文章编号：2095-2163(2021)09-0035-07 中图分类号：TP183 文献标志码：A

基于 VGG 的人脸表情识别与分类

周义飏
（北京师范大学 人工智能学院，北京 100088）

摘 要：为了使人脸表情识别更加快速、准确，以满足复杂社会情境中的需求，本文研究了基于深度卷积神经网络的人脸表情识别方法，实现了人脸不同离散表情识别分类。针对现有数据集数据量不足、深度网络计算易出现过拟合现象等问题，本文基于人脸图片关键点进行了剪裁，获得 64 个子区域，将数据扩充为 64 倍，以达到数据增强的目的；使用基于 VGG-19 网络模型的卷积神经网络，对动作单元进行分类与强度计算，使用 *Sigmoid* 函数，使网络具备多标签多分类能力，并在 VGG-19 网络的第四组卷积层之后加入一个加权处理层，提高准确率。结果显示，增强后叠加的人脸表情识别与分类基本能够完成，而引入加权处理层后的准确率则得到了显著提高。

关键词：人脸表情；面部动作编码系统；动作单元；卷积神经网络；数据增强；加权处理

Facial expression recognition and classification based on VGG

ZHOU Yiyang
（School of Artificial Intelligence, Beijing Normal University, Beijing 100088, China）

【Abstract】 In order to make facial expression recognition faster and more accurate to meet the needs in complex social situations, this paper studies the facial expression recognition method based on deep convolutional neural networks to realize the recognition and classification of different discrete facial expressions. First of all, in view of the insufficient amount of data in the existing dataset and the prone to over-fitting in deep network calculations, this paper cuts of the face image to obtain 64 sub-regions based on the key points and expand the data to 64 times to achieve the goal of data enhancement. Secondly, the convolutional neural network based on the VGG-19 network model is used to classify and calculate the strength of the action unit, and the Sigmoid function is used to make the network have multi-label and multi-classification capabilities. Finally, a weighted processing layer is added after the fourth group of convolutional layers of the VGG-19 network to improve accuracy. The results show that the facial expression recognition and classification after data enhancement can basically be completed, and the accuracy of the weighted processing layer has been significantly improved.

【Key words】 facial micro expression; facial action coding system; action unit; convolutional neural network; date augment; weighted processing

0 引 言

表情识别，是基于人的脸部特征信息进行身份识别的一种生物识别技术，在治安、刑侦、医疗、教育、零售等领域发挥着越来越重要的作用。自二十世纪七十年代以来，人们在以面部动作编码系统为基础的表情识别道路上，向高精度、高速率、大数据的方向不断前进。

1 研究背景

人脸表情识别从早期用于治安的道路监控、机场安检所用的基于人脸识别的身份确定，到用于审讯、心理治疗的表情识别与分析，已经成为了社会发展的一个重要课题。

从上世纪六十年代开始，人们已经在探索更精确、更系统的表情识别方法，其中最具代表性的为 Paul Ekman 提出的面部编码系统（Facial Action Coding System, FACS）[1]。FACS 的出现，使所有可能的面部表情都能被描述出来，并进行组合。

2 面部动作编码系统概述

二十世纪七十年代，Paul Ekman 与合作者通过对表情的观察和生物反馈实验，描述出了不同的脸部肌肉动作与不同表情的对应关系。FACS 将人脸分成了若干个动作单元（Action Units, AU），这些动作单元依据解剖学特点划分，相互独立但彼此间又具有联系。面部动作编码主要分为 3 大类：主要动作单元编码（见表 1）、头部动作单元编码、眼睛动作单元编码。本论文主要研究的是动作单元编码。

任何表情都能反映成若干 AU 的组合[2]（见表 2）。例如表示“快乐”情绪的表情通常表现为脸颊上抬和嘴角上扬，即 AU6 与 AU12 的组合。

作者简介：周义飏（1998-），男，本科生，主要研究方向：计算机科学与技术。
收稿日期：2021-07-27

表 1 主要动作单元编码

Tab. 1 Coding of main action units

AU	描述	AU	描述	AU	描述
AU1	内部眉毛抬起	AU13	拉动嘴角向上	AU25	双唇分开暴露牙齿
AU2	外部眉毛抬起	AU14	嘴角向牙齿收缩	AU26	双唇分开看见舌头
AU4	眉毛整体低垂	AU15	嘴角垂直向下拉动	AU27	双唇分开看见喉咙
AU5	抬起上眼皮	AU16	下嘴唇向下拉动	AU28	吸吮嘴唇覆盖牙齿
AU6	抬起脸颊	AU17	挤动下唇向上顶	AU41	微微低垂上眼皮
AU7	眼睛收缩	AU18	向中间皱起嘴巴	AU42	低垂上眼皮
AU9	收缩提起鼻子	AU20	嘴唇向后方拉扯	AU43	闭上眼靖
AU10	抬起上嘴唇	AU22	嘟起嘴唇成漏斗	AU44	下眼皮向上顶
AU11	加深中部鼻唇	AU23	收紧双唇成一字	AU45	双目眨眼
AU12	上扬嘴角	AU24	把双唇挤在一起	AU46	单目眨眼

表 2 7 种基本情绪与 AU 对应关系

Tab. 2 Correspondence between 7 basic emotions and AU

基本情绪	AU
无表情喜悦	AU6+AU12
悲伤	AU1+AU4+AU5
恐惧	AU1+AU2+AU4+AU5+AU7+AU20+AU26
惊讶	AU1+AU2+AU5+AU26
愤怒	AU4+AU5+AU7+AU23
厌恶	AU9+AU15+AU16

3 国内外研究现状

3.1 人脸检测

目前,广泛使用的几种深度人脸检测算法及其效率和性能的最低要求见表 3。

主动外观模型(Active Appearance Model, AAM)可从整体人脸外观和整体形状中优化所需的参数[3]。

表 3 几种深度人脸检测算法及其效率和性能的最低要求

Tab. 3 Several deep face detection algorithms and their minimum requirements for efficiency and performance

种类	算法	关键点个数	即时	表现
整体识别	主动外观模型(AAM)	68	否	概括性较差
部分识别	树木结构模型(MoT)	39/68	否	好
	判别式反应图拟合(DRMF)	66	否	
级联回归	监督下降方法(SDM)	49	是	非常好
	增量人脸对齐方法(IFA)	49	是	
深度学习	任务约束深度卷积网络(TCDCN)	5	是	
	多任务卷积神经网络(MTCNN)	5	是	非常好

在判别模型中[4],树木结构模型(Mixtures of Trees, MoT)和判别式反应图拟合(Discriminative Response Map Fitting, DRMF),通过每个人脸坐标周围的局部外观信息来表示人脸[5]。

3.2 人脸归一化

人脸归一化主要有两种常用的方法:照度归一化和姿态归一化。

照度归一化即通过操作,使一组人脸图像的照度和对比度相同。常用的照度归一化算法包括基于各向同性扩散(Isotropic Diffusion, IS)的归一化、基于离散余弦变换(Discrete Cosine Transform, DCT)的归一化和高斯差(Difference of GAUsian, DoG)。

姿态归一化即人脸正面化。文献[6]提出了一种方法,即在对面部关键点定位后,生成通用的 3D 纹理参考模型,这些 3D 纹理参考模型适用于所有人脸图片,能有效估计可见的人脸成分。通过将每个人脸图像,反投影到参考坐标系合成初始的正面人脸。值得一提的是,生成式对抗网络(Generative Adversarial Networks, GAN)在人脸图像处理中的运用次数正飞速增长。GAN 常用于生成大量的人脸图片作为训练集与数据集,一定程度上避免了以往因寻找足够的人脸图片而遇到的各种困难。

3.3　面部 AU 特征提取

3.3.1　基于外观特征的人脸 AU 特征提取

基于外观特征的人脸 AU 特征提取，通常会用到 Gabor 小波，其通过将面部图像与一组特定的具有不同方向和比例的 Gabor 滤波器进行卷积，来进行 Gabor 表示，从而提供面部图像的多尺度特征，反映像素之间的局部相邻关系。

文献[7]中通过在面部局部区域分别应用 Haar 小波分析，设计了自动 AU 检测系统，并使用 AdaBoost 来选择特征子集。与 Gabor 方法相比，Haar 和 AdaBoost 方法有着与 Gabor 方法相似的精度，但速度却提高了若干个数量级。

3.3.2　基于几何特征的人脸 AU 特征提取

基于几何的特征，描述了面部几何信息并基于几何形状将面部动作分类。几何信息可以是一组关键点连接起来的面部网格。一些研究中，利用面部分量的变形，表情和中性面部图像之间的基准点的位置或差异[8-9]来进行识别。但并非所有的 AU 都可以仅仅通过几何表示来识别，例如 AU6 的特征包括眼睛外角周围的皮肤起皱和脸颊隆起，这很难通过变形来识别。同时，几何特征也无法检测出细微的面部特征，例如皱纹或纹理变化[10]。

3.3.3　基于混合特征的人脸 AU 特征提取

一些研究整合了基于外观特征与基于几何特征两种方法[11-12]，并且结合了整体表示与局部表示、小波分析表示与直方图表示、低级表示与高级表示。文献[13]通过引入拓扑结构和关系约束提出多条件潜在变量模型。该模型将特征和模型级别的 AU，依赖项编码用于 AU 识别的学习中，对于 9 个 AU 进行操作，其最佳识别精度达到 92.7%。

4　数据预处理

4.1　人脸对齐

本方法对 1 268 张人脸图片进行识别与检测，在下巴、双眉、双眼、双唇、鼻梁、鼻尖 9 个部位返回 68 个关键点（不同部位的某些关键点可能重合）。图 1 为未处理的人脸图片，经过处理后可得到各关键点的坐标如图 2。

这些关键点与人脸各个部位的位置相对应，将在接下来的仿射变换中起到定位作用。本步骤运用仿射变换实现人脸对齐，对齐后的人脸图片如图 3 所示。

本文进行仿射变换的具体思路为：分别计算左、右眼中心坐标、计算左右眼中心坐标连线与水平方向的夹角、计算左右两眼整体中心坐标、以左右两眼整体中心坐标为基点，将图片逆时针旋转相应角度，使左右眼中心坐标连线与水平方向平行，确保人脸图片为视觉上的正向。

图 1　8 张未经处理的人脸图片

Fig. 1　8 unprocessed face pictures

图 2　含关键点的人脸图片

Fig. 2　Face pictures with key points

图 3　对齐后的人脸图片

Fig. 3　Face pictures after alignment

4.2　剪裁

实际上，对于执行过上一步骤的图片，CNN 可以较为精确地选取出图片中的人脸部分。但为了减少 CNN 的执行时间，需尽可能减少图片中的无效部分。根据 landmark 裁剪人脸到固定尺寸，水平方向以最靠左和最靠右的 landmark 中点为裁剪后图片的中心点，垂直方向上分为 3 部分：中部（双眼中心到双唇中心的像素距离）、底部和顶部（双眼中心到双唇中心的距离）。裁剪后的图片为边长为 138 像素的正方形，如图 4 所示。

4.3　数据增强

本文使用基于 68 张人脸图片关键点的“图像扩充”，即对每个关键点取子区域，使每个子区域能包含至少 2/3 的人脸区域，从而将数据量扩大至 69 倍，原本的 1 268 张人脸图片扩充为 87 492 张图片，但依然只反映 1 268 张人脸，如图 5 所示。子区域均为边长为 92 像素的正方形。

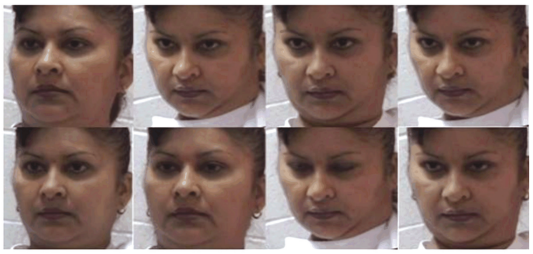

图 4 剪裁后的人脸图片

Fig. 4 Face pictures after cropping

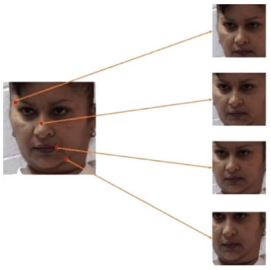

图 5 数据增强示例

Fig. 5 Example of Data enhancement

为了方便表示,建立 2 个平面直角坐标系记为:坐标系 *A* 与坐标系 *B*。分别以经上一步骤剪裁后的图片左上角顶点和每个关键点子区域左上角顶点为原点(如图 6)。以图 5 为例,各个关键点及其对应的坐标见表 4。

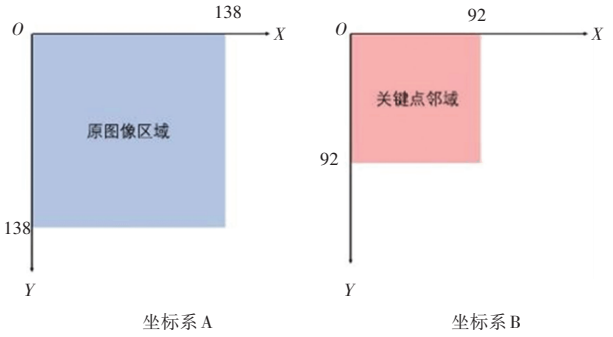

图 6 两个坐标系

Fig. 6 Coordinate systems

表 4 各个关键点及其对应的坐标

Tab. 4 Key points and their corresponding coordinates

部位	坐标系 A 中关键点坐标	坐标系 B 中关键点坐标
左眉	(111,33),(102,28),(91,28)(81,31)(72,36)	(58,10)
右眉	(21,33)(27,28)(36,28)(44,32)(52,37)	(32,10)
左眼	(78,49)(84,43)(92,43)(99,46)(93,50)(85,50)	(68,28)
右眼	(51,49)(44,43)(37,43)(31,46)(36,49)(44,50)	(24,28)
鼻	(62,47)(62,57)(62,67)(50,82)(55,85)(62,88)(69,85)(75,82)	(46,29)
嘴	(56,98)(62,99)(68,99)(68,103)(62,103)(56,102)(44,97)(47,98)(50,94)(56,93)(63,95)(62,95)(68,94)(77,95)(83,99)(87,99)(77,107)(68,110)(62,110)(56,109)(50,105)	(46,63)
左轮廓	(15,43)(15,58)(16,73)(19,88)(24,102)(31,114)(41,125)(51,143)(65,137)	(84,64)
右轮廓	(80,136)(95,129)(108,120)(118,108)(126,94)(129,77)(131,60)(131,42)	(8,64)

5 基于深度卷积网络的人脸表情识别与分类

5.1 卷积神经网络

CNN 是一类可进行卷积计算并且具有深度结构的前馈神经网络(Feedforward Neural Networks, FNN),是深度学习的代表算法之一。CNN 主要由输入层、池化层、激活函数、卷积层、全连接层 5 个部分组成。

深度卷积网络将小的神经网络串联起来,从而构成深度神经网络。以三维图进行卷积处理为例,如图 7 所示,同一卷积核对不同输入层进行卷积操作,得到一组输出,多个卷积核则得到多个输出。

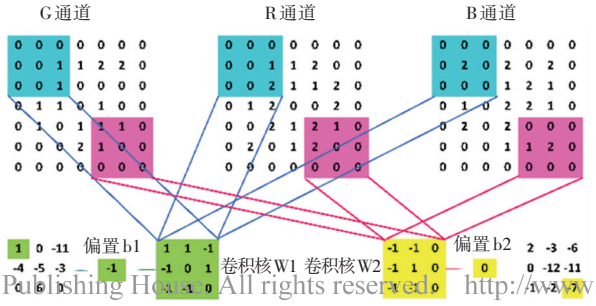

图 7 CNN 的卷积过程

Fig. 7 Convolution process of CNN

5.2　基于 VGG-19 的 AU 识别网络结构

VGG 网络是 Oxford Visual Geometry Group 于 2014 年提出的一种 CNN 模型，其采用连续的小卷积核代替较大卷积核，以获取更大的网络深度。例如，使用 2 个 3 * 3 卷积核代替 5 * 5 卷积核（图 7）。这种方法使得在确保相同感知野的条件下，VGG 网络具有比一般的 CNN 更大的网络深度，提升了神经网络特征提取及分类的效果。VGG 网络与其他几种常用的 CNN 模型对比见表 5。

表 5　几种常用的 CNN 模型

Tab. 5　Several CNN models

	AlexNet	VGGNet	GoogleNet	ResNet
提出年份	2012	2014	2014	2015
层数	5+3	13/16+3	21+1	151+1
内核大小	11,5,3	3	7,1,3,5	7,1,3,5
数据增强	是	是	是	是
Dropout	是	是	是	是
Inception	是	是	是	否
批量统一处理	是	是	否	是

本方法使用的 VGG-19 网络包含了 19 个隐藏层、16 个卷积层和 3 个全连接层。该网络模型使用的卷积核均为 3 * 3 卷积，池化层则采用 2 * 2 最大值池化（图 8）。

以往的研究中通常使用 *Soft - Max* 作为激活函数，损失函数则使用分类交叉熵，但这种方法仅适用于单标签分类。而本文方法不仅实验 AU 分类，更要对同一 AU 的不同强度进行识别和分类，因此需要进行多标签多分类。本文方法需要分类的 AU 为 12 个，每个 AU 分为 0~5 个强度等级，总共为 60 种分类项。因此，本文采用二分类叠加使用的方式，即先对不同 AU 种类进行二分类，再对单个 AU 所对应的不同强度进行二分类，最后将每种 AU 与对应强度结合形成对照表。采用 *Sigmoid* 函数作为激活函数，损失函数使用二进制交叉熵函数。

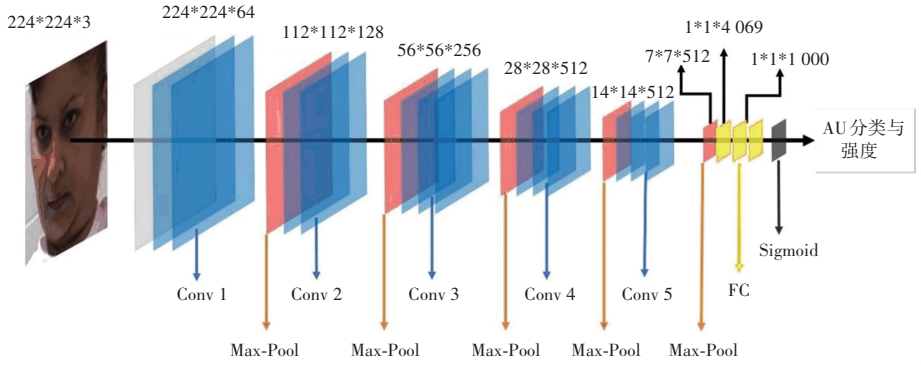

图 8　VGG-19 网络结构

Fig. 8　Structure of vgg-19 network

在后续的实验中，还将向 VGG19 网络中加入一个加权处理层。

5.3　叠加的人脸 AU 特征检测

5.3.1　数据来源

DISFA 是一个无姿势的面部表情数据库。该数据库包含具有不同种族的 27 位成人受试者（12 位女性和 15 位男性）的立体声视频。使用 PtGrey 立体成像系统以高分辨率（1 024×768）采集图像，由 FACS 专家手动对所有视频帧的 AU（0~5）强度进行评分。

5.3.2　AU 分类与强度计算

87 492 张图片分为两个数据集，其中 86 224 个子区域图片为训练集，1 268 张人脸图片为测试集，输入至 VGG-19 网络。

图 9 与图 10 分别展现了使用 VGG-19 网络训练 30 个 Epoch 和 60 个 Epoch 的效果。

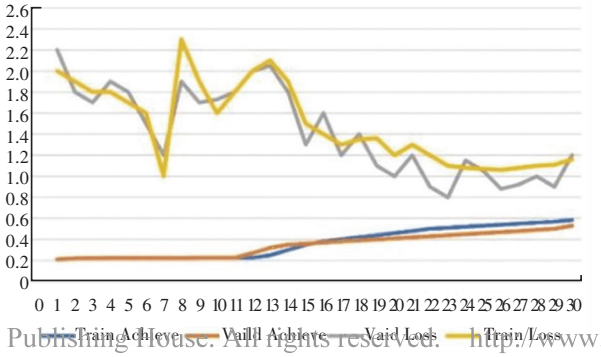

图 9　训练 30 个 Epoch 的效果

Fig. 9　Effect of training 30 epochs

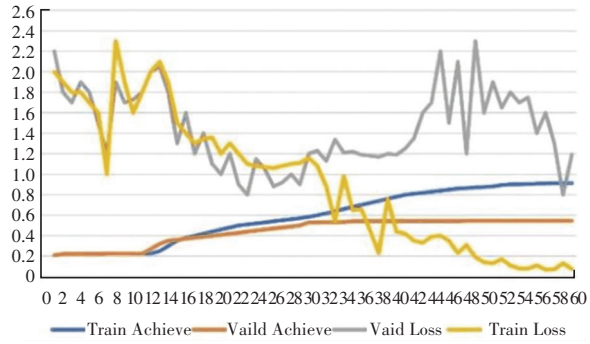

图 10 训练 60 个 Epoch 的效果

Fig. 10 Effect of training 60 epochs

可以看出,训练至第 30 个 Epoch 时测试集的准确率几乎不再发生改变,训练至第 60 个 Epoch 时测试集的准确率为 54.52%。

结果呈现为每个人脸图片 12 种 AU 的强度(0~5)。以图 11 为例,其 12 种 AU 的强度呈现在右侧表中所示。

AU	强度
AU1	0
AU2	5
AU4	1
AU5	2
AU6	5
AU9	2
AU12	4
AU17	1
AU20	3
AU25	0
AU26	0
AU43	0

图 11 人脸图片及其含有的 AU 强度

Fig. 11 Face image and its AU intensity

总体来看,该方法能基本满足人脸 AU 分类与强度计算,实现人脸表情分类。因此,为提高精度,将引入一个加权处理层。

6 人脸表情分类优化

为了提高 VGG 网络进行人脸表情分类时的精度,本文将通过在 VGG-19 网络中加入加权处理层,实现加权处理下的人脸表情分类。

6.1 加权处理层

加权处理层在 VGG 网络中的位置如图 12 所示。在这一层中,经过 4 组卷积层处理后的人脸图片会根据含有的 AU,被划为若干个子区域,子区域的划分是基于 AU 区域的中心。AU 中心为完成每个 AU 所需的面部器官对应关键点构成的矩形中心,而以这些中心为中心的边长 6 像素的正方形区域,为该 AU 中心的子区域。

在划分 AU 子区域后,对于子区域内的每个 1 * 1 像素块,计算其到 AU 中心的曼哈顿距离。

设 A 为权重,d 为该位置到 AU 中心的曼哈顿距离。由于经过第三个池化层和第四组卷积层处理后的图片大小为 28 * 28,子区域内每个位置的权重以该位置距离的 0.1% 进行衰减,即距离每增加 1 像素,权重减少 0.028。A 与 d 的关系如式(1)。

$$A = 1 - 0.028d \tag{1}$$

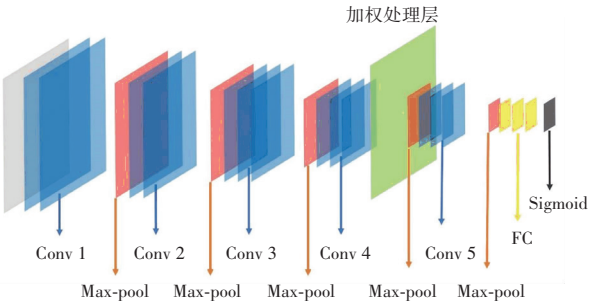

图 12 加权处理层在 VGG 网络中的位置

Fig. 12 Position of the weighted processing layer in the VGG network

6.2 结果比较

引入加权处理层后,训练 30 个 Epoch 和 60 个 Epoch 的效果分别如图 13 和图 14 所示。

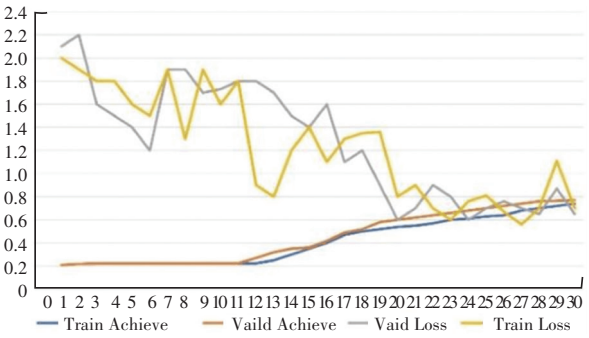

图 13 训练 30 个 Epoch 的效果

Fig. 13 Effect of training 30 epochs

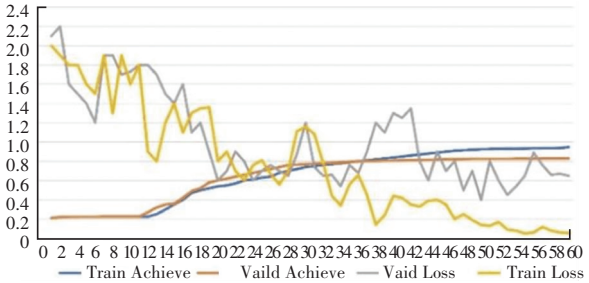

图 14 训练 60 个 Epoch 的效果

Fig. 14 Effect of training 60 epochs

结果显示,直接运用 VGG-19 网络进行表情分类的测试集准确率在训练 60 个 Epoch 时仅为 54.52%,而引入加权处理层后的准确率达到了 83.76%。即引入加权处理层能显著提高 VGG 网络进行表情分类的准确率。

7　结束语

本文采用了两种提高准确率的方法：一是在数据预处理阶段对图片进行二次剪裁，实现数据增强，在运用 VGG 网络进行训练时使用叠加后的数据集，提高了准确率，同时避免了以往研究中为获取庞大数据集而遇到的种种困难；二是在 VGG-19 网络的第四组卷积层和第五组卷积层之前加入一个加权处理层，从而提高准确率，最终使测试集准确率相比未引入加权处理层时提高了 53.63%。

在采用了提高准确率的新方法的同时，也存在一些有待改进之处，主要体现在人脸图片样本较为单一、缺少其它卷积神经网络模型的对比、加权处理层作用较为单一等问题，有待进一步研究解决。

参考文献

[1] WANG S J, YAN W J, LI X, et al. Micro-expression recognition using color spaces [C] //2015 IEEE Transactions on Image Processing, 2015,24(12):6034-6047.

[2] EL KALIOUBY R, ROBINSON P. Real-time inference of complex mental states from facial expressions and head gestures [M]//Real-time vision for human-computer interaction. Springer, Boston, MA, 2005: 181-200.

[3] COOTES T F, EDWARDS G J, TAYLOR C J. Active appearance models[J]. IEEE Transactions on pattern analysis and machine intelligence, 2001, 23(6): 681-685.

[4] ZHU X, RAMANAN D. Face detection, pose estimation, and landmark localization in the wild[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 2879-2886.

[5] ASTHANA A, ZAFEIRIOU S, CHENG S, et al. Robust discriminative response map fitting with constrained local models [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3444-3451.

[6] HASSNER T, HAREL S, PAZ E, et al. Effective face frontalization in unconstrained images [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4295-4304.

[7] WHITEHILL J, OMLIN C W. Haar features for FACS AU recognition [C]//7th international conference on automatic face and gesture recognition (FGR06). IEEE, 2006: 5-101.

[8] PANTIC M, ROTHKRANTZ L J M. Facial action recognition for facial expression analysis from static face images [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2004, 34(3): 1449-1461.

[9] TIAN Y I, KANADE T, COHN J F. Recognizing action units for facial expression analysis [J]. IEEE Transactions on pattern analysis and machine intelligence, 2001, 23(2): 97-115.

[10] VALSTAR M F, PANTIC M. Fully automatic recognition of the temporal phases of facial actions [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2011, 42(1): 28-43.

[11] NICOLLE J, BAILLY K, CHETOUANI M. Real-time facial action unit intensity prediction with regularized metric learning[J]. Image and Vision Computing, 2016, 52: 1-14.

[12] NICOLLE J, BAILLY K, CHETOUANI M. Facial action unit intensity prediction via hard multi-task metric learning for kernel regression [C]//2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). IEEE, 2015, 6: 1-6.

[13] ELEFTHERIADIS S, RUDOVIC O, PANTIC M. Joint facial action unit detection and feature fusion: A multi-conditional learning approach [J]. IEEE transactions on image processing, 2016, 25(12): 5727-5742.

(上接第 34 页)

[6] ZAHARIA M, CHOWDHURY M, FRANKLIN M J, et al. Spark: Cluster computing with working sets[J]. HotCloud, 2010, 10(10): 95.

[7] CHANG F, DEAN J, GHEMAWAT S, et al. Bigtable: A distributed storage system for structured data[J]. ACM Transactions on Computer Systems (TOCS), 2008, 26(2): 1-26.

[8] ZAHARIA M, CHOWDHURY M, DAS T, et al. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing [C]//Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12). 2012: 15-28.

[9] CODD E F. Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate[J]. http://www.arborsoft.com/papers/coddTOC.html, 1993.

[10] INMON W H. Building the data warehouse[M]. John wiley & sons, 2005.

[11] GRAY J, CHAUDHURI S, BOSWORTH A, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals[J]. Data mining and knowledge discovery, 1997, 1(1): 29-53.

[12] 赵本本, 殷旭东, 王伟. 基于 Scrapy 的 GitHub 数据爬虫[J]. 电子技术与软件工程, 2016 (6): 199-202.

[13] 汪文妃, 徐豪杰, 杨文珍, 等. 中文分词算法研究综述[J]. 成组技术与生产现代化, 2018 (3): 1.

[14] ZHANG D, ZHAI C, HAN J. Topic Cube: Topic modeling for olap on multidimensional text databases[C]//Proceedings of the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2009: 1124-1135.

[15] YANG Z, MA H, HE Z, et al. Finding maximal ranges with unique topics in a text database[J]. World Wide Web, 2018, 21(2): 289-310.