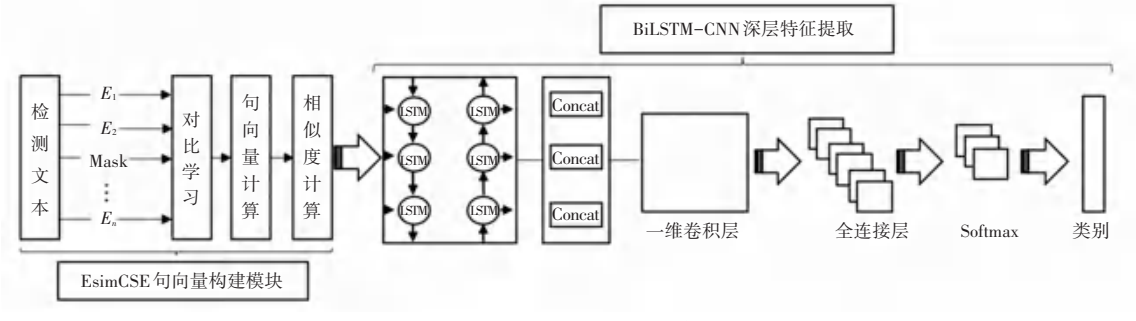


图 5 EsimCSE-BiLSTM-CNN 算法模型

Fig. 5 EsimCSE-BiLSTM-CNN algorithm model

EsimCSE-BiLSTM-CNN 混合网络模型结构的搭建步骤如下:

步骤 1 对输入的句子进行预处理,包括分词等,可以将句子转化为由词向量表示的序列,得到句子中每个词的表示;

步骤 2 采用对比学习的策略训练模型,辅以视角替换以及字符重复的方法,削弱正对等长带来的影响,使得同一语义下的句子在语义空间中更加接近,而不同语义下的句子则更加远离;

步骤 3 在对比学习的过程中,模型逐渐学习到句子的表示也就是句向量,通过相似度计算衡量句子之间的语义相似度,使得相似句子在向量空间中更加接近。

步骤 4 将 ESimCSE 句向量输入到 BiLSTM-

CNN 深层特征提取模块，经过 BiLSTM 循环神经网络，模型拥有了更全面的理解输入序列上下文的能力；由于 EsimCSE 卓越的句向量表示能力，建模时仅搭建了一层卷积网络，不仅提取了局部特征还防止了网络结构变得复杂；经过全连接层将特征连接起来，通过 Softmax 概率分布得到最后输出所属类别。

1.3　多模态特征层融合

从基于多头自注意力机制的 CNN-GRU 音频算法模型和基于 ESimCSE-BiLSTM-CNN 的文本算法模型的最后一层获取一个权重矩阵，对文本特征和音频特征进行线性变换，即将其特征统一到同一个输出空间，以便输入到二元交叉熵损失函数中。二元交叉熵损失函数可用下式表示：

$$Loss = -\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i)) \tag{10}$$

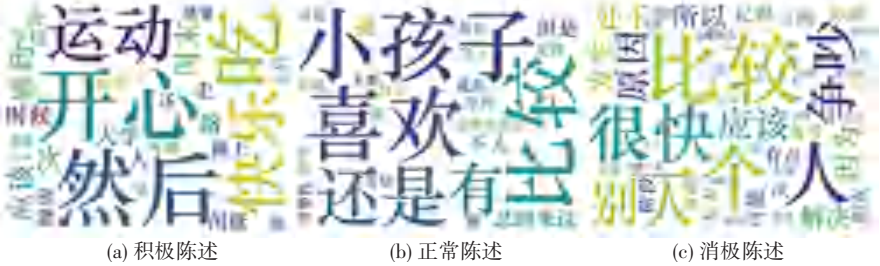
其中，y 是二元标签，$p(y)$ 为输出二元标签的概率。

通过将文本特征和音频特征的交叉熵损失值相加，得到了一个综合评估两者差距的损失值，并将其作为模型优化的指标。根据经过训练的模型权重向量，表示出不同模态的重要性，送入到全连接层从而得到分类概率。通过特征层融合的方式，使得模型在训练过程中可以同时考虑文本和音频特征的信息，进而提高了模型的训练效果和性能表现。

2　实验

2.1　数据集

为了验证模型性能的优劣，本文在公开的中文文本数据集 EATD-Corpus 进行性能验证，该数据集包含 162 名学生志愿者访谈中提取的积极陈述、正常陈述和消极陈述的音频和文本，根据 SDS(Self-Rating Depression Scale)得分是否大于等于 53 来判断一个志愿者是否为抑郁症志愿者，数据集中有 30 名抑郁志愿者和 132 名非抑郁志愿者。本文将数据集文本的部分情况用词云图表示，3 类情绪的云图表示如图 6 所示，可以看到抑郁症患者在表达不同情绪时的直观表现。



(a) 积极陈述　　(b) 正常陈述　　(c) 消极陈述

图 6　3 类情绪的云图表示

Fig. 6　Sample word cloud representation of three types of emotions

2.2　实验设置

实验采用的计算机处理器型号为 Intel Core i7-11700F，显卡为 NVIDIA GeForce GTX 3080Ti，文本模型的句向量维度、学习率、隐藏维度，训练轮次分别设置为 768、1e-4、128，120；音频模型的向量维度、学习率、训练轮次分别设置为 1 280、5e-6，45。两个单模态模型的优化器均选择了 AdamW。多模态融合模型的学习率和训练轮次设置为 5e-3 和 300，优化器选择了 Adam。

2.3　评价指标

本文采用 $F1$ 值、精确率和召回率 3 个指标作为测试集上文本单模态分类、音频单模态分类和音频文本多模态融合分类的评估标准。对于这些评价指标，其值越大性能越好。

2.4　实验结果

本文采用三折交叉验证的方法来验证实验模型的有效性，针对数据集不平衡问题，在模型训练前，将志愿者 3 种回答的顺序重新随机排列组合，使得抑郁类规模与非抑郁症类规模相当，实现数据增强。三折训练过程中基于混合神经网络的多模态抑郁症检测算法模型的准确率变化如图 7 所示，可见受制于中文数据的复杂性，准确率具有波动性，但是收敛速度达到理想水平，准确率在波动中逐渐提高。

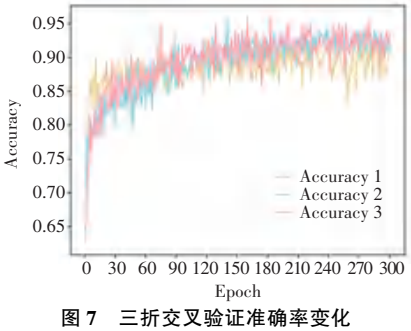


图 7　三折交叉验证准确率变化

Fig. 7　Change of accuracy rate of three-fold cross-validation

2.5 对比方法

本文将所提出的抑郁症音频单模态的算法与 Decesion Tree、SVM(Support Vector Machine)、Multi-modal LSTM[17], RF(Random Forest)、NetVLAD-GRU 算法进行了对比实验,实验结果见表 2,可见本文提出的算法在中文抑郁症音频检测领域的领先性能, $F1$ 值高出传统最优算法约 6%。

表 2 音频单模态实验结果对比

Table 2 Comparison of audio single-mode experimental results

特征	算法模型	$F1$ 值	召回率	精确率
音频	Decision Tree	0.45	0.44	0.47
	SVM	0.46	0.41	0.54
	Multi-modal LSTM	0.49	0.56	0.44
	RF	0.50	0.53	0.48
	NetVLAD-GRU	0.66	0.78	0.57
	本文	**0.72**	**0.89**	**0.60**

另外,将本文在抑郁症文本单模态的算法与其他算法如 SVM、RF、Decesion Tree、Multi-modal LSTM、Attention-BiLSTM、Bert-BiLSTM、ERNIE-BiLSTM-CNN、ESimCSE-BiLSTM、Bert-BiLSTM-CNN 进行了对比实验,Bert-BiLSTM 的句向量构建算法模型为 Bert-base-Chinese[18], ERNIE-BiLSTM-CNN 的句向量构建算法模型为 ERNIE-3.0-base-zh 模型[19],实验结果见表 3。可见本文提出的算法在中文抑郁症文本检测领域中取得了卓越的效果, $F1$ 值上相较于传统最优算法 Attention-BiLSTM 超出 23%,比 Bert-BiLSTM-CNN 高出 2%。

表 3 文本单模态实验结果对比

Table 3 Comparison of text single-mode experimental results

特征	算法模型	$F1$ 值	召回率	精确率
文本	SVM	0.64	**1.00**	0.48
	RF	0.57	0.53	0.61
	Decision Tree	0.49	0.43	0.59
	Multi-modal LSTM	0.57	0.63	0.53
	Attention-BiLSTM	0.65	0.66	0.65
	Bert-BiLSTM	0.68	0.71	0.64
	ERNIE-BiLSTM-CNN	0.69	0.63	0.76
	ESimCSE-BiLSTM	0.71	0.75	0.67
	Bert-BiLSTM-CNN	0.86	0.93	0.80
	本文	**0.88**	0.87	**0.90**

最后,将基于混合神经网络的多模态抑郁症检测算法模型与 Multi-modal LSTM 及 A GRU/BiLSTM-Model[20]算法模型进行了对比实验,结果见表 4。可以看到本文的算法设计与模态融合能力的不俗表现, $F1$ 值也达到了 0.909,与传统最优算法相比 $F1$ 值提高约 19%,召回率提高 3%。精确率提高 32%。

表 4 多模态融合实验结果对比

Table 4 Comparison of experimental results of multi-modal fusion

特征	模型	$F1$ 值	召回率	精确率
多模态融合	Multi-modal LSTM	0.57	0.67	0.49
	A GRU/BiLSTM-Model	0.71	0.84	0.62
	本文	**0.90**	**0.87**	**0.94**

2.6 分类效果可视化

在测试集上进行三折交叉验证的音频单模态、文本单模态,音频文本多模态的混淆矩阵情况如图 8 所示。通过混淆矩阵可见实验过程中不同折次、不同模态在数据增强的条件下测试数据的不同,泛化程度不一,同时可以直观感受到本文提出的算法模型在音频处理上相对还有改进空间,但经模态融合后,模态间极佳的互补性使得融合模型性能得到一定幅度的提升,最佳融合模型超过了任意单一模态的模型性能。

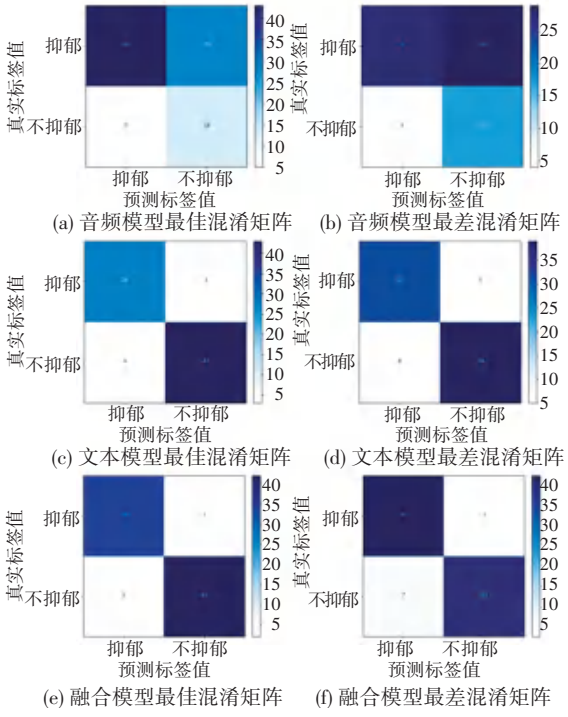


图 8 模型三折交叉验证混淆矩阵

Fig. 8 Model three-fold cross-validation confusion matrix

3 结束语

本文提出了一种基于混合神经网络的多模态融

合算法，针对中文文本特征创新性地设计了ESimCSE-BiLSTM-CNN 算法模型，通过对比学习优化句向量表示，并结合双向上下文建模与局部特征提取，显著提升了文本模态的语义表达能力。在音频模态中，提出的基于多头自注意力机制的 CNN-GRU 网络，通过融合动态特征聚合（GhostVLAD）、序列建模与跨子空间信息交互，有效捕捉了音频信号的非线性时序依赖与局部关键信息。通过特征层融合策略，模型进一步整合了文本与音频模态的互补性，较现有主流算法提升显著，期望为心理健康领域的客观诊断进一步发展提供有利支持。在未来的工作中，将进一步探索小样本场景下中文音频特征表达能力的提升策略，并引入更多模态（如生理信号等）以构建更全面的情感表征。

参考文献

[1] CLAYBORNECHONS Z M，VARIN M，COLMAN I. Systematic review and meta-analysis：Adolescent depression and long-term psychosocial outcomes - sciencedirect [J]. Journal of the American Academy of Child & Adolescent Psychiatry，2019，58(1)：72-79.

[2] PAPP M，CUBALA W J，SWIECICKI L，et al. Perspectives for therapy of treatment-resistant depression [J]. British Journal of Pharmacology，2022，179(17)：4181-4200.

[3] OZDAS A，SHIAVI R G，SILVERMAN S E，et al. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk [J]. IEEE Transactions on Biomedical Engineering，2004，51(9)：1530-1540.

[4] 李金鸣，付小雁. 基于深度学习的音频抑郁症识别[J]. 计算机应用与软件，2019，36(9)：161-167.

[5] CAI Hanshu，QU Zhitiao，LI Zhe，et al. Feature-level fusion approaches based on multimodal EEG data for depression recognition [J]. Information Fusion，2020，59：127-138.

[6] 郭威彤. 利用深度学习从面部表情和语音识别抑郁症方法的研究[D]. 兰州：兰州大学，2022.

[7] WILLIAMSON J R，GODOY E，CHA M，et al. Detecting depression using vocal，facial and semantic communication cues [C]//Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. 2016：11-18.

[8] AL HANAI T，GHASSEMI M M，GLASS J R. Detecting depression with audio/text sequence modeling of interviews [C]// Proceedings of the Annual Conference of the International Speech Communication Association，Interspeech. 2018：1716-1720.

[9] YUKUN C A O，JUNYI C. A short text semantic classification method for power grid service based on attention_gated recurrent unit（at_gru）neural network [C]//Proceedings of the 2018 5th International Conference on Systems and Informatics（ICSAI）. Piscataway，NJ：IEEE，2018：1105-1110.

[10] SIAMI-NAMINI S，TAVAKOLI N，NAMIN A S. The performance of LSTM and BiLSTM in forecasting time series [C]// Proceedings of the 2019 IEEE International Conference on Big Data. Piscataway，NJ：IEEE，2019：3285-3292.

[11] ZHONG Y，ARANDJELOVIĆ R，ZISSERMAN A. Ghostvlad for set-based face recognition [C]// Proceedings of the 14th Asian Conference on Computer Vision，Cham：Springer，2019：35-50.

[12] ARANDJELOVIĆ P，GRONAT P，TORII A，et al. NetVLAD：CNN architecture for weakly supervised place recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2018，40(6)：1437-1451.

[13] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need [C]//Proceedings of Advances in Neural Information Processing Systems. NIPS，2017：5998-6008. DOI：10.48550/arXiv.1706.03762.

[14] 孙志，王冠. 自监督对比学习的 CNN-GRU 语音情感识别算法[J]. 西安电子科技大学学报，2024，51(6)：182-193.

[15] WU X，GAO C，ZANG L，et al. Esimcse：Enhanced sample building method for contrastive learning of unsupervised sentence embedding [J]. arXiv preprint arXiv，2109.04380，2021.

[16] GAO T，YAO X，CHEN D. Simcse：Simple contrastive learning of sentence embeddings [J]. arXiv preprint arXiv，2104.08821，2021.

[17] AL HANAI T，GHASSEMI M M，GLASS J R. Detecting depression with audio/text sequence modeling of interviews [C]// Proceedings of the International Conference on Speech Communication. 2018：1716-1720.

[18] DEVLIN J，CHANG M W，LEE K，et al. BERT：Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. 2019：3498-4195.

[19] SUN Y，WANG S，FENG S，et al. Ernie 3.0：Large-scale knowledge enhanced pre-training for language understanding and generation [J]. arXiv preprint arXiv，2107.02137，2021.

[20] SHEN Y，YANG H，LIN L. Automatic depression detection：An emotional audio-textual corpus and a GRU/BiLSTM-based model [C]//Proceedings of the International Conference on Acoustics，Speech and Signal Processing（ICASSP）. Piscataway，NJ：IEEE，2022：6247-6251.