

基于MTCNN和Facenet的 Face Recognition Based on MTCNN and Facenet 人脸识别

刘长伟(中国联通智能城市研究院,北京 100048)

Liu Changwei(China Unicom Smart City Research Institute, Beijing 100048, China)

摘要:

人脸识别技术在智能城市建设中广泛应用,传统人脸识别算法依赖人工设定的特征,通常会带来不可期望的人为因素和误差。随着计算机算力的提升,基于神经网络的人脸识别方法由于其准确高效深受工业界偏爱。提出了基于多任务卷积神经网络(MTCNN——Multi-task Cascaded Convolutional Networks)和Facenet的人脸识别方法,并实现了从图像处理到识别结果输出的整个人脸识别系统。

Abstract:

Face recognition technology is widely used in the smart city construction. Traditional face recognition algorithms rely on artificial features, which usually bring unexpected human factors and errors. With the improvement of computer computing power, the face recognition method based on neural networks are favored by the industry because of its accuracy and efficiency. A face recognition method based on MTCNN (Multi-task Cascaded Convolutional Networks) and facenet is proposed, and the whole face recognition system from image processing to recognition result output is realized.

Keywords:

Face recognition; MTCNN; Facenet; System

关键词:

人脸识别; MTCNN; Facenet; 系统

doi: 10.12045/j.issn.1007-3043.2020.02.008

文章编号: 1007-3043(2020)02-0032-07

中图分类号: TP311

文献标识码: A

开放科学(资源服务)标识码(OSID):



引用格式: 刘长伟. 基于MTCNN和Facenet的人脸识别[J]. 邮电设计技术, 2020(2): 32-38.

0 引言

随着计算机及网络的发展,人脸识别技术在国家安全、公安、司法、电子商务、安保监控、安全检查等领域被越来越多地应用。传统人脸识别方法依赖于人工设计的特征(比如边和纹理描述量)与机器学习技术(比如主成分分析、线性判别分析或支持向量机)。人工设计的特征在无约束环境中的稳健性较差,这使得过去的研究者侧重研究针对不同变化类型的专用方法,比如能应对不同年龄的方法、能应对不同姿势的方法、能应对不同光照条件的方法等。近段时间,

传统的人脸识别方法已经被基于卷积神经网络(CNN)的深度学习方法代替。其主要优势是它可以用非常大型的数据集进行训练,从而学习到表征这些数据的最佳特征。本文提出了一种基于MTCNN和Facenet的人脸识别方法。

1 基础知识

本章对深度学习在图像领域的一些基础知识进行了简单描述。

1.1 计算机中图像的表达方式

在计算机中图像最简单的表达方式就是灰度图,如图1所示,在一张灰度图中,每个像素位置 (x, y) 对应一个灰度值 I ,图像在计算机中就存储为数值矩阵。

收稿日期: 2020-01-06

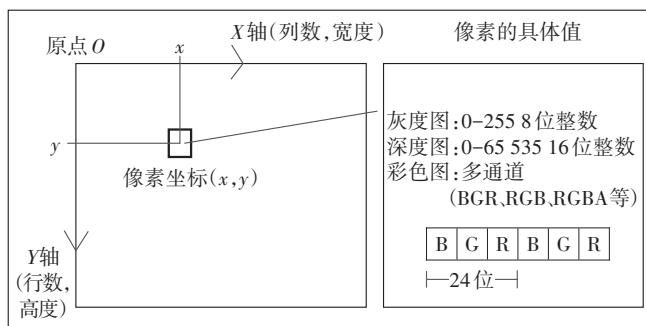


图1 图像在计算机中的表达方式示意图

一张宽度为640像素,高度为480像素分辨率的灰度图就可以表示为 480×640 的二维数组。数组的行数对应图像的高度,而列数对应图像的宽度,数组值对应着像素位置 (x, y) 的灰度值 I 。

彩色图,也就是RGB图,是通过红(R)、绿(G)、蓝(B)3个通道来表示一张图片,如图2所示,不同颜色可以通过红、绿、蓝3种颜色的组合表示,一张宽度为640像素,高度为480像素分辨率的RGB图就可以表示为 $480 \times 640 \times 3$ 的三维数组。数组的行数对应图像的高度,列数对应图像的宽度,而层数对应着红绿蓝不同颜色,数组值对应着像素位置 (x, y) 的值。

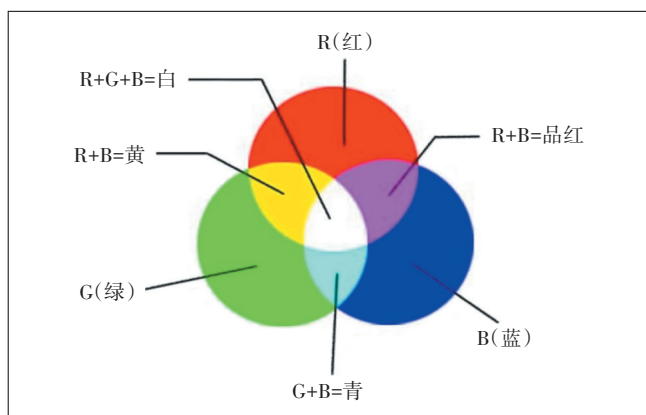


图2 RGB表示

1.2 图像卷积

图像卷积,有时也叫算子、滤波器。对于图像的每一个像素点,计算它的邻域像素和滤波器矩阵的对应元素的乘积,然后加起来,作为该像素位置的值。这样就完成了卷积过程。

如图3所示,在计算机中左边的矩阵代表图像,经过中间的 3×3 矩阵,也就是卷积核,与图像上的 3×3 矩阵对应元素的相乘,然后加和,最后得到目标值。在图像的行列上滑动卷积核进行相同操作便可得到整张图像经过卷积后的结果。

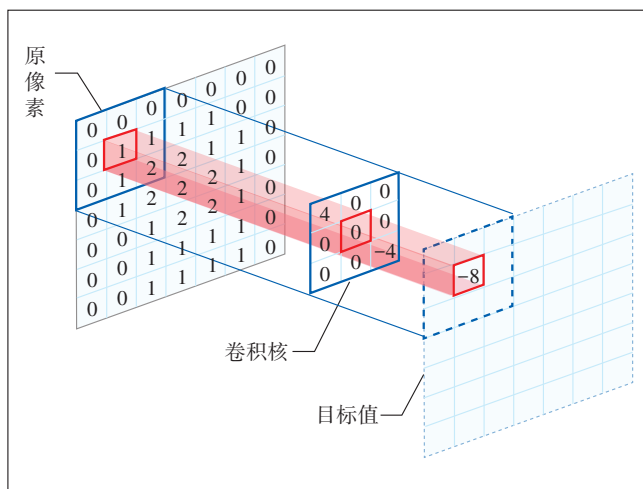


图3 图像卷积计算

图像处理中常用的卷积核有平均值滤波、高斯滤波、边缘检测、图像锐化等^[1]。

1.3 卷积神经网络

上节提到的高斯滤波等卷积核多为人为设定,而深度学习算法是通过学习得到卷积核。为了处理图像数据,就需要卷积神经网络(CNN——Convolutional Neural Network)。卷积神经网络实际上是层次模型(hierarchical model),原始输出通常经过卷积操作、池化操作和非线性激活函数映射等的层层组合。

CNN之父LeCun在1998年提出Lenet用于解决手写数字识别的视觉任务。自那时起,CNN的最基本的架构就定下来了:卷积层、池化层、全连接层^[2]。

池化是将输入数据进行缩小,减少像素信息,只保留重要信息,主要是为了减少计算量。主要包括最大池化和均值池化。

图4为最大池化,对 4×4 的矩阵进行 2×2 最大池化操作就是选出左边图中相同颜色的 2×2 矩阵的最大值,输出为右图所示矩阵。

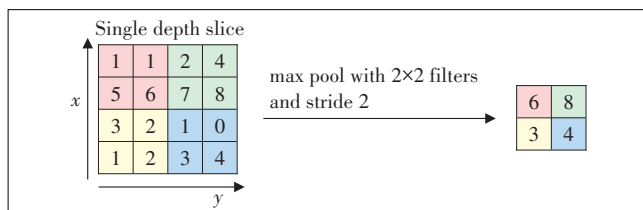


图4 最大池化

激活函数是用来给神经网络引入非线性。非线性变换是深度学习有效的原因之一。原因在于非线性相当于对空间进行变换,变换完成后相当于对问题空间进行简化,原来线性不可解的问题现在变得可解

了。常见的激活函数有 sigmoid、tanh 和 relu, 前两者常用在全连接层, relu 常见于卷积层^[3]。

其中 Sigmoid 函数表达式如下:

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

2 人脸识别

本文的人脸识别流程可分为图像数据采集→人脸检测→特征提取→信息对比4个步骤。

当新的人脸图片输入到系统中时,通过特征模型对图片进行特征提取后与人脸库中的特征数据对比,输出人脸识别结果,即可判别当前人脸是否在人脸库中,如果人脸库中存在的话,这个人匹配人脸库中的哪个人,这些都可以通过以上步骤实现。

其中进行人脸识别的摄像头为市面上常见的摄像头。下面分别对这4个步骤进行说明。

2.1 图像数据采集

图像数据采集包括获取摄像头传输图像、图像预处理等。获取摄像头传输图像经过相关接口处理即可。下面详细介绍一下图像预处理。

在现实环境下采集图像,由于图像受到光线明暗不同、脸部表情变化、阴影遮挡等外在因素的干扰,导致采集图像质量不理想,所以需要先对采集到的图像做预处理,如果图像预处理不好,会严重影响后续的人脸检测与识别。

预处理具体而言是指对系统采集到的人脸图像进行光线、旋转、切割、过滤、降噪、放大缩小等一系列的复杂处理来使得该人脸图像无论是从光线、角度、距离、大小等任何方面来看均能够符合人脸图像的特征提取的标准要求^[4]。

本文的图像预处理采用了 openCV 工具包中的方法,由于不是文章重点,不再过多赘述。

2.2 人脸检测

人脸检测的目标是找出图像中所有的人脸对应的位置,算法的输出是人脸外接矩形在图像中的坐标,可能还包括姿态如倾斜角度等信息。

虽然人脸的结构是确定的,由眉毛、眼睛、鼻子和嘴等部位组成,近似是一个刚体,但由于姿态和表情的变化,不同人的外观差异、光照、遮挡的影响,准确地检测处于各种条件下的人脸是一件相对困难的事情。

人脸检测要解决以下几个难点:

- a) 人脸可能出现在图像中的任何一个位置。
- b) 人脸可能有不同的大小。
- c) 人脸在图像中可能有不同的视角和姿态。
- d) 人脸可能部分被遮挡。

为解决自然场景下的人脸检测问题,发展出了各式各样的神经网络结构,例如 cascade-based 的 Cascade CNN^[5]、MTCNN、FaceCraft 等,联合人脸区域附近上下文信息的 HR、SSH、CMS-CNN 等,全新网络结构设计的 MSCNN、SRN、FANet^[6]等,基于通用目标检测模型,并结合进一步人脸优化策略的 Face R-CNN^[7]、Face-R-FCN、S3FD、ZCC 等,训练阶段采用全新数据增强策略的 DSFD、Pyramidbox、Improved SRN 等,结合特征图(feature map)注意力机制的 FAN^[8]、VIM-FD 等。

其中级联 CNN 系列是深度学习方法中速度最快的,CPU 都在 10 FPS 以上,级联 CNN 系列优化后可以轻松在 CPU 上实时检测人脸,全面优化后的 fast-MTCNN 可以在 ARM 上运行。

Faster R-CNN 系列的性能可以做到极高,但速度都很慢,甚至不能在 GPU 上实时检测人脸。

SSD/RPN 系列:有 SSH 和 SFD,都是目前 FDDB 和 WIDER FACE 上的最高水平,性能水平与 Faster R-CNN 系列不相上下,同时也可以保持 GPU 实时速度,SFD 的简化版 FaceBoxes 甚至可以 CPU 上实时检测人脸,极有潜力在 ARM 上运行。

文章采用的人脸及 Landmark 检测是多任务级联卷积神经网络(MTCNN)。下面章节会详细讲解。

2.3 特征提取

图片的特征提取是深度学习中最为基础也是最为重要的研究内容,特征提取可以简单地看作一个数据降维过程,其过程是将图片数据转化为具有一定区分度的一定维度的向量的图片特征的过程。

传统特征提取算法大体可分为以下几类:

- a) 基于颜色分类:颜色直方图、颜色集、颜色矩、颜色聚合向量等。
- b) 基于纹理特征:如 Tamura 纹理特征、自回归纹理模型、Gabor 变换、小波变换、MPEG7 边缘直方图等。
- c) 基于形状特征:如傅里叶形状描述符、不变矩、小波轮廓描述等。

基于神经网络的特征提取算法有 Inception、VGG、ResNet、Inception-ResNet^[9]等。特征提取的神经网络模型基本上以在 imagenet 的表现作为参考进行选择,也

可以自己设计神经网络架构,目前这些CNN网络模型在imagenet上的识别准确率及参数量如表1所示。

表1 模型效果

模型	Top-1 准确率	Top-5 准确率	参数
VGG16	0.713	0.901	138 357 544
VGG19	0.713	0.900	143 667 240
ResNet50	0.749	0.921	35 636 712
ResNet101	0.764	0.928	44 707 176
ResNet152	0.776	0.931	60 419 944
InceptionV3	0.779	0.937	23 851 784
Xception	0.790	0.945	22 910 480
InceptionResNetV2	0.803	0.953	55 873 736
MobileNet	0.704	0.895	4 253 864
MobileNetV2	0.713	0.901	3 538 984
DenseNetV121	0.750	0.923	8 062 504
DenseNetV169	0.762	0.932	14 307 880
DenseNetV201	0.773	0.936	20 242 984

Top-1 准确率就是预测的label取最后概率向量里面最大的那一个作为预测结果,如果预测结果中概率最大的那个分类正确,则预测正确,否则预测错误。Top-5 准确率就是取最后概率向量较大的前5个值作为预测结果,只要其中出现了正确概率即为预测正确,否则预测错误。

本篇文章通过Facenet对人脸图片进行特征提取,下面一章详细说明。

2.4 信息对比

完成特征提取后就是计算特征的相似度,特征就是结构化数据,不管是一维、二维还是 N 维,只需要计算出待识别图像的特征跟库中特征的相似度,就可以据此判断是否为同一个人,实现人脸识别。计算特征的相似度就是计算2组数据的相似度,方法有余弦相似度算法,具体如式(2)。

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (2)$$

余弦相似度是最常用的,最初用在计算文本相似度效果很好,由于余弦相似度表示方向上的差异,对距离不敏感,所以有时候考虑到距离上的差异,会先将每个待处理值减去均值,这样称为调整余弦相似度^[10]。

欧氏距离如式(3)。

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \\ = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (3)$$

欧氏距离更多考虑的是空间中两点之间的距离,而余弦相似度考虑的是空间夹角。欧氏距离能体现个体数值特征的绝对差异,更适用于需要从维度的数值大小中体现差异的分析。

当然除了以上2个常用的相似度算法之外,还有曼哈顿距离、明可夫斯基距离、皮尔森相关系数、斯皮尔曼等级相关系数等算法。

2.5 技术指标

2.5.1 人脸检测技术指标

目前人脸检测技术指标主要有以下4个:

a) 检测率:识别正确的人脸/图中所有的人脸。检测率越高,代表检测模型效果越好。

b) 误检率:识别错误的人脸/识别出来的人脸。误检率越低,代表检测模型效果越好。

c) 漏检率:未识别出来的人脸/图中所有的人脸。漏检率越低,代表检测模型效果越好。

d) 速度:从采集图像完成到人脸检测完成的时间。时间越短,检测模型效果越好。

例如,在摄像头某张抓拍图像中,一共有100张人脸,算法检测出80张人脸,其中75张是真实人脸,5张是把路标误识为人脸。在这个实际案例中:检测率=75/100,误检率=5/80,漏检率=(100-75)/100。

2.5.2 人脸识别技术指标

人脸识别则有以下技术指标:

a) 误识率(FAR——False Accept Rate):比较不同人的图像时,把其中的图像对当成同一个人图像的比例。误识率越小越好。

b) 拒识率(FRR——False Reject Rate):比较同一个人的图像时,把其中的图像对当成不同人图像的比例。误识率越小越好。

c) 速度:人脸识别耗费的时间。时间越短,模型效果越好。

2.6 应用领域

图5为人脸识别的产业链,目前人脸识别主要应用于如下行业:

a) 医疗行业。例如打击涉医犯罪,确保就诊安全,建立涉医犯罪人员布控库,与属地公安部门配合,进行实时布控。

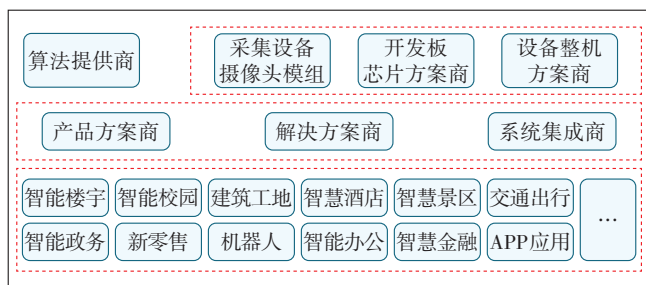


图5 人脸识别产业链

b) 新零售。例如减少突发事件的产生,门店遇到商品失窃的突发事件,通过数据分析,可以将不良客户拉入“黑名单”或是降低其信用水平。

c) 城市安防。例如重要场所的布控,对机场、车站、港口、地铁重点场所和大型商超等人群密集公共场所进行布控,以达到排查重点人员、抓捕逃犯等目的。

d) 公安机构。例如寻人寻亲,将老百姓或其他业务部门提供的照片直接送入系统进行比对、检索、筛选,最后人工确认。

以上示例并不能包含所有应用人脸识别的领域。

3 MTCNN与Facenet算法

本项目中人脸检测采用MTCNN算法,其速度较快,在CPU上可达到35 FPS,适用于实时人脸检测,特征提取算法采用Facenet算法,其准确率可达到99%以上,可靠性强。本章将详细介绍这2个算法的结构以及其实验效果。

3.1 基于MTCNN的人脸检测

3.1.1 算法详解

本项目采用了MTCNN算法^[11],该算法是一种从粗到细的处理方法,同时利用人脸检测任务和人脸分类任务来辅助人脸关键点检测。其总体框架分为3个阶段。3个阶段的输入为不同大小的图片,这是为了检测不同大小的人脸,如图6所示。

第1阶段,Proposal Network (P-Net):这个网络主要用于生成候选框(candidate window),作为第2阶段的输入。在不同size的图片上,使用一个FCN(Fully Convolutional Networks)来获取 candidate windows 的 bbox(边框)的坐标,同时使用 bbox 回归来修正这些 candidate,之后使用非极大值抑制(NMS)来聚合高度重叠的 candidate window。

因为P-Net使用了FCN,所以无需指定输入图片大小(图7的P-Net的Input size中的12×12×3为一个示

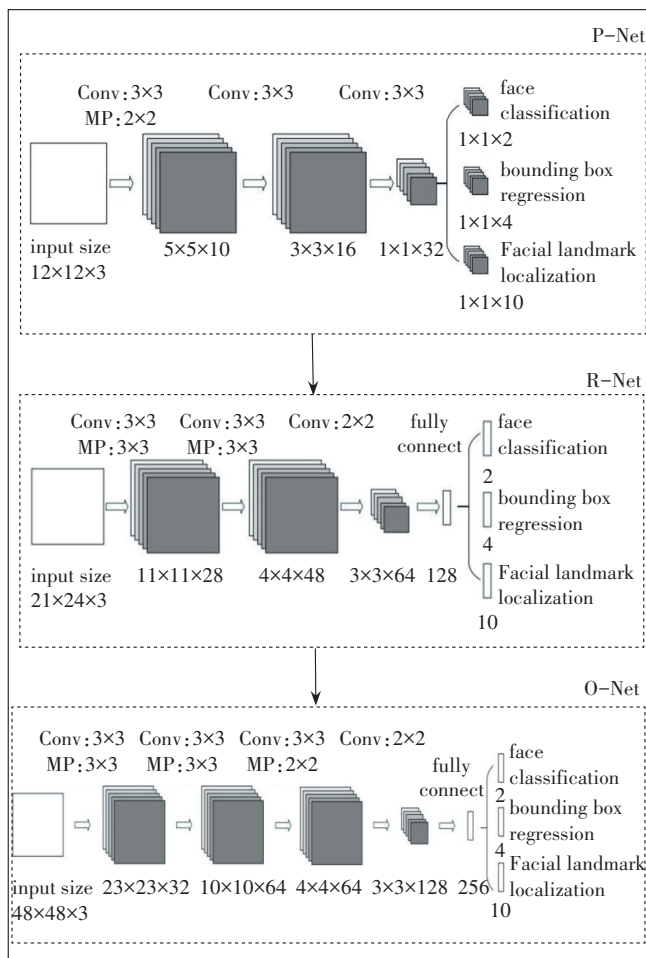


图6 MTCNN模型结构

例,可看成 anchor box)。它返回的是2个参数,1个人脸框的坐标,1个是是否为人脸的概率。

第2阶段,Refine Network (R-Net):将第1阶段生成的 candidates 作为另一个CNN的输入。它会进一步剔除错误的 candidates,对 bbox 进行校正,使用NMS合并 candidates。

R-Net接受的是P-Net输出的 bbox,其输入是固定的,因此需要对 bbox 进行 resize,中间的过程和R-Net类似,它返回的是2个参数,1个人脸框的坐标,1个是是否为人脸的概率。

第3阶段,Output Network (O-Net):这个网络接受的是R-Net返回的 bbox,也要 resize 到固定大小,这一阶段返回3个参数,第1个参数是人脸框,第2个参数是是否人脸的概率,第3个参数是5个关键点坐标。

在CNN的设计上,MTCNN把5×5的filter改为了3×3来减少计算量。

3.1.2 模型效果验证

图7为MTCNN的模型效果,左边的图为不同算法在Fddb数据集上的准确率,可见MTCNN优于face-ness、CCF等算法;右边的图为MTCNN关于眼口鼻等

关键点位置检测的平均错误率,MTCNN也低于CDM、ESR等人脸关键点检测算法。

算法在Fddb数据集上召回率达到93%,速度在

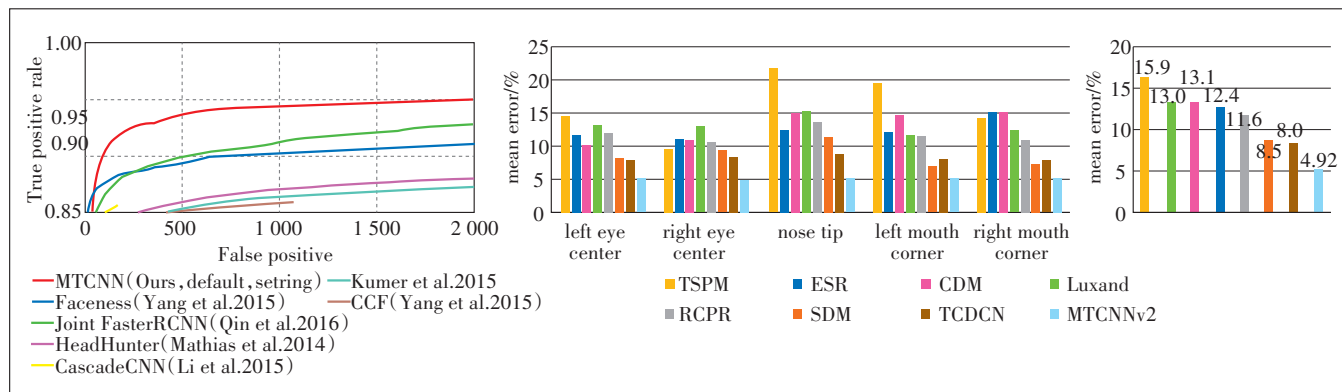


图7 MTCNN模型效果

CPU上约为35 FPS。

3.2 基于Facenet的特征提取

3.2.1 算法详解

本项目采用Facenet方法^[12]对图像进行特征提取,主体模型采用一个极深度网络InceptionResNetV2,InceptionResNetV2由3个带有残差连接的Inception模块和1个Inception V4^[13]模块组成。典型的Inception模块结构如图8所示。

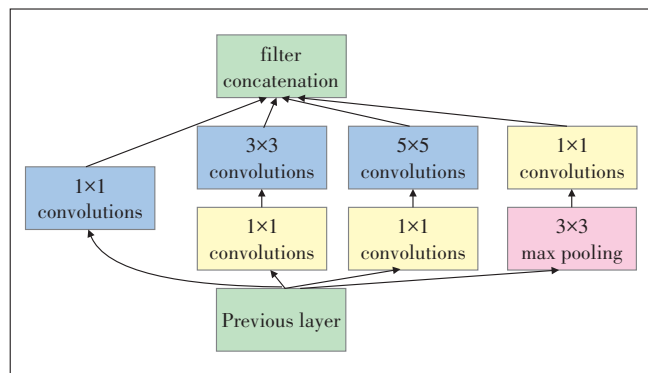


图8 Inception块

图8的模块由3组卷积核以及1个池化单元组成,它们共同接受来自前一层的输入图像,3种尺寸的卷积核和1个max pooling并行地对输入图像进行处理,然后将输出结果按照通道拼接起来。因为卷积操作接受的输入图像大小相等,而且卷积进行了padding操作,因此输出图像的大小也相同,可以直接按照通道进行拼接^[14]。

从理论上讲,Inception模块的目标是用尺寸更小的矩阵来替代大尺寸的稀疏矩阵,即用一系列小的卷

积核来替代大的卷积核,保证二者有近似的性能。

图9的卷积操作中,如果输入图像的通道数太多,则运算量太大,而且卷积核的参数太多,因此有必要进行数据降维。所有的卷积和池化操作都使用了1x1卷积进行降维,即降低图像的通道数。因为1x1卷积不会改变图像的高度和宽度,只会改变通道数。

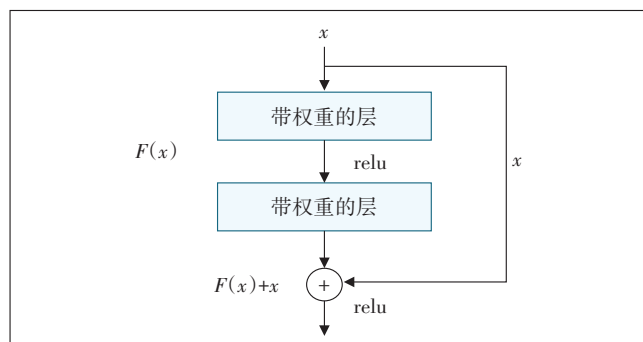


图9 残差块原理图

残差网络(Resnet——Residual Network)通过跳层连接和拟合残差来解决层次过多带来的问题^[15],这一结构的原理如图9所示。

残差网络并不是一个单一的超深网络,而是多个网络指数级的隐式集成,由此引入了多样性的概念,残差网络用来描述隐式集成的网络的数量;在预测时,残差网络的行为类似于集成学习。

3.2.2 损失函数

损失函数用来评价模型的预测值和真实值不一样的程度,损失函数越好,通常模型的性能越好。不同的模型用的损失函数一般也不一样。

常见的损失函数,包括MSE均方误差损失函数、SVM合页损失函数、Cross Entropy交叉熵损失函数、目标检测中常用的Smooth L1损失函数。

Facenet采用Triplet Loss损失函数^[16]。Triplet Loss是深度学习中的一种损失函数,用于训练差异性较小的样本,如人脸。训练数据包括锚(Anchor)示例、正(Positive)示例、负(Negative)示例,通过优化使锚示例与正示例的距离小于锚示例与负示例的距离,实现样本的相似性计算。

Triplet Loss的核心是锚示例、正示例、负示例共享模型,通过该模型,将锚示例与正示例聚类,远离负示例。

3.2.3 模型效果验证

表2为Facenet在不同数据集上取得的效果,所采用的特征提取框架为Inception ResNet v1,模型可在LFW数据集上达到99%以上的准确率。

表2 Facenet模型效果

LFW 准确率	训练集	模型
0.990 5	CASIA-WebFace	Inception ResNet v1
0.996 5	VGGFace2	Inception ResNet v1

综上所述,MTCNN人脸检测模块结合Facenet人脸识别模块可以满足日常生活中关于人脸识别的需求。

4 结束语

本文基于MTCNN和Facenet实现人脸识别。该方法通过MTCNN进行人脸检测,在Fddb数据集上召回率达到93%,速度在CPU上约为35 FPS。Facenet进行人脸识别时在LFW数据集上可达到99%以上的准确率。

现阶段人脸识别不单指身份识别,而是基于人脸特征数据进行相关操作和研究行为的宽泛概念,包括身份认证、美颜自拍、人脸关键点检测等。随着科技的发展,未来的人脸识别会越来越准确,应用场景会越来越广泛,关于人脸识别的相关规定也会越来越多。

参考文献:

- [1] 刘祝华. 图像去噪方法的研究[D]. 南昌:江西师范大学,2005.
- [2] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

- [3] LESHNO M, LIN V Y, PINKUS A, et al. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function[J]. Neural networks, 1993, 6(6):861-867.
- [4] 王虹,董雅丽. 用于人脸识别的图像预处理方法研究与实现[J]. 武汉理工大学学报(交通科学与工程版),2007(5):876-878.
- [5] QIN H, YAN J, LI X, et al. Joint training of cascaded CNN for face detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:3456-3465.
- [6] BEKMEZCI I, SAHINGOZ O K, TEMEL S. Flying ad-hoc networks (FANETs): A survey [J]. Ad Hoc Networks, 2013, 11(3):1254-1270.
- [7] JIANG H, LEARNED-MILLER E. Face detection with the faster R-CNN [C]// 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017:650-657.
- [8] CHENG Z, BAI F, XU Y, et al. Focusing attention: Towards accurate text recognition in natural images[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:5076-5084.
- [9] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning [C]// 31st AAAI Conference on Artificial Intelligence. 2017.
- [10] STRICKER M A, ORENGO M. Similarity of color images [C]// Storage and retrieval for image and video databases III. International Society for Optics and Photonics, 1995:381-392.
- [11] DAI J, HE K, SUN J. Instance-aware semantic segmentation via multi-task network cascades [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:3150-3158.
- [12] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:815-823.
- [13] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning [C]// Thirty-First AAAI Conference on Artificial Intelligence. 2017.
- [14] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:2818-2826.
- [15] TAI Y, YANG J, LIU X. Image super-resolution via deep recursive residual network [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017:3147-3155.
- [16] 何琼华. 基于Triplet-aware Center Loss的人脸识别算法研究与实现[D]. 成都:西南交通大学,2018.

作者简介:

刘长伟,工程师,硕士,主要从事智能城市建设中的算法研发工作。

