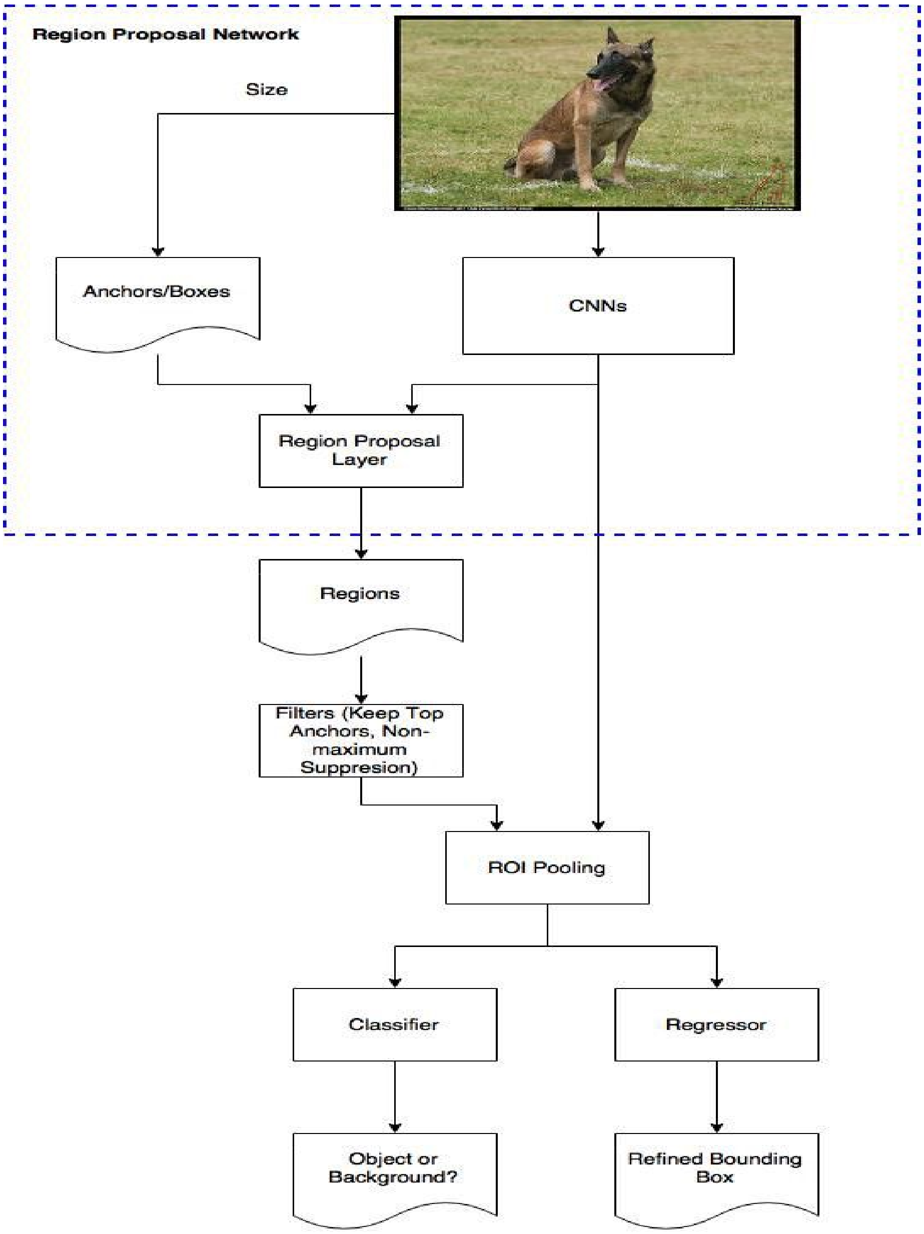
Faster R-CNN

Faster R-CNN has two network:

- RPN: region proposal network for generating region proposal
- A network using these proposals to detect objects.

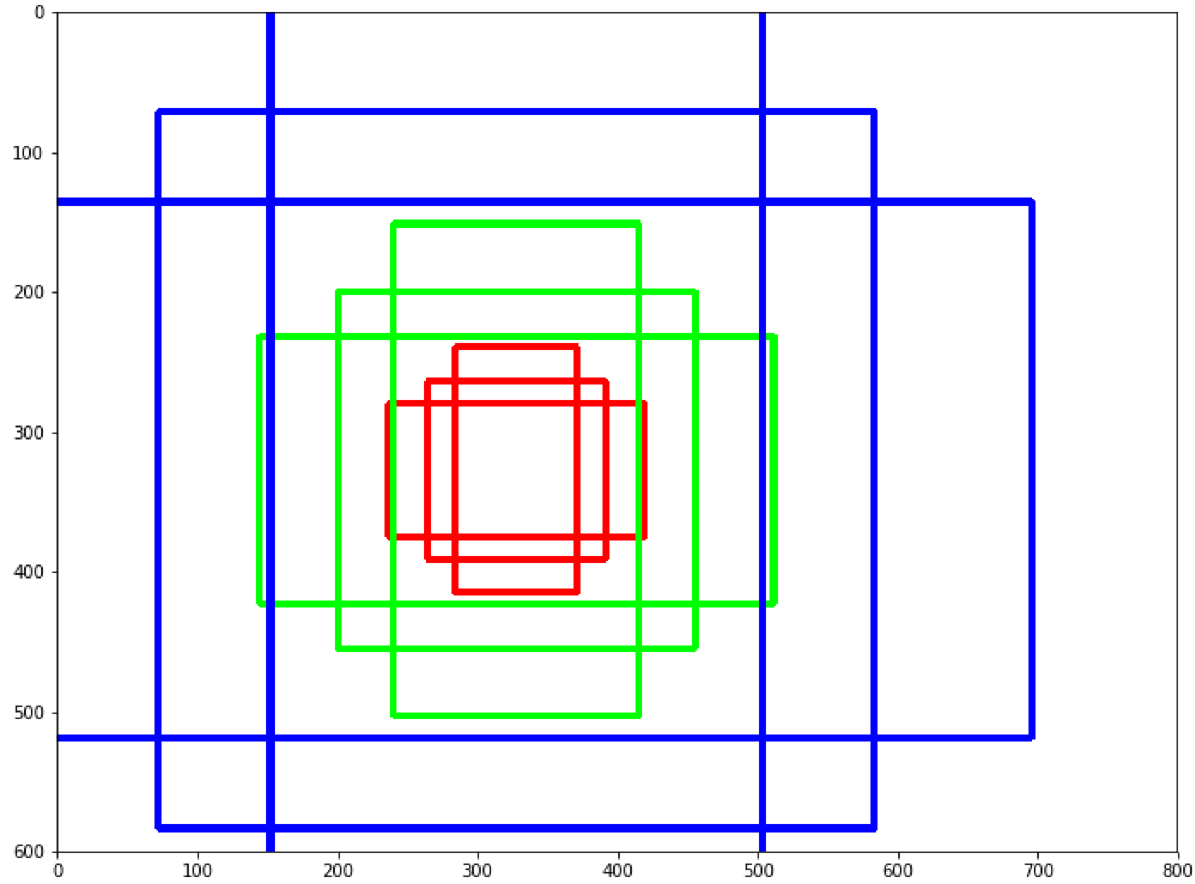
→ Briefly, RPN ranks region boxes (called anchors) and proposes the ones most likely containing objects.

There is the Fast R-CNN which uses selective search to generate region proposals. The time cost of generating region proposals is much smaller in RPN than selective search.



Anchors

Anchors play an important role in Faster R-CNN. An anchor is a box. In the default configuration of Faster R-CNN, there are 9 anchors at a position of an image. The following graph shows 9 anchors at the position (320, 320) of an image with size (600, 800).



1. Three colors represent three scales or sizes: 128x128, 256x256, 512x512.
2. For one color, the three boxes have height width ratios 1:1, 1:2 and 2:1 respectively.

Important comment:
→ You have the freedom to design different kinds of anchors/boxes. For example, you are designing a network to count passengers/pedestrians, you may not need to consider the very short, very big, or square boxes. A neat set of anchors may increase the speed as well as the accuracy.

Region proposal network:

RPN predicts the possibility of an anchor being background or foreground, and refine the anchor.

The output is a branch of anchors which will be examined by a classifier of background/foreground later.

The Classifier of Background and Foreground:

- Preparing a training dataset: Anchors + Images.
- Classification: the anchors with high overlaps as foreground, the anchors with low overlaps as background.
- Every position in the feature map has 9 anchors, and every anchor has two possible labels (background, foreground). If we make the depth of the feature map as 18 (9 anchors x 2 labels), we will make every anchor have a vector with two values (normal called logit) representing foreground and background.

ROI Pooling: (region of interest pooling)

We have different sized region after RPN which means different sized CNN feature maps.

→ Unlike Max-Pooling which has a fix size, ROI Pooling splits the input feature map into a fixed number (let’s say k) of roughly equal regions, and then apply Max-Pooling on every region. Therefore the output of ROI Pooling is always k regardless the size of input.

→ With the fixed ROI Pooling outputs as inputs, we have lots of choices for the architecture of the final classifier and regressor.