Speed/accuracy trade-offs for modern convolutional object detectors

Reporter

冯浩哲

提纲

论文立足于一个基本问题，即我们如何衡量object detector的性价比。先前的论文以及它们所提出的新结构都是基于如何提高准确率的问题，而唯一涉及到efficiency的话题就是*one stage or two stage detectors*.从一般经验上来说，*one stage detector*肯定比*two stage detector*要快，但是精度比后者高。但是这是一个定性的结论，在应用到具体问题的时候，往往不能提供具体的定量帮助。

本文其实也没有一个定量的帮助，但是好在它收集的数据足够多，画出图来也可以有一个比较直观的感觉。但是本文对比的效率是运行效率，也就是在做test的时候的效率而非训练一个网络的效率。总之，本文解决的问题是，不同的网络结构的精度与平均处理一张图的时间的比较。

为了方便起见，我将把论文拆解为以下2个问题并做报告：

1. 本文使用了哪些结构进行对比
2. 本文对比了哪几组数据，得出了什么结论
3. 本文适合有什么需求的人仔细阅读

本文使用了哪些结构进行对比

在主体部分，本文将网络结构分为3块：**Meta Architecture**,**Feature Extractor** and **Hyper Parameter**.

Meta Architecture

所谓**Meta Architecture**就是现在用于Object Detection的不同模型结构。作者将现有的模型结构分为三类：**Faster R-CNN**,**R-FCN**以及**SSD**。作者为什么要这么分呢？这是因为**Faster R-CNN**是*two stage detector*的代表模型，**SSD**是*one stage detector*的代表模型，而**R-FCN**虽然也是二阶段模型，但是它是在此基础上对**Faster R-CNN**系列的一个重要改进，即使得网络所生成的特征全部复用，不留单独的全连接层的一个模型，因此可以认为是*two stage*模型中偏向于速度的一个模型。作者把这个模型拿出来应该是想做一个中间量。

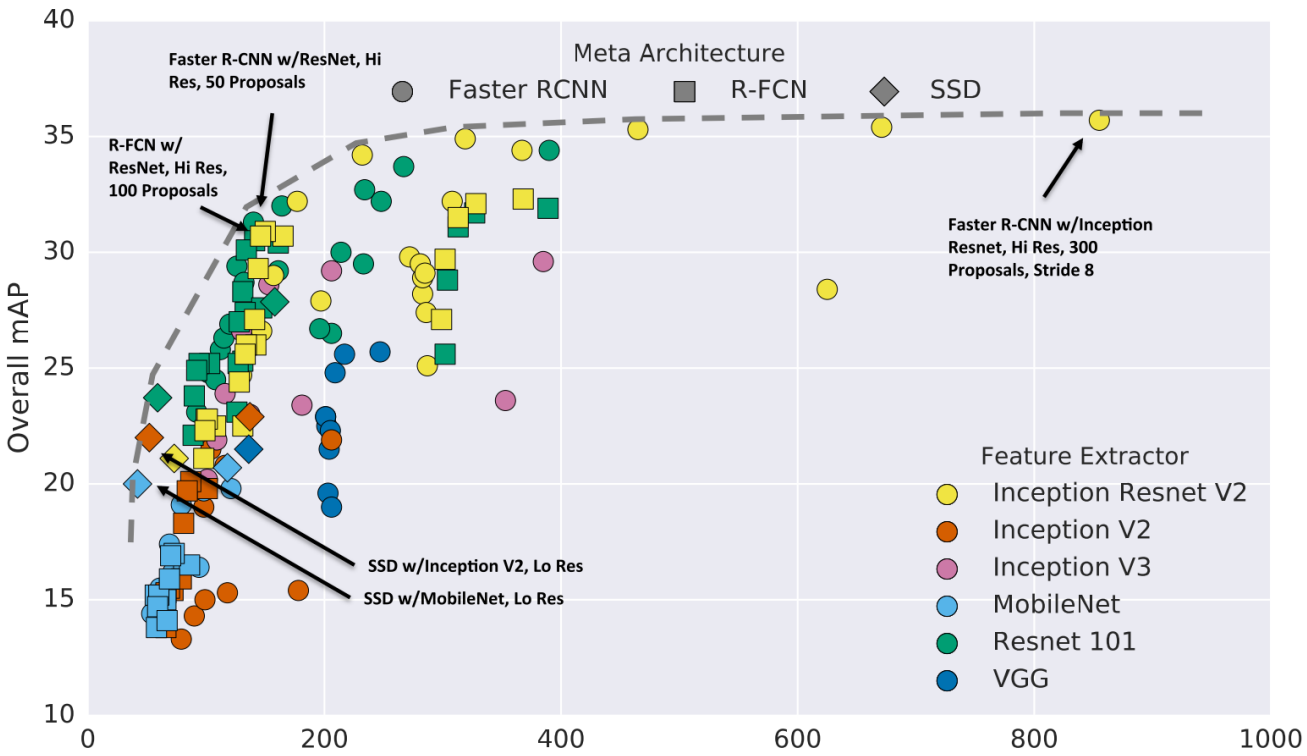
Feature Extractor

所谓**Feature Extractor**就是网络的特征提取部分，也就是head部分。作者选取了6个具有代表性的结构，分别为**Inception V2**,**Inception V3**,**VGG**,**Resnet101**,**Inception ResNet V2**,**MobileNet**.

那么作者为什么要从这么多网络结构中选出这6种网络结构呢？从CNN卷积神经网络架构综述中我们可以看出一些端倪。网络结构的发展有2个方向，一个是向更深的层数发展，代表就是**VGG**系列以及最终的**ResNet**,另一个方向是调整卷积核的大小，研究不同卷积核之间搭配的关系，代表就是**Inception**系列，而**GoogleNet**则是该系列的一个标杆。因此我们可以看出作者选取了2个方向的各2种网络，然后用**Inception ResNet V2**来对两种网络做一个优点融合(这种优点融合自然是很花计算量的)，最后作者再使用一个轻量级的网络**MobileNet**来代表牺牲精度换时间的一类网络，因此就构成了这6种网络结构。

Hyper Parameter

作者同时注意到了超参数的作用。这里的超参数是一个广义的概念，包括输入图像的大小，卷积核的stride，两阶段网络所propose的区域数目等。以上三个方面自由组合就构成了以下这张图：



本文对比了哪几组数据，得出了什么结论

其实从以上图中我们也可以直接看出一些端倪。具体而言，作者在以下9个方面做了一些介绍。

Accuracy vs time

Figure2的意思是横坐标是处理一张图片所需要的时间，而纵坐标则是mAP。每一个点代表着Meta Architecture与Feature Extractor的一个组合，很多相同的点其实背后是不一样的超参数。从图中我们可以得出的结论是R-FCN与SSD模型在平均意义上要快得多，但是Faster R-CNN旨在得到更慢但是更精确的模型。

同时，在这张图上，我们发现了一些有趣的点（用黑色箭头标出来的5个点）。这5个点的特点是，如果在此基础上需要提高一些精度，就要花很多很多的时间为代价，也就是说这5个点是性价比的一个边界点。

通过对边界点的研究，我们可知，SSD Net在需要更快的速度的时候也可以达到相对于其它结构更高的精度，而同时MobileNet也是最快的顶层架构。但是我们可以发现MobileNet在任何结构之下都没办法更进一步，它的精确度与速度始终处于左下角。同时，R-FCN在精确度上在前期可以与Faster R-CNN并驾齐驱，但是它的极限非常明显，也很容易达到。Faster R-CNN是唯一一个能够不断攀升更高准确度的网络。

同时我们在Feature Extractor上也可以看出，Resnet是平衡精度与速度的比较好的网络，Inception Resnet是速度很慢但是精确度潜力最大的网络，而MobileNet是速精比最高最快的网络。其它网络都表现不太好。

The effect of object size

注意到COCO数据集上关于mAP的标准有对大物体的检测，小物体的检测等等，我们可以看出，对于大物体而言，大部分都有很好的结果，但是SSD对于小物体检测，或者很多网络对于小物体检测都GG。

The effect of image size

输入图像的分辨率可以显著影响检测效果。作者给出的一个实验结果是，减少分辨率到原来的一半可以减少大概27.4%的时间，但是会损失原来的15.88%的精度。同时，加强精度对于小物体检测可以有很高的提升（一般是把精度扩大2倍），当然大物体也有。

The effect of the number proposals

对于二阶段网络，一个显著减少时间的方法就是减少propose的区域数目。作者发现，大大减少ROI数目不会很影响mAP，并给出了一个相对于300个ROI的一个好的数值，50个ROI，依然能达到原来96%的精度，但是减少了大概4倍时间。

Some other interesting results

这篇文章还研究了关于内存，浮点运算的一些内容，同时还有一个重要的结果是，在$IOU = 0.75$的时候所取得的mAP很高即意味着在所有IOU下都会取得好的结果，因此这里可以只计算一个有代表性的.75即可。

同时，本文还做出的一个贡献是，将以上训练出来的所有模型做了一个AdaBoost，然后超越了现在所有的算法，在COCO数据集上取得了非常好的成果。

这篇论文面向的对象

这篇论文在研究突破上有点水，但是确实是一份宝贵的经验手册。它里面详细叙述了各种超参数选取与可能的结果，同时对每个模型如何训练给出了详细的描述，可以作为一个模型手册来用，也可以当作了解现在模型的综述性文献阅读。

但是一方面它的重点是test速度而不是训练速度，因此在训练速度方面没有什么宝贵的成果。