TinaFace:强大但简单的人脸检测baseline

蔡鸿祥朱闫佳张舒涵†王晨好‡熊逸超§传媒智能科技有限公司

2023 年 1 月 15 日

摘 要

人脸检测近年来受到了广泛的关注。许多工作从模型架构、数据增强、标签分配等不同角度提出了大量的人脸检测专用方法，使得整个算法和系统变得越来越复杂。在本文中，我们指出了**人脸检测与一般目标检测之间没有差距**。在此基础上，我们提出了一种较强而简单的人脸检测baseline:TinaFace。我们的TinaFace以ResNet-50[11]为骨干网络，其中的所有模块和技术都是在现有模块上构造的，基于通用对象检测方法，易于实现。在最流行和最具挑战性的人脸检测基准**WIDER FACE**[48]的hard测试集上，单模型和单尺度上，我们的TinaFace达到了92.1%的平均精度，这超过了大多数最近的更大的人脸检测器的表现。在使用了TTA方法之后，我们的TinaFace比目前最先进的方法表现更好，达到了92.4%的AP。

1 Introduction

人脸检测是计算机视觉中一个非常重要的任务，它是人脸识别、验证、跟踪、对齐、表情分析等大多数任务和应用的第一步。因此，近年来在这一领域出现了许多不同角度的方法。一些文献[6,7,49]将注释信息作为额外的监督信号，另外一些文献[51,57,37,17,26,25,58]更加注重网络的设计。此外，还提出了一些新的损失设计[51,57,16]和数据增强方法[17,37],还有一些工作开始重新设计匹配策略和标签分配流程。显然，人脸检测似乎逐渐从一般的目标检测中分离出来，形成了一个新的领域。

直观地说，人脸检测实际上是通用目标检测的一种应用。在某种程度上，脸就是一个检测对象。所以自然就会出现一系列问题：比如“人脸检测与一般对象检测有什么区别?”“为什么不用一般对象检测技术来处理人脸检测?”“是否有必要另外设计处理人脸检测的特殊方法?”

首先，从数据的角度来看，人脸拥有的属性也存在于物体中，比如姿态、比例、遮挡、光照、模糊等。像表情和化妆这种面部的独特属性，也可以对应物体的扭曲和颜色。人脸检测所遇到的多尺度、小人脸、密集场景等挑战都存在于一般的目标检测中。因此，人脸检测似乎只是一般对象检测的一个子问题。为了更好地进一步回答上述问题，我们提供了一种基于通用对象检测的baseline，在WIDER FACE的hard测试集上胜过目前最先进的方法。

本文的主要贡献可以总结为:

- 说明人脸检测实际上是一类通用对象检测问题，可以通过通用对象检测技术进行处理。
- 提供了一种强大而简单的面部检测基线方

*相同贡献
†数据分析
‡数据分析
§通讯作者

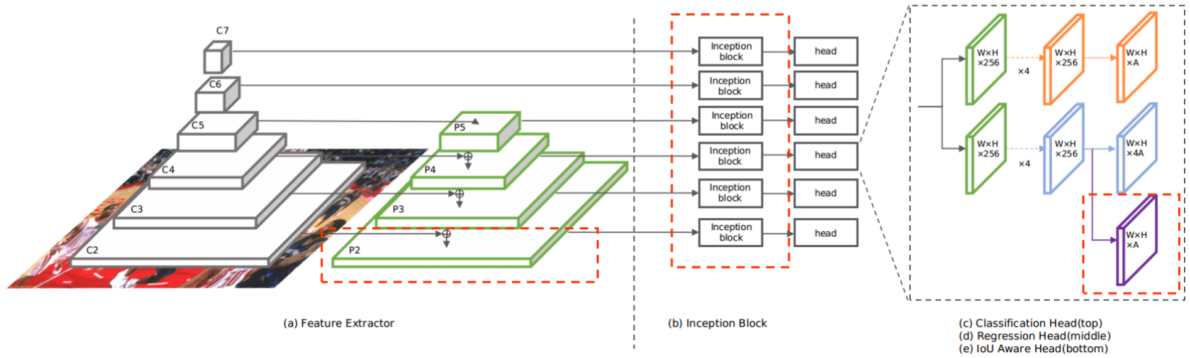

Figure 1: TinaFace的模型架构。(a)特征提取器:ResNet-50[11]和6级特征金字塔网络[18]，提取输入图像的多尺度特征。(b)Inception模块增强接受野。(c)分类头:5层FCN用于锚框的分类。 (d)回归头:5层FCN，用于锚框回归到地真框。(e) IoU感知头:用于IoU预测的单个卷积层。

法TinaFace。TinaFace中使用的所有思想和模块都是基于通用对象检测的。

- 在单尺度和单模型的情况下，我们在WIDER FACE测试子集下达到了92.1%的平均精度(AP)，这已经超过了当前大多数具有较大主干网络并使用TTA方法的模型。我们的模型最终在测试子集的中获得92.4%的AP，优于当前最先进的人脸检测方法。

2 Related Work

2.1 Generic Object Detection

，通用目标检测的目的是对给定图像中存在的目标进行定位和分类。在深度学习蓬勃发展之前，一般的目标检测主要是基于手工制作的特征描述算子，如SIFT[24]和HOG[5]。最成功的方法如DPM[8]将多尺度手工制作的特征、滑动窗口、可变形部件和SVM分类器相结合，形成通用的目标检测器。随着AlexNet[15]获得2012年大规模视觉识别挑战赛(ILSVRC2012)冠军，深度学习时代到来，通用目标检测迅速被深度学习方法所主导。两阶段方法从R-CNN[10]和Fast R-CNN[9]开始，很快，R-CNN[31]就提出了RPN网络，用预定义anchors代替selective search算法，成为最经典的基于anchors的通用目标检测方法。基于Faster R-CNN[31]，提出了很多新方法，如FPN[18]、Mask R-CNN[12]、级联R-CNN[1]等。为了克服两阶段方法的高潜伏期，出现了许多单阶段方法，如YOLO系列[30,28,29]、SSD[22]和RetinaNet[19]等。为了解决多尺度或小物体的问题，YOLOs[30,28,29]提出了新的锚点匹配策略，包括考虑建议反馈和一个地真对一个锚点，并对物体宽度和高度的回归进行重估。然后SSD[22]使用一个主干特征的层次结构，而FPN[18]使用特征金字塔。此外，SNIP[34]、SNIPER[35]系列、多尺度训练、多尺度测试也可以应对多尺度问题。除了通用目标检测中提出的新方法外，其他领域的发展，如归一化方法和深度卷积网络，也促进了通用目标检测。批处理归一化(BN)[14]沿通道维对批处理内的特征进行归一化，可以帮助模型收敛，使模型能够训练。为了处理batch size = BN的依赖关系，group normalization (GN)[44]将通道分成组，并在每组内计算归一化的平均值和方差。之后深卷积网络,AlexNet [15], VGG[33]增加深度使用架构和非常小的3×3卷积过滤器,GoogLeNet[36]介绍了《盗梦空间》模块使用不同数量的小过滤器并联形成的特性不同的接受域和帮助捕获对象以及上下文模型在多尺度, ResNet[11]展示了原始信息流的重要性，并提出了跳过连接来处理更深网络的退化（残差网络）。Face Detection 人脸检测作为通用目标检测的一种应用，其发展历史几乎是

相同的。在深度学习时代之前，人脸检测器也是基于Haar[39]等手工制作的特征。继[48]提出的最受欢迎和最具挑战性的人脸检测基准WIDER face dataset之后，人脸检测针对尺度、姿态、遮挡、表情、化妆、光照、模糊等极端和真实变化问题得到了快速发展。目前几乎所有的人脸检测方法都是从现有的通用目标检测方法发展而来的。基于SSD[22],年代3 FD [58] anchor-associated层延伸至C3阶段,提出了一种补偿规模锚匹配策略为了覆盖的小脸上,PyramidBox[37]提出PyramidAnchors (PA), 低级特征金字塔网络(LFPN),上下文敏感的预测模块(CPM)强调环境的重要性和data-anchor-sampling增加增加较小的面孔,DSFD[16]引入了改进锚匹配(IAM)和渐进锚丢失(PAL)的双镜头检测器。然后，RefineFace[57]基于视网膜et[19]，通过视网膜aface[6]人工标注人脸上的5个地标作为额外的监督信号，引入了选择性两步回归(Selective Two-step Regression, STR)、选择性两步分类(Selective Two-step Classification, STC)、尺度感知边缘损失(Scale-aware Margin Loss, SML)、特征监督模块(Feature supervision Module, FSM)和接受场增强(RFE) 5个额外模块，HAMBox[23]强调了一些不匹配锚的强大回归能力，提出了一种在线高质量锚挖掘策略(HAMBox)。此外，ASFD[51]采用神经体系结构搜索技术自动搜索体系结构，实现高效的多尺度特征融合和上下文增强。综上所述，人脸检测中的方法几乎涵盖了深度学习训练从数据处理到损失设计的每一个环节。很明显，所有这些方法都集中在小脸的挑战上。然而，实际上在通用对象检测中有很多方法可以解决这个问题，我们在前面提到过。因此，在这些方法的基础上，我们提出了TinaFace，一种强大但简单的人脸检测baseline方法。

3 TinaFace

我们在单阶段检测器RetinaNet[19]之前的工作上进行改进。TinaFace的架构如图1所示，红色虚线框显示了与RetinaNet[19]不同的部分。

3.1 Deformable Convolution Networks

卷积运算有其固有的局限性，即对采样位置的强先验是固定的、刚性的。因此，网络对复杂几何变换的学习和编码困难，模型的能力受到限制。为了进一步提高模型的性能，我们将DCN[4]应用到主干网络的第四阶段和第五阶段。

3.2 Inception Module

多尺度一直是通用目标检测中的一个难题。常用的处理方法有多尺度训练、FPN体系结构和多尺度测试。此外，我们在我们的模型中使用了inception模块[36]来进一步增强这种能力。 inception模块使用不同数量的3×3卷积层并行形成不同接受域然后将它们组合在一起，帮助模型在多个尺度上捕捉检测对象和上下文。

3.3 IoU-aware Branch

IoU-aware[43]是一种非常简单优雅的可以缓解单级目标检测器分类分数与定位精度不匹配的问题的方法，可以利用分类分数，抑制误报检测框(高分低IoU)。 IoU-aware的架构如图1所示，唯一不同的是紫色部分，一个平行头和一个用来预测被检测盒与对应的地真对象之间的IoU的回归头。而这个头部只有一个3×3的卷积层，然后是一个sigmoid激活层。在推理阶段，最终检测置信度计算公式如下:

$$score = p_i^{\alpha} IoU_i^{(1-\alpha)} \tag{1}$$

其中 p_i 和 IoU_i 是第i个检测盒的原始分类分数和预测 IoU ， $\alpha \in [0,1]$是控制分类分数和预测IoU对最终检测置信度贡献的超参数。

3.4 Distance-IoU Loss

在bbox回归中最常用的损失是smooth L1损失[9]，它回归四个坐标(box的中心及其宽度和高度)的参数。然而，这些优化目标与回归评价指标IoU并不一致，即损失越低并不等于IoU越高。因此，我们转向过去几年出现的不同IoU损失，直

接回归IoU度量，如 GIoU [32]、DIoU 和 CIoU [61]。我们之所以选择DIoU[61]作为我们的回归损失，是因为小人脸是人脸检测的主要挑战，在Widerface[48]中约有三分之二的数据属于小目标，而 DIoU [61]对小目标更友好。在实际应用中，DIoU 在MS COCO 2017[20]验证集APsmall上的性能较好。理论上，DIoU定义为:

$$L_{DIoU} = 1 - IoU + \frac{\rho^2(\mathbf{b}, \mathbf{b^{gt}})}{c^2} \tag{2}$$

其中 $boldb$ 和 $boldb_{gt}$ 表示预测框和地真框的中心点，$\rho()$ 是欧氏距离， c 是覆盖两个盒的最小外接盒的对角线长度。额外罚款项 $\frac{\rho^2(\mathbf{b}, \mathbf{b^{gt}})}{c^2}$ 提出对预测盒中心点与地面真盒之间的归一化距离进行最小化。相对于大目标，同样距离的小目标中心点会受到更多的惩罚，这有助于检测器在回归过程中更好的学习小目标。