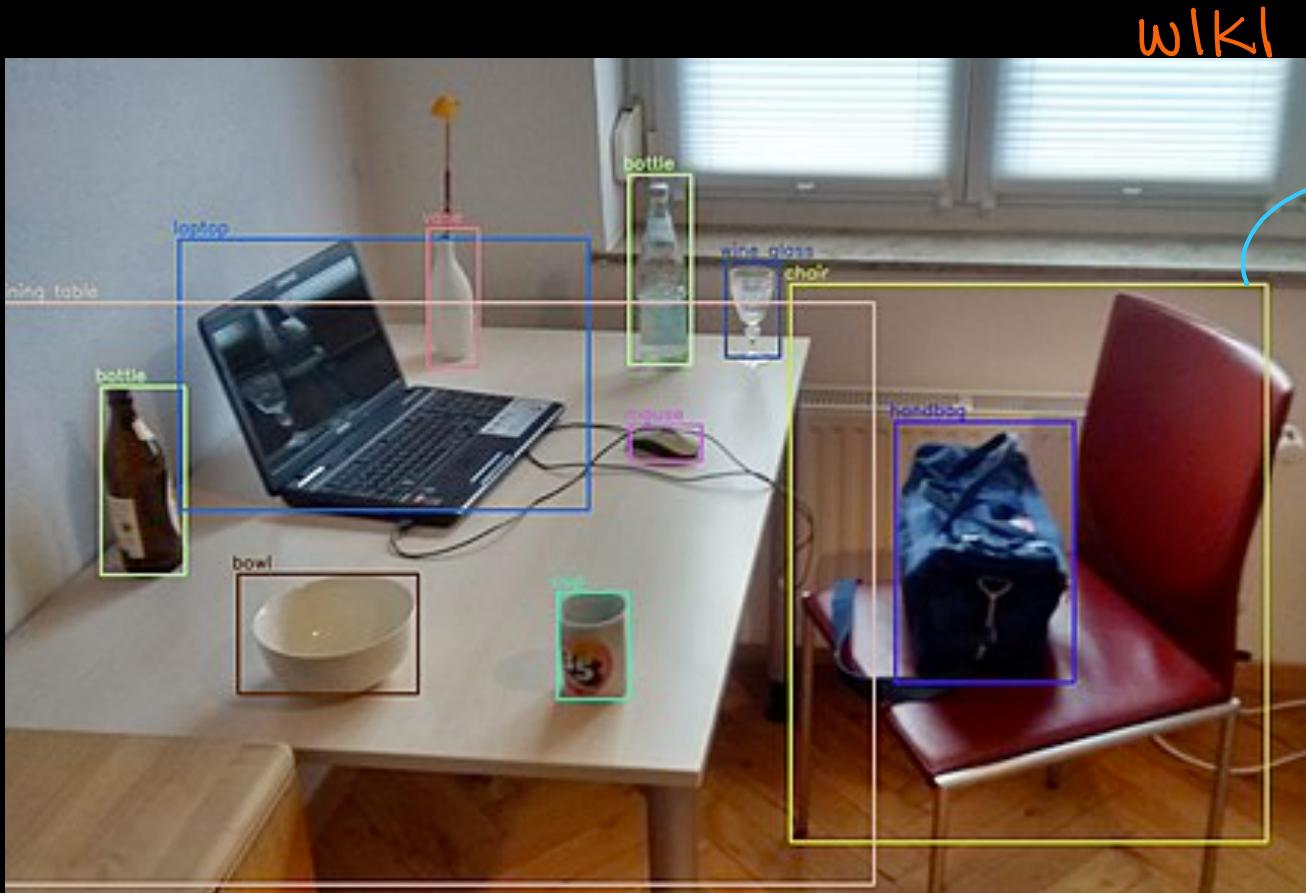


Object-detection

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

Using Azure-API

- Pre-req: Using Web-APIs in Python

Overview:

<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>

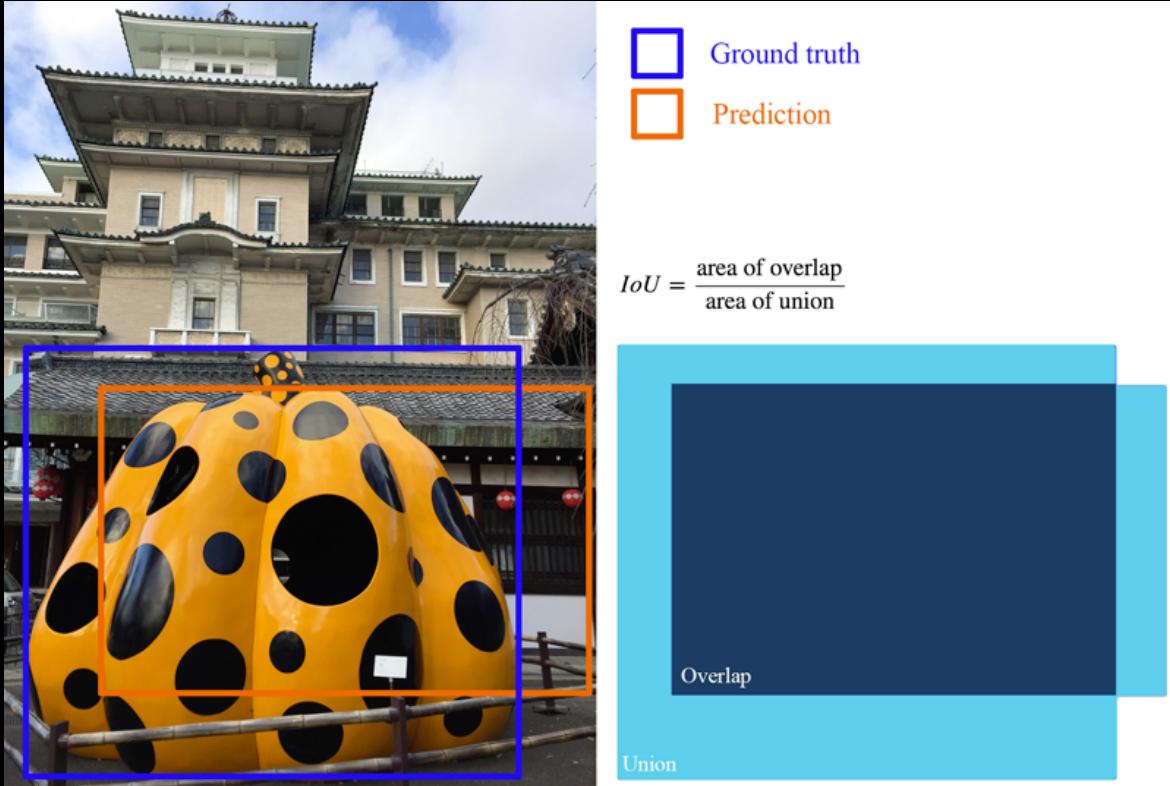
Python code:

<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

$IoU \geq 0.5$

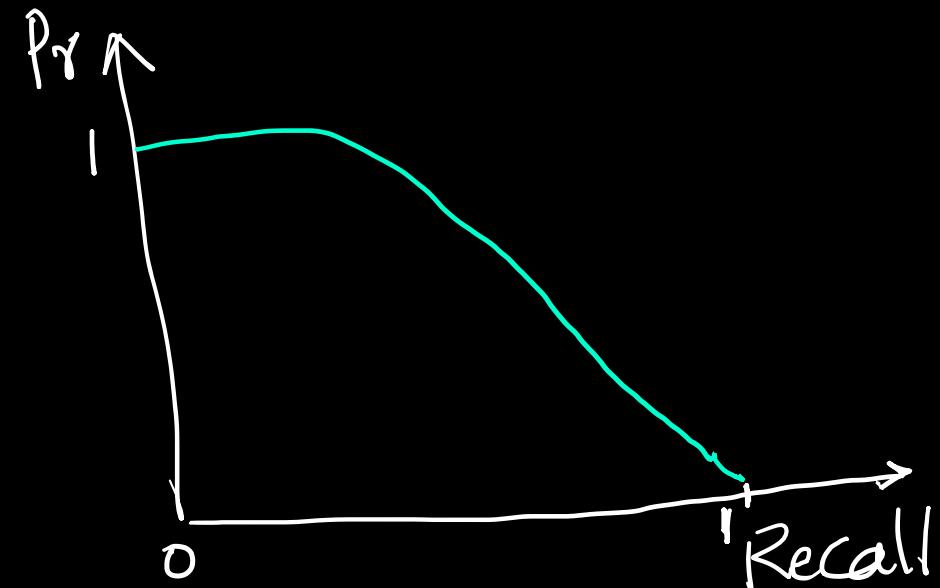
Performance - Metric :

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

compute avg-precision



AU-PR Curve



Performance - Metric :

mean - avg - precision (mAP) → NOT Max - A - posterior
(MAP)



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

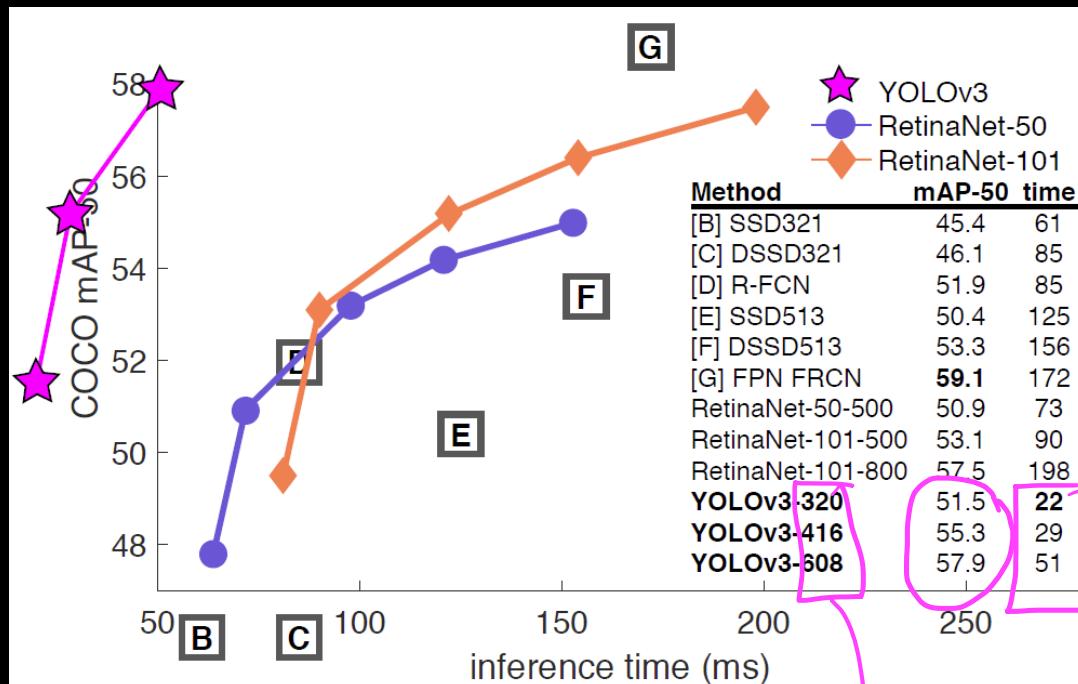
RetinaNet

YOLO V1

YOLO V2 / YOLO 9000

YOLO V3

APR-2018



input - image size ↗

Super-fast + v-good mAP

YOLO - V3: feature-extractor

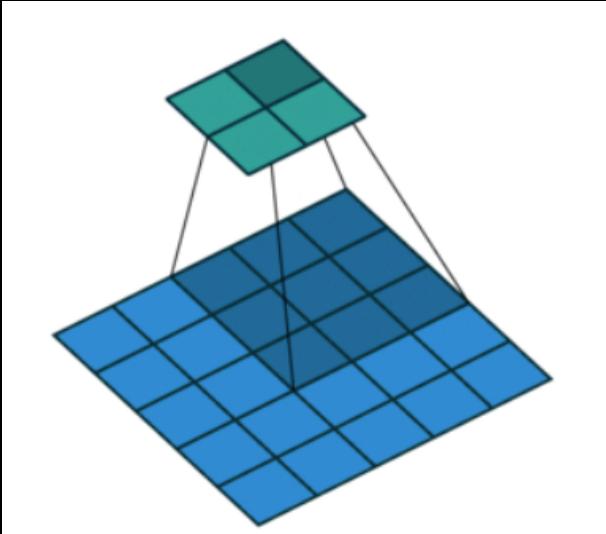
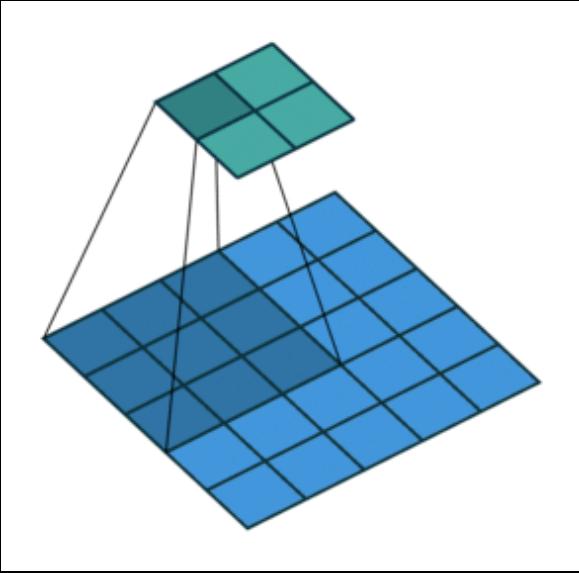
①

- fully convolutional Network (FCN)
- CONV - BatchNorm - LeakyReLU
- CONV with stride - downsample
- NO pooling
- pre-trained model

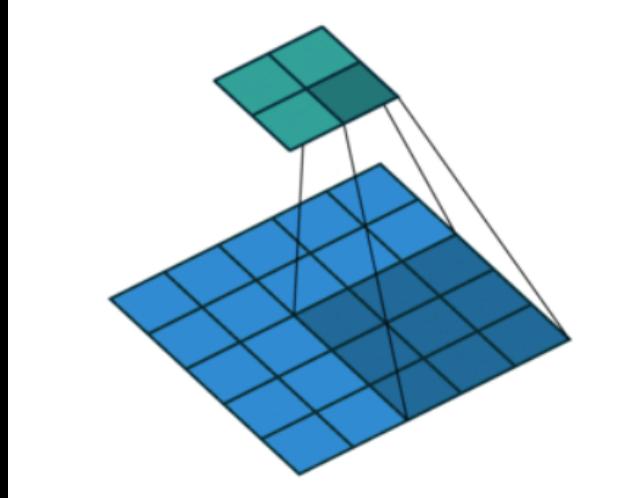
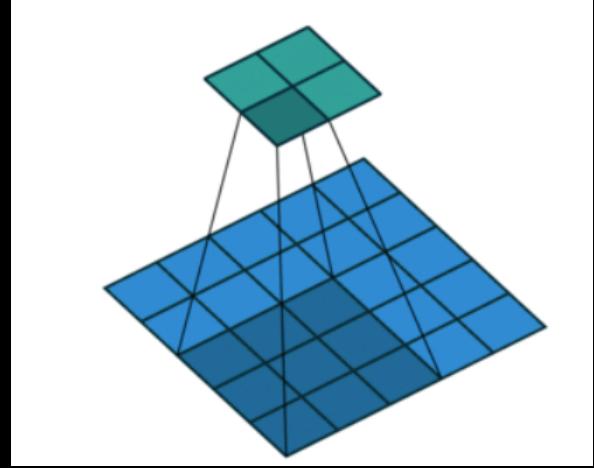
Type	Filters	Size	Output
Convolutional	32	3×3	256×256
Convolutional	64	$3 \times 3 / 2$	128×128
1x Convolutional	32	1×1	
1x Convolutional	64	3×3	
		Residual	128×128
Convolutional	128	$3 \times 3 / 2$	64×64
Convolutional	64	1×1	
2x Convolutional	128	3×3	
		Residual	64×64
Convolutional	256	$3 \times 3 / 2$	32×32
Convolutional	128	1×1	
8x Convolutional	256	3×3	
		Residual	32×32
Convolutional	512	$3 \times 3 / 2$	16×16
Convolutional	256	1×1	
8x Convolutional	512	3×3	
		Residual	16×16
Convolutional	1024	$3 \times 3 / 2$	8×8
Convolutional	512	1×1	
4x 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 $\xrightarrow{3 \times 3 \text{ kernels}} 2 \times 2$

https://github.com/vdumoulin/conv_arithmetic

a.k.a \rightarrow feature-extractor

Used
in
YOLOv2

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 \rightarrow V.good

Slowest (Darknet-19)
More Ops/sec

input - Images :

416 × 416 × 3

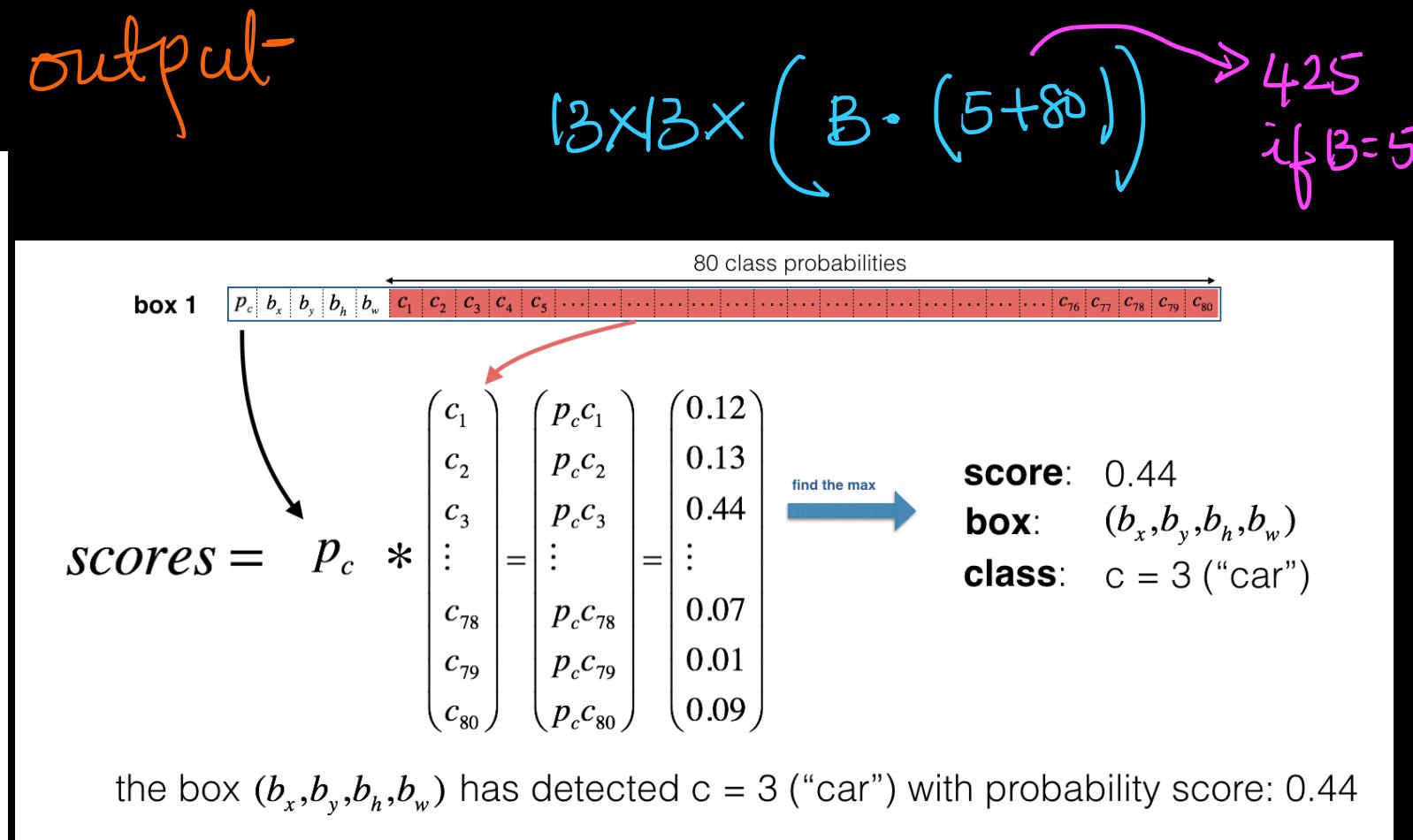
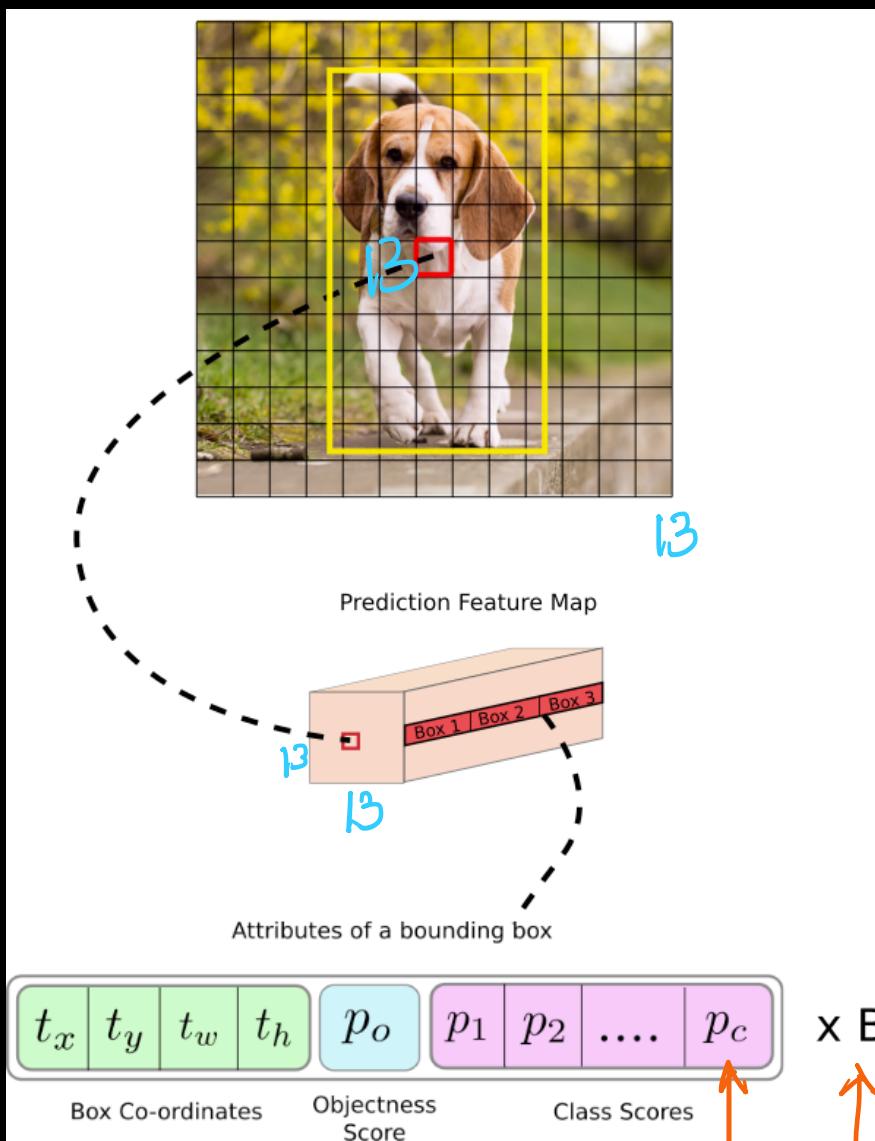
feature-extracted : 13 × 13 × 1024

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

Darknet-53 model

/32

② Bounding-boxes & output



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

feature-extractin

$13 \times 13 \times 1024$

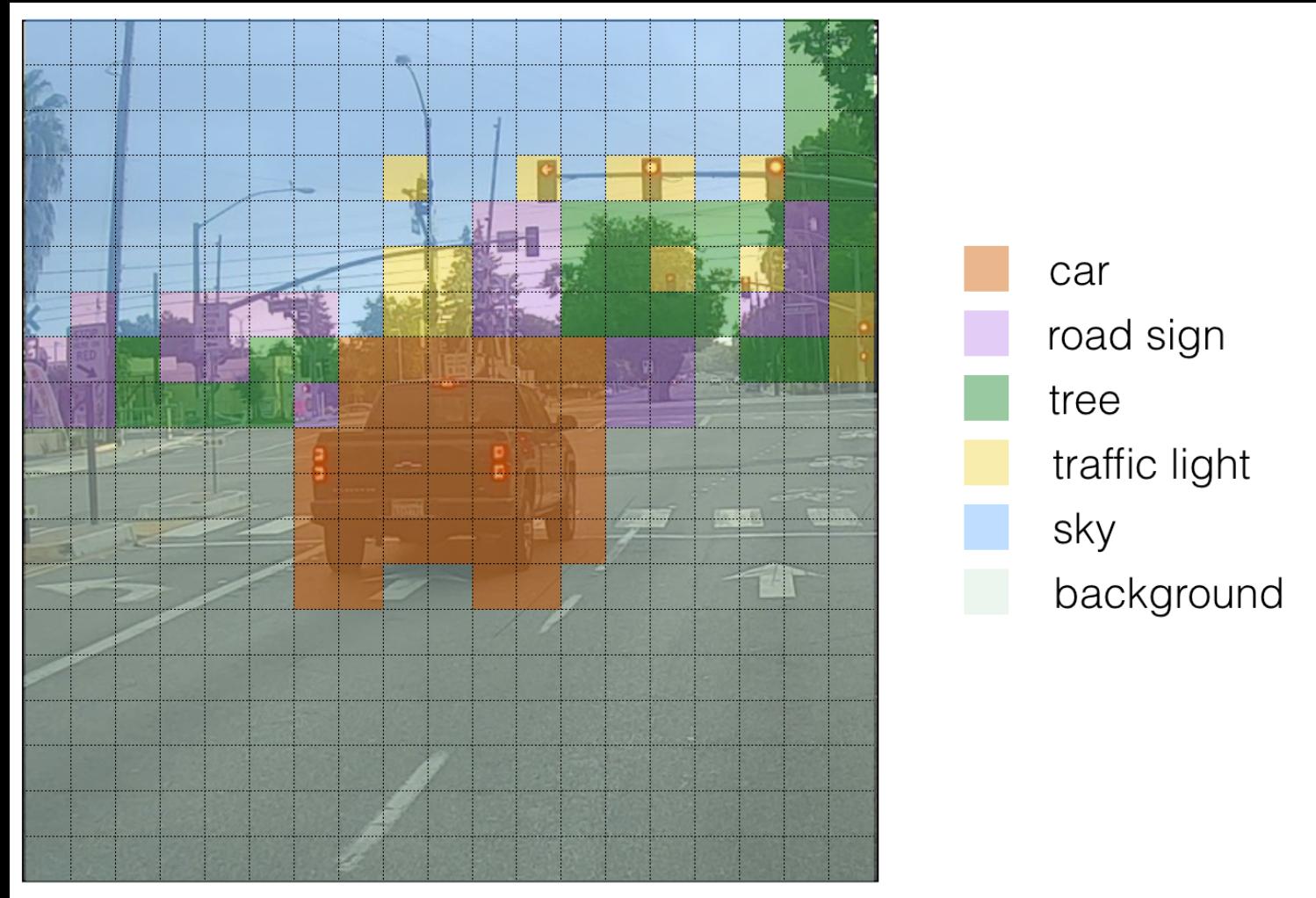
$1 \times 1 \text{ conv}$
425

$13 \times 13 \times 425$

res. output

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>

intermédiaire - result



bounding box representation

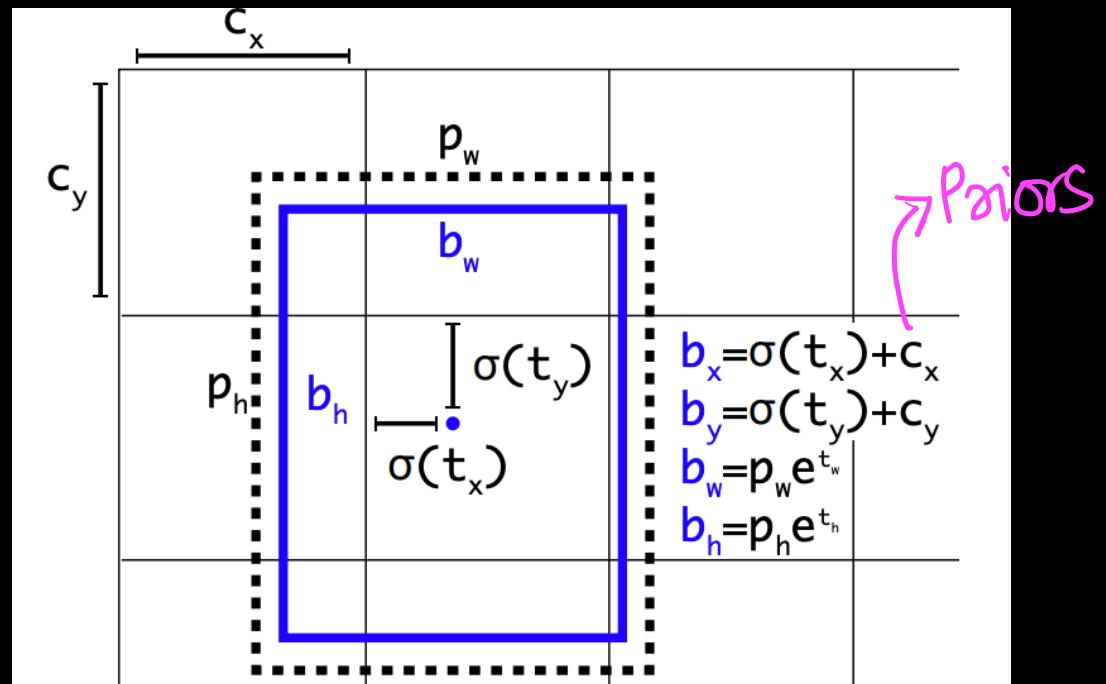
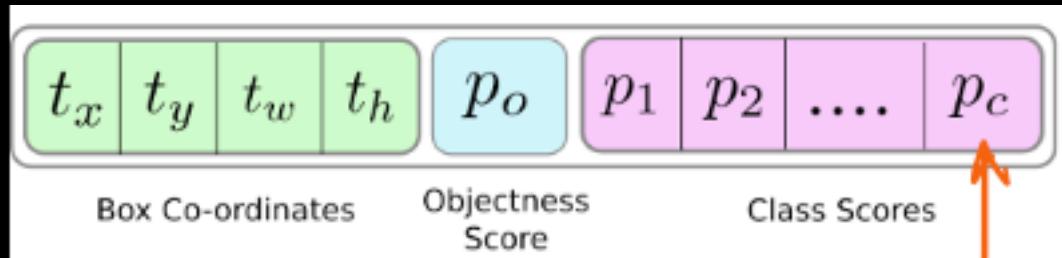


Figure 2. **Bounding boxes with dimension priors and location prediction.** We predict the width and height of the box as offsets from cluster centroids. We predict the center coordinates of the box relative to the location of filter application using a sigmoid function. This figure blatantly self-plagiarized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

ANCHOR-boxes:

$$[c_x, c_y, p_w, p_h]$$

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

ANCHOR-boxes $\{x_c, y_c, P_w, P_h\}$

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

→ $K=5 \Rightarrow$ 5 anchor-boxes



5 bounding boxes
per cell

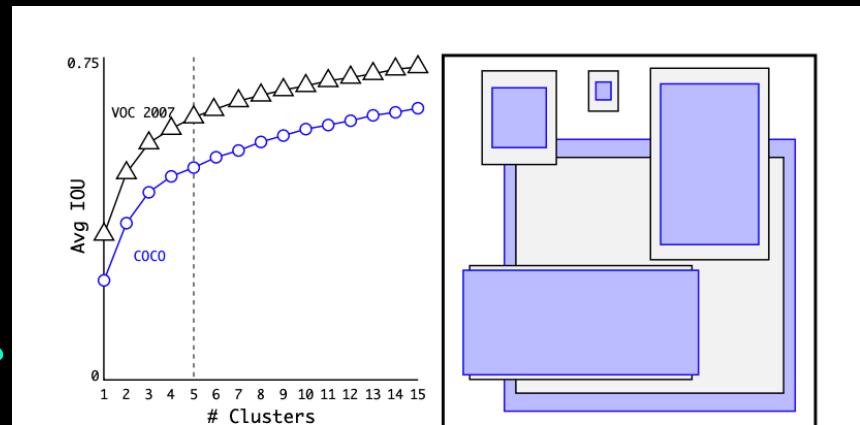
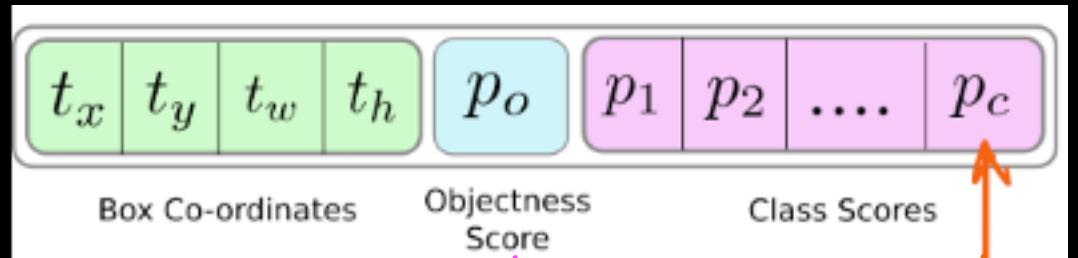


Figure 2: Clustering box dimensions on VOC and COCO. We run k-means clustering on the dimensions of bounding boxes to get good priors for our model. The left image shows the average IOU we get with various choices for k . We find that $k = 5$ gives a good tradeoff for recall vs. complexity of the model. 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



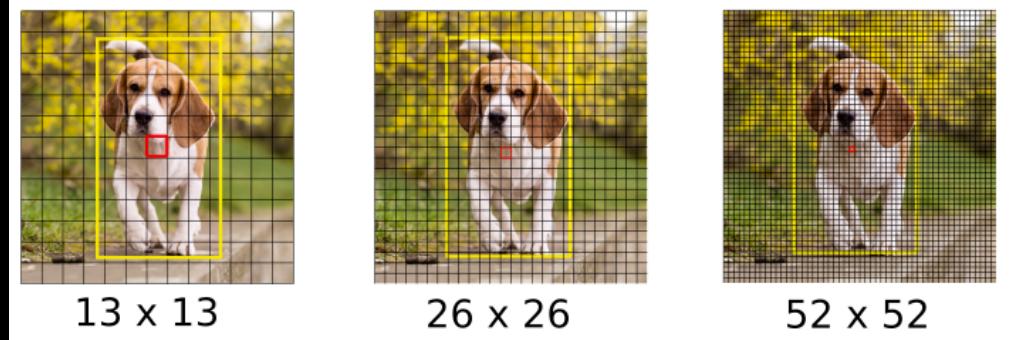
$P(\text{box contains } C_i \text{ object})$

$$= p_o \times p_i$$

Is there any
object at all

e.g.: person, woman
 C_{10} C_{13} [Softmax
does not work]

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	$3 \times 3 / 2$	128×128
1x Convolutional	32	1×1	
1x Convolutional	64	3×3	
Residual			128×128
Convolutional	128	$3 \times 3 / 2$	64×64
Convolutional	64	1×1	
2x Convolutional	128	3×3	
Residual			64×64
Convolutional	256	$3 \times 3 / 2$	32×32
Convolutional	128	1×1	
8x Convolutional	256	3×3	
Residual			32×32
Convolutional	512	$3 \times 3 / 2$	16×16
Convolutional	256	1×1	
8x Convolutional	512	3×3	
Residual			16×16
Convolutional	1024	$3 \times 3 / 2$	8×8
Convolutional	512	1×1	
4x Convolutional	1024	3×3	
Residual			8×8
Avgpool		Global	
Connected		1000	
Softmax			

Darknet-53 model

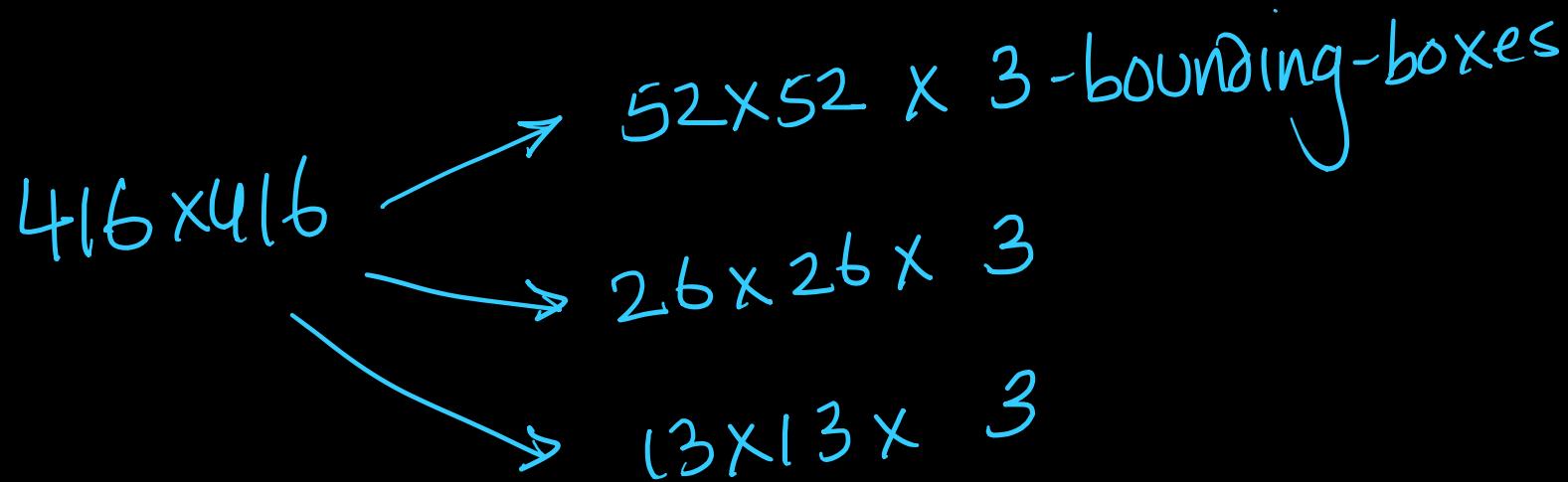
$\rightarrow 52 \times 52 \times 256$

$\rightarrow 26 \times 26 \times 512$

$\rightarrow 13 \times 13 \times 1024$

⑤

Combining boxes from various scales

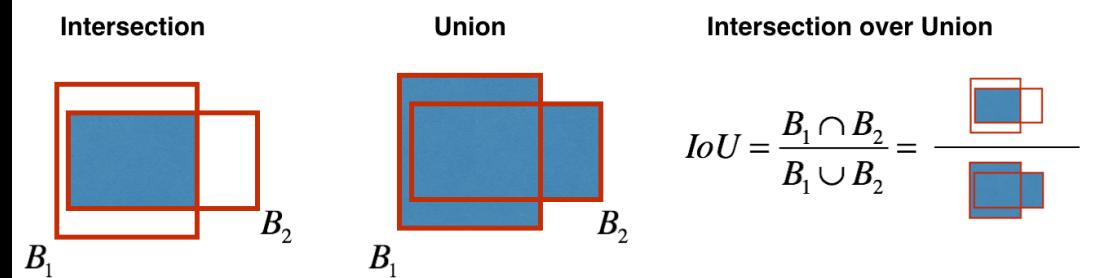


10,647 boxes

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

b) Non-max-Suppression (NMS)

PICK THE BOX
WITH MAX IoU



Code :

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

Keras-implementation : <https://github.com/qzwweee/keras-yolo3>
↳ not as-fast

