DEMYSTIFYING DEEP LEARNING: TUTORIAL SERIES

CHAPTER 6: CNN FOR OBJECT DETECTION

Vaibhav Verdhan

June 15 2020

AGENDA FOR THE SERIES

Session	Topic
Session 1	Introduction to Deep Learning
Session 2	Building blocks of Neural Network - 1
Session 3	Building blocks of Neural Network - 2
Session 4	Convolutional Neural Network
Session 5	CNN for Image Classification
Session 6	CNN for Object Detection
Session 7	Architectures like AlexNet, Inception etc. (June 18)
Session 8	Recurrent Neural Network
Session 9	NLP Applications of RNN

AGENDA FOR SESSION 6

- *Classification vs Detection in an Image*
- *One Shot Learning*
- *Object Detection*
- *YOLO*
- *SSD*
- *Region Proposals*
- *IOU*
- *Python Lab*

SHOW THE VIDEO

DIFFERENCE BETWEEN CLASSIFICATION & DETECTION

For example we want to identify a face vs recognise a face

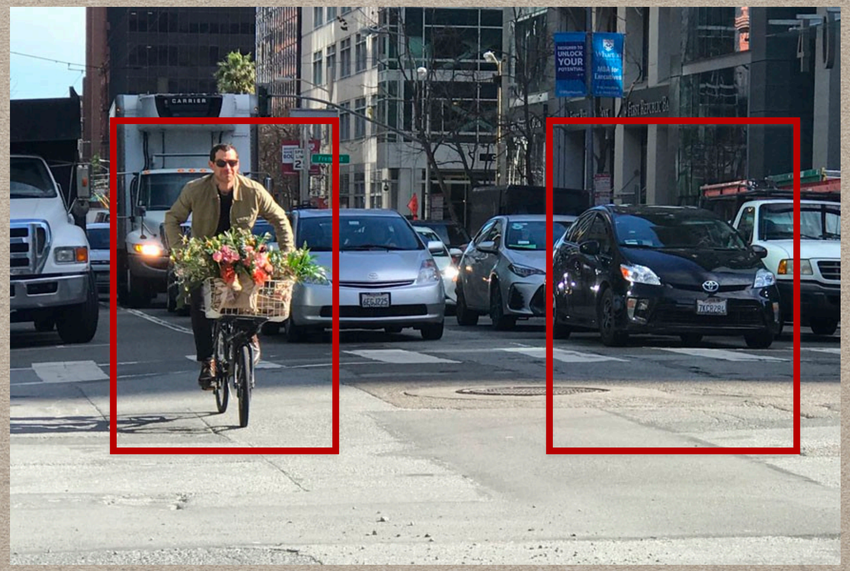
Is there a face

Is it of Vaibhav

Whose face is this

Have a database of people. There is a concept of one-shot learning

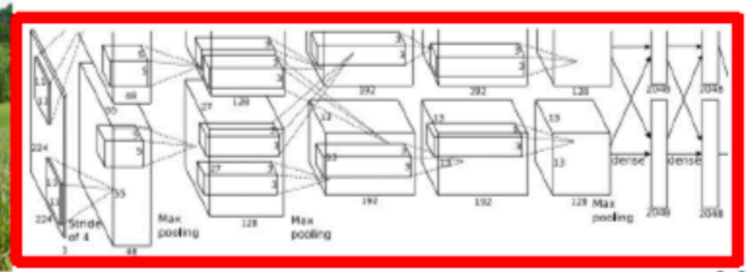
DIFFERENCE BETWEEN CLASSIFICATION & DETECTION



Not only classify the image but also identify the coordinates of the object present in the image i.e the locatilization

LOCATE THE CAT IN THE IMAGE

This image is CC0 public domain



Often pretrained on ImageNet
(Transfer learning)

Vector:
4096

Fully Connected:
4096 to 1000

Class Scores
Cat: 0.9
Dog: 0.05
Car: 0.01
...

Correct label:
Cat

Softmax Loss

Fully Connected:
4096 to 4

Box Coordinates
(x, y, w, h)

L2 Loss

Loss

Correct box:
(x', y', w', h')

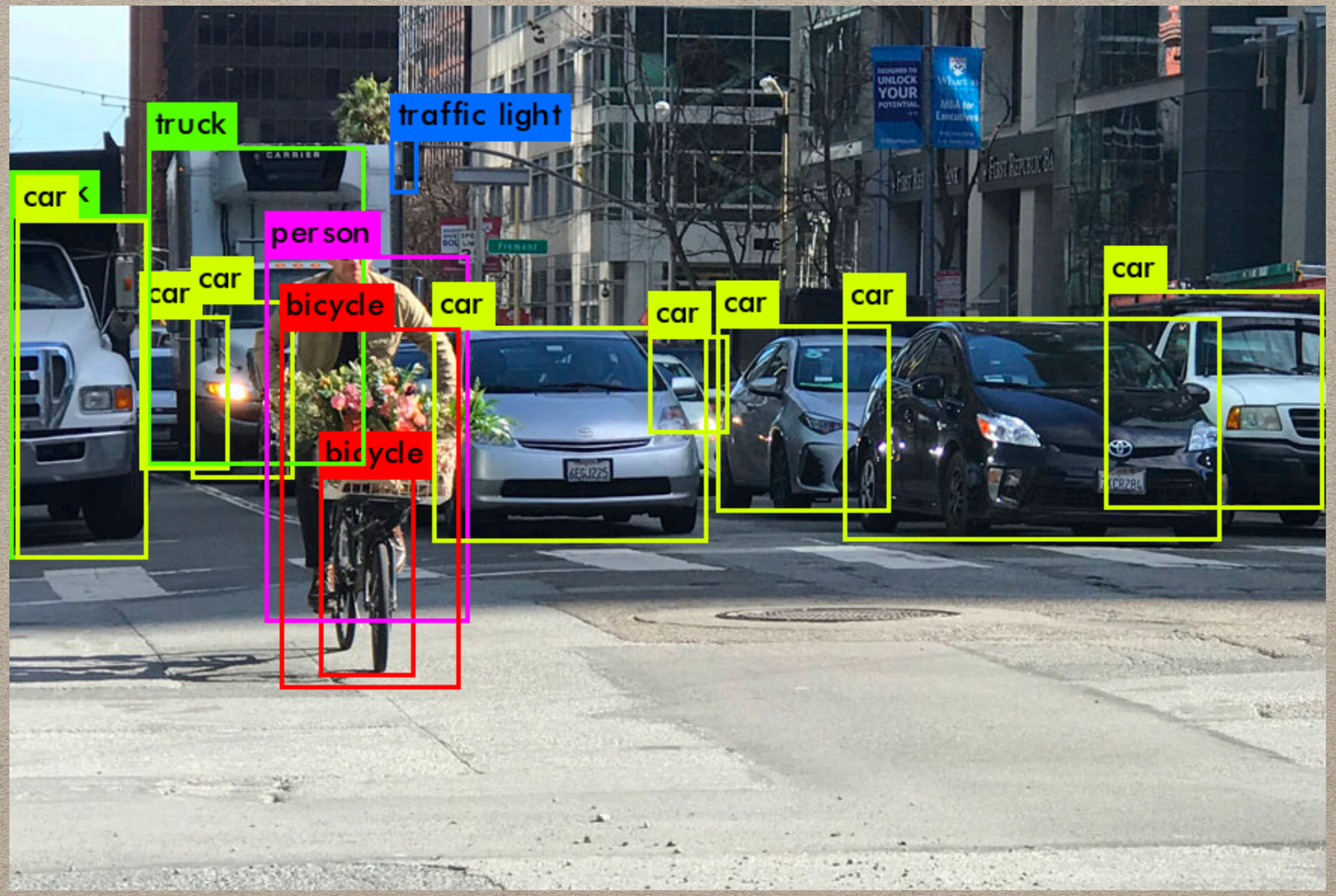
Treat localization as a regression problem!

OBJECT DETECTION

To make image classification we use a ConvNet with a Softmax attached to the end of it.

To make classification with localization we use a ConvNet with a softmax attached to the end of it and a four numbers bx, by, bh, and bw to tell you the location of the class in the image. The dataset should contain this four numbers with the class too

OBJECT DETECTION



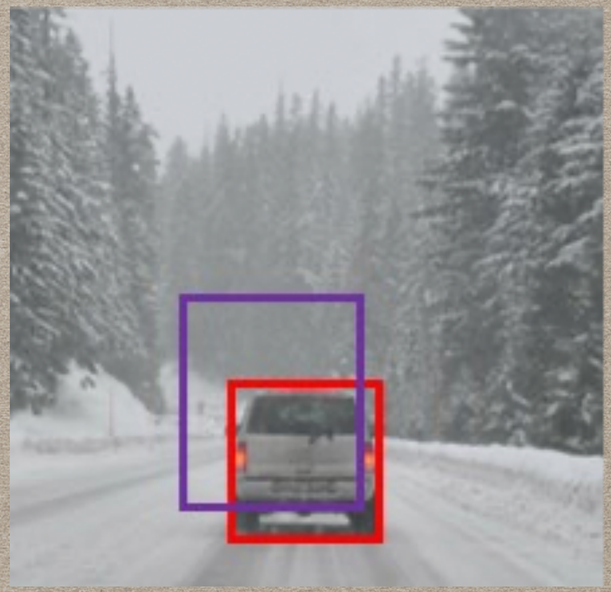
INTERSECTION OVER UNION

Intersection Over Union is a function used to evaluate the object detection algorithm.

It computes size of intersection and divide it by the union. More generally, IoU is a measure of the overlap between
two bounding boxes.

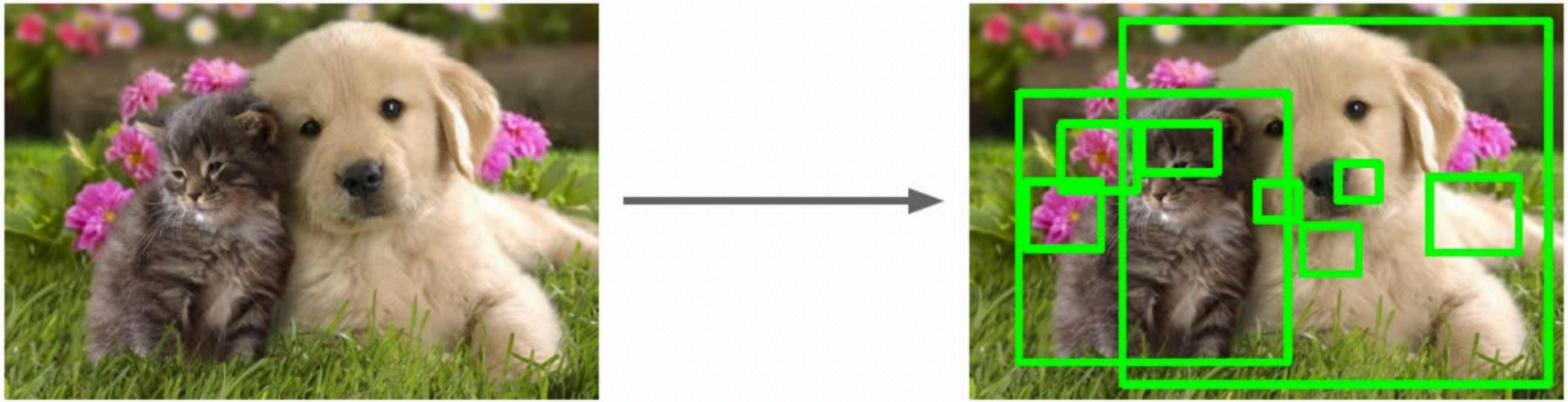
IOU = intersection area / Union area

If IOU >=0.5 then its good. The best answer will be 1.
The higher the IOU the better is the accuracy



REGION PROPOSALS

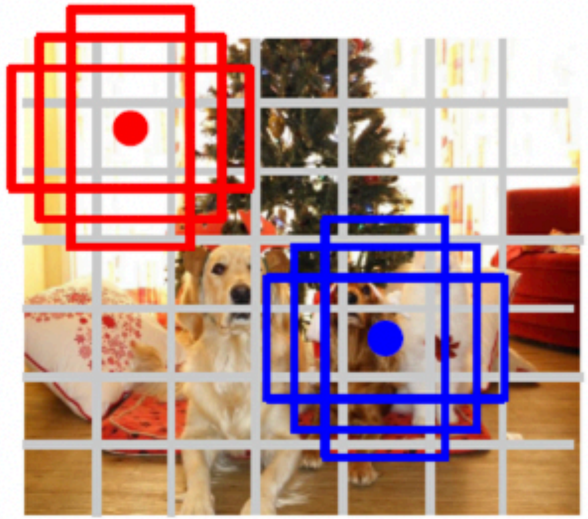
- Find “blobby” image regions that are likely to contain objects
- Relatively fast to run; e.g. Selective Search gives 1000 region proposals in a few seconds on CPU



DETECTION WITHOUT PROPOSALS

Go from input image to tensor of scores with one big convolutional network!

Input image
3 x H x W



Divide image into grid
7 x 7

Image a set of **base boxes** centered at each grid cell
Here B = 3

Within each grid cell:

- Regress from each of the B base boxes to a final box with 5 numbers:
  (dx, dy, dh, dw, confidence)
- Predict scores for each of C classes (including background as a class)

Output:
7 x 7 x (5 * B + C)

Redmon et al, "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016
Liu et al, "SSD: Single-Shot MultiBox Detector", ECCV 2016

YOLO WORKING METHODOLOGY

- YOLO actually divides the image into a grid of say, 15*15 cells (S=15)

- Each of these cells is responsible for predicting 5 bounding boxes (B=5) (A bounding box describes the rectangle that encloses an object)

- YOLO for each bounding box
  - outputs a confidence score that tells us how good is the shape of the box
  - the cell also predicts a class
- The confidence score of bounding box and class prediction are combined into final score -> probability that this bounding box contains a specific object

SINGLE SHOT DETECTION

By using SSD, we only need to take one single shot to detect multiple objects within the image

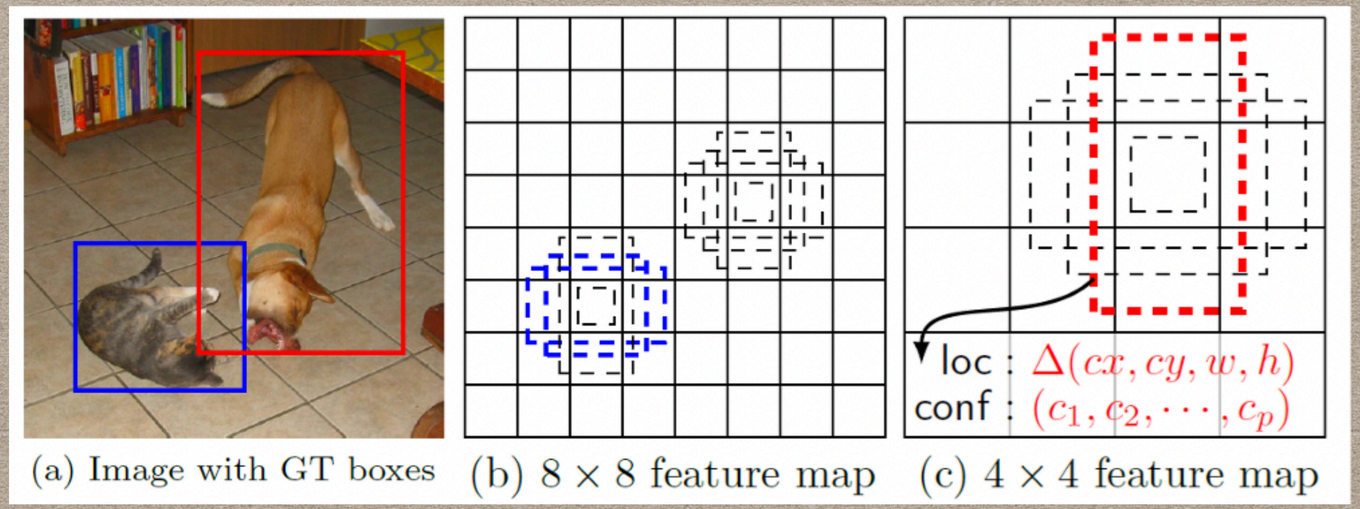
Regional proposal network (RPN) based approaches such as R-CNN, Fast R-CNN series need two shots, one for generating region proposals, one for detecting the object of each proposal.

SSD is much faster compared with two-shot RPN-based approaches.

The loss function consists of two terms: Lconf (confidence loss) and Lloc (localization loss)

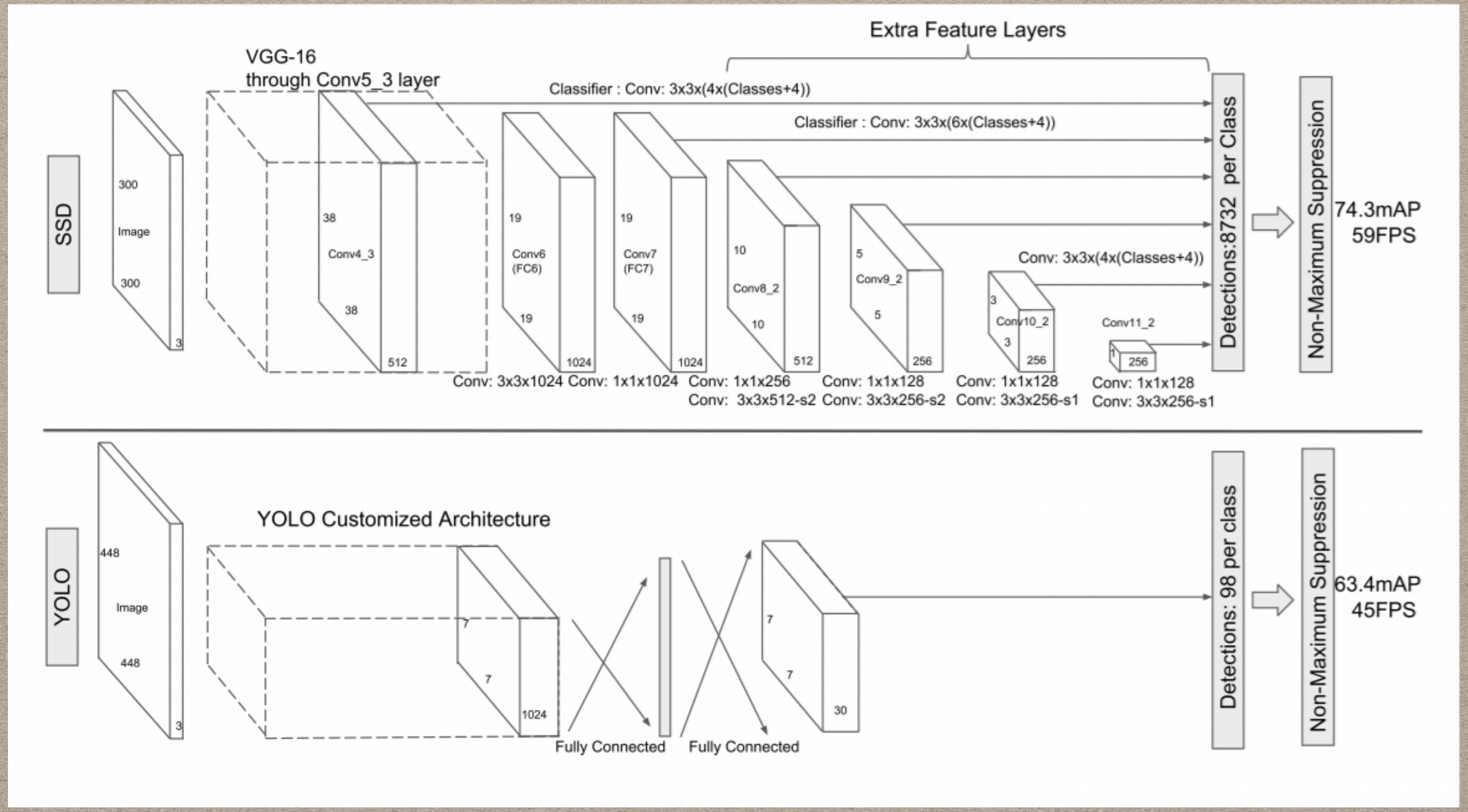
SINGLE SHOT DETECTION

- A feature layer of size m×n (number of locations) with p channels
- For each location, we got k bounding boxes
- For each of the bounding box, we will compute c class scores and 4 offsets relative to the original default bounding box shape.
- Thus, we got (c+4)kmn outputs.



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

SINGLE SHOT DETECTION



TIME TO HIT THE LAB

QUESTIONS PLEASE!

github.com/vverdhan

Thanks