Single Shot Detection Algorithm

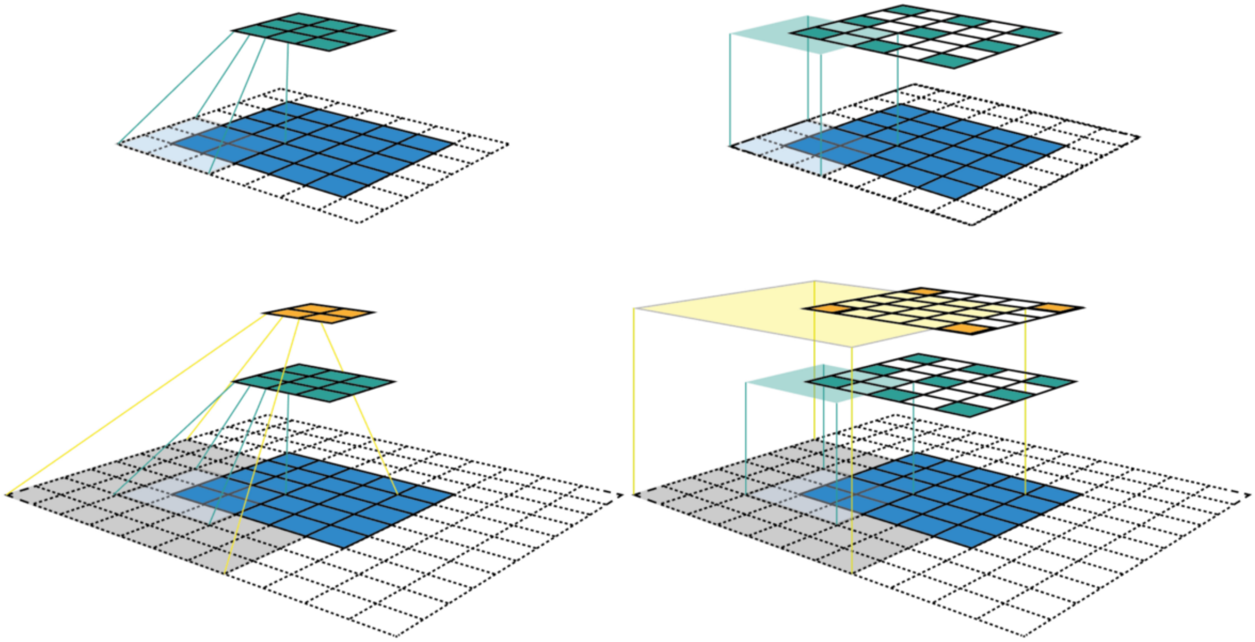
Resources

Undestand and Implement SSD

SSD Paper

Receptive fields

Receptive Fields in Detail



The input image is 5x5. Here we apply a convolutional filter of kernel **size k=3x3**, padding size **p=1x1**, and stride **s=2x2**. We end up with a **3x3 feature map**. If we apply the same convolution on top of the feature map, we end up with a **2x2 feature map**. The **receptive field** of a feature is defined as the area of the input image that is covered by that specific feature. If we use the fixed-size CNN visualization, we can see that the receptive filed of the second feature map is much larger than the receptive field of the first feature map.

Single Shot Detection

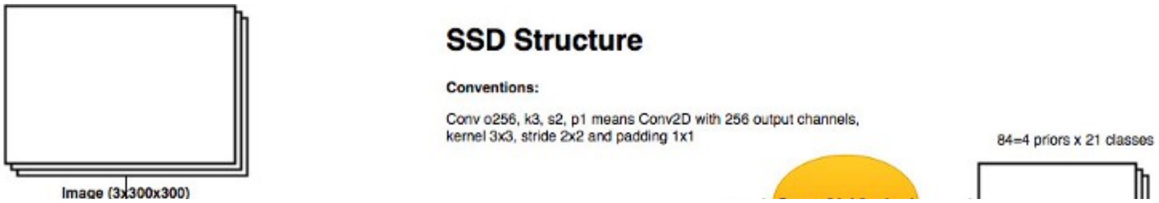
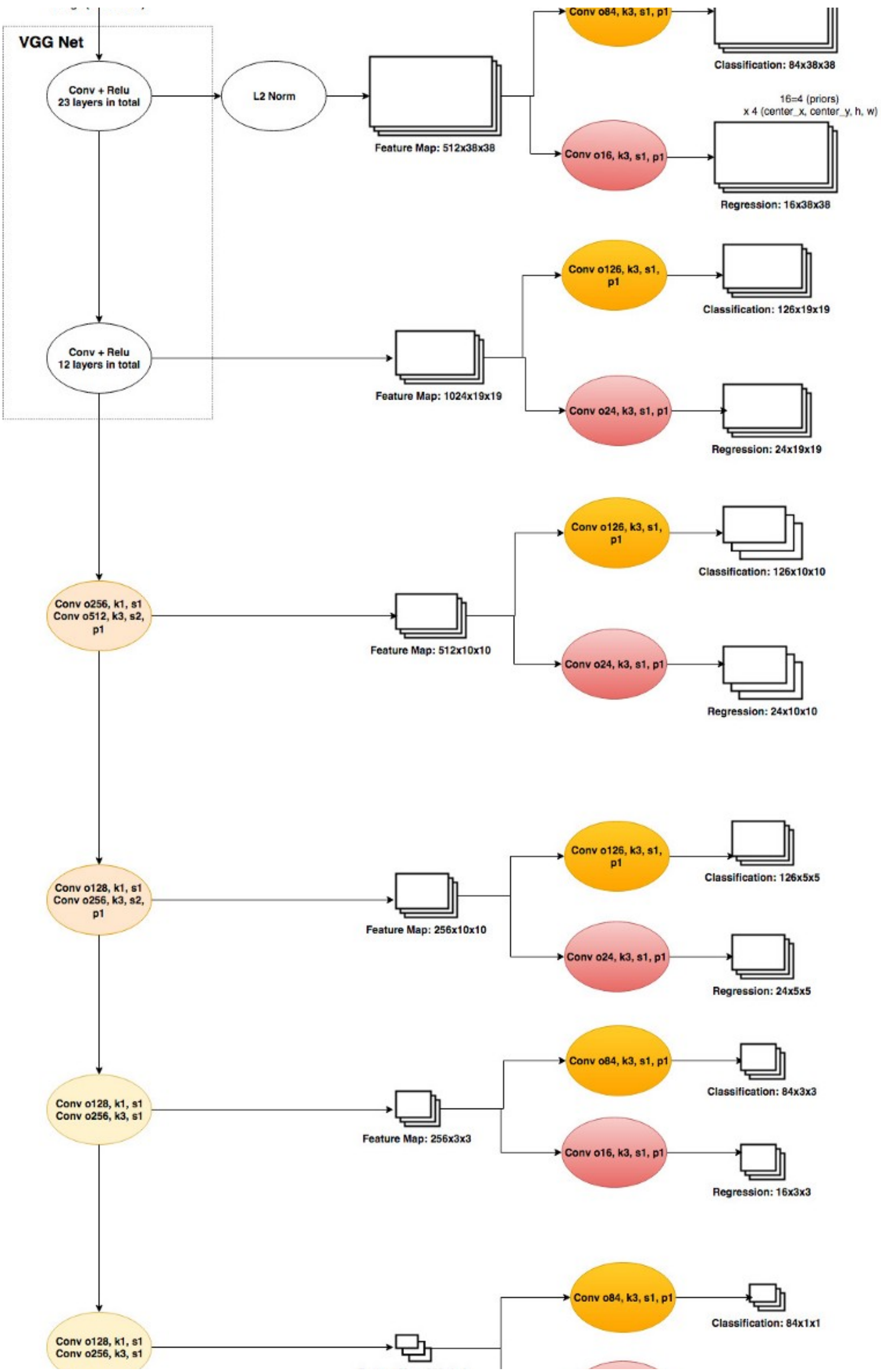
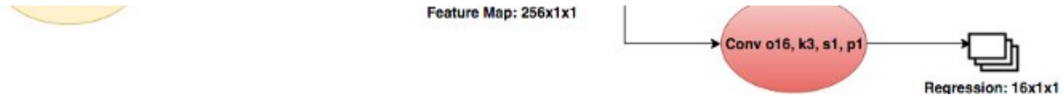


Image (3x300x300)
SSD Structure
Conventions:
Conv c256, k3, s2, p1 means Conv2D with 256 output channels, kernel 3x3, stride 2x2 and padding 1x1
84=4 priors x 21 classes


VGG Net
Conv + Relu
23 layers in total
L2 Norm
Feature Map: 512x38x38
Conv o84, k3, s1, p1
Classification: 84x38x38
16=4 (priors)
x 4 (center_x, center_y, h, w)
Conv o16, k3, s1, p1
Regression: 16x38x38
Conv + Relu
12 layers in total
Feature Map: 1024x19x19
Conv o126, k3, s1, p1
Classification: 126x19x19
Conv o24, k3, s1, p1
Regression: 24x19x19
Conv o256, k1, s1
Conv o512, k3, s2, p1
Feature Map: 512x10x10
Conv o126, k3, s1, p1
Classification: 126x10x10
Conv o24, k3, s1, p1
Regression: 24x10x10
Conv o128, k1, s1
Conv o256, k3, s2, p1
Feature Map: 256x10x10
Conv o126, k3, s1, p1
Classification: 126x5x5
Conv o24, k3, s1, p1
Regression: 24x5x5
Conv o128, k1, s1
Conv o256, k3, s1
Feature Map: 256x3x3
Conv o84, k3, s1, p1
Classification: 84x3x3
Conv o16, k3, s1, p1
Regression: 16x3x3
Conv o128, k1, s1
Conv o256, k3, s1
Conv o84, k3, s1, p1
Classification: 84x1x1

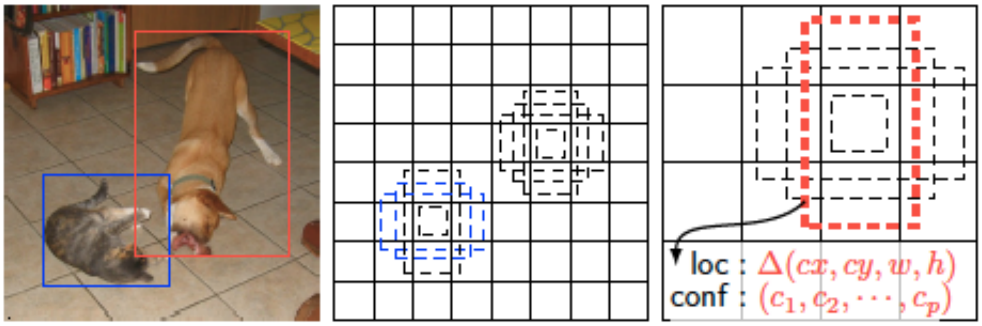


In the SSD architecture we take advantage of receptive fields. Shallow layers look at small portions of the input image, while deeper layers look at larger portions. Therefore, we can use shallow layers to classify small objects and deeper layers to classify larger ones.

Output

The model produces two outputs for each point in each feature map: a classification output and a regression output. The first one tells us which class belongs to the feature point, the second one tells us the coordinates of the corresponding bounding box.Ge

Generating Priors (Bounding Boxes)


loc : $\Delta(cx, cy, w, h)$
conf : $(c_1, c_2, \cdots, c_p)$

(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

For each feature map 4 bounding boxes is generated. The bounding boxes are sometimes referred to as priors or anchors.

Training

Match anchors with ground truth

During training each bounding box is matched with the ground truth. If the Intersection over Union (IoU) is more than 0.5 the bounding box becomes gets assigned to a class, otherwise it gets assigned to background.

Scale Default Boxes

Intuitively, it makes sense to scale the priors so that specific feature maps learn to be sensitive to particular scales of objects.

Hard Negative Mining

After matching, most default boxes are negative. This results in an imbalanced dataset. We can select only the background boxes with highest confidence loss to impose a fixed target/background equal to 1/3.

Loss Function

The loss function is a combination of a classification loss and a regression loss.

$$L(x,c,l,g) = \frac{1}{N}(L_{conf}(x,c) + \alpha L_{loc}(x,l,g))$$

N is the number of matched default boxes (or priors).

The classification loss is referred as confidence loss and is it a simple categorical cross entropy over the ground truth labels.

The regression loss is a smoothL1 loss between parameters of the predicted box parameters and the ones of the ground truth box. The parameters of the ground truth box (cx, cy, w, h) are offset by the ones of the matching deafult box.

$$L_{loc}(x,l,g) = \sum_{i}^{N} \sum_{m} smooth_{L1}(l_i^m - \hat{g}_i^m)$$

Data Augmentation

In order to make the model more robust towards object of various input sizes and shapes, the original dataset is augment by applying sample patching and photo-metric distortions.