DOI:10.19652/j.cnki.femt.2304691

基于Transformer与增强信息融合的双源情感识别*

闫 超 贾振堂
(上海电力大学电子与信息工程学院 上海 201306)

摘 要:为解决当前多模态情感识别效果不佳的问题,提出了一种基于Transformer与增强信息融合的双源情感识别模型,模型由音视频编码分支网络和双源增强特征融合模块组成。其中,视频编码分支利用MobileViTv2提取每帧视频的空间特征,并通过在Transformer编码器结构中内嵌残差结构,强化各帧短期关联语义信息的提取能力。在音频特征提取部分构建了维度匹配器,避免了潜在异构鸿沟,提高了模型训练的鲁棒性。在音视频特征融合处引入低参数量跨模态注意力机制,从两个角度同时增强特征融合能力。通过对比和消融实验证明了方法在多模态情感识别任务中的有效性。

关键词:情感识别;Transformer;注意力机制;多模态融合
中图分类号:TP391.4 **文献标识码**:A **国家标准学科分类代码**:520.6040

Dual-source emotion recognition based on transformer and enhanced information fusion

Yan Chao Jia Zhentang
(College of Electronics and Information Engineering, Shanghai University of Electric Power, Shanghai 201306, China)

Abstract: In order to solve the problem that the current multi-modal emotion recognition effect is not good, a dual-source emotion recognition model based on Transformer and enhanced information fusion is proposed. The model is composed of audio and video encoding and dual-source enhanced feature fusion modules. Among them, the video coding branch uses MobileViTv2 to extract the spatial features of each frame of video, and embeds the residual structure in the Transformer encoder structure to enhance the ability to extract short-term associated semantic information of each frame. A dimensionality matcher is built in the audio feature extraction part, which avoids potential heterogeneity gaps and improves the robustness of model training. A low-parameter cross-modal attention mechanism is introduced in the fusion of audio and video features to enhance the feature fusion ability from two perspectives. The effectiveness of our method in multimodal emotion recognition tasks is demonstrated by comparison and ablation experiments.

Keywords: emotion recognition; Transformer; attention mechanism; multimodal fusion

0 引 言

情绪已被证明对人类的日常生活至关重要,例如决策、行事动机和人际沟通等[1]。如何进一步提高人机交互体验,使计算机拥有可以实现复杂情感分析计算的功能,成为了近年来研究的热门问题。早期情感识别研究大多基于面部表情[2-3]、声音信号[4]、文本[5]等单一模态信息展开,容易受到噪声干扰,且无法总是准确地揭示人类真实情绪[6]。相比之下,结合声音,面部表情,动作姿态等多模态信息的情感识别系统,能够更充分地利用多源信息,分析各模态对真实情绪的贡献,实现人类情感的精准识别[7-8]。

多模态学习和深度学习技术的日益进步,有效弥补了单模态系统的不足之处,推动了人类情感识别领域的发展,研究者在这一领域已提出众多优秀的多模态情感识别方法。Tzirakis等[9]使用了端到端的卷积神经网络-长短期记忆(CNN-LSTM)结构捕捉音视频序列情感表征。姜明星等[10]提出SLTOP算法用于提取多尺度的时空特征,并使用基于云模型的决策层加权融合方法完成双模态特征的融合。刘天宝等[11]提出了一种嵌入自注意力的

收稿日期:2023-02-07
* **基金项目**:国家自然科学基金(62105196)项目资助

LSTM结构，并使用权重分配的方法决定各模态情感特征的贡献。Praveen等[12]提出一种联合交叉注意力融合方法，通过计算联合特征表征与各模态之间的注意力权重，得到模态内和模态间的关系，实现深层次的信息交互。Tsai等[13]提出多模态Transformer模型MULT，使用成对的Transformer交互各模态信息。Wang等[14]提出Husformer模型，它将所有模态的低层次特征映射到同一纬度空间融合，并与各模态高层次特征建立语义关联，从而实现多模态信息深度聚合。上述方法大多只在特征提取网络或融合算法一方面投入过多关注，忽视了两方面的有效结合，因此前述研究方法要么没有获取到足够的信息，导致模型检测较低，要么没有对两类信息进行有效结合，忽视了单一模态对检测效果的作用。

针对上述研究尚存的不足之处，本文提出了一种基于Transformer与增强信息融合的音视频情感识别模型。在视频特征提取部分，采用低参数量MobileViTv2[15]网络对于每帧人脸进行特征提取，随后将获取到的特征通过结合残差结构的Transformer编码模块实现全局信息的加强融合，并且构建帧与帧之间的信息互联；在音频部分将序列信息通过连续4个一维卷积层进行维度匹配避免出现异构鸿沟，加快了模型的收敛速度。通过内置残差结构的Transformer结构进行增强序列特征提取，并将全时域特征信息进行关联。通过在音视频特征融合结构中内嵌低参数量点乘注意力机制，加强双模态信息的融合能力。为验证本文方法的有效性，在公开的RAVDESS[16]数据上进行消融与对比试验，实验结果表明本文模型相比于其他方法，准确率更高。

1 基于Transformer与增强信息融合的双源情感识别模型

基于Transformer与增强信息融合的双源情感识别模型完整结构如图1所示，由视频编码分支、音频编码分支和特征融合模块3部分组成。其中，视频编码分支负责提取视频时空特征，音频编码分支用于提取音频时序特征，特征融合模块实现视频和音频情感特征的融合，最后通过线性层和Softmax输出预测情绪。

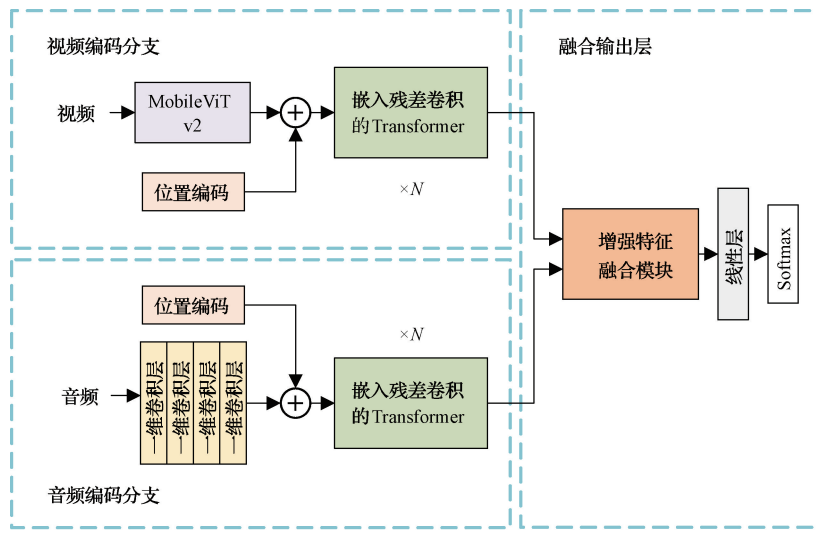

图1 双源情感识别网络模型结构

1.1 音频编码分支

人类能够通过声音中的语气变化表达出丰富的情感信息，但这些变化并不总是容易区分辨别的，为此本文设计了一种音频编码分支网络对音频信号进行处理，以便有效提取人类语音中的情感信息。该音频编码分支使用从音频中提取出的梅尔频率倒谱系数(Mel frequency cepstral coefficients，MFCC)作为输入特征，依次通过4个一维卷积层和改进的Transformer层从音频中提取有意义的情感信息。卷积网络每层由一维卷积运算、批归一化层、Gelu激活函数和最大池化层组成。其中，每个卷积核的大小设置为3，步长为1，填充为1，各层的卷积核数量设置为[64，128，128，C]，C代表嵌入残差卷积的Transformer网络的输入特征维度。逐层的卷积运算增大了输入特征维度，丰富了初步提取的语音特征中的表达样式。

随后为经过卷积网络处理得到的语音特征加入正弦位置编码[17]，然后将其送入内嵌残差卷积结构的Transformer中学习深层次的上下文语义信息。本文在Transformer的多头注意力层和前馈层之间嵌入了一个残差卷积层，用于增强网络局部信息学习能力。如图2所示，嵌入的卷积块使用了卷积核大小为1的卷积层进行特征维度映射；使用卷积核大小为3，步长为1，填充为1的卷积层学习全局信息中的局部关联，并在后续加入批归一化加快收敛；激活函数选用LeakyReLU函数。其中各层卷积核数量设置为[0.25×C，0.25×C，C]，这样可以降低嵌入模块的计算量，并且一维卷积操作能够引入额外的时序信息，补充了Transformer结构中缺乏的位置信息。

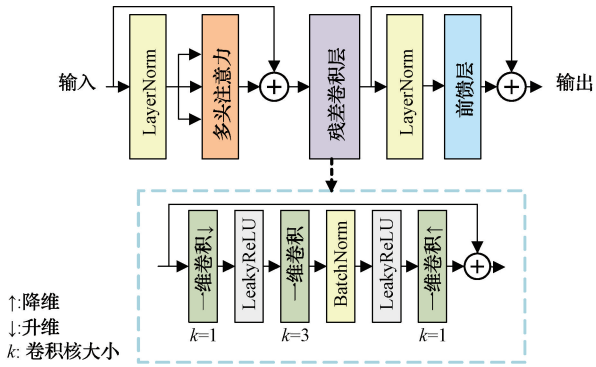

图 2　嵌入残差卷积的 Transformer 示意图

1.2　视频编码分支

视频是由连续静态图像序列组合而成的动态表征，准确地描述情感视频中的空间纹理信息以及时序上的纹理变化信息能够有效地实现人类情感的识别。CNN 网络在图像处理领域已被广泛应用，但随着 Transformer 网络在计算机视觉领域的成功应用，Vision Transformer[17] 已被证明在同样基于大量数据预训练的情况下性能明显优于 CNN 网络。为提高模型的时空信息提取能力，本文设计了一种视频编码分支网络，使用连续的人脸表情帧序列作为输入，引入 MobileViTv2 在空间维度上提取单帧视频中的人脸表情纹理特征，并使用嵌入残差卷积结构的 Transformer 网络在时间维度上学习表情帧序列的动态纹理变化信息，最终实现视频时空特征的提取。

MobileViTv2 网络使用多层 mobilenetv2 模块[15] 进行降采样操作，并经过多层 MobileViTv2 模块学习深层次空间特征。具体来说，对于输入向量 $\boldsymbol{X}_L \in \boldsymbol{R}^{B\times H\times W\times C\times T}$，为匹配空间特征提取器的输入维度，首先将批量维度 B 和时间维度 T 压缩合并得到 $\boldsymbol{X}_L \in \boldsymbol{R}^{F\times H\times W\times C}$，其中 $F = BT$。$\boldsymbol{X}_L$ 经过多层卷积降采样层后，被送入 MobileViTv2 模块。该模块首先利用 3×3 的深度可分离卷积处理输入特征图，并通过 1×1 卷积将输出映射到高维空间 d，随后将卷积输出特征图展开为 N 个不重叠的图像块 (patch) $\boldsymbol{X}_U \in \boldsymbol{R}^{F\times P\times N\times d}$。其中 $P = wh$，$N = HW/P$ 为图像块的数量，$h \leqslant n$，$w \leqslant n$ 为图像块的高和宽。$\boldsymbol{X}_U$ 中各图像块中的每个像素 $p \in \{1,\cdots,P\}$ 通过使用可分离自注意力的多层 Transformer 网络建模，得到 $\boldsymbol{X}_G \in \boldsymbol{R}^{F\times P\times N\times d}$。由于 $\boldsymbol{X}_U$ 已含有局部空间信息，在 Transformer 进行对各图像块编码时，每个像素点均含有邻近像素点编码信息，因此 $\boldsymbol{X}_G$ 中的所有像素都可以对输入 $\boldsymbol{X}_L$ 中的全局信息进行感知，因此，MobileViTv2 块的整体有效感受野为 $H\times W$。随后，对 $\boldsymbol{X}_G$ 进行折叠操作并利用 1×1 卷积将其特征维度还原到 C 维，得到输出 $\boldsymbol{X}_F \in \boldsymbol{R}^{F\times H\times W\times C}$。

最后，在 MobileViTv2 网络末尾部分，经过 1×1 卷积和池化操作，得到视频空间特征输出向量 $\boldsymbol{X}_S \in \boldsymbol{R}^{F\times m}$。为匹配时间维度特征的提取网络的输入尺度，将批量维度和时间维度展开得到视频帧序列向量 $\boldsymbol{X}_f \in \boldsymbol{R}^{B\times h\times T}$，其中 h 是嵌入残差卷积层的 Transformer 的输入隐藏层维度。将加入正弦位置编码的 $\boldsymbol{X}_f$ 输入内嵌残差卷积层的 Transformer 网络进行时序信息建模，捕捉动态纹理信息，最终输出视频时空特征 $\boldsymbol{X}_v \in \boldsymbol{R}^{B\times h\times T}$。

1.3　增强特征融合模块

为增强模型学习跨模态互补信息能力，提升模型特征融合效果，本文提出了基于跨模态注意力机制的增强特征融合模块如图 3 所示。该模块主要由双向信息增强模块，前馈层以及全局平均池化层组成。

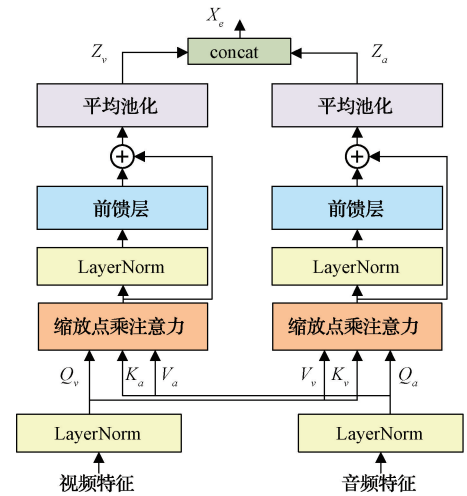

图 3　增强特征融合模块示意图

特征融合模块首先通过由两个缩放点乘注意力模块组成的跨模态双向信息增强模块交互输入音频和视频特征的跨模态信息。具体来说，在每个分支的输出处加入由缩放点乘注意力和前馈层组成的跨模态信息增强模块，建立当前分支模态和另一模态的共享语义空间。以视频编码分支连接的注意力模块为例，将视频编码分支输出特征 $\boldsymbol{X}_v$ 投影以获得查询向量 $\boldsymbol{Q}_v$，而键向量 $\boldsymbol{K}_a$ 和值向量 $\boldsymbol{V}_a$ 则从音频编码分支输出特征 $\boldsymbol{X}_a$ 处投影得到。随后 $\boldsymbol{Q}_v$ 和 $\boldsymbol{K}_a$ 被输入到缩放点积注意力层计算注意力权重，然后在 $\boldsymbol{V}_a$ 和注意权重之间进行矩阵乘法，最后通过位置前馈层得到含有跨模态互补信息的增强向量 $\boldsymbol{Z}_v$，计算公式如式(1)、(2)所示。

$$CA = Softmax\left(\frac{\boldsymbol{Q}_v\boldsymbol{K}_a^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_a \tag{1}$$

$$\boldsymbol{Z}_v = FFn(CA) \tag{2}$$

式中：CA 代表跨模态缩放点乘注意力权重；FFn 代表前馈层；d_k 为超参数。

同理另一分支可得到增强向量 $\boldsymbol{Z}_a$。双向信息增强模块的加入，使得注意力在两个方向上都被应用，有助于更好地捕捉跨模态情感特征。之后，在时间维度上对 $\boldsymbol{Z}_v$ 和 $\boldsymbol{Z}_a$ 进行平均池化，汇集跨时间特征，以获得用于分类的单一特征表示 $\overline{\boldsymbol{Z}}_v$ 和 $\overline{\boldsymbol{Z}}_a$。最后，将 $\overline{\boldsymbol{Z}}_v$ 和 $\overline{\boldsymbol{Z}}_a$ 融合得到音视频

情感向量 $\boldsymbol{X}_e$：

$$\boldsymbol{X}_e = concat(\overline{\boldsymbol{Z}}_v, \overline{\boldsymbol{Z}}_a) \tag{3}$$

式中：$concat$ 表示特征维度的向量拼接操作。

2 实验结果与分析

2.1 实验配置及数据集介绍

本文使用RAVDESS数据集进行实验，它是一个将情绪状态分为8类的公开数据集，包括平常、沉静、快乐、悲伤、愤怒、恐惧、惊讶和厌恶。该数据集共有1 440个有声视频，每种情绪的数据量均衡，均含有强烈和普通两种强度的表现。其中，音频采样率为48 000 Hz，视频清晰度为720 P，帧率为30 fps。本文选取其中1 200视频作为训练集，剩余240个视频作为验证集。

在Nvidia 4090，24 GB RAM的硬件设备上进行实验，使用Pytorch1.9搭建实验环境。对于实验数据，将所有视频裁剪至3.6 s的长度，随后从每个视频中均匀抽取16帧画面，之后采用MTCNN人脸检测算法[18]定位每帧图像中的人脸部分，并将图像裁剪至224×224尺寸。使用Cutout，随机水平翻转和随机左右旋转方法进行图像增强。对于分离出的音频数据，将其重采样至22 500 Hz频率，并提取10维MFCC作为音频输入。

MobileViTv2使用在Imagenet上预训练的权重初始化，内嵌残差卷积结构的Transformer层数均设置为4，多头注意力均设置为2头，所有查询、键和值向量的嵌入维度设置为64。

实验训练轮次固定为105轮，批量大小为4，学习率通过5轮预热(warm-up)到达0.000 2，随后使用余弦退火策略动态调整。优化器使用Adamw，权重衰退设置为0.000 1，损失函数使用带有标签平滑化[19]的交叉熵损失函数，超参数ε设置为0.1。

为评估模型效果，本文使用如下指标：准确率(accuracy，Acc)、精确率(precision，P)、召回率(recall，R)、模型的参数量(parameters，Params)以及模型接收的图像帧序列长度(frames，F)，计算公式如下：

$$Acc = \frac{TP + TN}{A + N} \tag{4}$$

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

式中：A 为正例样本总数；N 为负例样本总数。TP 为预测正确的正例样本数；FP 为预测错误的正例样本数；TN 为预测正确的负例样本数；FN 为预测错误的负例样本数。准确率、精确率以及召回率，用于验证模型的综合性能；模型参数量用于衡量模型的轻量化与否；模型接收的图像帧序列长度用于评估模型处理视频时序信息的能力。

2.2 音视频情感识别实验

图4所示为使用本文方法进行情感识别实验得到的混淆矩阵，矩阵左侧是各情绪类别的真实标签，下方是预测标签，主对角线上的值表示每个情绪类别的召回率，用于比较模型检测各类情绪的能力。从图4可以看出，绝大多数情绪都能够实现高精度检测，其中快乐和厌恶最容易检测出来。相比之下，恐惧情绪的检测能力较差，且容易被错误预测为悲伤情绪。

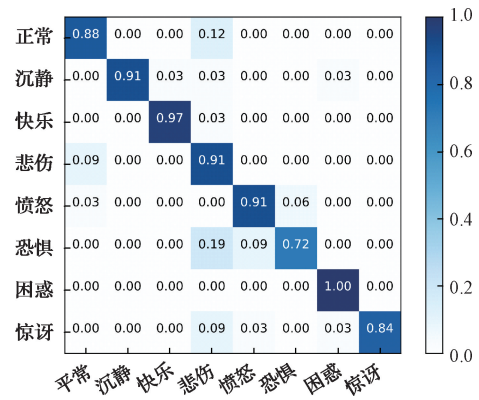

图4 音视频情感识别混淆矩阵

为探究这一现象是所提模型缺陷还是部分情绪识别难度较大的原因，与现有模型算法进行了各类别情绪识别率对比如表1所示。1)CNN14-biLSTM算法召回率在平常和悲伤类别表现最差，且悲伤类别的精确率表现远低于其他类别，此外，文献[20]指出了悲伤和恐惧情绪容易混淆的问题，并在时间维度上绘制后验概率图展示了误分类样本；2)Middya等[22]的算法平常类别的召回率最低，且悲伤和愤怒情绪的精确率偏低，原因在于模型仅使用单层LSTM处理图像帧序列时序信息，难以有效区分不同类别之间的动态特征差异；3) Luna-Jiménez等[23]同样发现了悲伤与恐惧情绪容易混淆的问题。综上可知，现有算法普遍存在悲伤情绪检测效果较差，且易与其他类别混淆，主要原因在于，悲伤与恐惧情绪在面部动态表现或是声调语气上存在相似之处，容易混淆。

为衡量所提模型性能，与现有方法进行比较，取多次试验的平均值作为最终结果，表2为本文模型与其他模型方法的性能对比结果。

分析表2可知，本文模型相较Luna-Jiménez等[23]的方法准确率提高了1.53%，普遍优于其他方法。主要原因如下：1)Luna-Jiménez等[23]的方法缺乏对于视频空间特征的深度挖掘，且过于重视音频特征的提取，基于WAV2VEC2.0模型的音频特征提取网络参数量过大，运算速度极慢；2)MULT方法通过成对的Transformer网络学习模态间的互补信息，但是网络深度较浅，对各模态自身特征提取不足，缺少模态间的差异信息；3)Chumachenko等[21]的方法忽视了视频时态特征的重要性，仅使用一维卷积构建的浅层时序特征提取网络，难以实现帧间动态纹理信息的有效捕捉以及音频时序特征的有效提取；

表 1 多模型各类情绪识别率结果对比 (%)

模型		平常	沉静	快乐	悲伤	愤怒	恐惧	困惑	惊讶
CNN14-biLSTM[20]	*P*	84.62	80.49	91.18	67.65	94.29	83.33	89.47	88.57
	R	68.75	91.67	86.11	65.71	94.29	83.33	91.89	88.57
Middya 等[22]	*P*	92.00	88.00	86.00	77.00	78.00	83.00	95.00	87.00
	R	73.00	100.00	100.00	82.00	82.00	80.00	88.00	80.00
Luna-Jiménez 等[23]	*P*	88.31	91.45	89.79	71.22	91.78	88.26	93.99	89.50
	R	82.25	85.63	89.25	81.88	92.37	83.62	88.50	87.87
本文	*P*	77.78	100.00	96.88	69.05	87.88	92.00	94.12	100.0
	R	87.50	90.62	96.88	90.62	90.62	71.88	100.0	84.38

表 2 音视频情感识别模型性能对比

模型	Acc/%	*P*/%	*R*/%	Params/($\times10^6$)	*F*/fps
MULT[13]	78.50	—	—	1.18	15
CNN14-BiLSTM[20]	80.08	84.95	83.79	85.7	—
Chumachenko 等[21]	81.58	—	—	2.05	15
Middya 等[22]	86.00	86.00	86.00	6.03	6
Luna-Jiménez 等[23]	86.70	88.04	86.42	319.48	—
本文	88.33	89.71	88.87	4.49	16

4)Middya 等[22]堆积多层卷积网络来放大模态间差异信息,但是短帧图像序列包含时序信息较少,且简单地将各模态特征展为一维向量并融合的方法不能实现跨模态特征的有效融合;5)CNN14-BiLSTM 方法使用支持向量机(SVM)算法作为模型输出分类器,难以得到高精度输出;6)本文方法设计的模型在保持低参数量的情况下,允许输入包含更多信息量的长图像帧序列,并同时注重音频和视频特征的提取,最后通过增强特征融合模块充分交互模态间信息,实现了模型识别精度的大幅提升。

2.3 特征融合方法对比实验

为证明本文增强特征融合方法的优越性,与现有的拼接、自适应权重分配、注意力机制融合方法进行比较。其中注意力融合方法在括号中进行模态标注,例如(A,V)代表使用音频模态映射得到查询向量 $\boldsymbol{Q}$,使用视频模态映射得到键向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$。

表 3 为不同融合方法结果对比,从表 3 可以看出,现有融合方法结果差距较小,其中基于注意力机制的融合方法分着占据最低和次高的准确率,说明了注意力机制融合方法根据查询,键和值向量的映射模态选择不同,结果会出现较大差异;自适应权重分配方法通过为各模态特征赋予不同的权重,得到融合向量,但是单纯的比例划分并不能完整描述各模态信息的贡献,限制了模型召回率表现;拼接方法直接将两个模态的输出特征全部融合到一起,虽然蕴含的信息丰富,但是无法充分利用,限制了模型的识别能力上限;本文的融合方法同时使用两个注意力模块组成双向通道,能够更加充分地学习跨模态关联信息,提升音视频特征融合效果,结果表现本文方法较其他融合方法,在准确率、精确率与召回率方面均有较大幅度提升。

表 3 特征融合方法性能对比 (%)

融合方法	Acc	*P*	*R*
拼接	82.13	83.66	82.17
自适应权重分配	81.67	82.63	80.11
注意力融合(V,A)	81.35	82.02	80.86
注意力融合(A,V)	82.50	84.58	83.21
增强特征融合法	88.33	89.71	88.87

2.4 消融实验

为证明在 Transformer 中的内嵌残差卷积结构和所提双向信息增强模块的有效性,在此进行消融实验。其中 AV 代表音视频双模态输入,A 和 V 分别代表了音频和视频单模态输入,由于单模态实验没有另一模态信息,因此音频和视频的单模态实验只进行残卷积差结构的消融实验。

表 4 为各模块对实验结果的影响,从表 4 可知,在音频和视频编码网络中,针对 Transformer 网络嵌入残差卷积结构后,音频和视频单模态情感识别结果在精确率方面提升明显,增长 2%左右,表明这一结构能够提升音视频编码分支提取特征的能力,提高模型识别准确率。同时,加入了双向信息增强模块的融合网络,能够利用跨模态互补信息增强类内特征表达,并结合不同模态间的差异信息增加类间区分度,提升模型识别能力、准确率、精确率和召回率的大幅度提升,验证了这一模块可以有效增强跨模态特征融合效果。

表4 消融实验结果 (%)

Module		Acc			*P*			*R*		
双向信息增强模块	残差卷积结构	AV	A	V	AV	A	V	AV	A	V
×	×	80.83	70.42	79.58	82.27	70.68	79.92	80.36	70.71	80.01
×	√	82.08	71.25	80.01	83.47	72.56	82.29	81.99	72.14	80.08
√	×	85.83	—	—	86.36	—	—	85.97	—	—
√	√	88.33	—	—	89.71	—	—	88.87	—	—

3 结 论

针对现行音视频双模态情感检测方法多局限于单模态性能或者特征融合结构,忽视了全局信息融合导致网络检测性能较差等不足,本文提出基于 Transformer 与增强信息融合的双源情感识别模型,通过结合残差结构对 Transformer 结构进行改进,同时增强音视频特征信息的提取能力以及特征融合处对局部有效信息的关注度,实现了较为准确的检测。实验表明本文模型对于音视频情感的识别检测准确率可达 88.33%,性能优于其他方法。后续将研究具有更多模态的情感识别,同时将针对 MobileViTv2 网络在处理长序列图像帧输入时存在运算量较大的问题进行轻量化改进。

参 考 文 献

[1] EYBEN F, WLLMER M, SCHULLER B. Opensmile: The munich versatile and fast open-source audio feature extractor[C]. 18th ACM International Conference on Multimedia(MM′10), 2010: 1459-1462.

[2] 焦亚萌,周成智,李文萍,等. 融合多头注意力的 VGGNet 语音情感识别研究[J]. 国外电子测量技术, 2022,41(1):63-69.

[3] 李翔,李昕,胡晨,等. 面向智能机器人的 Teager 语音情感交互系统设计与实现[J]. 仪器仪表学报,2013, 34(8):1826-1833.

[4] 牛慧,赵艳东. 基于改进 Gabor 小波变换的人脸情感识别[J]. 电子测量技术,2020,43(5):124-129.

[5] 郑啸,王义真,袁志祥,等. 基于卷积记忆神经网络的微博短文本情感分析[J]. 电子测量与仪器学报, 2018,32(3):195-200.

[6] OH G, RYU J, JEONG E, et al. Drer: Deep learning-based driver′s real emotion recognizer[J]. Sensors, 2021, 21(6): 2166.

[7] PORIA S, CAMBRIA E, BAJPAI R, et al. A review of affective computing: From unimodal analysis to multimodal fusion[J]. Information Fusion, 2017, 37: 98-125.

[8] 何俊,刘跃,何忠文. 多模态情感识别研究进展[J]. 计算机应用研究,2018(11):3201-3205.

[9] TZIRAKIS P, TRIGEORGIS G, NICOLAOU M A, et al. End-to-end multimodal emotion recognition using deep neural networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1301-1309.

[10] 姜明星,胡敏,王晓华,等. 视频序列中表情和姿态的双模态情感识别[J]. 激光与光电子学进展,2018, 55(7):071004.

[11] 刘天宝,张凌涛,于文涛,等. 基于嵌入注意力机制层级 LSTM 的音视频情感识别[J]. 激光与光电子学进展, 2021, 58(2):0210017.

[12] PRAVEEN R G, GRANGER E, CARDINAL P. Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention[J]. IEEE Transactions on Biometrics Behavior and Identity Science, 2022, DOI: 10.1109/TBIOM.2022.3233083.

[13] TSAI Y H H, BAI S, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]. Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL), 2019: 6558.

[14] WANG R, JO W, ZHAO D, et al. Husformer: A multi-modal transformer for multi-modal human state recognition[J]. Human-Computer Interaction V, 2022, DOI:10.48550/arXiv.2209.15182.

[15] MEHTA S, RASTEGARI M. Separable self-attention for mobile vision transformers[C]. Computer Vision and Pattern Recognition, 2022.

[16] LIVINGSTONE S R, RUSSO F A. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English[J]. PloS One, 2018, 13(5): e0196391.

[17] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]. Computer Vision and Pattern Recognition, 2020.

[18] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters,

2016，23(10)：1499-1503.

[19] MULLER R，KORNBLITH S，HINTON G. When does label smoothing help? [J]. Machine Learning，2019，arXiv：1906.02629.

[20] LUNA-JIMENEZ C，GRIOL D，CALLEJAS Z，et al. Multimodal emotion recognition on ravdess dataset using transfer learning[J]. Sensors，2021，21(22)：7665.

[21] CHUMACHENKO K，IOSIFIDIS A，GABBOUJ M. Self-attention fusion for audiovisual emotion recogni-tion with incomplete data[C]. Computer Vision and Pattern Recognition，2022.

[22] MIDDYA A I，NAG B，ROY S. Deep learning based multimodal emotion recognition using model-level fusion of audio-visual modalities [J]. Knowledge-Based Systems，2022，244：108580.

[23] LUNA-JIMENEZ C，KLEINLEIN R，GRIOL D，et al. A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset[J]. Applied Sciences，2021，12(1)：327.

作 者 简 介

闫超，硕士研究生，主要研究方向为深度学习图像处理。

E-mail：894553281@qq.com

贾振堂(通信作者)，博士，副教授，主要研究方向为智能视频监控、配电网故障定位。

E-mail：462458081@qq.com