UniGe | MaLGa

Object detection and segmentation: DL approaches

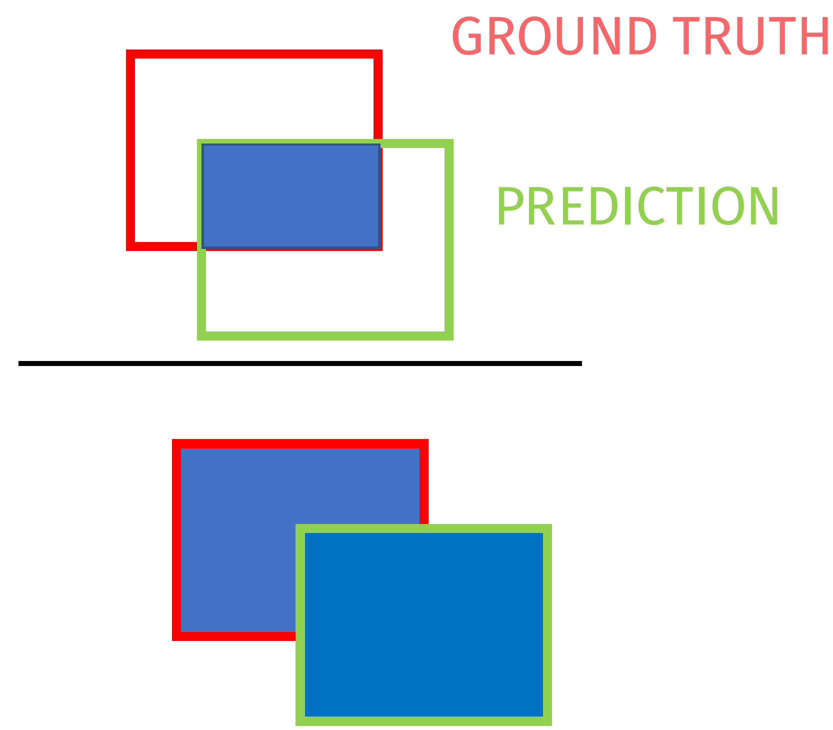
Matteo Moro & Francesca Odone matteo.moro@unige.it

Object detection evaluation

How are object detectors evaluated?

IOU or JACCARD METRIC

Intersection over Union=

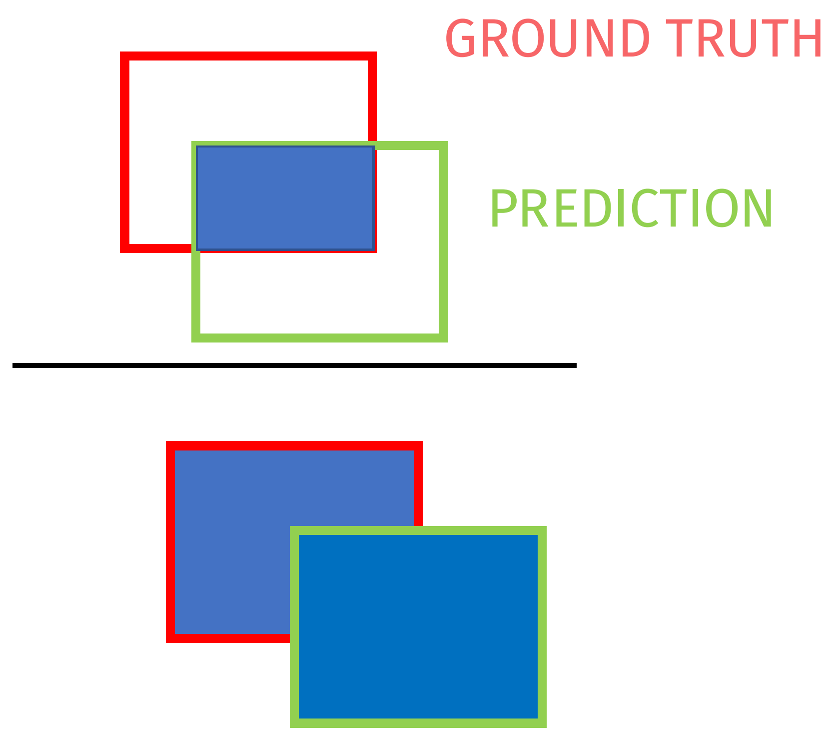


How are object detectors evaluated?

BINARY CASE

IOU METRIC

Intersection over Union=



TRUE POSITIVE
    if IOU >= THRESHOLD

FALSE POSITIVE
    if IOU < THRESHOLD

FALSE NEGATIVE
    A ground truth region with no associated prediction

TRUE NEGATIVE
    All the other regions

Threshold rule of thumb: greater than 0.5

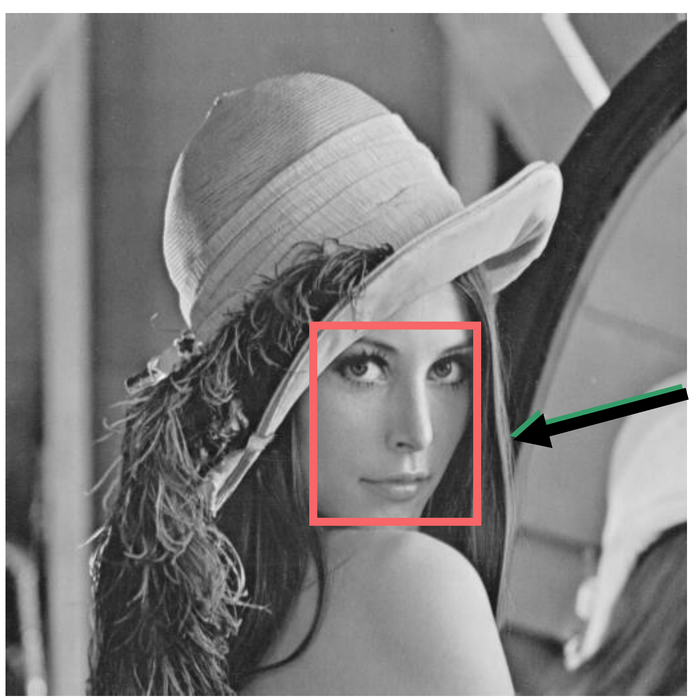
EXAMPLES

GROUND TRUTH

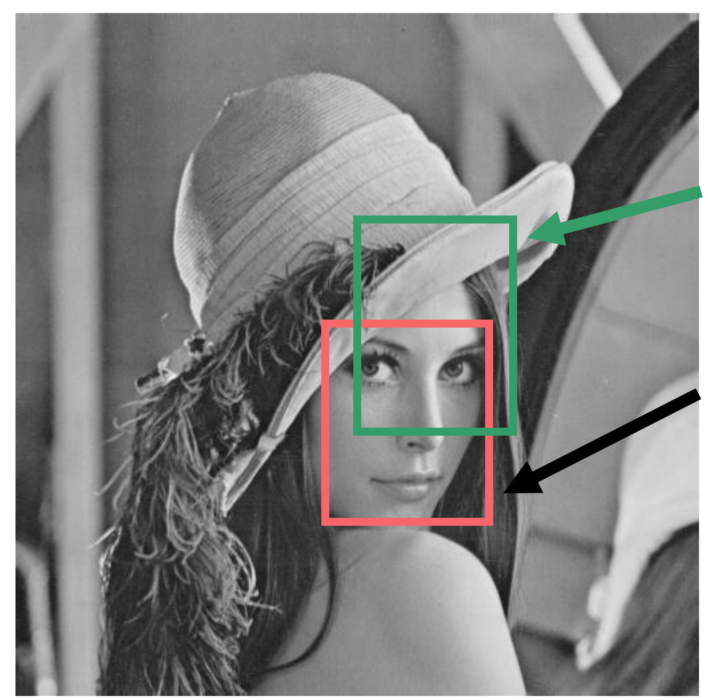
PREDICTION / DETECTION / ESTIMATION

FALSE POSITIVE

TRUE POSITIVE ✓



FALSE NEGATIVE



FALSE POSITIVE (low overlap)

FALSE NEGATIVE

Precision-Recall and F1 score

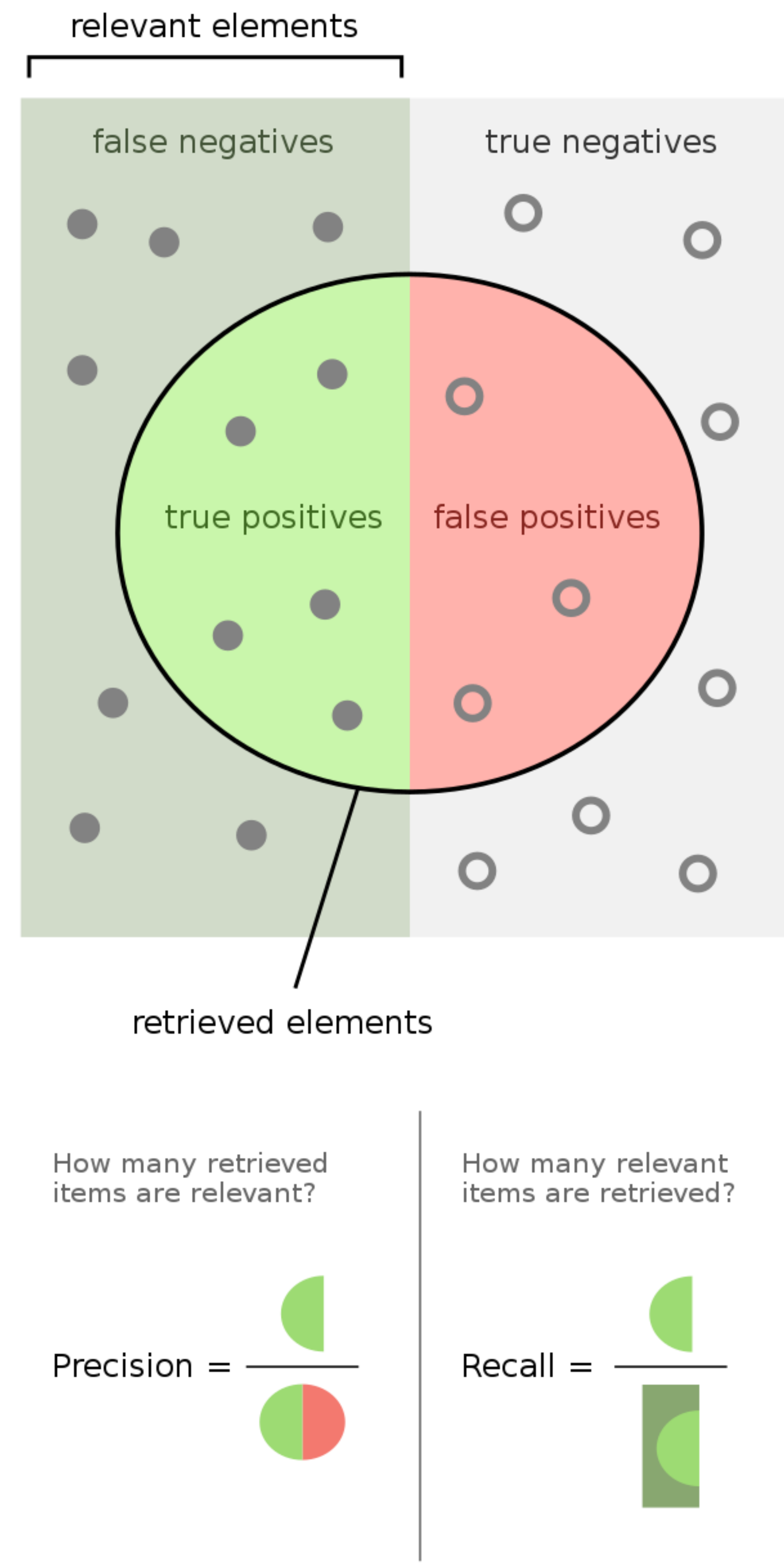
$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Estimated positives

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Real positives

$$F_1 = 2\frac{precision * recall}{precision + recall} = \frac{2TP}{2TP + FP + FN}$$



Multi-class evaluation

Confusion matrices

predicted

Ground truth

	C1	C2	C3	C4
C1	C1 true pos			
C2		C2 true pos		
C3			C3 true pos	
C4				C4 true pos

C1 false negatives

C2 false positives

Multi-class evaluation

Confusion matrices

predicted

Ground truth

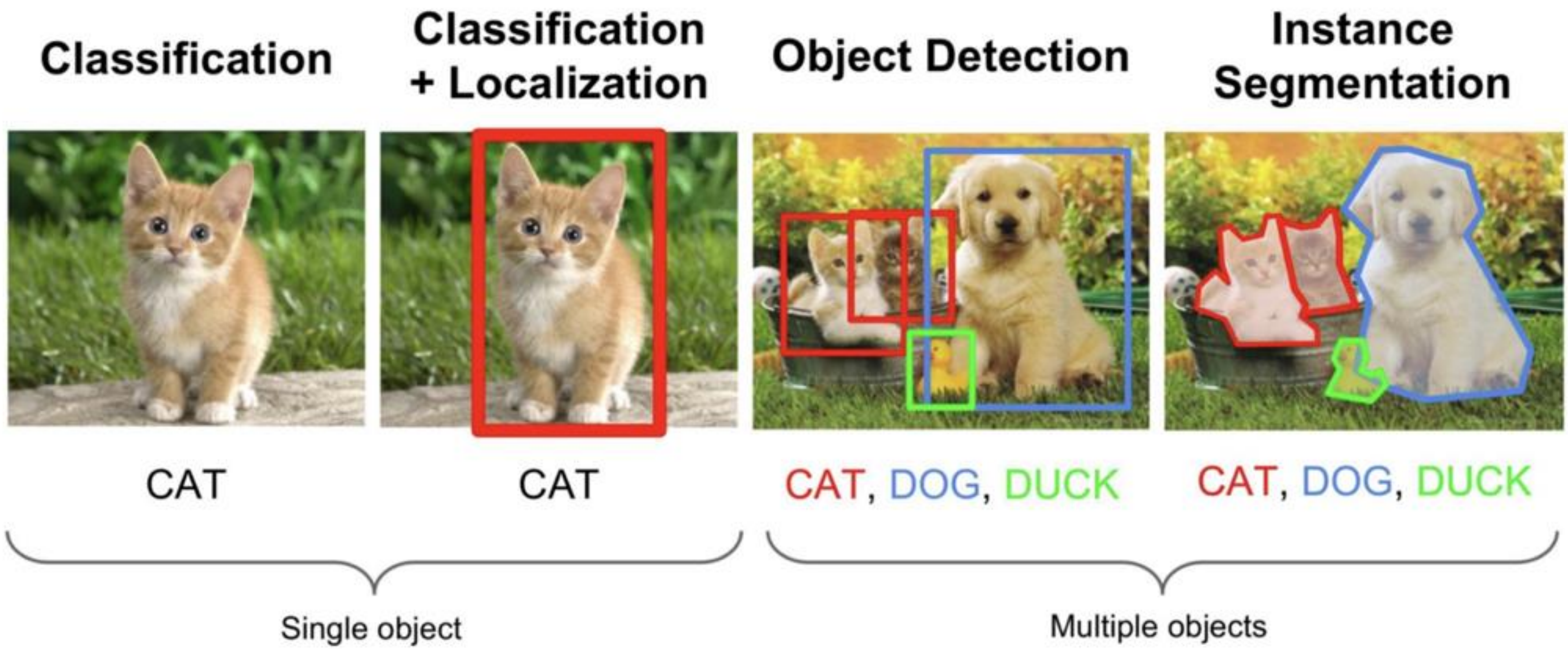
	C1	C2	C3	C4
C1	C1 true pos			
C2		C2 true pos		
C3			C3 true pos	
C4				C4 true pos

Region-based Convolutional Neural Networks – R-CNNs

- R. Girshick et al Region-based convolutional networks for accurate object detection and segmentation. TPAMI, 2015.
- Ren et al “Faster R-CNN. Towards Real-time object detection with Region Proposal Networks TPAMI 2017

Object detection

Different nuances



Object detection: sliding window basic idea

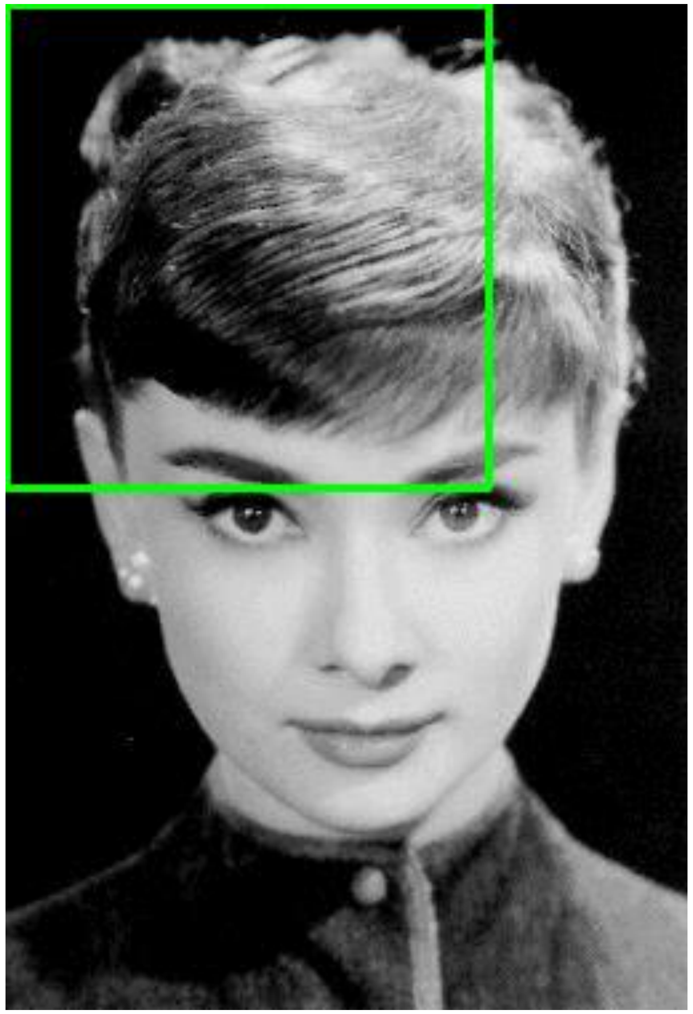
Slide a window across image and evaluate an object model at every location

that is perform
image classification!

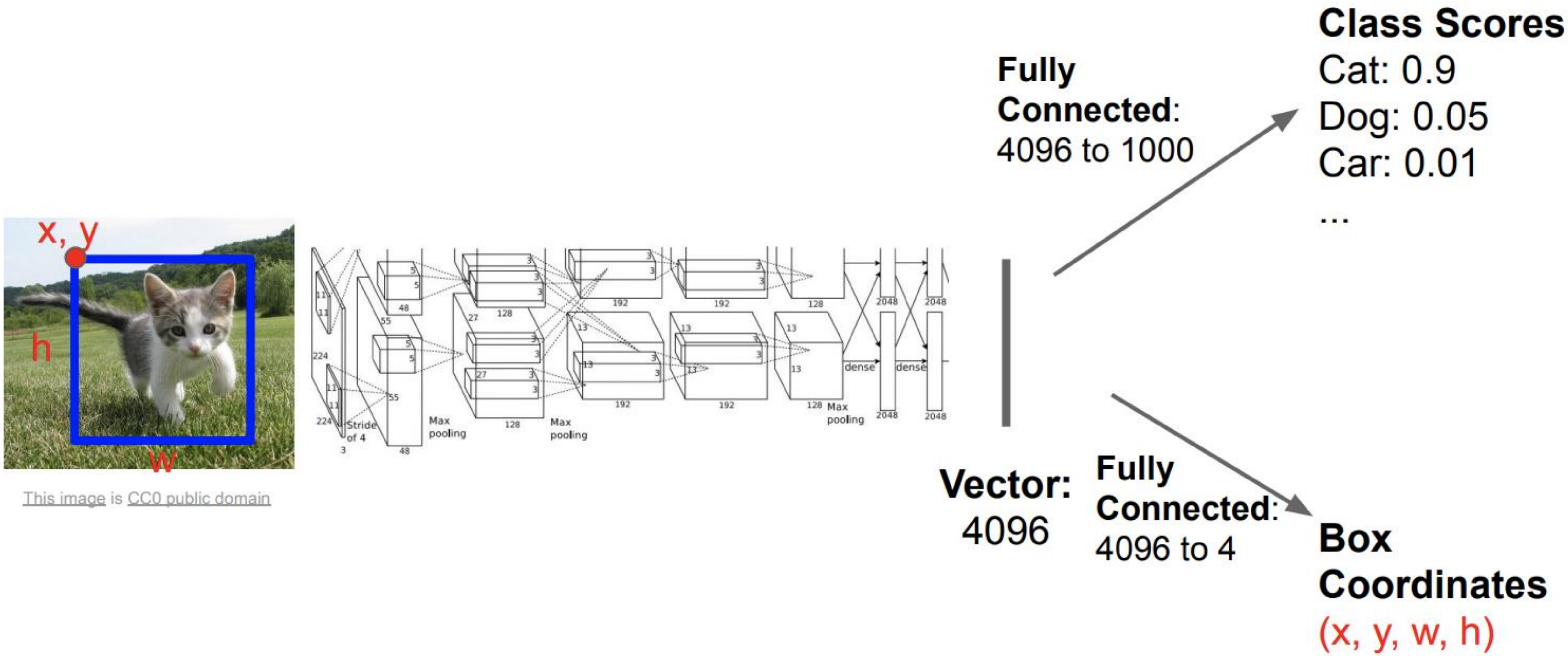
is it a face?

YES

NO

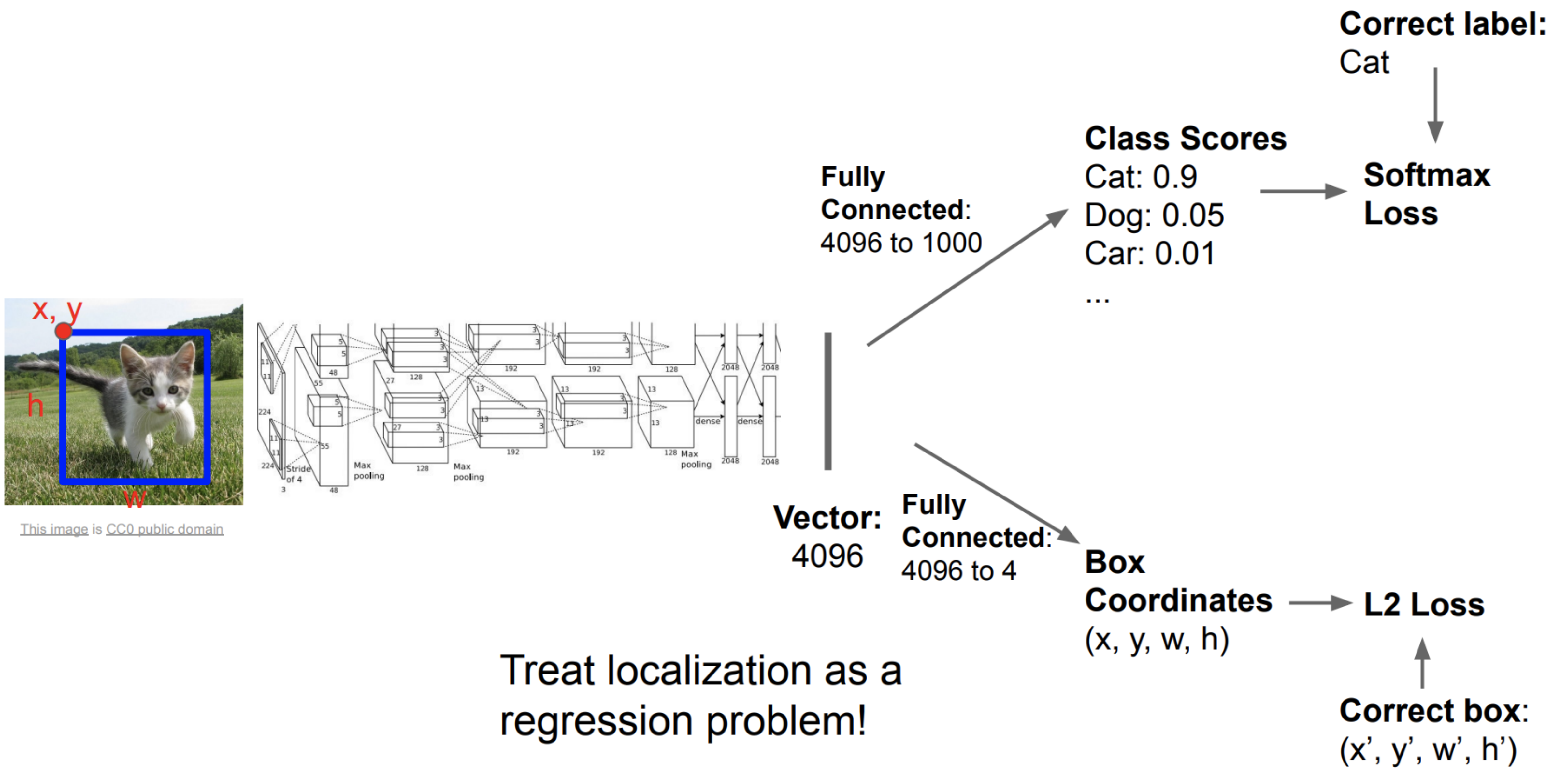


Classification + Localization



http://cs231n.stanford.edu/

Classification + Localization

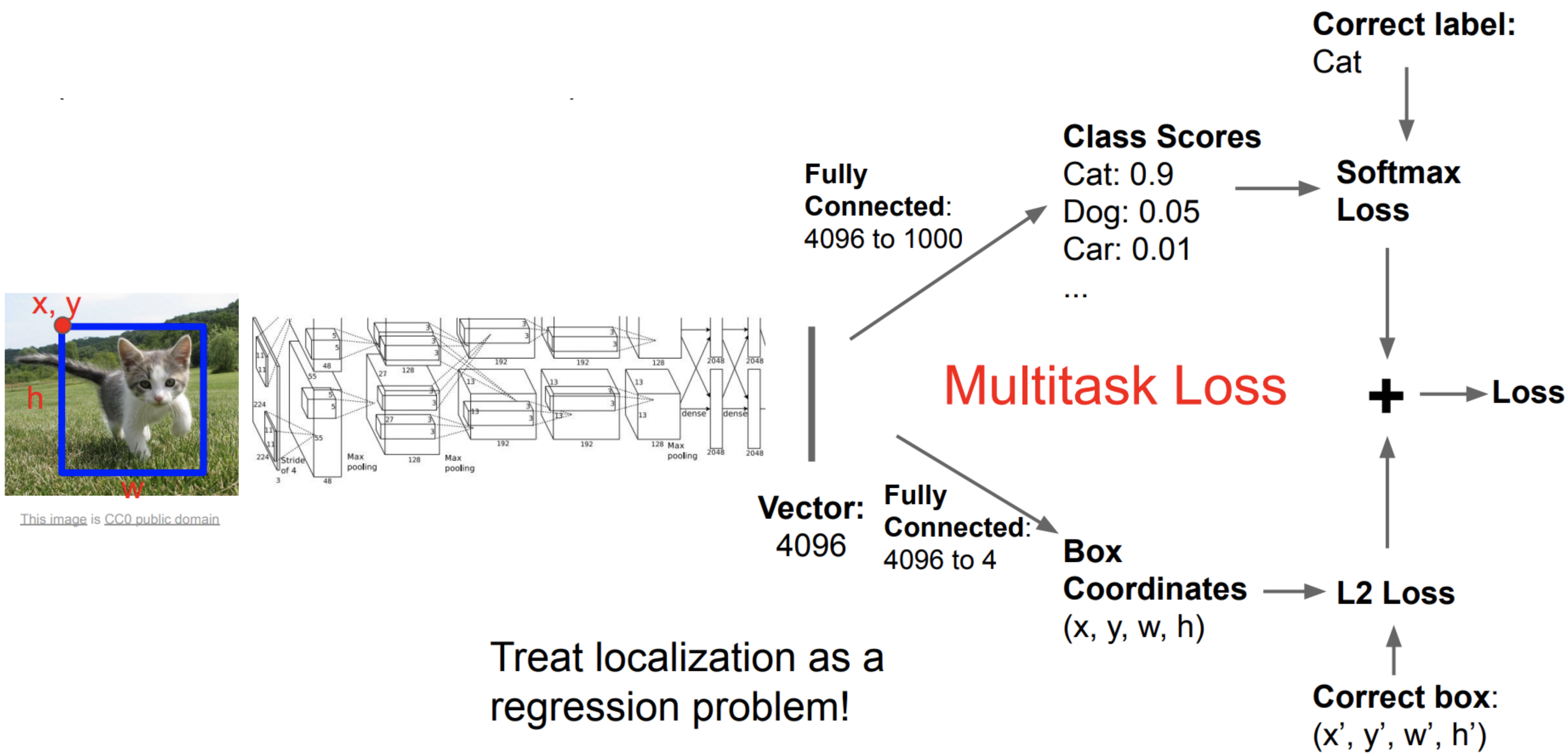


This image is CC0 public domain

Treat localization as a regression problem!

http://cs231n.stanford.edu/

Classification + Localization

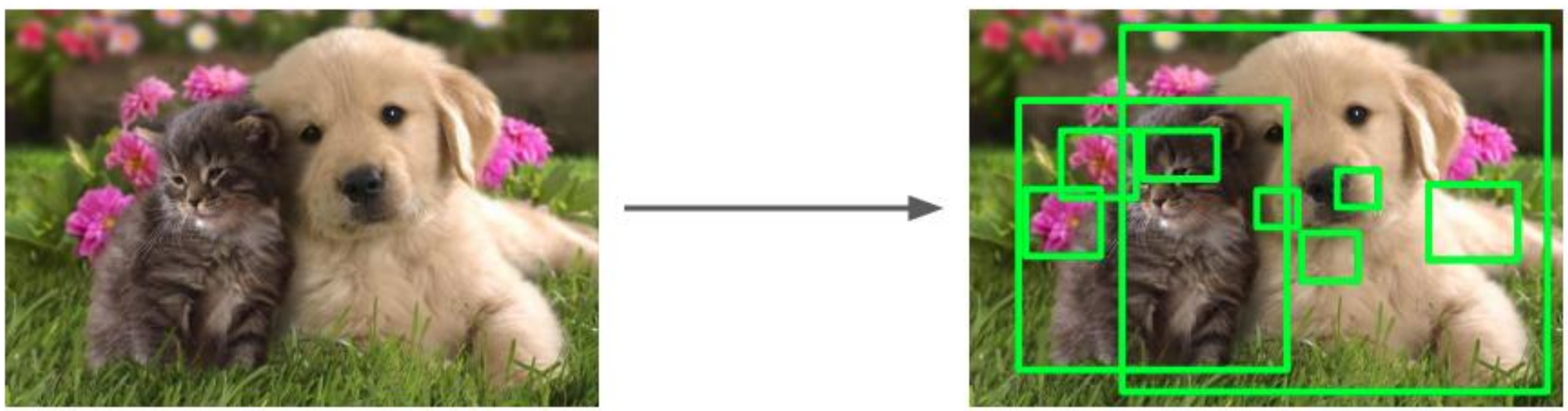


http://cs231n.stanford.edu/

Underlying concept

Region proposals

Find image regions that are likely to contain objects

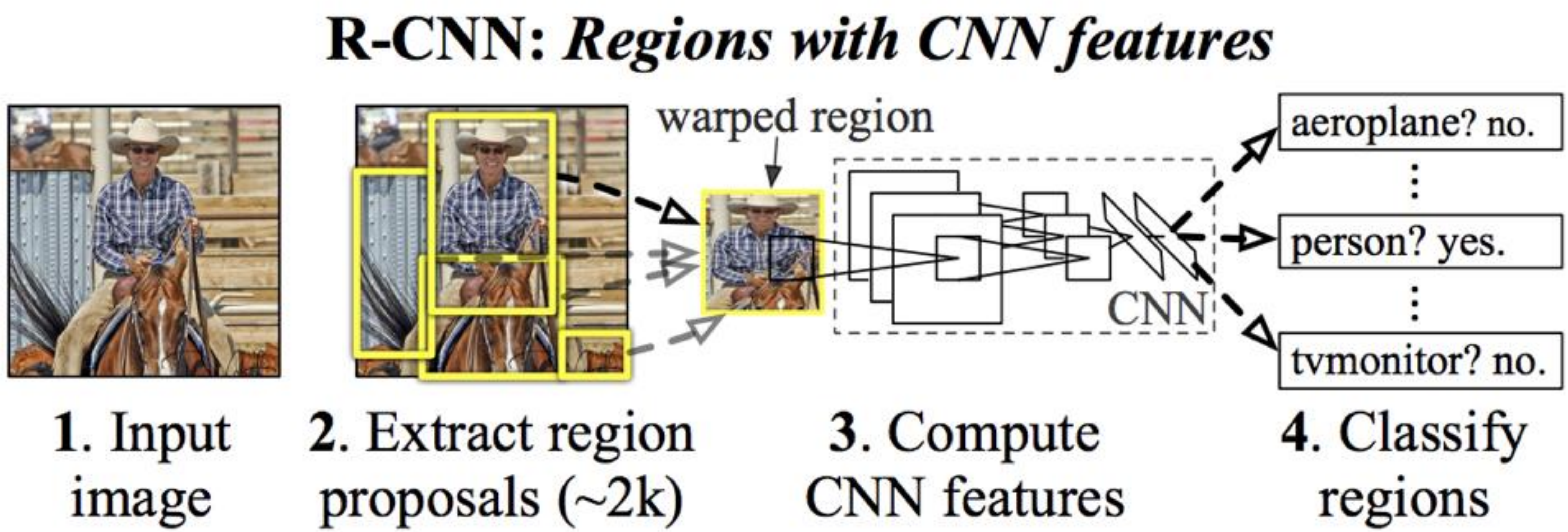


R-CNN

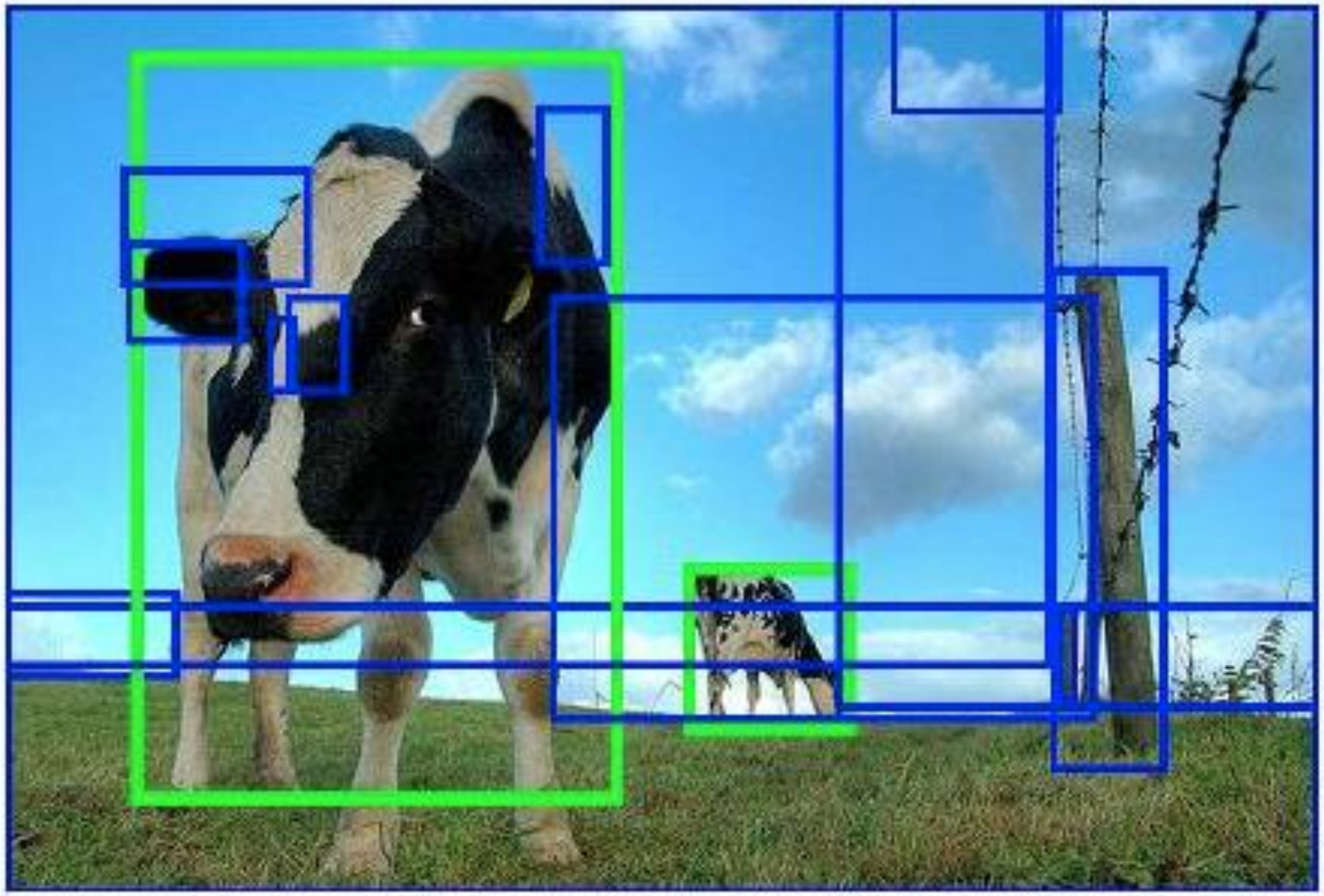
Two steps

- Selective search: identification of region proposals

- CNN (we may also use an external classifier, eg SVM)



Selective search



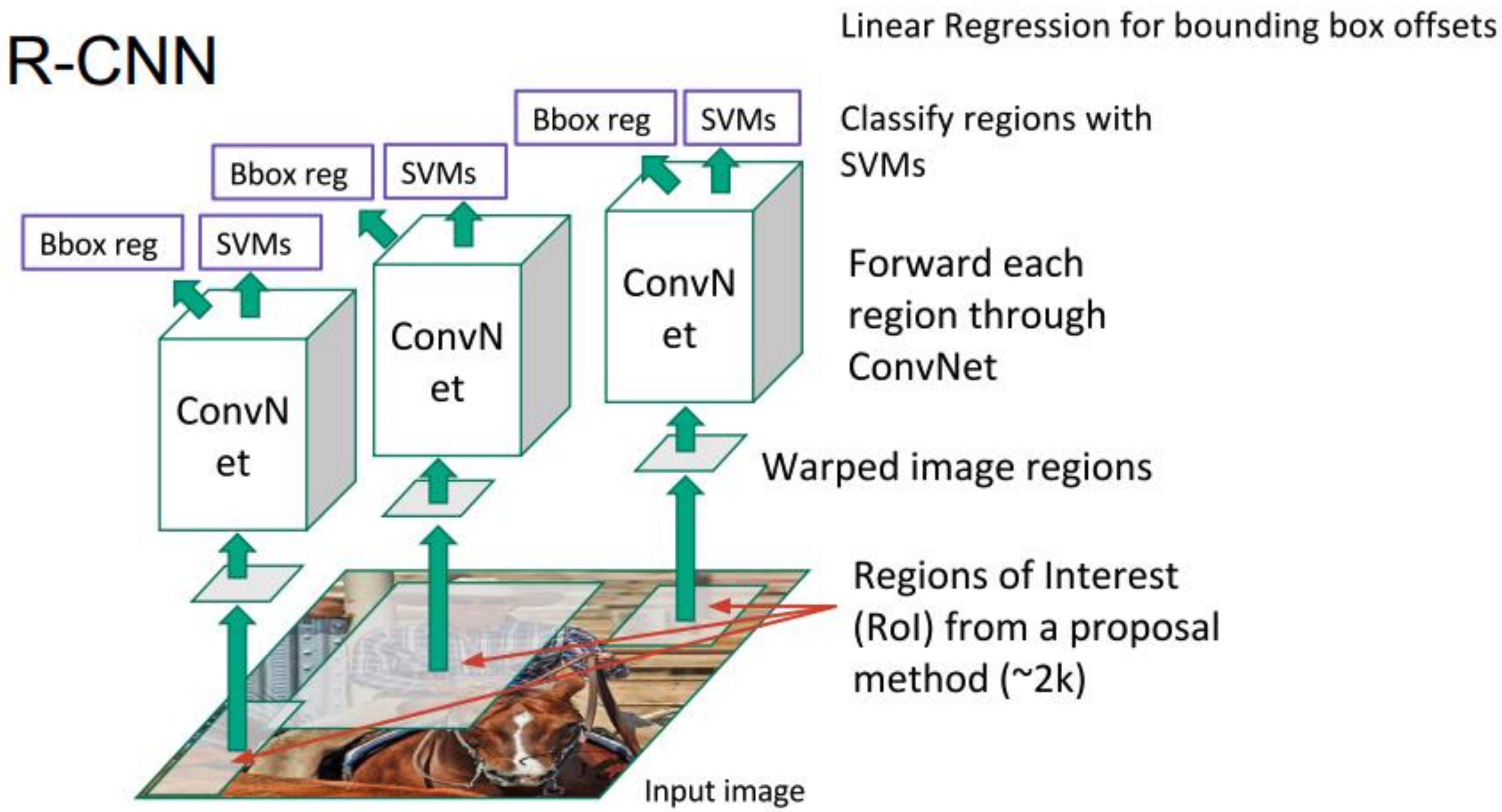
Designed to have high recall, but low precision: we have many false positive regions, but we are quite sure that they contain the object of interest

Selective search

Uijlings, Jasper RR, et al. "Selective search for object recognition." *International journal of computer vision* 104 (2013): 154-171.

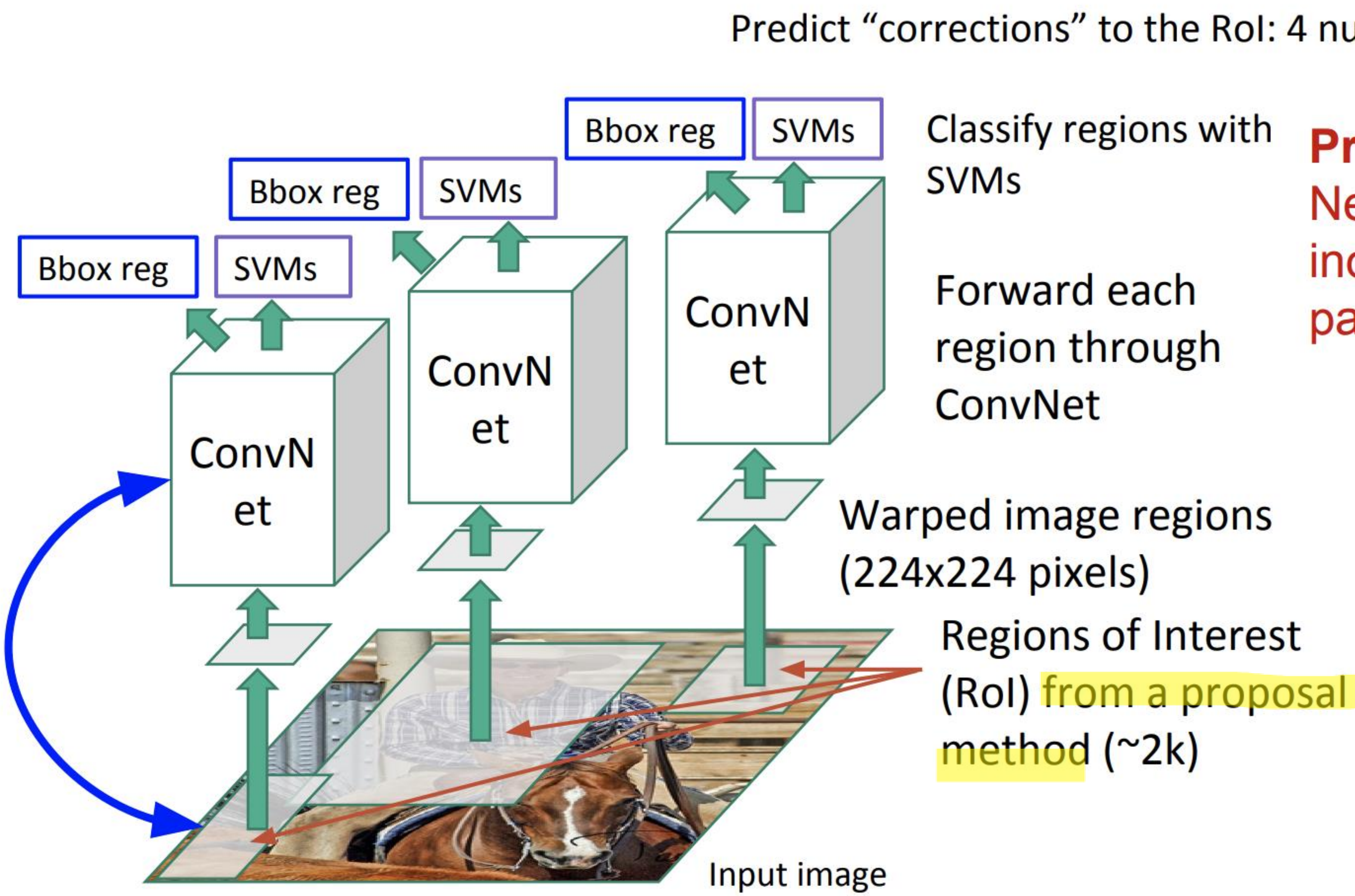
R-CNN

Two steps



http://cs231n.stanford.edu/

R-CNN

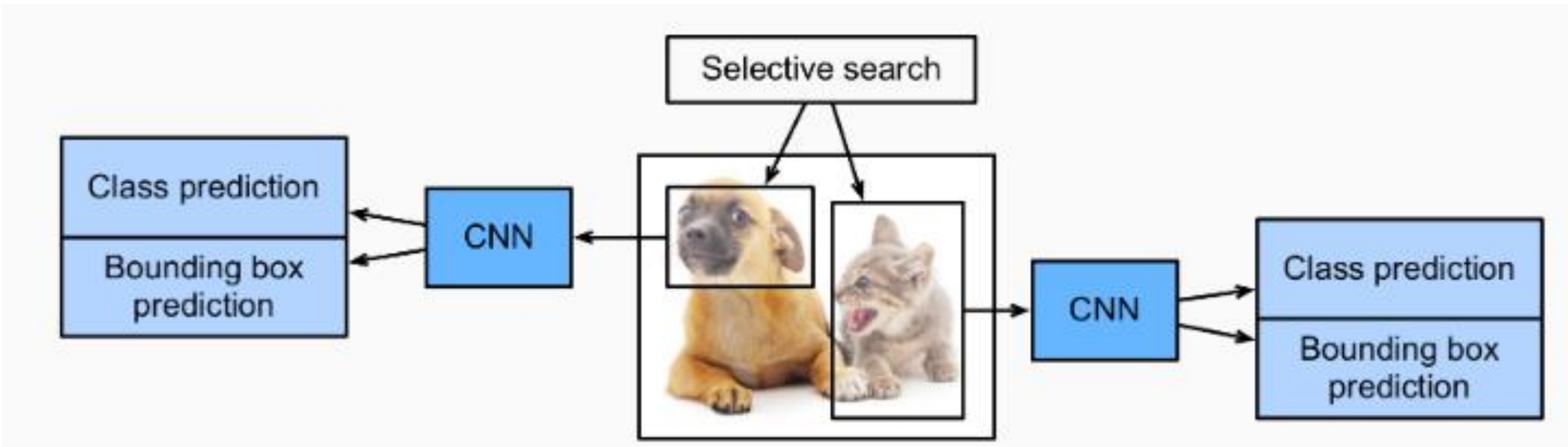


Problem: Very slow! Need to do ~2k independent forward passes for each image!

Idea: Pass the image through convnet before cropping! Crop the conv feature instead!

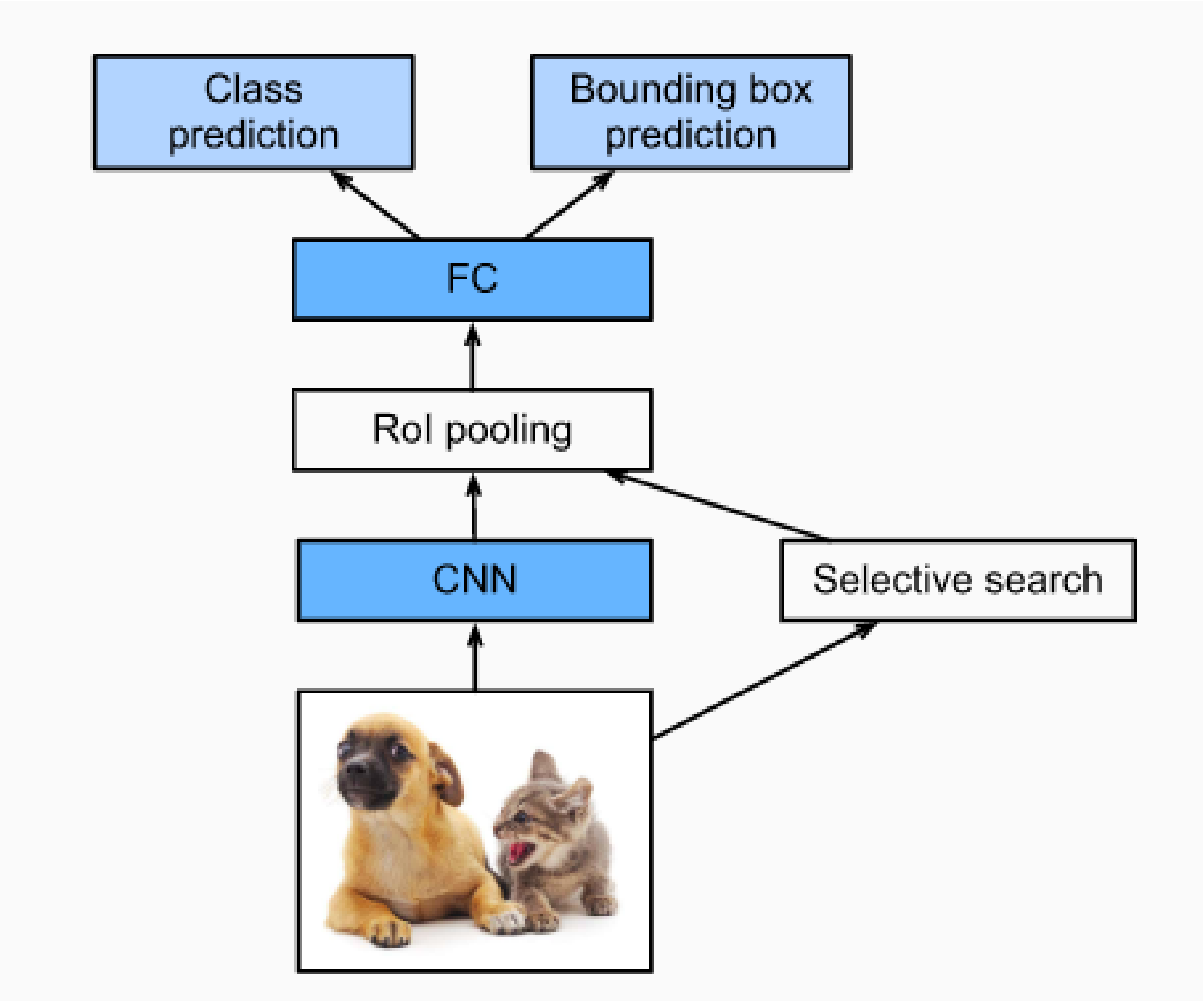
http://cs231n.stanford.edu/

R-CNN



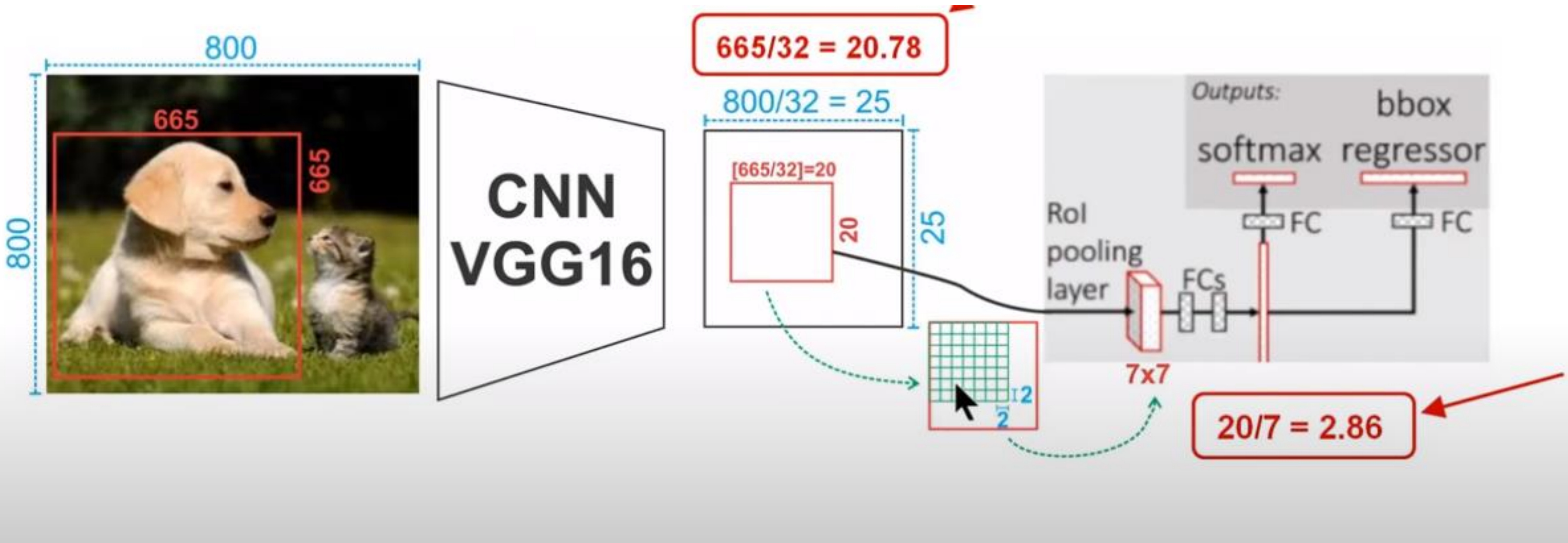
Selective search
Class prediction
Bounding box prediction
CNN
CNN
Class prediction
Bounding box prediction

Fast R-CNN



Class prediction
Bounding box prediction
FC
RoI pooling
CNN
Selective search

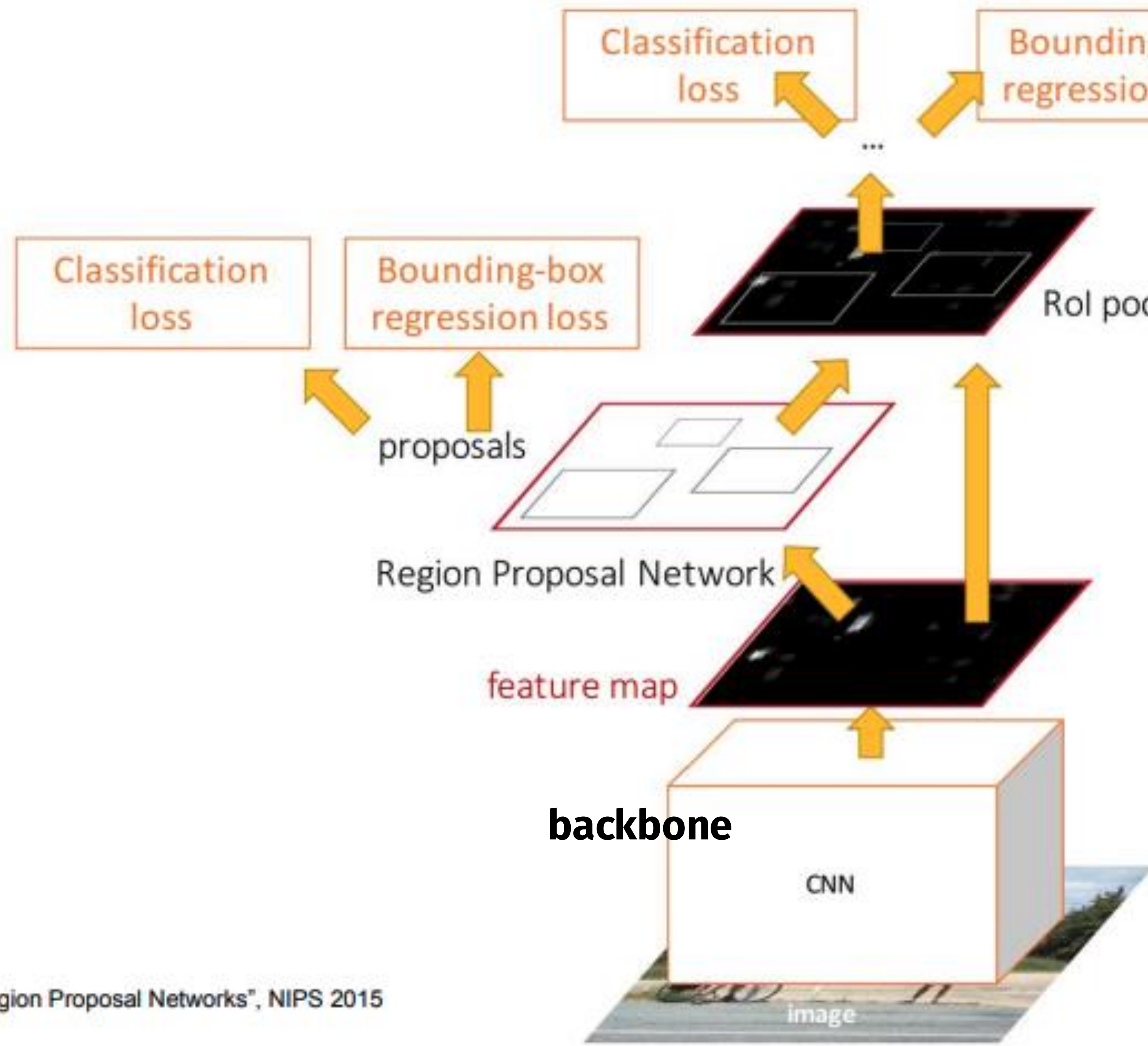
ROI pooling



665/32 = 20.78
800
800/32 = 25
665
665
800
CNN
VGG16
[665/32]=20
20
25
Outputs:
softmax
bbox
regressor
FC
FC
RoI
pooling
layer
FCs
7x7
2
2
20/7 = 2.86

Faster R-CNN

The network learns its proposals:
insert a **Region Proposal Network (RPN)** to predict proposals from features



Region Proposal Networks", NIPS 2015

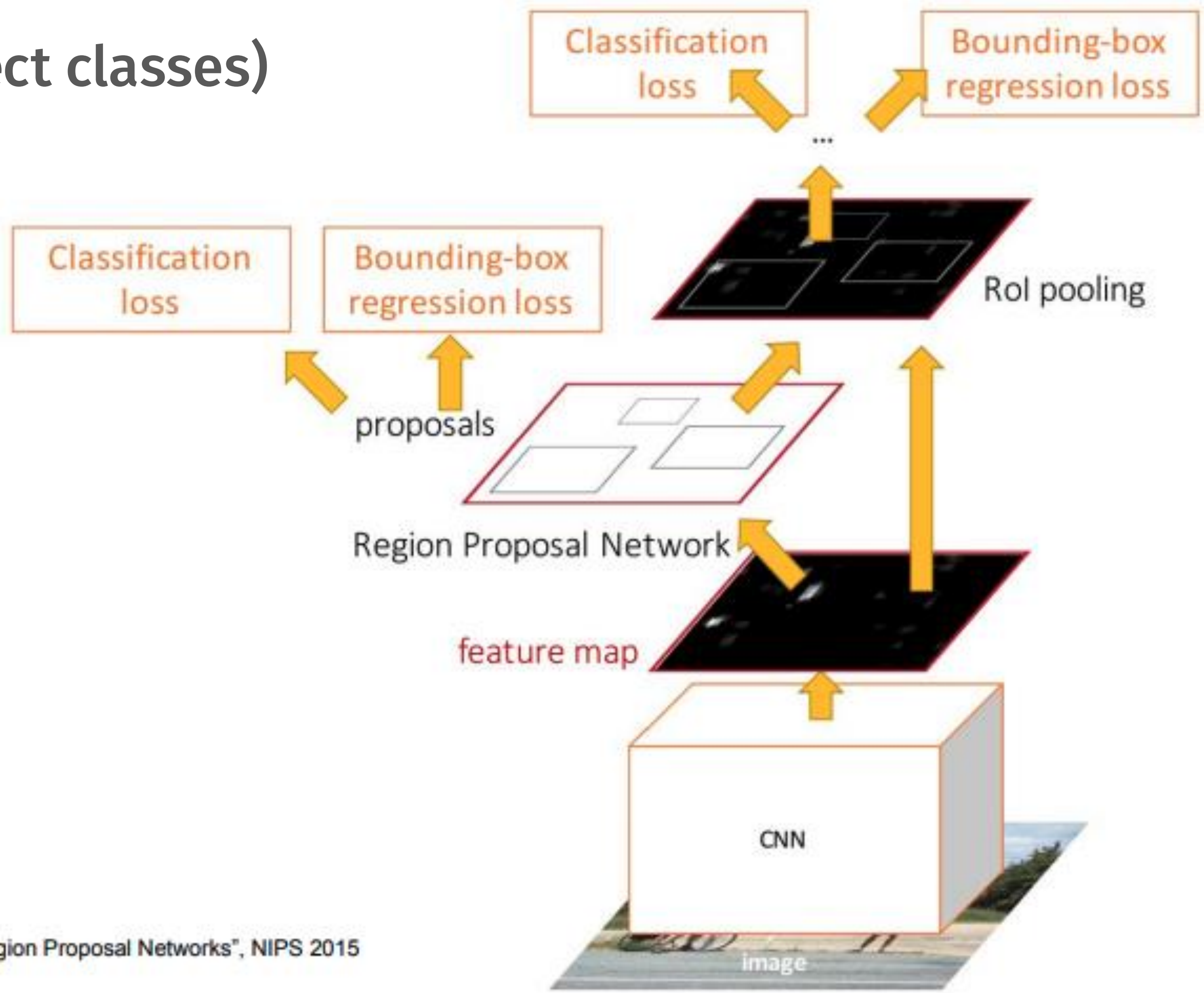
Faster R-CNN

The network learns its proposals:

insert a **Region Proposal Network (RPN)** to predict proposals from features

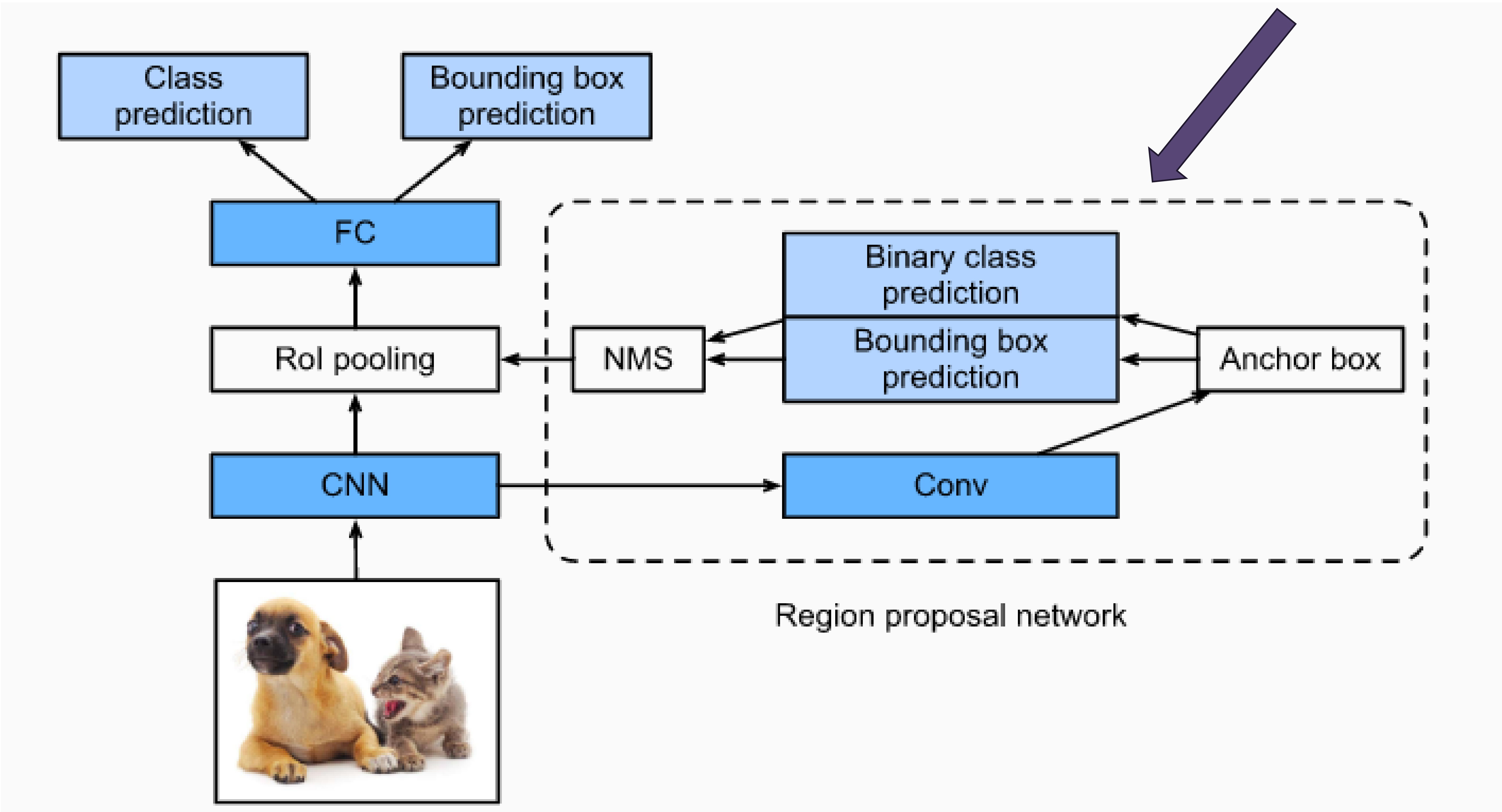
– Jointly train 4 losses

1. RPN classify object/non object
2. RPN regress box coordinates
3. Final Classification score (object classes)
4. Final Box coordinates

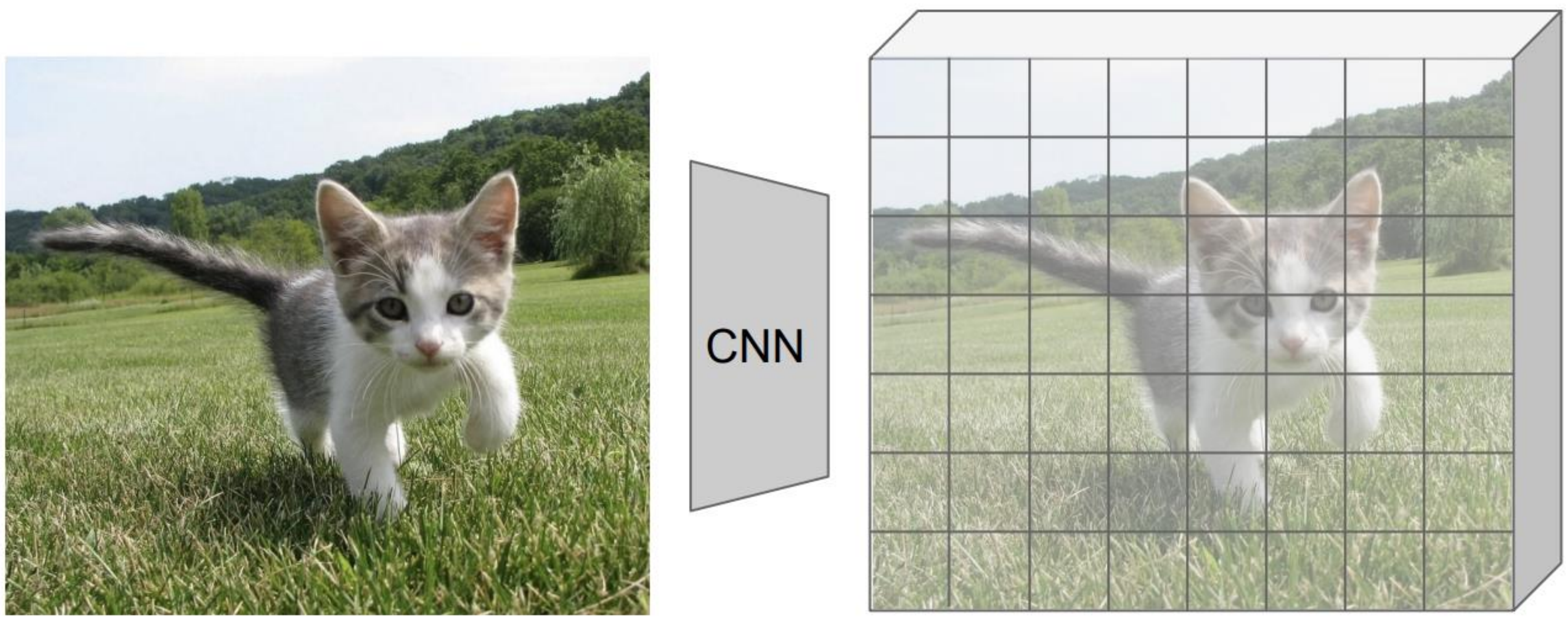


Region Proposal Networks", NIPS 2015

Faster R-CNN


Class prediction
Bounding box prediction
FC
RoI pooling
CNN
NMS
Binary class prediction
Bounding box prediction
Anchor box
Conv
Region proposal network

Region Proposal Network

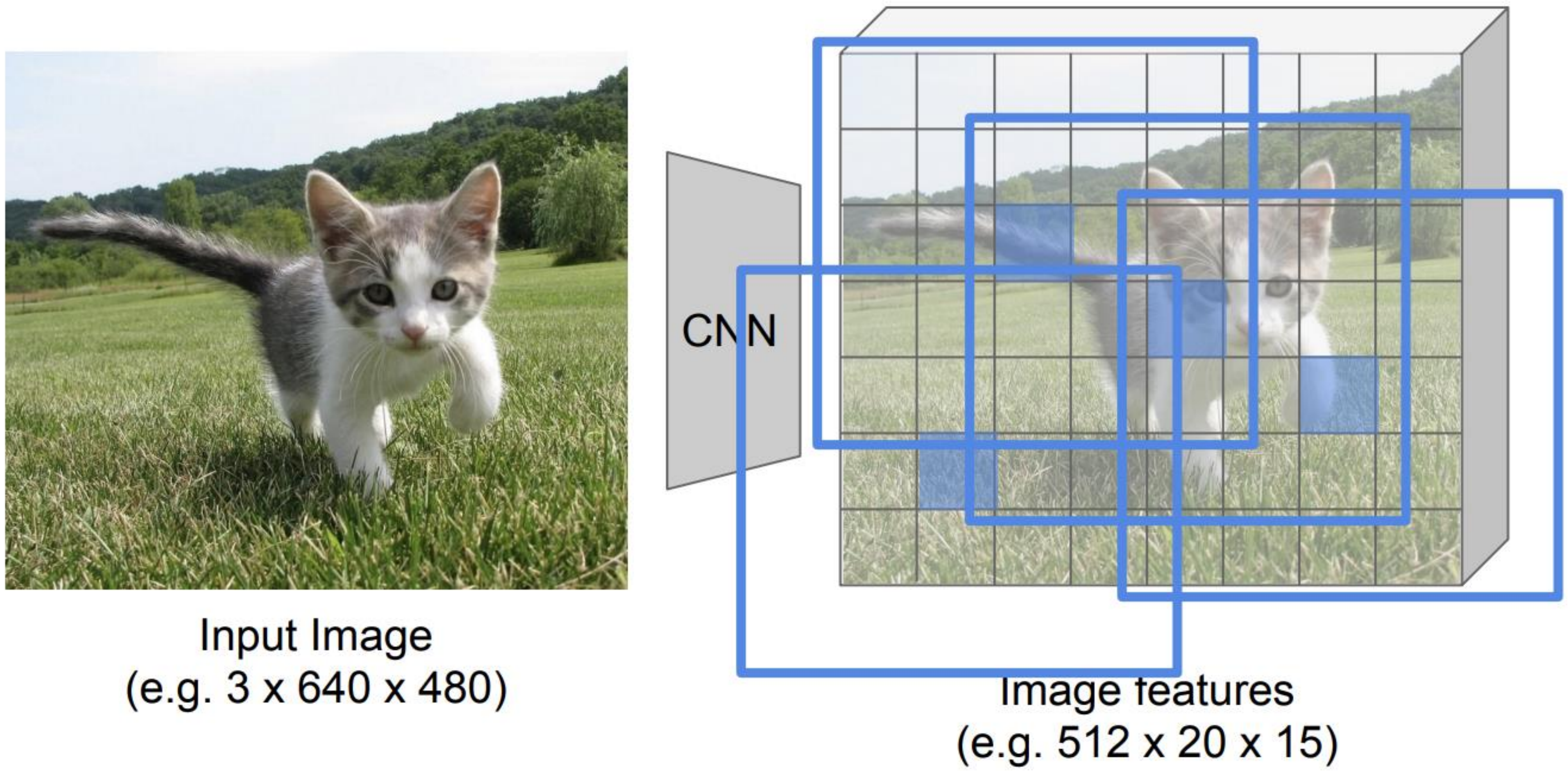


Input Image
(e.g. 3 x 640 x 480)

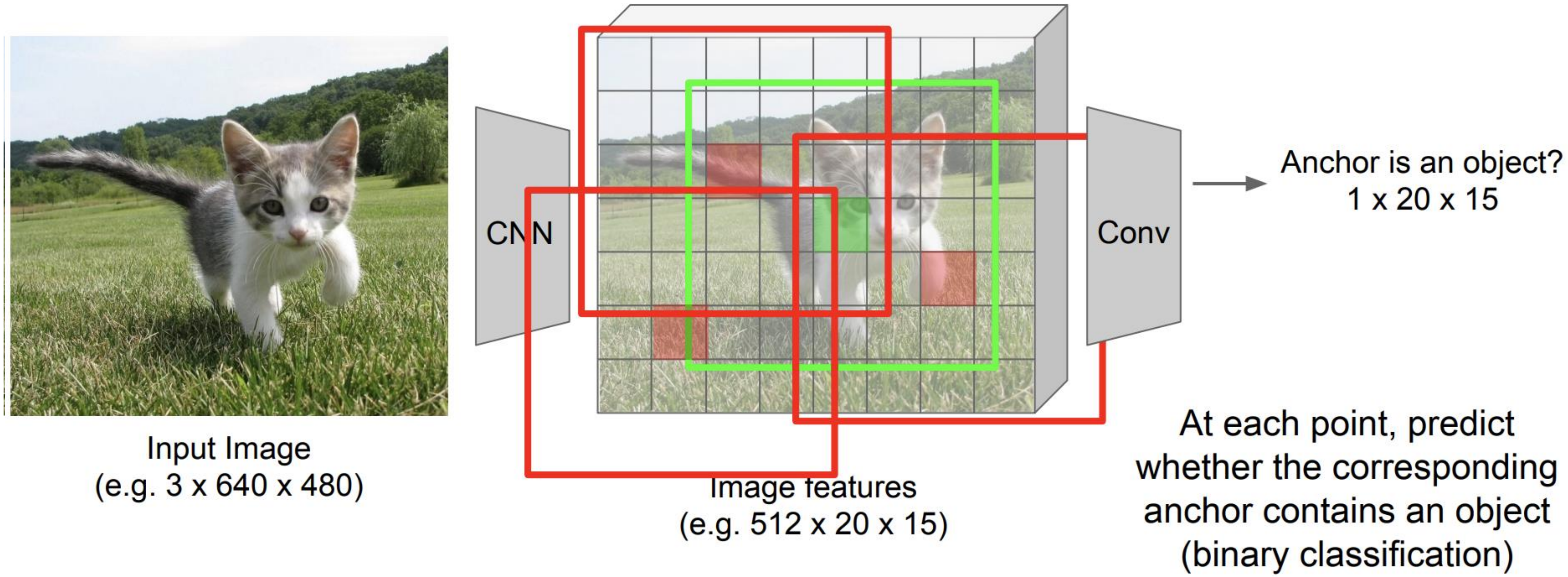
Image features
(e.g. 512 x 20 x 15)

Region Proposal Network

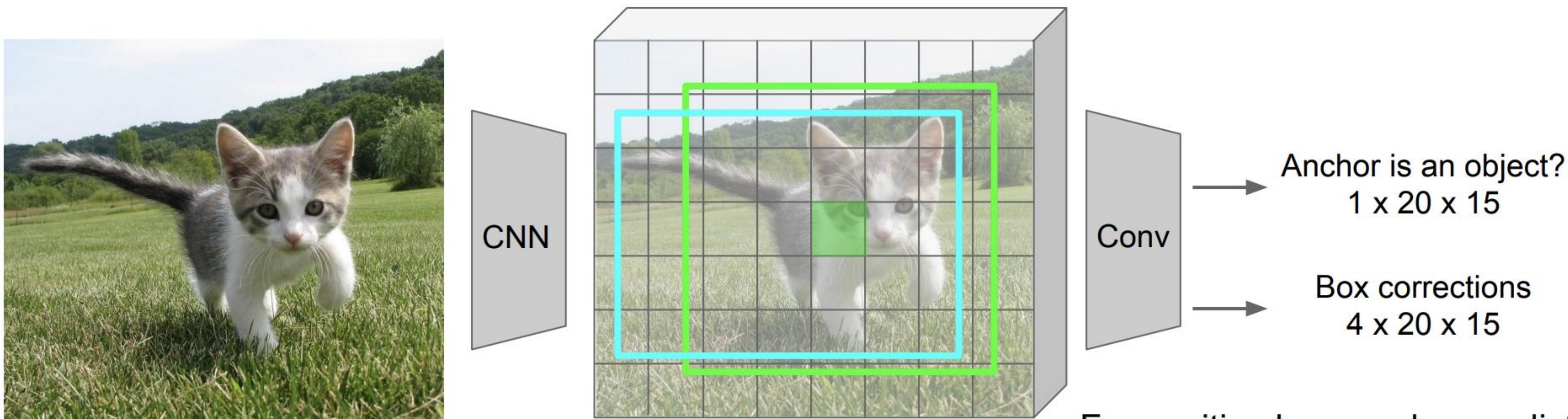
Anchor box of fixed size centered on
each point of the feature map



Region Proposal Network



Region Proposal Network



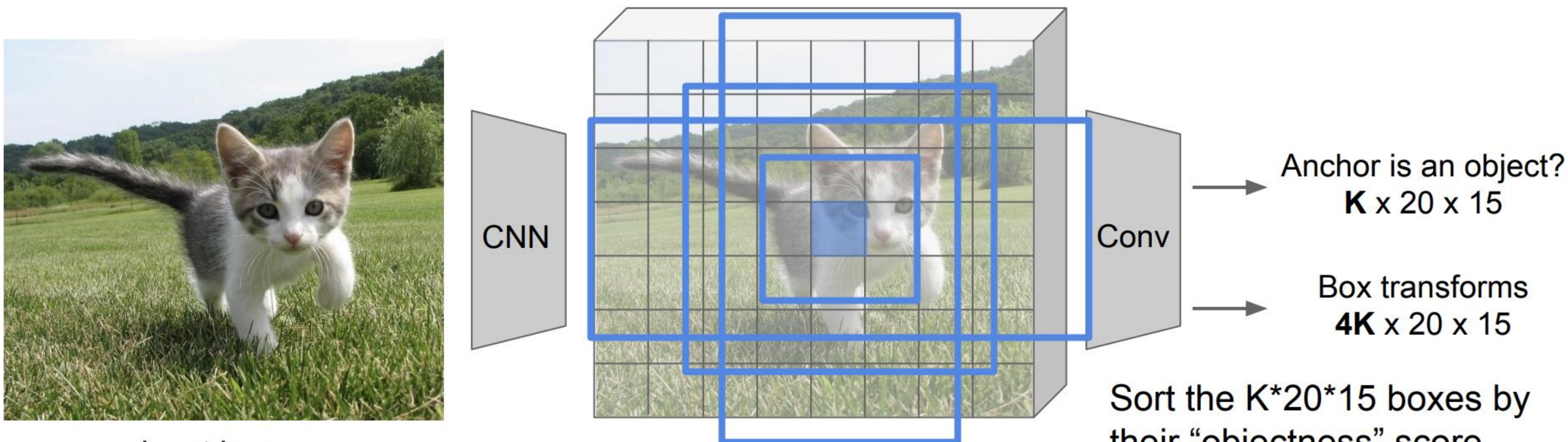
Input Image
(e.g. 3 x 640 x 480)

Image features
(e.g. 512 x 20 x 15)

For positive boxes, also predict a corrections from the anchor to the ground-truth box (regress 4 numbers per pixel)

Region Proposal Network

In practice use K different anchor boxes of different size / scale at each point



Input Image
(e.g. 3 x 640 x 480)

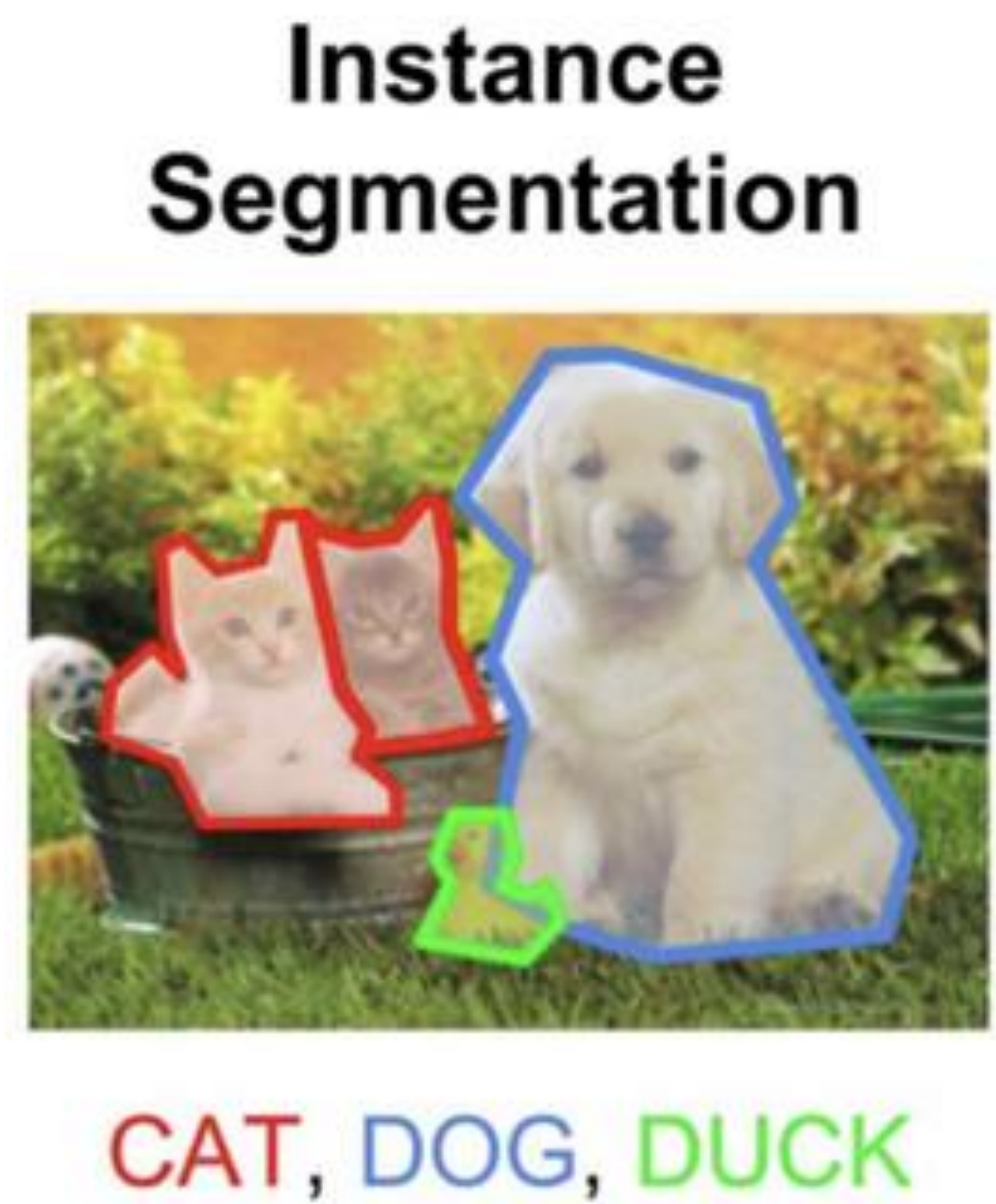
Image features
(e.g. 512 x 20 x 15)

Sort the K*20*15 boxes by their “objectness” score, take top ~300 as our proposals

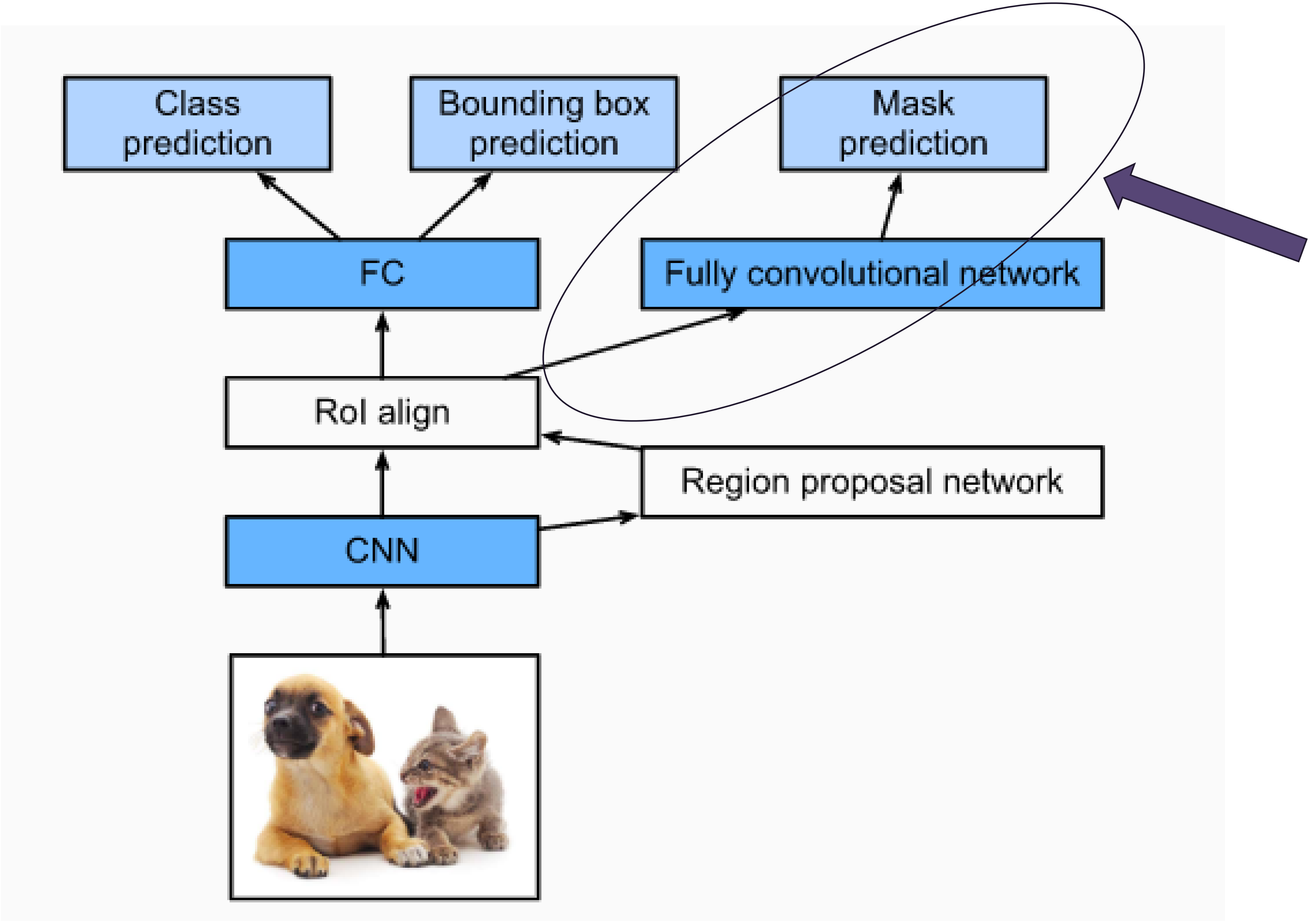
Mask R-CNNs

- He, K., Gkioxari, G., Dollár, P., & Girshick, R. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, 2017

Instance segmentation

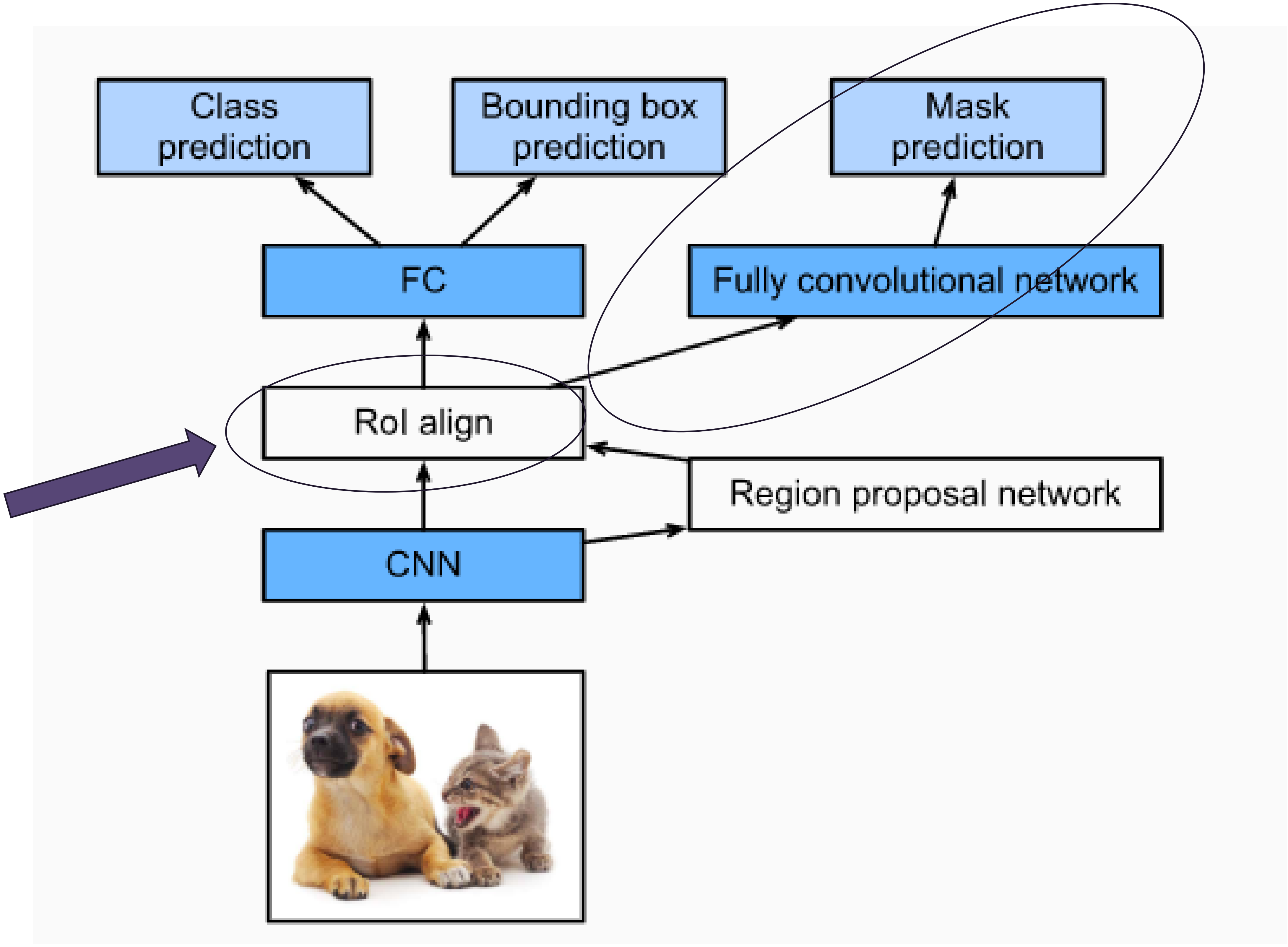


Mask RCNN

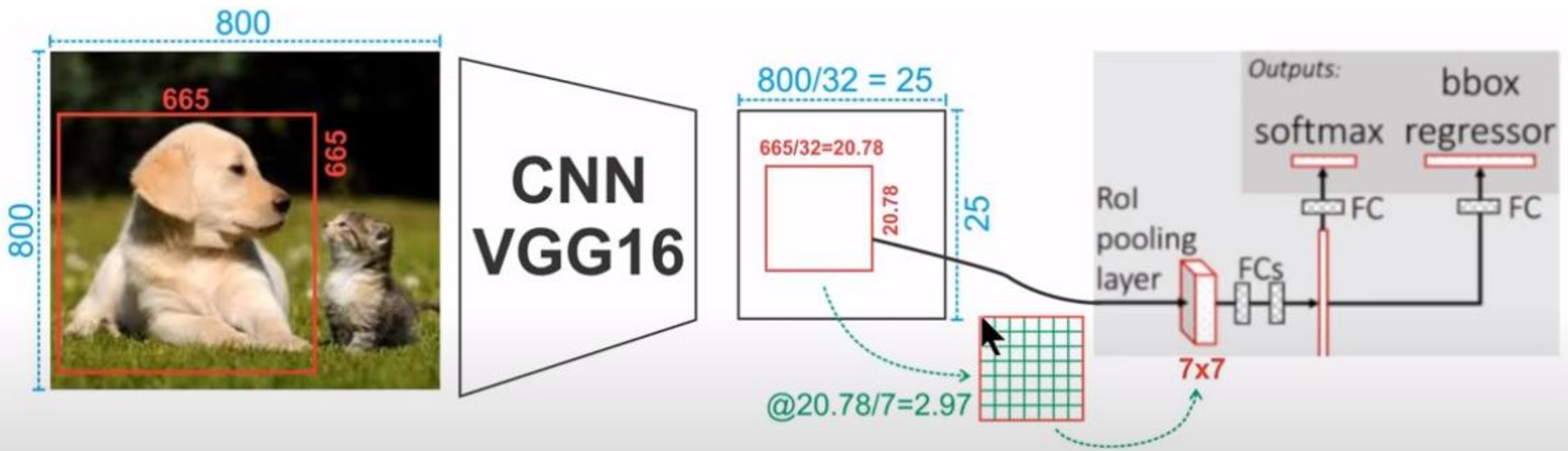


Class prediction
Bounding box prediction
Mask prediction
FC
Fully convolutional network
RoI align
Region proposal network
CNN

Mask RCNN

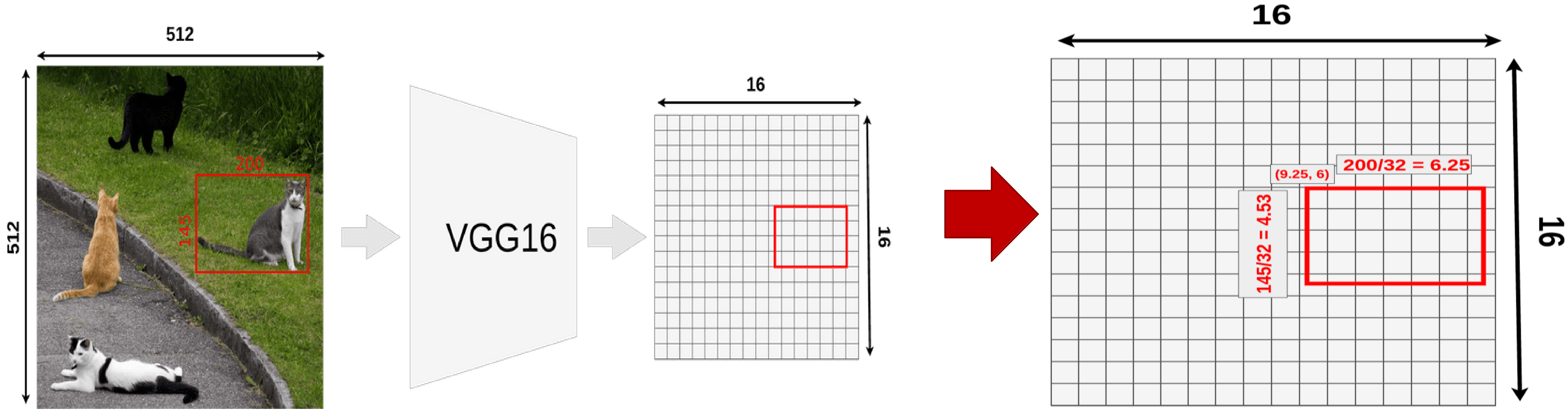

Class prediction
Bounding box prediction
Mask prediction
FC
Fully convolutional network
RoI align
Region proposal network
CNN

RoI align – no quantization



RoI align

➢ ***The main difference w.r.t ROI pooling is -> No Quantization***

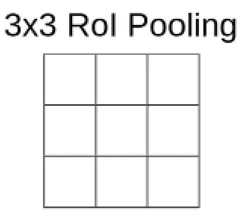


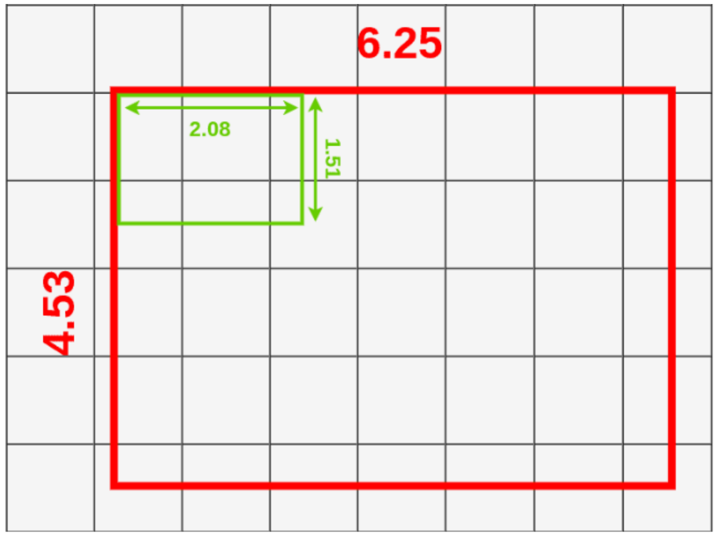
1.No quantization in mapping

Mask RCNN: Region of Interest Alignment

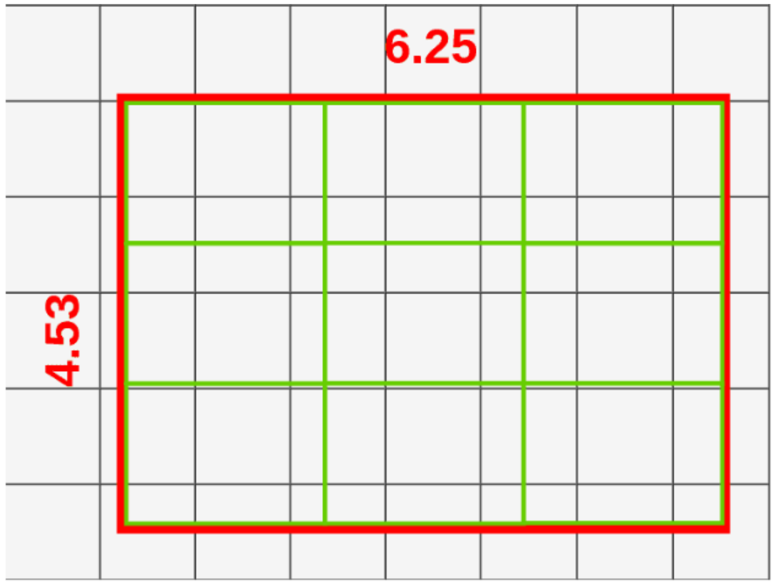
➢ *2 No Quantization in data pooling*

Pooling layer





RoI Box size



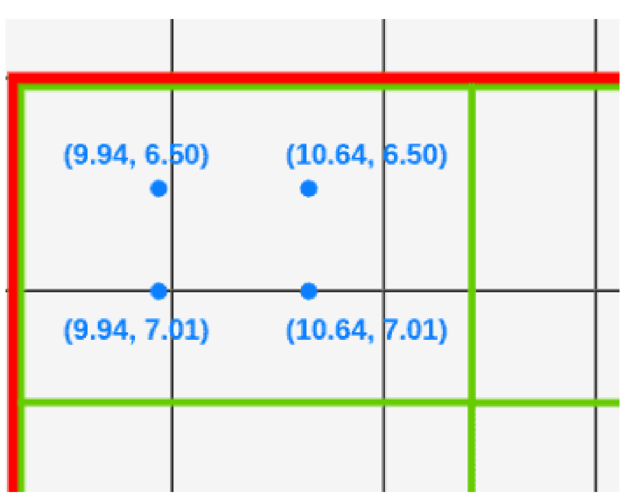
RoI divided into boxes

1. The pooling filter is 3x3 in the example;
2. We divide the mapped ROI in 9 boxes, with no quantization

Mask RCNN: Region of Interest Alignment

➢ *2 No Quantization in data pooling*

RoI divided into boxes



1. If we look at first box, it covers six pixels in The feature map
2. We sample 4 points inside the box (points Coordinates are chosen according to box size)

You can calculate where each of those points should be by **dividing height and width of the box by 3**.

In our case we're calculating first point (top left) coordinates like this:

- X = X_box + (width/3) * 1 = 9.94
- Y = Y_box + (height/3) * 1 = 6.50

To calculate the second point (bottom left) we have to change only the Y:

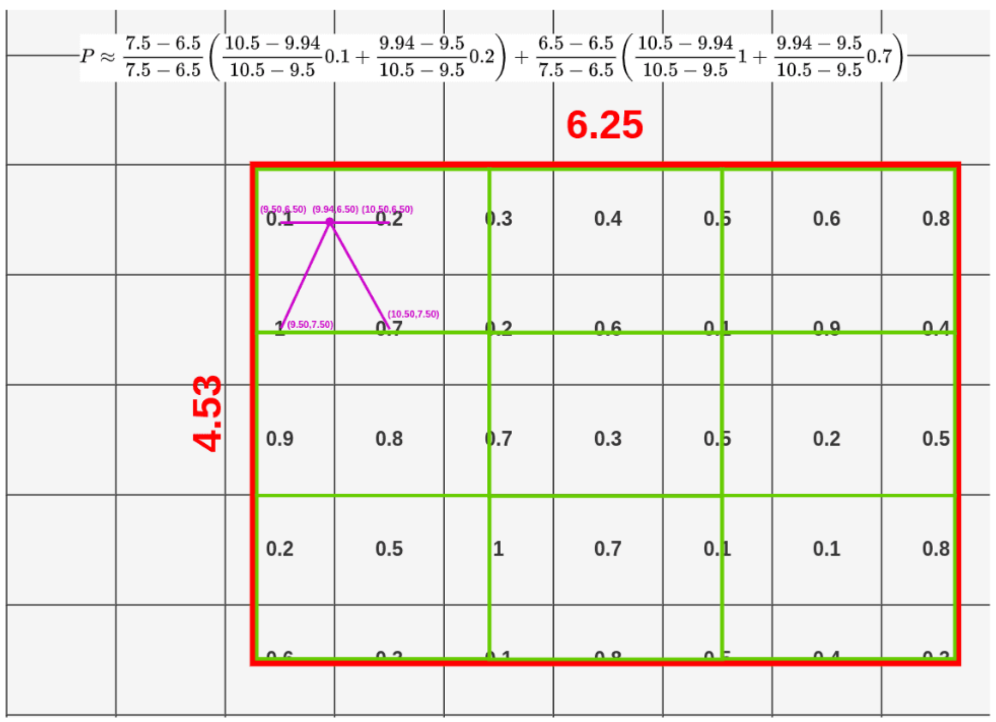
- X = X_box + (width/3) * 1 = 9.94
- Y = Y_box + (height/3) * 2 = 7.01

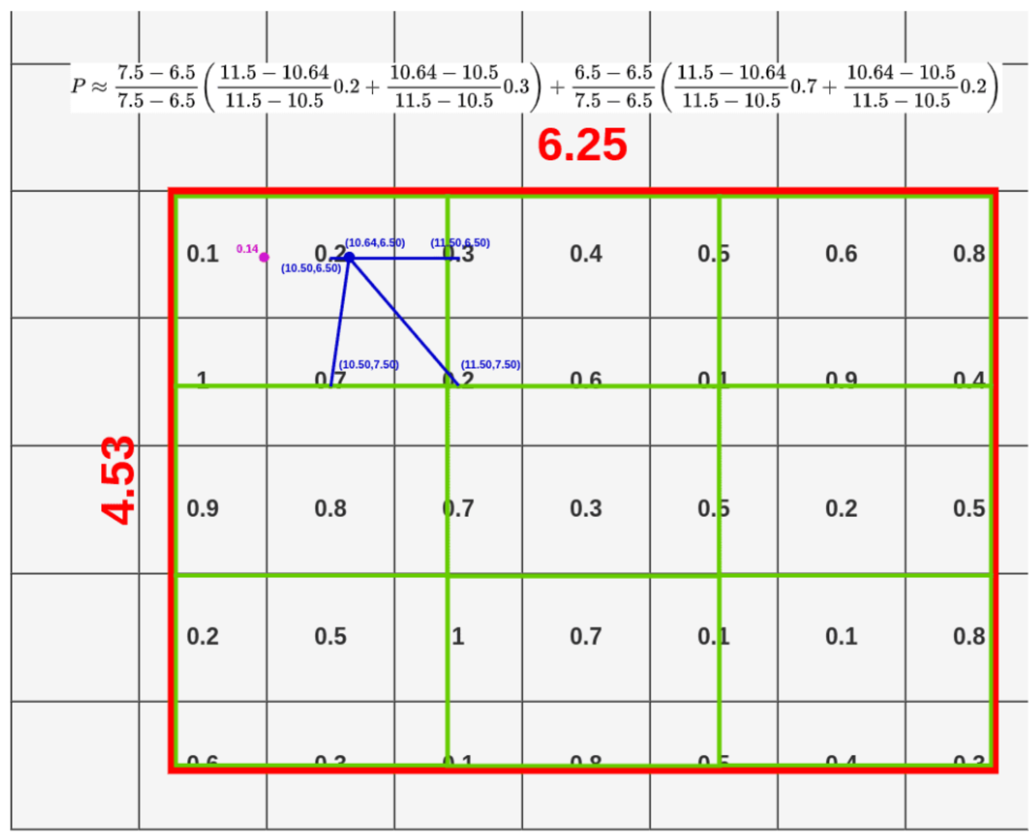
Mask RCNN: Region of Interest Alignment

Now we apply bilinear interpolation for each of the sampling points

$$P \approx \frac{y_2 - y}{y_2 - y_1}\left(\frac{x_2 - x}{x_2 - x_1}Q_{11} + \frac{x - x_1}{x_2 - x_1}Q_{21}\right) + \frac{y - y_1}{y_2 - y_1}\left(\frac{x_2 - x}{x_2 - x_1}Q_{12} + \frac{x - x_1}{x_2 - x_1}Q_{22}\right)$$

Bilinear Interpolation equation





UniGe

MaLGa