计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2024,33(11):27−37 [doi: 10.15888/j.cnki.csa.009682] [CSTR: 32024.14.csa.009682]

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

优化双线性 ResNet34 的人脸表情识别①

吕　军[1,2], 苌婉婷[1], 陈付龙[1,2], 王志伟[1]

[1](安徽师范大学 计算机与信息学院, 芜湖 241003)
[2](安徽师范大学 网络与信息安全安徽省重点实验室, 芜湖 241003)
通信作者: 吕　军, E-mail: lyujun@ahnu.edu.cn

摘　要: 为了能够更准确且快速地识别人脸表情, 提出了一种优化的基于 ResNet34 网络的双线性结构 (OBSR-Net) 来进行人脸表情识别. OBSR-Net 采用双线性网络结构作为整体框架, 主干网络使用 ResNet34 网络, 通过平移不变的方式对局部成对特征交互进行建模, 从而提取更加完整有效的特征, 同时采用迁移学习的策略来降低人脸表情小样本图像数据集对深度学习方法的限制. 此外, 在训练过程中使用一种新的通用优化技术, 即梯度集中. 该方法通过将梯度向量集中到零均值来直接对梯度进行操作, 可以看作是一种具有约束损失函数的投影梯度下降方法. OBSR-Net 在 Fer2013 和 CK+两个公开数据集上进行实验, 分别取得了 77.65% 和 98.82% 的识别准确率. 实验结果表明, 与其他先进的人脸表情识别方法相比, OBSR-Net 表现出较强的竞争力.

关键词: 人脸表情识别; 深度学习; 双线性结构; 迁移学习; ResNet34; 梯度集中

引用格式: 吕军,苌婉婷,陈付龙,王志伟.优化双线性 ResNet34 的人脸表情识别.计算机系统应用,2024,33(11):27–37. http://www.c-s-a.org.cn/1003-3254/9682.html

Facial Expression Recognition via Optimized Bilinear ResNet34

LYU Jun[1,2], CHANG Wan-Ting[1], CHEN Fu-Long[1,2], WANG Zhi-Wei[1]

[1](School of Computer and Information, Anhui Normal University, Wuhu 241003, China)
[2](Anhui Provincial Key Laboratory of Network and Information Security, Anhui Normal University, Wuhu 241003, China)

Abstract: An optimized bilinear structure based on ResNet34, termed OBSR-Net, is proposed for more accurate and quick facial expression recognition. OBSR-Net adopts a bilinear network structure as its overall framework and incorporates ResNet34 as the backbone network to model the local paired feature interaction by translation invariance, to extract more complete and effective features. At the same time, transfer learning mitigates the limitations imposed by small sample image data sets of facial expressions on deep learning. In addition, gradient concentration, a new general optimization technique, is utilized during the training process. This technique operates directly on gradients by concentrating gradient vectors to zero mean, which can be regarded as a projected gradient descent method with a constrained loss function. Experiments on two public datasets, namely Fer2013 and CK+, reveal that OBSR-Net achieves recognition accuracy of 77.65% and 98.82%, respectively. The experimental results show that OBSR-Net is more competitive than other advanced facial expression recognition methods.

Key words: facial expression recognition; deep learning; bilinear structure; transfer learning; ResNet34; gradient centralization (GC)

① 基金项目: 国家自然科学基金 (61972438); 教育部产学合作协同育人项目 (230803924042356)
收稿时间: 2024-04-25; 修改时间: 2024-05-20; 采用时间: 2024-06-04; csa 在线出版时间: 2024-09-27
CNKI 网络首发时间: 2024-09-30

人类是情绪化的生物, 而面部表情是人类表达情感状态和意图的最有力、最自然、最普遍的信号之一. 有关研究表明, 言语成分传达了人类交流的 30%, 而非言语成分则传达了 70%[1]. 人脸表情识别 (facial expression recognition) 是一种基于面部表情特征进行自动识别和情感分析的技术, 目前广泛运用在智慧教育[2]等领域, 可以实现自然交互和情感分析. 在智慧教育领域, 人脸表情识别技术可以快速捕捉教师和学生的表情, 从而能够对课堂效果进行总结, 进一步提高教学效率. 得益于深刻的现实意义和广泛的应用前景, 越来越多的研究者们开始从事人脸表情识别的研究.

虽然现在有很多学者对人脸表情识别进行了大量研究, 但大多数都没有考虑人脸表情图像自身的特点. 不同于其他的物体识别任务, 人脸表情识别有其独特的挑战. 首先, 人脸表情图像属于细粒度图像. 一方面, 虽然呈现出不同的表情, 但是一些不同类别的表情从视觉外观上来看是极其相似的; 另一方面, 受外在因素的影响, 同种类别的表情也会呈现出不同的视觉感官, 这些都会增加人脸表情识别的难度. 其次, 本文考虑使用基于深度神经网络的方法来完成人脸表情识别任务, 但是深度学习方法极其依赖大规模标注数据. 目前人脸表情图像数据集大都属于小样本数据集, 这会在一定程度上影响基于深度学习的识别方法的性能.

基于上述考虑, 本文采用了一种优化的基于 ResNet34 的双线性网络 OBSR-Net (optimized bilinear structure based on ResNet34 network) 用于人脸表情识别, 主要工作内容总结如下.

(1) 本文采用一种优化的基于迁移学习的双线性 ResNet34 网络 OBSR-Net, 该网络能够在端到端的训练方式上提取更加完整准确的局部判别性特征, 并且能够减少模型对数据的依赖, 提高模型的学习效率.

(2) 本文在网络模型中融合了一种新的通用网络优化技术, 即梯度集中. 它不仅可以平滑和稳定深度神经网络的训练过程, 而且可以提高模型的泛化性能.

(3) 本文在 Fer2013[3]和 CK+[4]两个公开数据集上验证本文所提的方法, 与 ResNet34 网络模型以及当前较为先进的方法进行了对比实验, 结果表明本文所提方法在人脸表情识别方面具有较强的竞争力.

1 相关工作

1.1 细粒度图像识别

细粒度图像识别是近年来计算机视觉、模式识别等领域中一个非常热门的研究课题, 其目的是对粗粒度的大类别进行更加细致的子类划分, 但由于子类别间细微的类间差异和较大的类内差异, 相较于普通的图像识别任务, 细粒度图像识别难度更大.

在细粒度图像识别领域, Zhang 等人[5]提出了序列多样化网络 (SDNs), 通过在骨干网络中构建多个轻量级子网络, 实现细粒度图像局部区域之间的信息交互. SDNs 共同促进了空间注意力的多样性, 对学习多样性表征有着极大的帮助. Niu 等人[6]从人类视觉识别机制的角度研究了注意学习过程, 其中注意区域通过注意转移机制在时间上被感知. 他们提出了基于注意力转移的深度神经网络 (AS-DNN) 来寻找注意力区域, 并对所发现的注意力区域之间的语义相关性进行迭代编码, 有效地提高了分类性能. Yu 等人[7]提出了一种用于弱监督细粒度图像识别的新型端到端可信多粒度信息融合 (TMGIF) 模型, 该方法可以自动提取细粒度图像的多粒度信息表示, 并进一步评估信息粒度的质量, 然后根据质量逐步融合多粒度信息以获得可靠且可解释的识别结果. Du 等人[8]重点研究了注意力区域的哪些粒度最有效, 并有效地融合了不同粒度的信息. 他们提出了一种渐进训练策略来融合多粒度的注意力特征, 并提出了一个随机拼图补丁生成器来使得模型关注特定粒度的注意力信息. Wang 等人[9]利用深度特征之间的协方差来构建表示, 并将矩阵幂归一化加入到全局协方差池化的学习中, 既提高了训练速度, 又降低了模型复杂度. Yong 等人[10]提出一种梯度集中 (gradient centralization, GC) 的优化技术, 并在各种应用上进行实验, 包括一般图像分类、细粒度图像分类、检测和分割, 结果表明 GC 可以提高深度神经网络 (deep neural network, DNN) 学习的性能. 双线性卷积神经网络 (B-CNN)[11]是第 1 个在细粒度识别任务上可以端到端训练的协方差池化网络模型, 其方法模型有效地提高了图像分类的准确率和效率.

1.2 人脸表情识别

深度学习 (deep learning, DL) 的概念起源于 Hinton 等人于 2006 年发表在《科学》杂志上的一篇论文[12]. 深度学习克服了传统算法依赖人工设计特征提取器的缺点. 基于此, Tang[13]使用线性支持向量机代替 Softmax 层并学习最小化基于边际的损失, 而不是交叉熵损失, 以此提高识别的准确率. Wen 等人[14]提出了一种基于概率融合的卷积神经网络方法, 有效地提高了表情识别的准确率. Yang 等人[15]提出了一种加权

混合深度神经网络结构来提取面部表情分类特征, 方法是创建浅层 CNN 和在 ImageNet 上预训练的 VGG16 网格模型, 从 LBP 和面部灰度图片中提取面部特征, 将两个特征加权融合, 并进行 Softmax 的表情分类输出. Zhu 等人[16]提出了一种级联注意力网络, 该网络将注意力机制与金字塔特征相结合, 由 3 个模块组成: 局部和多尺度立体空间上下文特征提取模块、级联注意力模块和时间序列特征提取模块. 该网络充分利用上下文信息来弥补空间特征的缺失, 增强了注意力机制的性能, 提高识别人脸的准确性. Yu 等人[17]提出一种基于多通道融合和轻量级神经网络的面部表情识别方法, 该方法通过将传统特征提取算法与深度学习特征提取算法相结合, 有效提取出更完整的图像特征, 提高了面部表情识别的准确性和鲁棒性.

2 方法模型

本文提出的 OBSR-Net 主要由迁移学习和双线性网络[11]结构模块两个部分组成, 同时在训练过程中使用了一种新的通用网络优化技术, 即梯度集中[10]. 整体网络结构如图 1 所示, 对于要训练的人脸表情图像会分别输入到两个主干网络中提取特征, 然后通过双线性池化函数将提取到的特征进行集成以获取最终的特征图.

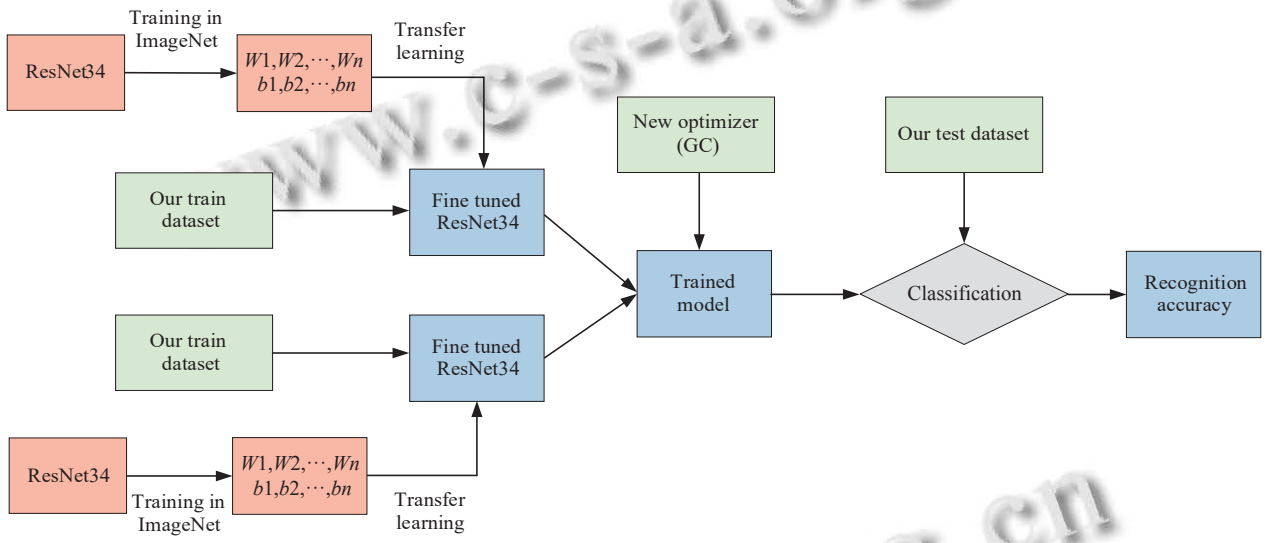


图 1 整体网络结构图

本文中, 主干网络选择的是 CNN 网络 ResNet34, 并在训练时使用从典型的海量数据集中学习到的预训练模型, 然后迁移到特定的人脸表情识别任务中. 同时, 在训练过程中使用一种新的优化器, 其通过引入对权重向量的新约束来约束损失函数, 该约束对权重空间进行了正则化, 从而提高了模型的泛化性能. 此外, 约束损失函数比原来的损失函数具有更好的 Lipschitzness, 这使得训练过程更加稳定和高效.

2.1 ResNet34 网络结构

针对表情识别任务, 本文选取的主干网络是 ResNet34[18]. ResNet34 网络中包含了 33 个卷积层和 1 个全连接层, 其网络结构如图 2 所示.

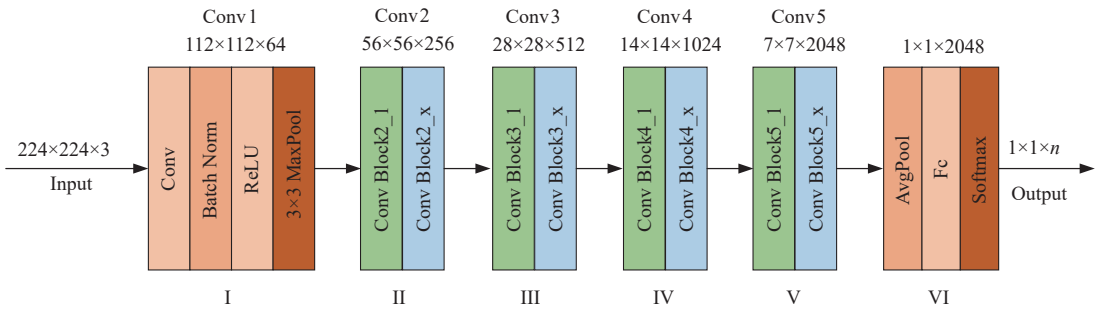


图 2 ResNet34 网络结构

ResNet34 网络可以分成 6 个部分, 第 1 部分不包含残差块, 主要对输入进行卷积、正则化、激活函数、最大池化的计算. 第 2–5 部分结构都包含了残差块, 所包含的残差块个数分别为 3、4、6、3 个, 每个残差模块包含两个卷积层和一个跳跃连接, 跳跃连接可以保证梯度在反向传播中能够通过整个网络结构传递, 避免了梯度消失问题. 一般情况下网络以 224×224×3 维的彩色图像作为输入数据, 经过前 5 部分的卷积计算, 输出 7×7×2048 维特征图, 接着第 6 部分会将其转化成一个特征向量, 最后通过分类器对这个特征向量进行计算并输出类别概率.

2.2 迁移学习

由于本文研究的面部表情图像识别领域获得的数据有限, 仅靠有限的数据集进行学习, 容易造成数据过拟合, 获得的模型不够稳健且训练成本较高. 本文采取迁移学习 (transfer learning)[19]的方法解决上述问题.

本文使用 ResNet34 网络在 ImageNet 数据集[20]的 1 000 个类别上预训练的模型, 采取微调的迁移策略: 首先将 ResNet34 在 ImageNet 数据集上训练好所有参数的模型用 ResNet34_F1 表示; 然后冻结模型 ResNet34_F1 中的参数; 最后对模型 ResNet34_F1 中的全连接层进行微调, 其参数是默认可学习的, 而其他网络层的参数均保持不变, 从而得到适合本文识别任务的网络结构 ResNet34_F2. 使用基于迁移学习的 ResNet34 网络进行面部表情识别的过程如图 3 所示.

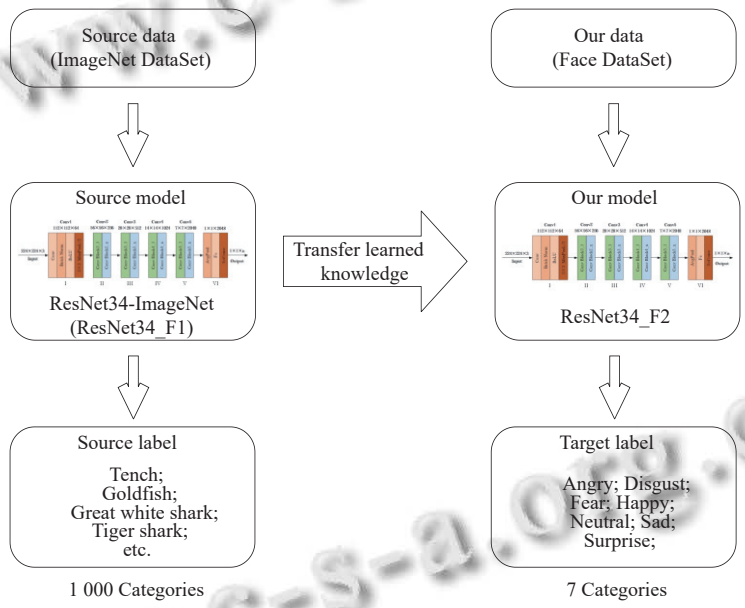


图 3 基于迁移学习的人脸表情识别过程

2.3 双线性网络结构

人脸表情图像属于细粒度图像, 这些图像具有较强的类间相似性和类内差异性, 例如图 4 中, neutral_1 和 neutral_2 均是中立的表情, 但由于角度的原因, 使得 neutral_1 和 fear_1 (害怕) 看起来更加相似, 这些会使得表情识别更加困难. 不仅如此, 对于一些细微的表情变化或者特殊场合下不便做出明显表情的情况, 面部中的特殊位置往往是能够准确识别表情的关键部位, 例如在细微的恐惧时, 眉头向中间聚拢、上扬, 上眼皮也随之上提. 而模型能够准确定位到这些位置并提取特征可以在一定程度上提升识别的整体性能.

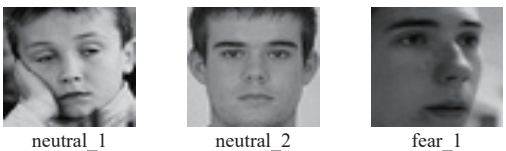


图 4 人脸表情图像的类间相似性和类内差异性

针对上述问题并为了实现端到端的训练和识别, 本文采用一种双线性网络结构, 并将微调后的 ResNet34 作为该结构的两个特征提取器, 其示意图如图 5 所示. 将图像输入到网络结构后, 两个提取器将分别提取特征, 这些特征通过双线性池化函数转化为双线性向量,

然后将双线性向量输入到 Softmax 函数中进行表情分类. 这种方法与人脑视觉处理的两条通路假说有关. 该假设表明人脑使用一条通路来定位物体, 并使用另一条通路来识别物体[21]. 具体到本文的表情识别任务中, 可以看作是两个特征提取器相互协调, 一个提取器用来定位到发生表情变化的关键部位, 另一个提取器则用于提取定位到的关键部位的特征. 该方法可以从图像中提取更多不同的特征, 局部特征通过双线性池化函数以线性方式进行集成. 因此该方法可以以平移不变的方式对局部成对特征之间的相互作用进行建模.

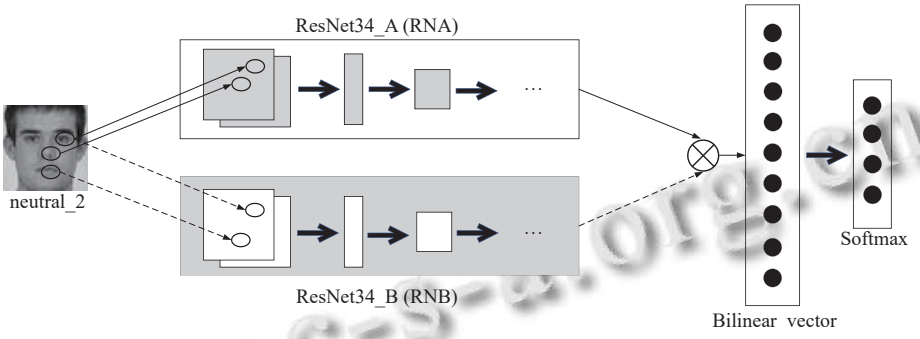


图 5 双线性 ResNet34 网络结构图

双线性 ResNet34 网络主要由 4 部分组成, 即 $\beta=(f_{RNA}, f_{RNB}, f_{BP}, f_C)$, 其中 f_{RNA} 和 f_{RNB} 代表特征提取器, 它们由微调的 ResNet34 构成, f_{BP} 是一个双线性池化函数, f_C 表示分类函数. 将图像 I 输入到该网络, 先分别使用 f_{RNA} 和 f_{RNB} 对图像 I 在 l 位置上提取两个特征 $f_{RNA}(I, l)$ 和 $f_{RNB}(I, l)$, 然后把同一位置上的两个特征进行双线性融合 (相乘), 得到矩阵 b, 其过程如式 (1) 所示; 接着对所有位置的矩阵 b 进行和池化 (sum pooling) 得到矩阵 M_I, 计算过程如式 (2) 所示.

$$b(I,l) = f_{RNA}^{T}(I,l) f_{RNB}(I,l) \tag{1}$$

$$M_I = \sum_l b(I,l) \tag{2}$$

然后, 将 M_I 作为双线性池化函数 f_{BP} 的输入, 经过计算得到一个图像表示向量 z, 它可以被描述为式 (3) 所示; 在该步骤中, M_I 首先通过式 (4) 被重塑为一个向量, 记为 x; 接着通过对 x 进行归一化得到最后融合后的特征向量 z, 归一化操作如式 (5) 和式 (6) 所示.

$$z = f_{BP}(M_I) \tag{3}$$

$$x = reshape(M_I) \tag{4}$$

$$y = \sin(x)\sqrt{|x|} \tag{5}$$

$$z = y/\|y\|_2 \tag{6}$$

最后, 将特征向量 z 输入到分类函数 f_C 中, 从而得到分类结果, f_C 定义如式 (7) 所示:

$$f_C(z_i) = \frac{e^{z_i}}{\sum_1^n e^{z_n}} \tag{7}$$

其中, z_i 表示第 i 个结点的输出值, n 表示输出节点的个数, 即分类任务中的类别总数.

2.4 新的优化技术—梯度集中

深度学习目前应用在很多领域之中, 其之所以能够如此成功, 很大程度上要归功于大规模数据集[22]、强大的计算资源 (如 GPU 和 TPU)、复杂的网络架构和优化算法的快速发展. 在这些因素中, 有效的优化技术, 如具有动量的随机梯度下降 (SGD)[23]、Adagrad[24] 和 Adam[25], 使得用大规模数据集训练深度的神经网络 (DNN) 成为可能, 从而在实践中提供更强大和稳健的 DNN 模型. 一个好的 DNN 优化器有两个主要目标: 加速训练过程和提高模型泛化能力. 第 1 个目标旨在花费更少的时间和成本来达到良好的局部最小值; 第 2 个目标旨在确保学习的 DNN 模型能够对测试数据做出更加准确的预测.

现阶段已经提出了很多优化技术来对激活、权值和梯度进行操作, 例如批量归一化 (batch normalization, BN)[26]使用一阶和二阶统计量对激活值进行 Z-score 标准化. 文献[27]表明 BN 降低了损失函数的 Lipschitz 常数, 使梯度更加 Lipschitz 平滑, 从而使优化现象变得更加平滑. 权重标准化 (weight standardization, WS)[28]通过权重向量上的 Z-score 标准化来降低损失函数的 Lipschitz 常数并平滑优化现象. BN 和 WS 分

别对激活和权重向量进行操作, 并且它们隐式地限制了权重的梯度, 从而提高了优化损失的 Lipschitz 性质. 与上述对激活或权重向量进行操作的技术不同, 本文使用的是一种新的优化技术, 即梯度集中 (GC), 它通过将梯度向量集中到零均值来直接对梯度进行操作. GC 可以看作是一种具有约束损失函数的投影梯度下降方法, 其可以正则化权重空间, 从而提高 DNN 的泛化性能. 此外, 约束损失函数比原来的损失函数具有更好的 Lipschitzness, 这使得训练过程更加稳定和高效.使用 GC 的示意图如图 6 所示, 其中 W 是权重, $\mathcal{L}$是损失函数, $\nabla_W\mathcal{L}$是权重的梯度, $\Phi_{GC}(\nabla_W\mathcal{L})$ 是集中梯度.

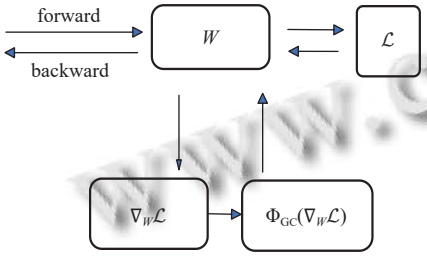


图 6 使用梯度集中 (GC) 的示意图

具体地, 对于卷积层或者全连接层的权重向量 W_i, 假设已经通过反向传播得到其梯度$\nabla_{W_i}\mathcal{L}$ $(i=1,2,\cdots,N)$, 将梯度中心化操作定义为 Φ_{GC}, 则 Φ_{GC} 如式 (8) 所示:

$$\Phi_{GC}(\nabla_{W_i}\mathcal{L})=\nabla_{W_i}\mathcal{L}-\mu\nabla_{W_i}\mathcal{L} \tag{8}$$

其中, $\nabla_{W_i}\mathcal{L}$表示的是梯度, 下标 i 表示当前是梯度矩阵中第 i 列的列向量. $\mathcal{L}$表示目标函数. $\mu\nabla_{W_i}\mathcal{L}=\frac{1}{M}\sum_{j=1}^{M}\nabla_{W_{i,j}}\mathcal{L}$, 表示权重矩阵第 i 列的梯度均值.

由于参数 $\mu\nabla_{W_i}\mathcal{L}$表示的是列方向上的均值, 根据矩阵的运算原理, 将其代入到式 (8) 中, 之后提出一个因子 P, 则得到式 (8) 的矩阵运算形式:

$$\Phi_{GC}(\nabla_W\mathcal{L})=P\nabla_W\mathcal{L},\ P=I-ee^{\mathrm{T}} \tag{9}$$

其中, $\nabla_W\mathcal{L}$表示$\mathcal{L}$相对于权值向量 W 的梯度, e 是一个 $M\times1$ 的单位向量, 单位向量的范数为 1, 因此 e 中每个元素的大小均为$1/\sqrt{M}$, $I\in\mathbb{R}^{M\times M}$表示一个单位矩阵, P 的物理意义是权重空间中, 法向量为 e 的超平面的投影矩阵.

可以看到, P 是单位矩阵和单位向量的线性结合, 所以直观上能够察觉到 P 是一个对称矩阵, 而根据矩阵乘法的性质, 可得 $P^2=P=P^{\mathrm{T}}$. 此外, P 还具有另一个比较重要的性质, 如式 (10) 所示:

$$e^{\mathrm{T}}P=e^{\mathrm{T}}(I-ee^{\mathrm{T}})=0 \tag{10}$$

式 (10) 可以较容易地得到证明, 则自然能够得到如式 (11) 所示的结论.

$$e^{\mathrm{T}}P\nabla_W\mathcal{L}=0 \tag{11}$$

其中, 若把$P\nabla_W\mathcal{L}$看成一个整体, 那么其表示的含义为$P\nabla_W\mathcal{L}$必定存在于一个和 e 正交的平面内, 这是因为 P 本身具有的性质就是和 e 正交. 这时再把 P 和$\nabla_W\mathcal{L}$分开, 则可看成矩阵$\nabla_W\mathcal{L}$在 P 所在平面上的投射. 也就是说, P 是一个中间矩阵, 它代表的是一个和单位向量 e 正交的平面, $P\nabla_W\mathcal{L}$就是将梯度投射到该超平面上.

现假设当前已执行了 t 次梯度下降的迭代, 每次迭代均使用了梯度中心化的方法处理, W^t 表示第 t 个时间步的权重, W 表示初始权重. 为了便于理解, 可直接把高维的权重 W^t 和 W 看成是向量, 由上面的结论可知:

$$e^{\mathrm{T}}(W-W^t)=0 \tag{12}$$

也就是说, 对于任何一个时间步 t, 其权重的变化将一直在这个超平面内, 从式 (12) 中可得出 $e^{\mathrm{T}}W^{t+1}=e^{\mathrm{T}}W^t=\cdots=e^{\mathrm{T}}W^0$, 即 $e^{\mathrm{T}}W$ 在训练期间是常量. 从数学上来讲, 对应的一个权重向量 W 的目标函数可以写成如式 (13) 所示:

$$\min_W\mathcal{L}(W),\ \text{s.t. } e^{\mathrm{T}}(W-W^0)=0 \tag{13}$$

这是一个关于权重向量 W 的约束优化问题, 正则化 W 的解空间, 从而减少过度拟合训练数据的可能性. 因此, 梯度中心化可以提升经过训练的 DNN 模型的泛化性能, 尤其是当训练样本数量有限时.

WS 对权值进行 $e^{\mathrm{T}}W=0$ 的约束, 当初始权重不满足约束时, 会直接修改权值来满足约束条件. 假设进行微调训练, WS 则会完全丢弃预训练模型的优势, 而 GC 可以适应任何初始权值, 因为它将初始化权重 W^0 考虑在约束中, 即 $e^{\mathrm{T}}(W^0-W^0)=0$ 永远成立. 这恰好适用于本文采取的迁移学习的方法.

GC 的另一个优点是防止梯度爆炸, 使得训练更加稳定, 作用原理类似于梯度裁剪[29], 过大的梯度会导致损失动荡, 难以收敛, 通过梯度中心化, 可以减小梯度矩阵的最大值, 加快收敛速度并使其训练过程更稳定.

3 实验评估

3.1 数据集

Fer2013 是一个包含困难的自然条件和挑战的特

定情绪识别数据集, 它在 2013 年的国际机器学习会议 (ICML) 上被引入. Fer2013 包含 7 种面部表情的图像, 分布为 angry (4 953)、disgust (547)、fear (5 121)、happy (8 989)、sad (6 077)、surprise (4 002) 和 neutral (6 198). 图 7 给出了 Fer2013 数据集中部分表情的示例图.

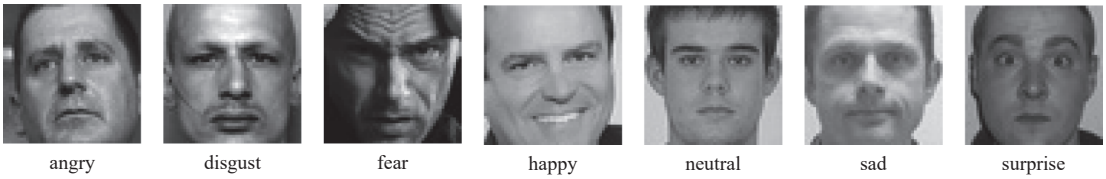


图 7 Fer2013 数据集部分示例图

CK+是用于人脸表情识别的公共数据集之一. 该数据集由美国卡内基梅隆大学 (CMU) 的 Cohn 等人创建, 于 2000 年开始发布.

该数据集包含了 123 位受试者的 593 个视频序列, 其中 327 个视频序列带有表情标签. 本实验选取除自然表情外的其他 7 种基本表情, 分布为 anger (135)、disgust (177)、fear (75)、happy (207)、sadness (84)、surprise (249) 和 contempt (54), 每张图像的原始尺寸为 640×490, 将其大小处理为 48×48. 图 8 给出了 CK+数据集中部分表情的示例图.

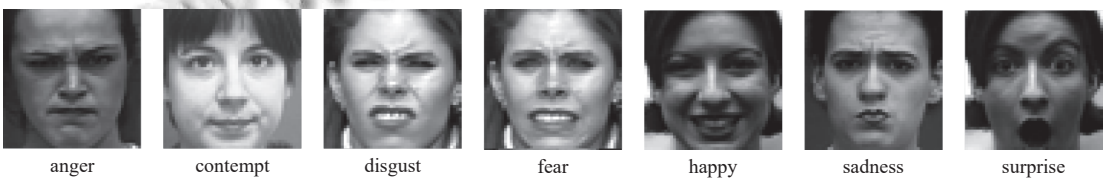


图 8 CK+数据集部分示例图

3.2 实验环境和参数设置

本文中的所有实验的硬件环境是: Intel(R) Core i7-10870H 221 GHz 处理器, 16 GB 内存; GeForce RTX 2060 的 GPU. 软件环境是: Windows 10 64 位操作系统, Python 3.8 为编程语言, 采用 PyTorch 1.11 框架搭建网络模型.

在实验中, 考虑到模型的训练效果和实验条件, 将 batch size 设置为 16, 因此每次输入模型的样本为 16 个; 模型的收敛速度由学习率决定, 本文中学习率设置为 1×10^{-4}; 在模型训练过程中, 使用交叉熵损失函数, 同时使用梯度集中这种新的优化技术进行优化, 通过实验, 本文在 Fer2013 数据集上使用将 GC 嵌入 AdamW[30] 的策略, 在 CK+数据集上使用将 GC 嵌入 Adam 的策略可以获得最佳效果; Epochs 设置为 200, 表示在训练过程中对全部样本数据进行 200 轮训练.

3.3 评价指标

针对多分类问题, 模型评估指标分为两大类: example-based metrics 和 label-based metrics. 评价指标常用来评估模型的好坏, 混淆矩阵是人工智能领域中最常用的模型评价方法, 常用的评价指标包括准确率、召回率等. 本文使用 label-based metrics 评估指标作为标准, 以准确率 (*Accuracy*)、精确率 (*Precision*)、召回率 (*Recall*) 和 *F*1 分数 (*F*1_*score*) 来验证所提方法的有效性. 其中, 准确率表示预测正确的样本数占样本总数的比例; 精确率表示预测所有正样本中判断正确的比例; 召回率表示预测正确的所有正样本占实际所有正样本的比例; *F*1_*score* 分数表示精确率与召回率的调和平均数. 4 项指标的计算过程如下:

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

$$Precision=\frac{TP}{TP+FP} \tag{15}$$

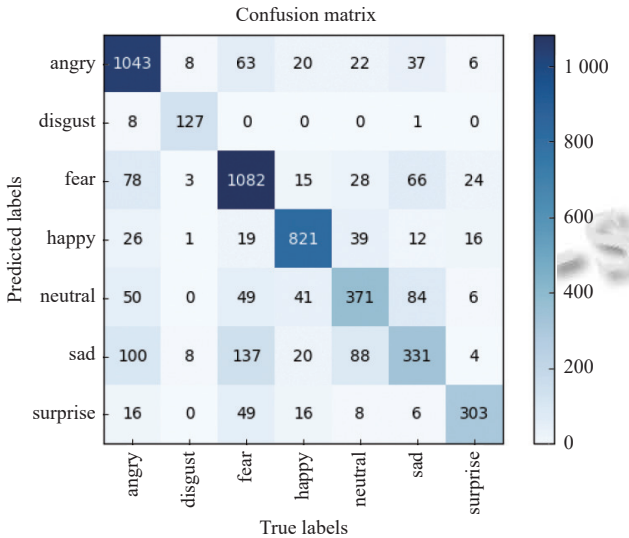
$$Recall=\frac{TP}{TP+FN} \tag{16}$$

$$F1\_score=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{17}$$

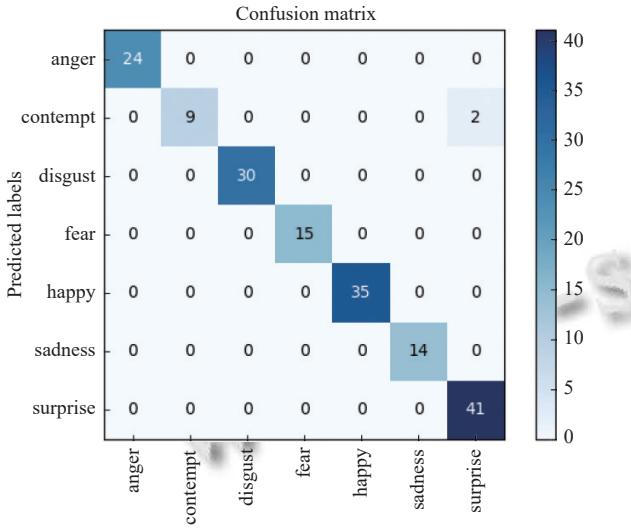
其中, *TP* 为预测为正例而且实际上也是正例; *FP* 为预测为正例, 然而实际上却是负例; *FN* 为预测为负例, 然而实际上却是正例; *TN* 为预测为负例而且实际上也是负例.

3.4 实验结果

将 Fer2013 和 CK+数据集应用到模型中, 经过大量的训练得到的结果如下所示. 图 9 表示的是模型在 Fer2013 和 CK+数据集测试的混淆矩阵; 图 10 表示的是模型在 Fer2013 和 CK+数据集上的损失曲线; 表 1 和表 2 分别表示的是模型在 Fer2013 和 CK+数据集上的测试精度、召回率、*F*1 分数.



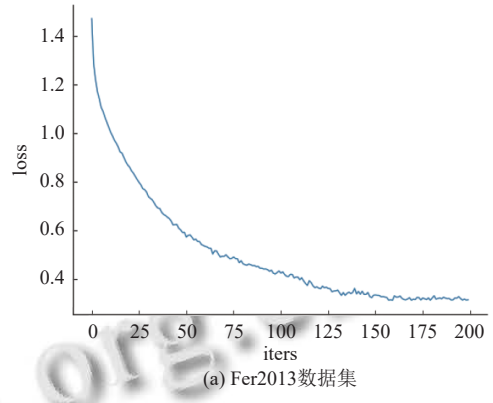
(a) Fer2013数据集



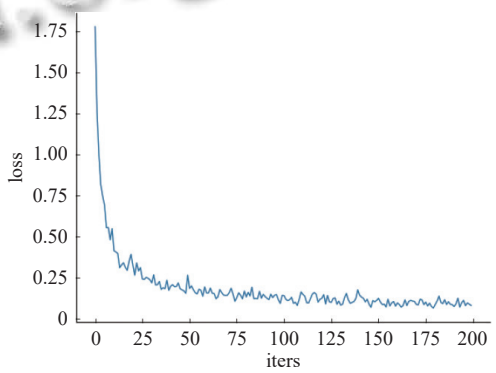
(b) CK+数据集

图 9 模型在 Fer2013 和 CK+数据集测试的混淆矩阵

由图 9 可知, (a) 表示的是在 Fer2013 数据集上的混淆矩阵, 由此可以得出 Fer2013 数据集的分类准确率为 77.65%; 而在 (b) CK+数据集的测试仅有部分 surprise 类表情被误识别为 contempt, 其他类的表情的测试准确率均为 100%, 可以得出 CK+数据集的分类准确率为 98.82%. 由此可以看出所提网络模型具有较好的分类性能.



(a) Fer2013数据集



(b) CK+数据集

图 10 模型在 Fer2013 和 CK+数据集上的损失曲线

表 1 模型在 Fer2013 数据集的测试精度、召回率、*F*1 分数

Type	*Precision*	*Recall*	*F1_score*
angry	0.870	0.790	0.828
disgust	0.934	0.864	0.898
fear	0.835	0.773	0.803
happy	0.879	0.880	0.879
neutral	0.617	0.667	0.641
sad	0.481	0.616	0.540
surprise	0.761	0.844	0.800

表 2 模型在 CK+数据集的测试精度、召回率、*F*1 分数

Type	*Precision*	*Recall*	*F1_score*
angry	1.0	1.0	1.0
contempt	0.818	1.0	0.9
disgust	1.0	1.0	1.0
fear	1.0	1.0	1.0
happy	1.0	1.0	1.0
sadness	1.0	1.0	1.0
surprise	1.0	0.953	0.976

由表 1 和表 2 可知, *Precision* 为模型判断为正样本的置信概率, 概率越高, 该模型判断出的正样本越可信; *Recall* 越高, 表示找的越全; 在 Fer2013 数据集中,

disgust 类表情测试的精确率最高, happy 类表情找的最全; CK+数据集大多数类的评价指标均为 1.0.

3.5 与主流方法的比较

本节将所提方法与较为先进的方法在 Fer2013 和 CK+两个公开数据集上进行比较并分析. 不同方法在 Fer2013 和 CK+数据集上的准确率分别如表 3 和表 4 所示.

表 3 不同方法在 Fer2013 数据集上的准确率 (%)

方法	准确率
ResNet34[22]	70.55
VGG16[31]	71.52
MobileNetV3[32]	68.24
ResMasking[33]	71.97
Deep-emotion[34]	70.02
GARAN[35]	72.14
VGNFL[36]	72.49
FERDL[37]	75.80
Ours	**77.65**

表 4 不同方法在 CK+数据集上的准确率 (%)

方法	准确率
AlexNet[20]	93.59
ResNet34[22]	97.81
VGG16[31]	97.12
DenseNet121[38]	94.69
ResMasking[33]	98.46
GARAN[35]	98.67
DeRL[39]	97.30
Ours	**98.82**

分析表 3 可知, 本文提出的方法在 Fer2013 标准数据集上取得了不错的识别效果. 由实验结果可以看出所提方法比原始主干网络提高 7.10%, 比其他主流的原始 CNN 模型至少提高 6.13%; 其次, 本文还对比了 5 种较为先进的人脸表情识别方法, 所提方法比这些较为先进的方法至少提高了 1.85%, 这些结果验证了所提方法在 Fer2013 数据集上的有效性.

分析表 4 可知, 由于 CK+数据集中的图像较为清晰, 类别标签也比较准确, 所以大多数方法在该数据集上均能达到超高的分类精度, 本文提出的方法在 CK+数据集上获得了 98.82% 的分类精度, 比原始主干网络提高 1.01%, 比现有较为先进的方法至少提高 0.15%, 表现出较强的竞争力.

3.6 消融实验

为了验证双线性网络结构以及所使用的新的优化器的有效性, 本文设计了一系列消融实验, 添加不同策略后训练相同的次数, 在 Fer2013 以及 CK+两个公开数据集上的准确率如表 5 所示. 其中, ResNet34 表示没有进行任何改进的基础网络, ResNet34_Pretrain 表示使用了迁移学习的 ResNet34 网络, ResNet34_Pretrain_BCN 表示基于迁移学习的双线性 ResNet34 网络, ResNet34_Pretrain_BCN_GC 表示在 ResNet34_Pretrain_BCN 的基础上添加了新的优化器模块.

表 5 不同模块识别准确率对比 (%)

方法	Fer2013	CK+
ResNet34 (baseline)	70.55	97.81
ResNet34_Pretrain	74.29	97.98
ResNet34_Pretrain_BCN	75.61	98.31
ResNet34_Pretrain_BCN_GC	**77.65**	**98.82**

从表 5 中的数据可以看出, 双线性池化网络结构、添加梯度中心化的 Adam 优化器都能够有效地提升人脸表情识别的准确率, 并且同时引入这几个模块对模型进行改进后的识别准确率提升最为显著, 在 Fer2013 和 CK+数据集上的表情识别准确率分别为 77.65% 和 98.82%, 与 ResNet34 基础网络相比分别提升了 7.10% 和 1.01%, 从而验证了本文所提方法的有效性.

4 结论与展望

本文提出一种使用新的优化技术的基于 ResNet34 网络的双线性网络结构来对人脸表情进行识别. 考虑到人脸表情图像的数据集属于小样本数据集, 基于迁移学习理论, 本文使用 ResNet34 网络模型在 ImageNet 大图像数据集进行预训练得到的模型, 然后将学习到的模型参数转移到人脸表情识别的特定任务中. 同时, 本文采用双线性网络结构作为整体框架, 双线性网络结构通过定位到人脸表情呈现的特殊位置后再提取该部分的特征, 从而使得模型能够提取到更加完整、更有效的特征. 此外, 在模型训练过程中使用一种新的优化技术, 即梯度集中 (GC), 其通过引入对权重向量的新约束来约束损失函数, 该约束对权重空间进行了正则化, 提高了模型的泛化性能, 而且约束损失函数比原始损失函数具有更好的 Lipschitzness, 使训练过程更加稳定和高效. 最后, 本文在 Fer2013 和 CK+两个公开数据集上验证了所提方法的有效性. 实验结果表明, 所提方法与现有较为先进的方法相比, 表现出较强的竞争力.

虽然人脸表情识别已经取得了较好的识别效果, 并且在科研项目上应用广泛, 但是光照、遮挡以及侧头等因素的影响依然较大, 为了克服这些外界因素, 未来的研究可以将表情识别转到更加复杂的场景, 例如课堂中, 识别教师和学生的面部表情评价教师的教学水平以及学生的听课情况. 同时人类的面部表情远不止这几种, 在未来, 我们将更加全面地研究人脸面部表情的识别.

参考文献

1 Hossain S, Umer S, Rout RK, *et al*. Fine-grained image analysis for facial expression recognition using deep convolutional neural networks with bilinear pooling. Applied Soft Computing, 2023, 134: 109997. [doi: 10.1016/j.asoc.2023.109997]

2 Ding ZP, Yun HX, Li EZ. A multimedia knowledge discovery-based optimal scheduling approach considering visual behavior in smart education. Mathematical Biosciences and Engineering, 2023, 20(3): 5901–5916. [doi: 10.3934/mbe.2023254]

3 Li S, Deng WH, Du JP. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2584–2593.

4 Lucey P, Cohn JF, Kanade T, *et al*. The extended Cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. San Francisco: IEEE, 2010. 94–101.

5 Zhang LB, Huang SL, Liu W. Learning sequentially diversified representations for fine-grained categorization. Pattern Recognition, 2022, 121: 108219. [doi: 10.1016/j.patcog.2021.108219]

6 Niu Y, Jiao Y, Shi GM. Attention-shift based deep neural network for fine-grained visual categorization. Pattern Recognition, 2021, 116: 107947. [doi: 10.1016/j.patcog.2021.107947]

7 Yu Y, Tang H, Qian J, *et al*. Fine-grained image recognition via trusted multi-granularity information fusion. International Journal of Machine Learning and Cybernetics, 2023, 14(4): 1105–1117. [doi: 10.1007/s13042-022-01685-6]

8 Du RY, Chang DL, Bhunia AK, *et al*. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 153–168.

9 Wang QL, Xie JT, Zuo WM, *et al*. Deep CNNs meet global covariance pooling: Better representation and generalization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(8): 2582–2597. [doi: 10.1109/TPAMI.2020.2974833]

10 Yong HW, Huang JQ, Hua XS, *et al*. Gradient centralization: A new optimization technique for deep neural networks. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 635–652.

11 Lin TY, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1449–1457.

12 Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507. [doi: 10.1126/science.1127647]

13 Tang YC. Deep learning using linear support vector machines. arXiv:1306.0239, 2013.

14 Wen GH, Hou Z, Li HH, *et al*. Ensemble of deep neural networks with probability-based fusion for facial expression recognition. Cognitive Computation, 2017, 9(5): 597–610. [doi: 10.1007/s12559-017-9472-6]

15 Yang B, Cao JM, Ni RR, *et al*. Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. IEEE Access, 2018, 6: 4630–4640. [doi: 10.1109/ACCESS.2017.2784096]

16 Zhu XL, He ZL, Zhao L, *et al*. A cascade attention based facial expression recognition network by fusing multi-scale spatio-temporal features. Sensors, 2022, 22(4): 1350. [doi: 10.3390/s22041350]

17 Yu YL, Huo H, Liu JQ. Facial expression recognition based on multi-channel fusion and lightweight neural network. Soft Computing, 2023, 27(24): 18549–18563. [doi: 10.1007/s00500-023-09199-1]

18 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

19 Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345–1359. [doi: 10.1109/TKDE.2009.191]

20 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks.

Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: NIPS, 2012. 1106–1114.

21 Goodale MA, Milner AD. Separate visual pathways for perception and action. Trends in Neurosciences, 1992, 15(1): 20–25. [doi: 10.1016/0166-2236(92)90344-8]

22 Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]

23 Qian N. On the momentum term in gradient descent learning algorithms. Neural Networks, 1999, 12(1): 145–151. [doi: 10.1016/S0893-6080(98)00116-6]

24 Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. The Journal of Machine Learning Research, 2011, 12: 2121–2159.

25 Kingma DP, Ba J. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2014.

26 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning. Lille: ACM, 2015. 448–456.

27 Santurkar S, Tsipras D, Ilyas A, *et al*. How does batch normalization help optimization? Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: ACM, 2018. 2488–2498.

28 Qiao SY, Wang HY, Liu CX, *et al*. Micro-batch training with batch-channel normalization and weight standardization. arXiv:1903.10520, 2019.

29 Pascanu R, Mikolov T, Bengio Y. Understanding the exploding gradient problem. arXiv:1211.5063, 2012.

30 Loshchilov I, Hutter F. Decoupled weight decay regularization. Proceedings of the 7th International Conference on Learning Representations. New Orleans: ICLR, 2019.

31 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.

32 Howard A, Sandler M, Chen B, *et al*. Searching for MobileNetV3. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 1314–1324.

33 Pham L, Vu TH, Tran TA. Facial expression recognition using residual masking network. Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Milan: IEEE, 2021. 4513–4519.

34 Minaee S, Minaei M, Abdolrashidi A. Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors, 2021, 21(9): 3046. [doi: 10.3390/s21093046]

35 闫河, 李梦雪, 张宇宁, 等. 面向表情识别的重影非对称残差注意力网络模型. 智能系统学报, 2023, 18(2): 333–340. [doi: 10.11992/tis.202201003]

36 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]

37 Khanzada A, Bai C, Celepcikay FT. Facial expression recognition with deep learning. arXiv:2004.11823, 2020.

38 Huang G, Liu Z, Van Der Maaten L, *et al*. Densely connected convolutional networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2261–2269.

39 Yang HY, Ciftci U, Yin LJ. Facial expression recognition by de-expression residue learning. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 2168–2177.

(校对责编: 张重毅)