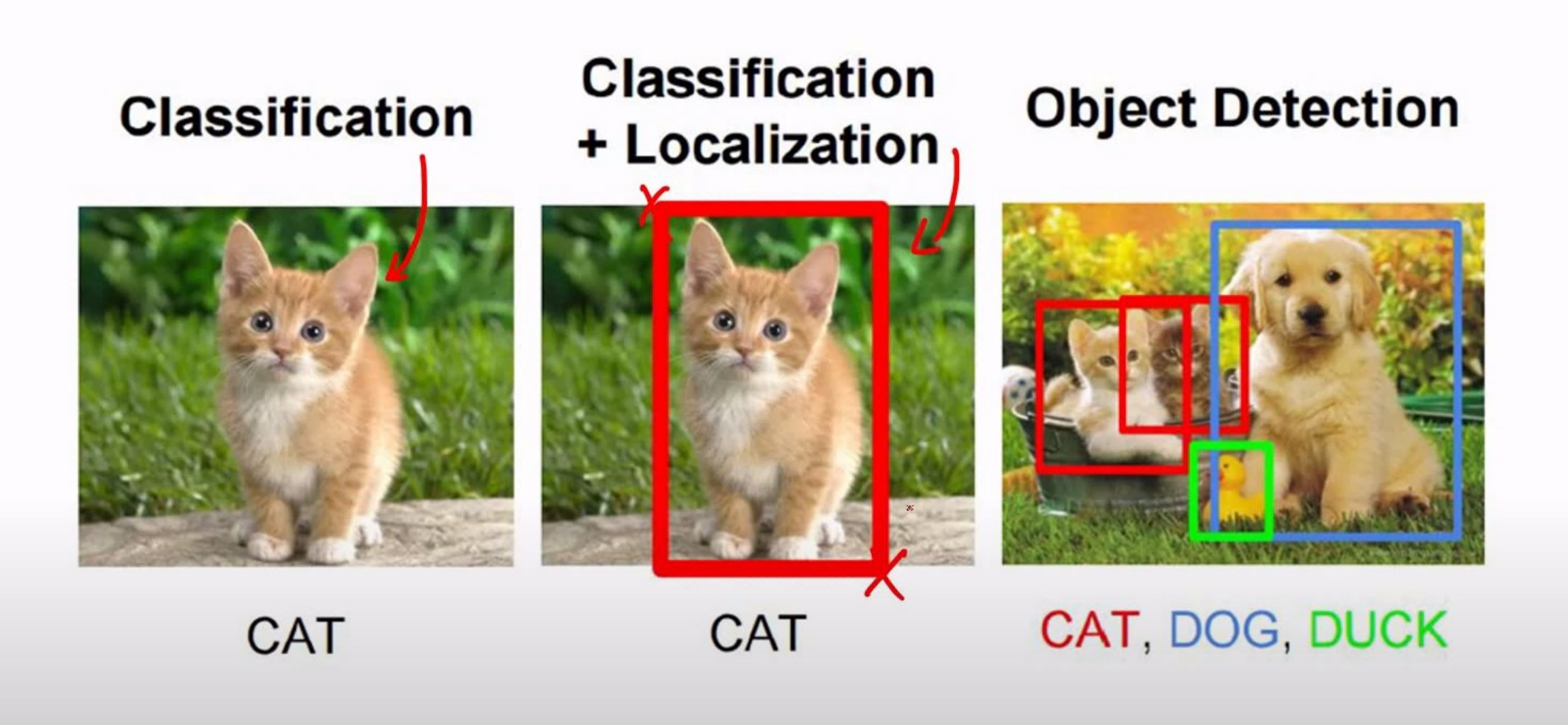
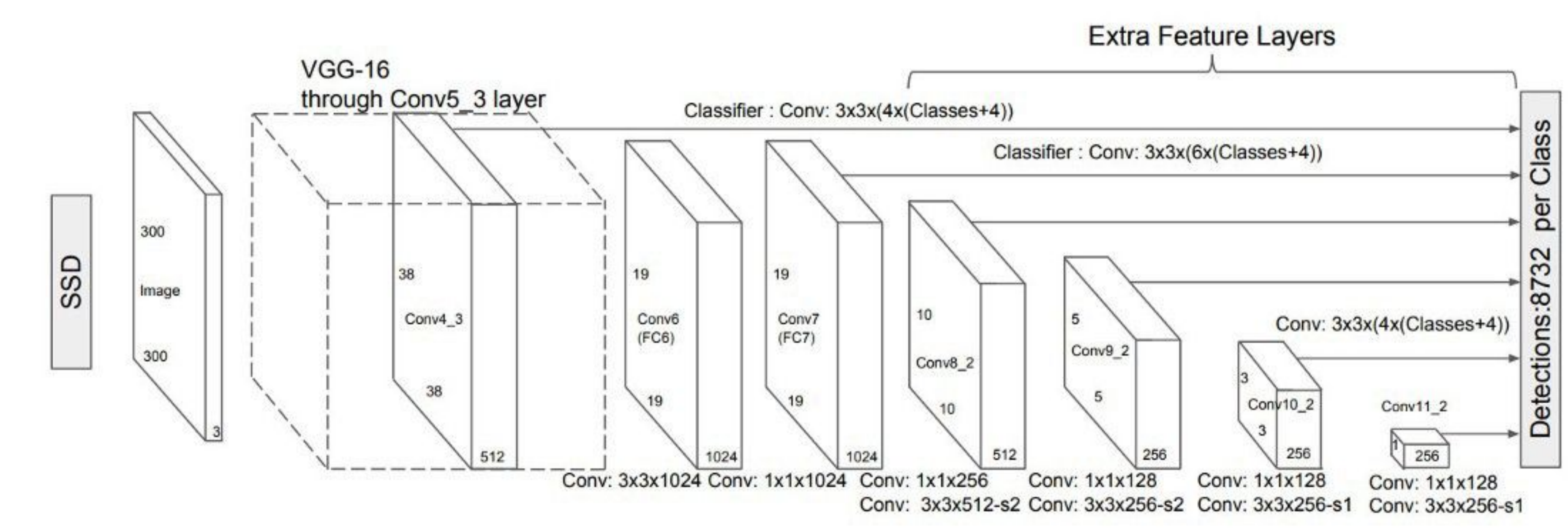
Single Shot Detector (SSD300)

Bahdah Shin
Matt Witman
Dan Moore

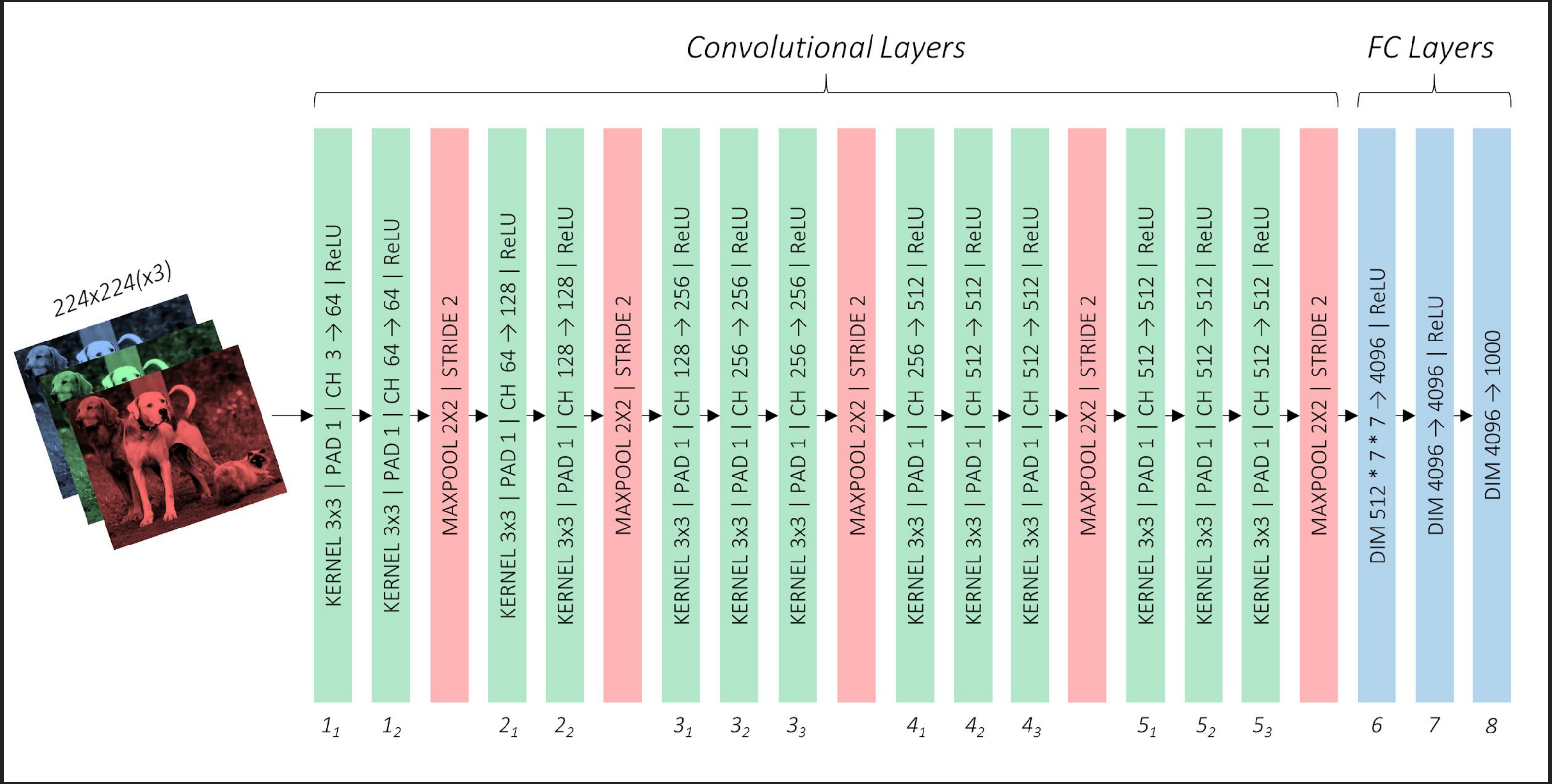
What is Object Detection?



What is SSD?


Extra Feature Layers
VGG-16
through Conv5_3 layer
Classifier : Conv: 3x3x(4x(Classes+4))
Classifier : Conv: 3x3x(6x(Classes+4))
SSD
300
Image
300
3
38
Conv4_3
38
512
19
Conv6
(FC6)
19
1024
19
Conv7
(FC7)
19
1024
10
Conv8_2
10
512
5
Conv9_2
5
256
Conv: 3x3x(4x(Classes+4))
3
Conv10_2
3
256
Conv11_2
1
256
Detections:8732 per Class
Conv: 3x3x1024
Conv: 1x1x1024
Conv: 1x1x256
Conv: 3x3x512-s2
Conv: 1x1x128
Conv: 3x3x256-s2
Conv: 1x1x128
Conv: 3x3x256-s1
Conv: 1x1x128
Conv: 3x3x256-s1

Base Convolutions: part 1 (vgg-16 architecture)



Convolutional Layers

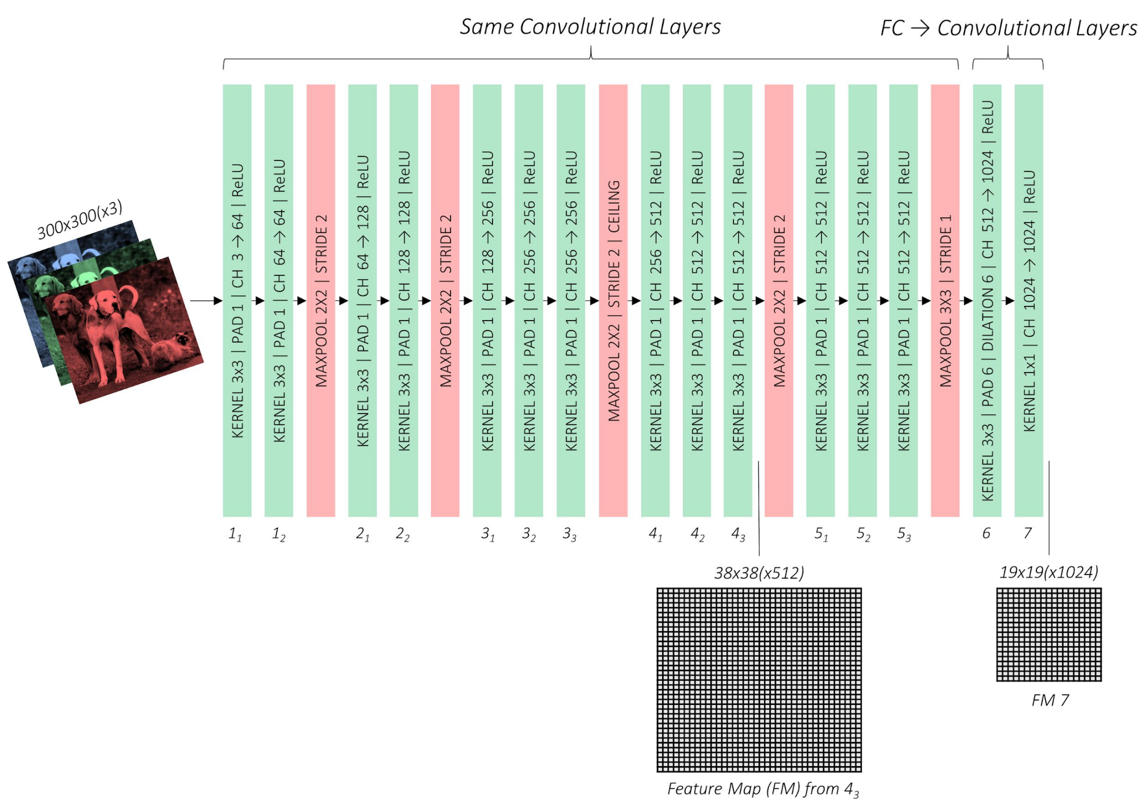
FC Layers

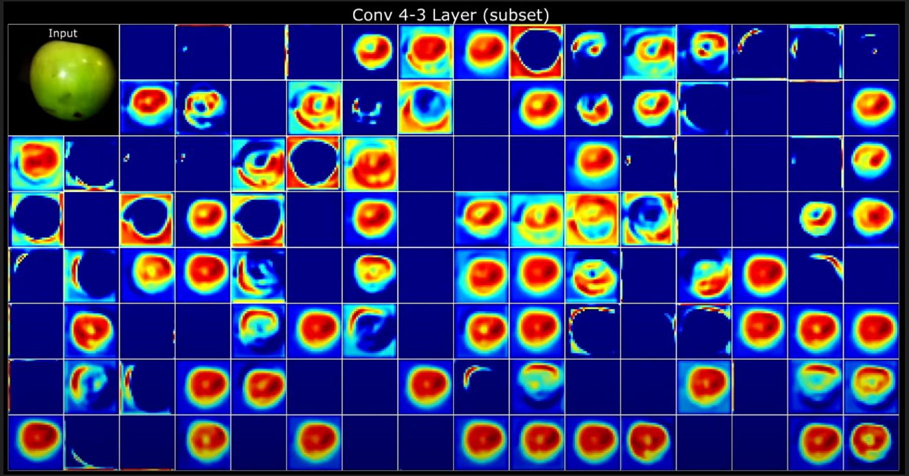
224x224(x3)

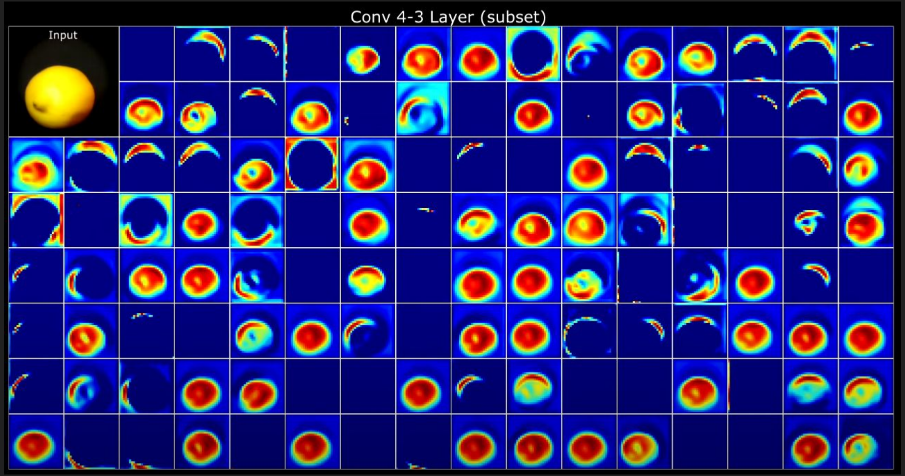
- KERNEL 3x3 | PAD 1 | CH 3 → 64 | ReLU
- KERNEL 3x3 | PAD 1 | CH 64 → 64 | ReLU
- MAXPOOL 2X2 | STRIDE 2
- KERNEL 3x3 | PAD 1 | CH 64 → 128 | ReLU
- KERNEL 3x3 | PAD 1 | CH 128 → 128 | ReLU
- MAXPOOL 2X2 | STRIDE 2
- KERNEL 3x3 | PAD 1 | CH 128 → 256 | ReLU
- KERNEL 3x3 | PAD 1 | CH 256 → 256 | ReLU
- KERNEL 3x3 | PAD 1 | CH 256 → 256 | ReLU
- MAXPOOL 2X2 | STRIDE 2
- KERNEL 3x3 | PAD 1 | CH 256 → 512 | ReLU
- KERNEL 3x3 | PAD 1 | CH 512 → 512 | ReLU
- KERNEL 3x3 | PAD 1 | CH 512 → 512 | ReLU
- MAXPOOL 2X2 | STRIDE 2
- KERNEL 3x3 | PAD 1 | CH 512 → 512 | ReLU
- KERNEL 3x3 | PAD 1 | CH 512 → 512 | ReLU
- KERNEL 3x3 | PAD 1 | CH 512 → 512 | ReLU
- MAXPOOL 2X2 | STRIDE 2
- DIM 512 * 7 * 7 → 4096 | ReLU
- DIM 4096 → 4096 | ReLU
- DIM 4096 → 1000

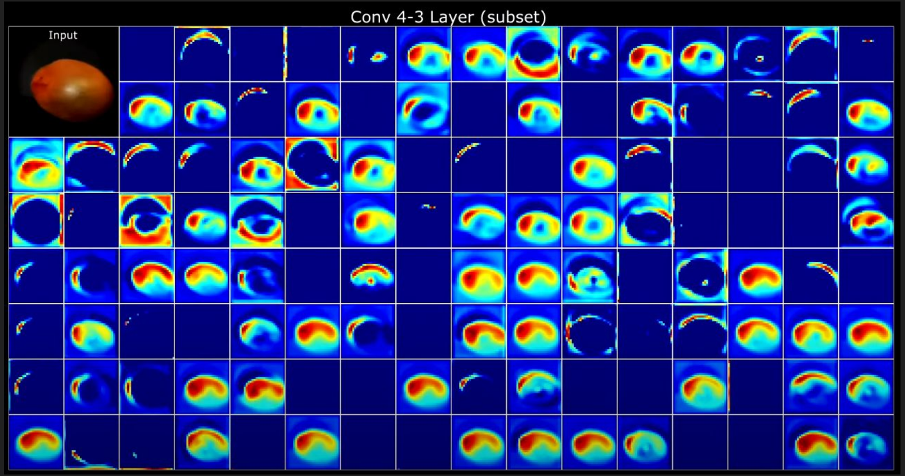
1_1 1_2 2_1 2_2 3_1 3_2 3_3 4_1 4_2 4_3 5_1 5_2 5_3 6 7 8

Base Convolutions: Part 2 (modified VGG-16)

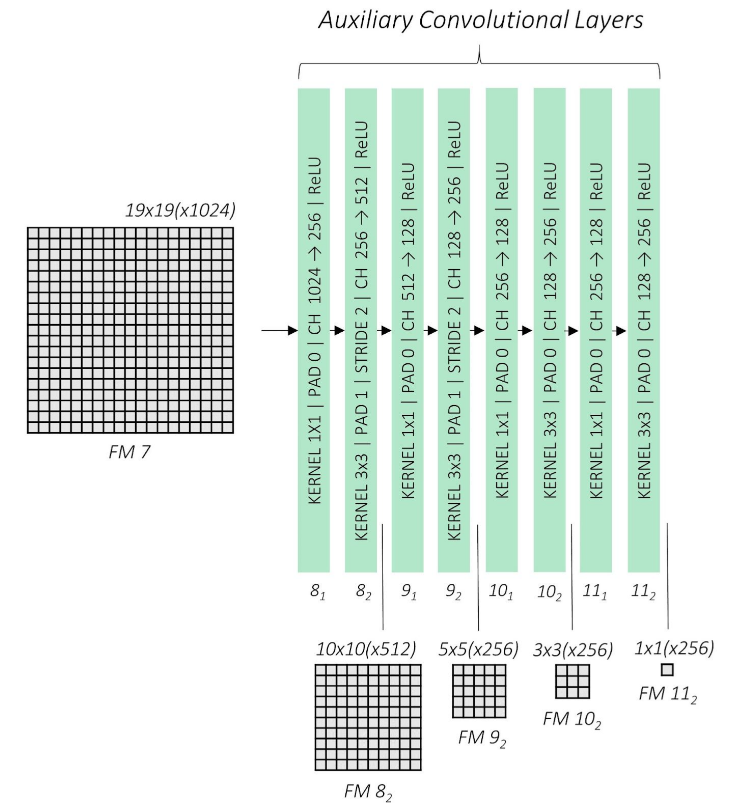

Same Convolutional Layers
FC → Convolutional Layers
300x300(x3)
KERNEL 3x3 | PAD 1 | CH 3 → 64 | ReLU
KERNEL 3x3 | PAD 1 | CH 64 → 64 | ReLU
MAXPOOL 2X2 | STRIDE 2
KERNEL 3x3 | PAD 1 | CH 64 → 128 | ReLU
KERNEL 3x3 | PAD 1 | CH 128 → 128 | ReLU
MAXPOOL 2X2 | STRIDE 2
KERNEL 3x3 | PAD 1 | CH 128 → 256 | ReLU
KERNEL 3x3 | PAD 1 | CH 256 → 256 | ReLU
KERNEL 3x3 | PAD 1 | CH 256 → 256 | ReLU
MAXPOOL 2X2 | STRIDE 2 | CEILING
KERNEL 3x3 | PAD 1 | CH 256 → 512 | ReLU
KERNEL 3x3 | PAD 1 | CH 512 → 512 | ReLU
KERNEL 3x3 | PAD 1 | CH 512 → 512 | ReLU
MAXPOOL 2X2 | STRIDE 2
KERNEL 3x3 | PAD 1 | CH 512 → 512 | ReLU
KERNEL 3x3 | PAD 1 | CH 512 → 512 | ReLU
KERNEL 3x3 | PAD 1 | CH 512 → 512 | ReLU
MAXPOOL 3X3 | STRIDE 1
KERNEL 3x3 | PAD 6 | DILATION 6 | CH 512 → 1024 | ReLU
KERNEL 1x1 | CH 1024 → 1024 | ReLU
1_1
1_2
2_1
2_2
3_1
3_2
3_3
4_1
4_2
4_3
5_1
5_2
5_3
6
7
38x38(x512)
Feature Map (FM) from 4_3
19x19(x1024)
FM 7


Conv 4-3 Layer (subset)
Input

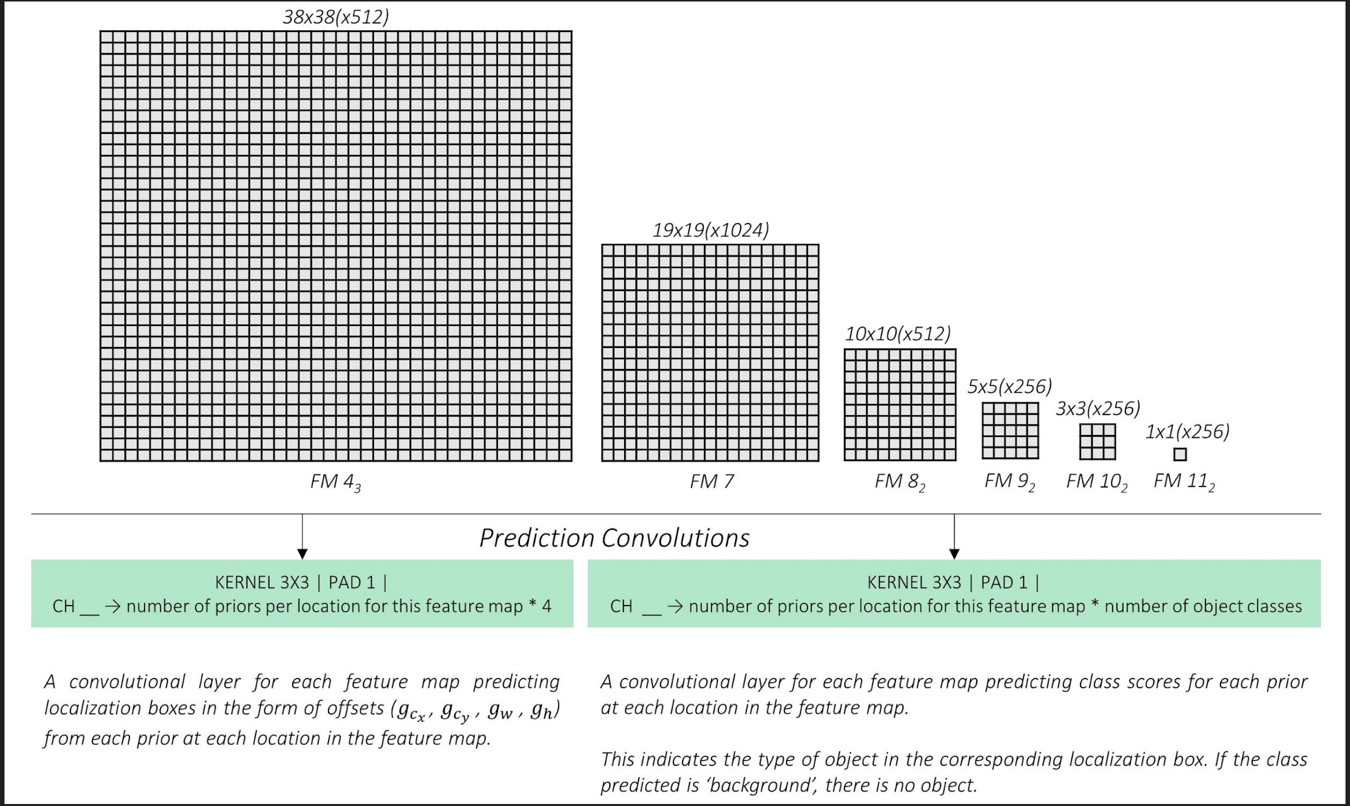

Conv 4-3 Layer (subset)
Input


Conv 4-3 Layer (subset)
Input

Auxiliary Convolutions


Auxiliary Convolutional Layers
19x19(x1024)
FM 7
KERNEL 1X1 | PAD 0 | CH 1024 → 256 | ReLU
KERNEL 3x3 | PAD 1 | STRIDE 2 | CH 256 → 512 | ReLU
KERNEL 1x1 | PAD 0 | CH 512 → 128 | ReLU
KERNEL 3x3 | PAD 1 | STRIDE 2 | CH 128 → 256 | ReLU
KERNEL 1x1 | PAD 0 | CH 256 → 128 | ReLU
KERNEL 3x3 | PAD 0 | CH 128 → 256 | ReLU
KERNEL 1x1 | PAD 0 | CH 256 → 128 | ReLU
KERNEL 3x3 | PAD 0 | CH 128 → 256 | ReLU
8_1
8_2
9_1
9_2
10_1
10_2
11_1
11_2
10x10(x512)
5x5(x256)
3x3(x256)
1x1(x256)
FM 11_2
FM 10_2
FM 9_2
FM 8_2

Prediction Convolutions



Prediction Convolutions

KERNEL 3X3 | PAD 1 |
CH __ → number of priors per location for this feature map * 4

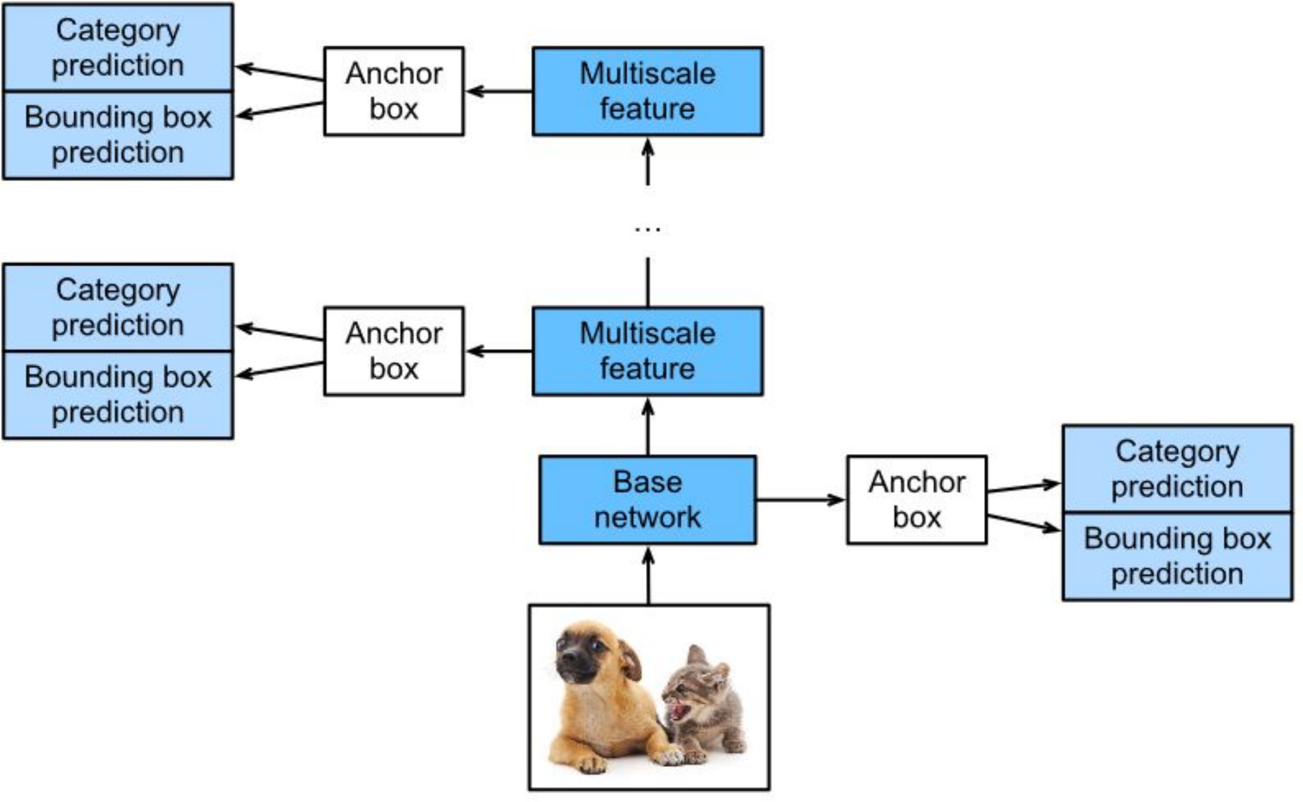
KERNEL 3X3 | PAD 1 |
CH __ → number of priors per location for this feature map * number of object classes

A convolutional layer for each feature map predicting localization boxes in the form of offsets (g_{c_x}, g_{c_y}, g_w, g_h) from each prior at each location in the feature map.

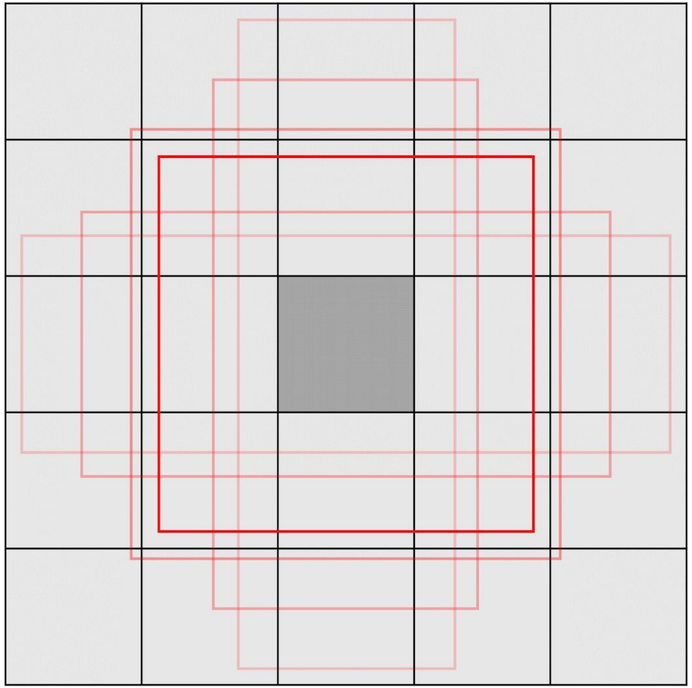
A convolutional layer for each feature map predicting class scores for each prior at each location in the feature map.

This indicates the type of object in the corresponding localization box. If the class predicted is 'background', there is no object.

Model: Anchor boxes (default box) (prior box)



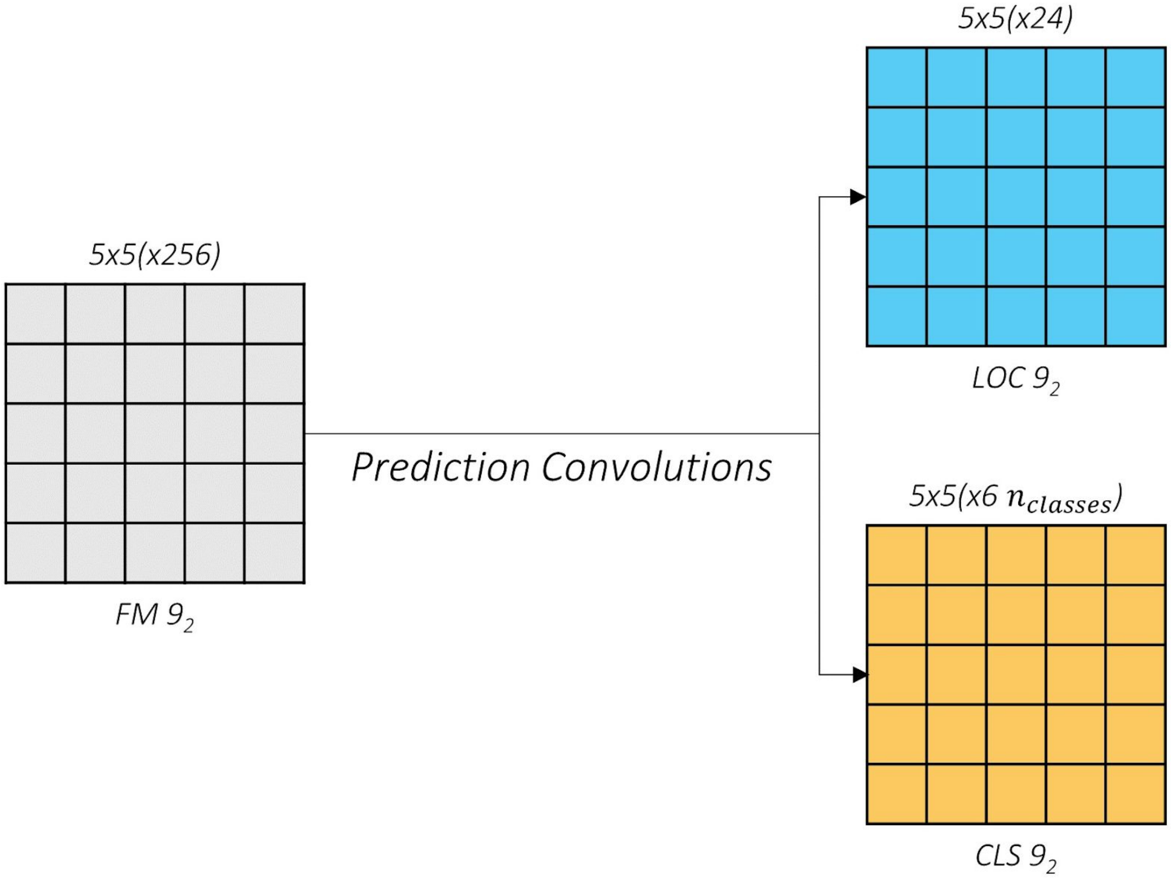
FM 9_2



At each location, there are 5 priors with aspect ratios 1, 2, 3, ½, ⅓ and areas equal to that of a square of side 0.55

Also, a 6th prior with aspect ratio 1 and of side 0.63

Output



5x5(x256)

FM 9_2

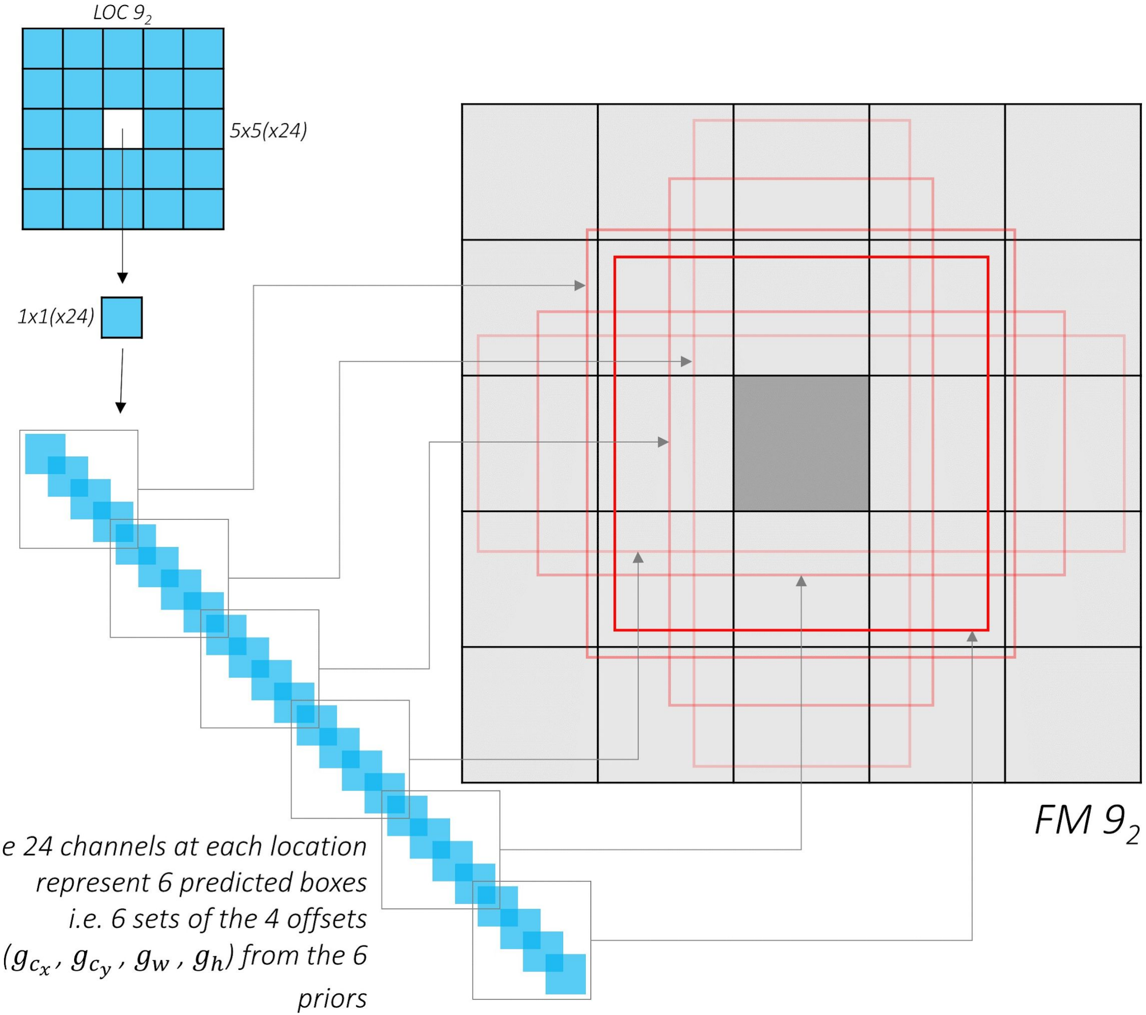
Prediction Convolutions

5x5(x24)

LOC 9_2

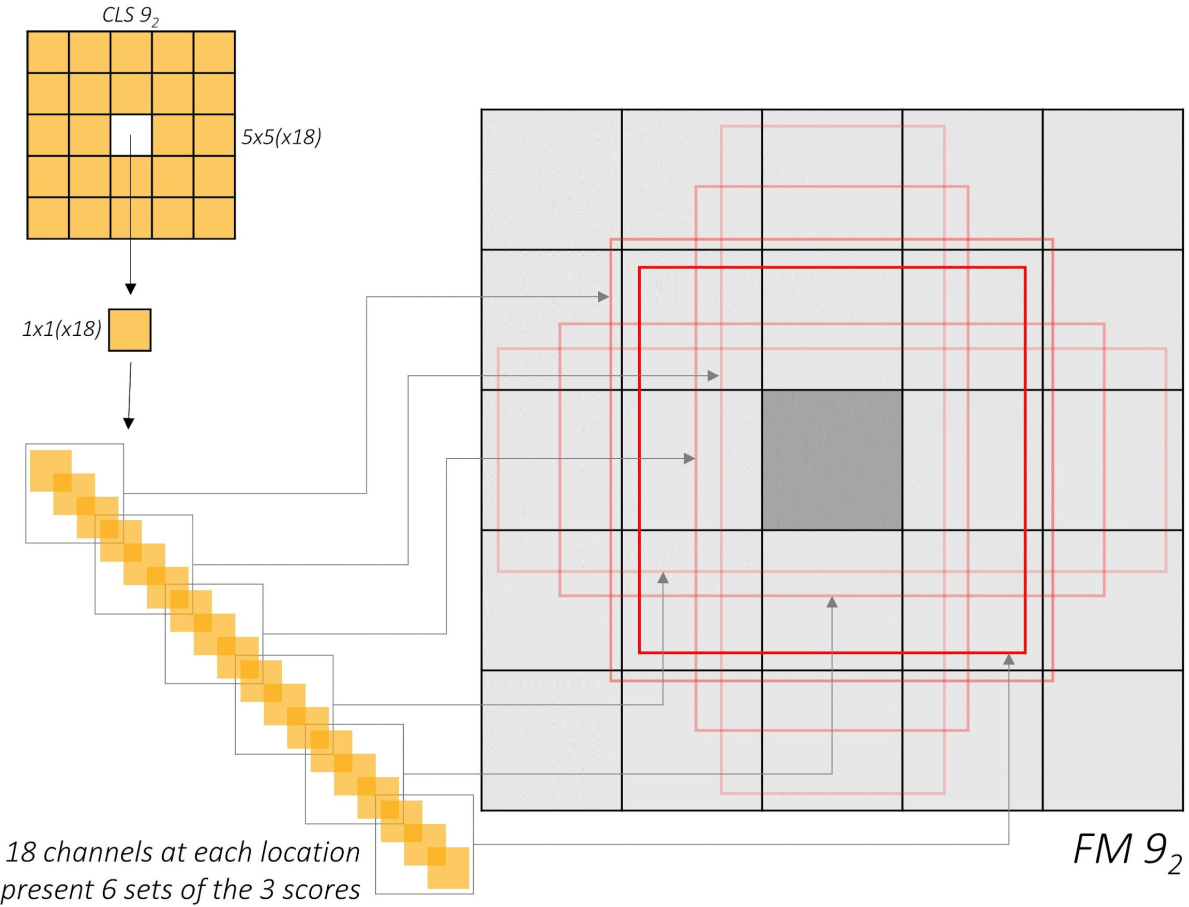
5x5(x6 $n_{classes}$)

CLS 9_2

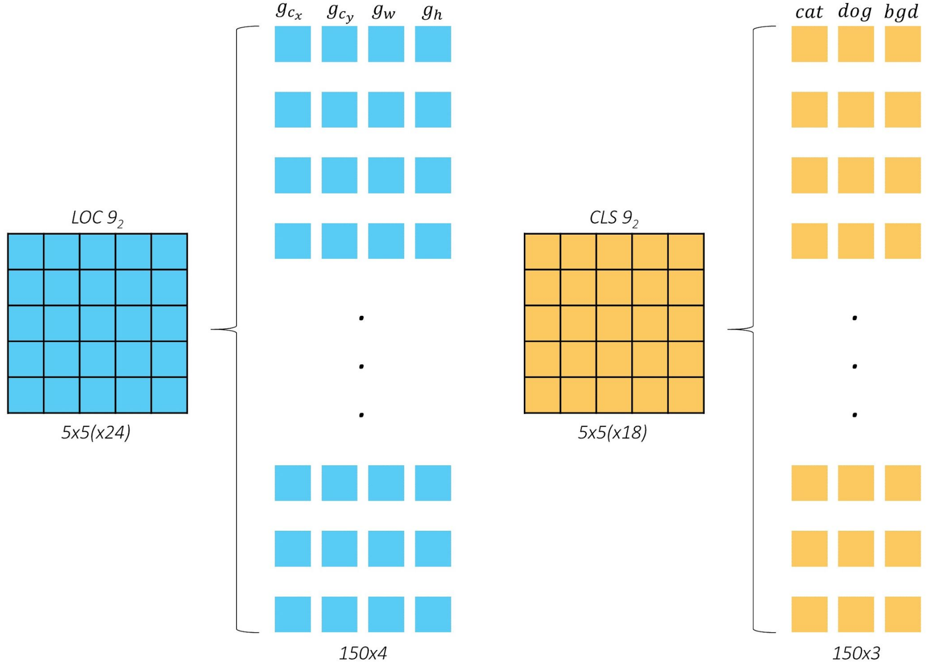


The 24 channels at each location represent 6 predicted boxes i.e. 6 sets of the 4 offsets $(g_{c_x}, g_{c_y}, g_w, g_h)$ from the 6 priors

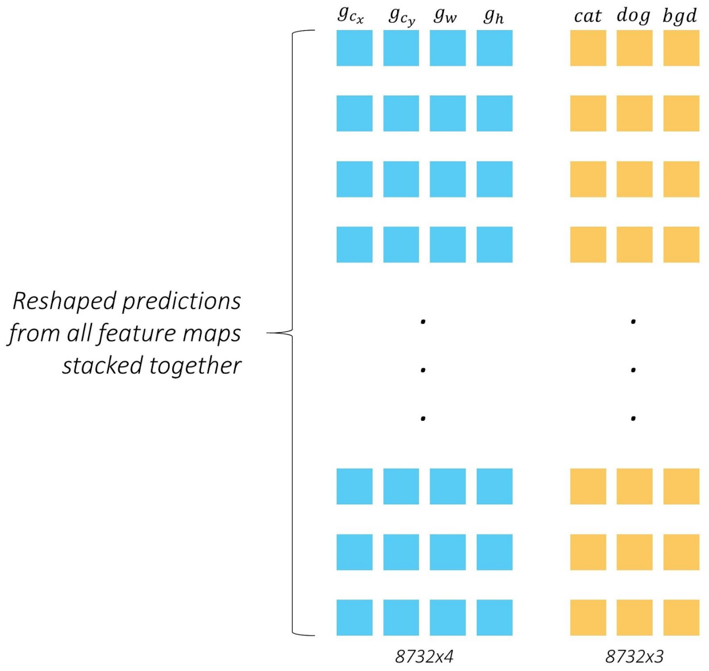
Assume $n_{classes} = 3\ (cat, dog, background)$



The 18 channels at each location represent 6 sets of the 3 scores $(\boldsymbol{cat}, \boldsymbol{dog}, \boldsymbol{bgd})$ *for the 6 priors*



Reshape predictions from FM 9_2 to represent offsets and class scores for the 150 predicted boxes


g_{c_x} g_{c_y} g_w g_h
cat dog bgd
Reshaped predictions from all feature maps stacked together
8732x4
8732x3

A total of 8732 predicted boxes

Conv4_3: 38 * 38 * 4 = 5776

Conv7: 19 * 19 * 6 = 2166

Conv8_2: 10 * 10 * 6 = 600

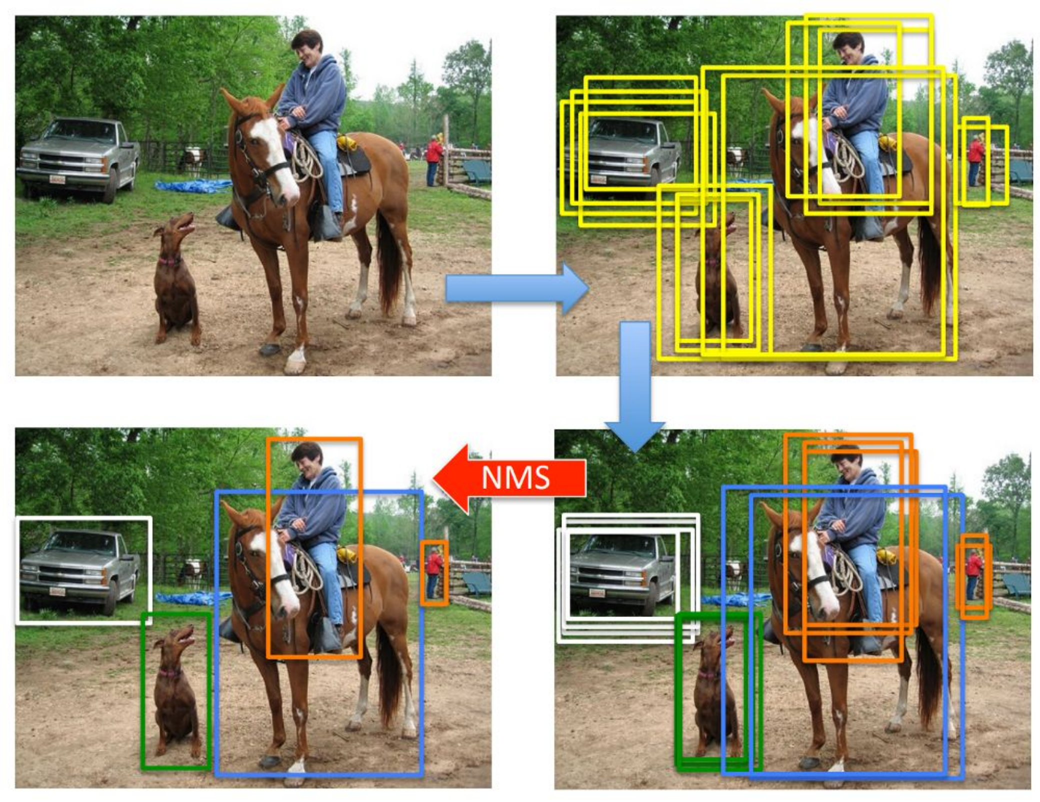
Conv9_2: 5 * 5 * 6 = 150

Conv10_2: 3 * 3 * 4 = 36

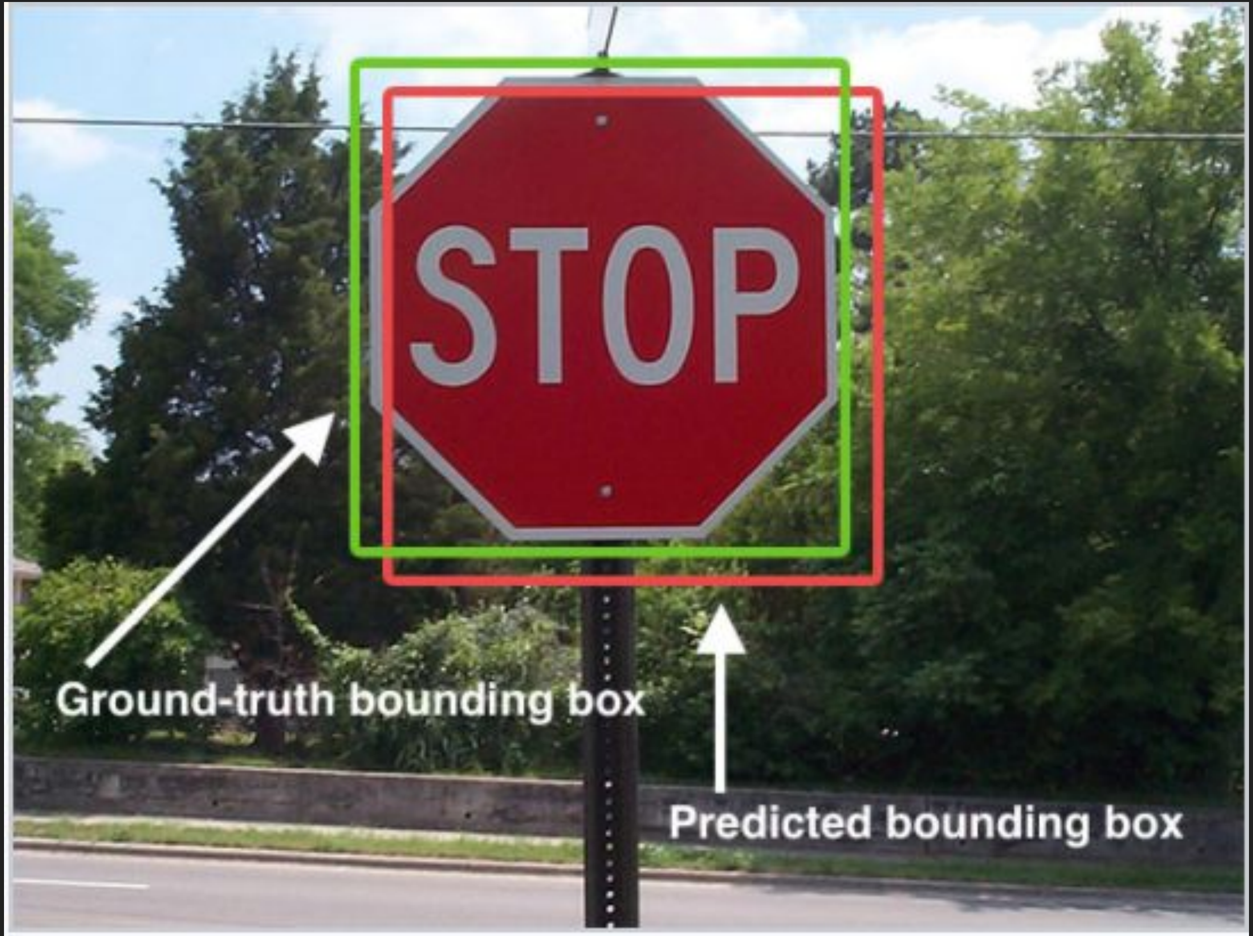
Conv11_2: 4

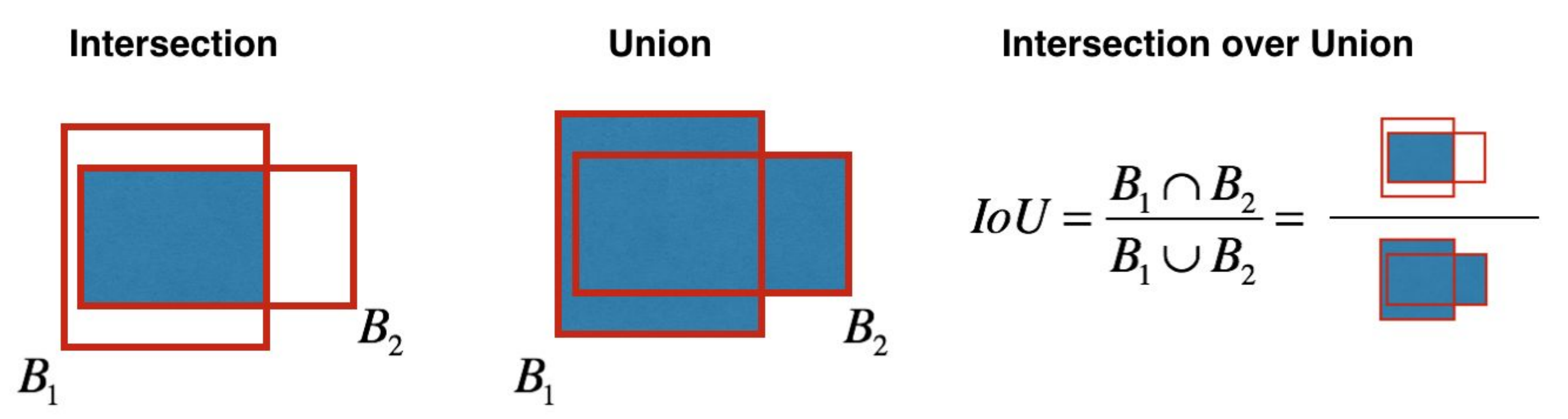
Total: 5776+ 2166 + 600 + 150 + 36 + 4 = 8732 priors

Non-Maximum Suppression

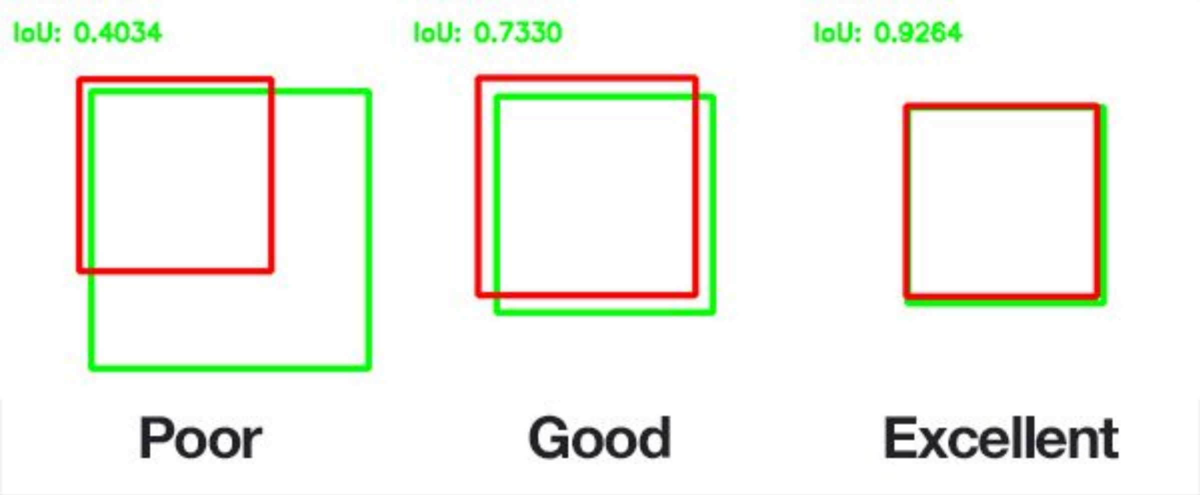


Training

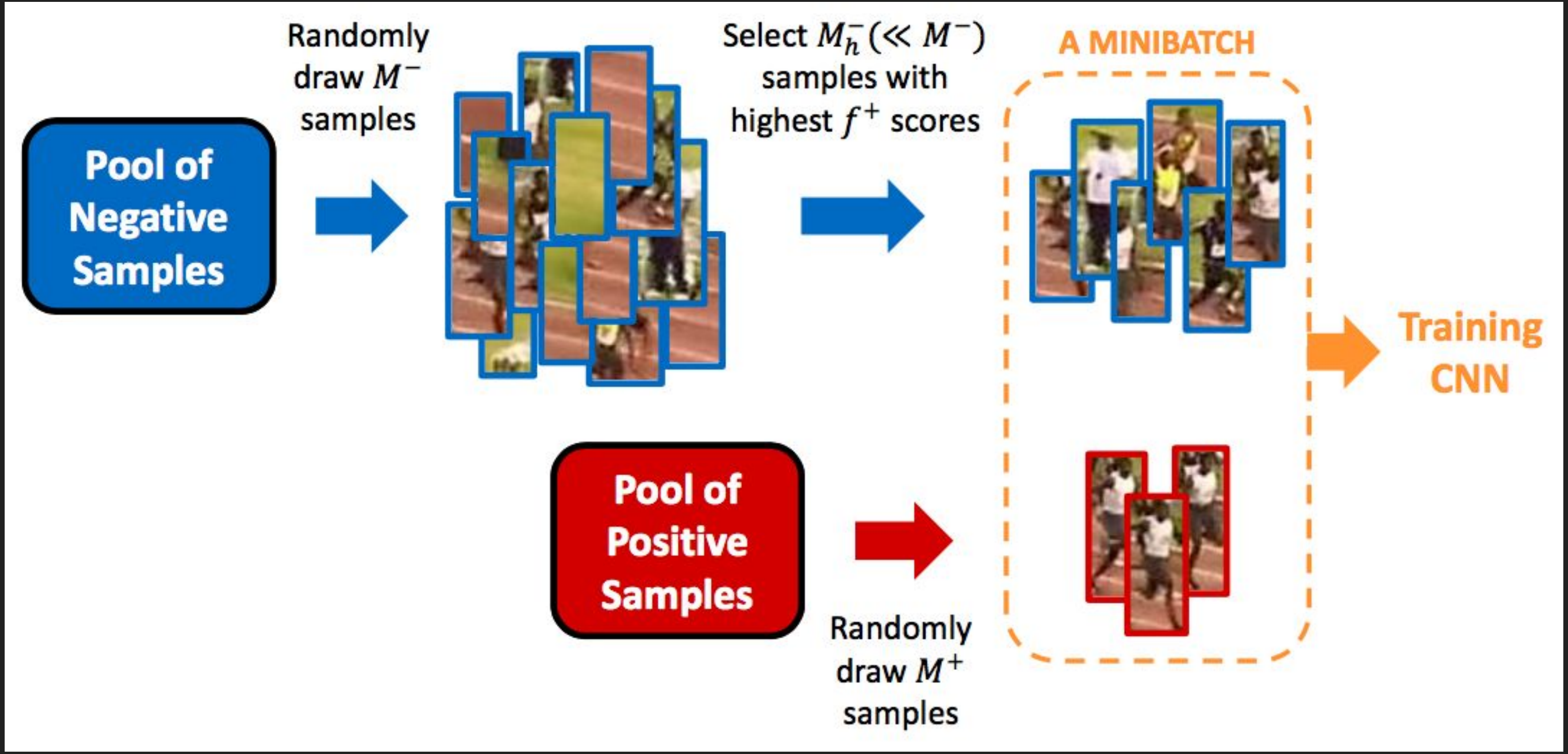

STOP
Ground-truth bounding box
Predicted bounding box



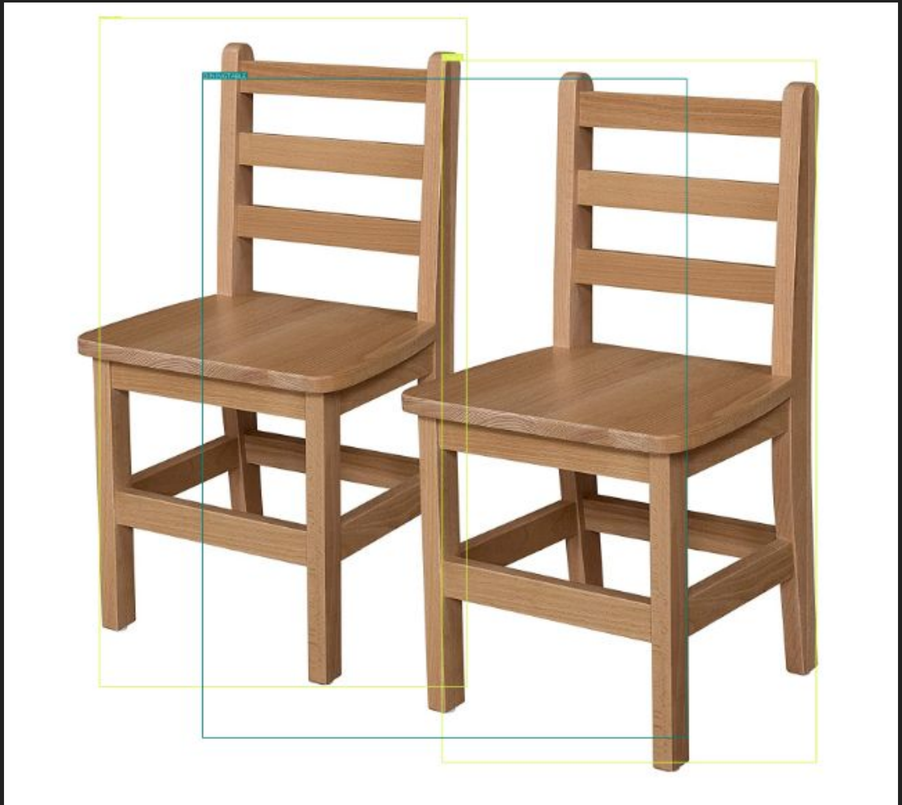
$$IoU = \frac{B_1 \cap B_2}{B_1 \cup B_2} =$$



Hard Negative Mining


Pool of Negative Samples
Randomly draw M^- samples
Select $M_h^- (\ll M^-)$ samples with highest f^+ scores
A MINIBATCH
Training CNN
Pool of Positive Samples
Randomly draw M^+ samples

Not always accurate



Sources

1. https://towardsdatascience.com/understanding-ssd-multibox-real-time-object-detection-in-deep-learning-495ef744fab
2. https://d2l.ai/chapter_computer-vision/ssd.html
3. https://towardsdatascience.com/object-detection-with-neural-networks-a4e2c46b4491
4. https://www.youtube.com/watch?v=RNnKtNrsrmg
5. https://arxiv.org/pdf/1512.02325.pdf
6. https://arxiv.org/pdf/1409.1556.pdf
7. https://arxiv.org/pdf/1409.1556.pdf