

三维动画与交互技术课程读书报告



题目 Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

作者姓名 姜峰

作者学号 21851057

指导教师 李启雷

学科专业 软件工程

所在学院 软件学院

提交日期 二○一八年十二月

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Qilei Li

By

Feng Jiang

Zhejiang University, P.R. China

2018.12

摘要

现如今最先进的目标检测网路都是基于区域提议算法去进行目标定位的。像SPPnet和Fast R-CNN虽然减少了检测网络的运行时间，但依旧没能摆脱区域提议计算时间的瓶颈。为此，我们提出了一种区域提议网络(RPN)，旨在通过使检测网络与特征提取卷积网络共享参数来使得计算区域提议的时间大大减少。RPN能够在每个位置上同时预测目标的socres和boxes的卷积神经网络。RPN可以经过端到端的训练生成高质量的候选区域，之后被用于Fast R-CNN做检测。我们通过共享卷积层的参数来将RPN与Fast R-CNN融合到一个网络中。用最近流行的“注意力机制”的术语来说，RPN会告诉你该往哪里去看。在很深的网络VGG16上，我们的检测系统在GPU上依旧有5fps的速率(包括所有步骤)。同时，我们也在PASCAL VOC 2007，2012比赛上达到了最先进的目标检测精确度，在MS COCO上我们的检测网络对于每一张图片仅仅只有300个候选区域。在ILSVRC和COCO 2015比赛上Faster R-CNN在多项任务上均稳居第一。代码现已公开发布。

**关键词**：目标检测，区域提议网络，卷积神经网络。

Abstract

State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet and Fast R-CNN have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. In this work, we introduce a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. The RPN is trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. We further merge RPN and Fast R-CNN into a single network by sharing their convolutional features—using the recently popular terminology of neural networks with “attention” mechanisms, the RPN component tells the unified network where to look. For the very deep VGG-16 model, our detection system has a frame rate of 5fps (including all steps) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007, 2012, and MS COCO datasets with only 300 proposals per image. In ILSVRC and COCO 2015 competitions, Faster R-CNN and RPN are the foundations of the 1st-place winning entries in several tracks. Code has been made publicly available.

**Keywords：**Object Detection, Region Proposal, Convolutional Neural Network.

**1引言**

**1.1研究问题**

如何从图像中解析出可供计算机理解的信息，是机器视觉的中心问题。深度学习模型由于其强大的表示能力，加之数据量的积累和计算力的进步，成为机器视觉的热点研究方向。那么，如何理解一张图片？根据后续任务的需要，有三个主要的层次。一是分类（Classification），即是将图像结构化为某一类别的信息，用事先确定好的类别(string)或实例ID来描述图片。这一任务是最简单、最基础的图像理解任务，也是深度学习模型最先取得突破和实现大规模应用的任务。其中，ImageNet是最权威的评测集，每年的ILSVRC催生了大量的优秀深度网络结构，为其他任务提供了基础。在应用领域，人脸、场景的识别等都可以归为分类任务。二是检测（Detection）。分类任务关心整体，给出的是整张图片的内容描述，而检测则关注特定的物体目标，要求同时获得这一目标的类别信息和位置信息。相比分类，检测给出的是对图片前景和背景的理解，我们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）。三是分割（Segmentation）。分割包括语义分割（semantic segmentation）和实例分割（instance segmentation），前者是对前背景分离的拓展，要求分离开具有不同语义的图像部分，而后者是检测任务的拓展，要求描述出目标的轮廓（相比检测框更为精细）。分割是对图像的像素级描述，它赋予每个像素类别（实例）意义，适用于理解要求较高的场景，如无人驾驶中对道路和非道路的分割。本篇文章关注的领域是目标检测，即图像理解的中层次。

**1.2 目标检测研究背景**

目标检测对于人类来说并不困难，通过对图片中不同颜色模块的感知很容易定位并分类出其中目标物体，但对于计算机来说，面对的是RGB像素矩阵，很难从图像中直接得到狗和猫这样的抽象概念并定位其位置，再加上有时候多个物体和杂乱的背景混杂在一起，目标检测更加困难。但这难不倒科学家们，在传统视觉领域，目标检测就是一个非常热门的研究方向，一些特定目标的检测，比如人脸检测和行人检测已经有非常成熟的技术了。普通的目标检测也有过很多的尝试，但是效果总是差强人意。传统的目标检测一般使用滑动窗口的框架，主要包括三个步骤：1.利用不同尺寸的滑动窗口框住图中的某一部分作为候选区域；2.提取候选区域相关的视觉特征。比如人脸检测常用的Harr特征；行人检测和普通目标检测常用的HOG特征等；3.利用分类器进行识别，比如常用的SVM模型。

传统的目标检测中，多尺度形变部件模型DPM（Deformable Part Model）是出类拔萃的，连续获得VOC（Visual Object Class）2007到2009的检测冠军，2010年其作者Felzenszwalb Pedro被VOC授予”终身成就奖”。DPM把物体看成了多个组成的部件（比如人脸的鼻子、嘴巴等），用部件间的关系来描述物体，这个特性非常符合自然界很多物体的非刚体特征。DPM可以看做是HOG+SVM的扩展，很好的继承了两者的优点，在人脸检测、行人检测等任务上取得了不错的效果，但是DPM相对复杂，检测速度也较慢，从而也出现了很多改进的方法。正当大家热火朝天改进DPM性能的时候，基于深度学习的目标检测横空出世，迅速盖过了DPM的风头，很多之前研究传统目标检测算法的研究者也开始转向深度学习。基于深度学习的目标检测发展起来后，其实效果也一直难以突破。2013年R-CNN诞生了，VOC 2007测试集的mAP被提升至48%，2014年时通过修改网络结构又飙升到了66%，同时ILSVRC 2013测试集的mAP也被提升至31.4%。

R-CNN是Region-based Convolutional Neural Networks的缩写，中文翻译是基于区域的卷积神经网络，是一种结合候选区域（Region Proposal）和卷积神经网络（CNN）的目标检测方法。Ross Girshick在2013年的开山之作“Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”奠定了这个子领域的基础，这篇论文后续版本发表在CVPR 2014，期刊版本发表在PAMI 2015。其实在R-CNN之前已经有很多研究者尝试用Deep Learning的方法来做目标检测了，包括OverFeat，但R-CNN是第一个真正可以工业级应用的解决方案，这也和深度学习本身的发展类似，神经网络、卷积网络都不是什么新概念，但在本世纪突然真正变得可行，而一旦可行之后再迅猛发展也不足为奇了。

R-CNN这个领域目前研究非常活跃，先后出现了R-CNN、SPP-net、Fast R-CNN、Faster R-CNN、R-FCN、YOLO、SSD等研究。Ross Girshick作为这个领域的开山鼻祖总是神一样的存在，R-CNN、Fast R-CNN、Faster R-CNN、YOLO都和他有关。这些创新的工作其实很多时候是把一些传统视觉领域的方法和深度学习结合起来了，比如选择性搜索（Selective Search)和图像金字塔（Pyramid）等。

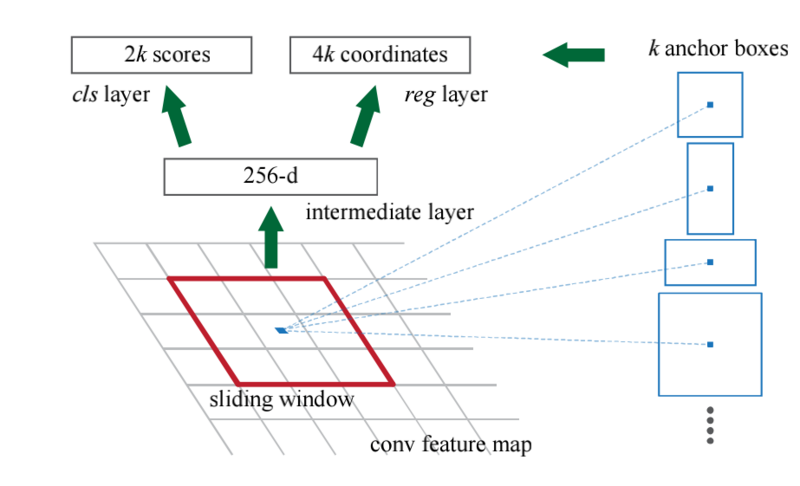
深度学习相关的目标检测方法也可以大致分为两派：1.基于候选区域的，如R-CNN、SPP-net、Fast R-CNN、Faster R-CNN、R-FCN；2.端到端（End-to-End），无需候选区域的，如YOLO、SSD。目前来说，基于候选区域的方法依然占据上风，但端到端的方法速度上优势明显，后续的发展拭目以待。

**2** **Faster R-CNN架构**

Faster R-CNN由两个模块组成。第一个模块是基于区域提议的深度全卷积网络，第二个模块是使用候选区域的Fast R-CNN检测器。整个系统是一个单个的，统一的目标检测网络。使用最近流行的“注意力”机制的神经网络术语，RPN模块告诉Fast R-CNN模块在哪里寻找。作者花了较大篇幅来介绍RPN网络。

**2.1 RPN网络**

RPN网络将一整幅图片（大小无所谓）作为输入，输出一些矩形框，且每一个都有背景或非背景得分，为了产生位置区域，作者在输出特征图的最后一层使用n\*n的子框从最后一层的卷积特征图上提取特征，之后将特征映射到256d的区域中，然后将这256d的数据送入的两个1\*1卷机层中分别用来判断这是否为背景的（cls）与边框回归（reg），在本文中作者设置n=3。过程大致如下：



**2.2 Anchors**

在窗口滑动的过程中，作者同时预测了多个区域 proposals，对于每个边框最大可能的形状有k种，因此CLS有2k个输出（即该形状的参考边框是否是背景），reg有4k个输出（及对该边框如何进行回归），同一个location的k种proposals，作者称之为anchor，而它可以通过中心点已经缩放度和长宽比来定义，通常使用3种缩放比例和3种长宽比来默认初始化一个anchor，那么整个特征图就有W\*H\*K个anchor了。

**2.3 Translation-Invariant Anchors**

在每个滑窗位置，同时预测k个region proposals，因此，reg层有4k个输出，以编码k个边界框坐标；cls层输出2k个scores，以估计每个proposal的object/non-object的概率。k个proposals相对于k个参考boxes进行参数化，记为anchors，其是一组矩形框。每个anchor在滑窗的中心，并与一个scale 和aspect ratio相关。这里采用3个scales和3个aspect ratios，在每个滑窗位置得到k=9个anchors。对于一个W×H(典型值∼2400)的conv feature map，会得到WHk个anchors。Faster R-CNN的一个重要属性是，对于anchors和计算相对于anchors的proposals的函数，都具有平移不变性。对比而言，MultiBox 方法采用k-means生成800个anchors，但不具有平移不变性。如果，平移图片中的一个object，对应的proposal也应该进行平移；相同的函数应该能预测该proposal。由于MultiBox anchors不具有平移不变性，其需要(4+1)×800维的输出层，而Faster R-CNN需要(4+2)×9维输出层。Faster R-CNN 参数更少，在小数据集上过拟合的风险更低。

**2.4 RPN训练及损失函数**

RPN 有两种类型的预测值输出：二值分类和边界框回归调整。训练时，对所有的anchors分类为两种类别。与ground-truth object 边界框的Intersection over Union(IoU)大于0.5的anchors作为foreground；小于0.1的作为background。然后，随机采样 anchors来生成batchsize=256的mini-batch，尽可能的保持foreground和background anchors的比例平衡。

RPN对mini-batch内的所有anchors采用binary cross entropy来计算分类loss. 然后，只对mini-batch内标记为foreground 的anchros计算回归loss。为了计算回归的目标targets，根据foreground anchor和其最接近的groundtruth object，计算将 anchor变换到object groundtruth的偏移值Δ。Faster R-CNN没有采用简单的L1或L2 loss用于回归误差，而是采用Smooth L1 loss。Smooth L1和L1基本相同，但是，当L1误差值非常小时，表示为一个确定值σ，即认为是接近正确的，loss就会以更快的速度消失。

采用动态batches是很有挑战性的。即使已经尝试保持background和foreground的anchors的平衡比例，也不总是可行的。根据图片中 groundtruth objects 和anchors的尺度与比例，很有可能得不到foreground anchors。这种情况时，将采用与groundtruth boxes具有最大IoU的anchors。这与理想情况相差很远，但实际中一般总能有foreground样本和要学习目标。

**2.5 RPN后处理-非极大值抑制(Non-maximum suppression)**

由于Anchors一般是有重叠的overlap，因此，相同object的proposals也存在重叠。为了解决重叠proposals问题，采用NMS算法处理，丢弃与一个score更高的proposal间IoU大于预设阈值的proposals。虽然NMS看起来比较简单，但IoU阈值的预设需要谨慎处理。如果 IoU值太小，可能丢失objects的一些proposals；如果IoU值过大，可能会导致objects出现很多proposals。IoU典型值为0.6。Proposal选择NMS处理后，根据score对topN个proposals排序。在Faster R-CNN论文中N=2000，其值也可以小一点，如50，仍然能的高好的结果。

**2.6 ROI pooling**

RPN 处理后，可以得到一堆没有class score的object proposals。待处理问题为，如何利用这些边界框bounding boxes，并分类。

一种最简单的方法是，对每个porposal，裁剪，并送入pre-trained base网络，提取特征；然后，将提取特征来训练分类器。但，这就需要对所有的2000个proposals进行计算，效率低，速度慢。Faster R-CNN则通过重用卷积特征图(conv feature map)来加快计算效率。即，采用RoI(region of interest)Pooling对每个proposal提取固定尺寸的特征图。R-CNN是对固定尺寸的特征图分类。

**2.7 RPN消融实验**

RPN上的消融实验。为了研究RPN作为区域建议的效果，作者进行了几项消融研究。首先，作者展示了在RPN和Fast R-CNN检测网络之间共享卷积层的效果。为此，作者在四步训练过程的第二步之后停止。使用单独的网络将结果略微降低至58.7％（RPN + ZF，非共享）。作者观察到这是因为在第三步中，当使用detectortuned特性来调整RPN时，区域质量得到了改善。接下来，作者分析了RPN对训练Fast R-CNN检测网络的影响。为此，作者通过使用2000个SS区域建议和ZF网络来训练Fast R-CNN模型。作者通过更改测试时使用的建议区域来检测此检测器并评估检测mAP。在这些消融实验中，RPN不与检测器共享特征。在测试期间用300个RPN提议替换SS得到了56.8％的mAP。mAP中的损失是因为训练/测试区域建议不一致。该结果作为以下比较的基准。有点令人惊讶的是，RPN在使用排名最高的前100个区域进行测试的时候仍然会得到一个很不错的结果（55.1％），这表明排名靠前的RPN区域建议是准确的。另一方面，使用排名最高的6000个RPN区域建议（无NMS）也可以得到很不错的mAP（55.2％），这表明NMS不会损害检测mAP并可能减少误报。接下来，作者通过在测试时关闭RPN的cls和reg输出来分别研究其作用。当在测试时移除cls层（因此不使用NMS /ranking）时，作者从未评分的区域中随机抽样N个区域建议。当 N=1000时的mAP（55.8％）几乎没有变化，但当N=100时降低到44.6％。这显示了cls分数能够体现排名靠前的区域建议的准确性。另一方面，当在测试时去除reg层（因此区域建议就是锚矩形框）时，mAP下降到52.1％。这表明高质量的建议区域主要是由于回归框限制。锚矩形框虽然具有多种比例尺和纵横比，但不足以进行准确的检测。作者还评估了更强大的网络对RPN区域建议质量的影响。作者使用VGG-16来训练RPN，并仍然使用SS+ZF的上述检测器。mAP从56.8％（使用RPN+ZF）提高到59.2％（使用RPN+VGG）。

**2.8 Faster R-CNN训练**

Faster R-CNN的训练，是在已经训练好的model（如VGG\_CNN\_M\_1024，VGG，ZF）的基础上继续进行训练。实际中训练过程分为6个步骤：

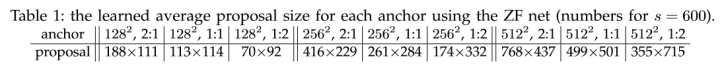
1. 在已经训练好的model上，训练RPN网络
2. 利用步骤1中训练好的RPN网络，收集proposals
3. 第一次训练Fast RCNN网络
4. 第二训练RPN网络
5. 再次利用步骤4中训练好的RPN网络，收集proposals
6. 第二次训练Fast RCNN网络

可以看到训练过程类似于一种“迭代”的过程，不过只循环了2次。至于只循环了2次的原因是应为作者提到："A similar alternating training can be run for more iterations, but we have observed negligible improvements"，即循环更多次没有提升了。

**3. 实验结果**

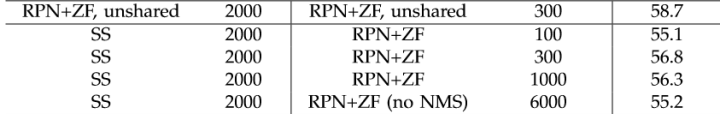
文章对于图像的预处理是放缩使得短边为600个像素，以VGG为例，经过四次的池化操作，每个卷积层像素的感受野对应于原图为16个像素。对于1000\*600的图片，其卷积特征层尺寸为60\*40，所以anchors数量为60\*40\*9，而去除在边缘的不完整的anchors，最后能得到6000个anchors，而这些anchors的特征是同时计算出来的。比如针对一个图像尺寸128\*128，其对应于卷积层特征尺寸为9\*9，然后对于所有的9\*9小块进行卷积得到的feature map就是所有的anchor对应的特征。RPN产生的region proposal重叠较多，通过NMS去除重叠后大约剩余2K个proposal，在测试时仅选用前300个proposal，效果也是相当的好。

不同尺度的anchors对应的真实的proposal大小：



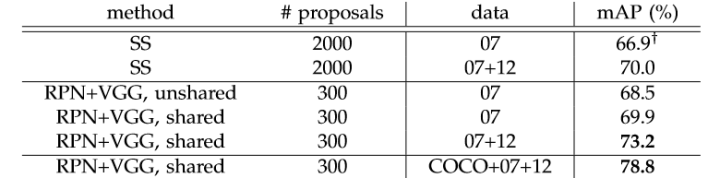
可以看到，我们的尺寸和比例还是起到了作用的，相应尺寸和比例的proposal也基本符合要求的。

proposal个数：



可以看到300个proposal其实并没有降低识别准确率，而且NMS对结果也没啥影响，所以测试时取前300个proposal，这样可以大幅提高速度。

分阶段权重共享的优点：



权重共享大约能够提升个1.5个百分点，还是很有效果的。同时我们也看到数据集扩充对于识别结果影响最大，通过COCO+07+12训练然后再07上测试能够达到78.8%，提高10个点。

时间性能：

