基于 Mask R-CNN 与 SSD 的口罩佩戴识别技术对比的研究

全世杰 李丹
四川大学锦城学院 四川 成都 610065

【摘 要】近些年，由于计算机性能的极大提升以及人们生活中的更多需求，目标检测越来越广泛地被运用在了我们的日常生活中，从而也出现了非常多的用于目标检测，语义分割以及实例分割的深度学习的神经网络算法，传统的目标检测算法精准度低，而且需要耗费巨大的时间来完成滑动窗口工作，R-CNN 以及 Fast R-CNN[4] 的成功提出使得基于深度机器学习的目标检测机器学习算法逐渐变得高效，使得基于目标价的检测算法的精度和性能都有了质的提升和飞越，而在之后的 Faster R-CNN[3] 被何凯明提出之后这样的基于目标价和深度机器学习的目标检测机器学习算法已经开始逐渐变得成熟，由何凯明和 Ross Girshick 等人提出的 Mask R-CNN[1] 的成功出现被认为是基于目标检测机器学习算法的一个重要里程碑，它不仅便利，速度快，而且精准度也非常高，能够非常高效地胜任目标检测、实例分割等工作。而 SSD 是另一种目标检测的算法，它是由 Wei Liu 等人提出 [2]，它结合了 yolo 与 Faster R-CNN 的优点，具有速度极快的特点，同时在具有高效的训练速度以及在高速的识别下仍具有极高的准确性。而且 Mask R-CNN 与 SSD 非常灵活，当下正值疫情期间，如果能对每个人口罩的佩戴进行检测那么会对防控工作有一定的促进，我想通过使用 Mask R-CNN 和 SSD 来对人体的口罩佩戴进行一个识别。因为我寻找了一些佩戴了口罩的人脸图片来作为训练的数据集，我希望能够训练出一个识别口罩佩戴的模型从而快速地完成口罩佩戴的识别工作。

【关键词】Mask R-CNN；识别技术；研究

引言

人体口罩佩戴的识别的目的是识别一张图像中的人是否有佩戴口罩，本文中是使用了 Mask R-CNN 以及 SSD 来完成该任务工作。而这样的算法的本质其实是一种目标检测的工作，目标检测在人工智能计算机视觉的领域是一种非常重要的任务，且近年来更是如火如荼，甚至愈演愈烈而未有消停之势。目标检测运用极为广泛，我将以此来完成人体的口罩是否佩戴的识别。

近期正值疫情，而佩戴口罩可以有效地抑制疫情的蔓延，正因如此检测每个人口罩的佩戴这样的一个工作变得非常重要，人工检测难免有所疏漏，而且利用传统的旧方法来完成智能化的口罩佩戴识别依然效率低下且不精准，每一次从图片中做特征提取都是一个非常耗时的工作，且不说提取结果的精准度不太高，在时间上的开销已经是我们不可接受的了，而且在此之后需要用到一个分类器来完成目标的分类而实现检测，不同的分类器的效率与时间消耗又是不同，一些相对精准的分类器如 SVM 等需要耗费大量的时间来完成分类工作，在结合特征提取的时间开销，这是无法在人流量密集的场所来使用，对计算机性能的刚需也成了无法大面积普及的缘由。采用 Mask R-CNN 以及 SSD 目标检测网络的方式来进行口罩佩戴识别可以有效地解决以上的一系列疑难问题。

1 相关工作

利用传统的技术来完成口罩识别任务需要对图像进行滑动窗口操作，然后得到的结果传入分类器中分类，最终得到识别的结果，而这样的工作是非常繁琐而麻烦的，操作时间非常长，而且精度很差，最终得到的结果往往强差人意，我希望能够使用一种深度学习的神经网络来完成口罩识别任务来使得效率更高结果更精准，即 Mask R-CNN 以及 SSD。

Mask R-CNN 与 SSD 都是结合了之前许多优秀的算法而提出的。在 2013 年，Ross Girshick 等人结合 Region proposals 和 CNNs 来提出了 R-CNN，R-CNN 在分类的时候采用了 SVM，为了优化这一点，在 2015 年，Ross Girshick 在此基础上提出了 Fast R-CNN，它在最后使用了全连接层来替代了 SVM 并且使用 softmax 来作为回归的方法，使得该方法在性能上有大幅提升。而为了省去在 Fast R-CNN 中 selective search 方法所带来的时间开销，

在2015年同年Shaoqing Ren,Kaiming He,Ross Girshick等人又提出了Faster R-CNN，它使用了一个RPN的全卷积网络来代替了selective search节省了每一次检测的时候的时间开销。R-CNN，Fast R-CMM和Faster R-CNN的出现使得目标检测算法在深度学习上的性能变得非常高效，而为了完成更准确的目标检测以及完成实例分割等计算机视觉任务，在2017年何凯明等将Faster R-CNN，Resnet以及FPN的优势结合起来推出一种新的算法即Mask R-CNN，他采用了Roi Align的方法来代替了Roi Pooling，极大地提升了准确度而使得实例分割成为了可能，并且提供了mask分支来完成mask的生成、解耦物体框等工作，Mask R-CNN的效率和精准度相较之前的算法更加优秀。

同样是基于Faster R-CNN，SSD也是一种在此基础上改良出的一种算法，SSD不仅为了提高精确度采用了FPN（特征金字塔网络）的思想同时在多个不同尺度的feature map上进行softmax的回归和分类操作，而且SSD参考了Faster R-CNN的Anchor思路提出一了Piror Box，这是一种新的检测框，用于进行高效的分类与回归。

2 基于Mask R-CNN的口罩佩戴识别技术

2.1 Mask R-CNN

自2017年以来Mask R-CNN被提出以后一直以高效以及高精准而闻名，它是在Faster R-CNN的基础之上提出的，相较于Faster R-CNN而言，Mask R-CNN为了使得有更高的精准率使用RoiAlign来代替了RoiPooling技术。在RoiPooling中，由于每一次在feature map上得到的roi并不是一定能够与像素点对齐，而为了进行pooling操作将会对得到的roi进行一些移动或者伸缩，而这样造成的直接结果就是降低了精确度，在这个时候何凯明等人提出了RoiAlign的方法，通过双线性插值运算来进行对没有完全与像素点对齐的roi来取点位从而以取得的点位来进行之后的Max Pooling操作，这样的方法不会对之前获得的roi进行位置以及大小的调整，从而不会有准确率的丢失，在不会影响效率的同时提升了精确度。其实Mask R-CNN与Faster R-CNN最大的区别是它可以用来做实力分割，它特别添加了一个mask分支来完成实例分割的操作。使用Mask R-CNN来进行口罩的佩戴识别能够相较之下获得更高的准确率。

2.2 SSD

引入SSD来与Mask R-CNN作对比旨在做精确度与速度的比较，Mask R-CNN与SSD同样是使用了特征金字塔思路的神经网络，且都是基于Faster R-CNN提出的，适合用与做比较。SSD是在yolo与Faster R-CNN的基础上提出的一种结合了FPN（特征金字塔网络）的一种用于目标检测的神经网络，相对于Faster R-CNN而言SSD具有更快的速度，而在Faster R-CNN的基础上，SSD基于其Anchor思想来提出了一种名为Piror Box的检测框，Piror Box可以直接在进行分类与回归操作。SSD在采用了特征金字塔网络的情况下在多种尺寸上的feature map上进行预测和分类，并且将每一个尺度的feature map预测出的尺寸不相同的bounding-box结合为一个bounding-box集合，继而由非极大值抑制的方式来将不合预期或者是不正确的bounding-box删除掉，从而最终得到一个分类的结果。

3 实验

3.1 数据集的制作

为了进行通过Mask R-CNN来实现人脸口罩佩戴识别的目的，我在各种途径中收集到了一系列佩戴口罩的人脸的图片，其中有不同环境中的佩戴口罩人脸，包含了医护人员，行人，工人，明星等，因此图片中所包含的环境涵盖了医院，街道，工地等众多常见的场所，而为了提高在一些有遮掩的情况下如佩戴了口罩时用手遮掩口鼻以及未佩戴口罩时以手遮掩口鼻等情况的图片，希望能够通过训练这样有遮掩的图片来实现一系列疑难情况下的口罩佩戴的正确识别。我为了使用Mask R-CNN来训练获得的图片数据集，我将每一张图片手动标记人脸并且制作为coco标准的数据集。由于适用于所配置网络的大部分权重更契合实例分割，所以我没有使用预设模型，而是从头来训练出一个适合口罩佩戴检测的Mask R-CNN模型。本次实验是在ubuntu环境下使用pytorch进行的，在训练的时候我采用了一张Nvidia 2080ti来进行GPU加速训练，总共训练了14000轮，耗时16小时，因为我没有屏蔽Mask R-CNN的实例分割功能，所以在训练的时间上会耗时久一点。在训练完成之后我发现得到的模型能够有效地对完整无遮掩的佩戴了口罩的人脸进行标记，且对一些有遮掩情况的人脸识也能达到不错的识别效果。而对于SSD我同样采用了14000轮的训练，同样是在一张2080ti的GPU加速下来进行训练，耗时2小时20分。

3.2 实验结果与对比

图1为Mask R-CNN与SSD在140000轮训练之后得到的AP值在matplotlib绘制的条形图下的对比。

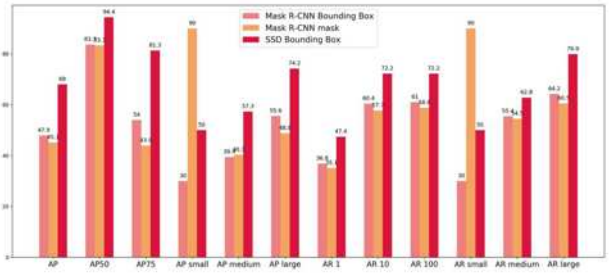

图 1 Mask R-CNN 与 SSD 的 AP

如图 1 所示，在同样经过了 140000 轮的训练之后，SSD 的准确率在大部分情况下都比 Mask R-CNN 更高，而 Mask r-cnn 对小目标识别的准确率却远高于 SSD，这意味着 Mask R-CNN 相较 SSD 而言能够对小目标进行更加精准的判断，这将会在密集人群环境的情况下进行人脸口罩佩戴检测会有更大的优势而不会导致目标的丢失，而 SSD 在对目标进行识别之后高准确率意味着识别成功之后不容易对当前人脸是否佩戴口罩进行错误的判断。

我首先做了一个两种不同的神经网络在训练速度上的比较，Mask R-CNN 每 20 轮耗时约 7 秒，而 SSD 每 20 轮耗时约 2 秒，在训练的速度上 SSD 具有明显的优势，在一张 2080ti 的 GPU 加速下 SSD 进行训练 140000 轮仅耗时 2 小时 20 分。而在检测速度上 Mask R-CNN 也要略微慢于 SSD，SSD 能够进行每秒 11 张图的检测，而 Mask R-CNN 每秒为 6 张。不过这样的差距在可控范围之内，SSD 的速度固然快速，但是 Mask R-CNN 的速度完全能够满足使用。

3.3 Mask R-CNN 与 SSD 对口罩佩戴的识别

Mask R-CNN 和 SSD 都是优秀的目标检测网络，而且都具有极佳的泛用性。在普通情况下，既正面完整显示的正确佩戴口罩的人脸、未佩戴口罩的人脸，Mask R-CNN 与 SSD 都能正确的识别是否佩戴了口罩。而在未佩戴口罩且用手挡住少部分脸部的情况下，Mask R-CNN 与 SSD 也能正确识别。在部分人脸模糊的情况下，两种模型同样能够完成正确识别口罩的配置与否。

（a）　（b）　（c）

（d）　（e）　（f）

图 2 普通场景下 Mask R-CNN 与 SSD 对口罩佩戴的识别

如图 2 所示，(a)(b)(c) 是 Mask R-CNN 识别的结果，(d)(e)(f) 是 SSD 识别的结果，在这样的场景下两种网络都能识别出人脸并且正确判断为是否口罩佩戴，Mask R-CNN 中我保留了蒙版的输出，将脸部的轮廓进行了一个包裹。

Mask R-CNN 与 SSD 对面部大面积遮挡的人脸以及多目标数量况下的口罩佩戴识别。

当场景稍微复杂，即图片中人数比较多的情况下，识别结果如图 3 所示。

（a）　（b）　（c）

（d）　（e）　（f）

图 3 在图片中人数较多及面板大面积遮挡的情况下 Mask R-CNN 与 SSD 对口罩佩戴的识别

如图 3，(a)(b)(c) 是 Mask R-CNN 识别的结果，(d)(e)(f) 是 SSD 识别的结果，可以看出，在图片中出现大量人脸以及人脸在佩戴了口罩的情况下被大面积遮挡时，SSD 可能会出现遗漏一些人脸没有识别出来的情况，而 Mask R-CNN 能够正确地将图片中的人脸全部标记出来并且正确地识别。如图 3(a)(b),(d)(e)，在人群密度大的时候，SSD 有出先人脸判断遗漏的情况，而未遗漏的人脸都能正确判断是否佩戴口罩，而 Mask R-CNN 本没有出现判漏的情况。如图 3 中 (c),(f) 所示，佩戴了口罩之后口罩部分被大面积遮挡的情况下，Mask R-CNN 能进行精准的识别，而在 (f) 中自行车后座的人脸因为被大面积遮挡而没有被 SSD 正确判断为人脸。由此可见在这样的复杂场景下，SSD 的精度不能满足于大面积遮挡情况下的人脸以及人群密度大时的口罩佩戴识别，但是 Mask R-CNN 能够精准识别出来，可见在口罩识别的场合中，Mask R-CNN 的精度相较 SSD 更高。

可见在容易识别的情况下两种网络都有优秀的检测效果，而在人群密度大以及脸部遮挡大的时候 SSD 却有一些不足，为了验证在这样的情况下 Mask R-CNN 相较于 SSD 的检测优势，我另外筛选了一些人群密集以及面板遮挡严重情况的图片来做测试，分别用两个网络训练的模型进行检测并且将最终得到的 bounding-box 结果的准确率与召回率做了一个比较，如图 4 所见，此时的准确率和召回率都是 Mask R-CNN 要更高，可见在人群密集的情况下 Mask R-CNN 的检测效果相较于 SSD 是有优势的。

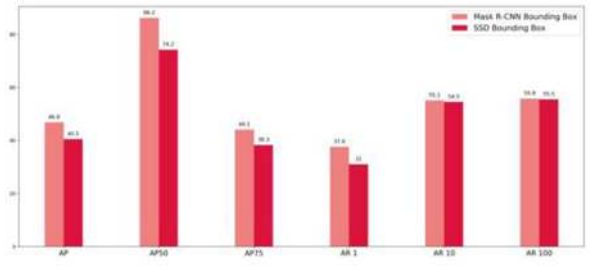

图 4 Mask R-CNN 与 SSD 在密集人群以及面部大规模遮挡情况下 bounding-box 的准确率与召回率比较

4 结束语

在 Mask R-CNN 与 SSD 的口罩佩戴识别的对比实验中可以看出，SSD 相较于 Mask R-CNN 具有更快的速度，而 Mask R-CNN 相较 SSD 具有更高的精度，SSD 适用于需要快速对口罩佩戴进行识别的场所，适合大量人群依次排队路过这样的情景，而 Mask R-CNN 适合于需要在复杂场景下对口罩佩戴进行精准识别的场景，而且 Mask R-CNN 还具有实例分割功能，能将每一个佩戴或未佩戴口罩的人脸进行区分，适用于对识别精细度有较高要求的场合。

【参考文献】

[1] Kaiming H , Georgia G,Piotr D . Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018:1-1.

[2] Liu W , Anguelov D , Erhan D. SSD: Single Shot MultiBox Detector[J].2016.

[3] Ren S , He K , Girshick R. 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.

[4] Ross Girshick. Scale-aware Fast R-CNN for Pedestrian Detection[J]. computer science, 2015.