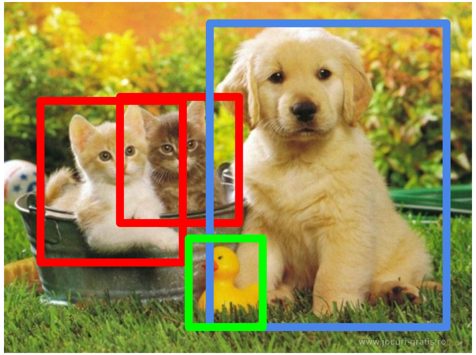
Deep learning and Artificial Neural Network for Object detection

Dr. Anthony Fleury - HDR
IMT Nord Europe

Dr. Sebastien Ambellouis
Ingénieur de recherche Université Gustave Eiffel
Associé à IMT Nord Europe

sebastien.ambellouis@univ-eiffel.fr
anthony.fleury@imt-nord-europe.fr

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

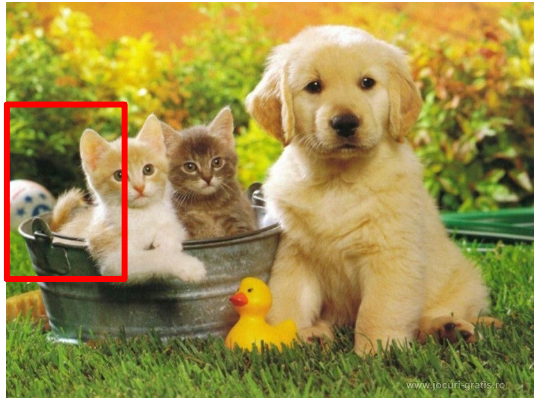
Classes = [cat, dog, duck]

Cat ? NO

Dog ? NO

Duck? NO

Object Detection as Classification



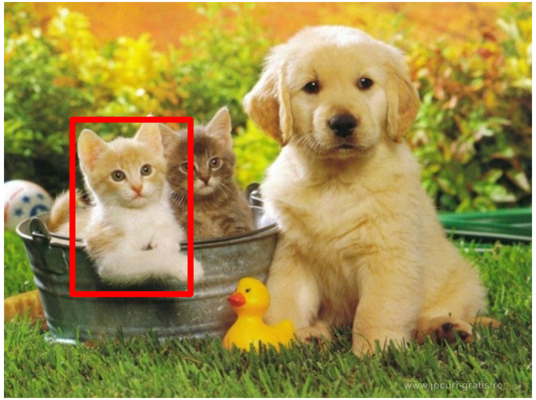
Classes = [cat, dog, duck]

Cat ? NO

Dog ? NO

Duck? NO

Object Detection as Classification



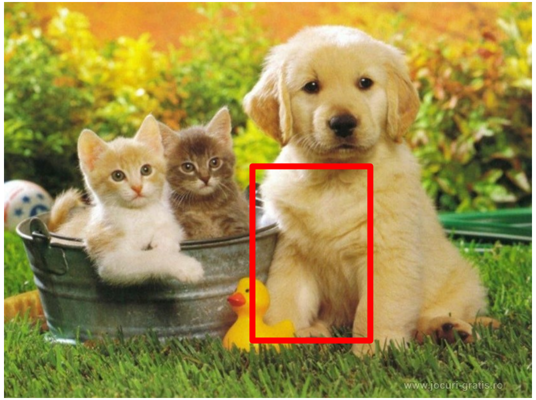
Classes = [cat, dog, duck]

Cat ? YES

Dog ? NO

Duck? NO

Object Detection as Classification



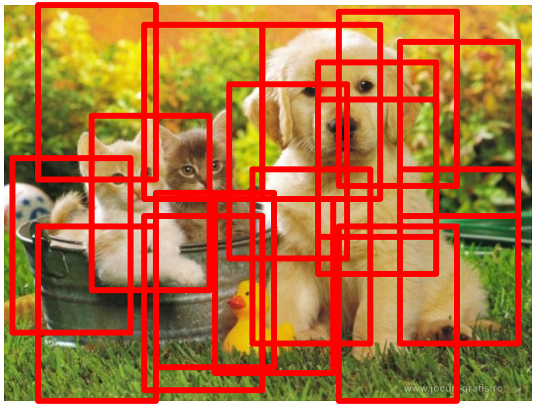
Classes = [cat, dog, duck]

Cat ? NO

Dog ? NO

Duck? NO

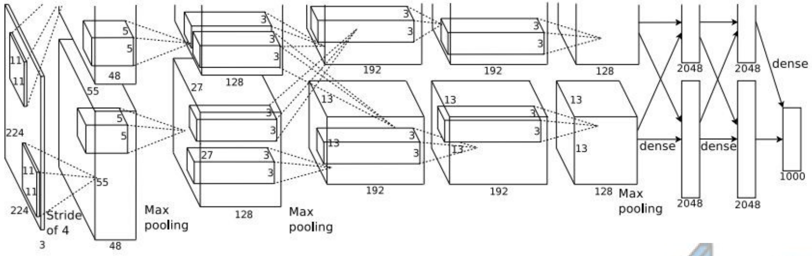
Object Detection as Classification



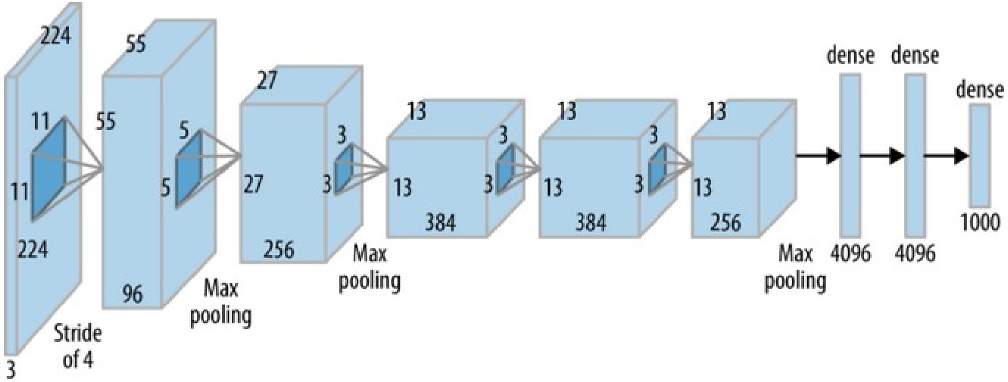
Problem:
Too many positions & scales to test

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

Object Detection with ConvNets?



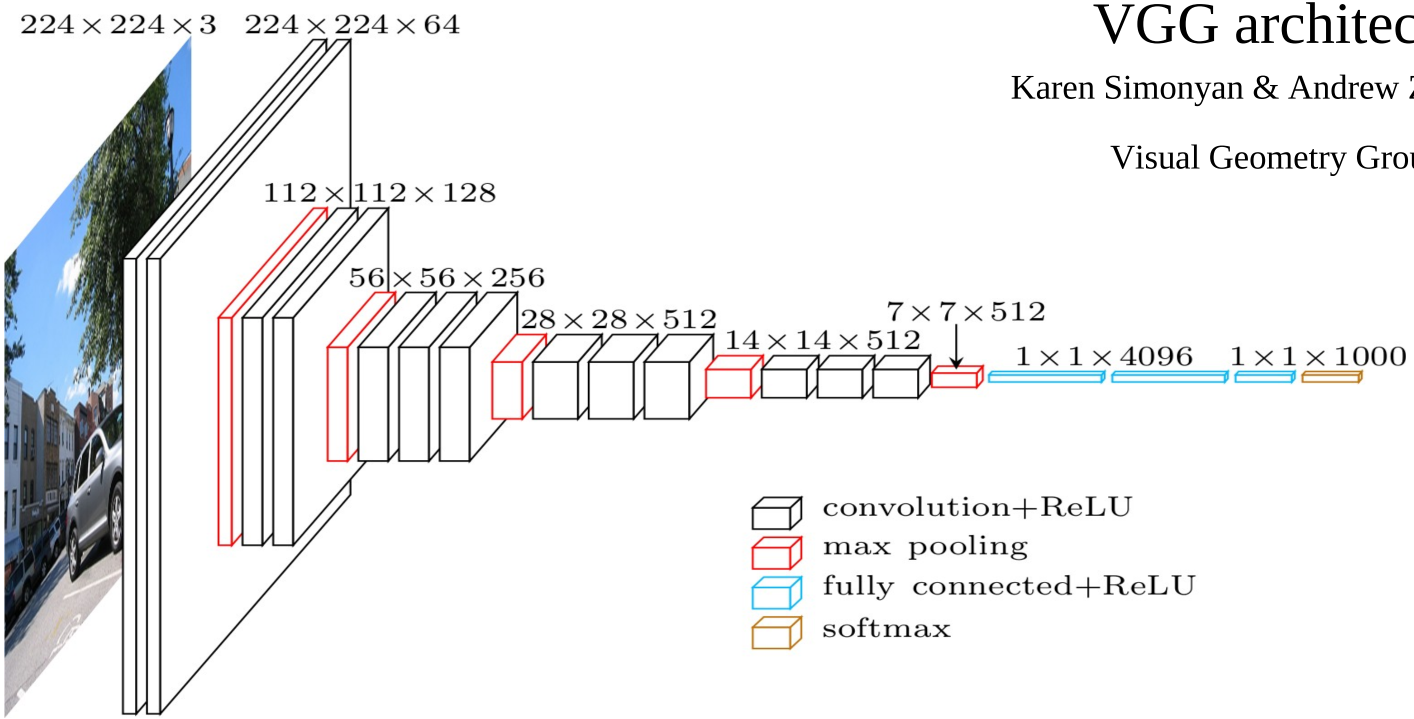
AlexNet



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

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

Object Detection with ConvNets?



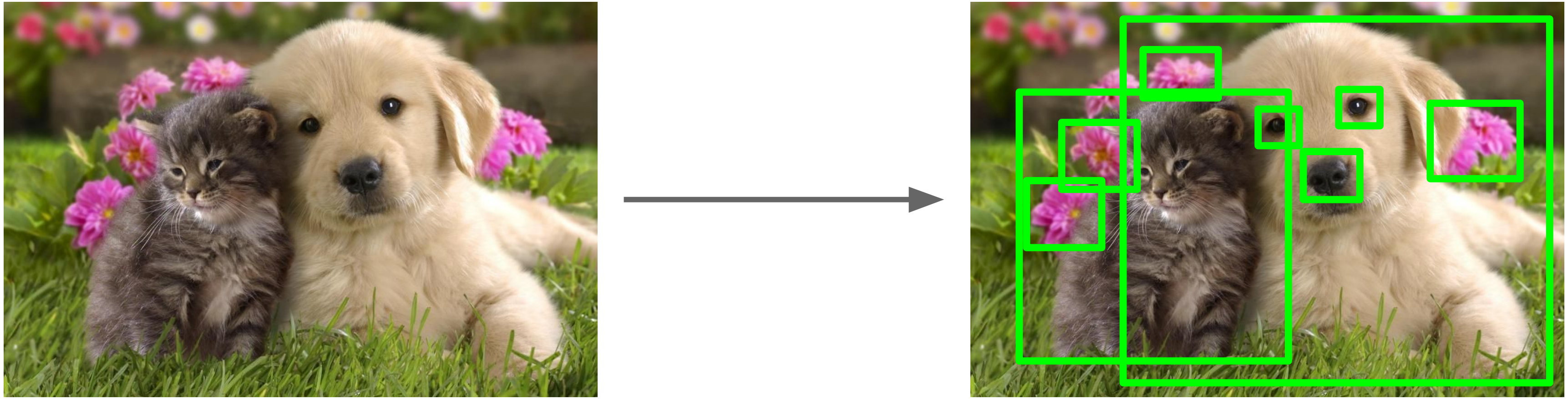
VGG architecture

Karen Simonyan & Andrew Zisserman

Visual Geometry Group,

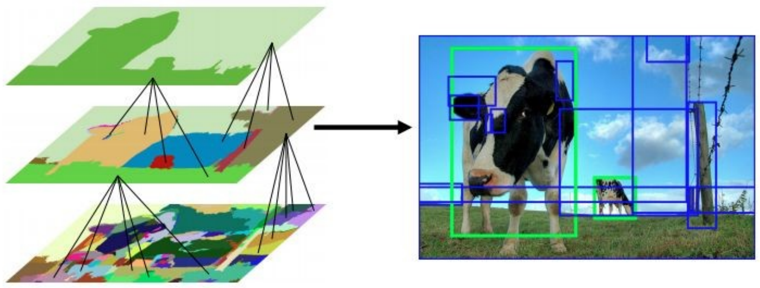
Region Proposals

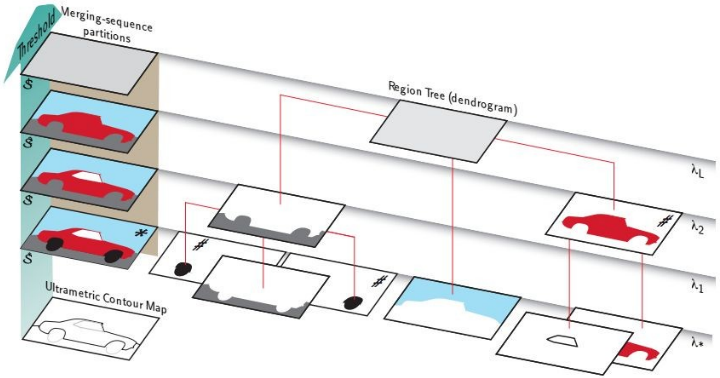
- Find “blobby” image regions that are likely to contain objects
- “Class-agnostic” object detector
- Look for “blob-like” regions



Slide Credit: CS231n

Region Proposals





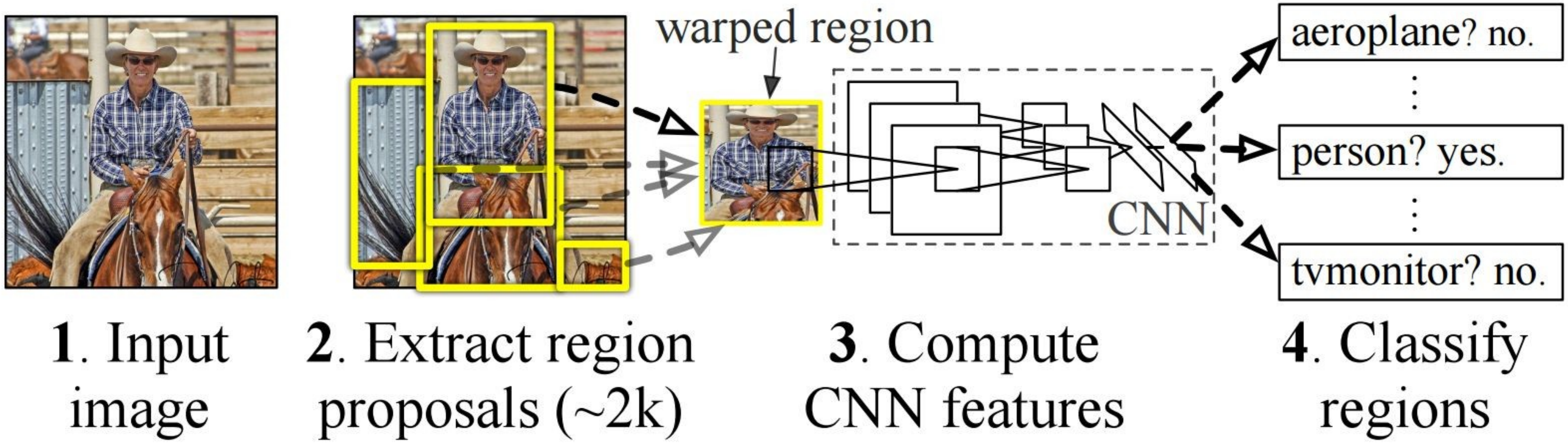
Selective Search (SS)

Multiscale Combinatorial Grouping (MCG)

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

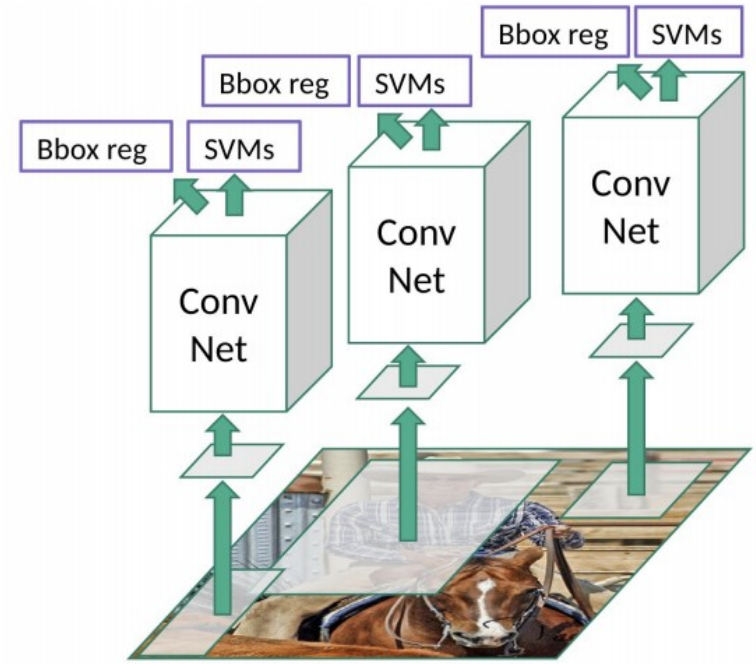
[MCG] Arbeláez, Pont-Tuset et al. Multiscale combinatorial grouping. CVPR 2014

Object Detection with Convnets: R-CNN



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

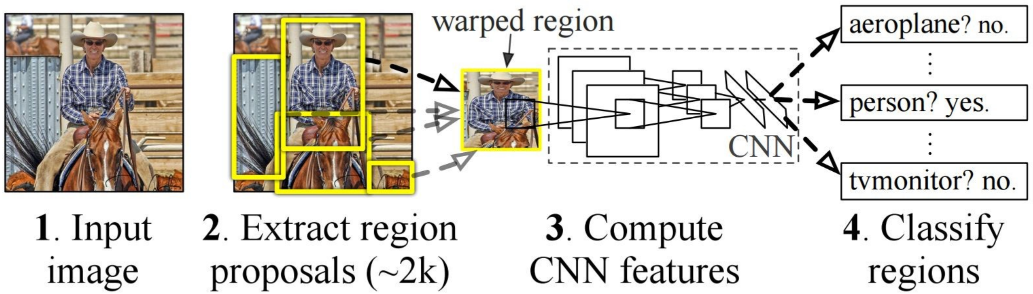
Object Detection with Convnets: R-CNN



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

R-CNN

1. Train network on proposals

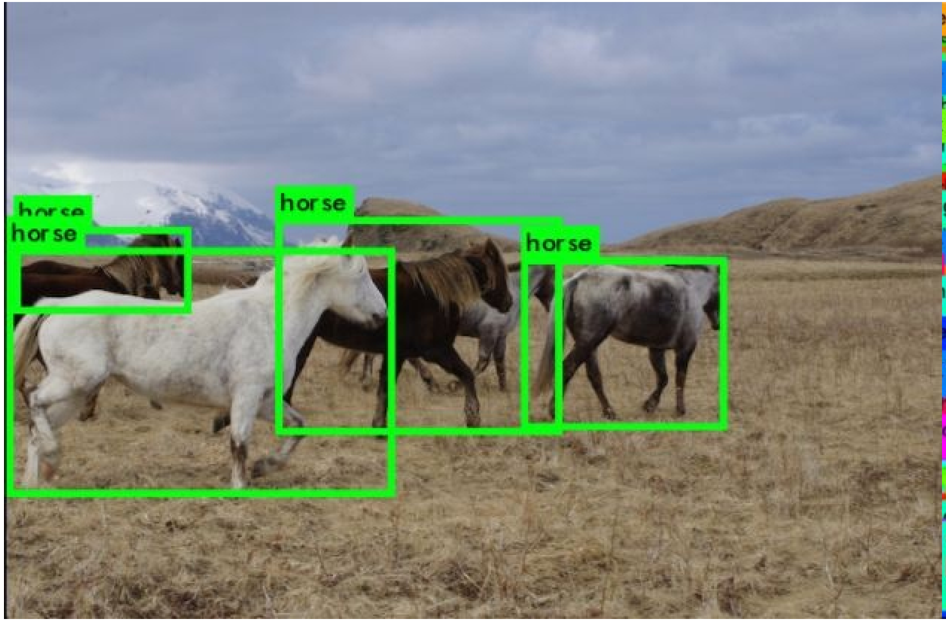


2. Post-hoc training of SVMs & Box regressors on fc7 features

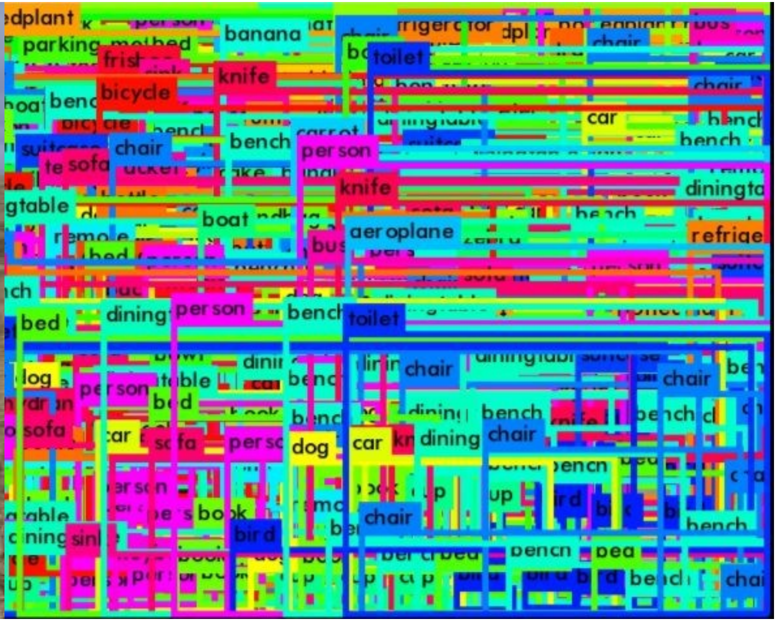
Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR 2014

R-CNN

We expect:

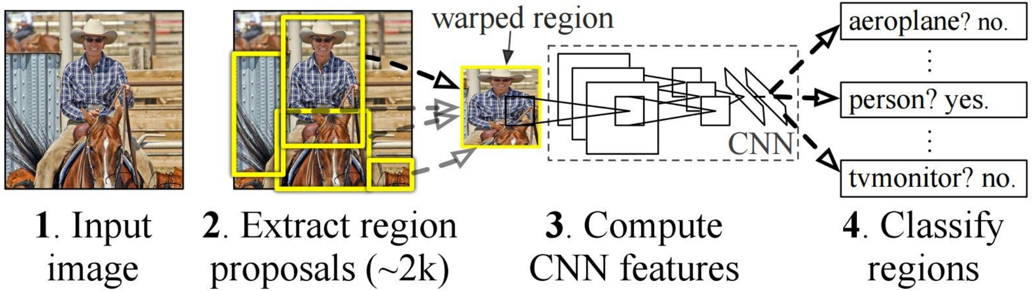


We get:



R-CNN

1. Train network on proposals

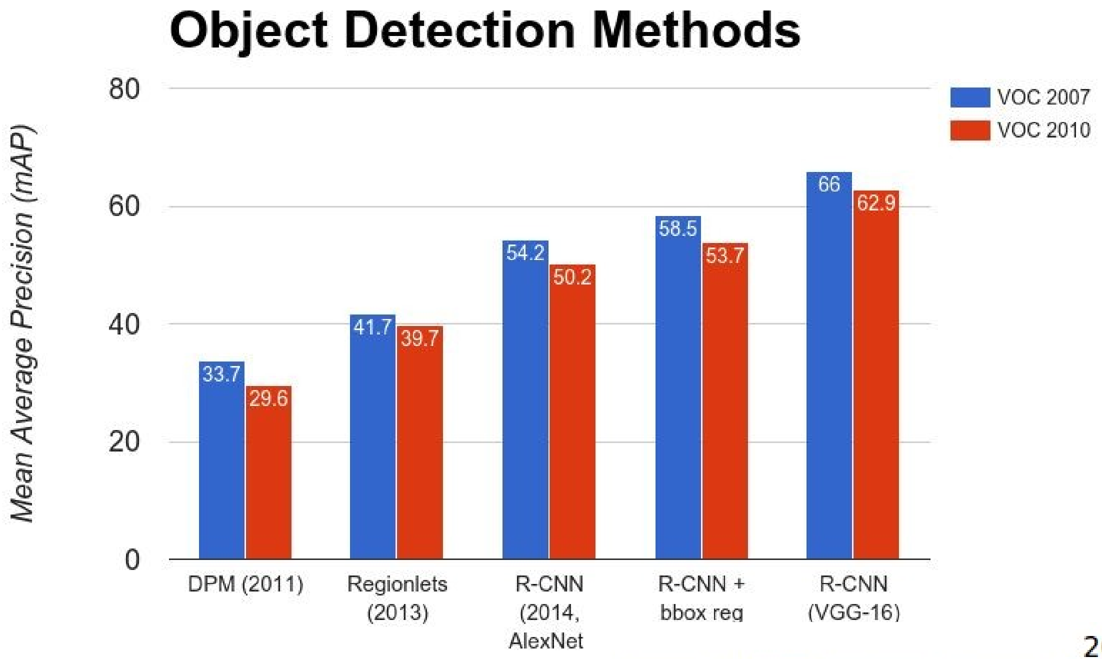


2. Post-hoc training of SVMs & Box regressors on fc7 features

3. Non Maximum Suppression + score threshold

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

R-CNN



The PASCAL Visual Object Classes (VOC) Dataset and Challenge

Mark Everingham
Luc Van Gool
Chris Williams
John Winn
Andrew Zisserman

PASCAL
Pattern Analysis, Statistical Modelling and Computational Learning

2012	20 classes. The train/val data has 11,530 images containing 27,450 ROI annotated objects and 6,929 segmentations.

20 classes

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

R-CNN: Problems

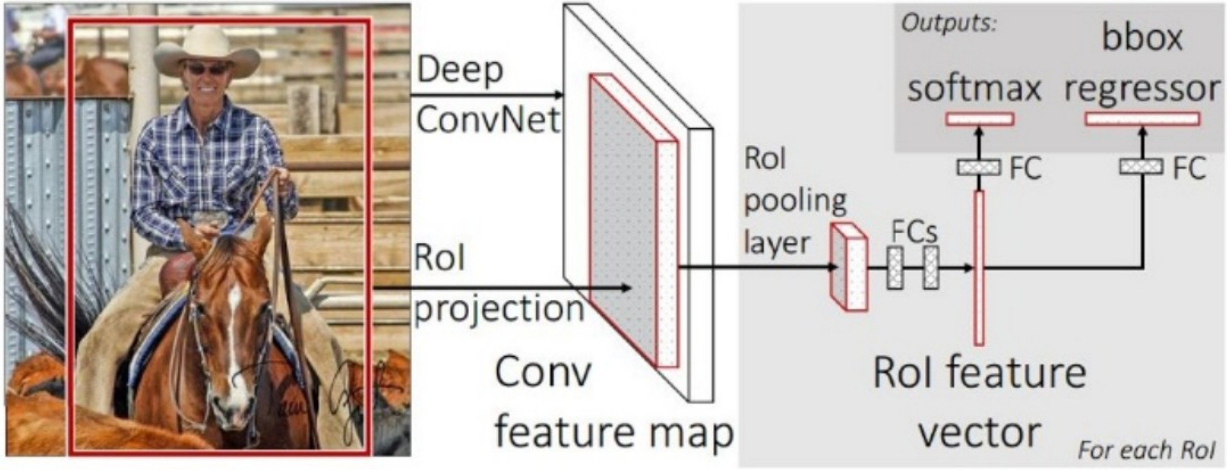
1. Slow at test-time: need to run full forward pass of CNN for each region proposal

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

3. Complex multistage training pipeline

Fast R-CNN – solution to problem 1

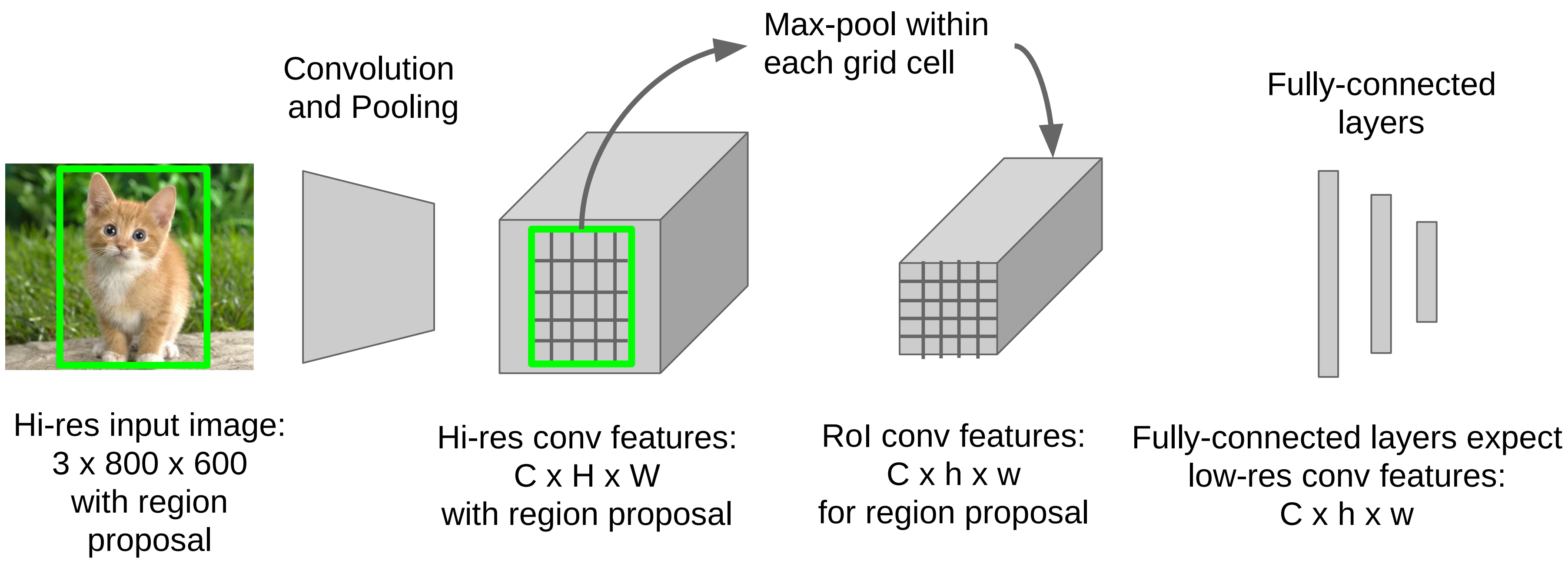
R-CNN : Slow at test-time – need to run full forward pass of CNN for each region proposal



Share computation of convolutional layers between region proposals for an image

Girshick [Fast R-CNN](). ICCV 2015

Fast R-CNN: Sharing features

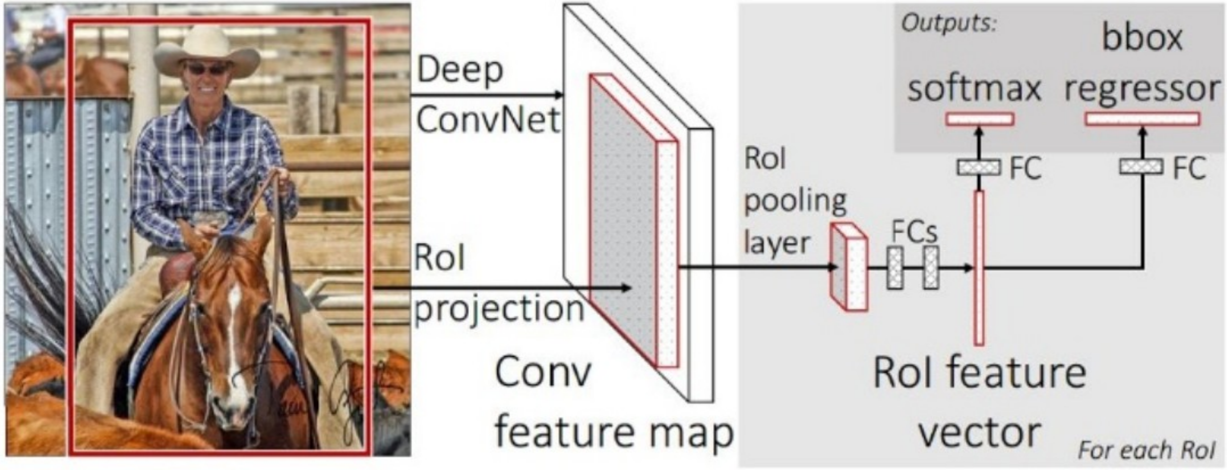


Girshick Fast R-CNN. ICCV 2015

Slide Credit: CS231n

Fast R-CNN – solution to problem 2

SVM and regressors are post-hoc and complex training



Training all 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

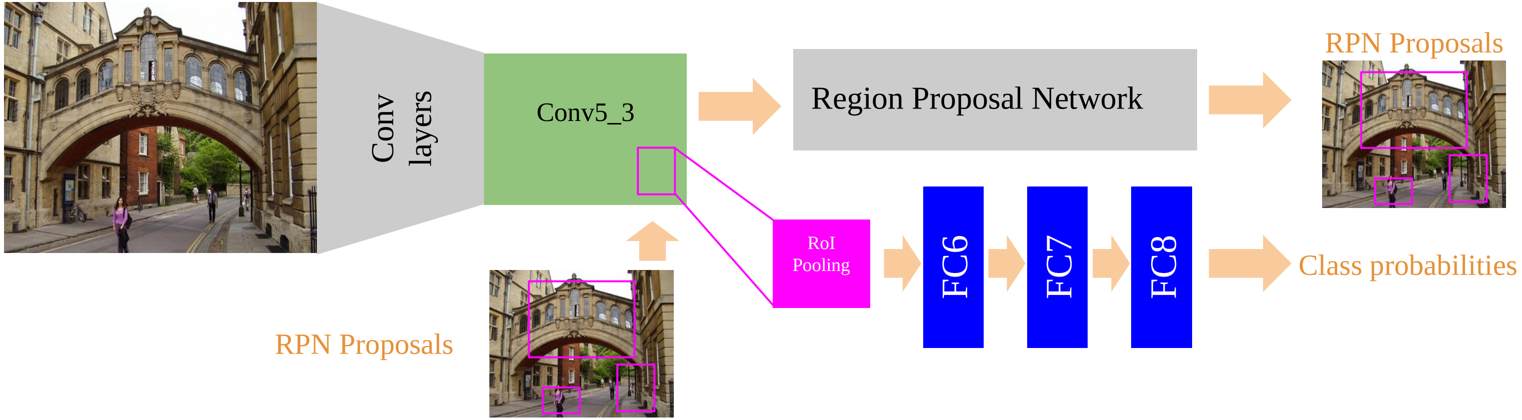
Test-time speeds don't include region proposals

	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

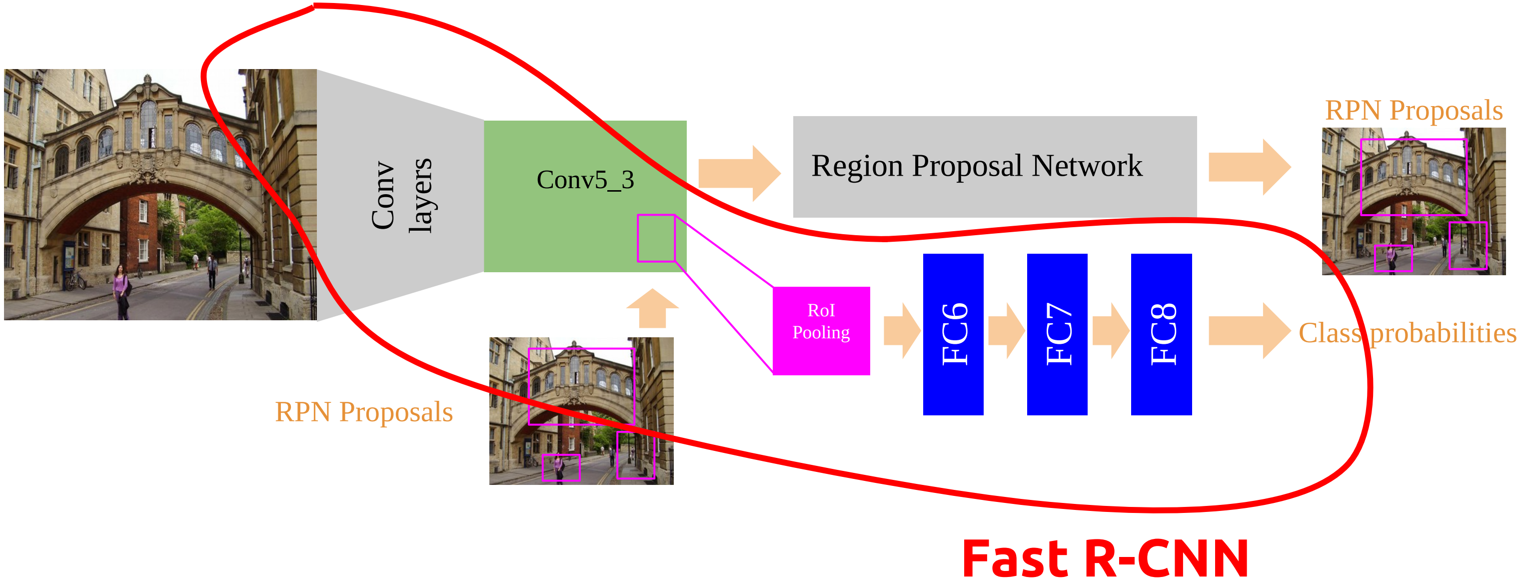
Learn proposals end-to-end sharing parameters with the classification network



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

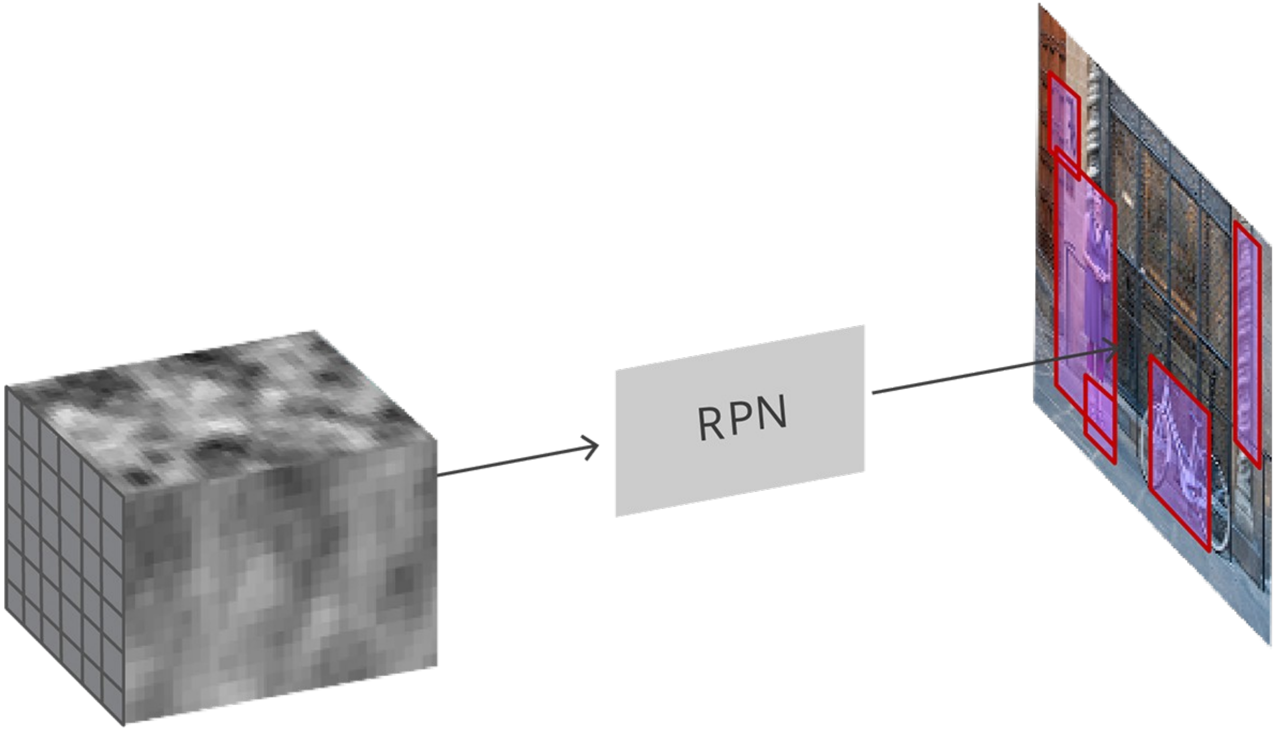
Faster R-CNN

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

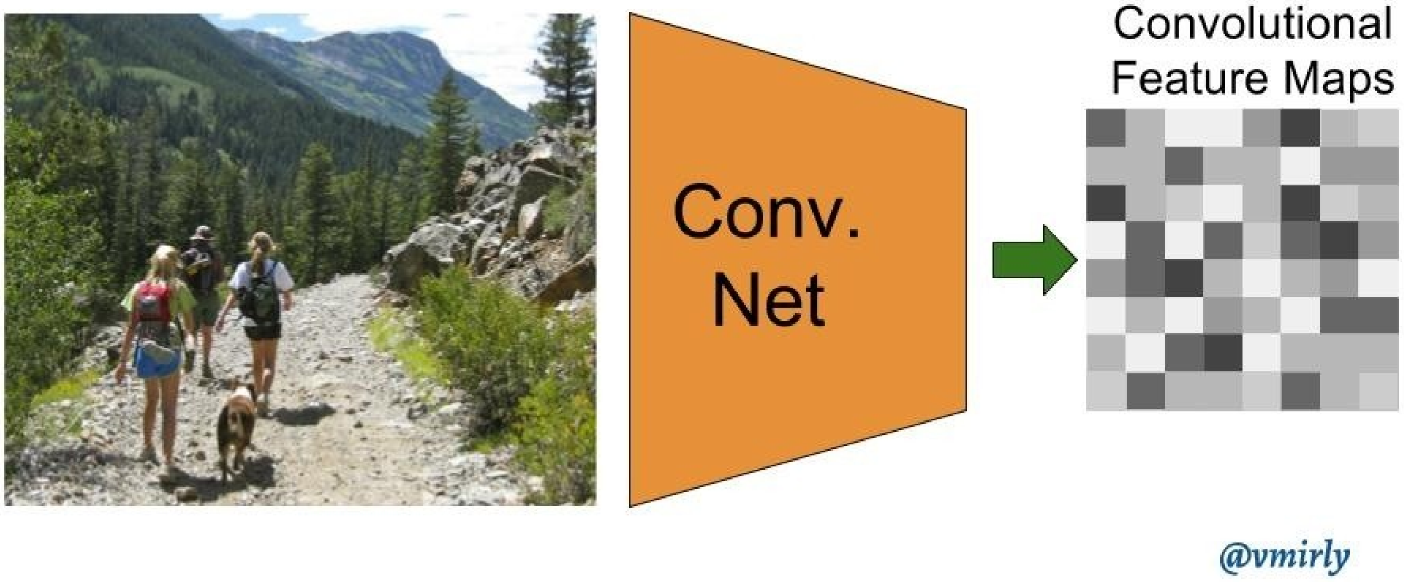
Region Proposal Network (more in details)



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

Region Proposal Network (more in details)

In the first step, the input image goes through a convolution network which will output a set of convolutional feature maps on the last convolutional layer

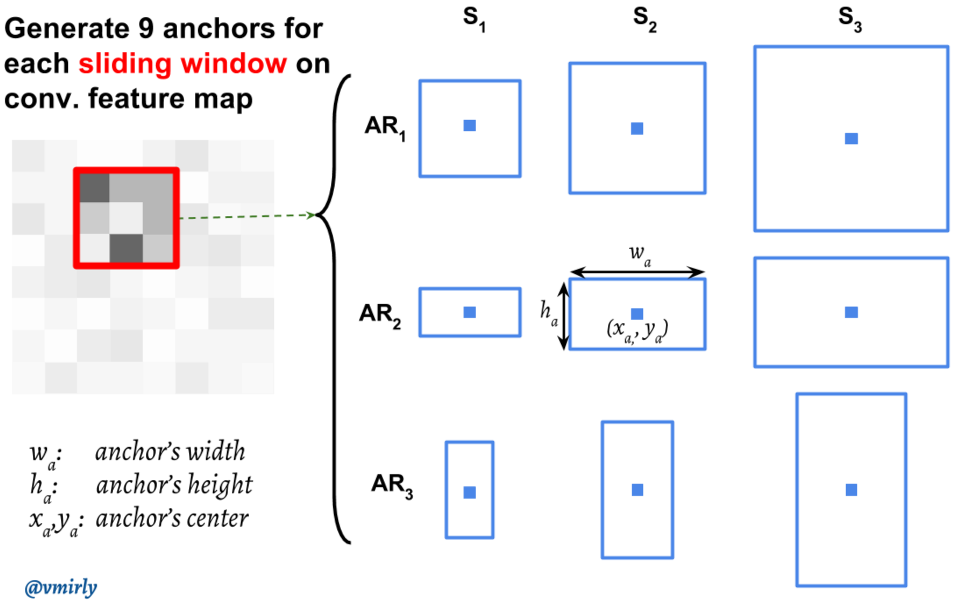


Ren et al. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS 2015

Region Proposal Network (more in details)

Then a n×n sliding window is run spatially on these feature maps and a set of 9 anchors are generated for each sliding windows (same center).

3 different aspect ratios and scales are used (in the input image).



Furthermore, for each of these anchors, a value p^* is computed which indicated how much these anchors overlap with the ground-truth bounding boxes.

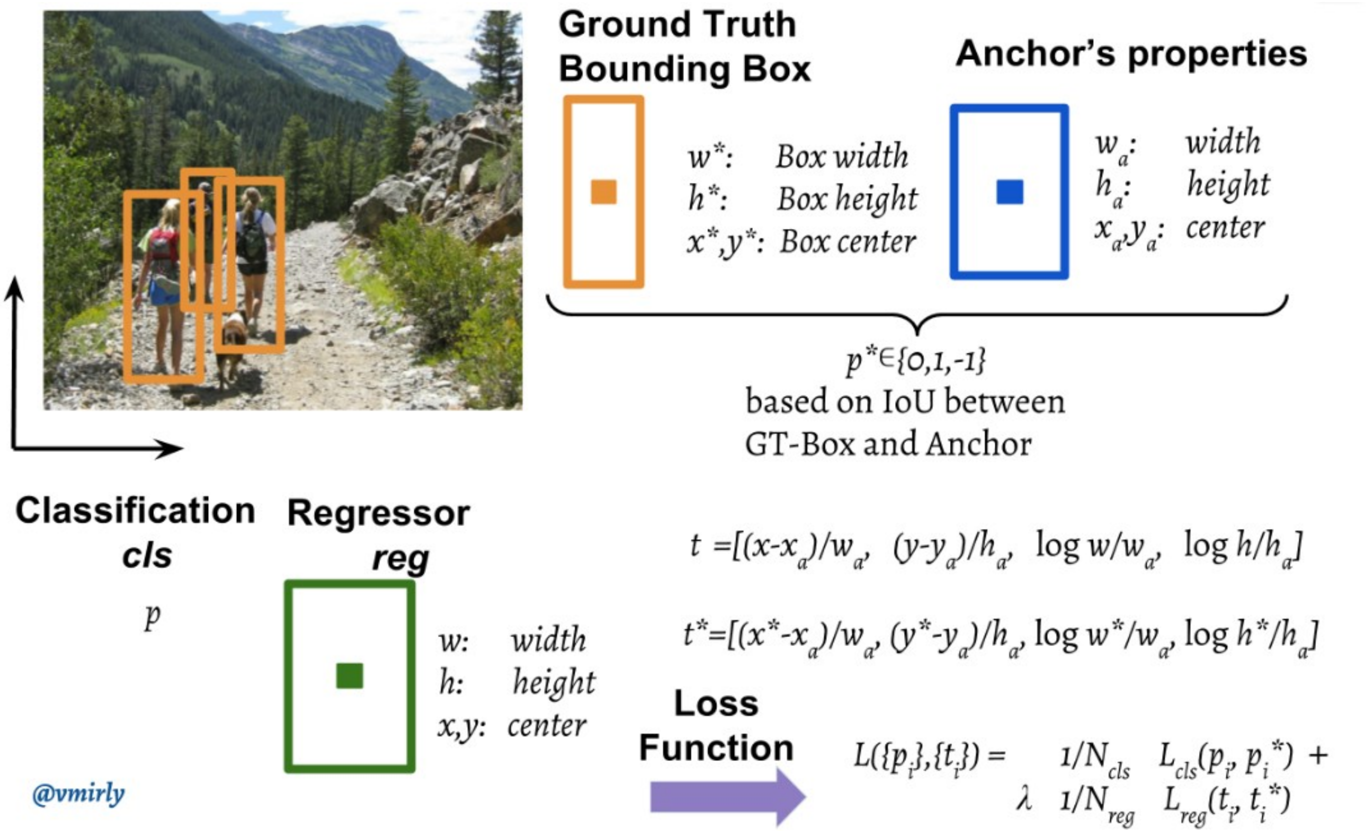
$$p^* = \begin{cases} 1 & if \quad IoU > 0.7 \\ -1 & if \quad IoU < 0.3 \\ 0 & \quad otherwise \end{cases}$$

where IoU is intersection over union and is defined below:

$$IoU = \frac{Anchor \cap GTBox}{Anchor \cup GTBox}$$

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

Region Proposal Network (more in details)

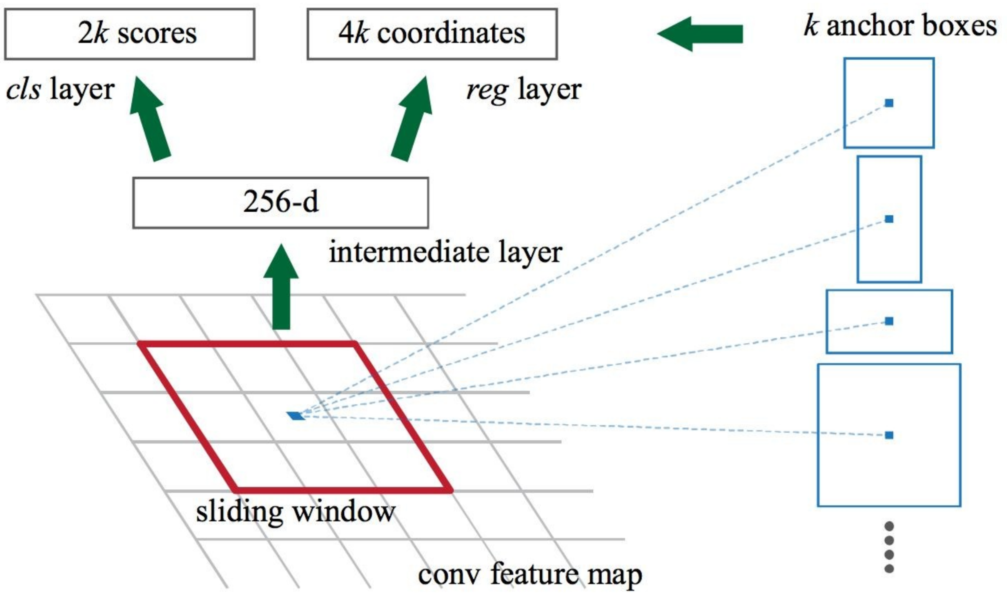


Ren et al. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS 2015

Region Proposal Network (more in details)

Bounding Box Regression

Objectness scores
(object/no object)



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

For the classification layer, we output two predictions per anchor: the score of it being background (not an object) and the score of it being foreground (an actual object).

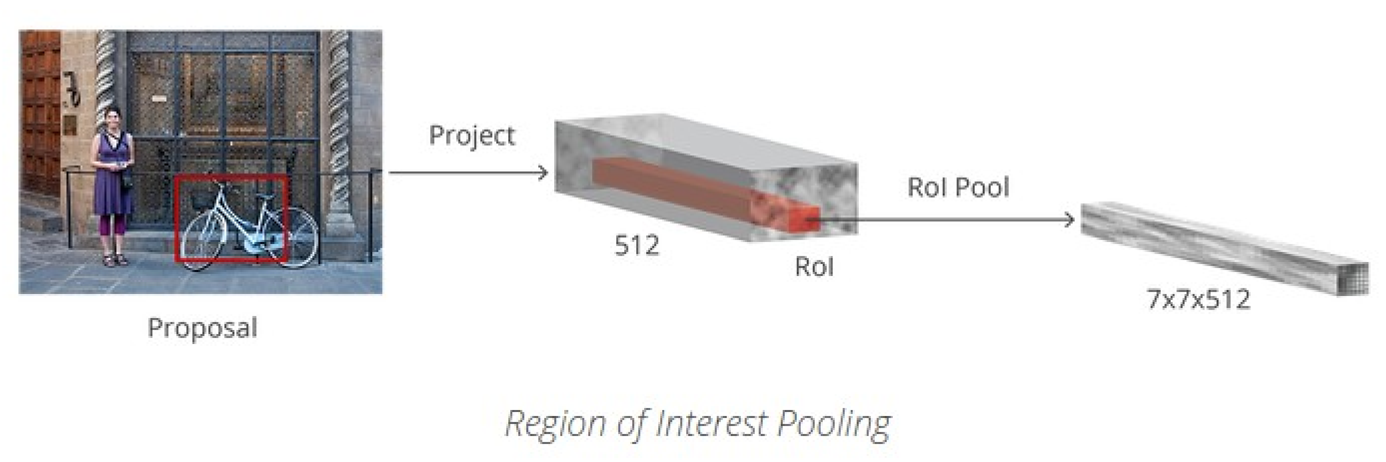
For the regression, or bounding box adjustment layer, we output 4 predictions: the deltas $\Delta_{x_{center}}, \Delta_{y_{center}}, \Delta_{width}, \Delta_{height}$ which we will apply to the anchors to get the final proposals.

Proposal selection After applying NMS, we keep the top N proposals sorted by score. In the paper $N = 2000$ is used, but it is possible to lower that number to as little as 50 and still get quite good results.

Test all proposals is slow … A solution is region of interest pooling

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

Faster R-CNN – Final steps



<u>Classification and bounding box adjusting</u>

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

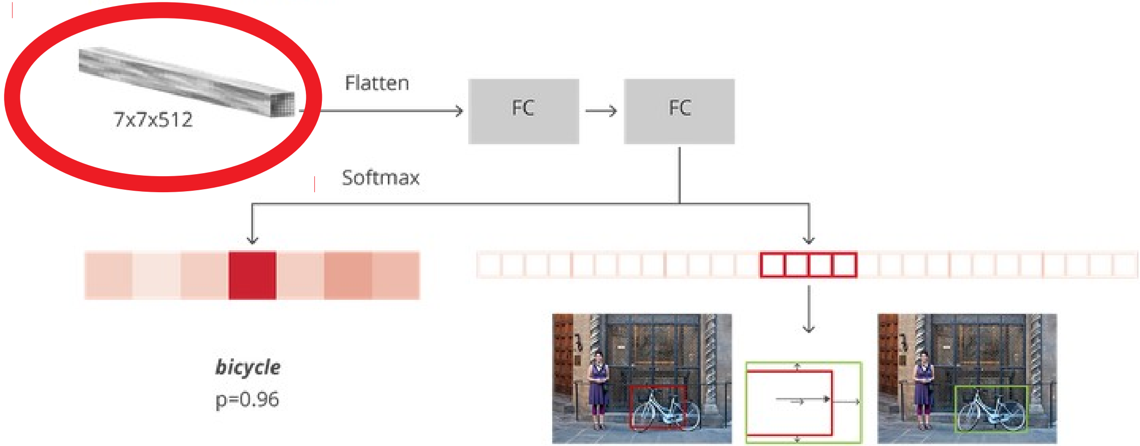
Faster R-CNN – Final steps

Classification and bounding box adjusting via classical R-CNN

Then, it uses two different fully-connected layers for each of the different objects:

- A fully-connected layer with $N + 1$ units where N is the total number of classes and that extra one is for the background class.
- A fully-connected layer with $4N$ units. We want to have a regression prediction, thus we need Δ_{center_x}, Δ_{center_y}, Δ_{width}, Δ_{height} for each of the N possible classes.

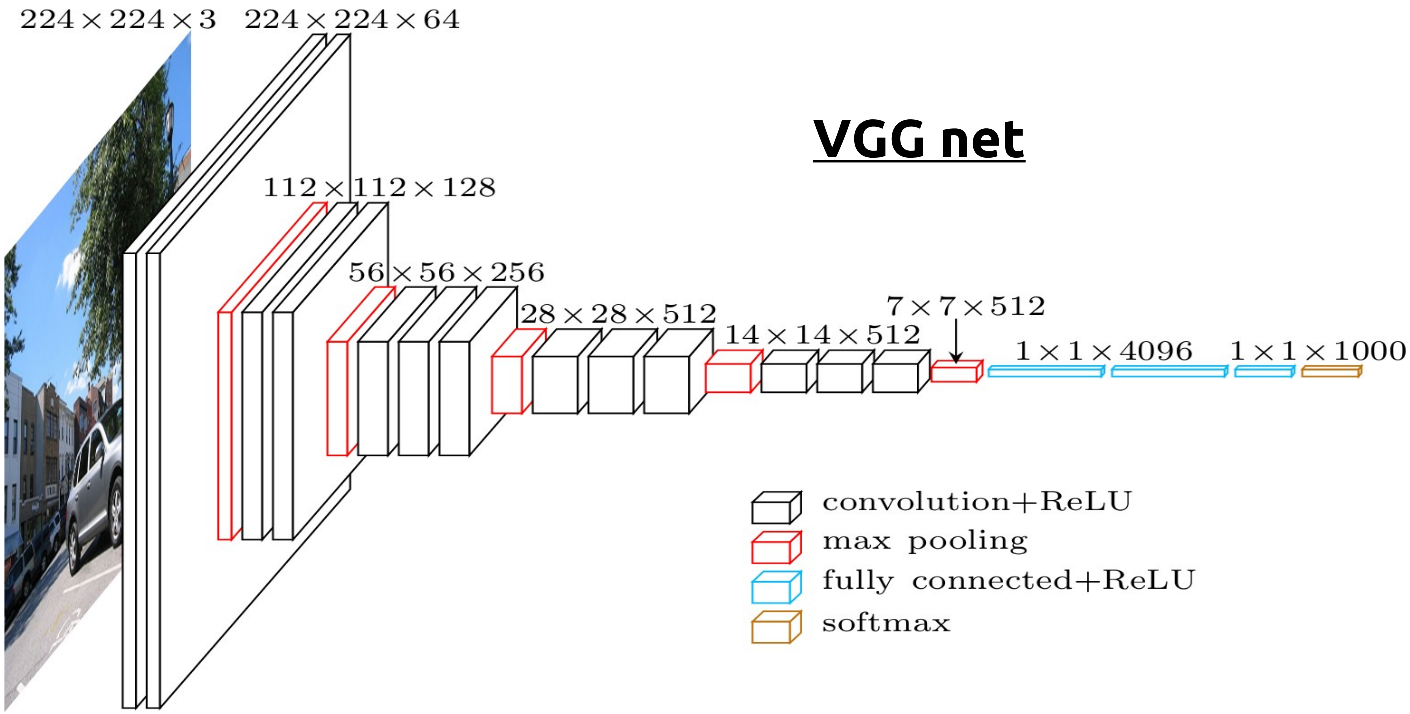
VGG net



R-CNN architecture

Faster R-CNN – Final steps

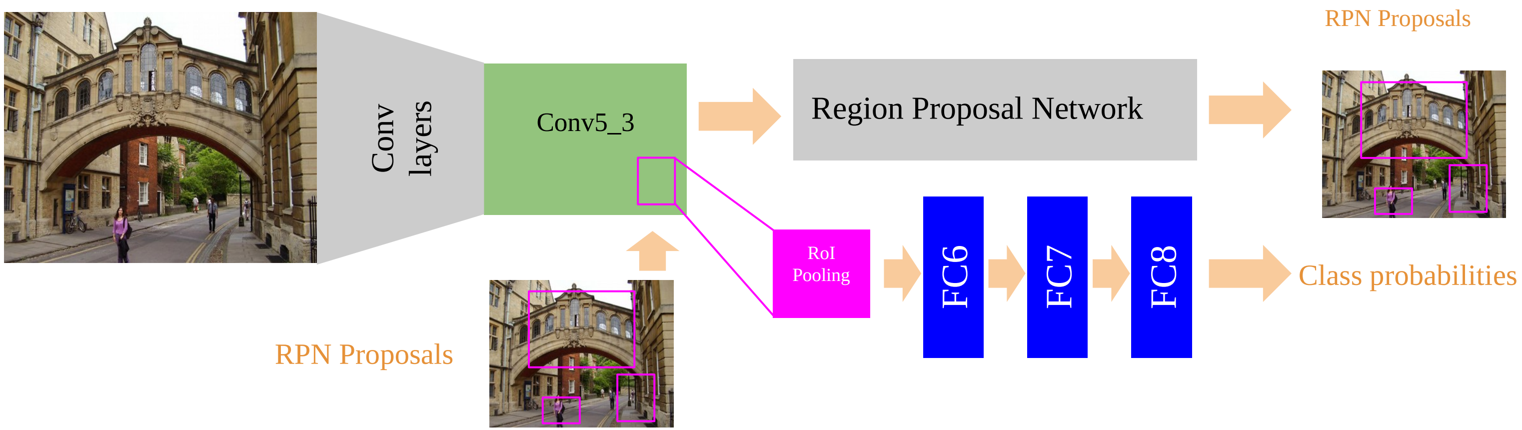
Classification and bounding box adjusting via classical R-CNN



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 (spoiler D3L6)



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

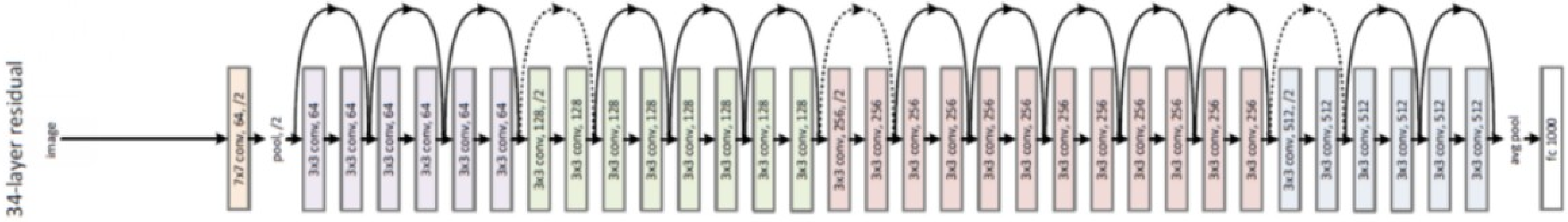
Faster R-CNN

	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**

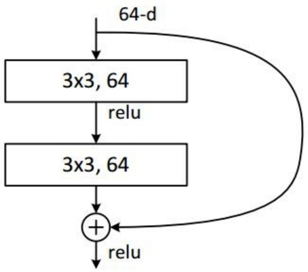
Ren et al. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS 2015

Faster R-CNN

Faster R-CNN (based on ResNet-101) is the winners of COCO challenge and ILSVRC 2015 and 2016



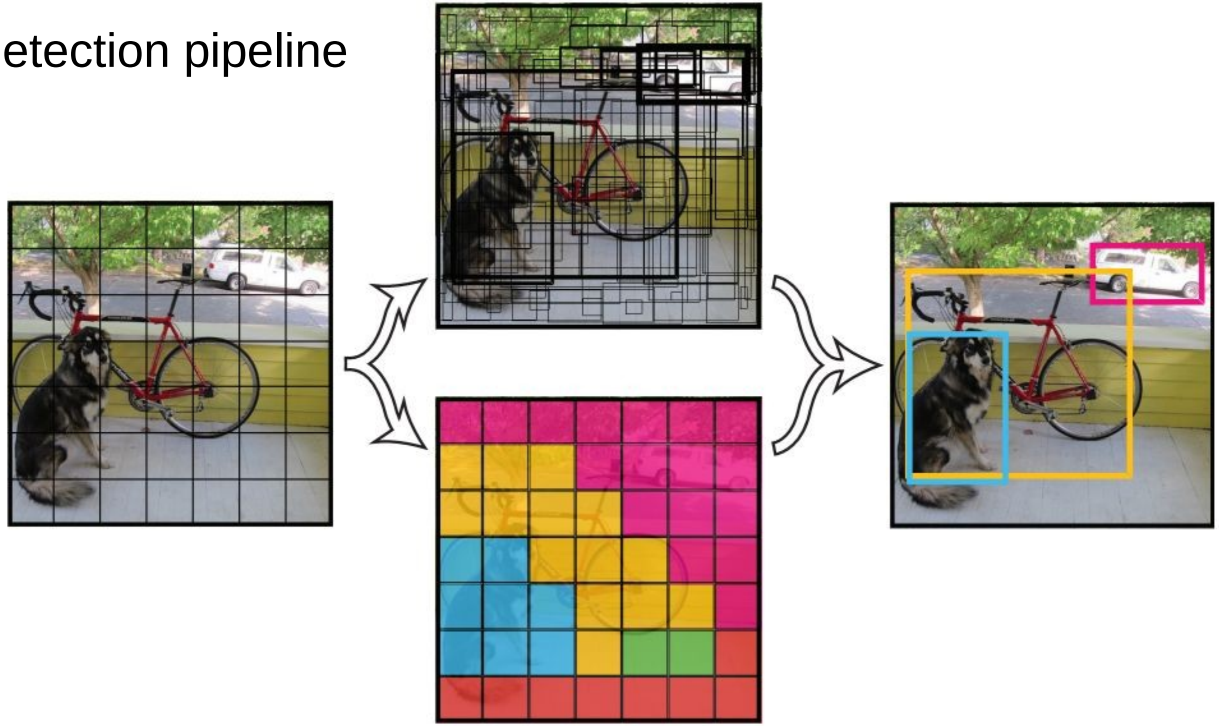
ResNet-xx



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

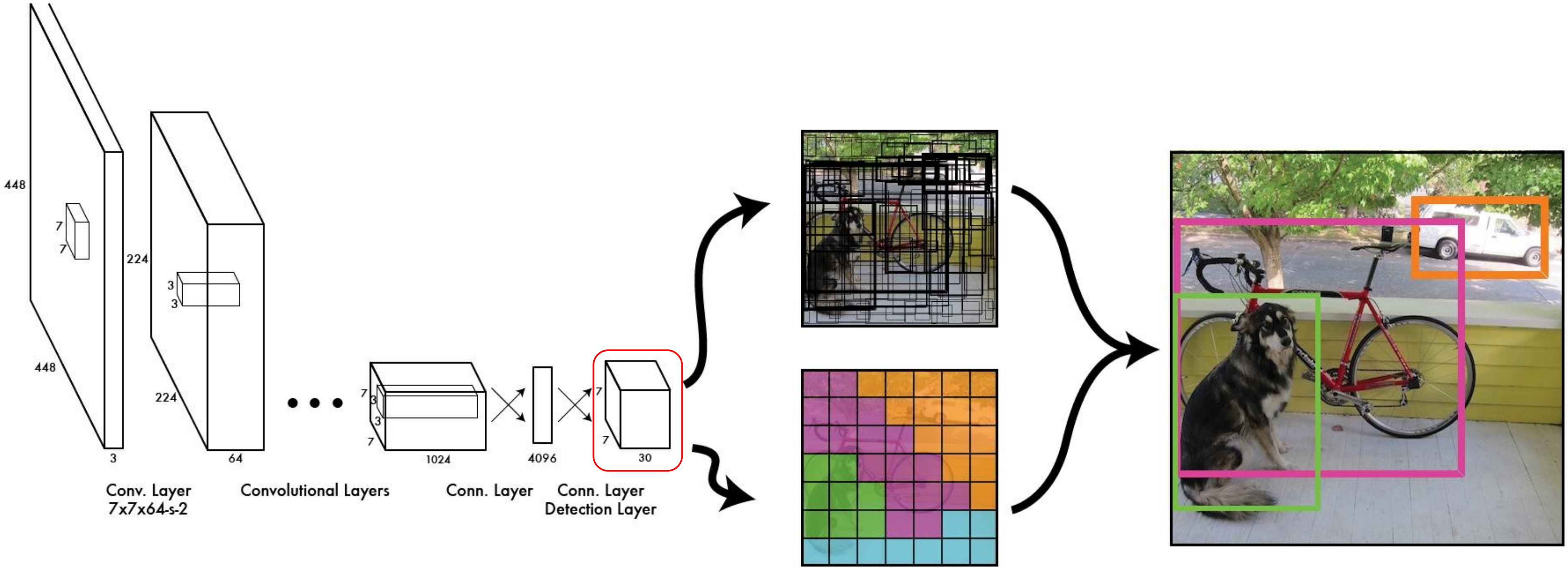
YOLO: You Only Look Once

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

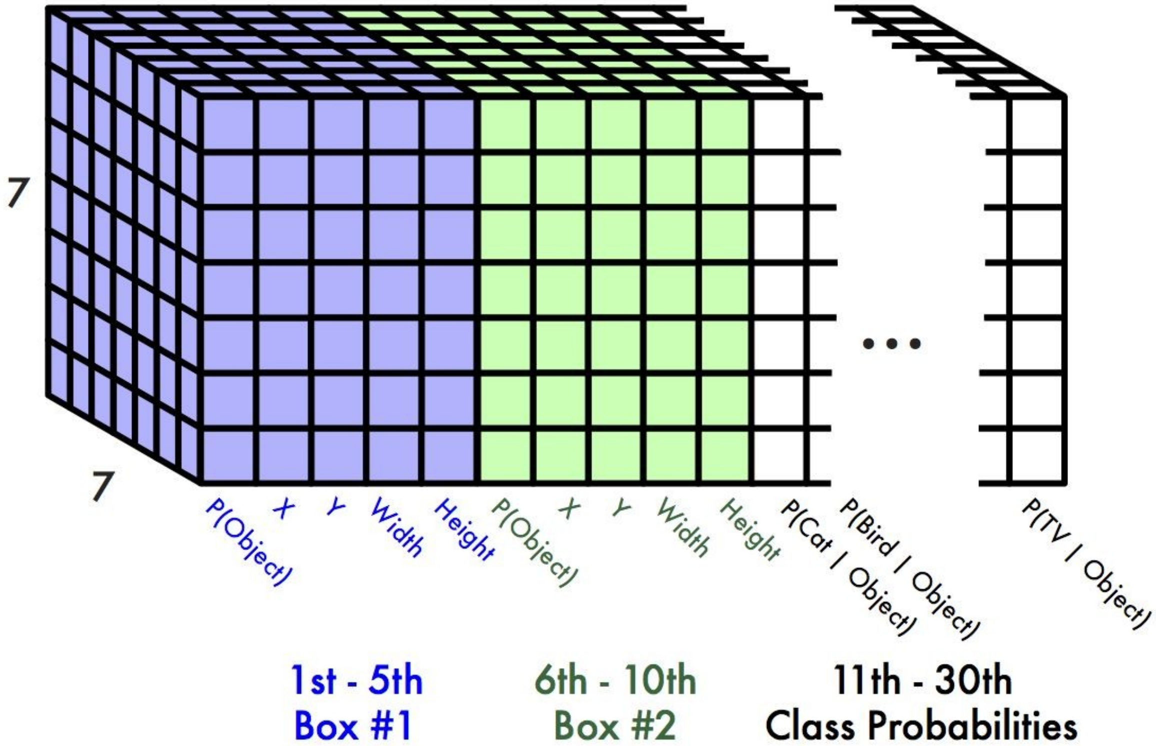
YOLO: You Only Look Once

Each cell predicts:

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

For Pascal VOC:

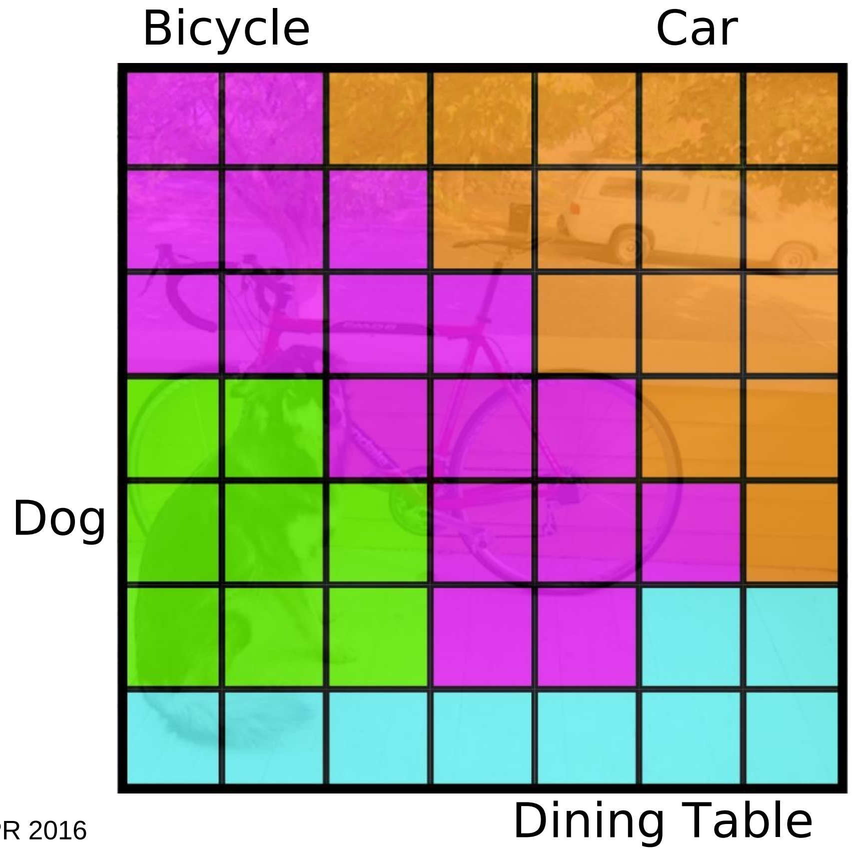
- 7x7 grid
- 2 bounding boxes / cell
- 20 classes



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

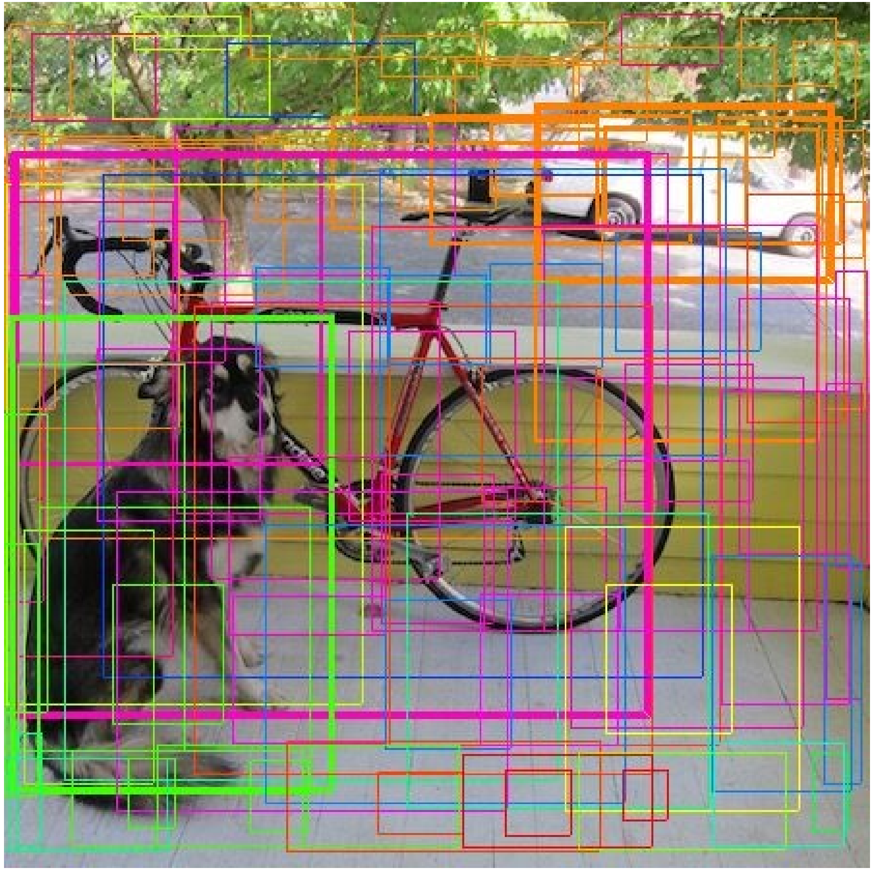
YOLO: You Only Look Once

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

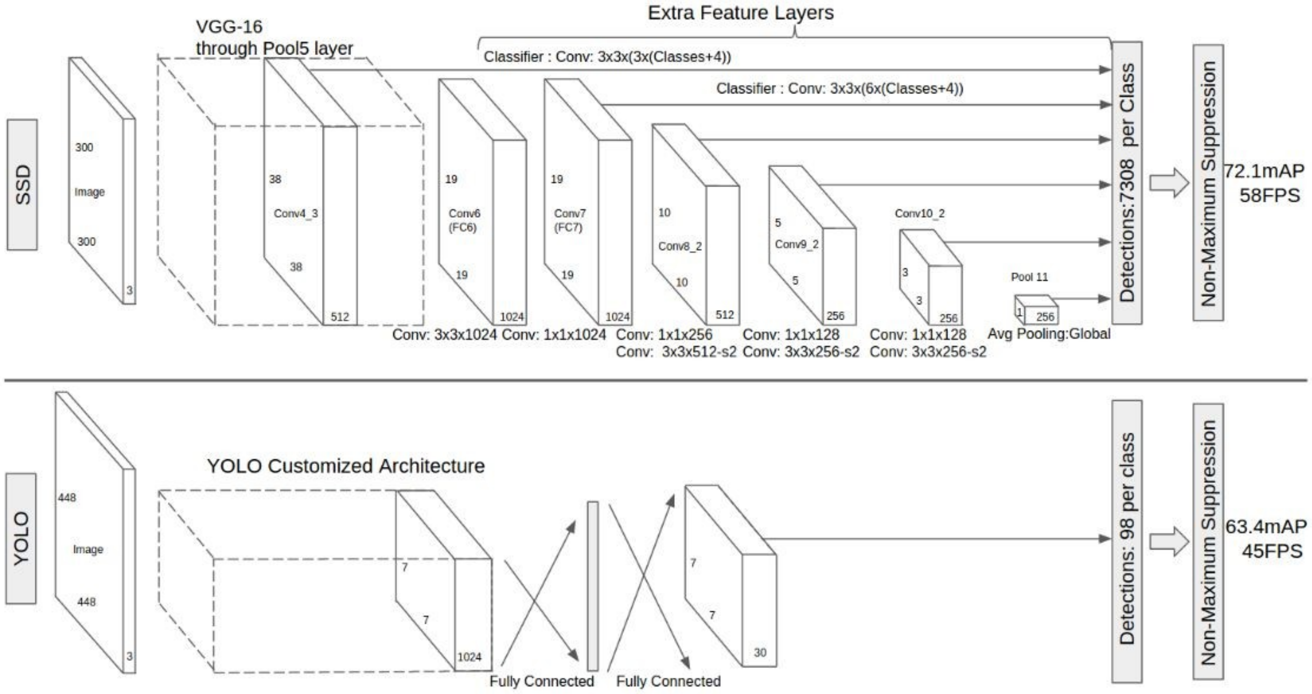


+ NMS
+ Score threshold

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

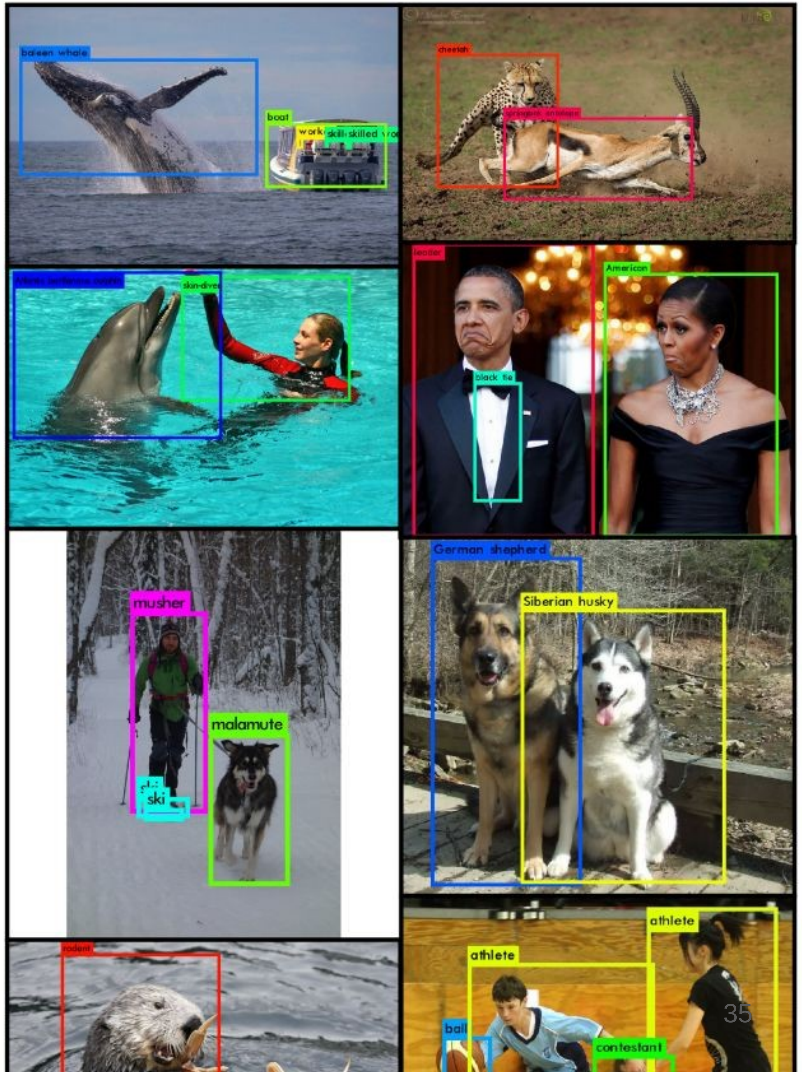


Liu et al. SSD: Single Shot MultiBox Detector, ECCV 2016

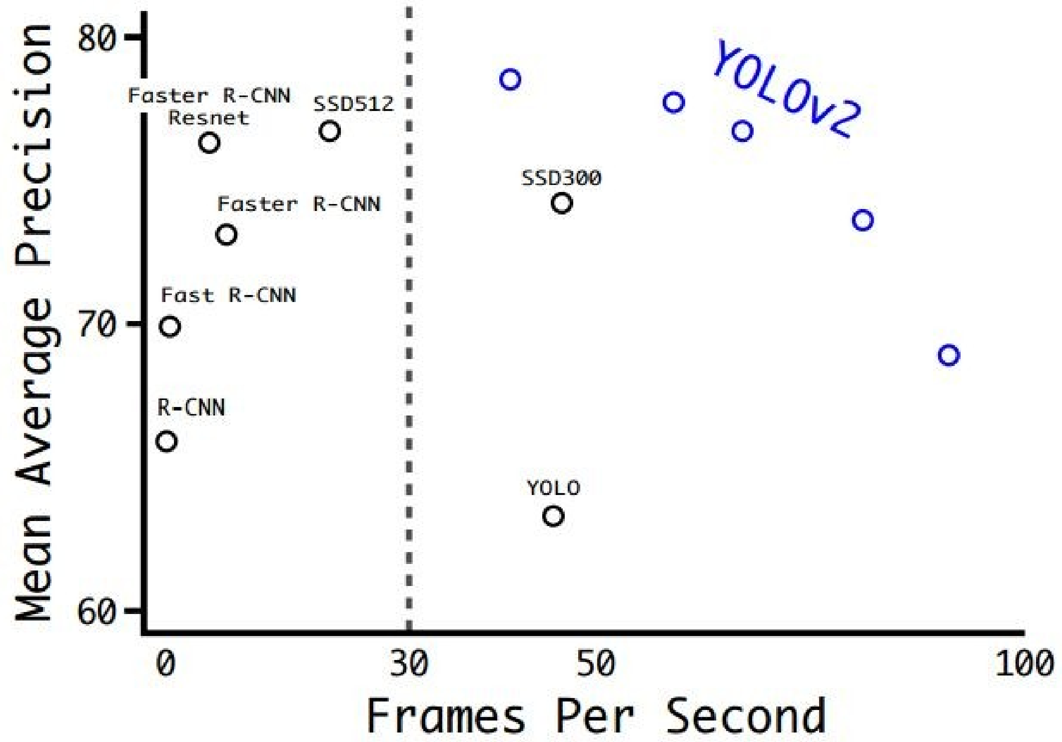
YOLOv2

	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

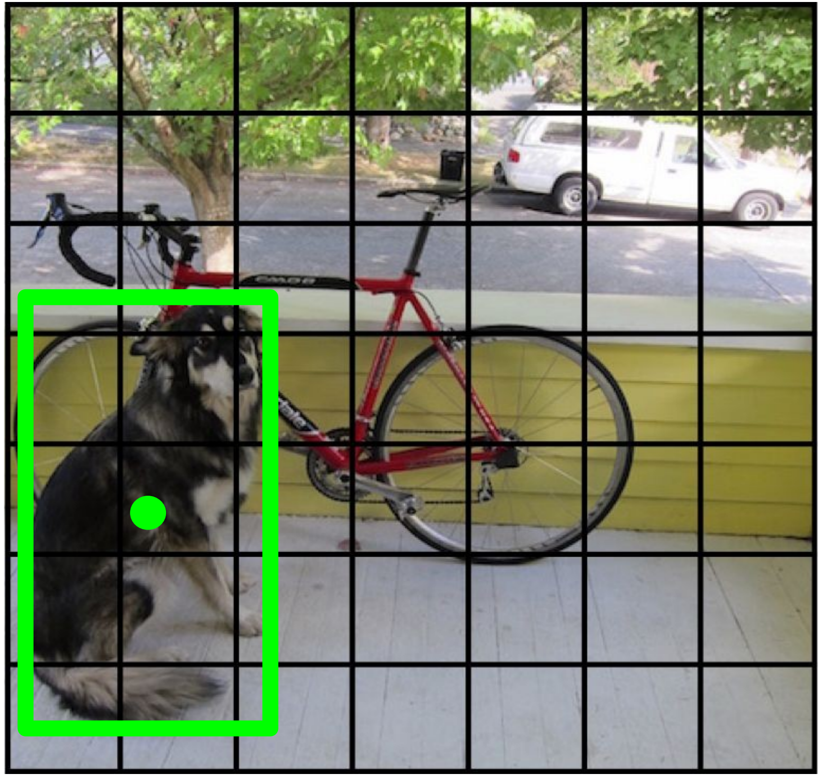
- R-CNN
- Fast R-CNN
- Faster R-CNN
- SPPnet
- R-FCN

Proposal-free methods

- YOLO, YOLOv2
- SSD

Questions?

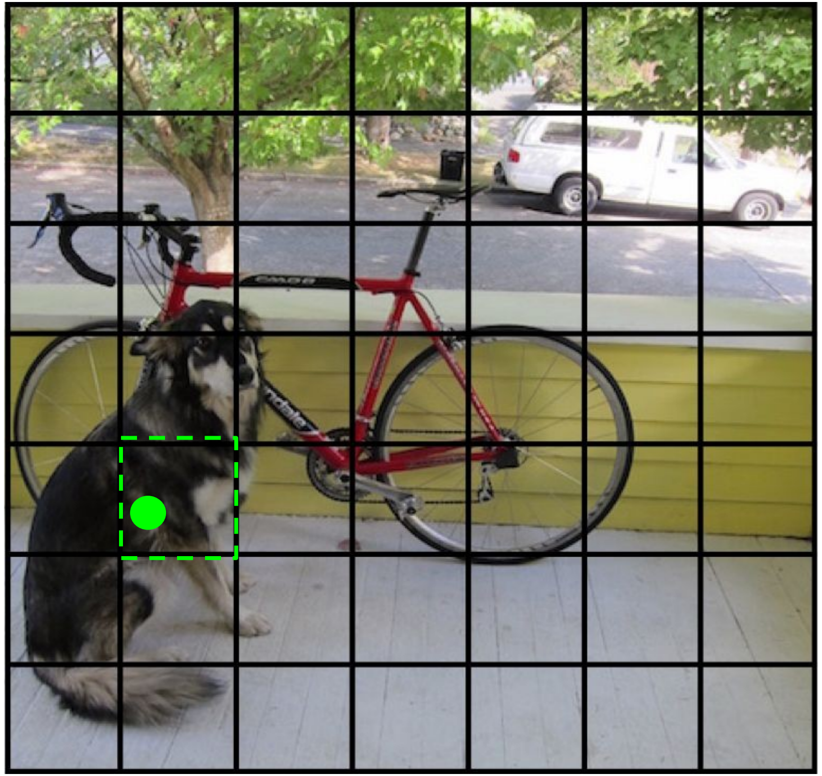
YOLO: Training



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

Slide credit: [YOLO Presentation @ CVPR 2016]

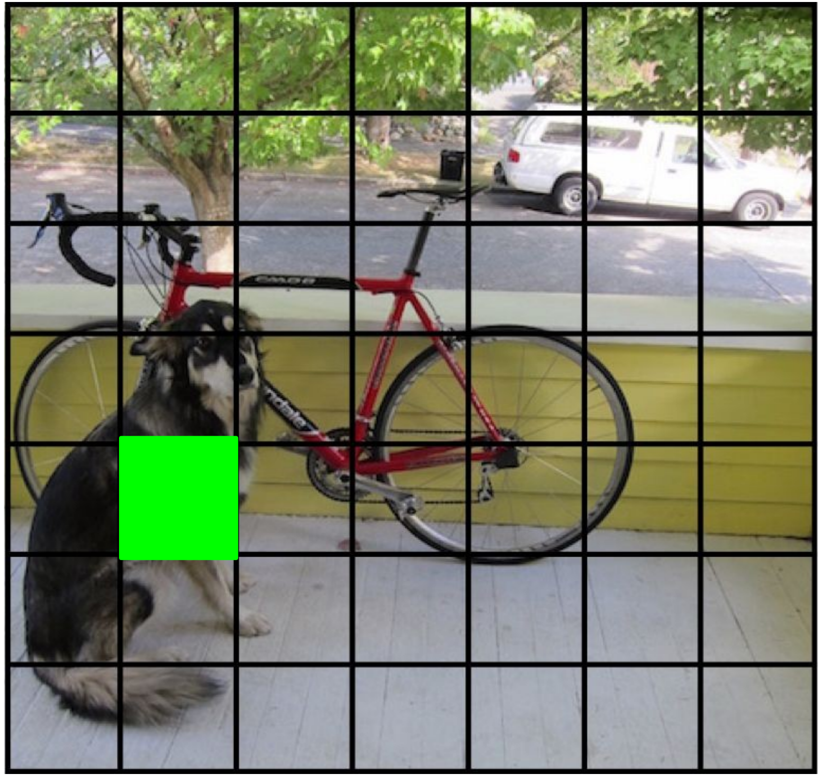
YOLO: Training



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

Slide credit: [YOLO Presentation @ CVPR 2016]

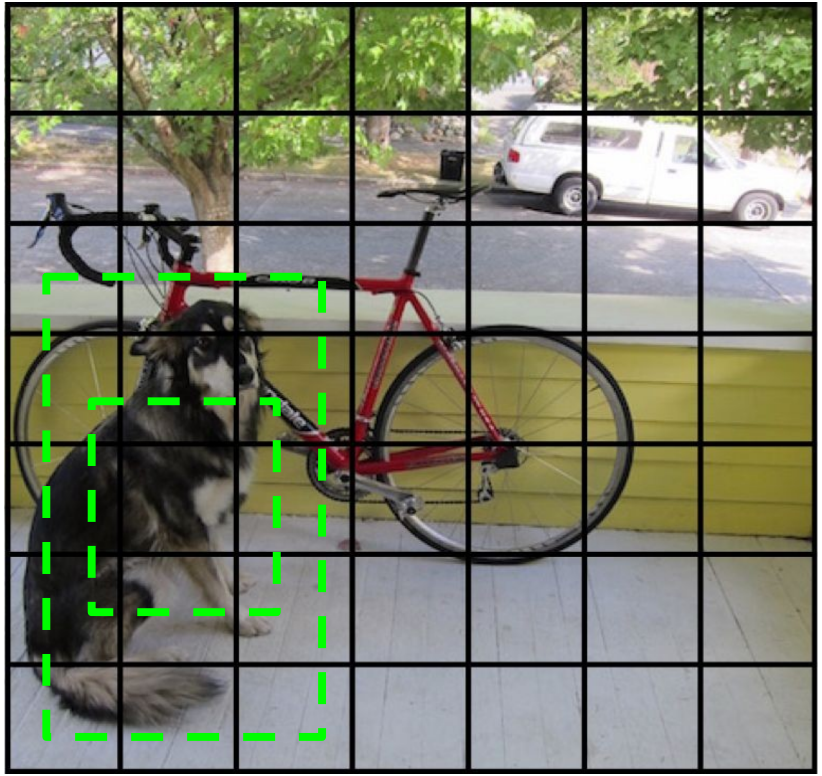
YOLO: Training



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

Slide credit: [YOLO Presentation @ CVPR 2016]

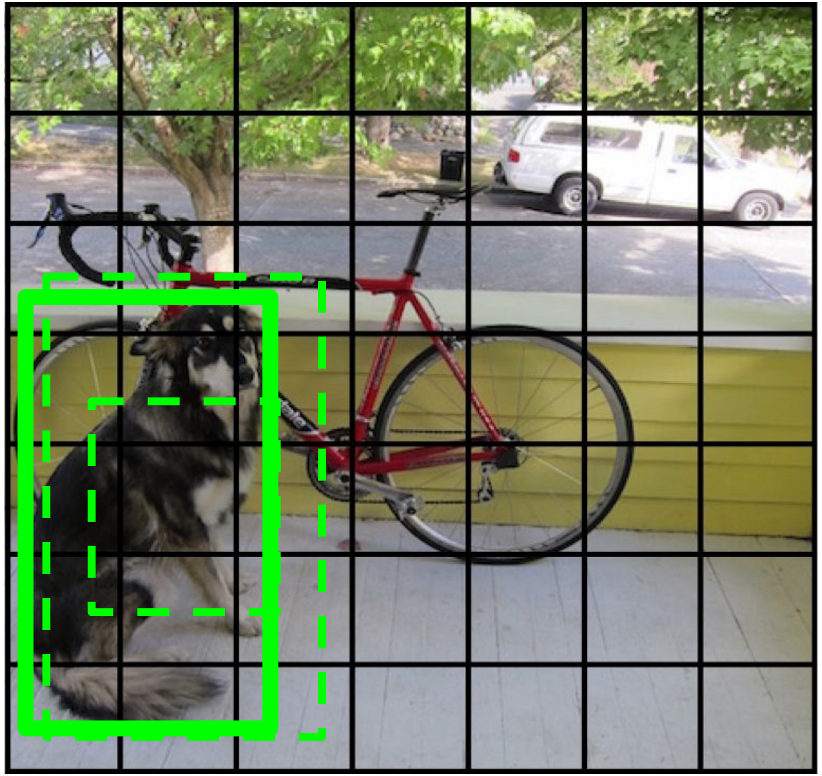
YOLO: Training



Predicted boxes for this cell

Slide credit: [YOLO Presentation @ CVPR 2016]

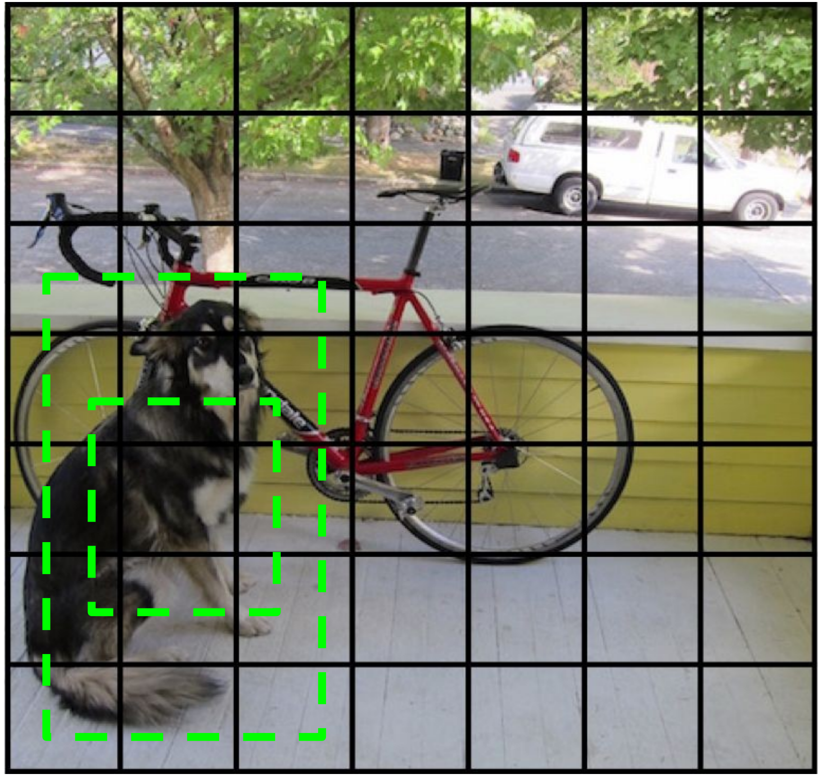
YOLO: Training



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]

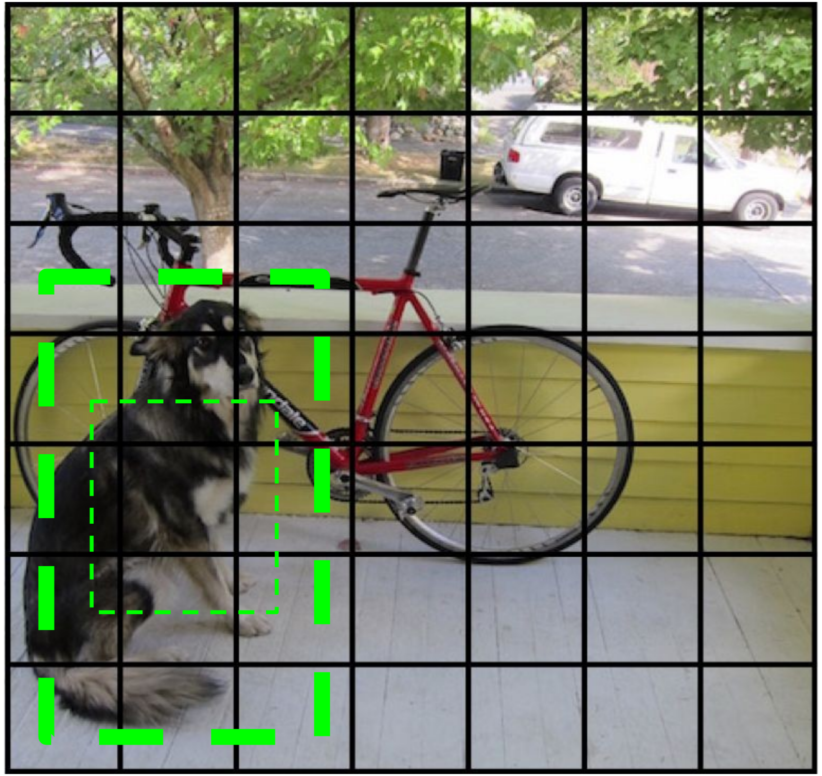
YOLO: Training



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

Slide credit: [YOLO Presentation @ CVPR 2016]

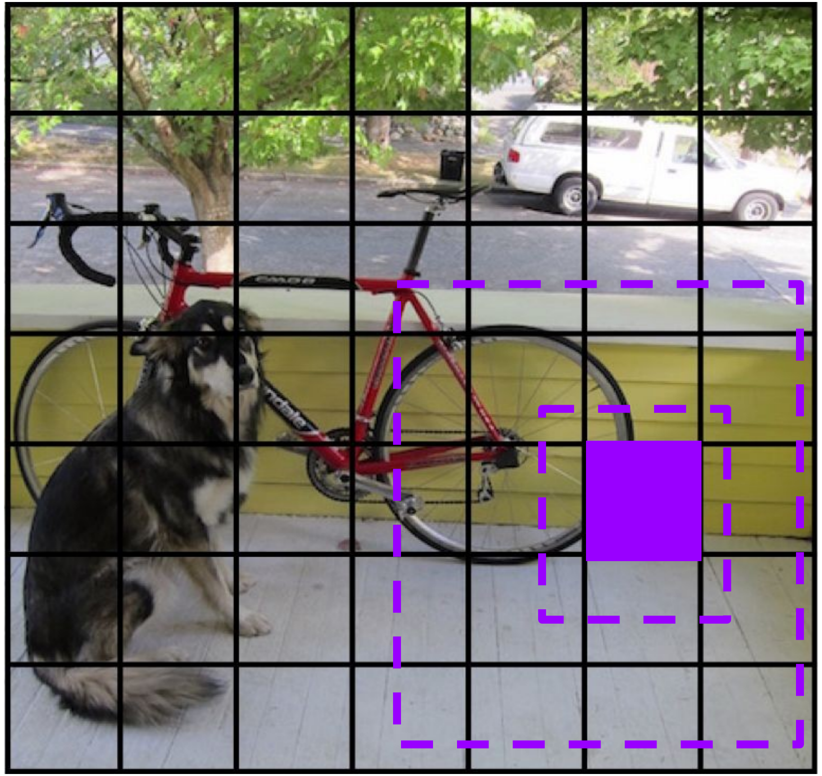
YOLO: Training



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

Slide credit: [YOLO Presentation @ CVPR 2016]

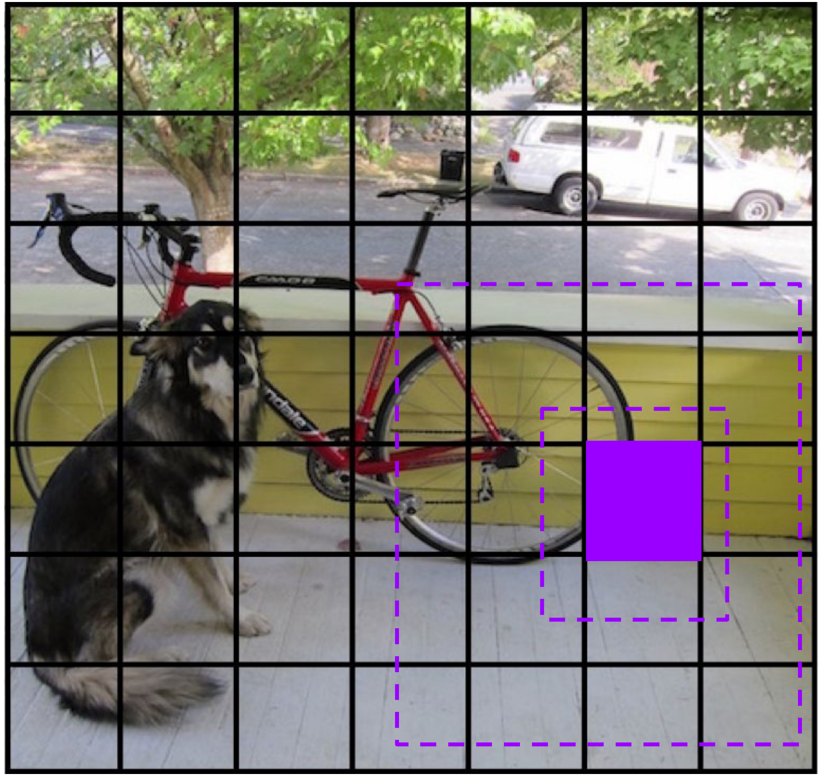
YOLO: Training



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

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

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

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]()