文章编号:2096-1472(2024)10-0068-06
DOI:10.19644/j.cnki.issn2096-1472.2024.010.014

一种基于多模态特征增强网络的抑郁症检测方法

赵小明[1,2]，范慧婷[1]，张石清[2]

(1.浙江理工大学信息科学与工程学院，浙江 杭州 310018；
2.台州学院智能信息处理研究所，浙江 台州 318000)
✉ tzxyzxm@163.com; courage_f@163.com; tzczsq@163.com

摘 要:针对传统的多模态融合方法在抑郁症检测中忽略了模态之间的交互性、未能充分提取出更全面的特征表示的问题,本研究提出一种基于多模态特征增强网络的抑郁症检测方法,该方法有效地集成了视频、音频和远程光电容积脉搏(photoplethysmographic,rPPG)信号3种模态,通过模态间Transformer、模态内Transformer和多头自注意力机制,共同学习输入模态序列每个时间步的模态内和模态间的动态关系,达到了特征增强的目的。最终,拼接3个模态增强后的特征获得全面特征表示。在AVEC2013公共数据集上的实验结果显示,该方法的平均绝对误差为7.07,优于单模态抑郁症检测,表明该方法有效促进了模态之间的交互,并实现了特征增强,在自动抑郁症检测任务中展现出显著的有效性。

关键词:多模态;深度学习;抑郁症检测;卷积神经网络;特征增强;多模态融合
中图分类号:TP391.41 **文献标志码**:A

A Depression Detection Method Based on Multimodal Feature Enhancement Network

ZHAO Xiaoming[1,2], FAN Huiting[1], ZHANG Shiqing[2]

(1.*School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou* 310018, *China*;
2.*Institute of Intelligent Information Processing, Taizhou University, Taizhou* 318000, *China*)
✉ tzxyzxm@163com; courage_f@163.com; tzczsq@163.com

Abstract: Traditional multimodal fusion methods tend to overlook the interactivity between modalities and fail to extract comprehensive feature representations in depression detection. To address these problems, this paper proposes a depression detection method based on a multimodal feature enhancement network, which effectively integrates three modalities: video, audio, and remote photoplethysmo graphy (rPPG) signals. By employing inter-modal Transformers, intra-modal Transformers, and a multi-head self-attention mechanism, the method learns the dynamic relationships both within and between modalities for each time step of the input modality sequence, achieving feature enhancement. Ultimately, the enhanced features from the three modalities are concatenated to obtain a comprehensive feature representation. Experimental results on the AVEC2013 public data set indicate that the proposed method achieves an average absolute error of 7.07, outperforming traditional unimodal depression detection methods. This demonstrates that the proposed method effectively facilitates interaction between modalities and enhances features, showing significant effectiveness in automated depression detection tasks.

Key words: multimodal; deep learning; depression detection; convolutional neural network; feature enhancement; multimodal fusion

0 引言(Introduction)

早期诊断抑郁症在促进治疗效果方面起着至关重要的作用。但是目前抑郁症的诊断依赖于主观行为,例如患者的自我报告评估和临床判断症状严重程度,而这些因素容易受到环境因素的影响。

如何有效地进行自动多模态抑郁症检测,以辅助医生实现

收稿日期:2024-02-27
基金项目:国家自然科学基金项目(62276180)

早期抑郁症的诊断，已成为当前一个既重要又具有挑战性的研究问题。因此，运用机器学习等技术进行抑郁症自动检测[1]的研究受到广大研究者的关注。然而，传统的融合方法通常直接采用简单的级联方式融合多模态特征，这种方式忽略了模态之间的交互性，无法充分提取出更全面的特征表示，从而影响了抑郁症的检测效果。

因此，本文探索了一种基于多模态特征增强网络的抑郁症检测方法，该方法融合了音频、视频及 rPPG 信号，其中 rPPG 模态作为一种附加模态，增强了多模态抑郁症检测的效果，通过堆叠多个模态间和模态内 Transformer，并配合多头自注意力机制，共同获取输入序列每个时间步的模态内和模态间的信息交互，以达到多模态特征增强的目的，从而提升抑郁症检测性能。

1　相关研究(Related research)

目前，主流的抑郁症检测方法主要可以分为 3 类：基于视频的检测、基于音频的检测和基于多模态的检测。

抑郁症患者常常表现出面部表情的减少或呆滞，他们的面部表情可能缺乏生动度和情感表达。研究者通过机器学习分析面部特征在辅助诊断抑郁症方面取得了比较大的进展[2]。例如，孙浩浩等[3]基于人脸图像的全局和局部特征，构建了一种融合通道层注意力机制的多支路卷积网络模型。音频作为传达情感的媒介，抑郁症患者和非抑郁症患者之间的言语模式存在明显的差异[4]。MA 等[5]提出了 DepAudioNet 深度模型，结合卷积神经网络(CNN)和长短期记忆(LSTM)，用于编码声道中的抑郁症相关特征，从而提供更全面的音频表示，取得了较好的检测效果。这些深度学习架构在提取有意义的音频或视频特征以及提高抑郁症检测的效果方面发挥着重要作用。然而，仅依赖音频或者视频特征可能会丢失测试对象的其他动态信息，从而限制了抑郁症检测的性能。

除了视频和音频模态，rPPG 信号也可用于抑郁症检测。rPPG 信号使用非接触式光学技术测量和分析心率和血流量等生理信息。一些研究通过提取 rPPG 信号并计算统计特征和心率变异性(HRV)特征，探讨了抑郁症与 HRV 之间的关系[6-7]。这些特征随后被输入基于随机森林和多层感知机(Multilayer Perceptron，MLP)的机器学习回归器中。这些发现证明了基于 rPPG 的抑郁症检测方法的潜力。然而，很少有研究关注和探索用于抑郁症检测的 rPPG 信号。

除了上述单模态方法，通过多模态信息融合方法整合多种模态在提高抑郁症检测性能方面也显示出不错的效果。HE 等[8]通过特征层融合将提取的音频和视频特征串联成一个高维特征向量，并使用支持向量回归(SVR)进行抑郁症预测。然而，这种方法容易产生高维特征表示，从而导致维度灾难。YANG 等[9]将获得的音频视频结果和文本结果进行决策融合，以获得最终的抑郁症检测结果。但是，决策层融合单独考虑不同模态，无法捕捉它们之间的内在关系。更多的研究者通过模型层融合考虑模态之间的关系。NIU 等[10]采用多模态注意力特征融合方法整合音频模态和视频模态。谷明轩等[11]结合了脑电信号和音频特征提出了基于全连接神经网络的多模态特征融合模型。但是，这些模型层融合方法在模态之间的交互性方面仍存在不足。近年来，Transformer[12]技术引起了广泛关注，Transformer 模型中的编码器和解码器组件利用多头自注意机制捕捉输入序列数据的长距离上下文信息。ILIAS 等[13]提出了一种将语言之外的信息融入基于 Transformer 的模型，用于社交媒体中抑郁症和压力检测，这一方法展现出良好的应用前景。

受到 Transformer 技术的优势和 rPPG 信号在抑郁症检测中潜力的启发，本研究提出一种基于多模态特征增强网络的抑郁症检测方法。首先，针对视频、音频和 rPPG 模态进行多模态特征提取；其次，通过基于 Transformer 的特征增强模块和多头自注意力机制，实现不同模态之间的交互；最后，利用多层感知机实现最终的抑郁症检测任务。

2　基于多模态特征增强网络的抑郁症检测方法(Depression detection method based on multimodal feature enhancement network)

基于多模态特征增强网络的抑郁症检测方法的整体结构如图 1 所示，该结构主要包括多模态特征提取、多模态特征增强和回归预测 3 个部分。(1)多模态特征提取：对于视频模态和音频模态，本文采用深度 CNN[14-15]提取高级视频和音频特征。对于 rPPG 模态，采用短时端到端 rPPG 估计框架[16]提取 rPPG 信号值。(2)多模态特征增强：模态之间的 Transformer 通过与其他模态之间进行信息交互，用于增强目标模态的特征。模态内 Transformer 聚焦于目标模态，对目标模态内部特征进行交互，关注到目标模态中最相关和有价值的信息。多头自注意力机制提取更丰富、更有用的特征，平均池化聚合目标模态特征。(3)回归预测：通过级联和自注意力机制处理增强后的特征并输入多层感知机网络进行最终的抑郁症预测。

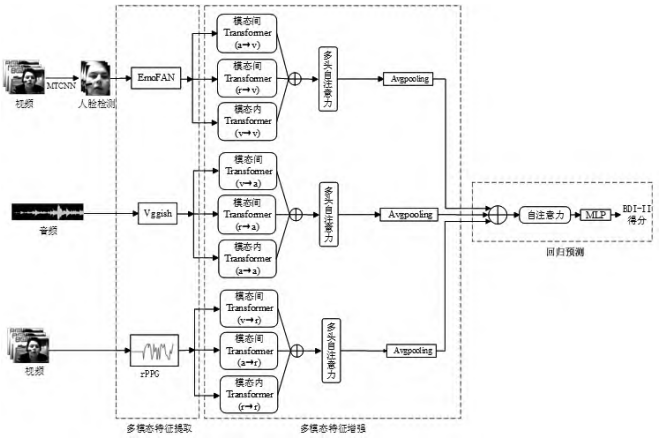

图 1　基于多模态特征增强网络的抑郁症检测方法的整体结构

Fig. 1 The overall structure of depression detection method based on multimodal feature enhancement network

2.1　多模态特征提取

对于视频模态，首先从每个视频样本中提取 100 个连续的帧，使用多任务级联卷积神经网络(MTCNN)[17]进行人脸检测任务，其次使用 EmoFAN[14]预训练深度卷积神经网络模型提取每个视频帧的面部特征。

对于音频模态，本文使用预训练的 VGGish[15]深度卷积神经网络模型进行特征提取。VGGish 模型在一百万个音频片段上进行了预训练，并为每个谱图段生成 128 维特征。

对于 rPPG 模态，本文使用短时端到端 rPPG 估计框

架[16]，该框架能够从视频流中检测到由血容量脉搏引起的微小颜色变化，进而实现 rPPG 的有效估计。具体来说，在人脸检测之后，首先使用类似 Unet[18] 的深度学习模型选择和跟踪感兴趣区域并进行皮肤和非皮肤像素的语义分割训练；其次计算皮肤分割像素的空间 RGB 通道均值，并将其投影到垂直于肤色的平面上，通过调整投影信号的 alpha 值获得 rPPG 信号值。

2.2 多模态特征增强网络

本节将详细介绍多模态特征增强网络的相关模块，该网络由多个模态间 Transformer（图 2）、模态内 Transformer（图 3）和多头自注意力机制组成，旨在共同捕捉输入序列每个时间步的模态内和模态间的动态关系，从而学习跨模态的渐进综合特征。

本文使用 v、a、r 分别表示视频模态、音频模态和 rPPG 模态。m 和 n 分别代表两种不同的模态。通过使用 Transformer 的注意力机制，逐步学习跨模态的全面特征。更具体地说，模态间 Transformer 通过与其他模态之间进行信息交互，学习到更全面的特征表示，以增强目标模态的特征。模态内 Transformer 聚焦于目标模态，对其目标模态内部特征进行交互，从而捕捉输入特征中不同位置之间的依赖关系，关注到目标模态中最相关和有价值的信息，并且这些 Transformer 使用编码器结构，从另一个模态序列 U_n 中提取当前模态序列 U_m 的信息。查询、键和值是 Transformer 编码器的 3 个输入。查询的来源是 U_m，键和值的来源是 U_n。因此，Transformer 编码器可以表示为

$$U_{n\to m}=Transformer(U_m,U_n,U_n)\in R^{T_m,d},m,n\in\{v,a,r\} \tag{1}$$

$$U_{m\to m}=Transformer(U_m,U_m,U_m)\in R^{T_m,d},m\in\{v,a,r\} \tag{2}$$

考虑到目标模态与不同模态交互后的模态特征存在差异性，为了后续进行更好的融合，本文将从模态内 Transformer 和模态间 Transformer 获得的所有特征进行连接，作为增强特征的输出，通过多头自注意力机制提取更丰富、更有用的特征，该过程定义为

$$U=Concat([U_{n_1\to m},U_{n_2\to m},U_{m\to m}])\in R^{T_m,3d} \tag{3}$$

$$U'_m=SA^{[i],mul}(U)\in R^{T_m,3d} \tag{4}$$

其中：$m\in\{v,a,r\}$，n_1、n_2 表示除 m 之外的其他两种模态，$SA^{[i],mul}$ 表示多头自注意力机制，i 表示第 i 个头部的注意力计算结果。考虑到跨模态的交互作用，期望增强的特征能够充分利用不同模态之间的互补性，生成更全面、更具表现力的特征表示。使用平均池化聚合目标模态特征，从而获得适用于下游任务的聚合特征。

2.2.1 模态间 Transformer

模态间 Transformer 利用跨模态注意力机制处理不同模态之间的交互，使得一个模态的信息能够影响另一个模态的表示，从而对目标模态进行特征增强，模态间 Transformer 网络结构如图 2 所示。模态间 Transformer 是建立在跨模态注意力机制上的，下文将详细介绍跨模态注意力机制的原理。

假设有两种模 m 和 n，分别对应着相应的序列 $X_m\in R^{(T_m\times d_m)}$ 和 $X_n\in R^{(T_n\times d_n)}$。其中，$T$ 和 d 分别表示序列的长度和特征维度。受到原始 Transformer[12] 模型的启发，本文使

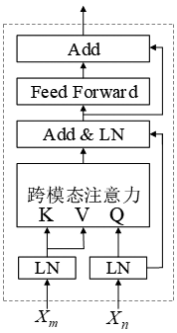

图 2 模态间 Transformer 网络结构

Fig. 2 Inter-model Transformer network

用 $m\to n$ 表示跨模态交互的潜在适应性，使用的模态包括视频、音频和 rPPG 信号。将查询、键、值矩阵分别定义为 $\boldsymbol{Q}_m=X_m\boldsymbol{W}_{Q_m}$、$\boldsymbol{K}_n=X_n\boldsymbol{W}_{K_n}$、$\boldsymbol{V}_n=X_n\boldsymbol{W}_{V_n}$，其中，$\boldsymbol{W}_{Q_m}\in R^{d_m\times d_k}$、$\boldsymbol{W}_{K_n}\in R^{d_n\times d_k}$ 和 $\boldsymbol{W}_{V_n}\in R^{d_n\times d_k}$ 是权重矩阵。$m\to n$ 的潜在适应性被定义为跨模态注意力，表示为

$$\begin{aligned}Y_m&=CM_{m\to n}(X_m,X_n)\in R^{T_m\times d_v}\\&=\text{softmax}\left(\frac{\boldsymbol{Q}_m\boldsymbol{K}_n^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_n\\&=\text{softmax}\left(\frac{X_m\boldsymbol{W}_{Q_m}\boldsymbol{W}_{K_n}^{\mathrm{T}}X_n^{\mathrm{T}}}{\sqrt{d_k}}\right)X_n\boldsymbol{W}_{V_n}\end{aligned} \tag{5}$$

其中：Y_m、$\boldsymbol{Q}_m$ 和 V_n 具有相同的长度（即 T_m 和 T_n），T 表示向量的转置操作。在公式（1）中，softmax 函数通过除以 $\sqrt{d_k}$ 进行缩放计算，进而得到一个得分矩阵 softmax(·)$\in R^{T_m\times T_n}$，其中第 (i,j) 个元素表示模态 m 的第 i 个时间步对模态 n 的第 j 个时间步的注意力。因此，Y_m 的第 i 个时间步是通过权重为 softmax(·)中第 i 行确定的对 V_n 进行加权汇总的结果。在这个意义上，公式(5)表示跨模态注意力。

基于跨模态注意力块，本文设计了模态间 Transformer，使一个模态可以从另一个模态获取交互信息。以将视觉(v)模态增强音频（a）模态为例，表示为 $v\to a$。每个模态间 Transformer 由 D 层跨模态注意力块组成，其中 $i=1,\cdots,D$ 表示层数。模态间 Transformer 公式具体如下：

$$\begin{aligned}X_{a\to v}^{[0]}&=X_a^{[0]}\hat{X}_{a\to v}^{[i]}\\&=CM_{a\to v}^{[i],mul}(LN(X_{a\to v}^{[i-1]}),LN(X_{a\to v}^{[0]}))+LN(X_{a\to v}^{[0]})X_{a\to v}^{[i]}\\&=f_{\theta_{a\to v}^{[i]}}(LN(\hat{X}_{a\to v}^{[i]}))+LN(\hat{X}_{a\to v}^{[i]})\end{aligned} \tag{6}$$

其中：X 是输入特征，f_θ 是由 θ 参数化的逐元素位置前馈子层，$CM_{a\to v}^{[i],mul}$ 表示多头注意力。LN 表示层归一化。在文中设置 $D=2$。

2.2.2 模态内 Transformer

设计的模态内 Transformer 采用多头自注意力机制，专注于对每种模态内部的动力学进行建模，获得单一模态内的时间依赖性，使模块能够理解每种模态的信息随时间的演变。基于原始 Transformer 中的自注意力机制，本文设计了模态内 Transformer，用于跨单个模态的特征增强，模态内 Transformer 网络结构如图 3 所示。同样，本文定义 $\boldsymbol{Q}_m=X_m\boldsymbol{W}_{Q_m}$、$\boldsymbol{K}_m=X_m\boldsymbol{W}_{k_m}$、$\boldsymbol{Q}_m=X_m\boldsymbol{W}_{V_m}$ 为查询、键、值矩阵，多头

自注意力表示为

$$\begin{aligned} Y_m &= SA_{m\to m}(X_m, X_m) \in R^{T_m \times d_v} \\ &= \text{softmax}\left(\frac{\boldsymbol{Q}_m \boldsymbol{K}_m^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_m n \\ &= \text{softmax}\left(\frac{X_m \boldsymbol{W}_{Q_m} \boldsymbol{W}_{K_m}^{\mathrm{T}} X_m^{\mathrm{T}}}{\sqrt{d_k}}\right) X_m \boldsymbol{W}_{V_m} \end{aligned} \tag{7}$$

以音频模态为例,模态内 Transformer 表示为

$$\begin{aligned} X_{a\to a}^{[0]} &= X_a^{[0]} \hat{X}_{a\to a}^{[i]} \\ &= SA_{a\to a}^{[i],mul}(LN(X_{a\to a}^{[i-1]}), LN(X_{a\to a}^{[0]})) + LN(X_{a\to a}^{[0]}) X_{a\to a}^{[i]} \\ &= f_{\theta_{a\to a}^{[i]}}(LN(\hat{X}_{a\to a}^{[i]})) + LN(\hat{X}_{a\to a}^{[i]}) \end{aligned} \tag{8}$$

其中,$SA_{a\to a}^{[i],mul}$ 表示多头自注意力。

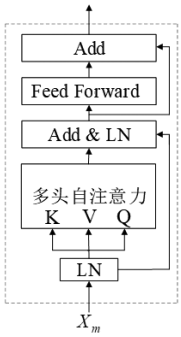

图 3　模态内 Transformer 网络结构

Fig. 3 Intra-model Transformer network

通过结合模态间和模态内的 Transformer 输出,并利用多头自注意力机制,能够创建输入序列更全面的表示。这种方法使模型能够更全面地理解输入数据,并进一步实现整体的特征增强。在每种模态都通过多头自注意力机制获得统一增强的特征后,使用平均池化方法对这些特征进行展平处理。

$$X'_m = \text{Avgpooling}(U'_m) \in R^k \tag{9}$$

其中:U'_m 是多头自注意力的输出,k 是输出特征的维度。

2.3　抑郁症预测

本文将 3 个模态的特征级联在一起后进行最终的抑郁症预测,并且使用自注意力机制过滤掉一些无用的特征,使用多层感知机(Multilayer Perceptron,MLP)进行最后的抑郁症分数预测。该过程可以表示为

$$Z = X'_v \oplus X'_a \oplus X'_r \in R^{3k} \tag{10}$$

$$\text{Attention}(Z) = \text{softmax}\left(\frac{\boldsymbol{Q}_m \boldsymbol{K}_m^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_m \in R^{3k} \tag{11}$$

$$y = MLP(Z) \tag{12}$$

其中,y 是最后的抑郁症得分。

3　实验与分析(Experiment and analysis)

3.1　数据集

本文在公开可用的抑郁症数据集——音频/视觉情感挑战 AVEC2013[19] 上进行了实验。该数据集中的每个视频都附带有从贝克抑郁量表问卷(BDI-Ⅱ)回答中获得的标签,该量表将得分划分为 0～63 的范围。根据 BDI-Ⅱ得分,抑郁的严重程度可以分为 4 个级别:最轻微(0～13)、轻度(14～19)、中度(20～28)和严重(29～63)。AVEC2013 数据集由 3 个不同的部分组成:训练、验证和测试,每个部分包含 50 个视频,总共有 150 个视频用于实验分析。

3.2　评估指标

平均绝对误差(MAE)和均方根误差(RMSE)是评估自动抑郁检测任务中方法性能的常用指标。

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{13}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{14}$$

其中:N 是样本数量,y_i 是样本标签,$\hat{y}_i$ 是预测值。

3.3　实验细节

对于视频模态,本文从每个视频中选择了 100 个连续帧,并使用 EmoFAN[14] 预训练模型提取 128 维的面部特征,维度为(100,128)。对于音频模态,本文使用 VGGish[15] 预训练模型提取 128 维的音频特征,维度为(*num_segments*,128),其中 *num_segments* 是分割后的频谱图段数。对于 rPPG 模态,获得了维度为(*num_seconds*,10)的特征,其中 *num_seconds* 是原始视频样本的持续时间。对于音频和 rPPG 模态,本文采用自适应平均池化[20]将提取的特征转换为(100,128)的固定特征维度供后续任务使用。使用的自适应平均池化[20]可以将具有任意空间维度的特征图转换为固定大小的表示。

所有深度学习方法都在 PyTorch 框架上进行,并使用 NVIDIA RTX 3090 GPU 进行计算。使用 Adam 优化器,初始学习率设置为 0.001,权重衰减设置为 0.000 05。采用批量大小为 4,并将最大训练轮数设置为 1 000。

3.4　实验对比结果和分析

3.4.1　AVEC2013 数据集实验结果

本研究在 AVEC2013 数据集上对比了单模态和多模态的结果。针对单模态情况,本文在模型中去除了模态间 Transformer 模块。如表 1 所示,在 AVEC2013 数据集上,视频模态的表现优于音频模态和 rPPG 模态。视频模态的 MAE 为 8.67,而音频模态和 rPPG 模态分别为 9.03 和 10.01。这一优势可能源于视频中的面部表情为抑郁症检测提供了更多的线索。对于多模态融合方法,综合考虑 3 个模态的结果优于仅考虑两个模态的结果,这表明 3 个模态在一定程度上相互补充,为抑郁症检测提供了更全面的线索。同时,验证了 rPPG 信号在多模态抑郁症检测中的有效性,为抑郁预测提供了额外的辅助信息。综上所述,将 3 个模态进行融合能够达到最佳的性能,验证了本文提出模型的有效性。

表 1　本文方法在 AVEC 2013 数据集上取得的结果

Tab.1 Results obtained by the proposed method on the AVEC2013 data set

模态	特征类型	*MAE*	*RMSE*
单模态	音频	9.03	12.46
	视频	8.67	11.07
	rPPG	10.01	12.95
多模态	音频+视频	8.06	10.65
	音频+rPPG	8.83	12.21
	视频+rPPG	8.39	10.87
	音频+视频+rPPG	**7.07**	**9.35**

3.4.2 消融实验

为了评估模型中每个多模态特征增强模块的有效性，在AVEC2013数据集上进行消融实验。本文进行了不同模块的组合实验，实验结果如表2所示。其中，模态间、模态内和多头自注意力分别表示模型多模态特征增强部分仅使用模态间Transformer、模态内Transformer或多头自注意力机制，“+”表示使用两个模块的组合。

表2中的结果显示：仅使用模态间Transformer的性能优于仅使用模态内Transformer，但低于这两个模块的组合使用。这表明，通过模态间Transformer中的跨模态注意力机制，能实现模态之间更有效地交互，从而对目标模态实现特征增强，同时，模态内Transformer也能关注到目标模态在时间上的变化信息。因此，将这两个模块结合使用能够获得更好的效果。此外，单独使用多头自注意力机制模型效果并不理想，然而当与模态间Transformer和模态内Transformer模块结合使用时，达到了本模型的最优效果。这表明，多头自注意力机制在一定程度上能够学习到模态交互后更全面的特征，从而实现整体的特征增强。

表2 不同模块对实验的影响

Tab.2 The impact of different modules on experiments

模块类型	*MAE*	*RMSE*
模态间	7.88	10.26
模态内	8.13	10.92
模态间+模态内	7.53	10.02
多头自注意力	8.49	11.68
多模态特征增强网络	**7.07**	**9.35**

3.4.3 不同模型对比结果

为了更全面地评估本文提出模型的有效性，在AVEC2013数据集上将其与目前较先进的方法进行了对比，对比结果如表3所示。根据所使用的模态数量，这些方法可以分为3个主要类别：基于音频的抑郁症检测方法、基于视频的抑郁症检测方法以及基于音频和视频的双模态抑郁症检测方法。以下是对一些具有代表性检测方法的简要介绍，更多的信息可参考表3中列出的相关文献。

基于音频的抑郁症检测方法：VALSTAR等[19]提取了LLD声学特征，并采用支持向量回归进行抑郁症检测。HE等[21]将深度音频特征与深度CNN和手工纹理特征相结合后，通过全连接层进行抑郁得分预测。NIU等[22]提取短时MFCC段的分段级特征并采用支持向量回归预测个体的抑郁水平。ZHAO等[23]提出了一种混合特征提取网络，将DCNN与自注意力网络集成，用于从语音信号中检测抑郁严重程度。

基于视频的抑郁症检测方法：ZHU等[24]提取了LPQ-TOP特征，并通过稀疏编码进行学习，以进一步提高抑郁症检测的准确性。JAZAERY等[25]使用3D卷积神经网络(3D-CNN)捕捉面部区域在两个不同尺度上的时空特征，并在决策层上进行融合。HE等[21]提出了一种名为DepNet的集成框架，用于捕捉视频中面部表情的时间动态特征，以进行抑郁症分析。

基于音频和视频的双模态抑郁症检测方法：MENG等[26]使用LLD声学特征对音频特征进行编码，并使用运动历史直方图捕捉面部区域内每个像素的运动，最终在决策层上融合音频和视频特征后，进行最终的抑郁症预测。NIU等[10]提出了一种时空注意网络和多模态注意特征融合策略，用于通过音频和视频预测个体的抑郁水平。

表3 AVEC2013数据集测试对比结果

Tab.3 Comparison of test results on AVEC2013 test set

模态类型	方法	*MAE*	*RMSE*
音频	文献[19]方法	10.35	14.12
	文献[21]方法	8.20	10.00
	文献[22]方法	7.48	9.79
	文献[23]方法	7.38	9.65
视频	文献[10]方法	7.14	9.50
	文献[24]方法	7.58	9.82
	文献[25]方法	7.37	9.28
	文献[21]方法	7.55	9.20
	文献[27]方法	7.52	9.22
音频+视频	文献[26]方法	8.72	10.96
	文献[28]方法	7.68	9.44
	文献[29]方法	9.09	11.19
音频+视频+rPPG	本文方法	**7.07**	**9.35**

4 结论(Conclusion)

针对自动抑郁症检测任务中传统的方法存在的问题，例如不能充分利用不同模态信息、未充分考虑多模态融合过程中模态间的交互等，本文提出了一种基于多模态特征增强网络的抑郁症检测方法。该方法通过与不同模态之间的交互，实现目标模态的特征增强，并融合了多种模态，将rPPG模态与视频模态和音频模态结合应用于多模态抑郁症检测任务。本文提出的方法利用模态间Transformer、模态内Transformer和多头自注意力机制逐步学习视频、音频和rPPG等不同模态的综合特征。在AVEC2013公共数据集上进行的大量实验证明，本文提出的方法在多模态抑郁症检测任务上展现出良好的性能。

本文提出的模型能更好地挖掘不同模态中的抑郁线索，为多模态融合提供了新思路。在未来工作中，我们将探索跨模态对齐的先进方法，以期进一步提高多模态融合效果。此外，研究发现，rPPG信号的性能并不优于视频模态和音频模态，这可能是由于提取rPPG信号值的方法不够精确。因此，使用更先进的方法提取更具表达力的生理信号，并将其应用于多模态融合具有重要的研究意义。

参考文献(References)

[1] 赵健，周莉芸，武孟青，等. 基于人工智能的抑郁症辅助诊断方法[J]. 西北大学学报(自然科学版)，2023，53(3)：325-335.

[2] 李欣，范青. 机器学习在抑郁症患者面部特征研究中的应用进展[J]. 上海交通大学学报(医学版)，2022，42(1)：124-129.

[3] 孙浩浩，邵珠宏，尚媛园，等. 融合通道层注意力机制的多支路卷积网络抑郁症识别[J]. 中国图象图形学报，2022，27(11)：3292-3302.

[4] 刘振焘，向春妮，刘陈陵，等. 基于语音的抑郁检测研究综述[J]. 信号处理，2023，39(4)：616-631.

[5] MA X C，YANG H Y，CHEN Q，et al. Depaudionet：an efficient deep model for audio based depression classification[C]//ACM.

Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. New York:ACM,2016:35-42.
[6] DAGDANPUREV S,SUN G H,SHINBA T,et al. Development and clinical application of a novel autonomic transient response-based screening system for major depressive disorder using a fingertip photoplethysmography sensor[J]. Frontiers in bioengineering and biotechnology,2018,6(2):64-74.
[7] CASADO C Á,CAÑELLAS M L,LÓPEZ M B. Depression recognition using remote photoplethysmography from facial videos[J]. IEEE transactions on affective computing,2023,14(4):3305-3316.
[8] HE L,JIANG D M,SAHLI H. Multimodal depression recognition with dynamic visual and audio cues[C]//IEEE. Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). Piscataway:IEEE,2015:260-266.
[9] YANG L,JIANG D M,SAHLI H. Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures[J]. IEEE transactions on affective computing,2021,12(1):239-253.
[10] NIU M Y,TAO J H,LIU B,et al. Multimodal spatiotemporal representation for automatic depression level detection[J]. IEEE transactions on affective computing,2023,14(1):294-307.
[11] 谷明轩,范冰冰. 基于多模态特征融合的抑郁症识别[J]. 计算机与现代化,2023,2023(10):17-22.
[12] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[J]. Advances in neural information processing systems,2017,30(2):225-235.
[13] ILIAS L,MOUZAKITIS S,ASKOUNIS D. Calibration of transformer-based models for identifying stress and depression in social media[J]. IEEE transactions on computational social systems,2024,11(2):1979-1990.
[14] TOISOUL A,KOSSAIFI J,BULAT A,et al. Estimation of continuous valence and arousal levels from faces in naturalistic conditions[J]. Nature machine intelligence,2021,3:42-50.
[15] HERSHEY S,CHAUDHURI S,ELLIS D P W,et al. CNN architectures for large-scale audio classification[C]//IEEE. Proceedings of the 2017 IEEE International Conference on Acoustics,Speech and Signal Processing. Piscataway:IEEE,2017:131-135.
[16] LI B,JIANG W,PENG J,et al. Deep learning-based remote-photoplethysmography measurement from short-time facial video[J]. Physiological measurement,2022,43(11):115-123.
[17] JIANG B,REN Q,DAI F,et al. Multi-task cascaded convolutional neural networks for real-time dynamic face recognition method[C]//LIANG Q,LIU X,NA Z,et al. International Conference in Communications, Signal Processing,and Systems. Singapore:Springer,2020:59-66.
[18] RONNEBERGER O,FISCHER P,BROX T. U-net:Convolutional networks for biomedical image segmentation [C]//MICCAI. Proceedings of the Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015:18th International Conference,Munich,Germany,October 5-9,2015,Proceedings,Part Ⅲ 18. Singapore:Springer International Publishing,2015:234-241.
[19] VALSTAR M,SCHULLER B,SMITH K,et al. Avec 2013:the continuous audio/visual emotion and depression recognition challenge[C]//ACM. Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge. New York:ACM,2013:3-10.
[20] HUANG Q H,HUANG C Q,WANG X Z,et al. Facial expression recognition with grid-wise attention and visual transformer[J]. Information sciences,2021,580:35-54.
[21] HE L,GUO C G,TIWARI P,et al. DepNet:an automated industrial intelligent system using deep learning for video-based depression analysis[J]. International journal of intelligent systems,2022,37(7):3815-3835.
[22] NIU M,TAO J,LIU B,et al. Automatic depression level detection via lp-norm pooling[C]//INTERSPEECH. Proceedings of Interspeech. Austria:INTERSPEECH,2019:4559-4563.
[23] ZHAO Z,LI Q,CUMMINS N,et al. Hybrid network feature extraction for depression assessment from speech[J]. International journal of intelligent systems,2020,6(3):8-12.
[24] ZHU Y,SHANG Y Y,SHAO Z H,et al. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics[J]. IEEE transactions on affective computing,2018,9(4):578-584.
[25] JAZAERY M A,GUO G D. Video-based depression level analysis by encoding deep spatiotemporal features[J]. IEEE transactions on affective computing,2021,12(1):262-268.
[26] MENG H Y,HUANG D,WANG H,et al. Depression recognition based on dynamic facial and vocal expression features using partial least square regression[C]//ACM. Proceedings of the 3rd ACM International WorksHop on Audio/visual Emotion Challenge. New York:ACM,2013:21-30.
[27] DU Z Y,LI W X,HUANG D,et al. Encoding visual behaviors with attentive temporal convolution for depression prediction[C]//IEEE. Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). Piscataway:IEEE,2019:1-7.
[28] KAYA H,ÇILLI F,SALAH A A. Ensemble CCA for continuous emotion prediction[C]//ACM. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. NewYork:ACM,2014:19-26.
[29] KÄCHELE M,GLODEK M ZHARKOV D,et al. Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression[J]. Recognition applications and methods,2014,1(1):671-678.

作者简介:

赵小明(1964-),男,硕士,教授。研究领域:模式识别,情感计算。
范慧婷(1998-),女,硕士生。研究领域:人工智能。
张石清(1980-),男,博士,教授。研究领域:模式识别,情感计算。