Insight
Centre for Data Analytics

DEEP LEARNING WORKSHOP

Dublin City University
21-22 May 2018

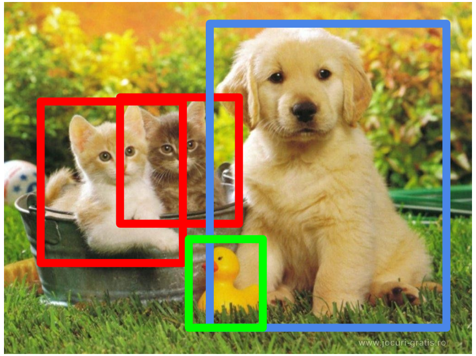
Day 2 Lecture 6

Object Detection

Amaia Salvador
amaia.salvador@upc.edu

PhD Candidate
Universitat Politècnica de Catalunya

Object Detection



CAT, DOG, DUCK

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

Object Detection: Datasets

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

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

COCO
Common Objects in Context

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

IMAGENET

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

Outline

Two-stage methods

One-stage methods

Object Detection as Classification

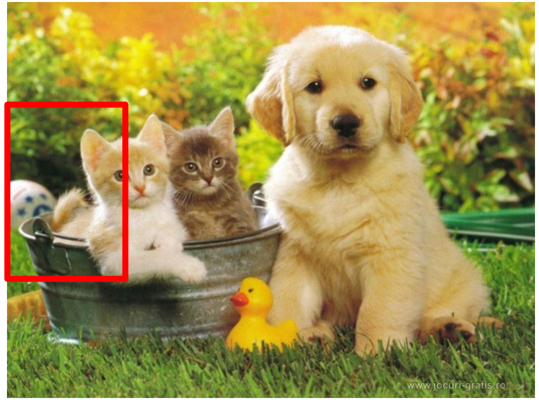
Classes = [cat, dog, duck]

Cat ? NO

Dog ? NO

Duck? NO

Object Detection as Classification



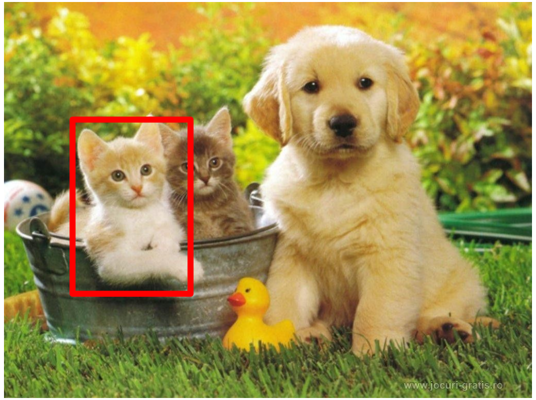
Classes = [cat, dog, duck]

Cat ? NO

Dog ? NO

Duck? NO

Object Detection as Classification



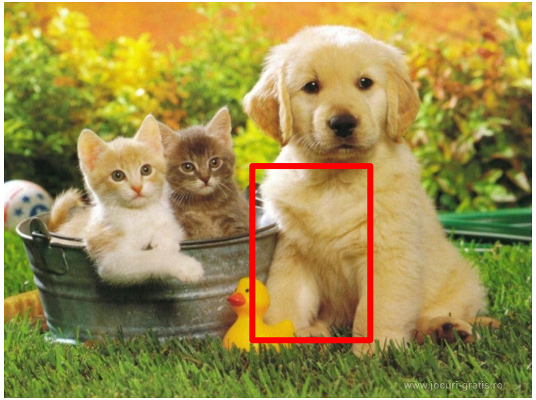
Classes = [cat, dog, duck]

Cat ? YES

Dog ? NO

Duck? NO

Object Detection as Classification



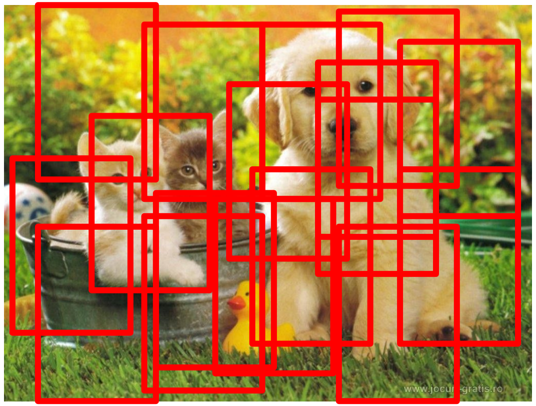
Classes = [cat, dog, duck]

Cat ? NO

Dog ? NO

Duck? NO

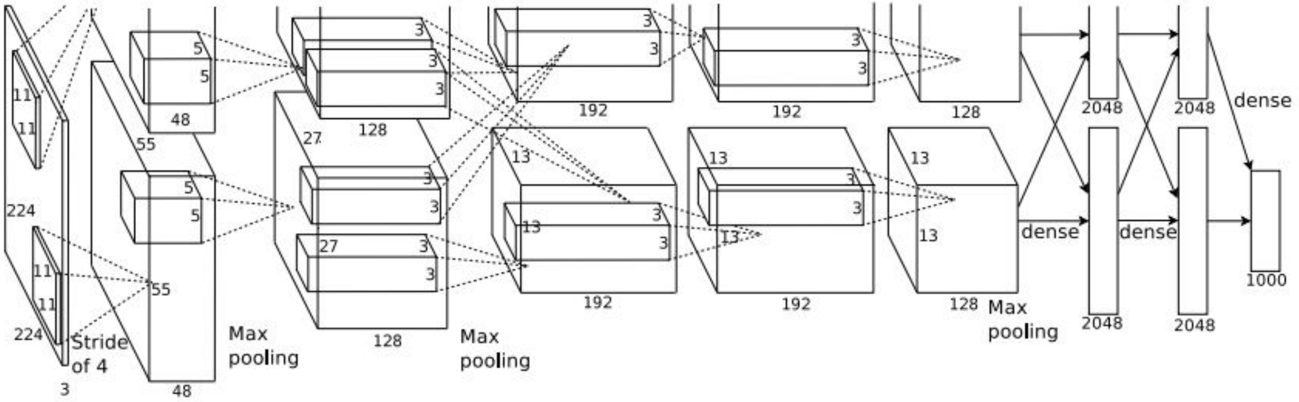
Object Detection as Classification



Problem:
Too many positions & scales to test

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

Object Detection with ConvNets?

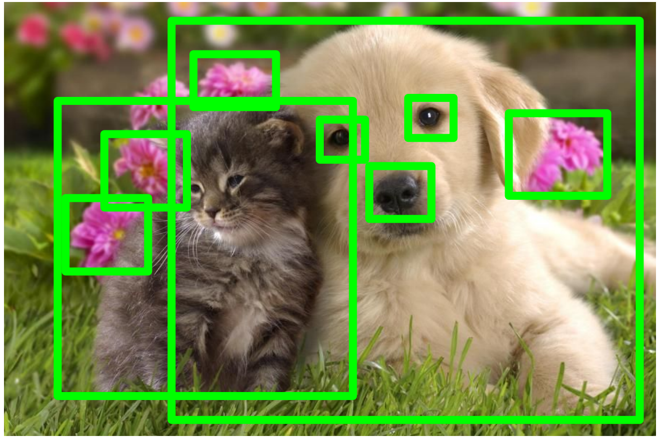


Convnets are computationally demanding. We can’t test all positions & scales !

Solution: Look at a tiny subset of positions & choose them wisely

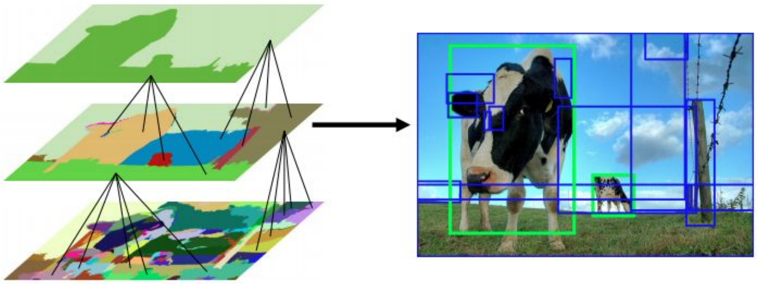
Region Proposals

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

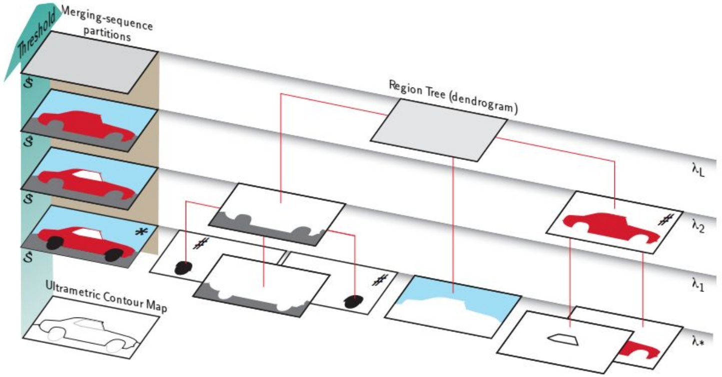


Slide Credit: [CS231n]

Region Proposals



Selective Search (SS)

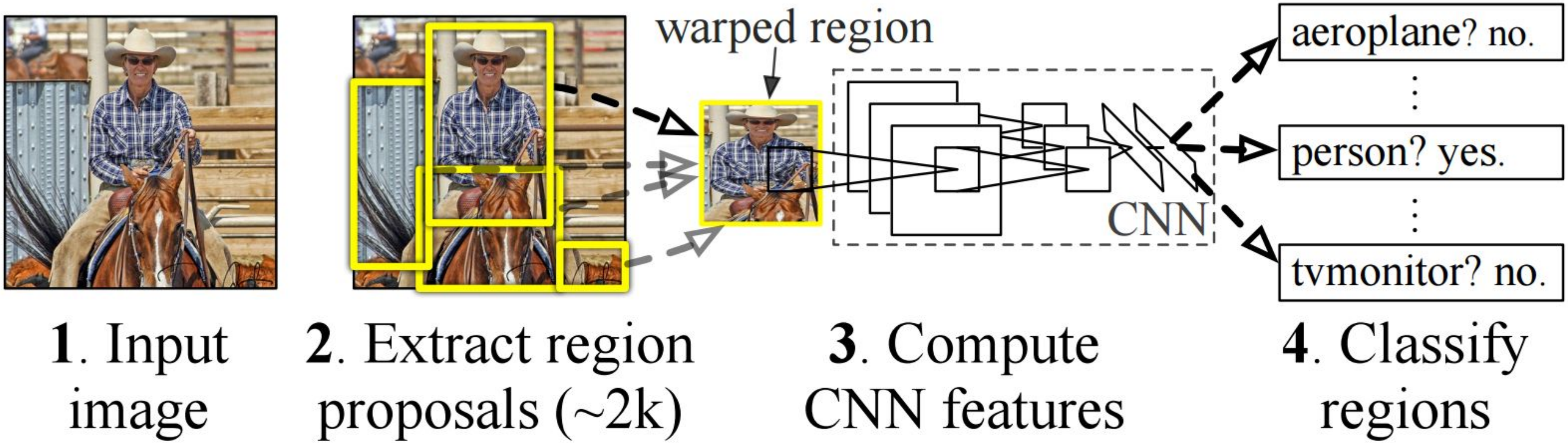


Multiscale Combinatorial Grouping (MCG)

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

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

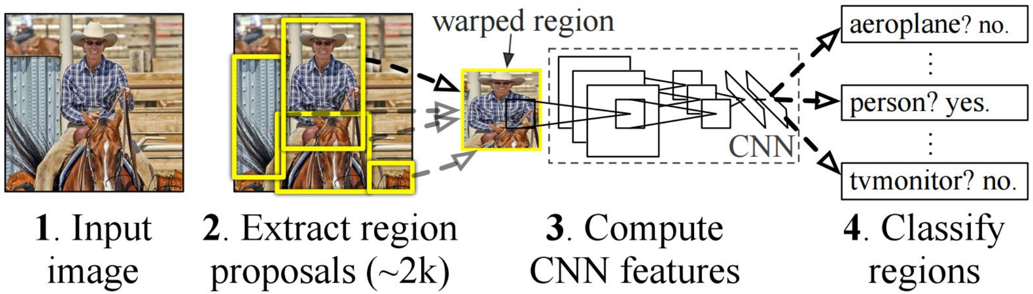
Object Detection with Convnets: R-CNN



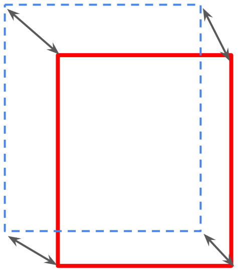
Girshick et al. [Rich feature hierarchies for accurate object detection and semantic segmentation](). CVPR 2014

R-CNN

1. Train network on proposals



2. Post-hoc training of SVM classifiers & bounding box regressors on fc7 features

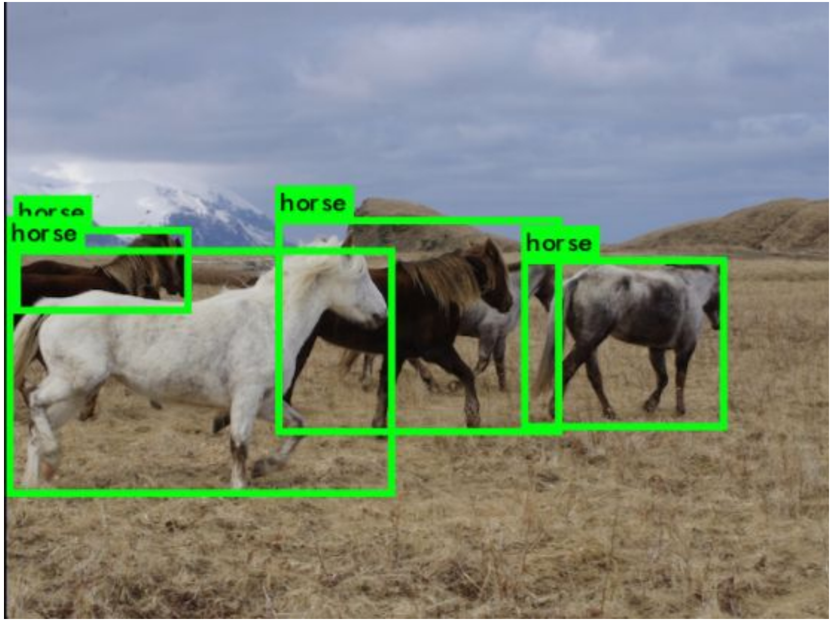


bounding box regressor predicts coordinate offsets

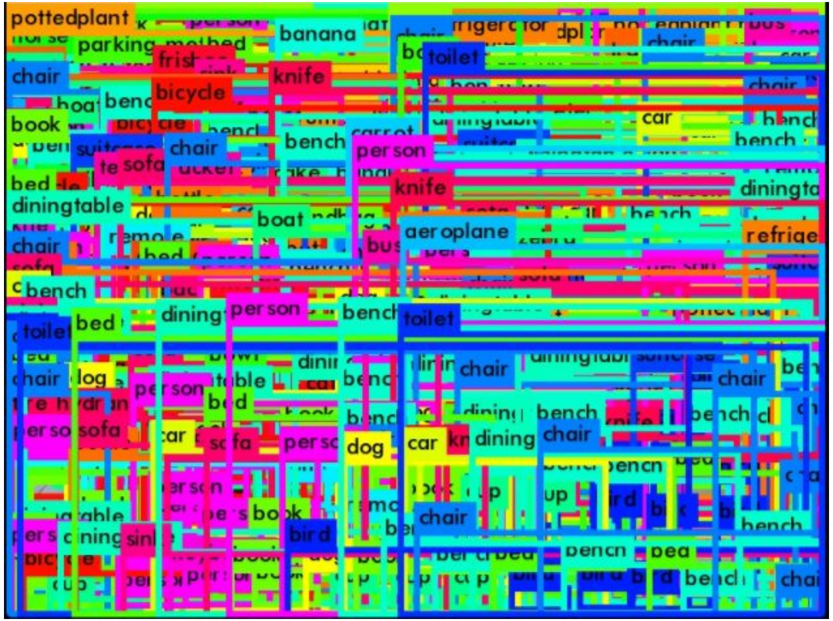
Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR 2014

R-CNN

We expect:

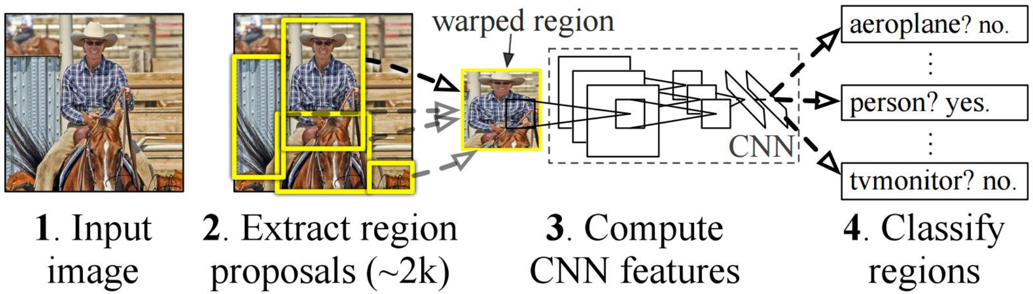


We get:



R-CNN

1. Train network on proposals

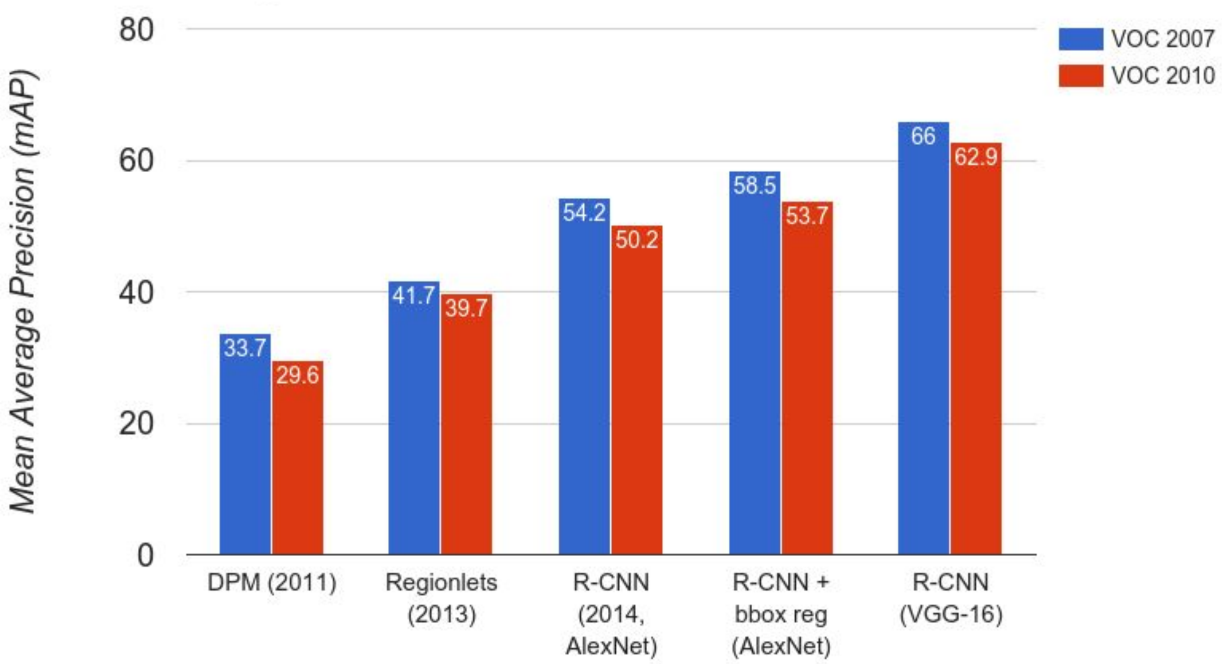


2. Post-hoc training of SVMs & Box regressors on fc7 features

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

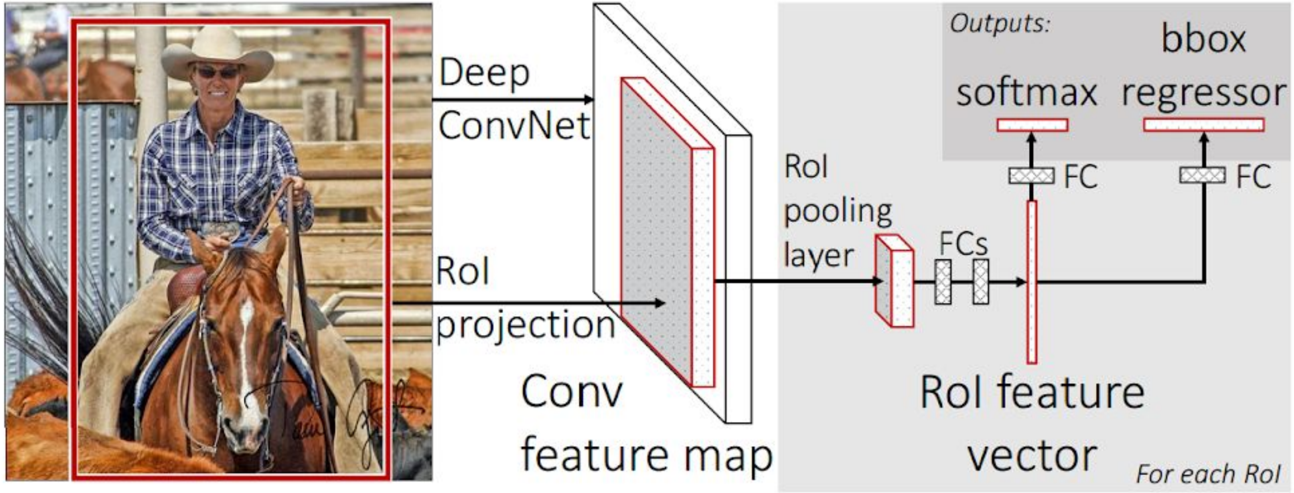
2. SVMs and regressors are post-hoc: CNN features not updated in response to SVMs and regressors

3. Complex multistage training pipeline

Slide Credit: [CS231n]

Fast R-CNN

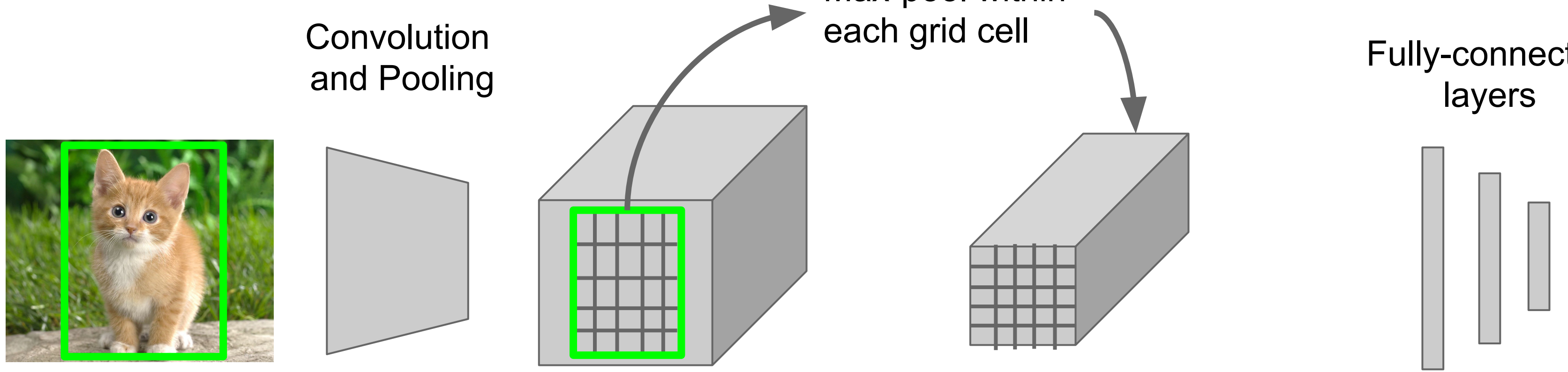
R-CNN Problem #1: Slow at test-time: need to run full forward pass of CNN for each region proposal



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

Girshick Fast R-CNN. ICCV 2015

Fast R-CNN: Sharing features

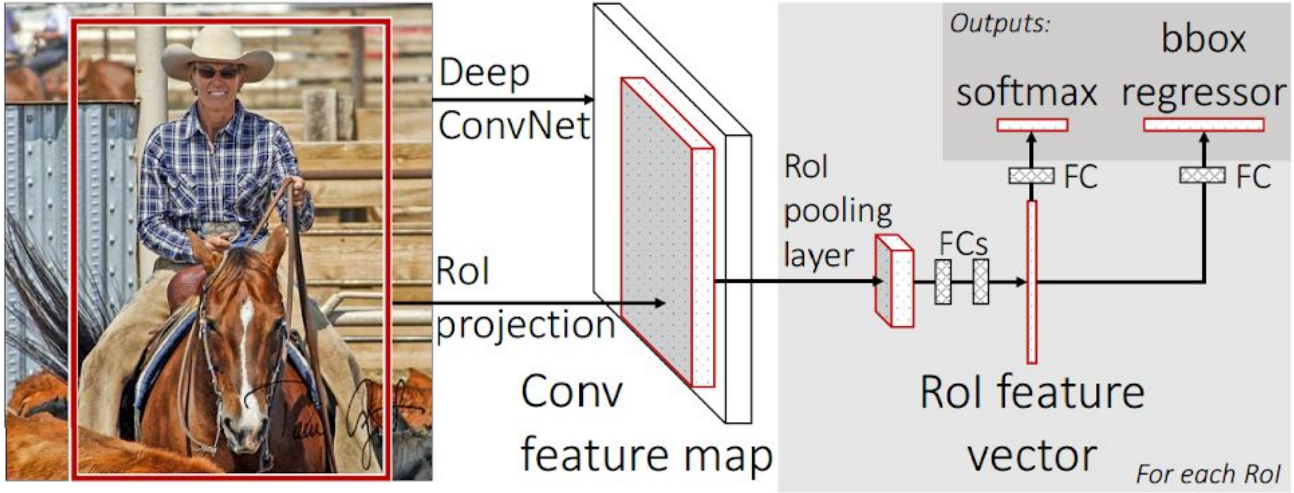


Girshick Fast R-CNN. ICCV 2015

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

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

Using VGG-16 CNN on Pascal VOC 2007 dataset

Slide Credit: [CS231n]

Fast R-CNN: Problem

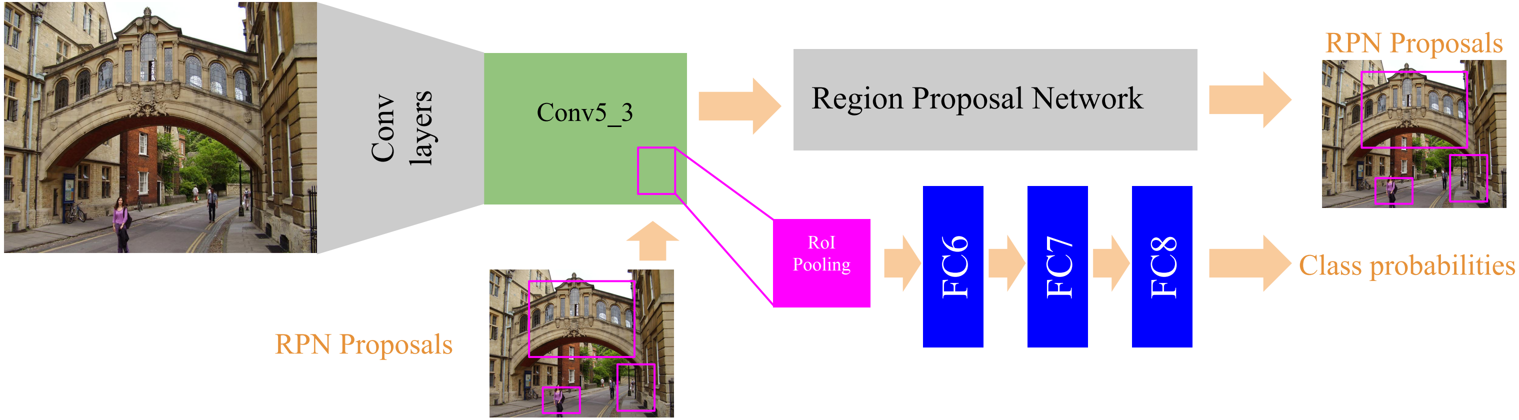
Test-time speeds don't include region proposals

	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

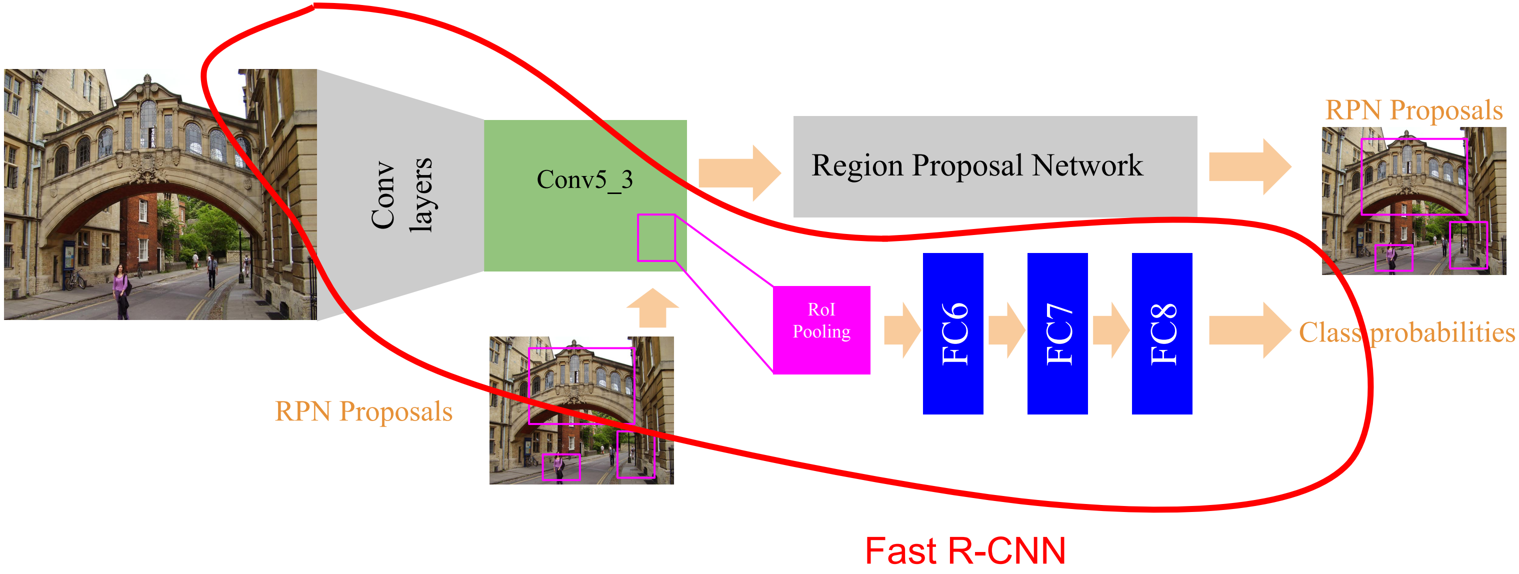
Learn proposals end-to-end sharing parameters with the classification network



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

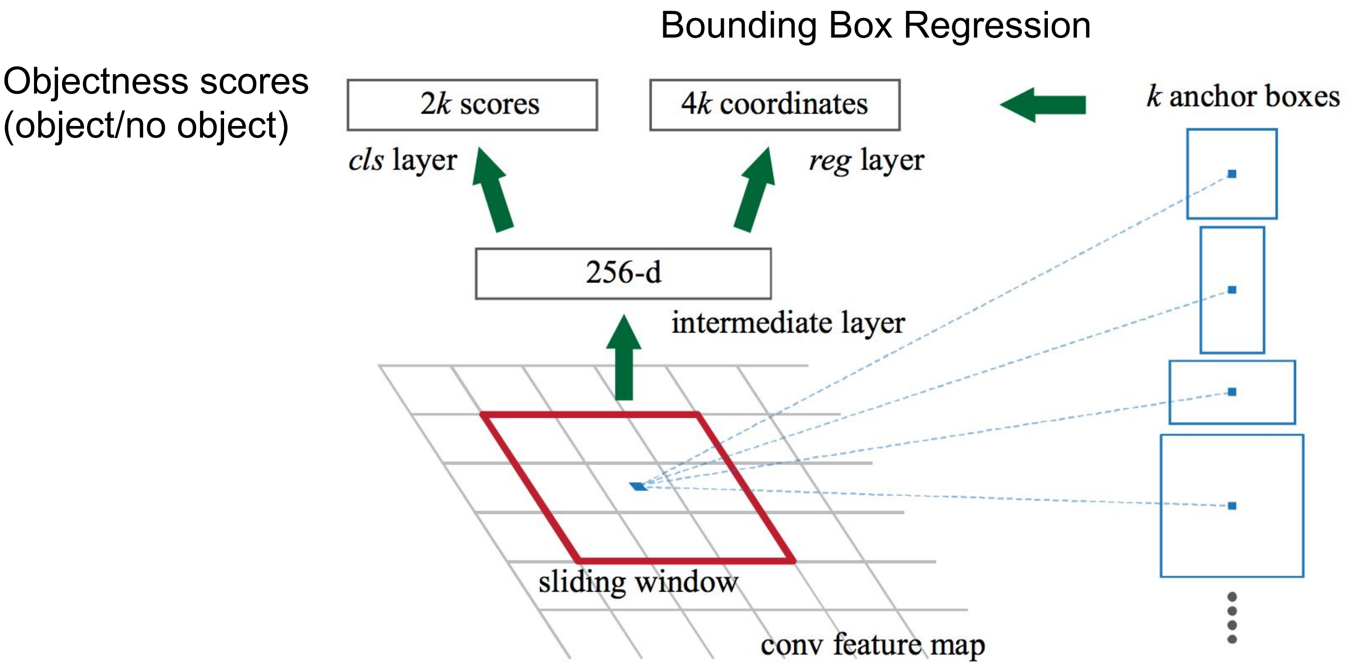
Faster R-CNN

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



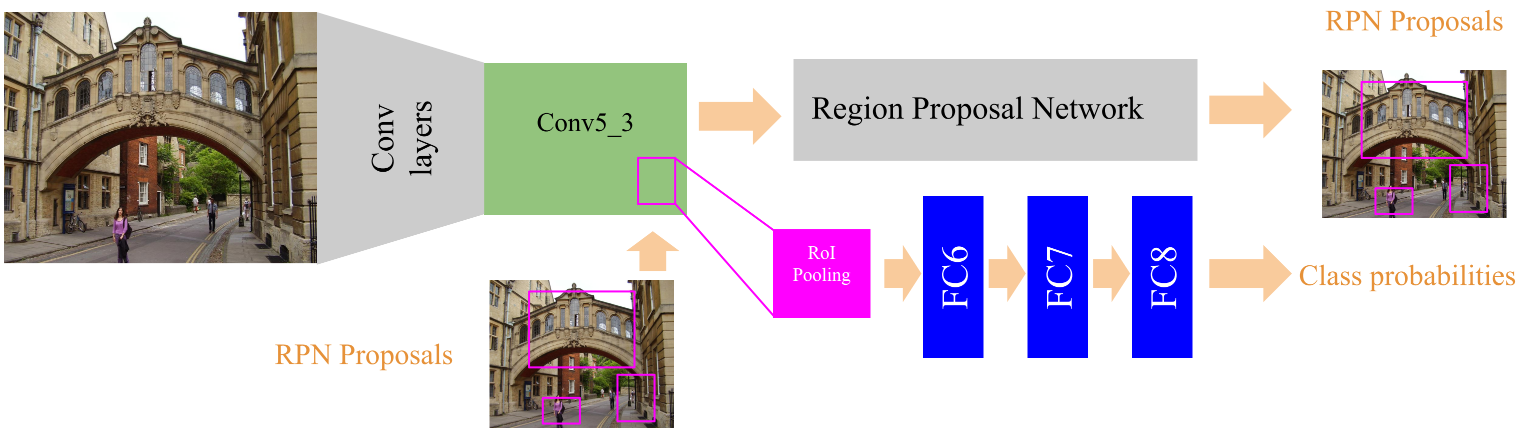
In practice, k = 9 (3 different scales and 3 aspect ratios) → 18k boxes for a 40x50 input feature map

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

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

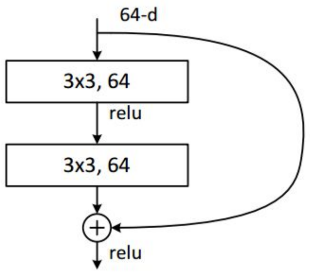
	R-CNN	Fast R-CNN	Faster R-CNN
Test time per image (with proposals)	50 seconds	2 seconds	**0.2 seconds**
(Speedup)	1x	25x	**250x**
mAP (VOC 2007)	66.0	**66.9**	**66.9**

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

Slide Credit: CS231n

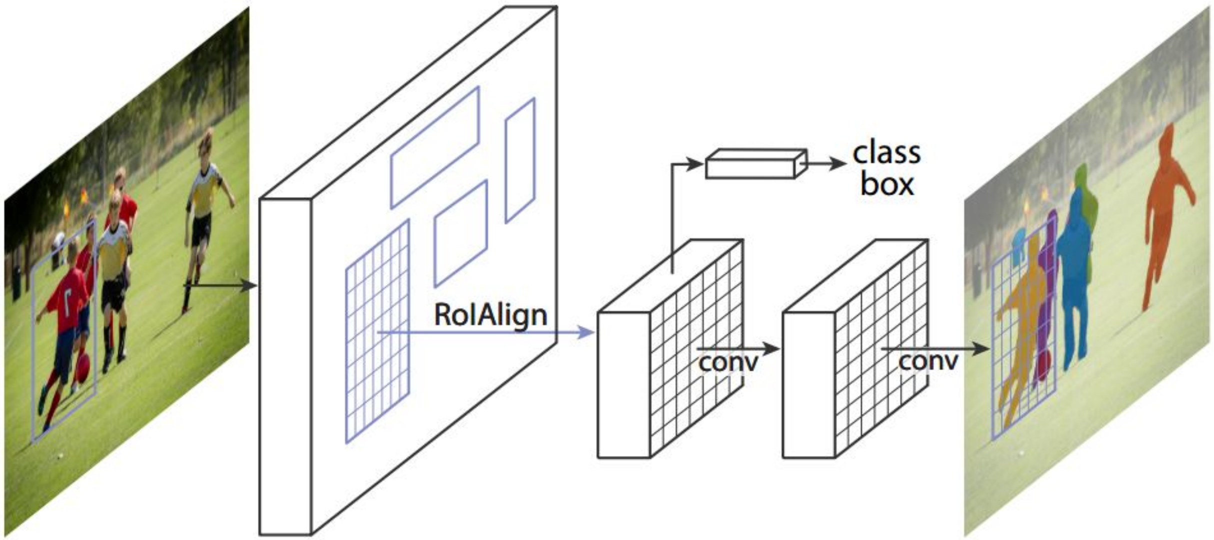
Better Encoder: ResNet

- Faster R-CNN was the basis of the winners of COCO and ILSVRC 2015&2016 object detection competitions.



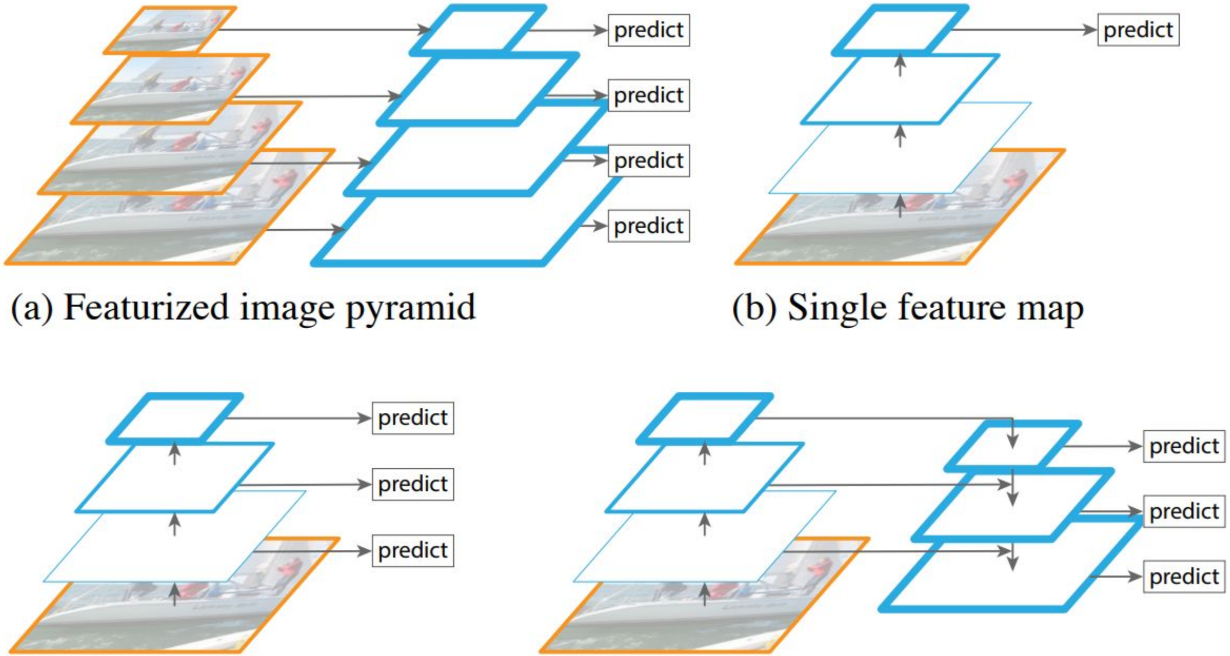
He et al. Deep residual learning for image recognition. CVPR 2016

Better Region Pooling : RoI Align



He et al. [Mask R-CNN](). ICCV 2017

Better Representations: FPN



(a) Featurized image pyramid (b) Single feature map

(c) Pyramidal feature hierarchy (d) Feature Pyramid Network

Lin et al. [Feature Pyramid Networks for Object Detection.](https://arxiv.org/abs/1612.03144) CVPR 2017

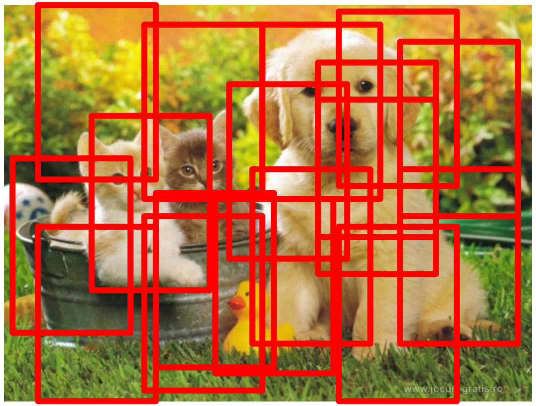
Outline

Two-stage methods

One-stage methods

One-stage methods

Previously… :

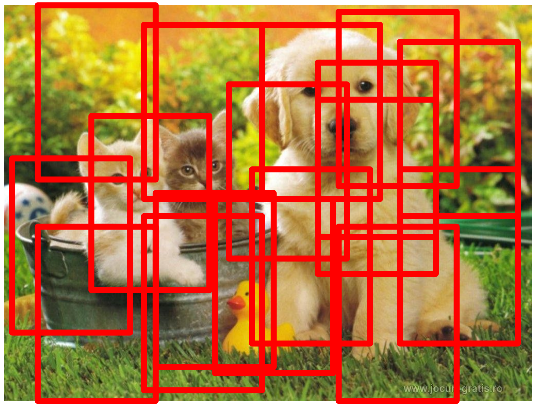


Problem:
Too many positions & scales to test

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

One-stage methods

Previously… :

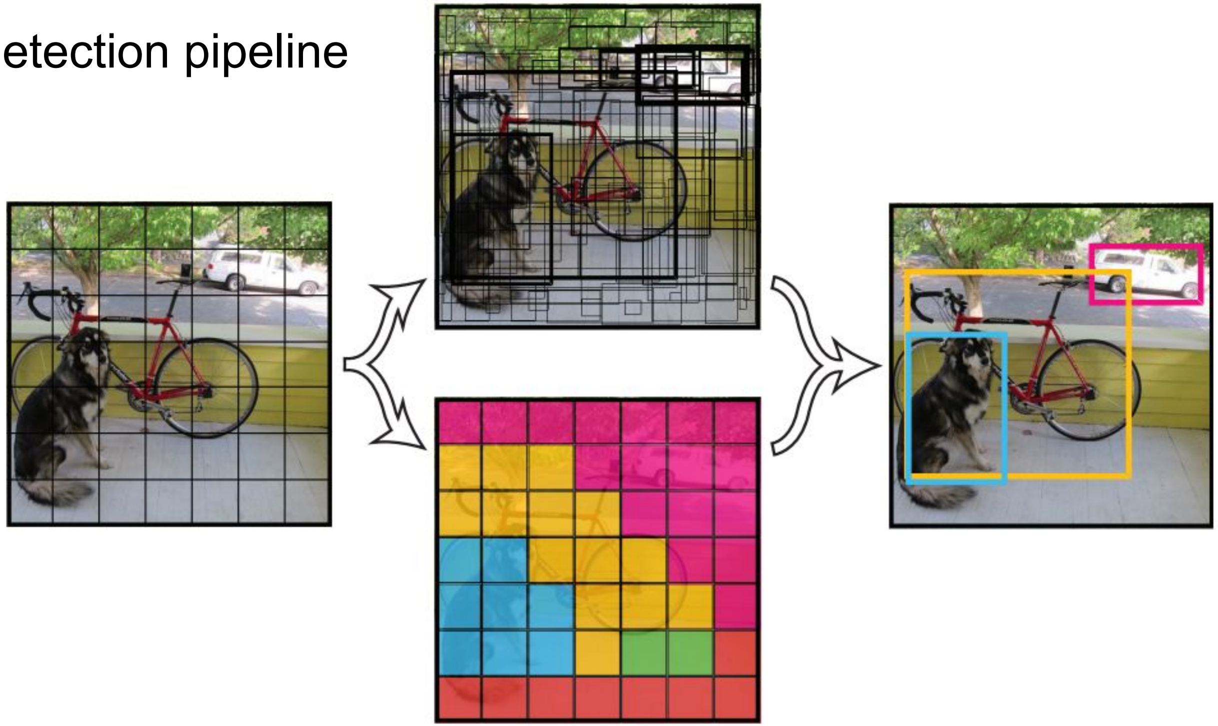


Problem:
Too many positions & scales to test

Modern detectors parallelize feature extraction across all locations.
Region classification is not slow anymore!

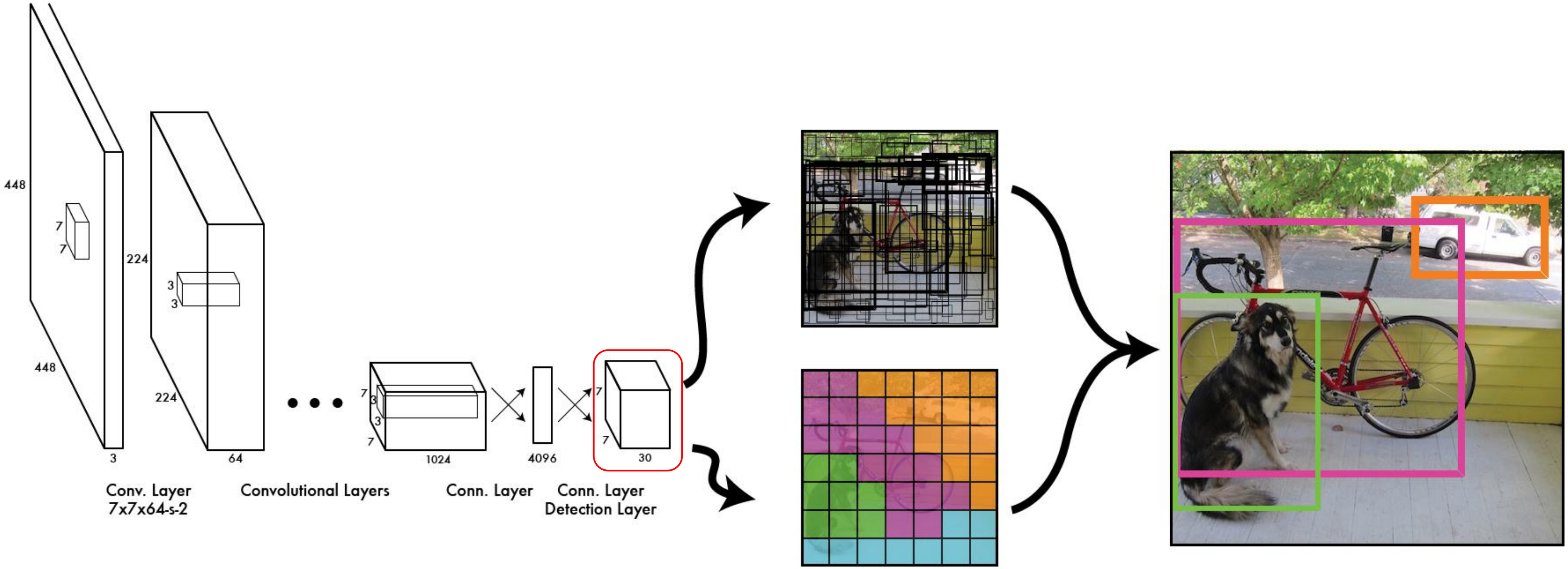
YOLO: You Only Look Once

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

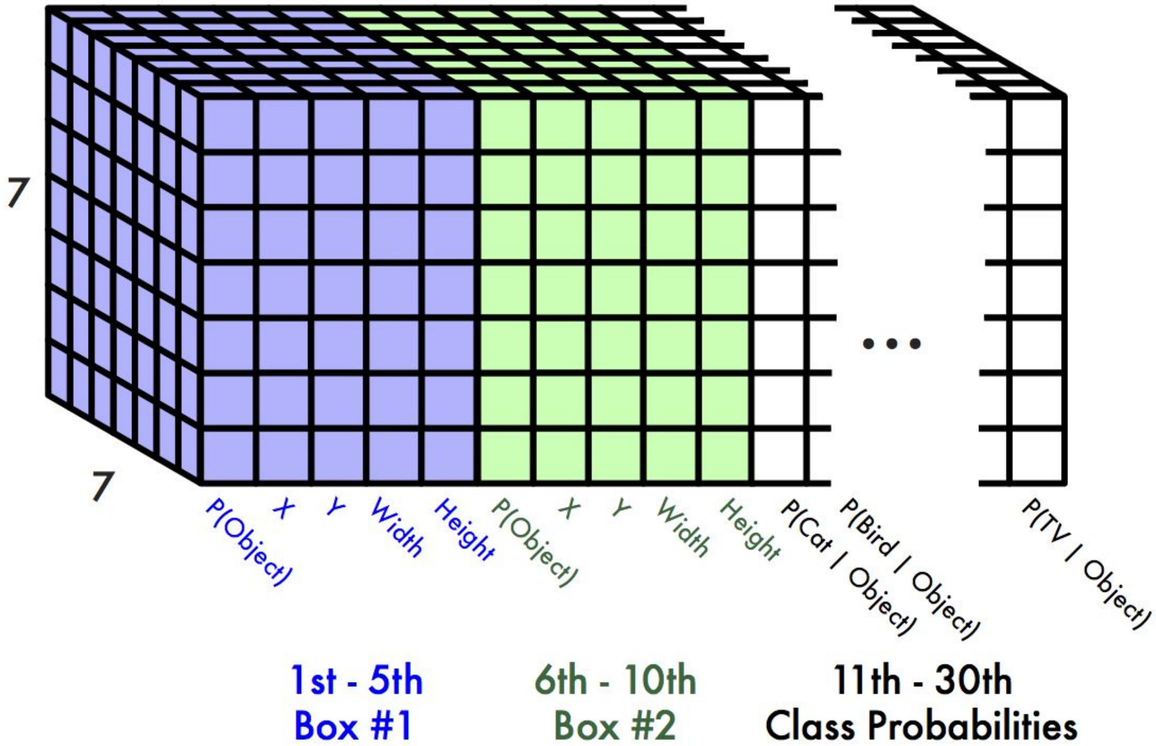
YOLO: You Only Look Once

Each cell predicts:

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

For Pascal VOC:

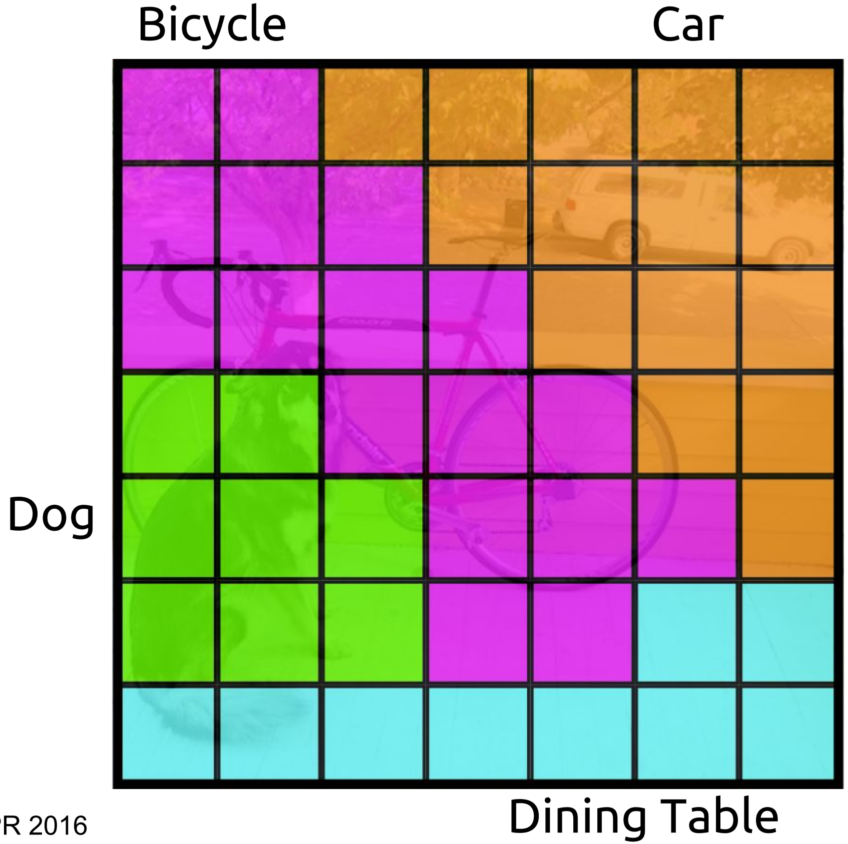
- 7x7 grid
- 2 bounding boxes / cell
- 20 classes



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

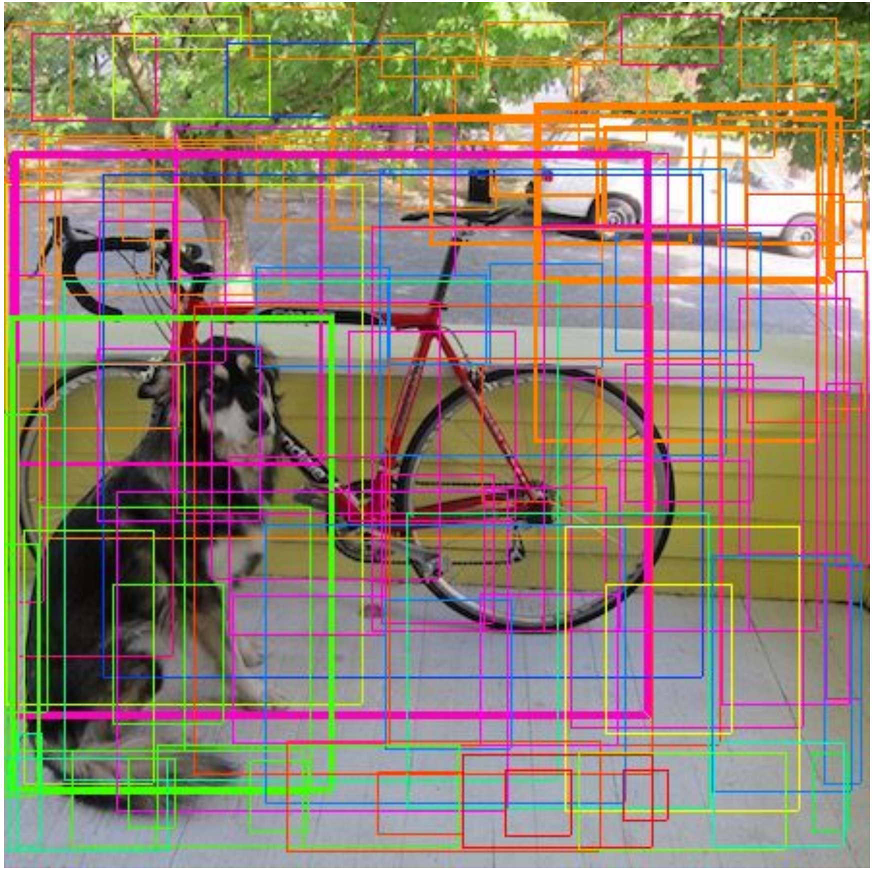
YOLO: You Only Look Once

Predict class probability for each cell



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

YOLO: You Only Look Once

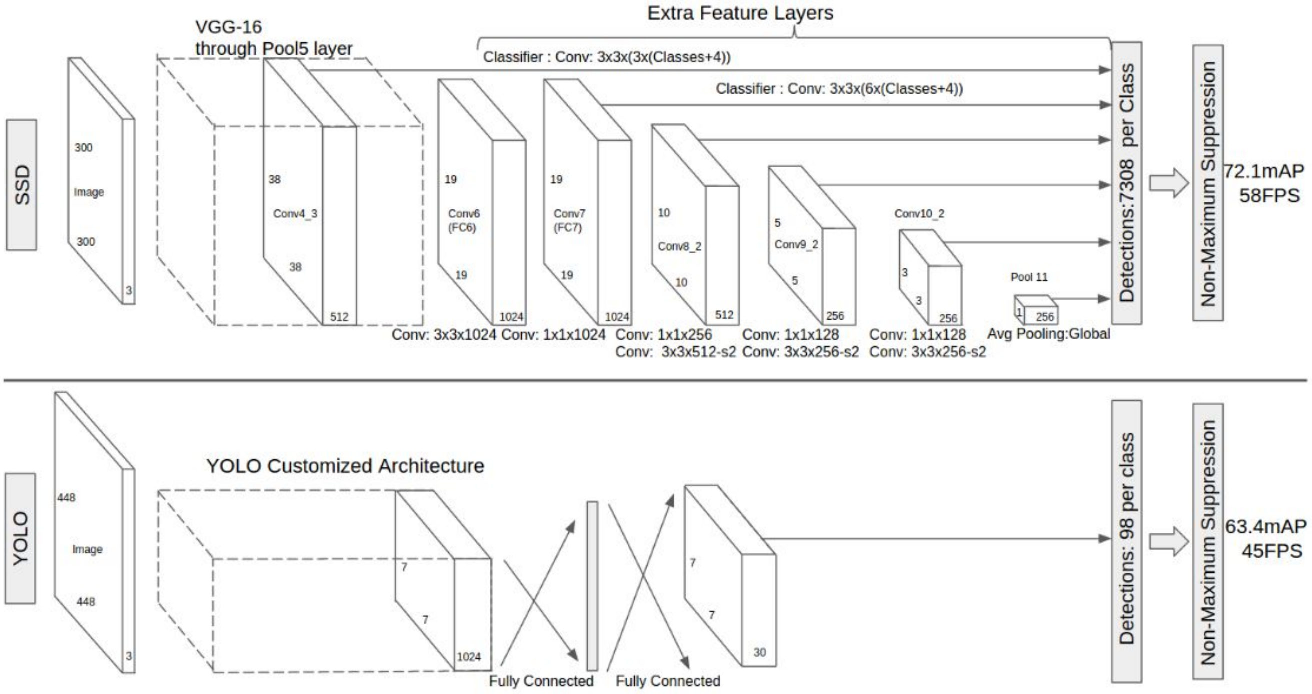


+ NMS
+ Score threshold

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

SSD: Single Shot MultiBox Detector

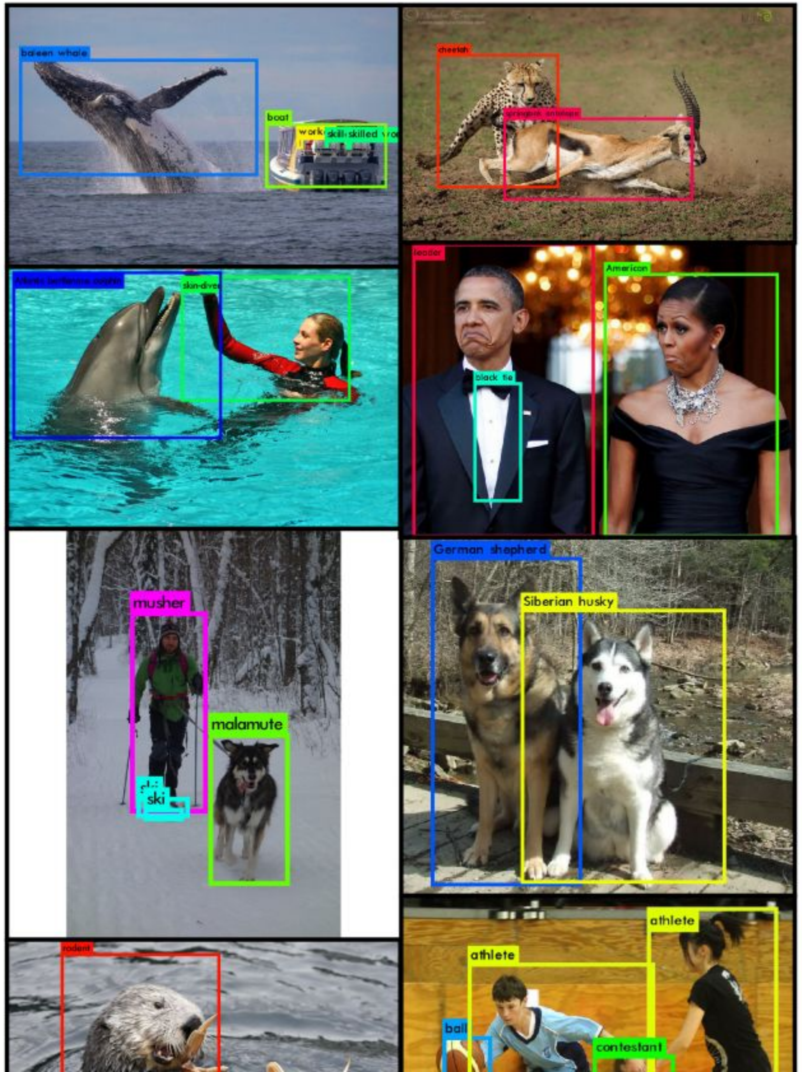
Same idea as YOLO, + several predictors at different stages in the network & uses anchor boxes



Liu et al. [SSD: Single Shot MultiBox Detector](), ECCV 2016

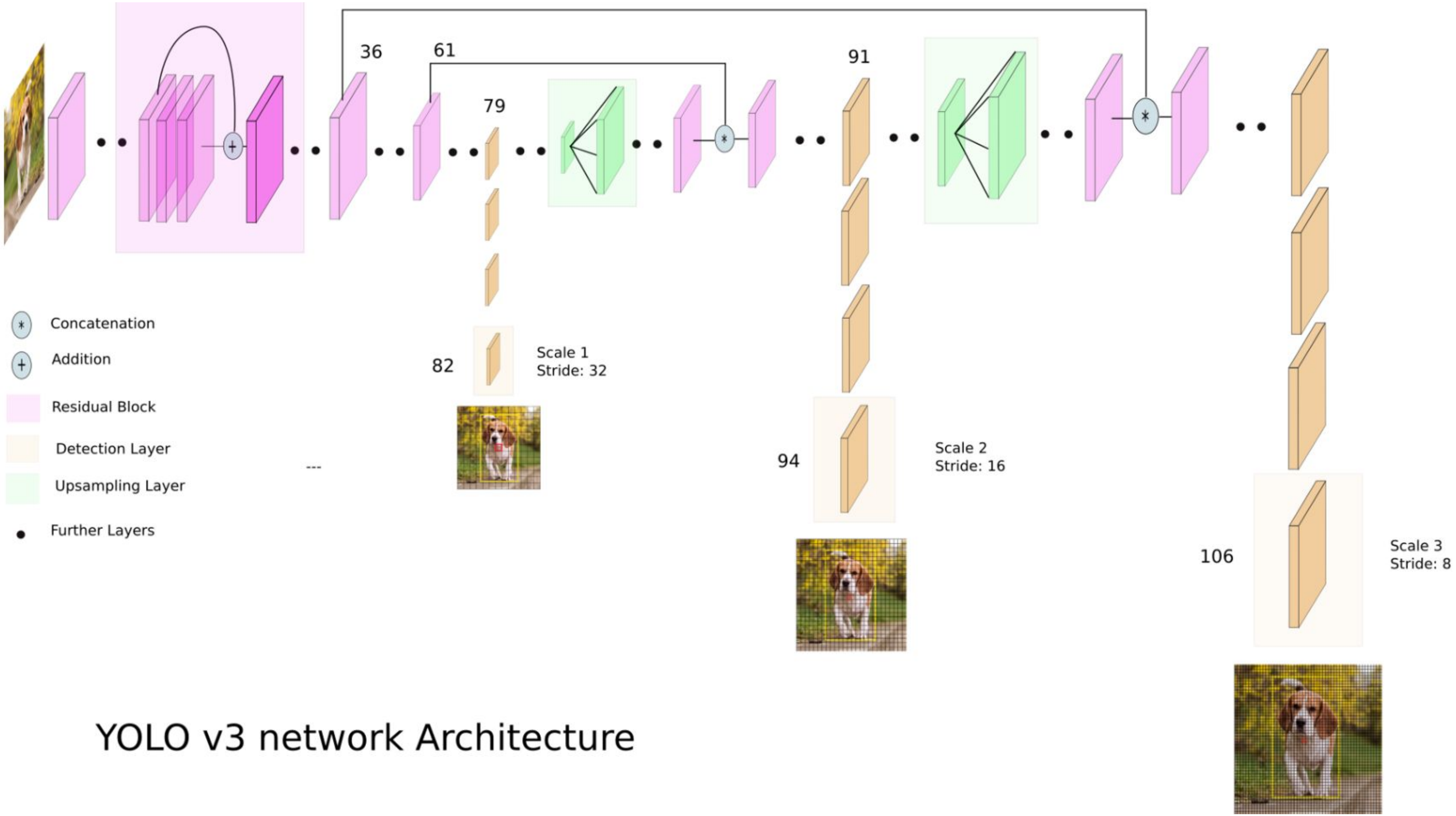
YOLOv2

	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

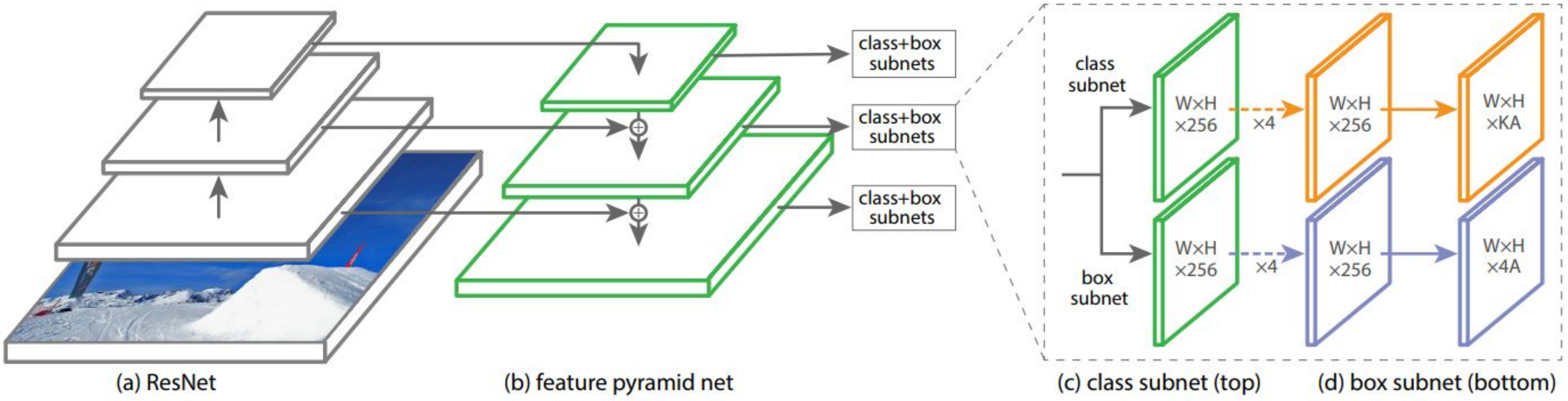
YOLOv3



YOLO v3 network Architecture

RetinaNet

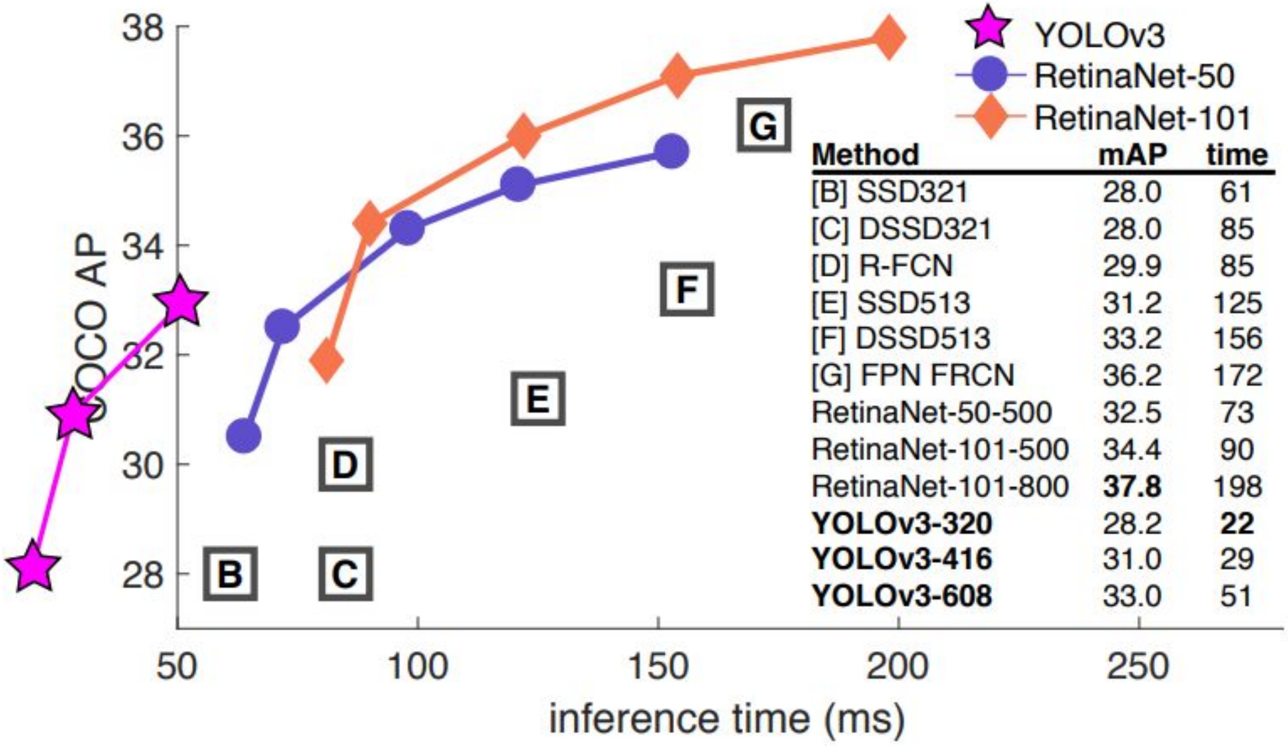
Matching proposal-based performance with a one-stage approach



Key idea is to lower loss weight for well classified samples, increase it for difficult ones

Lin et al. [Focal Loss for Dense Object Detection.] ICCV 2017

Overview



YOLOv3
RetinaNet-50
RetinaNet-101
Method mAP time
[B] SSD321 28.0 61
[C] DSSD321 28.0 85
[D] R-FCN 29.9 85
[E] SSD513 31.2 125
[F] DSSD513 33.2 156
[G] FPN FRCN 36.2 172
RetinaNet-50-500 32.5 73
RetinaNet-101-500 34.4 90
RetinaNet-101-800 37.8 198
YOLOv3-320 28.2 22
YOLOv3-416 31.0 29
YOLOv3-608 33.0 51
COCO AP
38
36
34
32
30
28
50
100
150
200
250
inference time (ms)
B
C
D
E
F
G

Summary

Two-stage methods

- R-CNN
- Fast R-CNN
- Faster R-CNN

One-stage methods

- YOLO
- SSD
- RetinaNet

Questions?

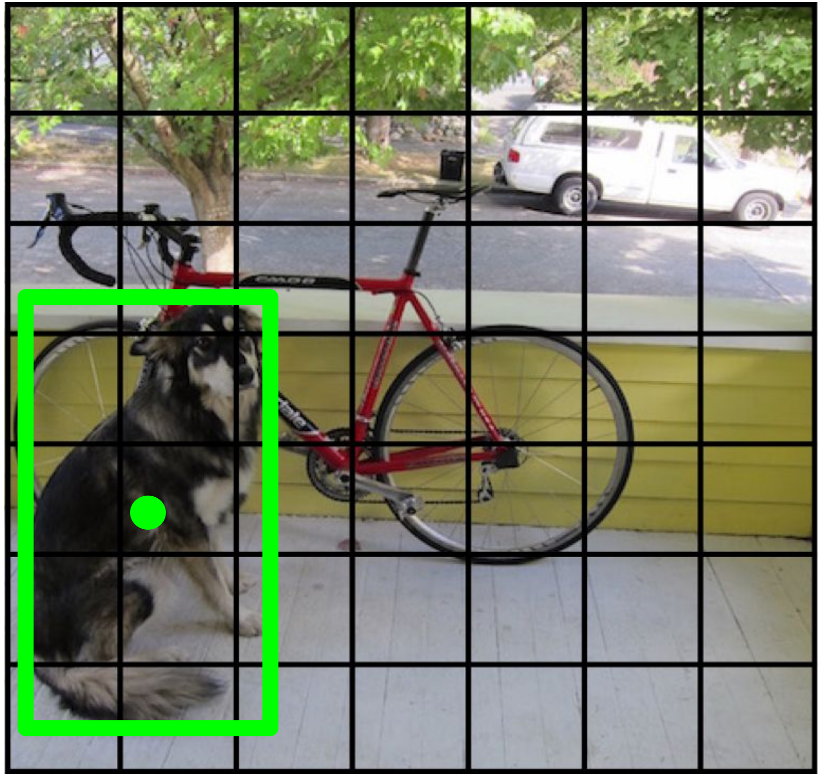
Resources

APIs including implementations to most popular detectors:

- Detectron: Facebook Object Detection API (Caffe2)
- Google's Tensorflow Object Detection API

Many unofficial ports to other frameworks !

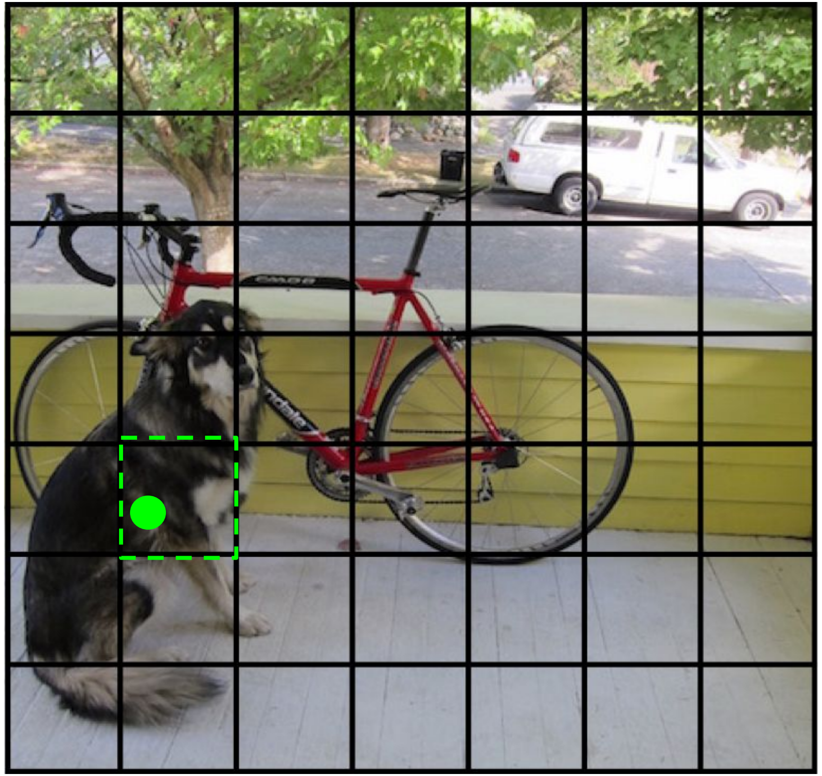
YOLO: Training



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

Slide credit: [YOLO Presentation @ CVPR 2016]

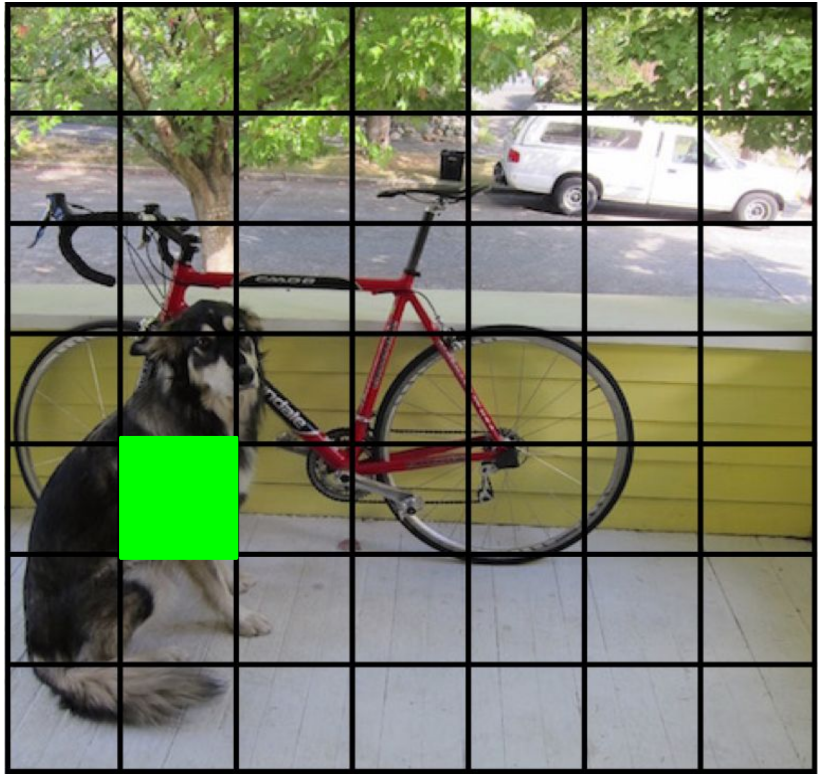
YOLO: Training



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

Slide credit: [YOLO Presentation @ CVPR 2016]

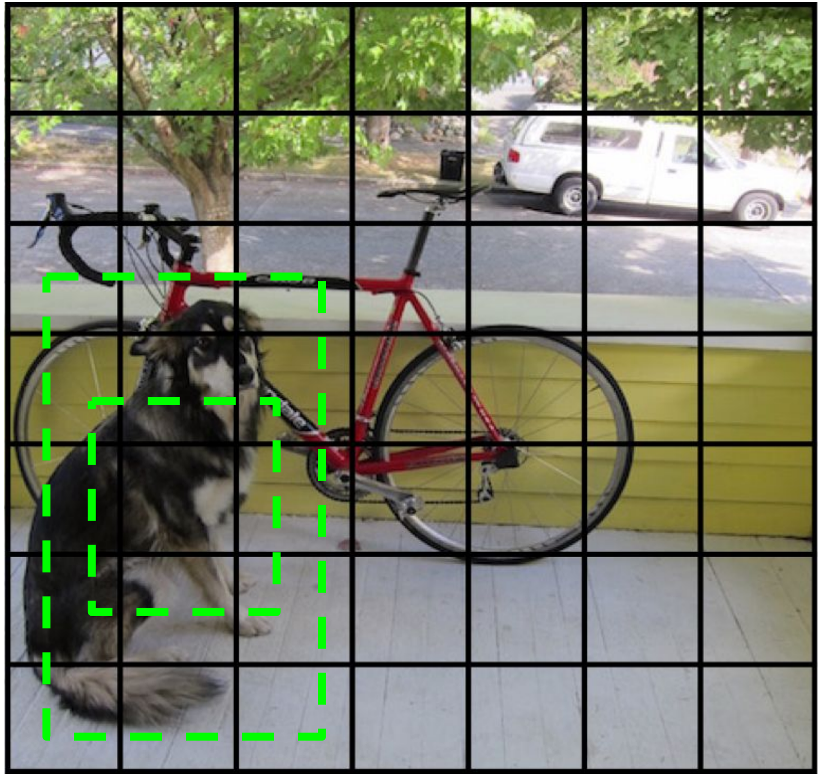
YOLO: Training



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

Slide credit: [YOLO Presentation @ CVPR 2016]

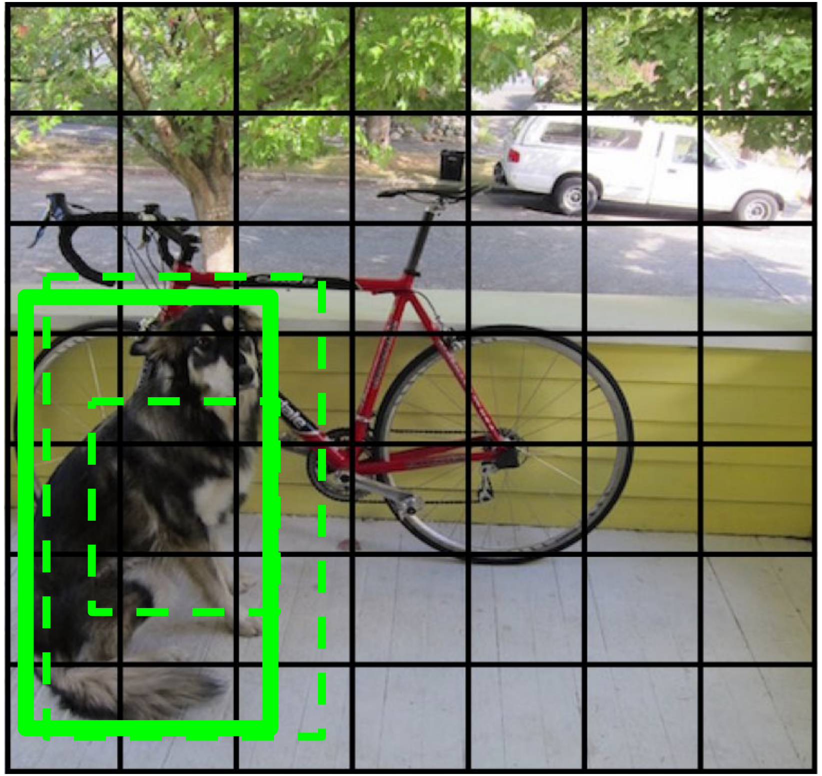
YOLO: Training



Predicted boxes for this cell

Slide credit: [YOLO Presentation @ CVPR 2016]

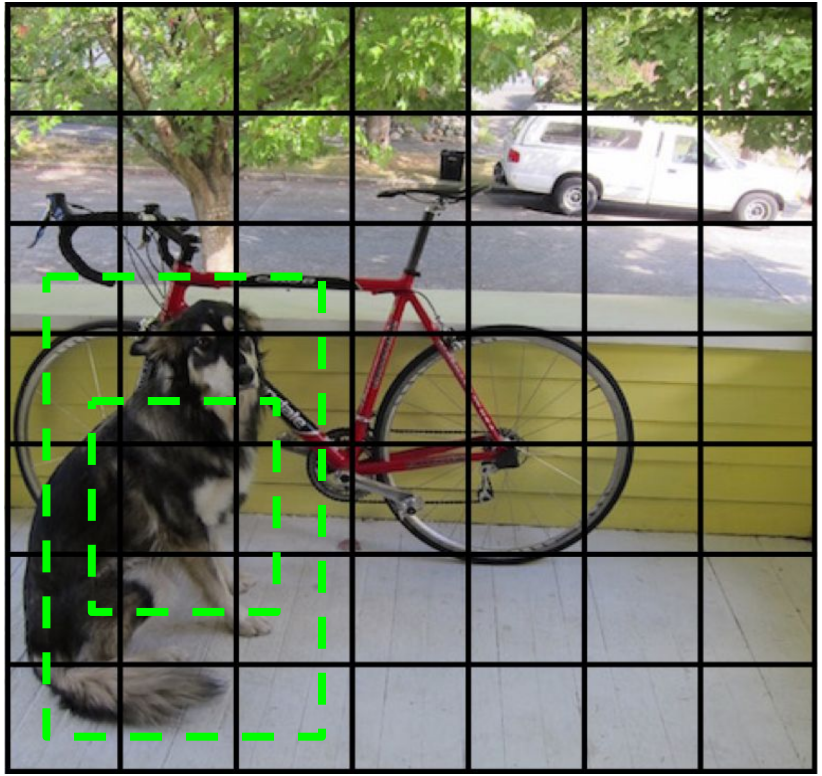
YOLO: Training



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]

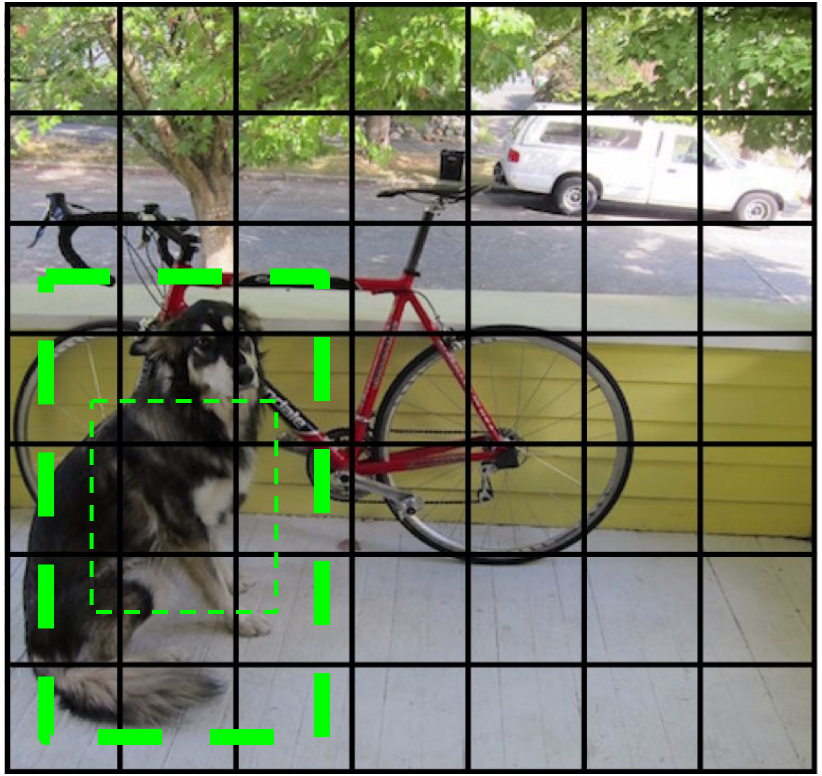
YOLO: Training



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

Slide credit: [YOLO Presentation @ CVPR 2016]

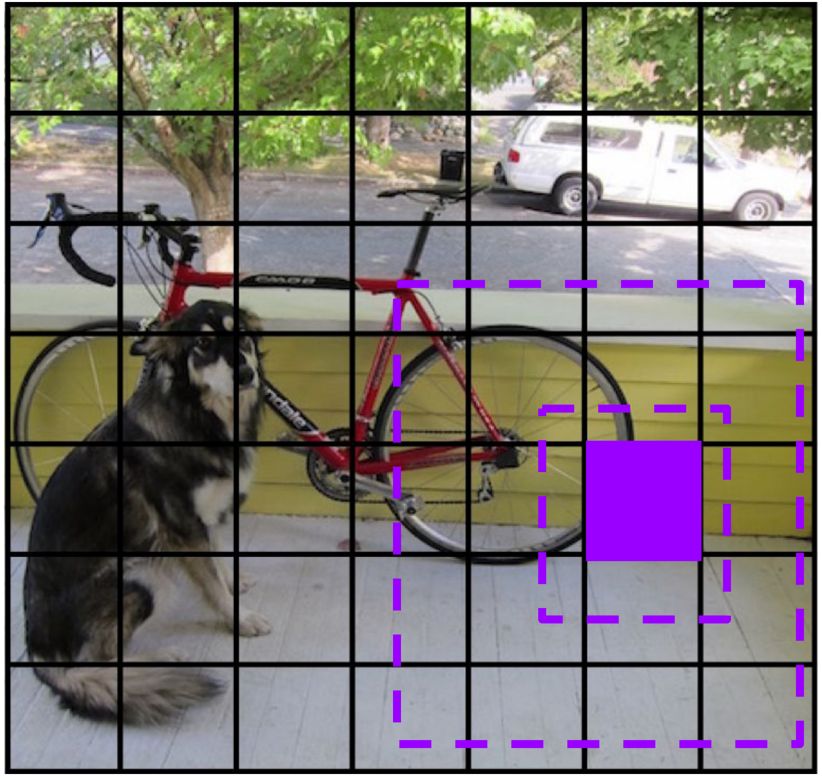
YOLO: Training



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

Slide credit: [YOLO Presentation @ CVPR 2016]

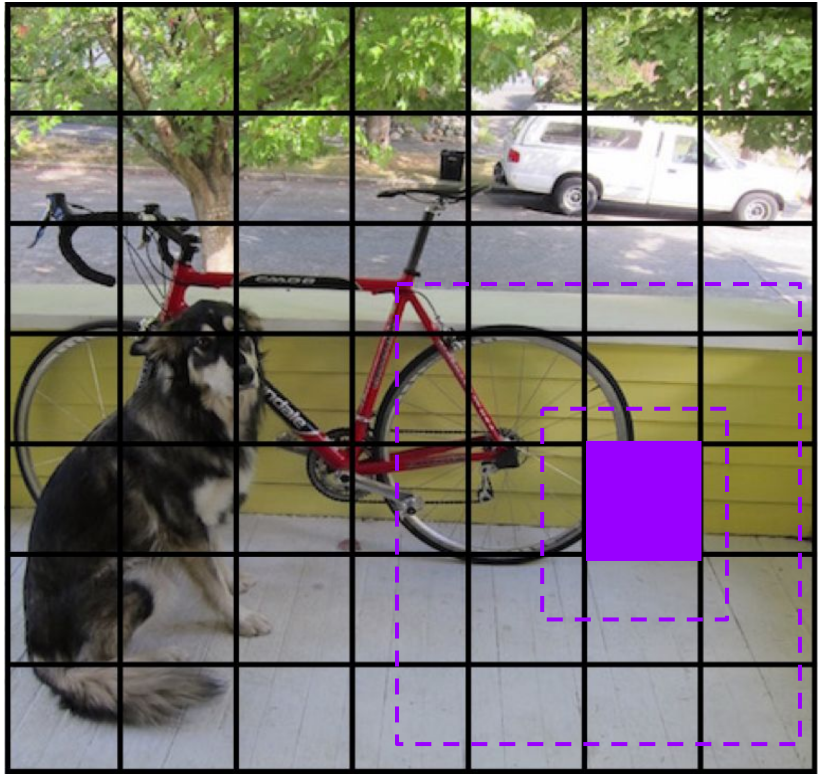
YOLO: Training



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

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

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

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

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

= 1 if cell *i* has an object present

Slide credit: [YOLO Presentation @ CVPR 2016]()

Fast R-CNN: RoI Pooling

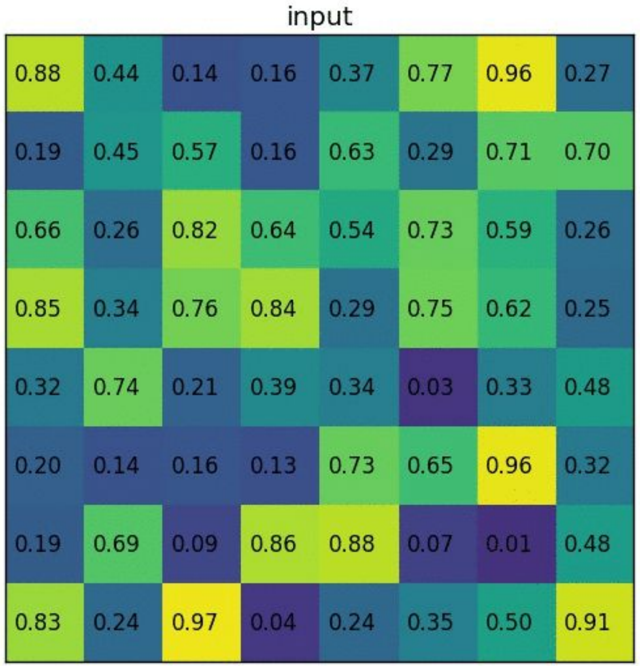


Image resource: [deepsense.ai]