计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(7):128−134 [doi: 10.15888/j.cnki.csa.008554]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Mixer Layer 的人脸表情识别①

简腾飞, 王　佳, 曹少中, 杨树林, 张　寒

(北京印刷学院 信息工程学院, 北京 102600)
通信作者: 王　佳, E-mail: wangjia@bigc.edu.cn

摘　要: 目前人脸表情识别研究多数采用卷积神经网络 (CNN) 提取人脸特征并分类, CNN 的缺点是网络结构复杂, 消耗计算资源. 针对以上缺点, 本文采用基于多层感知机 (MLP) 的 Mixer Layer 网络结构用于人脸表情识别. 采用数据增强和迁移学习方法解决数据集样本不足的问题, 搭建了不同层数的 Mixer Layer 网络. 经过实验比较, 4 层 Mixer Layer 网络在 CK+和 JAFFE 数据集上的识别准确率分别达到了 98.71% 和 95.93%, 8 层 Mixer Layer 网络在 Fer2013 数据集上的识别准确率达到了 63.06%. 实验结果表明, 无卷积结构的 Mixer Layer 网络在人脸表情识别任务上表现出良好的学习能力和泛化能力.

关键词: 深度学习; 迁移学习; 表情识别; Mixer Layer; 图像识别

引用格式: 简腾飞,王佳,曹少中,杨树林,张寒.基于 Mixer Layer 的人脸表情识别.计算机系统应用,2022,31(7):128–134. http://www.c-s-a.org.cn/1003-3254/8554.html

Facial Expression Recognition Based on Mixer Layer

JIAN Teng-Fei, WANG Jia, CAO Shao-Zhong, YANG Shu-Lin, ZHANG Han

(School of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China)

Abstract: At present, most facial expression recognition research uses a convolutional neural network (CNN) to extract facial features and classify them. The disadvantage of CNN is that its network structure is complex and consumes substantial computing resources. In response, this study uses the Mixer Layer network structure based on multilayer perceptron (MLP) for facial expression recognition. Data augmentation and transfer learning methods are employed to solve the problem of insufficient data set samples, and Mixer Layer networks with different layers are built. According to experimental comparison, the recognition accuracy of the 4-layer Mixer Layer network on CK+ and JAFFE data sets reach 98.71% and 95.93% respectively, and that of the 8-layer Mixer Layer network on Fer2013 data set is 63.06%. The experimental results show that the Mixer Layer networks without a convolution structure exhibit sound learning and generalization abilities in facial expression recognition tasks.

Key words: deep learning; transfer learning; expression recognition; Mixer Layer; image recognition

人脸表情是反映人类情感最普遍最重要的方式之一, 面部表情传达着人与人之间的社会和情感信息, 面部基本表情可分为 6 种 (快乐, 悲伤, 惊讶, 恐惧, 愤怒和厌恶). 随着人工智能和深度学习的兴起, 基于深度学习的人脸表情识别得到了广泛的发展和应用, 基于传统特征提取方法的人脸表情识别, 需要大量专业知识来设计提取器, 同时传统方法的泛化能力和鲁棒性相对于深度学习的方法略有不足. 神经网络可以获得表情图像中更抽象, 更复杂的特征, 使识别更加准确. 随着深度学习的发展, 基于卷积神经网络的人脸表情识别, 取得了巨大的进步.

Shi 等[1] 基于 ResNet 提出一种多分支交叉卷积神

① 基金项目: 北京市自然基金和北京市教委联合项目 (KZ202010015021); 北京市教育委员会科研计划 (KM201910015003); 北京印刷学院科研项目 (Ec202002, Eb202103)
收稿时间: 2021-09-18; 修改时间: 2021-10-29; 采用时间: 2021-11-07; csa 在线出版时间: 2022-05-31

经网络 (MBCC-CNN) 提高了每个感受野的特征提取能力, 在 CK+数据集上的识别准确率达到了 98.48%. Li[2] 利用 ResNet-101 使用文献 [3] 中的数据集识别准确率达到了 96.29%±0.78%. 魏赟等[4] 提出了一种引入注意力机制的轻量级 CNN 通道和卷积自编码器预训练通道的双通道模型, 在减少模型参数量的同时也保证了识别准确率. 江大鹏等[5] 提出局部二值模式 (LBP) 图像的卷积网络对 6 种面部表情识别, 通过 Viola-Jones 框架提取出面部表情感兴趣区域, 获得感兴趣区域的 LBP 图像, 再输入到卷积网络进行识别. 申毫等[6] 基于残差网络提出一种轻量卷积网络的多特征融合的人脸表情识别方法, 使用改进的倒置残差网络为基本单元, 搭建轻量级卷积网络, 用 11 层的卷积筛选网络中的浅层特征, 该模型的参数量仅有 0.2×10^6, 但在 RAD-DB 数据集上的识别准确率达到了 85.46%. 伊力哈木·亚尔买买提等[7] 提出了一种融合局部特征与深度置信网络 (DBN) 的人脸面部表情识别算法, 融合表情局部敏感质量分布图 (LSH) 非均匀光照不变特征和人脸面部表情的边缘局部细节纹理特征, 把融合后特征用于训练深度置信网络 (DBN) 模型, 在 JAFFE 数据集上达到了 97.56% 的识别率. 崔子越等[8] 通过改进 VGGNet 结合 Focal loss 的方法来处理面部表情数据集样本不均衡, 防止网络过拟合, 在数据集 CK+, JAFFE, Fer2013 上相比于传统的损失函数, 模型的准确率提升了 1%–2%, 模型的分类能力更加均衡. 在保证识别准确率的情况下, 张宏丽等[9] 通过优化剪枝 GoogLeNet 识别人脸表情, 以达到简化网络结构的参数量, 提高运行效率, 网络运行时间低于 200 ms. Dhankhar[10] 组合了 ResNet-50 和 VGG16 用于人脸表情识别, 在数据集 KDEF 上取得了较好的效果.

可以看出, 对于人脸表情识别的研究方法, 目前大多数是基于卷积神经网络, 同时对数据进行了一定预处理. 本文通过搭建无卷积结构的浅层神经网络对人脸表情进行识别, 该模型结构简单, 计算复杂度低.

1 人脸表识别方法

1.1 MLP-Mixer 网络结构

2021 年 Google 提出来一种无卷积和注意力机制的网络 MLP-Mixer[11], 网络结构如图 1[11] 所示.

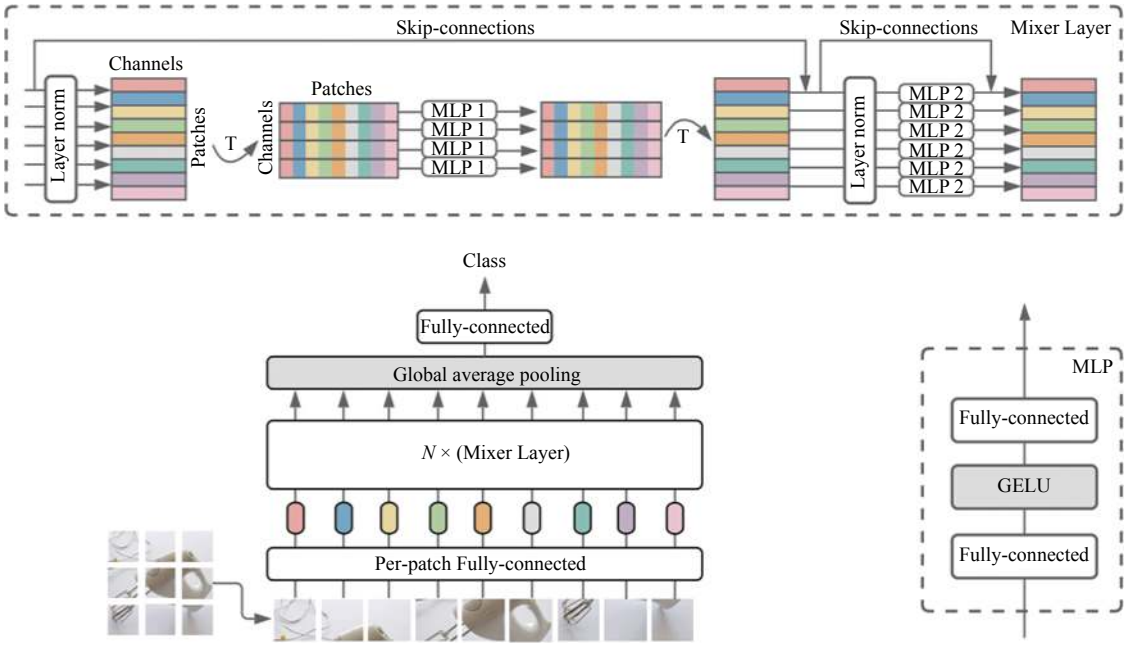


图 1 MLP Mixer 网络结构

图 1 展示了 MLP-Mixer 的网络结构, MLP-Mixer 网络的输入是一串不重复的图片块序列 S, 把每一个图片块映射成指定的维度 C, Mixer Layer 的输入维度为 $X\in\mathbb{R}^{S\times C}$. 假设输入的图片的分辨率为 (H, W), 每个图片块的分辨率为 (P, P), 则 $S=(H\times W)/P^2$. Mixer Layer 接受一系列的线性投影的图像块, 且输入输出形状保

持为 $X \in \mathbb{R}^{S \times C}$. Mixer Layer 由两种 MLP (多层感知机) 组成: token-mixing (MLP1) 和 channel-mixing (MLP2).

每个 MLP 包含两个全连接层. channel-mixing 将不同的通道之间联系起来, token-mixing 寻找图片上不同空间位置的关系. MLP-Mixer 的整体结构包括 Per-patch Fully-connected, Mixer Layer 和 Global Average Pooling. Per-patch Fully-connected 将分割的图片块映射为指定维度. 网络包含 GELU[12] 非线性激活函数, 跨越连接和 Layer Normal 等结构. Mixer Layer 可表示为式 (1).

$$\begin{cases} U_{*,i} = X_{*,i} + W_2\sigma(W_1\partial(X)_{*,i}),\ i = 1,2,\cdots,C \\ Y_{j,*} = U_{j,*} + W_4\sigma\left(W_3\partial(U)_{j,*}\right),\ j = 1,2,\cdots,S \end{cases} \tag{1}$$

其中, σ 表示 GELU 激活函数, W 为感知机权重, ∂ 为 Layer Normal. 分别用 D_C 和 D_S 表示感知机 channel-mixing 和 token-mixing 中全连接层的节点个数.

1.2 迁移学习

迁移学习是从源域传输信息提高目标域的学习训练效率, 迁移学习的源域和目标域担任的任务要相同, 在深度学习中, 迁移学习多用于解决数据量少, 训练样本不充分这一问题, 在图像识别领域被广泛运用.

用 Mixer Layer 代替 CNN , 使用 ExpW 数据集预训练主干网络, 将新的表情样本输入到网络中进行微调. 实验证明, 通过该方法训练完成的模型具有较好的表情识别效果, 具体步骤如图 2 所示.

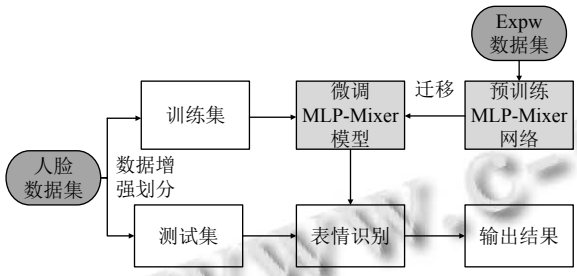


图 2 人脸表情识别方法结构图

2 实验过程

实验运行环境: Windows 10 (64 位) 操作系统, Intel(R) Xeon(R) Gold 6132 CPU, GPU 为 NVIDIA GeForce RTX 2080 Ti 显存大小为 11 GB, Python 版本为 3.7.0.

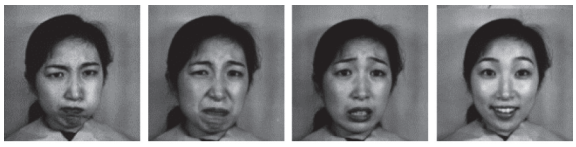
2.1 人脸表情数据集

为了说明该方法的有效性, 采用日本女性面部表情数据集 (JAFFE), CK+ (Extended Cohn-Kanada) 数据集和 Fer2013 数据集进行实验. 实验采用的样本数量分布如表 1 所示.

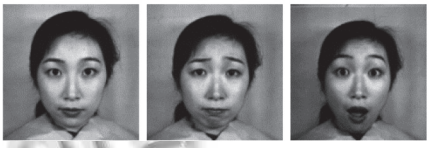
表 1 CK+、JAFFE、Fer2013 数据集实验样本选取数量分布表

数据集	Angry	Neutral	Disgust	Fear
JAFFE	30	30	29	32
CK+	135	0	177	75
Fer2013	4055	6189	546	5133
数据集	Happy	Sad/Sadness	Surprise	Contempt
JAFFE	31	31	30	0
CK+	207	84	249	54
Fer2013	8758	6074	3995	0

其中 JAFFE 数据集包含 10 位日本女性, 每个人做出 7 种表情, 一共包含 213 张大小为 256×256 的人脸正面图像, 共分为 angry, disgust, fear, happy, sad, surprise, neutral (愤怒, 厌恶, 恐惧, 高兴, 悲伤, 惊讶, 自然) 7 种标签. 该数据集的样本分布均匀, 标签准确, 如图 3 所示.



(a) Angry (b) Disgust (c) Fear (d) Happy



(e) Neutral (f) Sad (g) Surprise

图 3 JAFFE 数据集样例图

CK+ 数据集包含 123 个对象的 327 个被标记的表情图片序列, 包含 angry, disgust, fear, happy, sadness, surprise, contempt (愤怒, 厌恶, 恐惧, 高兴, 悲伤, 惊讶, 蔑视) 7 种标签. 每一个图片序列的最后一帧被提供了表情标签, 所以共有 327 个图像被标记. 该数据集样本分布较为不均匀, 如图 4 所示.

Fer2013 数据集总共有 35886 张人脸表情组成, 分为 angry, disgust, fear, happy, neutral, sad, surprise (愤怒, 厌恶, 恐惧, 高兴, 自然, 悲伤, 惊讶) 7 种表情, 其中包含训练集 28708 张, 共有验证集和私有验证集各 3589 张, 每张图片的固定大小为 48×48 的灰度图, 该样本数据分布不均衡且样本中包含了错误样本, 较为混乱, 分类难度大, 如图 5, 图 6 所示.

图 4 CK+ 数据集样例图

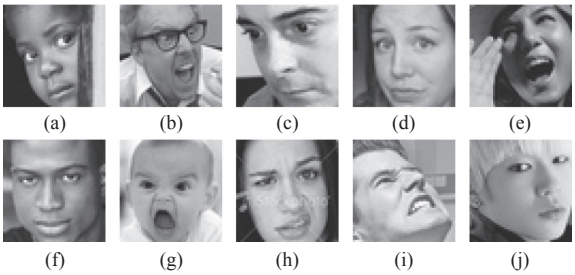


图 5 Fer2013 数据集样例图

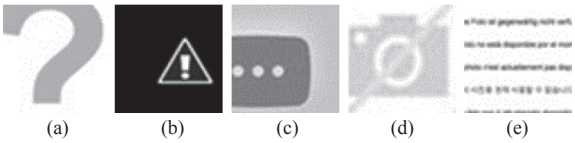


图 6 Fer2013 数据集错误样本样例图

2.2 数据增强

由表 1 可知 CK+和 JAFFE 数据集样本数量较少, 为了防止网络过拟合, 增加样本的复杂度, 在实验中使用了数据增强的方法, 如图 7 所示.

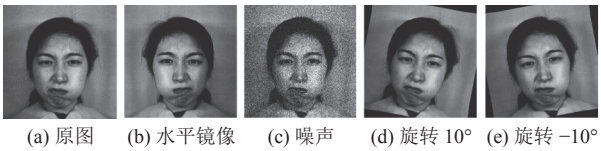


图 7 数据增强图

通过数据增强后 JAFFE 数据集一共有 907 张图片, CK+数据集一共有 4 905 张图片, 随机抽取数据集中 80% 作为训练集, 其余部分为验证集. 针对 Fer2013 数据集的特点, 本文实验剔除了数据集中不包含人脸样本, 并将所有样本混合, 随机抽取和原测试集样本同等数量的图片作为测试集, 其余部分为训练集.

2.3 预训练

为了防止网络过拟合, 在 Fully-connected 后加入了 Dropout. 如图 8 所示.

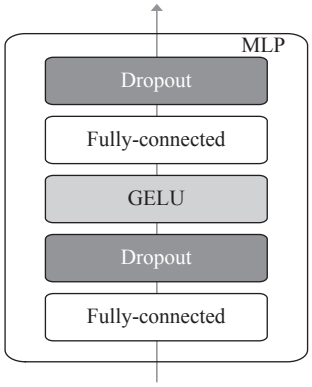


图 8 MLP 网络结构

Expression in-the-Wild 数据库 (ExpW) 包含使用 Google 图片搜索下载的 91 793 张面孔. 每个人脸图像都被手动注释为 7 个基本表情类别之一. 在注释过程中去除了非人脸图像. 如图 9 所示.

图 9 ExpW 数据集样例图

为保证预训练模型特征提取的正确性, 先从图片中提取出人脸, 再对人脸进行矫正, 去除样本中的错误样本, 剩余 87 305 张图片, 随机抽取 80% 作为训练集, 将 20% 的图片作为验证集, 验证模型的有效性. 在预训练过程中, 会将图像缩放为 48×48 大小的灰度图, 使用自适应矩估计 (Adam) 的策略, 设置学习率为 0.001, Batch size 为 64, Dropout 为 0.2, 使用交叉熵损失函数和 cosine learning rate decay[13] 学习率衰减策略, 训练至损失不再下降. Mixer Layer 网络参数如表 2 所示.

为了验证迁移学习的必要性, 使用 4 层 Mixer 网络, 对迁移前后准确率进行对比, 如表 3 所示.

表 2 Mixer 网络参数表

参数	值
Patch resolution $P\times P$	4×4
Hidden size C	512
Sequence length S	144
MLP dimension D_c	2048
MLP dimension D_s	256

表 3 数据集迁移学习前后准确率对比 (%)

数据集	迁移前	迁移后
JAFFE	90.0	95.93
CK+	97.9	98.71

由表 3 可以看出在训练小样本数据集时迁移学习的重要性. 通过迁移学习的方法将该模型在 JAFFE 数据集上的准确率提升了大约 5%, 在 CK+数据集上的准确率大约提升了 1%. 通过迁移学习, 能提高模型的识别准确率. 由于 Fer2013 数据集样本丰富, 因此该数据集不采取迁移学习策略.

3 实验设置与结果

使用无卷积的 Mixer 网络结构, 通过实验证明, 该网络同样具有提取人脸表情特征提取的能力, 在人脸表情识别达到了很好的识别效果. 同时, 在样本充足的数据集上训练过的 Mixer Layer 神经网络模型, 再对其结果进行调整和训练, 能够很好地迁移到其他小样本的数据集上.

3.1 训练过程

尝试了不同层数的 Mixer Layer 网络对 3 个数据集识别率的影响. 微调和训练网络时, 网络结构参数与表 2 保持一致, 其余参数如表 4 所示. 模型准确率如表 5 所示.

将增强后的目标数据集微调预训练好的网络, 综合考虑训练代价和识别准确率, 对数据集 CK+, JAFFE 采用含 4 层 Mixer Layer 网络. Fer2013 数据集采用含 8 层 Mixer Layer 网络. 训练精度和训练损失精度如图 10 所示.

从图中的准确率可以看出, 模型收敛快, 训练过程没有发生过拟合, 且在 CK+和 JAFFE 数据集上表现能力良好, 无卷积的 Mixer Layer 网络具有良好的学习能力和泛化能力. 将该方法与国内外优秀的人脸表情识别算法进行对比, 在 CK+ 数据集上准确率有 1%–4% 的提升, 在 JAFFE 数据集上有 1%–2% 的提升. Fer2013 数据集人为识别准确率为 (65±5)%, 8 层 Mixer Layer 模型的识别准确率达到了这一范围, 且准确率有 1%–2% 的提升. 验证了 Mixer Layer 结构在人脸表情识别上的有效性, 对比结果如表 6–表 8 所示.

表 4 微调和训练参数表

超参数	微调	训练
初始学习率	0.001	0.01
优化器	Adam	Adam
损失函数	交叉熵损失	交叉熵损失
Dropout	0.5	0.2
Batch size	64	64
策略	验证集损失不再下降时停止	验证集损失10 个epochs不下降时学习率下降为原来的0.1, 20个epochs不下降时停止训练

表 5 不同网络层数准确率

数据集	Mixer Layer层数	准确率 (%)
CK+	1	97.20
	4	**98.71**
	8	99.30
JAFFE	1	93.30
	4	**95.93**
	8	93.88
Fer2013	4	59.97
	8	**63.06**
	12	61.97
	20	60.92

为了进一步验证该算法, 根据 CK+和 JAFFE 数据集上的实验结果绘制混淆矩阵, 其中横坐标代表真实类别, 对角线代表该类样本预测正确的样本数, 其余为该类样本预测错误类别数, 该方法对于数据集 CK+和 JAFFE 法分类结果均匀, 各类表情样本更倾向于所属的类别, 具有良好的分类表现能力. 如图 11 所示.

4 结论与展望

本文基于 Mixer Layer 提出了一种结构简单的人脸表情识别方法. 针对数据集样本不足问题, 通过迁移学习和数据增强的方法提升了模型的识别准确率和泛化能力. 本文分别在 CK+, JAFFE 和 Fer2013 数据集上做了对比实验, 最终实验结果表明, 无卷积的 Mixer Layer 网络对人脸表情也有很好的识别性.

虽然基于 Mixer Layer 的网络在人脸表情识别取得了很好的识别效果, 但样本差异大, 有错误标注的数据集对网络识别准确率影响依然较大. 后续工作会在本文的基础上, 改进网络结构, 提升模型在复杂环境下的识别准确率.

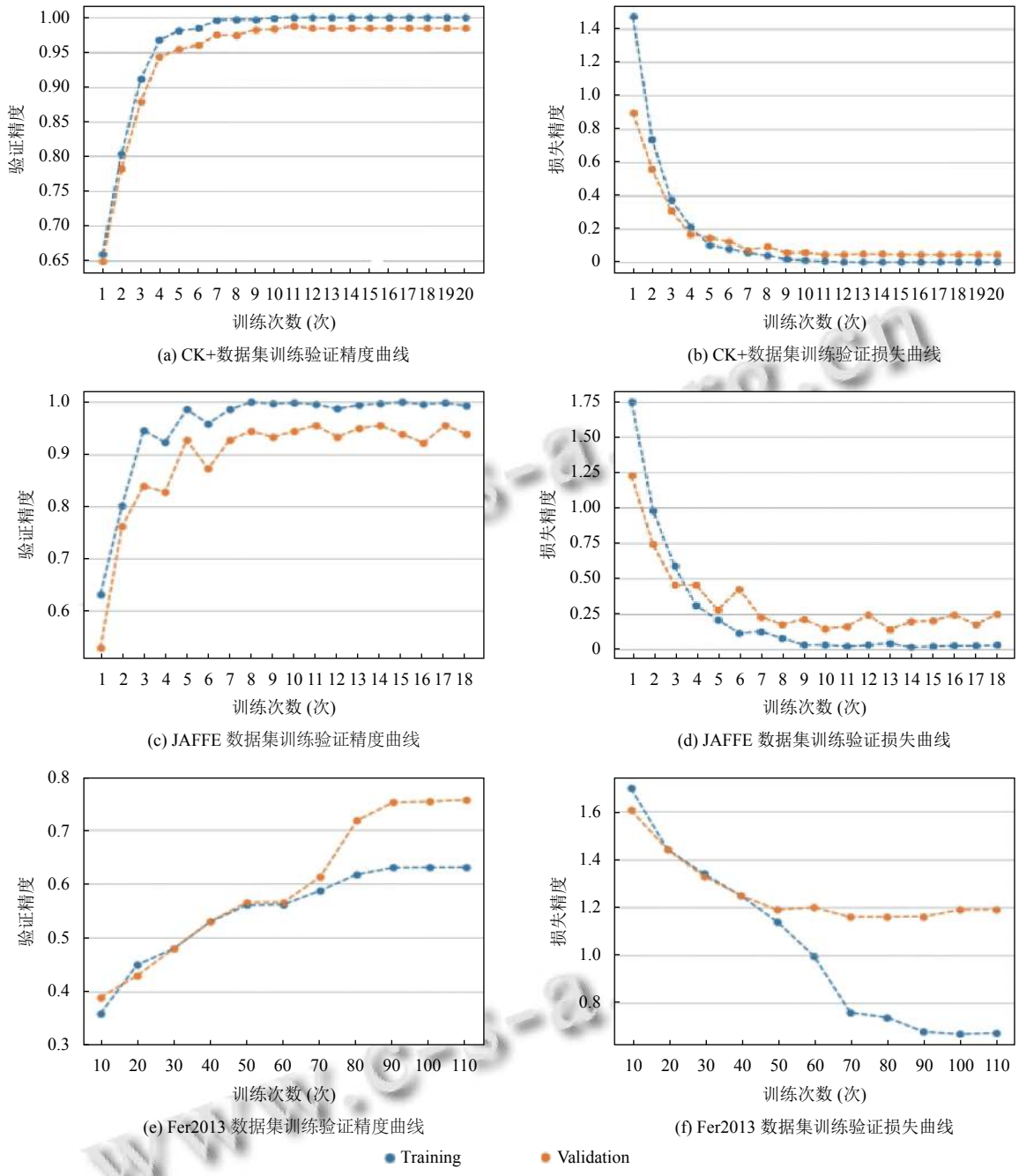


图 10 训练准确率和损失曲线

表 6 不同方法在 CK+数据集上识别准确率

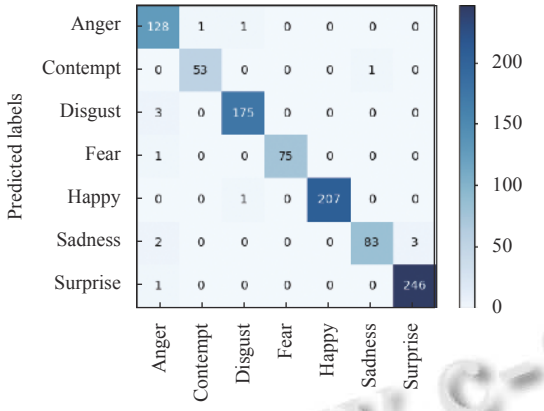
模型	准确率 (%)
MBCC-CNN[1]	98.48
剪枝GoogLeNet[9]	85.09
I2CNN[14]	96.2
改进AlexNet[15]	97.46
STM-ExpLet[16]	94.19
本文算法	**98.71**

表 7 不同方法在 JAFFE 数据集上识别准确率

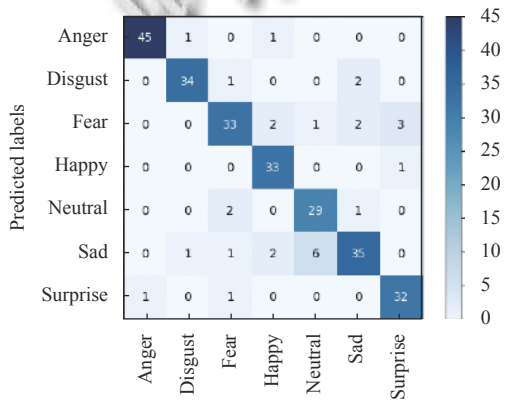
模型	准确率 (%)
LBP+CNN6[5]	93.48
剪枝GoogLeNet[9]	83.84
C-LetNet5[17]	94.37
本文算法	**95.93**

表 8 不同方法在 Fer2013 数据集上识别准确率

模型	准确率 (%)
MI+MII [18]	61.3
CNN5+ DAL结构[19]	61.59
本文算法	**63.06**



(a) CK+数据集混淆矩阵



(b) JAFFE 数据集混淆矩阵

图 11 数据集混淆矩阵

参考文献

1 Shi CP, Tan C, Wang LG. A facial expression recognition method based on a multibranch cross-connection convolutional neural network. IEEE Access, 2021, 9: 39255–39274. [doi: 10.1109/ACCESS.2021.3063493]

2 Li B. Facial expression recognition via transfer learning. EAI Endorsed Transactions on e-Learning, 2021, 7(21): e4. [doi: 10.4108/eai.8-4-2021.169180]

3 Zhang YD, Yang ZJ, Lu HM, *et al*. Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. IEEE Access, 2016, 4: 8375–8385. [doi: 10.1109/ACCESS.2016.2628407]

4 魏赟,李栋. 结合改进卷积神经网络与自编码器的表情识别. 小型微型计算机系统, 2022, 43(2): 387–392.

5 江大鹏, 杨彪, 邹凌. 基于 LBP 卷积神经网络的面部表情识别. 计算机工程与设计, 2018, 39(7): 1971–1977.

6 申毫, 孟庆浩, 刘胤伯. 基于轻量卷积网络多层特征融合的人脸表情识别. 激光与光电子学进展, 2021, 58(6): 0610005.

7 伊力哈木·亚尔买买提, 张伟. 一种新的人脸面部表情识别算法研究. 电子器件, 2021, 44(3): 616–623. [doi: 10.3969/j.issn.1005-9490.2021.03.020]

8 崔子越, 皮家甜, 陈勇, 等. 结合改进 VGGNet 和 Focal Loss 的人脸表情识别. 计算机工程与应用, 2021, 57(19): 171–178. [doi: 10.3778/j.issn.1002-8331.2007-0492]

9 张宏丽, 白翔宇. 利用优化剪枝 GoogLeNet 的人脸表情识别方法. 计算机工程与应用, 2021, 57(19): 179–188. [doi: 10.3778/j.issn.1002-8331.2102-0296]

10 Dhankhar P. ResNet-50 and VGG-16 for recognizing facial emotions. International Journal of Innovations in Engineering and Technology (IJIET), 2019, 13(4): 126–130.

11 Tolstikhin I, Houlsby N, Kolesnikov A, et al. MLP-Mixer: An all-MLP architecture for vision. arXiv: 2105.01601, 2021.

12 Hendrycks D, Gimpel K. Gaussian error linear units (GELUs). arXiv: 1606.08415, 2020.

13 He T, Zhang Z, Zhang H, *et al*. Bag of tricks for image classification with convolutional neural networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 558–567.

14 Meng ZB, Liu P, Cai J, *et al*. Identity-aware convolutional neural network for facial expression recognition. 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). Washington, DC: IEEE, 2017. 558–565. [doi: 10.1109/FG.2017.140]

15 石翠萍, 谭聪, 左江, 等. 基于改进 AlexNet 卷积神经网络的人脸表情识别. 电讯技术, 2020, 60(9): 1005–1012. [doi: 10.3969/j.issn.1001-893x.2020.09.002]

16 Liu MY, Shan SG, Wang RP, *et al*. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1749–1756. [doi: 10.1109/CVPR.2014.226]

17 李勇, 林小竹, 蒋梦莹. 基于跨连接 LeNet-5 网络的面部表情识别. 自动化学报, 2018, 44(1): 176–182.

18 Zeng GH, Zhou JC, Jia X, *et al*. Hand-crafted feature guided deep learning for facial expression recognition. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Xi'an: IEEE, 2018. 423–430. [doi: 10.1109/FG.2018.00068]

19 翟懿奎, 刘健. 面向人脸表情识别的迁移卷积神经网络研究. 信号处理, 2018, 34(6): 729–738.

(校对责编: 牛欣悦)