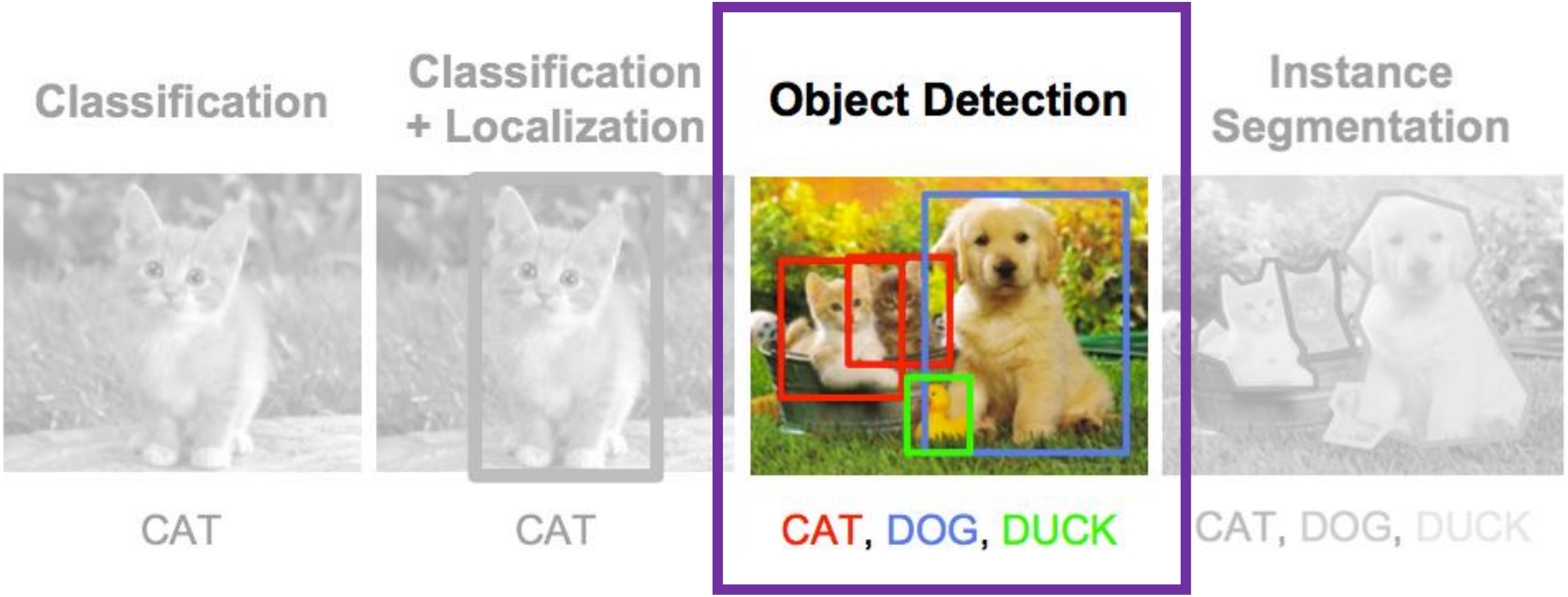
You Only Look Once

Joseph Redmon , Santosh Divvala, Ross Girshick , Ali Farhadi

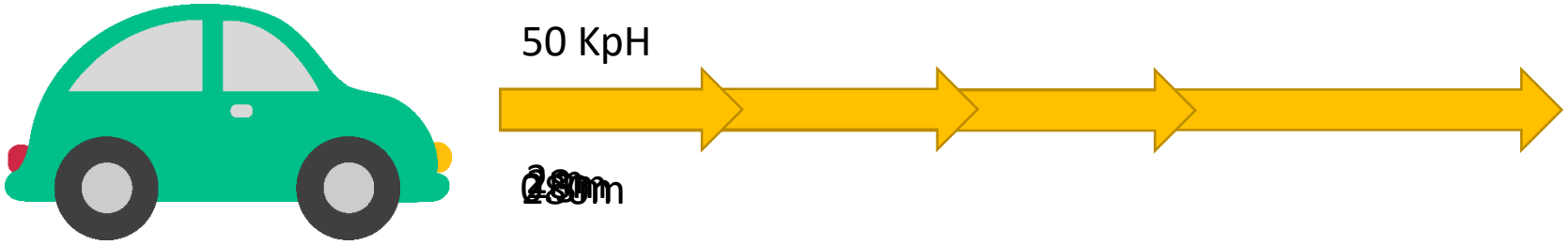
Hadar Schreiber and Lital Alyagon

Detection = Classification + Localization



Speed Improvement

	Pascal 2007 mAP	Speed	
R-CNN	66.0	.05 FPS	20 s/img
Fast R-CNN	70.0	.5 FPS	2 s/img
Faster R-CNN	73.2	7 FPS	140 ms/img
YOLO	69.0	45 FPS	22 ms/img




YOLO v2
http://pjreddie.com/yolo

Previous Classifiers

- Sliding window (VGGNet, Inception)
- Region Proposals: First predict which parts of the image contain interesting information
- In both approaches, we need to run the classifier many times

YOLO – You Only Look Once

Input image

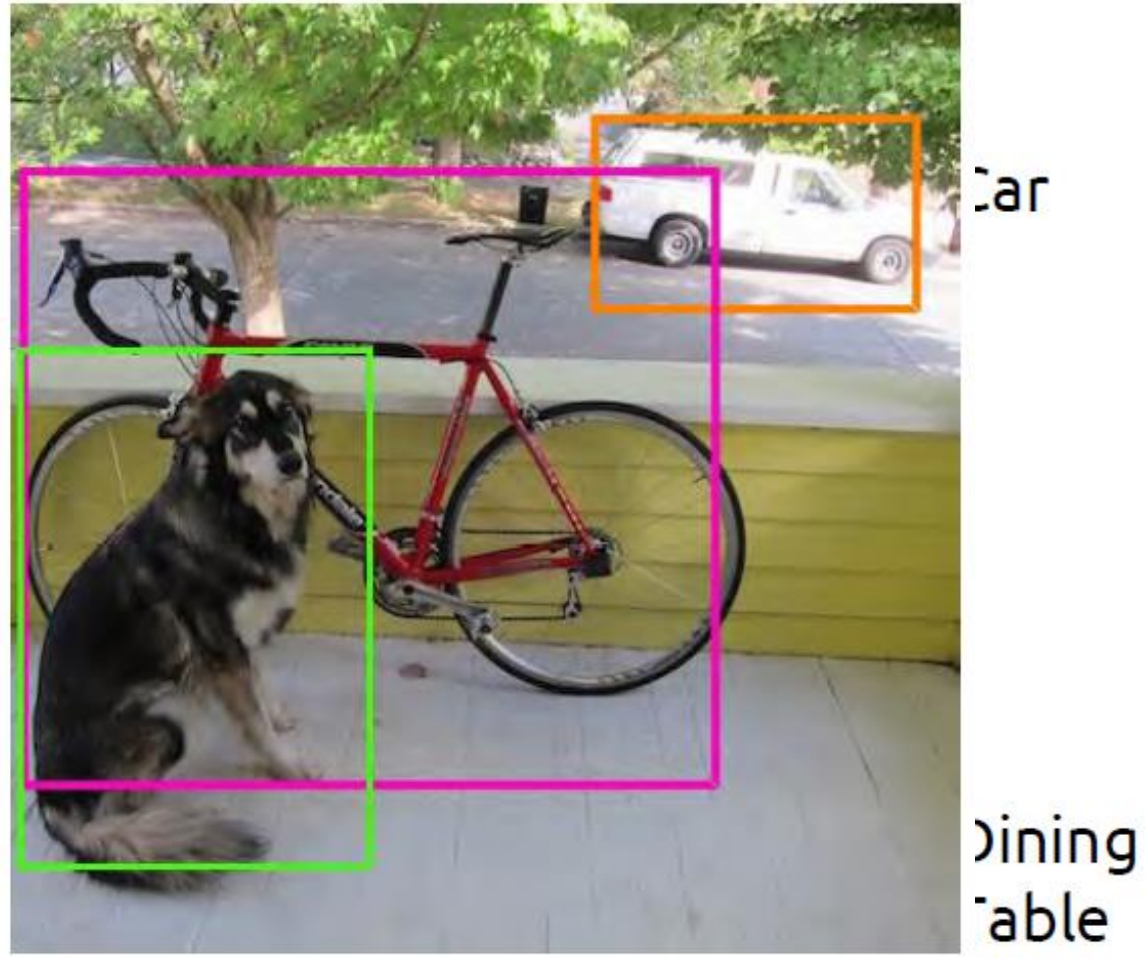
Split into grids

Create bounding boxes and predict confidence P(object) for each box

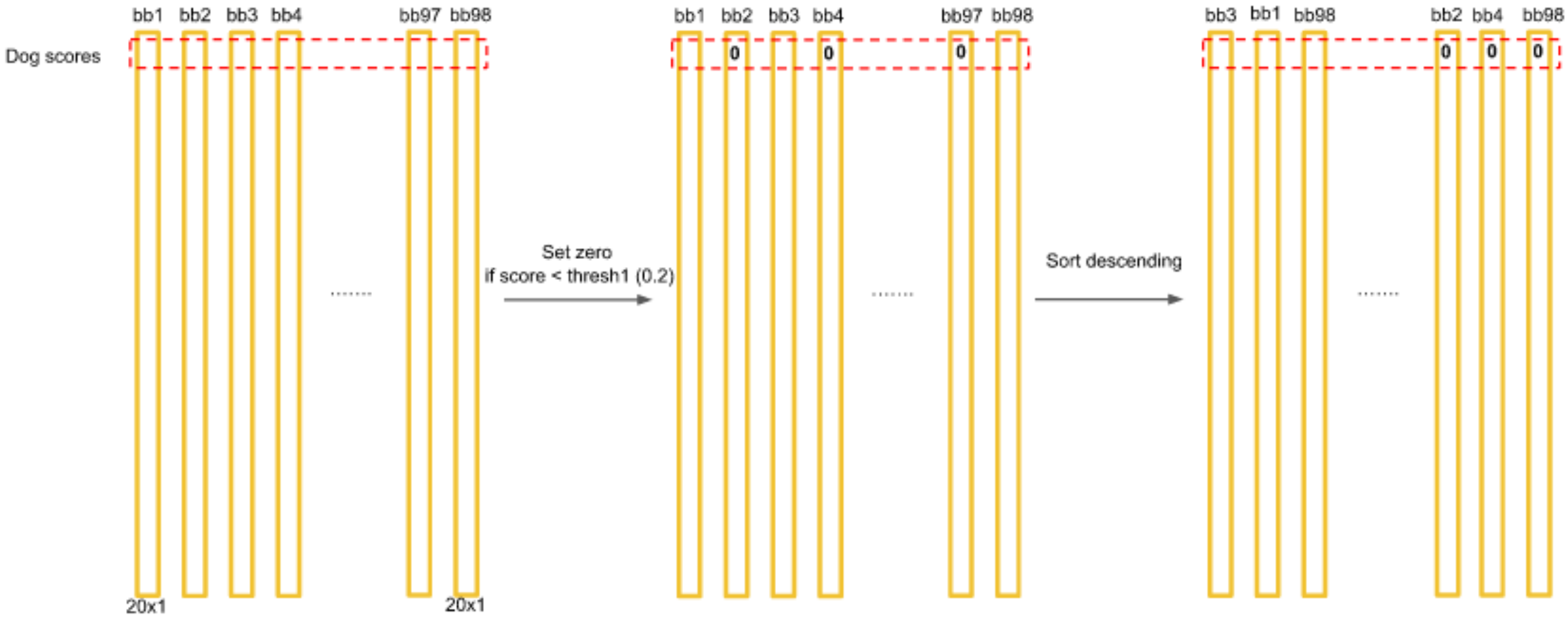
Predict class probability:
P(Class | Object) for each cell

Multiply the confidence value and the class probability

Choose best prediction using non-maximal suppression

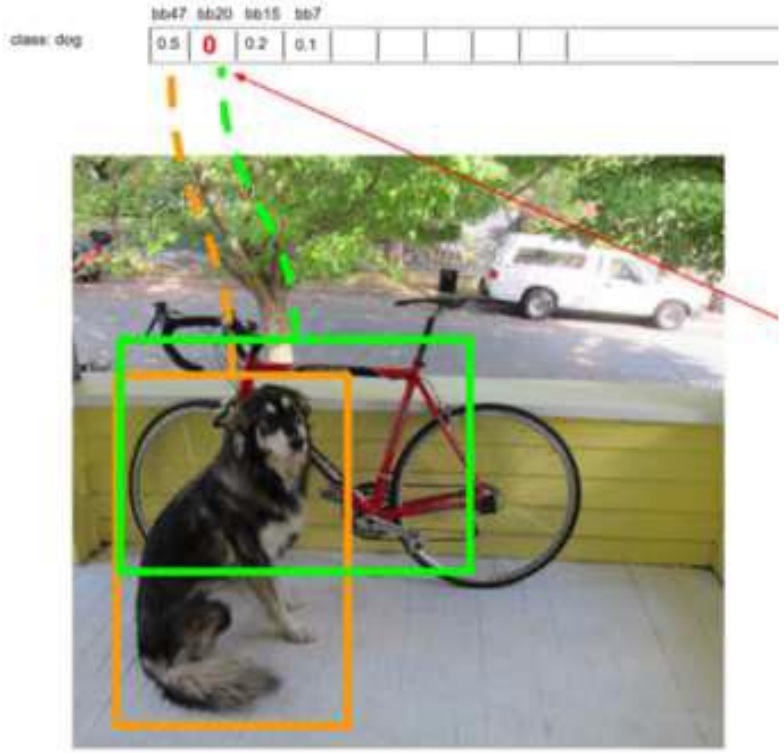


Non-maximal suppression



Non-maximal suppression

1. Start with the bounding box that has the highest score
2. Remove any remaining overlapping bounding boxes using IoU (Intersection over Union)
3. Go to step 1 until there are no more bounding boxes left



IoU > threshold (i.e 50%)

↓

Set to zero

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

YOLO – You Only Look Once

Input image

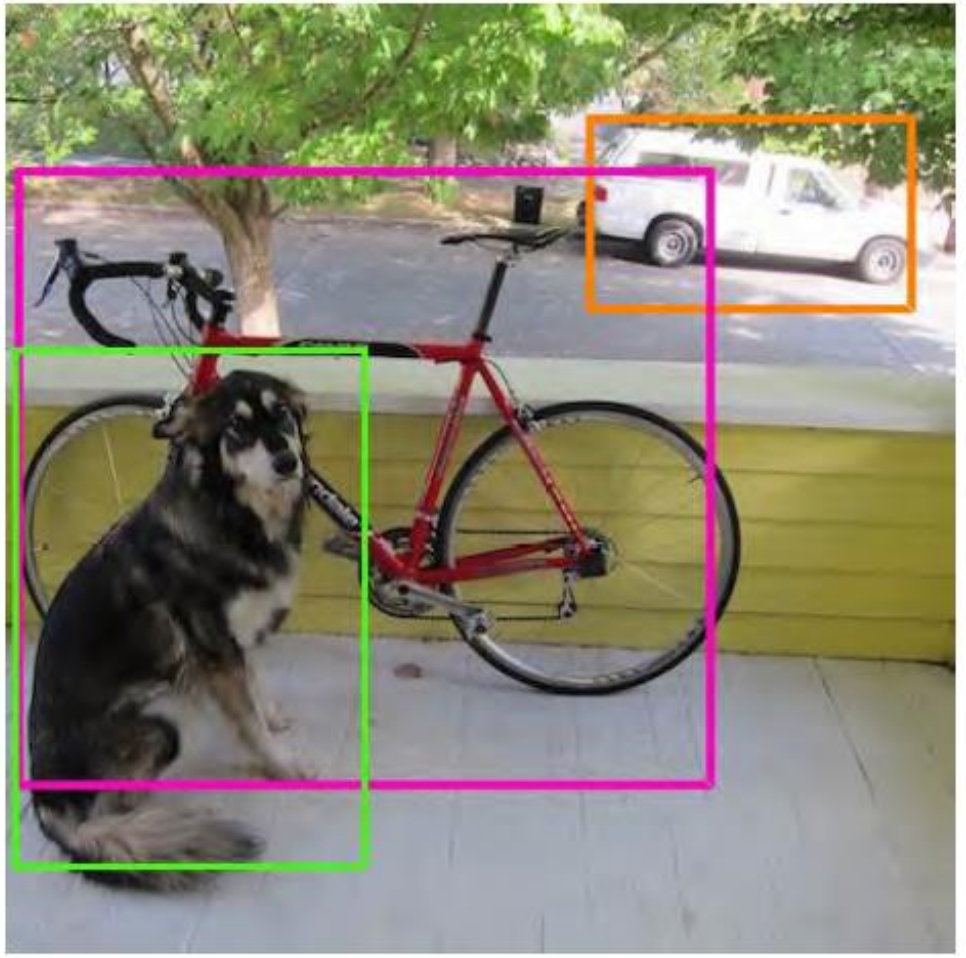
Split into grids

Create bounding boxes and predict confidence P(object) for each box

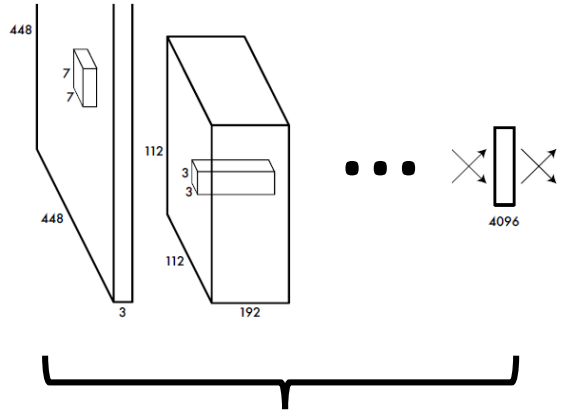
Predict class probability:
P(Class | Object) for each cell

Multiply the confidence value and the class probability

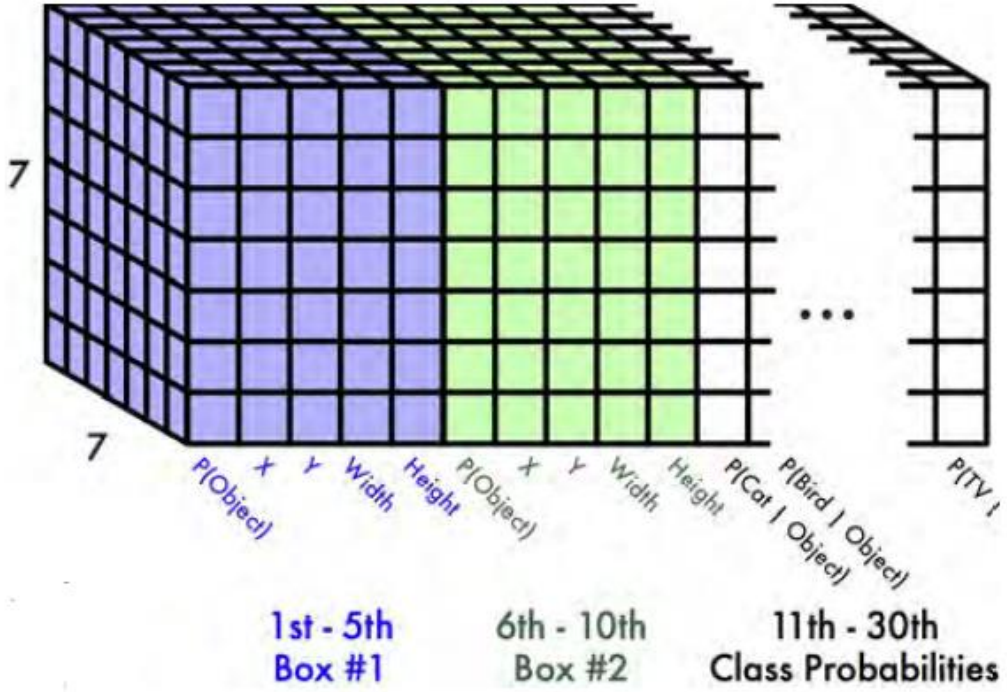
Choose best prediction using non-maximal suppression

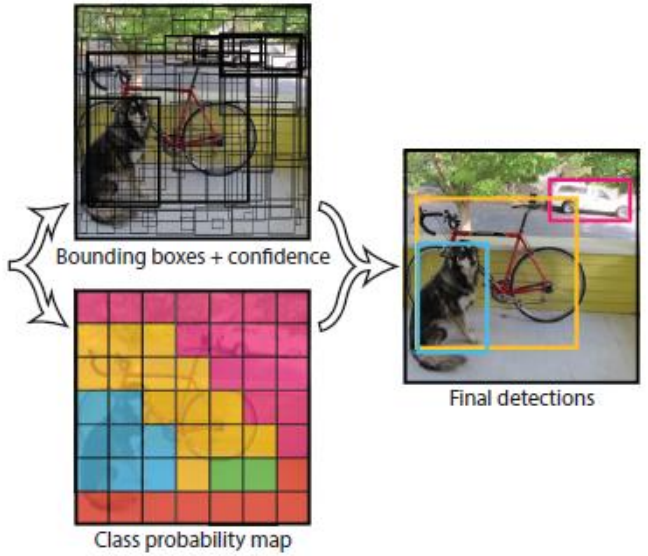


The Architecture



24 convolutional layers
2 fully connected layers





- 7 x 7 grid
- 2 bounding boxes for each cell
- 5 parameters (4 coordinates and confidence value)
- 20 classed

7 x 7 x (2 x 5 + 20) = **1470** parameters that the net needs to predict

YOLO Limitations

- Localization errors
- Low recall (the percent of the positive cases that we catch)

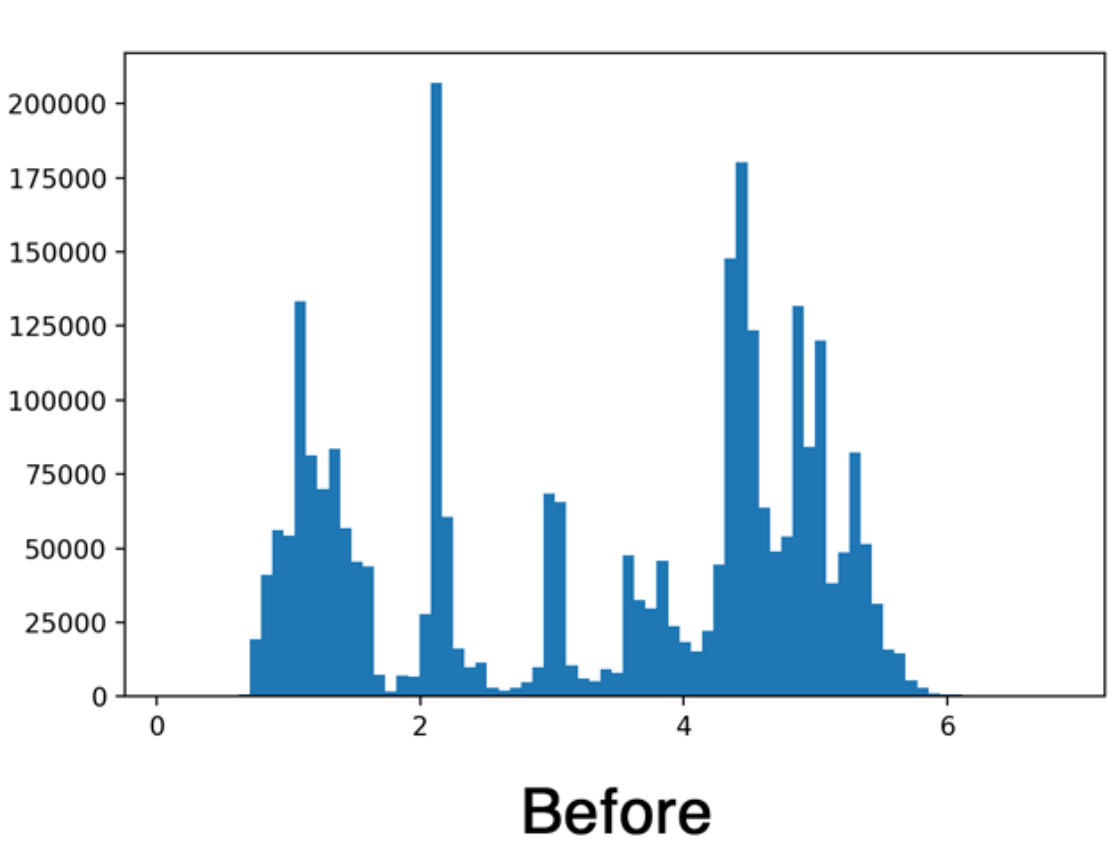
Goals:

- To create more accurate detector that still works fast
- To increase the number of detection classes

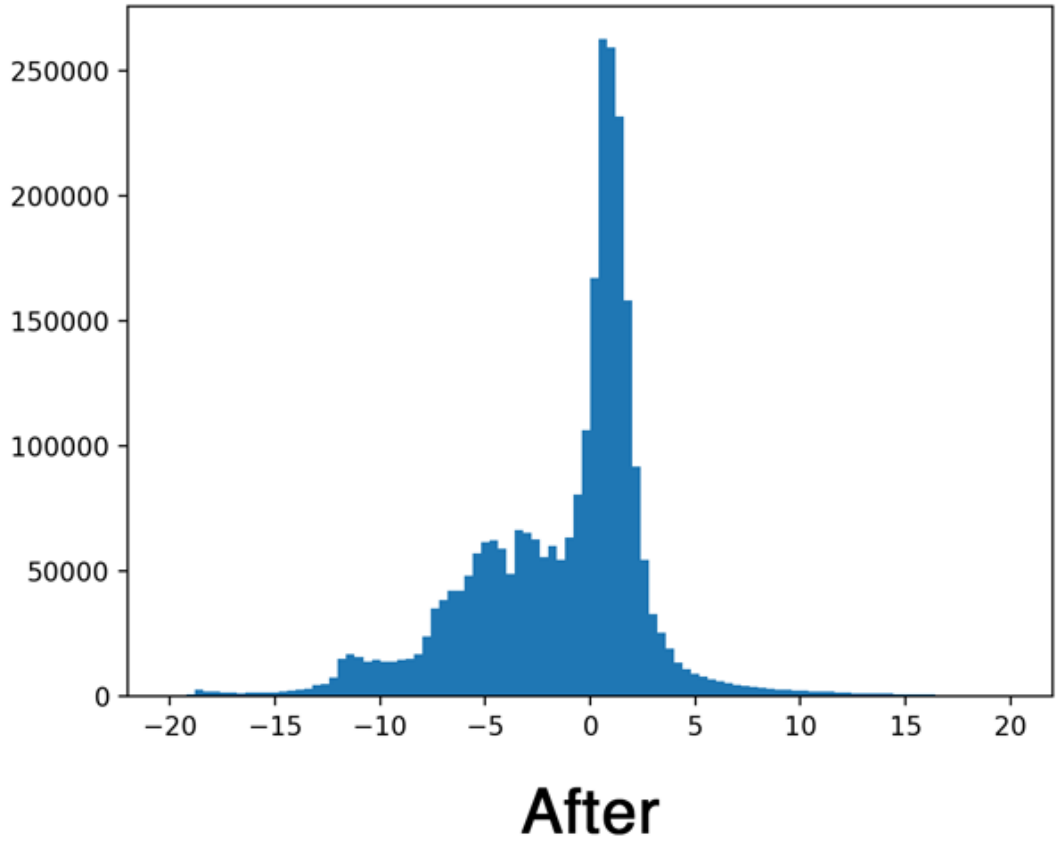

musher
malamute
JOSEPH
REDMON
ALI
FARHADI
RETURN IN.....
YOLO9000
Better, Faster,
Stronger

YOLO9000: **Better**, Faster, Stronger

- Batch normalization – increase of 4% mAP



Before



After

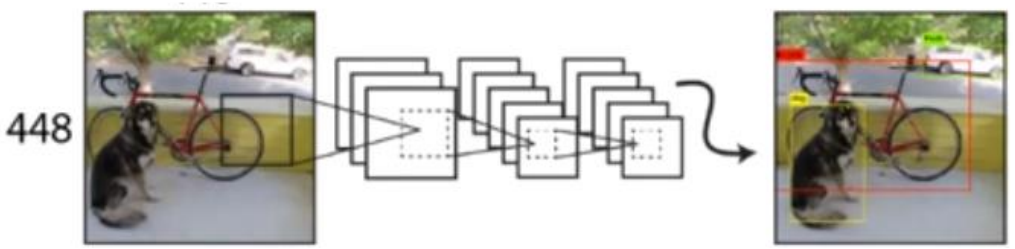
YOLO9000: **Better**, Faster, Stronger

- High resolution classifier

Train on ImageNet

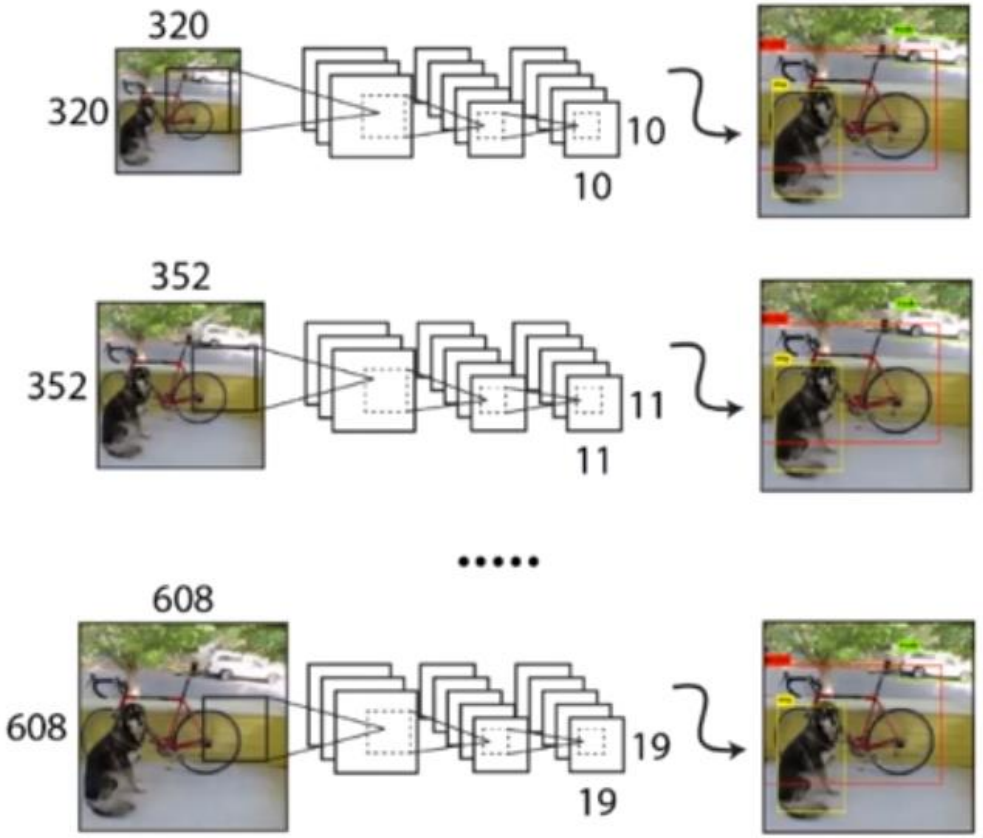
Original YOLO

Fine-tune on detection



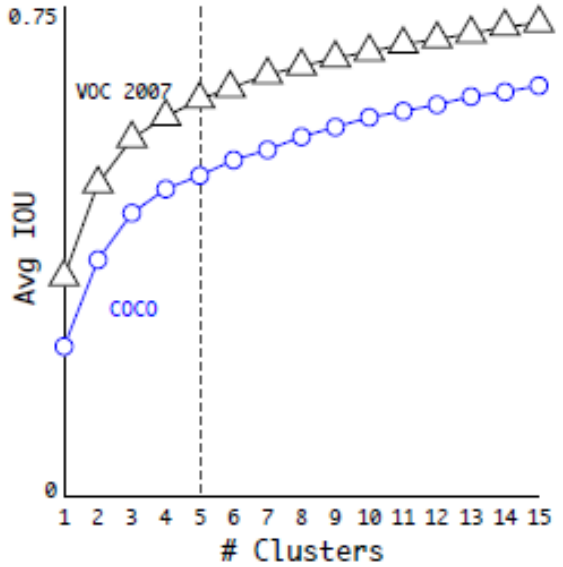
YOLO9000: **Better**, Faster, Stronger

- Multi-scale training

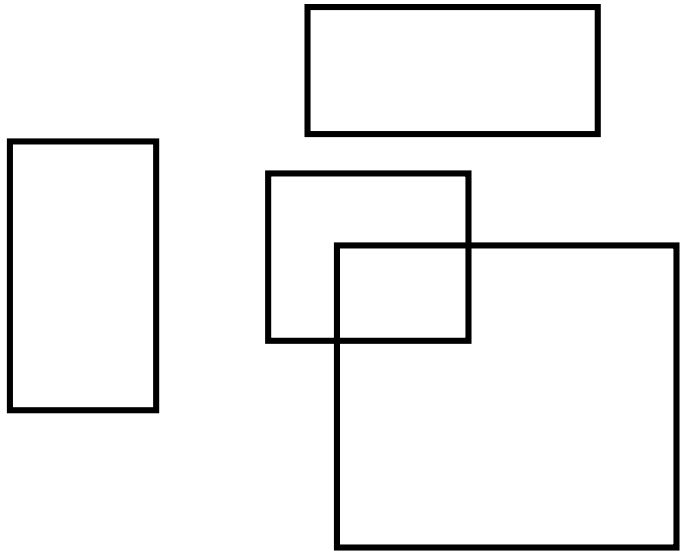


YOLO9000: **Better**, Faster, Stronger

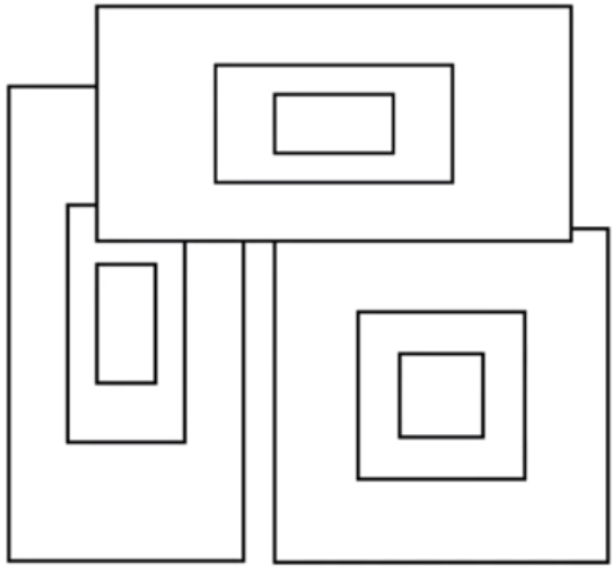
- Anchor boxes vs dimension clusters



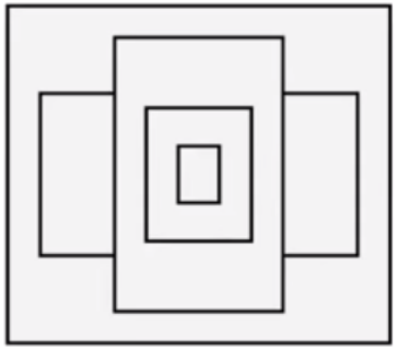
Original Yolo: Random Boxes



Anchor Boxes



Dimension Clusters



YOLO9000: Better, **Faster**, Stronger

- Darknet-19: Improved network and infrastructure
  - 19 convolutional layers
  - 5 max pooling
  - Fully connected layers removed

VGG-16: 30.69 billion FLOPs

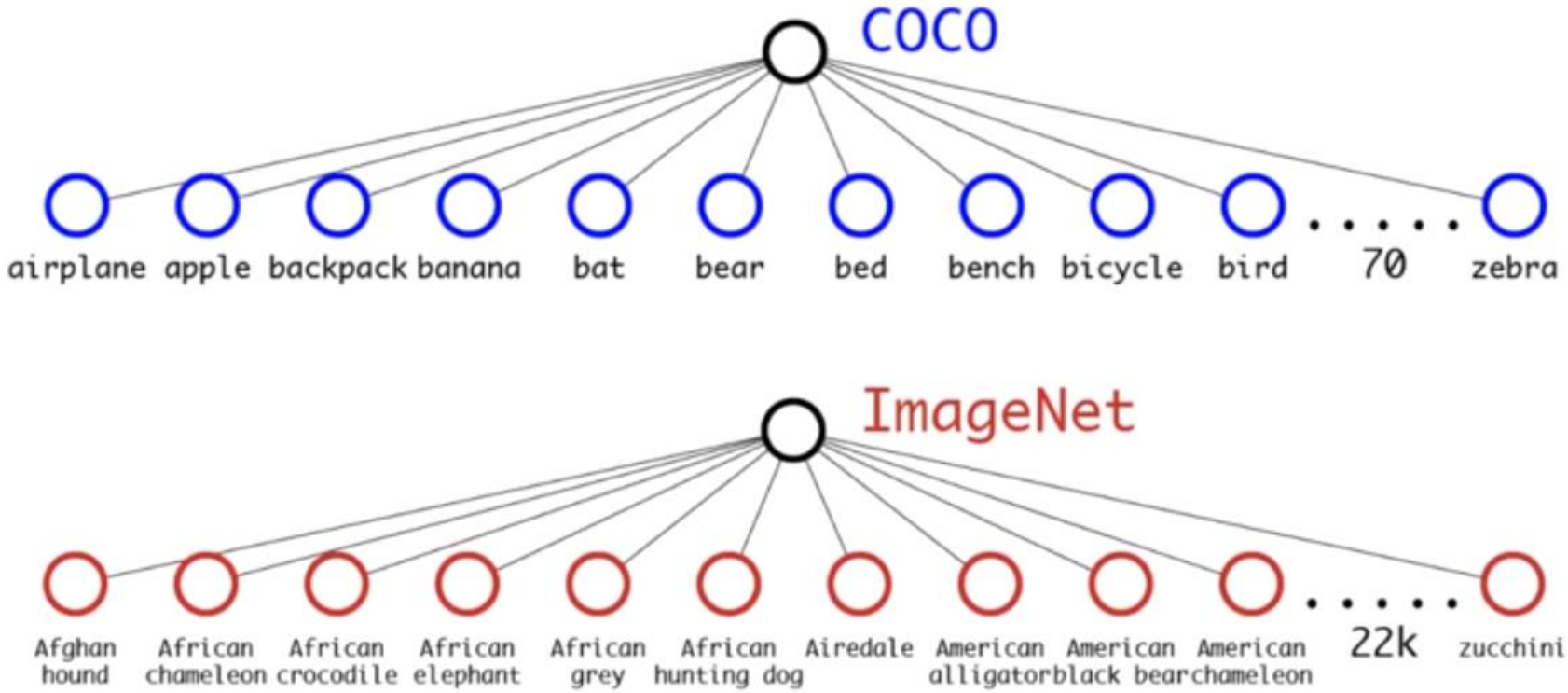
Darknet-19: 5.58 billion FLOPs

Type	Filters	Size/Stride	Output
Convolutional	32	3×3	224×224
Maxpool		$2 \times 2/2$	112×112
Convolutional	64	3×3	112×112
Maxpool		$2 \times 2/2$	56×56
Convolutional	128	3×3	56×56
Convolutional	64	1×1	56×56
Convolutional	128	3×3	56×56
Maxpool		$2 \times 2/2$	28×28
Convolutional	256	3×3	28×28
Convolutional	128	1×1	28×28
Convolutional	256	3×3	28×28
Maxpool		$2 \times 2/2$	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Maxpool		$2 \times 2/2$	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	1000	1×1	7×7
Avgpool		Global	1000
Softmax			

YOLO9000: Better, Faster, **Stronger**

Combine two data sets:

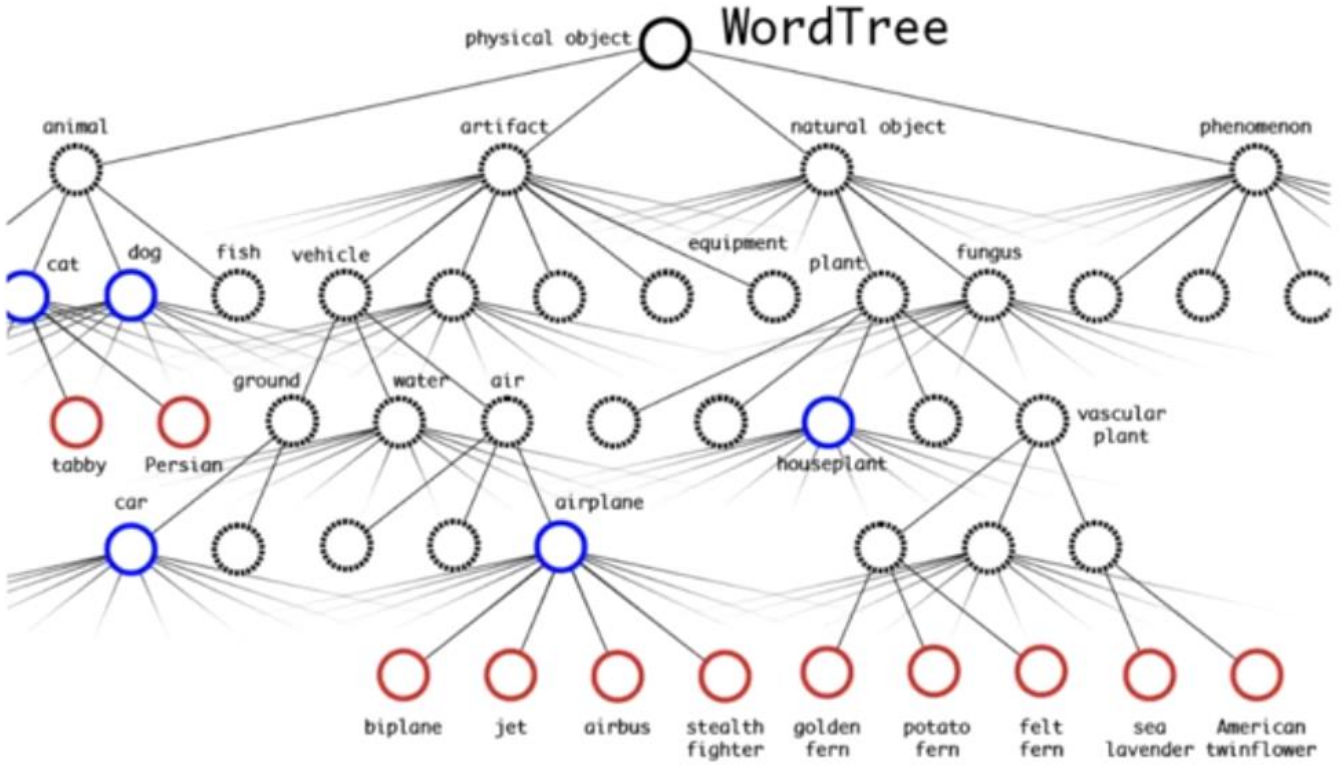
- Detection (COCO, 100K images, 80 classes)
- Classification dataset (ImageNet, 14 million images, 22k classes)



YOLO9000: Better, Faster, **Stronger**

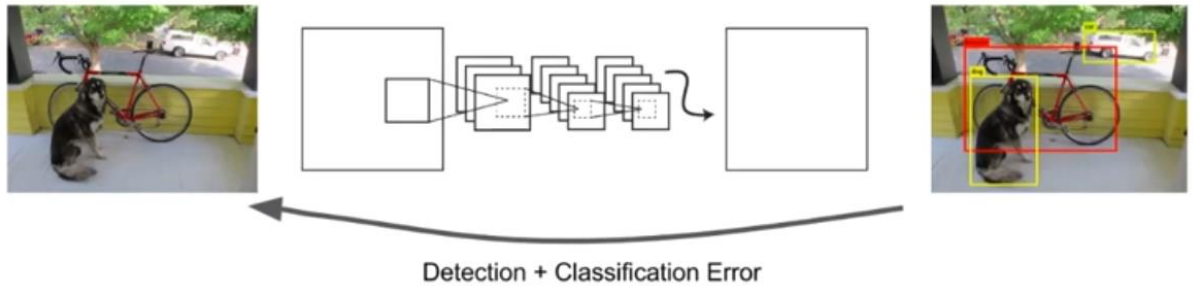
Combine two data sets:

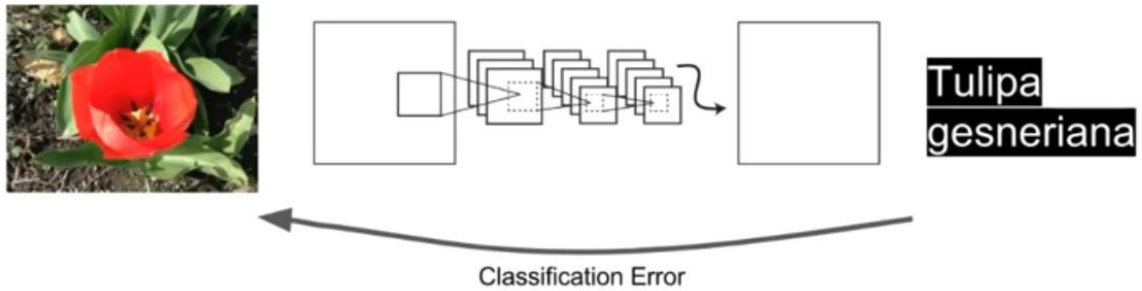
- Detection (COCO, 100K images, 80 classes)
- Classification dataset (ImageNet, 14 million images, 22k classes)



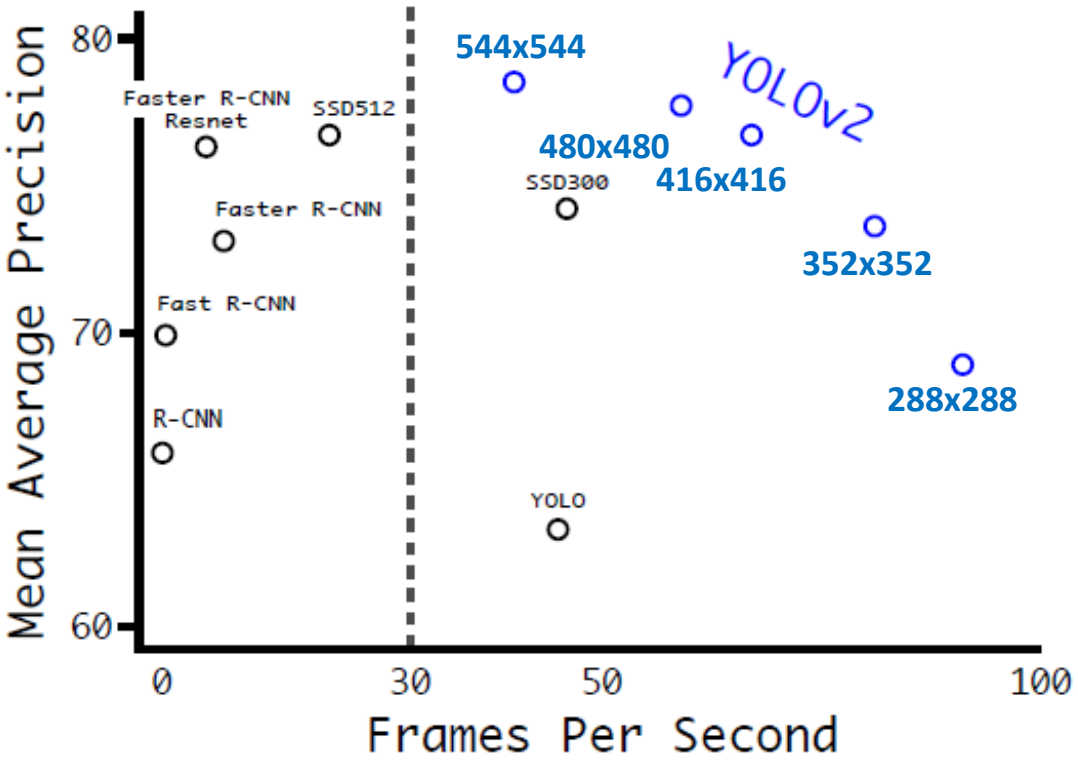
YOLO9000: Better, Faster, **Stronger**

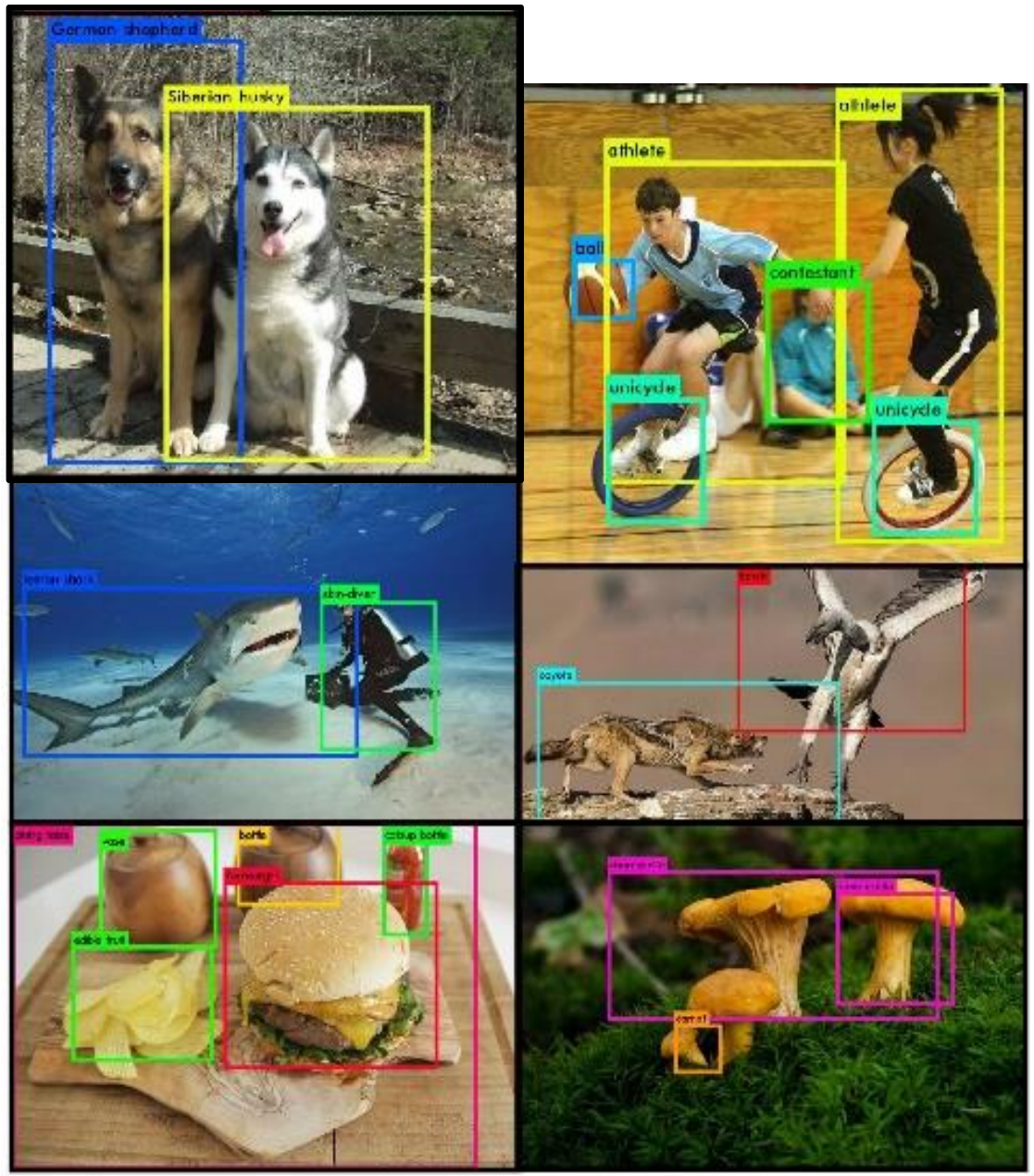
- Detection image: Backpropagate based on the full YOLOv2 loss function
- Classification image: Only backpropagate loss from the classification specific parts of the architecture





Results


Mean Average Precision
80
70
60
0
30
50
100
Frames Per Second
Faster R-CNN
Resnet
SSD512
Faster R-CNN
Fast R-CNN
R-CNN
SSD300
YOLO
544x544
480x480
416x416
352x352
288x288
YOLOv2


German shepherd
Siberian husky
athlete
athlete
ball
contestant
unicycle
unicycle

References

- Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, ***You Only Look Once: Unified, Real-Time Object Detection***, 2016 IEEE Conference on Computer Vision and Pattern Recognition
- Joseph Redmon, Ali Farhadi, ***YOLO9000: Better, Faster, Stronger***, CVPR 2017

Questions?


Coconut chocolate
White chocolate
Milk chocolate
Dark chocolate
Orange chocolate

Results

Detection Frameworks	Train	mAP	FPS
Fast R-CNN [5]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[15]	2007+2012	73.2	7
Faster R-CNN ResNet[6]	2007+2012	76.4	5
YOLO [14]	2007+2012	63.4	45
SSD300 [11]	2007+2012	74.3	46
SSD500 [11]	2007+2012	76.8	19
YOLOv2 288 × 288	2007+2012	69.0	91
YOLOv2 352 × 352	2007+2012	73.7	81
YOLOv2 416 × 416	2007+2012	76.8	67
YOLOv2 480 × 480	2007+2012	77.8	59
YOLOv2 544 × 544	2007+2012	**78.6**	40