基于改进ConvMixer和动态焦点损失的视听情感识别

师 硕,覃嘉俊,于 洋*,郝小可
(河北工业大学人工智能与数据科学学院,天津 300401)

摘 要: 视听双模态情感识别是情感计算领域的研究热点.目前情感识别方法存在无法同时提取视频局部和全局特征,多模态数据融合简单,损失函数在模型优化中无法关注错分样本等问题,导致情感识别结果精确度不高.本文提出一种基于改进的ConvMixer和动态权重焦点损失函数的视听情感识别方法.采用空间和时间邻接矩阵代替ConvMixer中的深度分离卷积,提取视频时域空域上的全局和局部特征.提出跨模态时间注意力模块,以对称结构捕捉模态间的时间相关性,提高特征融合效果.结合混淆矩阵计算具有动态权重的焦点损失函数,差异化地加大错分样本在损失中的占比,优化模型参数.在公开数据集上的实验结果表明,本文方法能提取到代表性特征,可有效优化网络结构,提高了情感识别的准确率.

关键词: 情感识别;ConvMixer;注意力机制;多模态特征融合;焦点损失函数
基金项目: 国家自然科学基金(No.61806071,No.62102129);河北省自然科学基金(No.F2020202025,No.F2021202030)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2024)08-2824-12
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221042

Improved ConvMixer and Focal Loss with Dynamic Weight for Audio-Visual Emotion Recognition

SHI Shuo, QIN Jia-jun, YU Yang*, HAO Xiao-ke
(*School of Artificial Intelligence*, *Hebei University of Technology*, *Tianjin* 300401, *China*)

Abstract: Audio-visual bimodal emotion recognition is a research hotspot in the field of emotion computing. At present, emotion recognition methods cannot simultaneously extract local and global features of video, multi-modal data fusion is simple, loss function can not pay attention to misclassification of samples in model optimization, resulting in low accuracy of emotion recognition results. In this paper, an audio-visual emotion recognition method based on improved ConvMixer and focus loss function with dynamic weight is proposed. Spatial and temporal adjacent matrices were used instead of deep separation convolution in ConvMixer to extract global and local features in video spatial and temporal domain. A cross-modal temporal attention module is proposed to capture the temporal correlation between modals with a symmetrical structure to improve the feature fusion effect. The focus loss function with dynamic weight was calculated by the confusion matrix, and the proportion of error samples in the loss was increased differentially to optimize the model parameters. Experimental results on public data sets show that the proposed method can extract representative features, optimize the network structure effectively, and improve the accuracy of emotion recognition.

Key words: emotion recognition; ConvMixer; attention mechanism; multi-modal feature fusion; focal loss function
Foundation Item(s): National Natural Science Foundation of China (No.61806071, No.62102129); Natural Science Foundation of Hebei Province (No.F2020202025, No.F2021202030)

收稿日期:2022-09-09;修回日期:2023-05-05;责任编辑:王天慧
*通讯作者:于洋

1　引言

情感计算,又称情感识别、情感分类,即针对人类的外在表现进行测量和分析,并能对情感施加影响的计算[1]. 情感计算开辟了计算机科学的新领域,其目标是使计算机拥有情感,能够像人类一样识别和表达情感,从而使人机交互更加自然. 目前已被广泛应用在医疗看护、行车安全、商业服务等领域[2].

情感计算研究中,相较于生理信号,音视频模态的情感数据具有开源、易收集、非接触式采集的优点,在情感识别领域被广泛使用. 目前,音视频特征常用卷积神经网络(Convolutional Neural Networks,CNN)中的残差网络(Residual Networks,ResNets)和循环神经网络(Recurrent Neural Networks,RNN)中的长短时记忆网络(Long Short Term Memory Networks,LSTM)进行提取. CNN能很好地提取数据的空间特征信息,LSTM以分段形式在提取数据的时间信息时表现优越. 但CNN和LSTM分别以逐像素和逐段的方式提取特征,在小范围空间和短时间维度内能直接提取特征,但不能进行大跨度的时空特征聚合,因而无法提取到全局特征. 随着图像分辨率提高和视频数据增长,采用CNN和LSTM的计算量和时间开销也呈线性增长.

音视频双模态数据的异构性使得两种模态数据融合也成为音视频双模态情感识别的关键环节. 目前,常用的融合方式分为特征融合和决策融合,具体融合策略分别是级联拼接[3]和加权[4]. 级联拼接通常将多模态特征直接拼接来增加特征维度,忽略了模态间的时间相关性,无法将多个模态的特征有机结合. 加权则需要大量实验,才能获得最优的超参数权重. 同时,多模态数据的差异,对模型泛化能力提出更高要求,而损失函数的设计对网络模型优化具有重要作用. 在情感计算领域常采用交叉熵损失(Cross Entropy Loss,CEL)函数[5]来衡量预测分布与真实分布之间的差距,但在模型训练后期,交叉熵损失对同一个batch的样本,既不能抑制分类正确样本,也不能提高错误分类样本在损失中的占比,导致模型参数优化欠佳.

基于音视频多模态数据的情感识别已引起研究者的广泛关注,并在人机交互方面具有广阔的应用前景,但依然是一项具有挑战性的任务,存在如下问题:(1)特征提取方面,现有的CNN或LSTM视频特征提取网络,无法同时提取视频时空域上的局部和全局特征,特征的情感表征能力不足;(2)特征融合方面,目前大多数方法只是将音视频两种特征进行简单的拼接或加权融合,这些方式过于通用化,难以体现不同模态数据的相关性和互补性;(3)现有方法对表情中难分样本的识别较差,且正确分类和错误分类样本无差别地作用于损失函数,导致情感识别精确度不高.

针对上述问题,本文提出一种基于改进的ConvMixer和动态权重焦点损失函数的视听情感识别方法. 主要贡献如下:

(1)提出结合邻接矩阵的ConvMixer(Adjacent Matrix-based ConvMixer,AMCM)视频特征提取网络. 用图神经网络(Graph Neural Networks,GNN)[6]的时间和空间邻接矩阵代替ConvMixer网络的深度分离卷积,将视频帧序列在空间和时间维度上分别划分patch,通过邻接矩阵权重自适应聚合所有patch特征,使模型能同时提取空间和时间维度上的局部和全局特征.

(2)设计两个呈对称结构的跨模态时间注意力模块(Cross-Modal Temporal Attention Module,CMTAM)进行特征融合. 通过两次交叉的时间相关性注意力计算,使两个模态相互提取时间特征并融合,丰富了各自模态的时间信息.

(3)提出动态权重焦点损失(Focal Loss with Dynamic Weight,FLDW)函数. 在训练过程中,以上个训练周期的混淆矩阵为依据,动态生成当前训练周期的焦点损失权重,提高模型对错分样本的关注度,根据不同错分情况计算得到差异化的损失权重,优化了模型的参数设置.

2　相关工作

2.1　特征提取方法

情感识别早期研究阶段,手工特征和机器学习相结合是主流的视频特征提取方法[7~9],但良好的手工特征需要通过大量实验选取得到. 随着数据集规模的不断增大,机器学习方法显现出数据处理能力的局限性. 目前,越来越多的研究学者采用深度学习方法进行视频特征提取. CNN和LSTM及其改进网络常用来分别提取视频数据的空间和时间特征,或二者结合共同提取时空特征[10]. 还有引入注意力机制(attention mechanism)来区分特征间的有效性差异,以此来突出特征的关键部分. Du等人[11]提出一个由时域卷积层堆叠得到的时域沙漏状卷积编码解码器,通过整合低层次的编码特征和高层次的解码特征,提取时间上下文关系. Liu等人[12]通过堆叠LSTM层和图卷积层提取视频帧的时间注意力特征,在CK+等数据集上得到最优分类结果. Zhao等人[13]把3D ResNet网络作为骨干网络,提出了一种由卷积层、全连接层和Softmax层组成的三层式注意力,并把三层式注意力堆叠,通过在每一层注意力层前转置特征矩阵,实现同种结构提取空间维度、时间维度和通道维度注意力特征的目的. Chen等人[14]结合方向不变的方向梯度直方图,提出基于全局极卷积(full polar convolution)和局部极卷积(local polar convolution)的深度特征,组成方向不变性特征,解决了人脸

偏转角度对情感识别精度的影响. Zhang 等人[15]基于自注意力机制(self-attention mechanism)提出一种跨模态时间注意力,通过独立的编码器获取音频和视频模态对应的 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 向量,再进行跨模态向量拼接实现跨模态自注意力计算. 文献[16]提出空间变换器(spatial transformer),经过预训练和微调,有效提取原始图像中的关键区域,达到快速拟合和提高分类准确率的目的.

对于音频数据,文献[17]引入 Wav2vec 模型替代文献[16]中使用的 CNN,从原始音频信号中直接提取特征,取得了更好的分类准确率. Tzirakis 等人[18]直接使用具有两层卷积层和两层最大池化层的 CNN 对音频原始一维信号进行特征提取. Hossain 等人[19]则将音频声谱图输入到 2D CNN 网络提取频域和时间信息. Wang 等人[20]从原始音频信号提取两个梅尔倒谱系数(Mel-scale Frequency Cepstral Coefficients, MFCC)谱图,再使用两个独立的 CNN 提取特征. Meng 等人[21]使用 MFCC 静态图、一阶差分图和二阶差分图作为 CNN 网络输入提取空间特征,得到的空间特征再输入到双向长短时记忆网络(Bidirectional Long Short-Term Memory, BiLSTM)提取时间特征. Song 等人[22]将音频分段,使用门控循环单元(Gate Recurrent Unit, GRU)提取到了很好的时间特征,但当数据样本持续时间较长时,GRU 串行计算的特性会导致计算时间成本明显增加.

Transformer 在自然语言处理领域取得了巨大成功,随后的 Vision Transformer (ViT)[23]模型开创了 Transformer 在计算机视觉领域的应用. 尤其对大规模训练集,ViT 取得了比 CNN 更优的性能. 由于 Transformer 中自注意力层的计算复杂度是 patch 数量的平方,在提取大分辨率图片特征时计算量剧增,随后,ViT 模型的各种变体,如 Video Vision Transformer、Swin Transformer 和 Video Swin Transformer 等相继被提出. 2022 年,ViT 团队使用参数更少的深度分离卷积和逐点卷积对 ViT 自注意力层进行改进,提出了结构更为简单的 ConvMixer[24]模型. 在 ImgNet 数据集上,ConvMixer 仅用 14.6×10^6 参数量实现了 77.7% 的准确率,与参数量为 86×10^6 的 ViT 模型得到的 77.9% 的准确率相当接近,超过了参数量为 25.6×10^6 的 ResNet-50 模型实现的 76.32% 的准确率. ConvMixer 模型在不增加计算量的基础上实现了模型分类性能的提升.

2.2 视听多模态融合与损失函数

视听双模态数据融合主要分为决策融合和特征融合. 决策融合通过超参数权重加权得到[25,26],特征融合则通过级联拼接[27],将所有特征直接拼接成一个高维特征向量. Simonyan 等人[28]提出空间流和时间流双流卷积神经网络并独立训练,在分类器前将两个网络通道的特征进行拼接融合. Birhala 等人[29]使用两个 CNN 分别对视频片段对应的静态帧和音频声谱图提取特征,然后将两个模态的特征进行拼接. Farhoudi 等人[30]采用特征拼接将音视频特征进行融合,将大脑情感学习(Brain Emotional Learning, BEL)作为融合与分类阶段的模块,学习模态之间的时间相关性. Liu 等人[31]和 Nie 等人[32]在模态融合阶段,分别改进和直接使用图卷积网络,利用图卷积网络聚合节点特征的特性,学习模态之间的相关性,增强了模态融合的效果. 但无论是加权求和还是拼接融合,都是将视频和音频视为独立存在的模态,均忽略了视频帧序列与音频属于同一视频样本而具有的时间相关性.

离散型情感识别通常使用传统交叉熵损失 CEL 及其改进作为模型优化的目标函数. 但交叉熵损失在模型训练过程中不能抑制正确分类样本的损失值,更无法关注被错误分类的样本,限制了模型泛化能力的提高;并且交叉熵损失的权重参数在训练前设置为固定值,不能根据模型的实际训练情况进行改变,影响了模型性能的提升. 为了克服交叉熵损失的上述不足,Zhao 等人[13]在传统交叉熵的基础上添加极性一致权重,当预测类别与真实类别极性不一致时,通过损失值与极性一致权重相乘使损失值增大,以此增强分类器的分类能力. Ghaleb 等人[33]和 Ma 等人[34]将 CNN 作为特征提取网络,均在分类损失基础上引入额外的相关性损失,拉近两个单模态特征之间的距离,增强了融合后特征分类效果. Zhong 等人[35]和 Zhu 等人[36]使用焦点损失(Focal Loss, FL)替代传统交叉熵作为损失函数,不仅避免了数据不均衡带来的性能损失,还使得模型更关注错误分类的样本. 李锵等人[37]在 FL 基础上,提出了加权焦点损失(Weighted Focal Loss, WFL)函数,通过加入权重调整因子,使网络依据每个类别的分类准确率合理分配权重,增大了难分样本的损失占比来提高其分类准确度. Bai 等人[38]提出了校准焦点损失(Calibrated Focal Loss, CFL),通过增加一个校准项,迫使网络输出正确分类预测概率分布,同时使概率分布具有更低的信息熵.

3 提出方法

本文提出的基于改进 ConvMixer 和动态权重焦点损失函数的视听情感识别方法,整体网络模型由 3 部分组成:结合邻接矩阵的 ConvMixer(AMCM)网络和残差网络(ResNet34)组成的双流视听模态特征提取网络;对称结构的跨模态时间注意力(CMTAM)的特征融合模块;用于优化模型参数的动态权重焦点损失(FLDW)函数. 其网络结构如图 1 所示.

首先,设计双流分支网络提取音视频特征,其中 AMCM 通过 Patch Embedding 层中的卷积层将视频帧序列划分为若干个 patch,提取每个 patch 的浅层特征,

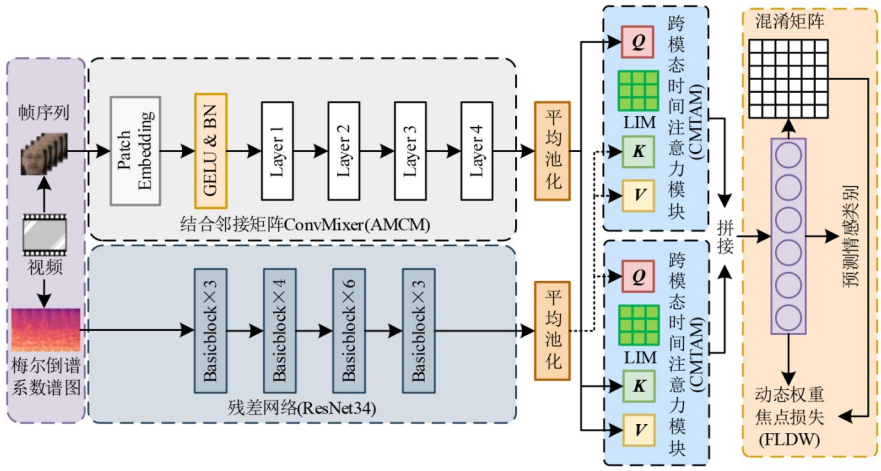

图1　本文方法网络结构

再输入到4层Layer中提取深层特征,和邻接矩阵结合提取视频帧序列特征;ResNet34网络提取音频的MFCC特征. 然后,构建两个对称结构的CMTAM进行特征融合,在CMTAM模块中构建可学习中间矩阵(Learnable Intermediate Matrix,LIM)将分属两种模态的$\boldsymbol{Q}$、$\boldsymbol{K}$特征向量融合,增强模态间的时间相关性. 最后,设计FLDW损失函数进行损失计算和反向传播,以上个训练周期生成的混淆矩阵为依据,针对每一个训练样本的分类情况,动态生成具有差异性的损失权重进行网络优化,最终得到精确的分类结果.

3.1　基于改进ConvMixer的特征提取

3.1.1　ConvMixer

ConvMixer保留了ViT模型中Patch Embedding层,使用深度卷积和逐点卷积替代ViT模型中复杂的注意力层,以更少的参数、更简单的结构实现接近ViT模型的性能,ConvMixer结构如图2所示.

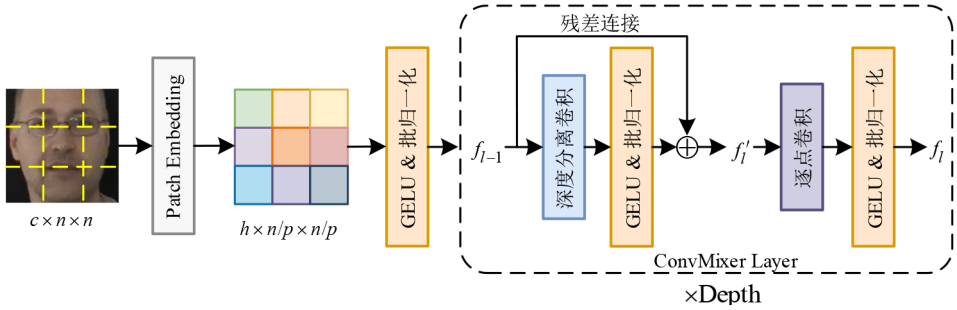

图2　ConvMixer结构

Patch Embedding层本质是一个卷积核大小与卷积核步长大小一致的卷积层,每一次卷积操作的像素区域不存在重叠,实现了将图片分块并进行特征嵌入. ConvMixer的Layer层使用深度分离卷积与逐点卷积替换ViT模型中复杂的注意力层,并且层中的特征通道数不会改变,固定为h. 深度分离卷积负责空间信息的建模,通过组数等于通道数h的分组卷积层实现;逐点卷积负责通道信息的建模,通过卷积核大小为1的卷积层实现. 深度分离卷积、逐点卷积操作分别如式(1)和式(2)所示:

$$f_l' = \mathrm{BN}\{\mathrm{GELU}[\mathrm{ConvDepthwise}(f_{l-1},\ \mathrm{kernel\_size}=k,\ \mathrm{group}=h)]\}+f_{l-1} \tag{1}$$

$$f_l = \mathrm{BN}\{\mathrm{GELU}[\mathrm{ConvPointwise}(f_l')]\} \tag{2}$$

其中,f_l'和f_l分别表示第l个Layer层的中间特征和输出特征,f_{l-1}表示第$(l-1)$个Layer层输出的特征. ConvDepthwise表示深度分离卷积,BN为批归一化层,GELU为激活函数,ConvPointwise表示逐点卷积.

在特征通道数h、卷积核尺寸k相同情况时,普通卷积参数量为$h\times k\times k\times h$,而深度分离卷积通过对特征按通道数分组,参数量仅为$k\times k\times h$. 由此可知,当卷积核尺寸相同时,深度分离卷积所用参数更少;当与普通卷积层参数相同时,可实现卷积核更大尺寸的卷积操作. 实验数据也表明ConvMixer[24]中深度分离卷积的卷积核尺寸越大,感受野越大,模型性能越好. 但当卷积核大小扩大到与输入特征图大小相同时,即使采用深度分离卷积,卷积的计算时间和空间成本也会明显增加.

3.1.2 结合邻接矩阵的ConvMixer

本文提出的AMCM保留了ConvMixer的网络结构和Patch Embedding层，在ConvMixer原有的Layer层采用GNN的邻接矩阵替换深度分离卷积. 邻接矩阵表示节点间的连接关系，与特征矩阵相乘的聚合操作，可实现近似于全局卷积的效果，提高了模型的感受力，使模型可提取到全局特征. 而且，AMCM在训练过程中，按有向图训练邻接矩阵的权重，减少的权重在聚合操作中实现了提取局部特征的效果.

AMCM在每个Layer层，将特征与对应的空间位置编码(Spatial Position Code，SPC)和时间位置编码(Temporal Position Code，TPC)进行级联拼接，保留了各个patch间的空间和时间信息. 再分别引入空间邻接矩阵(Spatial Adjacent Matrix，SAM)和时间邻接矩阵(Temporal Adjacent Matrix，TAM)代替原有ConvMixer Layer中使用的通道分离卷积层，进行先空间后时间的异步特征提取. AMCM的网络结构如图3所示.

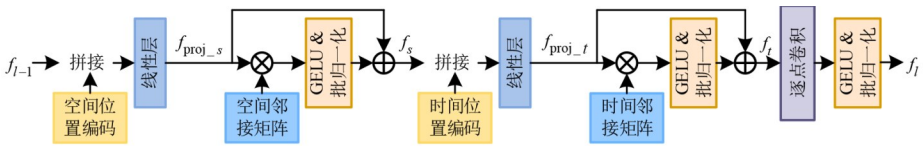

图3 AMCM网络结构

首先，构建位置编码和邻接矩阵. 根据输入Layer层的$f_{embedding}$的时间和空间维度，生成和建立对应大小的空间和时间坐标矩阵$\boldsymbol{S}\in\mathbb{R}^{2\times n\times n}$、$\boldsymbol{T}\in\mathbb{R}^{1\times t}$，其中$S_{ij}=[i,j]$，$\boldsymbol{S}$、$\boldsymbol{T}$如式(3)、式(4)所示：

$$\boldsymbol{S}=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{bmatrix} \tag{3}$$

$$\boldsymbol{T}=[0,1,\cdots,t] \tag{4}$$

空间坐标矩阵$\boldsymbol{S}$经过复制、拼接生成空间位置编码SPC，时间坐标矩阵$\boldsymbol{T}$经过复制、拼接生成时间位置编码TPC. 再根据Patch Embedding层对视频数据沿空间和时间维度划分patch个数的平方，设置对应维度的空间邻接矩阵SAM和时间邻接矩阵TAM. 与GNN一般训练过程中邻接矩阵权重固定不变不同，AMCM中的时间和空间邻接矩阵作为可训练的内部参数，统一初始化后按带权有向图不加任何限制地进行训练，以保留任意一对节点间可能不同的双向连接权重. 本文权重统一初始化为1时，训练完成后权重分布在0.8~1.3之间，呈现出差异性，再通过加权聚合操作得到的特征也是差异化的，提高了特征的多样性.

接下来，在Layer层进行特征提取. 首先，特征$f_{embedding}$与$\mathrm{SPC}\in\mathbb{R}^{2\times t\times n\times n}$进行特征通道维度的拼接，输入线性层Linear，得到空间映射特征$f_{proj_s}\in\mathbb{R}^{c\times t\times n\times n}$，如式(5)所示：

$$f_{proj\_s}=\mathrm{Linear}[\mathrm{Concat}(f_{embedding},\mathrm{SPC})] \tag{5}$$

再将f_{proj_s}从二维空间维度$n\times n$展平成一维n^2，得到特征$f\in\mathbb{R}^{c\times t\times n^2}$，输入到具有跳跃连接结构的空间特征提取模块，再与空间邻接矩阵SAM进行矩阵乘法操作，得到具有空间信息的特征$f_s\in\mathbb{R}^{c\times t\times n^2}$，如式(6)所示：

$$f_s=\mathrm{BN}[\mathrm{GELU}(f\times\mathrm{SAM})]+f \tag{6}$$

其中，$\mathrm{SAM}\in\mathbb{R}^{n^2\times n^2}$.

然后，特征f_s与$\mathrm{TPC}\in\mathbb{R}^{1\times t\times n^2}$进行特征通道维度的拼接，再输入到线性层Linear，得到时间映射特征$f_{proj_t}\in\mathbb{R}^{1\times t\times n^2}$，如式(7)所示：

$$f_{proj\_t}=\mathrm{Linear}[\mathrm{Concat}(f_s,\mathrm{TPC})] \tag{7}$$

再对f_{proj_t}进行特征维度变换，恢复空间维度，通过转置操作将时间维度t变化为特征的最后一个维度，得到特征$f'\in\mathbb{R}^{c\times n\times n\times t}$，同样输入到具有跳跃连接结构的时间特征提取模块，再与时间邻接矩阵TAM进行矩阵乘法操作，得到具有时间信息的特征$f_t\in\mathbb{R}^{c\times n\times n\times t}$，如式(8)所示：

$$f_t=\mathrm{BN}[\mathrm{GELU}(f'\times\mathrm{TAM})]+f' \tag{8}$$

其中，$\mathrm{TAM}\in\mathbb{R}^{t\times t}$.

最后，f_t输入逐点卷积层和GELU激活函数、批归一化层，进行通道间特征的建模，得到第l层输出的特征f_l，如式(9)所示：

$$f_l=\mathrm{BN}\{\mathrm{GELU}[\mathrm{ConvPointwise}(f_t)]\} \tag{9}$$

AMCM通过学习视频特征在空间维度中与情感表达有关的关键区域和时间维度中与情感表现程度相关的时间规律，增强了邻接矩阵中对应区域节点的特征权重. 邻接矩阵和特征矩阵中每个节点特征多次聚合过程中，增强了关键区域和关键时间段特征对聚合特征的贡献，同时提取到了全局和局部特征，增强了特征的情感辨别性.

3.2 视听双模态融合

3.2.1 自注意力机制

Transformer模型的自注意力机制，改进了RNN对先前神经单元隐状态信息的依赖，还将串行计算改为并行计算，如图4所示.

时间序列信息$\boldsymbol{X}=[x_1,x_2,\cdots,x_t]$，$x_i\in\mathbb{R}^{1\times d}$，$i=$

$1,2,\cdots,t$,分别与3个权重相乘得到3个向量$\boldsymbol{Q}=[q_1,q_2,\cdots,q_t]\in\mathbb{R}^{1\times d_q}$、$\boldsymbol{K}=[k_1,k_2,\cdots,k_t]\in\mathbb{R}^{1\times d_k}$和$\boldsymbol{V}=[v_1,v_2,\cdots,v_t]\in\mathbb{R}^{1\times d_v}$,$d_q$、$d_k$和$d_v$分别为对应的向量特征维度.向量$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$分别称为查询向量、键向量和值向量.注意力特征$f_{\text{att}}\in\mathbb{R}^{t\times d_v}$的计算过程如式(10)所示:

$$f_{\text{att}}=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{10}$$

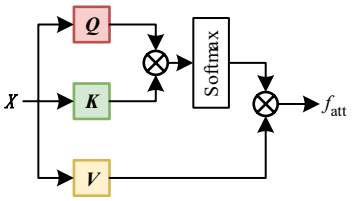

图4　自注意力机制

向量$\boldsymbol{Q}$与$\boldsymbol{K}$相乘再除以向量$\boldsymbol{K}$的维度数d_k的平方根,即为每个时间节点的信息对于当前时间位置信息的相关性权重,Softmax对权重进行归一化,再与向量$\boldsymbol{V}$中每个时间节点信息$v_1,v_2,\cdots,v_t$进行加权求和,最后得到注意力特征f_{att}.

3.2.2　跨模态时间注意力模块

受Transformer模型自注意力机制的启发,本文提出跨模态时间注意力模块CMTAM,具体结构如图5所示,两个呈对称结构的CMTAM组成特征融合模块.CMTAM中$\boldsymbol{Q}$向量与$\boldsymbol{K}$、$\boldsymbol{V}$向量分别属于两种模态,即当$\boldsymbol{Q}$向量属于视频模态,则$\boldsymbol{K}$、$\boldsymbol{V}$向量属于音频模态;当$\boldsymbol{Q}$向量属于音频模态,则$\boldsymbol{K}$、$\boldsymbol{V}$向量属于视频模态.为了表述简便,下文使用M和M'区别两种模态,不特指具体哪种模态.

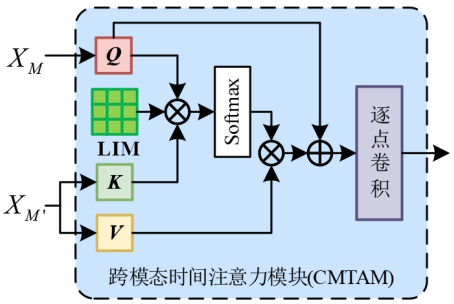

图5　CMTAM结构

CMTAM在自注意力机制基础上,在$\boldsymbol{Q}$、$\boldsymbol{K}$向量计算中引入一个可学习的中间矩阵**LIM**.**LIM**作为整个网络结构的参数矩阵,其参数在网络生成时被随机初始化,通过损失函数反向传播进行优化.模态M和M'的时间特征分别记为$\boldsymbol{X}_M=[x_1,x_2,\cdots,x_t]$,$x_i\in\mathbb{R}^{1\times d}$,$i=1,2,\cdots,t$和$\boldsymbol{X}_{M'}=[x'_1,x'_2,\cdots,x'_{t'}]$,$x'_{i'}\in\mathbb{R}^{1\times d'}$,$i=1,2,\cdots,t'$.根据时间特征$\boldsymbol{X}_M$计算得到查询向量$\boldsymbol{Q}=[q_1,q_2,\cdots,q_t]\in\mathbb{R}^{1\times d_q}$;根据时间特征$\boldsymbol{X}_{M'}$计算得到键向量$\boldsymbol{K}=[k_1,k_2,\cdots,k_t]\in\mathbb{R}^{1\times d_k}$和值向量$\boldsymbol{V}=[v_1,v_2,\cdots,v_t]\in\mathbb{R}^{1\times d_v}$,计算过程如式(11)~(13)所示:

$$\boldsymbol{Q}=\text{BN}[\text{Linear}(\boldsymbol{X}_M)] \tag{11}$$

$$\boldsymbol{K}=\text{BN}[\text{Linear}(\boldsymbol{X}_{M'})] \tag{12}$$

$$\boldsymbol{V}=\text{BN}[\text{Linear}(\boldsymbol{X}_{M'})] \tag{13}$$

其中,Linear为线性层.

$\boldsymbol{Q}$、$\boldsymbol{K}$向量和**LIM**$\in\mathbb{R}^{d_q\times d_k}$相乘,得到$t\times t'$的跨模态情感时间相关性权重,最后Softmax对权重进行归一化.然后将权重矩阵与$\boldsymbol{V}$向量相乘,得到跨模态的时间注意力特征f_{att},计算过程如式(14)所示:

$$f_{\text{att}}=\text{Softmax}\left(\frac{\boldsymbol{Q}\times\textbf{LIM}\times\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{14}$$

为了保留单模态M的时间信息,模态M的$\boldsymbol{Q}$向量通过跳跃连接结构与f_{att}进行加法操作,随后经过逐点卷积层进行通道信息的建模.计算过程如式(15)所示:

$$f_M=\text{ConvPointwise}(f_{\text{att}}+\boldsymbol{Q}) \tag{15}$$

两个CMTAM输出的特征f_M和$f_{M'}$经过拼接融合,输入全连接层进行分类.特征融合模块采用两个CMTAM对称组成,配合模块内的跳跃连接结构,不仅保留了两个单模态特征的信息,还融合了模态之间的时间相关特征.

3.3　损失函数

3.3.1　焦点损失

焦点损失(Focal Loss,FL)[39]函数能很好地解决目标检测中正负样本数量极不平衡的问题,如式(16)所示:

$$L=-\frac{1}{N}\sum_{i=1}^{N}(1-p_{i\hat{y}})^{\gamma}\log(p_{i\hat{y}}) \tag{16}$$

其中,N表示样本数,$p_{i\hat{y}}$表示模型对第i个样本的真实标签类别$\hat{y}$的预测概率.相较于常用的交叉熵损失,引入幂指数权重$(1-p_{i\hat{y}})^{\gamma}$对交叉熵损失值进行非线性压缩,指数$\gamma$可取0、0.5、1和2等数值.

3.3.2　动态权重焦点损失

在焦点损失基础上,为了使模型训练具有人类学习中关注错误,甚至更关注犯错频率高的特性,本文提出动态权重焦点损失FLDW函数,如式(17)所示:

$$\text{FLDW}=-\frac{1}{N}\sum_{i=1}^{N}\alpha(1-p_{i\hat{y}})^{\gamma}\log(p_{i\hat{y}}) \tag{17}$$

式(17)中权重α根据上个训练周期的混淆矩阵和样本分类情况动态生成,如式(18)所示:

$$\alpha=\begin{cases}1, & \hat{y}=y\\ 1+\dfrac{C_{\hat{y},y}}{\sum_{i=1}^{c}C_{\hat{y},i}-C_{\hat{y},\hat{y}}}, & \hat{y}\neq y\end{cases} \tag{18}$$

其中，$\hat{y}\in\{1,2,\cdots,c\}$，为样本的真实情感类别，$y\in\{1,2,\cdots,c\}$，为模型对样本的预测情感类别，$c$为数据集情感类别数. $C_{\hat{y},y}$和$C_{\hat{y},\hat{y}}$分别表示上个训练周期中模型把真实标签为$\hat{y}$的训练样本分类为y和分类正确的次数，$\sum_{i=1}^{c}C_{\hat{y},i}$为上一个训练周期中模型把真实标签为$\hat{y}$的训练样本分类为$i$的次数之和.

混淆矩阵不仅能反映模型的分类能力，还可计算出具体的预测错误频率. 假设共有6种表情分类的上个训练周期的混淆矩阵，如图6所示. 以真实标签为3的训练样本为例，在上个训练周期中，模型对其预测类别分别为2、3和5，即$\hat{y}=3$，错误分类时，$y=5$或$y=2$. 根据图6，$C_{\hat{y},\hat{y}}=87$，$\sum_{i=1}^{6}C_{\hat{y},i}=100$，当$y=5$，即被错分为5时，$C_{\hat{y},y}=8$，计算得到该样本FLDW函数的权重$\alpha$为1.62；当$y$=2，即被错分为2时，权重$\alpha$是1.38.

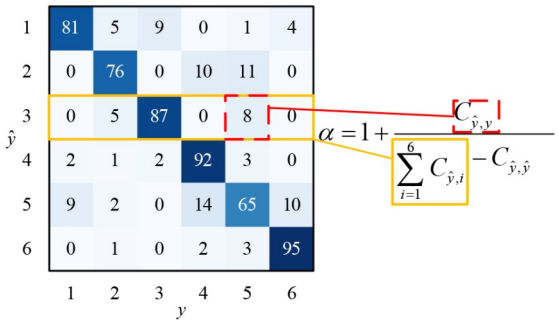

图6 权重α的计算过程

可见，权重α的取值与样本被错误分类次数正相关. 因此，FLDW能使模型对不同类型的错分样本分配不同的关注度，和人类学习相似，训练过程中关注错分样本，且更关注错分频率高的样本，以此提高模型的情感识别准确率.

3.4 训练细节

本文网络模型基于PyTorch开源框架，采用自适应矩估计(Adaptive moment estimation，Adam)优化算法，batch大小为8. 初始学习率为10^{-4}，衰减率为10^{-1}. 训练周期为120，每经过30个训练周期学习率进行一次衰减. 本文使用分类准确率(Accuracy，Acc)作为评价指标. 数据集划分成5份，进行交叉验证，取5折分类准确率平均值进行模型性能评估.

4 实验结果与分析

4.1 数据集与实验环境介绍

本文在eNTERFACE'05数据集[40]上进行各个模块的参数选择实验和消融实验；在eNTERFACE'05与RAVDESS数据集[41]上与其他先进方法进行对比实验.

eNTERFACE'05数据集记录了来自14个不同国家的42个受试者样本，受试者中81%为男性，其余19%为女性. 该数据集包含开心、难过、生气、厌恶、害怕和惊讶6种情感类别，共1 166个视频. 图7展示了eNTERFACE'05数据集6种情感类别对应的样本，同一行图片来自同一个视频.

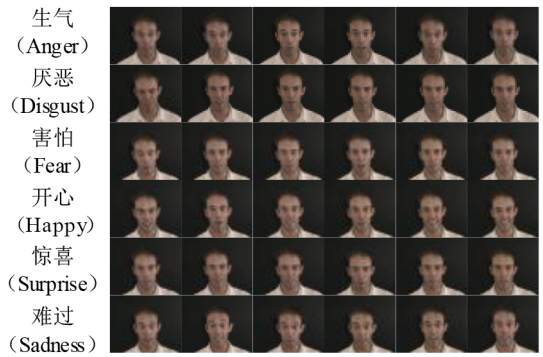

图7 eNTERFACE'05样本示例

RAVDESS数据集记录了24个受试者的样本，其中男女比例为1∶1. 该数据集包含中性、平静、开心、难过、生气、害怕、厌恶和惊喜8种情感类别. 数据集的情感表现包括说话和歌唱两种形式，分别包含1 440个和1 012个视频. 图8展示了RAVDESS数据集8种情感类别对应的样本，同一行图片来自同一个视频.

综合上述两个数据集，在表达难过、害怕和厌恶等消极情绪，或者表达开心、惊喜等积极情绪时，人脸的表现形式具有相似性.

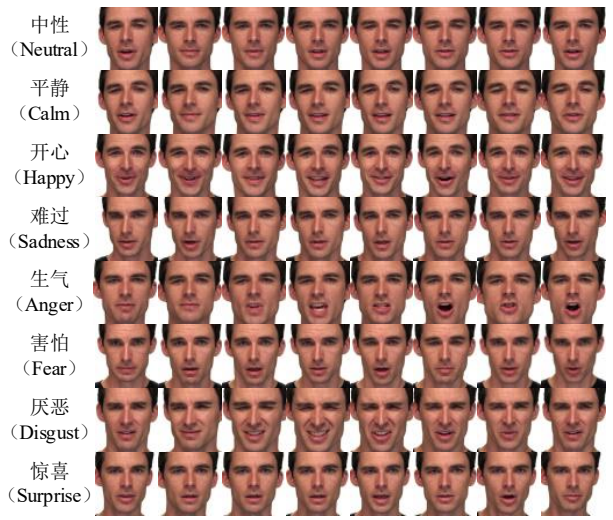

图8 RAVDESS样本示例

对原始数据进行音频和视频分离，提取视频的图像帧和音频的MFCC谱图. 再对图像帧进行人脸检测、对齐、裁剪及尺寸调整，最终得到112×112大小的人脸图像序列，并对图像序列采用随机裁切、旋转等数据增强. 所

有实验在Windows10 64位操作系统上完成,实验硬件设备CPU为Intel(R) Xeon(R) Gold6140 @2.30 GHz,GPU为NVIDIA GeForce GTX 1080 Ti.

4.2　结合邻接矩阵ConvMixer参数选择实验

AMCM模型中特征与邻接矩阵的聚合计算,是以先空间后时间异步方式进行.因此,在实验硬件条件允许的情况,对空间邻接矩阵和时间邻接矩阵的维度分别进行了4组和2组参数选择实验,以探究邻接矩阵维度对模型分类准确率的影响.选择空间邻接矩阵维度时,将时间邻接矩阵维度固定为8×8,空间邻接矩阵维度从784×784减小到64×64的过程中,模型分类准确率先上升后下降,当维度为256×256时,准确率达到最大值93.26%,如表1所示.因此,将空间邻接矩阵维度固定为256×256时,再对时间邻接矩阵维度分别进行16×16和8×8的实验.表1结果表明,当时间邻接矩阵维度为16×16时,AMCM的准确率最高达到94.86%.因此,选定空间邻接矩阵和时间邻接矩阵维度分别为256×256和16×16.

表1　邻接矩阵维度参数选择实验结果

邻接矩阵	维度	准确率/%
空间邻接矩阵	784×784	92.98
	256×256	93.26
	196×196	93.04
	64×64	90.51
时间邻接矩阵	16×16	94.86
	8×8	93.26

本文根据节点的位置关系,对AMCM网络中的空间邻接矩阵权重进行了空间位置还原,如图9所示,图中权重图呈网格化,每一小格代表其他节点指向该位置节点的有向连接权重.权重图颜色越深,代表越多的网格连接权重偏小,网格对应节点聚合得到的特征倾向于局部特征;反之,说明更多网格连接权重偏大,网格对应节点聚合得到的特征倾向于全局特征.即使空间邻接矩阵维度不同,但属于网络浅层的Layer 1层的空间邻接矩阵权重图颜色趋同,说明网络浅层中每个节点的连接情况大致相同,该层提取的特征不具有丰富的空间信息.随着网络层数加深,权重图中不同网格颜色发生变化,一部分网格颜色加深,一部分网格颜色变浅,说明网络深层可以同时提取空间局部特征和全局特征,丰富了特征的空间信息.当空间邻接矩阵维度为256×256时,Layer 4层的空间邻接矩阵权重图呈现出清晰的人脸轮廓,表明AMCM能使模型通过邻接矩阵对patch的空间位置信息进行准确地建模.其中,眼睛、眉毛和嘴巴等能表现情感的关键区域,包含更多的局部信息,可视化图中对应网格颜色更深,说明提取到了更多的空间局部特征;而其他非关键区域,则更加关注视频帧的全局信息,可视化图中对应网格颜色偏浅,则说明赋予了更多的全局特征.但其他维度的空间邻接矩阵,即使处于模型深层Layer也未能准确捕捉patch的空间位置信息,不能提取有用的空间特征.图9空间邻接矩阵权重可视化图与表1中的分类准确性一致,均说明AMCM选取256×256维度时,能提高分类的准确性.

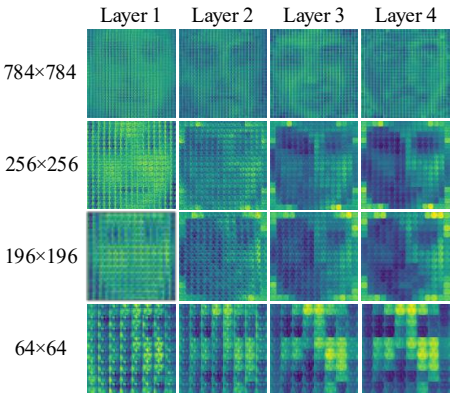

图9　空间邻接矩阵权重可视化

时间邻接矩阵权重可视化如图10所示,维度为16×16的时间邻接矩阵,随着网络加深,较大权重(颜色偏浅)集中于时间邻接矩阵的中部,符合情感从无到有,从微弱到强烈,最后减弱到无的规律,和处于中后部时间段的数据含有最丰富的情感信息一致.但维度为8×8的时间邻接矩阵,无法对时间特征进行准确提取,即使在深层网络结构中,甚至错误关注处于时间段末尾不含有情感信息的特征.

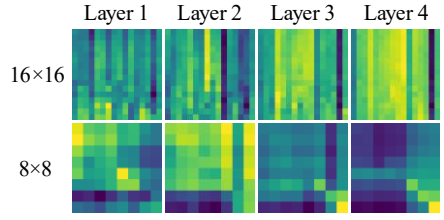

图10　时间邻接矩阵权重可视化

4.3　动态权重焦点损失参数选择实验

FLDW是在焦点损失(Focal Loss,FL)基础上,为其幂指数$(1-p_{i\hat{y}})^{\gamma}$项根据混淆矩阵赋值不同的动态权重.首先,针对FLDW的γ参数选取0.5、1、2和3分别进行参数选择实验.然后,本文提出的FLDW与交叉熵损失CEL、$\gamma=2$的FL、加权焦点损失WFL[37]和校准焦点损失CFL[38]进行对比,实验结果如表2所示.

根据表2可知:

(1)通过①~④可见,FLDW的准确率随参数γ改变,当$\gamma=2$时,分类准确率达到峰值96.22%;但当$\gamma=3$时,分类准确率却降低到94.46%,原因应该是$\gamma=3$时,对于正确分类样本的损失抑制幅度过大,反而使得模

表2 幂指数参数选择与对比实验结果

索引	损失函数	γ	准确率/%
①	FLDW	0.5	94.72
②	FLDW	1	95.13
③	FLDW	2	96.22
④	FLDW	3	94.46
⑤	CEL	—	94.86
⑥	FL	2	95.32
⑦	WFL	2	95.89
⑧	CFL	2	95.44

型在参数还未稳定时就开始忽略正确分类的样本，只关注错误分类的样本.

(2)综合③⑤⑥⑦⑧得到，无论是焦点损失还是其改进算法，都优于传统的交叉熵损失. 可见，通过抑制正确分类样本，增大错误分类样本在损失函数中的占比，模型参数优化方向更侧重将错误分类样本调整为正确分类样本，而不是使正确分类样本更接近其真实标签，从而增强了模型的泛化能力.

(3)根据③⑥⑦⑧，本文设计的FLDW的实验结果最好，FLDW出发点与WFL相同，都是根据先前训练周期的分类情况生成权重，但WFL从类别的整体分类准确率出发，迫使网络更加关注准确率低的类别样本，而FLDW则结合了训练样本所属类别、模型预测类别和上个训练周期的分类情况，更有针对性地生成权重，使网络动态地关注错误分类样本.

4.4 消融实验

为验证本文方法中不同模块的有效性，设计的消融实验如表3所示，表中√表示实验模型包含该模块，×表示不包含该模块. ①表示使用ConvMixer、特征拼接融合和交叉熵损失函数的方法，其实验结果为91.68%. 将①中ConvMixer替换为AMCM，分类准确率提升至94.86%，如②所示. 在②基础上，添加CMTAM模块，分类准确率进一步提高至95.07%，如③所示. 最后将损失函数由交叉熵损失替换为FLDW，分类准确率最终达到96.22%，如④所示.

表3 消融实验结果

索引	AMCM	CMTAM	FLDW	准确率/%
①	×	×	×	91.68
②	√	×	×	94.86
③	√	√	×	95.07
④	√	√	√	96.22

可见，AMCM对于分类精度提升的贡献最大，说明具有合适维度的空间和时间邻接矩阵，能够很好地从空间和时间两个维度对视频信息进行建模，提取更具判别力的局部和全局特征. 但CMTAM对于分类精度提升不太明显，可能由于自注意力机制原本从同模态的特征提取向量$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$，三者的时间维度相同，但本文视频时间特征和音频时间特征的维度差别过大，即使使用了可学习的中间矩阵，自注意力机制也不能非常准确地捕捉视频和音频模态间的时间相关性. 而使用FLDW损失函数，使模型在优化过程中适当增加了对错误分类样本的关注度，提升了模型的泛化能力，分类准确率提升了1.15%.

4.5 对比试验

本文在eNTERFACE'05和RAVDESS数据集上，与近4年涉及视频和音频两种模态的模型进行了对比. 在eNTERFACE'05数据集上的对比结果如表4所示，本文方法的分类准确率为96.22%，比端到端方法[8,9,27,30,31]的分类准确率有明显提升，比文献[8]的工作提高了5.4%. 但低于两个非端到端的方法[7,32]，分析与使用大量手工特征和更大规模人脸图像数据集进行预训练有关.

表4 eNTERFACE'05对比结果

对比文献	核心方法	年份	准确率/%
文献[7]	手工特征+PCA+特征拼接	2019	98.73
文献[27]	CNN+LSTM+特征拼接	2019	86.89
文献[8]	CNN+ PCA+典型相关分析	2020	90.82
文献[31]	CNN+CapsGCN+注意力	2021	80.23
文献[30]	CNN+LSTM+特征拼接+BEL	2021	81.70
文献[32]	CNN+LSTM+GCN	2021	97.07
文献[9]	K-mean+典型相关分析+SVM	2022	87.20
本文方法	AMCM+CMTAM+FLDW	—	96.22

本文方法在RAVDESS数据集上得到了97.55%的分类准确率，优于大部分对比工作的分类准确率，对比结果如表5所示. 但低于2020年Ma等人[34]工作的准确率，差距仅为0.02%，但本文方法的分类准确率是在数据集所有8种情感类别的样本上取得，比其方法多出中性(Neutral)和平静(Calm)两种情感类别.

表5 RAVDESS对比结果

对比文献	核心方法	年份	准确率/%
文献[33]	CNN+LSTM+相关性损失	2019	79.00
文献[34]	ResNet+相关性损失	2020	97.57
文献[16]	CNN+LSTM	2021	80.08
文献[17]	Transformer+LSTM	2021	86.70
文献[22]	VGG+GRU	2021	90.00
文献[10]	CNN+LSTM+特征拼接	2022	86.00
文献[5]	ResNet+自注意力机制	2022	87.89
Ours	AMCM+CMTAM+FLDW	—	97.55

本文方法在eNTERFACE'05和RAVDESS数据集上情感类别识别结果的混淆矩阵分别如图11和图12所示. 在eNTERFACE'05数据集上，本文方法在厌

恶、开心、难过和惊喜这4种情感上取得很好的识别效果,均在96%以上,开心情感类别的准确率达到了97.76%. 但生气和害怕这2种情感类别只得到了95.83%和94.95%的识别效果. 在RAVDESS数据集上,本文方法对平静、开心和害怕3种情感的识别效果好,均在98%以上. 中性、生气情感类别取得了98.41%和98.67%的分类准确率. 难过和惊喜2种情感类别的分类准确率也均在96%,但厌恶情感上的识别效果只有95.88%.

图11　eNTERFACE'05混淆矩阵

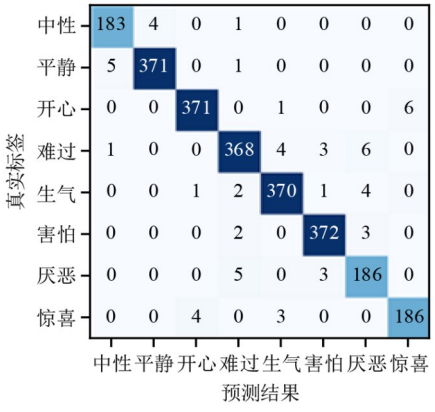

图12　RAVDESS混淆矩阵

两个混淆矩阵的结果表明,本文方法在两个数据集上对难过、害怕和厌恶3种消极情感类别的样本,容易错误分类为上述3种情感类别中的另外两种. 而开心和惊喜两种情感,倾向于错误分类为彼此的情感类别.

5　总结

针对目前情感计算领域中已有的深度学习模型卷积核尺寸受硬件条件限制、感受野小,多模态特征融合方式简单,损失函数忽略错误分类样本等问题,本文设计了结合邻接矩阵的ConvMixer和具有动态权重焦点损失函数的视听情感识别方法. 结合邻接矩阵的ConvMixer网络,通过空间和时间邻接矩阵,在空间和时间维度上对视频帧序列提取全局和局部特征,丰富了特征中的情感信息,增强了特征的情感辨别性. 两个跨模态时间注意力模块以对称的结构融合视频和音频的时间信息,捕捉跨模态的时间相关性,增强了模态间特征的时间信息. 提出动态权重焦点损失,模拟人类学习过程中更为关注错误的机制,提高模型的泛化能力. 在情感计算领域公开数据集上设计了参数选择实验、消融实验和对比实验,结果验证了本文方法及其组成模块的有效性,提高了情感分类的准确性. 但跨模态时间注意力模块对情感类别的分类性能提升效果不明显,可能由于视频和音频不是独立存在的,只通过中间矩阵把两个模态特征联系起来,依然不能很好地提取到跨模态的情感时间信息. 接下来需更多关注跨模态融合方式,深入研究多模态数据之间的相关性.

参考文献

[1] ROUAST P V, ADAM M T P, CHIONG R. Deep learning for human affect recognition: Insights and new developments[J]. IEEE Transactions on Affective Computing, 2019, 12(2): 524-543.

[2] 张瑞, 蒋晨之, 苏剑波. 基于稀疏特征挑选和概率线性判别分析的表情识别研究[J]. 电子学报, 2018, 46(7): 1710-1718.
ZHANG R, JIANG C Z, SU J B. Expression recognition based on sparse selection and plda[J]. Acta Electronica Sinica, 2018, 46(7): 1710-1718. (in Chinese)

[3] BIRHALA A, RISTEA C N, RADOI A, et al. Temporal aggregation of audio-visual modalities for emotion recognition[C]//2020 43rd International Conference on Telecommunications and Signal Processing (TSP). Piscataway: IEEE, 2020: 305-308.

[4] 李宏菲, 李庆, 周莉. 基于多视觉描述字及音频特征的动态序列人脸表情识别[J]. 电子学报, 2019, 47(8): 1643-1653.
LI H F, LI Q, ZHOU L. Dynamic facial expression recognition based on multi-visual and audio descriptor[J]. Acta Electronica Sinica, 2019, 47(8): 1643-1653. (in Chinese)

[5] MOCANU B, TAPU R. Audio-video fusion with double attention for multimodal emotion recognition[C]//2022 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP). Piscataway: IEEE, 2022: 1-5.

[6] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Net-

works, 2008, 20(1): 61-80.

[7] NOROOZI F, MARJANOVIC M, NJEGUS A, et al. Audio-visual emotion recognition in video clips[J]. IEEE Transactions on Affective Computing, 2019, 10(1): 60-75.

[8] WU M, SU W, CHEN L, et al. Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition[J]. IEEE Transactions on Affective Computing, 2022, 13(2): 805-817.

[9] CHEN L, WANG K, WU M, et al. K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition[J]. IEEE Transactions on Industrial Electronics, 2022, 70(1): 1016-1024.

[10] 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.

[11] DU Z, WU S, HUANG D, et al. Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition[J]. IEEE Transactions on Affective Computing, 2019, 12(3): 565-578.

[12] LIU D, ZHANG H, ZHOU P. Video-based facial expression recognition using graph convolutional networks[C]// 25th International Conference on Pattern Recognition (ICPR). Milan: IEEE, 2021: 607-614.

[13] ZHAO S, MA Y, GU Y, et al. An end-to-end visual-audio attention network for emotion recognition in user-generated videos[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 303-311.

[14] CHEN J, LUO Z, ZHANG Z, et al. Polar transformation on image features for orientation-invariant representations[J]. IEEE Transactions on Multimedia, 2018, 21(2): 300-313.

[15] ZHANG S, DING Y, WEI Z, et al. Continuo-us emotion recognition with audio-visual leader-follower attentive fusion[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 3567-3574.

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

[17] LUNA-JIMÉNEZ 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.

[18] 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.

[19] HOSSAIN M S, MUHAMMAD G. Emotion recognition using deep learning approach from audio-visual emotional big data[J]. Information Fusion, 2019, 49: 69-78.

[20] WANG J, XUE M, CULHANE R, et al. Speech emotion recognition with dual-sequence LSTM architecture[C]// 45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2020: 6474-6478.

[21] MENG H, YAN T, YUAN F, et al. Speech emotion recognition from 3D log-mel spectrograms with deep learning network[J]. IEEE Access, 2019, 7: 125868-125881.

[22] SONG Y, CAI Y, TAN L. Video-audio emotion recognition based on feature fusion deep learning method[C]// 2021 64th International Midwest Symposium on Circuits and Systems (MWSCAS). Piscataway: IEEE, 2021: 611-616.

[23] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]//International Conference on Learning Representations. Piscataway: IEEE, 2021: 1-21

[24] TROCKMAN A, KOLTER J Z. Patches are all you need? [EB/OL]. (2022-01-24)[2022-08-15]. https://arxiv.org/abs/2201.09792.

[25] 孙晓, 潘汀 . 基于兴趣区域深度神经网络的静态面部表情识别[J]. 电子学报, 2017, 45(5): 1187-1197.
SUN X, PAN T. Static facial expression recognition system using ROI deep neural network[J]. Acta Electronica Sinica, 2017, 45(5): 1189-1197. (in Chinese)

[26] BASBRAIN A M, GAN J Q, SUGIMOTO A, et al. A neural network approach to score fusion for emotion recognition[C]//2018 10th Computer Science and Electronic Engineering (CEE-C). Piscataway: IEEE, 2018: 180-185.

[27] MA F, ZHANG W, LI Y, et al. An end-to-end learning approach for multimodal emotion recognition: Extracting common and private information[C]//2019 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2019: 1144-1149.

[28] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 1(4): 568-576.

[29] BIRHALA A, RISTEA C N, RADOI A, et al. Temporal aggregation of audio-visual modalities for emotion recognition[C]//2020 43rd International Conference on Telecommunications and Signal Processing (TSP). Piscat-

away: IEEE, 2020: 305-308.
[30] FARHOUDI Z, SETAYESHI S. Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition[J]. Speech Communication, 2021, 127: 92-103.
[31] LIU J, CHEN S, WANG L, et al. Multimodal emotion recognition with capsule graph convolutional based representation fusion[C]//2021 46th International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2021: 6339-6343.
[32] NIE W, REN M, NIE J, et al. C-GCN: Correlation based graph convolutional network for audio-video emotion recognition[J]. IEEE Transactions on Multimedia, 2021, 23: 3793-3804.
[33] GHALEB E, POPA M, ASTERIADIS S. Multimodal and temporal perception of audio-visual cues for emotion recognition[C]//2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). Piscataway: IEEE, 2019: 552-558.
[34] MA F, ZHANG W, LI Y, et al. Learning better representations for audio-visual emotion recognition with common information[J]. Applied Sciences, 2020, 10(20): 7239.
[35] ZHONG Y, HU Y, HUANG H, et al. A lightweight model based on separable convolution for speech emotion recognition[C]//Interspeech 2020, 21st Annual Conference of the International Speech Communication Association. Lyon: ISCA, 2020: 3331-3335.
[36] ZHU Z, DAI W, HU Y, et al. Speech emotion recognition model based on Bi-GRU and focal loss[J]. Pattern Recognition Letters, 2020, 140: 358-365.
[37] 李锵, 赵启蒙, 关欣. 基于动态卷积的胸部X光片疾病分类算法[J]. 天津大学学报(自然科学与工程技术版), 2022, 55(9): 953-964.
LI Q, ZHAO Q M, GUAN X. Classification algorithm for chest X-ray diseases based on dynamic convolution[J]. Journal of Tianjin University (Science and Technology), 2022, 55(9): 953-964. (in Chinese)
[38] BAI H, CHENG J, SU Y, et al. Calibrated focal loss for semantic labeling of high-resolution remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 6531-6547.
[39] LIN T, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2980-2988.
[40] MARTIN O, KOTSIA I, MACQ B, et al. The enterface'05 audio-visual emotion database[C]//2006 22nd International Conference on Data Engineering Workshops (ICDEW'06). Piscataway: IEEE, 2006: 8-8.
[41] 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): 1-35.

作者简介

师　硕　女，1981年出生，河北保定人. 2006年在东北大学信息学院获得工学硕士学位，2014年在河北工业大学电子信息工程学院获得工学博士学位，现为河北工业大学人工智能学院副教授. 主要研究方向为情感计算、图像处理和计算机视觉.
E-mail: shishuo@hebut.edu.cn

于　洋　男，1981年出生，天津人. 2008年在河北工业大学人工智能与数据科学学院获得工学硕士学位，2012年在河北工业大学电子信息工程学院获得工学博士学位，现为河北工业大学人工智能与数据科学学院副教授. 主要研究方向为智能交通系统、表情与微表情识别、目标检测与重识别.
E-mail: yuyang@hebut.edu.cn