YOLO:
You Only Look Once
Unified Real-Time Object Detection

Tanvir Azhar
Lecturer, East Delta University

Outline

1. Review: R-CNN

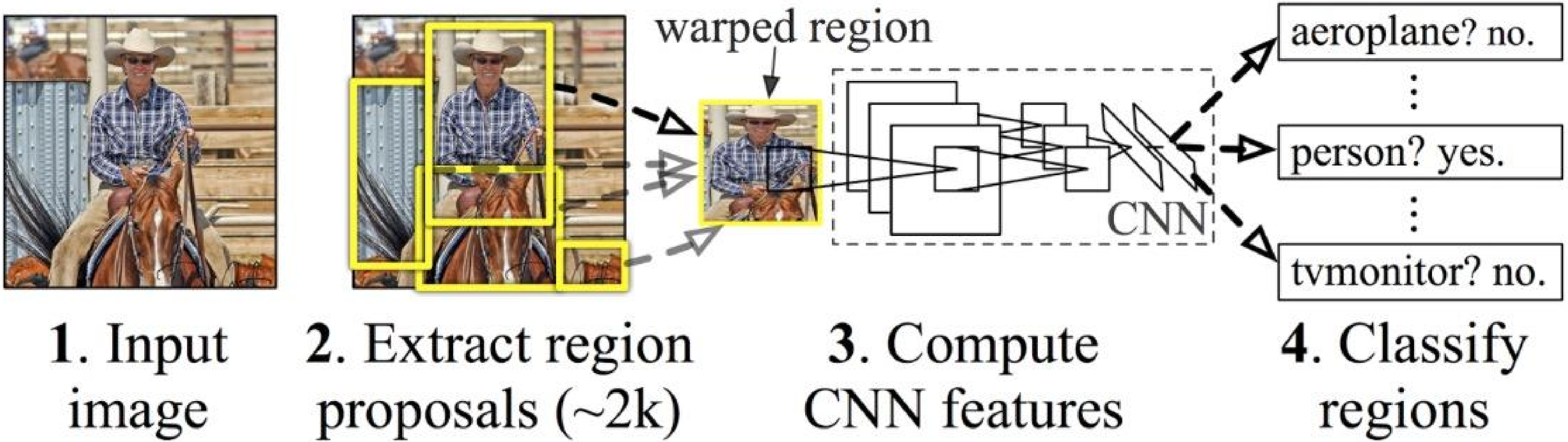
2. YOLO: -- Detection Procedure

   -- Network Design

   -- Training Part

   -- Experiments

R-CNN: *Regions with CNN features*

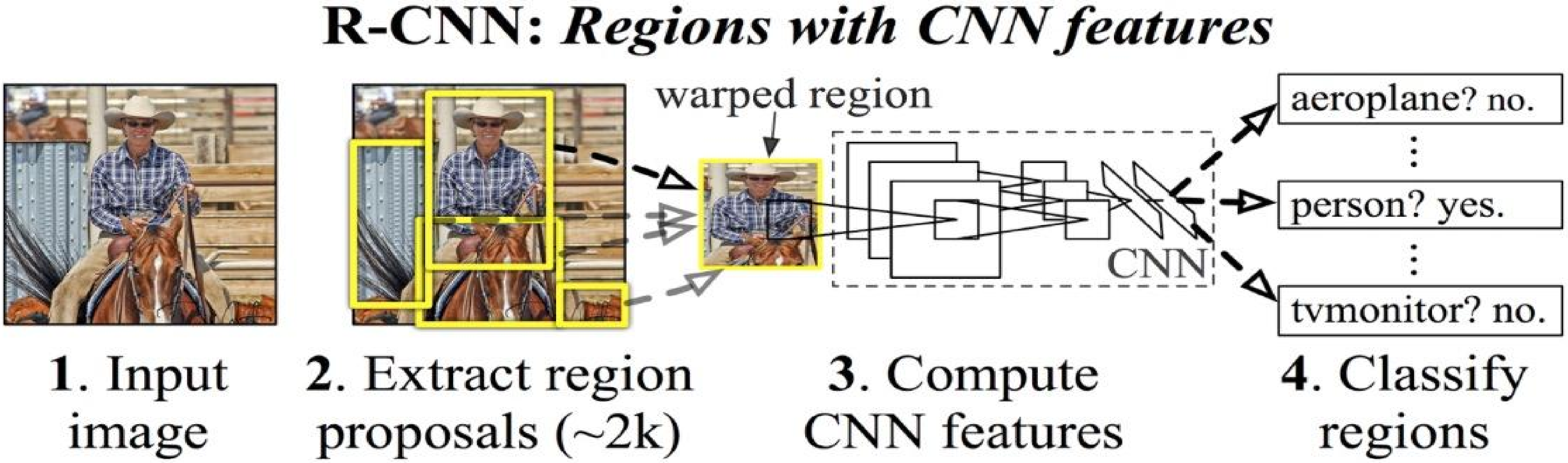


Proposal + Classification

Shortcoming:

1. Slow, impossible for real-time detection

2. Hard to optimize



WHAT’S NEW

Regression

YOLO Features：

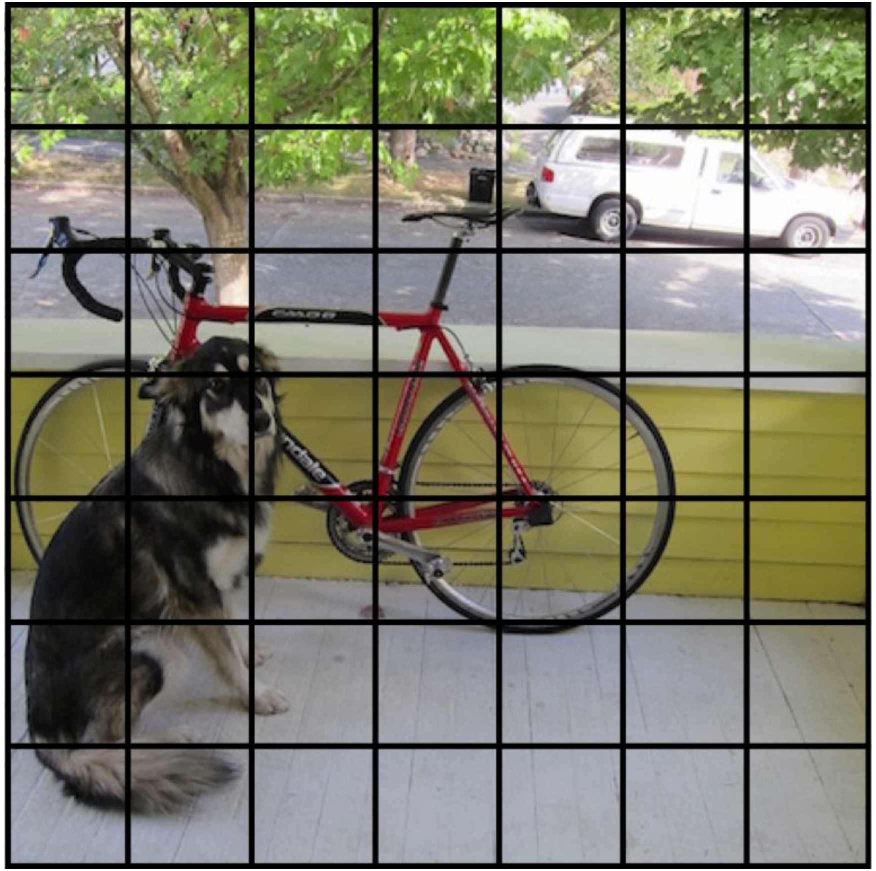
1. Extremely fast (45 frames per second)

2. Reason Globally on the Entire Image

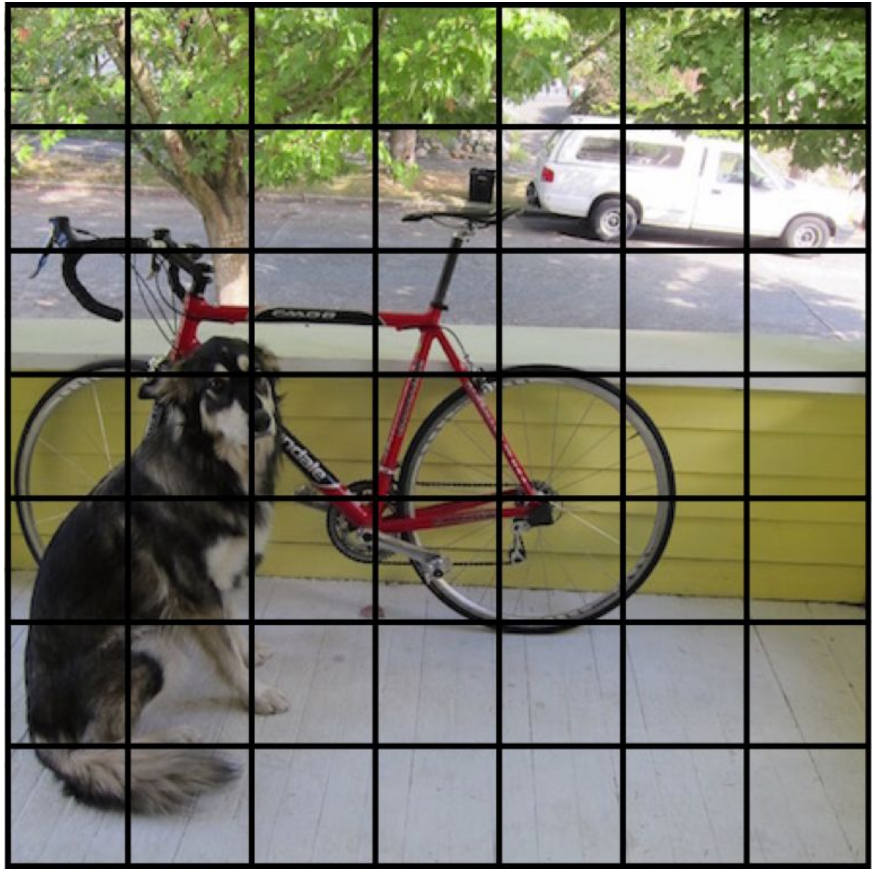
3. Learn Generalizable Representations

Detection Procedure

We split the image into an S*S grid

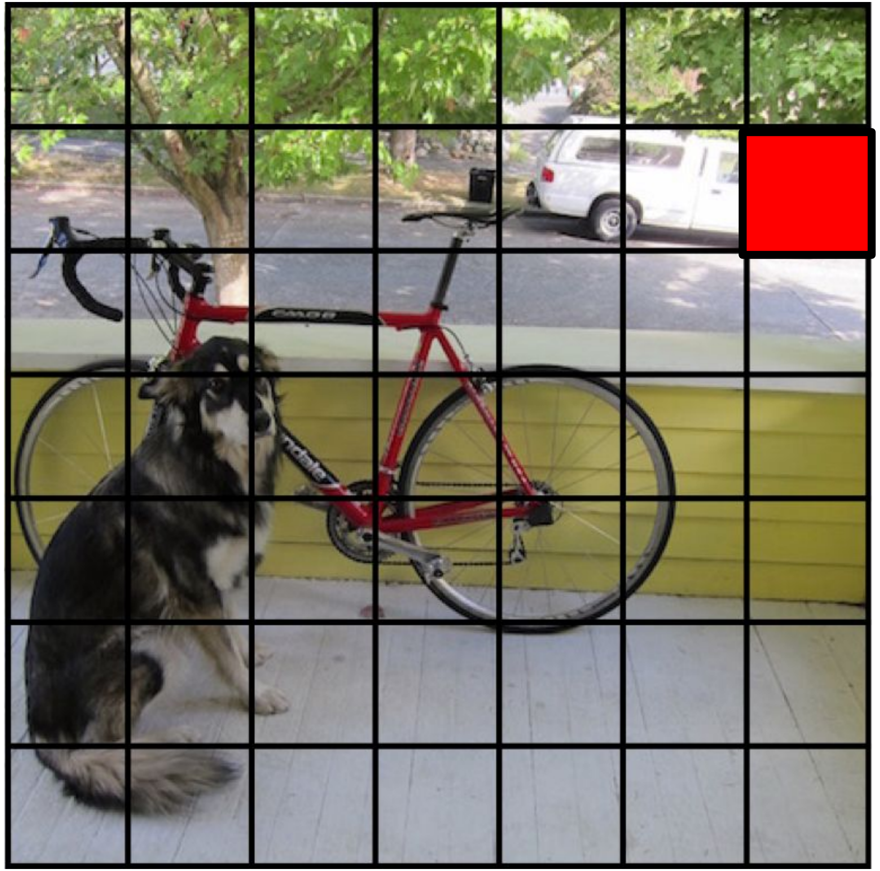


We split the image into an S*S grid



7*7 grid

Each cell predicts B boxes(x,y,w,h) and confidences of each box: P(Object)

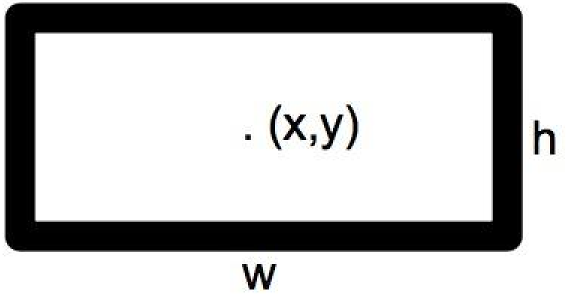


Each cell predicts B boxes(x,y,w,h) and confidences of each box: P(Object)

Each cell predicts B boxes(x,y,w,h) and confidences of each box: P(Object)

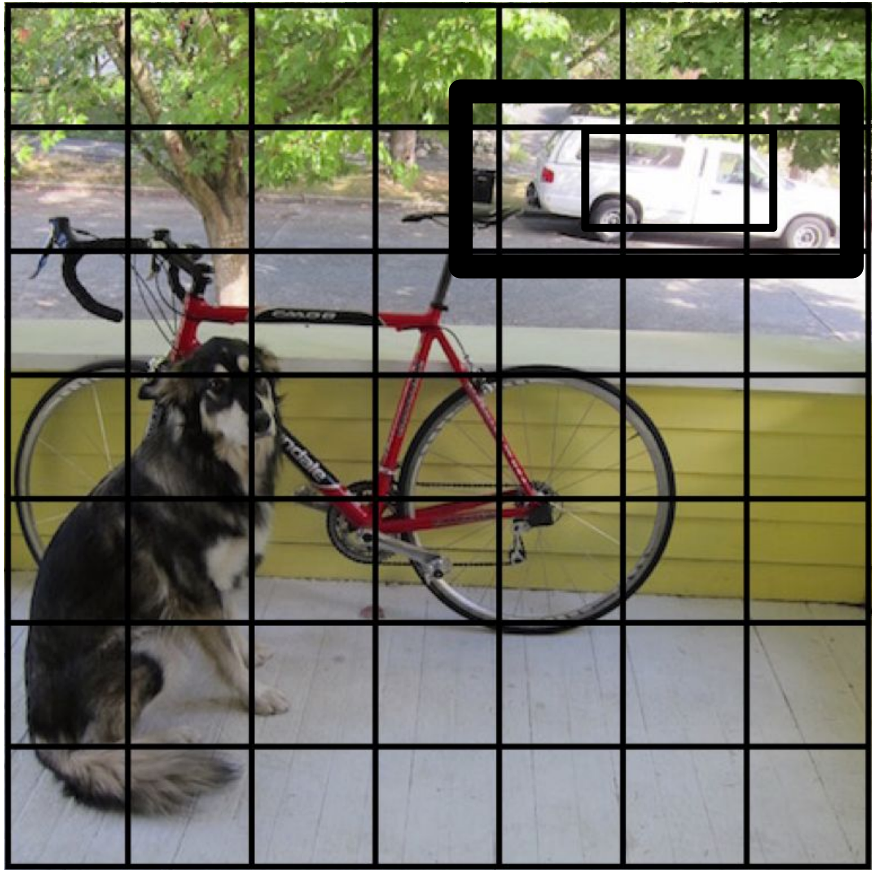
B = 2

each box predict:

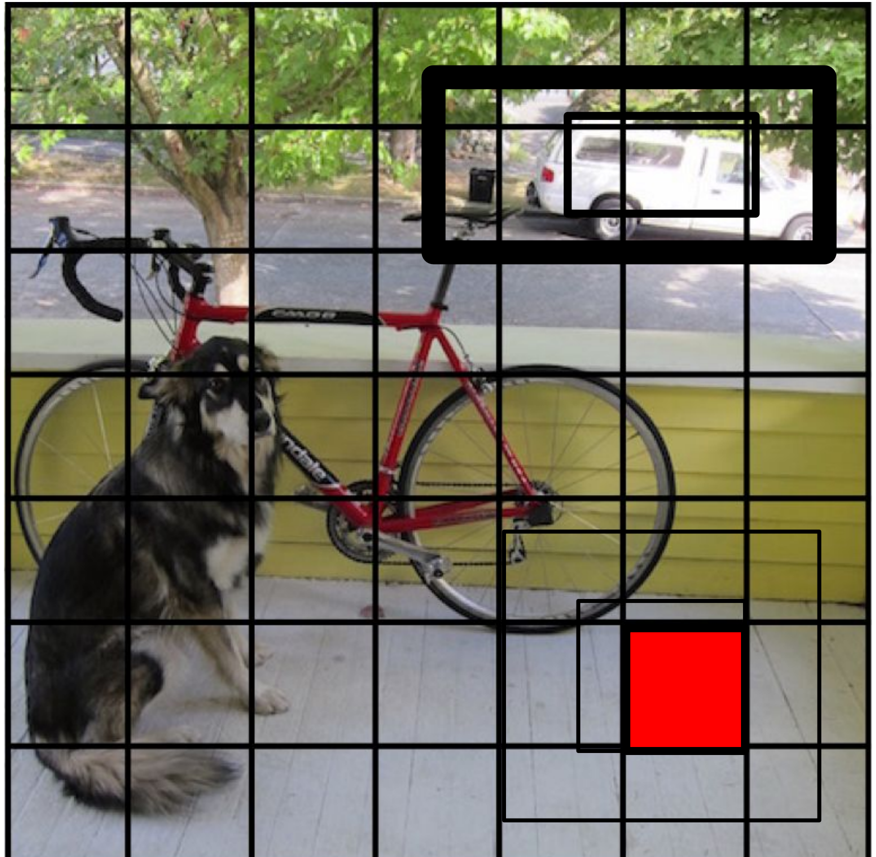


P(Object): probability that the box contains an object

Each cell predicts B boxes(x,y,w,h) and confidences of each box: P(Object)

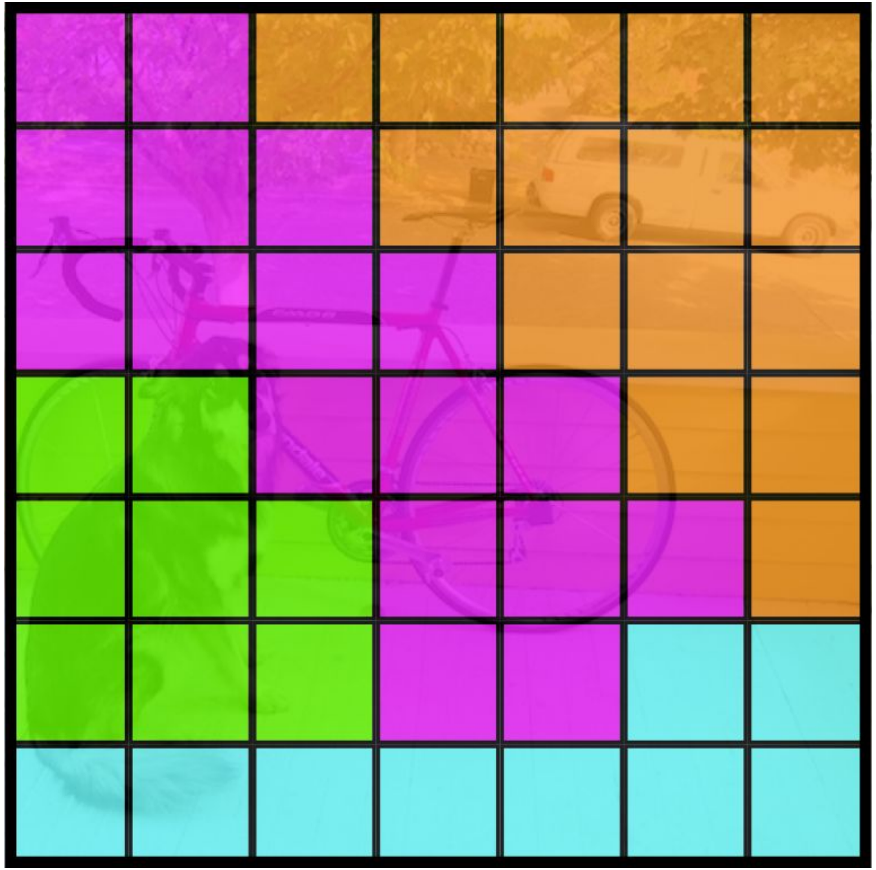


Each cell predicts B boxes(x,y,w,h) and confidences of each box: P(Object)



Each cell predicts boxes and confidences: P(Object)

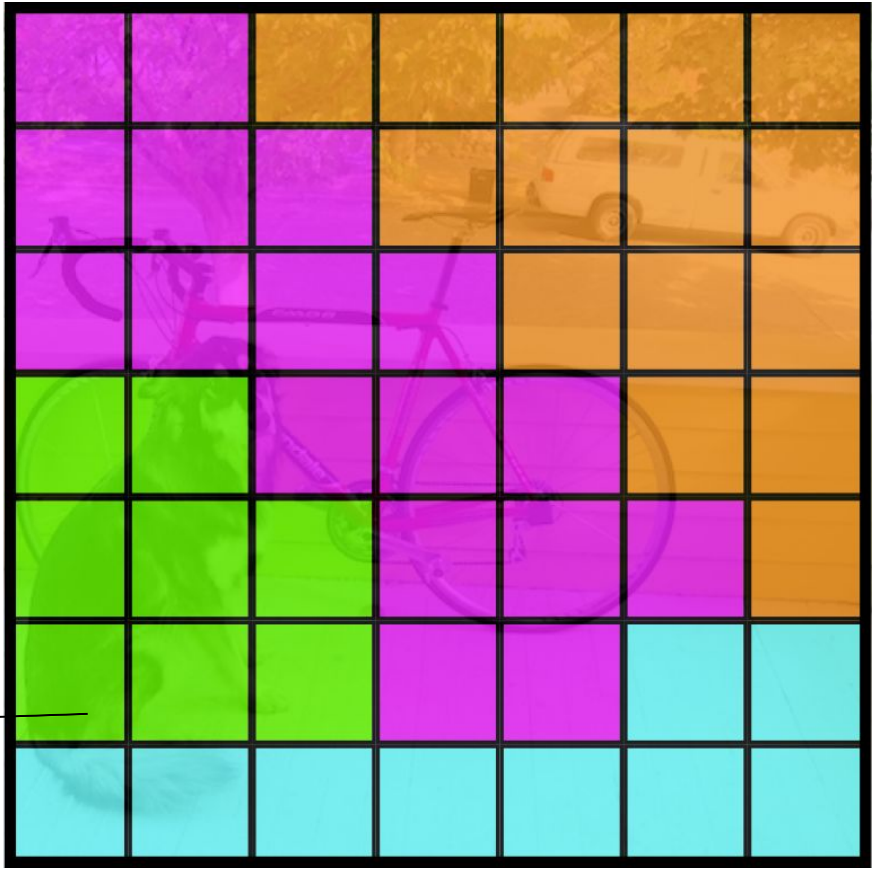
Each cell also predicts a class probability.



Conditioned on object: P(Car | Object)

Bicycle

Car

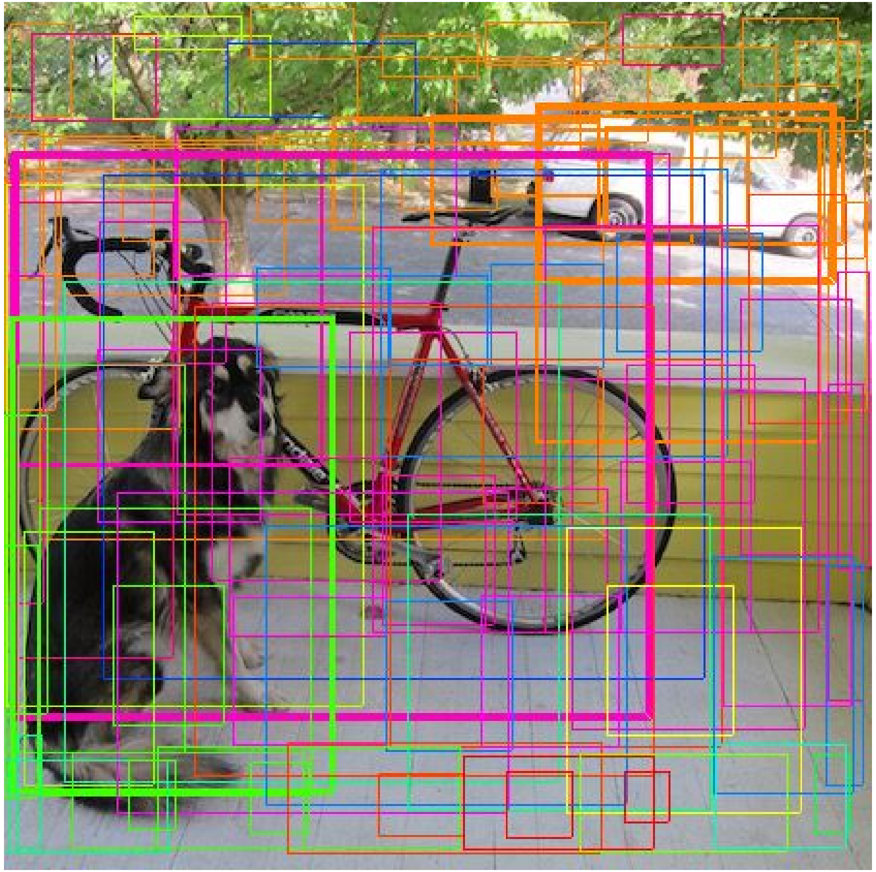


Dog

Eg.
Dog = 0.8
Cat = 0
Bike = 0

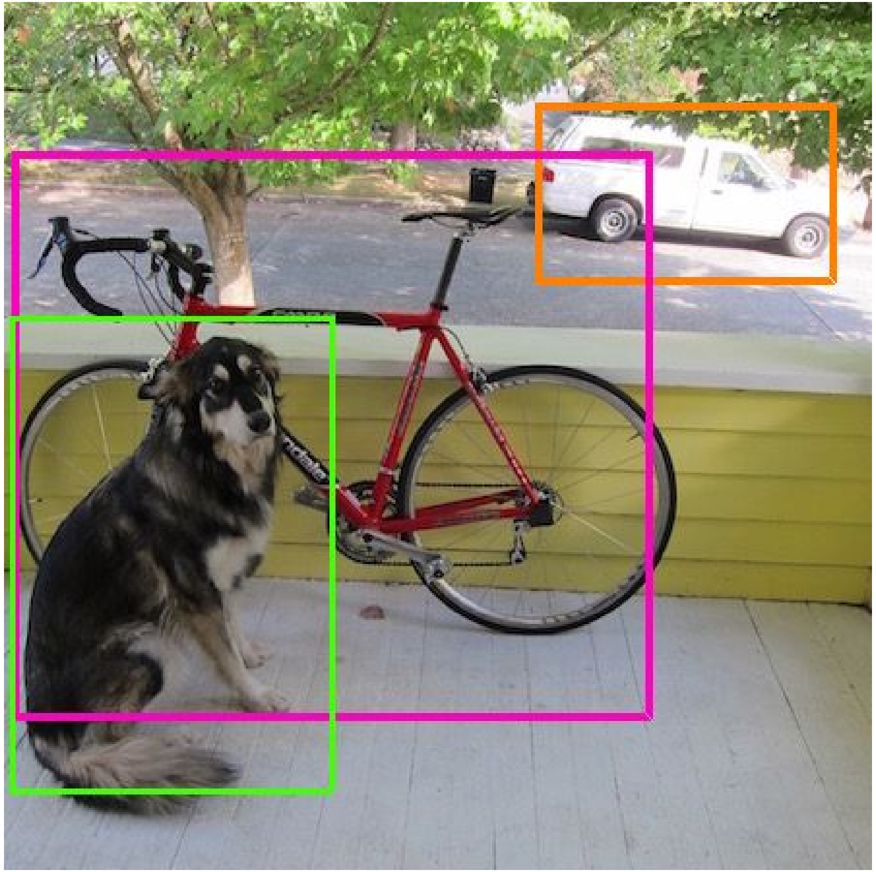
Dining
Table

Then we combine the box and class predictions.



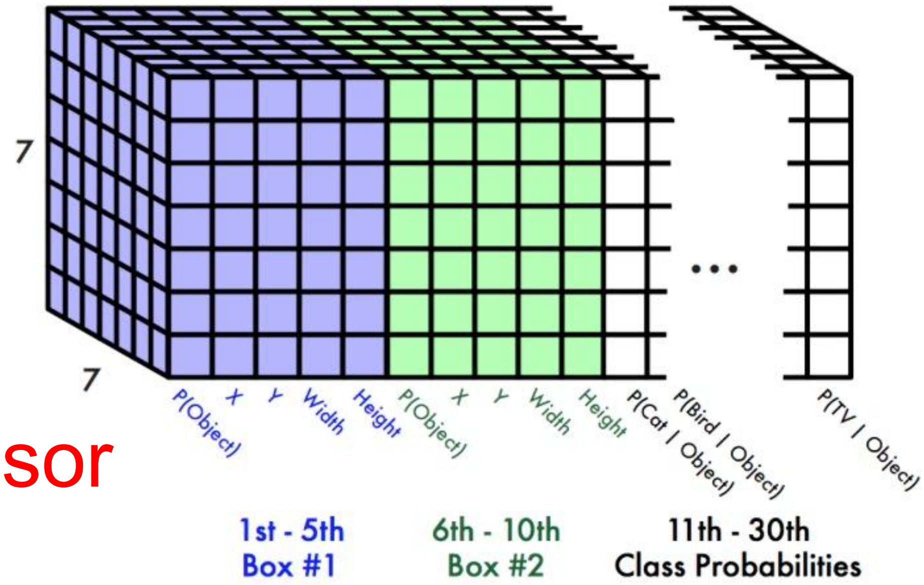
P(class|Object) * P(Object)
=P(class)

Finally we do threshold detections and NMS



Each cell predicts:

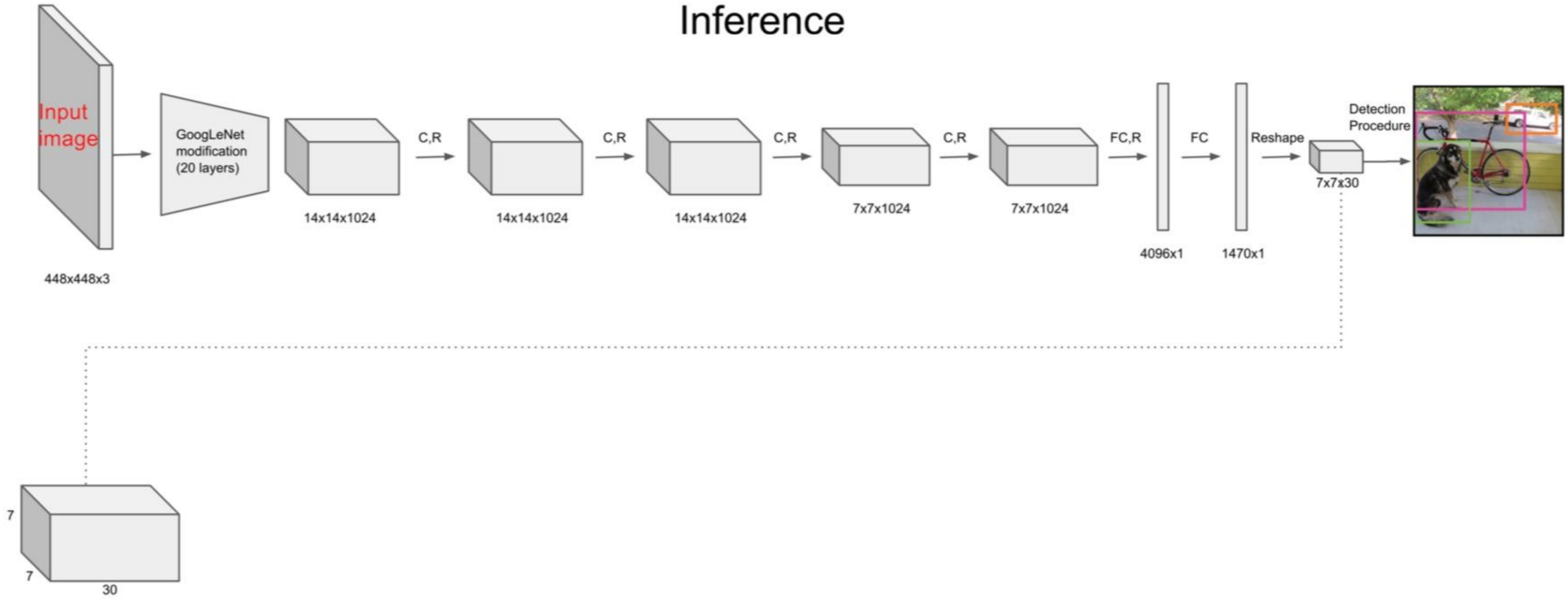
- For each bounding box:
  - 4 coordinates (x, y, w, h)
  - 1 confidence value
- Some number of class probabilities



S * S * (B * 5 + C) tensor

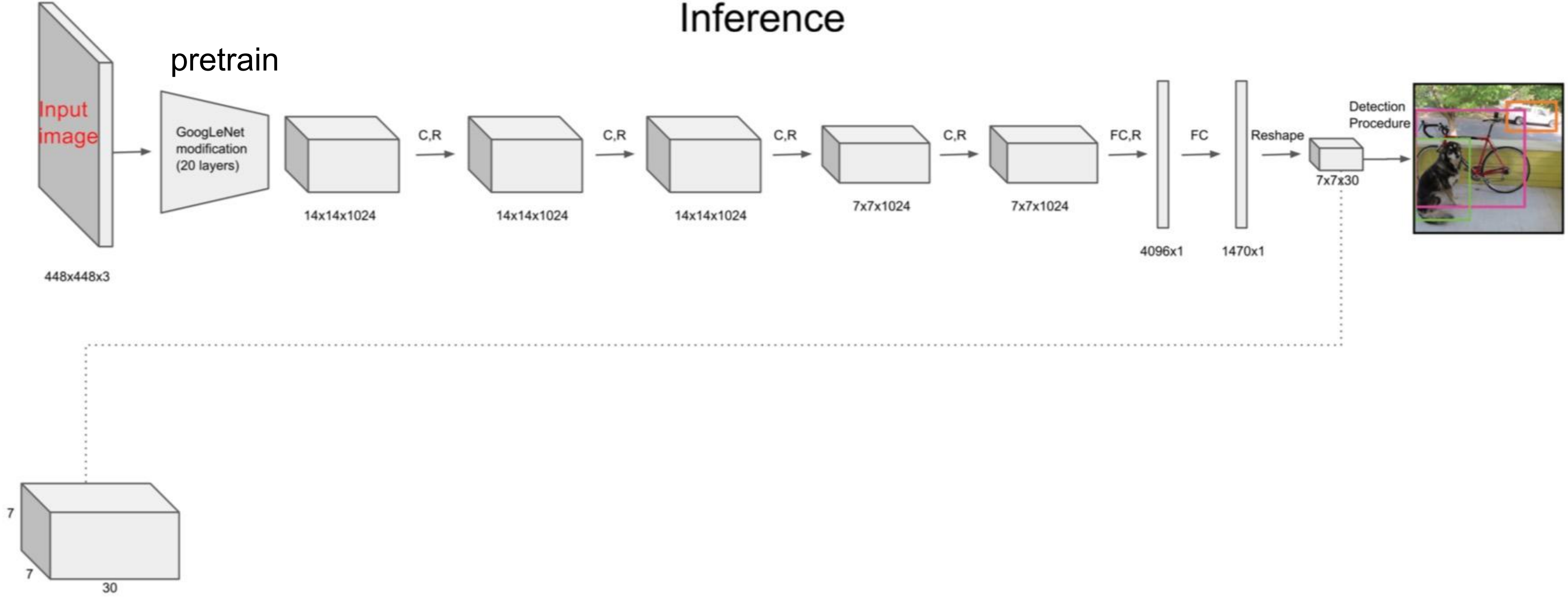
Network

Inference



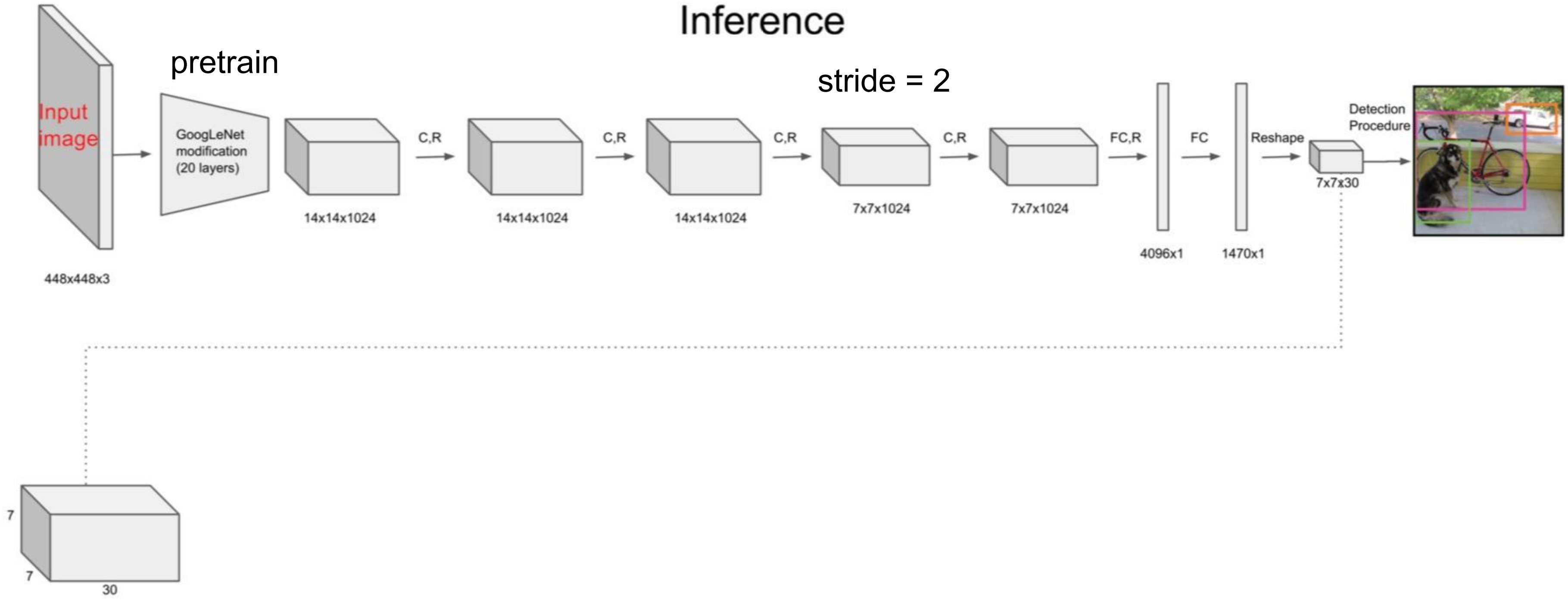
Input image
448x448x3
GoogLeNet modification (20 layers)
14x14x1024
C,R
14x14x1024
C,R
14x14x1024
C,R
7x7x1024
C,R
7x7x1024
FC,R
4096x1
FC
1470x1
Reshape
7x7x30
Detection Procedure
7
7
30

Inference



pretrain
Input image
448x448x3
GoogLeNet modification (20 layers)
14x14x1024
C,R
14x14x1024
C,R
14x14x1024
C,R
7x7x1024
C,R
7x7x1024
FC,R
4096x1
FC
1470x1
Reshape
7x7x30
Detection Procedure
7
7
30

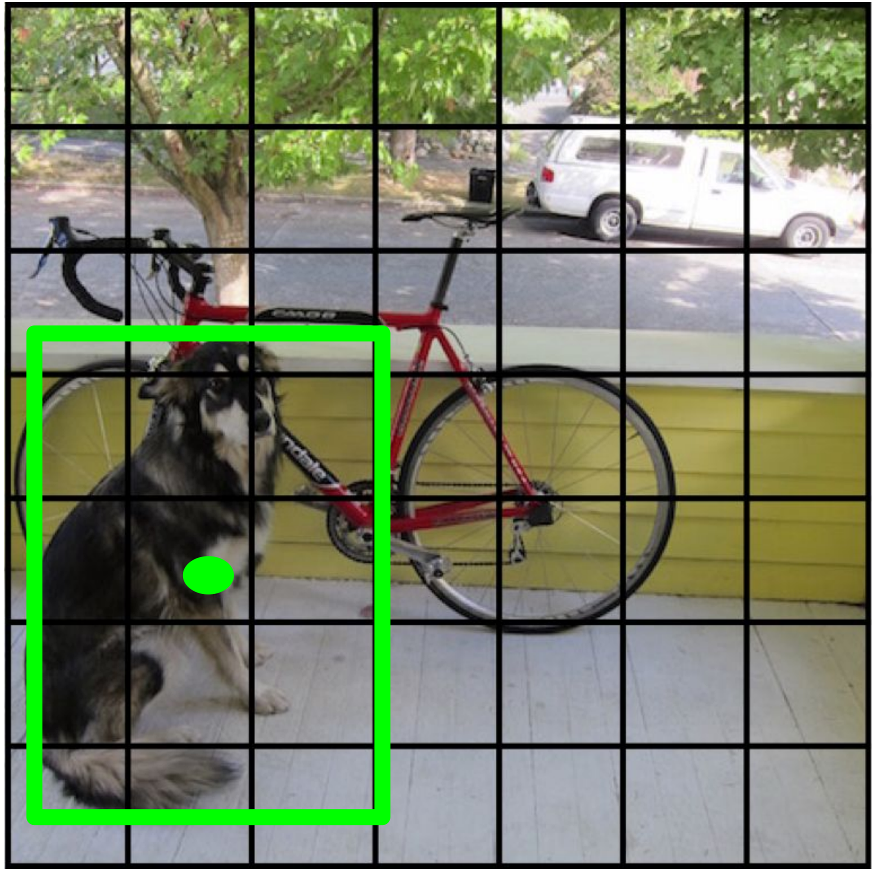
Inference



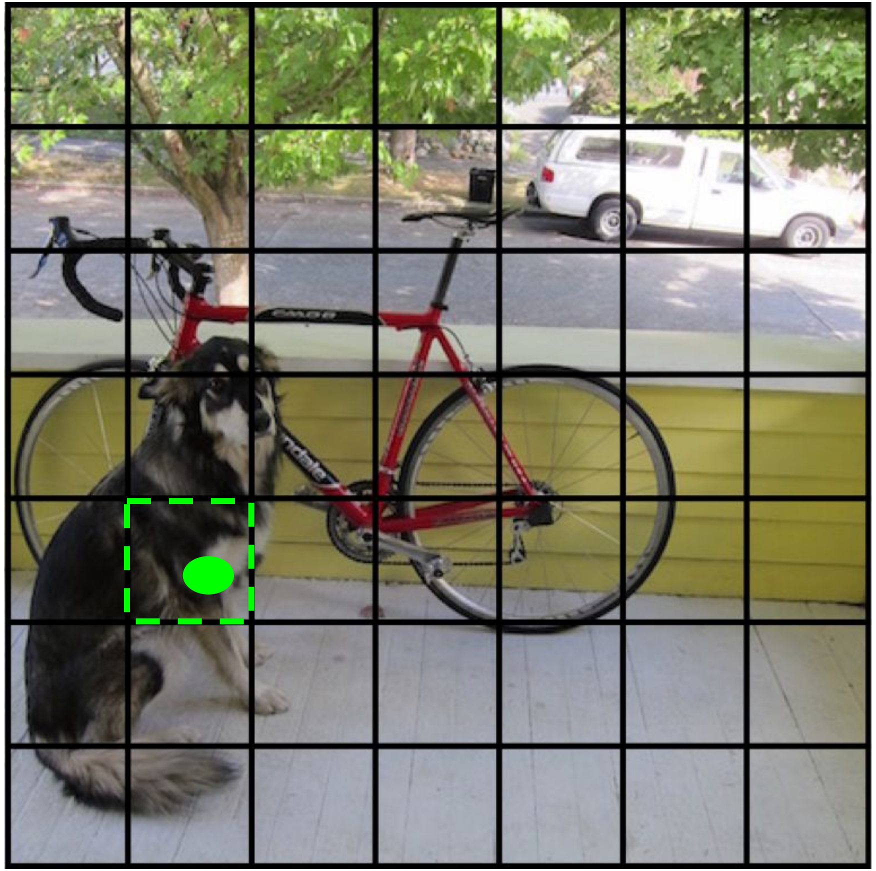
pretrain
stride = 2
Input image
448x448x3
GoogLeNet modification (20 layers)
14x14x1024
C,R
14x14x1024
C,R
14x14x1024
C,R
7x7x1024
C,R
7x7x1024
FC,R
4096x1
FC
1470x1
Reshape
7x7x30
Detection Procedure
7
7
30

Train

During training, match example to the right cell

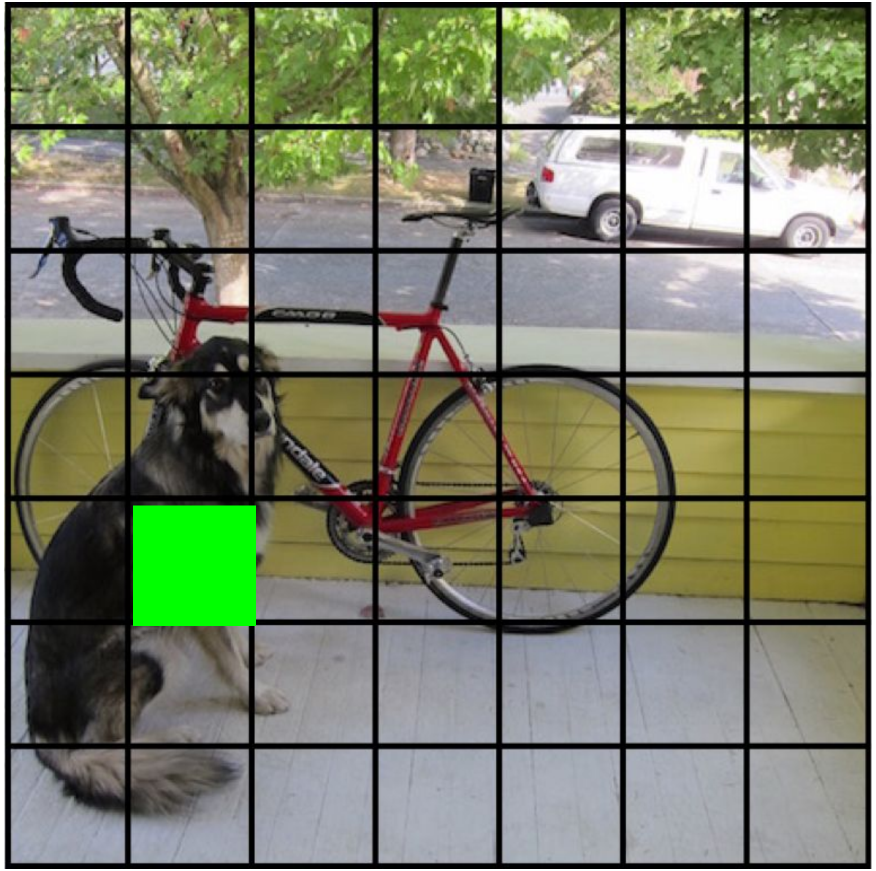


During training, match example to the right cell

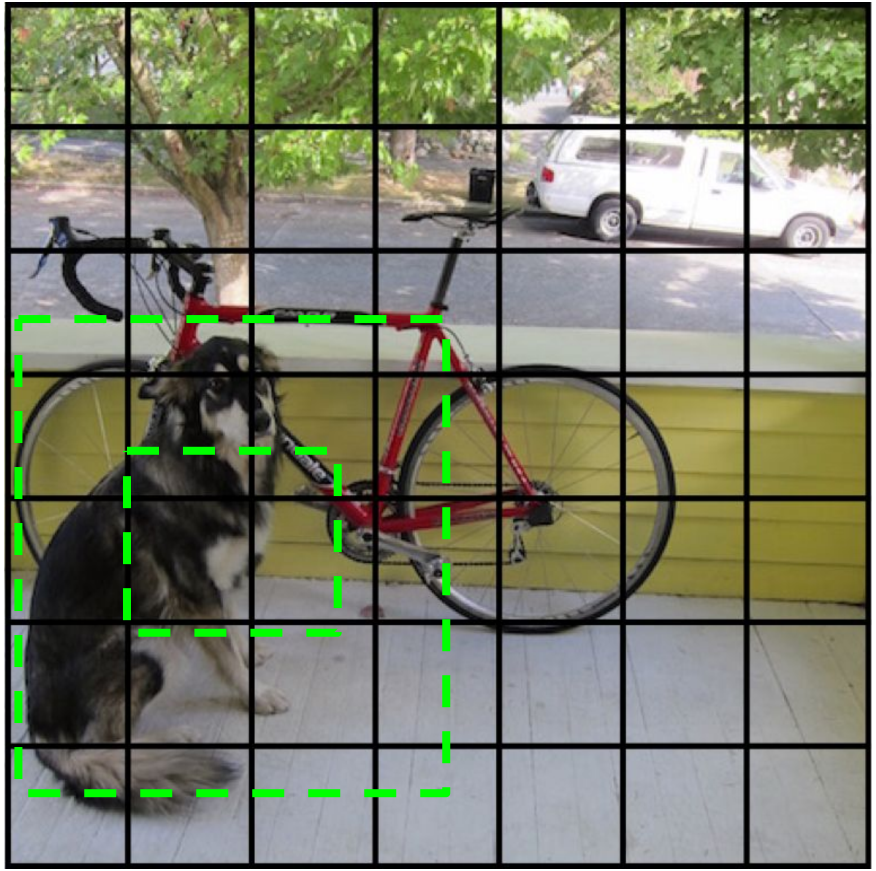


Adjust that cell's class prediction

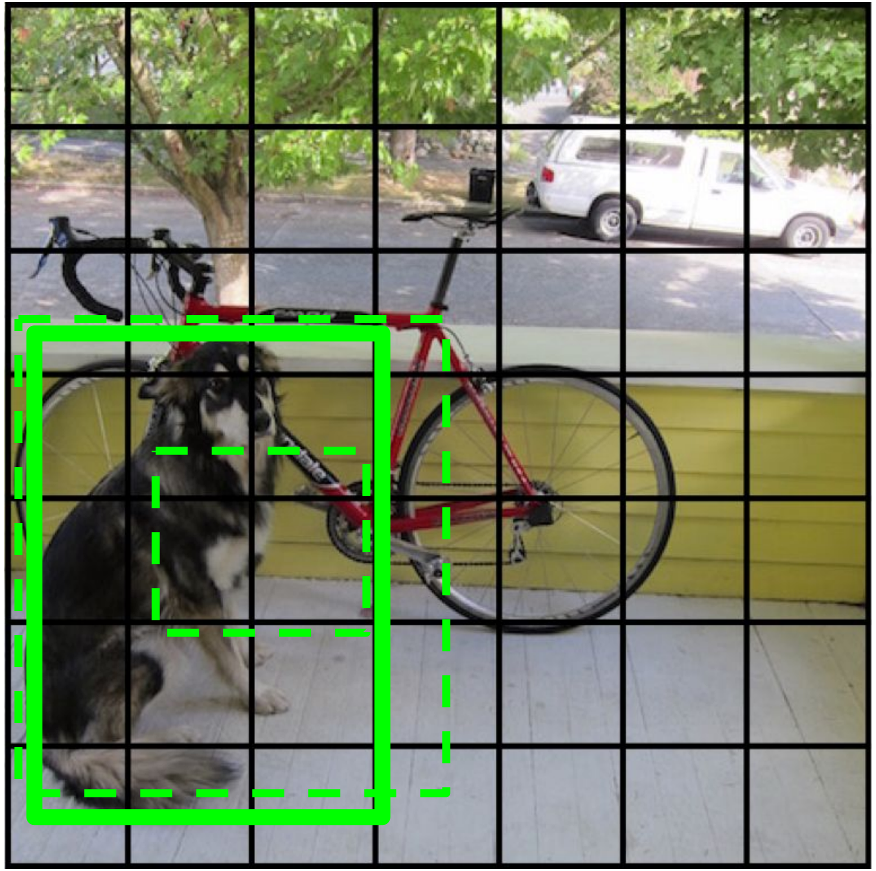
Dog = 1
Cat = 0
Bike = 0
...



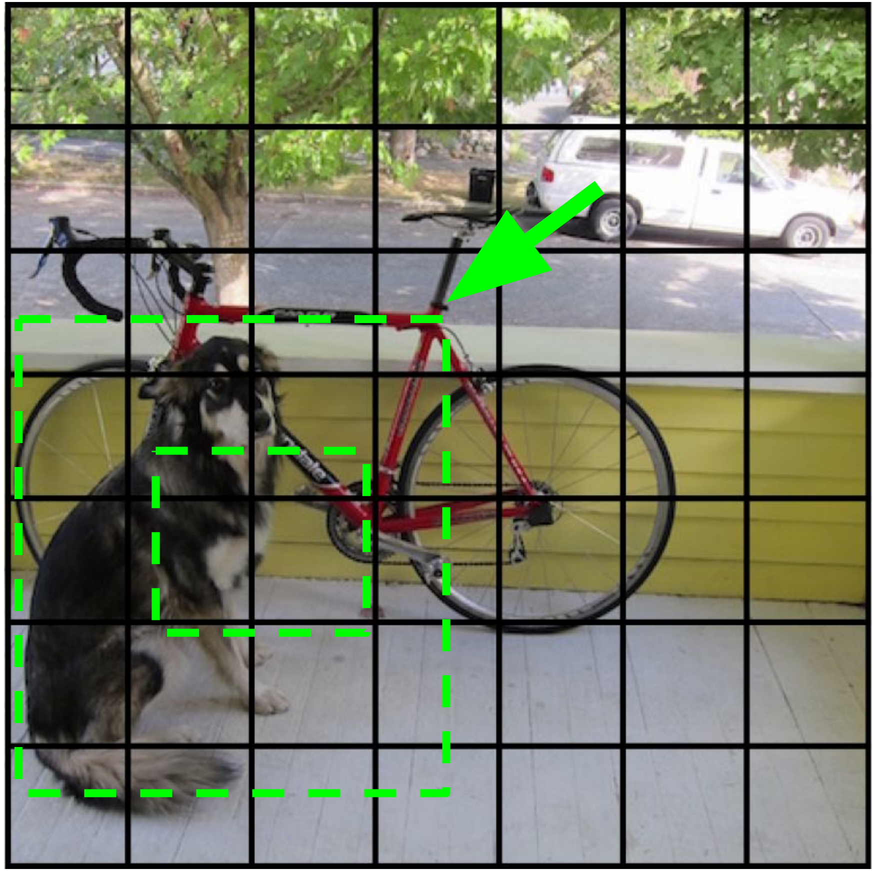
Look at that cell's predicted boxes



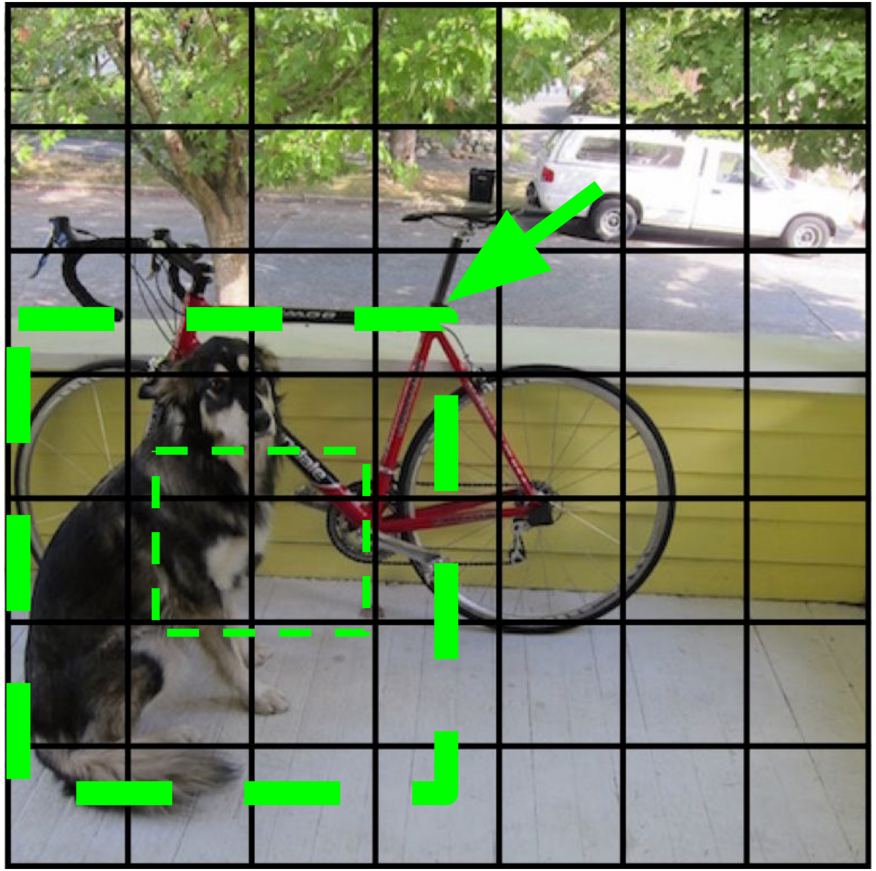
Find the best one, adjust it, increase the confidence



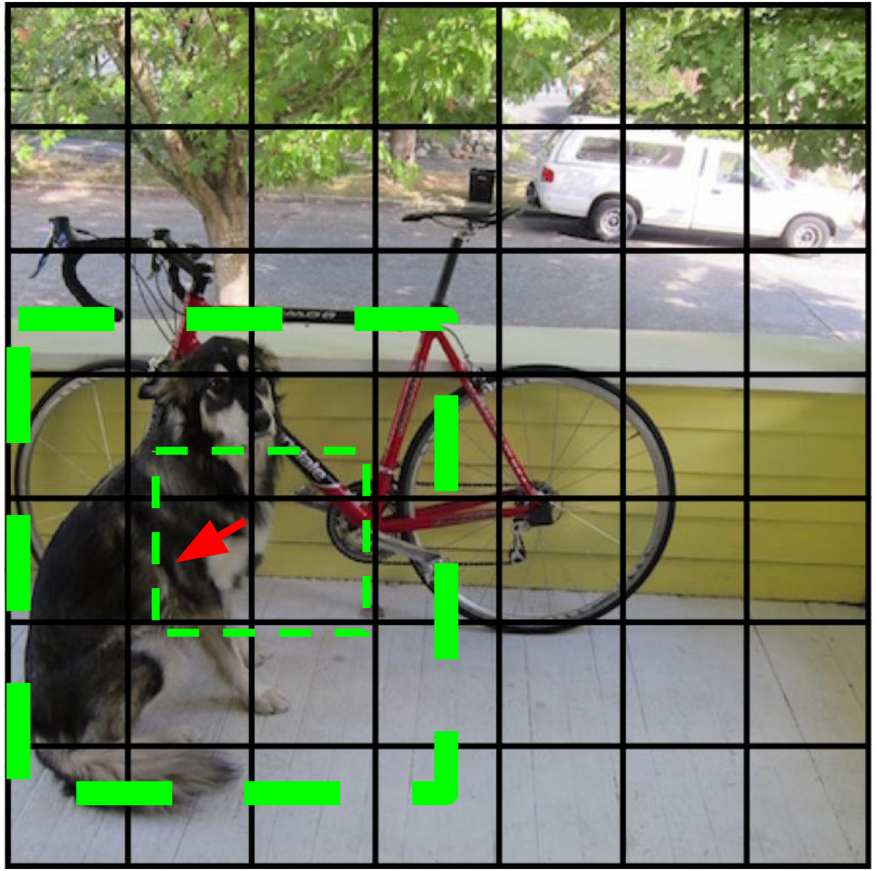
Find the best one, adjust it, increase the confidence



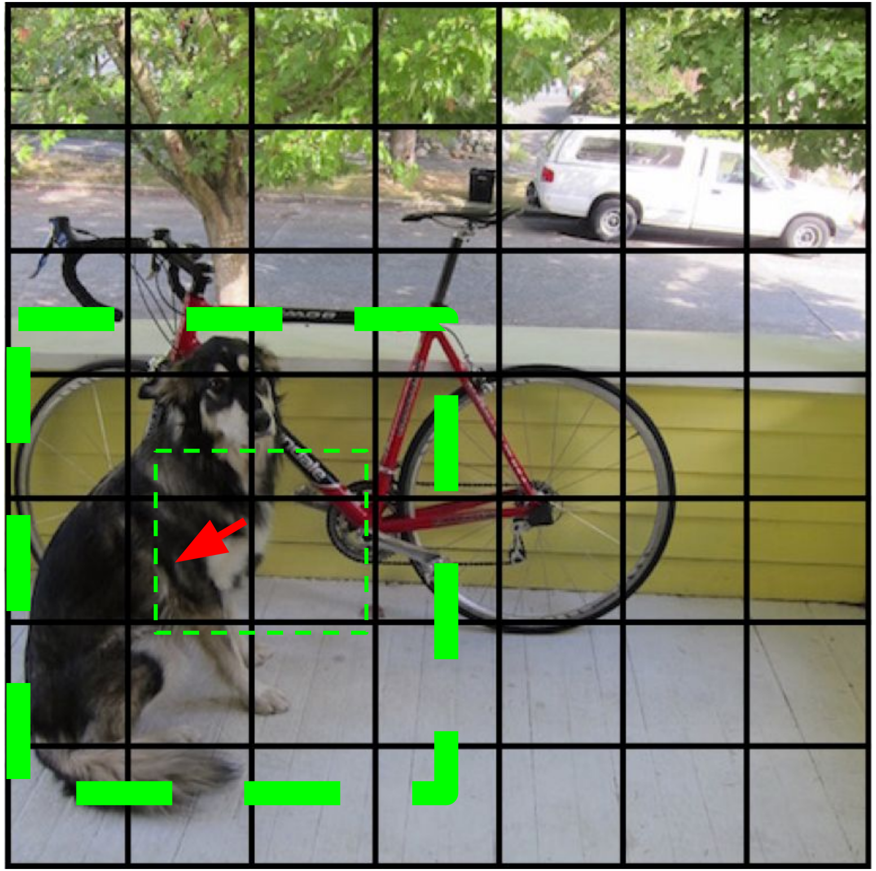
Find the best one, adjust it, increase the confidence



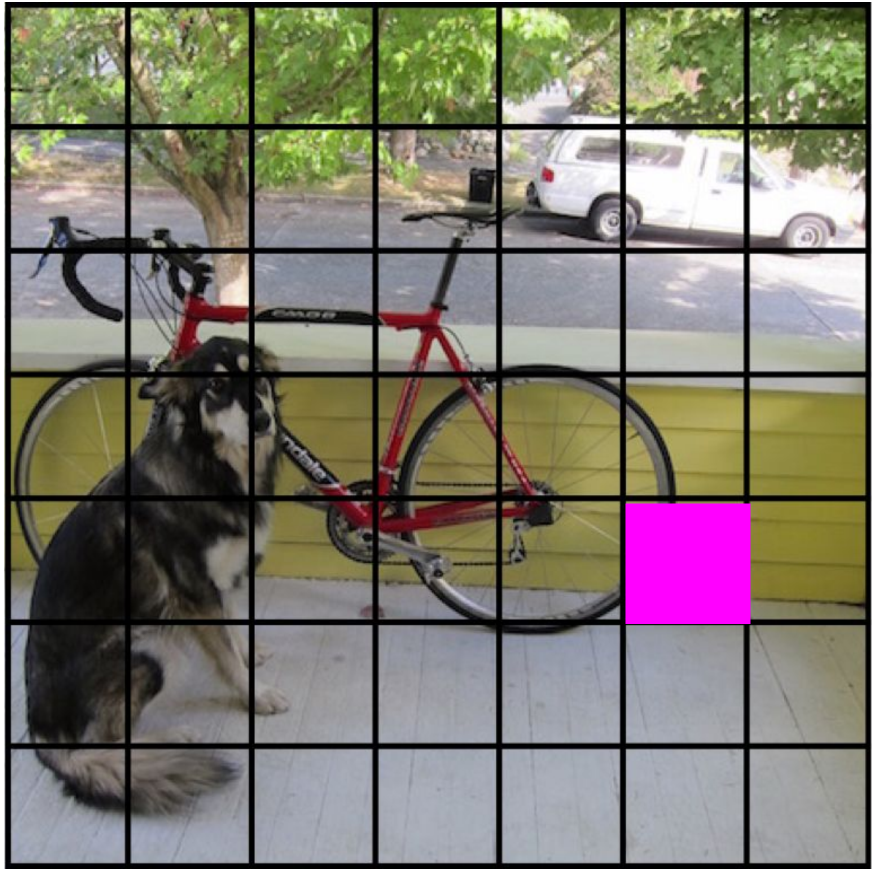
Decrease the confidence of the other box



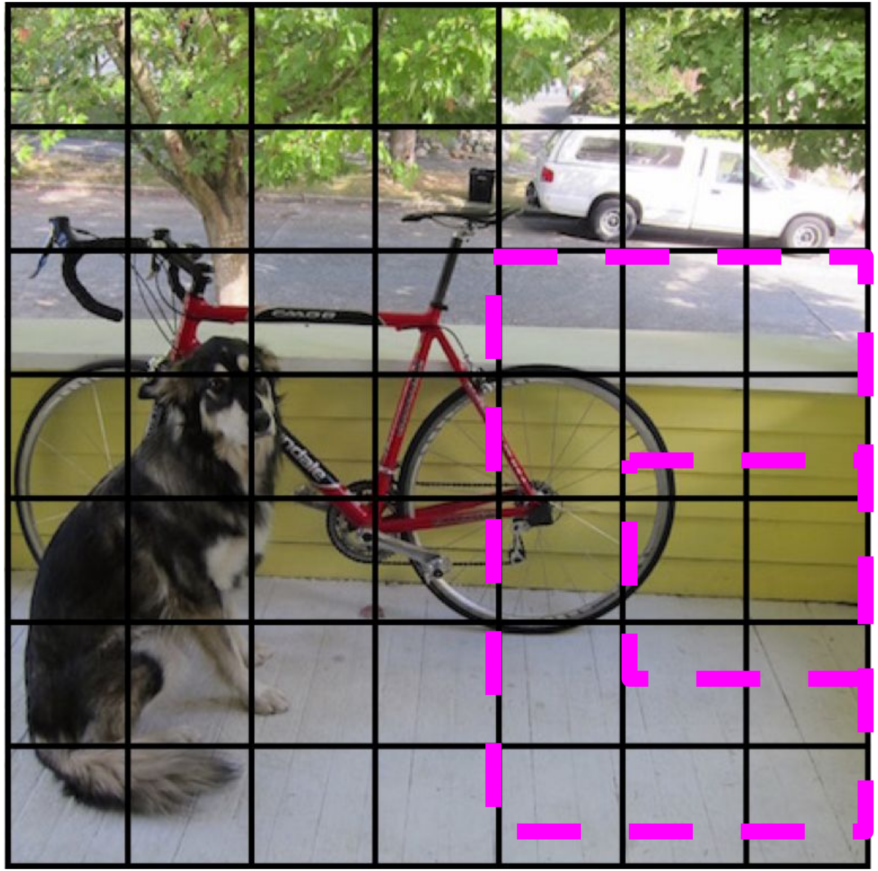
Decrease the confidence of the other box



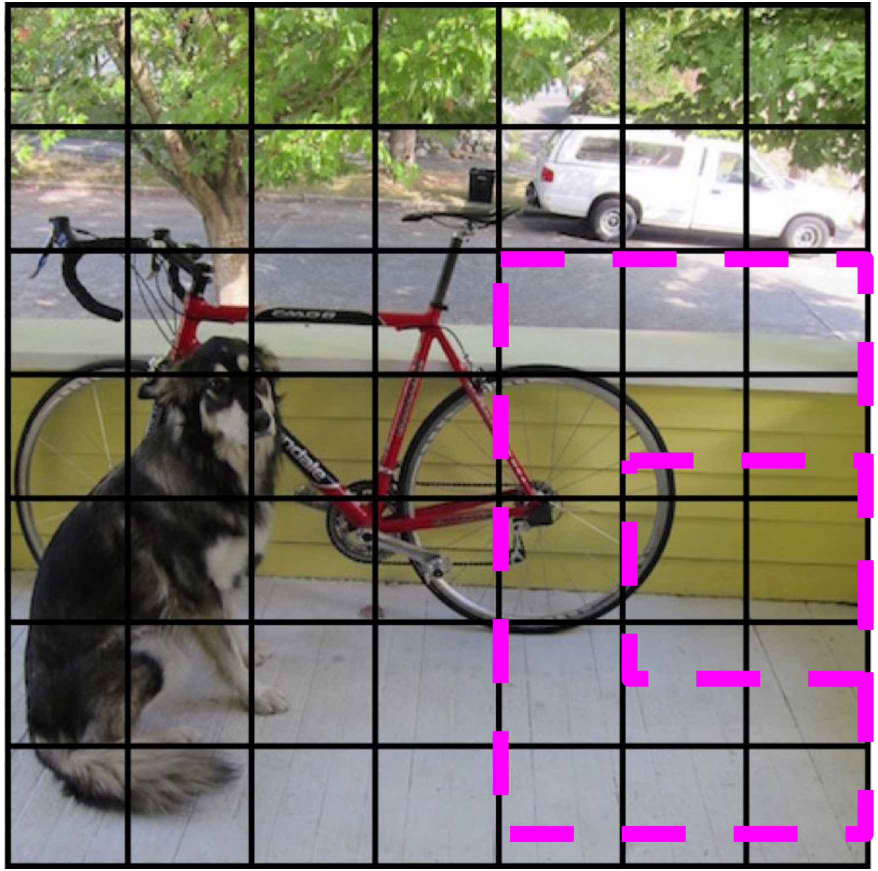
Some cells don’t have any ground truth detections!



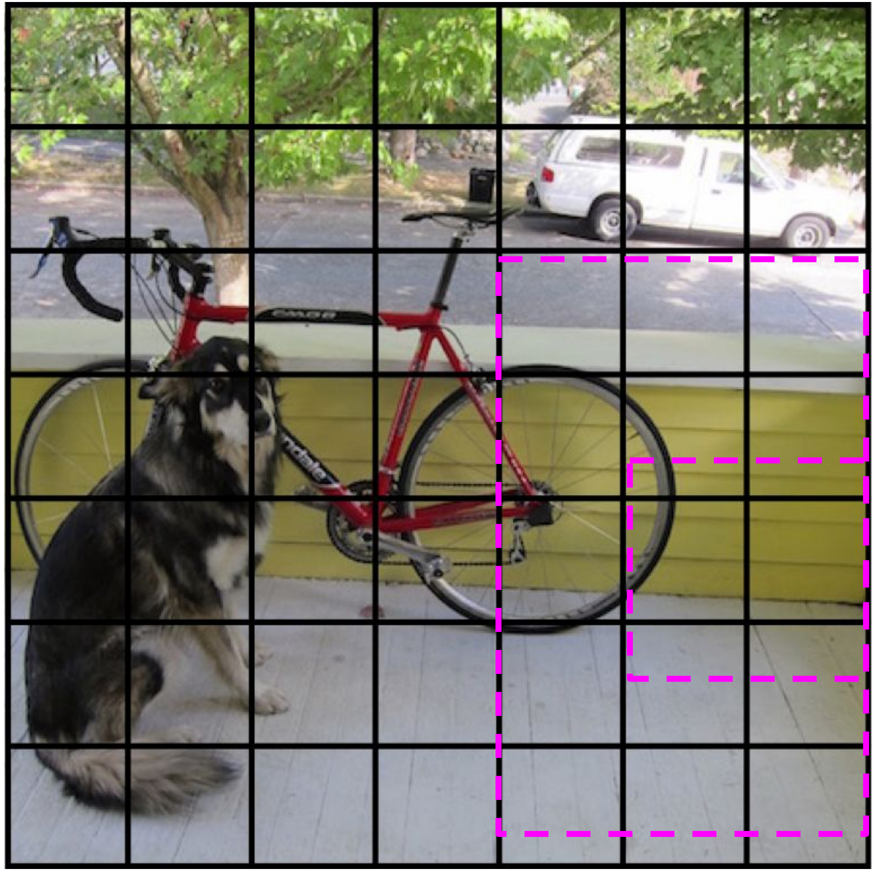
Some cells don’t have any ground truth detections!



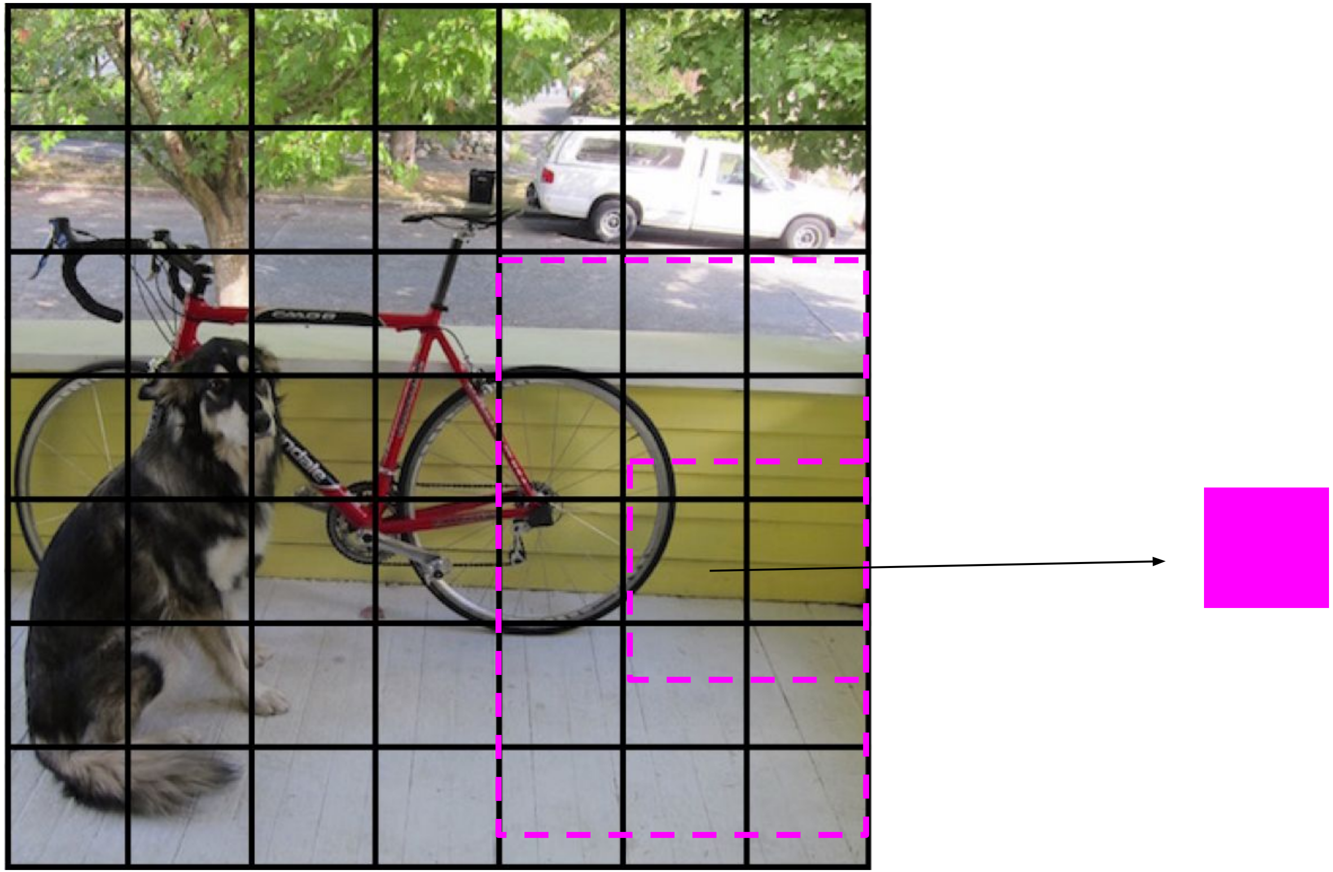
Decrease the confidence of boxes boxes



Decrease the confidence of these boxes



Don’t adjust the class probabilities or coordinates



Loss Function (sum-squared error)

loss function:

$$
\begin{gathered}
\lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
+ \lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \\
+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2 \\
+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2 \\
+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \qquad (3)
\end{gathered}
$$

model. We use sum-squared error because it is easy to optimize, however it does not perfectly align with our goal of maximizing average precision. It weights localization error equally with classification error which may not be ideal. Also, in every image many grid cells do not contain any object. This pushes the "confidence" scores of those cells towards zero, often overpowering the gradient from cells that do contain objects. This can lead to model instability, causing training to diverge early on.

To remedy this, we increase the loss from bounding box coordinate predictions and decrease the loss from confidence predictions for boxes that don't contain objects. We use two parameters, λ_{coord} and λ_{noobj} to accomplish this. We set $\lambda_{\text{coord}} = 5$ and $\lambda_{\text{noobj}} = .5$.

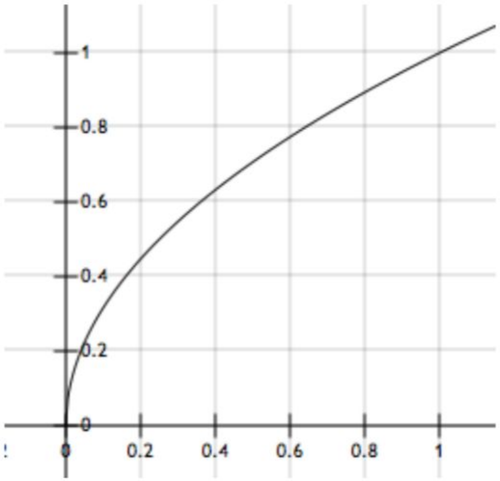
$$\lambda_{\text{coord}}=5,\ \lambda_{\text{noobj}}=0.5$$

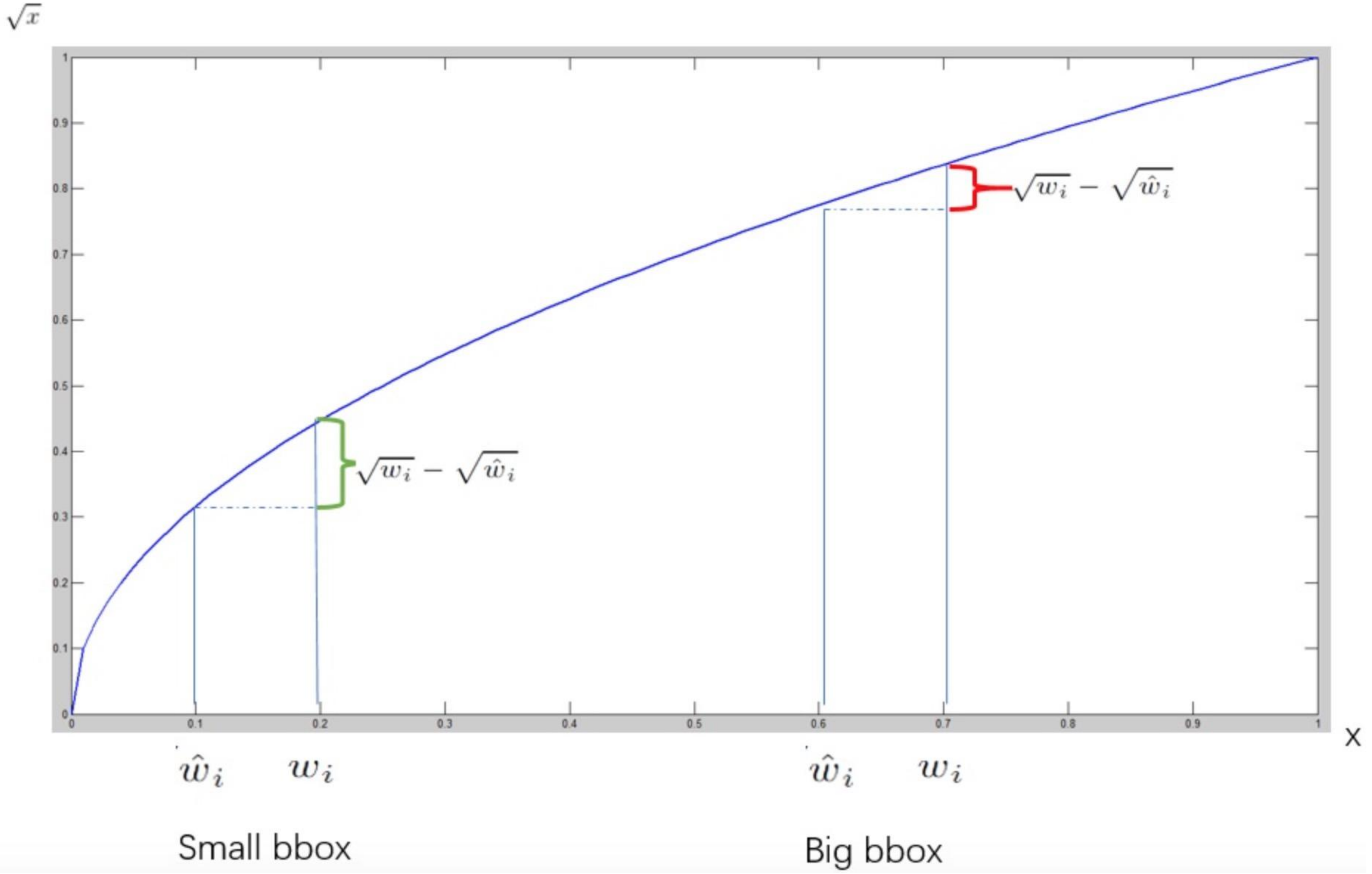
Loss Function (sum-squared error)

loss function:

$$
\begin{aligned}
&\lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
&+ \lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \boxed{\left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right]} \\
&+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2 \\
&+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2 \\
&+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2
\end{aligned} \qquad (3)
$$

Sum-squared error also equally weights errors in large boxes and small boxes. Our error metric should reflect that small deviations in large boxes matter less than in small boxes. To partially address this we predict the square root of the bounding box width and height instead of the width and height directly.





$\sqrt{x}$
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
$\sqrt{w_i} - \sqrt{\hat{w}_i}$
$\sqrt{w_i} - \sqrt{\hat{w}_i}$
x
$\hat{w}_i$
w_i
$\hat{w}_i$
w_i
Small bbox
Big bbox

Loss Function (sum-squared error)

loss function:

$$\lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \boxed{\mathbb{1}_{ij}^{\text{obj}}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right]$$
$$+ \lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \boxed{\mathbb{1}_{ij}^{\text{obj}}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right]$$
$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \boxed{\mathbb{1}_{ij}^{\text{obj}}} \left(C_i - \hat{C}_i \right)^2$$
$$+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \boxed{\mathbb{1}_{ij}^{\text{noobj}}} \left(C_i - \hat{C}_i \right)^2$$
$$+ \sum_{i=0}^{S^2} \boxed{\mathbb{1}_{i}^{\text{obj}}} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \qquad (3)$$

$\boxed{\mathbb{1}_{ij}^{obj}}$ **The *j*th bbox predictor** in ***cell i*** is "responsible" for that prediction

$\boxed{\mathbb{1}_{ij}^{noobj}}$

$\boxed{\mathbb{1}_{i}^{obj}}$ If object appears in ***cell i***

Note that the loss function only penalizes classification error if an object is present in that grid cell (hence the conditional class probability discussed earlier). It also only penalizes bounding box coordinate error if that predictor is "responsible" for the ground truth box (i.e. has the highest IOU of any predictor in that grid cell).

Experiments

- Datasets
- PASCAL VOC 2007 & VOC 2012

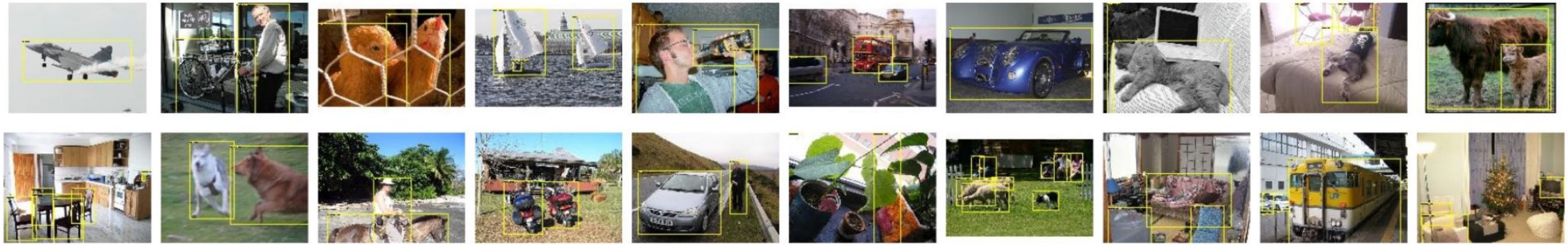
2007	20 classes: • *Person:* person • *Animal:* bird, cat, cow, dog, horse, sheep • *Vehicle:* aeroplane, bicycle, boat, bus, car, motorbike, train • *Indoor:* bottle, chair, dining table, potted plant, sofa, tv/monitor Train/validation/test: 9,963 images containing 24,640 annotated objects.

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

Experiments

- Datasets

20 classes



Accurate object detection is slow!

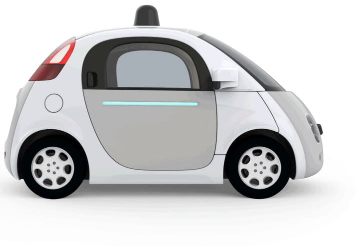
	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img

Accurate object detection is slow!

	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img

Accurate object detection is slow!

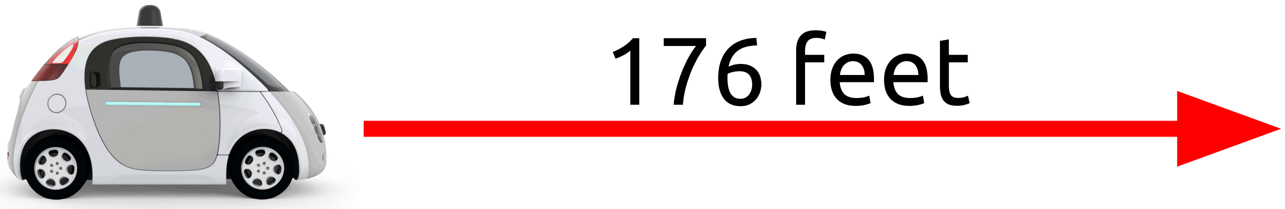
	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img



⅓ Mile, 1760 feet

Accurate object detection is slow!

	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img
Fast R-CNN	70.0	.5 FPS	2 s/img



Accurate object detection is slow!

	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img
Fast R-CNN	70.0	.5 FPS	2 s/img
Faster R-CNN	73.2	7 FPS	140 ms/img

Accurate object detection is slow!

	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img
Fast R-CNN	70.0	.5 FPS	2 s/img
Faster R-CNN	73.2	7 FPS	140 ms/img
YOLO	63.4	45 FPS	22 ms/img

2 feet

Error Analysis

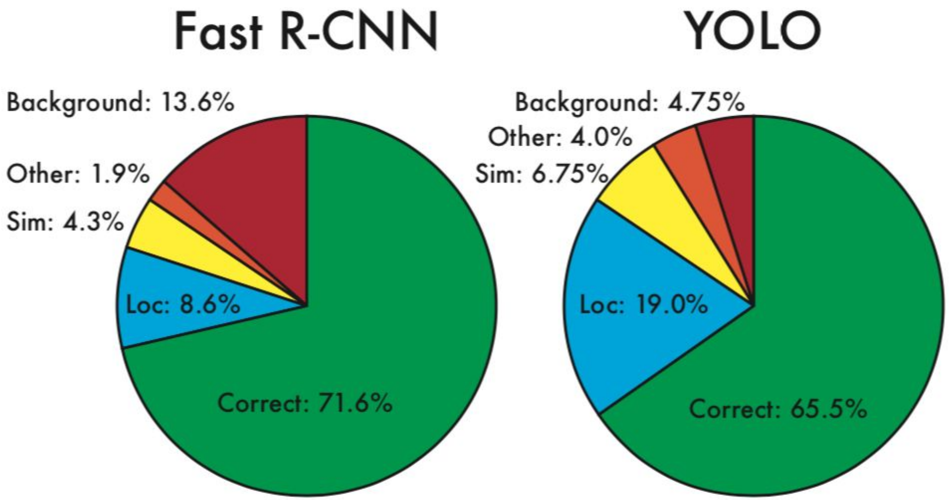


Figure 4: Error Analysis: Fast R-CNN vs. YOLO These charts show the percentage of localization and background errors in the top N detections for various categories (N = # objects in that category).

Loc: Localization Error

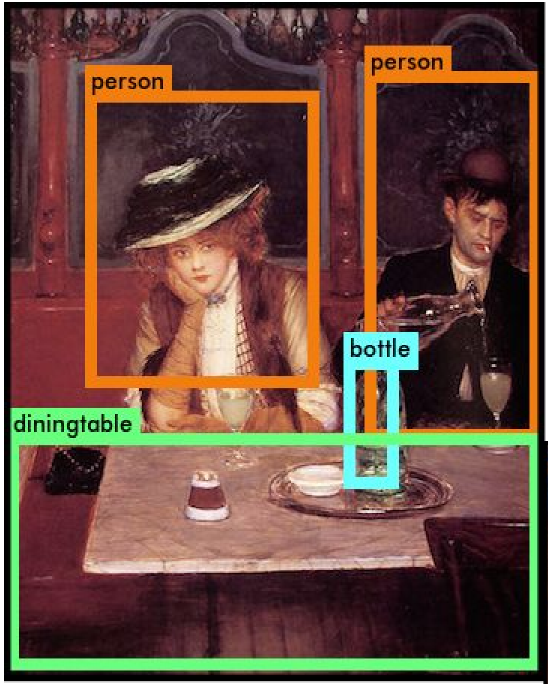
Correct class,

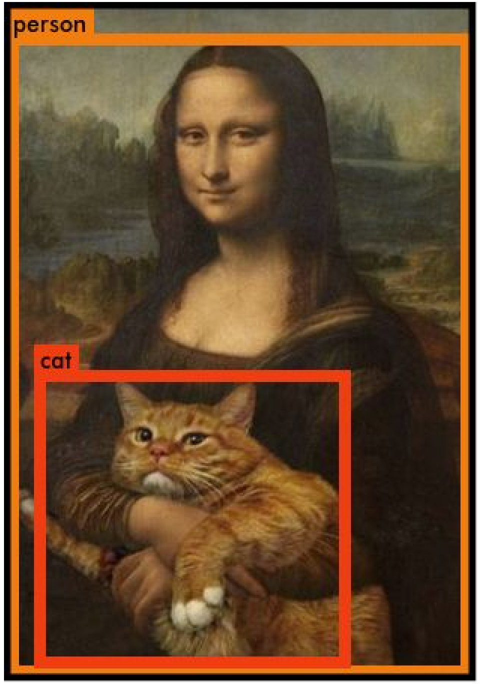
.1<IOU<.5

Background:

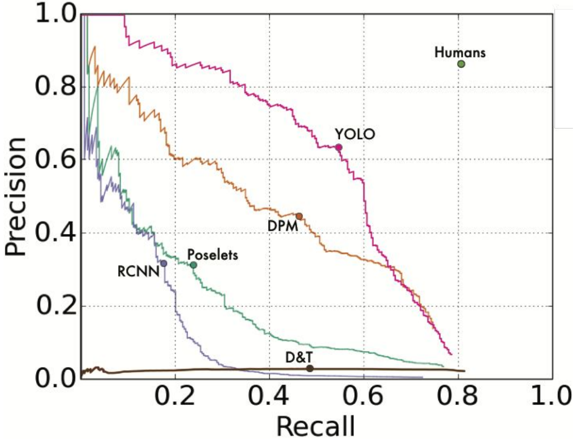
IOU<0.1

YOLO generalizes well to new domains (like art)





It outperforms methods like DPM and R-CNN when generalizing to person detection in artwork

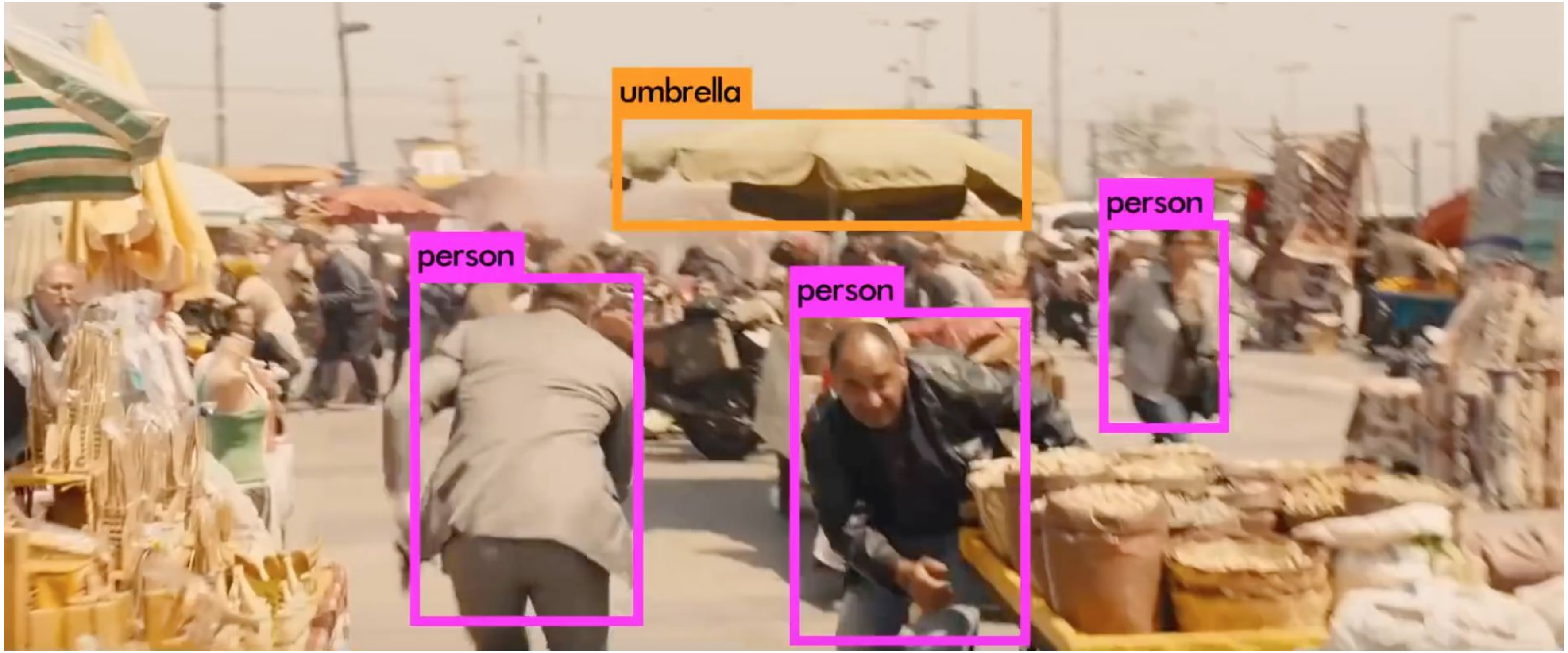


	VOC 2007	Picasso		People-Art
	AP	AP	Best F_1	AP
YOLO	**59.2**	**53.3**	**0.590**	**45**
R-CNN	54.2	10.4	0.226	26
DPM	43.2	37.8	0.458	32

S. Ginosar, D. Haas, T. Brown, and J. Malik. Detecting people in cubist art. In Computer Vision-ECCV 2014 Workshops, pages 101–116. Springer, 2014.

H. Cai, Q. Wu, T. Corradi, and P. Hall. The cross-depiction problem: Computer vision algorithms for recognising objects in artwork and in photographs.

Demo



Strengths and Weaknesses

- Strengths:
    - Fast: 45fps, smaller version 155fps
    - End2end training
    - Background error is low

Strengths and Weaknesses

- Weaknesses:
  - Performance is lower than state-of-art
  - Makes more localization errors