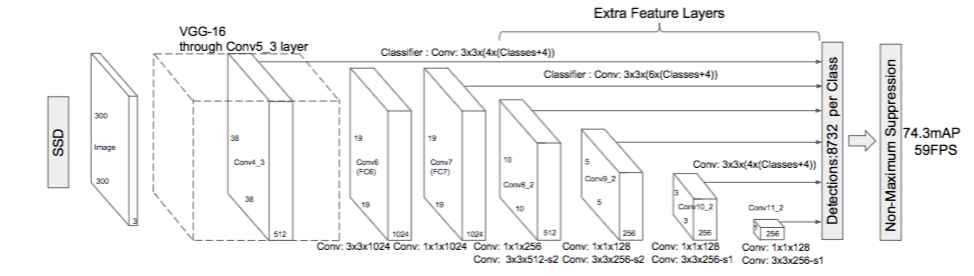
**SPRINT-4 (SSD\_Model Architecture)**



**Single Shot:**this means that the tasks of object localization and classification are done in a singleforward pass of the network.

SSD only needs an input image and ground truth boxes for each object during training. In a convolution manner. We evaluate a small of default boxes of different aspect ratios at each location in several feature maps with different scales (e.g. 8 × 8 and 4 × 4).

At training time, we first match these default boxes to the ground truth boxes. For example, we have matched two default boxes with the cat and one with the dog, which are treated as positives and the rest as negatives.

The model loss is a weighted sum between localization loss (e.g. Smooth L1 [6]) and confidence loss. The SSD normally start with a VGG on Resnet pre-trained model that is converted to a fully convolution neural network.

MultiBox’s loss function also combined two critical components that made their way into SSD:

**Confidence Loss**: this measures how confident the network is of the of the computed bounding box. Categorical cross-entropy is used to compute this loss.

**Location Loss:**this measures how far away the network’s predicted bounding boxes are from the ground truth ones from the training set.

In MultiBox, the researchers created what we call priors (or anchors in Faster-R-CNN terminology), which are pre-computed, fixed size bounding boxes that closely match the distribution of the original ground truth boxes.

There the conv4\_3 is responsible to detect the smallest objects while the conv11\_2 is responsible for the biggest objects*.*

Some activations are taken from the network and passed to a specialized sub-network that should work as a classifier and localizer.

During prediction we use a Non-maxima suppression algorithm to filter the multiple boxes per object that may appear.