CS6501: Deep Learning for Visual Recognition

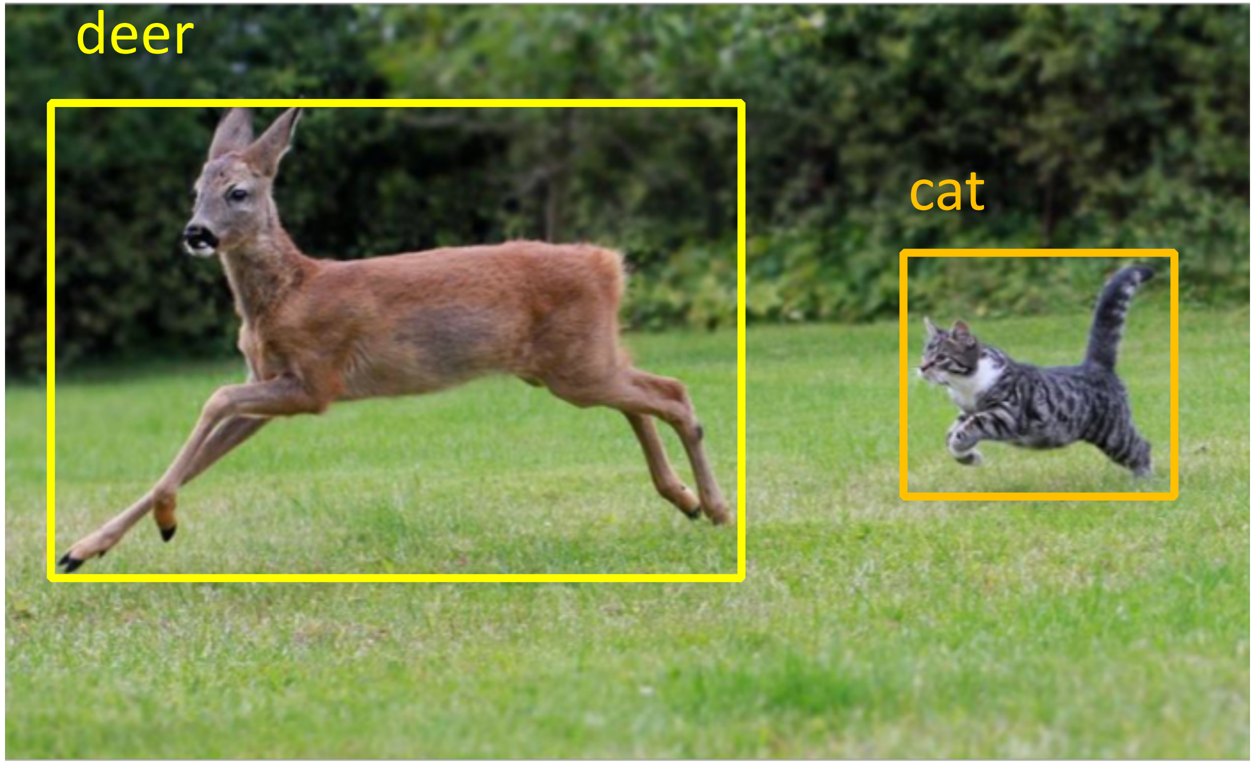
Object Detection I:

RCNN, Fast-RCNN, Faster-RCNN

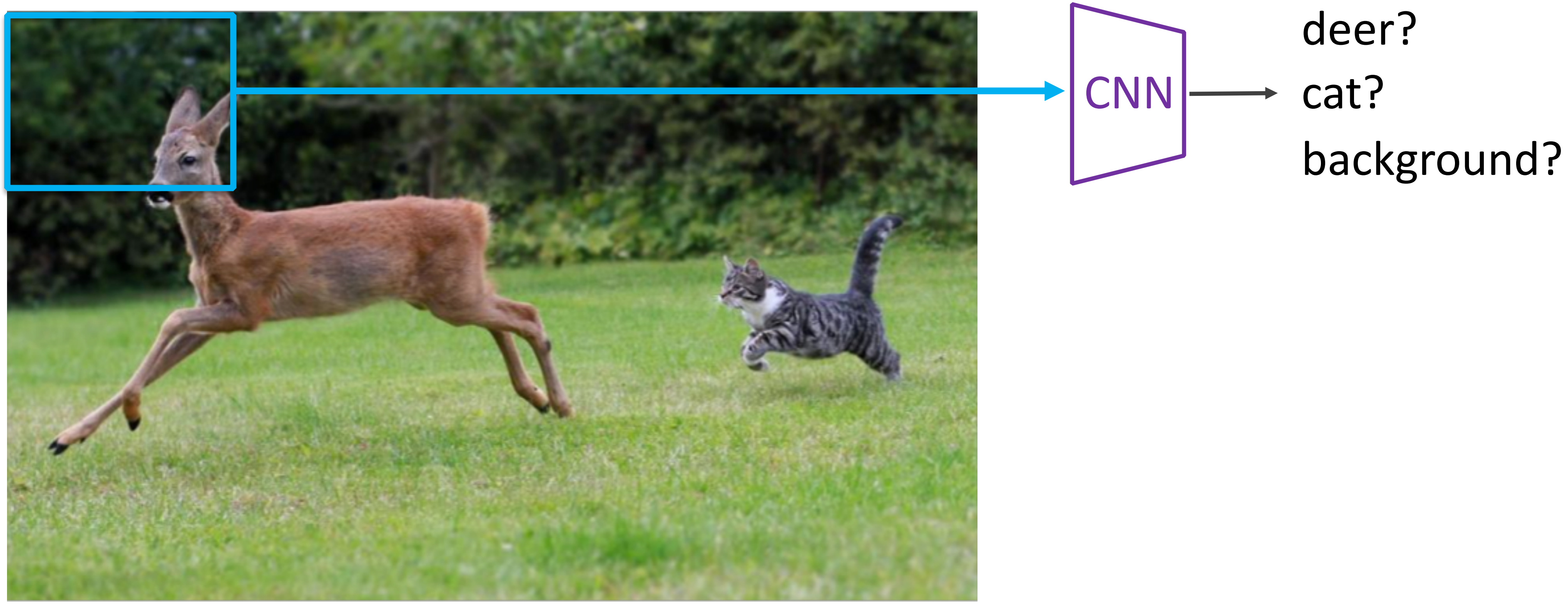
Today's Class

- Object Detection
- The RCNN Object Detector (2014)
- The Fast RCNN Object Detector (2015)
- The Faster RCNN Object Detector (2016)

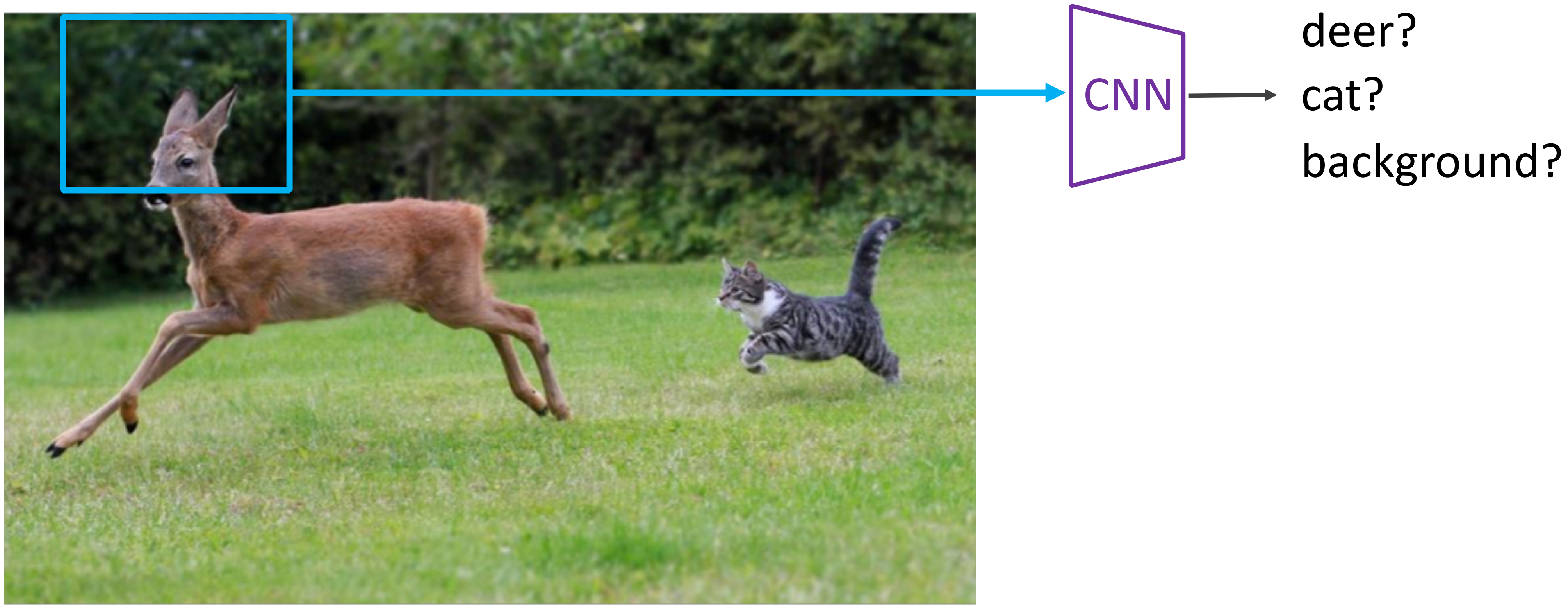
Object Detection



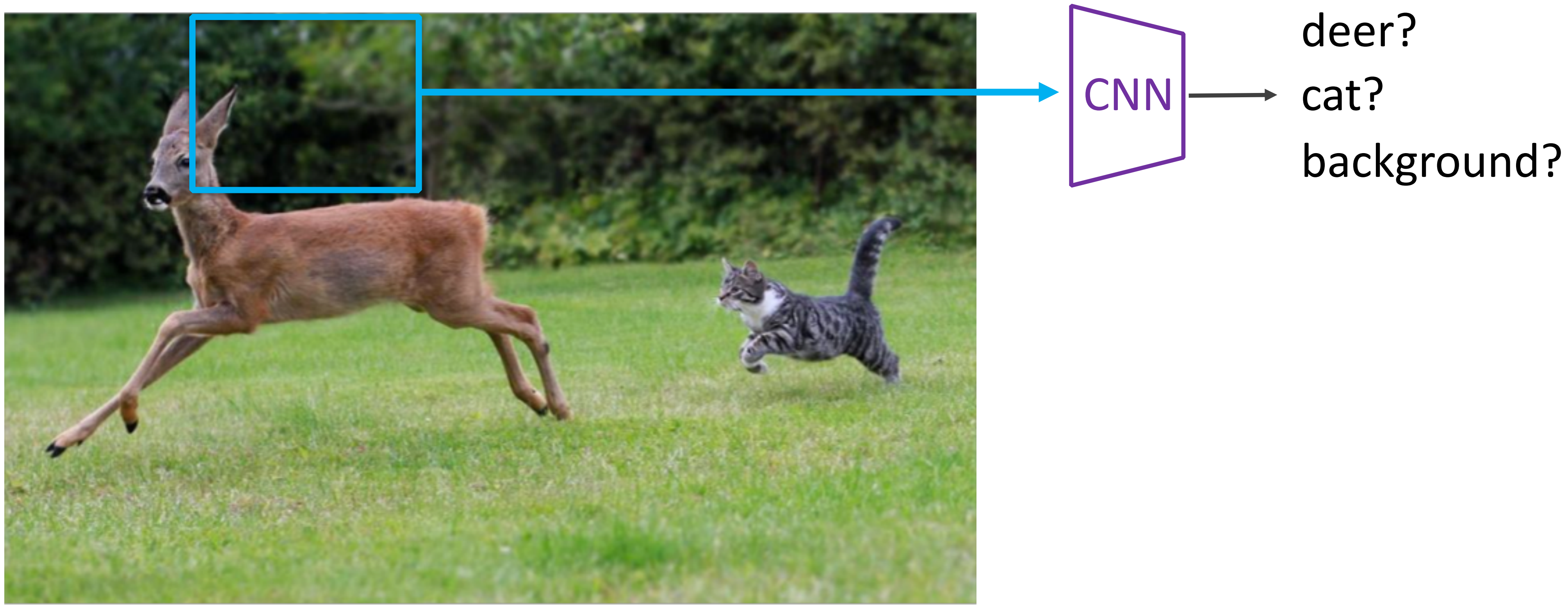
Object Detection as Classification



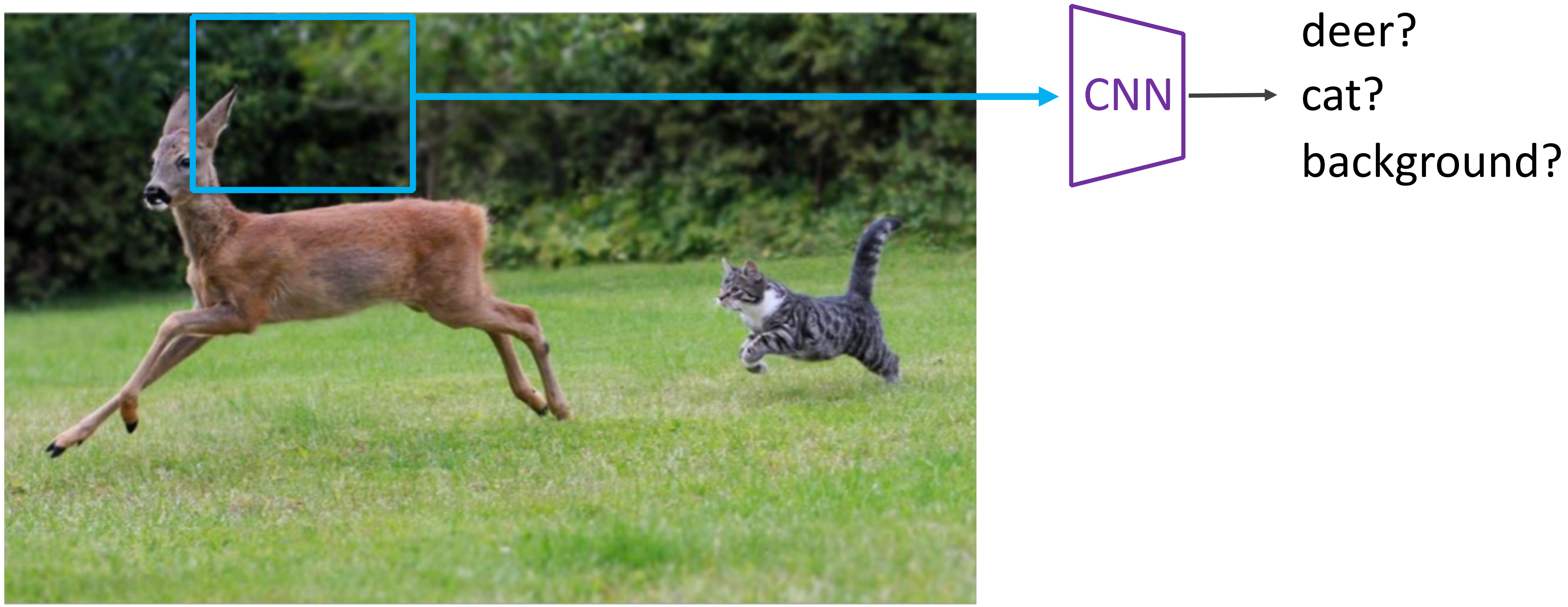
Object Detection as Classification



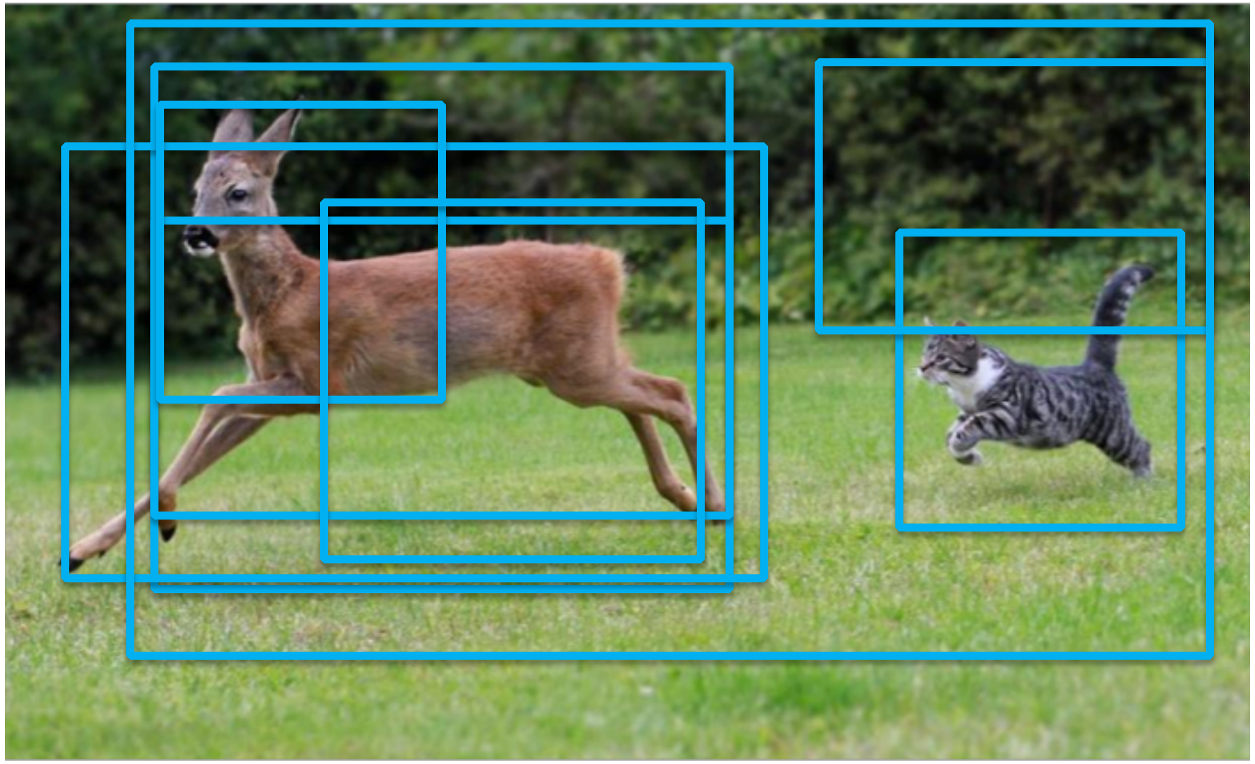
Object Detection as Classification



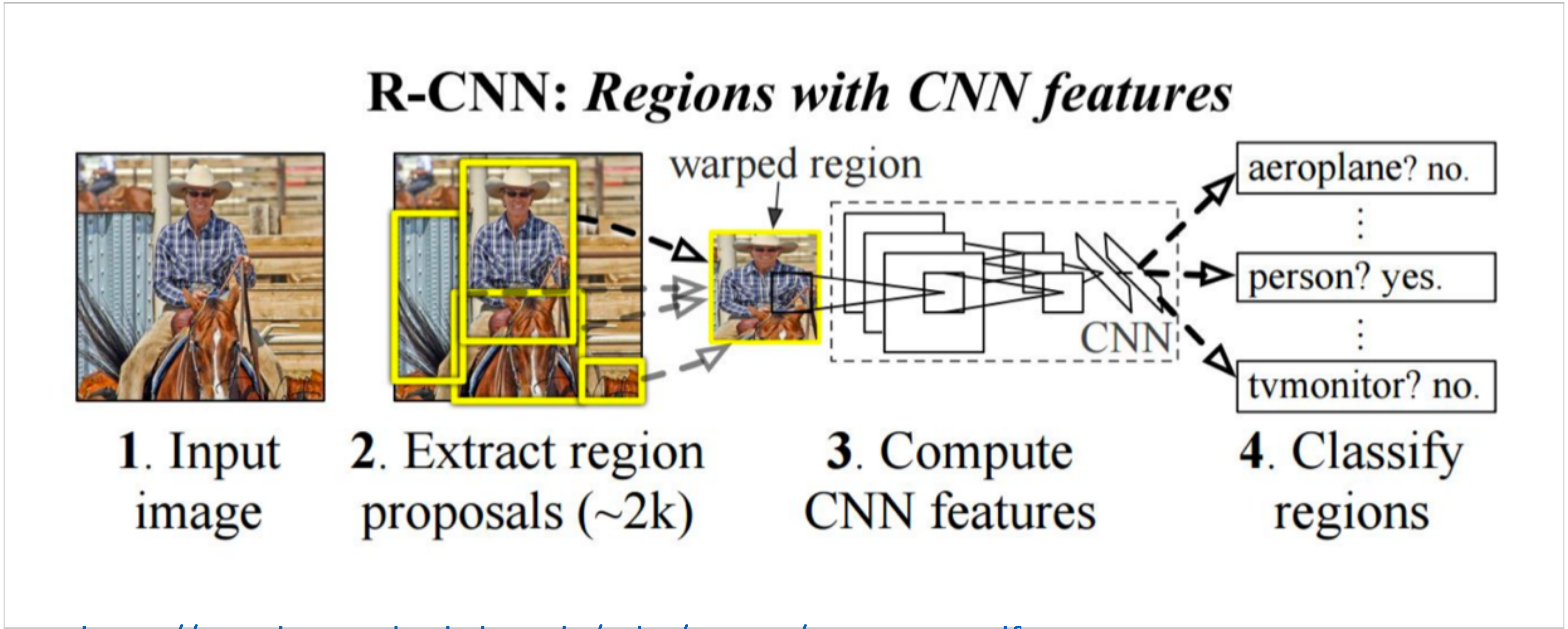
Object Detection as Classification with Sliding Window



Object Detection as Classification with Box Proposals



RCNN



https://people.eecs.berkeley.edu/~rbg/papers/r-cnn-cvpr.pdf

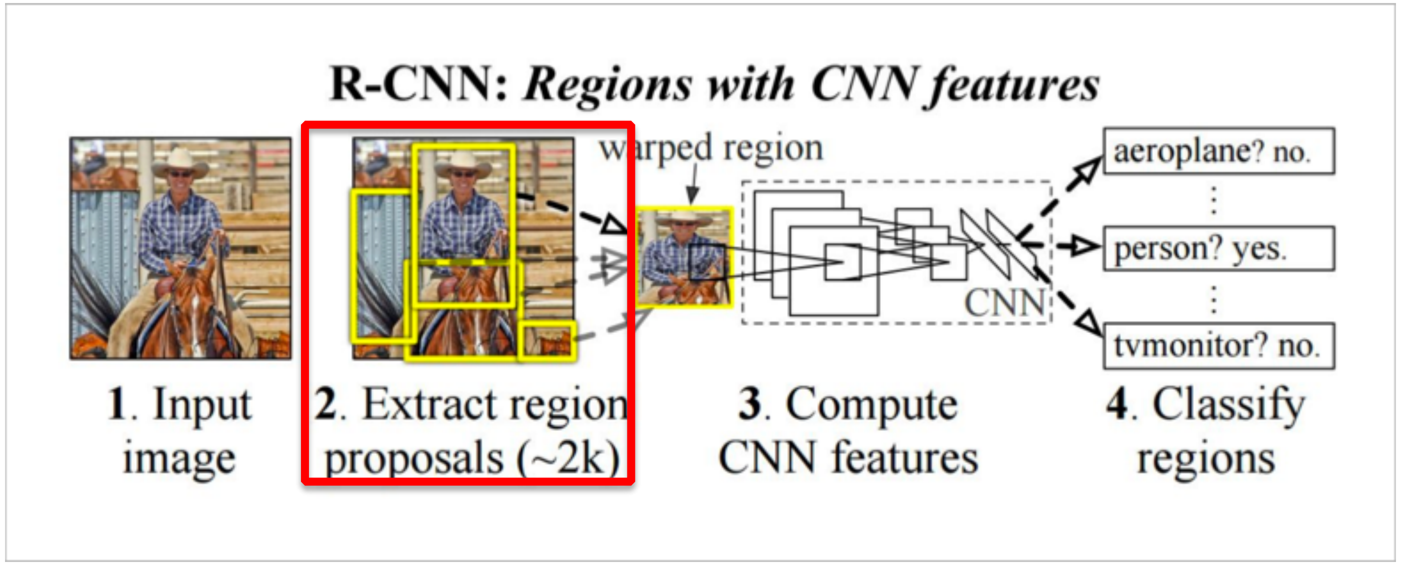
Rich feature hierarchies for accurate object detection and semantic segmentation.

Girshick et al. CVPR 2014.

RCNN

<u>First stage</u>: generate category-independent region proposals.

- 2000 Region proposals for every image



Selective Search: combine the strength of both an exhaustive search and segmentation. Uijlings et al. IJCV 2013.
ref

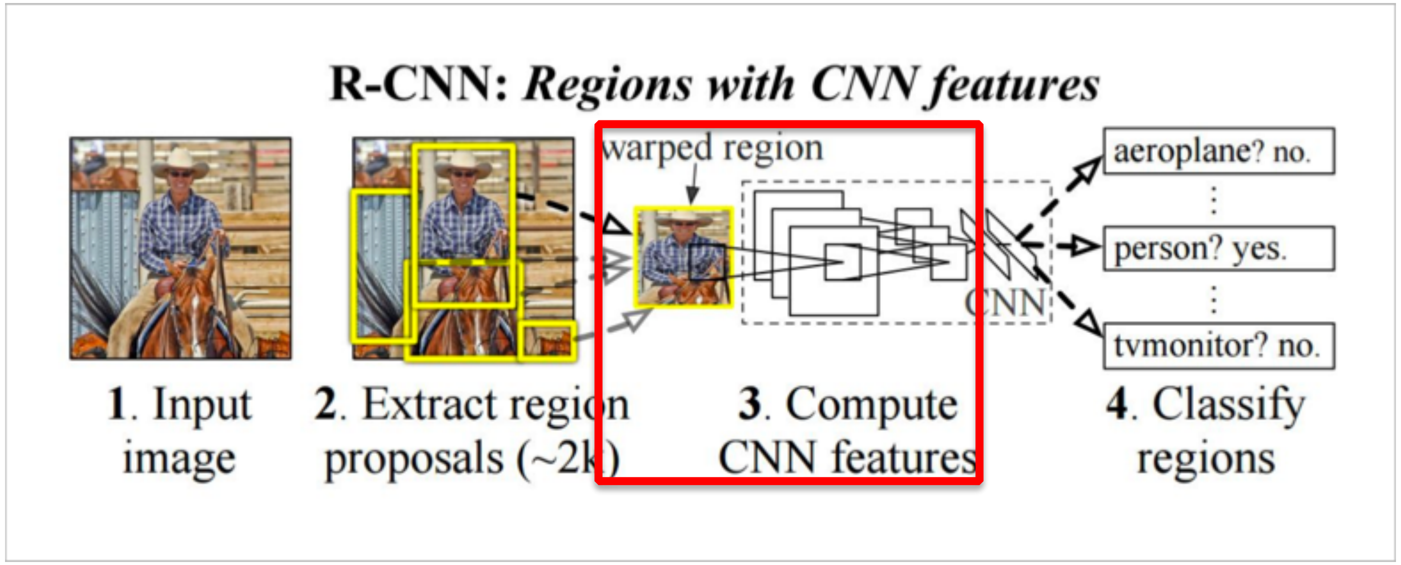
RCNN

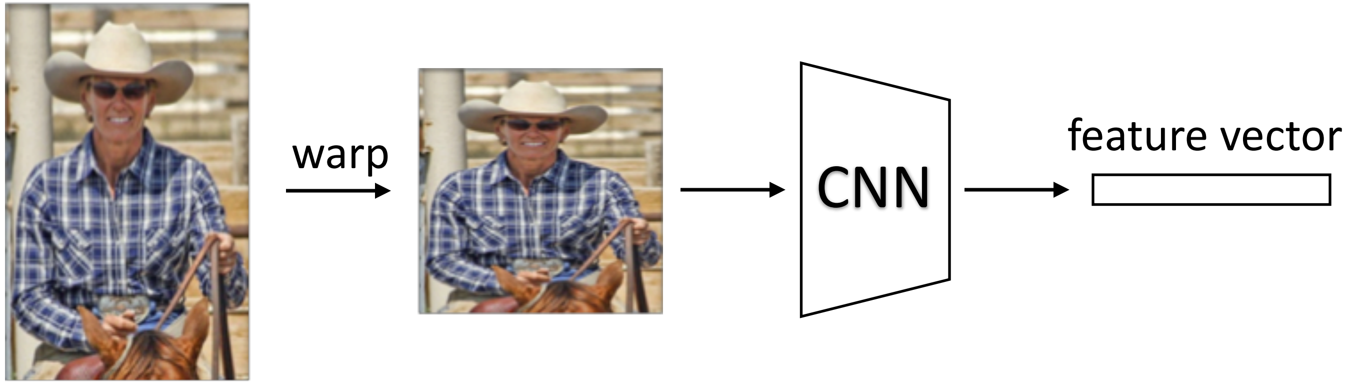
<u>First stage</u>: generate category-independent region proposals.

- 2000 Region proposals for every image

<u>Second stage</u>: extracts a fixed-length feature vector from each region.

- a 4096-dimensional feature vector from each region proposal





Arbitrary rectangles?
A fixed size input? 227 x 227

5 conv layers + 2 fully connected layers

RCNN

<u>First stage</u>: generate category-independent region proposals.

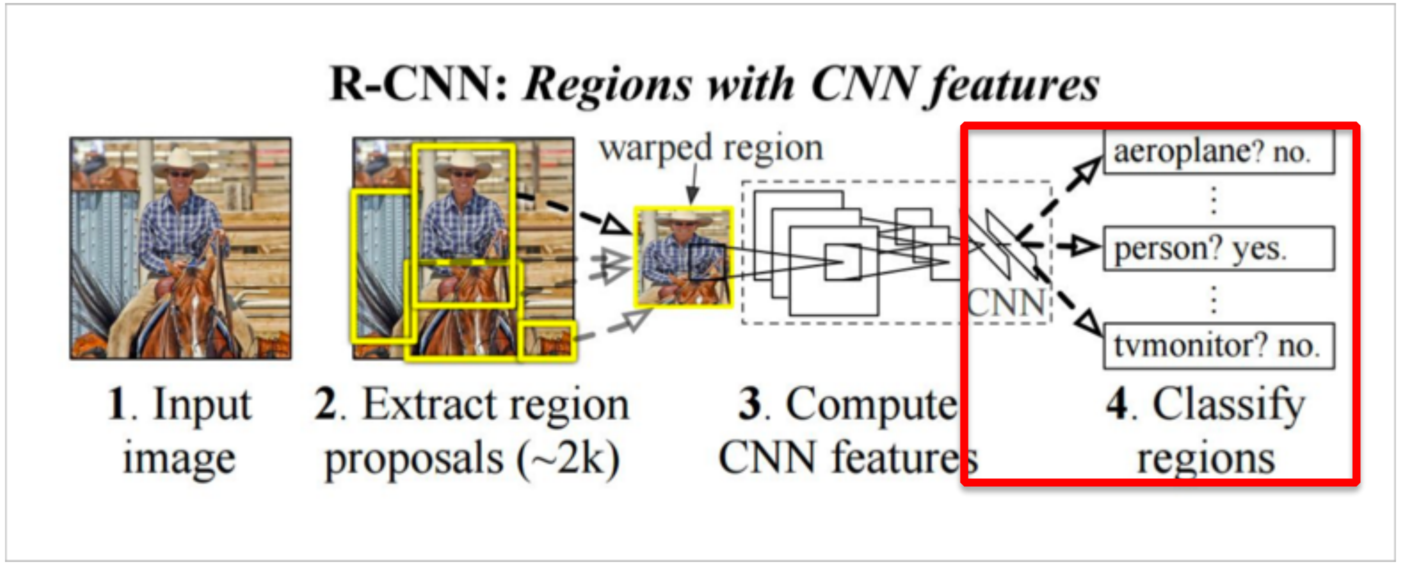
- 2000 Region proposals for every image

<u>Second stage</u>: extracts a fixed-length feature vector from each region.

- a 4096-dimensional feature vector from each region proposal

<u>Third stage</u>: a set of class- specific linear SVMs.

- object category and location



feature vector → linear svm → people? horse? background?

feature vector, proposal location → Bounding box regression → x y w h

proposal location

RCNN

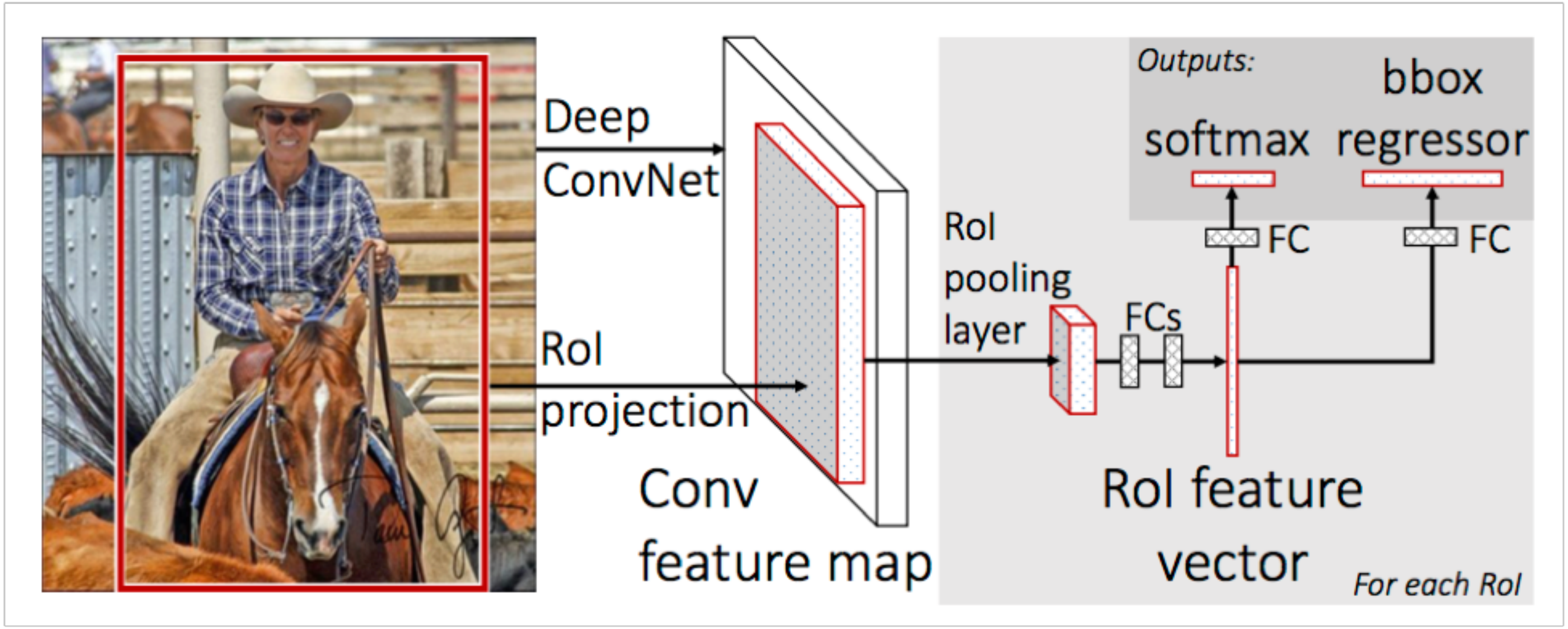
- Simple and scalable.
- improves mAP.

- A multistage pipeline.
- Training is expensive in space and time (features are extracted from each region proposal in each image and written into disk).
- Object detection is slow.

Fast-RCNN

?

Fast-RCNN

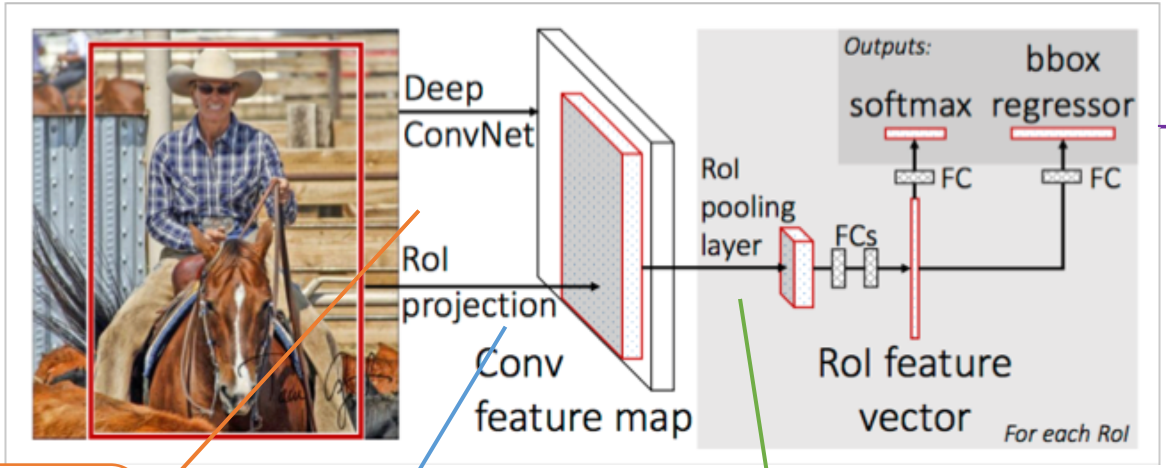


Idea: No need to recompute features for every box independently

https://arxiv.org/abs/1504.08083

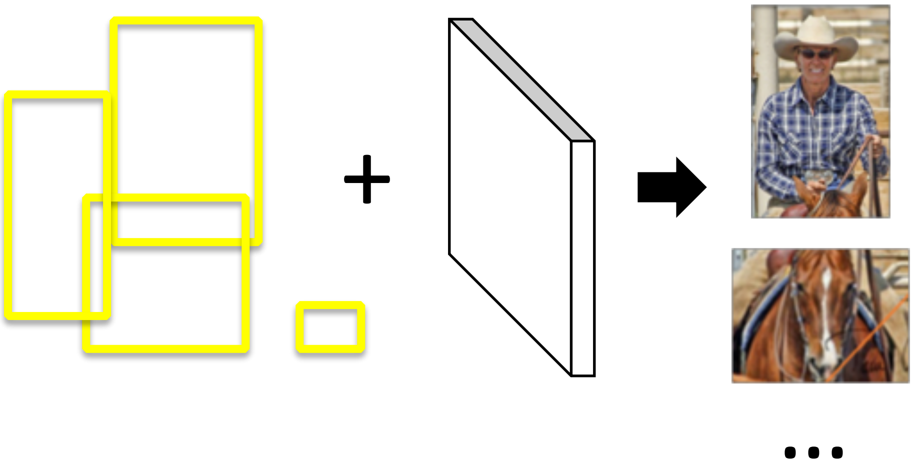
Fast R-CNN. Girshick. ICCV 2015.

Fast-RCNN

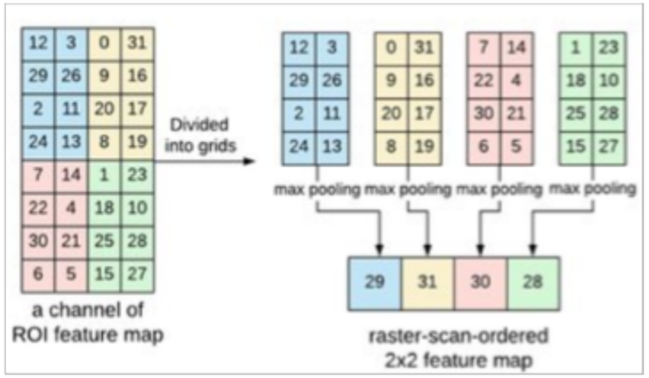


Process the whole image with several convolutional (*conv*) and max pooling layers to produce a conv feature map.

a region of interest (*RoI*) pooling layer extracts a fixed-length feature vector from the region feature map.



feature vector



FC+
softmax

K + 1 categories

FC+
regressor

four real-valued numbers for each of the K object classes.

RCNN

- Simple and scalable.
- improves mAP.

- A multistage pipeline.
- Training is expensive in space and time (features are extracted from each region proposal in each image and written into disk).
- Object detection is slow.

Fast-RCNN

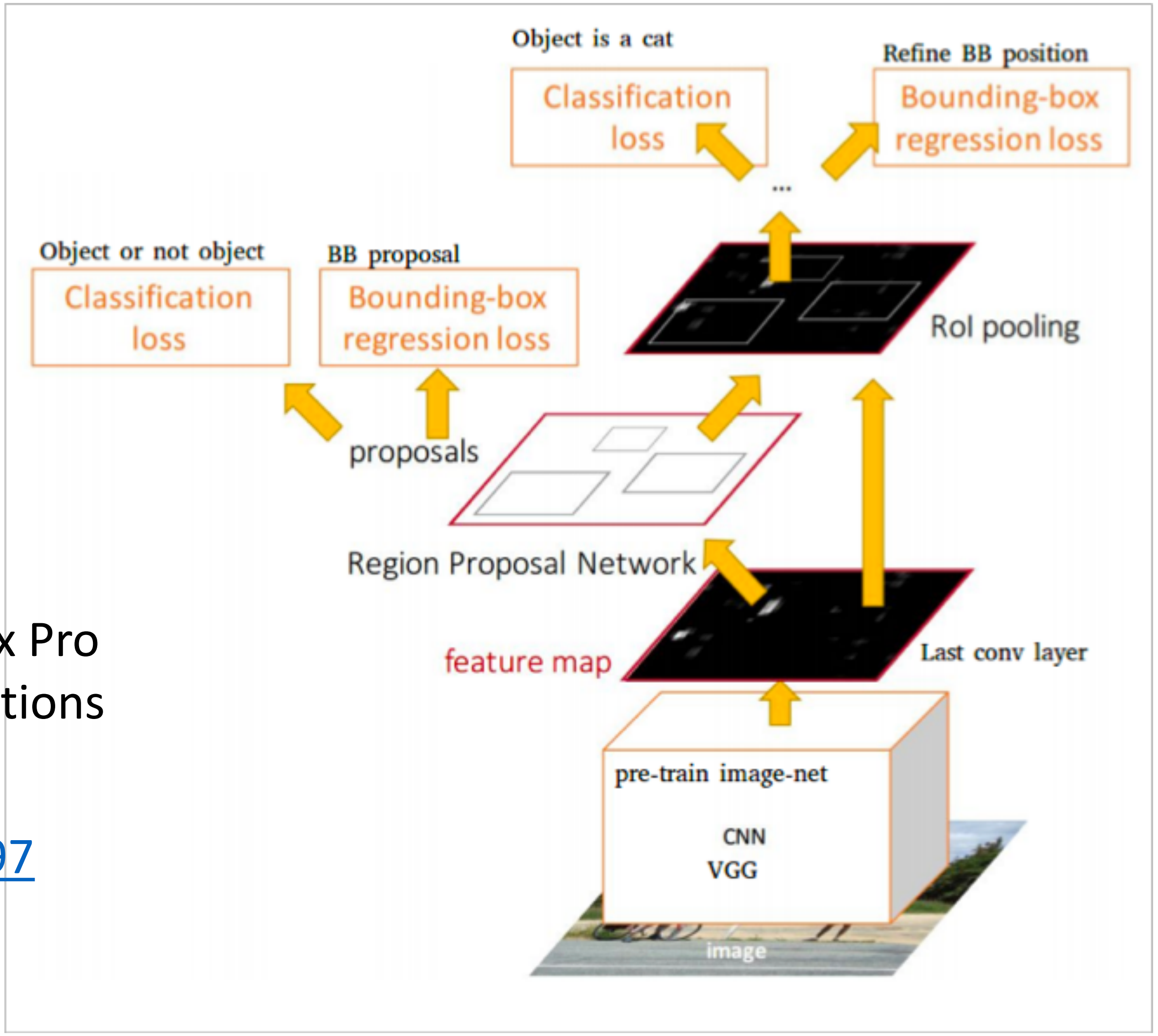
- Higher mAP.
- Single stage, end-to-end training.
- No disk storage is required for feature caching.

- proposals are the computational bottleneck in detection systems.

Faster-RCNN

?

Faster-RCNN

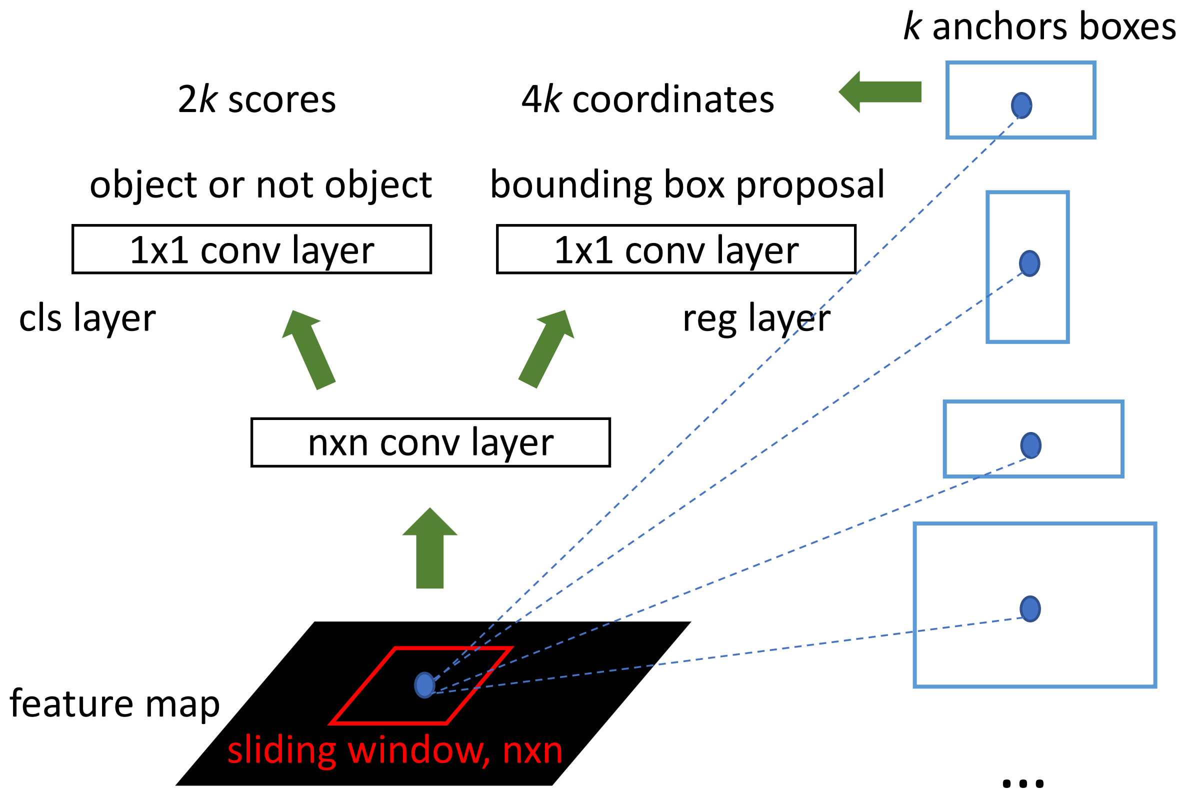


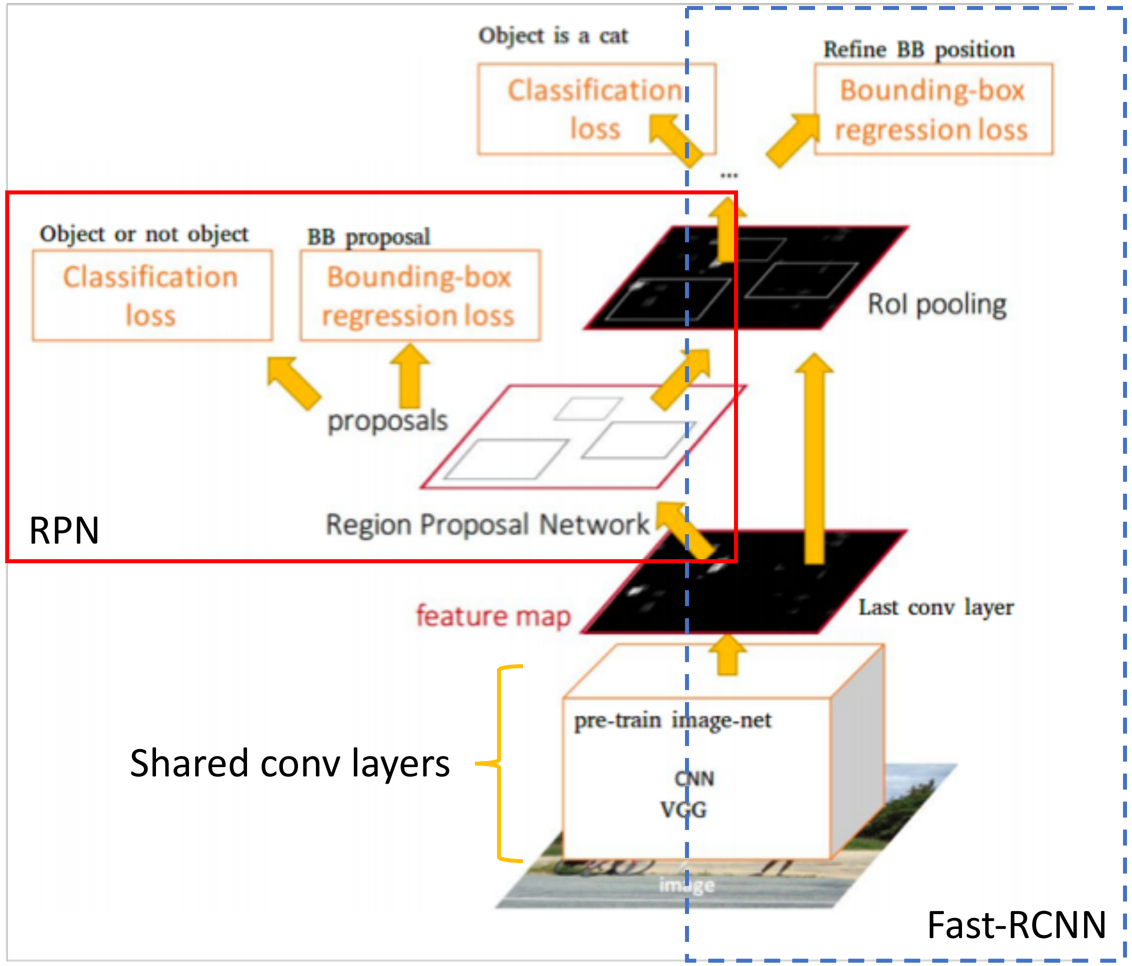
Idea: Integrate the Bounding Box Pro
posals as part of the CNN predictions

https://arxiv.org/abs/1506.01497
Ren et al. NIPS 2015.

Faster-RCNN

Region Proposal Networks:





RCNN

- Simple and scalable.
- improves mAP.

- A multistage pipeline.
- Training is expensive in space and time (features are extracted from each region proposal in each image and written into disk).
- Object detection is slow.

Fast-RCNN

- Higher mAP.
- Single stage, end-to-end training.
- No disk storage is required for feature caching.

- proposals are the computational bottleneck in detection systems.

Faster-RCNN

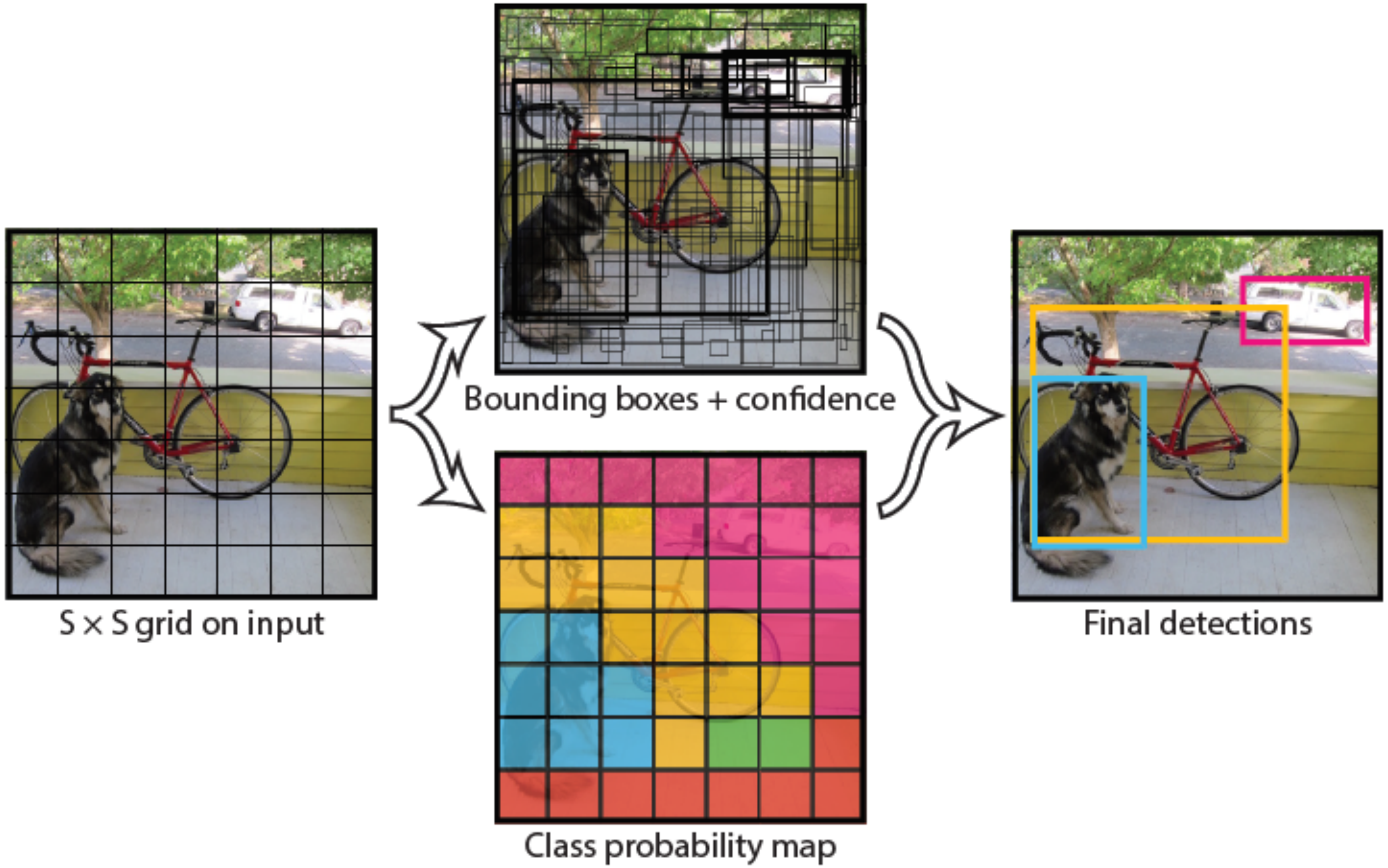
- compute proposals with a deep convolutional neural network --*Region Proposal Network* (RPN)
- merge RPN and Fast R-CNN into a single network, enabling nearly cost-free region proposals.

?

YOLO- You Only Look Once

Idea: No bounding box proposal. A single regression problem, stra ight from image pixels to boundi ng box coordinates and class pro babilities.

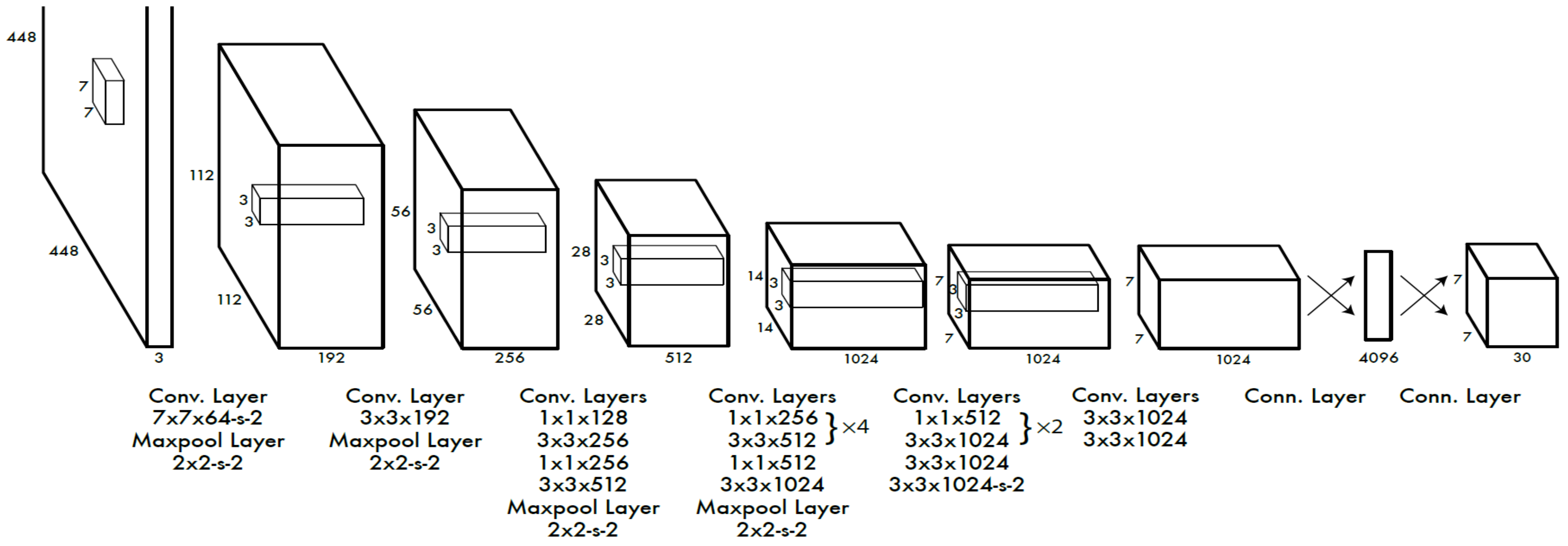
- extremely fast
- reason globally
- learn generalizable represent ations



https://arxiv.org/abs/1506.02640

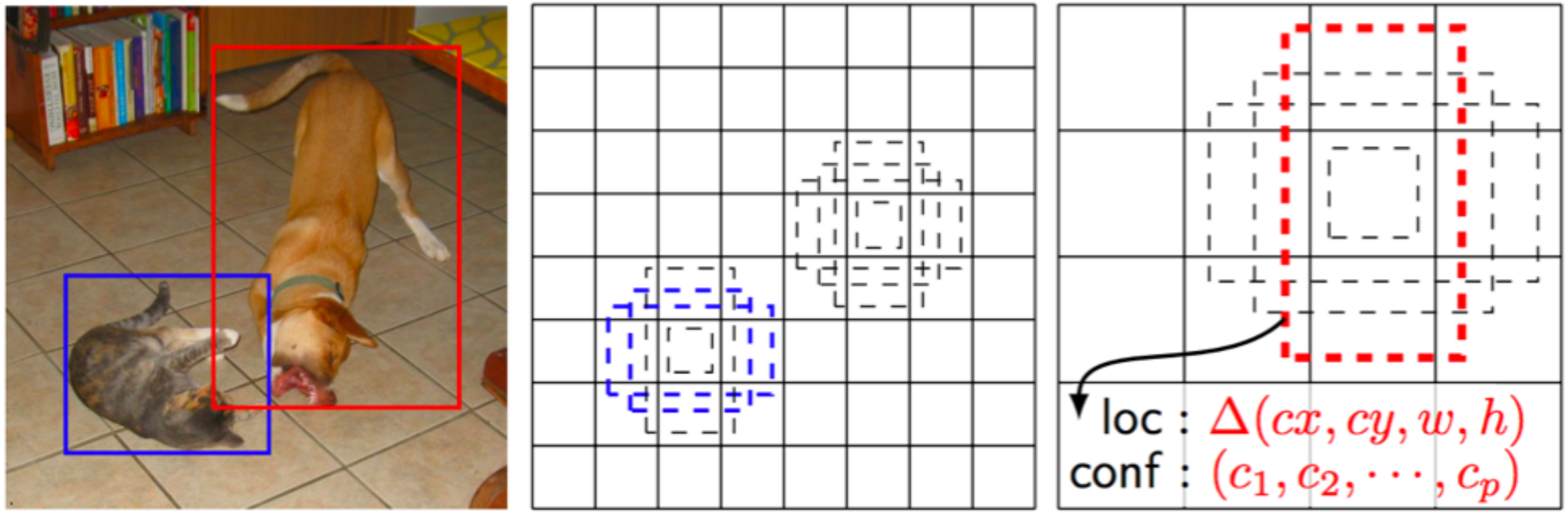
Redmon et al. CVPR 2016.

YOLO- You Only Look Once



Divide the image into 7x7 cells.
Each cell trains a detector.
The detector needs to predict the object's class distributions.
The detector has 2 bounding-box predictors to predict bounding-boxes and confidence scores.

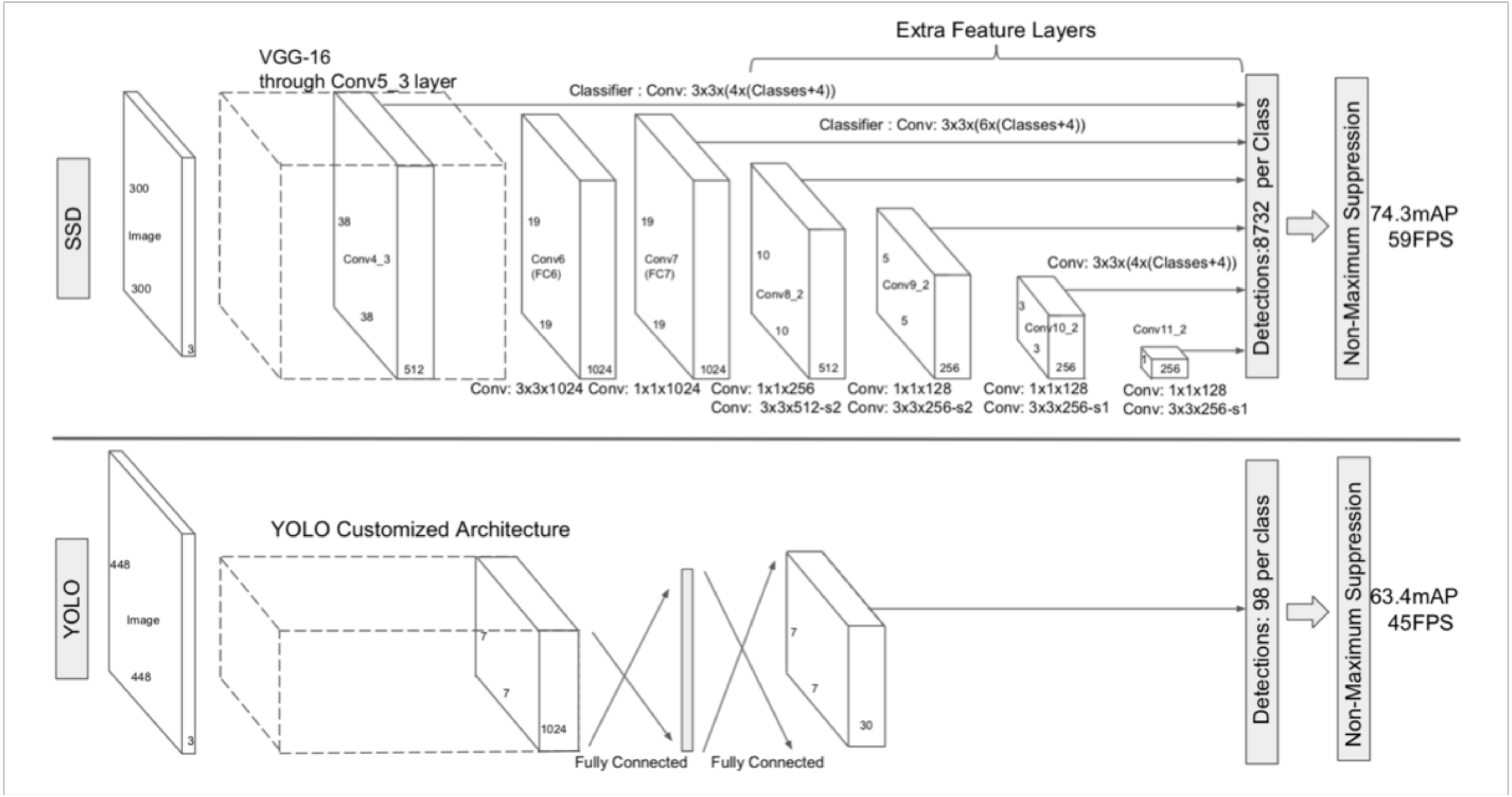
SSD: Single Shot Detector



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Idea: Similar to YOLO, but denser grid map, multiscale grid maps. + Data augme
ntation + Hard negative mining + Other design choices in the network.

Liu et al. ECCV 2016.


SSD
VGG-16
through Conv5_3 layer
Extra Feature Layers
Classifier : Conv: 3x3x(4x(Classes+4))
Classifier : Conv: 3x3x(6x(Classes+4))
Conv: 3x3x(4x(Classes+4))
300
Image
300
3
38
Conv4_3
38
512
19
Conv6
(FC6)
19
1024
19
Conv7
(FC7)
19
1024
10
Conv8_2
10
512
5
Conv9_2
5
256
3
Conv10_2
3
256
Conv11_2
1
256
Conv: 3x3x1024
Conv: 1x1x1024
Conv: 1x1x256
Conv: 3x3x512-s2
Conv: 1x1x128
Conv: 3x3x256-s2
Conv: 1x1x128
Conv: 3x3x256-s1
Conv: 1x1x128
Conv: 3x3x256-s1
Detections:8732 per Class
Non-Maximum Suppression
74.3mAP
59FPS
YOLO
448
Image
448
3
YOLO Customized Architecture
7
7
1024
Fully Connected
Fully Connected
7
7
30
Detections: 98 per class
Non-Maximum Suppression
63.4mAP
45FPS

Questions?