ConvNet

Ex. VGG16, ZF

Feature map

Proposals

Classifier

Class 1

Class 2

Class 3

Image
256x256

Feature
extractor
VGG16

Output stride = 16

Relu5_3
512x16x16

Last conv layer/
feature map

16x16 feature output
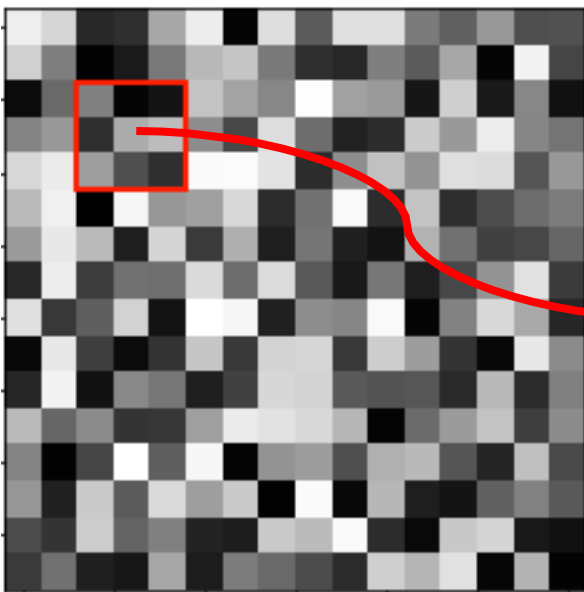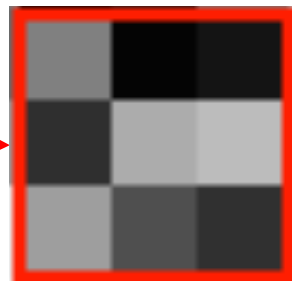
3x3 slidingwindow

512x3x3 feature map

3x3 conv

Output 512

1x1conv Object

Output 9x2

1x1conv Regre.

Output 9x4

16x16x9 proposals

Loss:
$$\frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*)$$
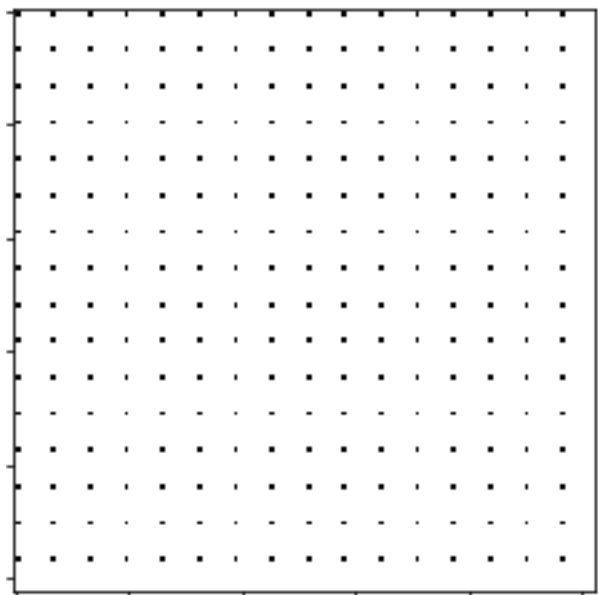
Proposal:
Objectscore
Boundary box coord

Loss:
$$\lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*)$$

Anchor centers in image

Each anchor=
3 scales
Ex. 10, 30, 50
3 aspect ratios
Ex 1:1, 1:2, 2:1

Anchors in image for 1 center

Relu5_3
512x16x16

Last conv layer/ feature map

>300 proposals

x, y, w, h

RoI pooling

Output: 512x7x7

Flatten

Output: 25088

fc6 4096

FC

fc7 4096

FC

FC Softmax

FC Regre.

Loss: $L_{cls}(p, k^*)$

Loss: $\lambda[k* \geq 0]L_{loc}(t, t^*)$

>300 proposals

Output stride = 2   Output stride = 4   Output stride = 8   Output stride = 16   Output stride = 32

Data | Conv1_1 | Conv1_2 | Pool1 | Conv2_1 | Conv2_2 | Pool2 | Conv3_1 | Conv3_2 | Conv3_3 | Pool3 | Conv4_1 | Conv4_2 | Conv4_3 | Pool4 | Conv5_1 | Conv5_2 | Conv5_3 | Pool5 | fc6 – drop6 | fc7 – drop7 | fc8 | Prop

Feature extractor network