RPN网络的目的是代替传统的selective search方法更好更快的产生region proposals，它的输入是原始的任意尺寸的图像，输出是许多region proposals。RPN网络的前部分结构和faster rcnn的特征提取结构是共享的，最后又有一个3*3卷积层（全连接层）和两个1*1卷积层（全连接层）。RPN网络也可以理解为只包含这两个卷积层，它的输入是特征提取结构输出的feature map。

对于每一个3*3的滑动窗口（即3*3卷积层）而言，对应到原始图像中是一个很大的局部感受野。那么怎样在每一个这种很大的感受野中产生良好的proposals呢?

RPN的思路是在每一个原图感受野中产生若干个anchors，这些anchors大小和宽高比都不相同，由于这些anchors选取的比较粗糙（既有包含前景的anchors也有包含背景的anchors，并且这些anchors还不能很好地与gt bboxes拟合），因此需要进一步对这些anchors进行过滤（排除掉背景置信度较高的anchors以及前景置信度较低的anchors）和回归（对保留下的前景置信度较高的anchors微调以更好地拟合gt bboxes），然后将这些处理过后的anchors作为proposals输出给接下来的ROIpooling层。因此可以看出，RPN网络产生的region proposals就是回归后的置信度较高的前景anchors。sliding windows就是RPN网络的输入卷积层。其实经过RPN网络处理后已经完成了图像检测的步骤，接下来的ROIpooling和Fast RCNN等网络结构完成图像识别的步骤。

具体地，RPN网络的3*3卷积层在feature map上滑动卷积时，将滑动窗口的中心位置映射回原始图像，以该原始图像位置为中心点，在原始图像上生成3种大小和3种宽高比共9个anchors，这9个anchors的中心位置重合。如果feature map大小为W*H*C，那么总共会生成W*H*9个anchors。也就是说，RPN网络在feature map的每个局部感受野（3*3窗口）上通过anchors学习区分前景特征和背景特征。

Mark Tang

每个anchor框对应到原图还是最后feature map上呢?

赞 回复 踩 举报

闲散人 (作者) 回复 Mark Tang

Anchor是在特征图上抽取的，要映射回原图计算iou

为什么不直接预测坐标而是预测偏移量？

我们的目标是寻找图片中的边框。这些边框是不同尺寸、不同比例的矩形。设想我们在解决问题前已知图片中有两个目标。那么首先想到的应该是训练一个网络，这个网络可以返回 8 个值：包含（xmin, ymin, xmax, ymax）的两个元组，每个元组都用于定义一个目标的边框坐标。这个方法有着根本问题，例如，图片可能是不同尺寸和比例的，因此训练一个可以直接准确预测原始坐标的模型是很复杂的。另一个问题是无效预测：当预测（xmin,xmax）和（ymin,ymax）时，应该强制设定 xmin 要小于 xmax，ymin 要小于 ymax。

另一种更加简单的方法是去预测参考边框的偏移量。使用参考边框（xcenter, ycenter, width, height），学习**预测偏移量（Δxcenter,Δycenter,Δwidth,Δheight）**，因此我们只得到一些小数值的预测结果并挪动参考变量就可以达到更好的拟合结果。

YOLO9000：

YOLO一代包含有全连接层，从而能直接预测Bounding Boxes的坐标值。 Faster R-CNN的方法只用卷积层与Region Proposal Network来预测Anchor Box的偏移值与置信度，而不是直接预测坐标值。作者发现通过预测偏移量而不是坐标值能够简化问题，让神经网络学习起来更容易。