计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(1):254−262 [doi: 10.15888/j.cnki.csa.009352]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于加权多头并行注意力的局部遮挡面部表情识别①

郭　胜[1,2], 蔡　姗[1,2], 邹　雪[1,2], 周珍胜[1,2], 王　林[1,2]

[1](贵州民族大学 数据科学与信息工程学院, 贵阳 550025)
[2](贵州民族大学 贵州省模式识别与智能系统重点实验室, 贵阳 550025)
通信作者: 王　林, E-mail: wanglin@gzmu.edu.cn

摘　要: 面部表情识别在诸多领域具有广泛的应用价值, 但在识别过程中局部遮挡会导致面部难以提取有效的表情识别特征, 而局部遮挡的面部表情识别可能需要多个区域的表情特征, 单一的注意力机制无法同时关注面部多个区域特征. 针对这一问题, 本文提出了一种基于加权多头并行注意力的局部遮挡面部表情识别模型, 该模型通过并行多个通道-空间注意力提取局部未被遮挡的多个面部区域表情特征, 有效缓解了遮挡对表情识别的干扰, 大量的实验结果表明, 本文的方法相比于很多先进的方法取得了最优的性能, 在 RAF-DB 和 FERPlus 上的准确率分别为 89.54%、89.13%, 在真实遮挡的数据集 Occlusion-RAF-DB 和 Occlusion-FERPlus 的准确率分别为 87.47%、86.28%. 因此, 本文的方法具有很强的鲁棒性.

关键词: 面部表情识别; 局部遮挡; 表情特征识别; 注意力机制; 加权多头并行注意力; 神经网络

引用格式: 郭胜,蔡姗,邹雪,周珍胜,王林.基于加权多头并行注意力的局部遮挡面部表情识别.计算机系统应用,2024,33(1):254–262. http://www.c-s-a.org.cn/1003-3254/9352.html

Facial Expression Recognition with Local Occlusion Based on Weighted Multi-head Parallel Attention

GUO Sheng[1,2], CAI Shan[1,2], ZOU Xue[1,2], ZHOU Zhen-Sheng[1,2], WANG Lin[1,2]

[1](School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China)
[2](Guizhou Province Key Laboratory of Pattern Recognition and Intelligent Systems, Guizhou Minzu University, Guiyang 550025, China)

Abstract: Facial expression recognition (FER) has widespread application significance in many fields, but it is difficult to extract effective FER features due to local occlusion during the recognition. FER with local occlusion may require expression features of multiple regions, and a single attention mechanism cannot focus on the features of multiple facial regions simultaneously. To this end, this study proposes a local occlusion FER model based on weighted multi-head parallel attention. The model extracts the expression features of multiple facial regions that are not occluded by multiple channels in parallel-spatial attention, alleviating the occlusion interference on expression recognition. A large number of experiments show that the proposed method yields the best performance compared with many advanced methods, and the accuracy on RAF-DB and FERPlus is 89.54% and 89.13%, respectively. On the occluded datasets Occlusion-RAF-DB and Occlusion-FERPlus, the accuracy is 87.47% and 86.28%, respectively. Therefore, this method has strong robustness.

Key words: facial expression recognition; local occlusion; expression feature recognition; attention mechanism; weighted multi-head parallel attention; neural network

① 收稿时间: 2023-06-24; 修改时间: 2023-07-27; 采用时间: 2023-08-11; csa 在线出版时间: 2023-11-24
CNKI 网络首发时间: 2023-11-28

© 中国科学院软件研究所 http://www.c-s-a.org.cn

1 引言

面部表情识别是人类表达情感的主要方式之一, 在人际交往过程中, 面部表情是人们相互传递和理解情绪状态的重要媒介之一. 随着人工智能的发展, 其在人机交互、自动驾驶、安全监控、心理健康评估、智能辅助等方面具有广泛的应用研究价值. 因此, 面部表情识别受到越来越多的关注, 我们把对面部表情 (中性、高兴、生气、悲伤、恐惧、厌恶、惊讶、轻蔑等) 的情绪状态研究称为面部表情识别. 虽然面部表情识别在许多领域展现出很高的应用价值, 但在识别过程中会遇到诸多不可避免的问题, 如局部遮挡. 局部遮挡导致面部难以提取有效的识别特征, 从而降低局部遮挡条件下面部表情识别的准确率. 如何在面部处于局部遮挡的状态下准确地识别面部表情, 仍然是一个亟待解决的重大研究课题.

早期传统的机器学习方法主要使用手工制作的特征或浅层学习, 如局部二值模式 (LBP)[1]、稀疏学习[2]和方向梯度直方图 (HOG)[3]. 然而, 在局部遮挡下这些手工制作的特征通常不够鲁棒和准确. 随着深度学习的不断发展, 许多研究者提出了各种深度学习网络模型[4–6]. 近年来, 局部遮挡的面部表情识别的注意机制也得到了更多的研究, 重点关注更有意义的表达区域[7,8]. Wang 等人[9]提出了区域注意力网络 (RAN) 来自适应的获取表情的关键区域, 缓解了局部遮挡变化带来的表情识别问题, 但采用固定位置裁剪、随机裁剪和基于标志点的裁剪方法, 可能导致局部区域的位置不确定性, 同时使用单一的注意力机制可能无法同时关注面部的多个区域特征. Zhao 等人[10]设计了全局多尺度和局部注意网络 (MA-Net) 来关注局部与全局显著特征, 以减少局部遮挡的干扰问题, 但单一的局部注意力块专注几个不重叠的区域, 可能会导致一些关键的信息丢失. Xue 等人[11]提出了能够学习丰富的关系感知局部表示的迁移模型, 引导该模型学习斑块内的不同信息, 并确定斑块之间的丰富关系, 该方法可以提高面部表情识别的准确性, 并且能够更好地捕捉到不同表情之间的关系, 在一定程度上缓解了局部遮挡面部表情识别的问题. Liu 等人[12]提出了自适应局部裁剪, 特别是对眼和嘴部分别进行裁剪, 引导模型找到更易区分的部分, 对局部遮挡变化具有鲁棒性, 但该方法无法准确裁剪与表情相关的关键区域, 可能会产生一些冗余的特征. Ju 等人[13]基于掩码的注意力并行网络, 利用关键地标检测提取的二值掩码构建基于掩码的注意力模块, 定位与表达相关的区域, 嵌入到并行网络提取特征, 将提取的并行特征从空间维度上分割成多个独立的块, 独立预测面部表情, 以解决局部遮挡问题. 单一的掩码注意力模块可能无法获取面部多个区域的特征, 导致获取的面部特征信息不够丰富. Gong 等人[14]提出一种基于多特征融合网络 (MFNet), 设计了浅层 Gabor 卷积网络, 增强了学习特征对方向和尺度变化的适应能力及局部细节特征的捕获能力, 提高了野外表情识别的准确率. 该方法利用局部和分层特征来缓解野外的面部表情识, 在遮挡的情况下, 提取特征具有一定的局限性. Ruan 等人[15]针对局部遮挡的面部, 提出了一种构建路径选择多网络模型的方法, 以实现面部局部遮挡场景下的表情识别. 文献[16]提出了一种遮挡面部表情识别框架 FER-AM, 结合注意力机制网络对不同的特征分配不同的权重, 进而能够使得 FER-AM 更关注人脸面部的关键特征, 如未被遮挡的区域, 最终能够有效地解决面部遮挡问题.

局部遮挡条件下的直接对面部进行特征提取仍然面临一定的困难, 因此, 许多学者尝试着利用注意力机制关注局部特征, 以获得关键区域的表达信息及鲁棒特征. 注意力机制是通过引导模型关注局部未被遮挡的面部区域, 有效选择面部局部区域特征, 缓解了遮挡条件下的面部表情识别问题. 然而, 这些方法都是通过单一的注意力机制来引导模型, 而局部遮挡的面部表情识别可能需要多个区域特征来识别表情. 因此, 为了解决局部遮挡条件下面部难以提取有用识别特征的问题, 本文提出了一种基于加权多头并行注意力的局部遮挡面部表情识别模型, 该模型能够同时关注局部未被遮挡的多个面部区域, 并且以较小的参数量和计算量提高了局部遮挡下面部表情识别的准确率.

2 模型框架

本文所提出的模型由特征提取器 (ResNet18)、加权多头并行注意力网络 (SE-multi-head parallel attention network, SE-MPAN) 和多注意力融合网络 (multi-attention fusion network, MAFN) 这 3 部分组成, 如图 1 所示. 首先通过特征网络提取面部表情的基本人脸特征; 然后, SE-MPAN 由多个单一的加权注意力网络 (SE-attention network, SE-AN) 组成, 构造多头注意子空间来同时关注面部多个区域特征; 最后, MAFN 来融合多个区域特征的分布, 同时减少重复区域的聚焦, 更全面地捕捉关键表达区域.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

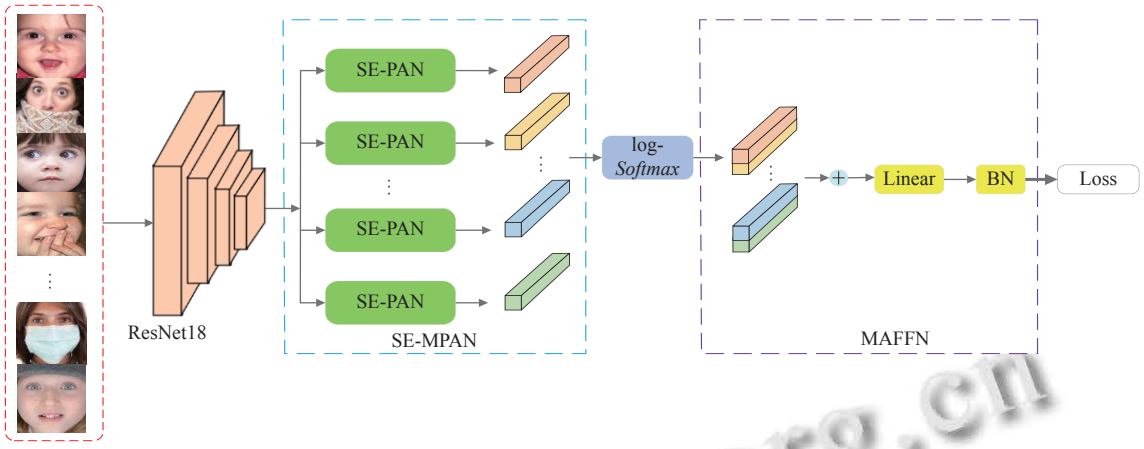


图 1 基于加权多头并行注意力的局部遮挡面部表情识别模型结构图

2.1 特征提取器

为了建立轻量级的网络模型, 本文采用轻量级的残差网络 ResNet18[17]作为骨干网络进行特征提取, 残差网络结构利用残差连接的方法能够很好地处理网络退化、梯度消失及爆炸问题. 考虑到面部表情存在类内与类间的变化问题, 为了更好地区分不同表情, 本文设计了内聚损失函数, 以优化面部表情的类内距离和类间距离, 并构建更精确的表情识别空间分布.

令 γ_r 是 ResNet18 骨干网络, p_r 是其参数, x_i 是输入特征向量, 则:

$$X_i' = \gamma_r(p_r, x_i) \tag{1}$$

对于 k 类的面部表情识别, 在 ResNet18 骨干网络提取特征过程中, 采用内聚损失函数进行优化类内距离和类间距离, 该损失可表述为:

$$L_C = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|x_i' - c_{y_i}\right\|_2^2}{\sigma_c^2} \tag{2}$$

其中, N 表示 batch 中训练样本的数量, c_{y_i} 表示相应的分类中心, σ_c^2 表示不同类的方差. 该损失函数能够更好地重构表情特征, 增强不同类表情的内聚性, 达到优化类内距离和类间距离的目的.

2.2 加权注意力网络

如图 2 所示, 加权注意力网络 (SE-attention network, SE-AN) 由通道注意单元和空间注意单元并行连接组成, 以获取感兴趣的面部特征区域.

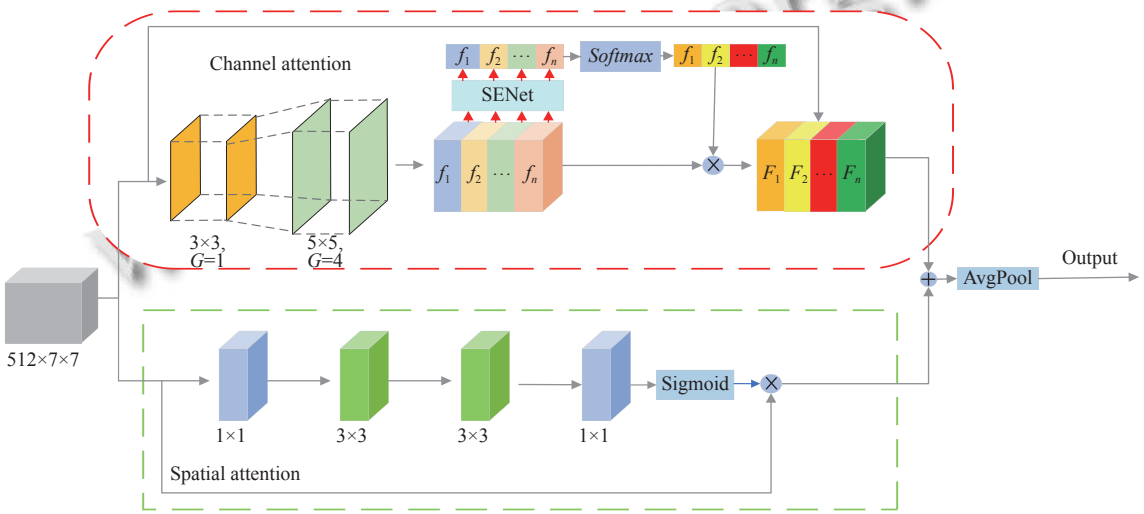


图 2 加权并行注意力网络 (SE-PAN)

通道注意单元由两个普通的3×3卷积、两个5×5的分组卷积、SEWeight 及 *Softmax* 组成. 具体流程如算法 1 所示. 首先通过骨干网络 ResNet18 提取的特征图S, $S \in R^{7\times7\times512}$, 其中 7 是空间大小, 512 是通道大

© 中国科学院软件研究所 http://www.c-s-a.org.cn

小. 所提取的特征图 S 经过通道注意单元来获取通道特征, 具体来说, 先用两个普通的3×3卷积、两个组为4 的5×5卷积来获得多尺度特征, 每个尺度下的特征通道数为原来总通道数C的$1/S$, 其中S为 4, 分组卷积能够减少卷积操作的计算量和参数量, 提高模型的运行速度和效率. 然后对多尺度特征$f \in (f_1, f_2, f_3, \cdots, f_n)$进行拼接, 该计算过程可表述为:

$$f = Concat([f_1, f_2, \cdots, f_n]),\ n = 1, 2, \cdots, S \tag{3}$$

为了增强每个通道间的关系, 使用轻量级的 SENet 模块[18]对获取的多尺度通道特征进行重新加权, 该模块称为 SEWeight. 首先 SEWeight 的 squeeze 采用全局池化操作, 先将$W \times H \times C$特征图压缩到$1 \times 1 \times C$, 这样就能获得全局感受野的信息. 之后是一个 excitation, 给每个部分一个相关性的权重值. 该计算过程可表述为:

$$Z_C = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} f(i, j) \tag{4}$$

$$Z = \sigma(W_2 \delta(W_1 Z)) \tag{5}$$

其中, $W \times H$为空间维度, W_1和W_2是用于降维和升维的两个全连接层的权重, δ是 ReLU 激活函数, σ是 Sigmoid 激活函数.

算法 1. SE-AN 通道单元模块

Input: 输入空间大小为H, W, 通道数为C的特征图$X_i' \in R^{7\times7\times512}$

1. conv_kernels=[3, 3, 5, 5], conv_groups=[1, 1, 4, 4] //两个普通的3×3卷积、两个组为 4 的5×5卷积

2. for S=1 to 4 do //将特征图X_i'的通道进行切分成 S=4 组

3. n=1, i=0

4. S_n=conv_kernels[i] //针对划分出来的每个通道特征图, 进行不同的卷积核提取多尺度特征

5. n=n+1, i=i+1

6. end

7. end for

8. $f \in (f_1, f_2, f_3, \cdots, f_n)$ // 提取的多尺度特征

9. Z_n=$SEWeight(f_n)$, n = 1, 2, 3, ···, S //利用 SEWeight 进行重新加权, 以加强通道内部间的交互

10. F_n=$f_n \times Softmax(Z_n)$ //通过 *Softmax* 函数获得更多自适应的注意力权重, 最后将加权重构的多尺度注意力权重与相应的尺度特征相乘

11. F_{output}=$X_i' \times Concat([F_1, F_2, \cdots, F_n])$ //将所有加权重构的特征进行组合, 得到最终的通道特征, 并与输入的特征X_i'相乘, 完成整个全局与局部交互

Output: 通道特征图F_{output}

对提取的多尺度特征的通道信息使用 SEWeight 进行重新加权, 以加强通道内部间的交互, 加权计算可表述为:

$$Z_n = SEWeight(f_n),\ n = 1, 2, 3, \cdots, S \tag{6}$$

其中, Z_n表示不同尺度下的通道注意权值, 将所有Z_n进行串联, 得到整个通道注意向量. 然后通过 *Softmax* 函数获得更多自适应的注意力权重, 最后将加权重构的多尺度注意力权重与相应的尺度特征相乘, 该计算过程表述为:

$$F_n = f_n \times Softmax(Z_n) \tag{7}$$

将所有加权重构的特征进行组合, 得到最终的通道特征, 并与输入的特征X_i'相乘, 完成整个全局与局部交互, 输出的F_{output}表述为:

$$F_{output} = X_i' \times Concat([F_1, F_2, \cdots, F_n]) \tag{8}$$

空间注意单元由两个1×1卷积、两个3×3卷积组成. 具体流程如算法 2. 具体来说, 先将特征$F \in R^{C\times H\times W}$投影降维得到$F \in R^{C/r\times H\times W}$, 该过程使用1×1卷积整合和压缩跨通道维度的特征映射. 然后, 使用与通道分支相同的衰减率$r = 16$, 衰减后, 利用两个3×3卷积来提取空间特征信息. 最后, 利用1×1卷积将特征再次简化为$R^{1\times H\times W}$空间注意力图. 每次卷积后都利用 BN 层进行特征归一化和 ReLU 进行激活. 卷积操作完成后通过 Sigmoid 激活函数得到每个通道下的权重. 通过将获得的权重与输入特征X_i'相乘, 最终得到空间注意特征F'_{output}, F'_{output}可表述为:

$$F'_{output} = X_i' \times \sigma(f_1^{1\times1}(f_1^{3\times3}(f_1^{3\times3}(f_1^{1\times1}(X_i'))))) \tag{9}$$

其中, X_i'是主干网络的输出特征, $f_1^{3\times3}(\cdot)$表示进行3×3卷积, $f_1^{1\times1}(\cdot)$表示进行1×1卷积, σ是 Sigmoid 激活函数.

算法 2. SE-AN 空间单元模块

Input: 输入空间大小为H, W, 通道数为C的特征图$X_i' \in R^{7\times7\times512}$

1. X_i'投影降维得到$R^{C/r\times H\times W}$ //r=16

2. f_1=$\delta(BN(f_1^{1\times1}(X_i')))$ //1×1 卷积整合和压缩跨通道维度的特征映射, δ是 ReLU 激活函数

3. $f_{3\times3}$=$\delta(BN(f_1^{3\times3}(f_1^{3\times3}(f_1))))$ //两个 3×3 卷积来提取空间特征信息

4. f_1=$\delta(BN(f_1^{1\times1}(f_{3\times3})))$ //利用 1×1 卷积将特征再次简化为$R^{1\times H\times W}$空间注意力图

5. F'_{output}=$X_i' \times \sigma(f_1)$ //σ是 Sigmoid 激活函数

6. end

Output: 空间特征图F'_{output}

最后, 将处理后的通道注意特征和空间注意特征行融, 经过全局平均池化 (GAP) 操作, 得到 SE-AN 注

© 中国科学院软件研究所 http://www.c-s-a.org.cn

意特征图.

$$Attention_i = F_{\text{output}} + F'_{\text{output}},\ i = 1,2,3,\cdots,h \tag{10}$$

其中, h是注意头的数量.

2.3 多注意特征融合

加权多头并行注意力 (SE-MPAN) 由多个加权注意力网络 (SE-AN) 组成, 具体流程如算法 3 所示. SE-MPAN 能够捕捉面部的多个区域, 形成多个注意图, 为了避免注意力重叠, 首先利用 log-*Softmax* 函数对注意图进行缩放, 强调最感兴趣的关键区域, 然后使用分区损失引导注意力集中关注不同区域, 最后将注意力图通过 BN 层归一化.

算法 3. SE-MPAN 模块

Input: SE-AN 注意特征图, $Attention_i=F_{\text{output}}+F'_{\text{output}}$, i=1, 2, 3,⋯, h, 假设$A=Attention_i\in R^{n\times c}$

1. A=log(*Softmax*(A)) //利用 log-*Softmax* 函数对注意图进行缩放, 强调最感兴趣的关键区域

2. A=*BN*(A) //归一化注意力图

3. A=*Summation*(A) //MAFN 融合归一化的注意力图

Output: 预测表情类别

假设$A = Attention_i \in R^{n\times c}$, 则 log-*Softmax* 函数可表述为:

$$\log(Softmax(A)) = \frac{\exp(A_i)}{\sum_{j=1}^{c}\exp(A_j)} \tag{11}$$

其中, A_i是注意图的第i个向量, A_j是注意图向量的第j个元素.

分区损失可表示为:

$$L_{\text{AF}} = \frac{1}{NC}\sum_{i=1}^{N}\sum_{j=1}^{C}\log\left(1+\frac{1}{\sigma_{i,j}^2}\right) \tag{12}$$

其中, C是注意图的通道大小, $\sigma_{i,j}^2$是第j个通道上第i个样本的方差.

2.4 交叉熵损失函数 (cross-entropy loss)

本文所使用的损失函数为 cross-entropy loss, 可表述为:

$$L_{\text{CE}} = -\frac{1}{N}\sum_{i=0}^{N-1}\log\frac{e^{W_{y_i}^{(k)\text{T}}v_i^{(k)}+b_{y_i}^{(k)}}}{\sum_{j=0}^{C-1}e^{W_{y_j}^{(k)\text{T}}v_j^{(k)}+b_{yj}^{(k)}}} \tag{13}$$

其中, N代表样本数, N代表类别数, $W^{(k)}$为FC层权重矩阵, $b^{(k)}$为FC层偏置项, $v_i^{(k)}$是第i个样本的输入, y_i是类标签.

最终的损失函数为:

$$L = \lambda_1 L_C + \lambda_2 L_{\text{AF}} + L_{\text{CE}} \tag{14}$$

其中, L_C表示内聚损失, L_{AF}表示注意分区损失, L_{CE}表示交叉熵损失, λ_1、λ_2表示L_C、L_{AF}的贡献值.

3 实验及结果

3.1 实验设置

在 RAF-DB 和 FERPlus 数据集上, 使用官方对齐的图像训练, 所有数据集的输入都将图像重塑为 224×224 大小的像素, 采用随机裁剪、水平翻转和擦除等数据增强的方法来避免过拟合. 实验采用 ResNet18 作为模型的基准网络, 为了公平评估模型, 使用在 MSCeleb-1M[19] 人脸识别数据集上预训练的 ResNet18 模型. 我们的方法是在 Linux/CentOS 7.6 操作系统下基于 PyTorch 代码实现的, 模型是在 A100-PCIE-40GB 的工作站上训练的. 模型的总参数是 18.13M, 浮点运算次数为 2.02 GFLOPs.

在模型训练过程中, 使用带动量优化器的 SGD 进行优化, 其中动量参数为 0.9, 权重衰减设置为 1E–4. 本文的方法在 RAF-DB 和 FERPlus 数据集上进行 100 epochs 的模型训练, 批量大小分别为 256 和 64, 初始学习率分别为 0.1 和 0.04, 每 10 个 epochs 衰减为原来的 0.1.

3.2 数据集

RAF-DB[20]是一个包含 29 672 张真实场景的面部图像数据集, 标注了 7 种基本表情和 11 种复合表情标签, 在本文实验中, 使用 7 种基本表情进行表情识别, 其中训练集 12 271 张图像, 测试集 3 068 张图像.

FERPlus[21]是在 FER2013 数据集上重新标记得到的真实场景中的表情数据集, 这些图像被重新标记为 10 类极度不平衡的表情. 本文实验选择在 7 种基本表情中增加了 contempt 类别表情进行测试. 其中包含 28 709 张训练图像, 3 589 张验证图像, 3 589 张测试图像, 大小为 48×48 像素.

Occlusion-RAF-DB 和 Occlusion-FERPlus 是 Wang 等人[9]为了检验真实遮挡场景下面部表情识别模型的性能, 分别从 RAF-DB 的测试集和 FERPlus 的测试集中建立的两个遮挡数据子集, 包含 735 和 605 张, 这些测试集标注了口罩、眼镜和手等不同类型的遮挡, 在本文的实验中, 使用 Occlusion-RAF-DB 和 Occlusion-

© 中国科学院软件研究所 http://www.c-s-a.org.cn

FERPlus 两个遮挡数据子集分别验证本文所提方法.

3.3 消融分析

(1) 验证 SE-MPAN 模块的数量对模型性能的影响. 如图 3, 横坐标表示注意头数, 纵坐标表示准确率, RAF-DB 准确率随着注意头数 (1−10) 的增加而变化, 当头数为 2 时, 达到了最好的准确率 89.54%, 实验表明, 提出的 MAFFN 优于单一的注意力模块. 当头数大于 2 时, 会产生精度上的降低或震荡. 对于精度上的降低可能是因为增加注意力头数会引入一些冗余的信息. 当不同的注意力头都注意到了相似的特征时, 模型可能会对这些特征进行过度关注, 从而导致信息的冗余, 降低了模型的精度. 对于精度上的震荡可能是因为增加注意力头数会增加模型的复杂度, 使得模型的训练变得更加困难, 造成模型训练不稳定, 导致精度的震荡. 为了验证 SE-MPAN 的有效性, 使用类激活映射 Grad-CAM++[22]进行可视化, 如图 4 所示, 数据集来自 RAF-DB, 第 1 行是 baseline 的可视化效果, 第 2、3 行是 SE-MPAN 可视化效果. 第 1、4 张图像为遮挡人脸, 模块只能聚焦局部未被遮挡的多个区域, 这与之前的假设是一致的. 第 2、3 张图像为正面人脸, 模块也能够聚焦面部局部多个显著区域.

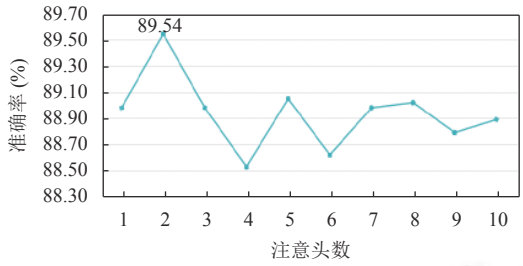


图 3 注意头数对模型的影响

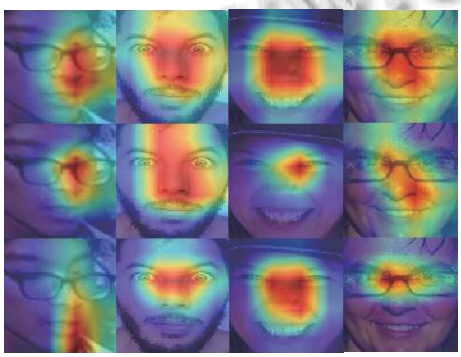


图 4 SE-MPAN 可视化效果

(2) 超参数λ_1、λ_2的影响. 如表 1 所示. 在 RAF-DB 及 FERPlus 数据集上评估λ_1、λ_2的影响对模型的影响. 首先将λ_2的值固定为 1.0, 然后依次从 0.1 到 1.0 验证λ_1的值. 结果表明, 当λ_1=0.2 时, MAFFN 在两个数据集上都达到了最高的准确率. 其次, 固定λ_1=0.2, λ_2的取值范围是 0.1–2.0. 当λ_2=0.5 时, 模型在两个数据集中获得最佳性能, 精度随着该值的增加而降低. 因此, λ_1、λ_2分别取值为 0.2、0.5.

3.4 实验结果

在本文中, 所提出的模型在公开数据集 RAF-DB 和 FERPlus 上与最先进的方法进行比较. 使用准确率 (*Accuracy*) 评估模型的性能, 准确率通常用于衡量分类模型的性能, 它表示模型正确预测的样本数与总样本数之间的比例, 可表示为:

$$Accuracy = \frac{T}{N} \times 100\% \tag{15}$$

其中, T表示为正确预测的样本数, N表示为总样本数.

RAF-DB: 从表 2 中的结果表明, 本文的方法优于许多先进的方法, 获得了 89.54% 的准确率. 与基于注意力的方法 RAN[9]、LANet[23]及 MAPNet[13]相比, 本文的方法在性能上优于三者. 相比于其他特征网络, 本文的方法仍然高于性能最好的 MFNet[14]方法, 提高了 1.01%.

FERPlus: 从表 3 中的结果表明, 本文的方法达到了 89.13% 的准确率. 对于基于注意力的方法 RAN[9], 我们的方法优于 RAN, 提高了 0.58%. 同时, 我们的方法仍然优于其他特征提取网络模型.

表 1 超参数λ_1、λ_2对模型的影响

超参数	数值	RAF-DB (%)	FERPlus (%)
λ_1	0.1	89.05	88.49
	0.2	**89.21**	**88.84**
	0.3	89.15	88.46
	0.4	88.66	88.52
	0.5	88.53	88.56
	1.0	88.92	88.74
λ_2	0.1	89.08	88.46
	0.25	89.02	88.94
	0.5	**89.54**	**89.13**
	1.0	89.21	88.74
	1.5	88.75	88.40
	2.0	89.05	88.72

训练和验证精度曲线: 本文分析了 RAF-DB 和 FERPlus 两个数据集的训练和验证精度曲线, 从图 5 可以看出, 在 RAF-DB 和 FERPlus 两个数据集上分别迭代 100 个 epochs, 相对较短的学习周期内, 模型收敛速度快, 第 18 个 epochs 之后, 模型趋近于稳定状态, 并达到最佳性能的准确率.

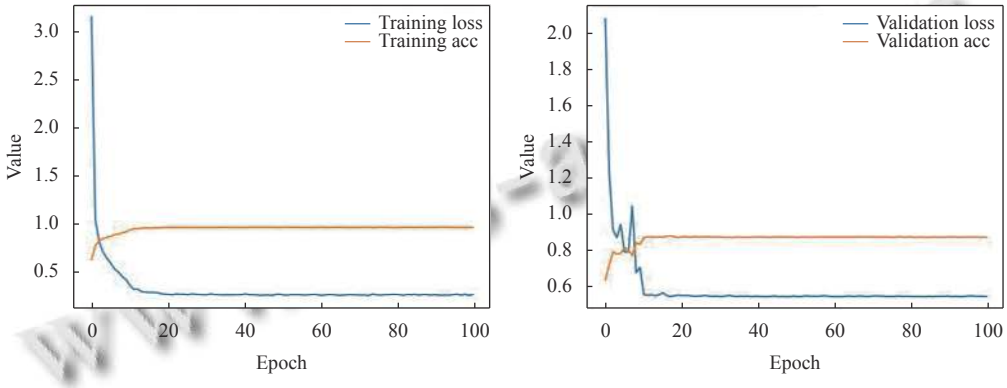
© 中国科学院软件研究所 http://www.c-s-a.org.cn

表 2 在 RAF-DB 上与最先进的方法进行比较 (%)

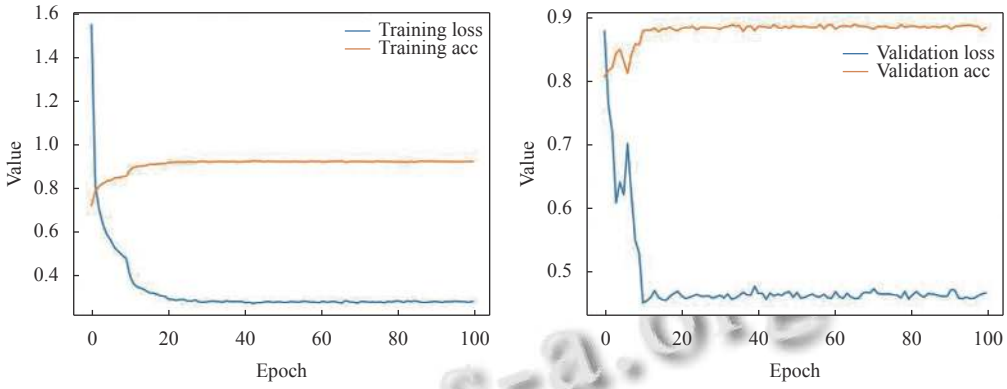
方法	年份	*Accuracy*
RAN[9]	2020	86.90
MA-Net[10]	2020	88.40
EfficientFace[24]	2021	88.36
LANet[23]	2021	86.70
MAPNet[13]	2022	87.26
MFNet[14]	2022	88.53
Ours	2023	**89.54**

表 3 在 FERPlus 上与最先进的结果进行比较 (%)

方法	年份	*Accuracy*
RAN[9]	2020	88.55
SCN[25]	2020	88.01
VTFF[26]	2021	88.81
ADC-Net[27]	2021	88.90
SpResNet-ViT[28]	2022	88.10
IPD-FER[29]	2022	88.42
Ours	2023	**89.13**



(a) RAF-DB



(b) FERPlus

图 5 RAF-DB、FERPlus 数据集的训练和验证精度曲线

3.5 真实遮挡的数据集上验证模型

为了评估本文模型在真实遮挡条件下的性能, 本文的方法在真实遮挡 Occlusion-RAF-DB 和 Occlusion-FERPlus 数据集上进行实验验证. 图 6 是一些真实遮挡数据样本图像. 从表 4 中的结果表明, 本文的方法在 Occlusion-RAF-DB 和 Occlusion-FERPlus 数据集上分别获得了 87.47%、86.28%, 这明显优于最先进的方法. 因此, 在真实遮挡的面部表情数据集上的实验结果表明, 本文的方法在真实遮挡变化的条件下具有很强的鲁棒性, 这对于实际应用来说, 解决面部局部遮挡变化问题, 提高模型在测试集上的识别准确率以及在计算机视觉领域内落地使用, 才是最具有价值的.

4 结论

本文针对在面部局部遮挡条件下难以提取有效的识别特征问题, 其中局部遮挡的面部表情识别可能需要多个区域特征识别表情, 而单一的注意力机制无法同时关注面部多个区域特征识别表情. 因此, 本文提出了一种基于加权并行注意力的局部遮挡面部表情识别模型, 该模型能够同时关注局部未被遮挡的多个面部区域, 有效缓解了遮挡对表情识别的干扰, 并且以较小的参数量和计算量提高了局部遮挡下面部表情识别的

© 中国科学院软件研究所 http://www.c-s-a.org.cn

准确率. 在 RAF-DB 和 FERPlus 公开数据集上的实验表明, 本文的模型相比与其他最先进的方法取得了最优的性能. 在真实遮挡条件下, 同样表现出最鲁棒的性能. 在未来的工作中, 将设计更鲁棒、更轻量化的注意力机制来提高局部遮挡条件下的面部表情识别, 提高其在诸多领域的应用价值具有重大意义.

图 6 真实遮挡数据样本

表 4 在真实遮挡的数据集 Occlusion-RAF-DB 和 Occlusion-FERPlus 上进行评估 (%)

数据集	方法	*Accuracy*
Occlusion-RAF-DB	ResNet18[9]	80.19
	RAN[9]	82.72
	EfficientFace[24]	83.24
	MA-Net[10]	83.65
	VTFF[26]	83.95
	MPCSAN[30]	86.26
	Ours	**87.47**
Occlusion-FERPlus	ResNet18[9]	73.33
	RAN[9]	83.63
	VTFF[26]	84.79
	MPCSAN[30]	86.12
	Ours	**86.28**

参考文献

1 Shan CF, Gong SG, McOwan PW. Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing, 2009, 27(6): 803–816. [doi: 10.1016/j.imavis.2008.08.005]

2 Zhong L, Liu QS, Yang P, *et al*. Learning active facial patches for expression analysis. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 2562–2569.

3 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587.

4 Wang JH, Ding HY, Wang SF. Occluded facial expression recognition using self-supervised learning. Proceedings of the 16th Asian Conference on Computer Vision. Macao: Springer, 2022. 1077–1092.

5 Liu C, Hirota K, Dai YP. Patch attention convolutional vision transformer for facial expression recognition with occlusion. Information Sciences, 2023, 619: 781–794. [doi: 10.1016/j.ins.2022.11.068]

6 Zhang XH, Zhang XM, Zhou JZ, *et al*. Occlusion-aware facial expression recognition based region re-weight network. Proceedings of the 18th Pacific Rim International Conference on Artificial Intelligence. Hanoi: Springer, 2021. 209–222.

7 Lu Y, Wang SG, Zhao WT, *et al*. WGAN-based robust occluded facial expression recognition. IEEE Access, 2019, 7: 93594–93610. [doi: 10.1109/ACCESS.2019.2928125]

8 Fan YR, Li VOK, Lam JCK. Facial expression recognition with deeply-supervised attention network. IEEE Transactions on Affective Computing, 2022, 13(2): 1057–1071. [doi: 10.1109/TAFFC.2020.2988264]

9 Wang K, Peng XJ, Yang JF, *et al*. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 2020, 29: 4057–4069. [doi: 10.1109/TIP.2019.2956143]

10 Zhao ZQ, Liu QS, Wang SM. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE Transactions on Image Processing, 2021, 30: 6544–6556. [doi: 10.1109/TIP.2021.3093397]

11 Xue FL, Wang QC, Guo GD. TransFER: Learning relation-aware facial expression representations with transformers. Proceedings of the 2021 IEEE/CVF International Conference

© 中国科学院软件研究所 http://www.c-s-a.org.cn

on Computer Vision. Montreal: IEEE, 2021. 3581–3590.
12 Liu HW, Cai HL, Lin QC, *et al*. Adaptive multilayer perceptual attention network for facial expression recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 6253–6266. [doi: 10.1109/TCSVT.2022.3165321]
13 Ju LZ, Zhao X. Mask-based attention parallel network for in-the-wild facial expression recognition. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022. 2410–2414.
14 Gong WJ, Wang CQ, Jia JL, *et al*. Multi-feature fusion network for facial expression recognition in the wild. Journal of Intelligent & Fuzzy Systems, 2022, 42(6): 4999–5011.
15 Ruan LH, Han YX, Sun JR, *et al*. Facial expression recognition in facial occlusion scenarios: A path selection multi-network. Displays, 2022, 74: 102245. [doi: 10.1016/j.displa.2022.102245]
16 张本文, 高瑞玮, 乔少杰. 新型融合注意力机制的遮挡面部表情识别框架. 重庆理工大学学报 (自然科学), 2023, 37(9): 217–226.
17 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.
18 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.
19 Guo YD, Zhang L, Hu YX, *et al*. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 87–102.
20 Li S, Deng WH, Du JP. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2852–2861.
21 Barsoum E, Zhang C, Ferrer CC, *et al*. Training deep networks for facial expression recognition with crowd-sourced label distribution. Proceedings of the 18th ACM International Conference on Multimodal Interaction. Tokyo: ACM, 2016. 279–283.
22 Zhou BL, Khosla A, Lapedriza A, *et al*. Learning deep features for discriminative localization. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2921–2929.
23 Ma H, Celik T, Li HC. Lightweight attention convolutional neural network through network slimming for robust facial expression recognition. Signal, Image and Video Processing, 2021, 15(7): 1507–1515. [doi: 10.1007/s11760-021-01883-9]
24 Zhao ZQ, Liu QS, Zhou F. Robust lightweight facial expression recognition network with label distribution training. Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI Press, 2021. 3510–3519.
25 Wang K, Peng XJ, Yang JF, *et al*. Suppressing uncertainties for large-scale facial expression recognition. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6897–6906.
26 Ma FY, Sun B, Li ST. Facial expression recognition with visual transformers and attentional selective fusion. IEEE Transactions on Affective Computing, 2023, 14(2): 1236–1248. [doi: 10.1109/TAFFC.2021.3122146]
27 Xia HY, Li CY, Tan YM, *et al*. Destruction and reconstruction learning for facial expression recognition. IEEE MultiMedia, 2021, 28(2): 20–28. [doi: 10.1109/MMUL.2021.3076834]
28 Gao WJ, Li L, Zhao HY. Facial expression recognition method based on SpResNet-ViT. Proceedings of the 2nd Asia-Pacific Conference on Communications Technology and Computer Science. Shenyang: IEEE, 2022. 182–187.
29 Jiang J, Deng WH. Disentangling identity and pose for facial expression recognition. IEEE Transactions on Affective Computing, 2022, 13(4): 1868–1878. [doi: 10.1109/TAFFC.2022.3197761]
30 Gong WJ, Qian YR, Fan YY. MPCSAN: Multi-head parallel channel-spatial attention network for facial expression recognition in the wild. Neural Computing and Applications, 2023, 35(9): 6529–6543. [doi: 10.1007/s00521-022-08040-4]

(校对责编: 孙君艳)

© 中国科学院软件研究所 http://www.c-s-a.org.cn