一、RCNN 的由来

Region CNN(RCNN)可以说是利用深度学习进行目标检测的开山之作。作者 Ross Girshick 多次在 PASCAL VOC 的目标检测竞赛中折桂，2010 年更带领团队获得终身成就奖，如今供职于 Facebook 旗下的 FAIR。 目前一系列目标检测算法：RCNN,Fast RCNN, Faster RCNN 代表当下目标检测的前沿水平，在 github 都给出了基于 Caffe 的源码。

二、RCNN 的功劳

RCNN 一举超过传统的目标检测方法，通过深度学习网络来学习目标特征代替传统的人为的设定学习特征（比如 HOG）省去很多麻烦的计算，另一方面预先选取一系列的可能是物体的候选区域代替传统的滑动窗口依次判断所有的可能的物体区域，极大的增加了目标检测的速度和正确率。论文发表的 2014 年，DPM 已经进入瓶颈期，即使使用复杂的特征和结构得到的提升也十分有限。本文将深度学习引入检测领域，一举将 PASCAL VOC 上的检测率从 35.1%提升到 53.7%

三、RCNN

1）准备工作

RCNN 将使用俩个数据库如下：

一个是较大的识别库(ImageNet ILSVC 2012):标定每张图片中物体的类别。一千万图像，1000 类。

一个是较小的检测库(PASCAL VOC 2007):标定每张图片中，物体的类别和位置。一万图像，20 类。

RCNN 采用识别库进行预训练，而后用检测库调优参数。最后在检测库上评测。

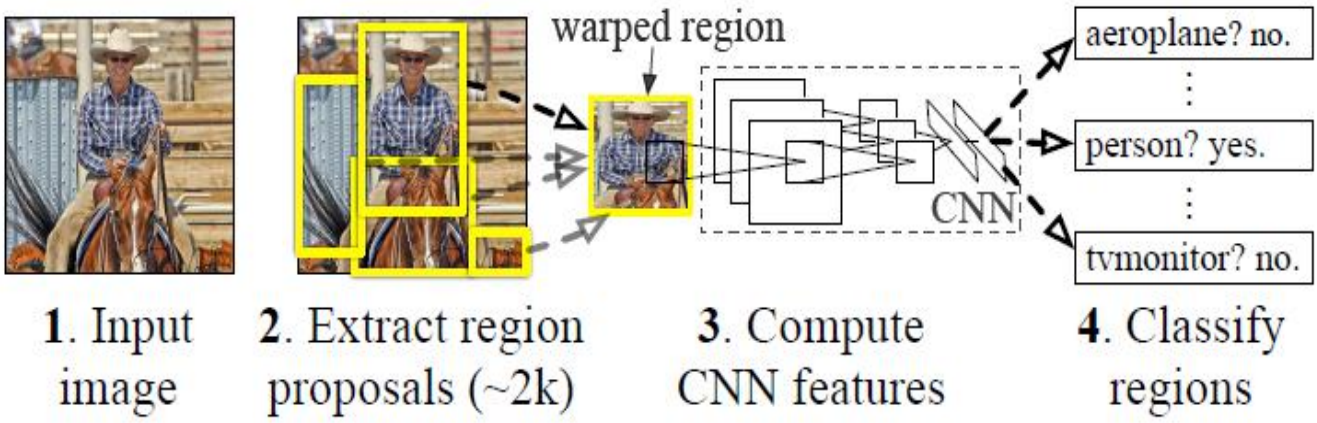
2）主要流程

1）将每张图片分成 2000～3000 个候选区域。

2）将每个候选区域通过 CNN 网络进行特征提取获得特征向量。

3）将获得的特征送人 svm 进行判别是否属于该类。

4）最后用回归器修正选框的位置。



3）候选区域的生成方法

候选区域的选择是相对独立的，一般来讲我们可以选择任意的方法进行选择，而传统的 RCNN 网络选用的是 Selective Search

方法，从一张图片中生成 2000~3000 个候选区域。提到这个方法让我想到了我曾经玩过的一个无聊的手机游戏，之所以无聊是因为他全程是将一个一个的小狗合成大狗，同样大小的够才能合成更高阶的狗，而正好和此方法很相似，我们使用一种过分割手段，将图像分割成小区域 ，查看现有小区域，合并可能性最高的两个区域，然后重复直到整张图像合并成一个区域位置 ，输出所有曾经存在过的区域即所谓候选区域。

不同的是游戏中合成的规则只有大小，而 Selective Search 的规则有以下几种：

- 颜色（颜色直方图）相近的
- 纹理（梯度直方图）相近的
- 合并后总面积小的
- 合并后，总面积在其 BBOX 中所占比例大的

-保证合并操作的尺度较为均匀，避免一个大区域陆续“吃掉”其他小区域。.

为尽可能不遗漏候选区域，上述操作在多个颜色空间中同时进行（RGB,HSV,Lab 等）。在一个颜色空间中，使用上述四条规则的不同组

合进行合并。所有颜色空间与所有规则的全部结果，在去除重复后，都作为候选区域输出。

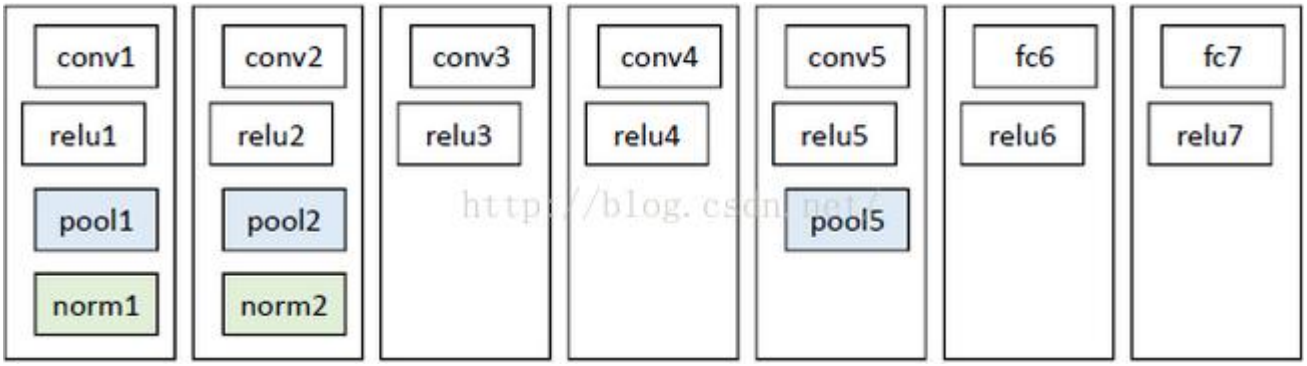
4）特征提取过程

正如经典的 CNN 网络一样，首先需要数据预处理，首先把候选区域归一化成同一尺寸 227×227。此处有一些细节可做变化：外扩的尺寸大小，形变时是否保持原比例，对框外区域直接截取还是补灰。会轻微影响性能。

接下来就是预训练：

网络结构

基本借鉴 Hinton 2012 年在 Image Net 上的分类网络，略作简化.



此网络提取的特征为 4096 维，之后送入一个 4096->1000 的全连接(fc)层进行分类。学习率 0.01。

训练数据

使用 ILVCR 2012 的全部数据进行训练，输入一张图片，输出 1000 维的类别标号。

调优训练

网络结构

同样使用上述网络，只不过将最后的全连接层输出数目由 1000 改为 21。

学习率 0.001，每一个 batch 包含 32 个正样本（属于 20 类）和 96 个背景。

训练数据

使用 PASCAL VOC 2007 的训练集，输入一张图片，输出 21 维的类别标号，表示 20 类+背景。

考察一个候选框和当前图像上所有标定框重叠面积最大的一个。然后考虑他们俩的重叠面积，如果重叠比例大于 0.5，则认为此候选框为此标定的类别；否则认为此候选框为背景。

5）类别判断

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。

正样本

本类的真值标定框。

负样本

考察每一个候选框，如果和本类所有标定框的重叠都小于 0.3，认定其为负样本

6）位置精修

目标检测问题的衡量标准是重叠面积 :许多看似准确的检测结果 ,往往因为候选框不够准确，重叠面积很小。故需要一个位置精修步骤。 回归器对每一类目标，使用一个线性脊回归器进行精修。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。将训练样本判定为本类的候选框修正为和真值重叠面积大于 0.6 的候选框。