硕士学位论文

(学术学位论文)

基于轻量化卷积神经网络的表情识别与可视化分析

FACIAL EXPRESSION RECOGNITION BASED ON LIGHTWEIGHT CONVOLUTIONAL NEURAL NETWORK AND VISUALIZATION

王聪荣

哈尔滨工业大学

2022 年 6 月

国内图书分类号：O24
国际图书分类号：519.6

学校代码：10213
密级：公开

硕士学位论文

基于轻量化卷积神经网络的表情识别与可视化分析

硕士研究生：王聪荣

导　　师：孙杰宝教授

申请学位：理学硕士

学科或类别：数学

所在单位：数学学院

答辩日期：2022年6月

授予学位单位：哈尔滨工业大学

Classified Index: O24
U.D.C: 519.6

Dissertation for the Master's Degree

FACIAL EXPRESSION RECOGNITION BASED ON LIGHTWEIGHT CONVOLUTIONAL NEURAL NETWORK AND VISUALIZATION

Candidate: Wang Congrong
Supervisor: Prof. Sun Jiebao
Academic Degree Applied for: Master of Science
Specialty: Mathematics
Affiliation: School of Mathematics
Date of Defence: June, 2022
Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

人脸表情识别作为人机交互的重要环节之一，在生活中的应用越来越广泛。卷积神经网络借鉴人类的视觉特性，通过构建多个卷积层实现端对端的特征提取和分类，在表情识别领域取得了较好的成绩。然而通常卷积神经网络模型规模较大，难以应用到移动端设备，且目前对表情识别模型输入与输出因果关系的研究较少，这些问题限制了表情识别模型的推广和改进。因此，本文提出一种基于卷积神经网络的轻量化表情识别模型，并使用可视化技术分析输入图像对表情识别模型决策的影响，具体研究内容如下：

首先，针对卷积神经网络规模较大难以应用到移动端设备的问题，提出轻量化卷积神经网络模型 lightResNetl 和 lightResNetm, 通过融合主流神经网络 ResNet 和轻量化结构（深度可分离卷积及倒置残差结构）大幅减少了模型的参数量和计算量，为移动端实时表情识别的实现提供技术基础。

其次，在实验室数据集 CK+ 和开放环境数据集 FER2013 上训练测试了本文模型 lightResNetl、lightResNetm 和基准模型 VGG19、ResNet18, 并针对开放环境数据集存在噪声图像和错误标签的问题提出改进数据集 FERm, 不同数据集上的实验结果均表明本文模型很好地达到了规模和性能的平衡。

最后，针对表情识别领域模型的输入与输出的因果关系不明的问题，使用 MP 方法和 Score-CAM 方法对本文模型 lightResNetl 进行可视化分析，验证模型的有效性：通过比较不同人脸表情图像的可视化结果，分析 lightResNetl 模型关注的人脸区域及区分不同表情的依据；通过对比表情图像变换前后的可视化结果，分析本文模型对不同图像变换的鲁棒性。

关键词：表情识别；轻量化；可视化；卷积神经网络

Abstract

Face expression recognition, as one of the important aspects of human-computer interaction, is increasingly used in life. Convolutional neural network which learns human visual characteristics and achieves end-to-end feature extraction and classification by constructing multiple convolutional layers, has achieved great results in the field of facial expression recognition.However, convolutional neural network models are usually large in scale and difficult to be applied to mobile devices, and there is little research on the causal relationship between input and output of expression recognition models, which limits the promotion and improvement of expression recognition models. Therefore, this paper proposes a lightweight facial expression recognition model based on convolutional neural network, and uses visualization technology to analyze the influence of input images on facial expression recognition model decisions. The specific work content is as follows:

Firstly, to address the problem that convolutional neural networks are too large to be applied to mobile devices, lightweight convolutional neural network models are proposed: lightResNetl and lightResNetm. By fusing the mainstream neural network ResNet and lightweight structures (deep separable convolution and inverted residual structures), the number of parameters and computation of the model are significantly reduced, providing a technical basis for the implementation of real-time expression recognition on mobile devices.

Secondly, lightResNetl and lightResNetm, as well as baseline models VGG19 and ResNet18, were tested for laboratory data set CK+ and open environment data set FER2013. An improved data set FERm was proposed to solve the problems of noisy images and wrong labels in open environment data set. Experimental results on different data sets show that the proposed model achieves a good balance between scale and performance.

Finally, to address the problem that the causal relationship between the input and output of the model in the expression recognition domain is unclear, two visualization methods, namely MP method and Score-CAM method, are used to analyze the lightResNetl model in this paper. By comparing the visualization results of different facial expression images,the face regions that the lightResNetl model focuses on and the basis for distinguishing different expressions are analysed. By comparing the visualization results before and after the expression image, the robustness of the proposed model for different

image transformation is analyzed.

Keywords: Facial Expression Recognition, Lightweight, Visualization, Convolutional Neural Network

目 录

第 1 章 绪论

1.1 课题背景及研究的目的和意义

1.1.1 课题的来源

人脸表情识别 (Facial Expression Recognition, FER) 技术来源于深度学习技术的计算机视觉领域，是指从输入的图像或视频中识别出特定的人类心理情绪，包括高兴、悲伤等基本表情和由基本表情复合而成的复杂情绪。随着计算机技术的迅猛发展和当今社会自动化程度的不断提高，表情识别作为计算机情感理解的基础，是当今学术界以及工业领域持续讨论的前沿课题。

目前，人脸表情识别技术的应用在生活中已随处可见，如司机疲劳驾驶监控、邢侦测谎技术、动画人物建模和心理健康监测等。然而深度学习模型受到使用设备的限制，难以推广到移动端设备，且表情识别领域模型的可解释性仍是一大难题。本文的主要研究目的即设计一个轻量化的人脸静态表情图像识别模型，探索实现模型性能和模型大小的平衡，并使用模型可视化方法尝试解释表情识别模型输入图像与分类结果的关系。

1.1.2 课题研究的背景和意义

面部表情是人际交流过程中的重要信息渠道之一，是传递自身情感最自然有力的信号。早在 20 世纪 70 年代，学者就开始了对表情识别的研究，Ekman 和 Friesen[1] 首先定义了六种基本表情：愤怒 (Anger)、厌恶 (Disgust)、恐惧 (Fear)、高兴 (Happiness)、悲伤 (Sadness) 和惊讶 (Surprise)，后来中立 (Neutral) 和蔑视 (Contempt) 也被加入到基本表情单位中[2]，表情识别领域的现有公开数据集大都以这些基本表情为标注。

随着数据时代的到来，人脸表情识别成为视觉领域的热点话题，在人机交互、医学治疗、疲劳驾驶监控、安防安全监控等场景中有着巨大的应用前景：在人机交互领域，表情识别可帮助人和机器完成情感交流的信息交互过程；在医学治疗领域，表情识别可帮助检测患者情绪，使医生更理解患者心理，这对自闭症儿童的治疗有重大意义；在安防安全监控领域中，表情识别可检测工人疲劳情绪，有助于排查安全隐患，保障工人健康。

人脸表情识别主要包括面部表情特征提取和特征分类两个步骤，特征提取步骤早期常采用人工提取特征的方法，如：Gabor 小波变换[3]、局部二值

模式 (Local Binary Pattern, LBP)[4] 和光流法等，传统的特征提取方法实现简单，需要的数据量少，然而对样本的要求较高，不易提取到深层次的特征，因此识别准确率极易受到外界环境的干扰。相比传统的特征提取算法，深度学习算法可经过训练后自动提取表情特征以及进行分类，实现了图像的端对端学习。随着计算机算力的飞速提升和数据集的迅猛扩充，基于深度学习的表情识别算法在表情分类精度和泛化性能上都要远超传统表情识别方法。因此深度学习技术逐渐取代传统算法，被广泛应用于表情识别研究领域。

基于深度学习的表情识别方法按照使用网络的不同大致可分为以下几类：基于卷积神经网络 (Convolutional Neural Network, CNN) 的表情识别方法、基于深度置信网络 (Deep Belief Net, DBN) 的方法、基于深度自编码器 (Deep Autoencoder, DAE) 的方法和基于循环神经网络 (Recurrent Neural Network, RNN) 的方法以及基于生成对抗网络 (Generative Adversarial Networks, GAN) 的方法[5]。其中 DBN 属于无监督的概率生成模型，可学习输入数据和对应标签之间的联合分布，但 DBN 缺少对图像二维空间结构信息的学习，科研工作者常将 DBN 与其他深度学习方法结合进行表情识别。RNN 方法相比其他网络可以更好地处理图像序列信息，在处理视频序列的表情识别领域具有独特的优势，然而在对静态图像的表情识别上效率不如 CNN. GAN 网络包含生成模型和判别模型两个部分，通过学习输入图像的潜在分布可生成多种多样的图像，对正面人脸图像面部遮挡问题的解决贡献尤为突出。

CNN 通常由输入层、卷积层、激活层、池化层和全连接层构成，按卷积的维度分类为 1-DCNN、2-DCNN 和 3-DCNN, 其中二维卷积在图像处理领域应用最为广泛。CNN 的实现简单，借鉴了人类的视觉特性，通过有监督的学习让机器自动提取图像特征，很好地学习了图像的位置信息和丰富的类别信息，具有平移不变性和自动学习图像局部相关性的特点，相比其他方法，极大程度地提高了表情识别的准确率。

随着手机、笔记本等移动终端对面部表情识别技术的需求不断增长，人们对面部表情识别的技术的效率提出了更高的要求。然而目前在实际应用中，基于卷积神经网络的表情识别方法出现细微表情特征提取困难，无法达到实时识别速度等复杂问题。因此，研究简单高效的表情识别模型在当下显得极为紧迫。本文设计了两种轻量化表情识别网络 lightResNetl 和 lightResNetm, 通过实验比较了本文网络和基准网络及先进网络的识别效果，并使用可视化方法对本文模型的表情识别性能进一步分析，解释表情识别模型输入与输出的对应关系。实验结果表明，本文提出的模型有效提升了表情识别速度和准确

率，且对图像变换具有较好的鲁棒性。

1.2 国内外研究现状及分析

1.2.1 基于卷积神经网络的表情识别研究现状

随着对表情识别研究的深入，卷积神经网络逐渐成为表情识别领域的主流方法，尤其是 Alexnet[6] 提出后，人们更加意识到 CNN 架构在分类问题上的独特优势。2013 年，Tang 等[7] 结合卷积神经网络方法和线性支持向量机进行表情识别，使用基于 SVM 算法的改进算法 L2-SVM 代替卷积神经网络的 Softmax 层，在 FER2013 数据集上取得了 71.2% 的准确率。对给定 K 分类问题的训练数据 (x_n, y_n), $n = 1, \cdots, N$, $x_n \in R^D$, $y_n \in R^K$, D 为输入数据的维度，N 为数据集样本数，L2-SVM 算法格式如下：

$$\min_{\omega,\xi_n} : \frac{1}{2}\omega^T\omega + C\sum_{n=1}^{N}\xi_n^2$$
$$s.t.\ \omega^T x_n y_n \geq 1 - \xi_n, \forall n$$

式中 ω 为网络倒数第二层至 SVM 层的权重，$\xi_n \in (1, 2, \cdots, N)$ 是松弛变量，用于最大化惩罚数据点和边界间的距离，C 为给定常数。

随着 VGGNet[8]、GoogleNet[9] 的提出，研究者发现通过加深卷积神经网络的层数和宽度可以使网络提取到更多深层信息，有助于图像的检测和分类等任务，由此表情识别领域出现了大量级联 CNN 和融合 CNN 的方法。2015 年，Yu 等[10] 设计级联 CNN 结构，结合自动分配级联 CNN 权重的优化算法，以 SFEW2.0 数据集上 61.29% 的识别准确率获得了 EmotiW2015 比赛中基于静态图像的表情识别亚军，通过实验证明了 CNN 架构在表情识别领域具有卓越的性能，文中提出两种级联网络的损失函数，分别为最大似然损失函数 (1-1) 和折页损失函数 (1-2)：

$$\min_{\omega} -\sum_{i=1}^{N}\log\sum_{k=1}^{K}P_k\left(y_i \mid X_i\right)\omega_k + \lambda\sum_{k=1}^{K}\omega_k^2$$
$$s.t.\sum_{k=1}^{K}\omega_k = 1,\ \omega_k \geq 0, \forall k \tag{1-1}$$

$$\min_{\omega}\sum_{i=1}^{N}\sum_{y\neq y_i}\left[1 - \frac{\sum_{k=1}^{K}\left(P_k^{i,y_i} - P_k^{i,y}\right)\omega_k}{\gamma}\right]_+ + \lambda\sum_{k=1}^{K}\omega_k^2$$

$$s.t.\sum_{k=1}^{K}\omega_k=1,\ \omega_k\geq 0,\forall k \tag{1-2}$$

其中 N 为训练集的样本数，$P_k(y_i \mid X_i)$ 表示第 i 个样本预测正确的概率，K 为训练的网络数量，ω_k 为第 k 个网络的权重，$P_k^{i,y} \triangleq P_k(y \mid X_i)$ 表示第 i 个样本预测为第 y 个标签的概率，λ, γ 为超参数。

与级联 CNN 方法类似，Kim[11] 等人针对静态表情识别问题提出一种深度卷积神经网络融合模型，如图 1-1 所示，文中设计了三种 CNN 架构，通过 CNN 架构、输入正则化和权重初始化方法的多样化组合，训练了 216 个卷积神经网络，接着通过基于有效精度的指数加权决策融合方法实现为各网络分配不同权重，最后使用最大投票法选择网络和使用加权平均法计算表情识别的结果。2016 年，Bargal[12] 等人采用标准的表情识别流程：人脸检测、图像预处理、深度特征提取、特征编码和 SVM 分类器对 AFEW 数据集进行训练测试，最终在测试集上达到 56.66% 的准确率，相比基准方法提升约 16%, 其中特征提取部分集成了改进的 VGG13、VGG16[8] 和 91 层的 ResNet[13] 网络，将不同网络提取的表情特征后通过全连接层进行融合。

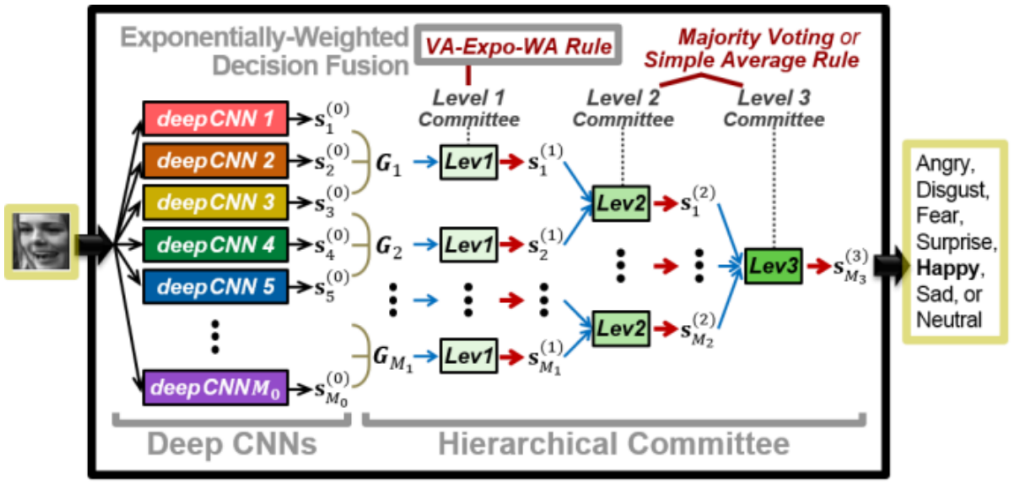


图 1-1 深度卷积神经网络融合模型[11]

尽管表情识别的研究历史悠久，相比细粒度图像分类和人脸识别问题，表情识别领域的数据集体量较小，难以训练大型 CNN 网络，因此学者们常使用图像分类或人脸识别数据集首先进行预训练，再将预训练模型在表情识别数据集上微调。然而，人脸识别数据集包含的信息与人脸表情识别需要的信息存在不小的差异，使用预训练模型微调的办法学习到的网络特征仍受人脸身份信息的影响，面部表情特征的提取有所欠缺。Ding[14] 等人提出了一种新

的两阶段网络训练方法 FaceNet2ExpNet: 第一阶段，使用训练好的 FaceNet 网络的深层特征训练表情识别网络的卷积层，具体实现方法为冻结 FaceNet 网络的参数，最小化第一阶段损失函数：

$$L_1 = \|g_\theta(I) - G(I)\|_p^p$$

式中 $G(I)$ 为冻结的 FaceNet 网络中某一层卷积层的输出，$g_\theta(I)$ 表示表情识别网络对应卷积层的输出，经过理论分析和实验测试，文中选定了最后一层池化层作为监督层，$\|\cdot\|_p$ 表示 L_p 范数；第二阶段，在训练好的表情识别网络卷积层后加上全连接层，使用高斯分布随机初始化全连接层的权重，使用表情识别数据集训练整个网络。FaceNet2ExpNet 在第一阶段学习到了人脸信息，弥补了表情识别数据集较小易过拟合的问题，在第二阶段继续训练了整个网络，相比微调方法，可以学习到更多表情识别相关的特征，实验结果表明，FaceNet2ExpNet 在 CK+ 数据集上的最佳识别准确率为 98.6%.

也有许多研究者试图设计专用于表情识别的卷积神经网络架构和损失函数，Cai 等[15] 提出了一种稀疏批归一化的 CNN 模型 (Sparse Batch Normalization Convolution Neural Network, 简称为 SBN-CNN), SBN-CNN 模型基于 VGGNet 进行改进，在浅层卷积层中使用大卷积核，结合批归一化技术和 dropout 技术避免模型训练过程中可能出现的问题。Cai[16] 等在 2018 年提出岛屿损失函数 (Island Loss), Island Loss 的格式如下：

$$L_{IL} = L_C + \lambda_1 \sum_{c_j \in \mathrm{N}} \sum_{\substack{c_k \in \mathrm{N} \\ c_k \neq c_j}} \left(\frac{c_k \cdot c_j}{\|c_k\|_2 \|c_j\|_2} + 1 \right)$$

$$L_C = \frac{1}{2} \sum_{i=1}^{m} \|x_i - c_{y_i}\|^2$$

式中 L_C 为中心损失函数，y_i 是第 i 个样本的类别标签，x_i 是第 i 个样本经过 CNN 全连接层输出的特征向量，$C_{y_i} \in \mathbb{R}^d$ 是标签均为 y_i 类的样本中心，m 是 batchsize 的大小。Island loss 通过中心损失函数使样本类内距离减小，再通过计算样本类间余弦距离使模型学习到更具区分性的特征。

2018 年，Kuo[17] 等提出了一种综合考虑模型性能和大小的表情识别网络模型，该网络模型在与通用网络性能相当的情况下，极大地减少了卷积核数量。2021 年，Cai[18] 等人设计了一个轻量级的卷积神经网络，模型包括 6 个卷积层和 3 个池化层，使用了深度可分离卷积降低模型大小，损失函数使

用了 Softmax loss 和中心损失函数结合的形式，在 FER2013 数据集上达到了 71.842% 的准确率。

近年来，研究者们尝试将卷积神经网络模型与其他网络结构结合，期望减少模型大小，并学习到更多有利于模型的泛化的特征。Georgescu 等[19]提出一种将多 CNN 模型特征与手工特征结合的表情识别方法，其中深度学习特征采用 VGG-Face, VGG-13, VGG-f 三种 CNN 网络模型经过稠密-稀疏-稠密 (Dense-Sparse-Dense, DSD) 训练得到，手工特征部分选择尺度不变特征 (Scale-Invariant Feature Transform, SIFT). 特征融合后，设计了一种局部学习框架进行特征分类，最终在 FERPlus 数据集上达到 86.71% 的精度。

大规模人脸表情数据集常受到表情模糊、低质量的人脸图像及标注者主观因素的影响存在许多错误标签的问题，为了有效地抑制表情识别数据集中的不确定性，2020 年 Wang 等人[20] 以 ResNet18 为基础，提出自修复网络 (Self-Cure Network, SCN). SCN 在网络结构中增加基于小批量样本的自注意力机制，对训练样本重新加权排序，进而识别出可能的错误标签并重新标注。

2021 年，北京大学的 Wu 等人[21] 为解决网络模型参数过多的问题，基于 DenseNet 模型[22] 设计了轻量化的表情识别网络，使用多尺度的深度可分离卷积提升卷积的感受野，丰富了卷积提取的特征，并结合通道注意力机制提高模型性能，最终在 CK+ 和 FER2013 数据集上分别达到 88.79% 和 71.77% 的准确率。

1.2.2 卷积神经网络的可视化技术

可视化技术是一种基于特征重要性的网络解释方法，主要研究模型内部的特征表示及这些特征与模型输入、输出之间的关系。可视化技术是最为直观的网络解释方法，将特征重要性结合输入图像生成显著图进行展示，可以帮助人们更好地理解卷积神经网络的内在机制和决策依据。本节将介绍常见的卷积神经网络中的可视化技术。

常用的卷积神经网络的可视化方法有基于扰动的方法、基于反向传播的方法、基于类激活映射的方法和基于注意力掩码的方法等。基于反向传播的方法将网络输出的特征通过特定规则反向传播至输入层，生成显著图。普通反向传播方法[23]（(Vanilla Backpropagetion, VBP) 为最简单的反向传播方法，通过直接计算网络输出特征关于输入数据的导数构成显著图。VBP 方法生成的显著图如图 1-2 所示，可见直接计算梯度生成的显著图含有较多噪声。

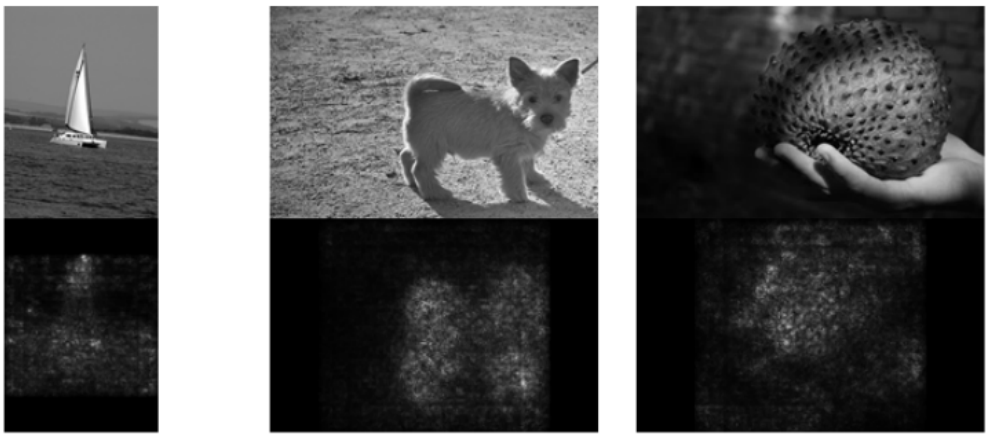


图 1-2 基于普通反向传播的可视化方法[23]

为得到更具区分性的显著图，2013 年，Zeiler 和 Fergus 提出了 ZFNet[24] 网络，他们针对 AlexNet 网络设计了卷积神经网络的逆过程即反卷积网络，将 ImageNet 分类数据集的特征通过反卷积网络进行可视化，发现浅层网络主要学习物体位置、轮廓、颜色和纹理等信息，而深层网络可学习到更多与类别整体有关的抽象信息。他们还提出特征可视化技术可以指导网络结构的设计，比如对 AlexNet 网络，在第一层使用更小的卷积核和步长可以提升网络性能。2021 年，Zhu[25] 等人进一步探索了反卷积可视化方法，他们将反卷积可视化方法结合流形几何研究了表情识别中 CNN 的分类机制。

随着 GPU 的不断发展和主流 CNN 网络层数的加深，研究者在应用深层网络时发现梯度消失问题更常出现，基于反向传播的方法受梯度的限制在深层网络中难以应用。Zhou 等人[26] 提出“随着 CNN 层数的加深，中间层特征图编码中与决策无关的信息越来越少，最后一层卷积层含有的高层语义信息最为丰富”，因此他们对 CNN 最后一层卷积特征图进行通道加权调整，生成类激活图 (Class Activation Mapping, CAM) 来可视化 CNN 识别不同类别时对图像的关注区域。

生成 CAM 图需要使用网络的全局平均池化层 (Global Average Pooling, GAP), 而一些常见的网络如 AlexNet、VGGNet 并没有。Selvaraju 等人[27] 为了能方便地对任意结构的 CNN 进行可视化，对 CAM 方法进行改进，提出了 Grad-CAM(Gradient-weighted CAM) 方法。Grad-CAM 使用反向传播中获取的通道梯度均值作为通道权重，克服了 CAM 方法对 GAP 层的依赖。近年来，还有许多研究者[28–30] 基于 CAM 方法进行了改进，基于类激活映射的方法可生成直观的可视化图像，相比基于梯度方法生成的显著图有更多具有类别区分

性的信息，然而基于类激活映射的方法大多受梯度影响大，且需要长时间复杂的迭代，计算复杂度较高。

基于扰动的可视化方法如图 1-3 所示，通过给输入图像添加特定的扰动，观察输出特征的变化，确定输入图像的重要区域。部分研究者采用简单的扰动例如使用随机值、蒙特卡洛抽样等方法生成扰动，文献[31] 提出一种有意义的扰动方法，结合常数扰动、高斯噪声扰动和高斯模糊扰动找到对图像决策有影响的重要区域。

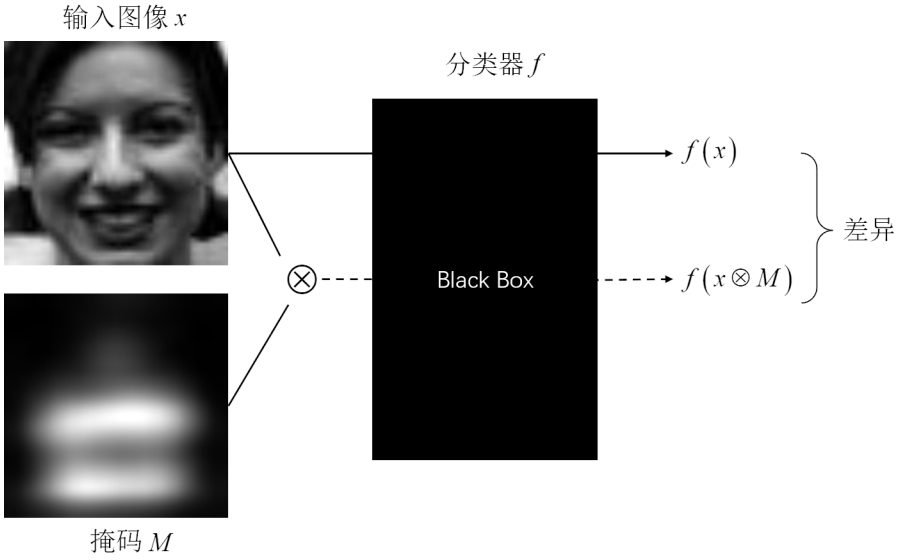


图 1-3 基于扰动的可视化方法流程图

Huang 等[32] 提出了一种对细粒度分类问题的可视化方法，结合注意力机制对 ResNet101 网络可视化，并给出了对可视化效果进行量化的方法，近年来对卷积神经网络的其他可视化方法见文献[33–35]。

1.2.3 国内外文献综述简析

目前，使用卷积神经网络进行表情识别的方法可分为以下几种：级联卷积神经网络和融合多卷积神经网络模型的方法、改进主流卷积神经网络结构和损失函数以适用于表情识别领域的方法、联合卷积神经网络和其他网络结构提取特征的方法，例如联合卷积神经网络特征和传统手工特征的方法以及联合卷积神经网络和注意力机制模块的方法。虽然这些算法在公开数据集上都取得了一些成果，但表情识别领域仍存在需要改进的地方，总结如下：

(1) 相比图像分类数据集和图像检测数据集，表情识别领域的公开数据集

体量较小，且在开放环境下采集的公开数据集存在非人脸噪声图像、水印以及错误标签的问题。

(2) 现有的基于卷积神经网络进行表情识别的方法为达到高准确率不断加深网络层，堆叠网络结构，甚至需要使用数百个CNN模型，导致算法结构复杂、参数量巨大，对训练和推理的设备要求很高，难以推广应用在移动端设备和生活场景中。因此，如何设计轻量化的网络结构，实现模型识别准确率和运行速度的平衡是表情识别领域急需解决的重要问题。

(3) 虽然理解卷积神经网络读取图像进行分类的原理的相关工作已有了一定的进展，但在表情识别领域，不同表情之间的变化更加细微，对人脸表情输入与输出相关关系的研究却很少，表情识别模型输入图像对模型决策有着怎样的影响仍是未知的。

1.3 主要研究内容

本文提出了针对表情识别的轻量化模型ligtResNetl和lightResNetm, 并使用可视化技术分析解释模型输入输出的关系，观察表情识别的重要特征，具体研究内容总结如下：

(1) 针对网络模型参数量巨大，使网络在训练时花费大量的时间，在推理时对设备性能要求很高的问题，本文融合ResNet模型和深度可分离卷积结构及倒置残差块结构，提出轻量化模型ligtResNetl和lightResNetm, 实验表明本文提出的模型很好地达到了模型准确率和效率的平衡，为实现移动端的实时表情识别提供技术基础。

(2) 在实验室环境下数据集CK+和开放环境下数据集FER2013上比较了本文模型与基准模型和先进模型的性能，针对开放环境数据集存在部分噪声图像的问题，对FERPlus数据集进行改进得到干净的基本表情数据集FERm, 实验表明在FERm数据集上的模型性能远高于FER2013数据集和FERPlus数据集。

(3) 本文使用可视化技术MP方法和Score-CAM方法对ligtResNetl模型进一步分析，通过观察引起特征激活的输入图像区域，探索网络输入和输出的相关性，也验证了本文模型对输入图像变换的可靠性。

1.4 本文结构

第 1 章，主要介绍本文课题的来源、研究目的和意义，对基于卷积神经网络的表情识别技术和网络可视化技术两个方面整理了国内外研究现状，分析总结出目前表情识别技术面临的主要问题，最后介绍了本文的主要研究内容。

第 2 章，主要研究卷积神经网络和人脸表情识别的相关预备知识，对卷积神经网络的基本结构简单介绍，并研究了通用表情识别系统的组成：人脸预处理、人脸对齐和检测、特征提取和表情分类。接着对常见表情识别数据集进行，最后研究了表情识别特征提取和分类方法。

第 3 章，主要研究改进的轻量化表情识别网络。首先介绍 ResNet 网络和轻量化网络模块：深度可分离卷积和倒置残差块，在此基础上，提出了本文的轻量化模型 lightResNetl 和 lightResNetm. 在 3.2 节中，介绍了本文提出模型的具体实现方法和训练细节，本文对 CK+、FER2013 和改进的数据集 FERm 进行了数据增强后，在各数据集上训练测试了模型 lightResNetl 和 lightResNetm, 同时训练测试了 VGG19 和 ResNet18 网络进行对比。在 3.3 节中，对 3.2 节的实验结果进行分析，通过比较各模型的准确率和混淆矩阵研究模型的识别性能，并将本文模型与目前先进的表情识别模型性能进行比较，最后给出了 FERm 数据集上各模型训练的损失函数变换曲线和准确率变化曲线。

第 4 章，主要研究了本文模型 ligtResNetl 的可视化。首先介绍了本文使用的可视化方法 MP 方法和 Score-CAM 方法，接着研究了本文模型对不同表情的可视化结果，并对不同图像变换前后可视化结果进行对比，分析表情识别模型进行决策的依据，验证了本文模型有较好的特征提取能力，且对图像亮度对比度等变化有较好的鲁棒性。

第 2 章 预备知识

2.1 卷积神经网络相关预备知识

卷积神经网络 (Convolutional Neural Network, CNN) 起源于 1980 年，日本学者福岛邦彦[36] 对猫的视觉系统进行研究，发现存在许多具有不同视觉感知功能的细胞，且神经元实际上存在局部感受区域 (Receptive Field), 也称感受野。福岛邦彦因此提出具有层级结构的神经网络，这种结构一直延用至今。

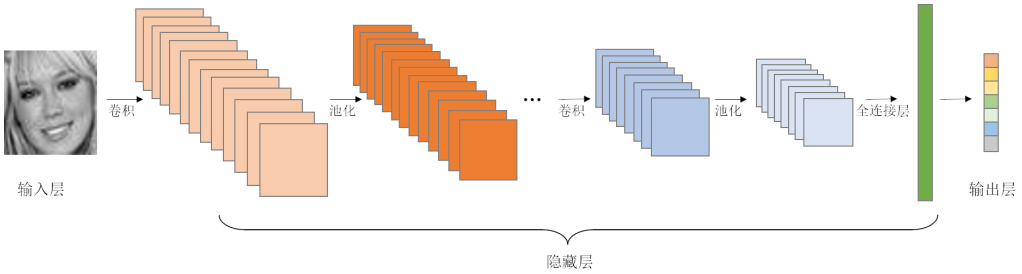


图 2-1 卷积神经网络结构图

卷积神经网络被定义为“在网络结构中至少一层使用卷积运算的神经网络[37]”，专门用来处理图像等数据，具有稀疏交互、权重共享和平移等变的特性，具体结构如图 2-1 所示，由输入层、隐藏层和输出层组成：输入层为输入样本的集合，多为灰度图像或彩色图像，灰度图像可看作是数值在 $[0,255]$ 范围内的矩阵，而彩色图像可看做是多维矩阵；输出层与具体下游任务有关，在表情识别任务中，输出为输入样本的预测表情类别；隐藏层由卷积层、池化层等模块堆叠而成，下面我们将对这些模块一一介绍。

(1) 卷积层

卷积层通过卷积运算对输入图像进行特征提取，对二维输入图像，卷积核窗口按从左往右，从上向下的顺序依次与输入图像矩阵做二维互相关运算，卷积操作可公式化为：

$$\begin{aligned} S(i,j) &= (I * K)(i,j) \\ &= \sum_{m,n} I(m,n) K(i-m, j-n) \end{aligned}$$

式中，$I(i,j)$ 为输入图像矩阵，K 表示大小为 (m,n) 的卷积核，$S(i,j)$ 是卷积后对应的输出特征。在卷积操作中，通过控制 $m \ll i, n \ll j$ 减少模型参数需求，相比传统神经网络卷积运算只需要很少的计算量。

对彩色图像，卷积运算如图 2-2 所示，此时卷积核深度与输入图像通道数相同，一个卷积核作用于输入图像得到一个输出的特征图。改变卷积核的参数可以实现对图像的不同特征提取功能，如：边缘检测、图片锐化及高斯模糊等。图 2-3 展示了对 Lena 图像进行卷积操作后的结果。

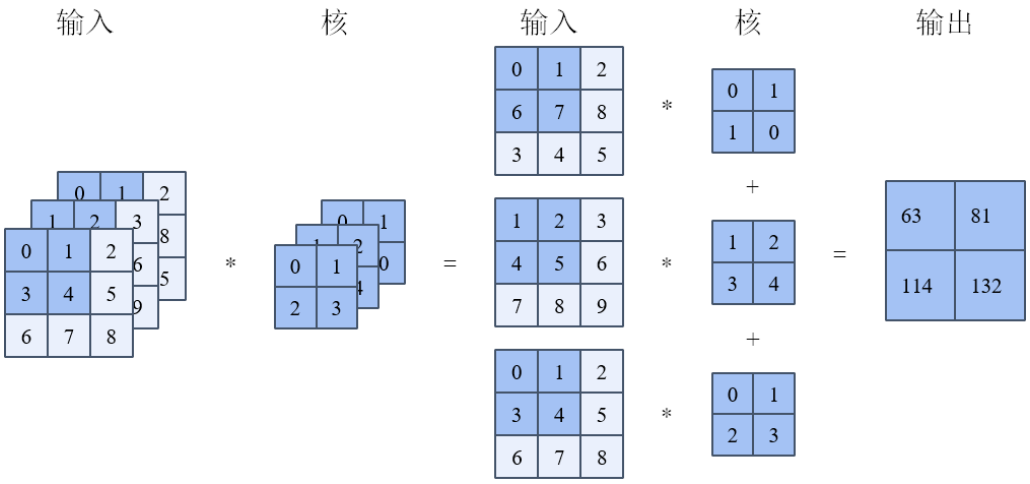


图 2-2 彩色图像卷积的运算

图 2-3 Lena 图像卷积效果图

(2) 池化层

池化层类似卷积运算，每次对输入数据选取固定大小状窗口，也就是池化窗口，在池化窗口内进行计算得到输出结果。不同于卷积的是，通常池化

窗口在特征矩阵上的滑动是非重叠的滑动，典型的池化操作有 L_p 池化、混合池化以及随机池化。

L_p 池化的形式为：

$$y_{i,j,k} = \left[\sum_{(m,n)\in R_{i,j}} (a_{m,n,k})^p \right]^{\frac{1}{p}}$$

式中，$y_{i,j,k}$ 是池化层在第 k 个特征图中位置 (i,j) 上的输出，$a_{m,n,k}$ 是输入第 k 个特征图上池化窗口中位置为 (m,n) 处的值。特别地，当 $p=1$ 时，表示平均池化，$p=\infty$ 时，表示最大池化。图 2-4 展示了池化窗口为 2×2 时，对 4×4 的特征图求最大池化和平均池化的示例，如图所示，由于采用的是非重叠的滑动，输出特征图大小为 2×2。

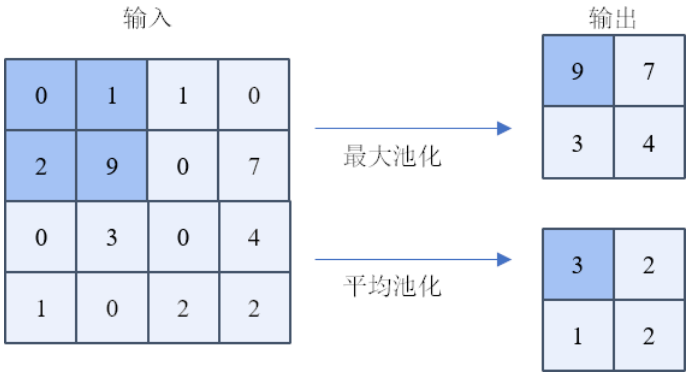


图 2-4 池化操作示意图

混合池化是指在池化时随机选择平均池化或最大池化，混合池化的公式如下：

$$y_{i,j,k} = \lambda \max_{(m,n)\in R_{i,j}} a_{m,n,k} + (1-\lambda)\frac{1}{|R_{i,j}|}\sum_{(m,n)\in R_{ij}} a_{m,n,k}$$

式中，λ 随机选择为 0 或 1.

随机池化根据多项式分布，在特征图的池化窗口中随机选择激活值，这样做可以更大限度地利用特征图信息，更加不容易受到极端值的影响，同时又不会增加太多计算量。随机池化首先对池化窗口 R_j 计算各位置的概率，采用的计算公式为：

$$p_i = \frac{a_i}{\sum_{k\in R_j}(a_k)}$$

接着根据分布 $P\left(p_1,\cdots,p_{|R_j|}\right)$ 对池化窗口进行随机采样，公式为：

$$y_j = a_l, l \sim P\left(p_1,\cdots,p_{|R_j|}\right)$$

(3) 全连接层

在 CNN 中，全连接层一般位于输出层之前，网络使用单层或多层的全连接层聚合特征向量并进行降维，达到指定维度的输出以用于下游任务。全连接层顾名思义，每个神经元都与上层所有神经元相连。公式为：

$$F(x)=\sigma\left(W^{T}x+b\right)$$

式中，x 为上一层输出的特征图，W 为权值矩阵，b 表示偏置向量，$\sigma(\cdot)$ 表示非线性函数。图 2-5 展示了全连接层和卷积层的区别。

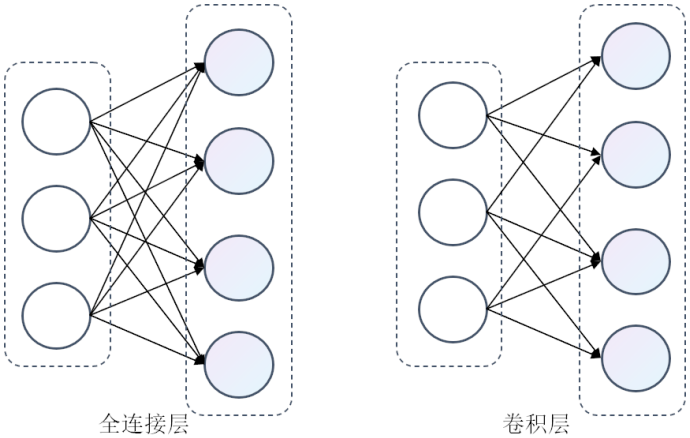


图 2-5 全连接层与卷积层的对比

(4) 激活函数

激活函数位于卷积层之后，使用非线性的函数避免 CNN 可简化为单一的线性函数，提高网络表达和学习特征的能力。激活函数通常具备以下特点：除零测集外连续可导；形式简单；激活函数及导数的值域有界，避免网络反向传播时发生梯度消失和梯度爆炸，影响训练。下面介绍几种常用的激活函数。

Sigmoid 型函数是两端饱和的 S 型函数，典型的 Sigmoid 型函数有 Logistic 函数和 Tanh 函数，形式如下：

$$\sigma(x)=\frac{1}{1+e^{-x}}$$

$$\tanh(x)=\frac{\sinh(x)}{\cosh(x)}=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$$

式中，$\sigma(x)$ 表示 Logistic 函数，$\tanh(x)$ 表示 Tanh 函数。

图 2-6 为 Logistic 函数和 Tanh 函数的图像，可见，Logistic 函数的值域为 $\sigma(x) \in (0,1)$, 且有 $\lim_{x\to-\infty}\sigma(x)=0, \lim_{x\to\infty}\sigma(x)=1$. Tanh 函数的值域为 $\tanh(x) \in (-1,1)$, 且有 $\lim_{x\to-\infty}\tanh(x)=-1, \lim_{x\to\infty}\tanh(x)=1$.

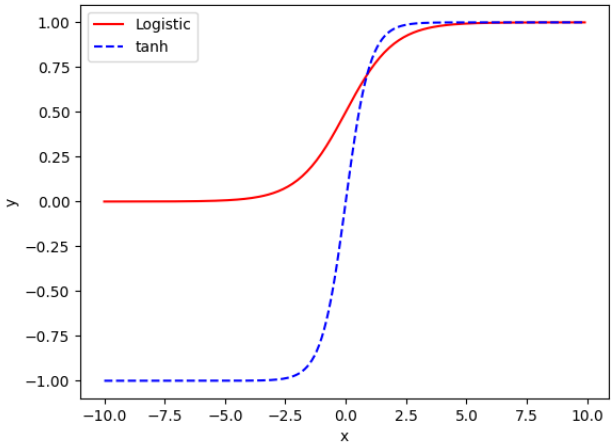


图 2-6 sigmoid 型函数

Logistic 函数和 Tanh 函数的导数形式为：

$$\begin{aligned}\sigma'(x) &= \frac{e^{-x}}{(1+e^{-x})^2} \\ &= \sigma(x)(1-\sigma(x)) \\ \tanh'(x) &= 1-\frac{(e^{x}-e^{-x})^2}{(e^{x}+e^{-x})^2} \\ &= 1-\tanh^2(x)\end{aligned}$$

可见，Logistic 函数和 Tanh 函数的导数均可用原式表出，且导数有界。

修正线性单元 (Rectified Linear Unit, ReLU) 是 CNN 中另一种常见的激活函数，定义为：

$$\mathrm{ReLU}(x)=\begin{cases}x, x\geq 0\\ 0, x<0\end{cases}$$

ReLU 函数启发于人脑神经元，只有输入 x 大于 0 时，才能启动激活，输入 x 小于 0 时，训练数据被置为 0. 相比 Sigmoid 型函数，计算更加简单，且由于 $\mathrm{ReLU}'(x)_{x>0}=1$, 可降低梯度消失的风险。ReLU 函数的图像如图 2-7 所示。

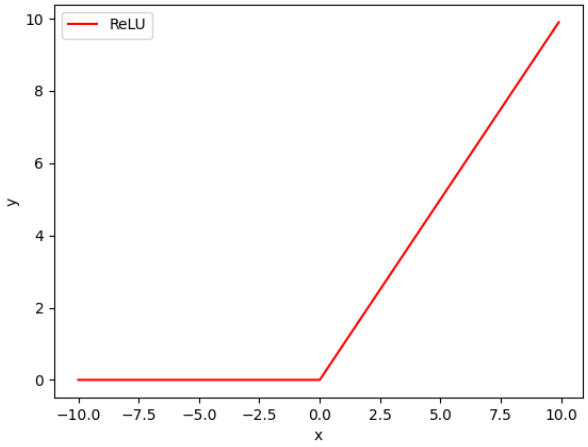


图 2-7 ReLU 函数图

2.2 表情识别相关预备知识

2.2.1 面部表情识别系统

表情识别系统通常包含四个环节，如图 2-8 所示，首先是获取输入样本，包括动态图像序列和静态图像，均可通过图像采集获得；接着对输入样本进行预处理，常见的预处理方法包括数据增强和对人脸照明及姿态的归一化处理，图像预处理对提高图像质量，降低噪声污染的影响，提升训练测试效率具有重要意义；预处理后，再使用人脸检测技术去除图像的非人脸区域；然后是特征提取与分类阶段，可使用传统特征提取的方法或基于深度学习的特征提取方法。本节中，我们将重点介绍人脸检测与特征提取中常用的算法。

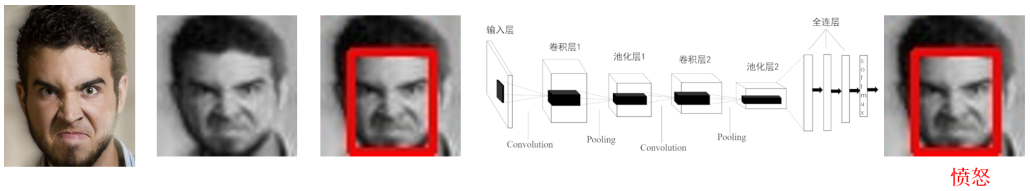


图 2-8 人脸表情识别系统

2.2.2 表情识别数据集

基于卷积神经网络的表情识别实验通常需要使用标注好的数据集进行训练和测试，目前已存在许多公开的表情识别数据集，如 CK+、FER2013、

AffectNet 等。表情识别数据集按采集来源可分为实验室采集和公开环境下采集，CK+ 数据集就是典型的实验室环境下采集的数据集，CK+ 数据集的示例如图 2-9 所示。开放环境下数据集有 FER2013、RAF-DB 数据集等，是从电影视频中剪辑获得或从网络上收集整理得到。图 2-10 展示了部分 RAF-DB 数据集图片，可见开放环境中获得的人脸图像光照条件有较大不同，且人脸姿态多样化，相比实验室环境，开放环境下的人脸表情识别仍是一个重要的挑战。

图 2-9 CK+ 数据集示例

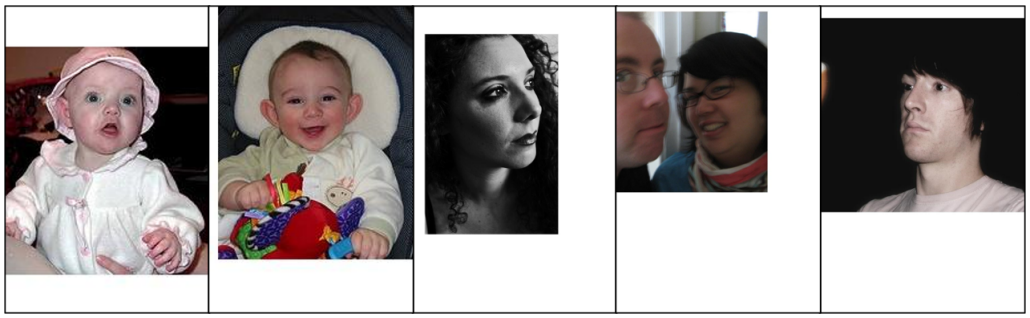


图 2-10 RAF-DB 数据集示例

2.2.3 人脸检测对齐算法

人脸检测是指从图像中检测出人脸所在位置，实现人脸定位。表情识别实验中可通过人脸检测技术去除图像的非人脸区域，避免表情识别算法受到噪声干扰。人脸对齐是指在人脸面部图像上通过关键点定位搜索人脸形状，通常人脸关键点是指眼睛、鼻子、嘴巴以及下巴，通过人脸关键点对图片进行一定的旋转、缩放可使人脸图像的角度保持一致。深度学习中大部分人脸检测算法可同时实现人脸关键点检测，常用的基于深度学习的人脸检测算法有 MTCNN(Multi-task Convolutional Neural Networks)[38]、CenterFace[39]、RetinaFace[40] 等。

CenterFace[39] 由 Xu 等人于 2020 年提出，具有网络结构简单、运行速度快且相比其他人脸检测模型识别精度较高的特点。CenterFace 基于多任务学习策略，将人脸检测问题转换为关键点估计和人脸框回归问题，可同时实现人脸检测和关键点定位。CenterFace 的网络结构如图 2-11 所示。

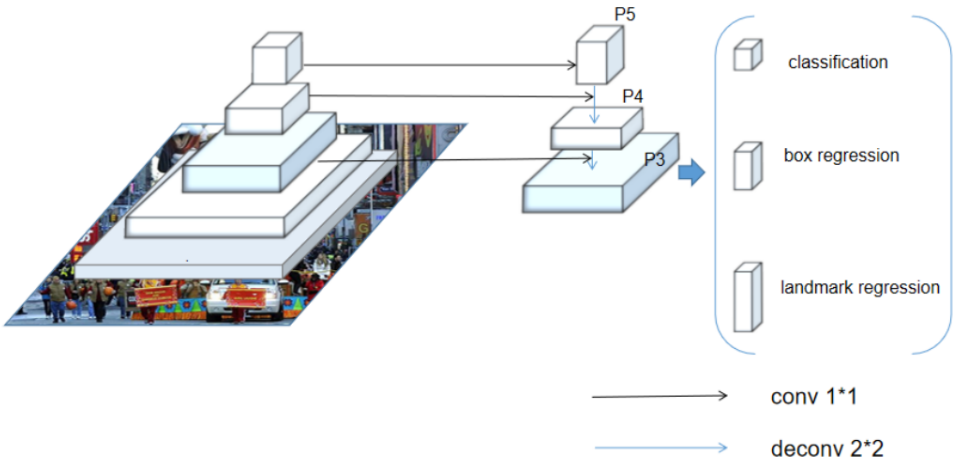


图 2-11 CenterFace 网络结构图[39]

图 2-12、2-13 展示了对 CK+ 数据集和 RAF-DB 数据集进行人脸检测与对齐后的处理效果，对比可见，数据集图片经过人脸检测对齐后，图像大小相同，且人脸角度保持一致。

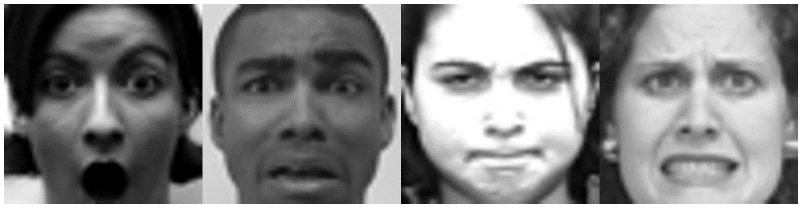


图 2-12 人脸检测对齐后的 CK+ 数据集示例

图 2-13 人脸检测对齐后的 RAF-DB 数据集示例

2.2.4 表情特征提取与分类

传统的表情识别方法往往是使用手工设计的特征或通过浅层学习进行人脸识别，随着科学技术的不断发展，得益于数据集规模的不断壮大及高性能计算机芯片的普及，深度学习方法在许多计算机视觉领域取得了极其良好

的性能。在特征提取与选择阶段，常用的基于机器学习的特征提取方法有局部二值特征、方向梯度直方图 (Histogram of Oriented Gradient, HOG)[41] 和小波变换方法等。

HOG 是一种传统的图像特征提取方法，使用图像的梯度特征作为图像描述子，因此 HOG 特征更注重图像边缘信息，可以较好地描述局部信息，HOG 特征提取的算法步骤如下：

(1) 首先将输入图像 $I(x)$ 灰度化，并利用 gamma 校正法将图片归一化；

(2) 计算图像各像素点的梯度，包括梯度幅值和方向，具体公式为：

$$\begin{aligned} G_x(x,y) &= I(x+1,y) - I(x-1,y) \\ G_y(x,y) &= I(x,y+1) - I(x,y-1) \\ G(x,y) &= \sqrt{G_x^2 + G_y^2} \\ \theta(x,y) &= \arctan\frac{G_y}{G_x} \end{aligned}$$

(3) 将图像均匀划分为多个小窗口 (cell), 在每个窗口内计算方向梯度直方图。首先把梯度方向平均分为 N 个方向角度区域 (Bin), 计算对每个像素点 (x,y) 关于对第 n 个 Bin 的权值 $\lambda_n(x,y)$:

$$\lambda_n(x,y) = \begin{cases} G(x,y), \theta(x,y) \in \mathrm{Bin}_n \\ 0, \theta(x,y) \notin \mathrm{Bin}_n \end{cases}$$

则每个窗口的梯度直方图向量为 $\left[H_1^l, H_2^l, \cdots, H_N^l\right]$, 其中 $H_n^l = \sum_{(x,y)\in \mathrm{cell}} \lambda_n(x,y)$, $1 \le n \le N$.

(4) 将相邻的 m 个窗口组成一个计算单元，每个计算单元内的梯度特征表示为 ν, 并进行 L_2 模标准化：

$$\begin{aligned} \nu &= \left[H_1^l, \cdots, H_N^l, \cdots, H_1^m, \cdots, H_N^m\right] \\ \nu &\leftarrow \frac{\nu}{\sqrt{(|\nu| + \epsilon)^2}} \end{aligned}$$

(5) 将图像中所有计算单元的梯度特征串联得到图像的 HOG 特征描述子。

图 2-14 为一张 48×48 的人脸灰度图像的 HOG 特征，设置梯度方向为 9 个方向角度区域，计算单元大小为 (12,12), 滑动步长为 6, 每个窗口的大小为

(6,6), 则图片提取的 HOG 特征维度为 1764 维。可见在人脸的面部较为平坦的区域，HOG 特征值较小，而在面部非平坦区域，如人脸嘴角处，HOG 特征值较大，且展示出了梯度方向为 90 度。尽管 HOG 方法实现简单，但提取的特征维度较大，计算量大，且只能提取到图像的浅层信息，无法处理如图像遮挡等复杂问题，故不再适用于实际应用场景。

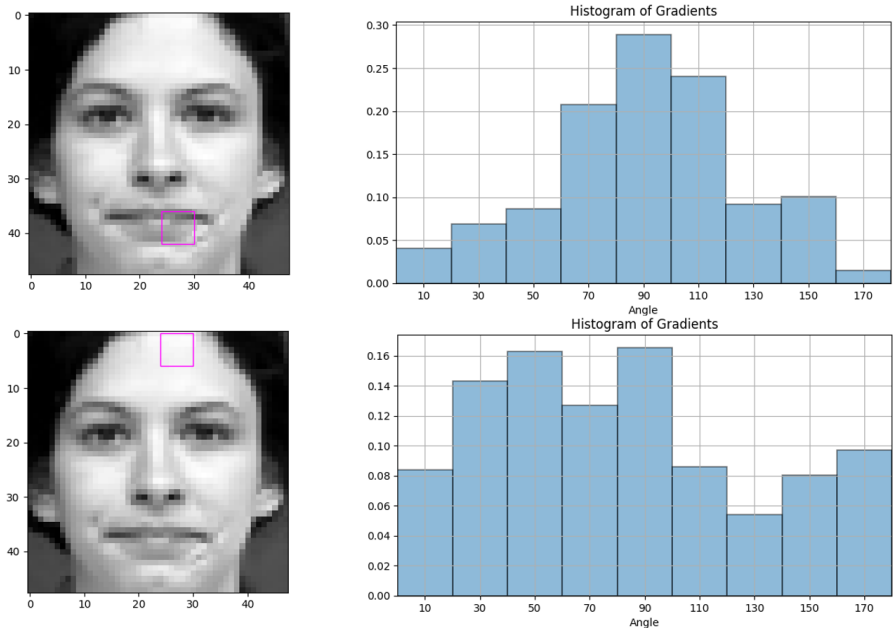


图 2-14 人脸图像的 HOG 特征

基于深度学习的人脸特征提取方法可以避免由于手工设计的特征描述带来的提取误差，通过仿照人脑的学习过程，在没有人为干预的情况下将原始图像提取为高维特征。卷积神经网络由于其稀疏连接、权重共享等特性，在表情识别领域的应用最为广泛，但是基于 CNN 进行表情识别的方法仍存在需要数据量大、训练和部署时对设备要求高及可解释性差的问题。

2.3 本章小结

本章主要介绍了表情识别领域的基础知识，首先研究了卷积神经网络的基本结构，包括卷积层、池化层、全连接层和激活函数，接着介绍了通用的面部表情识别系统，对表情识别流程中的每个步骤进行研究，主要是介绍常见的公开表情识别数据集、人脸检测和关键点定位算法以及人脸特征提取分类算法。

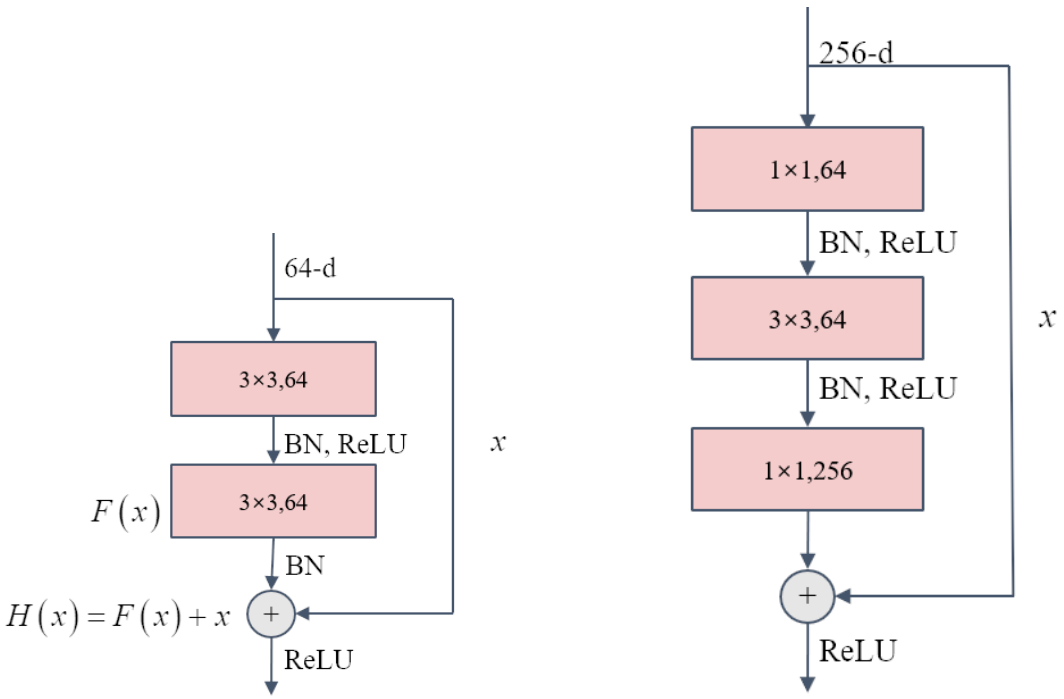
第 3 章 基于改进轻量化网络的人脸表情识别

本章提出了一个轻量化的表情识别网络架构，首先通过融合经典的 ResNet 网络和轻量化结构即深度可分离卷积和倒置残差块提出轻量化模型 lightResNetl 和 lightResNetm, 接着对实验中使用的数据处理方法和训练设置进行介绍，最后给出了多个数据集上的实验结果。

3.1 基于 ResNet 模型的轻量化改进

3.1.1 ResNet 网络

随着深度学习的发展，一个重要的趋势是网络模型不断变深，这是因为深层网络可以学习到更多抽象的特征表示，有利于图像的特征提取。然而，简单的堆叠网络层会出现随着网络深度的增加，模型精度达到饱和后急速下降的现象，这就是网络的退化问题。为解决深层网络的退化问题，ResNet[13] 在网络中引入“恒等映射连接”，提出如图 3-1(a) 所示的残差块结构。



(a) 残差块结构　　(b) 瓶颈结构

图 3-1 ResNet 结构图

其中“64-d”表示输入特征图通道数为64，“$3\times3, 64$”表示64个尺寸为3×3的卷积核构成的卷积层，ReLU是非线性激活函数，BN为批归一化。假设在神经网络的某一个层结构中，输入为x, 期望的输出值为$H(x)$, 则在ResNet的残差结构中，网络学习的不再是$H(x)$, 而是学习期望输出$H(x)$与输入x的残差，即学习$F(x)$, 其中$F(x)=H(x)-x$. 这种残差学习的思想没有引入额外的参数，因此不会增加网络的计算复杂度，且浅层网络的信息可以不经过变换直接传输到深层网络中，对原始浅层网络的学习能力没有影响，还可避免深层网络可能存在的冗余信号传递。

在大型网络如ResNet50、ResNet101中，为减少训练复杂度，采用瓶颈结构代替残差块结构，瓶颈结构示意图如图3-1(b)所示，其中“256-d”表示输入特征图通道数为256, 首先经过1×1卷积层对特征图进行降维，最后经过1×1卷积层对特征图进行升维。这样做使得3×3卷积层的输入输出特征图维度均较小，可减少模型参数量和计算量。表情识别领域中常用的ResNet网络为ResNet18, 具体结构如图3-5所示，ResNet18包含5个部分，第一部分为快速下采样层，在第二至五部分中均采用残差块结构或瓶颈结构进行堆叠，且和VGGNet类似，均采用3×3小卷积核，最后通过池化层和Softmax进行分类。

3.1.2 深度可分离卷积和倒置残差块

深度可分离卷积[42]的主要思想是将传统标准卷积操作分解为：深度卷积(Depthwise Convolution, DW)和逐点卷积(Pointwise Convolution, PW). 图3-2展示了传统标准卷积和深度可分离卷积的对比，可见，标准卷积作用于特征图的所有通道上，同时实现了卷积以及联合各图通道信息。深度卷积的通道数均为1, 卷积核个数数等于输入特征图的通道数，因此深度卷积只改变特征图的尺寸，不改变特征图通道数。逐点卷积长宽尺寸均为1, 用于联合深度卷积输出特征图中跨通道的信息，并得到要求的输出特征图通道数。

假设对深度网络某一层结构，输入特征图尺寸为$D\times D\times M$, 输出特征图尺寸为$D_1\times D_1\times N$, 则对应的标准卷积尺寸为$D_k\times D_k\times M\times N$. 若使用深度可分离卷积，深度卷积尺寸为$D_k\times D_k\times M$, 逐点卷积尺寸为$1\times1\times M\times N$. 因此，深度可分离卷积和标准卷积的计算量比值为式(3-1)所示，参数量比值为式(3-2)所示：

$$\frac{D_k\times D_k\times M\times D\times D+M\times N\times D\times D}{D_k\times D_k\times M\times D\times D\times N}=\frac{1}{N}+\frac{1}{D_k^2} \tag{3-1}$$

$$\frac{D_k \times D_k \times M + M \times N}{D_k \times D_k \times M \times N} = \frac{1}{N} + \frac{1}{D_k^2} \tag{3-2}$$

当 D_k 为 3 时，深度可分离卷积的计算量和参数量均约为标准卷积的九分之一，且深度可分离卷积的参数量主要集中在逐点卷积处。

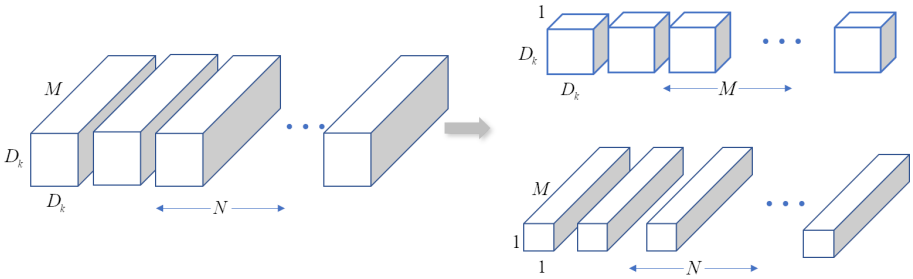
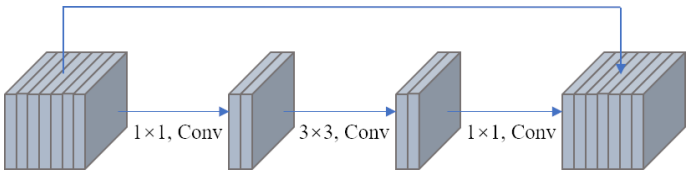
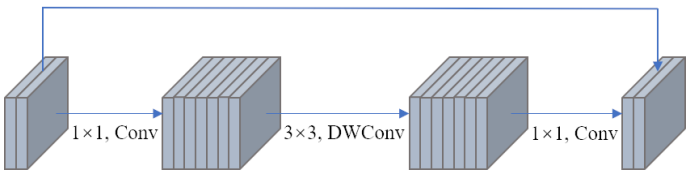


图 3-2 标准卷积与深度可分离卷积

倒置残差结构在 MobileNet V2[43] 中提出，结构与 ResNet 中瓶颈结构相同，均采用 $1 \times 1 \rightarrow 3 \times 3 \rightarrow 1 \times 1$ 的模式，不同点在于 ResNet 使用标准卷积提取特征，而倒置残差结构使用 DW 卷积提取特征。故 ResNet 中特征图通过 1×1 卷积先降维再进行卷积，减少网络的复杂度，而倒置残差结构中特征图通过 1×1 卷积先升维再卷积，提升倒置残差结构提取特征的能力，两者结构的对比如图 3-3 所示。



(a) 残差结构



(b) 倒置残差结构

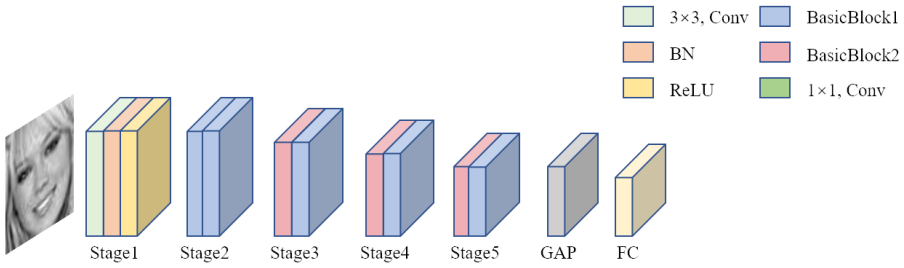
图 3-3 残差结构与倒置残差结构

3.1.3 ResNet 网络的轻量化改进

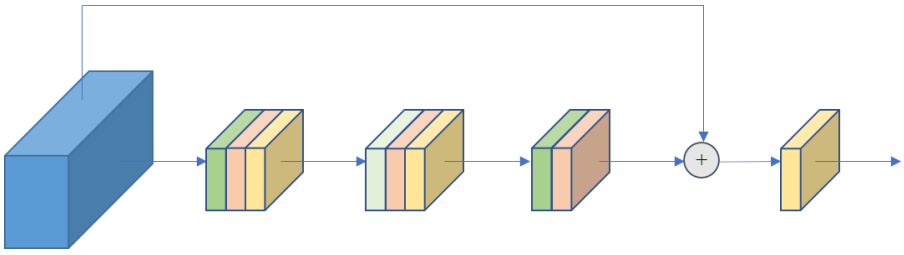
由于目前主流网络模型参数量巨大，使得网络在训练时花费大量的时间，在推理时对设备性能要求很高，对主流网络架构的轻量化改进是网络结构优

化的重要方向。本文将 ResNet 网络和轻量化结构进行融合，具体做法为在网络中用深度可分离卷积代替传统的普通卷积，引入倒置残差结构代替残差结构，提出轻量化的网络结构 lightResNets.

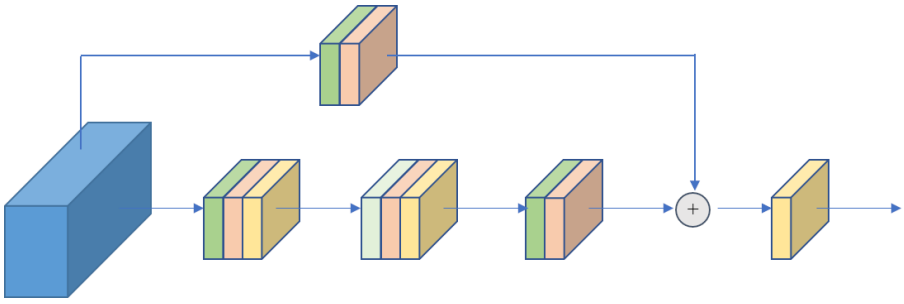
lightResNets 的网络结构如图 3-4(a) 所示，网络由 5 个基本 Stage 构成，每个 Stage 均由两种 BasicBlock 组合而成，BasicBlock1 和 BasicBlock2 的结构如图 3-4(b) 和图 3-4(c) 所示。BasicBlock1 输入和输出的特征图尺寸不变，BasicBlock2 相比 BasicBlock1 在恒等映射连接中增加了 1×1 卷积，可实现网络的下采样。



(a) lightResNets



(b) BasicBlock1



(c) BasicBlock2

图 3-4 lightResNets 结构图

基于 lightResNets 的网络结构设置，为比较网络宽度设置对模型复杂度和模型性能的影响，本文提出了两种具体的轻量化模型，分别为 lightResNetl 和 lightResNetm, 二者的网络框架结构见图 3-5.

Stage	ResNet18	lightResNetl	lightResNetm	Output size
Stage1	7×7, 64, stride 2 3×3 max pool, stride 2	3×3, 64, stride 1		44×44
Stage2	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,64\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,64\end{bmatrix}\times2$	44×44
Stage3	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,128\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,128\end{bmatrix}\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,128\end{bmatrix}$	22×22
Stage4	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,256\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,256\end{bmatrix}\begin{bmatrix}1\times1,1024\\3\times3,1024\\1\times1,256\end{bmatrix}$	11×11
Stage5	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,1024\\3\times3,1024\\1\times1,512\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,1024\\3\times3,1024\\1\times1,512\end{bmatrix}\begin{bmatrix}1\times1,2048\\3\times3,2048\\1\times1,512\end{bmatrix}$	6×6
	Average pool, 1000-d fc, Softmax	Global Average pool, 7-d fc		

图 3-5 ResNet18、lightResNetl 和 lightResNetm 网络框架对比图

本文模型与 ResNet18 的网络框架对比如图 3-5 所示，ResNet18 模型主要包含一个快速下采样结构和四个残差块结构，其中快速下采样结构采用了一个 7×7 卷积层和一个最大池化层，在所有模块中，卷积层后都添加了 BatchNorm 和 ReLU 激活层。由于实验中训练表情识别模型的输入图片大小均处理为 48×48, 因此本文模型在快速下采样结构中采用了更小的 3×3 卷积核，且令步长为 1, 以保留更多的图像信息。为最大限度保留网络提取特征的能力，同时减少网络的参数和计算量，在每个残差块结构中采用倒置残差结构，设置网络中每个 Stage 中的 Blocks 数量与 ResNet18 相同，即 $(2,2,2,2)$. 每个倒置残差块的扩展因子分别为 $(2,4,4,4)$. 在网络的最后一层中，lightResNetl 和 lightResNetm 均采用全局平均池化代替 ResNet18 中的平局池化，理论分析和实验结果均表明采用全局平均池化可以有效扩大网络的感受野，并使网络最终的全连接层不受输入图像尺度的影响[43]。

为更直接地体现本文设计模型在参数量和计算量上的大幅减少，在表 3-1 中列出了 lightResNetl、lightResNetm 与 ResNet18 的模型参数量和计算量对比，其中计算量均以输入图像大小为 48×48 为例。可见 lightResNetl 的参数量约为 ResNet18 的四分之一，VGG19 的二分之一；计算量约为 ResNet18 的三分之一，VGG19 的七分之一。lightResNetm 的参数量则约为 ResNet18 的五分之二，VGG19 的五分之一；而计算量仅为 ResNet18 的五分之二，VGG19 的三分之二。

表 3-1 不同模型参数量和计算量对比图

模型	FLOPs	Params
VGG19	677 413 888	20 038 983
ResNet18	1 100 677 120	11 172 423
lightResNetl	**342 251 520**	**2 678 087**
lightResNetm	446 557 184	4 079 431

3.2 数据预处理与模型训练

本文采用的表情识别数据有 CK+ 和 FER2013, 并在 FERPlus 数据集的基础上进行改进，得到 FERm 数据集，下面对这几个数据集进行简单介绍。

CK+ 数据集是在实验室环境中得到，通过采集 123 个对象的 593 个图像序列，获得共 981 张表情图片，训练前将 CK+ 数据集分为 882 张训练图片和 99 张验证图片，CK+ 数据集中各类表情的数量分布如表 3-2 所示。

表 3-2 CK+ 数据集分布

	Angry	Disgust	Fear	Happy	Sad	Surprise	Contempt	Total
Train	123	159	66	186	75	225	48	882
Valid	12	18	9	21	9	24	6	99

表 3-3 FER2013 数据集分布

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total
Train	3 995	436	4 097	7 215	4 830	3 171	4 965	28 709
Valid	467	56	496	895	653	315	607	3 589
Test	491	55	528	879	594	316	626	3 589

FER2013 数据集是从互联网收集的视频中截取的面部表情图像，常用作表情识别分类挑战，表 3-3 为数据集中各类表情具体的分布。由于 FER2013 数据集是从开放环境中获得，不仅有人脸图片，还有部分动画图片，且存在标签错误、水印等问题，FER2013 数据集中部分噪声图像如图 3-6 所示。

图 3-6 FER2013 数据集中部分噪声图片

FERPlus 数据集是在 FER2013 数据集的基础上，人工修改图片标签，扩充为 10 分类数据集，包括蔑视，未知和非人脸标签。对所有图片采用复合标签，提高了标签准确性。

本实验中去除 FERPlus 数据集的“未知”和“非人脸类别”，使用最大投票方法修改了原数据集的复合标签，考虑到“蔑视”和“厌恶”类别图片数量较少且较为接近，将“蔑视”和“厌恶”类别结合，处理后的数据集命名为 FERm, FERm 数据集的表情类别分布如表 3-4 所示。

表 3-4 FERm 数据集分布

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total
Train	2 100	238	532	7 287	3 014	3 149	8 740	25 060
Valid	287	37	62	865	351	415	1 182	3 199
Test	273	34	83	893	384	396	1 090	3 153

由于实验中使用的数据集较小，尤其是 CK+ 数据集中只有约 1000 张图片，为防止模型过拟合，并且提升模型的泛化能力，实验采用了一系列在线数据增强方法，包括：随机旋转、随机翻转、亮度调整、对比度调整、10-crop 随机裁减等，在线数据增强效果如图 3-7 所示。

图 3-7 数据增强示例图

实验中使用的操作系统是Ubuntu(20.04版)，由于实现深度学习网络架构需要较高性能的计算机，且实验使用数据集的计算量较大，故使用了专门用于大型计算的图像处理器(GPU), 使用GPU的型号为RTX3090, 安装Pytorch1.7版，与Pytorch相适应的CUDA为11.0版，CUDNN为8.0.5版，实验使用的编程语言为python3.8版。此外，根据实验需要，还需安装如表3-5所示的python模块：

表3-5 实验所需模块

numpy	h5py	psutil	opencv-python
matplotlib	torchvision	scikit-learn	scipy

本文在CK+、FER2013和FERm数据集上分别训练了VGG19、ResNet18、lightResNetl和lightResNetm模型，通过消融实验，确定实验中的各超参数，下面介绍各数据集上的训练参数设置。CK+数据集中，设置batchsize为32, 学习率初始值为0.01, 从训练的第二十轮开始每轮进行衰减，学习率衰减率为0.8, 总共训练60轮。FER2013和FERm数据集中batchsize为128, 学习率初始值为0.01, 从训练的第八十轮开始衰减，每五个epoch衰减一次，学习率衰减率为0.9, 总共训练200轮。使用的loss函数为交叉熵损失函数，则带有正则化的目标函数公式如下：

$$Obj = L + \frac{\lambda}{2}\sum_{i=1}^{m}\omega_i^2$$

$$L = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_{ij})\log(q(x_{ij}))$$

式中L为交叉熵损失，m为batchsize大小，n为类别数，$p(x_{ij})$表示样本的真实标签，$q(x_{ij})$表示i类别样本被识别为j类别的概率，λ为正则化参数，实验中设置为$5e-4$.

在CK+、FER2013和FERm数据集上均采用随机梯度下降法，公式如下：

$$\nabla\omega = \frac{1}{m}\sum\frac{\partial Obj}{\partial\omega}$$

$$\nu_t = \mu\nu_{t-1} + \nabla\omega$$

$$\omega = \omega - \nu_t\alpha$$

式中α为学习率，μ为动量，本文实验设置动量为0.9.

3.3 实验结果及分析

3.3.1 CK+ 数据集上的实验结果及分析

在表情识别算法中，通常使用平均识别准确率 (Accracy, 简记为 Acc) 和混淆矩阵来衡量算法的性能。平均识别准确率是图像分类算法中评估模型性能最常用的方式之一，人脸表情识别算法的平均识别准确率是指各表情类别中预测正确的图像数量占测试集中总数量的百分比平均值，计算公式如下：

$$\mathrm{Acc} = \frac{1}{m}\sum_{i=1}^{m}\frac{n_i}{N_i}$$

式中 m 表示表情类别数，实验中设置 $m = 7$, N_i 为第 i 类表情的测试样本数量，n_i 表示模型对第 i 类表情预测正确的图片数量。

表 3-6 为 CK+ 数据集上本文模型 lightResNetl 和 lightResNetm 和基准模型 VGG19、ResNet18 的识别准确率对比，可见由于实验室数据集人脸姿态统一，背景单调，VGG19、ResNet 和本文模型的识别准确率都很高。

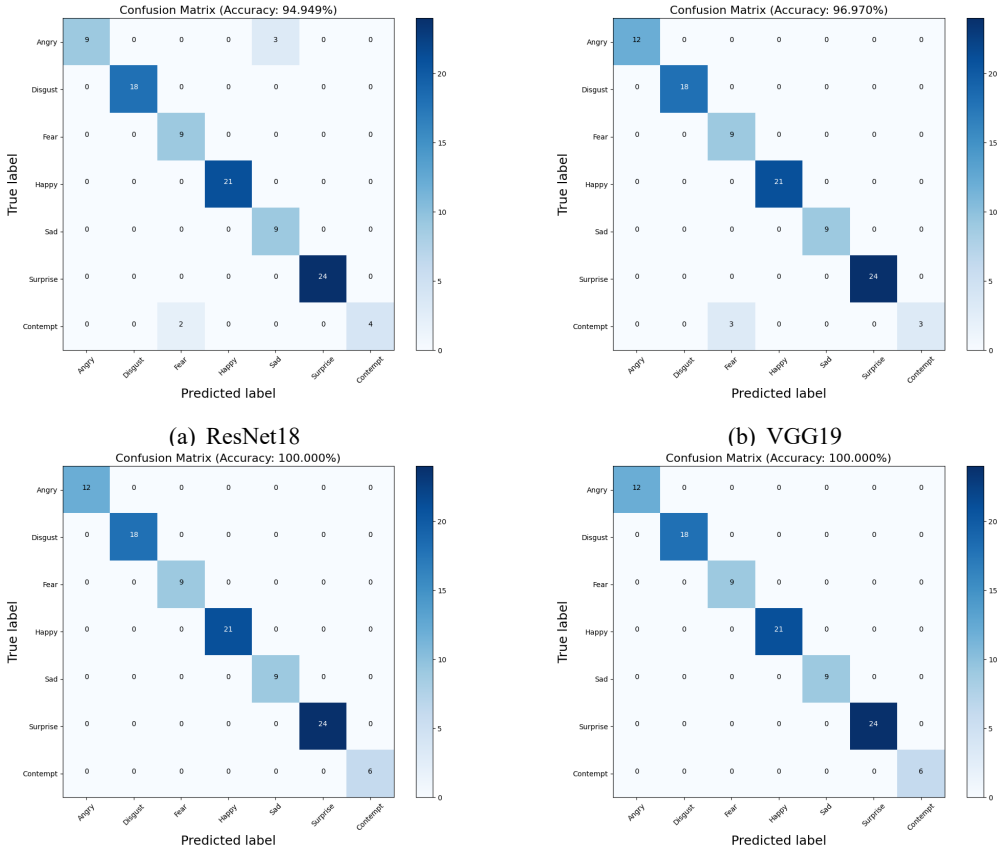
表 3-6 CK+ 上本文模型与基准模型的对比

模型	识别准确率
VGG19	96.970%
ResNet18	94.949%
lightResNetl	**100%**
lightResNetm	**100%**

虽然平均识别准确率在分类算法中最为常用，但是平均识别准确率不能体现模型在数据集中不同类别的识别效果，对类别不均衡的数据集，使用平均识别准确率易加大误差，因此，通常会结合混淆矩阵来解决这类问题。混淆矩阵以 m 行 m 列表格的形式来表示类别真实值与预测值的关系，其中 m 为数据集的类别总数，也被称为误差矩阵，可以详细地展现分类算法在数据集各个类别上的识别准确率。混淆矩阵的对角线元素表示对应类别正确识别样本的数量占比，其他位置的元素表示该类别被错误识别成其他类别的概率，表情识别任务中，混淆矩阵可以直观地展现模型对不同表情的识别性能。

图 3-8 展示了各模型在 CK+ 测试数据集上的混淆矩阵。可见对基准模型，

恐惧类别表情和厌恶类别表情最容易混淆，而本文模型较好地区分了各种表情。



(a) ResNet18　(b) VGG19

(c) lightResNetl　(d) lightResNetm

图 3-8 CK+ 数据集上各模型的混淆矩阵

表 3-7 CK+ 上本文模型与先进模型的对比

模型	准确率
LBP-CNN[44]	95.29%
VMCNN-LSTM[45]	97.5%
MBCC-CNN[46]	98.48%
SACNN-ALSTM[47]	99.07%
lightResNetl	**100%**
lightResNetm	**100%**

将本文提出的 lightResNetl 和 lightResNetm 模型在 CK+ 数据集上的识别准确率与先进模型准确率的比较结果如表 3-7 所示。可见本文模型有较好的识别性能。

3.3.2 FER2013 数据集上的实验结果及分析

虽然本文模型在 CK+ 数据集上达到了很好的识别准确率，但考虑到 CK+ 数据集中的图片数较少，实验室环境下数据集环境和人脸姿态均较统一，且本文模型较基准模型层数更深，可能产生模型的过拟合现象。因此我们进一步分析本文模型在开放环境数据集 FER2013 上的识别性能。

表 3-8 FER2013 上本文模型与基准模型的对比

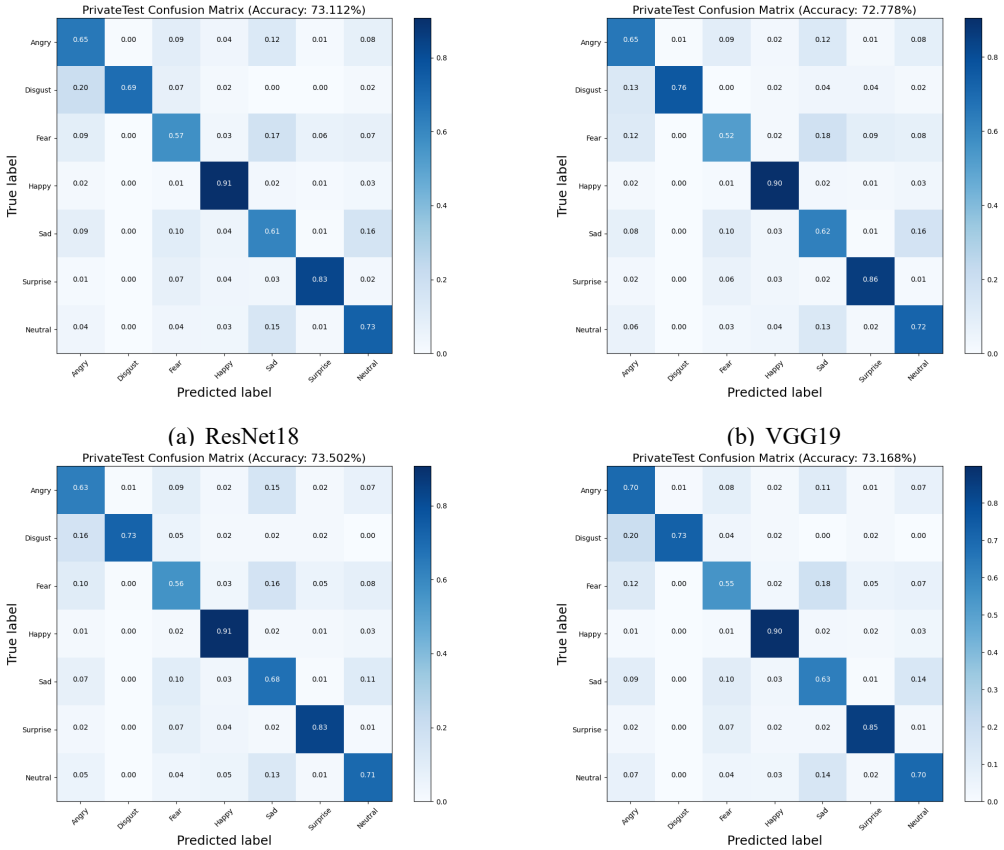
模型	识别准确率
VGG19	72.778%
ResNet18	73.112%
lightResNetl	**73.447%**
lightResNetm	72.168%

表 3-9 FER2013 上本文模型与先进模型的对比

模型	准确率
Conv+Inception[48]	66.4%
SCN[20]	72.67%
Deep-Emotion[49]	70.02%
SACNN-ALSTM[47]	71.31%
RUL[50]	**73.75%**
lightResNetl	**73.502%**
lightResNetm	**72.806%**

FER2013 数据集上本文模型和基准模型以及先进模型的识别准确率对比分别如表 3-8、3-9 所示，可见，由于 FER2013 数据集是在开放环境下采集得到的图片集合，数据集中人脸姿态多元，背景、光照条件多样化，不同模型的

识别准确远不如对 CK+ 数据集的识别准确率。但本文提出的模型 lightResNetl 和 lightResNetm, 尤其是 lightResNetl 仍具有优越的分类性能。



(a) ResNet18　　(b) VGG19

(c) lightResNetl　　(d) lightResNetm

图 3-9 FER2013 数据集上各模型的混淆矩阵

不同模型对 FER2013 数据集的混淆矩阵如图 3-9 所示。可见模型对高兴和惊讶表情的识别准确率最高，而愤怒、恐惧和悲伤表情易被混淆。高兴和惊讶表情的人脸嘴部变化较大，因此更容易被区分出来，而愤怒、恐惧和悲伤表情的人脸动作变化集中在眉心、鼻翼部位，变化较为细微，对模型提取细微特征的能力有更高的要求。

3.3.3 FERm 数据集上的实验结果及分析

FERm 数据集上本文模型和基准模型的识别准确率对比如表 3-10 所示，还选取了部分 FERPlus 数据集上先进模型的实验结果和本文模型进行对比，结果见表 3-11 . 可见本文模型在 FERm 数据集上的识别性能均高于其他模型，

而且，对比 FERm 数据集上的准确率和 FER2013 数据集上的识别准确率可发现，本文对数据集做的改进明显提升了各模型的性能。

表 3-10 FERm 上本文模型与基准模型的对比

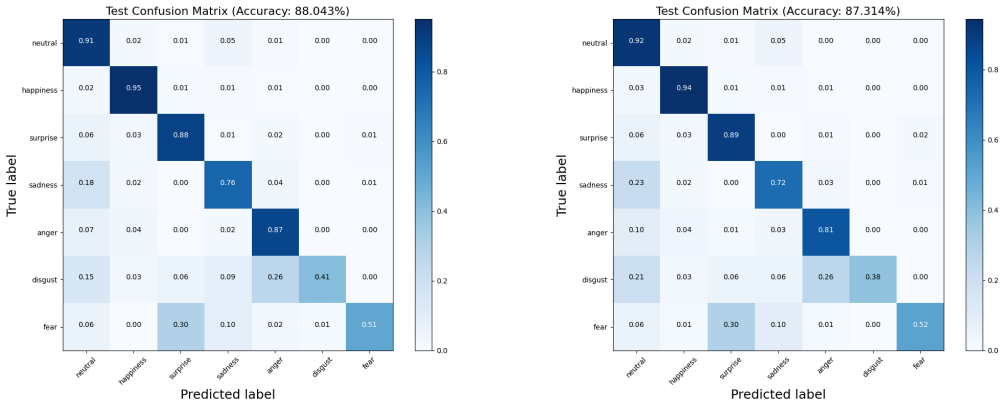
模型	识别准确率
VGG19	87.314%
ResNet18	88.043%
lightResNetl	**88.107%**
lightResNetm	**88.107%**

表 3-11 FERm 上本文模型和先进模型的对比

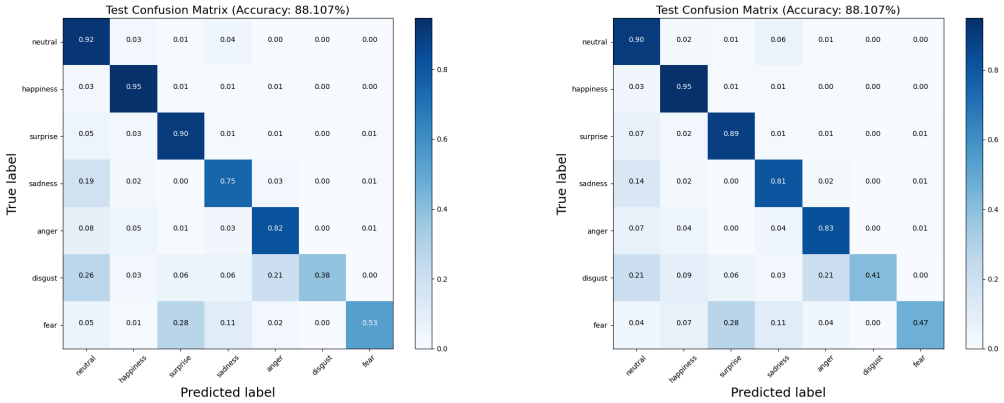
模型	准确率
Densenet[51]	86.54%
Fan[52]	87.6%
MBCC-CNN[46]	88.1%
lightResNetl	**88.107%**
lightResNetm	**88.107%**

不同模型对 FERm 数据集的混淆矩阵如图 3-10 所示。由图 3-10 可知，在 FERm 数据集上，模型对厌恶、恐惧类别表情识别效果最差，而对高兴、中性表情识别效果最好，对比数据集中不同类表情的分布，见表 3-4，可知数据集中高兴、中性表情图片最多，而厌恶、恐惧类别表情图片最少。因此模型对厌恶、恐惧类别表情学习不充分导致识别准确率低。其中厌恶类别图片最容易与中性、愤怒类别图片混淆，恐惧类别图片最容易与惊讶类别图片混淆。

实验中 FERm 数据集上各模型训练的准确率和损失函数随着训练轮数的变化曲线如图 3-11 和图 3-12 所示。可见，在 FERm 数据集上，各模型在训练集上的准确率高，损失函数值低，说明各模型均已训练充分，但在测试集上准确率降低，损失函数值增大，因此各模型仍存在一定的过拟合问题。

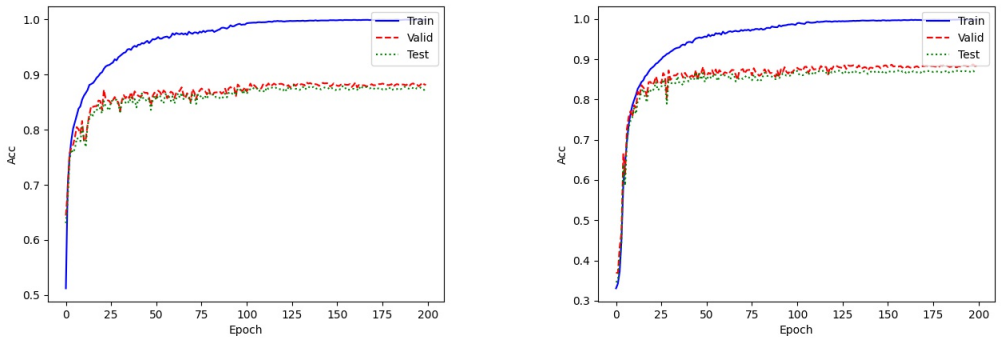


(a) ResNet18　　(b) VGG19



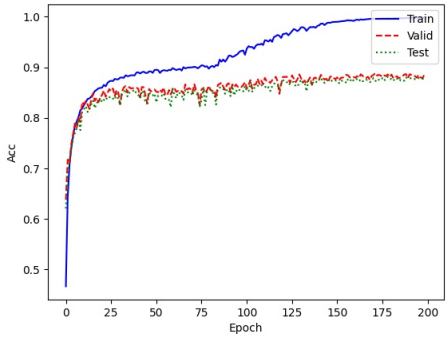
(c) lightResNetl　　(d) lightResNetm

图 3-10 FERm 数据集上各模型的混淆矩阵

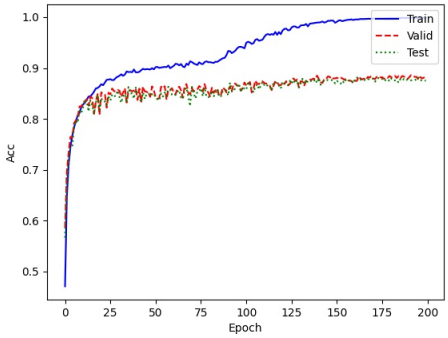


(a) ResNet18　　(b) VGG19

图 3-11 FERm 数据集上训练的准确率曲线

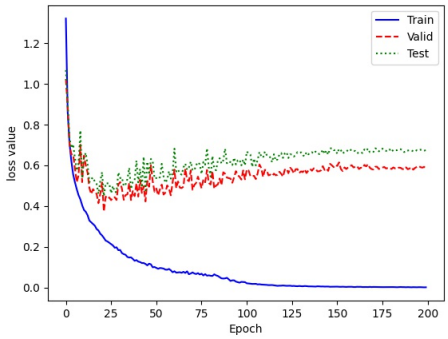


(a) lightResNetl

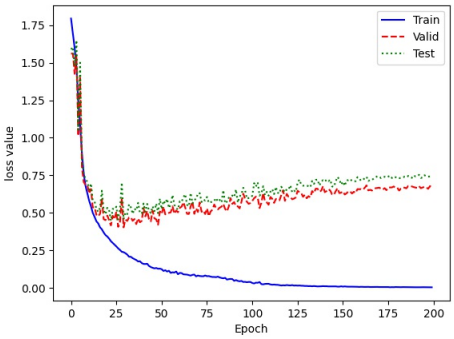


(b) lightResNetm

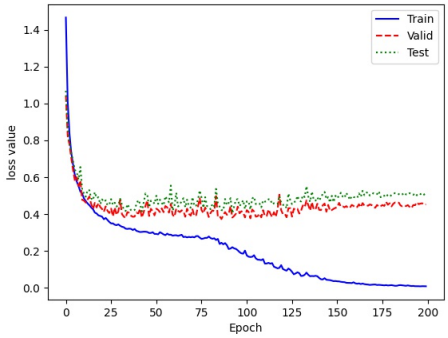
图 3-11 （续图）



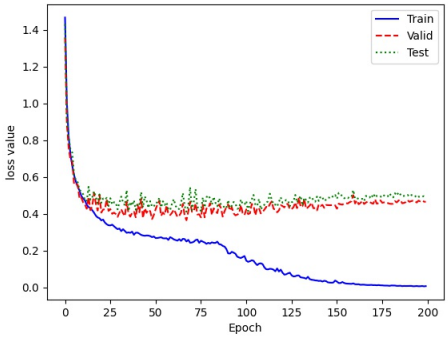
(a) ResNet18



(b) VGG19



(c) lightResNetl



(d) lightResNetm

图 3-12 FERm 数据集上训练的损失函数曲线

3.4 本章小结

本章首先介绍了经典的 ResNet 网络，ResNet 网络中使用残差学习及瓶颈结构很大程度上解决了深层网络的退化问题，接着将 ResNet 模型融合深度可

分离卷积和倒置残差块提出轻量化模型 lightResNetl 和 lightResNetm. 在3.2节介绍了实验所需环境配置和数据集的预处理方法及效果，最后在三个表情识别数据集上对比了本文模型和基准模型的识别性能，并给出了本文模型与先进模型识别准确率的比较，结果表明，本文模型在参数量和计算量远小于主流模型的情况下，在不同数据集上均具有较好的识别效果。

第 4 章 表情识别模型特征可视化分析

基于卷积神经网络进行表情识别的方法由于其高精度、端对端的特点被广泛使用，本文设计了两种轻量化的表情识别模型：lightResNetl 和 lightResNetm, 并在第3章中从识别准确率和混淆矩阵两个方面验证了模型的有效性。然而，CNN 是如何识别不同表情特征的，模型的输入、内部特征以及输出三者之间具有怎样的关系仍是一个难题。

因此，使用 CNN 的特征可视化方法对模型输入、特征编码及输出之间的因果关系以视觉的方式呈现，进一步理解 CNN 内部表征和决策的关系具有重要意义。本章使用两种可视化方法对 lightResNetl 模型进行可视化分析，通过对比不同表情图像以及不同图像变换前后的可视化结果来分析本文模型关注的人脸区域。

4.1 可视化算法

4.1.1 基于有意义的扰动特征可视化方法

基于有意义的扰动特征可视化方法是 Fong 等人[31] 于 2017 年提出，简称为 MP 方法 (Meaningful Perturbation). MP 方法通过给输入图像添加特定的扰动，包括常值扰动、噪声扰动和模糊扰动，将找出图像中与预测结果最相关部分的问题转换为优化问题，即找到最小的删除掩码，使图像经过模型的全连接层后，关于目标类别的得分显著下降。

下面介绍 MP 方法的有关公式及定义。设输入的图像为 $x_0 : \Lambda \to \mathbb{R}^3, \Lambda = \{1, \cdots, H\} \times \{1, \cdots, W\}$ 是一个离散域，$u \in \Lambda$ 是输入图像的每个像素点。扰动算子定义为：

$$[\Phi(x_0; m)](u) = \begin{cases} m(u)x_0(u) + (1 - m(u))\mu_0, & \text{constant} \\ m(u)x_0(u) + (1 - m(u))\eta(u), & \text{noise} \\ \int g_{\sigma_0 m(u)}(v - u)x_0(v)dv, & \text{blur} \end{cases}$$

式中 $m : \Lambda \to [0, 1]$ 为图像掩码，作用于每个像素值得到标量 $m(u)$, μ_0 是输入图像的平均像素值，$\eta(u)$ 是对每个像素点的高斯噪声采样，g_σ 是高斯模糊核，σ_0 是最大各向同性的高斯核标准差。为了找到使图像分类分数明显变化的最小图像区域，即图像的类别重要区域，定义如下目标函数：

$$m^* = \underset{m \in [0,1]^\Lambda}{\operatorname{argmin}} \lambda_1 \|1 - m\|_1 + \lambda_2 \sum_{u \in \Lambda} \|\bigtriangledown m(u)\|_\beta^\beta + \mathbb{E}_\tau \left[f_c \left(\Phi \left(x_0 \left(\cdot - \tau \right), m \right) \right) \right]$$

式中 $\lambda_1, \lambda_2, \beta, \tau$ 为超参数，目标函数第一项控制 m 使扰动区域尽可能小，第二项为正则化项，主要目的为防止掩码中伪影的产生，$f_c(\Phi(x_0; m))$ 为添加干扰后的图片输入网络获得的全连接层分数。

本文实验中使用Adam优化算法，Adam方法综合了SGD-M算法和AdaGrad的思想，将梯度的一阶动量和二阶动量结合起来，实现简单，计算效率高，且对超参数的调节少，具体公式为：

$$
\begin{aligned}
g_t &= \nabla_m L_t(m_{t-1}) \\
\nu_t &= \beta_1 \nu_{t-1} + (1-\beta_1) g_t \\
\eta_t &= \beta_2 \eta_{t-1} + (1-\beta_2) g_t^2 \\
\hat{\nu}_t &= \frac{\nu_t}{1-\beta_1^t} \\
\hat{\eta}_t &= \frac{\eta_t}{1-\beta_2^t} \\
m_t &= m_{t-1} - \alpha \cdot \frac{\hat{\nu}_t}{\sqrt{\hat{\eta}_t} + \varepsilon}
\end{aligned}
$$

式中，t 表示当前迭代次数，L_t 为目标函数，β_1, β_2 为超参数，m_t 为当前迭代更新的扰动区域。

4.1.2 分数加权的类激活图可视化方法

2020 年，Wang 等人[28] 提出了一种无梯度分数加权的类激活图可视化方法，即 Score-CAM 方法。Score-CAM 方法不同于其他类激活映射方法采用池化层和全连接层的神经元权重或反向传播梯度作为激活图的权重，而是通过激活图在目标类别上的前向传播分数来获得权重，从而摆脱了可视化方法对梯度的依赖。这样做既避免了反向传播和模型梯度消失可能产生的噪声，又极大地减少了方法的计算量。

下面我们将详细介绍 Score-CAM 方法的有关定义和公式。设输入图像为 X_0, 输入图像的目标类别为 c, 图像经过模型的输出是 $f(x)$, 基本图片用 X_b 表示，l 是模型的最后一层卷积层。$L^c_{\text{Score-CAM}}$ 表示 Score-CAM 算法得到的显著图，则 Score-CAM 的算法流程如表 4-1 所示。

表 4-1 Score-CAM 的算法流程

Score-CAM 算法流程
输入：$X_0, X_b, f(X), c$ 和 l；
输出：$L^c_{\text{Score-CAM}}$；
初始化：$M \leftarrow [], A_l \leftarrow f_l(X)$；
其中 A_l 为网络最后一层卷积层的特征图，记 C 为 A_l 的特征图通道数；
for k in $[0, \cdots, C-1]$ do:
$M_l^k \leftarrow \text{Upsample}\left(A_l^k\right)$
$M_l^k \leftarrow s\left(M_l^k\right), s\left(M_l^k\right) = \frac{M_l^k - \min M_l^k}{\max M_l^k - \min M_l^k}$
M.append$\left(M_l^k \circ X_0\right)$
end
$M \leftarrow \text{Batchify}(M)$
$S^c \leftarrow f^c(M) - f^c(x_b)$
$\alpha_k^c \leftarrow \frac{\exp\left(S_k^c\right)}{\sum_k \exp\left(S_k^c\right)}$
$L^c_{\text{Score-CAM}} \leftarrow \text{ReLU}\left(\sum_k \alpha_k^c A_l^k\right)$

表中，$\text{Batchify}(\cdot)$ 表示将 M 根据通道数切片，$f^c(\cdot)$ 表输入至模型中得到的目标类别的分数，最后将得到的显著图通过 ReLU 操作是为了保留与输出正相关的像素，去除与输出负相关的像素点。

4.2 对表情识别模型 lightResNetl 的可视化分析

4.2.1 对不同表情图像的可视化分析

本节使用 MP 方法和 Score-CAM 方法对本文提出的模型 lightResNetl 进行可视化分析，通过对比不同人脸表情图像的可视化结果，解释说明表情识别模型重点关注了人脸的哪些部位，在验证本文模型是否可以通过人脸面部动作的不同来区分表情的同时，也帮助理解表情识别模型输入图像与决策之间的相关性。

本实验在 Ubuntu 操作系统下进行，采用的显卡为 GeForce GTX 1060, 内存

10G, 使用的深度学习框架为 PyTorch1.7.0+cu110. 实验中各超参数设置参考文献[28,31]，即设置 MP 方法中，Adam 方法的学习率为 0.1, 迭代次数为 500. 设置 Score-CAM 方法中基本图像为全黑图像，上采样方法为双线性插值法。

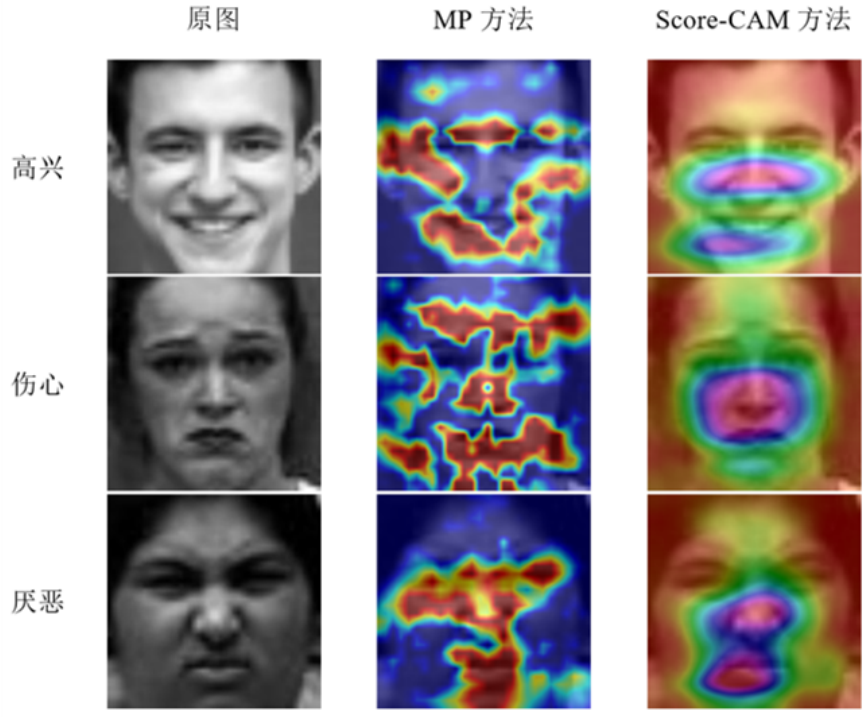


图 4-1 不同表情可视化结果对比

不同表情图像在 lightResNetl 模型上的可视化结果如图 4-1 所示。实验中主要选取了高兴、伤心和厌恶三种表情的可视化结果进行对比，可见，对同一表情图像，MP 方法和 Score-CAM 方法识别出的重要区域基本保持一致，表情识别模型 lightResNetl 主要学习了人脸图像的局部特征，即人脸的眉心、鼻翼和嘴巴部分重要程度高于人脸其他部位，这与人类的视觉习惯相符，这一结果说明本文模型 lightResNetl 的确学习到了表情识别的关键性特征。

此外，对比不同表情的可视化结果，lightResNetl 识别高兴表情的重要区域集中在嘴巴和眼角，识别厌恶表情的重要区域相比高兴表情增加了鼻翼部分，而识别伤心表情的重要区域则为眉心和嘴巴位置，这些重要区域与人类识别不同表情区分的位置基本保持一致，可见本文模型对人脸重要区域的定位有良好的性能，这也验证了本文模型对表情识别的有效性。

4.2.2 基于可视化方法对输入图像变换的可靠性分析

本节旨在通过可视化方法分析本文模型对输入图像变换的鲁棒性，从而进一步验证本文模型 lightResNetl 有效地学习了人脸表情特征。本节分析的图像变换包括图像亮度、对比度变换、图像旋转、水平翻转和垂直翻转，由第3章可知，lightResNetl 模型在训练过程中采用的数据增强方法有图像亮度变换、对比度变换、图像旋转以及水平翻转，因此，本文模型应对以上数据增强方法具有较好的鲁棒性，但由于模型训练中没有学习到图像垂直翻转的变换，本文模型对垂直翻转的人脸图像识别效果应不如其他图像变换。

1、原图和不同亮度表情图像实验结果对比

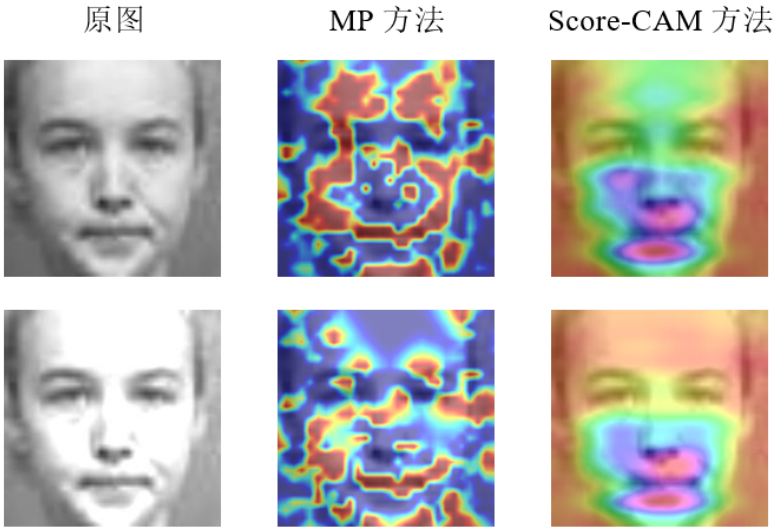


图 4-2 蔑视表情图像与改变亮度图像可视化结果对比

图 4-2 是蔑视图像原图和改变亮度图像的可视化结果对比，其中原图以 0.943 的概率被识别为蔑视，改变亮度后，图像以 0.908 的概率被识别为蔑视。可见本文模型对图像的亮度变换具有一定的鲁棒性，亮度变换并没有影响图像的分类结果，且从可视化图像上看，改变亮度后，模型对图像识别的重要区域更加集中在脸的鼻子和嘴巴部位，因此，由本实验可得对蔑视表情的分类，眉心处的特征也具有较高的重要性。

2、原图和不同对比度表情图像实验结果对比

图 4-3 为高兴表情的原图和改变对比度后的图像可视化结果对比，其中原图以 0.99996 的概率被识别为高兴表情，改变对比度图像以 0.99997 的概率被识别高兴表情，因此本文模型对图像的对比度变换也具有一定的鲁棒性，

从可视化结果看，高兴表情的重要区域为鼻子和嘴巴部位，眉心处的特征对分类结果没有较大影响，且改变对比度的表情图像可视化结果对人脸边缘处的关注度降低，去除了冗余特征，也提升了高兴表情的分类概率。

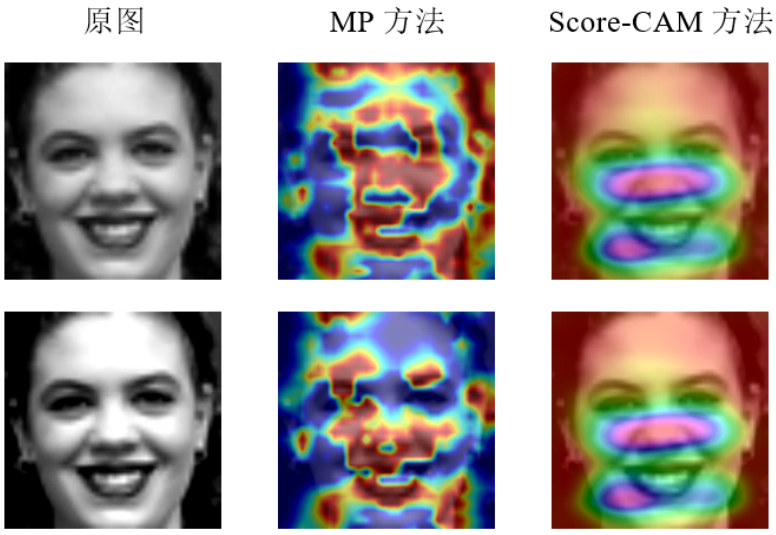


图 4-3 高兴表情图像与改变图像对比度可视化结果对比

3、原图和旋转表情图像实验结果对比

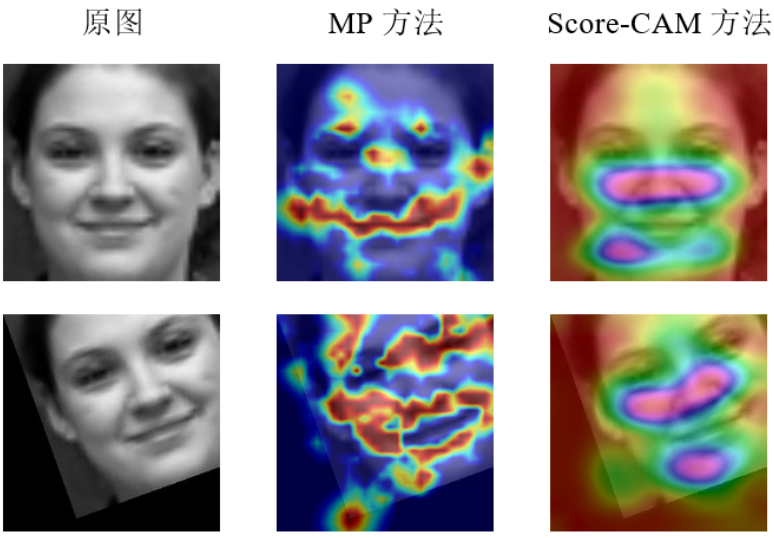


图 4-4 高兴表情图像与旋转图像可视化结果对比

图 4-4 为高兴表情和其旋转图像的可视化实验结果对比，其中原图以 0.99 的概率被识别为高兴表情，旋转图像以 0.96 的概率被识别为高兴表情。因此，本文模型对图像的旋转变换具有一定的鲁棒性，且从可视化结果来看，

模型对高兴表情的识别也与我们的分析相符，模型更关注人脸的嘴鼻部位特征，眉心等部位的特征对高兴表情的识别重要程度不大。

4、原图和水平翻转表情图像实验结果对比

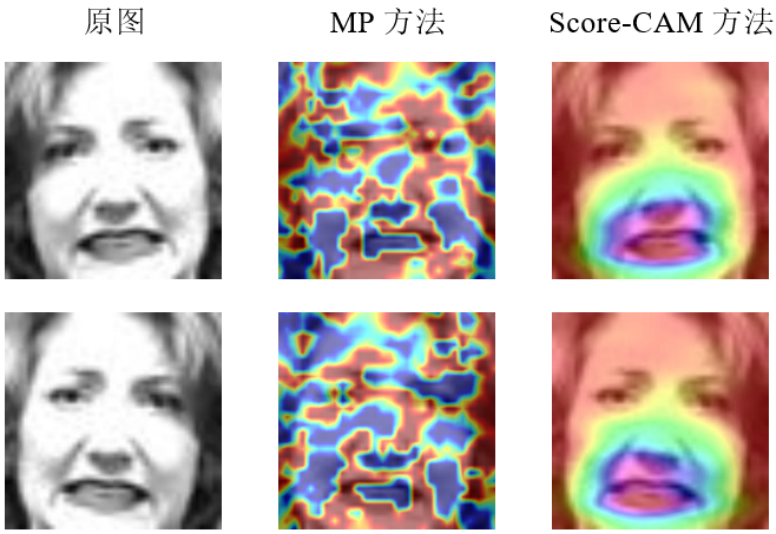


图 4-5 恐惧表情图像与水平翻转图像可视化结果对比

图 4-5 为恐惧表情和其水平翻转的人脸图像可视化结果对比，其中原图以 0.902 的概率被识别为恐惧，改变对比度后人脸图像以 0.906 的概率被识别为恐惧，从可视化结果看，本文模型对图像的水平翻转具有一定的鲁棒性，图像水平翻转前后的重要区域基本保持一致，且对恐惧表情，人脸眼睛处的重要性相比其他表情要更加明显。

5、原图和垂直翻转表情图像实验结果对比

如图 4-6 为厌恶表情图像原图和垂直翻转的可视化结果对比，其中原图以 0.99 的概率被分类为厌恶，而垂直翻转图像以 0.64 的概率被分类为愤怒，结合可视化结果来看，虽然图像垂直翻转后，模型仍能较好地定位到人脸的口鼻位置，但由于训练时模型没有学习到垂直翻转图像的特征，且数据集中厌恶和愤怒的人脸面部动作改变相差不大，数据集中厌恶和愤怒的表情图像对比如图 4-7 所示，因此模型不能将垂直翻转后的人脸图像准确分类。由此可知，人脸不同表情图像的变化较小，要使模型能对表情进行准确地分类，不仅需要增大训练数据集的数量，对训练数据集做多样的数据增强也是必要的。

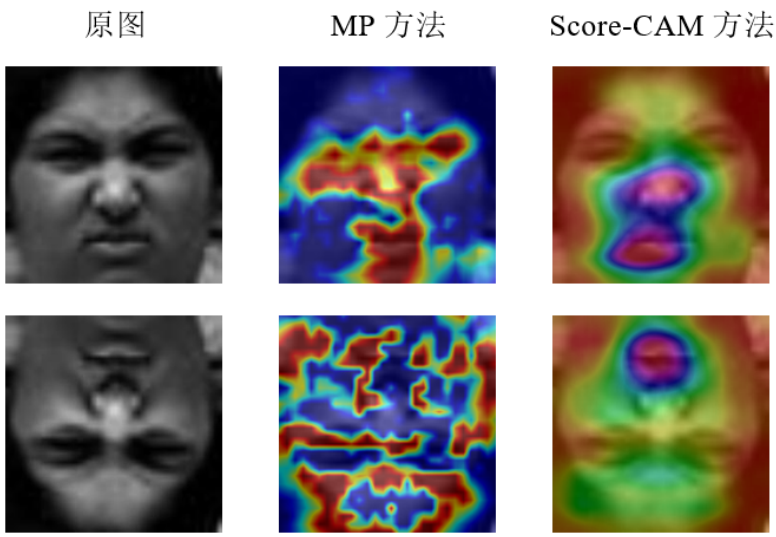


图 4-6 厌恶表情图像与垂直翻转图像可视化结果对比

图 4-7 厌恶与愤怒表情图像对比

4.3 本章小结

本章通过基于有意义的干扰和分数加权的类激活图可视化方法对本文提出的表情识别模型进行解释，首先分析了高兴、伤心和厌恶表情的可视化结果，以此验证本文模型对表情识别的有效性，也说明了表情识别模型学习人脸的局部特征，对人脸的主要关注区域在面中部位，对不同的表情模型侧重区域也有所不同。接着分析了本文模型对输入图像的各种图像变换的鲁棒性，结果表明，本文模型对图像亮度、对比度变化、旋转和水平翻转都有较

好的适应性，但模型对垂直翻转的人脸图像分类结果不准确，由此验证了对表情识别这种需要关注细微特征的分类任务，对图像进行数据增强的必要性，也进一步验证了模型 lightResNetl 经过训练和测试有效地学习到了表情识别的关键特征。

结　论

针对表情识别问题，本文设计了一个轻量化的卷积神经网络，达到了模型速度和精度的平衡，又对本文设计的表情识别模型进行可视化分析，探索了表情识别模型输入与输出的因果关系，进一步验证了本文模型的有效性和可靠性。具体来说，本文取得的研究成果有以下两点：

第一，针对表情识别领域模型较大，难以推广应用到移动端设备的问题，本文将主流卷积神经网络 ResNet 和轻量化模块（深度可分离卷积以及倒置残差结构）融合得到轻量化网络 lightResNetl 和 lightResNetm. 在实验室数据集 CK+ 以及公开环境数据集 FER2013 上进行表情识别测试，并针对 FER2013 数据集存在标签错误和噪声图片的问题，本文对数据集 FERPlus 做了进一步的改进得到数据集 FERm, 实验结果表明，本文提出的轻量化模型在模型参数量和计算量远小于主流模型的前提下，识别准确率达到甚至超过基准模型和先进模型。

第二，针对表情识别领域模型可解释性不强，输入图片与模型决策的因果关系不明的问题，本文采用 MP 方法和 Score-CAM 方法对 lightResNetl 模型进行可视化分析。通过对不同表情图像的可视化分析表明表情识别模型重点关注了人脸的眉心、口鼻位置，且对不同表情的侧重区域有所不同，这也说明了本文模型对表情识别的有效性。通过对表情图像变换前后的可视化结果对比可知，训练过程中的数据增强确实可以增强模型对图像变换的鲁棒性，这也说明了本文模型对表情识别任务的可靠性。

尽管本文对表情识别模型的轻量化和可视化分析进行了深入的研究，还有一些问题值得进一步地思考：

首先，相比细粒度图像分类、目标检测等问题来说，大多表情识别领域的数据集体量较小，这造成了在训练过程中模型易产生过拟合问题。尽管本文模型采用了批归一化、Dropout 等方法防止模型的过拟合，受数据集的限制，过拟合现象仍然存在。一方面，未来可以尝试开发使用更大规模的表情识别数据集，另一方面，可以研究改进针对小规模数据集的训练方法。

第二，从本文对表情识别模型的可视化结果来看，模型只关注人脸图像的局部特征，因此对表情的准确分类有赖于模型对细微特征的提取能力。而表情图像的变换均对模型的特征提取有一定的干扰，在实际应用中，如何减少环境、光照等外在因素对模型特征提取的影响仍有待研究。

参考文献

[1] Ekman P, Friesen W V. Constants Across Cultures in The Face and Emotion.[J]. Journal of Personality and Social Psychology, 1971, 17(2): 124.

[2] Matsumoto D. More Evidence for the Universality of a Contempt Expression[J]. Motivation and Emotion, 1992, 16(4): 363–368.

[3] Hernandez-Matamoros A, Bonarini A, Escamilla-Hernandez E, et al. Facial Expression Recognition with Automatic Segmentation of Face Regions Using a Fuzzy Based Classification Approach[J]. Knowledge-Based Systems, 2016, 110: 1–14.

[4] Ojala T, Pietikainen M, Harwood D. Performance Evaluation of Texture Measures with Classification Based on Kullback Discrimination of Distributions[C]// Proceedings of 12th International Conference on Pattern Recognition. IEEE, 1994, 1: 582–585.

[5] 党宏社, 王淼, 张选德. 基于深度学习的面部表情识别方法综述[J]. 科学技术与工程, 2020, 20(24): 9724.

[6] Krizhevsky A, Sutskever I, Hinton G E. Imagenet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems, 2012, 25(2).

[7] Tang Y. Deep Learning Using Support Vector Machines[J]. CoRR, abs/1306.0239, 2013, 2: 1.

[8] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-scale Image Recognition[J]. International Conference on Learning Representations, 2015.

[9] Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2015: 1–9.

[10] Yu Z, Zhang C. Image Based Static Facial Expression Recognition with Multiple Deep Network Learning[C]// Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. 2015: 435–442.

[11] Kim B K, Lee H, Roh J, et al. Hierarchical Committee of Deep CNNs with Exponentially-weighted Decision Fusion for Static Facial Expression Recognition[C]// Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. 2015: 427–434.

[12] Bargal S A, Barsoum E, Ferrer C C, et al. Emotion Recognition in the Wild from Videos Using Images[C]// Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016: 433–436.

[13] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770–778.

[14] Ding H, Zhou S K, Chellappa R. Facenet2expnet: Regularizing A Deep Face Recognition Net for Expression Recognition[C]// 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017: 118–126.

[15] Cai J, Chang O, Tang X L, et al. Facial Expression Recognition Method Based on Sparse Batch Normalization CNN[C]// 2018 37th Chinese Control Conference (CCC). IEEE, 2018: 9608–9613.

[16] Cai J, Meng Z, Khan A S, et al. Island loss for learning discriminative features in facial expression recognition[C]// 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018: 302–309.

[17] Kuo C M, Lai S H, Sarkis M. A Compact Deep Learning Model for Robust Facial Expression Recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018: 2121–2129.

[18] Cai Y, Gao J, Zhang G, et al. Efficient Facial Expression Recognition Based on Convolutional Neural Network[J]. Intelligent Data Analysis, 2021, 25(1): 139–154.

[19] Georgescu M I, Ionescu R T, Popescu M. Local Learning with Deep and Handcrafted Features for Facial Expression Recognition[J]. IEEE Access, 2019, 7: 64827–64836.

[20] Wang K, Peng X, Yang J, et al. Suppressing Uncertainties for Large-scale Facial Expression Recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6897–6906.

[21] Wu Y, Jia K, Sun Z. Lightweight Facial Expression Recognition Based on Multi-Scale Dense Convolutional Neural Network[C]// 2021 10th International Conference on Computing and Pattern Recognition. 2021: 5–11.

[22] Huang G, Liu Z, Van Der Maaten L, et al. Densely Connected Convolutional Networks[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 4700–4708.

[23] Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps[J]. Computer Science, 2013.

[24] Zeiler M D, Fergus R. Visualizing and Understanding Convolutional Networks[C]// European Conference on Computer Vision. Springer, 2014: 818–833.

[25] Zhu Y, Fan H, Yuan K. Classification Mechanism of Convolutional Neural Network for Facial Expression Recognition[C]// International Conference on Pattern Recognition. Springer, 2021: 717–729.

[26] Zhou B, Khosla A, Lapedriza A, et al. Learning Deep Features for Discriminative Localization[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921–2929.

[27] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual Explanations from Deep Networks via Gradient-Based Localization[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 618–626.

[28] Wang H, Wang Z, Du M, et al. Score-CAM: Score-Weighted Visual Explanations For Convolutional Neural Networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 24–25.

[29] Patro B N, Lunayach M, Namboodiri V P. Uncertainty Class Activation Map (U-CAM) Using Gradient Certainty Method[J]. IEEE Transactions on Image Processing, 2021, 30: 1910–1924.

[30] Omeiza D, Speakman S, Cintas C, et al. Smooth Grad-Cam++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models[J]. arXiv preprint arXiv:1908.01224, 2019.

[31] Fong R C, Vedaldi A. Interpretable Explanations of Black Boxes by Meaningful Perturbation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 3429–3437.

[32] Huang Z, Li Y. Interpretable and Accurate Fine-Grained Recognition via Region Grouping[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8662–8672.

[33] Wagner J, Kohler J M, Gindele T, et al. Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9097–9107.

[34] Oramas J, Wang K, Tuytelaars T. Visual Explanation by Interpretation: Improving Visual Feedback Capabilities of Deep Neural Networks[J]. CoRR, 2017, abs/1712.06302.

[35] Agarwal C, Schonfeld D, Nguyen A. Removing Input Features via A Generative Model to Explain Their Attributions to Classifier's Decisions[J]. CoRR, 2019, abs/1910.04256.

[36] Fukushima K, Miyake S. Neocognitron: a Self-Organizing Neural Network Model for A Mechanism of Visual Pattern Recognition[C]// Proceedings of the US–Japan joint seminar. 1982: 267–85.

[37] 邱锡鹏. 神经网络与深度学习[M]. 机械工业出版社, 2020.

[38] Xiang J, Zhu G. Joint Face Detection and Facial Expression Recognition with MTCNN[C]// 2017 4th International Conference on Information Science and Control Engineering (ICISCE). IEEE, 2017: 424–427.

[39] Xu Y, Yan W, Yang G, et al. CenterFace: Joint Face Detection and Alignment Using Face as Point[J]. Scientific Programming, 2020, 2020: 1–8.

[40] Deng J, Guo J, Ververas E, et al. Retinaface: Single-Shot Multi-Level Face Localisation in the Wild[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5203–5212.

[41] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005, 1: 886–893.

[42] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[43] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted Residuals and Linear Bottlenecks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4510–4520.

[44] Shao J, Qian Y. Three Convolutional Neural Network Models for Facial Expression Recognition in the Wild[J]. Neurocomputing, 2019, 355: 82–92.

[45] Zhang H, Huang B, Tian G. Facial Expression Recognition Based on Deep Convolution Long Short-Term Memory Networks of Double-Channel Weighted Mixture[J]. Pattern Recognition Letters, 2020, 131: 128–134.

[46] Shi C, Tan C, Wang L. A Facial Expression Recognition Method Based on a Multi-branch Cross-Connection Convolutional Neural Network[J]. IEEE Access, 2021, 9: 39255–39274.

[47] Liu C, Hirota K, Ma J, et al. Facial Expression Recognition Using Hybrid Features of Pixel and Geometry[J]. IEEE Access, 2021, 9: 18876–18889.

[48] Mollahosseini A, Chan D, Mahoor M H. Going Deeper in Facial Expression Recognition Using Deep Neural Networks[C]// 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016: 1–10.

[49] Minaee S, Minaei M, Abdolrashidi A. Deep-emotion: Facial Expression Recognition Using Attentional Convolutional Network[J]. Sensors, 2021, 21(9): 3046.

[50] Zhang Y, Wang C, Deng W. Relative Uncertainty Learning for Facial Expression Recognition[J]. Advances in Neural Information Processing Systems, 2021, 34.

[51] Miao S, Xu H, Han Z, et al. Recognizing Facial Expressions Using a Shallow Convolutional Neural Network[J]. IEEE Access, 2019, 7: 78000–78011.

[52] Fan X, Deng Z, Wang K, et al. Learning Discriminative Representation for Facial Expression Recognition from Uncertainties[C]// 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020: 903–907.

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于轻量化卷积神经网络的表情识别与可视化分析》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：王聪荣　　日期：2022年 6 月 20 日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其它复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：王聪荣　　日期：2022年 6 月 20 日

导师签名：[illegible]　　日期：2022 年 6 月 20 日

致　谢

时光飞逝，两年的研究生生涯在此便要落下帷幕，在这两年里，我成长了许多，在本论文完成之际，衷心感谢两年中帮助过我的老师们和同学们。正因为有着许多良师益友的陪伴，我才得以拥有这段美好而充实的校园时光。

首先我要由衷地感谢我读研期间的导师孙杰宝教授，孙老师知识渊博，治学严谨，教学风趣，平易近人，读研期间一直鼓舞我、激励我，他的言传身教使我终生受益。

感谢郭志昌老师和姚文娟老师，在完成学位论文的整个过程中，从选题、开题到定稿，两位老师时刻关注着我的写作进展，耐心地给我提出指导建议，本文的成型离不开两位老师的精心指导，再次感谢两位老师的指导与鼓励，在未来的学习生活中我会更加努力，认真学习老师们严谨的科研态度与精神。

感谢吴勃英老师，吴老师待人友善，作为数学学院科学与工程计算课题组的总负责人，为我们开展了许多学术讲座，给我们提供了良好的学习环境和实验环境。感谢课题组的张达治老师、李佳老师和李爻老师，初入研究生学习阶段，是课题组的老师和同学们的帮助让我尽快地适应了高强度的学习，走出了对身份转变的迷茫。这两年来我不仅学习了知识，科研能力得到了很大的提高，也学会了许多待人处事的道理，收获了视野和阅历。

感谢课题组的所有老师以及研究生期间所有的任课老师，“规格严格，功夫到家”在老师们身上得到了最好的诠释。课堂上感受到的老师们对数学的热爱与激情是我学习期间最宝贵的财富。

感谢我的室友们和数学系的同学们在学习和生活上对我的帮助，祝愿大家在未来的日子里一切顺利。

感谢吴龙浩同学，感谢他的陪伴和鼓励。最后，感谢我的父母与亲人，因为他们的支持，我才得以一往无畏。我坚信阳光总在风雨后，祝愿大家平安喜乐。未来的日子，我将心怀感恩，更加努力地生活。