

人脸口罩检测

郑家瀚 2019 医学院 生物医学工程 @mails.tsinghua.edu.cn

姚非凡 2019312571 医学院 生物医学工程 yff19@mails.tsinghua.edu.cn

摘要 本文利用深度学习技术对人脸口罩进行识别，构建了 3 种不同的模型，分别是一阶段的 SSD 以及二阶段的 Faster RCNN 以及 Focal Loss，来做口罩目标的检测。同时也针对结果最优秀的 Faster RCNN 做了一个电脑摄像头的实时检测。

Keywords—目标检测 SSD Focal Loss Faster RCNN

I. 简介

在 2020 年即将到来之时，新冠肺炎的病毒侵入了中国的武汉，随之席卷了全中国乃至全球。在抗击疫情的过程中，人脸口罩检测 (Face Mask Detection) 是其中一项必要的工作，通过对口罩的检测，可以加快后续的人脸识别操作，也是对医疗卫生、公共安全的保障。

口罩的识别任务主要依据的是目标检测 (Object Detection) 技术。目标检测关注是在图片中特定的物体目标，要求同时获得单个目标或多个目标的类别信息和位置信息。目标检测给出的是对图片前景和背景的理解，需要从背景中分离出感兴趣的目标，并确定这一目标的描述 (类别和位置)，因而，检测模型的输出是一个列表，列表的每一项使用一个数据组给出检出目标的类别和位置 (常用矩形检测框的坐标表示)。而我们则是需要在佩戴口罩的图片中识别出口罩是否存在；存在遮挡物的话，是否是口罩；存在的话，给出对应的位置。

A. 单阶段 (1-stage) 检测模型

单阶段模型没有中间的区域检出过程，直接从图片获得预测结果，也被成为 Region-free 方法。

1) YOLO

YOLO[1] 是单阶段方法的开山之作。它将检测任务表述成一个统一的、端到端的回归问题，并且以只处理一次图片同时得到位置和分类而得名。YOLO 将图片缩放，划分为等分的网格，每个网格按跟 Ground Truth 的 IoU 分配到所要预测的样本，其卷积网络由 GoogLeNet 更改而来，每个网格对每个类别预测一个条件概率值，并在网格基础上生成 B 个 box，每个 box 预测五个回归值，四个表征位置，第五个表征这个 box

含有物体（注意不是某一类物体）的概率和位置的准确程度（由 IoU 表示）。测试时，分数如下计算：

$$Pr(Class_i|Object) * Pr(Object) * IOU_{pred}^{truth} = Pr(Class_i) * IOU_{pred}^{truth}$$

等式左边第一项由网格预测，后两项由每个 box 预测，以条件概率的方式得到每个 box 含有不同类别物体的分数。因而，卷积网络共输出的预测值个数为 $S \times S \times (B \times 5 + C)$ ，其中 S 为网格数，B 为每个网格生成 box 个数，C 为类别数。在后处理上，YOLO 使用 NMS (Non-Maximum Suppression，非极大抑制) 过滤得到最后的预测框。损失函数被分为三部分：坐标误差、物体误差、类别误差。为了平衡类别不均衡和大小物体等带来的影响，损失函数中添加了权重并将长宽取根号。

YOLO 的主要优点是快；全局处理使得背景错误相对少，相比基于局部（区域）的方法，泛化性能好。

2) SSD: Single Shot Multibox Detector

SSD 算法 [2] 在传统的基础网络（比如 VGG）后添加了 5 个特征图尺寸依次减小的卷积层，对 5 个特征图的输入分别采用 2 个不同的 3×3 的卷积核进行卷积，一个输出分类用给的 confidence，每个 default box (default box，是指在 feature map 的每个小格 (cell) 上都有一系列固定大小的 box) 生成 21 个类别的 confidence；一个输出回归用的 localization，每个 default box 生成 4 个坐标值，最后将 5 个特征图上的结果合并 (Contact)，送入 loss 层。

SSD 在基础网络 (VGG) 后添加了辅助性的层进行多尺度卷积图的预测结果融合，提出了类似 Anchor 的 Default boxes，解决了输入图像目标大小尺寸不同的问题，同时提高了精度，可以理解为一种特征金字塔；SSD

提出了一个彻底的 end to end 的训练网络，保证了精度的同时大幅度提高了检测速度，且对低分辨率的输入图像的效果很好。

B. 两阶段（2-stage）检测模型

两阶段模型因其对图片的两阶段处理得名，也称为基于区域（Region-based）的方法。

1) R-CNN

R-CNN[3] 将检测抽象为两个过程，一是基于图片提出若干可能包含物体的区域（即图片的局部裁剪，被称为 Region Proposal），文中使用的是 Selective Search 算法 CNN 对输入图像的大小有限制，所以在将候选区域输入 CNN 网络之前，要将候选区域进行固定尺寸的缩放，缩放分为两大类：各向同性缩放，长宽放缩相同的倍数与各向异性缩放，长宽放缩的倍数不同；二是在提出的这些区域上运行当时表现最好的分类网络（AlexNet），对 CNN 输出的特征用 SVM 进行打分，得到每个区域内物体的类别，针对每个类，通过计算 IoU 指标，采取非极大性抑制，以最高分的区域为基础，剔除掉那些重叠位置的区域，并将 CNN 对候选区域提取出的特征输入训练好的线形回归器中，得到更为精确的位置定位，实现时加入了 log/exp 变换来使损失保持在合理的量级上，可以看做一种标准化（Normalization）操作。

R-CNN 将检测任务转化为区域上的分类任务，是深度学习方法在检测任务上的试水。模型本身存在的问题也很多，如需要训练三个不同的模型（proposal, classification, regression）、重复计算过多导致的性能问题等。

2) Fast R-CNN

Fast R-CNN[4] 指出 R-CNN 耗时的原因是 CNN 是在每一个 Proposal 上单独进行的，没有共享计算，便提出将基础网络在图片整体上运行完毕后，再传入 R-CNN 子网络，共享了大部分计算，故有 Fast 之名。

图片经过 feature extractor 得到 feature map，同时在原图上运行 Selective Search 算法并将 RoI (Region of Interest, 实为坐标组，可与 Region Proposal 混用) 映射到 feature map 上，再对每个 RoI 进行 RoI Pooling 操作便得到等长的 feature vector，将这些得到的 feature vector 进行正负样本的整理（保持一定的正负样本比例），分 batch 传入并行的 R-CNN 子网络，同时进行分类和回归，并将两者的损失统一起来。Fast R-CNN

将 Proposal, Feature Extractor, Object Classification 和 Localization 统一在一个整体的结构中，并通过共享卷积计算提高特征利用效率。

Faster R-CNN 是 2-stage 方法的奠基性工作，提出的 RPN 网络取代 Selective Search 算法使得检测任务可以由神经网络端到端地完成。RPN 网络将 Proposal 这一任务建模为二分类（是否为物体）的问题。第一步是在一个滑动窗口上生成不同大小和长宽比例的 anchor box (如上图右边部分)，取定 IoU 的阈值，按 Ground Truth 标定这些 anchor box 的正负。于是，传入 RPN 网络的样本数据被整理为 anchor box (坐标) 和每个 anchor box 是否有物体（二分类标签）。RPN 网络将每个样本映射为一个概率值和四个坐标值，概率值反应这个 anchor box 有物体的概率，四个坐标值用于回归定义物体的位置。最后将二分类和坐标回归的损失统一起来，作为 RPN 网络的目标训练。由 RPN 得到 Region Proposal 在根据概率值筛选后经过类似的标记过程，被传入 R-CNN 子网络，进行多分类和坐标回归，同样用多任务损失将二者的损失联合。

3) Focal Loss

Focal Loss 的模型，Facebook 团队称之为 Retinanet，是由 Facebook 团队研发出来的，其中来自广州的作者——何恺明，便曾经参与了 Faster-RCNN 模型的开发，所以也不难猜到这两个模型是有其相似性的，而实际上 Focal Loss 模型便是基于 Faster-RCNN 改进开发出来的。

迄今为止，最高精度的目标检测模型是基于 R-CNN 普及的两阶段方法，它们同样是可以使用在高度稀疏的图像检测上的。相反的，如果应用在重复性高的采样，一阶段检测器可能会变得更快更简单，但到目前为止，它的精度已经落后于二阶段检测器。这是因为，在检测器密集的训练过程中，遇到了极端前景/背景类别的失衡问题。而 Retinanet 通过重塑标准交叉熵来解决此类不平衡问题，从而降低把权重分配给容易分类的示例的可能性。这是个在当时很富开创性的工作，它让模型能够把重点放在稀疏、困难的示例上，并防止在培训过程中，受到太多 easy negative 的影响。它的效能当时是超越了 Faster-RCNN 的算法模型的，而且运算速度可以跟一阶段检测模型比拟，其准确度也比它们要高。

C. 深度学习目标检测的发展

深度学习对目标检测的研究上不断发展，诞生了大量的工作。

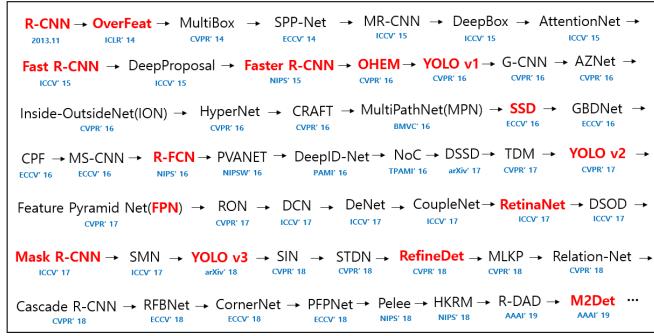


图 1. 2013 至 2019 深度学习目标检测发展

在这个基础上，一系列技术也得到了进步 [5]：

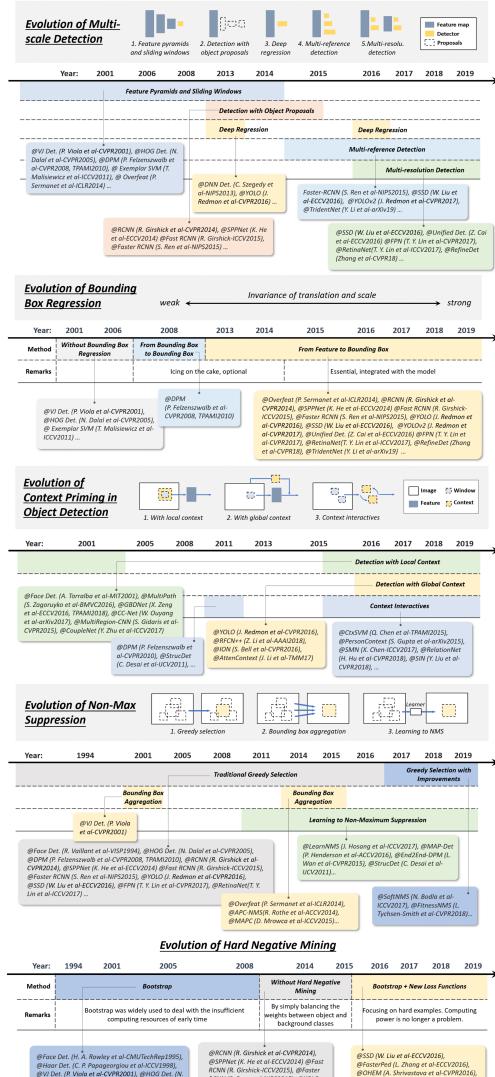


图 2. Evolution of techniques in object detection from 2001 to 2019

D. 本文贡献

本文利用深度学习技术检测对人脸口罩进行识别。本文构建了 3 种不同的模型，分别是一阶段的 SSD 以及二阶段的 Faster RCNN 以及 Focal Loss，来做口罩目标的检测。同时也针对结果最优秀的 Faster RCNN 做了一个电脑摄像头的实时检测。

II. 数据整理

采用数据为公开数据集：AIZOO 的 FaceMaskDetection 数据集。

AIZOO 的 FaceMaskDetection 数据集 (<https://github.com/AIZOOTech/FaceMaskDetection>) 开源了人脸口罩检测的主流框架的相应模型，并提供了相应的推理代码。该作者开源了如表格所示的 7,959 张人脸标注图片，数据集来自于 WIDER Face 和 MAFA 数据集，并重新修改了标注和校验。

我们使用的 AIZOO 数据集，可以发现该数据集有两类的数据，分别为口罩和非口罩人像，且数据相对平衡，数据集包含对每张照片的注释，注释信息包含图片的类别、目标的位置，该数据集适合作为训练与测试，该数据分布如下：

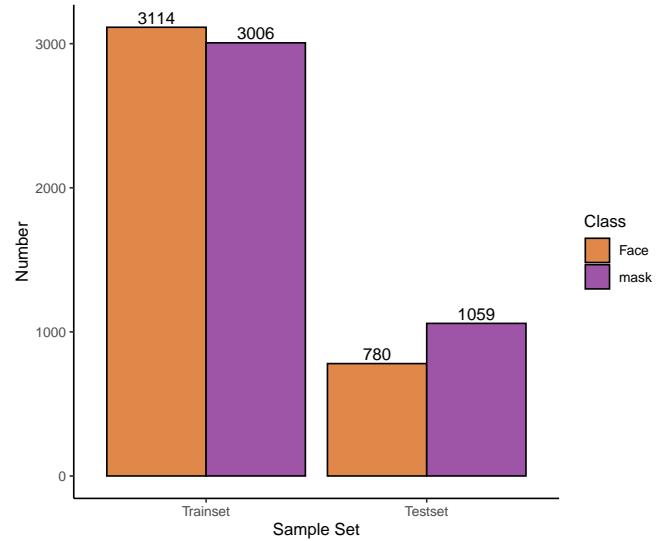


图 3. 数据集分布可视化

在把数据应用在模型训练之前，我们务必要搞清楚要使用的标签类别与格式。主要的标签类别有 Bounding Boxes、Polygonal Segmentation、Semantic Segmentation 以及 3D cuboids，在此一模型我们想当然用的就是用得最广泛的 Bounding boxes 的格式，而刚好 AIZOO 的数据类型也是属于这一类。在数据标签

格式方面的话，AIZOO 是属于 Pascal VOC 的标签格式，而我们在 SSD 使用便是这一类型；在 Faster RCNN 以及 Focal Loss 则是采用了 COCO 的标签格式。之所以会采用两种不同的数据格式，这是因为考虑到原模型使用了这样的标签格式，为了避免成果相差太大，我们采用了符合原模型的数据类型。

在数据整理的主要工作，便是要把 Pascal VOC 转换到 COCO。Pascal VOC 格式是一个照片对应一个同名的 xml 文件，而 COCO 则是所有照片对应一个 json 文件。在转换过程中也有发现一些错字，比如 face_mask 打成了 face_nask。在格式转换过程中，最大的难点在于原文件的数据格式并不一致，有者少了 pascal voc 的 path 的数据，有者则是没有图像大小的数据，而且数字序号并不统一，这些都是必须要解决的。毕竟 COCO 格式是在假定名称后面的数字序号是不重复的，因此可以直接通过序号来搜索资料。在转换过程中，统一把 path 的资料只留下图像的名称，序号也重新计算，格式转换是通过 python 完成，也一并附在参考代码里了。

III. 模型设计

A. SSD

使用了 SSD 类型的架构，本模型输入大小为 260x260，主干网络只有 8 个卷积层，加上定位和分类层，一共只有 24 层（每层的通道数目基本都是 32、64、128），只有 101.5 万参数。八个卷积层是主干网络，也就是特征提取层，20 层是定位和分类层。训练目标检测模型，最重要的合理的设置 anchor 的大小和宽高比，笔通过统计数据集的目标物体的宽高比和大小来设置 anchor 的大小和宽高比，因为人脸的一般是长方形的，而很多图片是比较宽的，人脸的宽度和高度归一化后，有很多图片的高度是宽度的 2 倍甚至更大。从上图也可以看出，归一化后的人脸高宽比集中在 12.5 之间。根据数据的分布，我们将五个定位层的 anchor 的宽高比统一设置为 1, 0.62, 0.42。（转换为高宽比，也就是约 1, 1.6: 1, 2.4:1）。

为了避免使用手挡住嘴巴就会欺骗部分口罩检测系统的情况，在数据集中加入了部分嘴巴被手捂住的数据，另外在训练的过程中，随机的往嘴巴部分粘贴一些其他物体的图片，从而避免模型认为只要露出嘴巴的就是没戴口罩，没露出嘴巴的就是带口罩这个问题，通过这两个规避方法，解决了非口罩遮挡物被当作口罩的误

判。后处理部分主要就是非最大抑制（NMS），我们使用了单类的 NMS，也就是戴口罩人脸和不戴口罩人脸两个类别一起做 NMS，从而提高速度。

迭代下模型 Loss 如下：

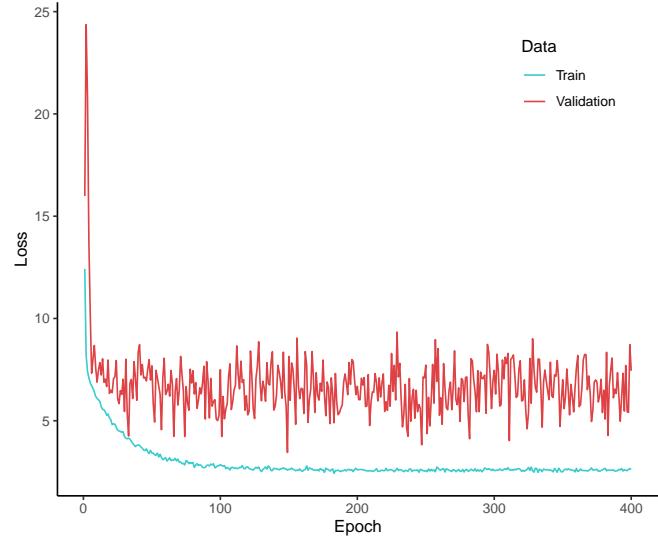


图 4. 数据集分布可视化

IV. 实验设计及结果

数据集将训练集分为训练样本和验证样本，并对测试集中的样本进行测试。

下表中展示各个模型每个类别的 mAP@.5, mAP@.7, mAP@.9, mAP@[.5:.95]。

表 I
不同模型下的 MAP

Model	Class	mAP@.5	mAP@.7	mAP@.9	mAP@[.5:.95]
SSD	Mask	0.86	0.78	0.22	0.65
	Face	0.81	0.78	0.29	0.66
Focal Loss	Mask	0.90	0.78	0.14	0.70
	Face	0.87	0.84	0.34	0.71
Faster RCNN	Mask	0.91	0.82	0.25	0.72
	Face	0.91	0.89	0.47	0.73

SSD 模型下 IoU 阈值取 0.5, 0.7, 0.9 时每类的 Precision-Recall 曲线如下：

V. 实验结果分析

SSD 模型结构：

不同模型的检测时间不同，对测试集的所有图片进行测试，再进行平均可以发现 SSD 模型的检测时间最短，其次是 Focal Loss，最长的是 Faster RCNN：

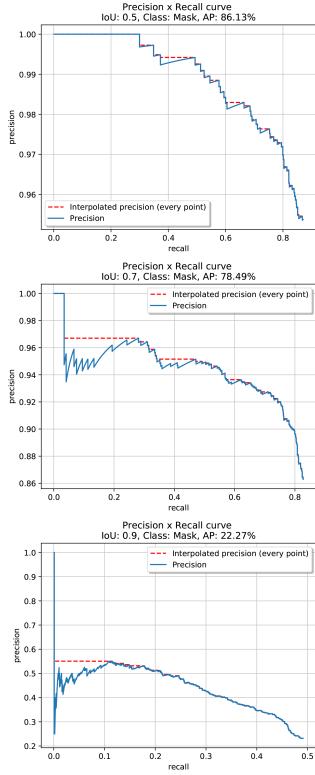


图 5. SSD 模型的 Precision-Recall 曲线（左侧从上到下分别是口罩目标检测在 $\text{IoU}=0.5$ 、 0.7 、 0.9 时的 Precision-Recall 曲线；右侧为脸目标检测。）

VI. 小组成员贡献

郑家瀚：

姚非凡：

VII. 参考代码

见附件中。

VIII. 结论

通过对口罩进行目标检测，发现三种模型的检测结果中，Faster RCNN 的准确率最高，不仅如此，其对脸的目标检测也最优。时间上 SSD 最快，但对脸的目标检测最为不佳。

参考文献

- [1] Redmon, J. , Divvala, S. , Girshick, R. , & Farhadi, A. . (2015). You only look once: unified, real-time object detection.
- [2] Liu, W. , Anguelov, D. , Erhan, D. , Szegedy, C. , Reed, S. , & Fu, C. Y. , et al. (2016). Ssd: single shot multibox detector.
- [3] Girshick, R. , Donahue, J. , Darrell, T. , & Malik, J. . (2013). Rich feature hierarchies for accurate object detection and semantic segmentation.
- [4] Ren, S. , He, K. , Girshick, R. , & Sun, J. . (2015). Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence.

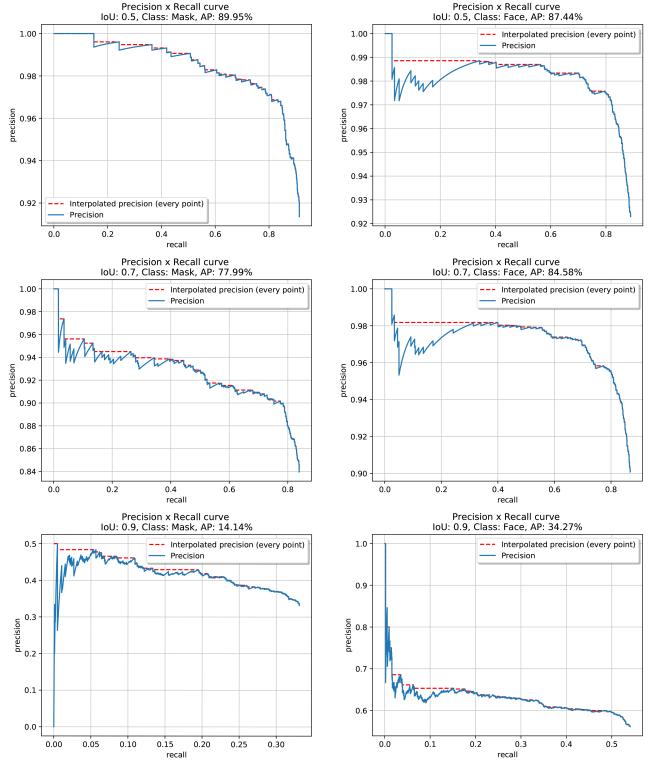


图 6. Focal Loss 模型的 Precision-Recall 曲线（左侧从上到下分别是口罩目标检测在 $\text{IoU}=0.5$ 、 0.7 、 0.9 时的 Precision-Recall 曲线；右侧为脸目标检测。）

[5] Zou, Z. , Shi, Z. , Guo, Y. , & Ye, J. . (2019). Object detection in 20 years: a survey.

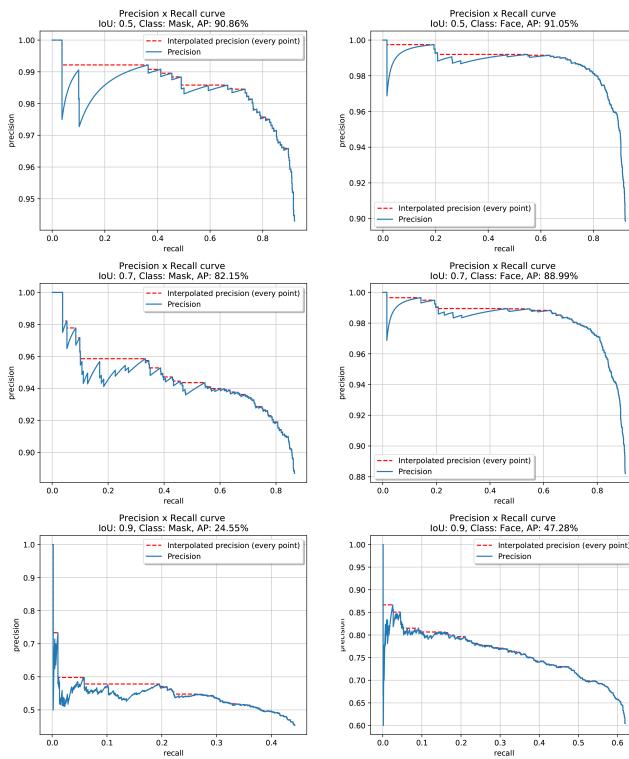


图 7. Faster RCNN 模型的 Precision-Recall 曲线（左侧从上到下分别是口罩目标检测在 IoU=0.5、0.7、0.9 时的 Precision-Recall 曲线；右侧为脸目标检测。）

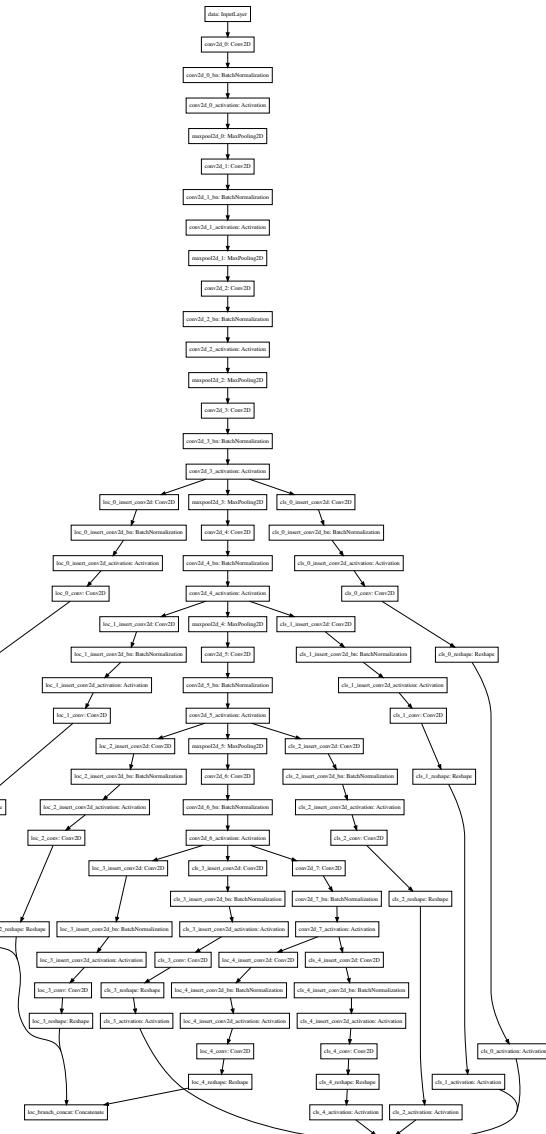


图 8. SSD 网络结构

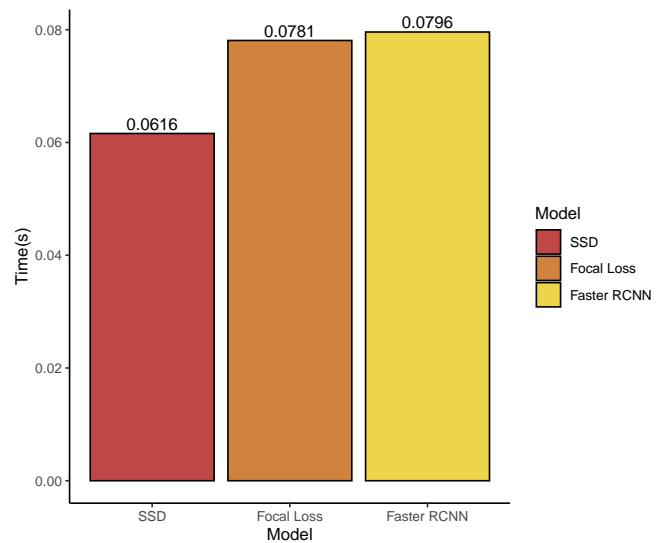


图 9. 不同模型检测的时间