Region-based CNN (R-CNN)

Ref: [https://arxiv.org/abs/1311.2524]

1. Main Steps

- 生成候选框图

  R-CNN 使用了 Selective Search 的方法进行bounding box的生成，这是一种Region Proposal的方法。
  生成的2000个Bbox使用NMS计算IoU指标剔除重叠的位置。

- 针对候选图作embedding的抽取
  Bbox直接Resize为227*227供AlexNet的输入，再Resize之前对所有BBox进行padding。

- 使用分类器对embedding训练和分类

2. Training

使用TL，在VOC数据集上进行fine-tune。原始ImageNet上训练的网络能预测1000类，这里采用了20类加背景一共21类的输出方式。

3. IoU Threshold

IoU的threshold在本文被设置为0.3，如果一个区域与ground truth的IoU低于0.3，这个区域被视作Negative。

4. Hard Negative Mining

*Hard Negative Mining*和 *Hard Negative Example*：

- *Hard Negative Example*：由于根据IoU生成的bbox正样本远远少于负样本，可以IoU<0.1的样本为负样本或者使用随机抽样使正负样本比为1:3
- *Hard Negative Mining*：指一种训练手段： 在bootstrapping中，首先使用初始的较小的正负样本集训练一个分类器，随后将负样本中的错误分类的样本(hard negative)放入负样本集继续训练分类器。

5. Bounding Box Regression

当输入的Proposal box和Ground truth的IoU较大时($IoU>0.6$)，可以认为二者之间存在线性变换。这里BBox Reg即给定输入的BBox特征向量(x,y,w,h)，使用y=Wx学习到的W来使P框能接近G框。

给定的学习的变换形式为：

$$
\begin{aligned}
G_x &= P_x + P_w d_x(P) \\
G_y &= P_y + P_h d_y(P) \\
G_w &= P_w * e^{d_w(P)} \\
G_h &= P_h * e^{d_h(P)}
\end{aligned}
$$

但是在R-CNN中，实际上不是使用的框的坐标进行回归，而是使用pool5层的输出作为feature：

$$d_i = w_i^T \phi_{5i}$$

$$loss = \sum_i^N (t_i - w_i^T \phi_{5i})^2 + \lambda \|w_i\|^2$$

Fast R-CNN

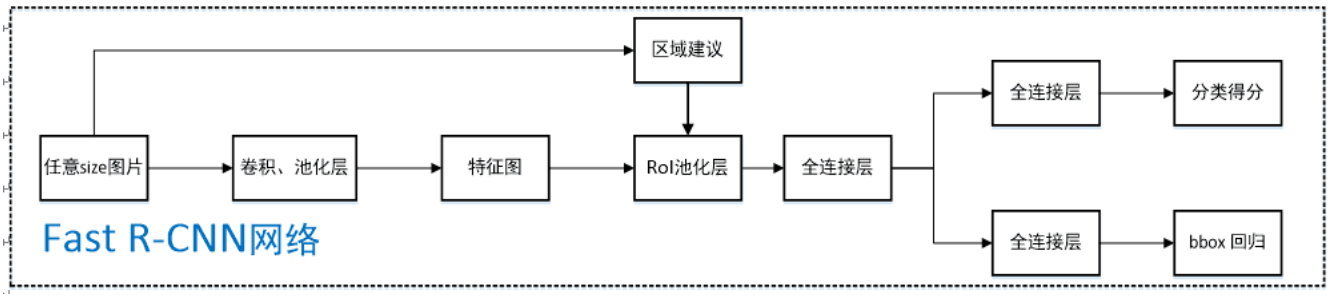
Ref: [https://arxiv.org/abs/1504.08083]

1. Highlights

- R-CNN中用CNN对每一个BBox反复提取特征，而2000个Bbox之间有大量重叠，造成算力的浪费。
- Fast R-CNN提出将目标分类Classification和Bbox Regression统一，形成Multi-task模型。

2. Main Steps

如图所示：



首先将原始图像通过conv extractor (本文使用了VGGNet) 得到一个Conv Feature map，并且将所有的Proposal Bbox 映射到这张feature map上。

使用ROI Pooling层使h x w 大小的RoI窗口降为H x W大小的小的feature map。

3. Region of Interest (ROI)

RoI pooling layer 将所有备选BBox通过划分为$H X W$的网络，对每一个网络做Max Pooling，
转换为统一大小的feaeture map (e.g. $H x W$)。