上海理工大學

研究生课程（论文类）试卷

2016 / 2017 学年第 1 学期

课程名称：图像处理与分析

课程代码：12010048

论文题目：Mask R-CNN

学生姓名：沈天马

专业,学号：173800801

学　院：光电信息与计算机工程学院

课程（论文）成绩：
课程（论文）评分依据（必填）： 任课教师签字： 日期：　　年　　月　　日

摘　要

我们组主要负责 computer vision 中 Segmentation 这一块，我之前三个组员负责讲解了传统方法和 2014 年提出的 FCN 全卷积神经。所以我承接以上内容讲解目前最新的 Instance Segmentation 的最新模型——Mask R-CNN。这篇文章不仅荣获今年 2017 的 best paper，而且在 Segmentation 和 objective detection 的工业界应用（亚马逊，谷歌，Facebook 等）发生了全新的变化。因此不难看出 Mask R-CNN 比以前算法的优越度提高了不少。正如之前所示，Mask R-CNN 涉及到机器视觉两个领域 Segmentation 和 objective detection。所以本人在介绍 Mask R-CNN 的同时，顺带介绍一下 R-CNN（objective detection）从 2013 年到如今的演变。（Segmentation 由我们组员顾天飞介绍过了）

从 2013 年到 2017 年，R-CNN 的发展史上有几个最为关键的里程碑：R-CNN（2013），SPP-net(2015.4), Fast R-CNN(2015.5), Faster R-CNN(2016.1), YOLO(2016.3), SSD(2016.12), Mask R-CNN(2017.4)。因为本次报告本人用 Latex 写的，所以封面格式会有点变动望老师理解。

关键字：　Mask R-CNN　Segmentation　detection

目录

一 R-CNN 的发展

1.1 R-CNN（2013）

1.1.1 Introduction of R-CNN（2013）

R-CNN 的结构如图 1.1 所示（所有结构图都是本人手动画制，并非来自论文），也是第一次运用卷积神经网络后战胜了传统机器视觉算法（objective detection）。不难发现，R-CNN 所有的结构与当下主流的 objective detection 相差很大，但是对于当时的 computer vision 的算法来说已经是 CNN 的突破性应用了。

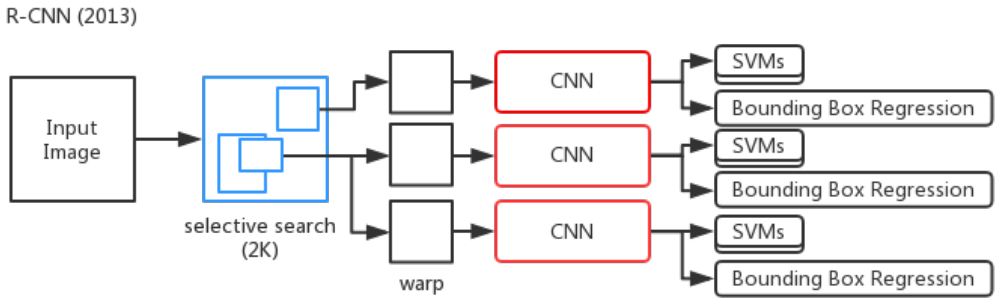

图 1.1: R-CNN (2013)

R-CNN 的缺点也是非常明显（针对于当下的算法）：

- region proposal 的算法还是基于 computer vision 的 selective search。
- 每一张图片都有将近 2K 个预选区域，分别经过 CNN 时间开销非常庞大。
- warp 强制转换图片的大小丢失细节。
- 采用 SVM 来分类，而针对于多个类别的时候性能远差于 softmax。

1.1.2 selective search

selective search，简称为 SS。是 computer vision 里 region proposal 的算法，原理是通过扫描临近的像素通过阈值的设定，来确定是否邻近的像素点是同一类。最终实现将原图分割成一块一块。这个思想和原理主要依据的是，针对于同一个物体来说颜色应该是比较相近的。

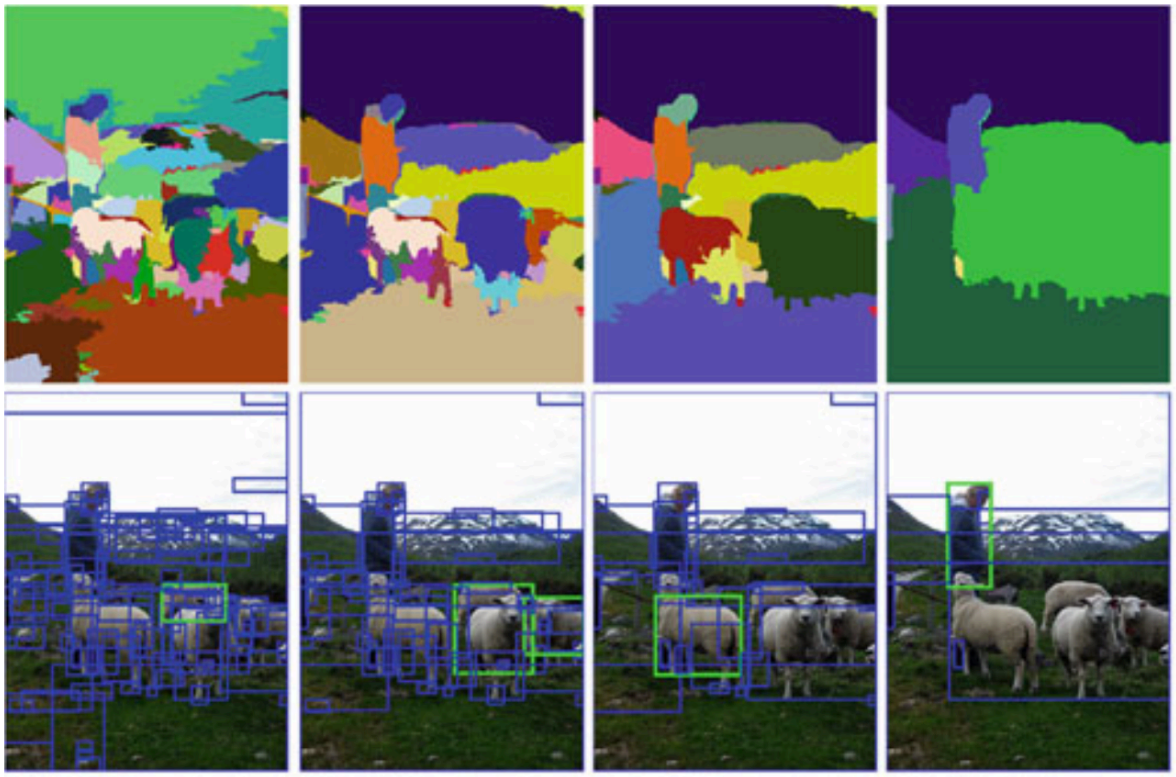

图 1.2: Selective Search

1.2 SPP-net（2015.4）

1.2.1 Introduction of SPP-net（2015.4）

其实对于 R-CNN 时间开销最大一处就是对于每一张图片都要经过相同 CNN 中，所以后人将想到 CNN 的共享。因为对于同一个图片针对同一个卷积和的结果是很大一部分冗余，完全可以合并后在对 feature maps 进行 region proposal 的处理。（feature maps 就是卷积后的输出，因为卷积后基本提取出图片的特征，所以取这个名字）SPP-net 就是修改了上述的这一点，并且不光如此，它还提出了取消 warp，通过 SPP 层 (Spatial Pyramid Pooling) 来

做到 FC 连接层的维度统一问题。判别分类的地方也采用了 softmax 改进。

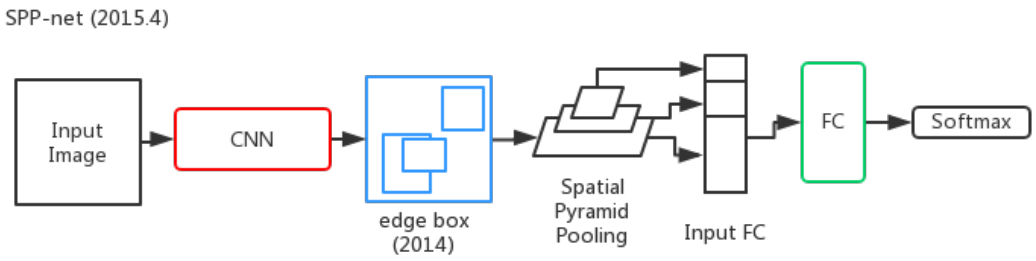

图 1.3: SPP-net（2015.4）

1.2.2 edge box

SPP-net 采用了全新一种 region proposal 的算法——edge box。这算法思想是基于所有特征物体的启发式起点应该在物体的轮廓，针对于边缘轮廓的提取之后进行聚类。最终效果图如 1.4 所示。

1.3 Fast R-CNN（2015.5）

Fast R-CNN 在 region proposal 的算法选择上重新选择了 selective search，在 paper 中作者对比了主流的方法，最终作者选择在所有数据集效果均值最好的算法（作者称为算法的稳定性）即 SS（selective search）。如图 1.5 所示，我将 SPP-net 和 Fast R-CNN 的结构放在一起，能更加容易的发现区别。在原先 SPP 层 (Spatial Pyramid Pooling)，作者将其改成 ROI（only one Pyramid level）来代替原先的三层，并且加入 bounding box regression 来更好的得到矩形框。

1.4 Faster R-CNN（2016.1）

Faster R-CNN 可以算 R-CNN 这一领域新的篇章，因为之前的算法无论如何都必须通过传统 computer vision 的算法来实现 region proposal。这就导致在处理 region proposal 这算

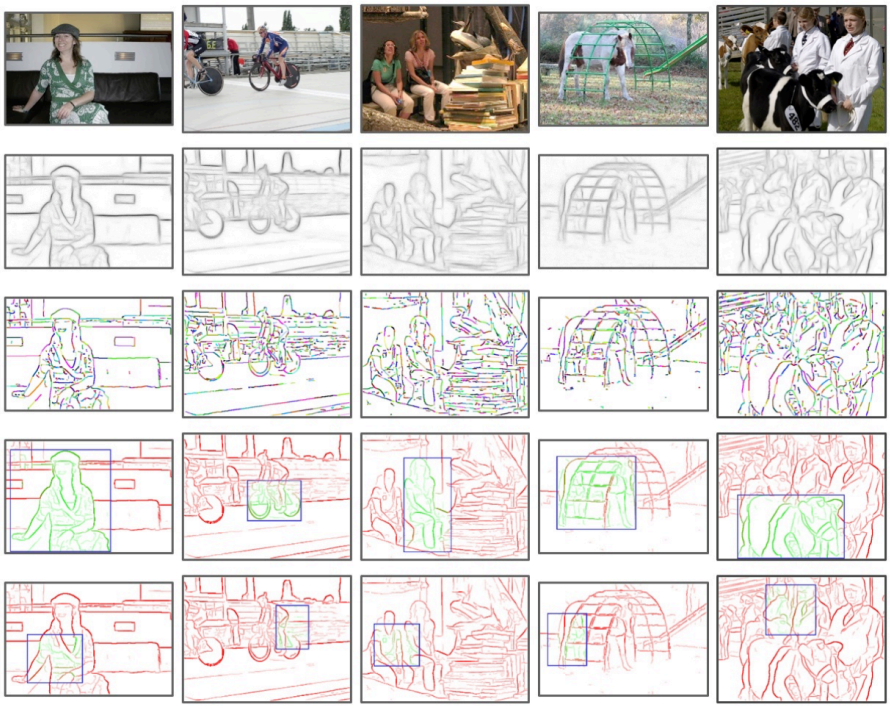

图 1.4: Edge Box

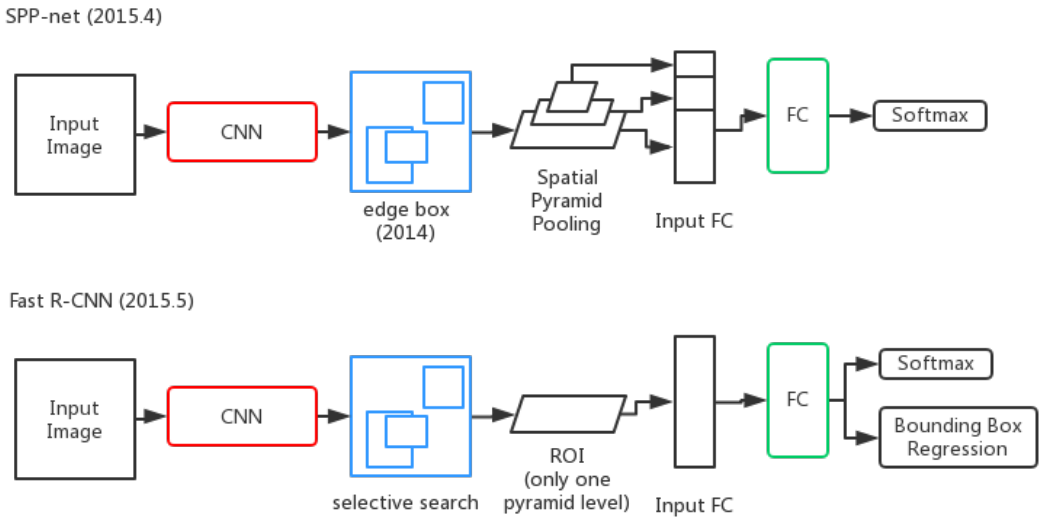

图 1.5: Fast R-CNN（2015.5）

法执行无法和 CNN 一起放入 GPU 训练，使得训练时间太慢（针对于现在，之前的 region proposal 的算法都是在 CPU 上执行，速度远低于 GPU）。

因此 Faster R-CNN 提出 region proposal 也可以通过神经网络来代替，作者取名为 RPN（region proposal network）。在 RPN 中的 softmax 并不是类别的分类器，它只是分类是否是自己感兴趣的物体（物体与背景的二分类）。针对 CNN 输出的 feature maps，作者提出 anchor 的感念。因为如果想扫描 feature maps 的每一块区域，我们有两种做法：1、同一大小的窗口滑动，改变输入图片的大小。2、不改变输入图片的大小，改变滑动窗口的比例和大小。很明显实验结果倾向于第二种，即作者提出 anchor（就是不同大小尺寸的窗口）。在选择 anchor 尺寸上，在 paper 中是将数据集的窗口进行聚类从而选择合适的。

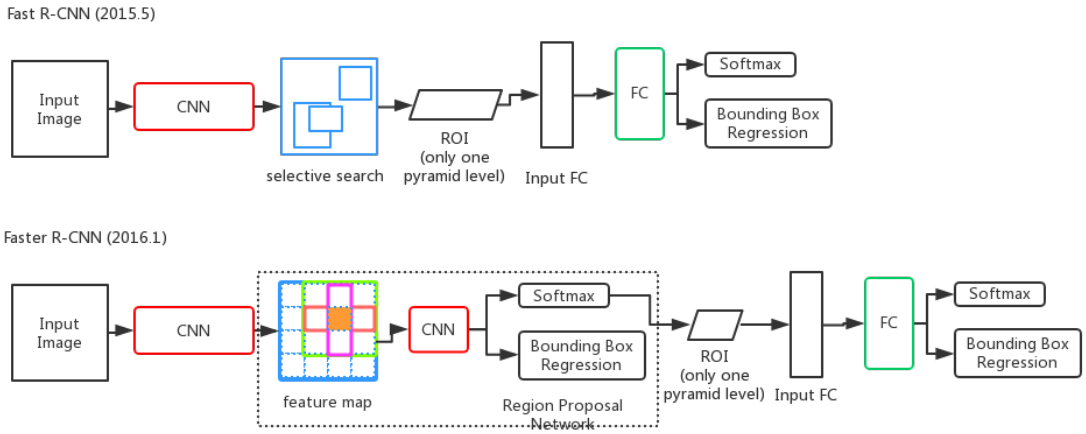

图 1.6: Faster R-CNN（2016.1）

1.5 YOLO(2016.3) and SSD(2016.12)

在 2016 年期间，虽然 Faster R-CNN 已经取得很好的效果，但是是否还能进一步加快，或者牺牲一部分准确度提高大幅度的速度。这个目的为出发的论文就孕育而生——YOLO（you only look once）和 SSD（single shoot detetion）。

YOLO 的思想其实很简单，通过 CNN 将 feature maps 映射到原始图像，即将原始图近似看成分割成块状。这样对于 feature map 每个像素点直接进入全连接 FC，来输出：BOOL（判别是否有物体）、bounding box 的位置、类别的分类。速度效果很明显，但是缺点也很致命:1、因为 feature maps 像素低，所以对于细小，重叠的物体来说无法检测 2、最后使用 FC 代价太大。所以随后就出现 YOLO-v2 和 YOLO9000，一部分是结构近似于 SSD，改进了内部优化，如 batch normalization 等；另一部分 word tree 数据集融合的体现。（YOLO-v2 和

9000 就一笔带过了，写起来太多了）

SSD 的研究和 YOLO 其实是并行的，他的想法是将 softmaax 合二为一，这样就可以直接将 Faster R-CNN 的后半部分融合在一起。并且为了能检测更小的物体特征，将第一次得到的 feature map 在此循环进入此结构（看代码 paper 只执行了 3 次），从而在提高速度的同时也保证了精度。

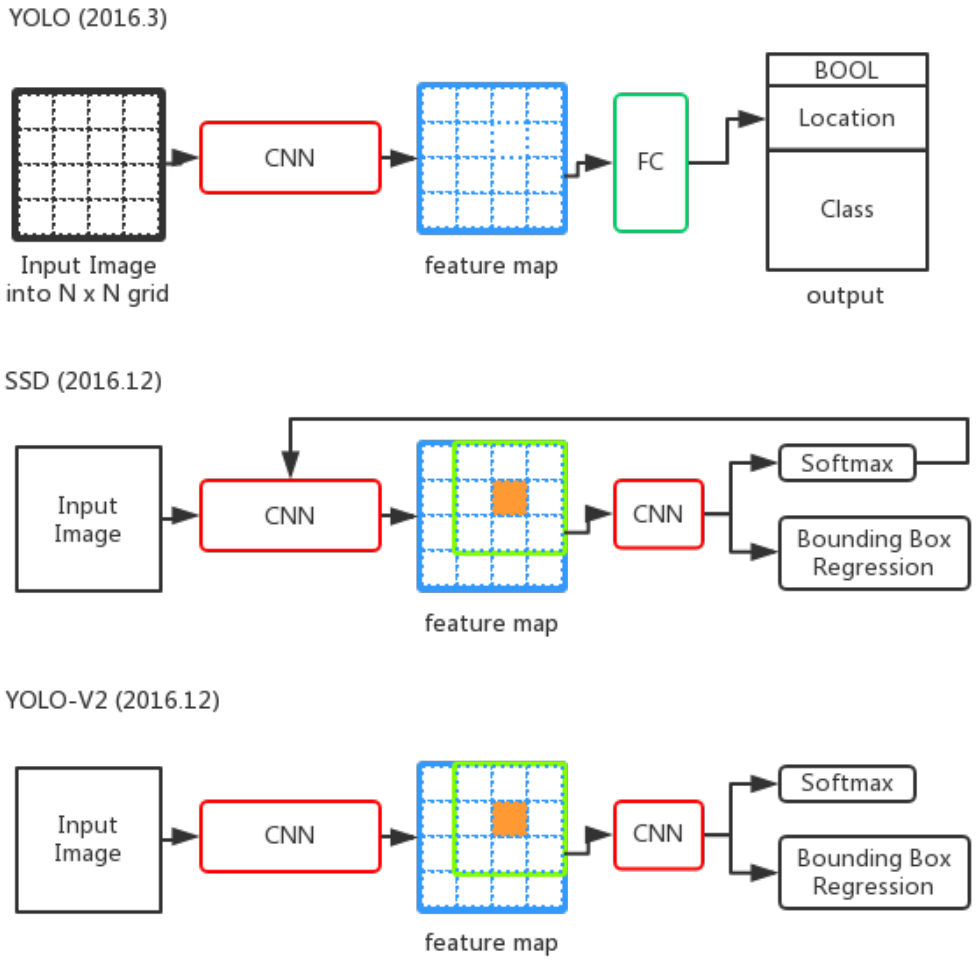

图 1.7: YOLO(2016.3)，SSD(2016.12) and YOLO-V2（2016.12）

二 Mask R-CNN

2.1 Introduction of Mask R-CNN（2017.4）

Mask R-CNN 从结构上看不是特别复杂，而且从算法的突破性也不是很大。但是他是第一篇将 objective detection 和 Segmentation 将结合在一起的论文。从图 2.1 所示，不难看出，Mask R-CNN 的结构就是在 Faster R-CNN 上加入了 FCN 来提高模型的精确度。

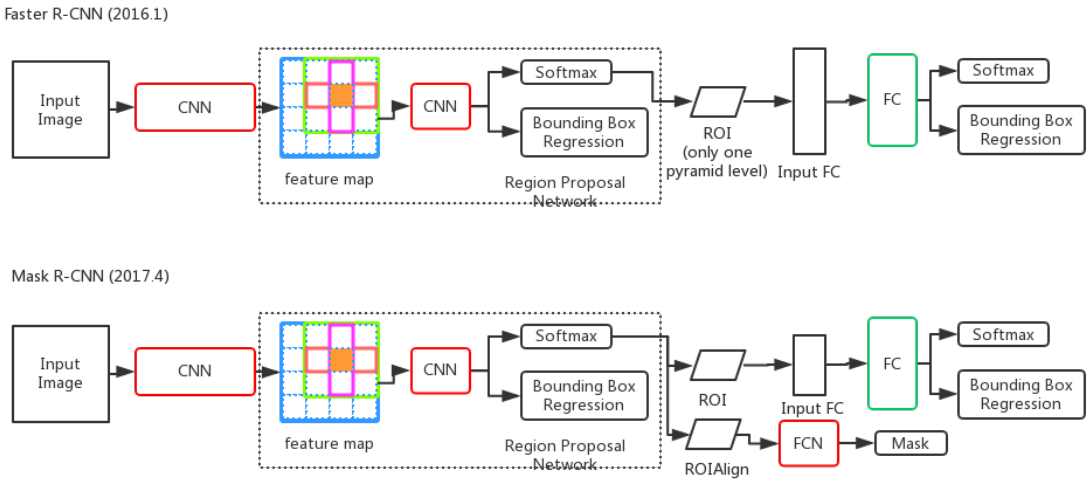

图 2.1: Mask R-CNN（2017.4）and Faster R-CNN（2016.1）

2.2 ROIAlign

这篇 paper 也用一种很讨巧的方法，来改进 FCN 得到的 mask 和原始图像像素直接的偏差。论文采用双向性插值的方法（先前是 bounding box 无需很高的精度，所以直接是按比例取整的）表达式如下。

$$f(x,y) = \frac{1}{(x_2-x_1)(y_2-y_1)} \begin{pmatrix} x_2-x & x-x_1 \end{pmatrix}$$

$$\begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} y_2 - y \\ y - y_1 \end{pmatrix}$$

另外就是用到了并行 ROI 的结构，来实现 FCN 的输入。

2.3 FCN

这部分我们组员顾天飞已经讲解，我就不多描述了。

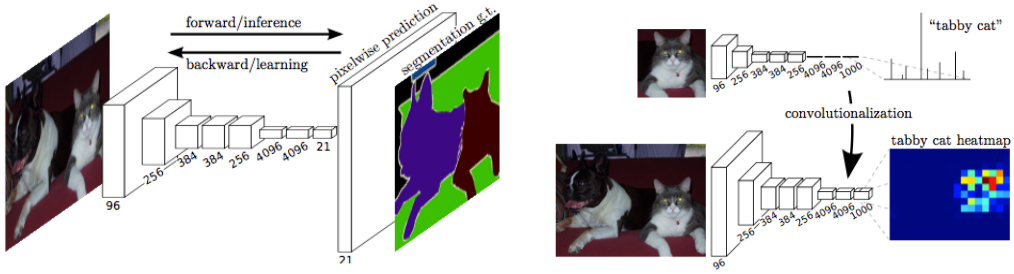

图 2.2: FCN: from image to pixels