Journal of Computer Applications
计算机应用,2020,40(4):1002 - 1008
ISSN 1001-9081
CODEN JYIIDU
2020-04-10
http://www.joca.cn

文章编号:1001-9081(2020)04-1002-07
DOI:10.11772/j.issn.1001-9081.2019091535

面向移动平台人脸检测的FaceYoLo算法

任海培,李　腾*
(安徽大学 电气工程与自动化学院,合肥 230601)
(*通信作者电子邮箱liteng@ahu.edu.cn)

摘　要:针对移动平台上人脸检测实时性不强的问题,提出了一种基于深度学习的FaceYoLo实时人脸检测算法。首先,在YoLov3检测算法的基础上,加入快速消化卷积层(RDCL)缩小输入空间,然后加入多尺度卷积层(MSCL)丰富不同检测尺度的感受野,最后加入中心损失和致密化策略加强模型的泛化能力和鲁棒性。实验结果表明,在GPU上测试时,该算法较YoLov3算法在速度上提高至原来的8倍,每幅图像的处理速度可达0.002 8 s;精度提高了2.1个百分点;在Android平台上测试时,该算法较最好的MobileNet模型在检测速率上从5 frame/s提升到10 frame/s。通过实验结果可知,该算法能有效提高人脸检测在移动平台上的实时性能。

关键词:卷积神经网络;人脸检测;深度学习;移动平台;Android
中图分类号:TP391.41　　**文献标志码**:A

FaceYoLo algorithm for face detection on mobile platform

REN Haipei, LI Teng*
(*School of Electrical Engineering and Automation*, *Anhui University*, *Hefei Anhui* 230601, *China*)

Abstract: Concerning the problem of low real-time performance of face detection on mobile platform, a FaceYoLo real-time face detection algorithm based on deep learning was proposed. Firstly, based on the YoLov3 detection algorithm, the Rapidly Digested Convolutional Layers (RDCL) were added to reduce the input space size, then Multiple Scale Convolutional Layers (MSCL) were added to enrich the receptive fields of different detection scales, and finally the central loss and densification strategy were added to strengthen the generalization ability and robustness of the model. The experimental results show that, when tested on the GPU, the proposed algorithm improves the speed by nearly 8 times compared with the YoLov3 algorithm, has the processing time of each image reached 0.002 8 s, and increases the accuracy by 2.1 percentage points; when tested on the Android platform, the proposed algorithm has the detection rate increased from 5 frame/s to 10 frame/s compared with the best MobileNet model, demonstrating that the algorithm can effectively improve the real-time performance of face detection on mobile platform.

Key words: Convolutional Neural Network (CNN); face detection; deep learning; mobile platform; Android

0　引言

面向移动平台的人脸检测任务是人脸技术应用过程中的重要环节。移动平台检测任务的质量不仅和检测算法的性能有关,还受到移动嵌入式设备硬件性能的限制。然而目前影响深度学习算法在嵌入式平台中应用的主要原因还是实时性达不到要求。影响因素主要和输入图片大小、选用的网络模型以及训练模型的数据量等有关。较大的搜索空间和大小不一的对象尺度进一步增加了模型的搜索时间,图1是一个典型的搜索空间比较大、人脸数量较多的对象场景。为了提高人脸检测在嵌入式移动设备端的检测速度,保证应用软件(APPlication, APP)运行流畅,本文主要从算法的模型优化入手。

人脸检测是一种特殊的目标检测案例。现代人脸检测方法大致可分为两类:一类是传统的检测算法,另一类是基于卷积神经网络(Convolutional Neural Network, CNN)的算法。

最早的目标检测用的是级联分类器框架,由Viola等[1]在2001年电气电子工程师学会举办的计算机视觉和模式识别领域的顶级会议(IEEE conference on Computer Vision and Pattern Recognition, CVPR)中提出来,该算法第一次使目标检测成为现实。接着就是Haar检测,但是这种检测算法只适合刚性物体检测,无法检测行人等非刚性目标,所以又提出了方向梯度直方图(Histogram of Oriented Gradients, HOG)组合支持向量机(Support Vector Machine, SVM)结构。接着传统算法就一直围绕着特征器作改进。而针对人脸方面的检测则由开创性的Viola-Jones人脸检测器提出将Haar特征[2]、AdaBoost学习[3]和级联推理相结合用于人脸检测,因此后续的许多工作都被提出用于实时人脸检测,如新的局部特征、新的增强算法和新的级联结构。

CNN检测算法主要包含两大家族,分别是区域卷积神经

收稿日期:2019-09-05;**修回日期**:2019-10-12;**录用日期**:2019-10-15。
基金项目:国家重点研发计划项目(2018YFB1305804);安徽省杰出青年基金资助项目(1908085J25)。
作者简介:任海培(1994—),男,安徽芜湖人,硕士研究生,主要研究方向:模式识别、深度学习;　李腾(1980—),男,安徽淮南人,教授,博士,CCF会员,主要研究方向:模式识别、深度学习。

网络(Regions with CNN features, RCNN)[4]家族和YoLo家族。RCNN家族的核心贡献就是构造了区域建议网络(Region Proposal Network, RPN)结构,瞬间提高了检测精度和特征提取效率。尤其是Faster RCNN的出现无疑是检测发展史上的第一次高潮,该算法在各大数据集上测试都展现了极其优越的性能;但是Faster RCNN算法缺点也很明显,即检测速度不够快。YoLo家族包括YoLov1、YoLov2、YoLov3,以及单发多盒探测器(Single Shot multibox Detector, SSD)。该家族共性都是产生Proposal[5-6]的同时进行Classification加Regression[7],一次性完成,即所谓的One-shot[8]。

YoLov1使用了1×1卷积层[9]和3×3的卷积层替代了Inception结构,且在最后使用了全连接层进行类别输出。这样做缺点也很明显:1)输入尺寸固定。由于输出层为全连接层,因此训练模型只支持与训练图像相同的输入分辨率。2)小目标检测效果不好。虽然每个格子可以预测B个Bounding box,但是最终只选择重叠度(Intersection Over Union, IOU)得分最高的Bounding box作为物体检测输出[10]。YoLov2则引入了Faster RCNN中Anchor box的思想[11],对网络结构的设计进行了改进,输出层使用卷积层替代YoLov1的全连接层。YoLov2中借鉴了残差网络(Residual Network, ResNet)思想[12-13],在网络中设计了跨层跳跃连接[14],解决了输入图像分辨率不统一的问题,加深了对浅层特征的学习;但是缺点也依旧存在,某些小尺度对象的特征学习问题还是得不到很好的解决。而YoLov3算法则注重解决多尺度检测问题,设计了Darknet并向里面加入了多尺度预测结构[15];同时YoLov3没有使用Softmax对每个框进行分类,而是采用了独立的多个Logistic分类器替代[16];但是该算法在模型参数大小以及速度上还是存在一定的适用局限性,模型参数较大、速度不够快仍是它的问题。

相对于RCNN家族的两个阶段,YoLo家族速度占优,准确率和召回率较低。而深度学习人脸检测追溯可从中国计算机学会(China Computer Federation, CCF)在CNN特征的基础上使用增强来进行人脸检测,至Farfade等[17]将微调CNN模型训练在1k ImageNet分类任务上进行人脸和非人脸分类任务和Qin等提出联合训练级联神经网络实现端到端优化[18],最终形成成熟的检测体系。

而移动端的人脸检测技术发展相对较晚,主要原因与移动平台的开发和集成技术有关。伴随着近几年来集成技术以及人工智能领域的发展,移动端的人脸检测技术逐渐成为了未来人工智能技术落地的应用趋势。人脸检测技术主要结合人脸识别技术应用在金融、安防等各个领域,其中应用最广泛的是利用深度学习方法中的卷积神经网络来训练优异的检测模型。如目前较好的由中国科学院深圳先进技术研究院在2016年提出来的多任务卷积神经网络(Multi- Task Convolutional Neural Network, MTCNN)人脸检测任务的多任务神经网络模型,该模型采用三个级联的网络以及候选框加分类器的思想,进行快速高效的人脸检测,将训练好的模型移植到Android平台,但是这种方法对硬件性能的要求比较高;接着就是2017年谷歌提出的MobileNet以及2018年提出的MobileNet V2更是移动端人脸检测技术发展的高潮,它是一种专门针对硬件性能要求不高的仅支持CPU平台的人脸检测技术。该技术特点就是利用拆分卷积的思想使得训练参量极大地降低,使得算法模型在性能较差的CPU上也能有较高的检测速率。

但这些检测算法仍然在速度上存在局限性,而本文研究的目的就是要保证检测效果足够好的前提下继续优化检测算法,提高复杂场景下人脸检测的实时性能。本文在YoLov3的框架基础上进行了结构创新与拓展,主要包括网络结构优化与人脸聚类方法的改进。

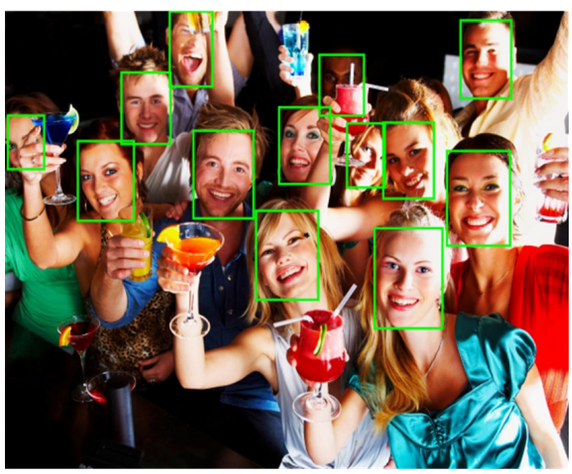


图1　人脸检测示意图

Fig. 1　Schematic diagram for face detection

1　相关工作

传统的人脸检测算法中特征提取使用Haar-like特征进行检测,然后利用积分图对Haar-like特征求值进行加速,通过AdaBoost算法训练区分人脸和非人脸的强分类器,使用筛选式级联把强分类器级联到一起,以提高准确率。Haar分类器最早来源于Viola和Jones发表的经典论文,在AdaBoost算法的基础上,使用Haar-like小波特征和积分图方法进行人脸检测,他们不是最早提出使用小波特征的,但是他们设计了针对人脸检测更有效的特征,并对AdaBoost训练出的强分类器进行级联,形成了Viola-Jones检测器。同一时期, Lienhart等[19]对这个检测器进行了扩展,最终形成了开源计算机视觉库(Open Source Computer Vision Library, OpenCV)和现在的Haar分类器。

现代的CNN检测算法中,YoLov3算法是最受欢迎的方法之一。YoLov3设计了特定的53层暗网络(Darknet with 53 convolutional neural network, Darknet53),通过设计多尺度网络和分类损失提高了检测性能。然而YoLov3算法在大量人脸数据集的工程应用下,检测速度还是达不到要求。Darknet53网络自带的多尺度策略是根据三个不同输出特征层上的输出特征再融合来加深对小目标的特征学习,通过研究表明这样做忽视了不同特征层输出Anchor密度不均匀带来的人脸召回率低下等原因,导致YoLov3算法对小目标检测的鲁棒性还不够高。为了进一步提高算法的工程应用水平,本文设计了特定的基于快速消化卷积和多尺度卷积的暗网络结构(darknet network structure based on Rapidly digested convolutional layers and Multiple scale convolutional layers, RM-darknet),增加了快速消化卷积层(Rapidly Digested

Convolutional Layers，RDCL)和多尺度卷积层(Multiple Scale Convolutional Layers，MSCL)两个卷积结构以及 Anchor 致密化和中心损失(Center Loss)两个策略以提高 YoLov3 算法的检测速度。其中 RDCL 结构通过设计合适的内核大小，快速缩小了输入空间。RDCL 结构中新设计的串联整流线性单元(Concatenated Rectified Linear Unit，CReLU)激活函数减少了输出通道数，提高了算法的检测速度，同时也简化了本研究的算法模型。MSCL 结构通过离散 Anchor 到多个不同分辨率的层上处理不同尺寸的人脸，再结合致密化策略的协调配合，巧妙地增加了算法对小目标检测的鲁棒性。

同时此次研究还考虑到特征学习中的类内紧性对特征学习的作用。有关研究结果表明，算法中加入新的中心损失函数不仅能提高类内距离，而且还有助于提高类间距离，达到类似 Softmax 函数一样的作用。因此研究的算法在原始的 Softmax 损失后又加了中心损失函数，通过本项目的实验结果也表明确实提高了人脸特征的识别能力。为了进一步加快梯度收敛、防止过拟合问题，本次项目中巧妙地结合批量正则化(Batch Normalization，BN)结构解决了上述问题；并且为了进一步增强本次研究工作的说服力，与比较流行的移动端网络(Mobile neural Network，MobileNet)算法进行了比较，分别对两者训练的模型进行了移动端的测试，结果也很好地验证了本文算法的有效性。整个实验结果测试采用的是常用的人脸检测评价标准，对人脸检测算法来说，评价的一个重要标准是模型是否具有较高的正确检测率和较小的漏检率。

正确检测率(Correct Detection Rate，CDR)的计算公式如下：

$$r_{\mathrm{CDR}}=\frac{n_{\mathrm{c}}}{N}\times 100\%;\ n_{\mathrm{c}}\leqslant N \tag{1}$$

漏检率(Missed Detection Rate，MDR)的计算公式如下：

$$r_{\mathrm{MDR}}=\frac{n_{\mathrm{f}}}{N}\times 100\%;\ n_{\mathrm{f}}\leqslant N \tag{2}$$

其中：n_{c}和n_{f}分别为正确检测到的人脸数量和漏检人脸数量；N为待检测人脸数量。

本文工作主要分为以下几个方面：

1)在设计的网络中加入了快速消化卷积层，使人脸检测在速度上实现了飞跃以及模型参数的减少。

2)传统的 Yolov3 在小目标的识别性能上差强人意。为了应对复杂条件下的人脸识别挑战，本文方案在检测结构中引入多尺度卷积操作(MSCL)和 Anchor 致密化操作。为保证小尺度的人脸检测精度，通过丰富接收域和分层离散锚点来处理不同尺度的检测对象。

3)与传统的以 Softmax 损失为主的 YoLov3 检测训练优化方案不同，对于人脸检测任务，本文同样引入度量学习的策略，即中心损失，鼓励网络学习过程在扩大类间变化的同时最小化类内变化的判别特征，进一步提高模型对更具区分性特征的理解，提高检测性能。

4)在 Wider Face 人脸数据集[20]上进行训练，在 FDDB 人脸基准数据集上进行了测试，验证算法的效果。

5)利用新算法训练的模型在移动端进行应用测试。

2 算法设计分析

本文在 YoLov3 算法的基础上加以改进，设计了 RM-darknet 网络用来提高人脸检测的实时性能，如图 2 所示。YoLov3 中的 Darknet53 网络由一系列 1×1 和 3×3 卷积层组成，每个卷积层后都有一个 BN 层和一个带泄露修正线性单元(Leaky Rectified Linear Unit，Leaky ReLU)层[21]。Darknet53 网络一共有 53 层，包括 52 个卷积层和 1 个全连接层(除 Residual 层外)，将最后的 Softmax 层更换成 Logistic 分类器，其中回归损失用的是平滑损失(Smooth L1 Loss Layer，SmoothL1)函数。为了进一步提高人脸检测算法的实时性，加快检测速度，本文在网络结构中加入 RDCL 结构，实现快速缩小输入空间尺寸、减少输出通道数目的功能。然后在 RDCL 层后添加 MSCL 结构以丰富不同检测尺度的感受野，同时提高网络对小目标多个尺度上的检测灵敏性。通过在 MSCL 层引入致密化策略用来保证不同离散尺度 Anchor 密度的一致性。最后为了加强特征的类内紧密性，提高人脸的检测精度，本项目在网络最后的 Softmax 后加入了中心损失，对于加入的结构在后文中做了灼烧实验以论证理论的可实施性。

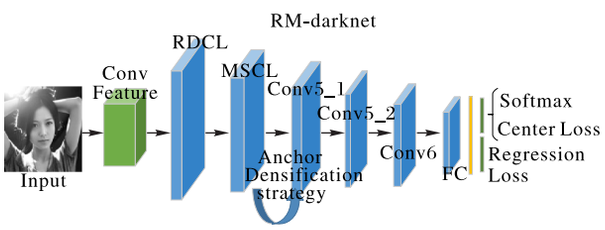


图 2 RM-darknet 网络结构

Fig. 2 RM-darknet network structure

下面依次介绍设计算法的几个主要模块：快速消化卷积层、多尺度卷积层、Anchor 致密化和 Center Loss 策略。

2.1 快速消化卷积

在复杂网络的实际应用中，很少会使用具有较大卷积核尺寸的滤波器，例如 5×5、7×7，研究[22]认为用多层 3×3 就能达到较大内核尺寸滤波器同等大小的感受野，例如两层 3×3 可以达到一层 5×5 的感受野，3 层 3×3 可以达到一层 7×7 的感受野。用多个小卷积核尺寸的滤波器对数据进行处理，不仅可以有效减少计算量，而且可以加深网络深度和层次，优化网络整体的非线性性能。但是在高效网络里，较小的卷积核尺寸在有些情况下并不一定具有更好的性能。如在图像分辨率足够的情况，较大的卷积核对数据的表征能力更强，压缩空间的速度更快。大多数人脸检测需要考虑时间成本，尤其偏向于大场景下的人脸检测应用对速度的要求更为苛刻，检测器的速度直接影响产品能否成功落地。

如图 3 所示，RDCL 结构的核心是通过快速缩小空间尺寸，减小特征图(Feature Map)的大小，达到快速提高检测速度的目的。为此对两个卷积层和两个池化层的步长分别设置为 4、2、2 和 2，这样一个总步长为 32 的输入尺度成功将输入空间大小压缩至原输入的 1/32。

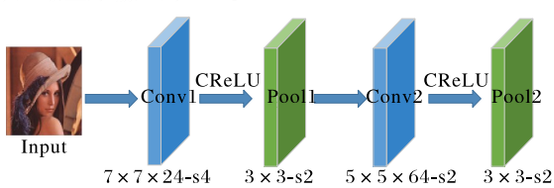


图 3 RDCL 结构

Fig. 3 RDCL structure

另外为了保持网络的高效和有效性,选取 7 × 7、5 × 5 作为 Conv1 和 Conv2 的卷积核大小,3 × 3 作为所有池化层的卷积核大小。通过 RDCL 第一个卷积层,核为 7 × 7,步长为 4,默认 padding 为 3,输出特征大小计算为 256 × 256;通过第一个池化层,核为 3 × 3,步长为 2,默认 padding 为 1,输出特征计算大小为 128 × 128;通过 RDCL 第二个卷积层,核为 5 × 5,步长为 2,默认 padding 为 2,输出特征大小计算为 64 × 64;通过第二个池化层,核为 3 × 3,步长为 2,默认 padding 为 1,输出特征计算大小为 32 × 32。这个设计借鉴了 MobileNet 网络设计中轻量化的思想[23],在网络结构中摒弃了 Pooling Layer 而直接采用 Stride 为 2 进行卷积运算。

另外,此次研究选择 CReLU 激活函数[24]来减少输出通道的数量。研究表明 CNN 较低层中的滤波器会形成匹配对(滤波器具有相反的效应),网络的底层接收冗余滤波器来提取输入的正负相位信息的可能性,因此可以考虑采用适当的操作移除这些冗余滤波器。文献[25]中提出了 CReLU 结构,如图 4 所示,将激活函数的输入额外做一次取反,等价于将输入相位旋转 180°。这种策略可以看作在网络中加入相位的先验知识。这样 CReLU 可以通过在 ReLU 之前简单地 Concat 否定的输出来使输出通道的数量加倍,再经过后面的池化层达到快速缩小输入空间的目的。CReLU 的使用显著提高了速度,而且精度基本不会下降,在本文的实验部分进行了验证。

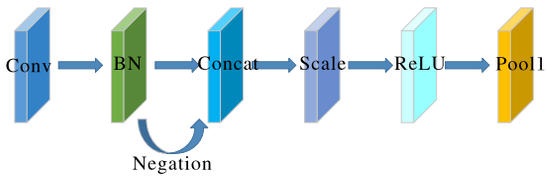


图 4　CReLU 内部结构

Fig. 4　Internal structure of CReLU

2.2　多尺度卷积

MSCL 的设计源自对 RPN[24]的理解,RPN 是一种多类别目标检测场景中不依赖于类的卷积运算,如图 5 所示,RDCL 结构通过 BN 层以及不同卷积核的卷积分支组成,并通过 ReLU 线性激活单元避免不同输入尺寸人脸特征的梯度消失问题。

其次,Anchor 相关层负责检测一定尺度范围内的人脸,但它只有一个单一的接受域,无法匹配不同尺度的人脸。针对以上问题,本文设计的 MSCL 结构中将 Anchor 离散到多个不同分辨率的层上,自然处理不同尺寸的人脸。并且通过 Inception Modules 增加网络宽度[26],实现不同 Anchor 层的输出特征对应于不同大小的接受域,获取不同尺度的人脸。表 1 是 MSCL 结构中部分网络层的数值。

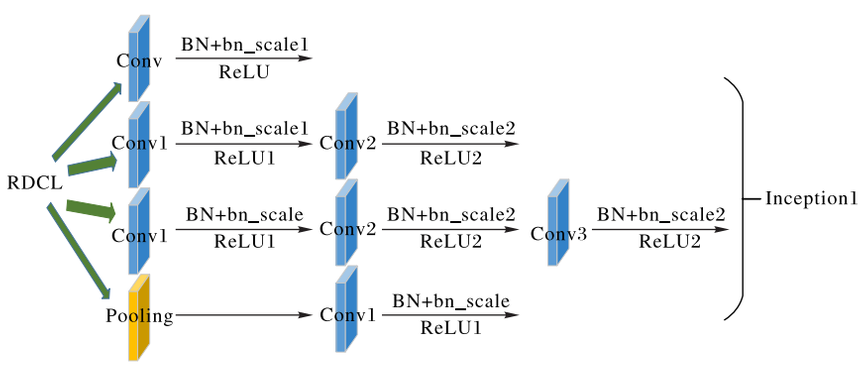


图 5　MSCL 中的 Inception 结构

Fig. 5　Inception structure in MSCL

表 1　MSCL 结构相关层参数

Tab. 1　Parameters of the relevant layers in MSCL structure

卷积层	默认尺寸	接受域
Inception3	32 × 32, 64 × 64, 128 × 128	143 × 143, 207 × 207, 271 × 271, 335 × 335, 527 × 527
Conv3_2	256 × 256	271 × 271, 335 × 335, 527 × 527, 591 × 591, 655 × 655
Conv4_2	512 × 512	527 × 527, 591 × 591, 655 × 655, 783 × 783, 911 × 911

2.3　训练策略

2.3.1　Anchor 致密化策略

Anchor 致密化策略是为了解决不同尺度 Anchor 之间存在密度不均匀导致的小面孔召回率[26]太低的问题。通过提出 Anchor 致密化策略,在接受域中心均匀地平铺一个数值为 n^2 的 Anchor,然后对不同尺度的 Anchor 进行致密化,保证不同尺度的 Anchor 在图像上具有相同的密度[27]。本文定义锚的平铺密度:

$$A_{\text{density}} = \frac{A_{\text{scale}}}{A_{\text{interval}}} \tag{3}$$

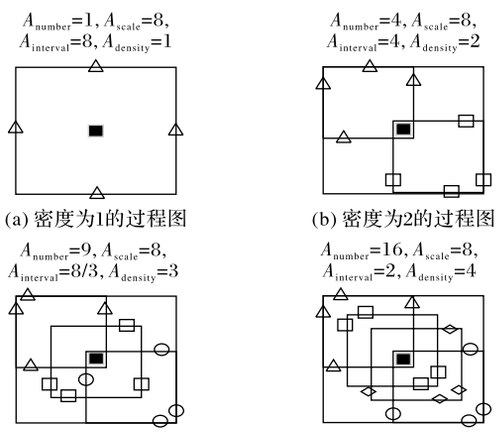
其中:A_{density} 是不同尺度 Anchor 对应的密度;A_{scale} 是 Anchor 的尺寸;A_{interval} 是 Anchor 的平铺区间。默认锚的平铺间隔分别为 32、32、32、64 和 128。例:如果 A_{interval} 值取 32,对应的密度分别为 1、2、4、4 和 4,对应密度的模拟图致密化过程如图 6 所示(A_{number} 为致密后的量化单位)。显然在不同尺度上 Anchor 的密度不均衡。

2.3.2　中心损失策略

文献[28]提出的中心损失函数在人脸识别任务中取得了非常好的效果。中心损失策略的基本思想是鼓励网络学习在扩大类间变化的同时最小化类内变化的判别特征。中心损失的公式为:

$$L_c(x) = \frac{1}{2\sum_{i=1}^{m}\left\|\boldsymbol{x}_i - \boldsymbol{c}_{\text{yi}}\right\|_2^2} \tag{4}$$

其中：$\boldsymbol{x}_i$为输入特征向量；$\boldsymbol{c}_{y_i}$为第i类中心，是一个特征矢量。在每次迭代中，中心都基于小批量更新，因此可以很容易地使用标准随机梯度下降（Stochastic Gradient Descent，SGD）进行培训。在人脸检测任务中，只有两个中心分别代表人脸和非人脸。本文研究的目的是最小化类内部的变化，需要注意的是中心损失应该与Softmax损失[29]一起优化。研究表明，中心损失在最大限度地减少类内变化方面是非常有效的，而Softmax损失在最大限度地提高学习特征的类间变化方面具有一定的优势。因此，采用中心损失和Softmax损失相结合的方法来追求判别特征是非常合理的。



(a) 密度为1的过程图 (b) 密度为2的过程图

(c) 密度为3的过程图 (d) 密度为4的过程图

图6 致密化过程模拟图

Fig. 6 Simulation of densification process

3 实验与结果分析

3.1 数据集

实验采用Wider Face数据集作为训练集，该数据集包含32 203张图片，共包含393 703张人脸，分别在图像尺度和人脸姿势以及曝光度等方面表现出了大的变化。这对本文算法学习不同情境中的人脸特征来说十分有利。整个项目中测试集采用了FDDB数据集，该数据集中包含2 845张图片，共有5 171张人脸，对于本文算法测试十分有说服力。同时，测试过程中本文从网上找来了包含200张人脸的自组织数据集进行进一步验证，该数据集内容主要涉及证件照、毕业照、行人照、生活照以及朋友圈里的人脸图片，整个数据集相对复杂且更能体验模型的泛化能力。

3.2 软硬件环境

实验的软硬件环境为：人脸检测模型的训练采用五舟高性能计算集群（High Performance Computing，HPC），显卡采用的是NVIDIA Tesla V100，算法的设计通过Darknet框架实现；移动端测试使用的是芯片型号为rk3326的嵌入式开发板，并使用墨子（深圳）人工智能技术有限公司提供小墨机器人（基于本文算法的产品）进行调试。

3.3 实验结果与分析

使用YoLov3算法和本文算法FaceYoLo进行了对比实验，过程中参考相关算法的实验设计方案[12,16]，并选用YoLov3不同的网络结构分别做了实验。实验过程中也比较了移动端比较流行的MobileNet算法，通过训练两者的算法模型并移植到准备好的开发板中测试。在模型的训练过程中，每次实验迭代次数设置为500 000，并且每5万次保存一个模型。

表2为不同算法在FDDB数据集上的实验结果。对比YoLov3算法使用darknet19、darknet152和darknet53等网络结构进行的模型测试结果可看出，darkent53的模型质量更优。与MTCNN和FaceBoxes算法等其他对比算法的结果可看出，本文FaceYoLo算法不仅保证了平均检测精度为92.6%，而且检测速度有了大幅度提高；比原始算法YoLov3使用darknet53网络的组合，本文算法在精度上提高了2.1个百分点，速度上提升至原来的8倍。

表2 不同算法在FDDB数据集上的实验结果

Tab. 2 Experimental results of different algorithms on FDDB dataset

模型	GPU	mAP/%	检测速度/(frame·s^{-1})	模型大小/MB
YoLov3+darknet19	V100	88.6	93	240.00
YoLov3+darknet152	V100	90.4	25	280.00
YoLov3+darknet53	V100	90.5	43	228.00
FaceBoxes	V100	96.0	125	—
MTCNN	V100	94.0	108	—
FaceYoLo	V100	92.6	334	2.28
MobileNet	Tablet	—	5	—
FaceYoLo	Tablet	—	10	—

表3为验证FaceYoLo算法各个模块进行的单一对照实验，实验为该算法依次去除RDCL、MSCL、Center Loss以及Anchor致密化模块与原算法进行的对比结果。结果显示，RDCL结构对速度的提升影响显著，MSCL结构对平均检测精度的影响较为显著。

表4是人脸评价指标实验，通过验证不同数据集一定数量人脸的正检率和漏检率测试模型的鲁棒性。其中FDDB测试集中共5 171张人脸，正确检测率达到92.6%，漏检查率为4.8%，误检率为0.4%；另外随机抽取的200张自组织人脸数据集测试中，正确检测率达到了96%，漏检率为4%，误检率为0.99%。数据结果表明，FaceYoLo模型不仅在标准数据集上可以达到良好的检测质量，而且面对不确定的检测对象依旧能够保证良好的检测性能，进一步验证了模型的泛化能力和鲁棒性。图7是部分测试结果展示，前两行是FDDB数据集的测试结果，最后一行是在网上挑选的三张比较经典的图片对应的测试结果，从结果来看，模型的效果是不错的。

表3 不同模块的灼烧实验结果对比

Tab. 3 Comparison of experimental results of different burning modules

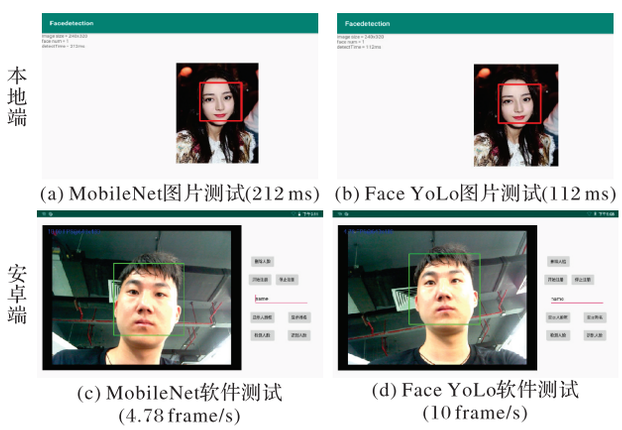
方法	显卡	mAP/%	检测速度/(frame·s^{-1})
去除RDCL	V100	90.1	261
去除MSCL	V100	89.9	340
去除Center Loss	V100	87.5	335
去除Anchor	V100	88.2	330
FaceYoLo	V100	92.6	334

图7　测试结果展示

Fig. 7　Demonstration of test results

为验证模型部署在移动端上的可行性,在人脸检测APP中,用FaceYoLo模型替换原始的MobileNet模型,并完成了两项实验。移动端测试过程如下:首先随机选定一张图片放在本地,测试Android应用在不同模型下返回人脸框的速度;其次,将实际摄像头传回来的帧送入模型,查看处理速度。图8比较的是MobileNet与FaceYoLo算法分别在本地端和APP端上的测试结果图,对比结果显示最终FaceYoLo模型速度大约提升了1倍左右。另外实验中也比较了MobileNet与其他算法的模型,其他模型在移动端的效果都次于它,所以之后只与MobileNet进行了比较。

实验结果表明,FaceYoLo算法检测速度对于原始的YoLov3算法从43 frame/s提高到334 frame/s,同时训练权重参数相对于YoLov3算法减少至原来的1%,模型大小也从228 MB减少至2.28 MB,平均检测精度达到92.6%。在移动端测试中,将MobileNet训练出的模型和本设计算法的模型放到同一台嵌入式设备中进行测试,从测试结果显示,FaceYoLo算法的检测速度接近MobileNet速度的2倍;与此同时,在安卓平台下也进行了二次模型测试,安卓平台下摄像头端实时检测人脸的速度也从5 frame/s提升至10 frame/s,这对于移动端的开发来说是很大的进步。从整个实验结果来看,新算法在保证高精度的情况下具备更快的检测速度,可以很好地适用于移动平台的人脸检测任务。



(a) MobileNet图片测试(212 ms)　(b) Face YoLo图片测试(112 ms)

(c) MobileNet软件测试 (4.78 frame/s)　(d) Face YoLo软件测试 (10 frame/s)

图8　移动平台的测试结果

Fig. 8　Test results on mobile platform

表4　人脸检测评价指标实验结果

Tab. 4　Experimental results of face detection evaluation indexes

指标	待检测人脸数	正确检测人脸数	漏检人脸数	误检人脸数	漏检率/%	正确检测率/%	误检率/%
自组织人脸数据集	200	192	8	2	4.00	96.00	0.99
FDDB数据集	5 171	4 788	383	21	4.80	92.60	0.40

4　结语

本文通过对YoLov3算法的研究,深入学习了YoLo系列网络的全连接层结构特点和Darknet网络技术。同时也深入分析了算法的网络结构对模型产生的影响,例如采用不同尺寸的小卷积核可以实现类似Inception结构的功能效果;比较BN结构与Dropout[30]结构在卷积运算过程中提取特征信息的丢失情况;ResNet跨层跳跃连接后输出和输入的特征可视化特点以及RDCL结构的步长选择问题[31]。

但此次研究工作中也有一定的局限性,比如实验中测试的硬件开发设备来自客户公司提供的移动平台,模型移植到其他不同硬件的Android设备上是否也满足实时性要求,还缺乏更多的对比实验。但是此次工作在Android手机上测试了打包的APK文件,检测人脸的速度可以达到40 frame/s,基本可以满足实时性要求。另外相对于一些专门的人脸检测算法,如FaceBoxes[32]和MTCNN[33]来说,本文算法虽然在一定的内存、耗时等条件下展现了极强的速度优越性,但是在单纯人脸的检测精度上来看仍然需要提高,这也是后面工作中针对算法精度需要考虑解决的问题。

参考文献 (References)

[1] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features [C]// Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2001: 1-9.

[2] LIAO S, JAIN A K, LI S Z. A fast and accurate unconstrained face detector [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 211-223.

[3] SCHMIDHUBER J. Deep learning in neural networks: an overview [J]. Neural Networks, 2015, 61: 85-117.

[4] YOU Q, LUO J, JIN H, et al. Building a large scale dataset for image emotion recognition: the fine print and the benchmark [C]// Proceedings of the 30th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2016: 308-314.

[5] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[6] WANG P, LIN L, SHEN C, et al. Multi-attention network for one shot learning [C]// Proceeding of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 6212-6220.

[7] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [J]. IEEE Transactions on Pattern Analysis and

Machine Intelligence, 2020, 42(2): 318-327.

[8] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications [EB/OL]. [2018-09-20]. https://arxiv.org/pdf/1704.04861.pdf.

[9] KINGMA D P, DHARIWAL P. Glow: Generative flow with invertible 1x1 convolutions[EB/OL]. [2019-02-10]. http://papers.nips.cc/paper/8224-glow-generative-flow-with-invertible-1x1-convolutions.pdf.

[10] SIMONVAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2018-05-10]. https://arxiv.org/pdf/1409.1556.pdf.

[11] WANG F, LIU W, LIU W, et al. Additive margin Softmax for face verification [J]. IEEE Signal Processing Letters, 2018, 25(7): 926-930.

[12] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [C]// Proceedings of the 2014 European Conference on Computer Vision, LNCS 8693. Cham: Springer, 2014: 740-755.

[13] KARPATHY A, LI F. Deep visual-semantic alignments for generating image descriptions [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 3128-3137.

[14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// Proceeding of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[15] GONG C, TAO D, LIU W, et al. Label propagation via teaching-to-learn and learning-to-teach [J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(6): 1452-1465.

[16] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[17] FARFADE S S, SABERIAN M J, LI L J. Multi-view face detection using deep convolutional neural networks [C]// Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. New York: ACM, 2015: 643-650.

[18] 孙劲光,孟凡宇. 基于深度神经网络的特征加权融合人脸识别方法[J]. 计算机应用, 2016, 36(2): 437-443. (SUN J G, MENG F Y. Face recognition based on deep neural network and weighted fusion of face features [J]. Journal of Computer Applications, 2016, 36(2): 437-443.)

[19] LIENHART R, MAYDT J. An extended set of Haar-like features for rapid object detection [EB/OL]. [2019-02-10]. http://www.staroceans.org/documents/ICIP2002.pdf.

[20] MISHKIN D, SERGIEVSKIY N, MATAS J. Systematic evaluation of convolution neural network advances on the ImageNet [J]. Computer Vision and Image Understanding, 2017, 161: 11-19.

[21] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[22] 牛连强,陈向震,张胜男,等. 深度连续卷积神经网络模型构建与性能分析[J]. 沈阳工业大学学报, 2016, 38(6): 662-666. (NIU L Q, CHEN X Z, ZHANG S N, et al. Modeling and performance analysis of deep continuous convolutional neural network [J]. Journal of Shenyang University of Technology, 2016, 38(6): 662-666.)

[23] REN S, HE K, GIRSHICK R, et al. Object detection networks on convolutional feature maps [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(7): 1476-1481.

[24] 熊咏平,丁胜,邓春华,等. 基于深度学习的复杂气象条件下海上船只检测[J]. 计算机应用, 2018, 38(12): 3631-3637. (XIONG Y P, DING S, DENG C H, et al. Ship detection under complex sea and weather conditions based on deep learning [J]. Journal of Computer Applications, 2018, 38(12): 3631-3637.)

[25] SHANG W, SOHN K, ALMEIDA D, et al. Understanding and improving convolutional neural networks via concatenated rectified linear units [EB/OL]. [2019-02-10]. https://arxiv.org/pdf/1603.05201.pdf.

[26] UIJLINGS J R R, VAN DE SANDE K E A, GEVERS T, et al. Selective search for object recognition [J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[27] QU Y, LIN L, SHEN F, et al. Joint hierarchical category structure learning and large-scale image classification [J]. IEEE Transactions on Image Processing, 2017, 26(9): 4331-4346.

[28] WEN Y, ZHANG K, LI Z, et al. A discriminative feature learning approach for deep face recognition [C]// Proceedings of the 2016 European Conference on Computer Vision. Cham: Springer, 2016: 499-515.

[29] LI J, MEI X, PROKHOROV D, et al. Deep neural network for structural prediction and lane detection in traffic scene [J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(3): 690-703.

[30] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates, 2017: 6000-6010.

[31] GUO Y, ZHANG L, HU Y, et al. MS-Celeb-1M: a dataset and benchmark for large-scale face recognition [C]// Proceedings of the 2016 European Conference on Computer Vision, LNCS 9907. Cham: Springer, 2016: 87-102.

[32] ZHANG S, ZHU X, LEI Z, et al. FaceBoxes: a CPU real-time face detector with high accuracy [C]// Proceedings of the 2017 IEEE International Joint Conference on Biometrics. Piscataway: IEEE, 2017: 1-9.

[33] XU K, BA J L, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention [C]// Proceedings of the 32nd International Conference on International Conference on Machine Learning. New York: Journal of Machine Learning Research, 2015: 2048-2057.

This work is partially supported by the National Key Research and Development Program of China (2018YFB1305804), the Anhui Provincial Outstanding Youth Foundation (1908085J25).

REN Haipei, born in 1994, M. S. candidate. His research interests include pattern recognition, deep learning.

LI Teng, born in 1980, Ph. D., professor. His research interests include pattern recognition, deep learning.