
Computer vision

Computer Vision

Lecture 7: Object Detection

Dr. Dina Khattab

dina.khattab@cis.asu.edu.eg

Scientific Computing Department

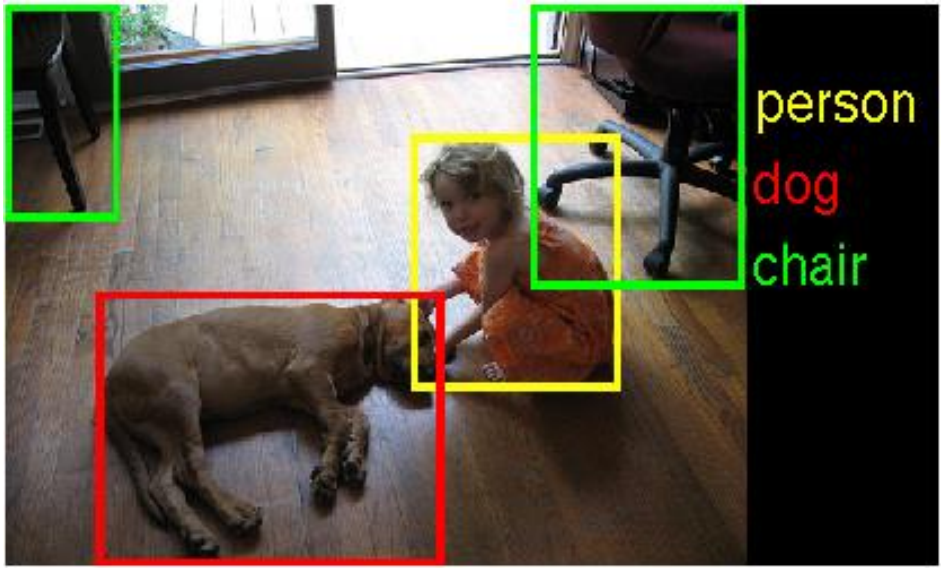
Instructor:	Dr. Dina Khattab
Email:	dina.khattab@cis.asu.edu.eg
Office:	Main Building – 4th floor – Room 302
Office Hours:	Monday 12:00 - 2:00 PM Thursday 11:00 AM to 12:00 PM

Agenda

- Object Detection Evaluation
- Sliding Window Object Detection
- Region-based Detectors
  - R-CNN
  - Fast R-CNN

Object Detection

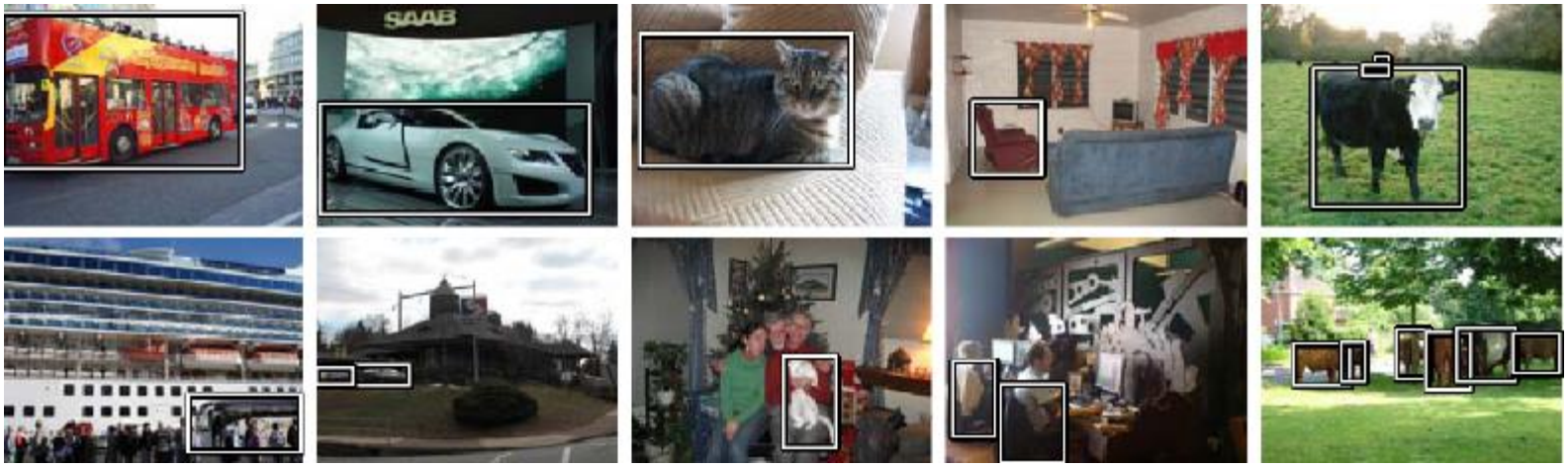
- **Problem**: Detecting and localizing generic objects from various categories, such as cars, people, etc.



- Challenges:
  - Illumination,
  - viewpoint,
  - deformations,
  - Intra-class variability

Object Detection Benchmarks

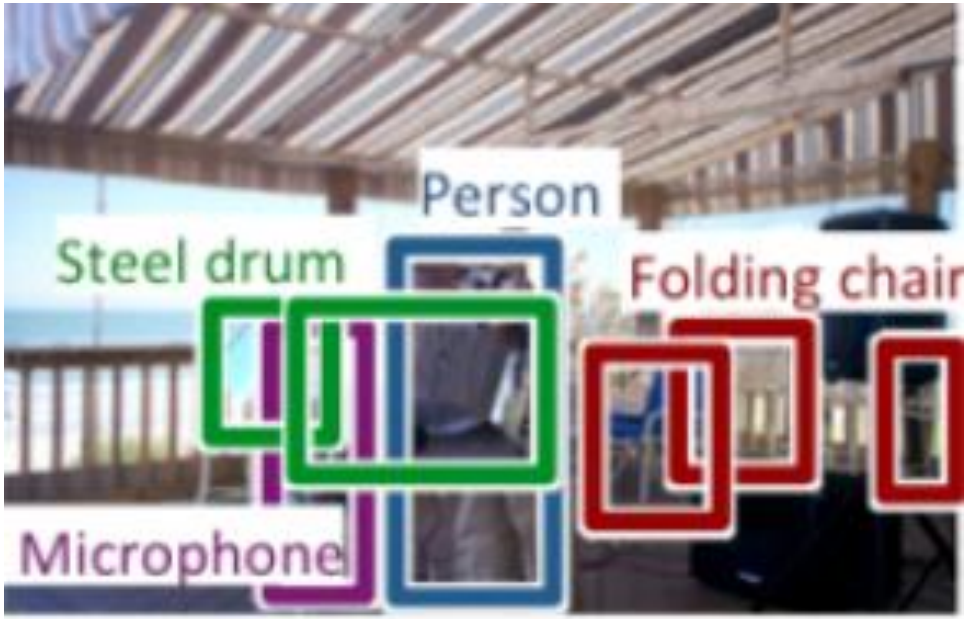
- PASCAL VOC Challenge



- 2005 to 2012
- Tested 20 categories
- High quality benchmark with high variability within each category

Object Detection Benchmarks

- ImageNet Large Scale Visual Recognition Challenge (ILSVR)
- 200 Categories for detection
- Had more variability in the object types
- Had many objects in a single image



Object Detection Benchmarks

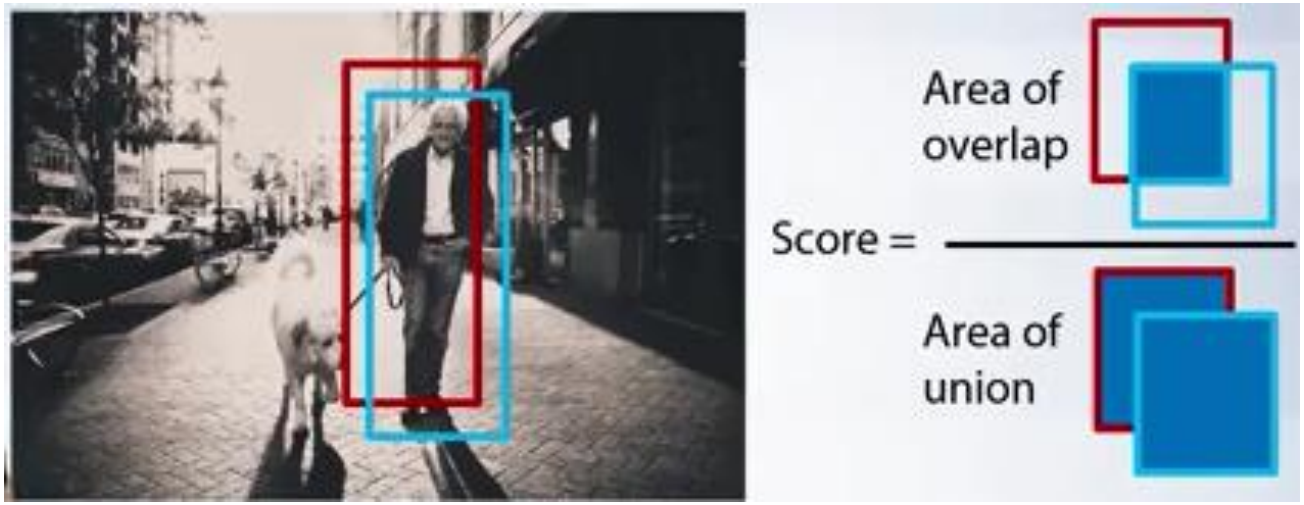
- Common Objects in Context (COCO)
- 80 Object categories
- Tests for segmentation in addition to detection with detailed bounding box

Object Detection Benchmarks

	PASCAL VOC (2010)	ImageNet Detection (ILSVRC 2014)	MS-COCO (2014)
Number of classes	20	**200**	80
Number of images (train + val)	~20k	**~470k**	~120k
Mean objects per image	2.4	1.1	**7.2**

Object Detection Evaluation
Intersection over Union — IoU

- Computes intersection over the union of the two bounding boxes of the ground truth and the predicted box by algorithm
- IoU is 1 if the predicted and the ground-truth bounding boxes overlap perfectly.



Object Detection Evaluation

- The ground truth is provided by humans who manually classify and locate objects in the images.

predictions

ground truth

Object Detection Evaluation

predictions

ground truth

True positive:

- The overlap of the prediction with the ground truth is MORE than 0.5

Object Detection Evaluation

predictions

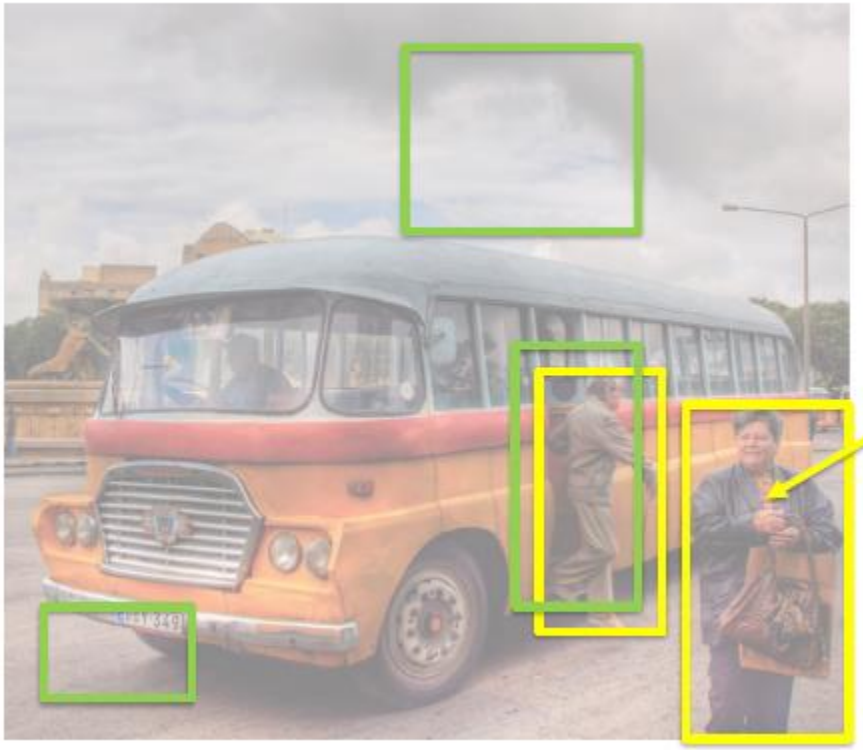
ground truth

True positive:

False positive:

- The overlap of the prediction with the ground truth is LESS than 0.5

Object Detection Evaluation



predictions

ground truth

True positive:

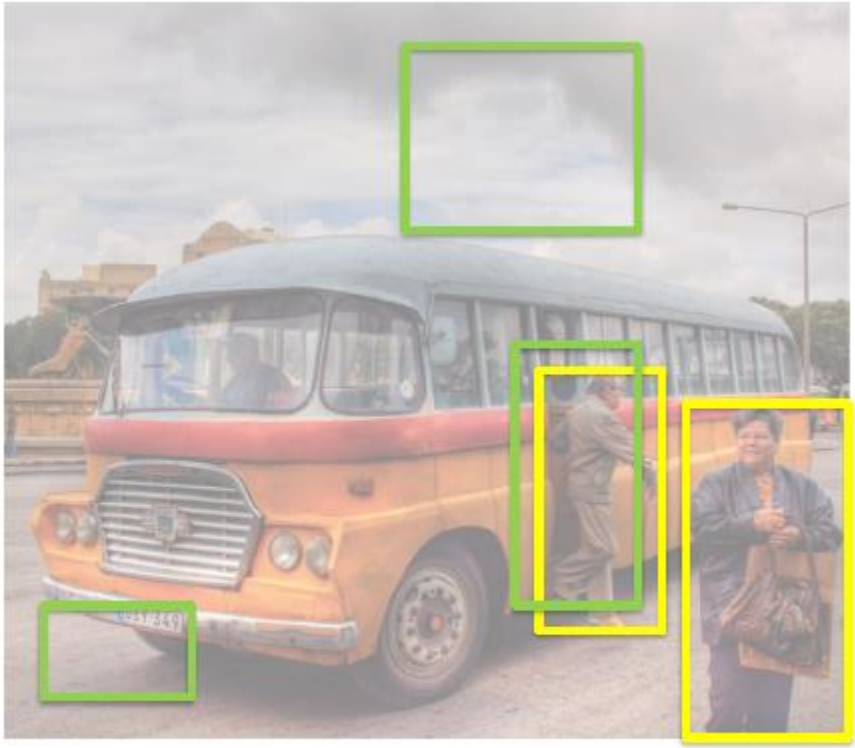
False positive:

False negative:

- The objects that our model doesn't find

Object Detection Evaluation

- True negatives are anywhere our algorithm didn't produce a box and the annotator did not provide a box.



predictions

ground truth

True positive:

False positive:

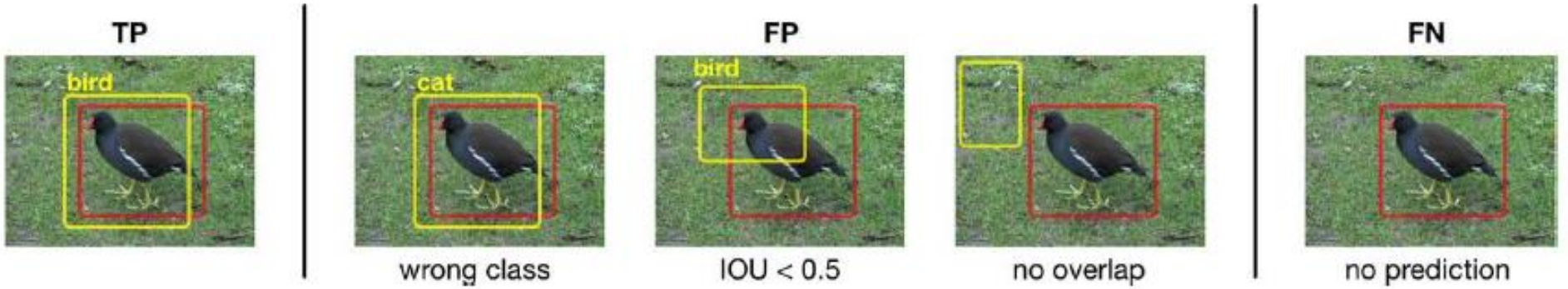
False negative:

- The objects that our model doesn't find

What is a True Negative?

Object Detection Evaluation

- True positives
- False positives
- False negatives



Object Detection Evaluation

- Precision: how many of the object detections are correct?
- Recall: how many of the ground truth objects can the model detect?

	Predicted 1	Predicted 0
True 1	true positive	false negative
True 0	false positive	true negative

	Predicted 1	Predicted 0
True 1	TP	FN
True 0	FP	TN

	Predicted 1	Predicted 0
True 1	hits	misses
True 0	false alarms	correct rejections

$$precision = \frac{TP}{TP + FP}$$

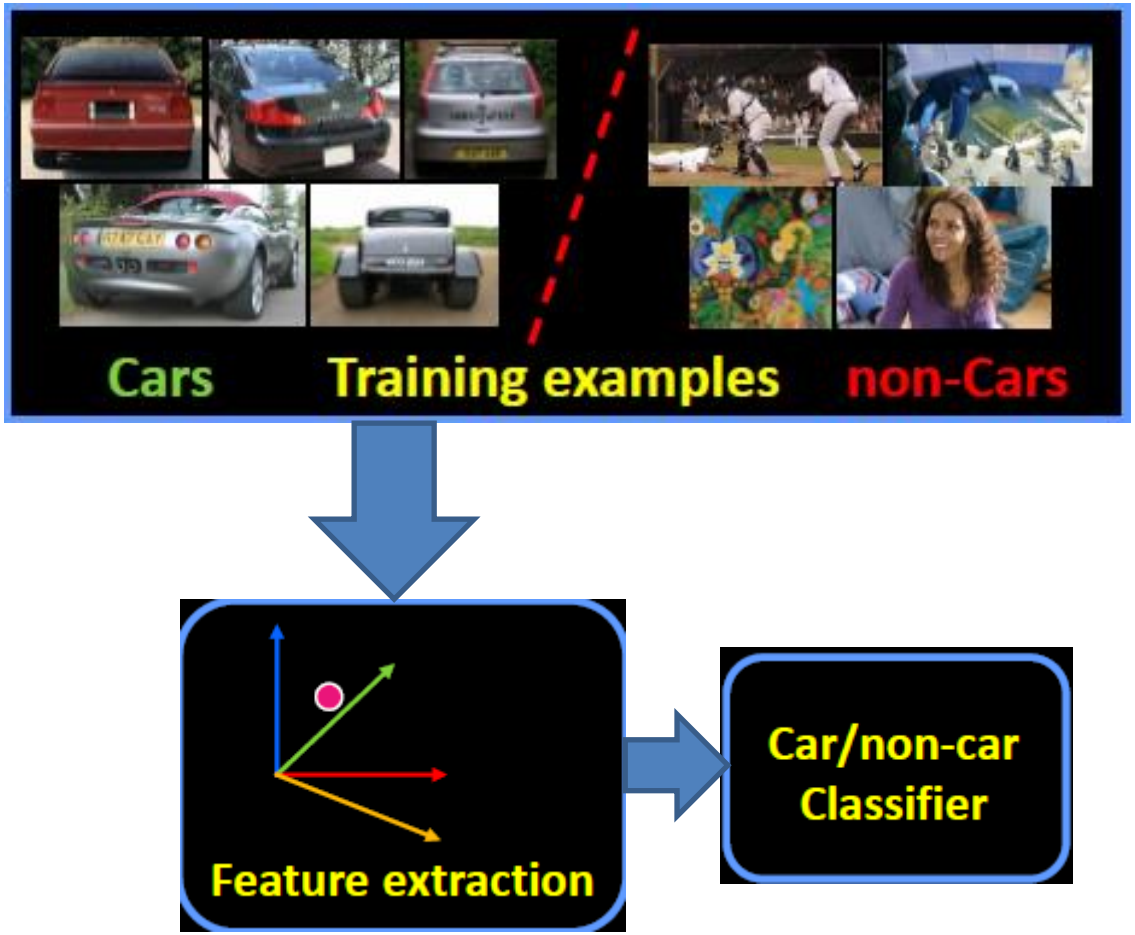
$$recall = \frac{TP}{TP + FN}$$

SLIDING WINDOW OBJECT DETECTION

Sliding Window-based object detection

Training:

1. Obtain training data
2. Define features
3. Train classifier

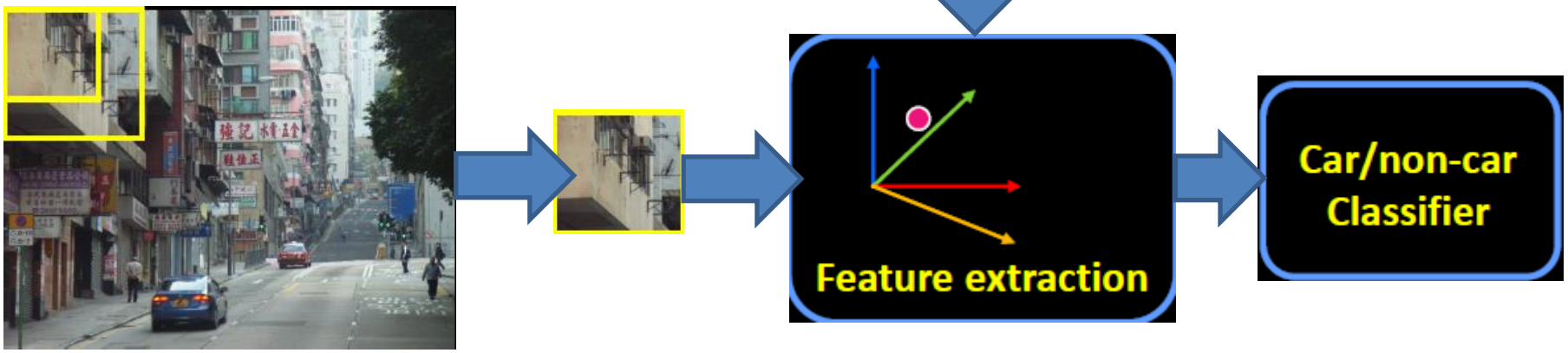


Sliding Window-based object detection

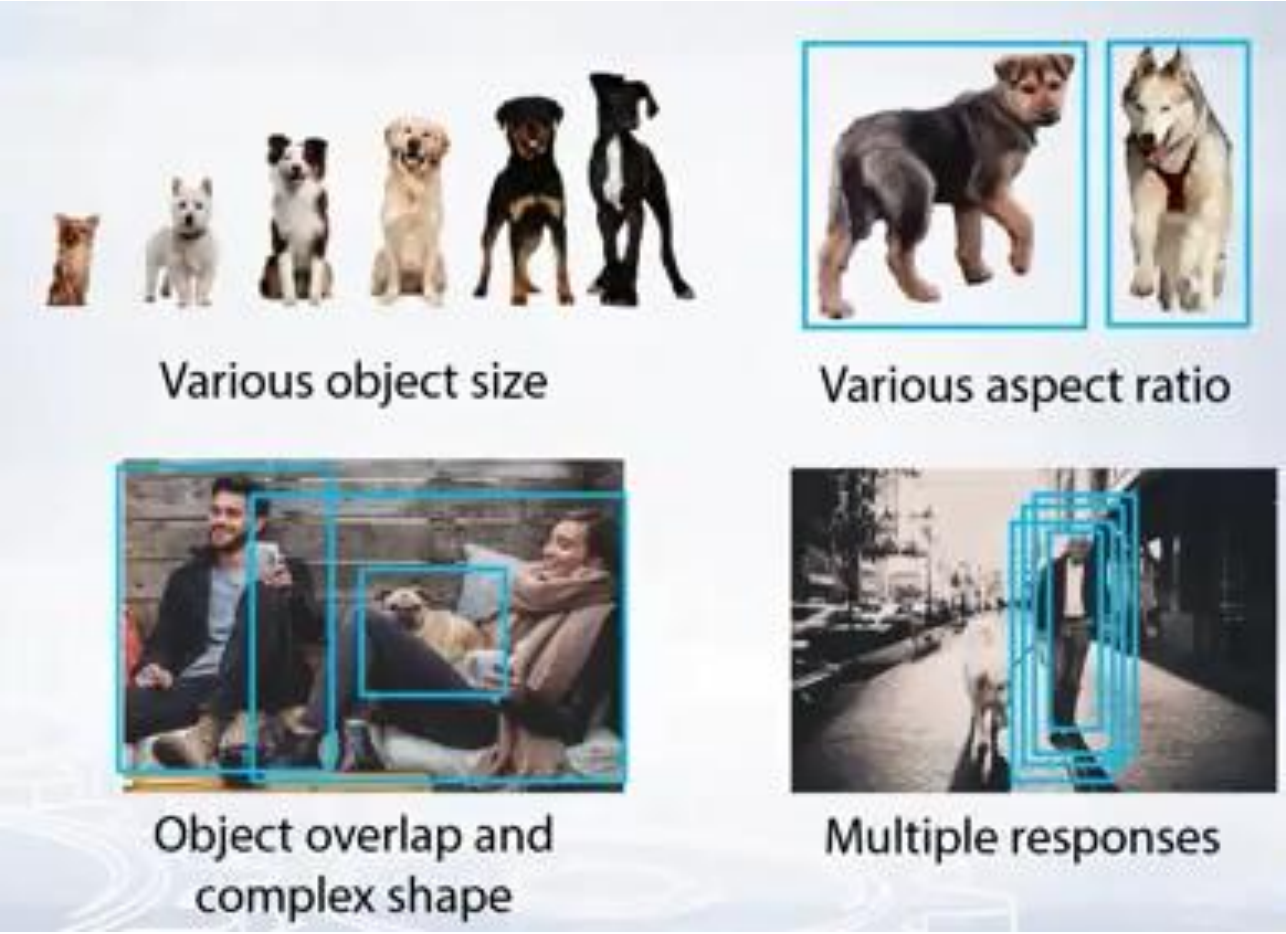
Given new image:

1.Slide window

2.Score by classifier



Problems of Sliding Window

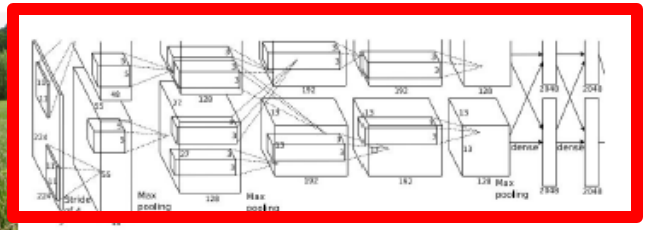


REGION-BASED CNNS

Object Detection: Single Object

(Classification + Localization)

This image is CC0 public domain



Vector: 4096

Fully Connected: 4096 to 1000

Class Scores
Cat: 0.9
Dog: 0.05
Car: 0.01
...

Softmax Loss

Correct label: Cat

Fully Connected: 4096 to 4

Box Coordinates (x, y, w, h)

L2 Loss

Correct box: (x', y', w', h')

Multitask Loss

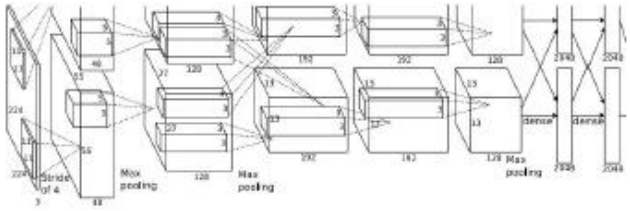
+ → Loss

Often pretrained on ImageNet (Transfer learning)

Treat localization as a regression problem!

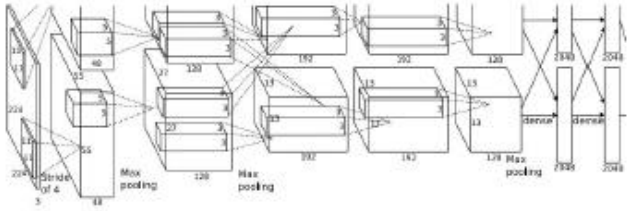
Object Detection: Multiple Objects

Each image needs a different number of outputs!



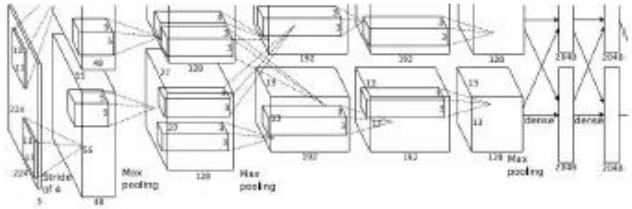
CAT: (x, y, w, h)

4 numbers



DOG: (x, y, w, h)
DOG: (x, y, w, h)
CAT: (x, y, w, h)

16 numbers

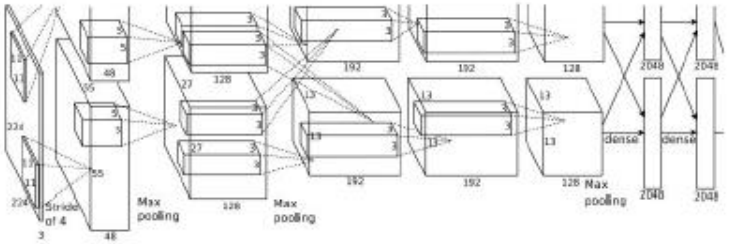


DUCK: (x, y, w, h)
DUCK: (x, y, w, h)
....

Many numbers!

Object Detection: Multiple Objects

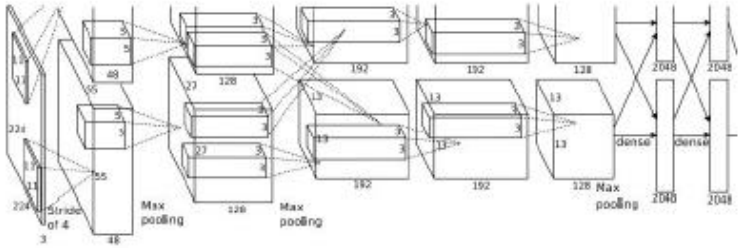
Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? NO
Cat? NO
Background? YES

Object Detection: Multiple Objects

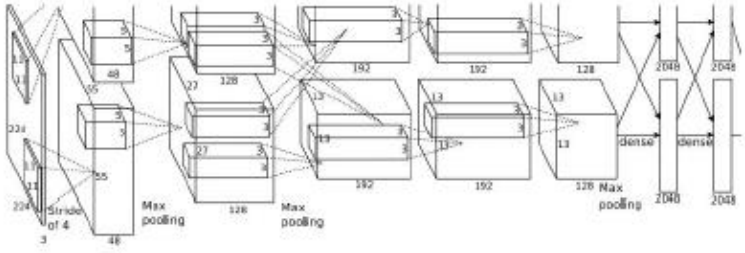
Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? YES
Cat? NO
Background? NO

Object Detection: Multiple Objects

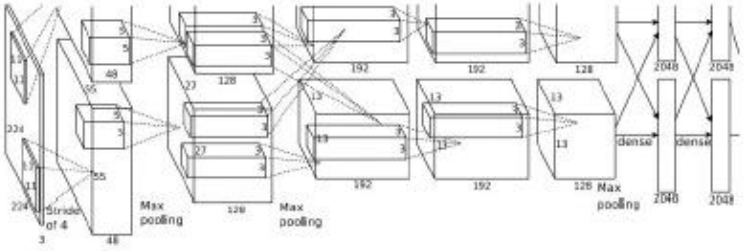
Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? YES
Cat? NO
Background? NO

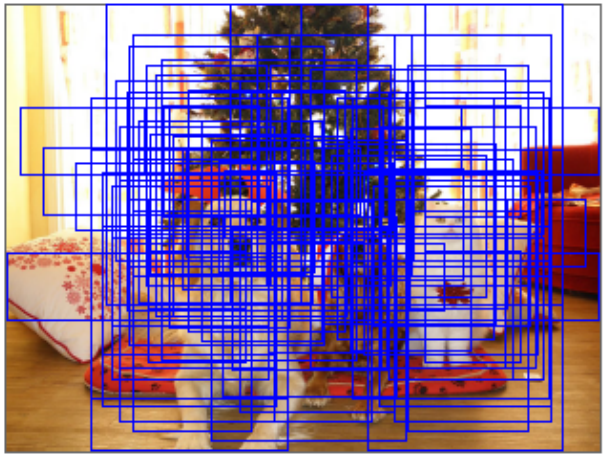
Object Detection: Multiple Objects

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

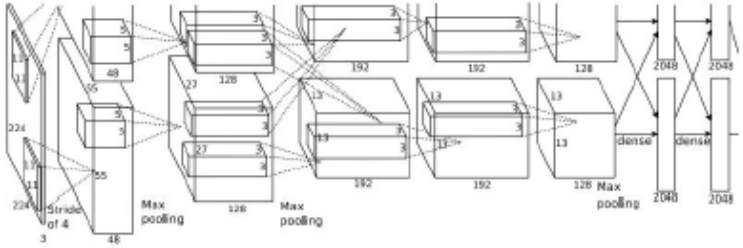


Dog? NO
Cat? YES
Background? NO

Object Detection: Multiple Objects



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

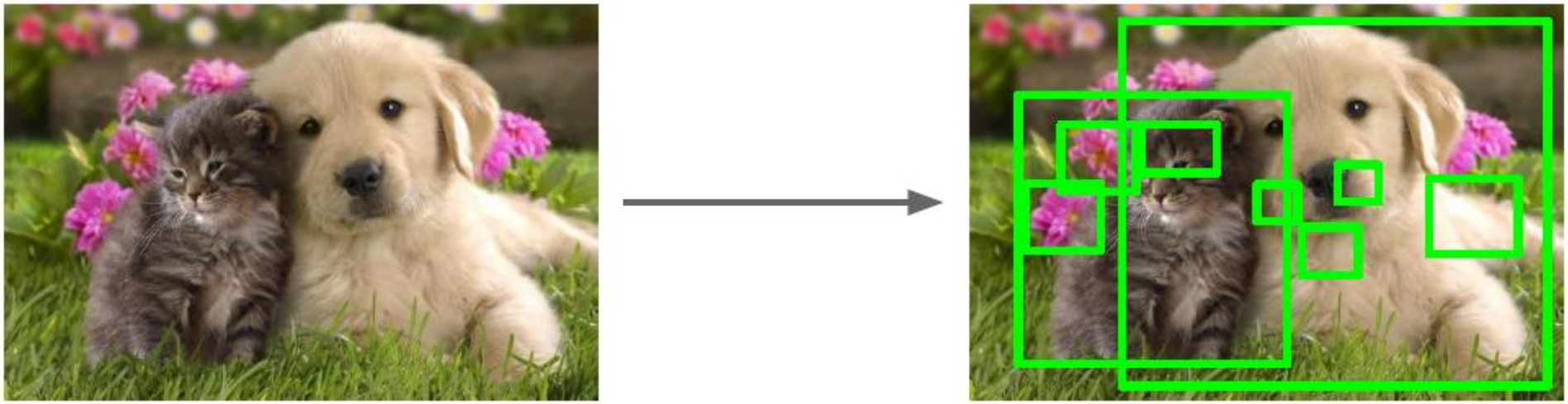


Dog? NO
Cat? YES
Background? NO

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

Region Proposals: Selective Search

- Find “blobby” image regions that are likely to contain objects
- Relatively fast to run; e.g. Selective Search gives 2000 region proposals in a few seconds on CPU



Alexe et al, “Measuring the objectness of image windows”, TPAMI 2012
Uijlings et al, “Selective Search for Object Recognition”, IJCV 2013
Cheng et al, “BING: Binarized normed gradients for objectness estimation at 300fps”, CVPR 2014
Zitnick and Dollar, “Edge boxes: Locating object proposals from edges”, ECCV 2014

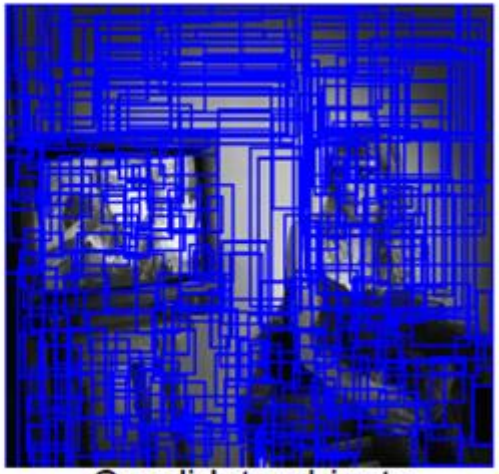
Selective Search

Step 1: Generate initial sub-segmentation

- Goal: Generate many regions, each of which belongs to at most one object.

Input Image

Segmentation



Candidate objects

JRR Uijlings et. al, Selective Search for Object Recognition, IJCV 2013

Selective Search

Step 2: Recursively combine similar regions into larger ones.

Greedy algorithm:

1. From set of regions, choose two that are most similar.
2. Combine them into a single, larger region.
3. Repeat until only one region remains.

- This yields a hierarchy of successively larger regions, just like we want.

Input Image | Initial Segmentation | After some iterations | After more iterations

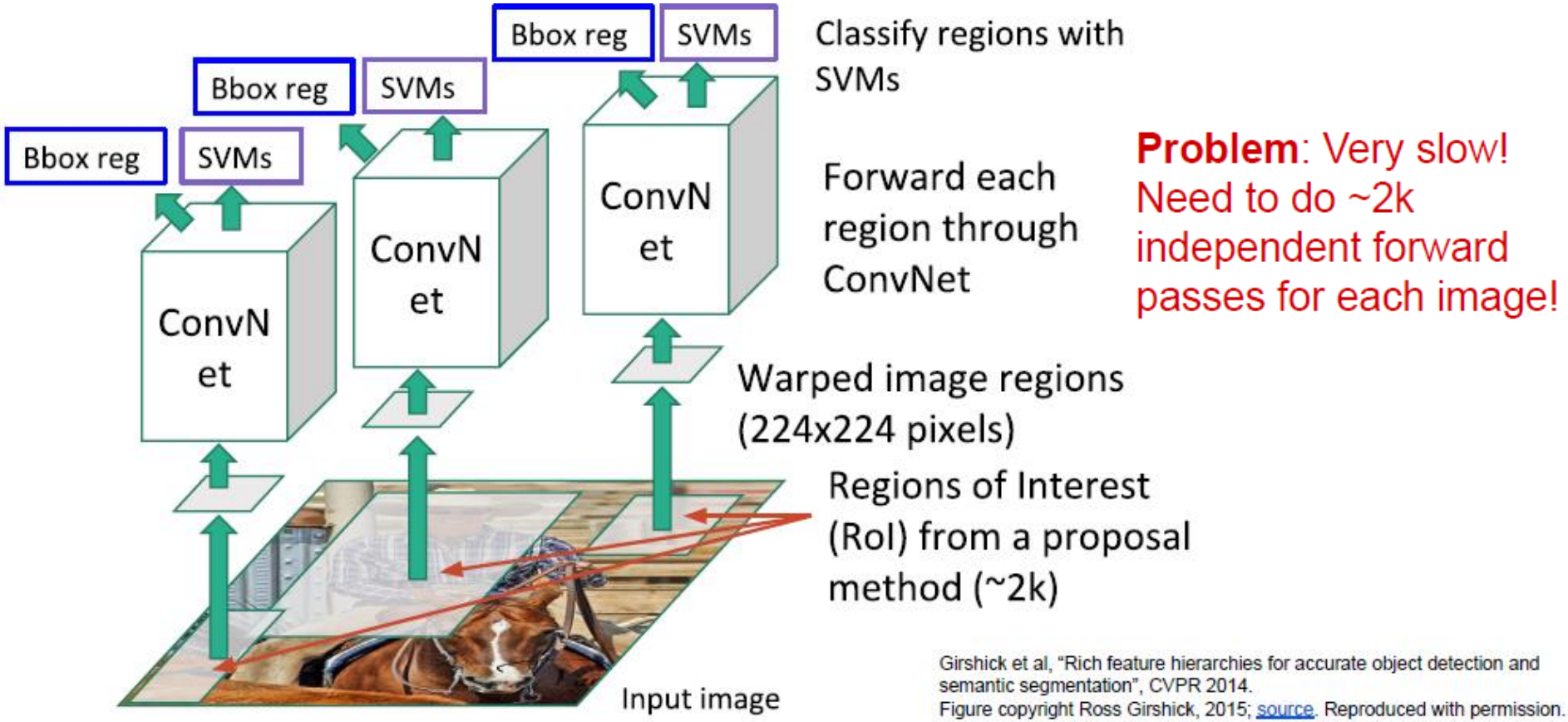
Selective Search

Step 3: Use the generated regions to produce candidate object locations.

Input Image

R-CNN

- R-CNN uses 2000 regions for an image based on selective search



Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

RCNN

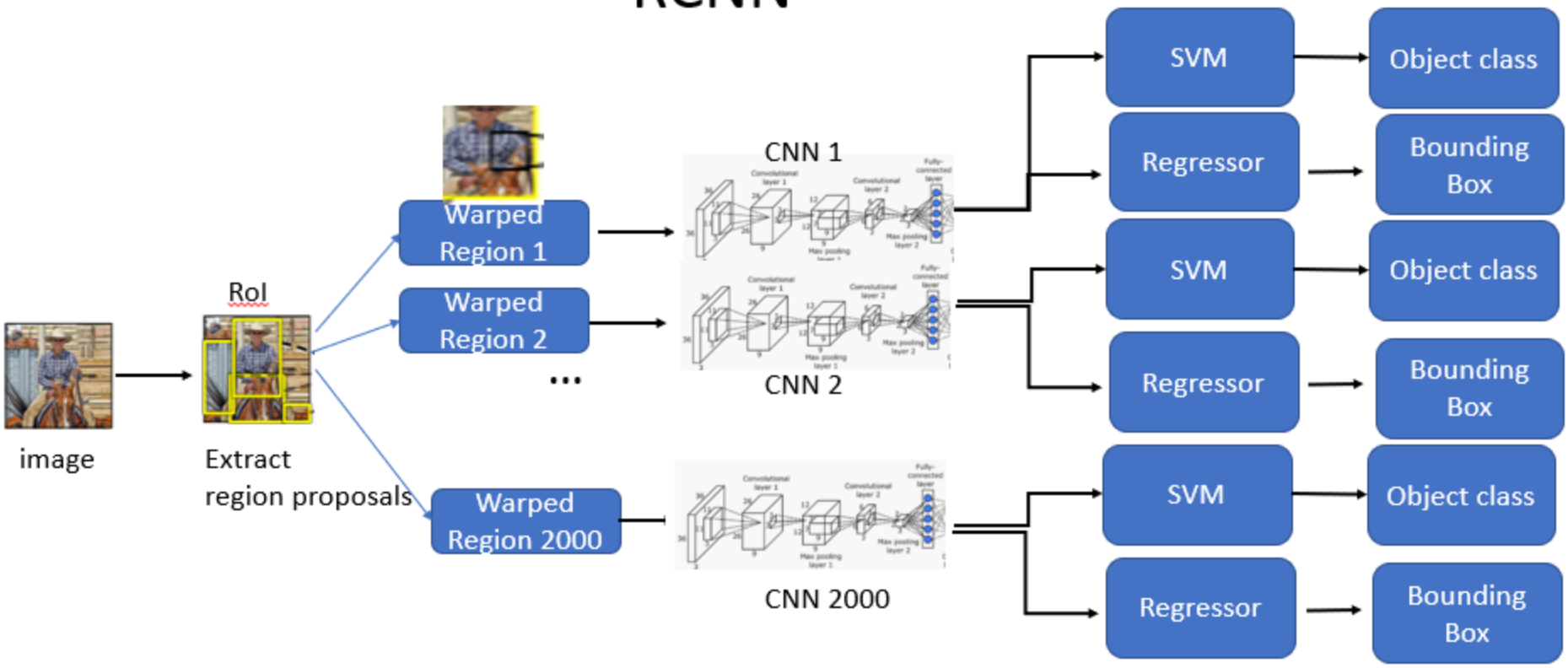
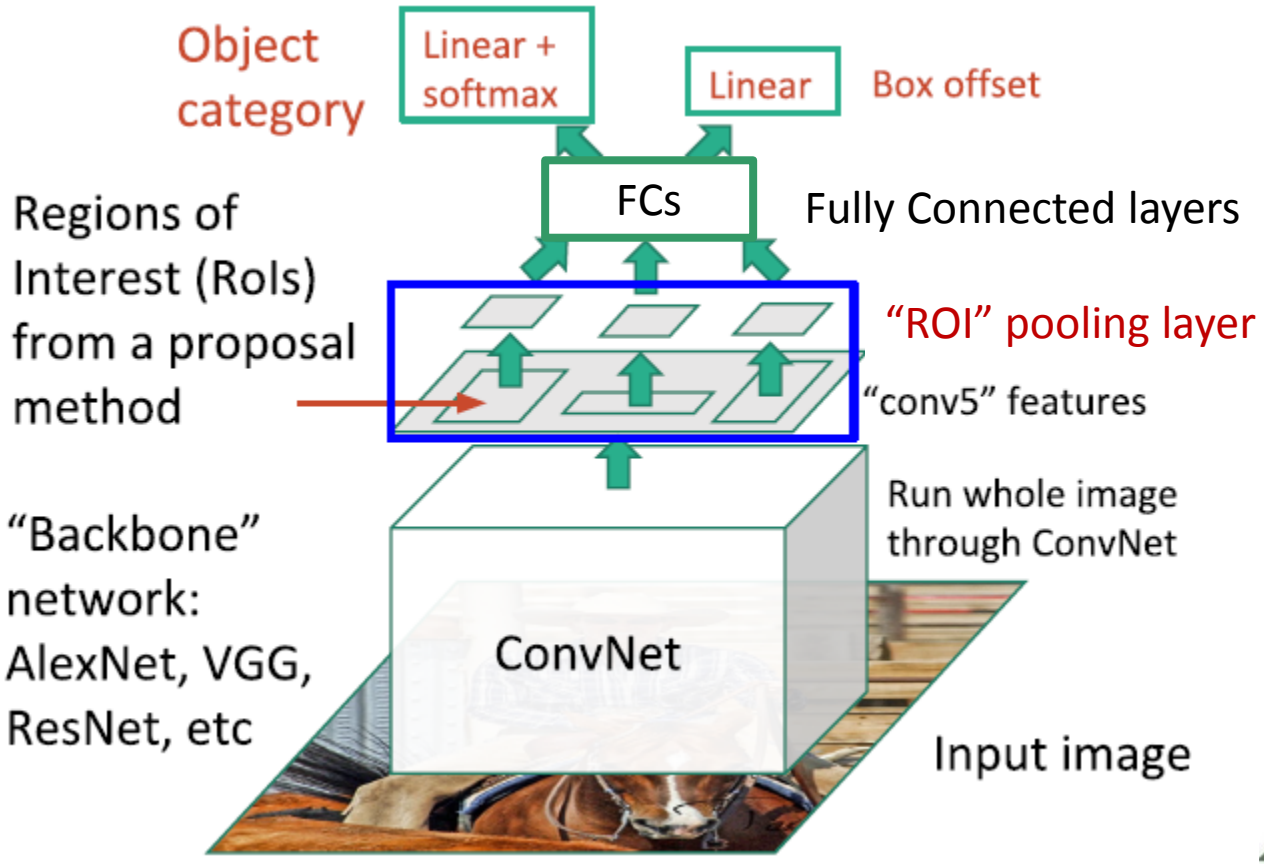
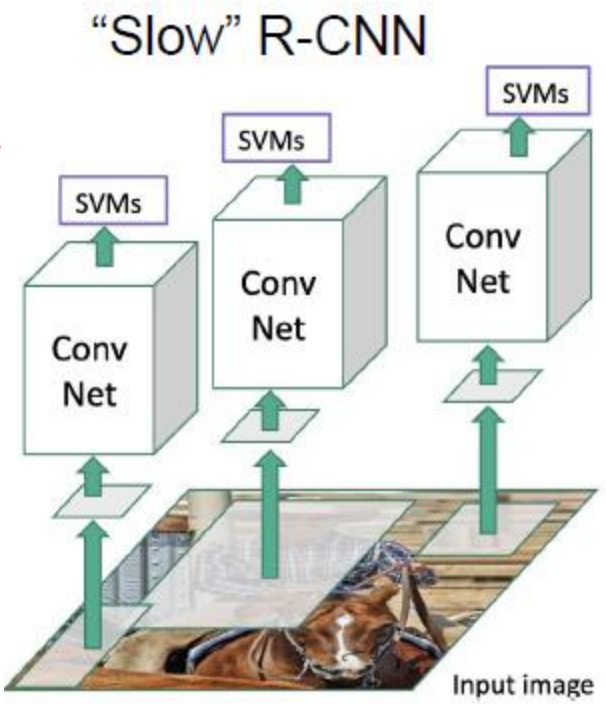


image
RoI
Extract region proposals
Warped Region 1
Warped Region 2
...
Warped Region 2000
CNN 1
CNN 2
CNN 2000
SVM
Regressor
SVM
Regressor
SVM
Regressor
Object class
Bounding Box
Object class
Bounding Box
Object class
Bounding Box

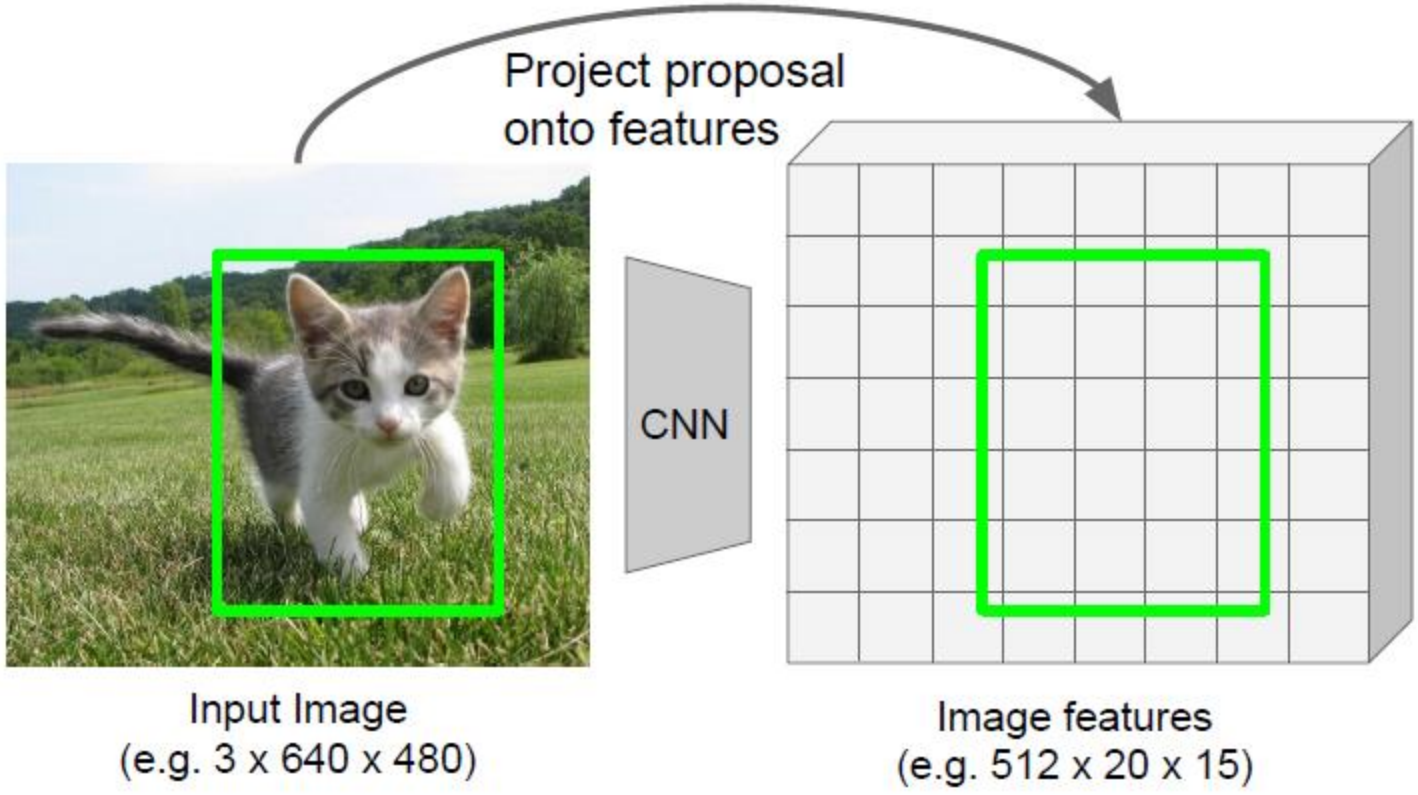
Fast R-CNN



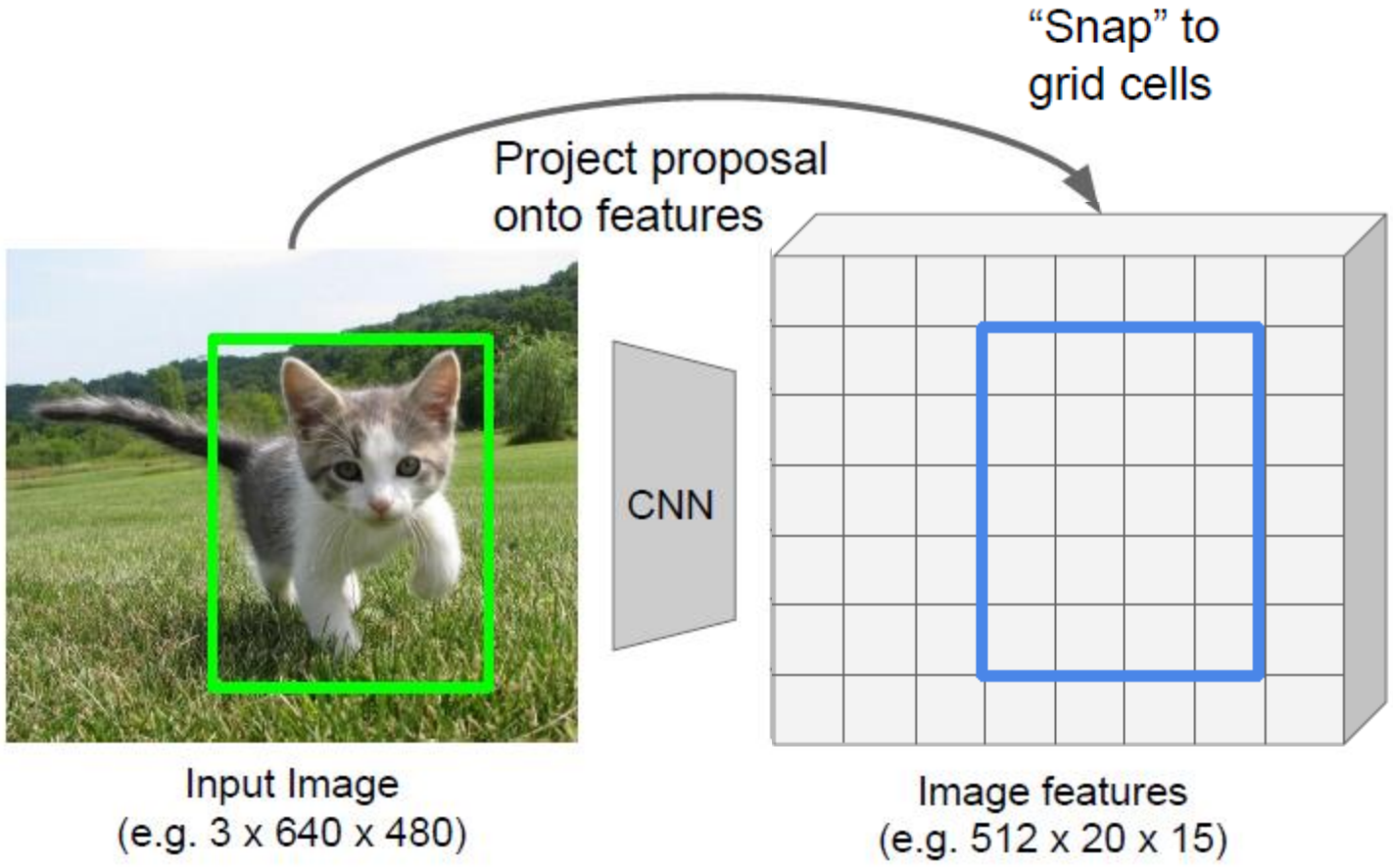


Girshick, "Fast R-CNN", ICCV 2015. Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

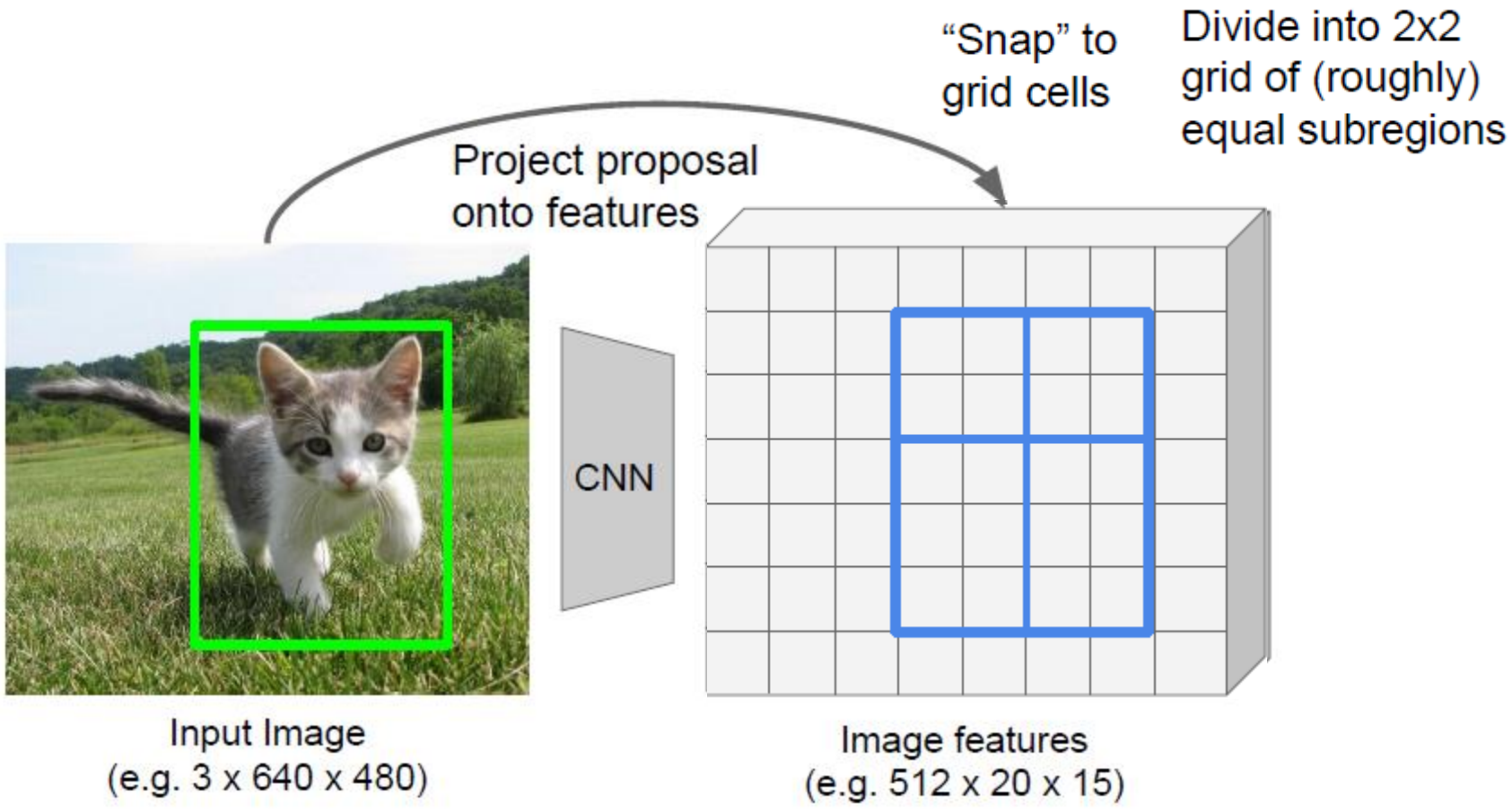
Cropping Features: RoI Pool



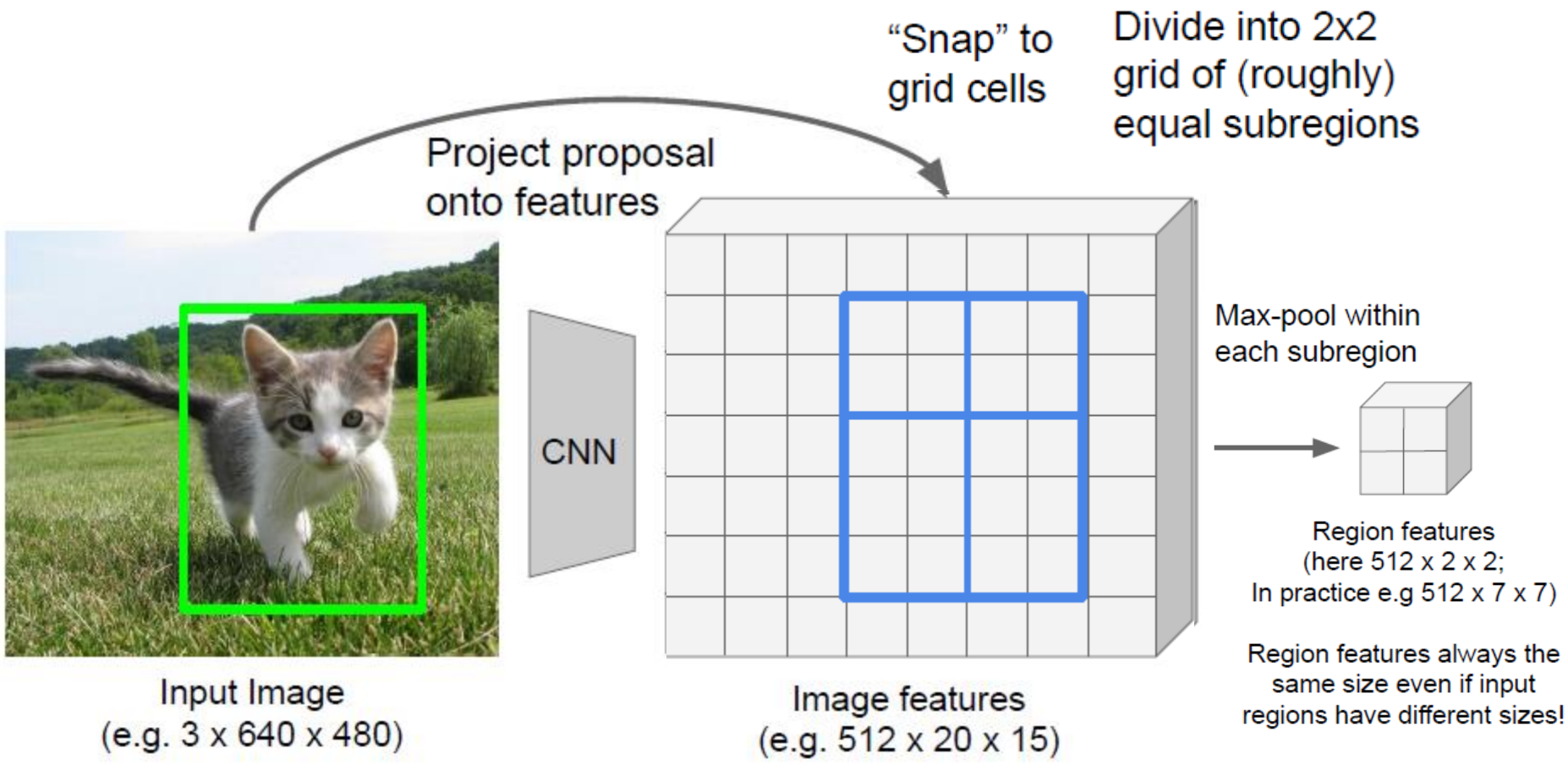
Cropping Features: RoI Pool



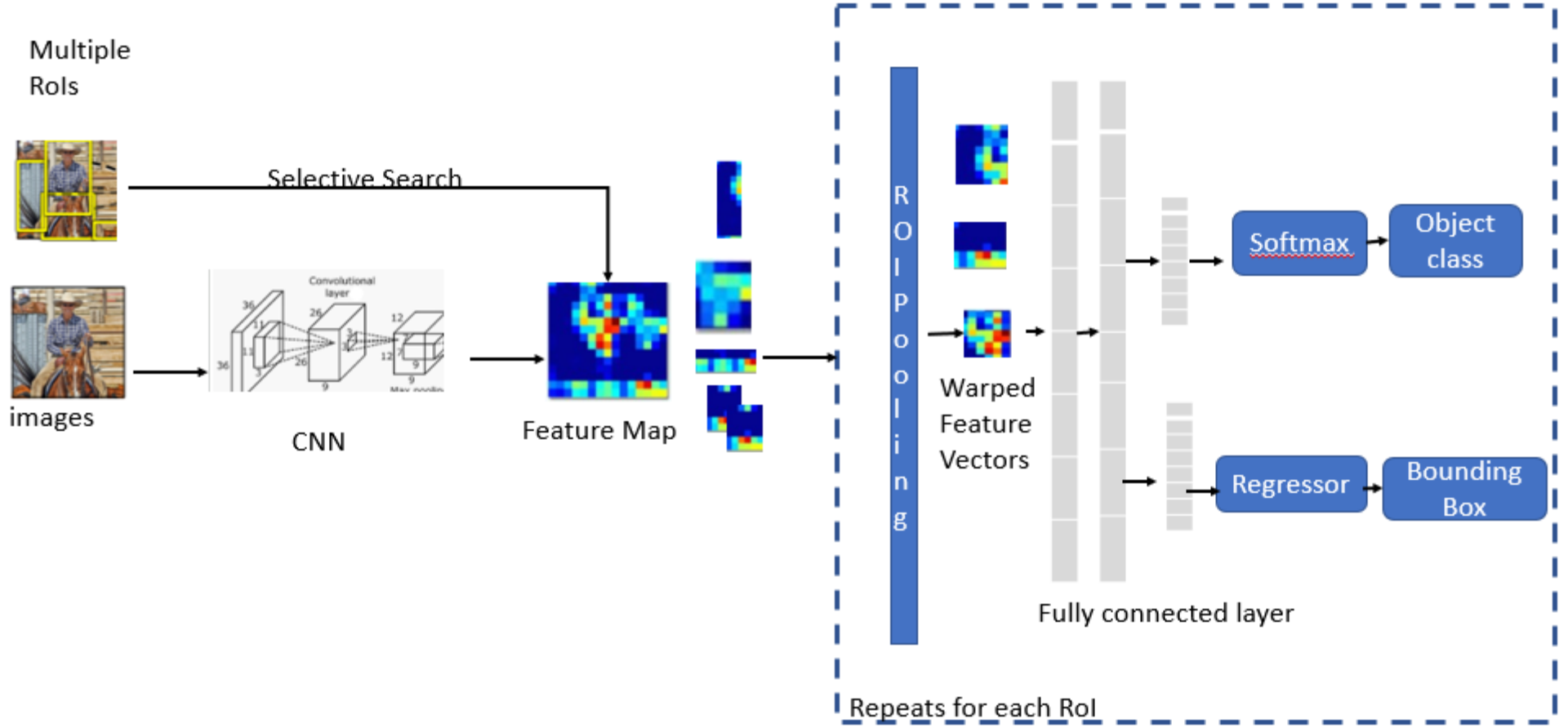
Cropping Features: RoI Pool



Cropping Features: RoI Pool



Fast RCNN


Multiple RoIs
Selective Search
images
Convolutional layer
CNN
Feature Map
ROI Pooling
Warped Feature Vectors
Softmax
Object class
Regressor
Bounding Box
Fully connected layer
Repeats for each RoI

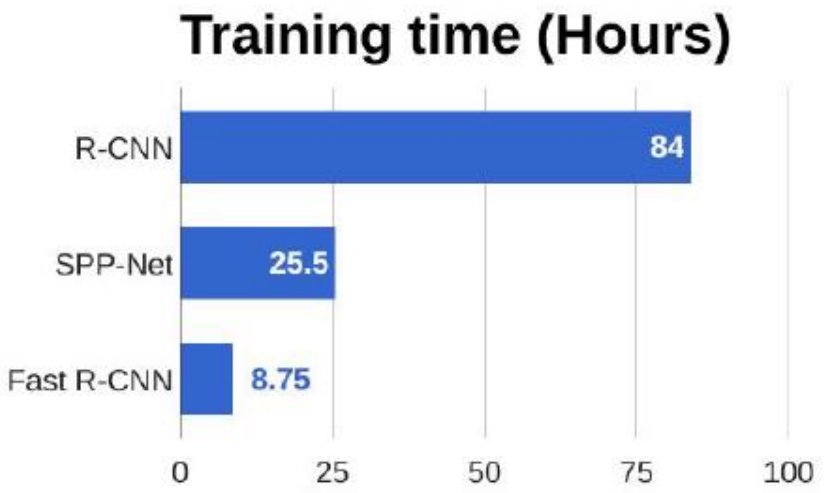
Key differences between R-CNN and Fast R-CNN

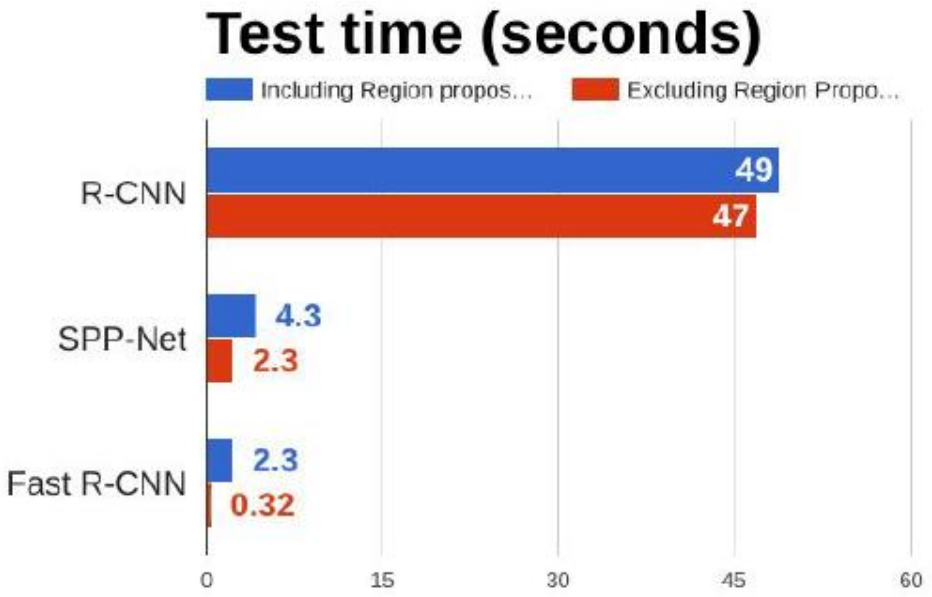
- **Fast R-CNN uses single Deep ConvNet for feature extractions.** A single deep ConvNet speeds us the image processing significantly unlike R-CNN that uses 2000 ConvNets for each region of the image.

- **Fast R-CNN uses softmax for object classification instead of SVM used in R-CNN**. Softmax slightly outperforming SVM for objection classification

- **Fast R-CNN uses multi task loss to achieve an end to end training of Deep ConvNets increases the detection accuracy.**

R-CNN vs Fast R-CNN





Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
He et al, "Spatial pyramid pooling in deep convolutional networks for visual recognition", ECCV 2014
Girshick, "Fast R-CNN", ICCV 2015

Credit for

CS131 "Computer Vision: Foundations and Applications" by University of Stanford (Fall 2019)

CS231n "Convolutional Neural Networks for Visual Recognition" by University of Stanford

(Lecture 11)