基于改进的 ResNet 网络的人脸表情识别

周 婕[1],马明栋[2]
(1. 南京邮电大学 通信与信息工程学院,江苏 南京 210003;
2. 南京邮电大学 地理与生物信息学院,江苏 南京 210003)

摘 要:近几年来,人工智能的热度一直居高不下,其中作为人机交互的一种重要方法—人脸表情识别已经成为计算机视觉研究的热点。从传统的机器学习算法到现在的深度学习,识别效率也在不断地提高,为了进一步提高人脸表情识别率,在传统的卷积神经网络的基础上,提出了一种基于改进的 ResNet 卷积神经网络的表情识别方法。该方法基于 ResNet 网络的基本结构,采用的中间卷积部分是前后各一个卷积核为 1 * 1 的卷积层,中间是卷积核大小为 3 * 3 的卷积层,同时将下采样移到后面的 3 * 3 卷积层里面去做,减少信息的流失,并用 PReLU 替代 ReLU 激活函数。与 ResNet 模型相比,改进的网络结构可以减少计算量,提高识别速度和识别率。利用 Tensorflow 构建经过改进的 ResNet 卷积神经网络框架,并在增强的 Fer2013 数据集上进行了训练,得到了准确且高效的人脸表情识别模型,最后再结合 OpenCV 中的人脸检测分类器,从视频中抓取人脸进行识别,实现了实时识别人脸表情效果的输出。实验结果表明,改进的 ResNet 卷积神经网络模型较其他的人脸表情识别方法在识别率上有了一定的提高。

关键词:表情识别;深度残差网络;深度学习;OpevCV;人脸检测
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)01-0025-05
doi:10.3969/j.issn.1673-629X.2022.01.005

Facial Expression Recognition System Based on Improved ResNet

ZHOU Jie[1], MA Ming-dong[2]
(1. School of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. School of Geographical and Biological Information, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In recent years, the popularity of artificial intelligence has remained high. Among them, as an important method of human-computer interaction, facial expression recognition, has become a hotspot in computer vision research. From traditional machine learning algorithms to the current depth learning, recognition efficiency is constantly improving. In order to further improve the recognition rate of facial expressions, on the basis of convolutional neural network, an expression recognition method based on the improved ResNet convolutional neural network is proposed. This method is based on the basic structure of the ResNet network. The middle convolution part is a 1 * 1 convolution layer before and after the convolution kernel, and the middle is a convolution layer with a convolution kernel size of 3 * 3, and the downsampling is shifted do it in the following 3 * 3 convolution to reduce the loss of information and replace the ReLU function with PReLU. Compared with the ResNet model, the improved network structure can reduce the amount of calculation and increase the recognition speed and recognition rate. The Tensorflow is used to build an improved ResNet framework and trains it on the enhanced Fer2013 data set to obtain an accurate and efficient facial expression recognition model, and finally combines the face detection classifier in OpenCV to grab faces from the video for recognition, thus realizing the output of real-time recognition of facial expressions. The experimental results based on the data set show that the recognition rate of the improved ResNet convolutional neural network model is indeed improved compared with other facial expression recognition methods.

Key words: expression recognition; ResNet; deep learning; OpenCV; face detection

---

收稿日期:2020-12-28 **修回日期**:2021-04-28
基金项目:江苏省自然科学基金-青年基金项目(BK20140868)
作者简介:周 婕(1995-),女,硕士研究生,CCF 会员(F3964G),研究方向为图像处理;马明栋,博士,教授,研究方向为地理信息系统平台软件设计与开发等。

0 引 言

众所周知,人脸表情作为非语言交际的一种形式,包含着丰富的情感信息,同时传达出一些有关人的认知行为、性格和心理情绪,虽然显示出的信息是比较隐晦的,但更能实时地、真实地反映出人的内心活动,真实性更高,且这种信息表达方式不能被其他方式所替代,因此人脸表情在人们的日常交流中占据着重要地位。随着计算机技术的快速发展,人们对人工智能的研究更加深入,希望通过计算机能模拟人类行为,提高人类的生活质量,造福人类。因此人脸表情识别技术作为通过计算机来预测人类心理状态的一种方式具有广阔的应用前景,比如在教育、医学、心理学、商业、安全驾驶等各大领域都有对此技术的研究。

人脸表情识别的关键就在于人脸不同表情特征点的提取,然而传统的特征提取算法[1-2],如尺度不变特征变换(SIFT)、局部二值模式(LBP)等,不仅设计方法比较困难,而且特征点提取不完全,从而导致效率低下。因此,研究人员将卷积神经网络如 AlexNet[3]、VGGNet[4]、GoogleNet[5]等用于人脸表情识别。卷积神经网络以其能够共享卷积核,对高维数据处理无压力且特征分类效果好的独特优势,在图像、语音处理方面得到广泛的应用。但随着网络深度的加深,学习能力的加强,反而造成了梯度爆炸和梯度消失,从而出现了所谓的“退化”问题,即优化效果越来越差,测试数据和训练数据的准确率也越来越低。基于这样的背景,He Kaiming 等人提出了 ResNet[6] 网络模型,与其他网络最主要的区别就是在卷积神经网络中引入了残差的思想,主要是通过添加 shortcut 连接,把通过跳层连接的梯度更新成一样,解决了网络变深之后,前面层次的网络权值得不到更新,从而导致梯度消失的问题。

为了更加高效且准确地识别出人脸表情,该文提出了基于改进的 ResNet 卷积神经网络,该网络的参数量和运算量更少,能够更快更好地提取人脸表情特征,且保存了“最有辨识力”的信息。同时为了能够实时识别人脸表情,直接利用 OpenCV 中的基于 Haar 特征的人脸检测分类器,实现了从视频中抓取人脸,并加载训练好的模型,最终实现对视频人脸表情的实时识别系统。

1 相关技术

1.1 OpenCV

OpenCV 是开源的计算机视觉和机器学习软件库,可以运行在不同的操作系统上。它主要由 C 函数和 C++类组成,但也提供了其他语言接口。OpenCV 在图像处理方面提供了很多通用算法,因此大大提高了图像处理的效率。该文主要使用到它的视频处理模块和人脸检测分类器。

视频处理模块主要是从视频序列读取帧,因此只需创建一个 cv::VideoCapture 类的实例,然后在一个循环中提取并显示视频的每帧就可以读取数据进行处理了。

人脸检测分类器[7] 主要是检测并分割出人脸,本系统选用的是 Haar 特征加上 Adaboost 级联分类器的组合。Haar 特征[8] 是一种反映图像灰度变化,像素分模块求差值的特征,主要利用黑色和白色这两种矩形组成特征模板,再用两种矩形像素和的差值作为该特征模板的特征值。通常,人脸的眼睛要比脸颊颜色要深,鼻梁两侧比鼻梁颜色要深,嘴巴比周围颜色要深等,因此,人脸的一些特征就可以由 Haar 特征来简单描述。然而如果对一幅图像采用全部的特征去检测,效率必定非常低,因此 Haar 特征一般结合 Adaboost 级联分类器,将全部特征分为各个阶段,并且每个阶段的特征是逐渐增加的,使得人脸检测效率大大地提高了。

1.2 经典的 ResNet 网络

1.2.1 ResNet 网络的提出

深度卷积神经网络起源于 AlexNet 网络,后来针对此网络的不足,研究人员又提出了 VGGNet、GoogleNet 等网络,不难看出随着网络深度的增加,网络的表达能力越来越强大,识别的速度和准确率也不断上升,这是因为网络越深,所能获取的信息越多,提取的特征也越丰富。然而实验表明不断加深的网络深度并没有得到人们预期的识别结果,反而出现了“退化”现象,优化效果变差,测试数据和训练数据的准确率也降低了,这是因为网络的加深会造成梯度爆炸和梯度消失的问题。针对这个现象,对输入数据和中间层的数据进行归一化操作,这种方法可以保证网络在反向传播中采用随机梯度下降(SGD)[9],从而让网络达到收敛。但是归一化操作只能解决深度为几十层的网络的梯度消失问题,如果网络深度再加深的话,这种方法就无效了。

为了让更深的网络也能训练出好的效果,He Kaiming 等人提出了新的网络结构—ResNet,通过使用 Residual Unit 成功训练了深度为 152 层的卷积神经网络,并在 ILSVRC 2015 比赛中获得了冠军。ResNet 网络在获得低误差率,需要较小的参数量和计算量的同时,也加快了模型训练的速度,使得训练模型的效果非常突出。

ResNet 较之于其他网络,最主要的区别就是在卷积神经网络中引入了残差函数。ResNet 网络在内部的残差块使用了跳跃连接,这样做的好处是缓解了在卷积神经网络中增加深度带来的梯度消失的问题,使得 ResNet 网络容易优化,即能够通过增加网络的深度

来提高准确率。

1.2.2　ResNet 网络结构

ResNet 网络见图 1,主要由输入部分、卷积部分以及输出部分组成,其中卷积部分又分为四个阶段[10]。

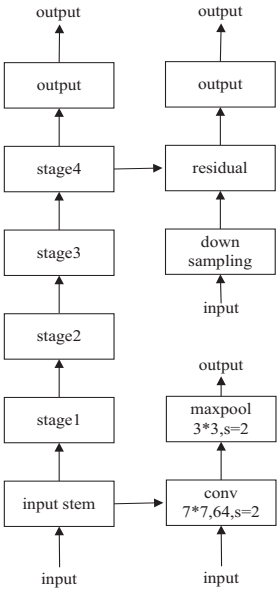

图 1　ResNet 网络结构

由图 1 可知,ResNet 网络的输入部分主要由大卷积核和最大池化[11]这两个部分组成,这一步的目的是为了将大像素的输入图像变成小像素的特征图像,在尽量保留含有信息的特征点的同时,也减少了存储所需的大小。

中间卷积部分是此网络结构的核心,引入的残差块将输入数据分成两条路,如图 2,一条路经过 2 个卷积核为 3 * 3 的卷积层,另一条路则直接短接(shortcut),通过 shortcut 将输入和输出进行一个 element-wise 的加叠,这个简单的加法不仅不会为网络增加额外的参数和计算量,而且还可以加快模型的训练速度,提高模型的训练效果,这样做可以有效地解决梯度爆炸和梯度消失的问题。最终两条路相加并经过 ReLU 激活函数处理后输出。这一步的目的是为了实现特征信息的提取。

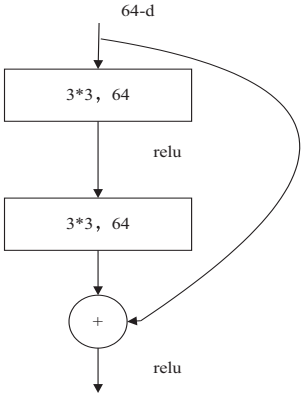

图 2　ResNet 残差块

最后的输出部分先是通过全局自适应平滑池化,然后接全连接层[12]输出,这样做的好处是首先通过 GAP 减少了参数的数量,降低过拟合的发生几率,再连接 FC 就是高度提纯的特征,方便交给最后的分类器。

1.3　改进的 ResNet 网络

1.3.1　残差块的优化

Basicblock 结构如图 2,将输入数据分成两条路,一条路经过卷积层,另一条路则直接短接(shortcut),最终两条路相加并经过 relu 激活函数处理后输出。Bottleneck 结构如图 3,对残差块做了计算优化,即将两个 3 * 3 的卷积层替换为两个 1 * 1 的卷积层加上一个 3 * 3 的卷积层,虽然在原来的结构上增加了一个卷积层,但是通过第一个 1 * 1 卷积层的降维处理后又在最后一个 1 * 1 卷积层下进行了还原处理,这样做既保持了精度又减少了计算量。直接计算来计较一下两个结构的计算量,比如说,对于 256 维的输入特征,Bottleneck 结构的参数数目为 $1\times1\times256\times64+3\times3\times64\times64+1\times1\times64\times256=69\ 632$,Basicblock 结构的参数数目为 $(3\times3\times256\times256)\times2=1\ 179\ 648$,计算量简化了约 6%。

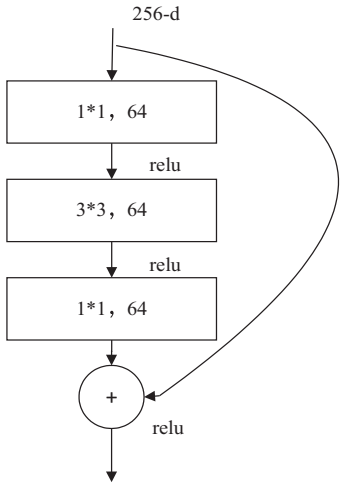

图 3　ResNet 残差块的优化

1.3.2　下采样部分的改进

原本的下采样[13]是在每个阶段的第一个卷积下去做的,这样做的后果就是输入数据会通过一个 stride =2 的 1 * 1 卷积,直接使得特征图的尺寸缩小了一半,大量的特征信息丢失,使训练的模型不够精确,从而导致识别率降低。因此,该文将下采样这一步骤转移到 3 * 3 的卷积里面,这样做的好处就是避免大量的信息流失,使特征信息的提取更加完整。

1.3.3　激活函数的改进

ResNet 网络是在每个卷积之后都加入了 ReLU 激活函数[14],主要目的是为了引入非线性因素,将神经网络可以应用到非线性模型中,提高神经网络对模型

的表达能力。

ReLU 的函数公式是 $f(x) = \max(0,x)$，函数图像如图 4 所示。

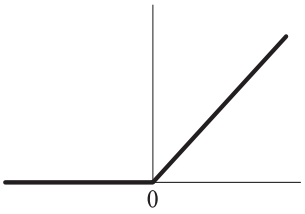

图 4 ReLU 激活函数图像

ReLU 函数能够加快计算与收敛速度，而且在一定程度上缓解梯度消失的问题，但是从 ReLU 的函数图像中可以看出如果输入小于 0 的话，经 ReLU 函数激活后输出为 0，这相当于完全没有激活，这个函数也是“死掉的”，即产生所谓的“dying relu”问题，导致后面的权值不再更新，影响到网络的表达能力。因此尝试用 PReLU 激活函数[15]替代 ReLU 函数。

PReLU 的函数公式是 $f(x) = \max(a \times x,x)$，函数图像如图 5 所示。

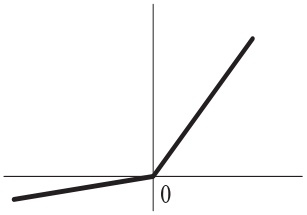

图 5 PReLU 激活函数图像

从图中可以看出在负数区域内，PReLU 有一个很小的斜率，这样既保留了 ReLU 函数的优点，同时又能避免“dying relu”问题。同时调整 PReLU 的位置，将相加后的激活函数移入残差块内部，加强模型的表达能力。

2 系统整体流程设计

基于改进的 ResNet 网络的人脸表情识别系统的整体流程设计如图 6 所示。该系统主要由三个部分组成，分别是视频数据的读取、人脸检测与人脸图像提取、人脸表情的预测及结果输出。

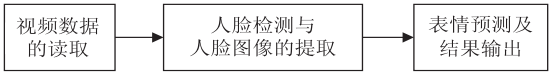

图 6 系统流程

视频数据的读取：由于视频或摄像头的实时画面是由一帧一帧的图像组成，因此动态的数据读取本质上是图像的读取。该文使用 OpenCV 的 VideoCapture 函数读取摄像头数据也就是当前帧图像，将获取的实时画面数据存放在定义的 Mat 数据容器(frame)中，并判断 frame 是否为空，若不为空则使用窗口显示读取到的图像。

人脸检测与人脸图像提取：这步的作用主要是定位到人脸图像并截取出来，为之后的人脸表情识别做准备。本系统利用 OpenCV 自带的 Haar 特征人脸检测器。Haar 特征主要是根据人脸的立体感造成的灰度变化而通过像素分模块求差值，这样，人脸的一些特征就可以由 Haar 特征来简单描述，再结合 Adaboost 级联分类器，将全部特征分为各个阶段，提高人脸检测效率。OpenCV 已经包含许多用于脸部、眼睛、嘴巴等的预先分类器，这些 XML 文件存储在 opencv/data/haarcascades/文件夹中。首先加载所需的 XML 分类器，XML 中存放的是训练后的特征池，其中特征大小是根据训练时的参数而定，然后以灰度模式加载输入图像(或视频)，最后就可以在图像中定位人脸位置，如果找到人脸的话，它会以坐标形式返回检测到的脸部的位置从而提取出人脸图像。整个人脸检测过程的原理是将存放在 XML 中的每个固定大小的特征与输入图像同样大小的区域进行对比，如果相符则记录此矩形区域的位置，然后滑动窗口，重复以上步骤检测图像的其他区域。

人脸表情预测及结果输出：以 Tensorflow2.0 深度学习框架为基础实现改进后的卷积神经网络来训练模型，再利用训练好的模型完成对人脸表情的预测，并判断抓取的人脸表情属于哪个标签，最后输出识别结果。

3 实 验

3.1 数据集选取

该文选取 Fer2013 数据集[16]作为人脸表情识别研究的数据集，虽然该数据集的测试集存在许多标签的错误，导致测试精度不是很高，但本身已划分了训练集、验证集和测试集，因此选用该数据集，有利于在相同条件下将文中方法与其他相关方法进行比较。

Fer2013 人脸表情数据集由 35 886 张人脸表情图片组成，其中，训练图(Training)28 708 张，公共验证图(PublicTest)和私有验证图(PrivateTest)各 3 589 张，每张图片是由大小固定为 48 * 48 的灰度图像组成，共有 7 种表情，分别对应于数字标签 0 ~ 6，具体表情对应的标签和中英文如下：0 anger 生气；1 disgust 厌恶；2 fear 恐惧；3 happy 开心；4 sad 伤心；5 surprised 惊讶；6 normal 中性。

数据集并没有直接给出图片，而是将表情、图片数据、用途的数据保存到 csv 文件中，第一列表示表情标签，第二列为原始图片数据，最后一列为用途。这样处理数据的目的是为了方便训练时读取数据。

3.2 数据增强

一般来说，训练的数据量越大，系统的识别率也越精确，因此，为了得到一个比较成功的神经网络，就需

要大量的参数。然而,实际情况中,并没有这么多的数据可以用于训练,因此,在将数据提供给模型之前进行扩充即增强数据[17]。数据集增强主要是为了加大训练的数据量,提高模型的泛化能力及模型的鲁棒性,减少网络的过拟合现象。该文通过对训练图片进行如随机缩放、翻转、平移、旋转等变换操作来增强数据,使数据集的数据量增加了数十倍。

3.3　实验结果及分析

将增强后的 Fer2013 数据集分别放入 ResNet 和改进后的 ResNet 网络中进行训练和测试,可以得到如表 1 所示的准确率。ResNet 在 Fer2013 数据集上的准确率为 70.3%,改进后的 ResNet 在相同数据集上的准确率为 73.2%,准确率提高了将近 3%,说明改进后的 ResNet 网络确实能够提高人脸表情的识别率。

表 1　不同模型在数据集上的准确率对比

数据集	模型	准确率/%
FER2013	ResNet	70.3
FER2013	改进后的 ResNet	73.2

得到训练模型后,加载 OpenCV 自带的 Haar 特征的人脸检测器和训练好的模型,先是通过摄像头按帧读取图像,然后从图像中检测出并截取出人脸,利用训练好的模型完成对人脸表情的预测,判断抓取的人脸表情属于哪个标签,最后输出识别结果。改进方法的效果如图 7 所示,结果可以接受。

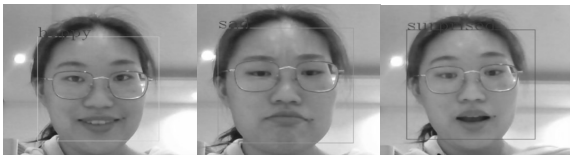

图 7　识别结果

4　结束语

本系统实现了结合深度学习来进行人脸表情识别的输出,主要是基于传统 ResNet 网络的基本结构,并对其进行了优化。输入和输出部分仍保持原来的结构,主要是对中间的卷积部分进行了改进:将中间卷积部分改为前后各一个卷积核为 1 * 1 的卷积层,中间是卷积核大小为 3 * 3 的卷积层,这样做既可以减少计算量又可以保持精确度不下降;将下采样移到后面的 3 * 3卷积里面去做,目的是为了减少信息的流失,最大程度地保证有信息量的特征点保留下来;用 PReLU 替代 ReLU 函数,同时调整激活函数的位置,可以在提高神经网络对模型表达能力的同时避免出现“dying relu”的问题。最后的实验结果表明,与传统 ResNet 模型相比,改进的网络结构减少了计算量,提高了识别速度以及识别率。

参考文献:

[1] TARIQ U,LIN K H,LI Z,et al. Recognizing emotions from an ensemble of features[J]. IEEE Transactions on Systems, Man,and Cybernetics,Part B (Cybernetics),2012,42(4):1017-1026.

[2] GIRSHICK R,DONAHUE J,DRRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//IEEE conference on computer vision and pattern recognition (CVPR). Columbus, OH, USA: IEEE, 2014:580-587.

[3] KRIZHEVSKY A,SUTSKEVER I,HINTON G E,et al. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM,2017,60(6):84-90.

[4] ZENG N Y,ZHANG H,SONG B Y,et al. Facial expression recognition via learning deep sparse autoencoders[J]. Neurocomputing,2018,273:643-649.

[5] SZEGEDY C,WEI L,JIA Y,et al. Going deeper with convolutions[C]//IEEE conference on computer vision and pattern recognition (CVPR). Boston, MA, USA: IEEE, 2015: 1-9.

[6] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//IEEE conference on computer vision and pattern recognition (CVPR). Las Vegas, NV, USA: IEEE,2016:770-778.

[7] 罗明刚,李一民,曾素娣. 基于 AdaBoost 算法的人脸检测研究[J]. 计算机工与数字工程,2007,35(2):7-8.

[8] SUN X,LV M. Facial expression recognition based on a hybrid model combining deep and shallow features[J]. Cognitive Computation,2019,11(4):587-597.

[9] 王功鹏,段　萌,牛常勇. 基于卷积神经网络的随机梯度下降算法[J]. 计算机工程与设计,2018,39(2):441-445.

[10] 吴宇豪,陈晓辉. 基于改进的 ResNet 的人脸表情识别系统[J]. 信息通信,2020(7):37-39.

[11] 卢官明,何嘉利,闫静杰,等. 一种用于人脸表情识别的卷积神经网络[J]. 南京邮电大学学报:自然科学版,2016,36(1):16-22.

[12] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large scale image recognition[J]. Computer Science,2014,52(3):1-14.

[13] 余　锐. 基于深度学习的人脸表情特征分析[J]. 现代计算机,2018(13):49-53.

[14] 王红霞,周家奇,辜承昊,等. 用于图像分类的卷积神经网络中激活函数的设计[J]. 浙江大学学报:工学版,2019,53(7):1363-1373.

[15] 田　娟,李英祥,李彤岩. 激活函数在卷积神经网络中的对比研究[J]. 计算机系统应用,2018,28(7):43-49.

[16] 娄　洋,李　丹. 基于 FER2013 数据集的人脸表情识别[J]. 计算机系统网络和电信,2020(4):22-24.

[17] 张晓峰,吴　刚. 基于生成对抗网络的数据增强方法[J]. 计算机系统应用,2019,28(10):201-206.