Insight
Centre for Data Analytics

DEEP LEARNING WORKSHOP

Dublin City University
27-28 April 2017

DCU

Day 2 Lecture 5

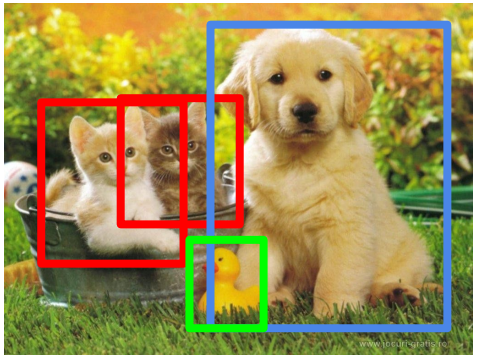
Object Detection

Amaia Salvador
amaia.salvador@upc.edu

PhD Candidate
Universitat Politècnica de Catalunya

UPC

Object Detection



CAT, DOG, DUCK

The task of assigning a **label** and a **bounding box** to all objects in the image

Object Detection: Datasets

20 categories
6k training images
6k validation images
10k test images

80 categories
200k training images
60k val + test images

200 categories
456k training images
60k validation + test images

Object Detection as Classification

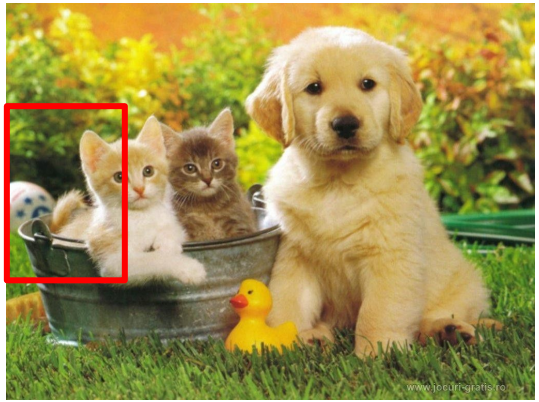
Classes = [cat, dog, duck]

Cat ? NO

Dog ? NO

Duck? NO

Object Detection as Classification



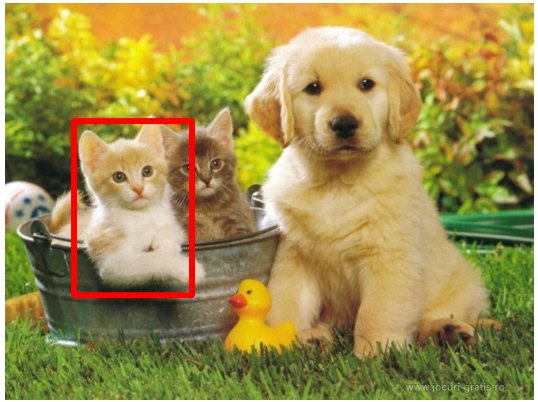
Classes = [cat, dog, duck]

Cat ? NO

Dog ? NO

Duck? NO

Object Detection as Classification



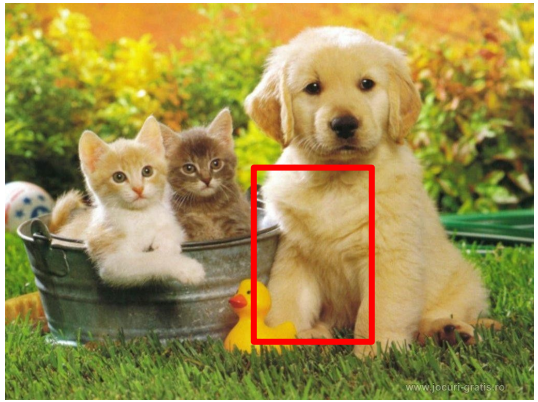
Classes = [cat, dog, duck]

Cat ? YES

Dog ? NO

Duck? NO

Object Detection as Classification



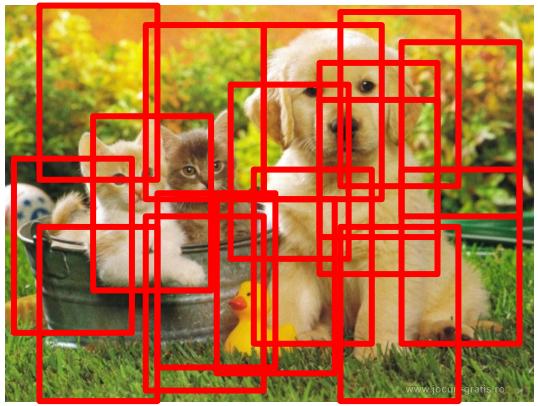
Classes = [cat, dog, duck]

Cat ? NO

Dog ? NO

Duck? NO

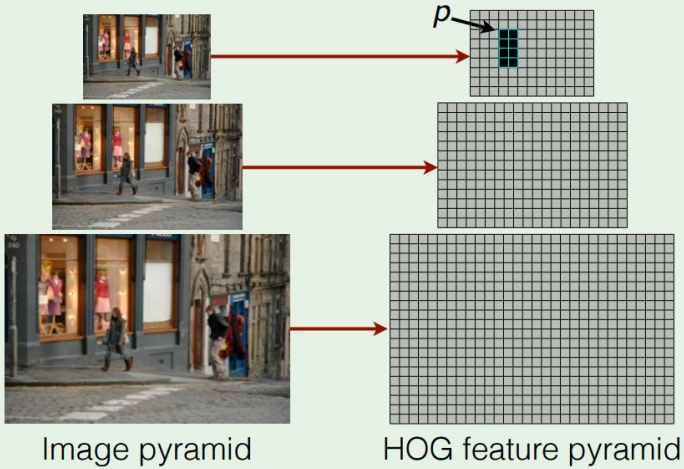
Object Detection as Classification



Problem:
Too many positions & scales to test

Solution: If your classifier is fast enough, go for it

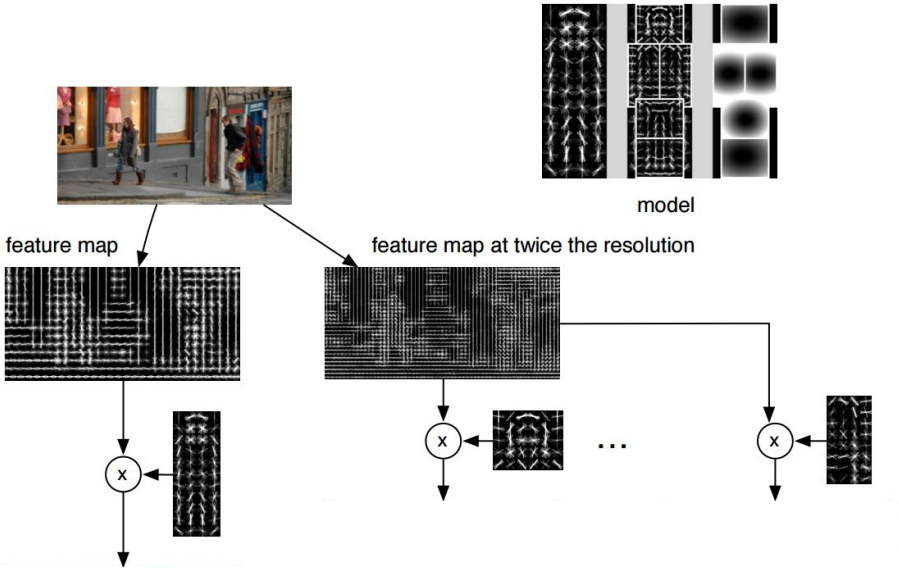
HOG: Histogram of Oriented Gradients

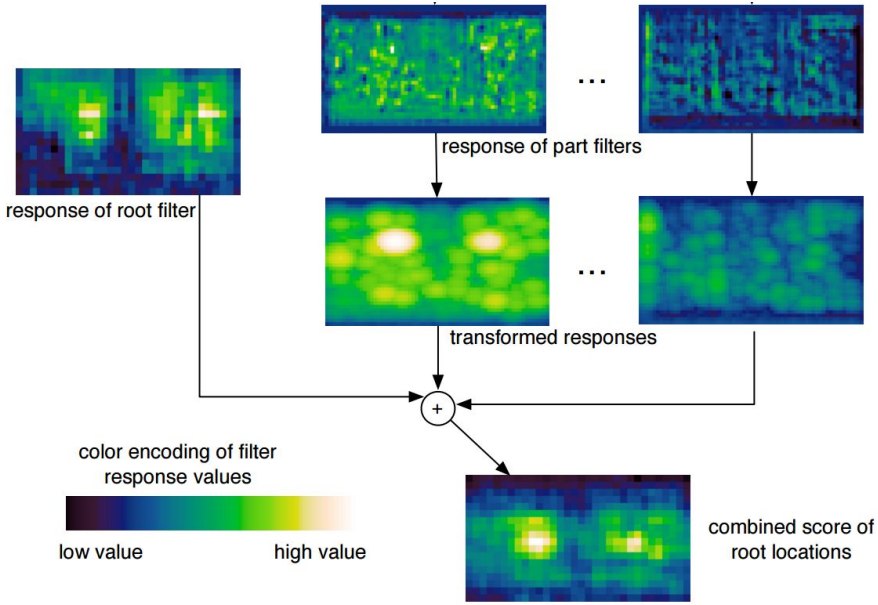


- Compute HOG of the whole image at multiple resolutions
- Score every subwindow of the feature pyramid
- Apply non-maxima suppression

Dalal and Triggs. Histograms of Oriented Gradients for Human Detection. CVPR 2005

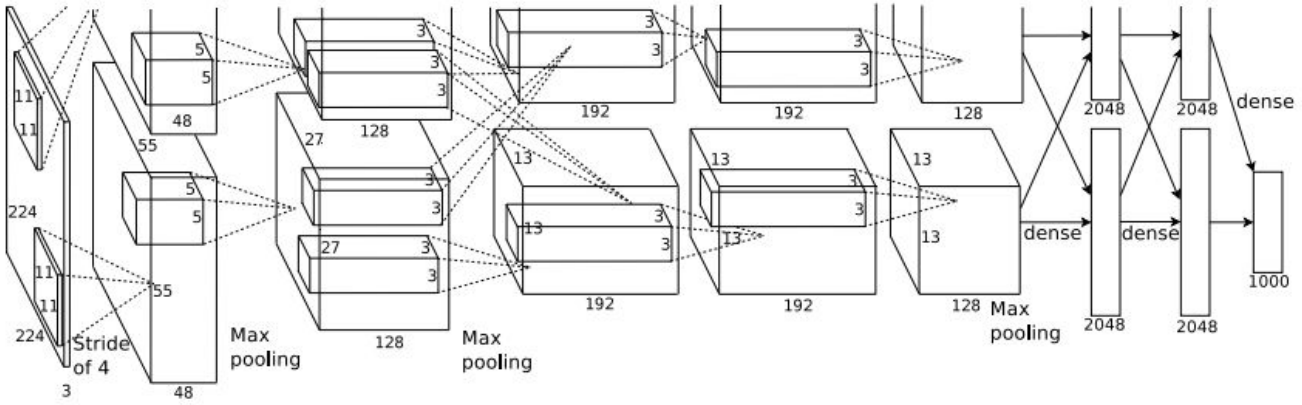
Deformable Part Model





Felzenszwalb et al, Object Detection with Discriminatively Trained Part Based Models, PAMI 2010

Object Detection with ConvNets?

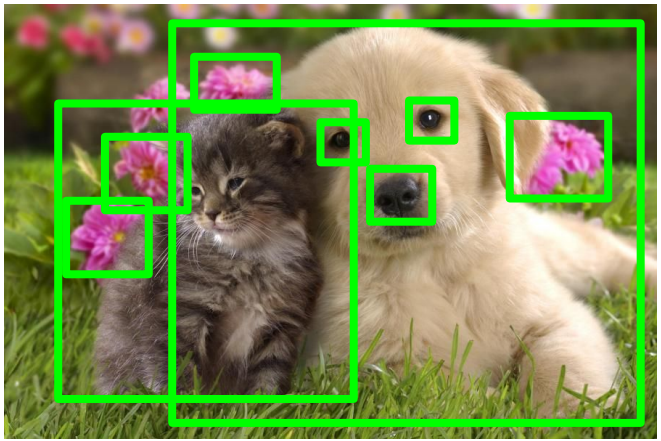


Convnets are computationally demanding. We can’t test all positions & scales !

Solution: Look at a tiny subset of positions. Choose them wisely :)

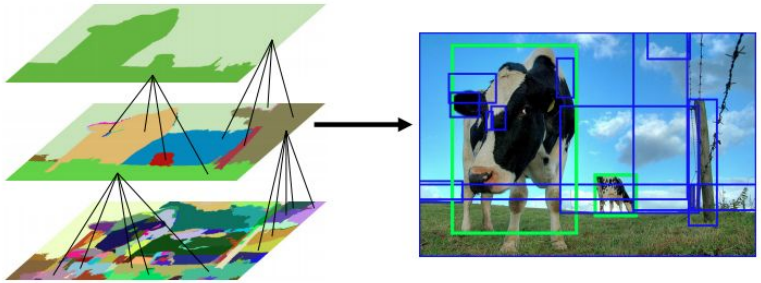
Region Proposals

- Find “blobby” image regions that are likely to contain objects
- “Class-agnostic” object detector
- Look for “blob-like” regions

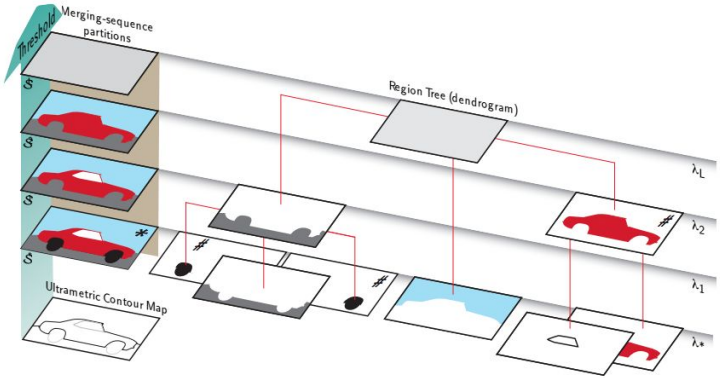


Slide Credit: CS231n

Region Proposals



Selective Search (SS)

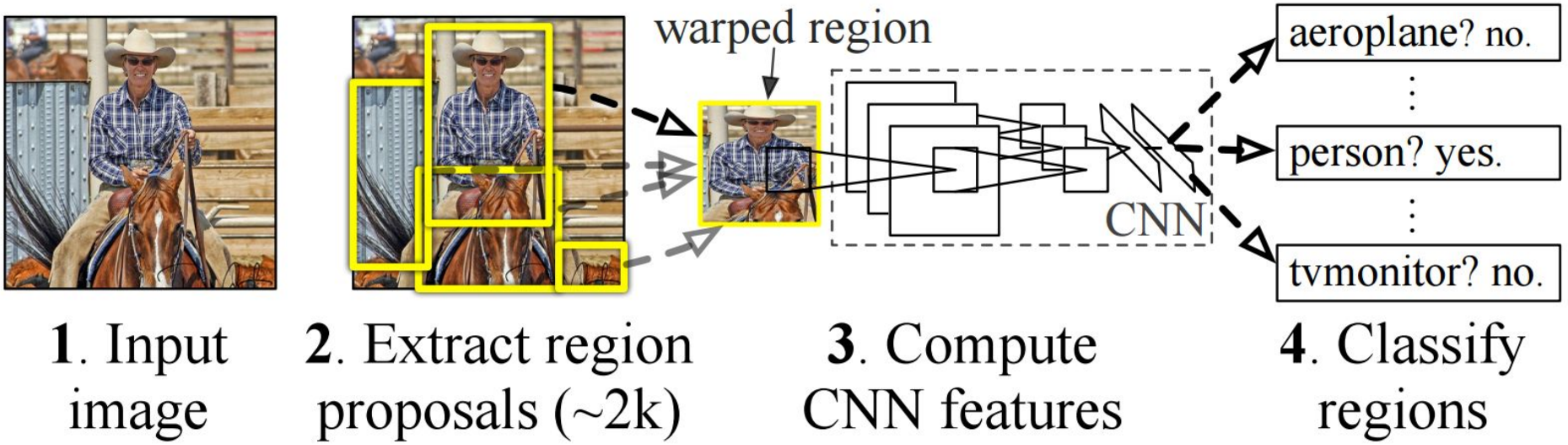


Multiscale Combinatorial Grouping (MCG)

[SS] Uijlings et al. Selective search for object recognition. IJCV 2013

[MCG] Arbeláez, Pont-Tuset et al. Multiscale combinatorial grouping. CVPR 2014

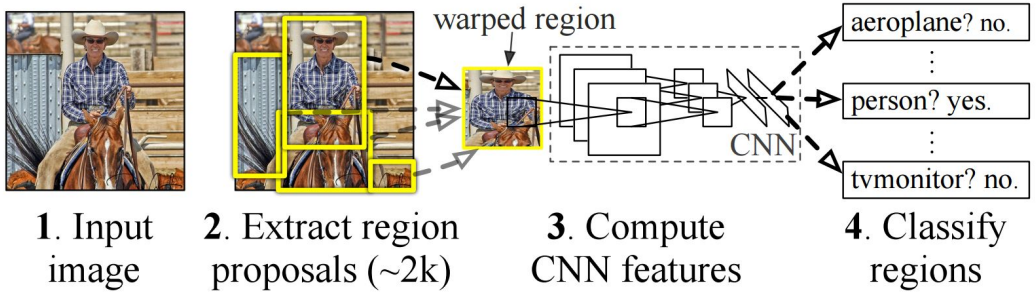
Object Detection with Convnets: R-CNN



Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR 2014

R-CNN

1. Train network on proposals

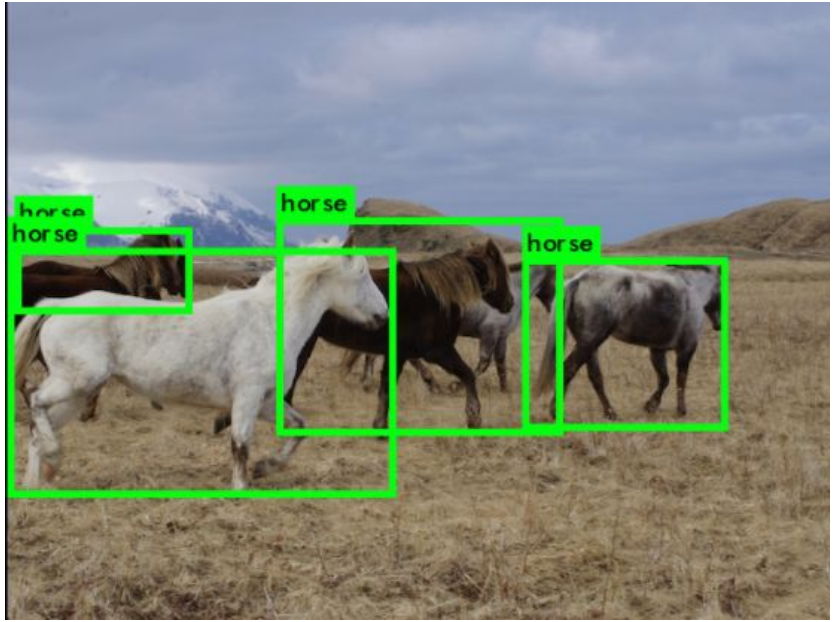


2. Post-hoc training of SVMs & Box regressors on fc7 features

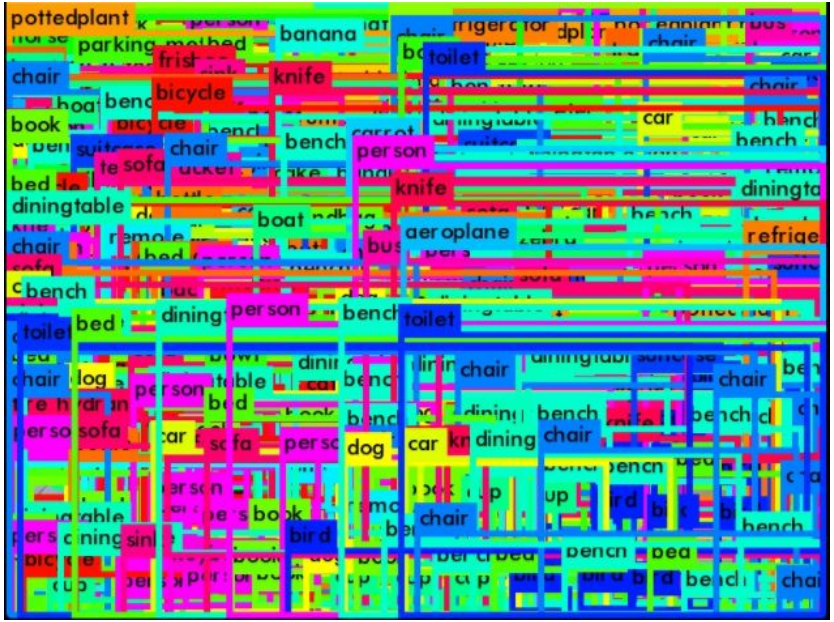
Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR 2014

R-CNN

We expect:

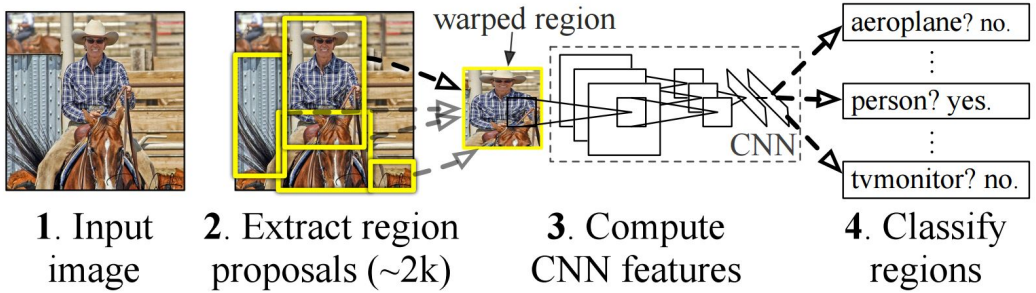


We get:



R-CNN

1. Train network on proposals

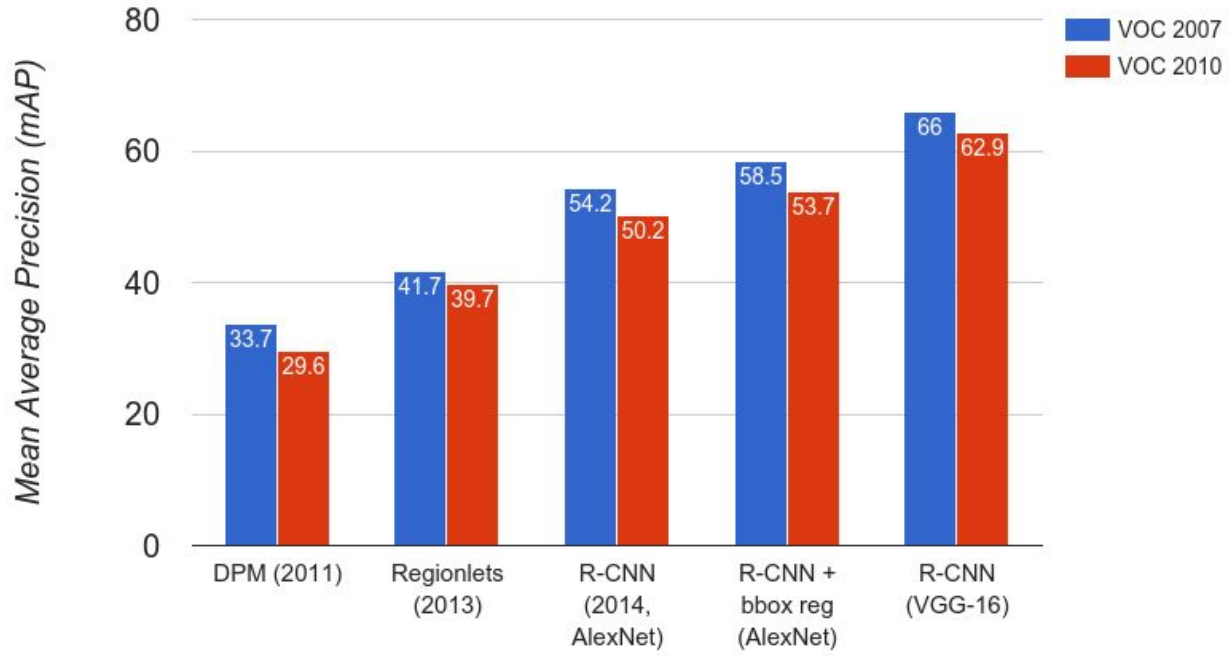


2. Post-hoc training of SVMs & Box regressors on fc7 features

3. **Non Maximum Suppression + score threshold**

Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR 2014

R-CNN



Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR 2014

R-CNN: Problems

1. Slow at test-time: need to run full forward pass of CNN for each region proposal

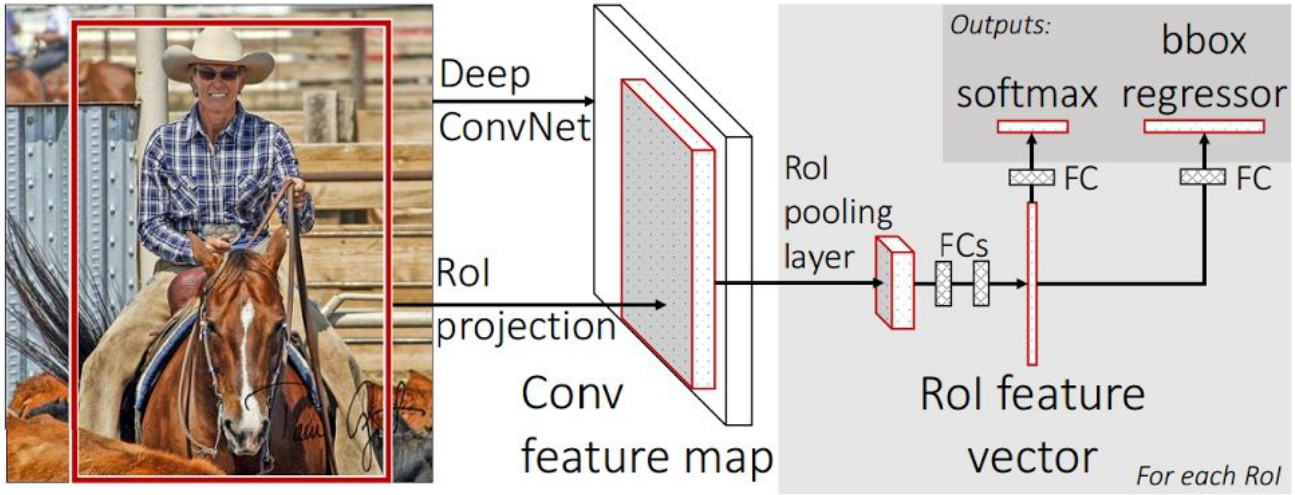
2. SVMs and regressors are post-hoc: CNN features not updated in response to SVMs and regressors

3. Complex multistage training pipeline

Slide Credit: CS231n

Fast R-CNN

R-CNN Problem #1: Slow at test-time: need to run full forward pass of CNN for each region proposal



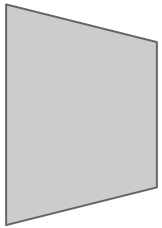
Solution: Share computation of convolutional layers between region proposals for an image

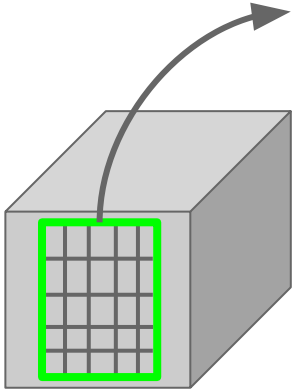
Girshick Fast R-CNN. ICCV 2015

Fast R-CNN: Sharing features

Hi-res input image:
3 x 800 x 600
with region
proposal

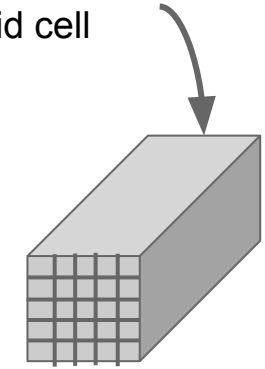
Convolution
and Pooling





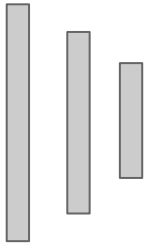
Hi-res conv features:
C x H x W
with region proposal

Max-pool within
each grid cell



RoI conv features:
C x h x w
for region proposal

Fully-connected
layers



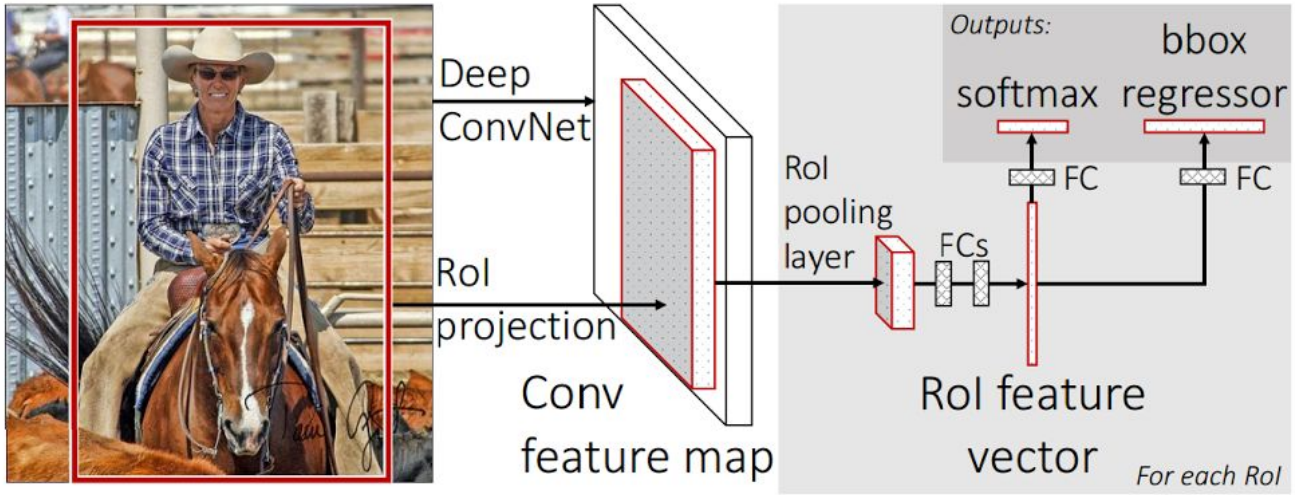
Fully-connected layers expect
low-res conv features:
C x h x w

Girshick Fast R-CNN. ICCV 2015

Slide Credit: CS231n

Fast R-CNN

R-CNN Problem #2&3: SVMs and regressors are post-hoc. Complex training.



Solution: Train it all at together E2E

Girshick Fast R-CNN. ICCV 2015

Fast R-CNN: End-to-end training

True box coordinates

Predicted box coordinates

Only for positive boxes

$$L(p, u, t^u, v) = L_{\text{cls}}(p, u) + \lambda[u \geq 1] L_{\text{loc}}(t^u, v),$$

Predicted class scores

True class scores

Log loss

Smooth L1 loss

Girshick Fast R-CNN. ICCV 2015

Fast R-CNN: Positive / Negative Samples

Positive samples are defined as those whose IoU overlap with a ground-truth bounding box is > 0.5.

Negative examples are sampled from those that have a maximum IoU overlap with ground truth in the interval [0.1, 0.5).

25%/75% ratio for positive/negative samples in a minibatch.

Girshick Fast R-CNN. ICCV 2015

Fast R-CNN

Faster!

FASTER!

Better!

	R-CNN	Fast R-CNN
Training Time:	84 hours	**9.5 hours**
(Speedup)	1x	**8.8x**
Test time per image	47 seconds	**0.32 seconds**
(Speedup)	1x	**146x**
mAP (VOC 2007)	66.0	**66.9**

Using VGG-16 CNN on Pascal VOC 2007 dataset

Slide Credit: CS231n

Fast R-CNN: Problem

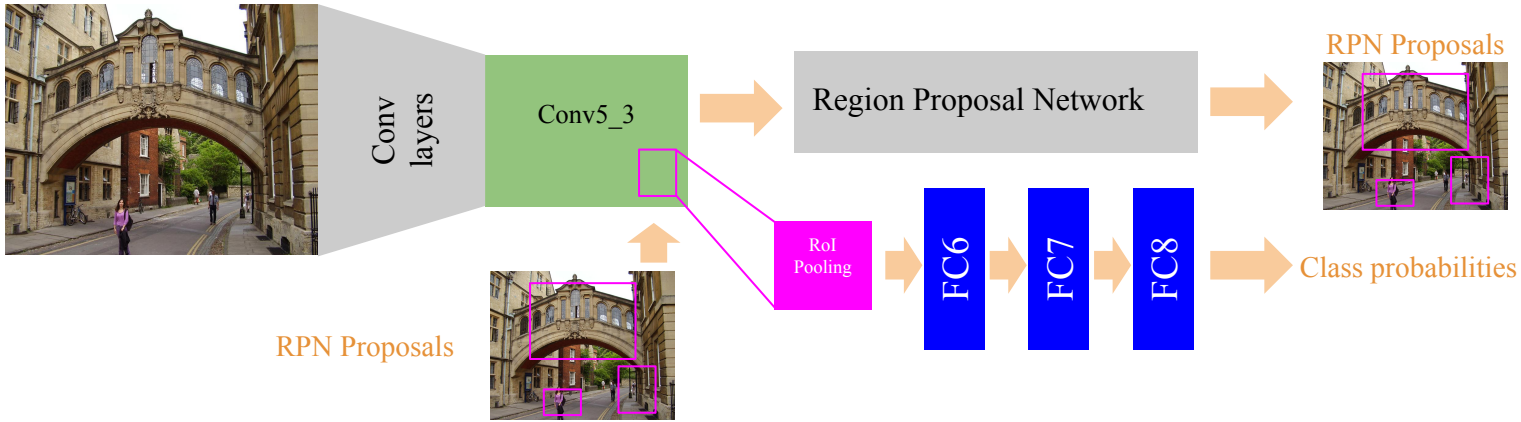
Test-time speeds don't include region proposals

	R-CNN	**Fast R-CNN**
Test time per image	47 seconds	**0.32 seconds**
(Speedup)	1x	**146x**
Test time per image with Selective Search	50 seconds	**2 seconds**
(Speedup)	1x	**25x**

Slide Credit: CS231n

Faster R-CNN

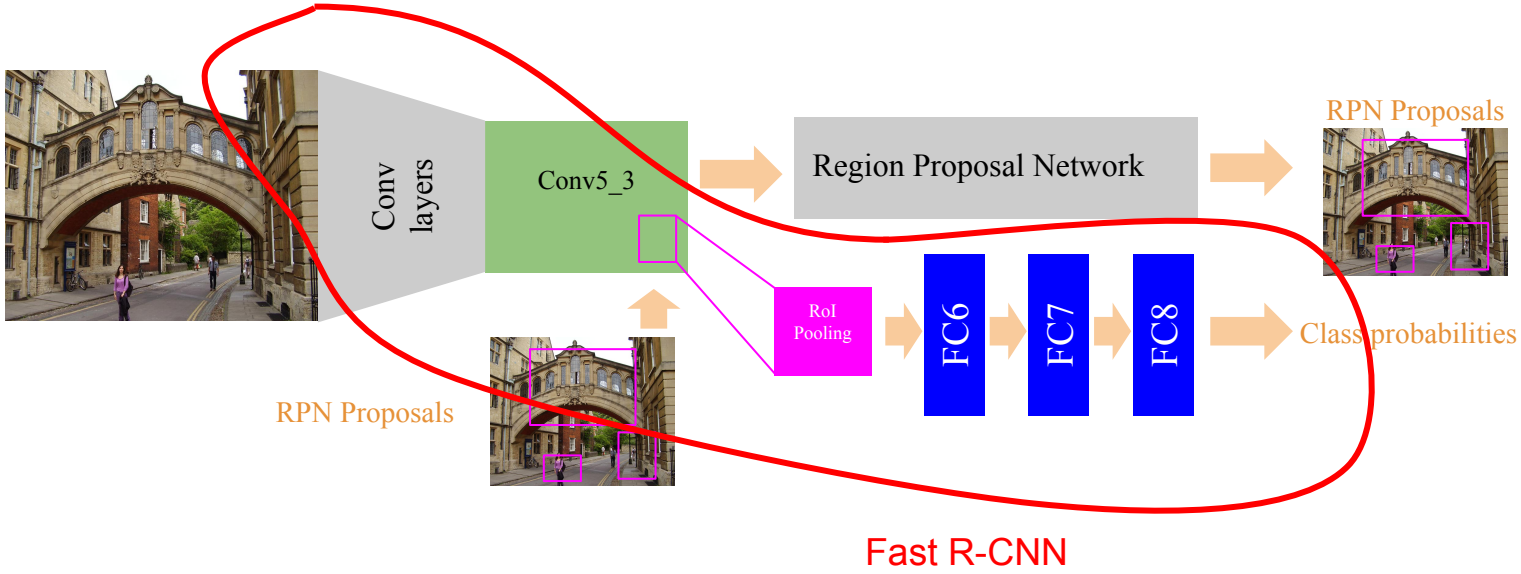
Learn proposals end-to-end sharing parameters with the classification network



Ren et al. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS 2015

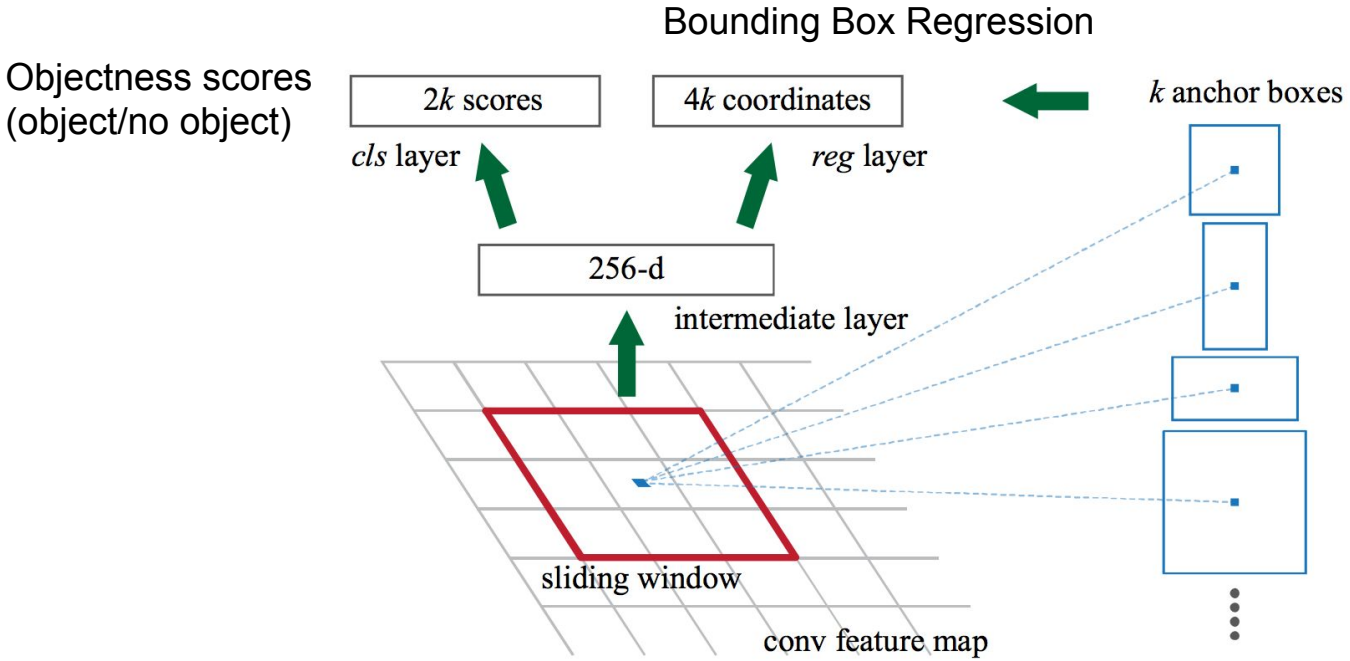
Faster R-CNN

Learn proposals end-to-end sharing parameters with the classification network



Ren et al. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS 2015

Region Proposal Network



In practice, k = 9 (3 different scales and 3 aspect ratios)

Ren et al. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS 2015

Region Proposal Network: Loss function

i = anchor index in minibatch

Log loss

Ground truth objectness label

Smooth L1 loss

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*).$$

True box coordinates

Coordinates of the predicted bounding box for anchor *i*

Predicted probability of being an object for anchor *i*

In practice $\lambda = 10$, so that both terms are roughly equally balanced

N_{cls} = Number of anchors in minibatch (~ 256)
N_{reg} = Number of anchor locations (~ 2400)

Region Proposal Network: Positive / Negative Samples

An anchor is labeled as positive if:

(a) the anchor is the one with highest IoU overlap with a ground-truth box

(b) the anchor has an IoU overlap with a ground-truth box higher than 0.7

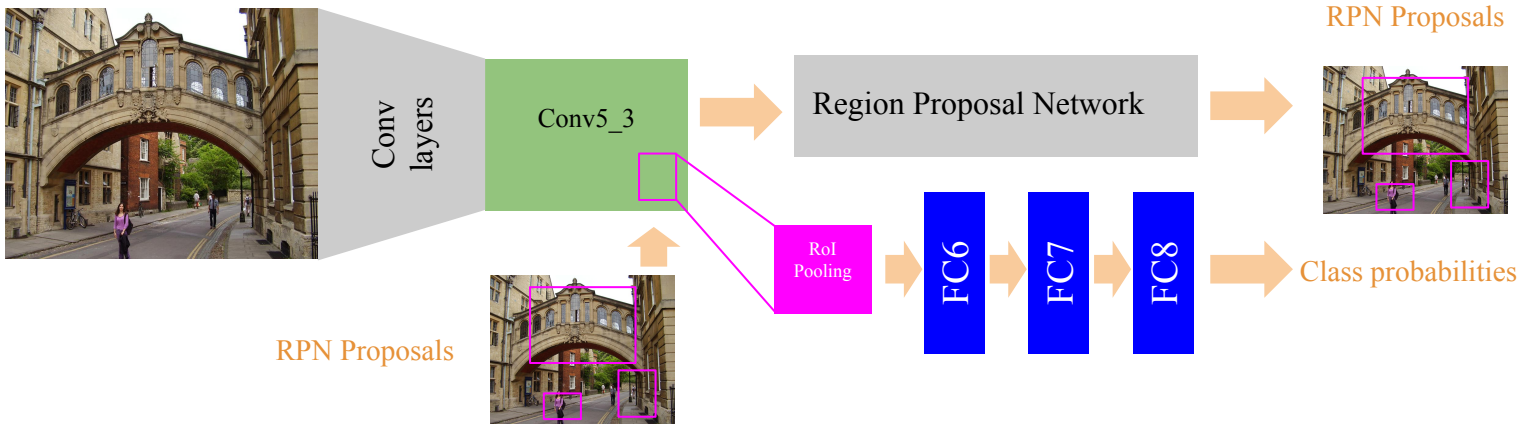
Negative labels are assigned to anchors with IoU lower than 0.3 for all ground-truth boxes.

50%/50% ratio of positive/negative anchors in a minibatch.

Faster R-CNN: Training

RoI Pooling is not differentiable w.r.t box coordinates. Solutions:

- Alternate training
- Ignore gradient of classification branch w.r.t proposal coordinates
- Make pooling function differentiable



Ren et al. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS 2015

Faster R-CNN

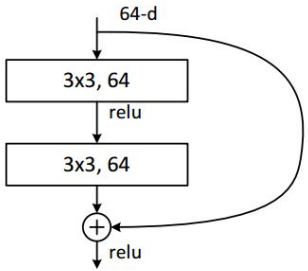
	R-CNN	Fast R-CNN	Faster R-CNN
Test time per image (with proposals)	50 seconds	2 seconds	**0.2 seconds**
(Speedup)	1x	25x	**250x**
mAP (VOC 2007)	66.0	**66.9**	**66.9**

Ren et al. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS 2015

Slide Credit: CS231n

Faster R-CNN

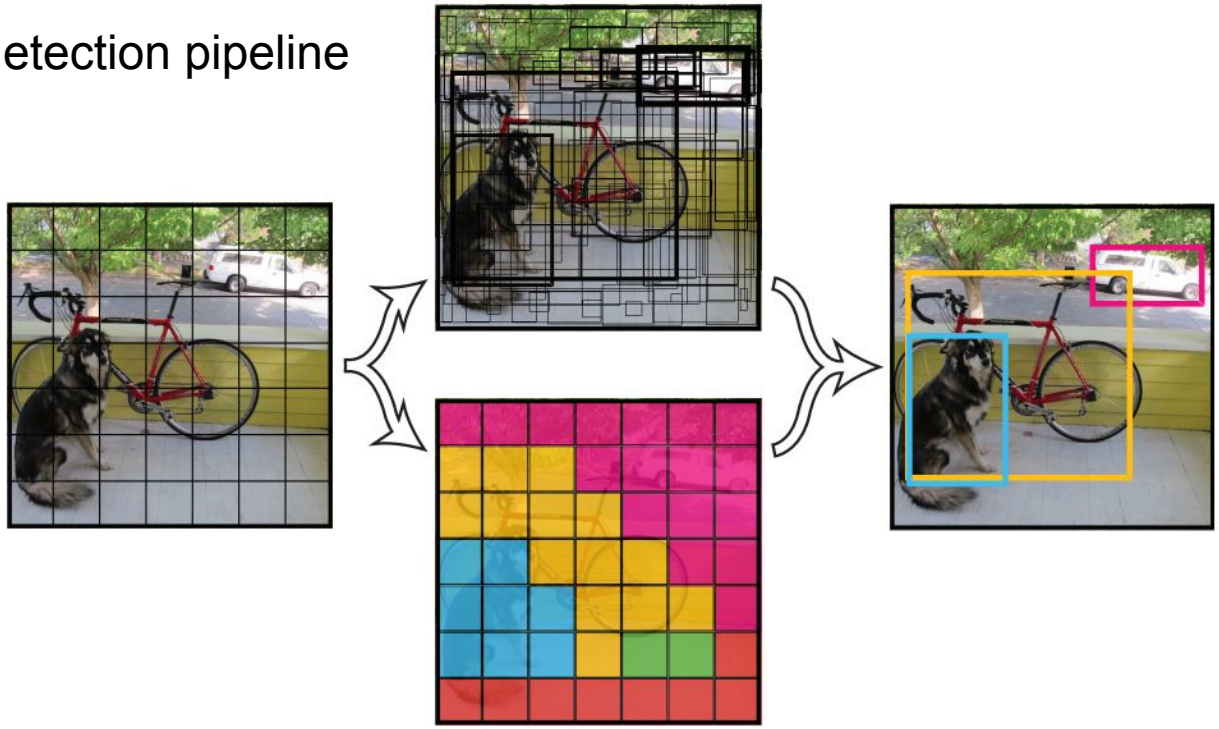
- Faster R-CNN is the basis of the winners of COCO and ILSVRC 2015 object detection competitions.



He et al. Deep residual learning for image recognition. CVPR 2016

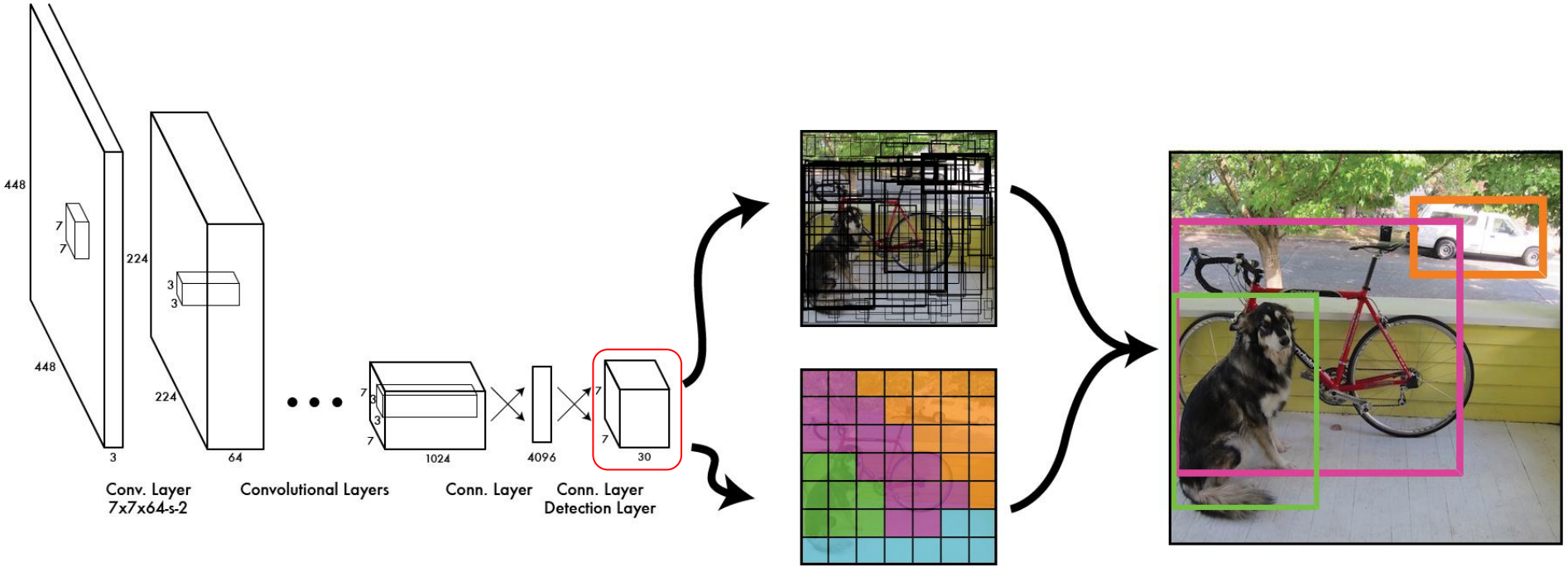
YOLO: You Only Look Once

Proposal-free object detection pipeline



Redmon et al. [You Only Look Once: Unified, Real-Time Object Detection,](#) CVPR 2016

YOLO: You Only Look Once



Redmon et al. [You Only Look Once: Unified, Real-Time Object Detection,](#) CVPR 2016

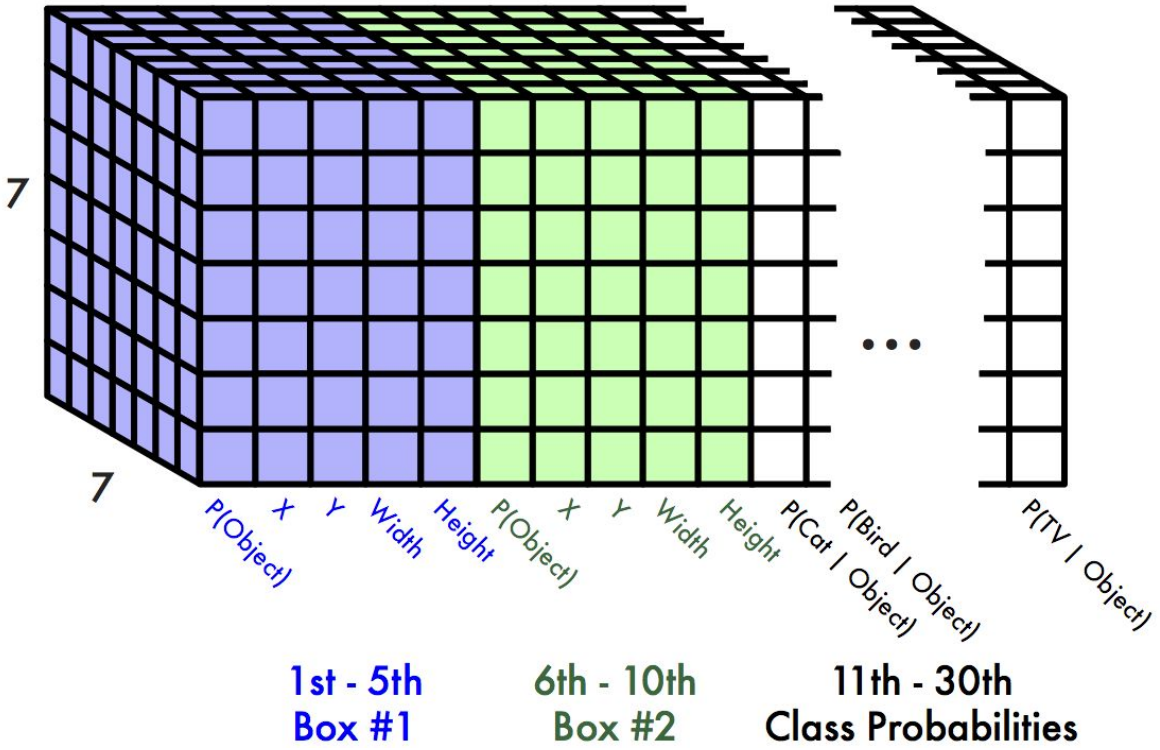
YOLO: You Only Look Once

Each cell predicts:

- For each bounding box:
  - 4 coordinates (x, y, w, h)
  - 1 confidence value
- Some number of class probabilities

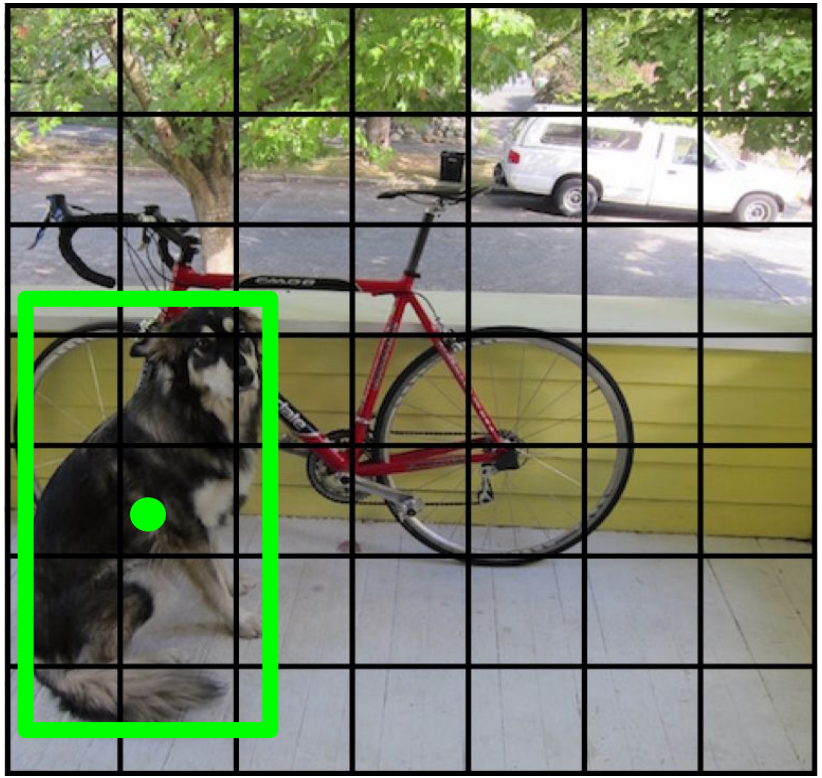
For Pascal VOC:

- 7x7 grid
- 2 bounding boxes / cell
- 20 classes



7 x 7 x (2 x 5 + 20) = 7 x 7 x 30 tensor = **1470 outputs**

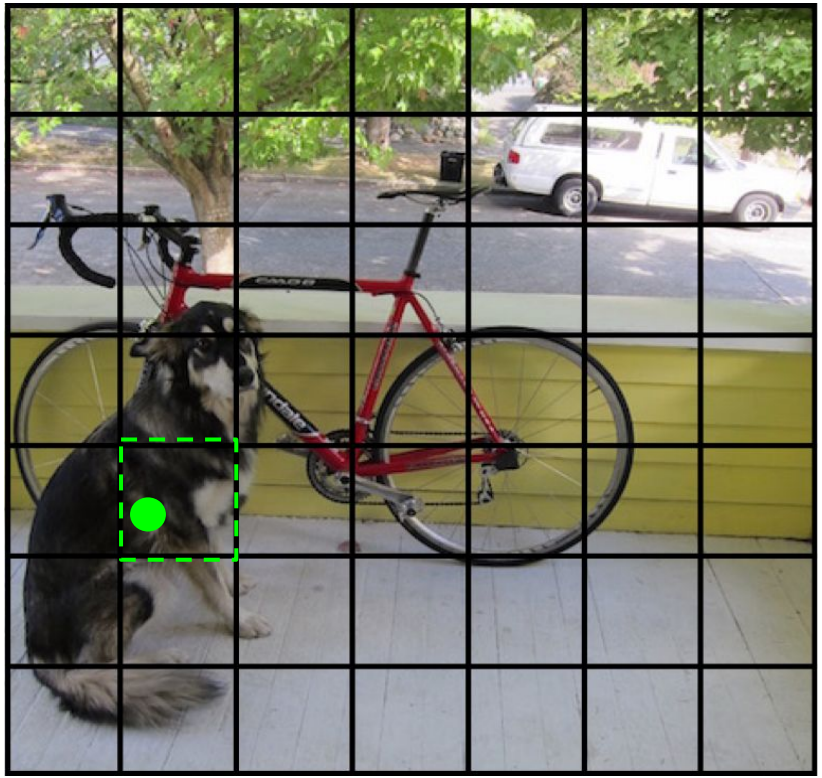
YOLO: Training



For training, each ground truth bounding box is matched into the right cell

Slide credit: YOLO Presentation @ CVPR 2016

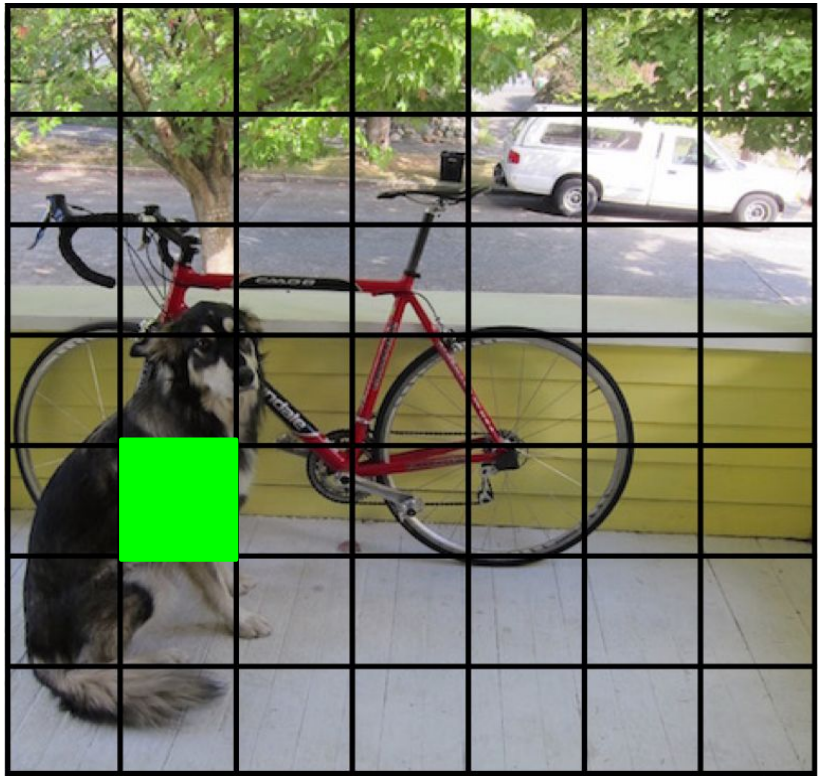
YOLO: Training



For training, each ground truth bounding box is matched into the right cell

Slide credit: YOLO Presentation @ CVPR 2016

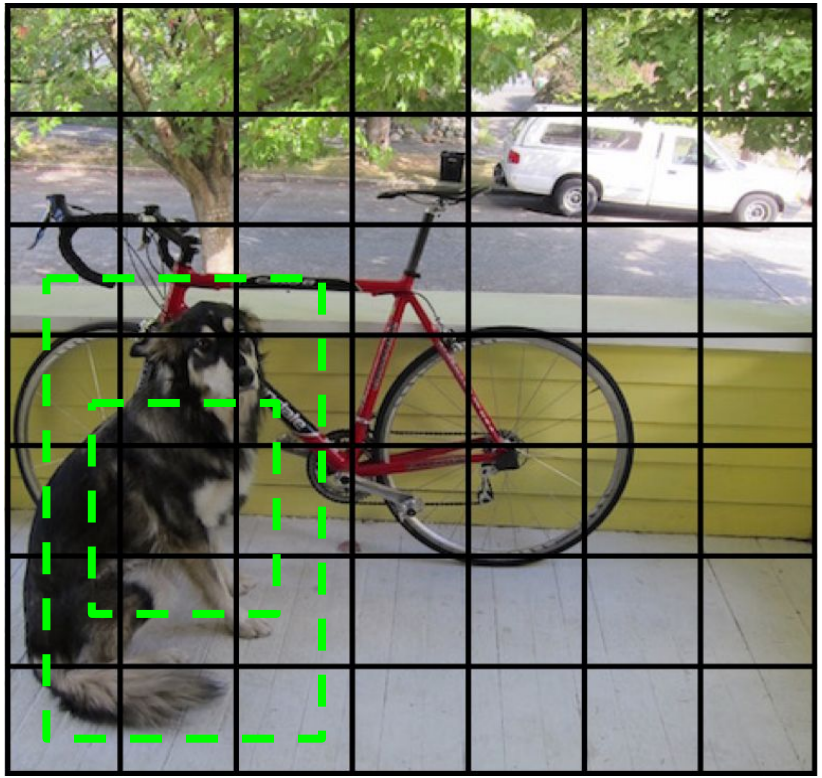
YOLO: Training



Optimize class prediction in that cell:
dog: 1, cat: 0, bike: 0, ...

Slide credit: YOLO Presentation @ CVPR 2016

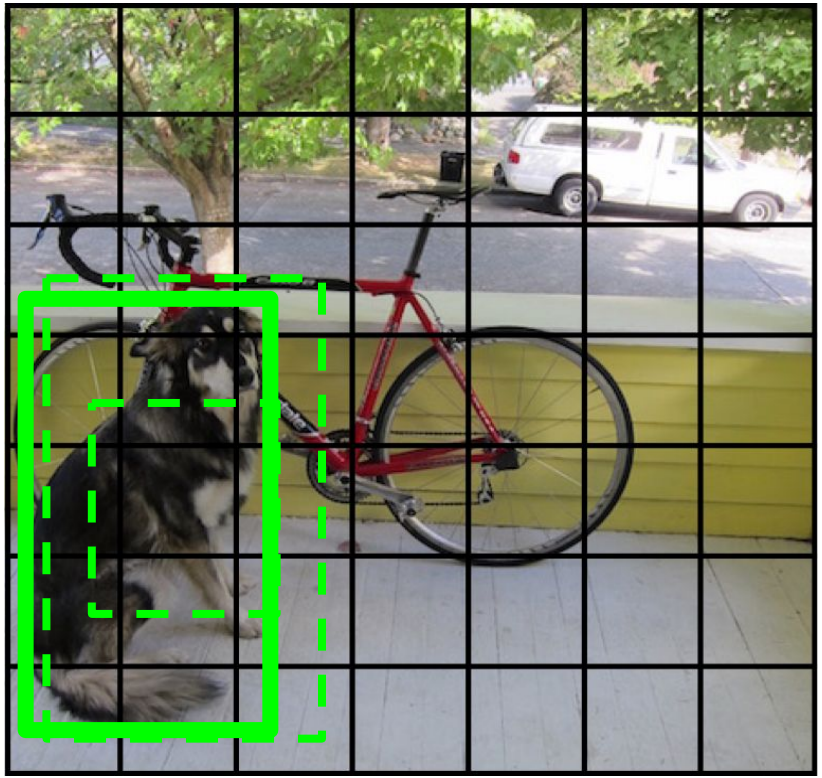
YOLO: Training



Predicted boxes for this cell

Slide credit: YOLO Presentation @ CVPR 2016

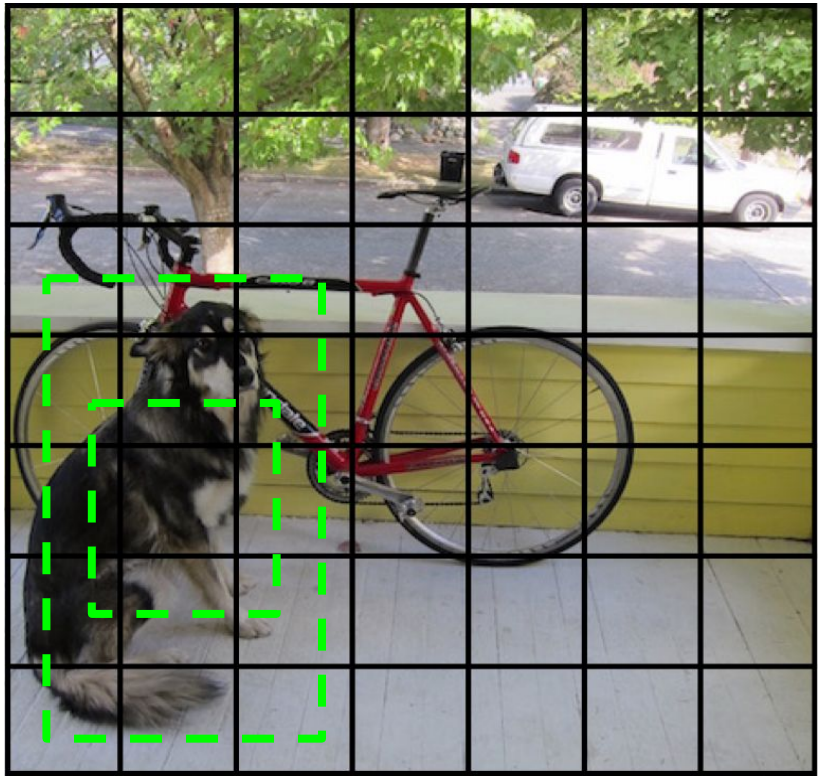
YOLO: Training



Find the best one wrt ground truth bounding box, optimize it (i.e. adjust its coordinates to be closer to the ground truth's coordinates)

Slide credit: YOLO Presentation @ CVPR 2016

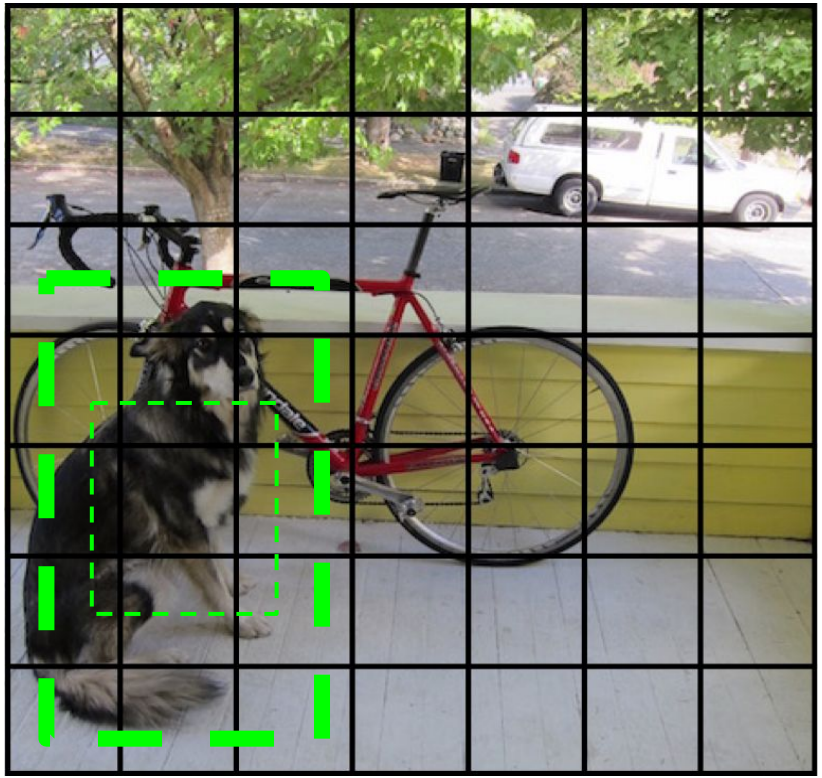
YOLO: Training



Increase matched box's confidence, decrease non-matched boxes confidence

Slide credit: YOLO Presentation @ CVPR 2016

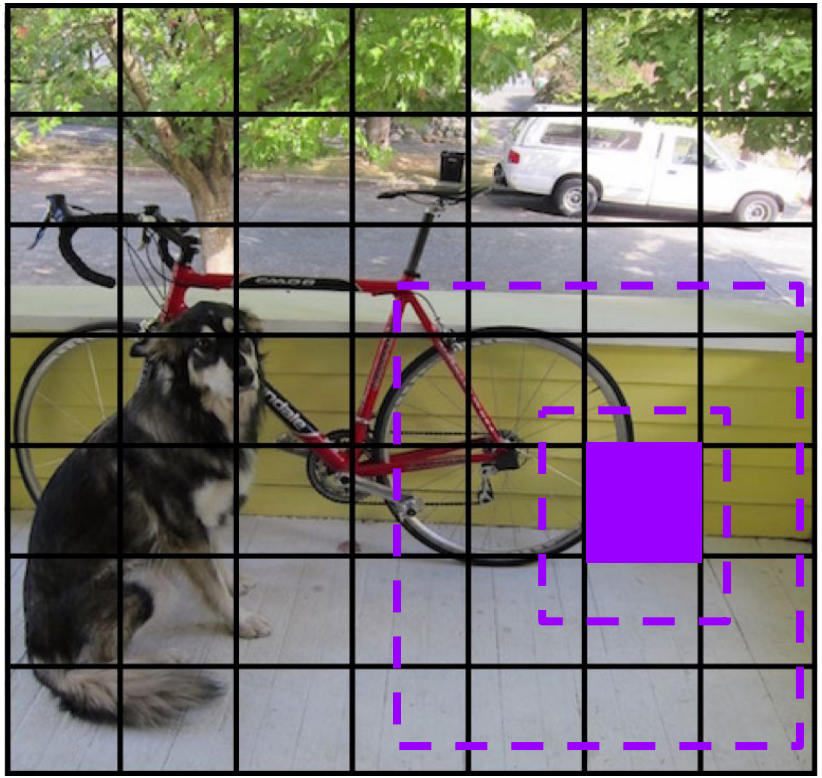
YOLO: Training



Increase matched box's confidence, decrease non-matched boxes confidence

Slide credit: YOLO Presentation @ CVPR 2016

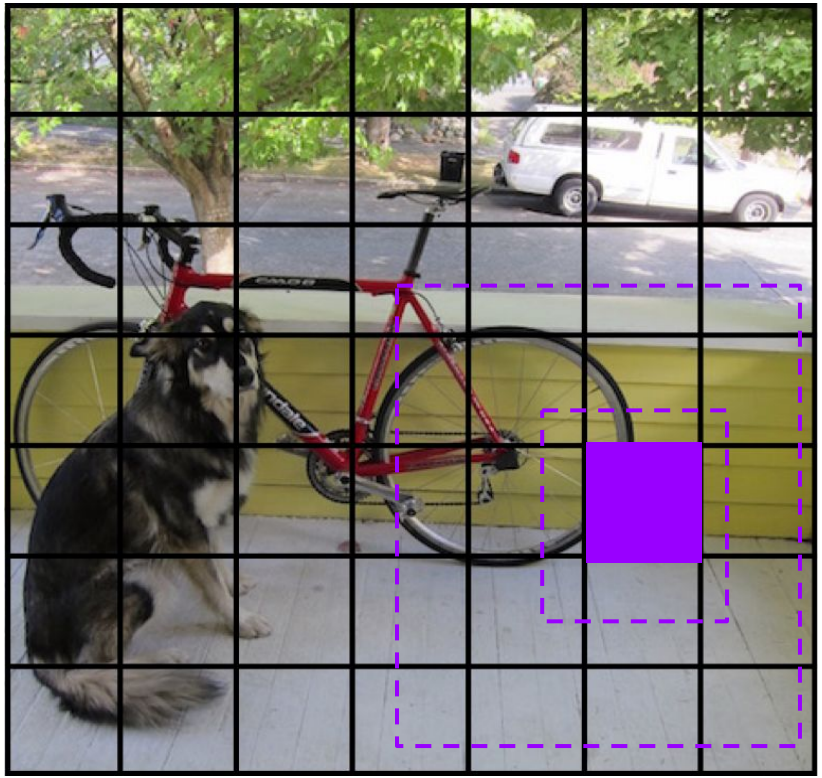
YOLO: Training



For cells with no ground truth detections, confidences of all predicted boxes are decreased

Slide credit: YOLO Presentation @ CVPR 2016

YOLO: Training



For cells with no ground truth detections:

- Confidences of all predicted boxes are decreased
- Class probabilities are not adjusted

Slide credit: YOLO Presentation @ CVPR 2016

YOLO: Training, formally

= 1 if box *j* and cell *i* are matched together, 0 otherwise

Bounding box coordinate regression

$$\lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right]$$

$$+ \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right]$$

Bounding box score prediction

$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2$$

= 1 if box *j* and cell *i* are NOT matched together

$$+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2$$

Class score prediction

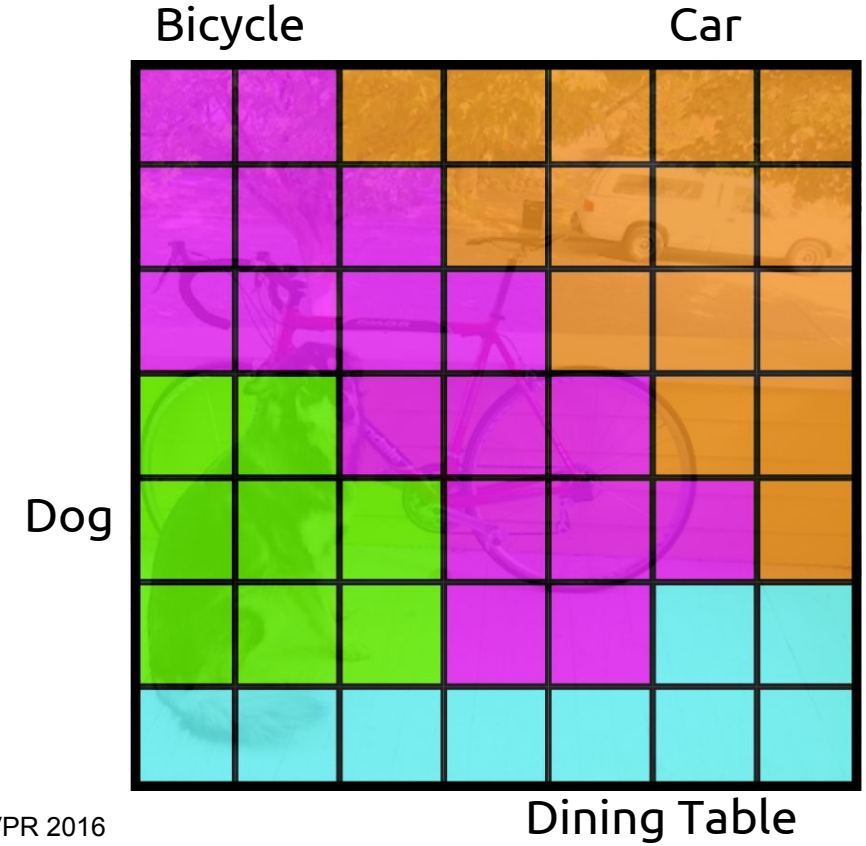
$$+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} \left(p_i(c) - \hat{p}_i(c) \right)^2$$

= 1 if cell *i* has an object present

Slide credit: [YOLO Presentation @ CVPR 2016]

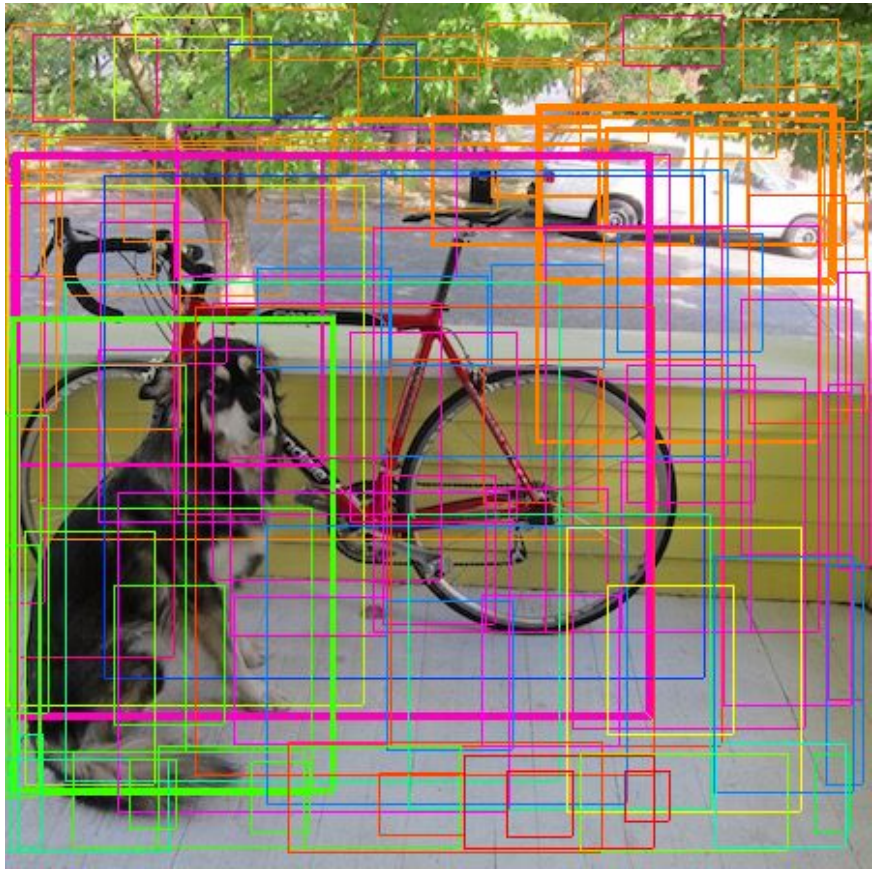
YOLO: You Only Look Once

Predict class probability for each cell
(conditioned on object P(car | object))



Redmon et al. [You Only Look Once: Unified, Real-Time Object Detection,](#) CVPR 2016

YOLO: You Only Look Once

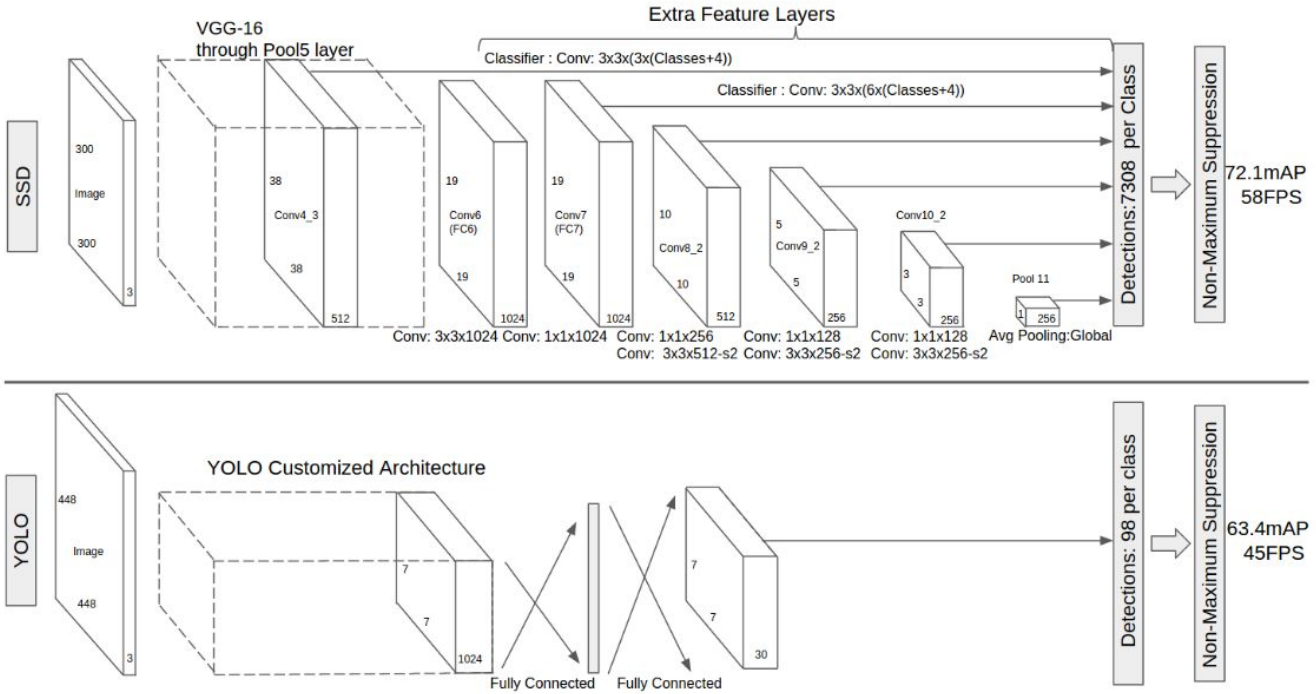


+ NMS
+ Score threshold

Redmon et al. You Only Look Once: Unified, Real-Time Object Detection, CVPR 2016

SSD: Single Shot MultiBox Detector

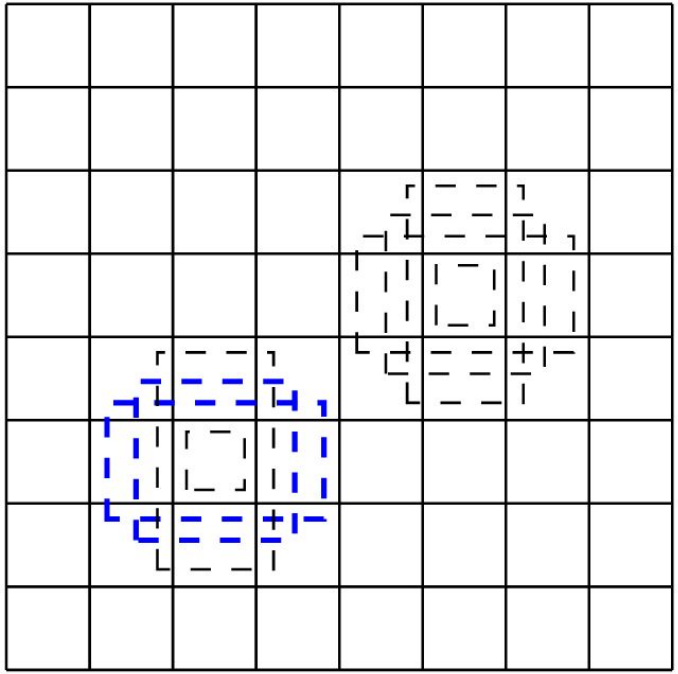
Same idea as YOLO, + several predictors at different stages in the network



Liu et al. SSD: Single Shot MultiBox Detector, ECCV 2016

SSD: Single Shot MultiBox Detector

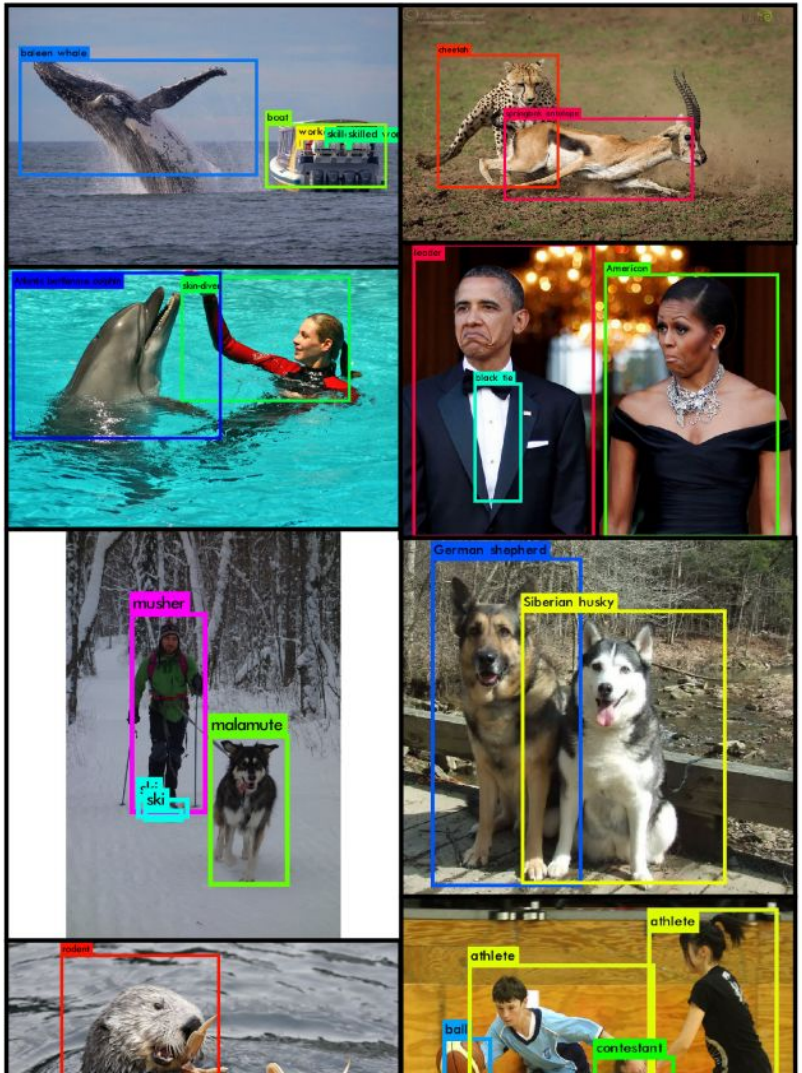
Similarly to Faster R-CNN, it uses box anchors to predict box coordinates as displacements



Liu et al. SSD: Single Shot MultiBox Detector, ECCV 2016

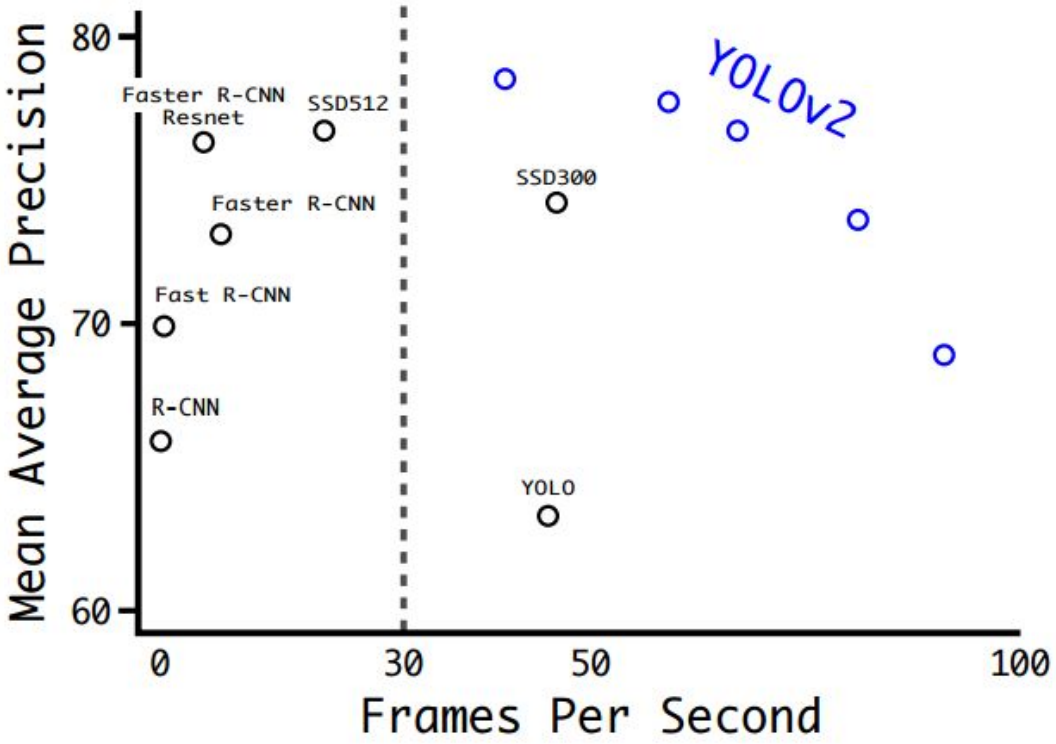
YOLOv2

	YOLO							YOLOv2
batch norm?		✓	✓	✓	✓	✓	✓	✓	✓
hi-res classifier?			✓	✓	✓	✓	✓	✓
convolutional?				✓	✓	✓	✓	✓
anchor boxes?				✓	✓			
new network?					✓	✓	✓	✓
dimension priors?						✓	✓	✓
location prediction?						✓	✓	✓
passthrough?							✓	✓
multi-scale?								✓
hi-res detector?								✓
VOC2007 mAP	63.4	65.8	69.5	69.2	69.6	74.4	75.4	76.8	**78.6**



Redmon & Farhadi. YOLO900: Better, Faster, Stronger. CVPR 2017

YOLOv2



Results on Pascal VOC 2007

YOLOv2

		0.5:0.95	0.5	0.75	S	M	L	1	10	100	S	M	L
Fast R-CNN [5]	train	19.7	35.9	-	-	-	-	-	-	-	-	-	-
Fast R-CNN[1]	train	20.5	39.9	19.4	4.1	20.0	35.8	21.3	29.5	30.1	7.3	32.1	52.0
Faster R-CNN[15]	trainval	21.9	42.7	-	-	-	-	-	-	-	-	-	-
ION [1]	train	23.6	43.2	23.6	6.4	24.1	38.3	23.2	32.7	33.5	10.1	37.7	53.6
Faster R-CNN[10]	trainval	24.2	45.3	23.5	7.7	26.4	37.1	23.8	34.0	34.6	12.0	38.5	54.4
SSD300 [11]	trainval35k	23.2	41.2	23.4	5.3	23.2	39.6	22.5	33.2	35.3	9.6	37.6	56.5
SSD512 [11]	trainval35k	**26.8**	**46.5**	**27.8**	**9.0**	**28.9**	**41.9**	**24.8**	**37.5**	**39.8**	**14.0**	**43.5**	**59.0**
YOLOv2 [11]	trainval35k	21.6	44.0	19.2	5.0	22.4	35.5	20.7	31.6	33.3	9.8	36.5	54.4

Results on COCO test-dev 2015

Summary

Proposal-based methods

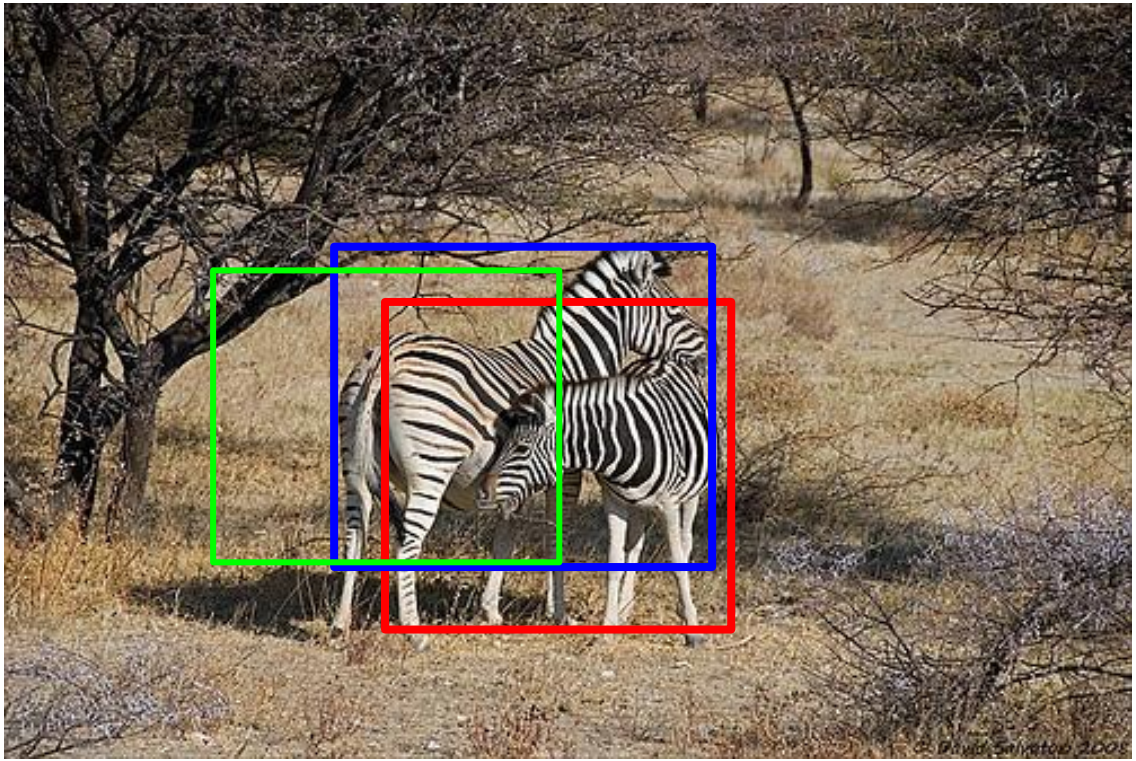
- R-CNN
- Fast R-CNN
- Faster R-CNN
- SPPnet
- R-FCN

Proposal-free methods

- YOLO, YOLOv2
- SSD

Questions ?

A note on NMS



Objectness scores

0.8

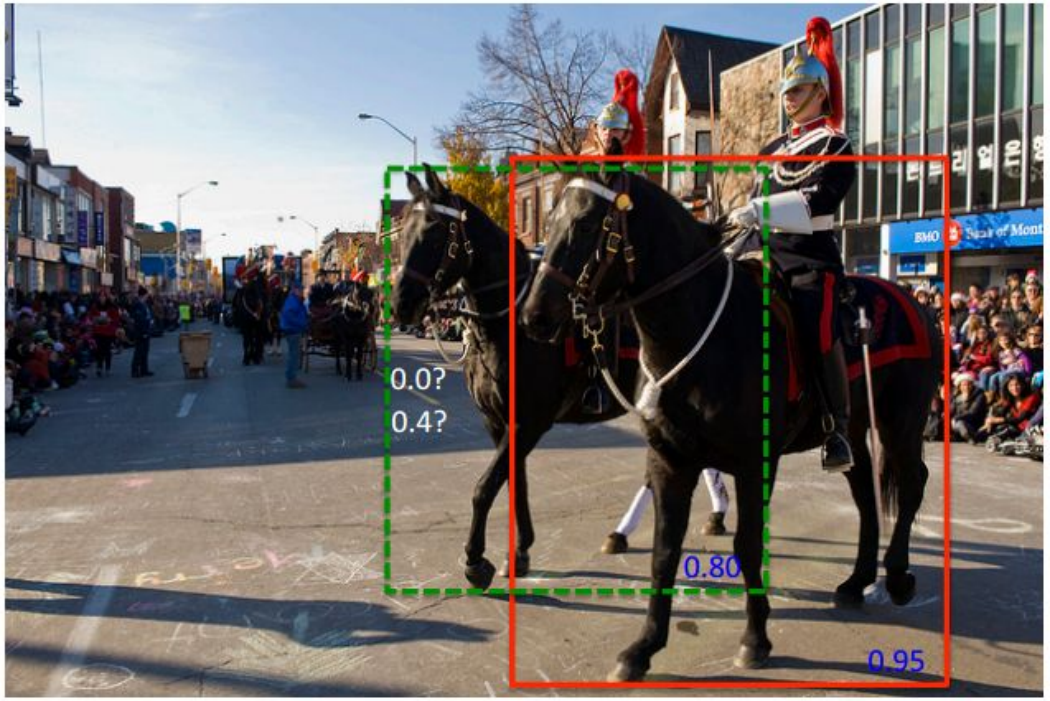
0.7

0.6

Tradeoff between recall/precision

Soft NMS

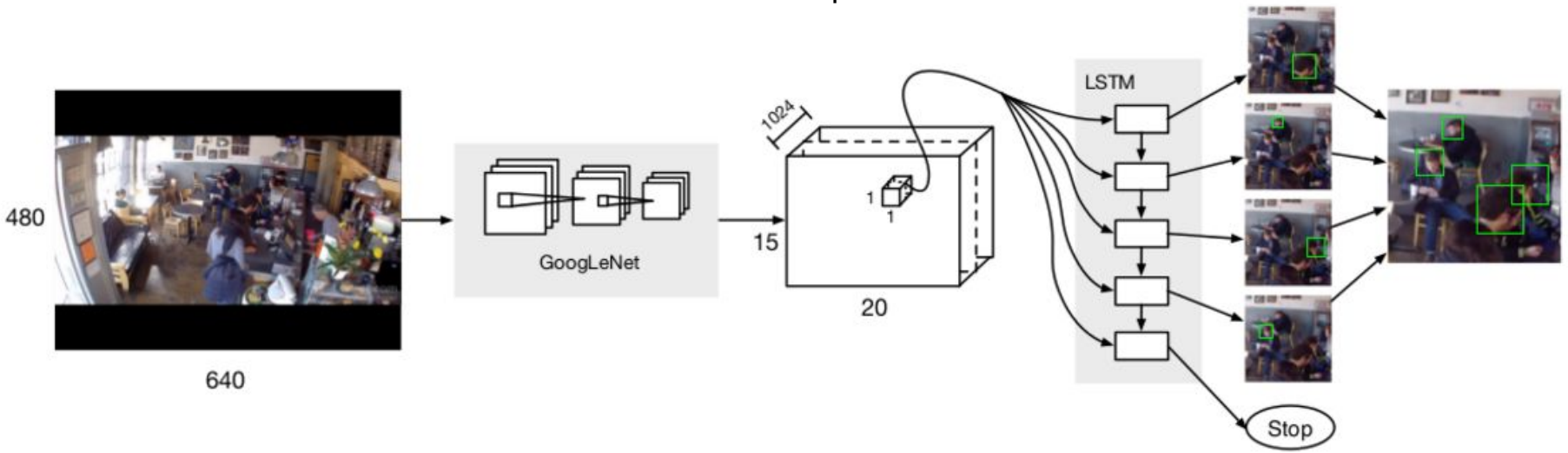
Decay detection scores of contiguous objects instead of setting them to 0



Bodla, Singh et al. Improving Object Detection With One Line of Code. arXiv Apr 2017

Avoid NMS: Sequential box prediction

Predict boxes one after the other and learn when to stop



Stewart et al. End-To-End People Detection in Crowded Scenes. CVPR 2016