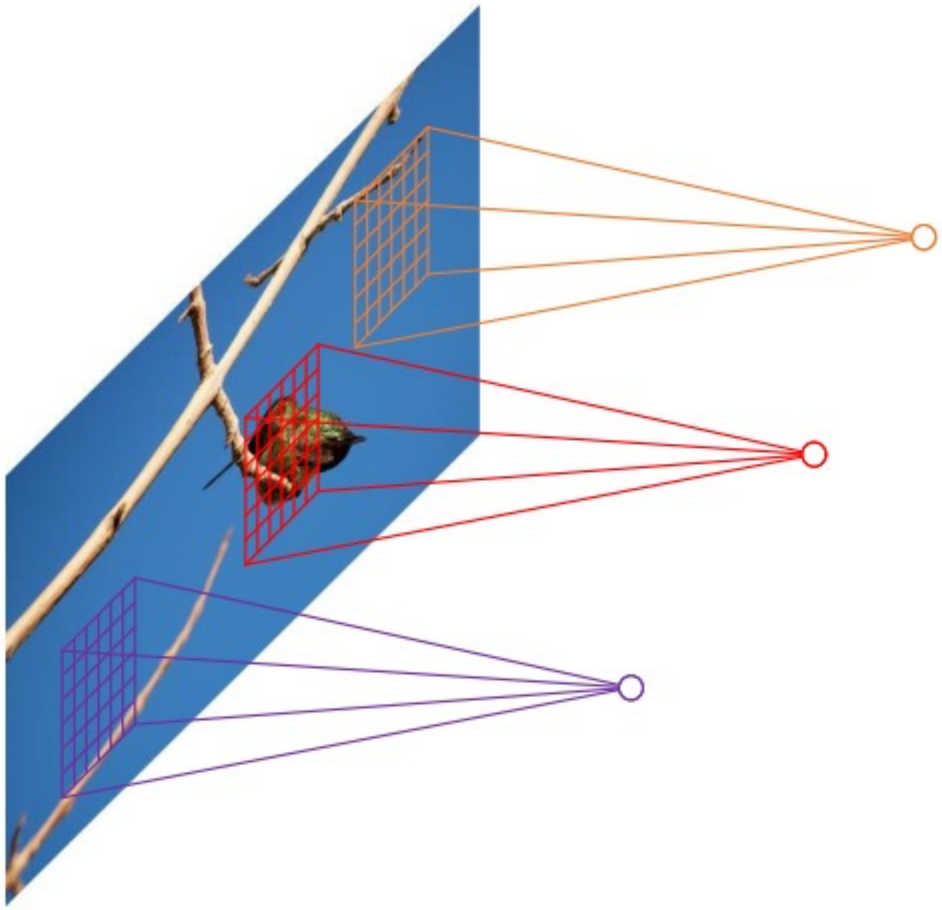
Object Detection

Chetan Arora

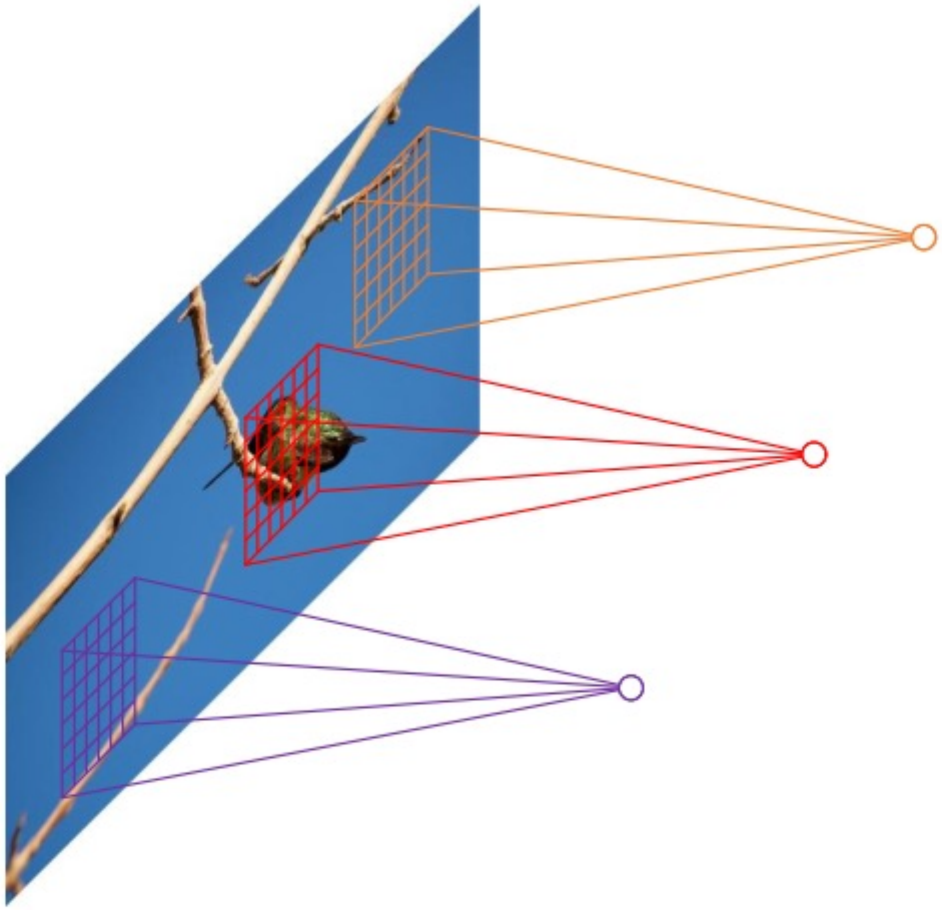
Convolutional Layers



Convolutional layers are locally connected

- a filter/kernel/window slides on the image or the previous map
- the position of the filter explicitly provides information for localizing
- local spatial information w.r.t. the window is encoded in the channels

Convolutional Layers

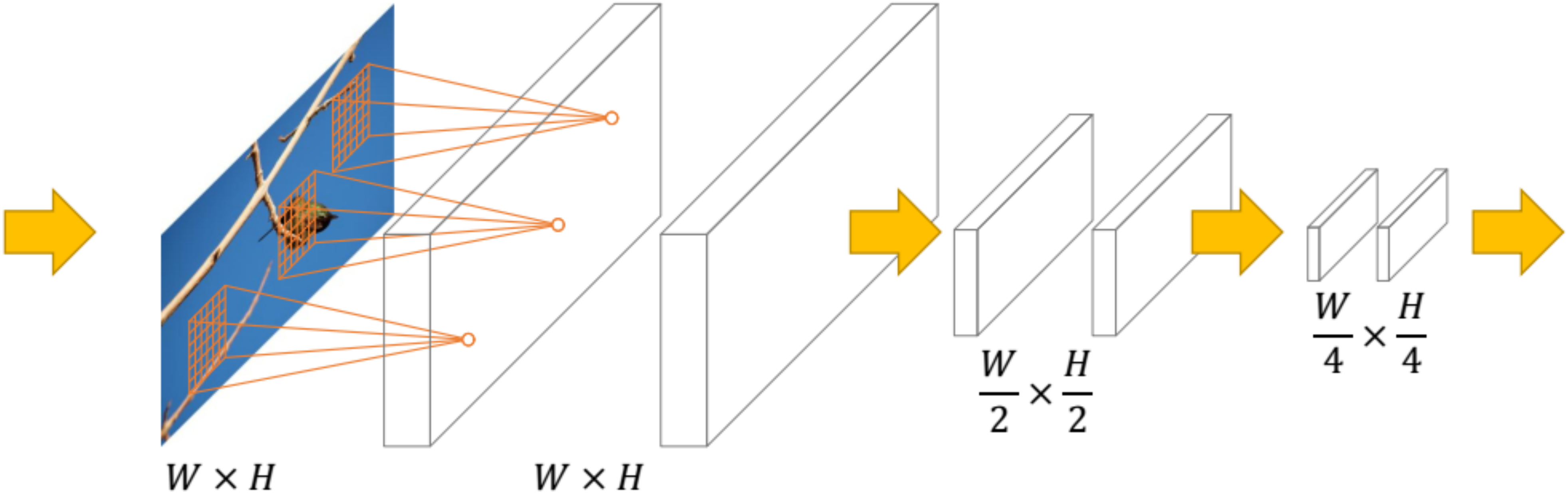


Convolutional layers share weights spatially: translation-invariant

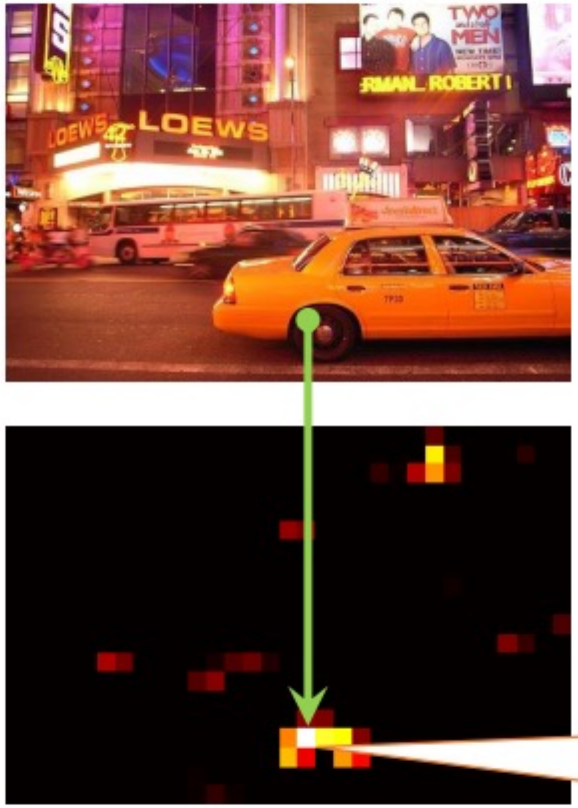
- Translation-invariant: a translated region will produce the same response at the correspondingly translated position
- A local pattern's convolutional response can be re-used by different candidate regions

Convolutional Layers

- Convolutional layers can be applied to images of any sizes, yielding proportionally-sized outputs



Feature Maps = features and their locations



one feature map of conv_5
(#55 in 256 channels of a model trained on ImageNet)

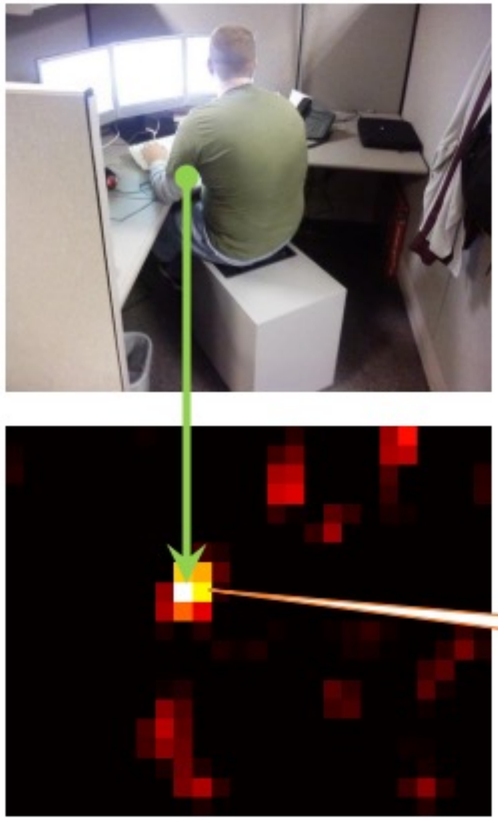
ImageNet images with strongest responses of this channel

Intuition of *this* response:
There is a "circle-shaped" object (likely a tire) at this position.

What **Where**

Slide Credit: Kaiming He

Feature Maps = features and their locations



one feature map of conv_5
(#66 in 256 channels of a model trained on ImageNet)

ImageNet images with strongest responses of this channel

Intuition of *this* response:
There is a “λ-shaped” object (likely an underarm) at this position.

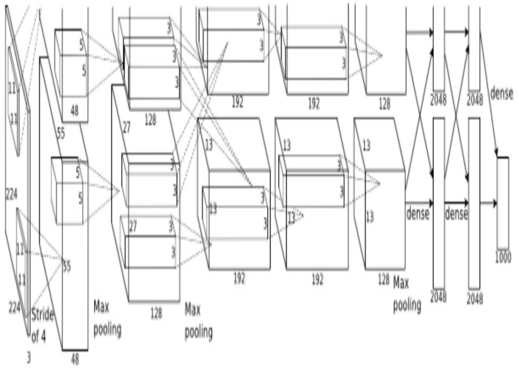
What **Where**

Slide Credit: Kaiming He

Classification + Localization

Cat

Localization as Regression Problem



Output of last fully connected layer: 4096 dim vector

4096 -> 1000

Class Scores

Cat: 0.9
Dog: 0.1
Car: 0.05

Correct Label

Softmax Loss

4096 -> 4

Box Coordinates

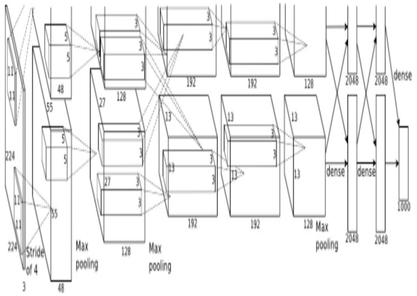
x, y, w, h

L2 Loss

Correct Box

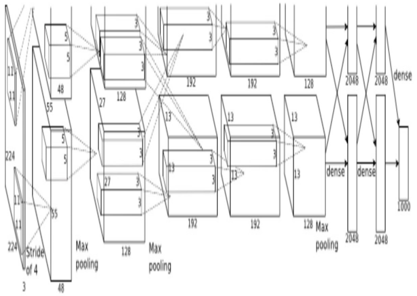
Total Loss

Localization as Regression: Problem?

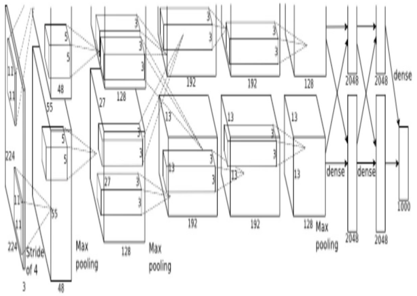


Cat: (x,y,w,h)

Localization as Regression: Problem?

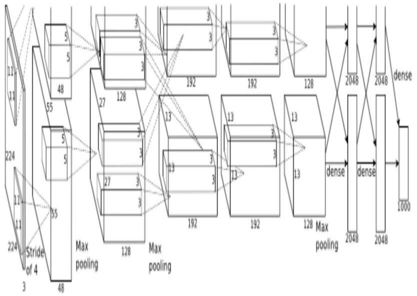


Cat: (x,y,w,h)

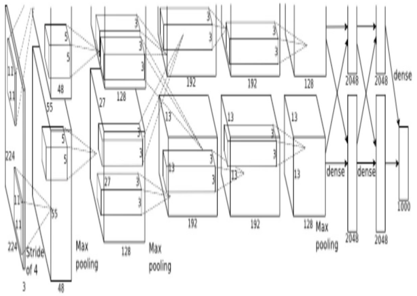


Dog: (x,y,w,h)
Cat: (x,y,w,h)

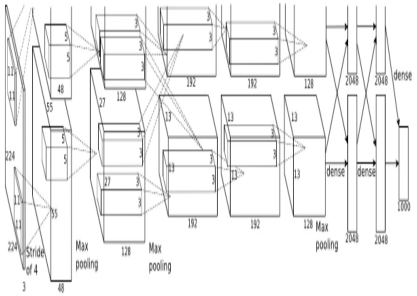
Localization as Regression: Problem?



Cat: (x,y,w,h)

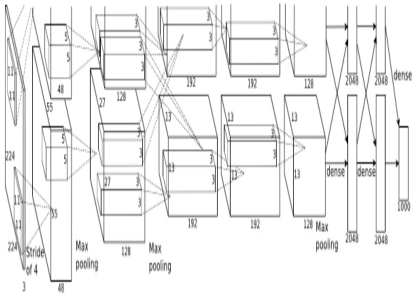


Dog: (x,y,w,h)
Cat: (x,y,w,h)

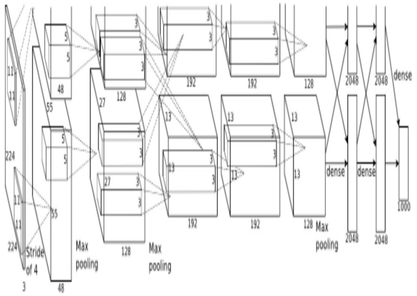


Duck: (x,y,w,h)
Duck: (x,y,w,h)
…

Localization as Regression: Problem?

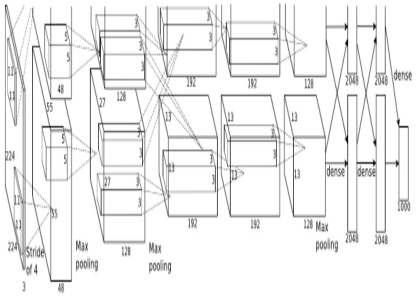


Cat: (x,y,w,h)



Dog: (x,y,w,h)
Cat: (x,y,w,h)

Each image needs a different number of outputs!

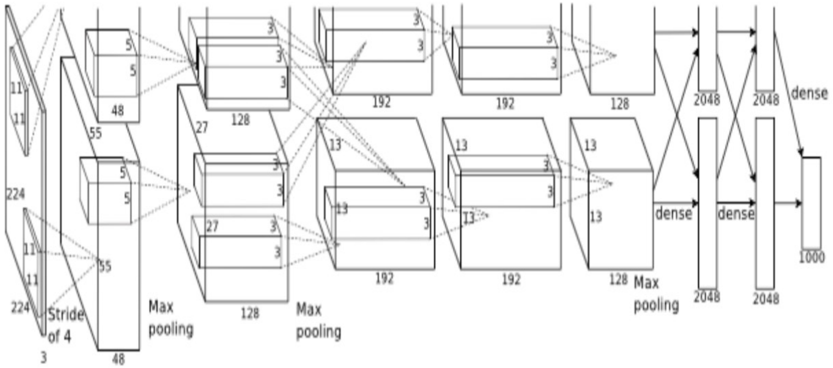


Duck: (x,y,w,h)
Duck: (x,y,w,h)
…

Chetan Arora
Computer Vision and Graphics Lab, IIT Delhi

Object Detection as Classification

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

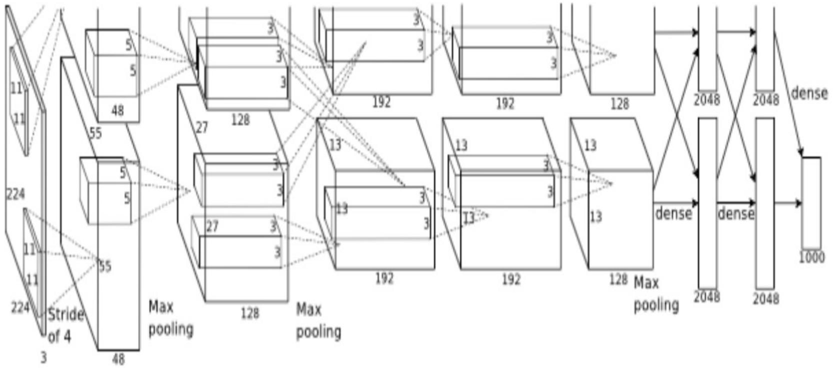


Dog: No
Cat: No
BG: Yes

Problem: Need to apply CNN to huge number of locations and scales, very computationally expensive.

Object Detection as Classification

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background

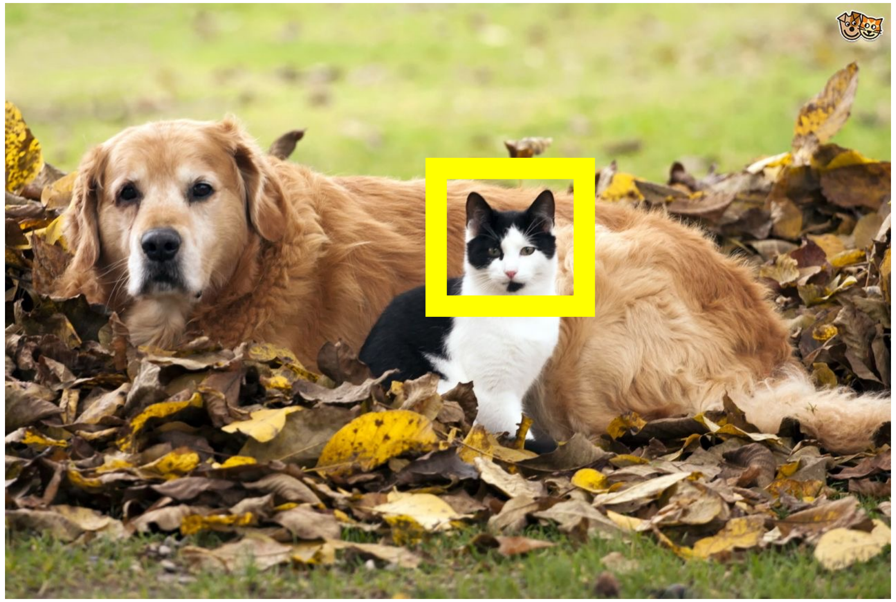


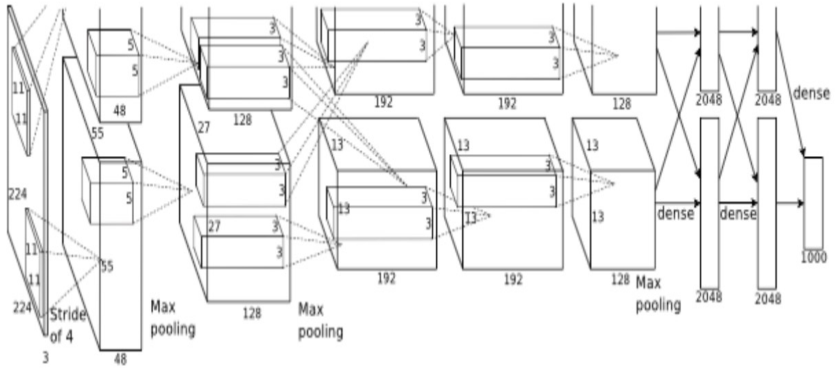
Dog: Yes
Cat: No
BG: No

Problem: Need to apply CNN to huge number of locations and scales, very computationally expensive.

Object Detection as Classification

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background





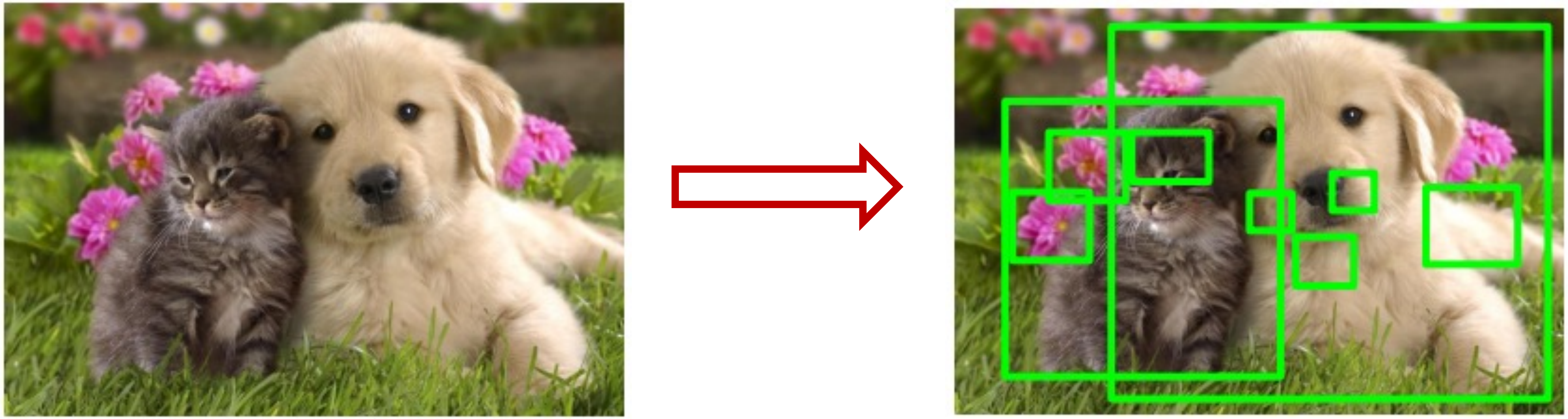
Dog: No
Cat: Yes
BG: No

Problem: Need to apply CNN to huge number of locations and scales, very computationally expensive.

Chetan Arora
Computer Vision and Graphics Lab, IIT Delhi

Region Proposals: “Objectness” Detection

- Find “blobby” image regions that are likely to contain object
- Relatively fast to run; e.g. Selective Search gives 1000 region proposals in a few seconds on CPU



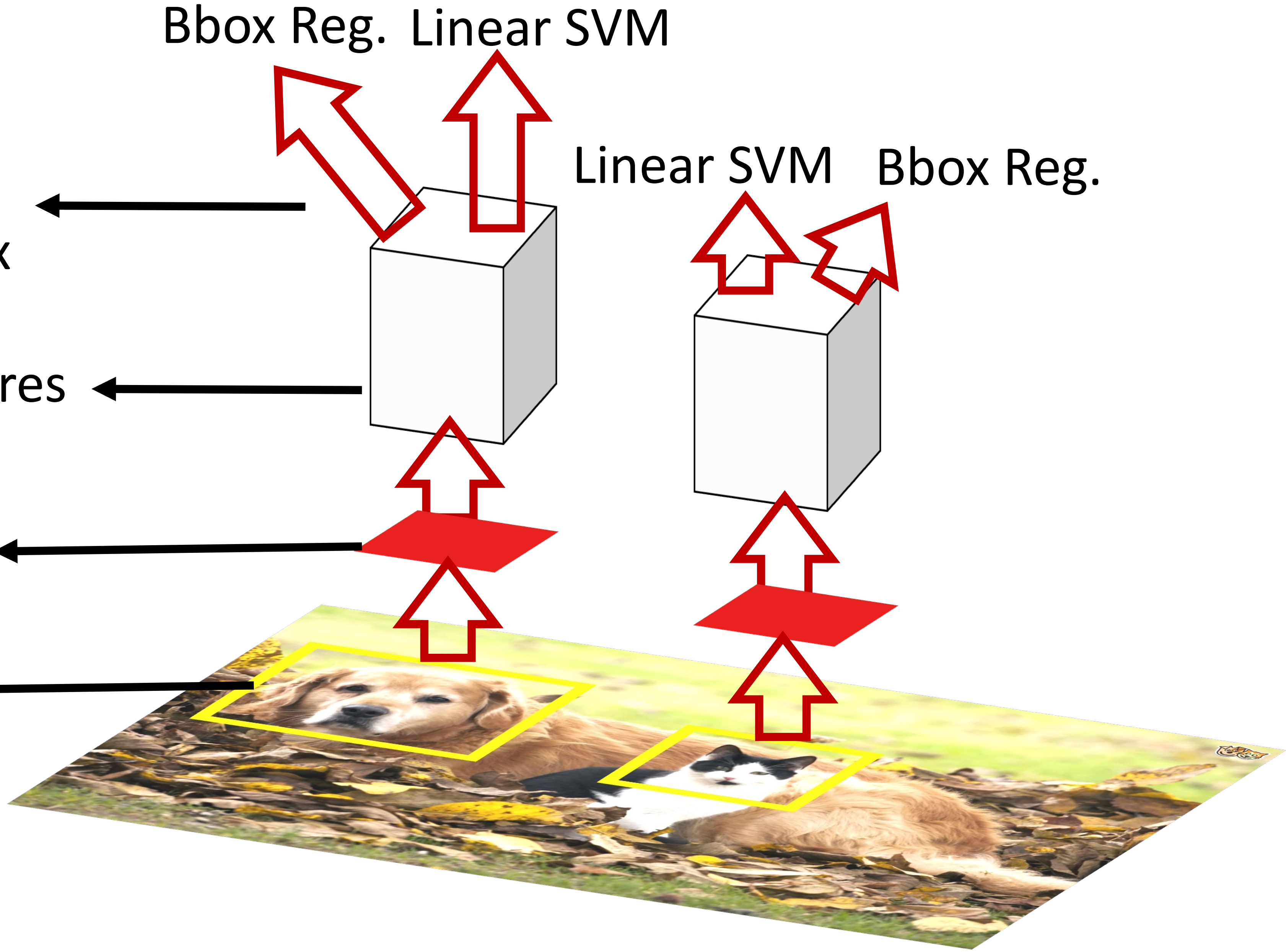
RCNN

Classify regions, and
linear regression for bbox

Find Convolutional Features

Resize to a standard size

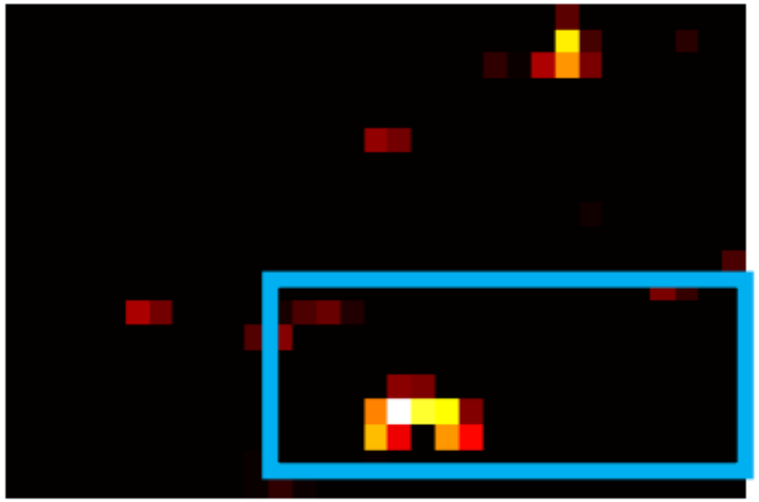
Region Proposals



Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik
Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR 2014

Regions on Feature Maps

- Compute convolutional feature maps on the entire image only once.
- Project an image region to a feature map region (using correspondence of the receptive field center)
- Extract a region-based feature from the feature map region…

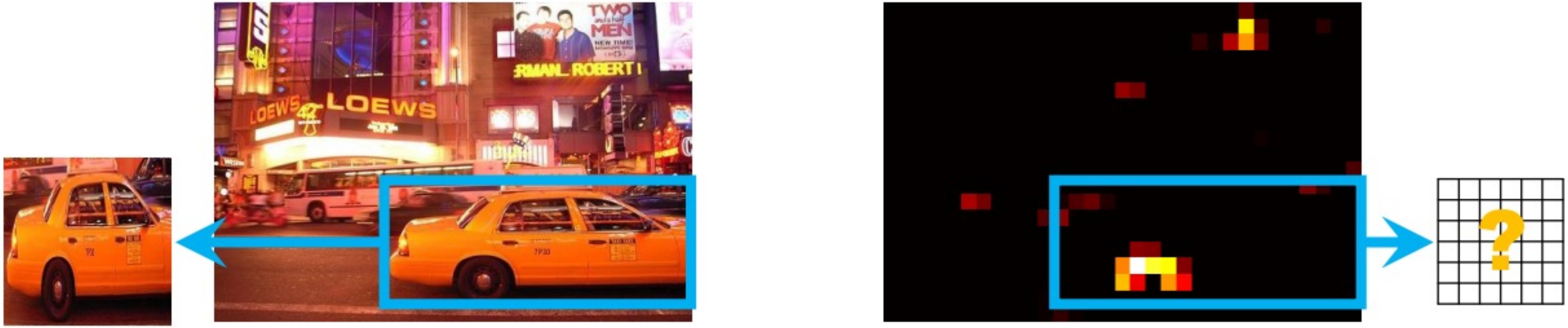


Slide Credit: Kaiming He

Regions on Feature Maps

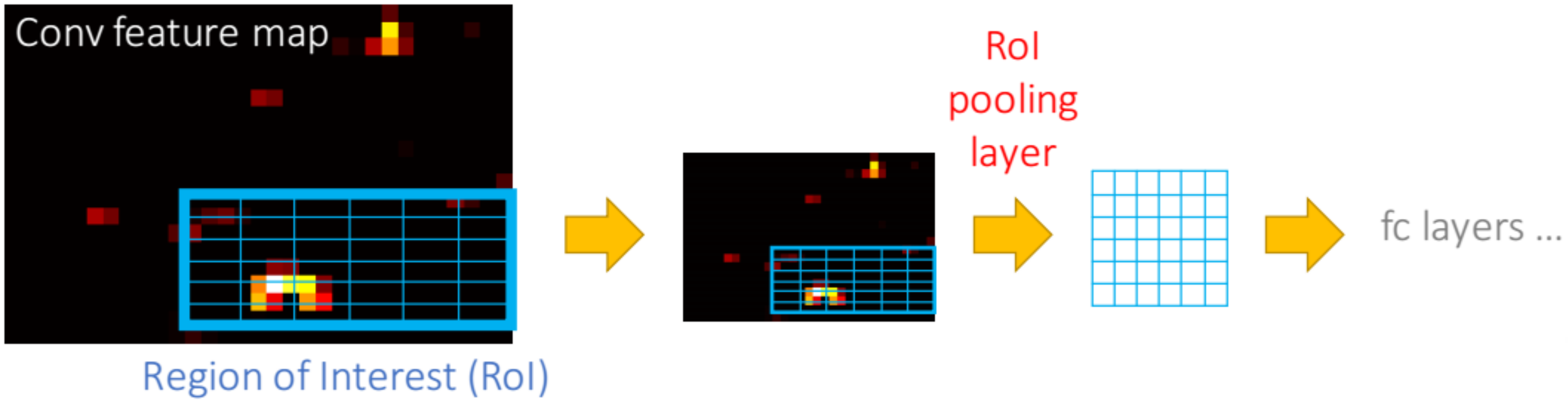
- Fixed-length features are required by fully-connected layers or SVM

- But how to produce a fixed-length feature from a feature map region?



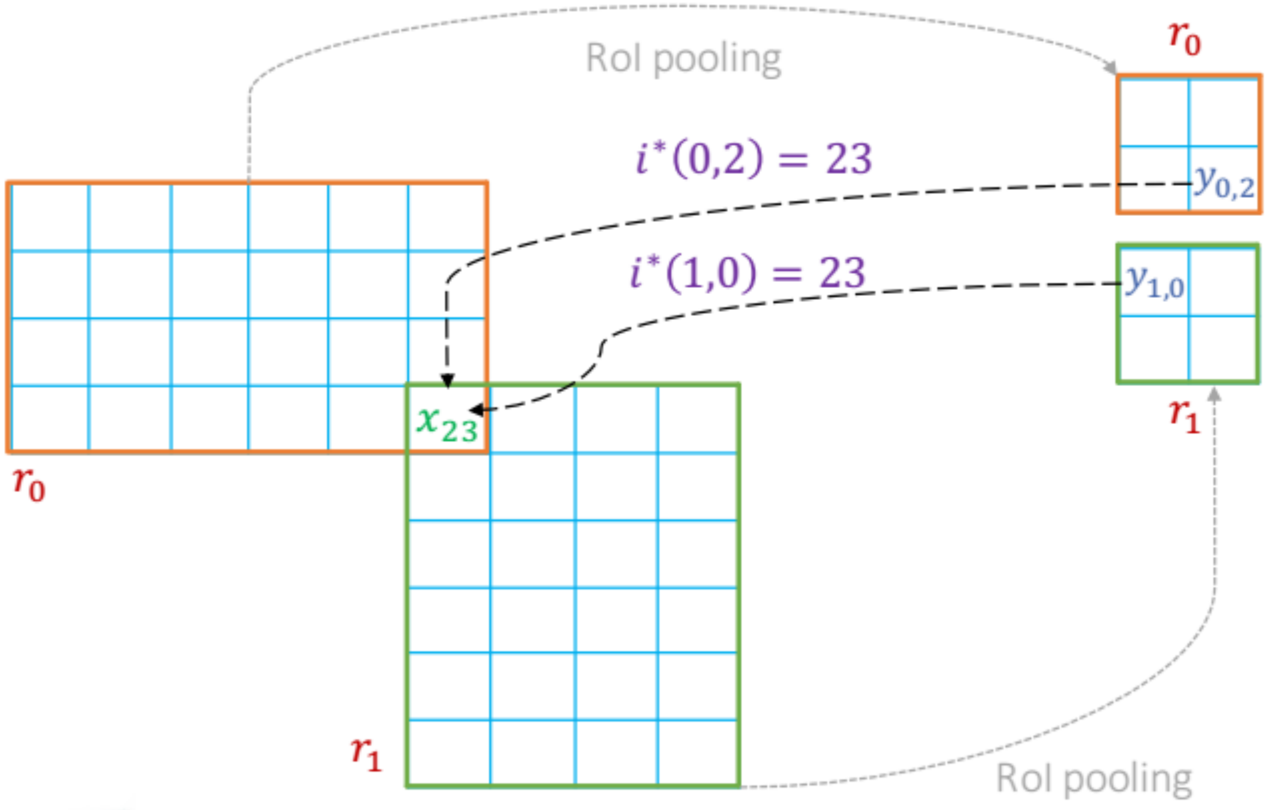
Slide Credit: Kaiming He

ROI Pooling Layer



Differentiable ROI Pooling

RoI pooling / SPP is just like max pooling, except that pooling regions overlap



1 if r, j "pooled" input i; 0 o/w

$$\frac{\partial L}{\partial x_i} = \sum_r \sum_j [i = i^*(r,j)] \frac{\partial L}{\partial y_{rj}}$$

Partial for x_i

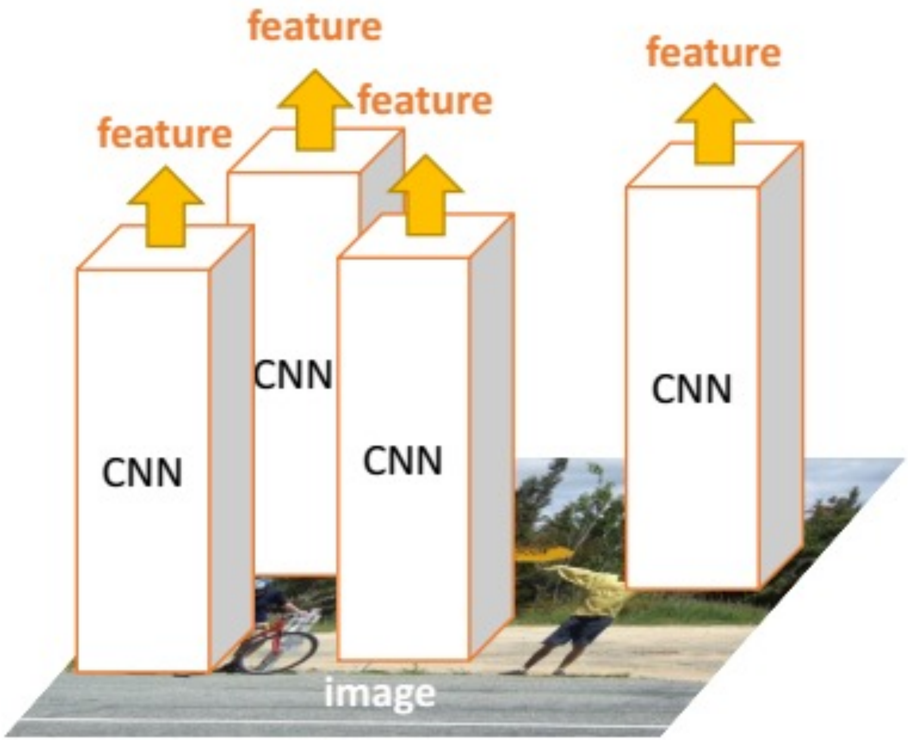
Over regions r, locations j

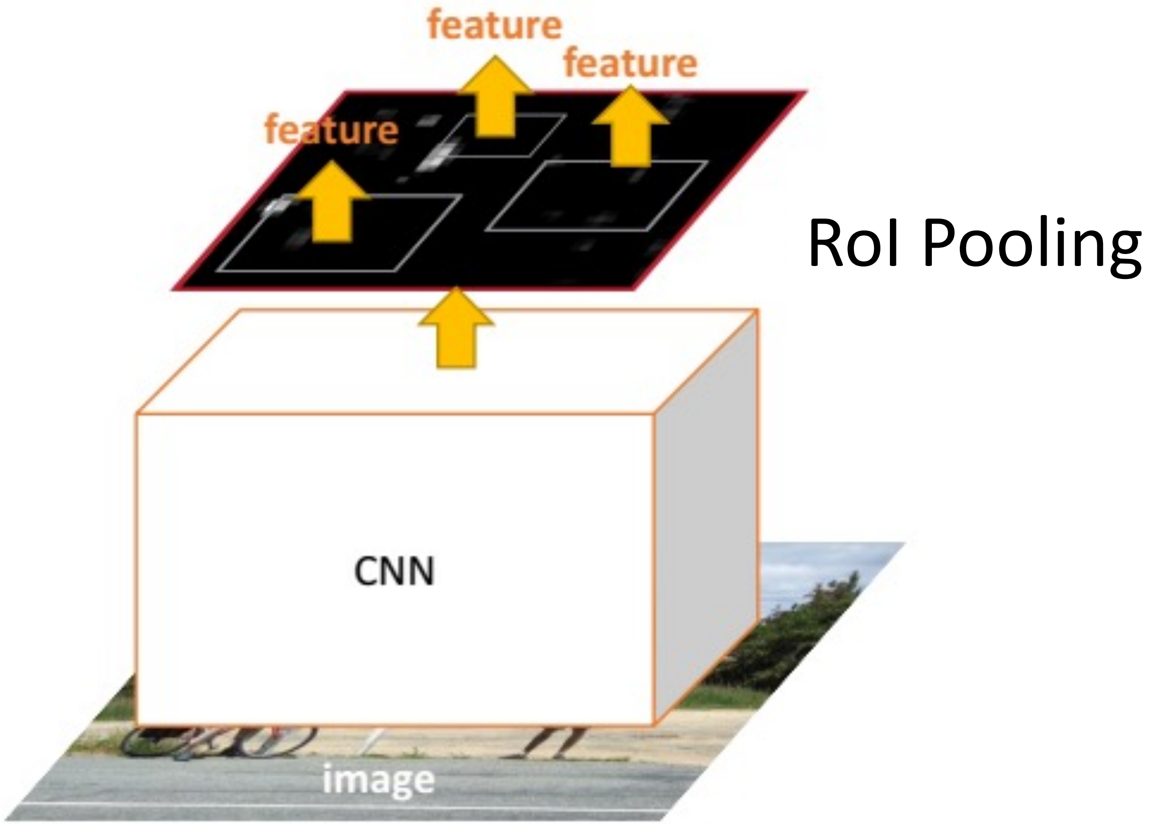
Partial from next layer

←--- max pooling "switch" (*i.e.* argmax back-pointer)

Slide Credit: Kaiming He

R-CNN vs. Fast R-CNN





R-CNN

- Extract image regions
- Classify using CNN inference.
- 1 CNN inference per proposal (2000 proposals)

Fast R-CNN

- 1 CNN on the entire region
- Extract features from feature map regions.
- Classify region based features

Slide Credit: Kaiming He

Region Proposal from Feature Maps

- Object detection networks are fast (0.2s)…

- But what about region proposal?
  - Selective Search [Uijlings et al. ICCV 2011]: 2s per image
  - EdgeBoxes [Zitnick & Dollar. ECCV 2014]: 0.2s per image

- Can we do region proposal on the same set of feature maps?

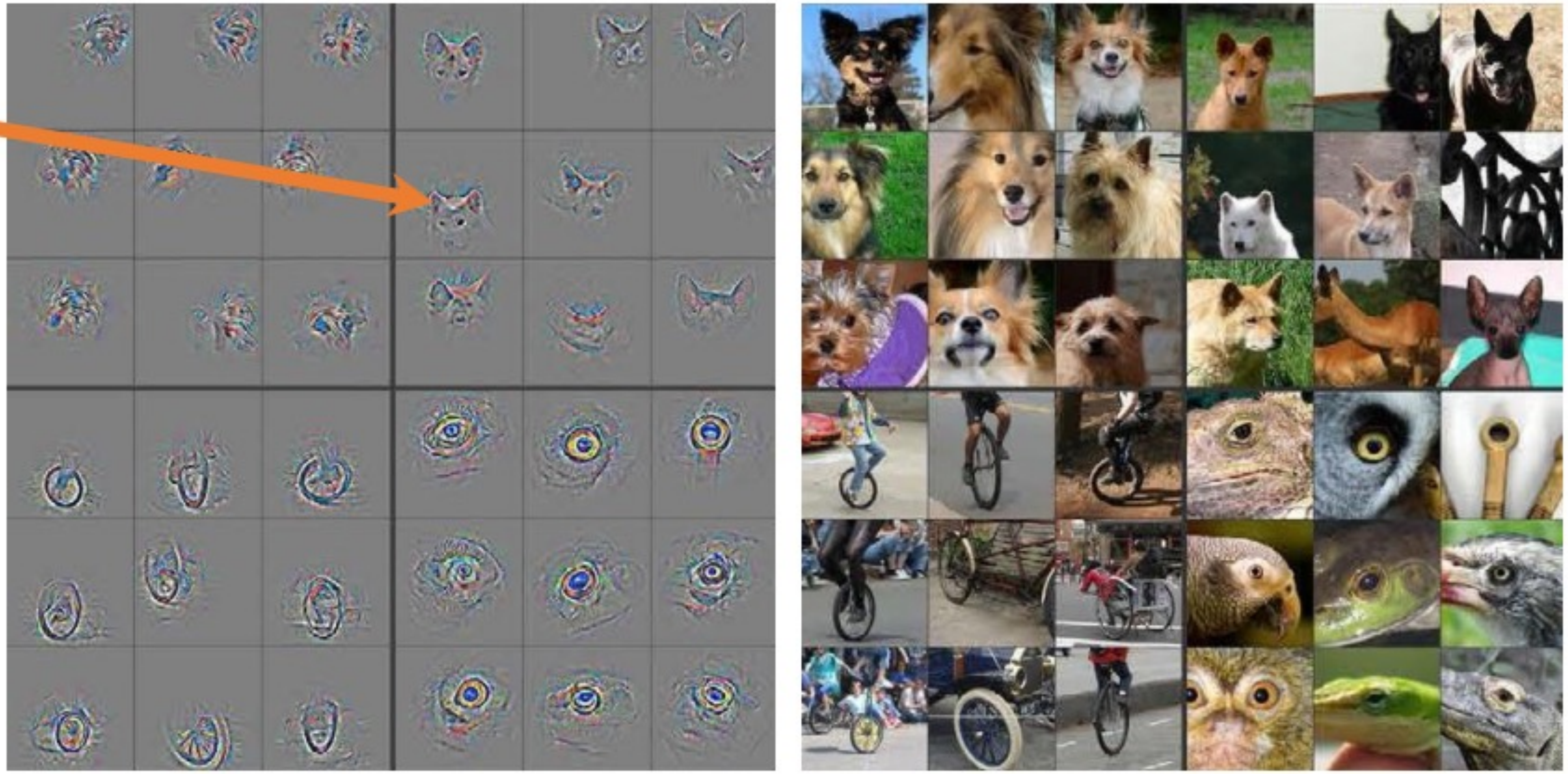
Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun
Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS 2015

Region Proposal from Feature Maps

- By decoding one response at a single pixel, we can still roughly see the object outline*
- Finer localization information has been encoded in the channels of a convolutional feature response
- Extract this information for better localization…

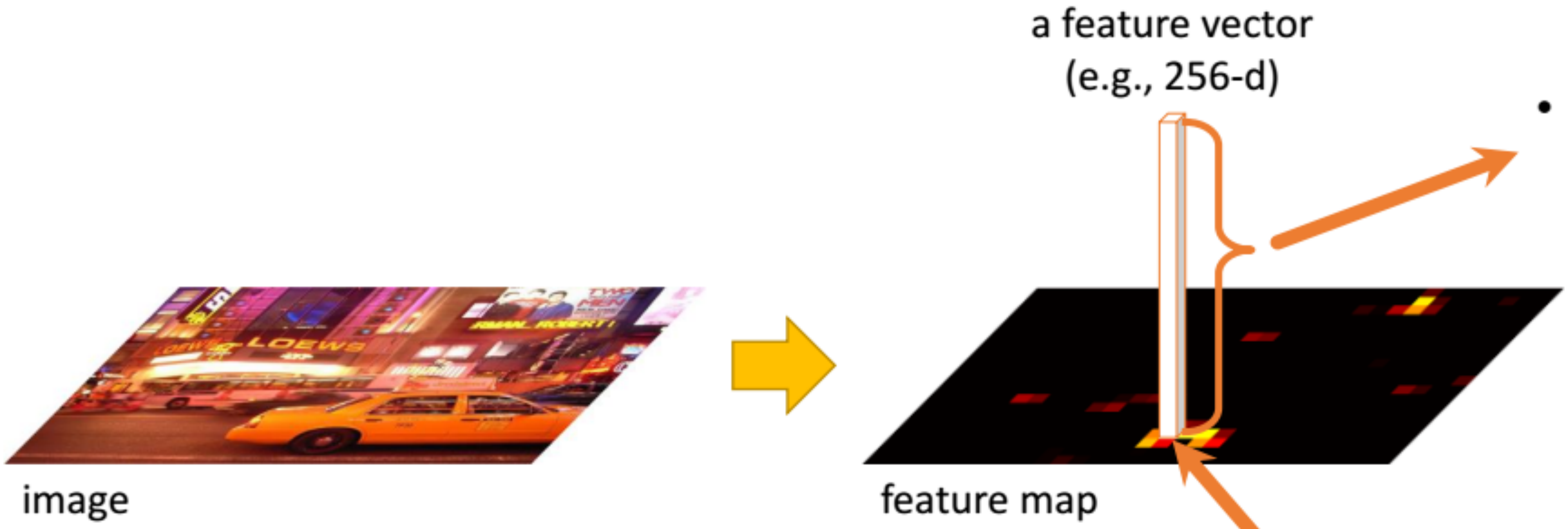
* Zeiler & Fergus's method traces unpooling information so the visualization involves more than a single response. But other visualization methods reveal similar patterns.

Revisiting visualizations from Zeiler & Fergus



Slide Credit: Kaiming He

Region Proposal from Feature Maps

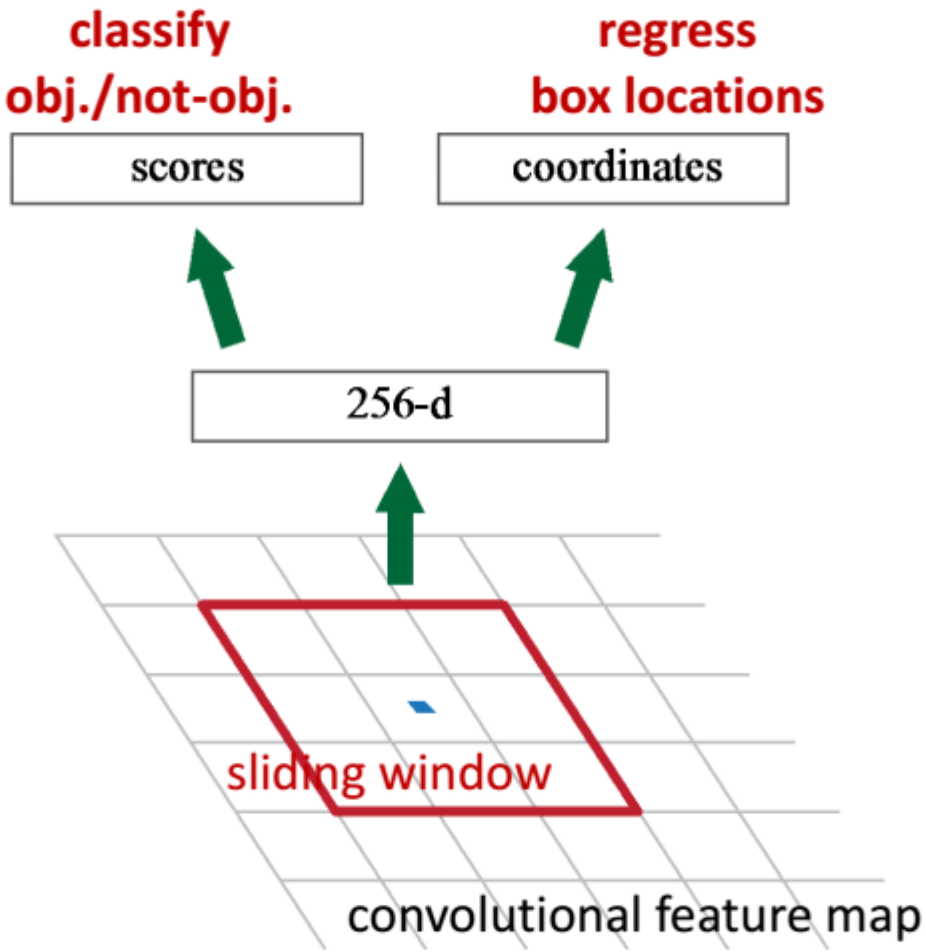


- The channels of this feature vector encodes finer localization information

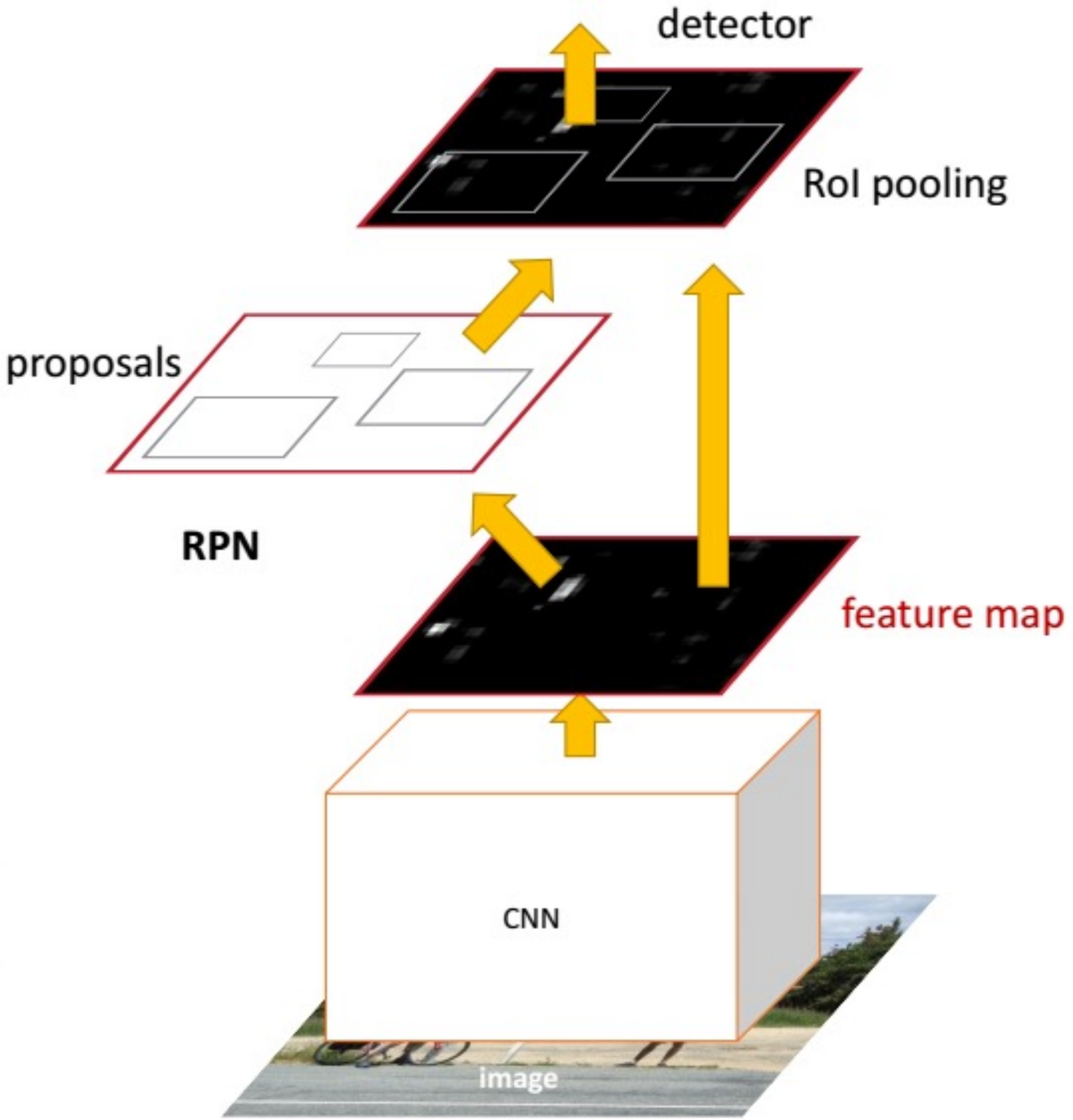
- The spatial position of this feature provides coarse locations

Region Proposal Network

- Slide a small window on the feature map
- Build a small network for:
  - classifying object or not-object, and
  - regressing bbox locations
- Position of the sliding window provides localization information with reference to the image
- Box regression provides finer localization information with reference to this sliding window

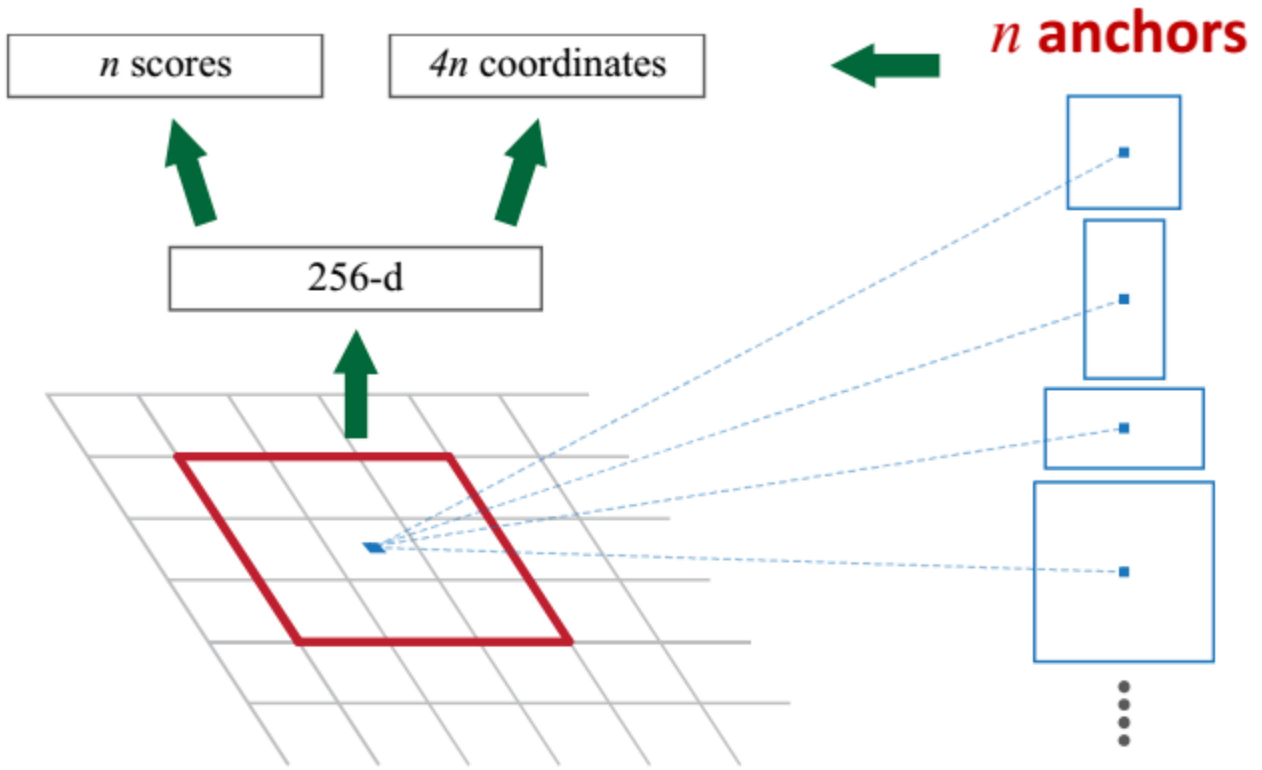


Faster RCNN



Anchors: Pre-defined Reference Boxes

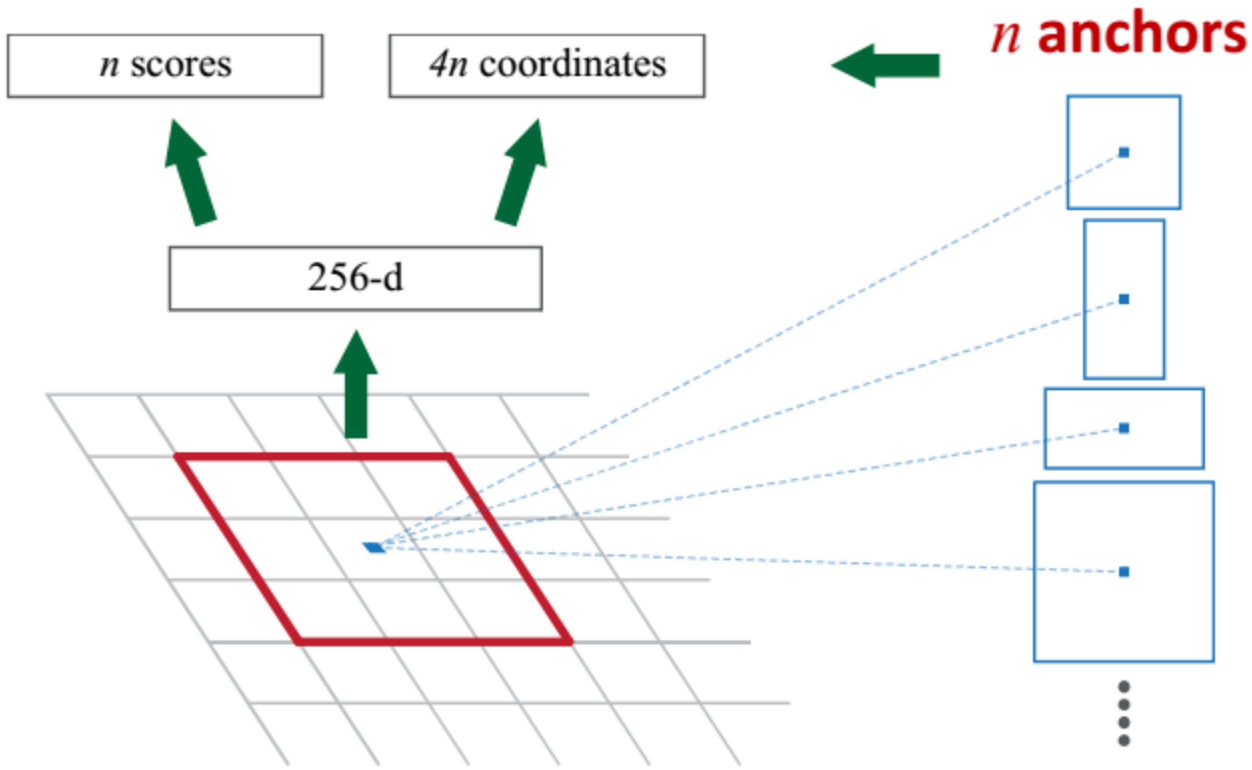
- Box regression is with reference to anchors: regressing an anchor box to a ground-truth box
- Object probability is with reference to anchors, e.g.:
  - anchors as positive samples: if IoU > 0.7, or IoU is max (encourages anchor specialists)
  - anchors as negative samples: if IoU < 0.3



Anchors: Pre-defined Reference Boxes

Multi-scale/size anchors:

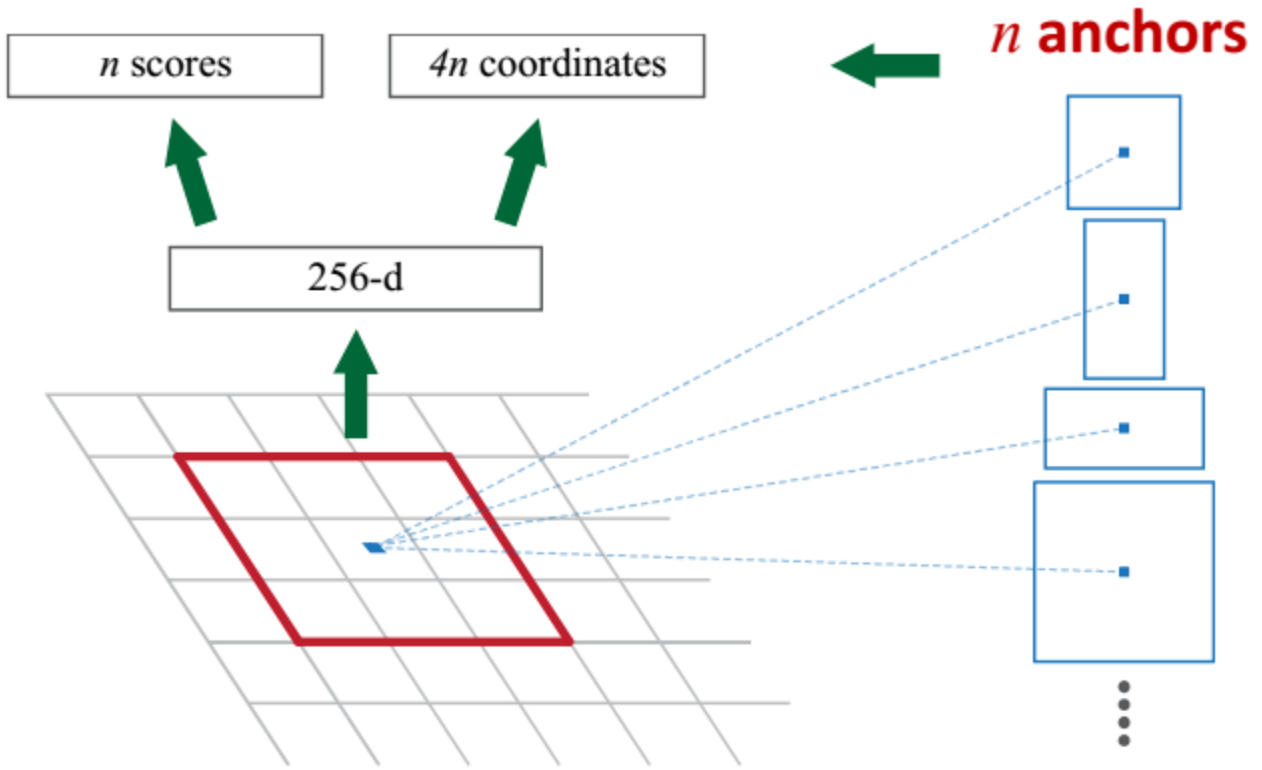
- multiple anchors are used at each position: e.g., 3 scales $(128^2, 256^2, 512^2)$ and 3 aspect ratios (2:1, 1:1, 1:2) yield 9 anchors
- each anchor has its own prediction function
- single-scale features, multi-scale predictions



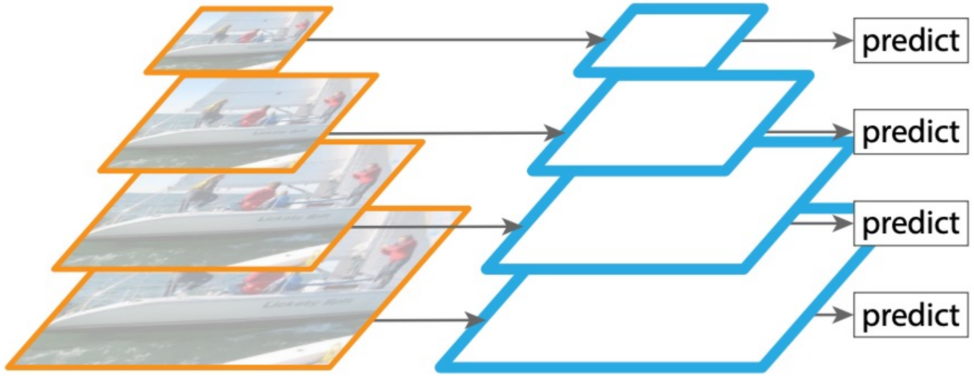
Anchors: Pre-defined Reference Boxes

Translation-invariant anchors:

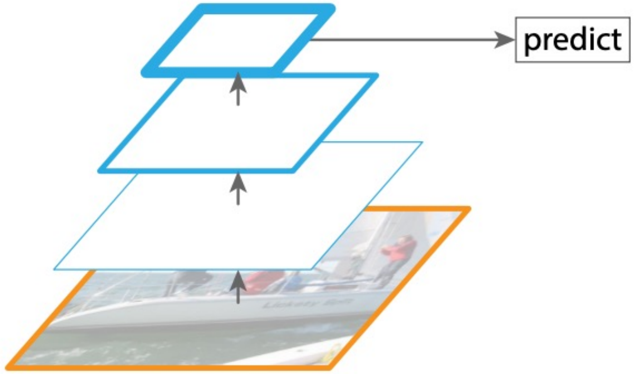
- the same set of anchors are used at each sliding position
- the same prediction functions (with reference to the sliding window) are used
- a translated object will have a translated prediction



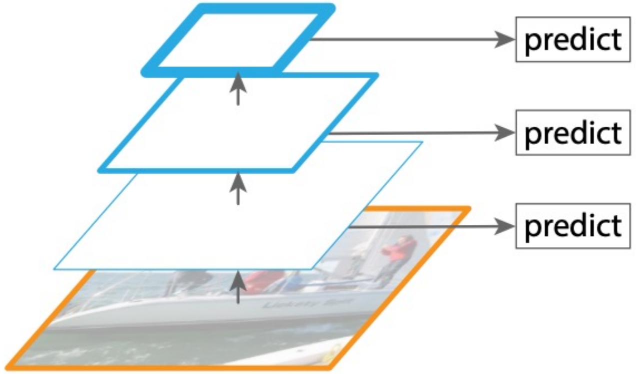
Pyramid Strategies for Object Detection



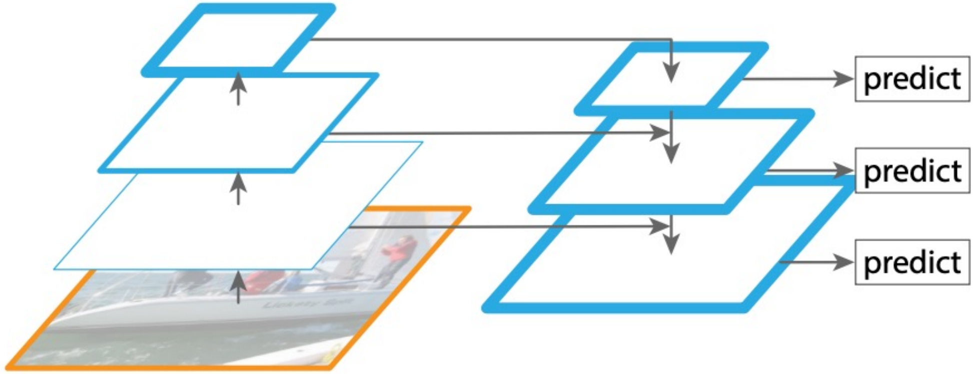
(a) Featurized image pyramid



(b) Single feature map



(c) Pyramidal feature hierarchy

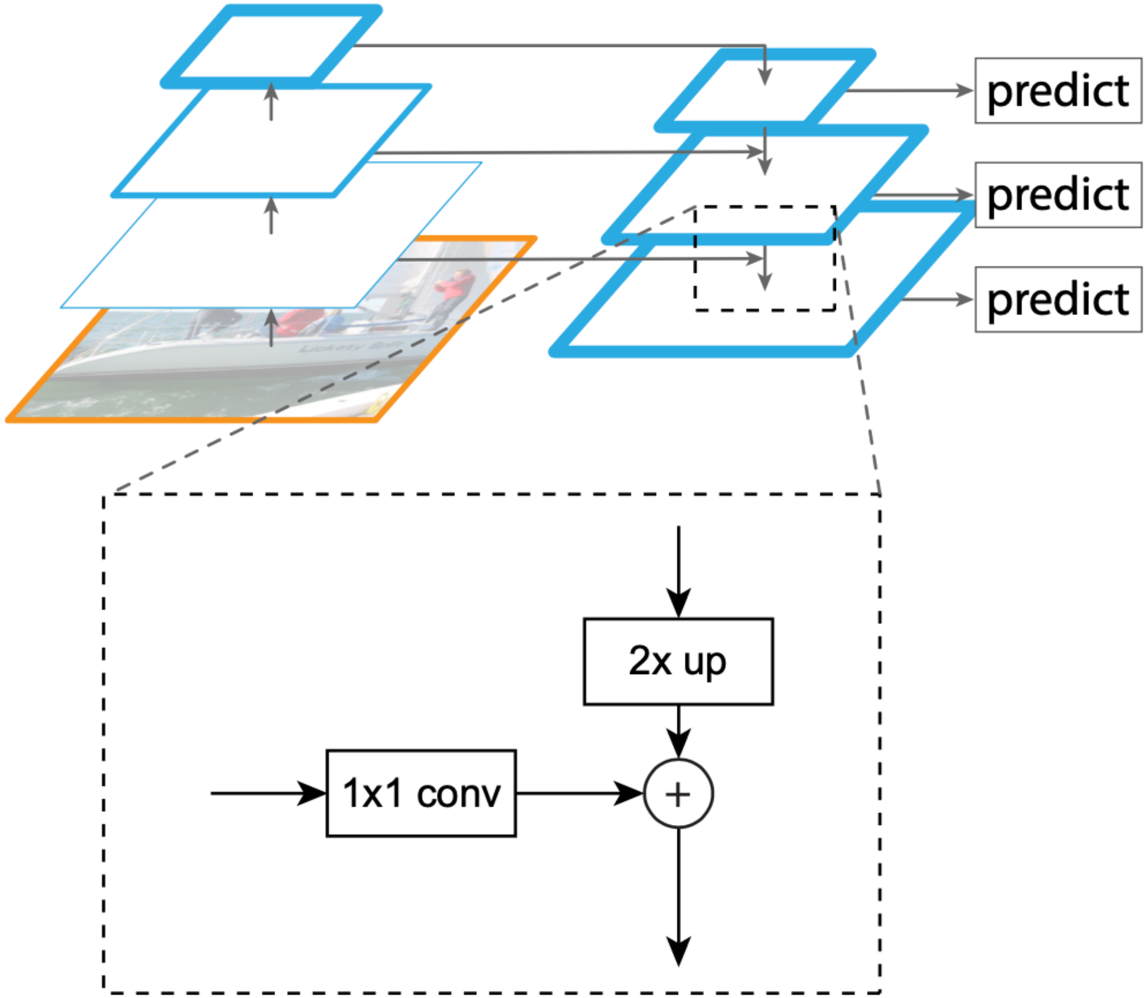


(d) Feature Pyramid Network

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie
Feature Pyramid Networks for Object Detection. CVPR 2017

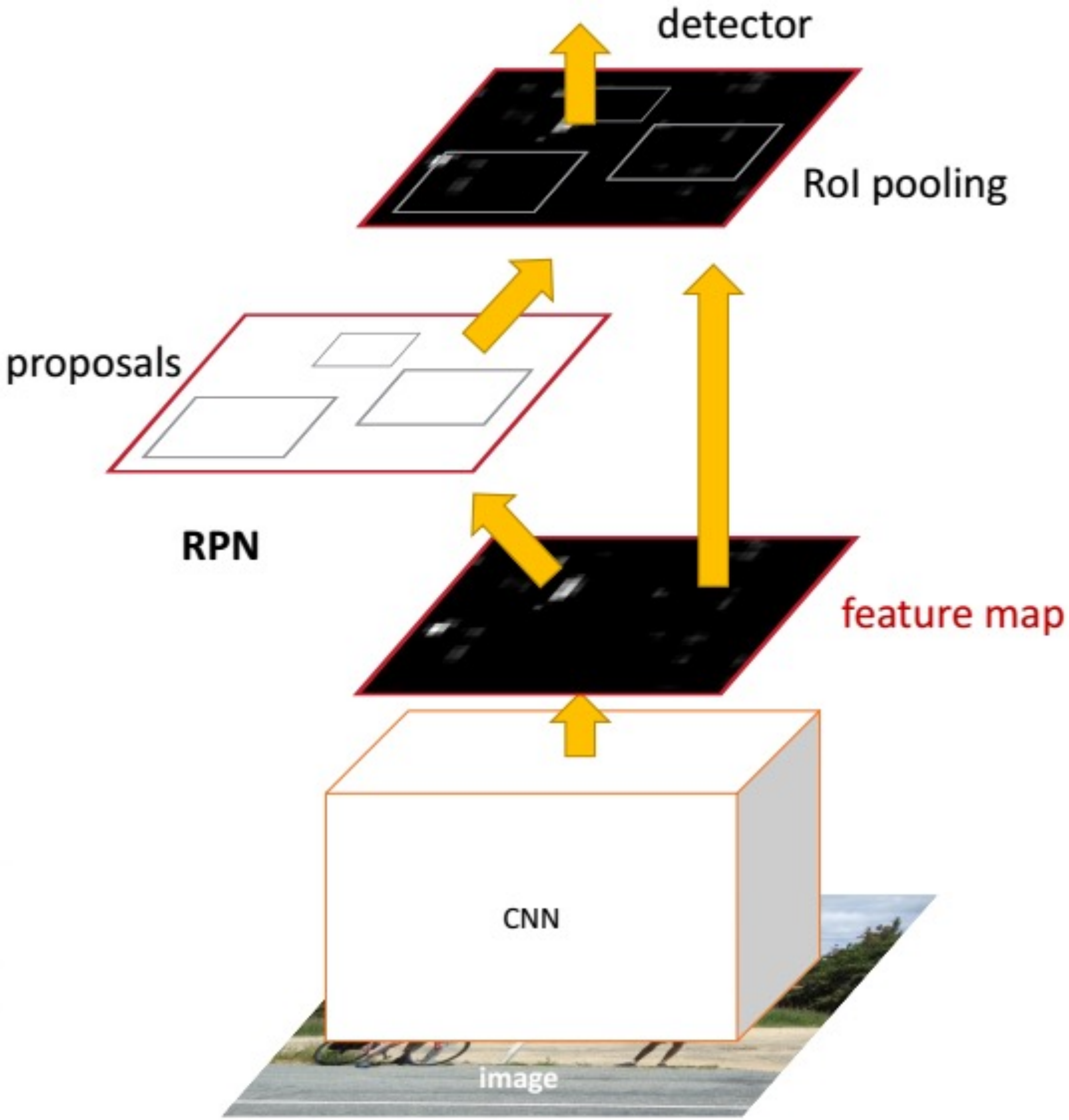
Feature Pyramid Networks

- Combine low-resolution, semantically strong features with high-resolution, semantically weak features via a top-down pathway and lateral connections



Chetan Arora
Computer Vision and Graphics Lab, IIT Delhi

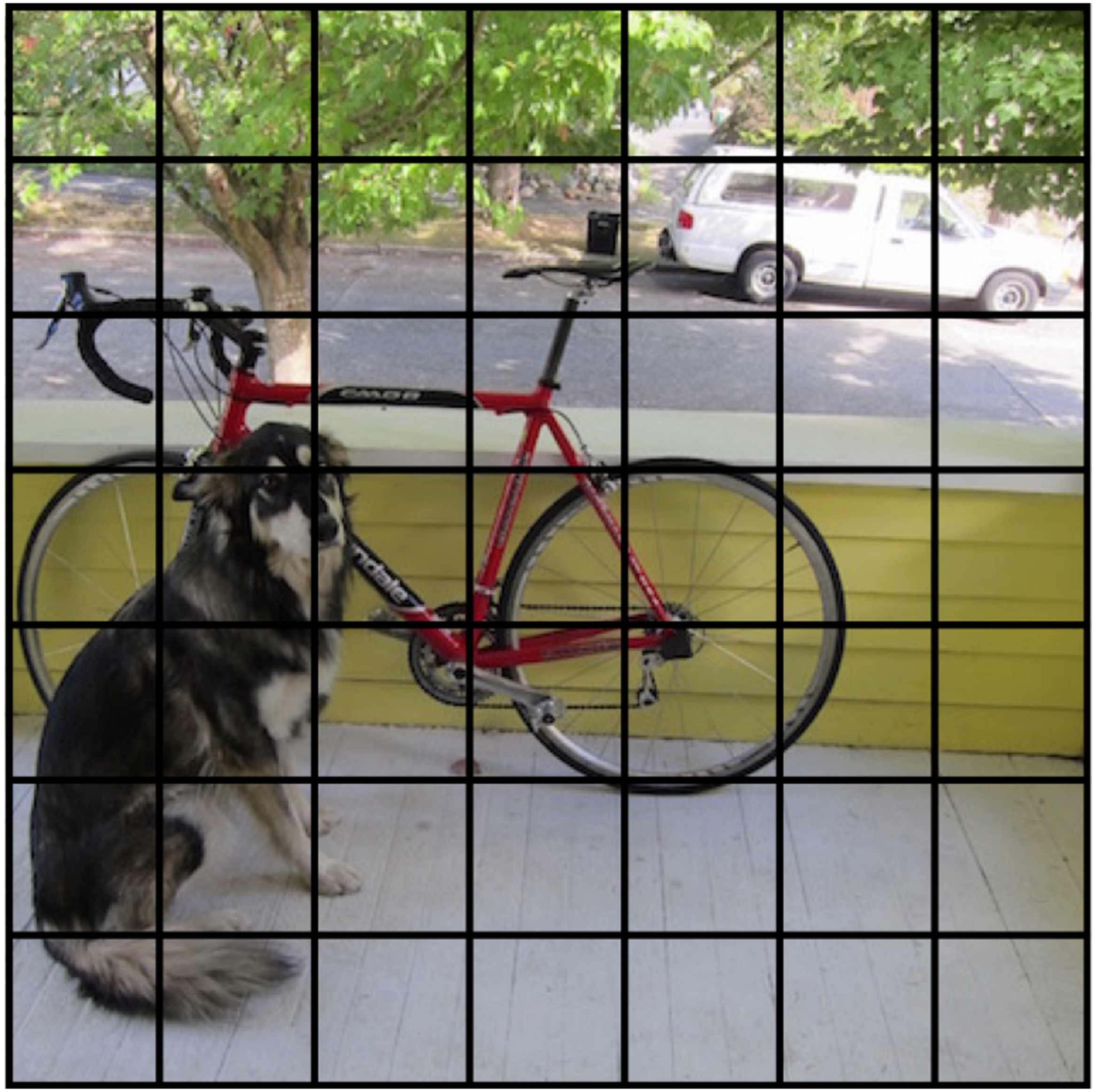
Recap: Faster RCNN



YOLO

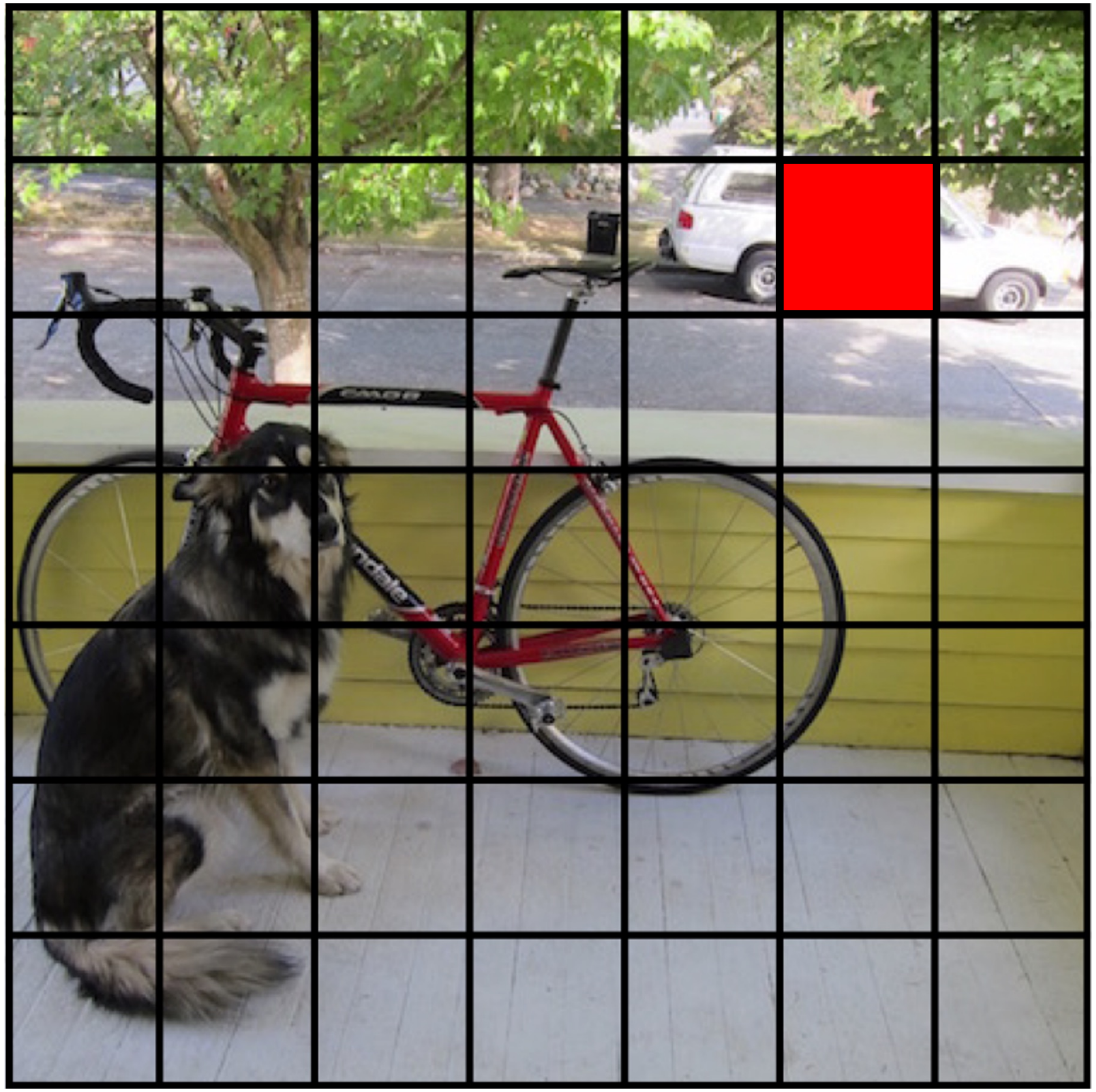
YOLO

- Split the image into a grid



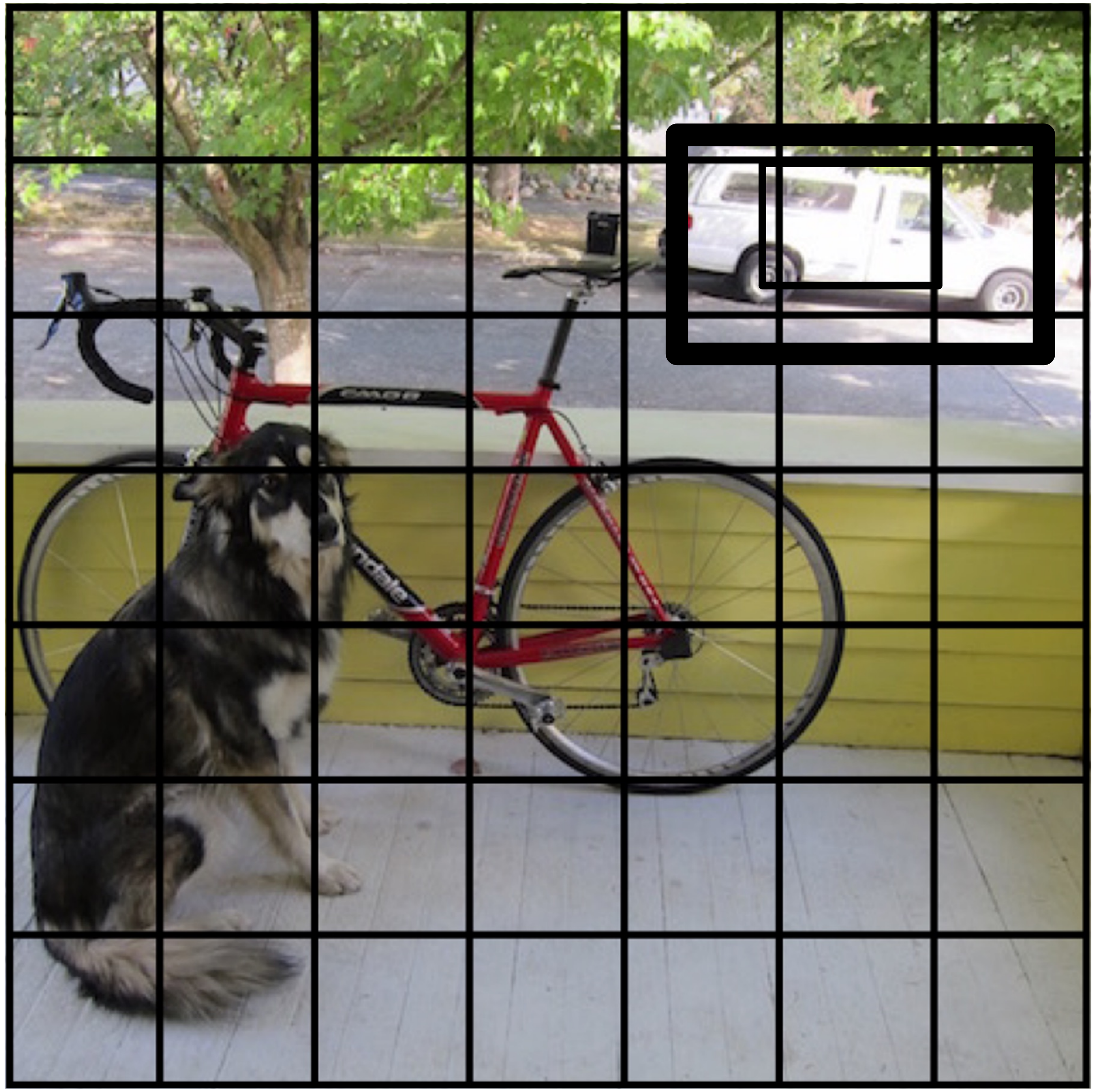
YOLO

- Each cell predicts boxes and confidences: P(Object)



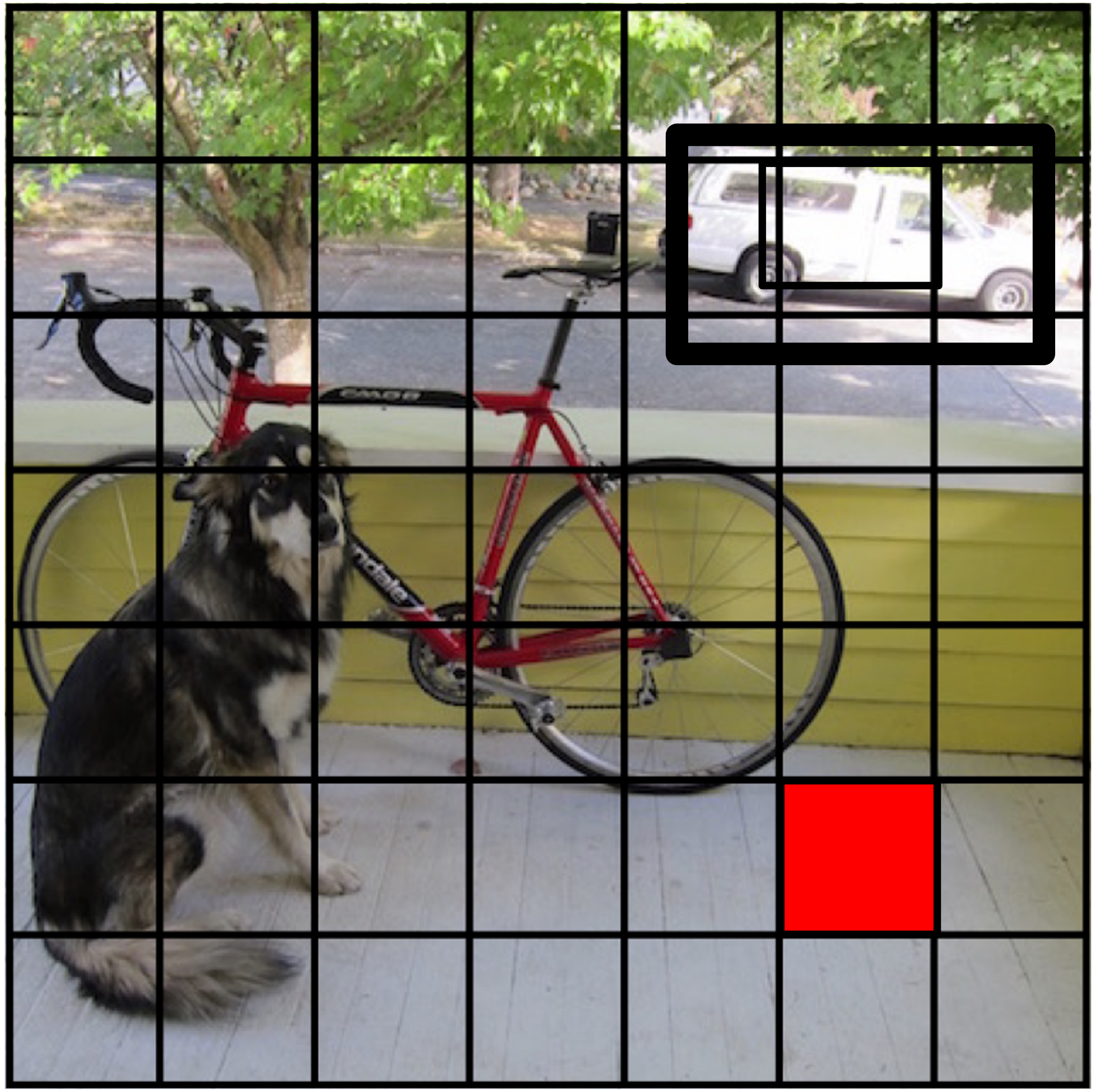
YOLO

- Each cell predicts boxes and confidences: P(Object)



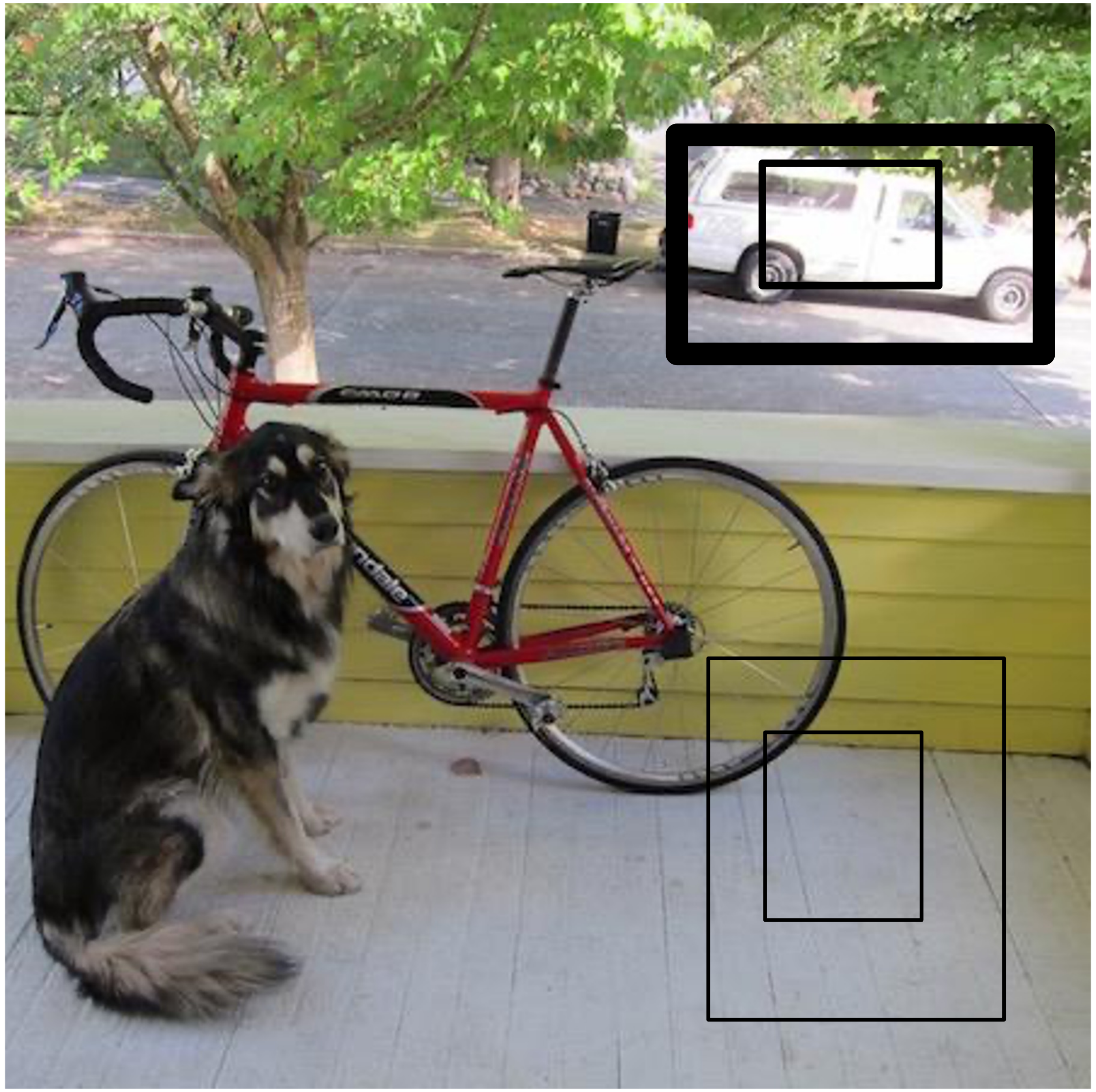
YOLO

- Each cell predicts boxes and confidences: P(Object)



YOLO

- Each cell predicts boxes and confidences: P(Object)

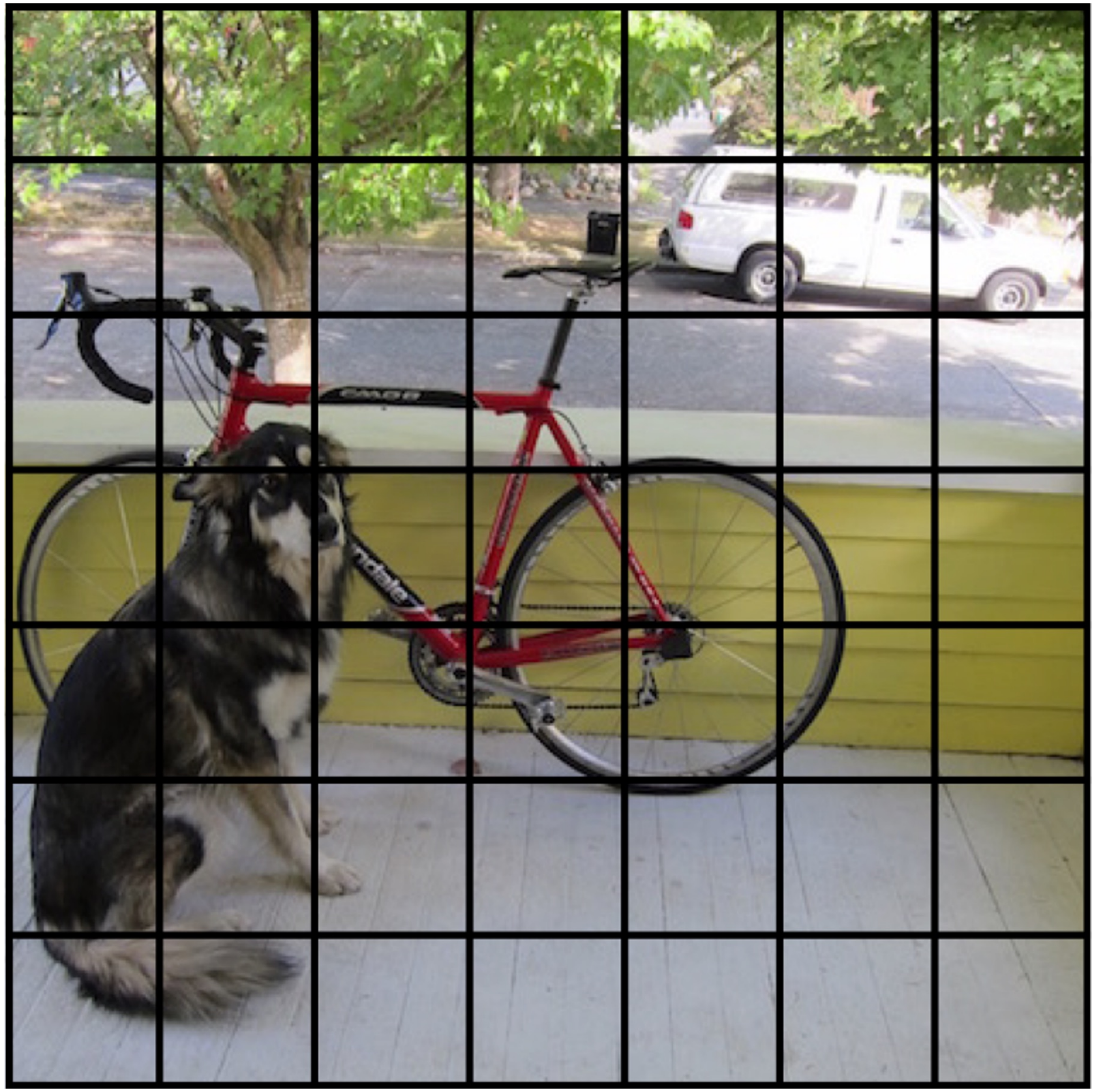


YOLO

- Each cell predicts boxes and confidences: P(Object)

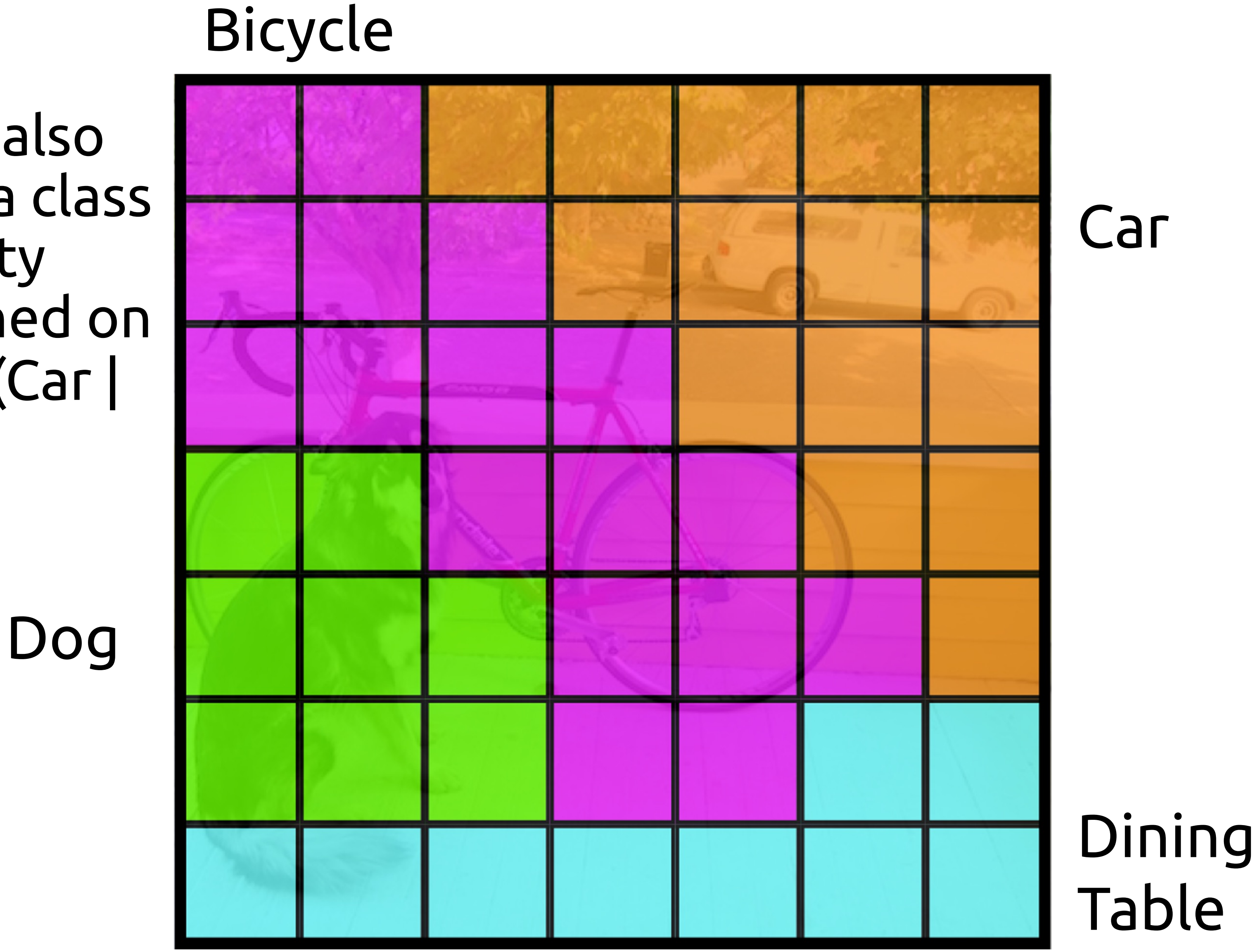
YOLO

- Each cell also predicts a class probability.



YOLO

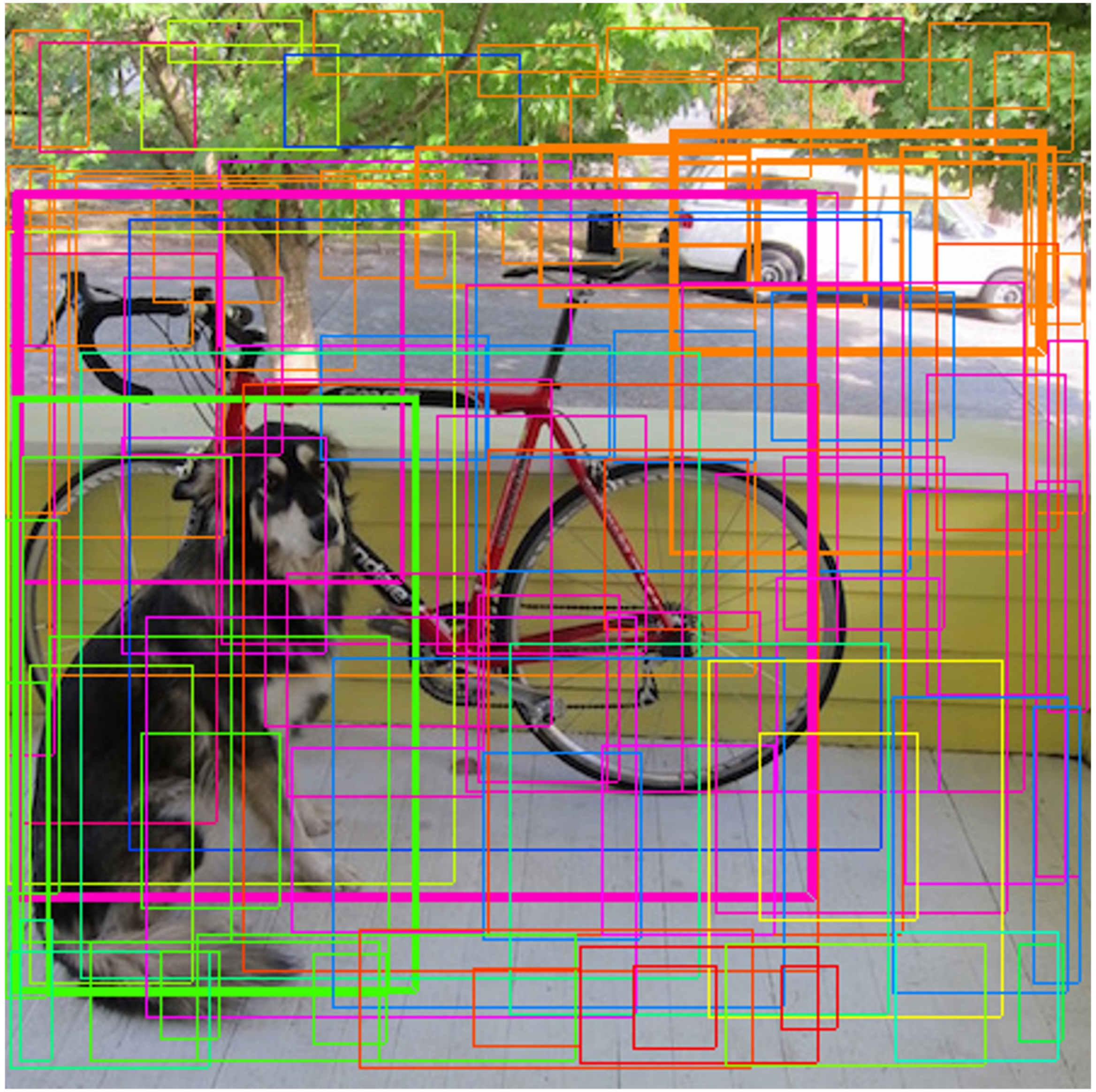
- Each cell also predicts a class probability conditioned on object: P(Car | Object)



Chetan Arora
Computer Vision and Graphics Lab, IIT Delhi

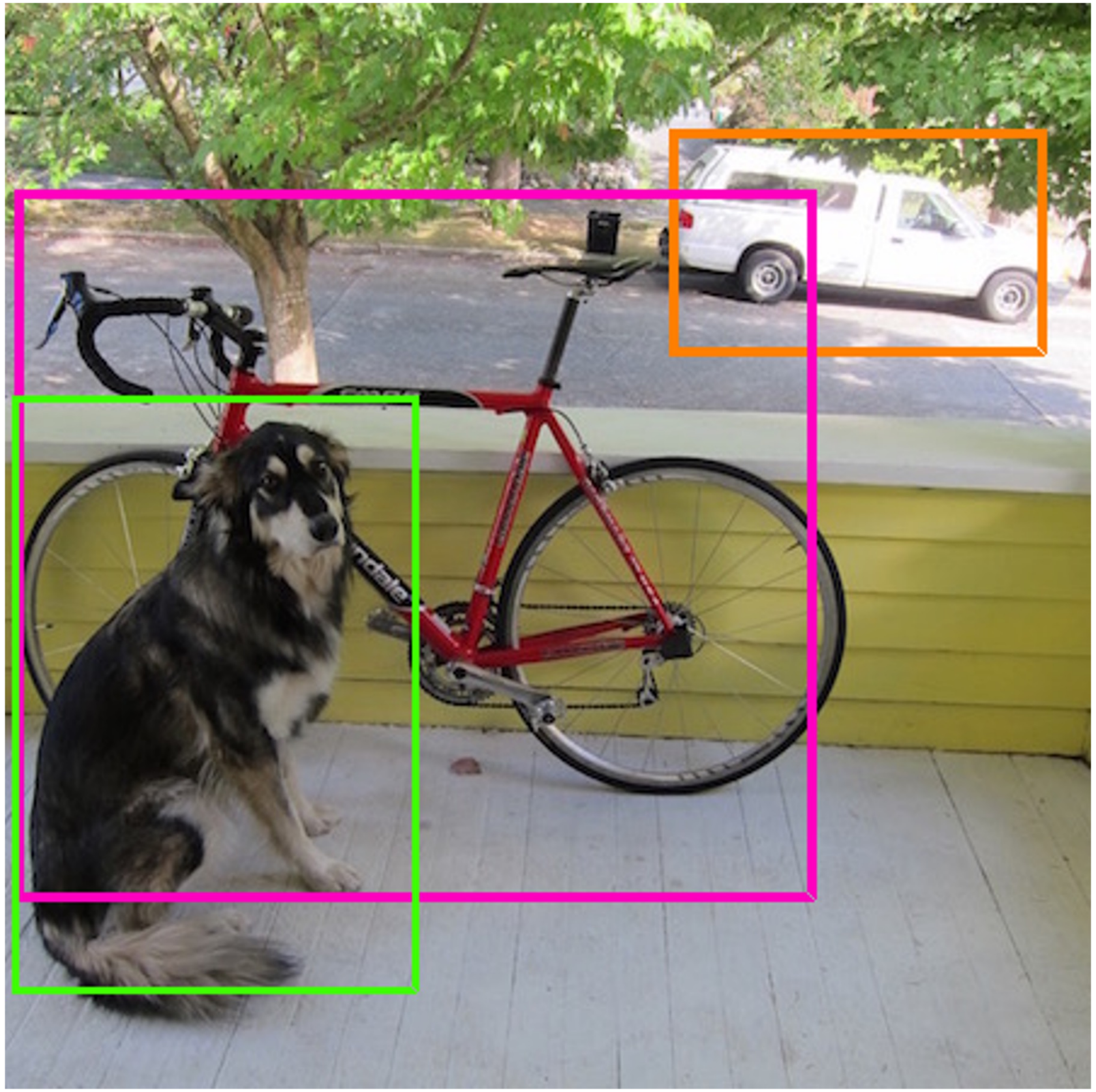
YOLO

- Then combine the box and class predictions.



YOLO

- Finally do NMS and threshold detections

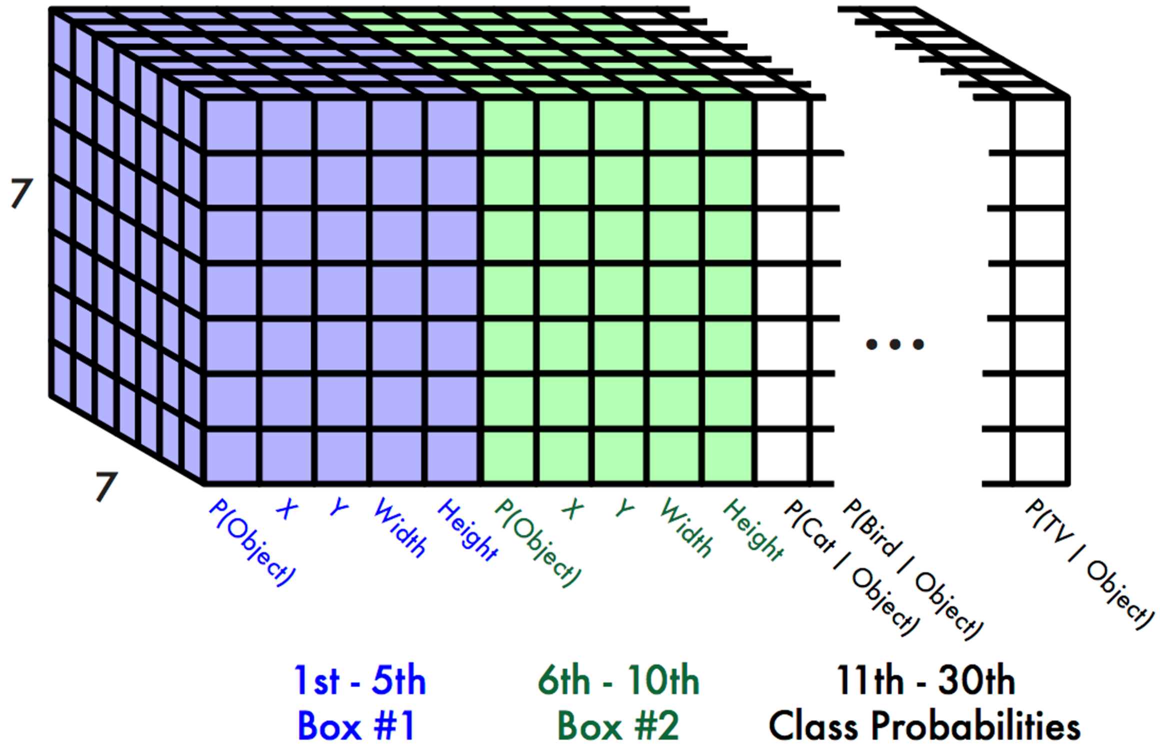


YOLO

This parameterization fixes the output size

Each cell predicts:

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



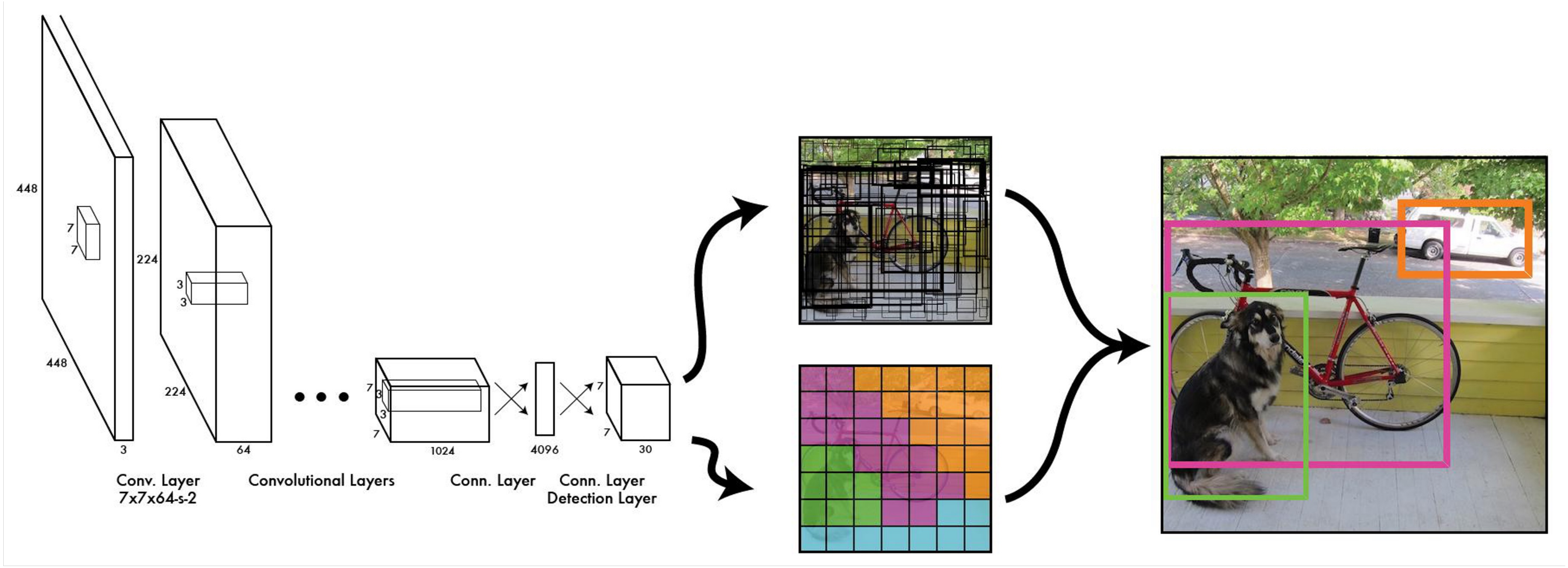
For Pascal VOC:

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

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

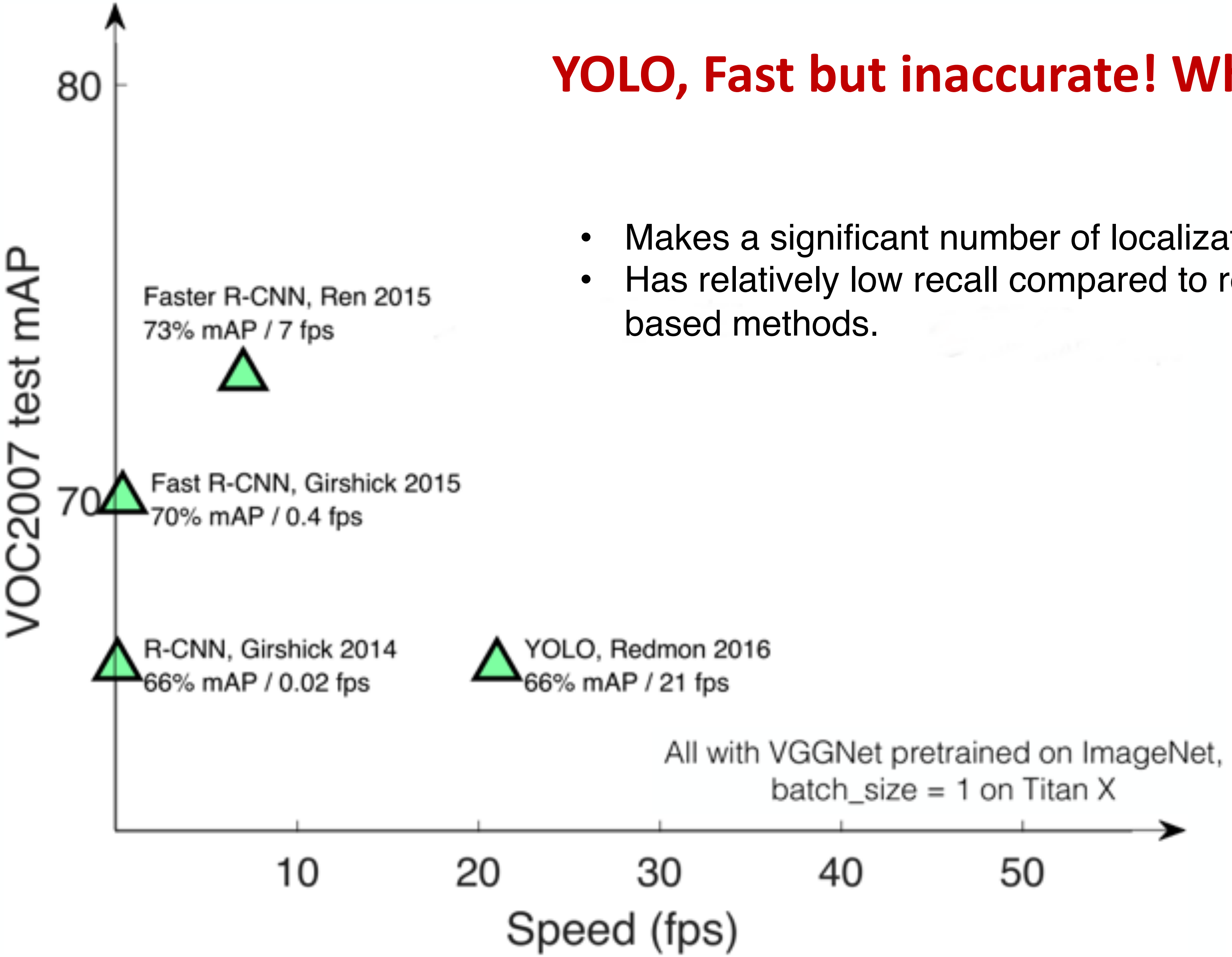
YOLO

- Thus we can train one neural network to be a whole detection pipeline



Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi
You Only Look Once: Unified, Real-Time Object Detection. CVPR 2016

YOLO, Fast but inaccurate! Why?



- Makes a significant number of localization errors.
- Has relatively low recall compared to region proposal-based methods.