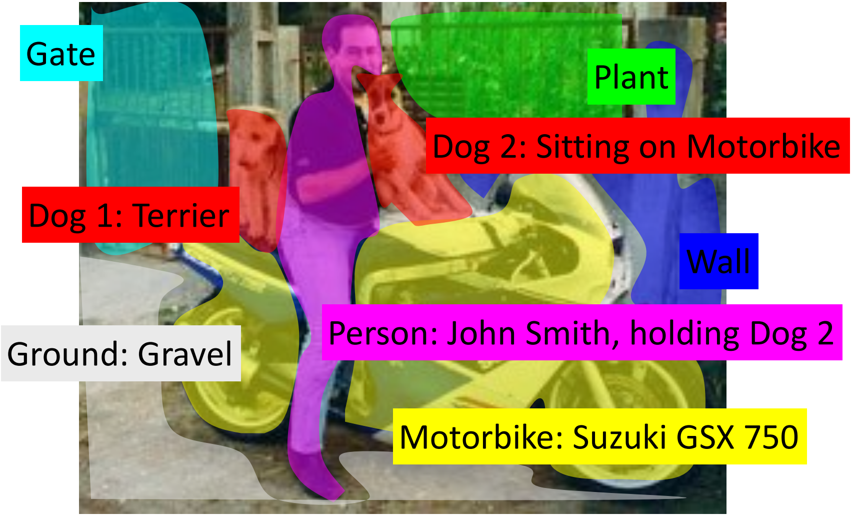
Computer Vision

CS-E4850, 5 study credits

Lecture 9: Object category detection

What we would like to be able to do…

- Visual scene understanding
- What is in the image and where
- Object categories, identities, properties, activities, relations,…



Recognition tasks

- Image classification
  - Does the image contain an aeroplane?

Recognition tasks

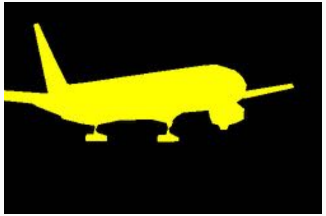
- Image classification
  - Does the image contain an aeroplane?
- Object class detection/localization
  - Where are the aeroplanes (if any)?

Recognition tasks

- Image classification
  - Does the image contain an aeroplane?

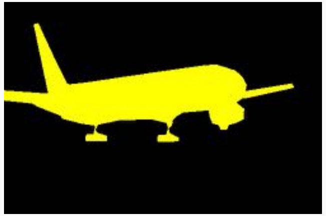
- Object class detection/localization
  - Where are the aeroplanes (if any)?

- Object class segmentation
  - Which pixels are part of an aeroplane?



Recognition tasks

- Image classification
  - Does the image contain an aeroplane?
- Object class detection/localization
  - Where are the aeroplanes (if any)?
- Object class segmentation
  - Which pixels are part of an aeroplane?



Challenges and Applications

Background clutter

Occlusions and truncation

Intra class variation

Preview of typical results

Aeroplane

Bicycle

Car

Cow

Horse

Motorbike

Credits: A Zisserman

Preview of tracking by detection

Detect to track and track to detect, Feichtenhofer, Pinz, Zisserman, ICCV 2017

Application: collision prevention

www.mobileye.com

Application: Funny Nikon ads

"Nikon S60 detects up to 12 faces."

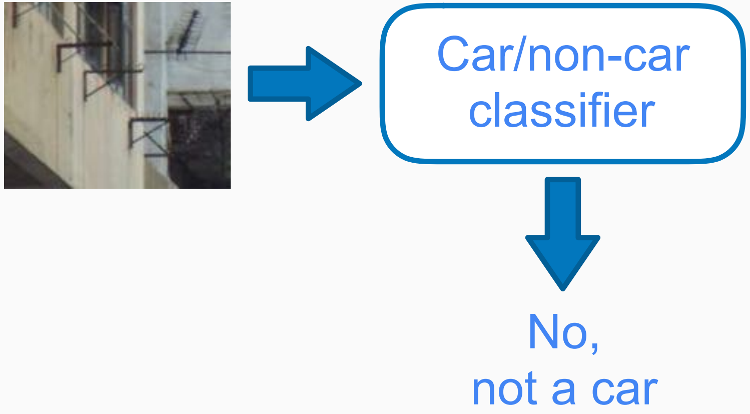
Sliding window detector

Problem of background clutter

- Use sub window:
  - At correct position, no clutter is present
  - Slide window to detect objects
  - Change size of the window to search over scales

Detection by classification

- Basic component: binary classifier



Detection by classification

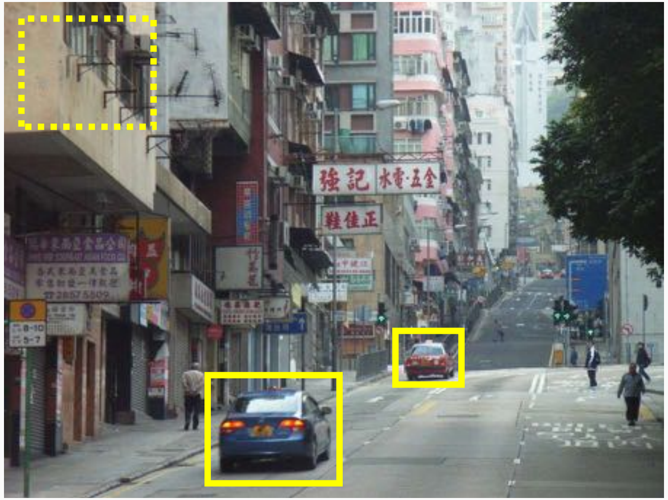
- Detect objects in clutter by **search**

Sliding window: exhaustive search over position and scale

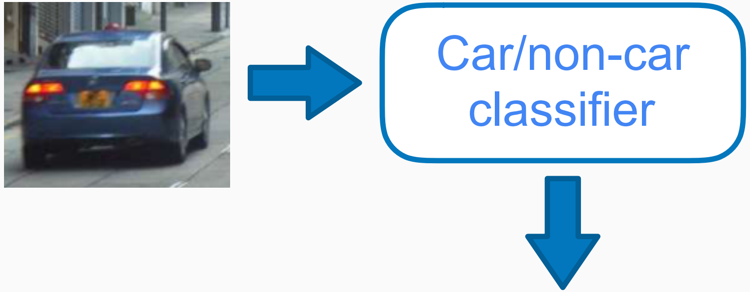
Car/non-car classifier

Detection by classification

- Detect objects in clutter by **search**

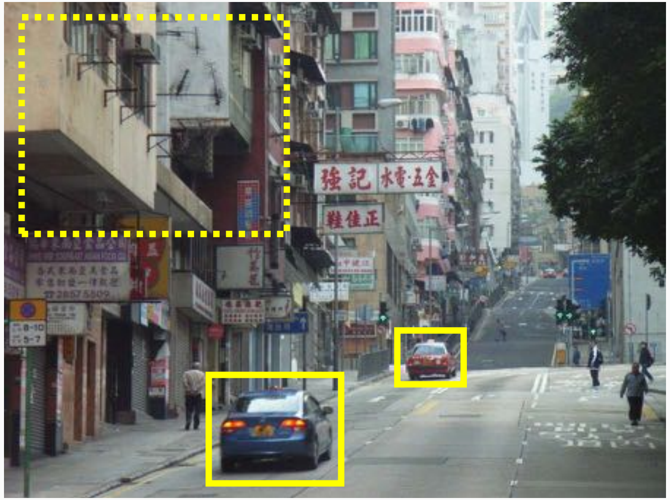


Sliding window: exhaustive search over position and scale



Detection by classification

- Detect objects in clutter by **search**



Sliding window: exhaustive search over position and scale

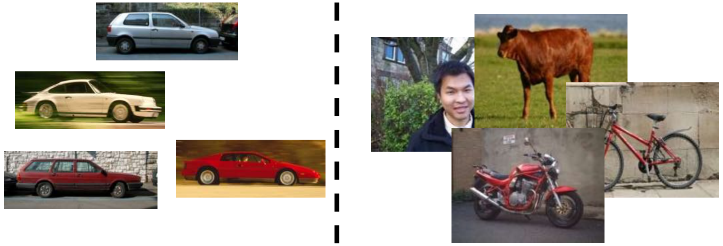
Car/non-car classifier

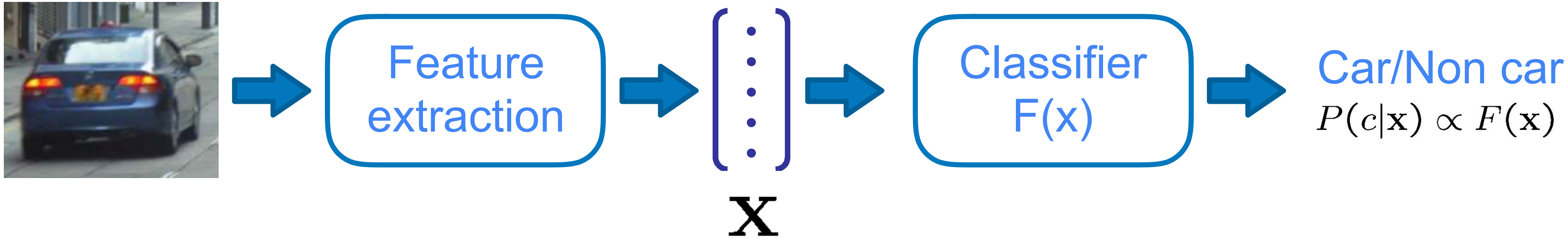
In practice one can use same window size over spatial pyramid

Window (image) classification

- Features usually engineered
- Classifier learned from data

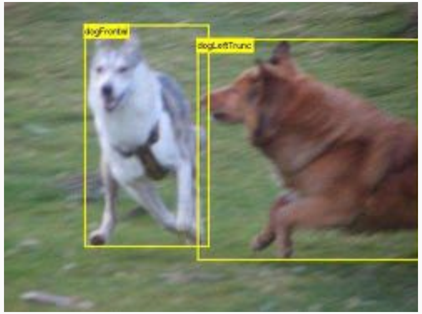
Training data

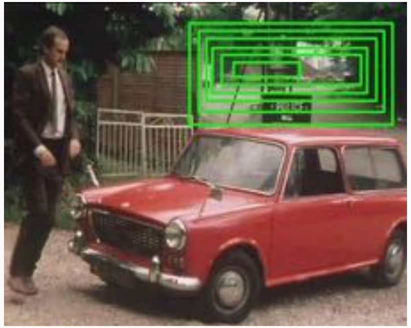




Problems with sliding windows

- Aspect ratio
- Granularity (finite grid)
- Partial occlusions
- Multiple responses
  -> Non-maximum suppression

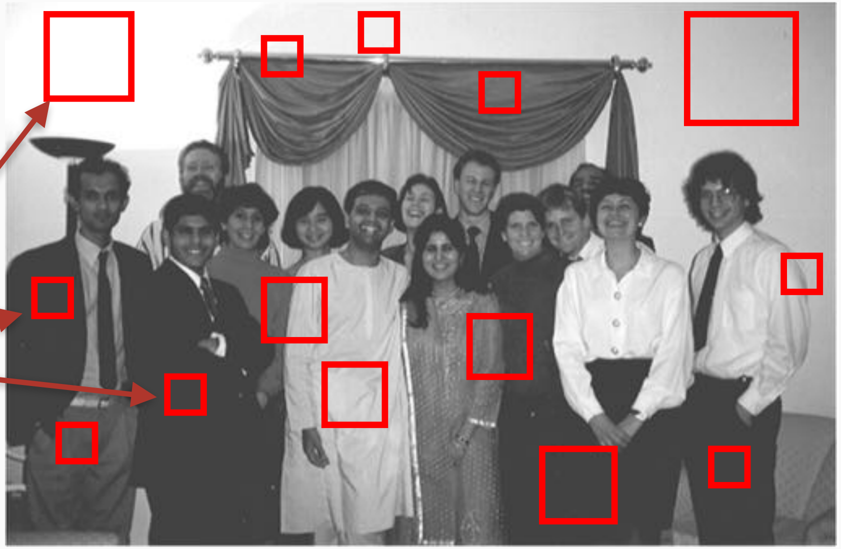




Accelerating sliding window search

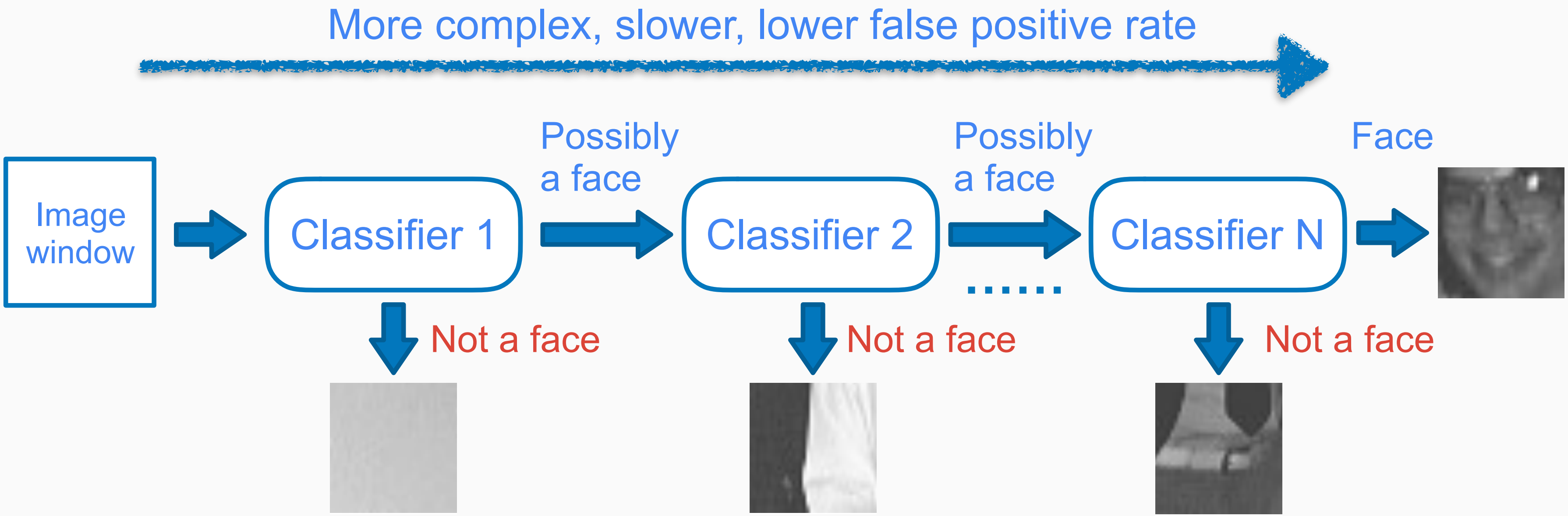
Accelerating sliding window search

- Sliding window search is slow since many windows are needed
- m x n x scale = 100 000 windows for 320 x 240 image
- Most windows are clearly negative
- Is it possible to seed up the search?



Example: face detection

Cascaded classification



Reject easy non-objects using simpler and faster classifiers

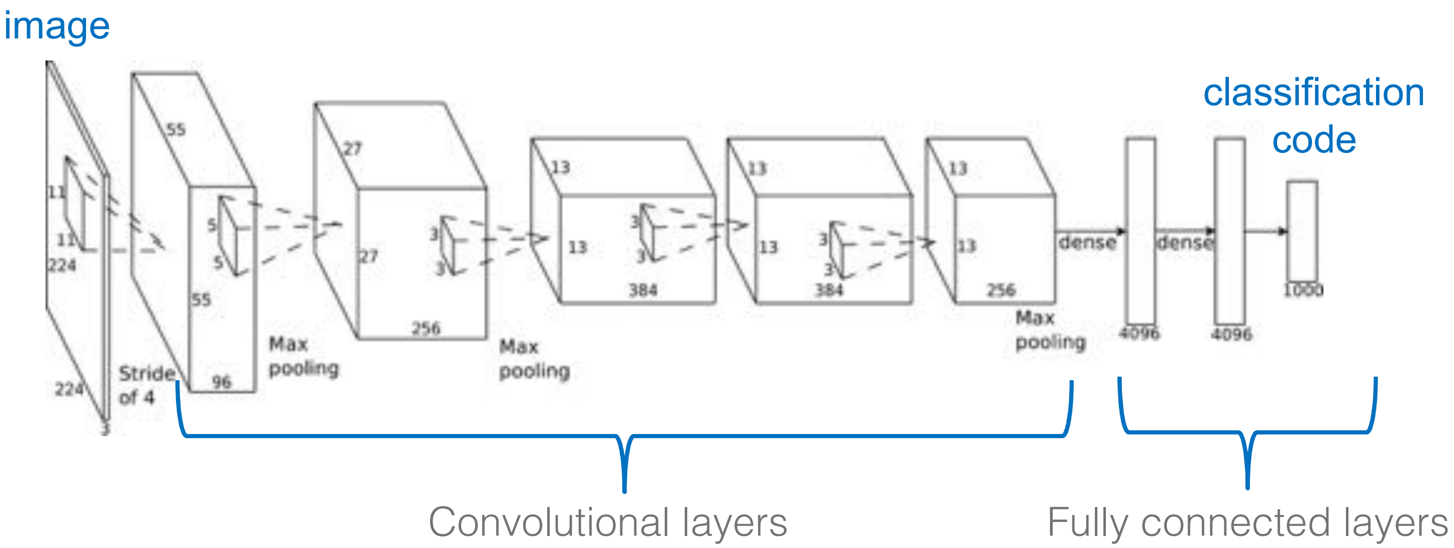
Cascaded classification

- Slow and expensive classifiers only applied to a few windows
  -> significant speedup
- Controlling complexity vs. speed: number of features, number of parts..

Deep networks for object detection

Reminder: Classification CNNs

AlexNet (Krizhevsky et al. 2012)

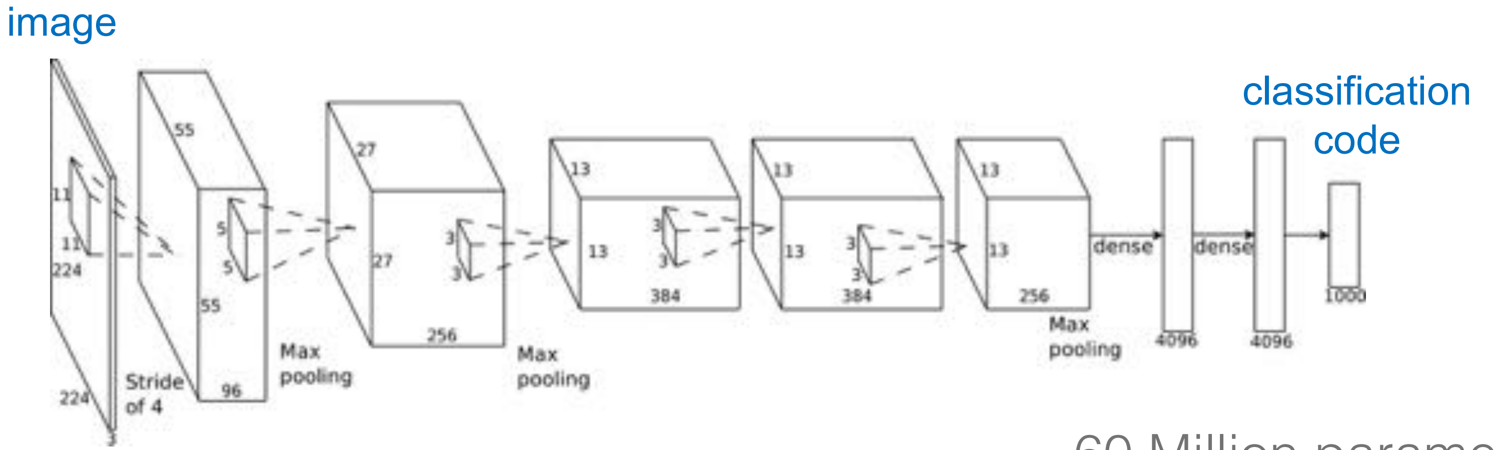


60 Million parameters

ImageNet classification challenge

- 1000 categories
- 1000 images from each category for training (approx. 1M images)
- 100k images for testing

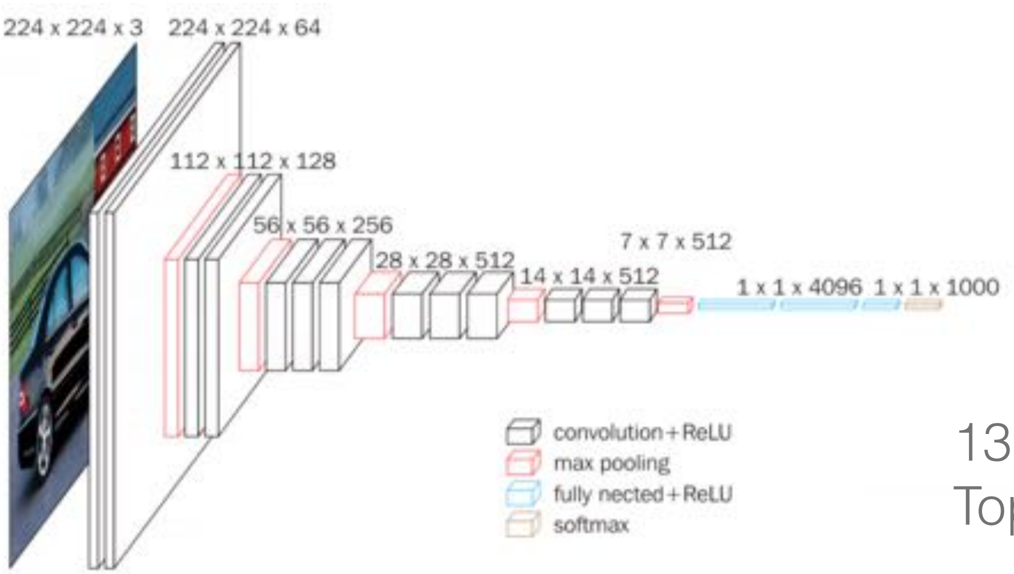
AlexNet (Krizhevsky et al. 2012)



60 Million parameters
Top-5 error 16%

ImageNet classification with deep convolutional neural networks, Krizhevsky et al. NIPS 2012

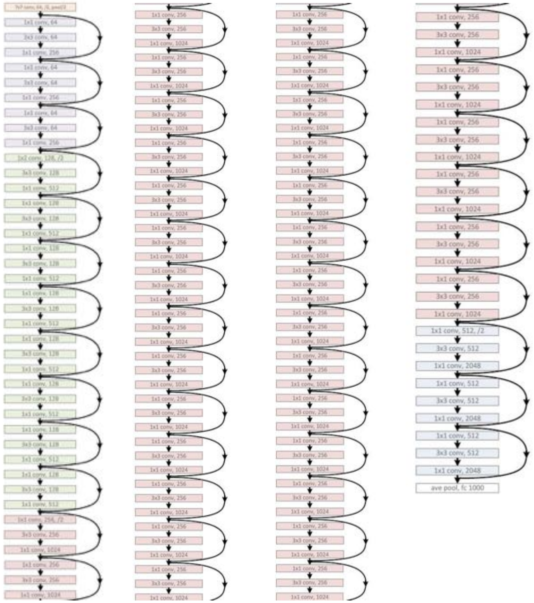
VGG-16 (Simonyan & Zisserman 2014)



138 Million parameters
Top-5 error 7%

Very deep convolutional networks for large-scale image recognition, Simonyan et al. arXiv 2014

ResNet (He et al. 2015)



152 layers (60 Million parameters)
Top-5 error 4%

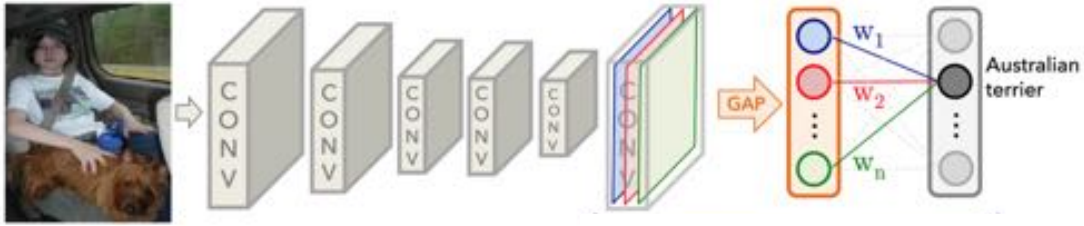
Deep residual learning for image recognition, He et al. CVPR 2016

ImageNet classification results (CLS)

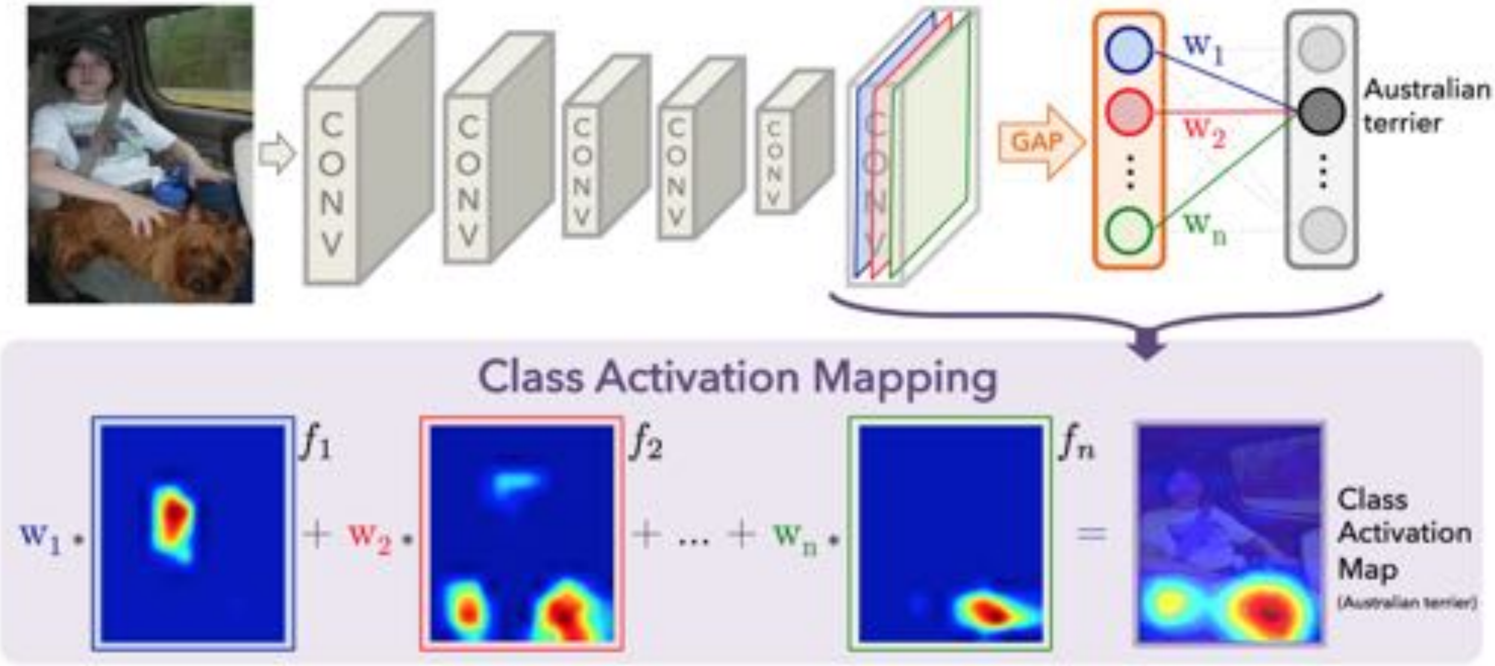
Classification error

CNNs for detection - intuition I

- Modern classification architectures, such as ResNet or Inception, use convolutional layers throughout
  - No fully connected layers
  - Less parameters
  - Feature vector by spatial pooling

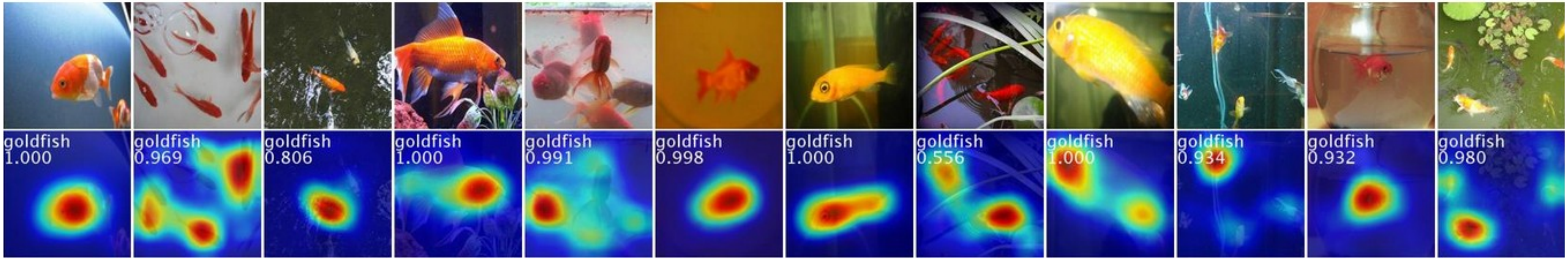


CNNs for detection - intuition II



Is object localisation for free?-weakly-supervised learning with convolutional neural networks, Oquab et al. CVPR 2015
Learning deep features for discriminative localisation, Zhou et al. CVPR 2016

CNNs for detection - intuition II



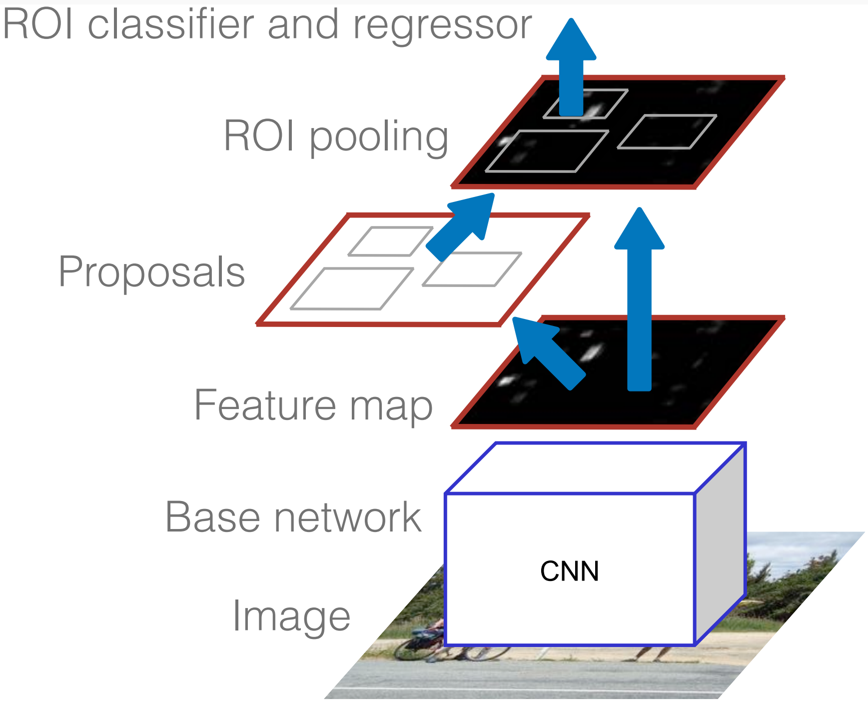
Faster R-CNN

Classical object detectors

- Two stage procedure:
    1. Propose class agnostic regions in the image (sliding window or proposals)
    2. Classify regions into object classes or background
- Can this be captured in a deep network?

Faster R-CNN

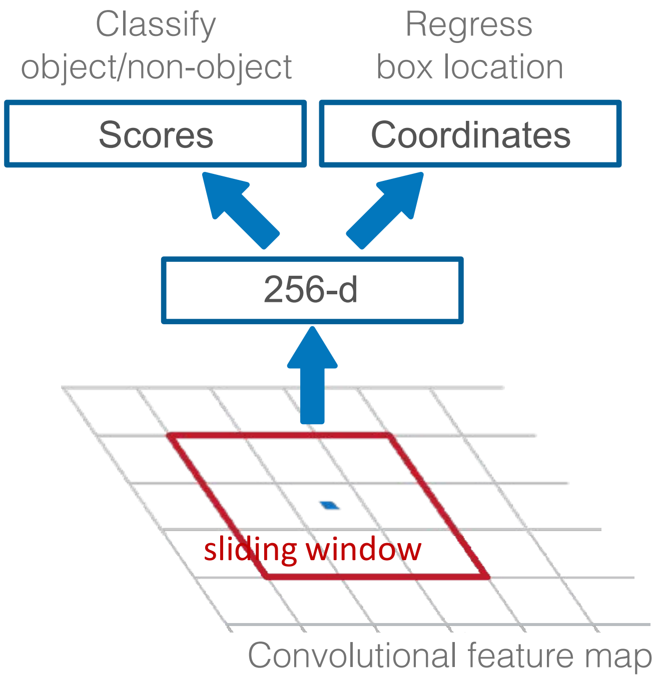
- Two stage system:
  - Region proposal network (RPN)
  - Classification/regression network
- Base network VGG16



Faster R-CNN: Towards real-time object detection with region proposals, Ren et al. NIPS 2015

Region proposal network (RPN)

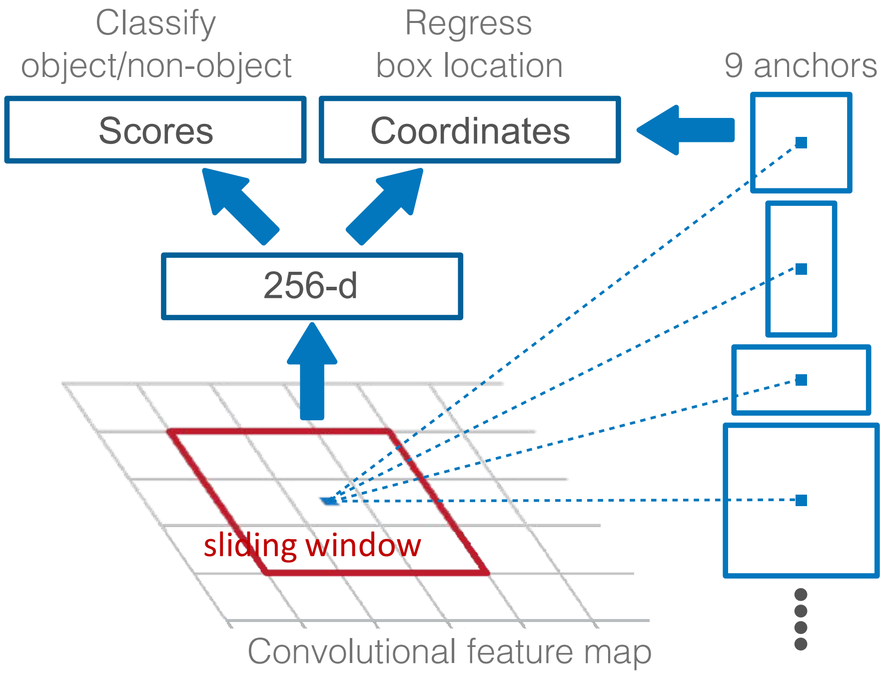
- Slide a small window on feature map
- Window position provides localisation **with reference to the image**
- Box regression provides finer localisation **with reference to window**



Faster R-CNN: Towards real-time object detection with region proposals, Ren et al. NIPS 2015

“Anchors”: predefined candidate regions

- Multi-scale/size anchors are used at each position: 3 scales x 3 aspect ratios yields 9 anchors
- Each anchor has its own prediction function
- **Single-scale** features, multi-scale predictions



Faster R-CNN: Towards real-time object detection with region proposals, Ren et al. NIPS 2015

Training data: positive and negative boxes

- Label training boxes based on overlap with ground truth box
- Pre-train VGG16 CNN on ImageNet classification task

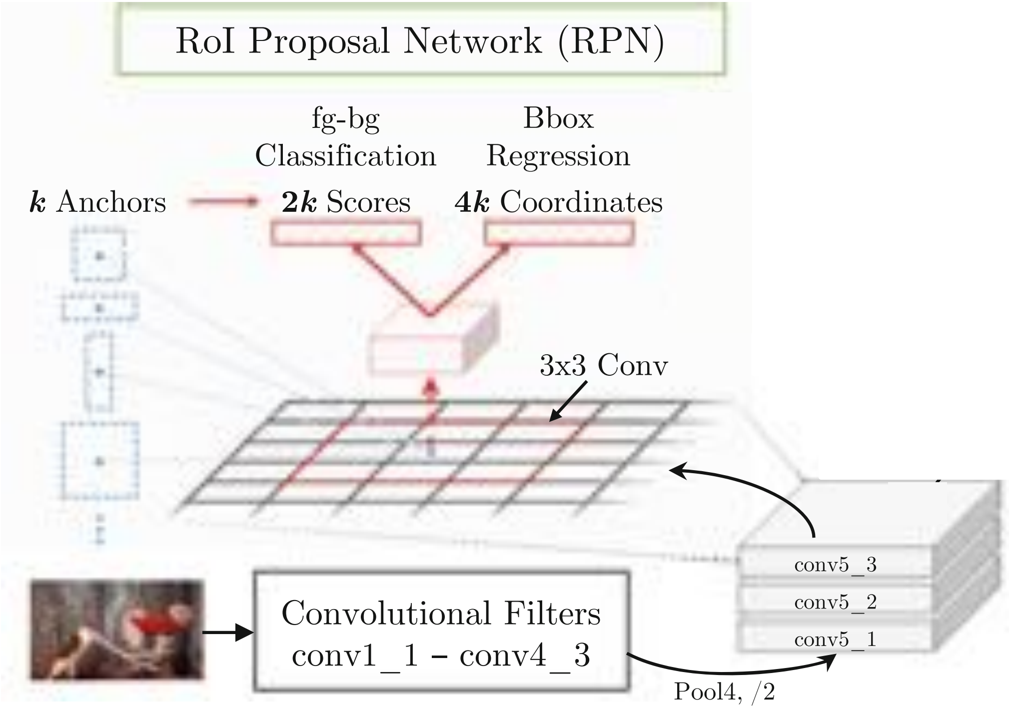
ground truth region

a **positive** training region overlap > 70%

a **negative** training region overlap < 30%

Faster R-CNN: Towards real-time object detection with region proposals, Ren et al. NIPS 2015

Faster R-CNN



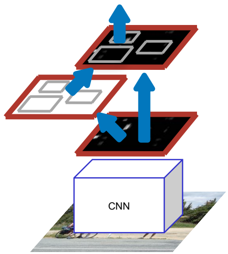
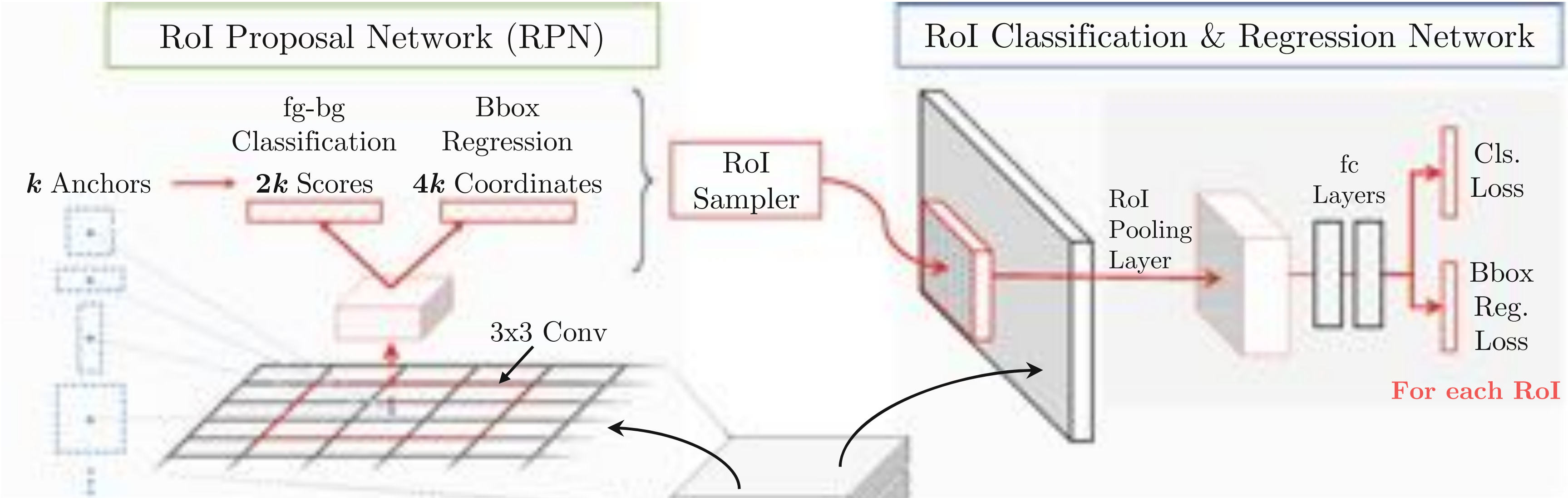
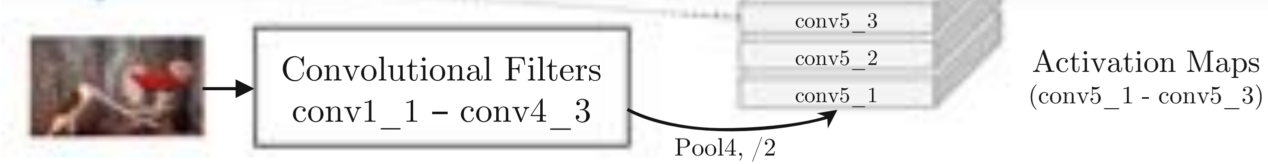


Figure from: Contextual priming and feedback for Faster R-CNN, Srivastava et al., ECCV 2016

Faster R-CNN





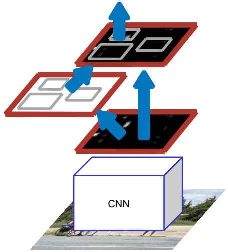
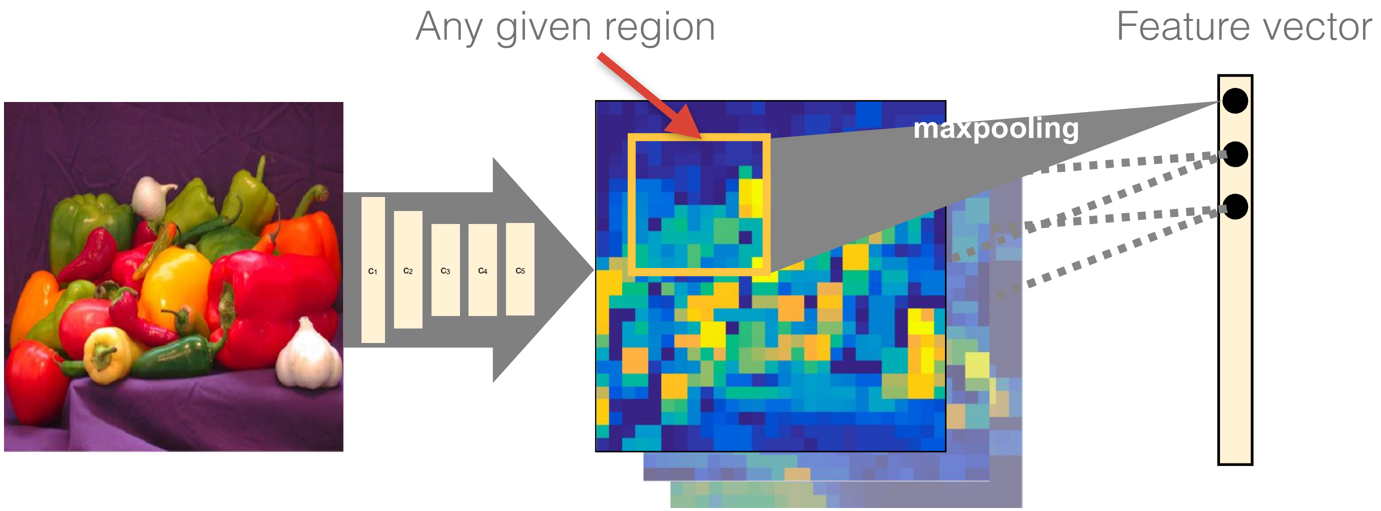


Figure from: Contextual priming and feedback for Faster R-CNN, Srivastava et al., ECCV 2016

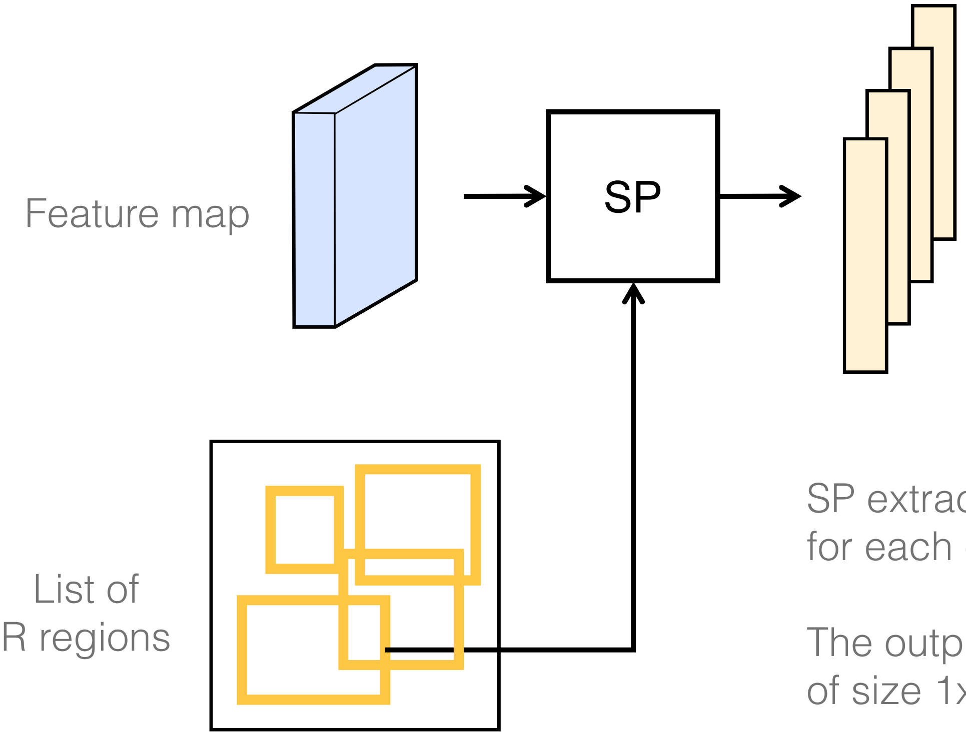
The Spatial Pooling (SP) layer

- Performs max-pooling for the feature responses in a given region
- Can be used to extract many region-specific feature vectors using same convolutional feature output



Spatial pyramid pooling (SPP) in deep convolutional networks for visual recognition, He et al. ECCV 2014

The Spatial Pooling (SP) layer as a building block



R region-specific feature vectors

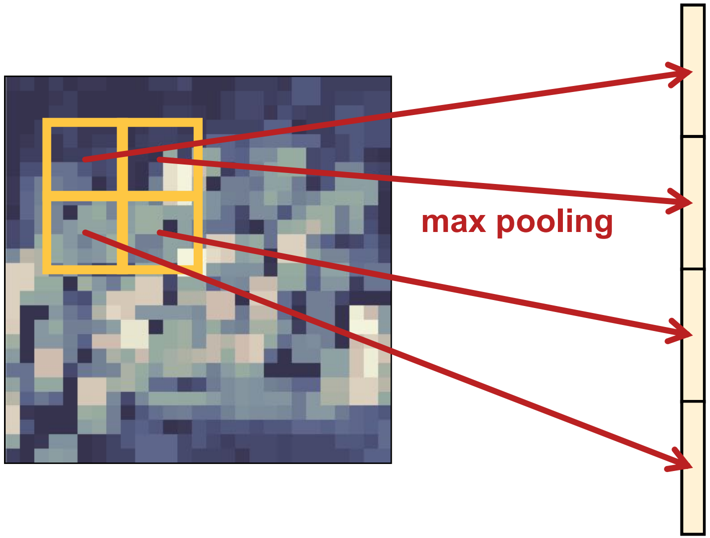
SP extracts a feature vector for each of the R regions

The outputs are R tensors of size 1x1xC

Spatial pyramid pooling (SPP) in deep convolutional networks for visual recognition, He et al. ECCV 2014

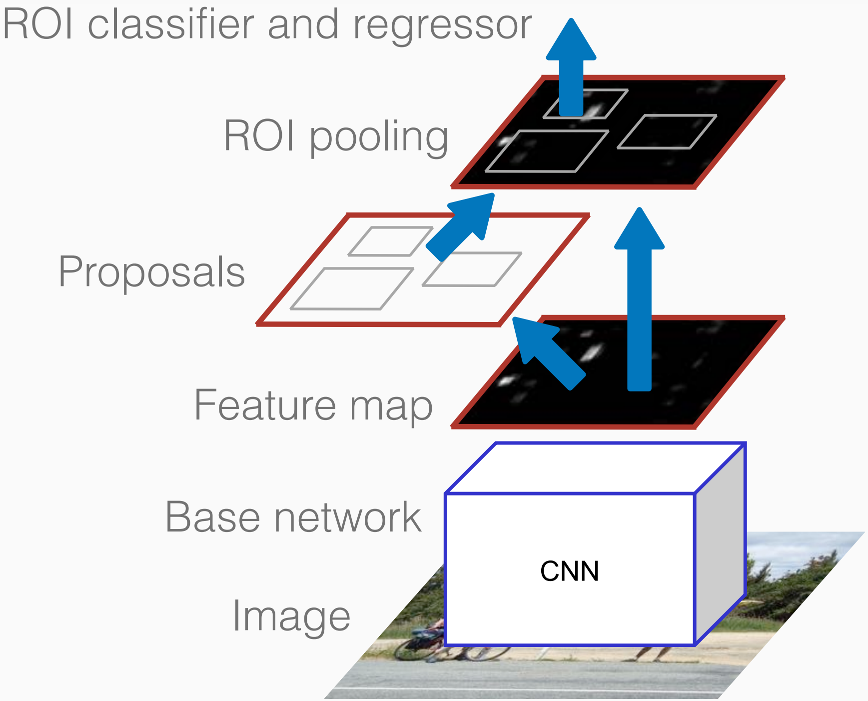
The Spatial Pyramid Pooling (SPP) layer

- Similar to SP, but pools features in tiles of a grid-like subdivision of the region (SP with multiple subdivisions)
- Feature vector **captures the spatial layout** of the original region
- Converts the region to a **fixed size vector**



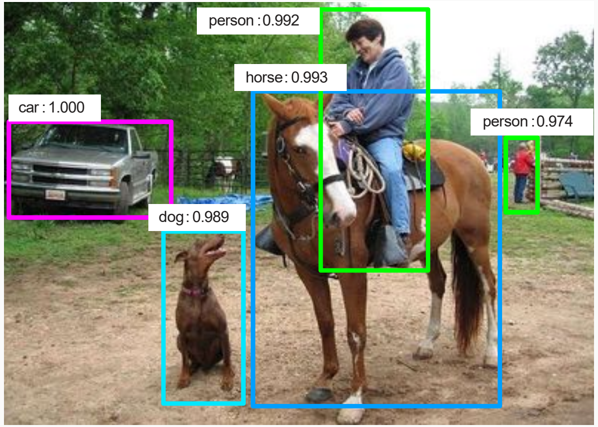
Faster R-CNN

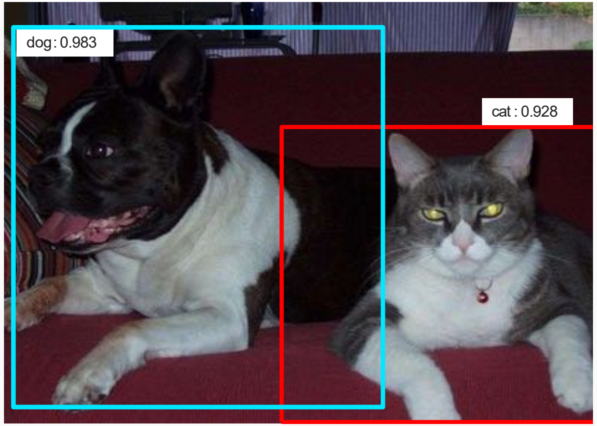
- Same CNN conv5 features used for:
  - The region proposal network
  - Classifying/regressing the regions
- Thus CNN runs only once on image
- Trained end-to-end
- Base network VGG16

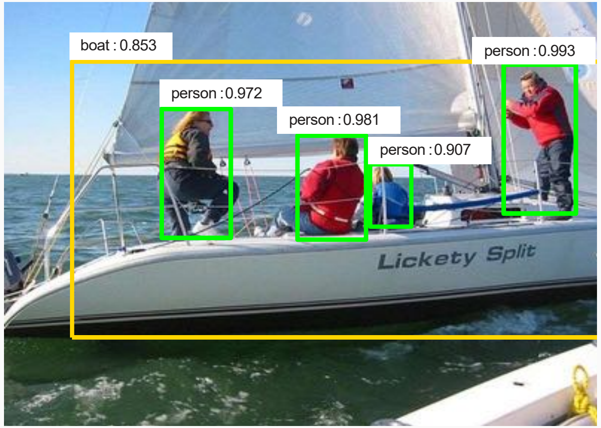


Faster R-CNN: Towards real-time object detection
with region proposals, Ren et al. NIPS 2015

Example detections

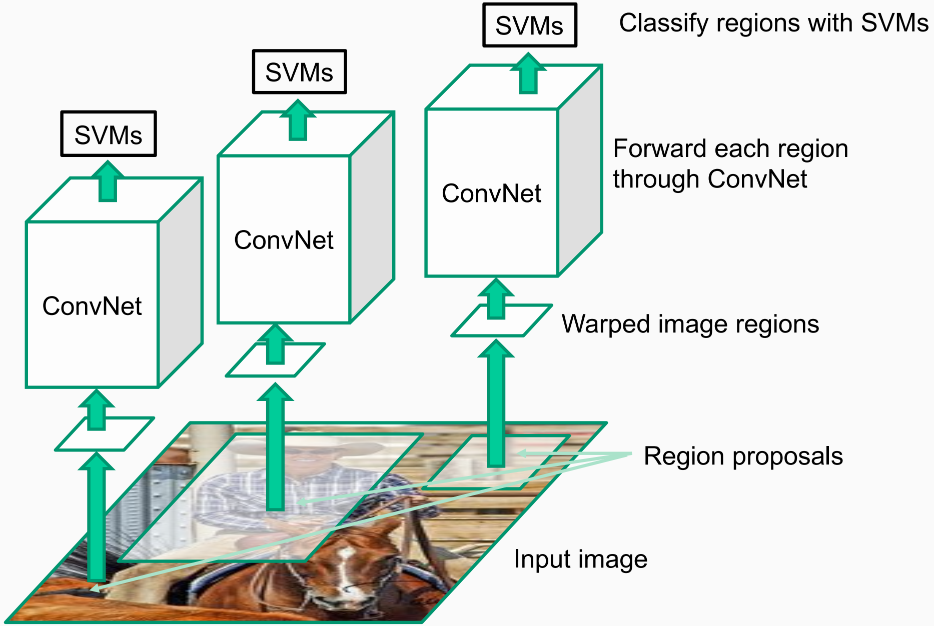






Why “Faster R-CNN”?

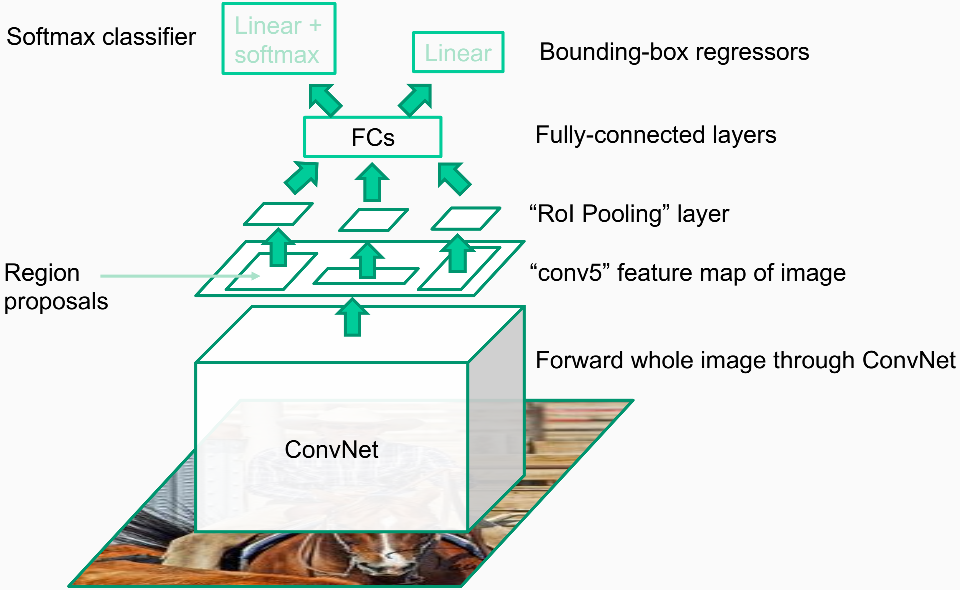
- First: R-CNN
- Inference time approx. 50s per image



Rich feature hierarchies for accurate object detection and semantic segmentation, Girshick et al., CVPR 2014

Why “Faster R-CNN”?

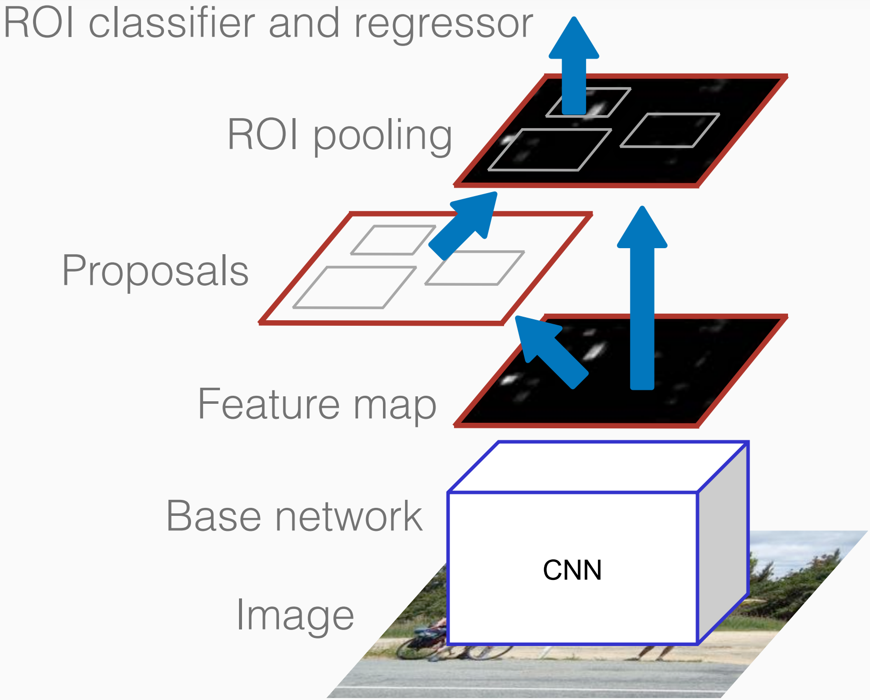
- Second: Fast R-CNN
- Inference time approx. 2s per image



Fast R-CNN, Girshick., ICCV 2015

Why "Faster R-CNN"?

- Third: Faster R-CNN
- Inference time approx. 198ms per image



Faster R-CNN: Towards real-time object detection with region proposals, Ren et al. NIPS 2015

Evaluating object detectors

Evaluating object detectors

- Classical benchmark:

  The PASCAL Visual Object Classes (VOC) dataset and Challenge 2007-2012

  Mark Everingham, Luc Van Gool, Chris Williams, John Winn, Andrew Zisserman

PASCAL VOC dataset content

- Objects from 20 classes:
  aeroplane, bicycle, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, train, TV
- Real world images downloaded from Flickr (not filtered for “quality”)
- Complex scenes, multiple scales, lighting, occlusions,….

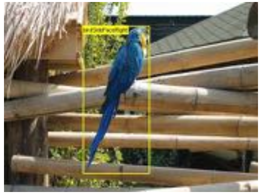



Examples

Aeroplane

Bicycle

Bird



Boat

Bottle

Bus

Car

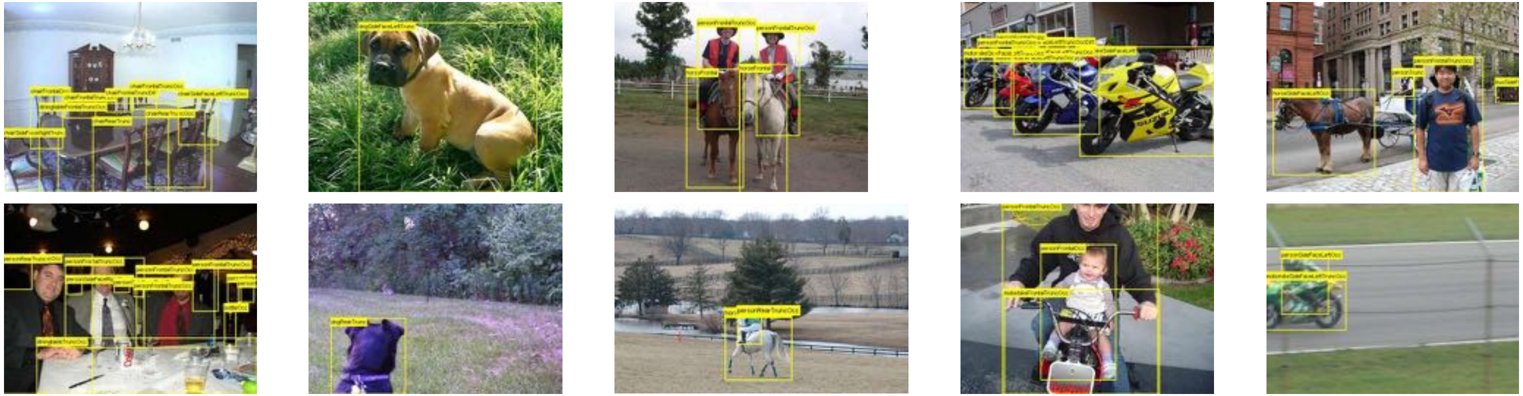
Cat

Chair

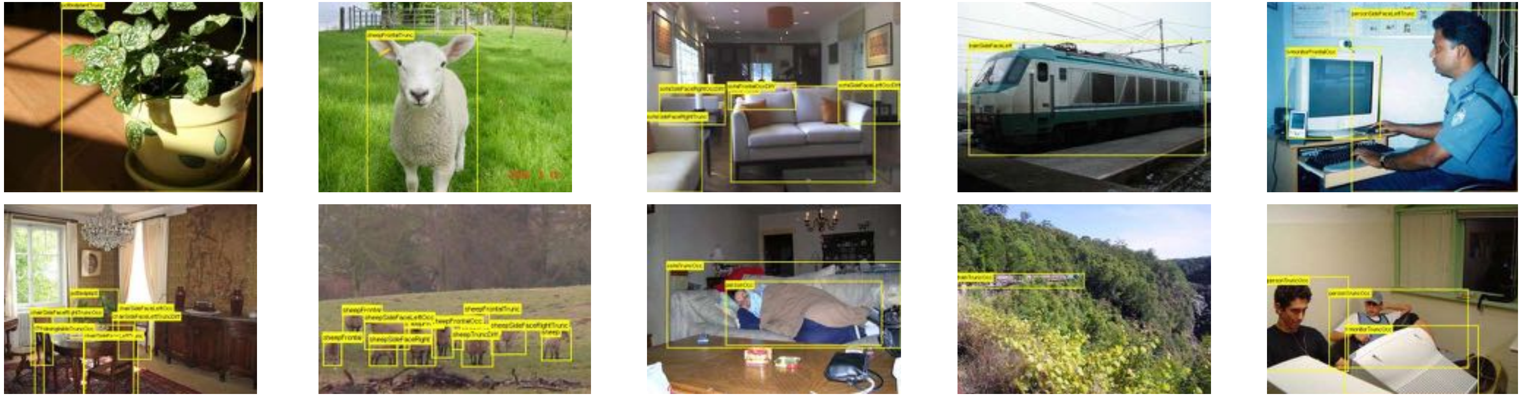
Cow

Examples

Dining Table | Dog | Horse | Motorbike | Person



Potted Plant | Sheep | Sofa | Train | TV/Monitor

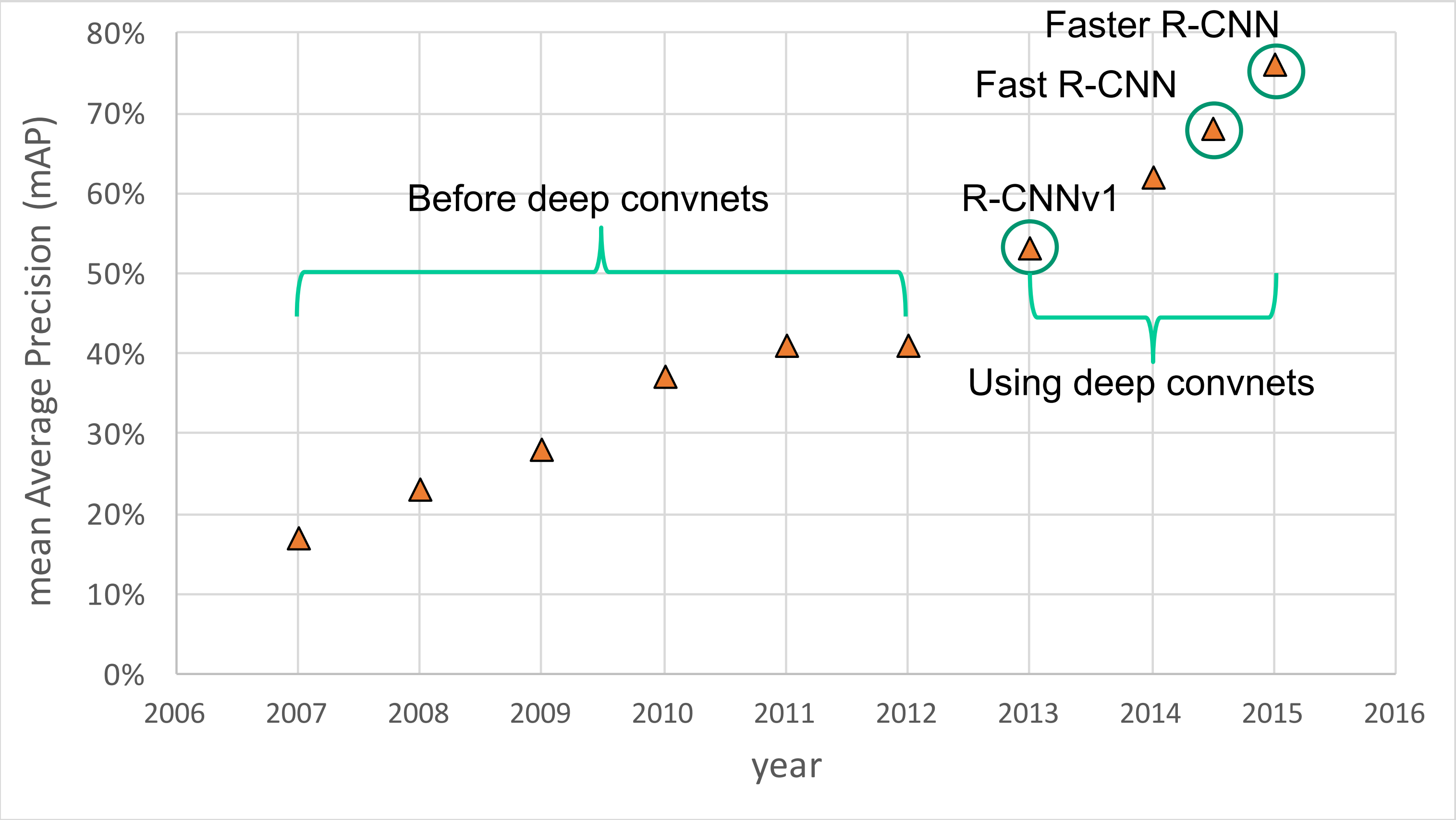


PASCAL VOC statistics

- Minimum 600 training objects per category
- Approx. 2000 cars, 1500 dogs, 8500 people
- Approximately similar distribution across training and test sets

	Training	Testing
Images	11,540	10,994
Objects	27,450	27,078

Progress in object detection (PASCAL VOC)



Application: Faster R-CNN face detector

- VGG16 pre-trained on ImageNet
- Detector trained on the WIDER dataset (12k images 160k faces)

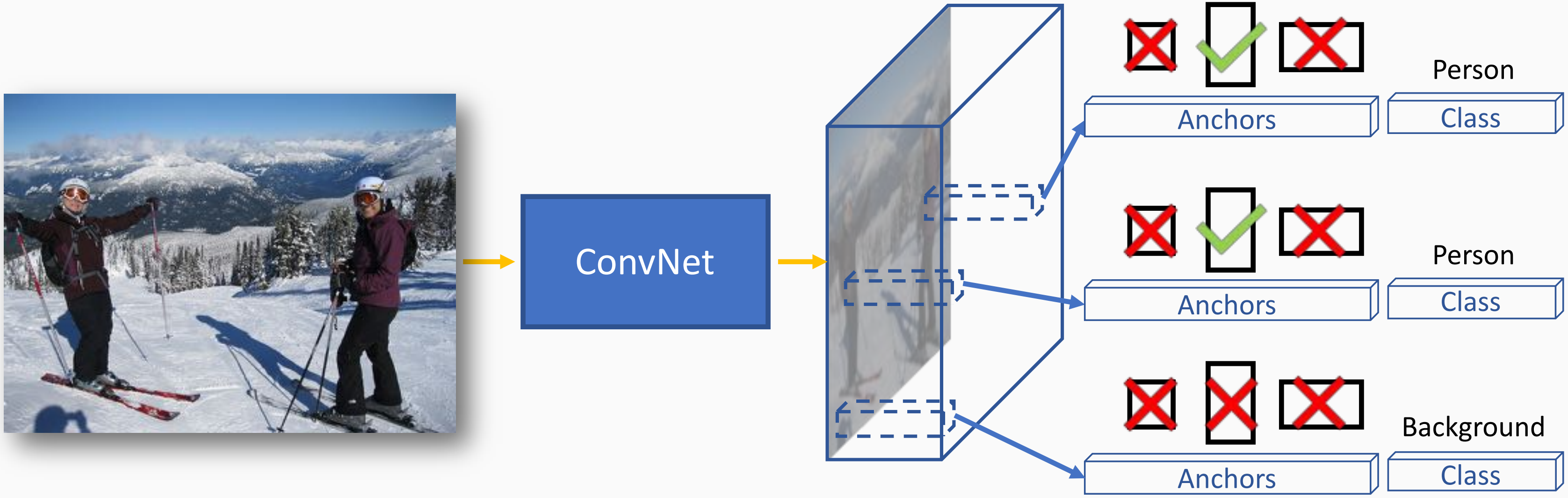
Credits: Sam Albanie, Qiong Cao

Single stage detectors

Two strands of detection architectures

- Detectors using region proposal networks (RPN)
  - Two stages: 1) RPN, followed by 2) features from regions for classification and regression of box
  - Possibly slow due to two steps
  - Examples: Faster RCNN, R-FCN
- Detector using unified framework (no explicit RPN)
  - Regions are build into the architecture (convolutional layers) -> possibly fast
  - Examples: YOLO, SSD, TinyFaces

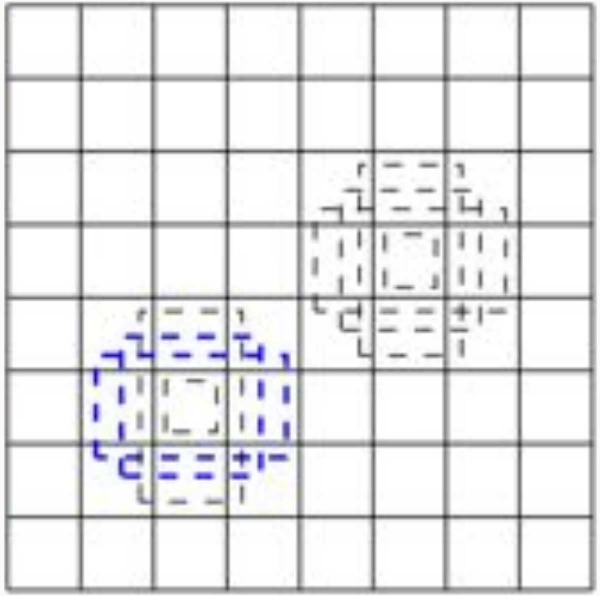
One-stage detectors



Redmond et al. CVPR 2017, Shen et al. ICCV 2017, Liu et al. ECCV 2016,
Fu et al. arXiv 2017, Lin et al. ICCV 2017, Zhang et al. CVPR 2018

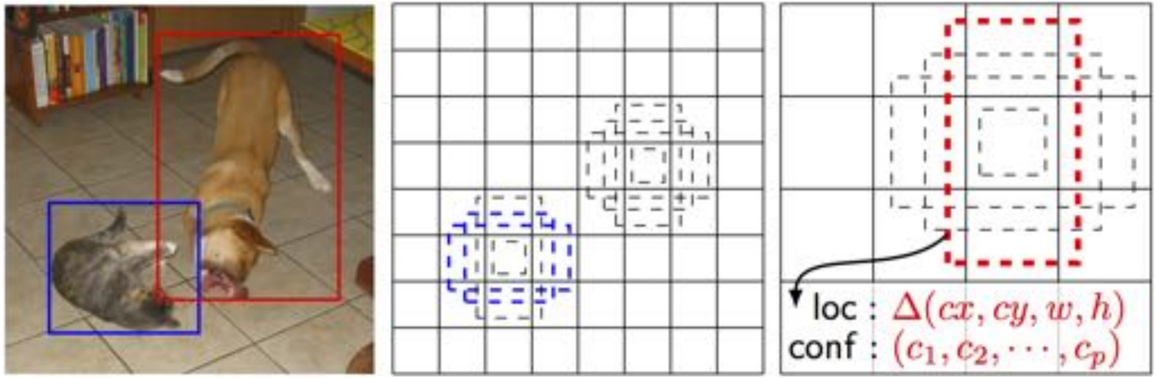
Single Shot MultiBox Detector (SSD)

- Fully convolutional detector (no RPN)
- Pre-defines regions:
  - Predict categories and box offsets
  - Multiple aspect ratios per cell
  - Similar to Faster R-CNN anchor boxes

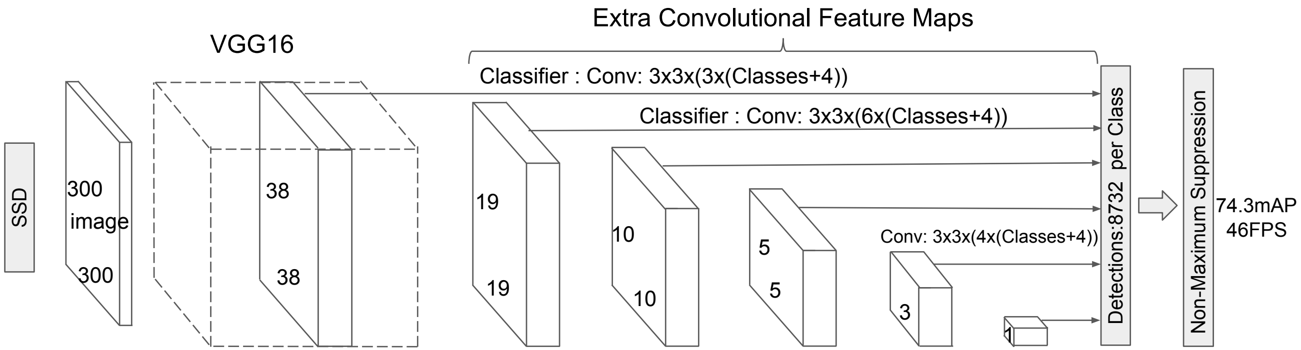


SSD: Single Shot MultiBox Detector, Liu et al., ECCV 2016

Single Shot MultiBox Detector (SSD)



SSD runs at 59 fps
cf. 7 fps for Faster R-CNN

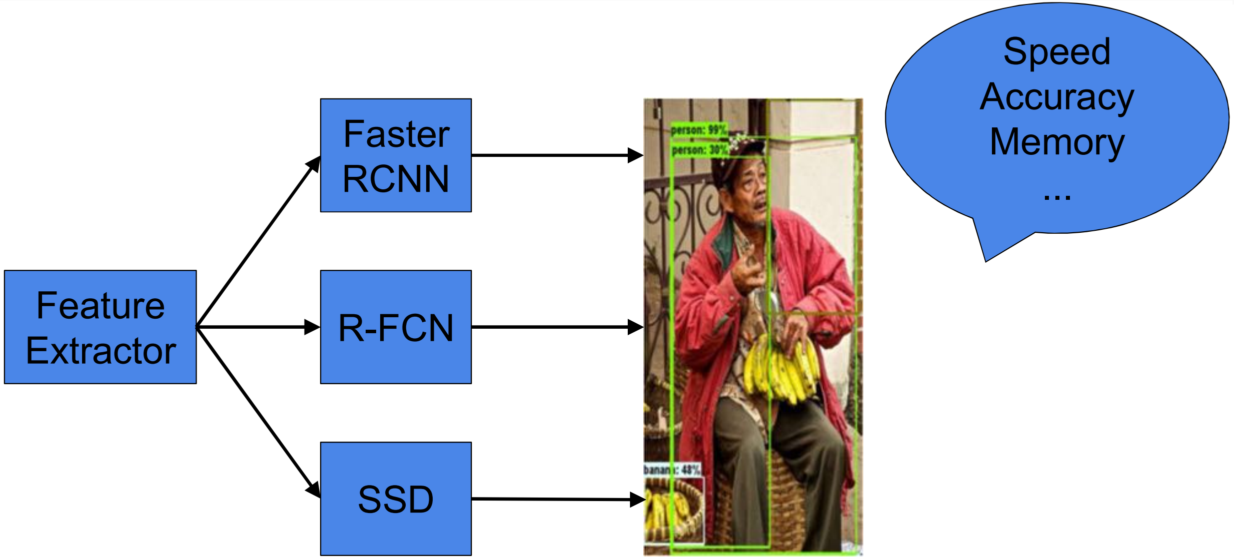


SSD: Single Shot MultiBox Detector, Liu et al., ECCV 2016

Single Shot MultiBox Detector - video example

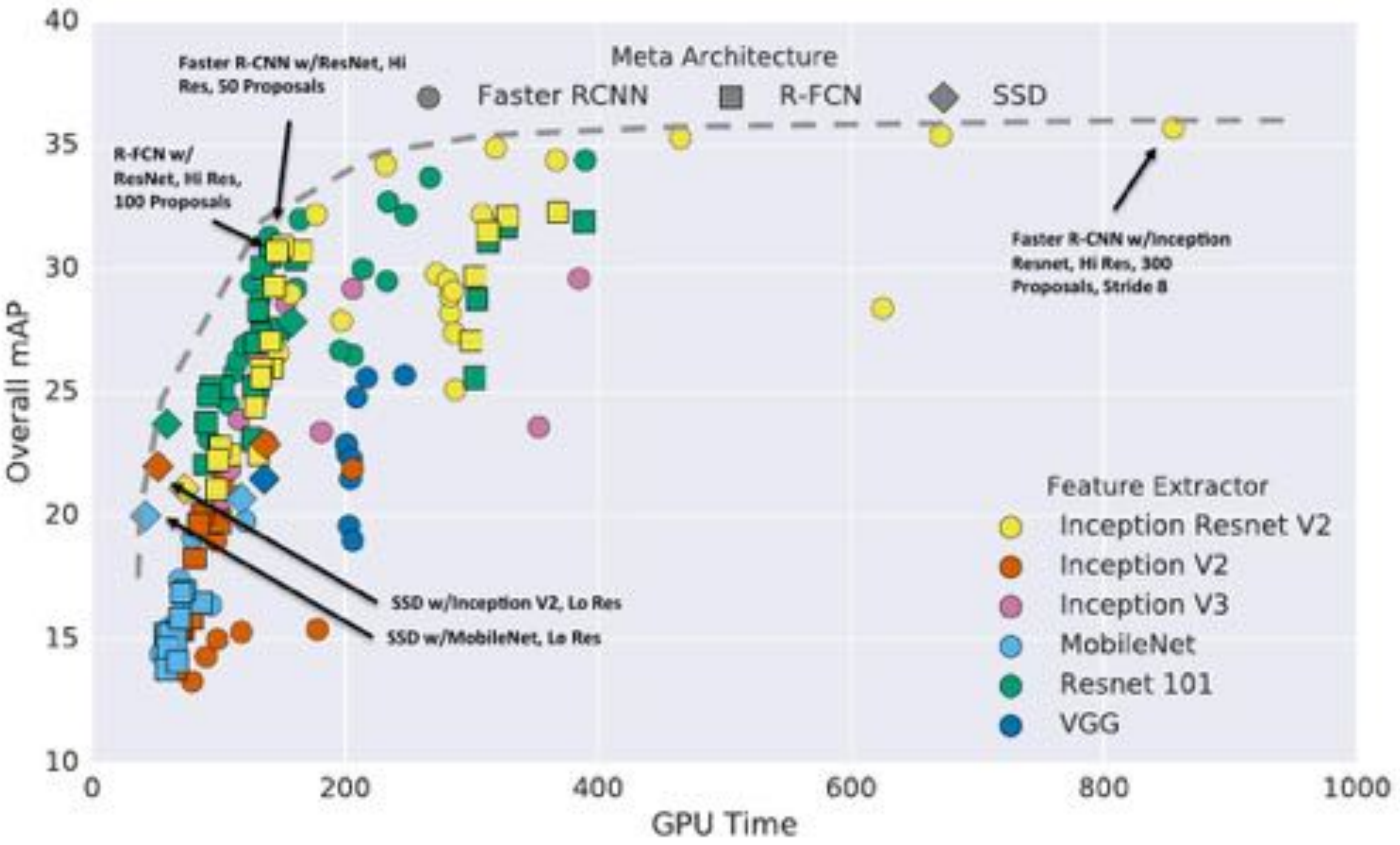
SSD: Single Shot MultiBox Detector, Liu et al., ECCV 2016

Summary and comparison



Speed/accuracy trade-offs for modern convolutional object detectors, Huang et al. CVPR 2017
Unified tensor flow architecture for comparing speed, accuracy, and memory usage

Accuracy vs speed (COCO)

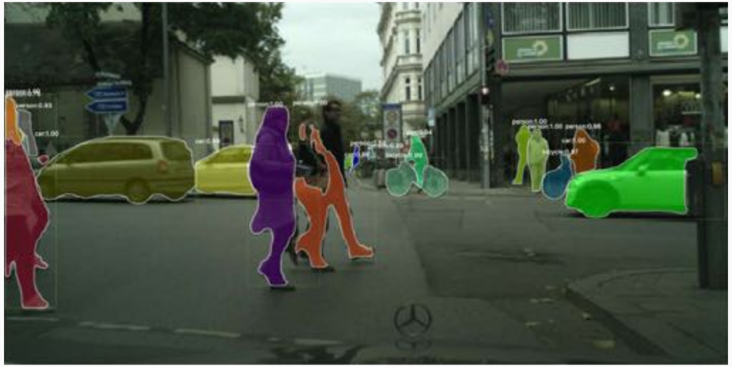


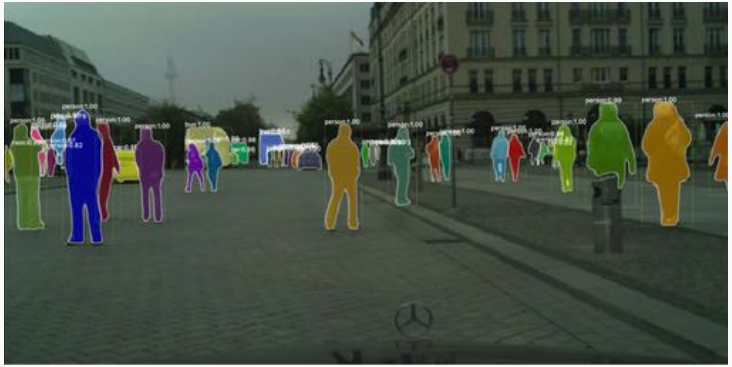
Speed/accuracy trade-offs for modern convolutional object detectors, Huang et al. CVPR 2017

Object instance segmentation

Instance segmentation

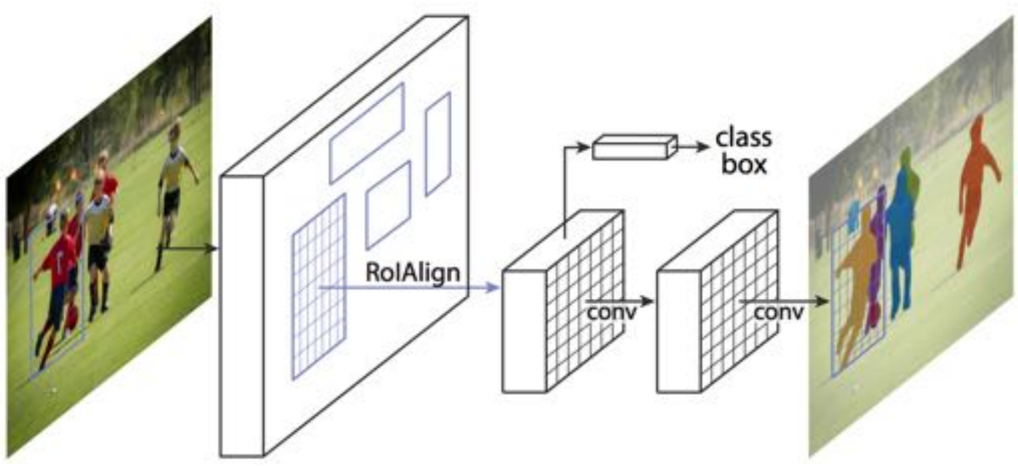
- Given an image produce instance-level segmentation
  - Which class does each pixel belong to?
  - Which instance does each pixel belong to?





Mask R-CNN

- Extend Faster R-CNN to predict mask as well as a box



Mask R-CNN, He et al., CVPR 2017

Mask R-CNN - video example

Mask R-CNN, He et al., CVPR 2017

https://www.youtube.com/watch?v=UWtac4cFERM