Note of Object Detection Paper Reading

Region-based CNN (R-CNN)

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

Main Steps

1. 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。

2. 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*。

3. 使用分类器对 *embedding* 训练和分类

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。

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*。

Hard Negative Mining

首先请区分 Hard Negative Mining 和 Hard Negative Example 的概念：
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 $1:3$。
而 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。

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：

$$G_x = P_x + P_w d_x(P)$$

$$G_y = P_y + P_h d_y(P)$$

$$G_w = P_w * e^{d_w(P)}$$

$$G_h = P_h * e^{d_h(P)}$$

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 ϕ_5 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]

Main Steps

如图所示 [Structure of Fast R-CNN](“./image/fastrcnn.png”,“fast R-CNN”)

Region of Interest (ROI)

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。