西北大学学报(自然科学版)
2024 年 4 月,第 54 卷第 2 期,Apr. ,2024,Vol. 54,No. 2
Journal of Northwest University (Natural Science Edition)

·人工智能情感计算·

嵌入混合注意力机制的 Swin Transformer 人脸表情识别

王坤侠[1,2],余万成[1],胡玉霞[1,2]

(1. 安徽建筑大学 电子与信息工程学院,安徽 合肥 230601;
2. 安徽省古建筑智能感知与高维建模国际联合研究中心,安徽 合肥 230601)

摘要 人脸表情识别是心理学领域的一个重要研究方向,可应用于交通、医疗、安全和刑事调查等领域。针对卷积神经网络(CNN)在提取人脸表情全局特征的局限性,提出了一种嵌入混合注意力机制的 Swin Transformer 人脸表情识别方法,以 Swin Transformer 为主干网络,在模型 Stage3 的融合层(Patch Merging)中嵌入了混合注意力模块,该方法能够有效提取人脸面部表情的全局特征和局部特征。首先,层次化的 Swin Transformer 模型可有效获取深层全局特征信息。其次,嵌入的混合注意力模块结合了通道和空间注意力机制,在通道维度和空间维度上进行特征提取,从而让模型能够更好地提取局部位置的特征信息。同时,采用迁移学习方法对模型网络权重进行初始化,进而提高模型的精度和泛化能力。所提方法在 FER2013、RAF-DB 和 JAFFE 这 3 个公共数据集上分别达到了 73.63%、87.01%和 98.28%的识别准确率,取得了较好的识别效果。

关键词 表情识别;Transformer;注意力机制;迁移学习
中图分类号:TP391.4 **DOI**:10.16152/j.cnki.xdxbzr.2024-02-003

Facial expression recognition in Swin Transformer by embedding hybrid attention mechanism

WANG Kunxia[1,2], YU Wancheng[1], HU Yuxia[1,2]

(1. School of Electronic and Information Engineering, Anhui Jianzhu University, Hefei 230601, China;
2. Anhui International Joint Research Center for Ancient Architecture Intellisencing and Multi-Dimensional Modeling, Hefei 230601, China)

Abstract Facial expression recognition is an important research domain in psychology that can be applied to many fields such as transportation, medical care, security, and criminal investigation. Given the limitations of convolutional neural networks (CNN) in extracting global features of facial expressions, this paper proposes a Swin Transformer method embedded with a hybrid attention mechanism for facial expression recognition. Using the Swin Transformer as the backbone network, a hybrid attention module is embedded in the fusion layer (Patch Merging) in the model of Stage3, which can effectively extract global and local features from facial expressions. Firstly, the hierarchical Swin Transformer model can effectively obtain deep global features. Second-

收稿日期:2023-10-18
基金项目:国家自然科学基金青年项目(62105002);安徽省住房城乡建设科学技术计划项目(2023-YF113,2023-YF004);安徽建筑大学智能建筑与建筑节能安徽省重点实验室开放课题(IBES2022ZR02)。
第一作者:王坤侠,女,博士,副教授,从事人工智能情感计算研究,kxwang@ahjzu.edu.cn。

ly, the embedded hybrid attention module combines channel and spatial attention mechanisms to extract features in the channel dimension and spatial dimension, which can attain better local features. At the same time, this article uses the transfer learning method to initialize the model network weights, thereby improving the recognition performance and generalization ability. The proposed method achieved recognition accuracies of 73.63%, 87.01%, and 98.28% on three public datasets (FER2013, RAF-DB, and JAFFE) respectively, achieving good recognition results.

Keywords expression recognition; Transformer; attention mechanism; transfer learning

人脸表情是人类传递情感和意图最直接有效的方式之一。人脸表情识别(facial expression recognition, FER)可以通过机器分析识别人脸图像中的不同表情种类[1]。在人脸表情识别过程中,特征提取尤为重要,一种好的特征提取方法将有效提高表情识别的准确率。在使用深度学习技术进行面部表情特征提取时,目前大多数研究工作倾向于使用卷积神经网络(CNN)进行特征提取[2-4]。一些经典的 CNN 模型,例如 ResNet[3] 在图像分类任务中取得了较好的效果。文献[4]在 ResNet 基础上提出了 NA-Resnet 模型,该模型利用 NA 模块提取表面特征来辅助人脸表情识别。

卷积神经网络具有共享卷积核和平移不变性等优点,但 CNN 模型对于全局上下文信息的建模能力相对较弱,不能很好地提取全局特征。Transformer[5] 中的自注意力机制能有效获取全局信息,并且可以通过多头自注意力机制将所获得的特征信息映射到多个空间,从而增强模型的全局感知能力。目前,研究人员已将 Transformer 广泛应用在计算机视觉领域[6],并取得了较好的效果。在 2020 年,Google 团队提出的 Vision Transformer(ViT)模型[7] 在图像分类领域取得了显著的成果。ViT 是一种基于 Transformer 架构的图像分类模型,它将图像分割成小的图块,然后通过将这些图块转换为序列传入 Transformer 中进行特征提取。然而 ViT 需要在大规模数据集上进行训练,并需要更多的算力资源支持,为了解决 ViT 的训练困难特性,PVT[8]、CvT[9] 和 Swin Transformer[10] 等模型都采用了不同的优化策略。同时,许多研究人员也将 Transformer 成功应用于人脸表情识别,并取得了较好的效果。其中,文献[11]介绍了 PACVT 人脸表情识别模型,该模型通过利用 PAU 模块提取局部特征,同时采用 Transformer 提取全局特征,最后将这 2 种特征进行融合,用于人脸表情识别任务。文献[12]提出了 FST-MWOS 人脸表情识别模型,该模型以 Swin Transformer 为基础,加入了多重权重优化机制,以提高模型识别精度。文献[13]将自监督学习与 Vision Transformer 进行联合预训练,提出了 SSF-ViT 模型用于人脸表情识别。

此外,注意力机制能够有效地提取局部特征信息。近年来,随着注意力机制的流行,出现了多种类型的注意力机制[14],如空间注意力机制 STN[15]、通道注意力机制 ECA-Net[16] 和混合注意力机制 CBAM[17] 等。其中,通道注意力机制 ECA-Net 致力于对通道维度特征进行自适应的重要性加权,以增强网络对重要通道信息的关注,从而提高特征提取的能力。空间注意力机制 STN 则专注于对特征图的空间变换和注意力调整。通过对空间位置的显式建模,STN 可以对模型感兴趣区域进行准确地提取和调整,从而增强对局部特征的提取能力。混合注意力模块 CBAM 结合了通道和空间注意力机制,使得模型网络能够同时在通道维度和空间维度上进行特征提取和加权。注意力网络也在人脸表情识别得到应用,文献[18]提出了空时注意力网络用于表情识别。

为更有效地提取人脸表情特征,本文将 Transformer 与注意力机制相结合,提出了一种嵌入混合注意力机制的 Swin Transformer 人脸表情识别方法。该方法在 Swin Transformer 网络基础上进行了改进,在模型的内部 Patch Merging 层中嵌入了混合注意力模块 CBAM,并利用迁移学习的方法对权重进行初始化,以提高模型训练的速度和人脸表情识别的准确度。

1　模型设计

1.1　Swin Transformer 模型

经典的 Transformer 架构对 N 个 token 进行自注意力计算,模型的计算复杂度为 $O(N^2)$,而 Swin Transformer 采用了一种分而治之的优化思想,将模型的计算复杂度降低为 $O(N)$。因此,本

文中选用了 Swin Transformer 作为人脸表情识别模型的骨干网络。同时,层次化的 Swin Transformer 模型能够从多种尺寸和维度的特征图中提取特征信息,该模型主要由 4 个 Stage 组成,如图 1 所示。在 Swin Transformer 模型中,主要由 Patch Merging 层和 Swin Transformer Block 串联组成。Patch Merging 层能够根据设定的下采样倍率对人脸表情特征图进行下采样操作,在该层中嵌入注意力模块,可以有效地提取多维度的人脸表情特征信息。

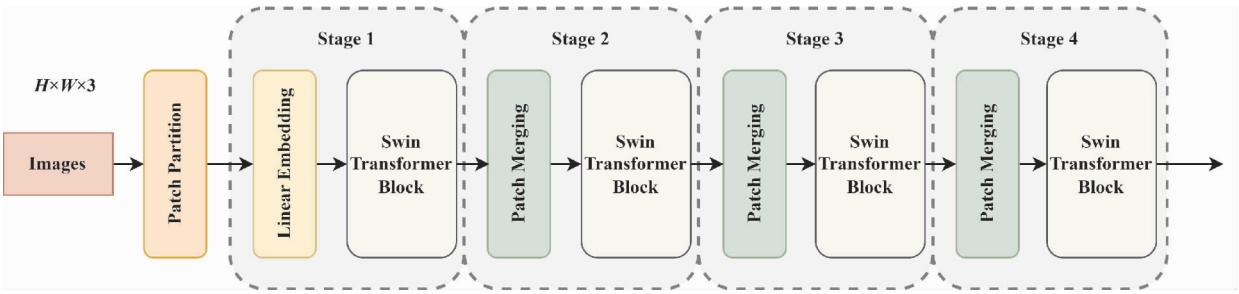


图 1 Swin Transformer 网络框架

Fig. 1 Swin Transformer network structure

在 Swin Transformer Block 中,核心部分包括窗口多头自注意力(W-MSA)和移动窗口多头自注意力(SW-MSA),具体结构如图 2 所示。W-MSA 能够将自注意力的计算限制在窗口内,从而可以有效地降低模型内部的计算量,但这种方式会存在一个明显的问题,窗口之间的连接缺失可能会导致全局信息的丢失,并限制模型对全局特征的建模能力。因此,在 SW-MSA 中引入了基于移动窗口机制的跨窗口操作,以增加窗口之间的信息交互。在窗口移位和分割之后,使用循环移位和反向循环移位处理窗口的数量增大和大小不一致问题。通过这种方法,可以实现相邻窗口之间的信息交互,从而扩大模型的全局感受野,获取图像更高层的语义信息。这样能够更好地提取人脸表情的全局语义特征,使得模型在表情识别任务中能够更加准确地识别不同的表情种类。

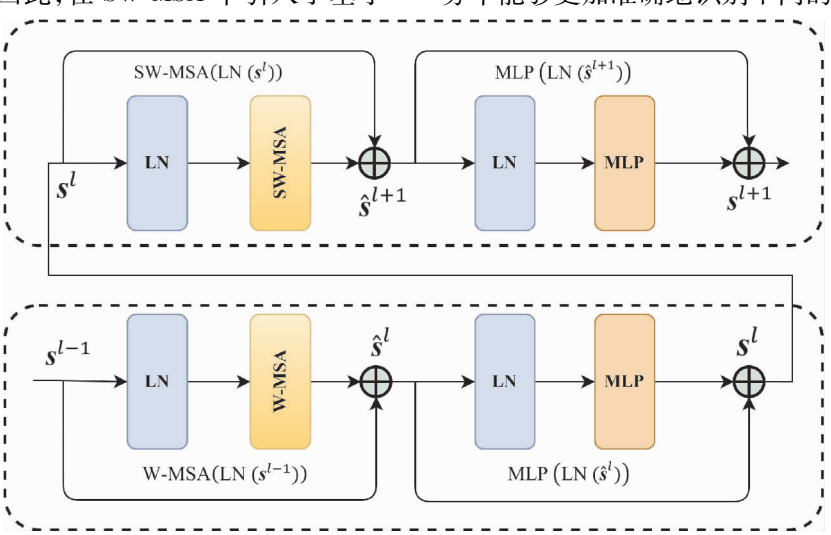


图 2 Swin Transformer Block 框架

Fig. 2 Swin Transformer Block module structure

在图 2 中,$\hat{s}^l$ 表示(S)W-MSA 输出的特征信息;s^l 表示 MLP 模块输出的特征信息;LN 表示 Layer Norm 层;W-MSA 表示基于窗口的多头自注意力模块;SW-MSA 表示移动窗口的多头自注意力模块;MLP 表示多层感知机。如式(1) ~ (4) 所示。

$$\hat{s}^l = \text{W-MSA}(\text{LN}(s^{l-1})) + s^{l-1} \tag{1}$$

$$s^l = \text{MLP}(\text{LN}(\hat{s}^l)) + \hat{s}^l \tag{2}$$

$$\hat{s}^{l+1} = \text{SW-MSA}(\text{LN}(s^l)) + s^l \tag{3}$$

$$s^{l+1} + \text{MLP}(\text{LN}(\hat{s}^{l+1})) + \hat{s}^{l+1} \tag{4}$$

1.2 CBAM 注意力机制

CBAM 注意力机制能够帮助模型更加关注人脸表情的重要特征信息,并忽略目标周围的干扰因素,从而提高人脸表情识别模型的准确性。CBAM 注意力模块是一种混合型注意力机制,由

2 个独立部分组成:通道注意力模块和空间注意力模块。通过引入通道和空间注意力机制,CBAM 能够自适应地调整不同通道和空间位置上的特征权重,使得模型能够更好地捕捉和利用局部特征信息。相比于只有单通道注意力机制的 SE-Net[19],CBAM 能够取得更好的识别效果,CBAM 总体网络框架如图 3 所示。

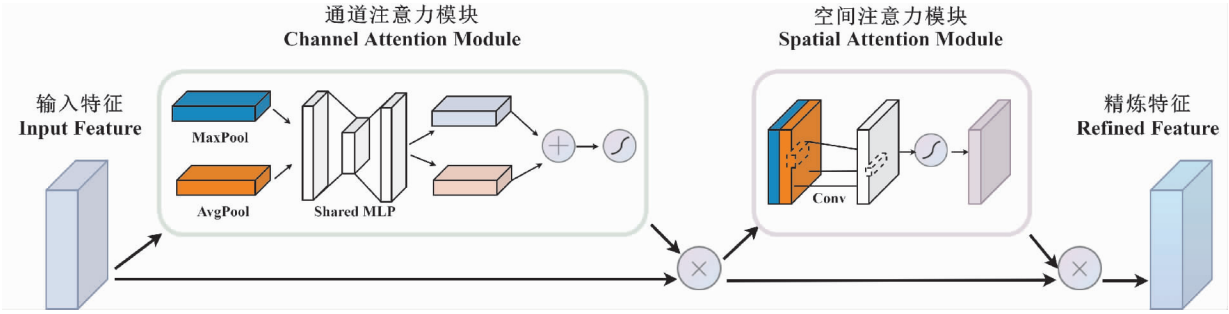


图 3　CBAM 网络框架

Fig. 3　CBAM network structure

通道注意力模块首先对输入的特征图进行全局平均池化(AvgPool)和最大池化(MaxPool)操作,分别得到描述特征 $\boldsymbol{F}_{c\_avg}$ 和 $\boldsymbol{F}_{c\_max}$。然后,这 2 个特征经过共享多层感知机进行特征相加操作,并通过 Sigmoid 函数进行处理,得到 $M_c(\boldsymbol{F})$。最后,将通道权重系数 $M_c(\boldsymbol{F})$ 与输入的特征图 $\boldsymbol{F}$ 进行相乘,得到通道注意力特征图 $\boldsymbol{F}'$。如式(5)、(6)所示。

$$M_c(\boldsymbol{F}) = \sigma(\mathrm{MLP}(\boldsymbol{F}_{c\_avg}) + \mathrm{MLP}(\boldsymbol{F}_{c\_max})) \tag{5}$$

$$\boldsymbol{F}' = M_c(\boldsymbol{F}) \otimes \boldsymbol{F} \tag{6}$$

空间注意力模块再对输入的特征图 $\boldsymbol{F}'$ 进行平均池化和最大池化操作,分别得到描述特征 $\boldsymbol{F}_{s\_avg}$ 和 $\boldsymbol{F}_{s\_max}$。然后,将这 2 个特征进行横向拼接聚合,传入大小为 7 × 7 的卷积核进行操作($f^{7\times7}$)。接着,将得到的结果经过 Sigmoid(σ) 操作,得到 $M_s(\boldsymbol{F}')$。最后,通过将空间权重系数 $M_s(\boldsymbol{F}')$ 与输入的特征图 $\boldsymbol{F}'$ 进行相乘,得到混合注意力特征图 $\boldsymbol{F}''$。如式(7)、(8)所示。

$$M_s(\boldsymbol{F}') = \sigma(f^{7\times7}([\boldsymbol{F}_{s\_avg};\boldsymbol{F}_{s\_max}])) \tag{7}$$

$$\boldsymbol{F}'' = M_s(\boldsymbol{F}') \otimes \boldsymbol{F}' \tag{8}$$

1.3　嵌入混合注意力机制的 Swin Transformer 模型

基于上述的 Swin Transformer 模型和 CBAM 混合注意力模块,本文提出了嵌入混合注意力机制的 Swin Transformer 人脸表情识别模型。该模型以 Swin Transformer 作为骨干网络,并嵌入了 CBAM 混合注意力模块。具体结构如图 4 所示。

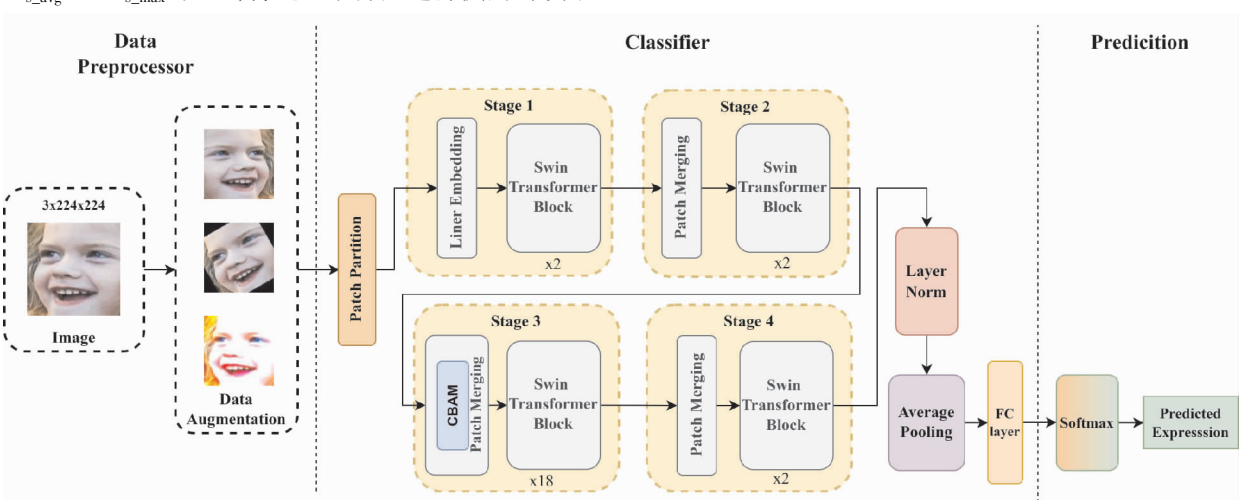


图 4　模型具体结构

Fig. 4　Specific structure of the model

在该模型中,先将人脸表情图像通过 Patch Partition 层将图像分割成多个 Patch。然后,通过 Stage1 中的 Liner Embedding 层,将划分好的 Patch 进行线性映射后传入 Swin Transformer Block 中,以便更好地提取特征信息。特征提取完成后再输入到下一个 Stage。本文在 Stage3 中的 Patch Merging 层嵌入了混合注意力模块 CBAM,该模块的嵌入能够有效地提升模型对局部特征的捕捉能力,并且能够抑制特征周围不必要区域的影响,从而加强模型的感知能力并提高人脸表情识别的准确率。本文模型的主要思想是利用 Transformer 模型提取全局特征信息,并运用混合注意力机制获取局部特征信息,进而在模型训练中对全局特征和局部特征进行融合,以实现对人脸表情特征更精准的识别。

为了将 Swin Transformer 模型更好地应用于人脸表情分类任务,本文对模型结构进行了微调。在模型顶层中添加了 LayerNorm 层、自适应平均池化层、全连接层和 Softmax 层。

2 实验与结果分析

2.1 数据集介绍

为了验证本文模型的有效性,我们选用了 3 个被广泛应用的公共数据集和 1 个私有数据集作为实验数据集。公共数据集包括 FER2013 数据集[20]、JAFFE 数据集[21]和 RAF-DB 数据集[22]。其中,所选的数据集包含了自然环境下的人脸表情数据集和实验室环境下的标准人脸表情数据集。图 5 展示了这 3 个公共数据集中各类表情的部分图像样本。

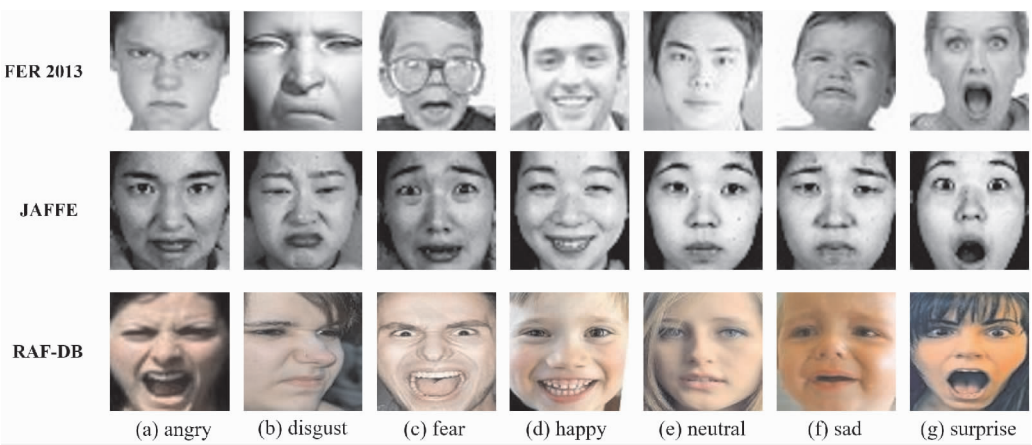


图 5 部分数据集样本

Fig. 5 A partial sample of datasets

1) FER2013 数据集。该数据集样本数量较大,包含真实人脸表情图像和卡通人物表情图像,具有丰富的样本。该数据集共包含 35 887 张表情图像。

2) JAFFE 数据集。该数据集的图像来源于实验室环境中的 10 名日本女性,她们按照指示做出了多种不同种类的表情,所采集的表情图像较为清晰,具有很高的识别率。该数据集共包含 213 张图像,每个人展示 7 种表情。

3) RAF-DB 数据集。该数据集共有 29 672 张人脸表情图像,图像质量相对较高,收集的人脸表情都来源于自然场景,因此表情更自然,更加接近真实人脸的表情。本文的实验主要研究该数据集中的 7 类基本表情图像。

2.2 实验环境

本文实验基于 PyTorch 1.7.0 框架进行训练和测试。实验环境如下:Ubuntu18.04,Cuda 版本为 11.0,显卡为 NVIDIA RTX 3080 Ti(12 GiB)。在实验中,首先将人脸表情图像的大小缩放到 224 × 224,并进行数据增强操作,包括随机旋转、图像对比度增强等。在训练过程中,批量大小设为 32,损失函数选用交叉熵损失函数,并使用 AdamW 优化器进行模型的反向传播优化,同时设置权重衰减为 5E-2,以帮助控制模型的复杂度并提高泛化性能。

2.3 评价标准

在图像分类任务中,通常使用准确率、混淆矩阵和召回率等指标来评估分类模型的性能。对于本文的人脸表情识别任务,为了更好地评估模型

和每个表情类别的识别精度,可以采用准确率和混淆矩阵作为评价标准,准确率(Accuracy,式中简记 R_{ACC})的计算公式为

$$R_{ACC}=\frac{N_{TP}+N_{TN}}{N_{TP}+N_{FP}+N_{TN}+N_{FN}} \tag{9}$$

式中:N_{TP}和 N_{TN}代表模型正确预测的正例和负例的样本数量;N_{FP}和 N_{FN}分别代表模型错误预测的正例和负例的样本数量。通过混淆矩阵,可以更直观地展示每个表情类别的预测结果。混淆矩阵中的对角线元素表示模型正确分类的样本数量,即真正例 N_{TP}和真负例 N_{TN}。对角线上的值越高,说明模型的分类效果越好。而非对角线上的元素表示模型错误分类的样本数量,即假正例 N_{FP}和假负例 N_{FN}。

2.4　CBAM 嵌入位置验证

为了验证 CBAM 混合注意力模块在不同 Stage 中对最终识别效果的影响,本文将 CBAM 模块嵌入到 Swin Transformer 模型的不同 Stage 中,并进行对比实验。由于不同 Stage 中的特征图尺寸和维度不同,CBAM 模块所处理的特征图信息也会有所差异,这可能就会对模型最终的识别效果产生影响。为了评估这种影响,本文在 3 个不同阶段分别嵌入了 CBAM 模块,以及在 3 个阶段中都嵌入了 CBAM 模块进行了对比实验。实验结果详见表 1。

在 3 个公共数据集(JAFFE、RAF-DB、FER-2013)和 1 个私有数据集上进行对比实验的结果表明,将混合注意力模块 CBAM 嵌入到 Stage3 中 Patch Merging 层所获得的实验效果最好,平均准确率达到了 80.54%,模型的参数量为 48.814×10^6。因此,本文选择在 Stage3 中嵌入混合注意力模块更具有一定的科学性及有效性。

表 1　在不同阶段将 CBAM 嵌入 Patch Merging 层中的准确性

Tab. 1　Accuracy of CBAM being placed in Patch Merging at different stages

Stage	准确率/%				平均准确率/%	参数量/M
	JAFFE	RAF-DB	FER2013	Private Dataset		
Stage2	95.01	86.24	73.11	58.95	78.32	48.801
Stage3	97.42	86.96	73.63	64.17	80.54	48.814
Stage4	96.01	84.97	73.32	63.39	79.42	48.869
Stage2 + Stage3 + Stage4	92.89	85.47	70.24	55.19	75.95	48.892

2.5　消融实验

为了验证在模型中嵌入 CBAM 混合注意力模块的有效性,本文进行了消融实验,分别在 JAFFE、RAF-DB、FER2013 以及 1 个私有数据集上进行了实验验证,对比了有无嵌入混合注意力模块对实验结果的影响,具体实验结果详见表 2。通过表 2 可以看出,嵌入混合注意力模块的模型在 3 个公共数据集和 1 个私有数据集上的识别准确率均有所提升。

表 2　嵌入混合注意力模块对实验结果的影响

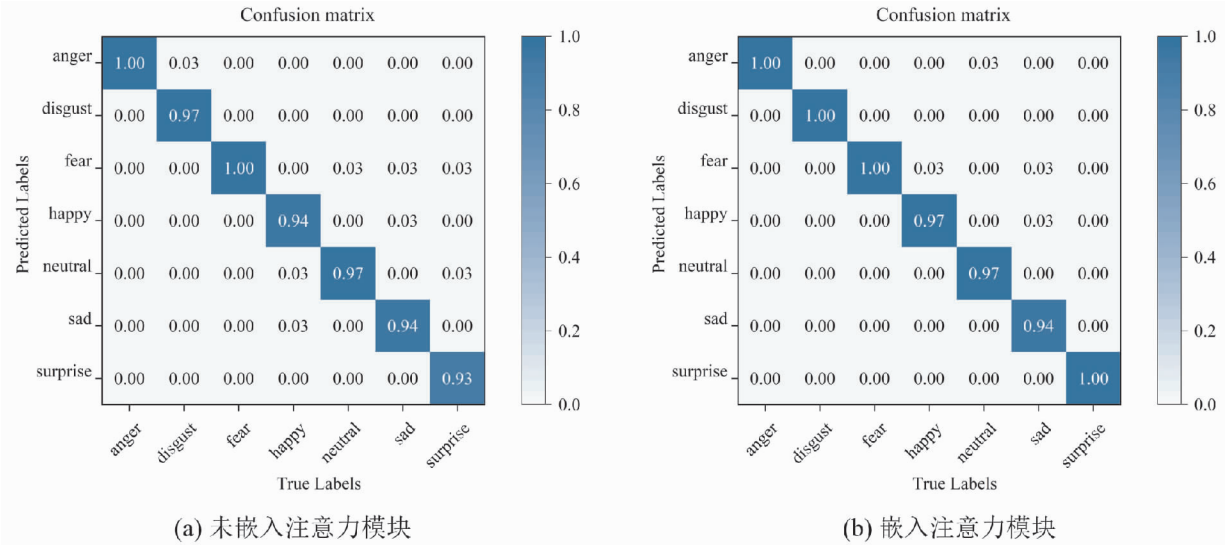
Tab. 2　Effect of embedding attention modules on experimental results

模型	准确率/%			
	JAFFE	RAF-DB	FER2013	Private Dataset
Swin Transformer	96.56	85.89	73.27	63.25
Swin + CBAM	97.42	86.96	73.63	64.17

相较于未嵌入混合注意力模块的情况,将 CBAM 混合注意力模块嵌入 Swin Transformer 中,能够有效提高模型对人脸表情的识别精度。图 6 展示了在 JAFFE 数据集上,有无嵌入 CBAM 混合注意力模块的混淆矩阵验证结果。从图 6 中能够观察到对于高兴、厌恶和惊讶等表情类别,模型的识别准确率都有所提升。

同时,本文在数据集规模较大的 FER2013 数据集上进行了预训练,并将训练好的预训练权重与在 ImageNet 数据集上的预训练权重进行了对比。结果表明,不同的预训练数据集会对模型的表情识别精度产生影响。在实验过程中,我们发

现使用 FER2013 数据集上的预训练权重可以有效地提高模型对表情识别的准确率,具体实验结果详见表 3。



(a) 未嵌入注意力模块 (b) 嵌入注意力模块

图 6 JAFFE 数据集上的混淆矩阵验证结果

Fig. 6 Confusion matrix validation results on JAFFE

表 3 预训练对实验结果的影响

Tab. 3 Effect of pre-training on experimental results

模型	准确率/%			
	pre-train	JAFFE	RAF-DB	Private Dataset
Swin + CBAM	ImageNet	97.42	86.96	64.17
Swin + CBAM	FER2013	98.28	87.01	66.53

2.6 方法比较

为了进一步验证本文方法的有效性,与其他多种网络模型进行了比较。表 4 展示了本文方法与其他模型在 RAF-DB 和 FER2013 数据集上的对比结果。在 RAF-DB 数据集上,本文方法与 RAN[23]、Twins[24]、POSTER[25]和 SPWFA-SE[26]等算法模型进行了比较;在 FER2013 数据集上,本文方法与 MoEffNet[27]、Efficient-CapsNet[28]、Auto-FERNet[29]和 Inception-V3[30]等模型进行了比较。通过实验对比,本文方法在 RAF-DB 和 FER2013 这 2 个公共数据集上的准确率明显优于表 4 中其他算法模型。

2.7 可视化实验结果

为了更直观地展示嵌入 CBAM 混合注意力模块后的效果,本文采用了 Grad-CAM[31]技术,用于生成分类网络中最后一层的热力图。图 7 展示了本文在 7 类不同表情上的热力图效果。热力图能够验证网络对图像区域的关注程度,颜色越鲜艳则意味着该区域的内容对于网络的识别越重要。这些可视化实验结果表明,嵌入 CBAM 混合注意力模块后,模型能够将注意力集中在表情特征的重点区域,从而更精准地识别人脸表情种类。

表 4 对比其他方法的识别率

Tab. 4 Comparison with the recognition rates of other methods

模型	准确率/%	
	RAF-DB	FER2013
RAN[23]	86.90	–
Twins[24]	86.15	–
POSTER[25]	86.03	–
SPWFA-SE[26]	86.31	–
MoEffNet[27]	–	71.02
Efficient-CapsNet[28]	–	72.94
Auto-FERNet[29]	–	73.10
Inception-V3[30]	–	73.09
本文方法	87.01	73.63

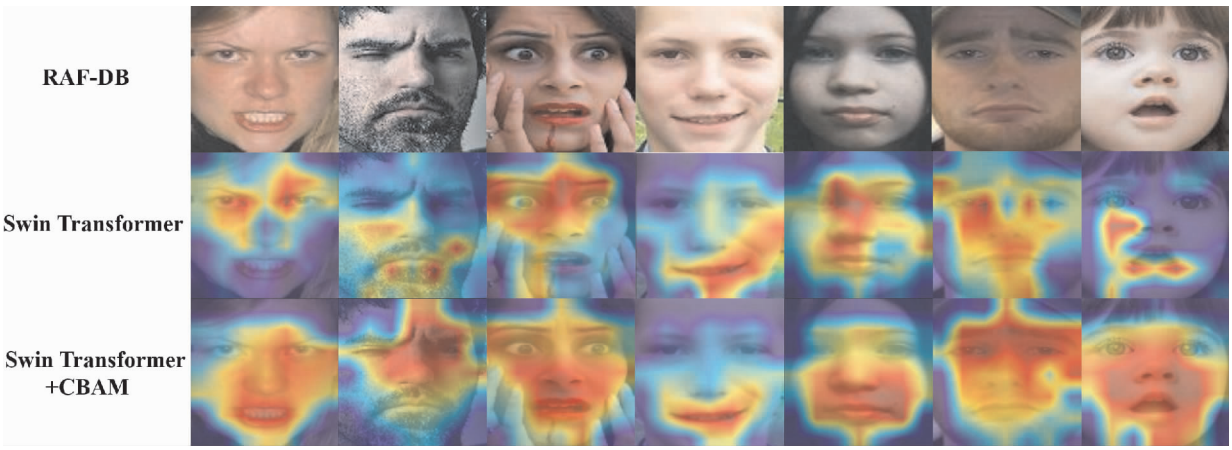


图 7　Grad-CAM 热力图对比

Fig. 7　Comparison of Grad-CAM activation maps

3　结语

针对人脸表情识别,本文提出了一种嵌入混合注意力机制的 Swin Transformer 人脸表情识别方法。该方法在模型的 Patch Merging 层中嵌入了 CBAM 混合注意力模块,并通过迁移学习方法进行训练。相较于传统卷积神经网络,Swin Transformer 能够更好地获取图像的全局语义信息。同时,CBAM 模块的嵌入能够使模型更多地关注局部的重要表情特征信息,并抑制无用信息的干扰,将有限的计算资源聚焦分配给权重较大的重要区域,从而加快模型的收敛速度并提高表情识别性能。实验结果表明,在模型的 Stage3 中嵌入 CBAM 混合注意力模块能够取得最佳效果。最后,本文所提出的方法在 FER2013、RAF-DB 和 JAFFE 数据集上分别获得了 73.63%、87.01% 和 98.28% 的准确率。在之后的研究中,可以考虑采用更轻量级结构的 Transformer 模型,以解决模型过大和参数量过多等问题。

参考文献

[1] 李珊,邓伟洪. 深度人脸表情识别研究进展[J]. 中国图象图形学报,2020,25(11):2306-2320.
LI S,DENG W H. Deep facial expression recognition: A survey[J]. Journal of Image and Graphics,2020,25(11):2306-2320.

[2] ADYAPADY R R, ANNAPPA B. A comprehensive review of facial expression recognition techniques[J]. Multimedia Systems, 2023, 29(1): 73-103.

[3] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[4] QI Y F, ZHOU C Y, CHEN Y X. NA-Resnet: Neighbor Block and optimized attention module for global-local feature extraction in facial expression recognition [J]. Multimedia Tools and Applications, 2023, 82(11): 16375-16393.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017: 6000-6010.

[6] MA T L, MAO M Y, ZHENG H H, et al. Oriented object detection with transformer[EB/OL]. (2021-06-06) [2023-09-20]. http://arxiv.org/abs/2106.03146.

[7] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03) [2023-09-20]. http://arxiv.org/abs/2010.11929.

[8] WANG W H, XIE E Z, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]// 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 548-558.

[9] WU H P, XIAO B, CODELLA N, et al. CvT: Introducing convolutions to vision transformers[C]// 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 22-31.

[10] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows [C]// 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9992-10002.

[11] LIU C, HIROTA K, DAI Y P. Patch attention convo-

lutional vision transformer for facial expression recognition with occlusion[J]. Information Sciences, 2023, 619(C): 781-794.

[12] FENG H Q, HUANG W K, ZHANG D H, et al. Fine-tuning swin transformer and multiple weights optimality-seeking for facial expression recognition[J]. IEEE Access, 2023, 11: 9995-10003.

[13] CHEN X C, ZHENG X W, SUN K, et al. Self-supervised vision transformer-based few-shot learning for facial expression recognition[J]. Information Sciences, 2023, 634(C): 206-226.

[14] 祁宣豪,智敏. 图像处理中注意力机制综述[J]. 计算机科学与探索,2024,18(2):345-362.
QI X H,ZHI M. Review of attention mechanisms in image processing[J]. Journal of Frontiers of Computer Science and Technology,2024,18(2):345-362.

[15] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2. Montreal: ACM, 2015: 2017-2025.

[16] WANG Q L, WU B G, ZHU P F, et al. ECA-net: Efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11531-11539.

[17] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]// European Conference on Computer Vision (ECCV). Cham: Springer, 2018: 3-19.

[18] 冯晓毅,黄东,崔少星,等. 基于空时注意力网络的面部表情识别[J]. 西北大学学报(自然科学版),2020,50(3):319-327.
FENG X Y,HUANG D,CUI S X. Spatial-temporal attention network forfacial expression recognition[J]. Journal of Northwest University(Natural Science Edition). 2020,50(3):319-327.

[19] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132-7141.

[20] GOODFELLOW I J, ERHAN D, CARRIER P L, et al. Challenges in representation learning: A report on three machine learning contests[C]// The 20th International Conference on Neural Information Processing. Daegu: Springer, 2013:117-124.

[21] LYONS M, AKAMATSU S, KAMACHI M, et al. Coding facial expressions with Gabor wavelets[C]// Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. Nara: IEEE, 2002: 200-205.

[22] LI S, DENG W H, DU J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2584-2593.

[23] WANG K, PENG X J, YANG J F, et al. Region attention networks for pose and occlusion robust facial expression recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 4057-4069.

[24] CHU X X, TIAN Z, WANG Y Q, et al. Twins: Revisiting the design of spatial attention in vision transformers[EB/OL]. (2021-09-30)[2023-09-20]. http://arxiv.org/abs/2104.13840.

[25] ZHENG C, MENDIETA M, CHEN C. POSTER: A pyramid cross-fusion transformer network for facial expression recognition[C]//2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Paris:IEEE, 2023: 3138-3147.

[26] LI Y J, LU G M, LI J X, et al. Facial expression recognition in the wild using multi-level features and attention mechanisms[J]. IEEE Transactions on Affective Computing, 2023,14(1):451-462.

[27] SINGH R, SHARMA H, MEHTA N K, et al. Efficientnet for human fer using transfer learning[J]. ICTACT Journal on Soft Computing, 2023,13(1): 2792-2797.

[28] WANG K X, HE R X, WANG S, et al. The Efficient-CapsNet model for facial expression recognition[J]. Applied Intelligence, 2023,53(13): 16367-16380.

[29] LI S Q, LI W, WEN S P, et al. Auto-FERNet: A facial expression recognition network with architecture search[J]. IEEE Transactions on Network Science and Engineering, 2021, 8(3): 2213-2222.

[30] MEENA G, MOHBEY K K, KUMAR S. Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach[J]. International Journal of Information Management Data Insights, 2023, 3(1): 100174.

[31] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 618-626.

(编 辑 李 静)