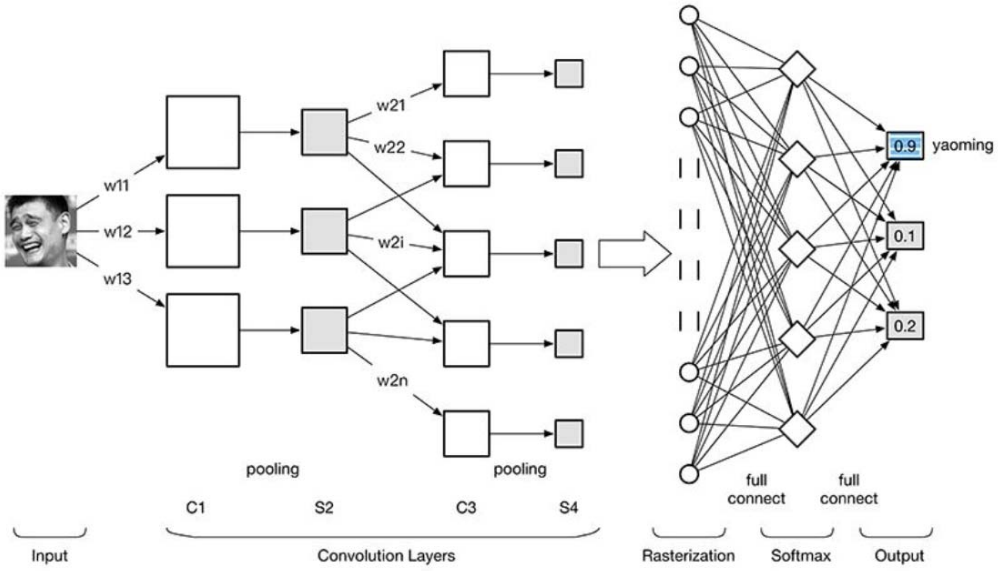
2018 年秋-第 8 周

本周主要看了 Faster-RCNN 系列以及 SSD 的一些资料，整理了一些有助于理解的部分。

CNN 基础



由于以前用 matlab 跑过 alex，所以简单摘录了一些。

卷积神经网络 ConvNet 可以分为 4 大层：

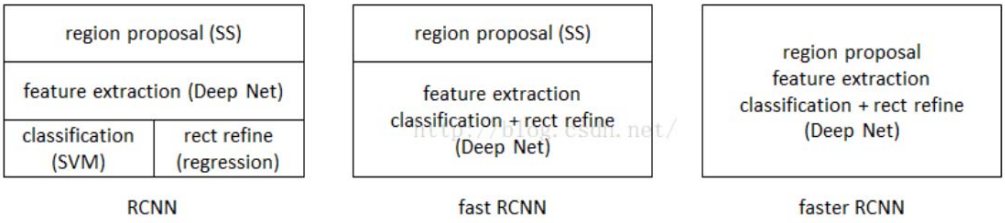
1. 图像输入 Image Input：为了计算方便一般需要归一化。

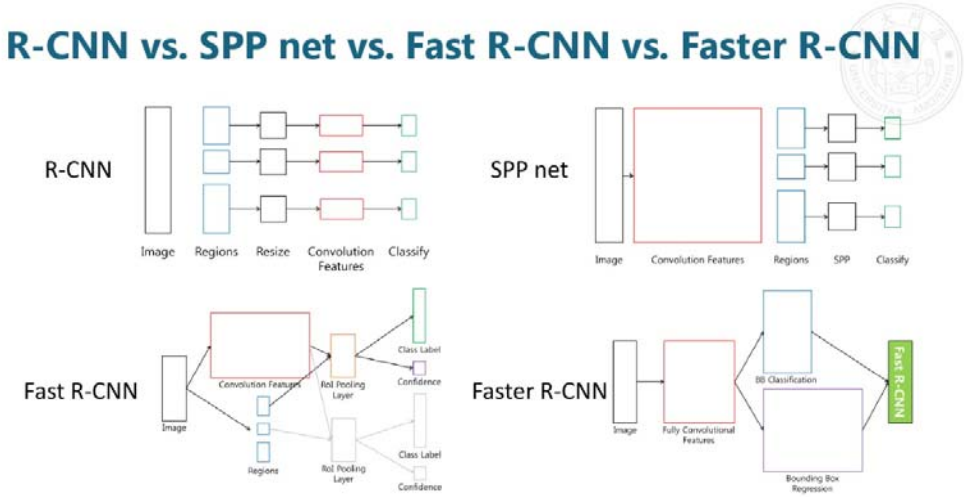
2. 卷积层(Convolution Layer)：特征提取层(C 层) - 特征映射层(S 层)。将上一层的输出图像与本层卷积核(权重参数 w)加权值，加偏置，通过一个 Sigmoid 函数得到各个 C 层，然后下采样 subsampling 得到各个 S 层。C 层和 S 层的输出称为 Feature Map(特征图)。

3. 光栅化(Rasterization)：为了与传统的多层感知器 MLP 全连接，把上一层的所有 Feature Map 的每个像素依次展开，排成一列。

4. 多层感知器(MLP)：最后一层为分类器，一般使用 Softmax，如果是二分类，当然也可以使用线性回归 Logistic Regression，SVM，RBM。（疑问：样本）

R-CNN、Fast R-CNN、Faster R-CNN 三者关系：



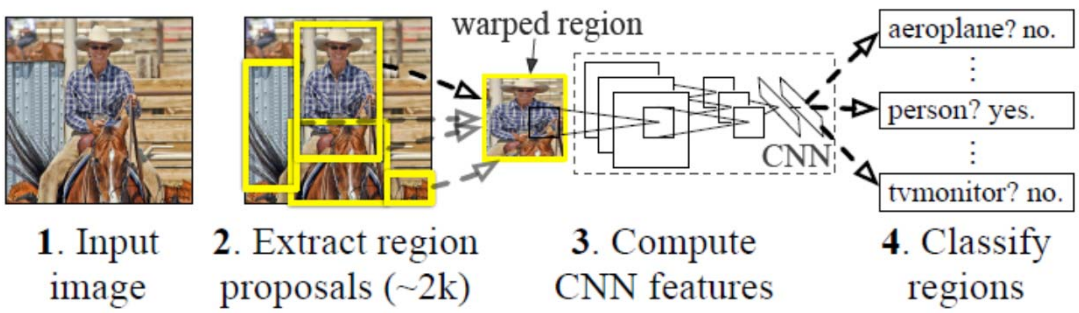


RCNN 解决的是：用 CNN 做 classification

Fast R-CNN 解决的是：一起输出 bounding box 和 label

Faster R-CNN 解决的是：去掉 selective search

RCNN



RCNN 的检测算法是基于传统方法来找出一些可能是物体的区域，再把该区域的尺寸归一化成卷积网络输入的尺寸，最后判断该区域到底是不是物体，是哪个物体，以及对是物体的区域进行进一步回归的微微调整学习，使得框的更加准确。

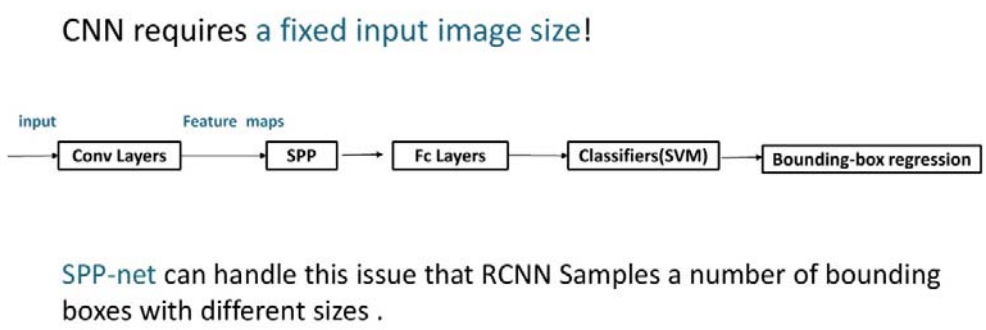
可以把 RCNN 看成四个部分：ss 提 proposals，深度网络提特征，训练分类器，训练对应回归器。

特别的：

（分类器） 对每一类目标，使用一个线性 SVM 二类分类器进行判别。输入为深度网络输出的 4096 维特征,输出是否属于此类。由于负样本很多,使用 hard negative mining 方法。

（回归器） 对每一类目标，使用一个回归器进行精修。正则项 λ=10000。输入为深度网络 pool5 层的 4096 维特征，输出为 xy 方向的缩放和平移。

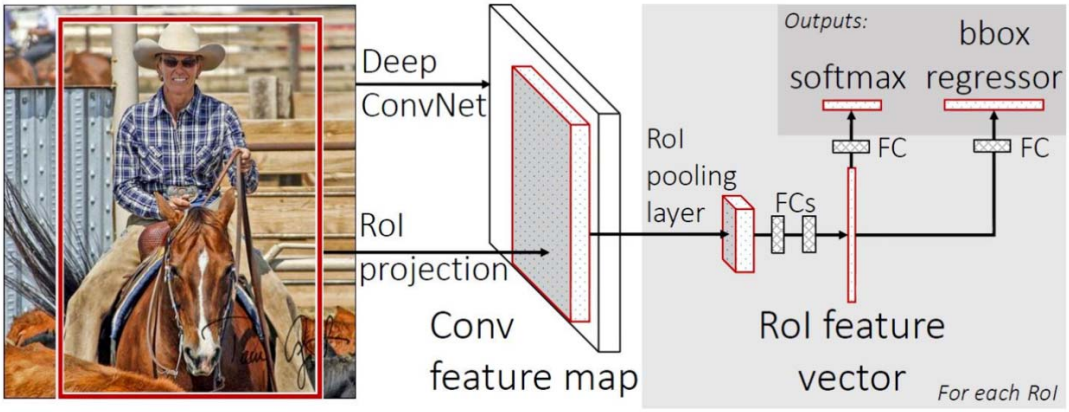
SPP



SPP 网络主要是解决深度网络固定输入层尺寸的这个限制，用不同尺度的 pooling 来 pooling 出固定尺度大小的 feature map，这样就可以不受全链接层约束任意更改输入尺度了。

SPP 网络的核心思想：通过对 feature map 进行相应尺度的 pooling，使得能 pooling 出 4×4, 2×2, 1×1 的 feature map，再将这些 feature map concat 成列向量与下一层全链接层相连。这样就消除了输入尺度不一致的影响。

Fast RCNN



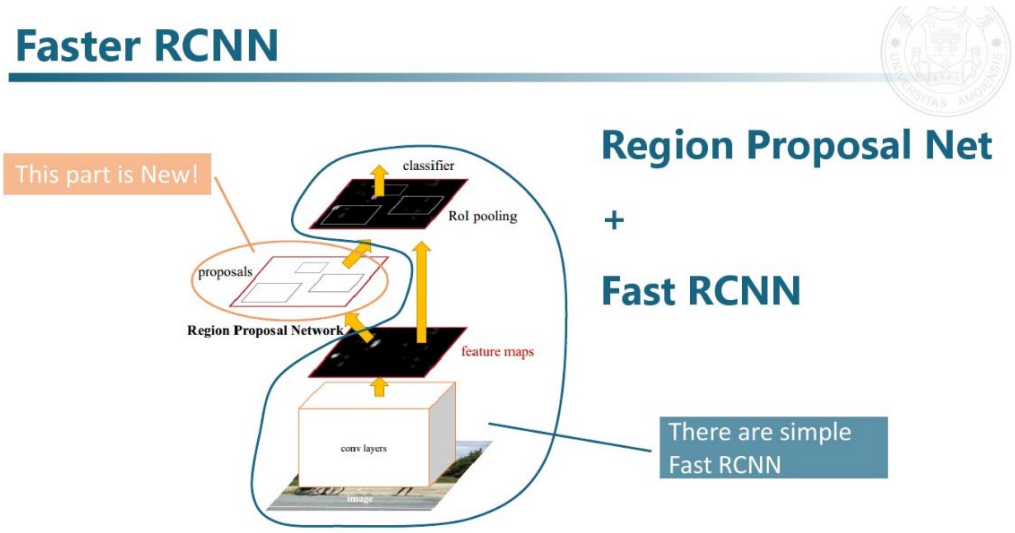
提出了一个特殊的层 RoI，这个实际上是 SPP 的变种，SPP 是 pooling 成多个固定尺度，而 RoI 只 pooling 到一个固定的尺度（6×6）

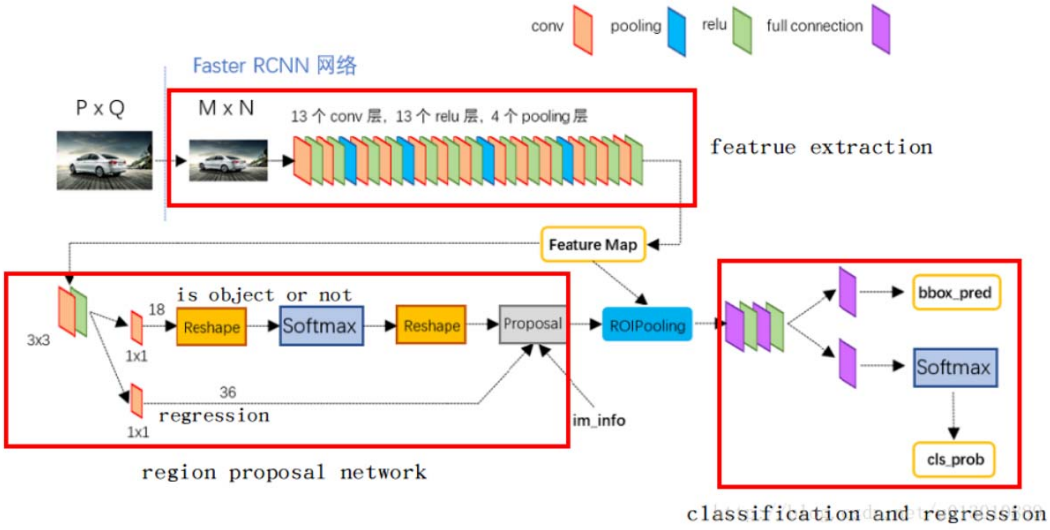
网络结构与之前的深度分类网络（alex）结构类似，不过把 pooling5 层换成了 RoI 层，并把最后一层的 Softmax 换成两个，一个是对区域的分类 Softmax（包括背景），另一个是对 bounding box 的微调。

这个网络有两个输入，一个是整张图片，另一个是候选 proposals 算法产生的可能 proposals 的坐标。

Fast R-CNN 提取建议区域的方法依然是 select search

Faster RCNN



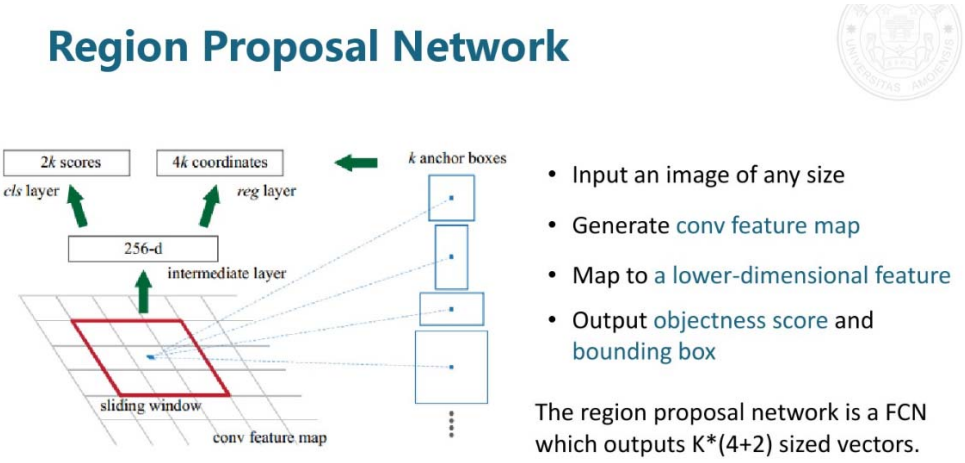


主要解决两个问题：

1、提出区域建议网络 RPN，快速生成候选区域；

2、通过交替训练，使 RPN 和 Fast-RCNN 网络共享参数。

RPN 网络：



anchor 机制

需要确定每个滑窗中心对应感受野内存在目标与否。由于目标大小和长宽比例不一，需要多个尺度的窗。Anchor 即给出一个基准窗大小，按照倍数和长宽比例得到不同大小的窗。

RPN 网络训练

1、假如某 anchor 与任一目标区域的 IoU 最大，则该 anchor 判定为有目标；

2、假如某 anchor 与任一目标区域的 IoU>0.7，则判定为有目标；

3、假如某 anchor 与任一目标区域的 IoU<0.3，则判定为背景。

所谓 IoU，就是预测 box 和真实 box 的覆盖率，其值等于两个 box 的交集除以两个 box 的并集。其它的 anchor 不参与训练。

联合四步训练法：

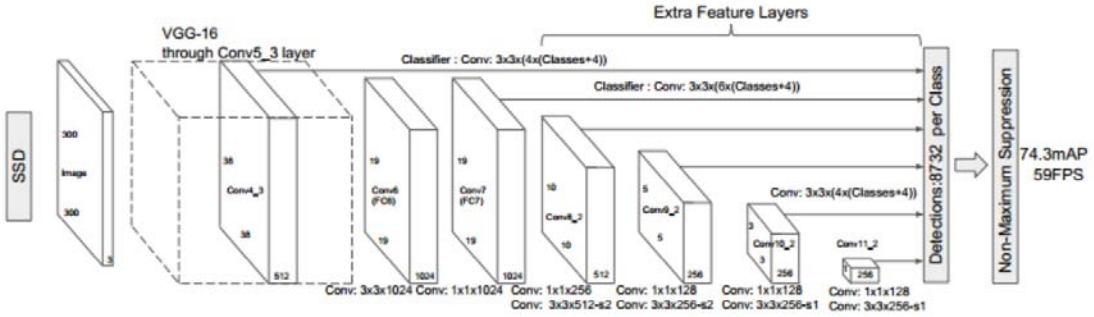
1、 单独训练 RPN 网络，网络参数由预训练模型载入；

2、 单独训练 Fast-RCNN 网络，将第一步 RPN 的输出候选区域作为检测网络的输入。具体而言，RPN 输出一个候选框，通过候选框截取原图像，并将截取后的图像通过几次 conv-pool，然后再通过 roi-pooling 和 fc 再输出两条支路，一条是目标分类 softmax，另一条是 bbox 回归。

3、 再次训练 RPN，此时固定网络公共部分的参数，只更新 RPN 独有部分的参数；

4、 那 RPN 的结果再次微调 Fast-RCNN 网络，固定网络公共部分的参数，只更新 Fast-RCNN 独有部分的参数。

SSD



SSD，其主要思路是均匀地在图片的不同位置进行密集抽样，抽样时可以采用不同尺度和长宽比，然后利用 CNN 提取特征后直接进行分类与回归

采用多尺度特征图用于检测

采用大小不同的特征图，CNN 网络一般前面的特征图比较大，后面会逐渐采用 stride=2 的卷积或者 pool 来降低特征图大小

- scale: 假定使用 m 个不同层的feature map 来做预测，最底层的 feature map 的 scale 值为 $s_{min}=0.2$ ，最高层的为 $s_{max}=0.95$ ，其他层通过下面公式计算得到
  $s_k = s_{min} + \frac{s_{max}-s_{min}}{m-1}(k-1), k \in [1,m]$
- ratio: 使用不同的 ratio值 $a_r \in \left\{1, 2, \frac{1}{2}, 3, \frac{1}{3}\right\}$ 计算 default box 的宽度和高度：
  $w_k^a = s_k\sqrt{a_r}$ ， $h_k^a = s_k/\sqrt{a_r}$。另外对于 ratio = 1 的情况，额外再指定 scale 为 $s_k' = \sqrt{s_k s_{k+1}}$ 也就是总共有 6 中不同的 default box。
- default box中心：上每个 default box的中心位置设置成 $(\frac{i+0.5}{|f_k|}, \frac{j+0.5}{|f_k|})$ ，其中 $|f_k|$ 表示第k个特征图的大小 $i,j \in [0,|f_k|)$ 。

采用卷积进行检测

SSD 直接采用卷积对不同的特征图来进行提取检测结果

设置先验框

每个单元设置尺度或者长宽比不同的先验框，预测的边界框（bounding boxes）是以这

些先验框为基准的，在一定程度上减少训练难度。一般情况下，每个单元会设置多个先验框。对于每个单元的每个先验框，其都输出一套独立的检测值，对应一个边界框，主要分为两个部分。第一部分是各个类别的置信度或者评分，第二部分就是边界框的 location

训练过程

Matching strategy：

如何将 groundtruth boxes 与 default boxes 进行配对，以组成 label 呢？

在开始的时候，用 MultiBox 中的 best jaccard overlap 来匹配每一个 ground truth box 与 default box，这样就能保证每一个 groundtruth box 与唯一的一个 default box 对应起来。

但是又不同于 MultiBox ，本文之后又将 default box 与任何的 groundtruth box 配对，只要两者之间的 jaccard overlap 大于一个阈值，这里本文的阈值为 0.5。

数据增强策略

- 使用整张图片
- 使用IOU和目标物体为0.1, 0.3，0.5, 0.7, 0.9的patch （这些 patch 在原图的大小的 [0.1,1] 之间， 相应的宽高比在[1/2,2]之间）
- 随机采取一个patch

当 ground truth box 的 中心（center）在采样的 patch 中时，我们保留重叠部分。在这些采样步骤之后，每一个采样的 patch 被 **resize** 到固定的大小，并且以 0.5 的概率随机的 水平翻转（horizontally flipped）。用数据增益通过实验证明，能够将数据mAP增加8.8%。