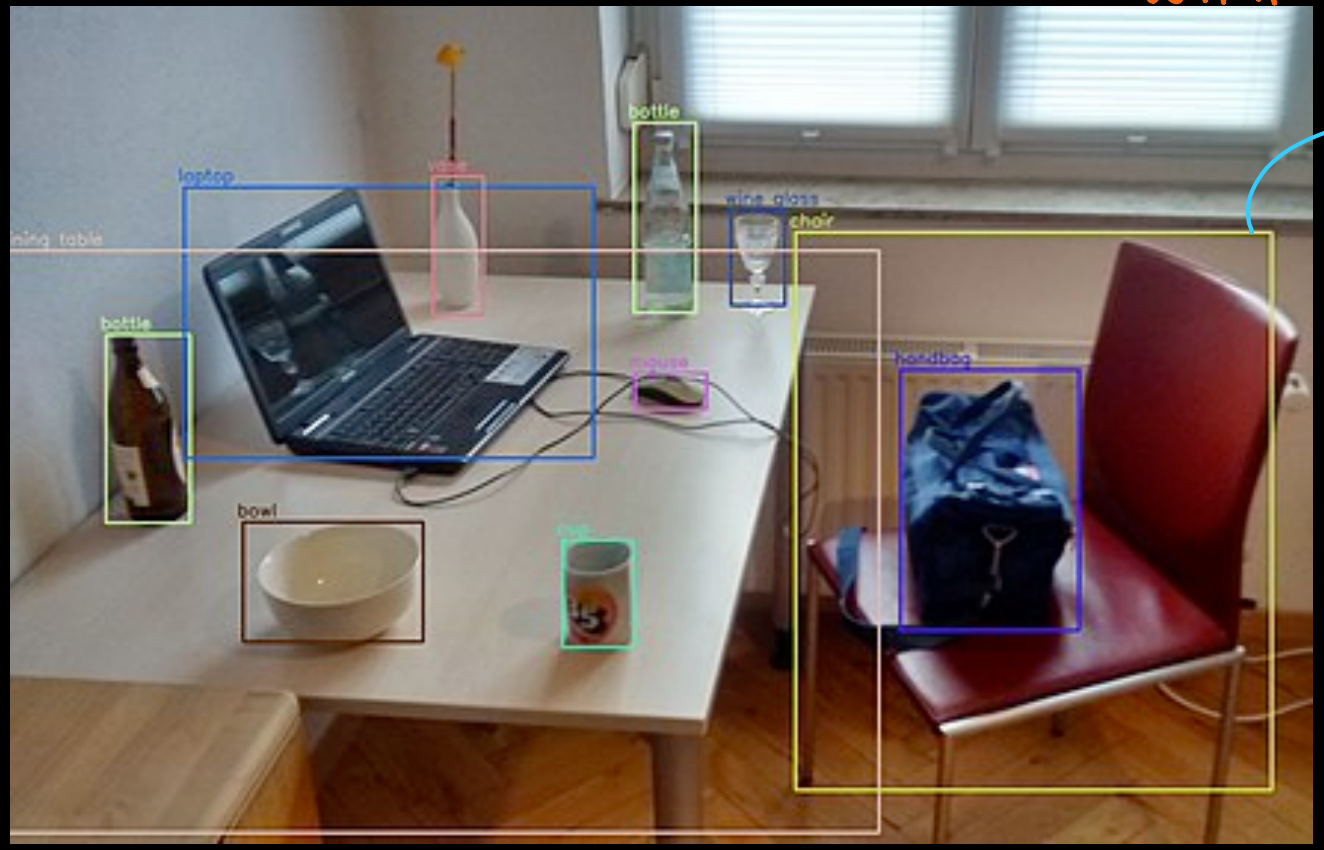
Object - detection

Agenda

1. Problem formulation
2. Azure API-based (+code)
3. Overview of Deep-Learning based models
4. YoLO V3 dive-deep
    a. Theory
    b. C-code

Problem - formulation :

wiki



→ bounding boxes

(NOT pixels)

↳ multiple-objects in single image

APPLIED COURSE

Using Azure-API

– Pre-req: Using Web-APIs in Python

Oveerview:
https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-object-detection

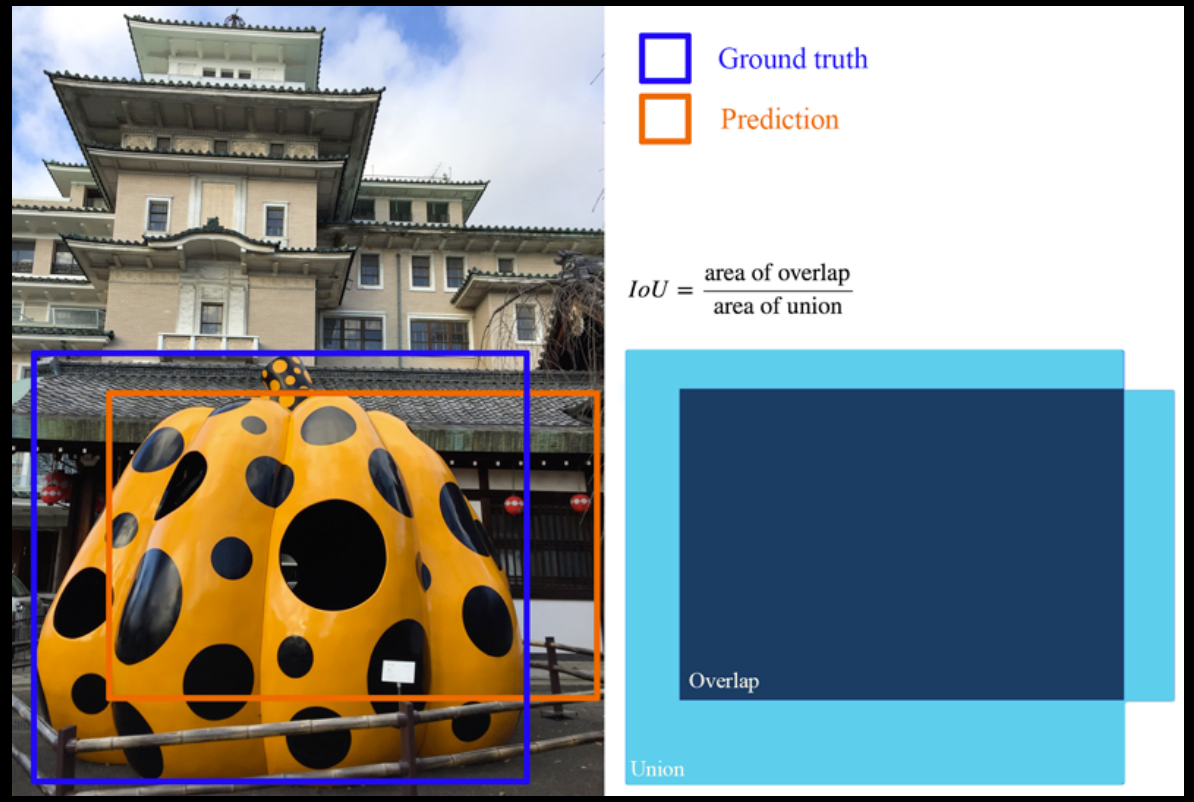
Documentation:
https://westcentralus.dev.cognitive.microsoft.com/docs/services/5adf991815e1060e6355ad44/operations/56f91f2e778daf14a499e1fa

Python code:
https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/quickstarts/python-analyze

```
print(json.dumps(response.json()))
```

(no 'objects' in visualFeatures)

Performance - metric :



https://medium.com/@jonathan_hui/map-mean-average-precision-for-object-detection-45c121a31173

Prediction is correct if

$IoU \geq 0.5$

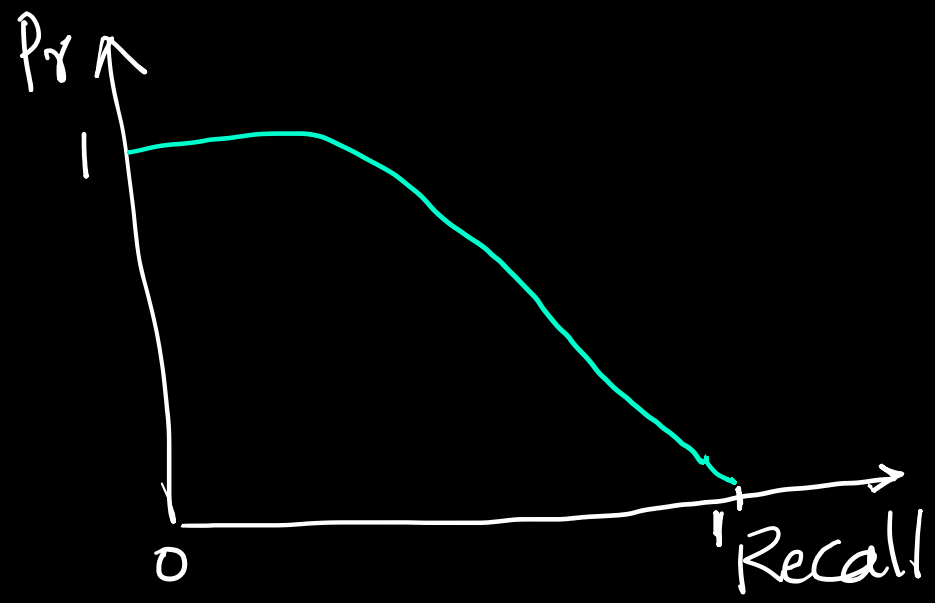
Performance - metric :

Now, for each class (chair, person, ...),

compute avg-precision

↓

AU - PR Curve



Performance - Metric :

mean-avg-precision (mAP) → NOT Max-A-posterion (MAP) in stats.

↓

mean across all classes

Deep-Learning - models:

input : image / video

output : bounding-box-1, object-class-1

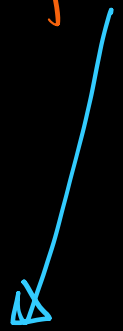
bounding-box-2, object - class-2

⋮

Dataset : COCO http://cocodataset.org/#home [80-class-data]

Trade off : Speed VS mAP

medical diagnosis



self-driving cars

real-time face-detection

R-CNN

Fast R-CNN

Faster R-CNN

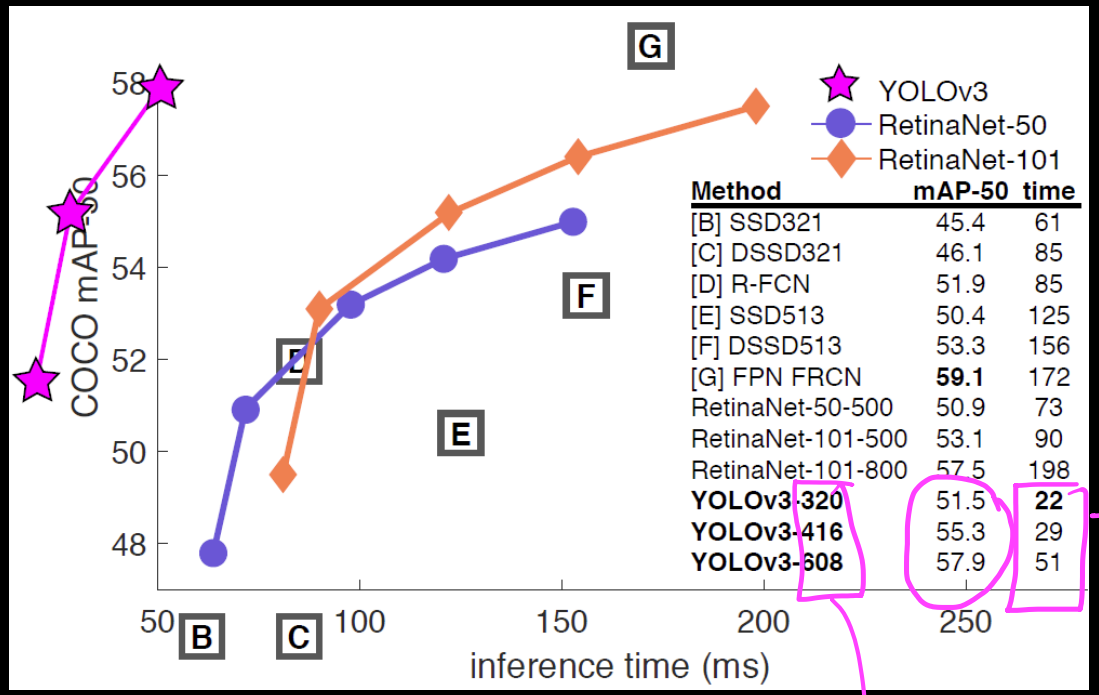
SSD

RetinaNet

YOLO V1

Yolo V2 / YOLO 9000

Yolo V3 — APR-2018



Method	mAP-50	time
[B] SSD321	45.4	61
[C] DSSD321	46.1	85
[D] R-FCN	51.9	85
[E] SSD513	50.4	125
[F] DSSD513	53.3	156
[G] FPN FRCN	**59.1**	172
RetinaNet-50-500	50.9	73
RetinaNet-101-500	53.1	90
RetinaNet-101-800	57.5	198
YOLOv3-320	51.5	**22**
YOLOv3-416	55.3	29
YOLOv3-608	57.9	51

→ Super-fast + V. good mAP

input - image-size

APPLIED COURSE

YOLO - V3: feature-extractor (1)

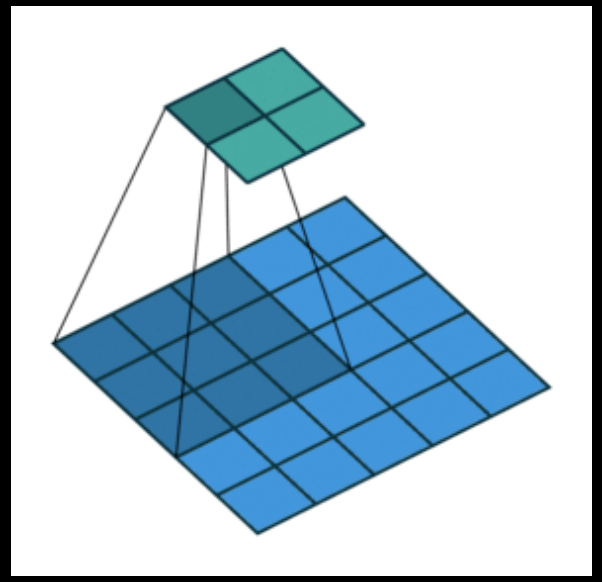
- Fully convolutional Network (FCN)
- conV - BatchNorm - LeakyReLU
- conv with stride - downsample
- no pooling
- pre-trained model

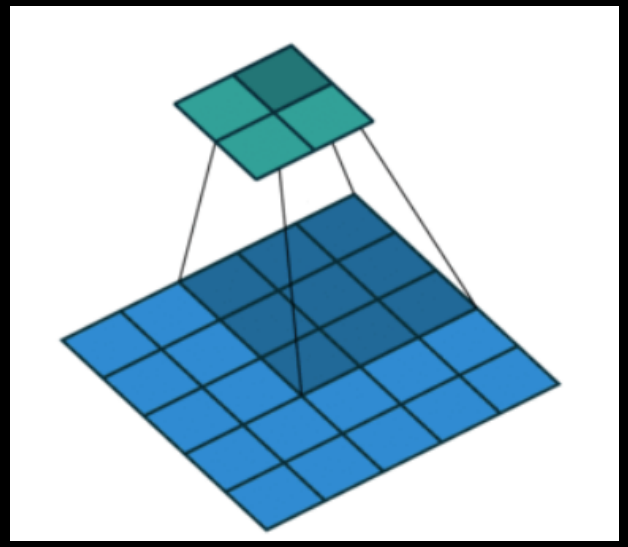
→ stride-2-conv

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

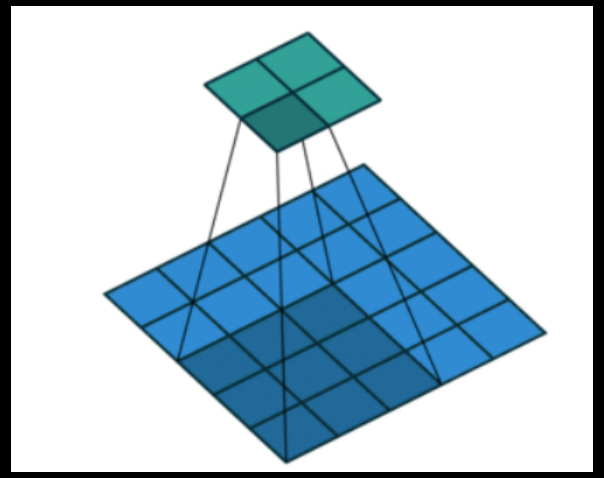
Darknet-53 model

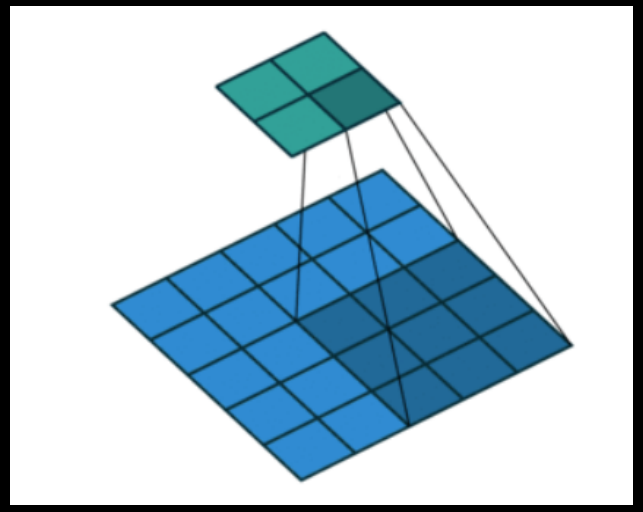
https://pjreddie.com/media/files/papers/YOLOv3.pdf





Stride-2 convolutions





5x5 —3x3 kernels→ 2x2

https://github.com/vdumoulin/conv_arithmetic

a.k.a → feature-extractor

Backbone	Top-1	Top-5	Bn Ops	BFLOP/s	FPS
Darknet-19 [15]	74.1	91.8	7.29	1246	**171**
ResNet-101[5]	77.1	93.7	19.7	1039	53
ResNet-152 [5]	**77.6**	**93.8**	29.4	1090	37
Darknet-53	77.2	**93.8**	18.7	**1457**	78

used in YOLOV2

Slower than Darknet-19

more op/sec

V.good

input = images :

416 x 416 x 3

feature-extracted : 13 x 13 x 1024

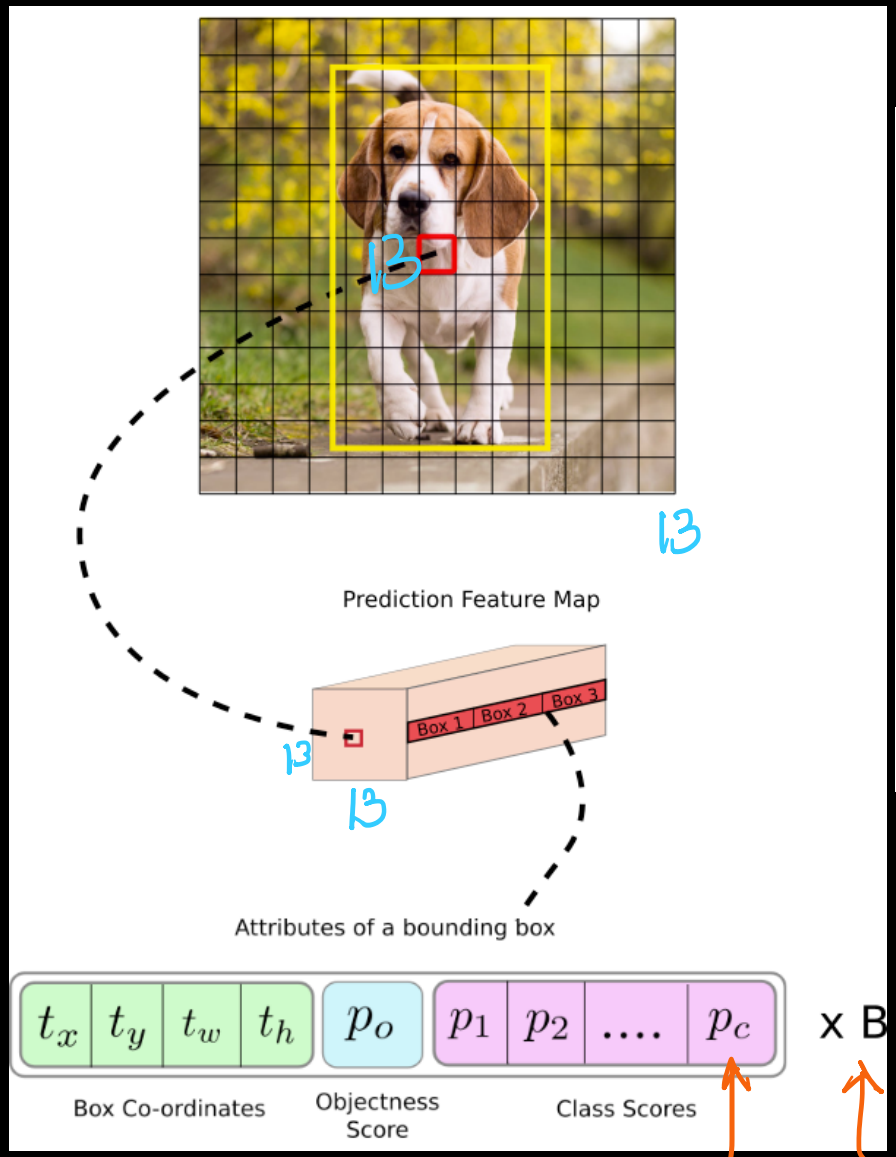
	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

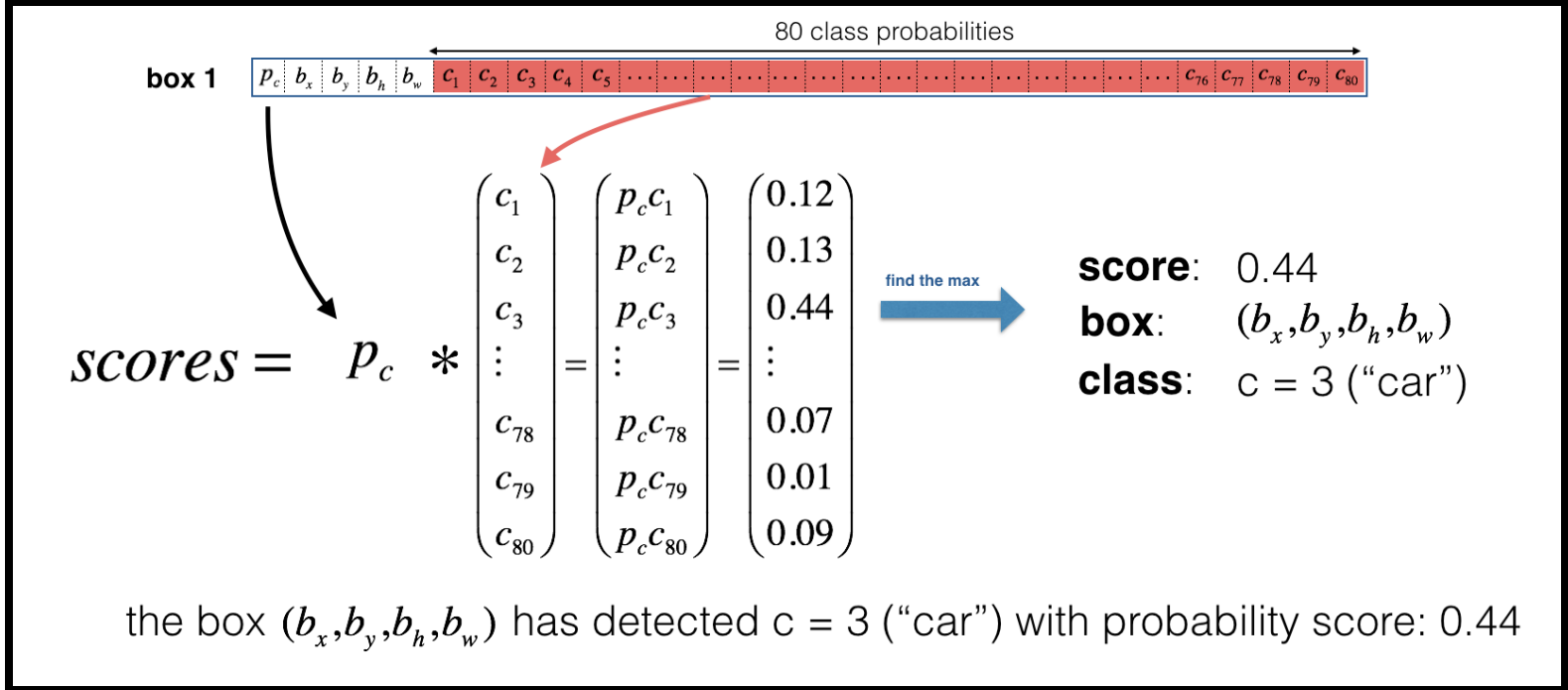
Darknet-53 model

/32

② Bounding-boxes & output

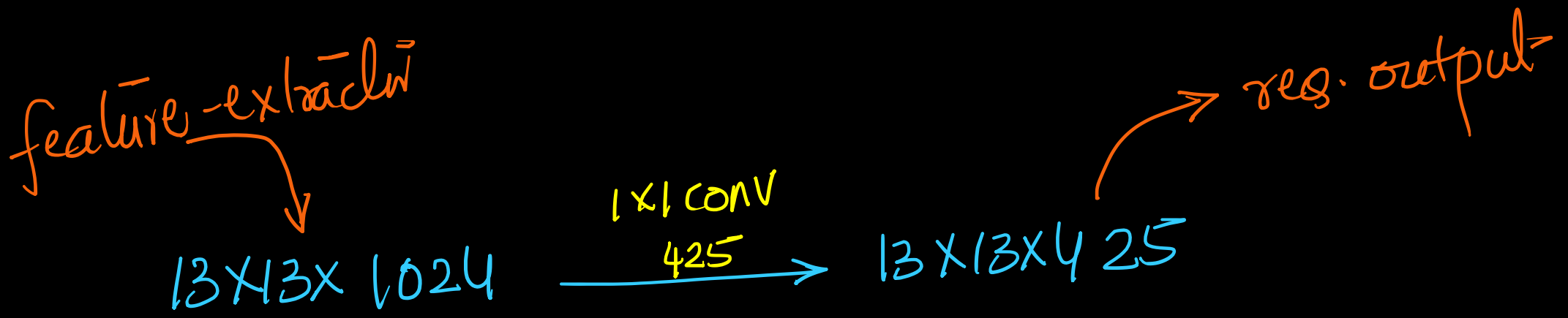
$13 \times 13 \times (B \cdot (5+80)) \rightarrow 425$ if $B=5$





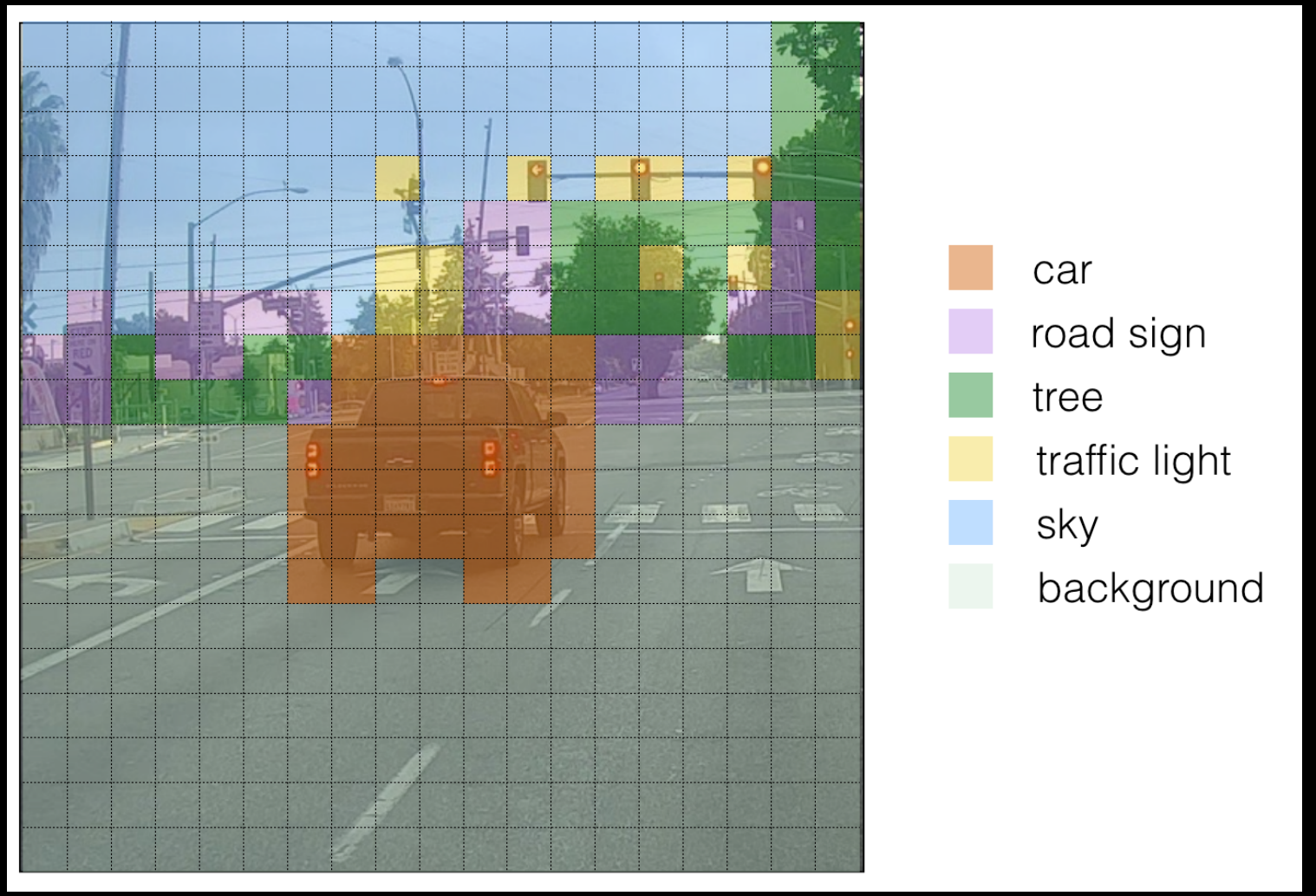
https://pylessons.com/YOLOv3-introduction/

APPLIED COURSE

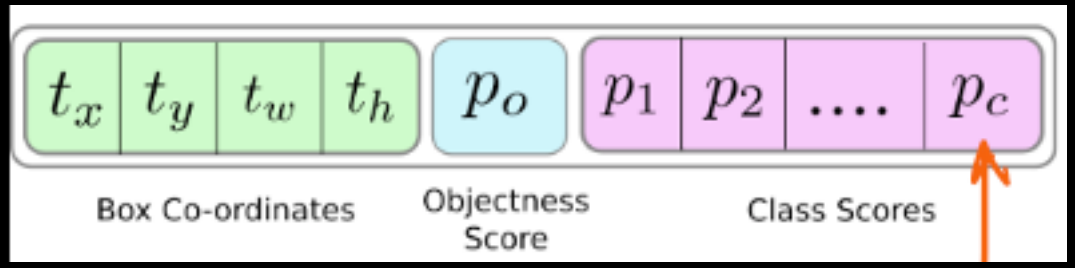


Refer: https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3412/inception-network/8/module-8-neural-networks-computer-vision-and-deep-learning

intermediate - result



bounding box representation



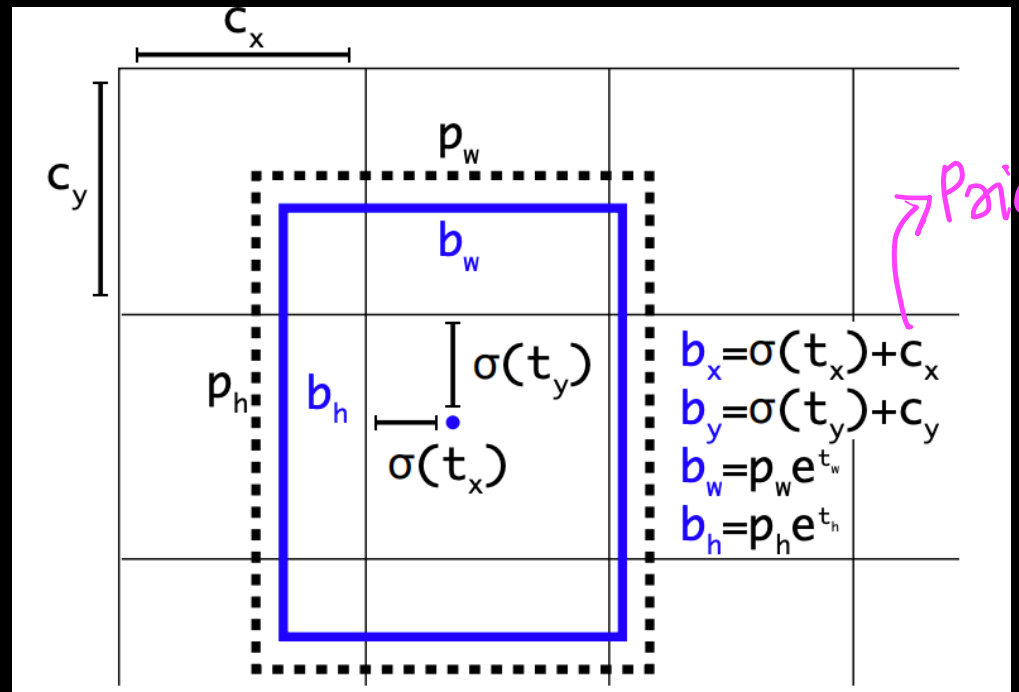


Figure 2. **Bounding boxes with dimension priors and location prediction.** We predict the width and height of the box as offsets from cluster centroids. We predict the center coordinates of the box relative to the location of filter application using a sigmoid function. This figure blatantly self-plagiarized from [15].

https://pjreddie.com/media/files/papers/YOLOv3.pdf

ANCHOR-boxes:

[Cx, Cy, Pw, Ph]

APPLIED COURSE

ANCHOR - boxes [Cx, Cy, Pw, Ph]

→ K-means Clustering on train data's bounding boxes.

→ K=5 ⇒ 5 anchor-boxes

↓

5 bounding boxes per cell

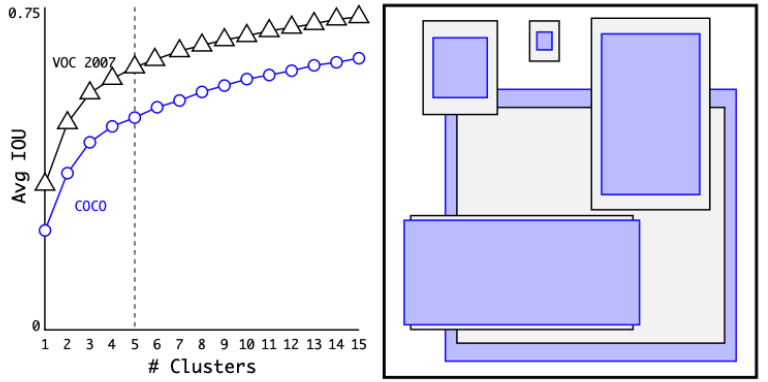
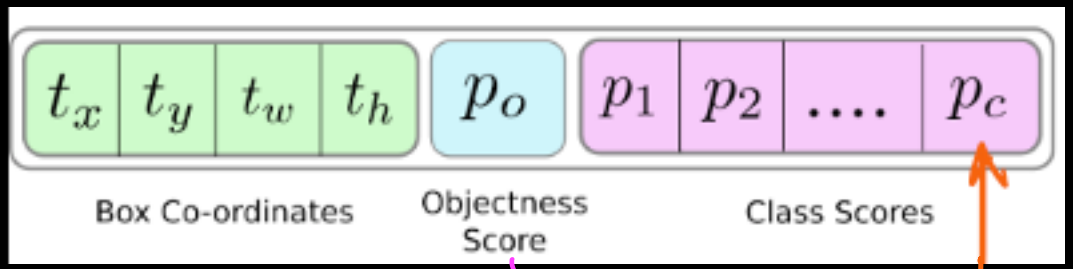


Figure 2: Clustering box dimensions on VOC and COCO. We run k-means clustering on the dimensions of bounding boxes to get good priors for our model. The left image shows the average IOU we get with various choices for k. We find that $k = 5$ gives a good tradeoff for recall vs. complexity of the model. The right image shows the relative centroids for VOC and COCO. Both sets of priors favor thinner, taller boxes while COCO has greater variation in size than VOC.

③ Per-class Sigmoids



p (box contains C_i object)

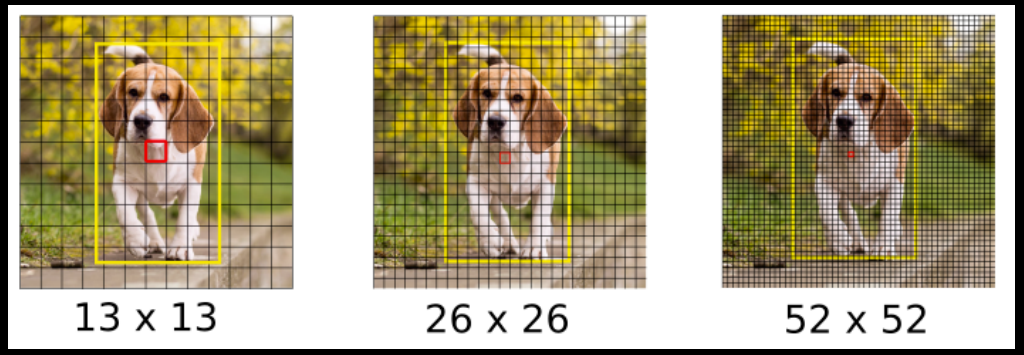
$= p_o \times p_i$

Is there any object at all

e.g: person C_{10}, woman C_{13} [softmax doesn't work]

APPLIED COURSE

④ Multi-scale Prediction



→ detect smaller objects

→ ideas from feature pyramid networks (FPN)

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Darknet-53 model

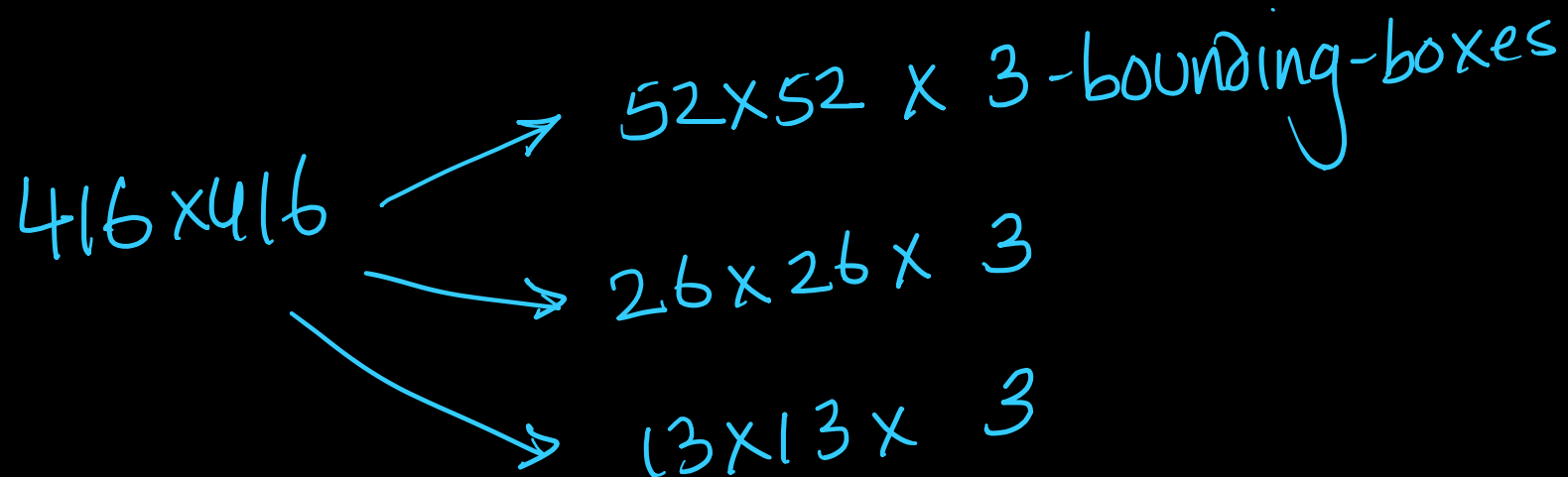
→ 52x52x256

→ 26x26x512

→ 13x13x1024

APPLIED COURSE

(5) Combining boxes from various scales

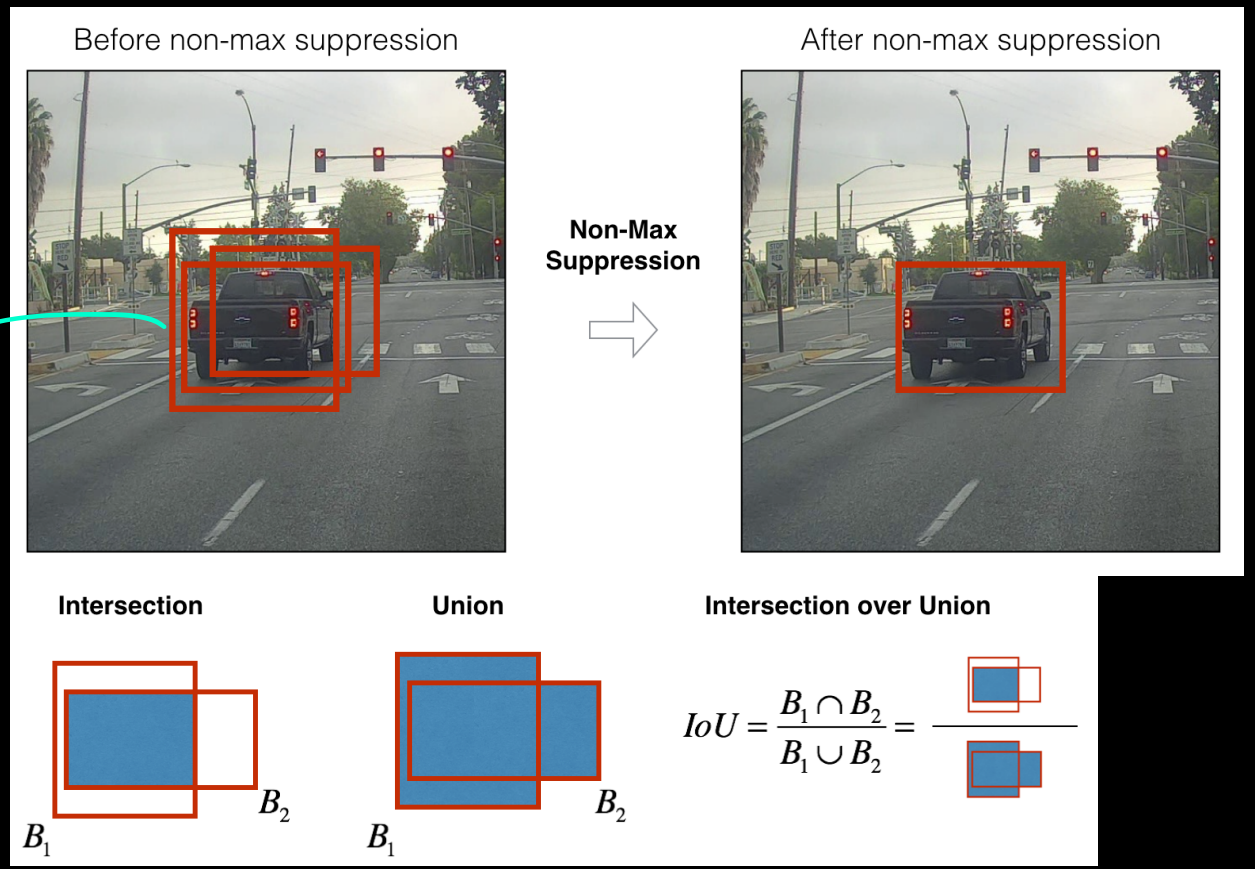


10,647 boxes

ⓐ Filter : boxes with $\max(p_o \cdot p_i) < 0.5$ are ignored

ⓑ Non-max-Suppression (NMS)

pick the box with max IoU



Code :

https://pjreddie.com/darknet/yolo/ → fast C- implementation

Keras- implementation : https://github.com/qqwweee/keras-yolo3

↳ not as-fast