



基于多尺度分层双线性池化网络的细粒度表情识别模型

苏志明, 王 烈, 蓝峥杰

(广西大学 计算机与电子信息学院, 南宁 530004)

摘 要: 人脸表情细微的类间差异和显著的类内变化增加了人脸表情识别难度。构建一个基于多尺度双线性池化神经网络的识别模型。设计3种不同尺度网络提取人脸表情全局特征,并引入分层双线性池化层,集成多个同一网络及不同网络的多尺度跨层双线性特征以捕获不同层级间的部分特征关系,从而增强模型对面表情细微特征的特征及判别能力。同时,使用逐层反卷积融合多层特征信息,解决神经网络通过多层卷积层、池化层提取特征时丢失部分关键特征的问题。实验结果表明,该模型在FER2013和CK+公开数据集上的识别率分别为73.725%、98.28%,优于SLPM、CL、JNS等人脸表情识别模型。

关键词: 卷积神经网络;细粒度表情识别;多尺度网络;分层双线性池化;多层特征融合

开放科学(资源服务)标志码(OSID):



中文引用格式: 苏志明,王烈,蓝峥杰.基于多尺度分层双线性池化网络的细粒度表情识别模型[J].计算机工程,2021,47(12):299-307,315.

英文引用格式: SU Z M, WANG L, LAN Z J. Fine-grained expression recognition model based on multi-scale hierarchical bilinear pooling network[J]. Computer Engineering, 2021, 47(12): 299-307, 315.

Fine-Grained Expression Recognition Model Based on Multi-Scale Hierarchical Bilinear Pooling Network

SU Zhiming, WANG Lie, LAN Zhengjie

(School of Computer and Electronic Information, Guangxi University, Nanning 530004, China)

[Abstract] Facial expressions are characterized by subtle differences between expression classes and significant changes within a class, which increases the difficulty of expression recognition. To address the problem, a neural network model is proposed based on multi-scale bilinear pooling. The global features of facial expressions are extracted by using three networks with different scales. Then a hierarchical bilinear pooling layer is introduced, and multi-scale cross-layer bilinear features of the same network and different networks are integrated to capture some feature relationships between different levels, thus enhancing the ability of the model to represent and recognize subtle features of facial expressions. Multilayer feature information is fused by layer deconvolution, so the loss of key features that occurs when the neural network extracts features through multiple convolution layers and the pooling layer is solved. The experimental results show that the proposed model achieves a 73.725% recognition accuracy on FER2013 and 98.82% on CK+public data sets, outperforming SPLM, CL, JNS and other facial expression recognition algorithms.

[Key words] Convolution Neural Network (CNN); fine-grained expression recognition; multiple-scale network; hierarchical bilinear pooling; multilayer feature fusion

DOI: 10.19678/j.issn.1000-3428.0060133

0 概述

人脸表情识别(Facial Expression Recognition, FER)旨在通过识别人脸表情使机器能够理解人的

内心感受。该技术在远程教育、辅助医疗、安全驾驶、人机交互、公共安全等多个领域具有广泛应用^[1],相关人脸表情识别研究已成为人工智能主要研究热点之一。

基金项目: 广西自然科学基金(2013GXNSFAA0019339)。

作者简介: 苏志明(1994—),男,硕士研究生,主研方向为深度学习;王 烈,教授;蓝峥杰,硕士研究生。

收稿日期: 2020-11-30 **修回日期:** 2020-12-31 **E-mail:** lwang@gxu.edu.cn

早期的表情识别基于传统特征提取方法,大体上可分成3种:基于线性变换,如主成分分析法^[2](Principal Component Analysis, PCA);基于纹理特征,如局部二值模式法(Local Binary Pattern, LBP)^[3];基于几何,如主动形状法(Active Shape Models, ASM)^[4]和主动外观模型(Active Appearance Model, AAM)^[5]。但这些方法存在特征提取不充分导致识别率低的问题。由于深度学习可以从端到端地学习更多差异化的面部表情特征,且与传统方法相比具有更高识别率,因此研究人员致力于将深度学习应用于面部表情识别,基于深度学习的人脸表情识别算法也层出不穷。文献[6]改进了 AlexNet,引入多尺度卷积提取多尺度特征和利用全局平均池化将低层特征降维跨连到全连接层分类,在 CK+人脸表情数据集的准确率达到 94.25%。文献[7]提出利用小尺度核卷积代替大尺度核卷积的神经网络模型,在 FER2013 数据集上取得了 73.39% 的识别率。LIU 等^[8]将课程学习策略应用到卷积神经网络训练阶段,在 FER2013 数据集上达到 72.11% 的识别准确率。LI 等^[9]改进了经典模型 LeNet-5,通过将池化层的特征跨连到全连层,有效融合了高低层特征分类,在 JAFFE 和 CK+ 这 2 个公开人脸表情数据集的识别率分别达到 94.37%、83.74%,但改进后的模型人脸表情识别率仍然有待提高。ZHANG 等^[10]提出一种基于注意力分层双线性池化残差网络的表情识别方法。该方法在 ResNet-50 的框架基础上嵌入有效通道注意力机制,并引入分层双线性池化网络以交互同一网络不同层级间的特征,取得了不错的人脸表情分类效果。但该方法仅交互同一网络中来自 3 个不同层级间的特征,缺乏不同网络不同跨层的多尺度特征表达,因此

面部表情细微特征表征能力有待进一步提升。

目前,国内外在表情识别领域已取得较大进展,但人脸表情识别算法仍面临众多挑战。总体而言,人脸表情识别研究仍需要解决复杂环境下的表情识别、模型层间交互和模型多层特征融合等问题。

本文设计并训练应用于人脸表情分类的 3 种粗细尺度网络,并构建一个基于多尺度双线性池化卷积神经网络的识别模型。通过分层双线性池化捕捉不同网络的多尺度特征,挖掘神经网络对嘴巴、眉毛、眼睛等面部表情关键区域细微变化的辨别力,同时提出一种多层信息融合的方法获取有用的低频信息,从而提高人脸表情分类性能。

1 卷积神经网络

1.1 神经网络结构

本文提出多尺度分层双线性池化网络(Multi-scale Hierarchical Bilinear Pooling Network, MHBP)模型如图 1 所示。3 列网络分别使用卷积尺度核为 3、5、7 的不同粗细尺度网络,以提取更为精细的人脸表情特征。每列网络有 9 个卷积层、3 个最大池化层。3 列网络有共同的人脸表情图像输入,通过分层双线性池化网络交互 3 列网络最后同一深度位置的后 3 个卷积层的特征图,集成同一网络以及不同网络的不同跨层特征,捕获不同层级间的部分联系,以便于后续人脸表情特征分类。因集成的特征维度过高,冗余特征多,直接用以分类并不适用,所以需添加 2 层全连接层过滤特征以实现表情分类。图中仅给出了网络最后一层卷积层输出的特征图跨层交互简略示意图,忽略了其他层,具体交互的机制见 1.3.2 节。

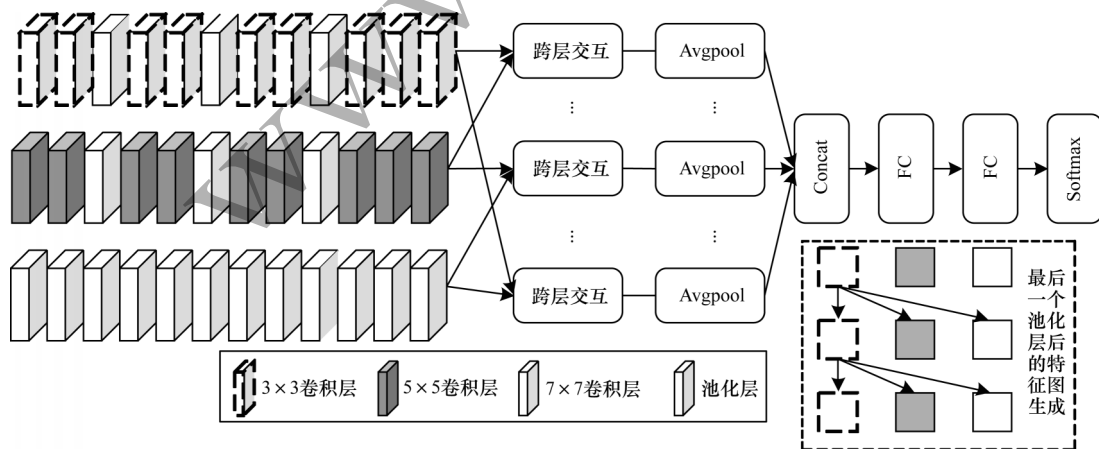


图1 多尺度分层双线性池化网络结构

Fig.1 Multi-scale hierarchical bilinear pooling network structure

为更好地利用主干网络的不同尺度特征,本文提出了多尺度注意力交互模块,如图2所示。

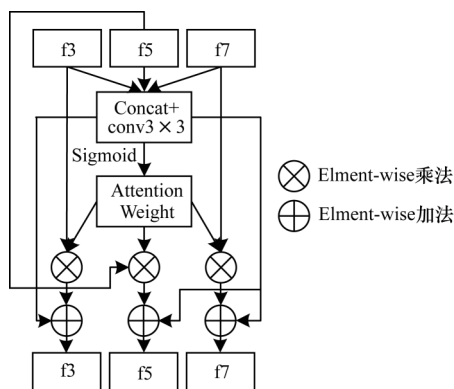


图2 多尺度注意力交互模块

Fig.2 Multi-scale attention interaction module

该模块先将3个粗细尺度网络的特征图 f_3 、 f_5 和 f_7 经 3×3 融合,一支支经 Sigmoid 函数激活生成特征权重后分别与特征图 f_3 、 f_5 和 f_7 元素相乘得到重新标定的特征图,最后分别与另一分支经 PReLU 函数^[11]激活后的融合特征元素相加得到最终的各自输出。该模块可根据反向传播自我更新学习,自动选择各支路需要融合的多尺度特征。该模块在本网络训练测试时加入位置为网络前3个池化层前的2个卷积层之间,共添加3个。

1.2 模型参数配置

MHBP 模型的具体参数配置如表1和表2所示。表1和表2省略了主干网络多尺度特征融合模块。根据输出特征图分辨率的不同,可分为4个阶段,每个网络的前3阶段均有2个卷积,后一个阶段有3个卷积。3个网络拥有共同输入为 48×48 大小的人脸表情图像灰度图。 3×3 、 5×5 、 7×7 网络的卷积核尺寸分别为 3×3 、 5×5 、 7×7 ,步长和填充均为1,卷积核个数均为32。Maxpool中的 $(3, 2, 1)$ 表示滤波器尺寸为 3×3 ,步长为2,填充为1。输出特征图为 $h \times w \times c$,其中: h 、 w 分别为特征图的高和宽; c 为特征图数量即卷积核的数量。通过双线性池化集成了18组512维人脸表情双线性特征向量。为避免模型过拟合,通过提高模型的泛化能力进一步提高人脸图像分类准确率。在每个池化层后加入 Dropout 网络,丢弃概率为0.1。同样地,在汇合分类阶段,2层全连接层后均加入 BN (Batch Normalization) 和 Dropout 网络,其中 Dropout 网络丢弃的概率为0.5,目的是加强网络挖掘隐藏特征的能力,从而提升模型性能。第1个全连接层与1024个神经元完全连接,而第2个全连接层与512个神经元完全连接。输出层是由7个神经元组成的 Softmax 层,用以预测7种表情的输出。

表1 MHBP 模型参数配置1

Table 1 Parameter configuration of MHBP model 1

阶段	类型	3×3 网络	5×5 网络	7×7 网络	输出
一阶段	Conv1_0	3×3	5×5	7×7	48×48×32
	Conv1_1	3×3	5×5	7×7	48×48×32
	Maxpool1	(3, 2, 1)	(3, 2, 1)	(3, 2, 1)	24×24×32
二阶段	Conv2_0	3×3	5×5	7×7	24×24×32
	Conv2_1	3×3	5×5	7×7	24×24×32
	Maxpool2	(3, 2, 1)	(3, 2, 1)	(3, 2, 1)	12×12×32
三阶段	Conv3_0	3×3	5×5	7×7	12×12×32
	Conv3_1	3×3	5×5	7×7	12×12×32
	Maxpool3	(3, 2, 1)	(3, 2, 1)	(3, 2, 1)	6×6×32
四阶段	Conv4_0	3×3	5×5	7×7	6×6×32
	Conv4_1	3×3	5×5	7×7	6×6×32
	Conv4_2	3×3	5×5	7×7	6×6×32

表2 MHBP 模型参数配置2

Table 2 Parameter configuration of MHBP model 2

阶段	类型	分层双线性池化层	输出
汇合分类	FC1	输入 512×18	$1 \times 1 \times 1024$
	FC2	输入 $1 \times 1 \times 1024$	$1 \times 1 \times 512$
	Output	输入 $1 \times 1 \times 512$	$1 \times 1 \times 7$

1.3 分层双线性池化

分层双线性池化^[12]对局部成对特征的交互式建模已被证明是解决细粒度识别问题的有力工具。为获得更好的表情特征表达,提出了一种在细粒度识别任务背景下的表情挖掘方法。通过在不同跨层中交互建模,融合同尺度网络 and 不同尺度网络不同卷积层的中间层特征,实现表情识别。

1.3.1 工作原理

分层双线性池化模型基于分解双线性池化模型^[13]构建。因子分解双线性池化的过程为:将经粗细尺度主干网络提取的特征图记为 $X \in \mathbb{R}^{h \times w \times c}$,其中 h 、 w 、 c 分别为特征图的高度、宽度、通道数。令 $\mathbf{x} = [x_1, x_2, \dots, x_c]^T$ 为 X 上的一个空间位置 c 维描述符。双线性模型的定义如下:

$$z_i = \mathbf{x}^T \mathbf{W}_i \mathbf{x} \quad (1)$$

其中: z_i 为双线性模型的输出; $\mathbf{W}_i \in \mathbb{R}^{c \times c}$ 为投影矩阵。双线性模型可将 \mathbf{W}_i 分解为低阶外积运算得到输出特征:

$$z = P^T (U^T \mathbf{x} \circ V^T \mathbf{x}) \quad (2)$$

其中: $P \in \mathbb{R}^{d \times o}$ 为分类矩阵; d 为决定嵌入维度的超参数; o 为图像分类类别总数; $U \in \mathbb{R}^{c \times d}$ 和 $V \in \mathbb{R}^{c \times d}$ 为从 c 维特征向量中获得 d 维池化特征向量的投影矩阵; \circ 为哈达玛积。

双线性池化捕获成对表征关系,是细粒度识别的重要技术。因为判断人脸表情属性的重点区域只

有眼睛、眉毛、鼻子、嘴角附近区域,属于精细工作,因此可借助双线性池化完成。但若只关注单一卷积层,完全忽略信息的跨层交互,将导致人脸表情分类效果不佳。这是因为单个卷积层的激活不完整,每个表情均有多个属性,例如嘴的形状、嘴角的弧度等,而这些对表情的细微变化至关重要。不同卷积层之间的层间特征相互作用能够帮助捕获细微表情的区别性特征。利用跨层双线性池集成更多中间卷积层,进一步增强表情特征的表征能力。通过独立的线性映射(1×1 卷积)将来自不同卷积层的特征扩展到多维空间。集成不同跨层表情特征的输出表达式为:

$$O_o = P^T \text{concat}(U^T x \circ V^T y, U^T x \circ S^T z, V^T y \circ S^T z, \dots) \quad (3)$$

其中: $U \in \mathbb{R}^{c \times d}$, $V \in \mathbb{R}^{c \times d}$, $S \in \mathbb{R}^{c \times d}$, \dots 分别为需要的交互跨层卷积层特征 x, y, z, \dots 的投影矩阵。将聚合的来自不同跨层的表情特征输入至全连接层和Softmax分类中,Softmax分类损失函数定义如下:

$$L_{\text{softmax}} = -\frac{1}{m} \sum_{i=1}^m \log_a \frac{\exp[W_{y_i}^T x_i + b_{y_i}]}{\sum_{j=1}^n \exp[W_j^T x_i + b_j]} \quad (4)$$

其中: m 为样本数; n 为总类别数,因本文需识别7种面部表情,故取值为7; x 为分类前全连接层的输入特征向量; b 为偏置量; $W_{y_i}^T x_i + b_{y_i}$ 表示第 i 个样本全连接层输出矩阵中预测类别为真实类别的目标判定。

1.3.2 交互机制

为捕获不同尺度层间特征关系,本文分层双线性池化跨层融合了来自同一网络及不同网络的不同卷积层特征,需要融合的层为经PReLU函数激活的不同尺度网络最后3层卷积层(Conv4_0, Conv4_1, Conv4_2)见表3。PReLU4_0 $_j$, $j=0, 1, 2$ 。其中: j 为第几列网络标号;0为 3×3 网络;1为 5×5 网络;2为 7×7 网络。

表3 双线性交互层列表

Table 3 Bilinear interaction layer list

卷积层	3×3网络	5×5网络	7×7网络	通道数
Conv4_0	PReLU4_0_0	PReLU4_0_1	PReLU4_0_2	32
Conv4_1	PReLU4_1_0	PReLU4_1_1	PReLU4_1_2	32
Conv4_2	PReLU4_2_0	PReLU4_2_1	PReLU4_2_2	32

现将双线性汇合不同跨层特征分为以下3类:

1) 同一网络不同层级特征。将3种网络最后3个卷积层经PReLU函数激活后的特征图分别在同一网络内两两交互,得到共9组双线性特征。

具体交互的双线性特征计算表达式如下:

$$(\text{PReLU4_0_0} * \text{PReLU4_1_0} + \text{PReLU4_0_0} * \text{PReLU4_2_0} + \text{PReLU4_1_0} * \text{PReLU4_2_0}) + (\text{PReLU4_0_1} * \text{PReLU4_1_1} + \text{PReLU4_0_1} * \text{PReLU4_2_1} + \text{PReLU4_1_1} * \text{PReLU4_2_1}) + (\text{PReLU4_0_2} * \text{PReLU4_1_2} + \text{PReLU4_0_2} * \text{PReLU4_2_2} + \text{PReLU4_1_2} * \text{PReLU4_2_2}) \quad (5)$$

图3所示为同一网络不同层特征交互示意图,其中,每列特征分别对应 3×3 、 5×5 、 7×7 网络的最后3个卷积层经激活函数激活后的输出特征每个虚线框代表一组特征两两交互。

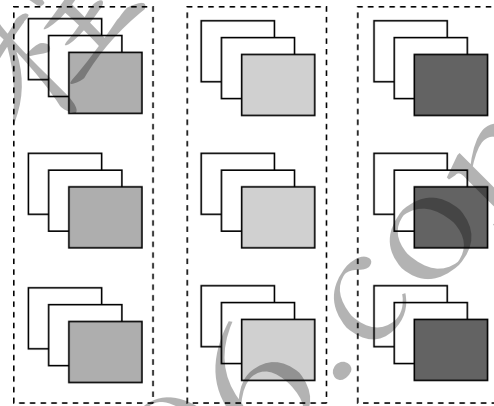


图3 同一网络不同层的特征交互

Fig.3 Feature interaction between different layers of the same network

2) 不同网络不同层级特征。为使不同尺度网络的不同层特征交互,增加了2个限制条件:

(1) 每个网络最后一个卷积层必须参与交互。这是因为一方面目前主流的神经网络分类模型通过最后一层卷积提取的特征直接展平为一维向量,或者先通过全局平均池进行降维,然后平铺为一维向量分类。另一方面,神经网络的最后一层通常包含输入图像的高频和全局特征信息,适合分类。

(2) 不同网络的不同层存在互斥。以不同网络不同层做为一组,分为3组,每组有3个层两两交互,共得9组双线性交互特征。

具体交互的双线性特征计算表达式如下:

$$(\text{PReLU4_0_0} * \text{PReLU4_2_1} + \text{PReLU4_0_0} * \text{PReLU4_1_2} + \text{PReLU4_2_1} * \text{PReLU4_1_2}) + (\text{PReLU4_1_0} * \text{PReLU4_0_2} + \text{PReLU4_1_0} * \text{PReLU4_2_2} + \text{PReLU4_0_2} * \text{PReLU4_2_2}) + (\text{PReLU4_2_0} * \text{PReLU4_1_1} + \text{PReLU4_2_0} * \text{PReLU4_0_2} + \text{PReLU4_1_1} * \text{PReLU4_0_2}) \quad (6)$$

如图4所示,3种不同颜色的箭头指向的不同层级特征图为3组需要交互的特征。其中,每种颜色箭头包函3个不同网络不同层的特征。

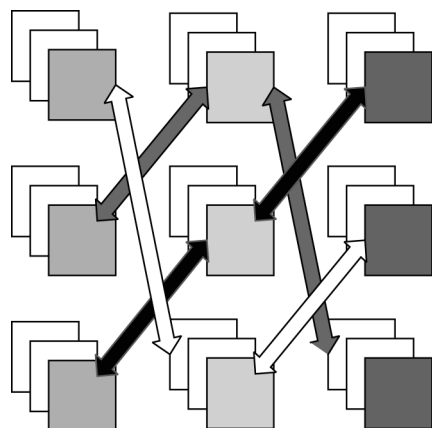


图4 不同网络不同层的特征交互

Fig.4 Feature interaction between different layers of different networks

3)不同网络同一深度位置特征。以同一深度层做为一组,可分为3组,每组内有3个层两两交互,共得9组双线性交互特征。具体交互的双线性特征计算表达式如下:

$$\begin{aligned} & (\text{PReLU4_0_0} * \text{PReLU4_0_1} + \text{PReLU4_0_0} * \text{PReLU4_0_2} + \\ & \text{PReLU4_0_1} * \text{PReLU4_0_2}) + (\text{PReLU4_1_0} * \text{PReLU4_1_1} + \\ & \text{PReLU4_1_0} * \text{PReLU4_1_2} + \text{PReLU4_1_1} * \text{PReLU4_1_2}) + (\text{PReLU4_2_0} * \text{PReLU4_2_1} + \\ & \text{PReLU4_2_0} * \text{PReLU4_2_2} + \text{PReLU4_2_1} * \text{PReLU4_2_2}) \end{aligned} \quad (7)$$

不同网络同一深度特征交互示意图如图5所示。

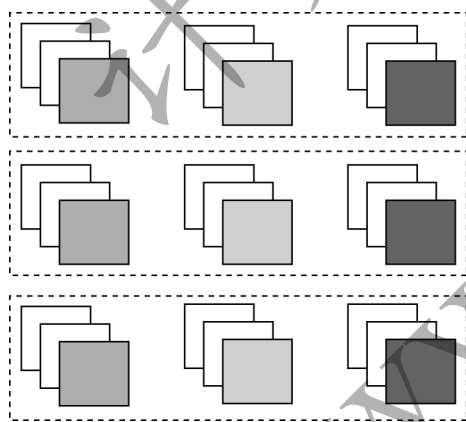


图5 不同网络同一深度的特征交互

Fig.5 Feature interaction in the same depth of different networks

1.4 多层信息融合

当卷积神经网络向前传播时,通过逐层卷积获得高频信息,将最后一层提取的特征输入到全连接层并进行分类。逐层过滤将丢失一些低频特征信息,如纹理、边缘等细节信息,导致信息无法得到充分利用。为获取有用的低频信息,从而提高人脸表情图像的识别率,本文提出一种多层信息融合(Multi-layer Information Fusion, MIF)方法。该方法

通过反卷积将当前卷积层输出的激活值转换为新的激活值,并逐层融合、逐层降维,最后将其输入全连接层分类,如图6所示。

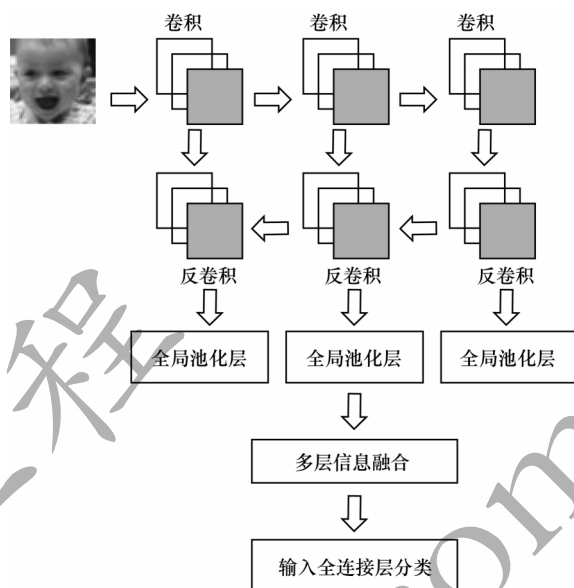


图6 多层信息融合

Fig.6 Multi-layer information fusion

若将没池化前的卷积层记为一个阶段,则MHBP网络共有3个同一深度位置的池化层,故可将卷积分为4个阶段。图6中的卷积为MHBP网络一个阶段的卷积,前3个阶段的每个阶段均有6个卷积层,最后一个阶段共有9个卷积层。多卷积层特征具体融合的过程如下:

步骤1 将最后一个阶段 n 的9个卷积层输出的特征图激活值通过`torch.cat`拼接,通过 1×1 卷积融合降维为上一个阶段所有卷积层特征图的维数(这里取 $32 \times 6 = 192$),并通过BN和PReLU函数激活得到激活值。接着将其输入到反卷积层得到特征图激活值。

步骤2 将步骤1中得到的激活值和拼接后上一阶段 $n-1$ 的所有卷积层的激活值做加性融合。融合后的特征图 $c \times h \times w$ 共有2分支操作,一分支通过全局平均池化降维得到压缩特征图 $c \times 1 \times 1$,另一分支继续通过反卷积得到特征图 $c \times h \times w$ 激活值。降维的目的是减少参数,加快网络运行。

步骤3 执行步骤1一次,重复步骤2两次可得到3组降维特征图,将其拼接融合后展开为一维向量,则共有 $32 \times 6 \times 3 = 576$ 维表情特征向量,将其添加到MHBP网络的全连接层进行人脸表情分类。

2 实验

2.1 实验环境

FER2013实验在型号为GTX2080Ti的Pytorch框架上进行, `batch_size=128`,初始学习效率为0.01,

60个epoch后每10个epoch衰减0.9倍。模型优化器为SGD,动量为0.9,权重衰减为0.001。为提高模型的泛化能力,引入一种数据增强方式Mixup^[14],该方式主要用于图像分类,可用来提高模型的表情识别率。从训练样本中随机抽取2个样本进行简单随机加权求和,样本的标签也对应于加权求和。通过加权求和,计算预测结果与标签之间的损失,并用逆导数更新参数,计算式如式(8)和式(9)所示:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \quad (8)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \quad (9)$$

其中: x_i 和 x_j 为原始图像输入向量; y_i 和 y_j 为one-hot标签编码。其数据增强方式与文献[15]类似,区别是增加了随机旋转,角度为0.5。计算数据集的均值和方差后,再归一化输入到网络训练和测试。使用Xavier系统进行初始化,训练时将表情图片尺寸预处理为52×52,再随机剪切为48×48,训练和测试时采用TenCrop方法,将图像沿左上角、右上角、左下角、右下角、中心剪切并水平翻转,取人脸表情识别率的平均值做为模型最终表情分类准确率。

2.2 数据集

FER2013^[16]数据集共有35 888张人脸表情图像,其中训练样本28 709张,公开测试样本和私有测试样本各3 589张。采用私有测试样本测试。FER2013数据集由Python爬虫获得,存在人脸角度、遮挡、光照条件、头部姿态变化、噪声等复杂环境,满足本文研究要求。

CK+数据集^[17]是国内外研究人员常用的面部表情识别研究课题的基础数据库,具有较高的认可度,共593个图像序列,其中带标签的表情序列有327个,从每个序列中提取最后3个帧,共981张。CK+实验去掉了中性表情,取剩下的生气、厌恶、害怕、高兴伤心、惊讶、蔑视等表情做为实验数据。2个数据集人脸表情示意图如图7所示。

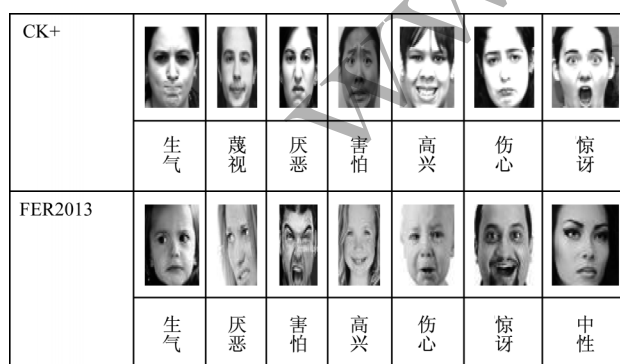


图7 不同数据集人脸表情

Fig.7 Facial expressions in different data sets

2.3 实验结果与分析

实验1 不同方法的有效性验证。不同方法的表情识别率如表4所示。前3个方法实验网络架构的池化层前卷积只有一个,网络使用2×2池化且池化后没有加入dropout网络,在FER2013人脸表情公开数据集上的识别率仅为72.081%,而使用3×3重叠池化且加入dropout后,识别率为72.75%,提升了0.669个百分点。这说明dropout在防止模型过拟合的同时还能增强网络的学习能力。延长训练次数识别率有所提高,说明网络还没有完全收敛。当将每个最大池化前的卷积添加到2个时,网络的表情特征提取能力得到增强,表情识别率达到了73.725%,提升了0.724个百分点。由于硬件资源的限制,本文算法没有继续通过增加卷积层数来探索网络性能。

表4 不同方法的表情识别率对比

Table 4 Comparison of expression recognition rates of different methods %

方法	准确率
2×2池化,无dropout	72.081
3×3池化 dropout(0.1)	72.750
延长训练 epoch 至 700	73.001
添加卷积层	73.725

实验2 不同层的特征交互有效性验证。实验2在实验1的基础上进行,将同一网络不同层级的特征交互记为 x_1 ,将不同网络不同层级特征交互记为 x_2 ,将不同网络同一深度位置特征交互记为 x_3 。在FER2013数据集上通过分层双线性池化集成不同跨层的双线性特征对比实验如表5所示。当集成同一网络不同层级特征和不同网络不同层级特征时,表情识别率最高为73.725%,说明模型捕获了不同跨层间特征的部分联系,也说明了多尺度双线性池化的有效性。模型集成特征 x_1 、 x_2 与单集成特征 x_1 ,集成特征 x_1 、 x_2 、 x_3 相比,表情识别率分别提高了0.494、0.278个百分点。通过双线性池化集成过多的不同跨层特征容易导致冗余特征过多,不利于面部表情分类。

表5 FER2013数据集不同层特征交互结果

Table 5 Interaction results of different layer features of FER2013 data set %

特征维度	准确率
x_1	73.231
$x_1 + x_2$	73.725
$x_1 + x_2 + x_3$	73.447

实验 3 不同高维空间的模型性能分析。将不同层次的特征升级到高维空间后采用分层双线性池的方法提取双线性特征,将对人脸表情识别的准确性有一定的影响。FER2013 测试集的准确率随不同维度空间变化的曲线如图 8 所示。随着维数的增加,人脸表情识别的准确率逐渐提高(不同维度空间中 0~200 的低点在误差范围内),当维数为 400~600 时达到最大值,之后随着维数的增加识别率逐渐降低。

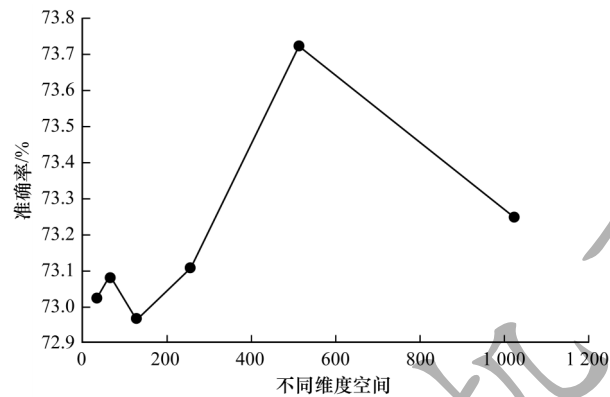


图 8 不同维度空间对识别率的影响
Fig.8 Influence of different dimension space on recognition rate

为更直观地看到不同维度空间的表情识别率,表 6 给出了具体的数值。当尺寸从初始值 32 上升到 64、128、256 时,精度呈现非线性提高但幅度有限。当维数升级到高维空间值 512 时,模型的识别率可高达 73.725%,当维数升级到 1 024 个超高维空间时,模型的识别率下降到 73.252%。这说明将维数提升到合适的高维空间可以提高模型的表情识别率,且当维数过低时,由于缺乏有效的人脸表情特征分类导致分类性能低下;但维数过高时,过多的冗余特征将影响分类性能。

表 6 不同维度空间的识别率	
Table 6 Recognition rate of different dimension spaces	%
维度	准确率
32	73.029
64	73.084
128	72.973
256	73.112
512	73.725
1 024	73.252

实验 4 不同通道数的模型性能分析。通过实验 3 发现,当把当前维度升维到高维空间时表情识别率有所提升,在升维 16 倍时取得最高值。实验 4 将研究不同通道数及升维 16 倍到高维空间时模型表情的识别性能,性能分析如表 7 所示。

表 7 不同通道数的模型性能分析
Table 7 Performance analysis of models with different channel numbers

通道数	维度	准确率/%	扩维	准确率/%
32	512	73.752	—	—
48	512	73.642	768	73.558
64	512	73.252	1 024	73.140

当维度保持 512 不变时,随着通道数增加,人脸表情识别准确率逐渐降低,说明基础通道数对模型性能影响不大。当通道数为 48、维度为 512 时,表情准确率为 73.642%,而将通道数扩展 16 倍变成 768 时,准确率有所下降,但幅度不大。同样地,将维度通道数 64 拓展到 1 024 特高空间时,准确率同样出现略有下降的趋势,原因可能是当维度扩大时,全连接层的输出参数不变。

实验 5 不同尺度网络的识别率比较。由图 9 曲线可知,网络模型大概在 400 个 epoch 时开始收敛,最后趋于稳定。在表情识别率方面,不同尺度网络 MHBP>5×5>7×7>3×3。其中 3、5、7 均为单一尺度网络,5×5 网络较 7×7 网络识别率高,这是因为网络最后 3 层卷积的特征图大小为 6×6,大尺度核卷积较小尺度核移动次数少,导致缺失重叠卷积部分特征。单一尺度 5 和 7 网络均比单一尺度 3 网络的识别率高,说明适当增大尺度核尺寸可提高识别率。MHBP 网络集成了不同的跨层多尺度特征,可捕获同一尺度及不同尺度的不同跨层特征联系,加大对表情细微特征表征对象的利用。因此,本文提出多尺度 MHBP 网络在人脸表情识别率方面优于其他单一尺度网络算法。

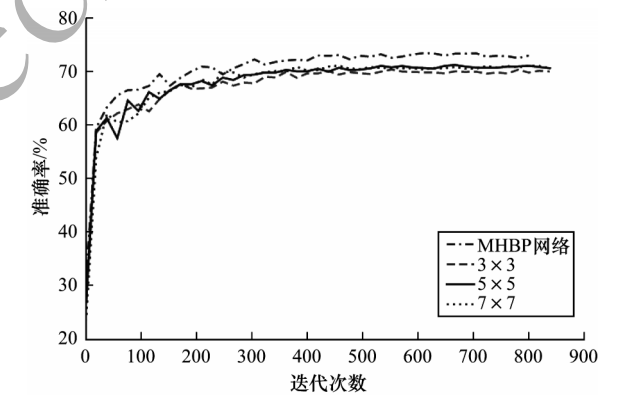


图 9 不同尺度网络对识别率的影响
Fig.9 Influence of different scale networks on recognition rate

如图 10 所示,4 种不同尺度网络的损失值均在 0.8~0.9 之间。MHBP 网络的损失值最小;5×5 网络和 7×7 网络次之,且两者损失值相近;3×3 网络损失值最大。此外,MHBP 网络损失值下降速度最快,收敛也快,且波动幅度不大,这进一步说明了 MHBP 网络集成不同跨层特征的优越性。

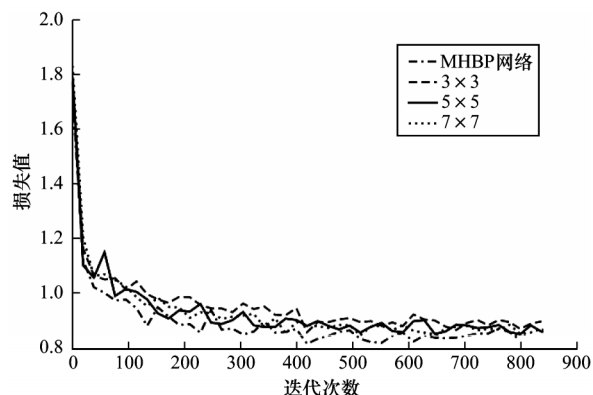


图10 不同尺度网络损失值比较

Fig.10 Comparison of network loss values at different scales

由表8可知, MHBP网络在FER2013公开人脸表情数据集的识别准确率比单一 3×3 、 5×5 、 7×7 网络分别提高了3.037、2.034、2.173个百分点。这说明单一尺度网络虽然集成了网络最后3层卷积层的特征,但缺乏多尺度表情特征信息,因此并不能准确地对表情做出判断。同时也证明端到端学习集成能够提高多跨层多尺度双线性人脸表情特征的辨识度,从而提高模型分类准确率。

表8 不同尺度网络的表情识别率对比

Table 8 Comparison of expression recognition rate of different scale networks %

不同尺度方法	准确率
3×3 网络	70.688
5×5 网络	71.691
7×7 网络	71.552
MHBP网络	73.725

实验6 多层信息融合的有效性分析。在MHBP网络添加多层信息融合FER2013的实验发现表情识别率并没有提升。这可能是受FER2013数据集存在标签错误、光照不一、头部姿势各异等复杂背景因素影响。为排除外界非人脸因素的影响,本节实验选择实验室环境的CK+表情数据集验证多层信息融合的有效性。将CK+数据集按照9:1划分为训练集和测试集,采用十折交叉训练,选择优化器为AdaBound^[18]。优化器参数设置如下:学习率为0.001,并在250 epoch后每2个epoch衰减0.9倍,amsbound参数设置为True。如表9所示,在MHBP网络中加入MHBP+MIF及MHBP+MIF多层信息融合后,7种表情的平均识别率提高了1个百分点。其中,悲伤和蔑视的识别率分别比MHBP网络提高了0.05和0.09个百分点,而其他表情的识别率基本相同。实验结果表明,通过反卷积对多层信息进行融合分类并恢复丢失的低频特征信息,能够提高表情识别率,此结果验证了多层信息融合的有效性。

表9 在CK+数据集上多层信息融合的性能分析

Table 9 Performance analysis of multi-layer information fusion on CK+ data set

算法	生气	厌恶	害怕	开心	伤心	惊讶	蔑视	识别率/%
MHBP	0.94	1.00	0.99	1.00	0.91	1.00	0.83	97.273
MHBP+MIF	0.95	1.00	1.00	1.00	0.96	0.99	0.92	98.283

2.4 不同算法的识别率比较

为更好地评估本文方法的有效性,选取几个较新的算法在CK+和FER2013数据集上做比较,结果如表10所示。TURAN等^[19]提出了一种新的更有效的流形学习方法—软局部保持映射(Soft Locality Preserving Map, SLPM),该方法旨在控制不同类的扩散水平,能有效降低特征向量的维数,并增强所提取网络对表情识别的区分能力。ZHOU等^[20]改善了Softmax层,使识别率得到了一定的提升。YANG等^[21]提出了一种基于残差表情的人脸表情识别方法。残差学习法用于生成模型中间层的残差,该残差包含输入表情图像任何生成模型的表情成分。实验结果证明了从模型中间层提取表情成分的有效性。TIAN等^[22]提出一种新的基于类别感知容差和孤立点抑制的Triplet损失函数。根据特征距离分布,对每一对表情,如快乐、恐惧等分配自适应容差参数,以剔除异常Triplet。SHAO等^[23]提出3种不同架构的新型卷积神经网络(CNN)模型:6个深度可分离的残差模块构成的浅网络,双分支并行提取传统LBP特征和深度学习特征的CNN模型和采用转移学习技术设计了预训练的CNN模型。实验结果具有竞争力和代表性。FENG等^[7]提出小尺度核网络,LIU等^[8]引入课程学习(Curriculum Learning, CL)到卷积神经网络,这些均使识别率得到了一定程度的提升。LAN等^[24]提出了联合过滤器响应正则化和批量正则化(Joint Normalization Strategy, JNS)训练模型,弥补了单一正则化的不足,提高了表情识别率。

表10 不同算法在CK+和FER2013数据集上的识别率对比

Table 10 Comparison of recognition rates of different algorithms on CK+ and FER2013 data sets %

算法	CK+识别率	算法	FER2013识别率
TURAN ^[19]	96.10	ZHOU ^[20]	70.91
YANG ^[21]	97.30	TIAN ^[22]	72.64
SHAO ^[23]	95.29	SHAO ^[23]	71.14
FENG ^[7]	94.65	FENG ^[7]	73.39
LIU ^[8]	98.18	LIU ^[8]	72.11
LAN ^[24]	94.90	LAN ^[24]	73.56
MHBP(ours)	98.28	MHBP(ours)	73.75

FENG等^[7]和LAN等^[24]对FER2013数据集的识别率较高,但对CK+数据集的识别率较低。相反,LIU等^[8]对CK+数据集的识别率较高,但对FER2013数据集的识别率较低。这说明以上算法均不具备普适性。然而,本文方法在CK+和FER2013这2个数据集上均取得了较好的效果。这是因为本文方法集成了多跨层的多尺度表情特征,能够捕捉表情深层次的微妙变化,且通过反卷积融合多层特征,恢复了表情图像逐层传递过程中的特征信息损耗,从而解决了模型层间交互以及多层特征融合的问题,因此更适用于表情分类。

3 结束语

本文设计3种不同尺度的网络作为提取人脸表情特征的主干网络,通过在网络中加入多尺度特征融合模块实现主干网络多尺度特征的自主融合,同时引入分层双线性池化网络集成同一网络及不同网络的跨层表情特征,以获取有区分度的细腻表情特征属性。在此基础上,进一步探究不同通道数及维度空间对所提MHBP算法的影响,提出一种多层信息融合方法。实验结果表明,多层特征融合方法能有效利用丢失的信息,提高表情分类精度,且基于多尺度双线性池化的网络能捕获具有明显辨识度的人脸表情特征,提高人脸表情识别率。下一步将设计轻量级神经网络,并利用金字塔池化改进多层特征融合的方式,以获得更高的运行效率和更好的识别效果。

参考文献

- [1] LI S, DENG W. Deep facial expression recognition: a survey [J]. *IEEE Transactions on Affective Computing*, 2018, 3(9): 1-10.
- [2] WOLD S, ESBENSEN K, GELADI P. Principal component analysis [J]. *Chemometrics and Intelligent Laboratory Systems*, 1987, 2(3): 37-52.
- [3] OJALA T, PIETIKAINEN M, HARWOOD D. A comparative study of texture measures with classification based on featured distributions [J]. *Pattern Recognition*, 1996, 29(1): 51-59.
- [4] COOTES T F, TAYLOR C J, COOPER D H, et al. Active shape models-their training and application [J]. *Computer Vision and Image Understanding*, 1995, 61(1): 38-59.
- [5] COOTES T F, EDWARDS G J, TAYLOR C J. Comparing active shape models with active appearance models [EB/OL]. [2020]. https://www.researchgate.net/publication/221259802_Comparing_Active_Shape_Models_with_Active_Appearance_Models.
- [6] 杨旭, 尚振宏. 基于改进 AlexNet 的人脸表情识别 [J]. *激光与光电子学进展*, 2020, 57(14): 243-250.
YANG X, SHANG Z H. Facial expression recognition based on improved AlexNet [J]. *Advances in Laser and Optoelectronics*, 2020, 57(14): 243-250. (in Chinese)
- [7] 冯杨. 基于小尺度核卷积的人脸表情识别研究 [D]. 武汉: 华中师范大学, 2020.
FENG Y. Facial expression recognition based on small-scale kernel convolution [D]. Wuhan: Central China Normal University, 2020. (in Chinese)
- [8] LIU X Q, ZHOU F Y. Improved curriculum learning using SSM for facial expression recognition [J]. *The Visual Computer*, 2020, 36(6): 1-15.
- [9] 李勇, 林小竹, 蒋梦莹. 基于跨连接 LeNet-5 网络的面部表情识别 [J]. *自动化学报*, 2018, 44(1): 176-182.
LI Y, LIN X Z, JIANG M Y. Facial expression recognition based on cross-connect LeNet-5 network [J]. *Acta Automatica Sinica*, 2018, 44(1): 176-182. (in Chinese)
- [10] 张爱梅, 徐杨. 注意力分层双线性池化残差网络的表情识别 [J]. *计算机工程与应用*, 2020, 56(23): 161-166.
ZHANG A M, XU Y. Attention hierarchical bilinear pooling residual network for expression recognition [J]. *Computer Engineering and Applications*, 2020, 56(23): 161-166. (in Chinese)
- [11] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification [C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*. Washington D. C., USA: IEEE Press, 2015: 1026-1034.
- [12] YU C J, ZHAO X Y, ZHENG Q, et al. Hierarchical bilinear pooling for fine-grained visual recognition [C]//*Proceedings of 2018 European Conference on Computer Vision*. Berlin, Germany: Springer, 2018: 574-589.
- [13] KIM J H, ON K W, LIM W, et al. Hadamard product for low-rank bilinear pooling [EB/OL]. [2020-10-29]. <https://arxiv.org/abs/1610.04325v1>.
- [14] ZHANG H Y, CISSE M, DAUPHIN Y N, et al. Mixup: beyond empirical risk minimization [EB/OL]. [2020-10-29]. <https://arxiv.org/abs/1710.09412>.
- [15] YANG S Z, GONG Z, YE K, et al. EdgeCNN: convolutional neural network classification model with small inputs for edge computing [EB/OL]. [2020-10-29]. https://www.researchgate.net/publication/336147679_EdgeCNN_Convolutional_Neural_Network_Classification_Model_with_small_inputs_for_Edge_Computing.
- [16] GOODFELLOW I J, ERHAN D, CARRIER P L, et al. Challenges in representation learning: a report on three machine learning contests [C]//*Proceedings of 2013 International Conference on Neural Information Processing*. Berlin, Germany: Springer, 2013: 117-124.
- [17] LUCEY P, COHN J F, KANADE T, et al. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression [C]//*Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-workshops*. Washington D. C., USA: IEEE Press, 2010: 94-101.

(下转第315页)

(上接第 307 页)

- [18] LUO L C, XIONG Y H, LIU Y, et al. Adaptive gradient methods with dynamic bound of learning rate[EB/OL]. [2020-10-29]. https://www.researchgate.net/publication/331371132_Adaptive_Gradient_Methods_with_Dynamic_Bound_of_Learning_Rate.
- [19] TURAN C, LAM K M, HE X. Soft locality preserving map for facial expression recognition[EB/OL]. [2020-10-29]. <https://arxiv.org/abs/1801.03754>.
- [20] ZHOU J C, JIA X, SHEN L L, et al. Improved softmax loss for deep learning-based face and expression recognition[J]. Cognitive Computation and Systems, 2019, 1(4): 97-102.
- [21] YANG H Y, CIFTCI U, YIN L J. Facial expression recognition by de-expression residue learning [C]// Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA; IEEE Press, 2018: 2168-2177.
- [22] TIAN Y, WEN Z W, XIE W C, et al. Outlier-suppressed triplet loss with adaptive class-aware margins for facial expression recognition [C]// Proceedings of 2019 IEEE International Conference on Image Processing. Washington D. C. , USA; IEEE Press, 2019: 46-50.
- [23] SHAO J, QIAN Y S. Three convolutional neural network models for facial expression recognition in the wild[J]. Neurocomputing, 2019, 355(25): 82-92.
- [24] 兰凌强, 李欣, 刘淇缘, 等. 基于联合正则化策略的人脸表情识别方法[J]. 北京航空航天大学学报, 2020, 46(9): 1797-1806.
- LAN L Q, LI X, LIU Q Y, et al. Facial expression recognition method based on a joint normalization strategy [J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1797-1806. (in Chinese)

编辑 赖玉玲