UniGe | MaLGa

Understanding the semantics of an image: introduction

Francesca Odone francesca.odone@unige.it

Image and semantics

Image classification: associate a label to an image

$$(x_1, y_1), \ldots, (x_n, y_n)$$

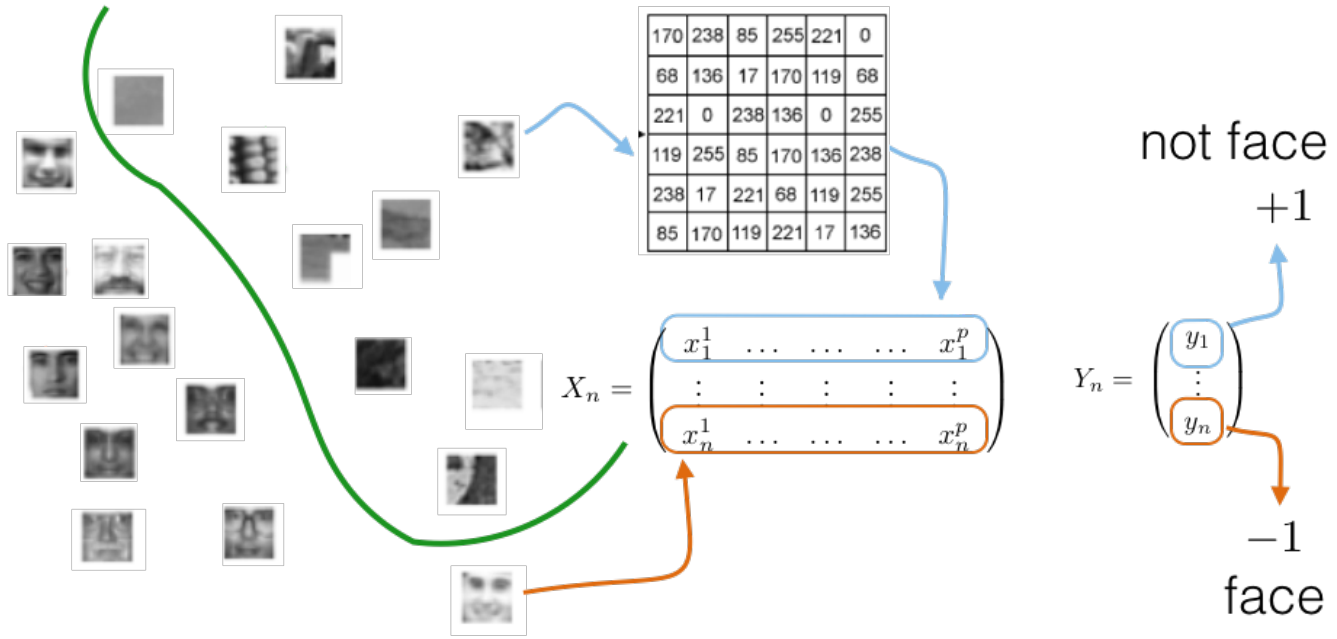
$$x_i \in \mathbb{R}^p \quad \text{and} \quad y_i \in Y, \quad i = 1, \ldots, n$$

$$X_n = \begin{pmatrix} x_1^1 & \ldots & \ldots & \ldots & x_1^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^1 & \ldots & \ldots & \ldots & x_n^p \end{pmatrix} \qquad Y_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Image and semantics

Image classification: associate a label to an image

a classical **binary classificatio**n problem



at the end of the training/validation procedure
you derive a model for *faces*

Image semantics

Multi-class classification: the set of possible labels grows

Image classification vs object recognition

Similar concepts, if the image contains objects

Image classification

Image classification

Here there is also a main “object”
(I could also talk about object recognition)

Object recognition vs instance recognition

notice the difference

instance recognition - “his car” “that bicycle” “my mug”

category recognition - generic object recognition
“cars”, “bicycles”, “mugs”

Image representations

What are they, exactly?

$$(x_1, y_1), \ldots, (x_n, y_n)$$

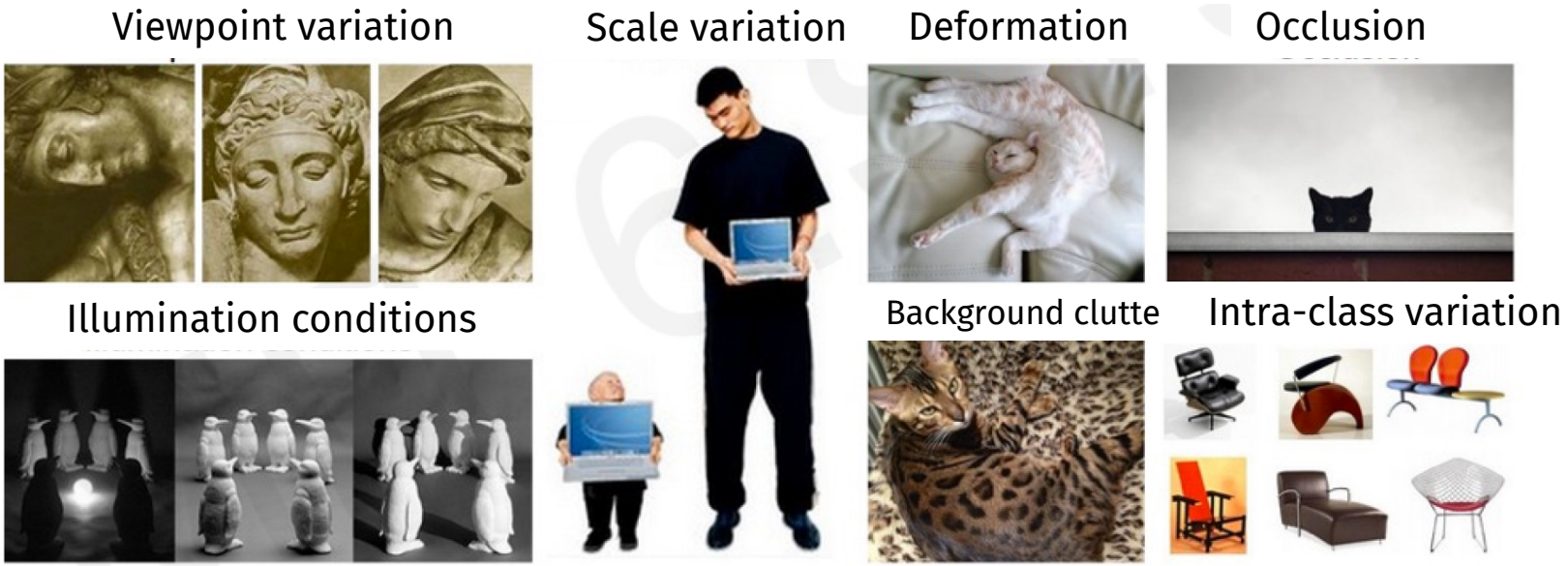
$$x_i \in \mathbb{R}^p \quad \text{and} \quad y_i \in Y, \quad i = 1, \ldots, n$$

$$X_n = \begin{pmatrix} x_1^1 & \ldots & \ldots & \ldots & x_1^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^1 & \ldots & \ldots & \ldots & x_n^p \end{pmatrix} \qquad Y_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Image representations

They must be robust to intra-class variations of different types

They must describe the whole image (or the part we are about to classify)



(global) image representations

The list of possibilities is huge

A list incorporating what we have seen so far:

- Pixels (not so robust)
- Graylevels /color /gradient histograms (not too descriptive)
- Local keypoints, in Bag of keypoints (still too “hand-crafted”)

.....CNN-features (learnt in a supervised or unsupervised way)

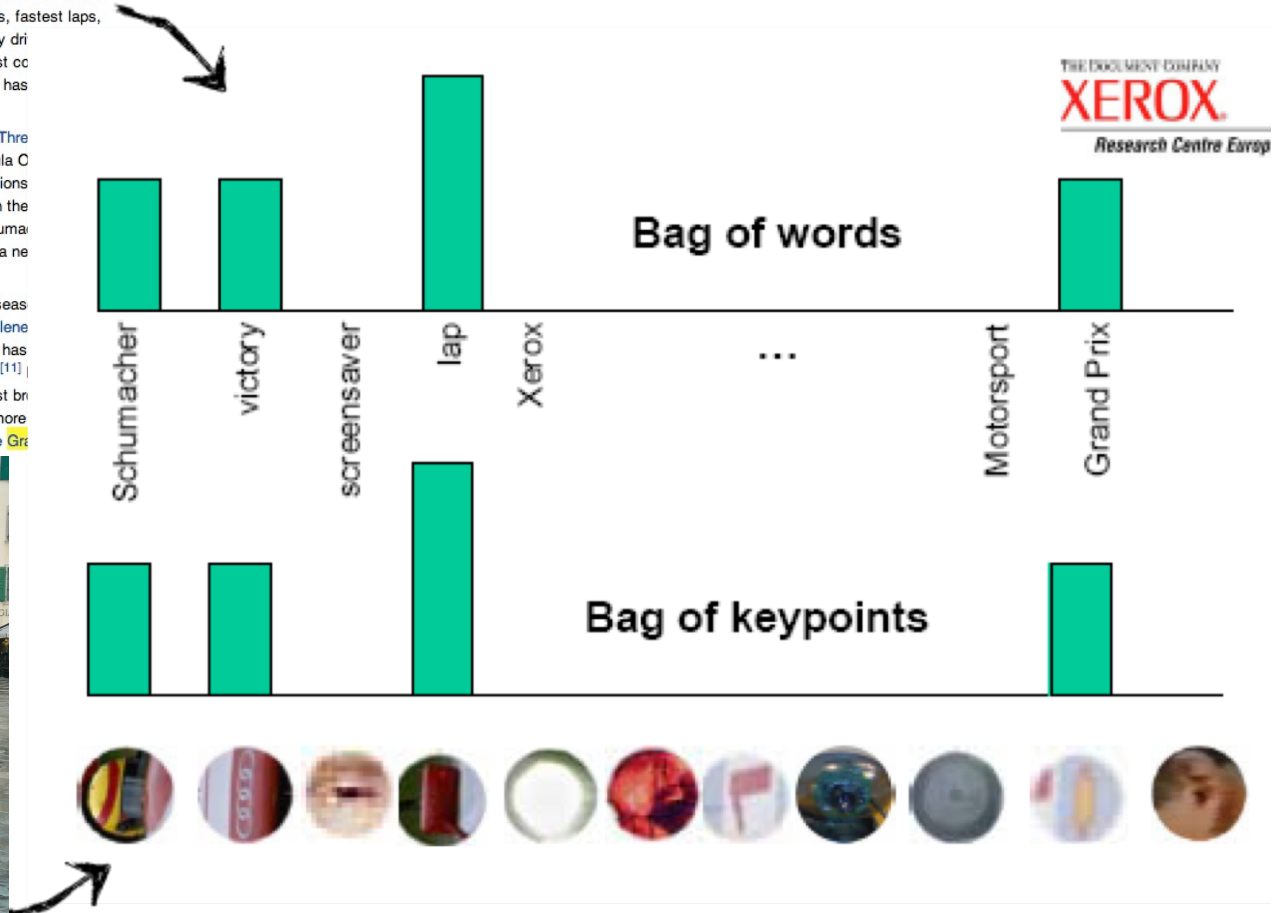
Finding a global feature vector from local keypoints

Bag of words: the inspiration comes from text analysis

Michael Schumacher (German pronunciation: ['mɪçaʔeːl 'ʃuːmaxɐ] (listen); born 3 January 1969) is a retired German racing driver. Schumacher is a seven-time Formula One World Champion and is widely regarded as one of the greatest F1 drivers of all time.[1][2][3][4] He holds many of Formula One's driver records, including most championships, race victories, fastest laps, pole positions, points scored and most races won in a single season – 13 in 2004. In 2002 he became the only dri Formula One history to finish in the top three in every race of a season and then also broke the record for most cc podium finishes. According to the official Formula One website he is "statistically the greatest driver the sport has seen".[5]

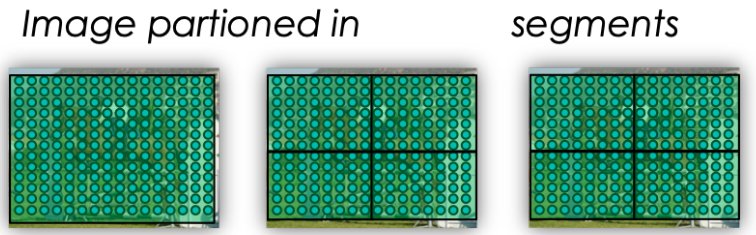
After beginning with karting, Schumacher won German drivers' championships in Formula König and Formula Thre joining Mercedes in the World Sportscar Championship. After one Mercedes-funded race for the Jordan Formula C Schumacher signed as a driver for the Benetton Formula One team in 1991. After winning consecutive champions Benetton in 1994/5, Schumacher moved to Ferrari in 1996 and won another five consecutive drivers' titles with the 2000 to 2004. Schumacher retired from Formula One driving in 2006 staying with Ferrari as an advisor.[6] Schuma to return for Ferrari part-way through 2009, as cover for the badly injured Felipe Massa, but was prevented by a ne He later signed a three-year contract to drive for the new Mercedes GP team starting in 2010.[7][8][9]

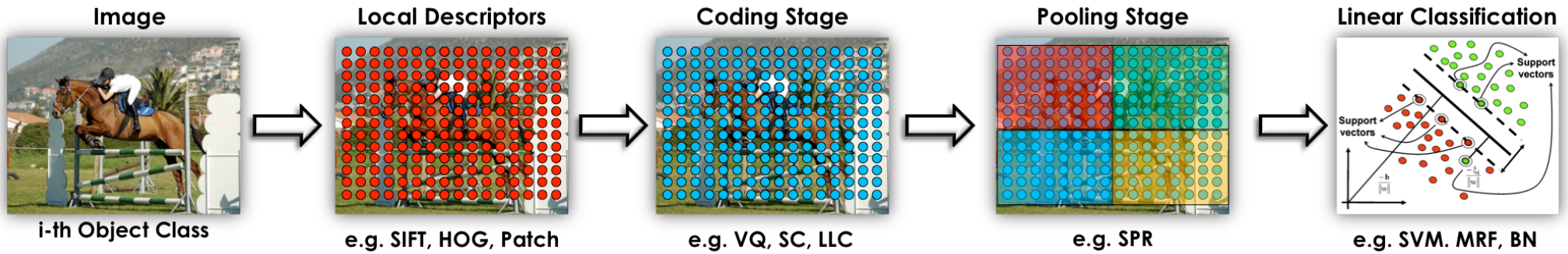
His career has not been without controversy, including being twice involved in collisions in the final race of a seas determined the outcome of the world championship, with Damon Hill in 1994 in Adelaide, and with Jacques Villene in Jerez.[10] Off the track Schumacher is an ambassador for UNESCO and a spokesman for driver safety. He has involved in numerous humanitarian efforts throughout his life and donated tens of millions of dollars to charity.[11] his younger brother Ralf Schumacher are the only brothers to win races in Formula One, and they were the first br finish 1st and 2nd in the same race, in Montreal in 2001. The two brothers repeated this achievement in four more 2001 French Grand Prix, the 2002 Brazilian Grand Prix, the 2003 Canadian Grand Prix and the 2004 Japanese Gra



coding - pooling image classification

pre-defined partitioning...





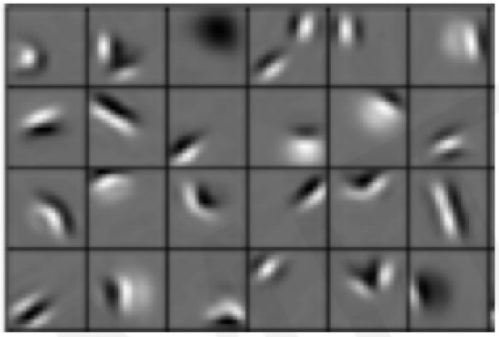
unsupervised

supervised

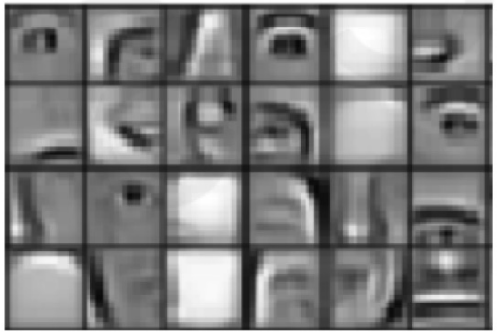
Towards image classification

- Is it possible to learn the most appropriate representation, possibly with a hierarchy, directly from the data?

Low-level features



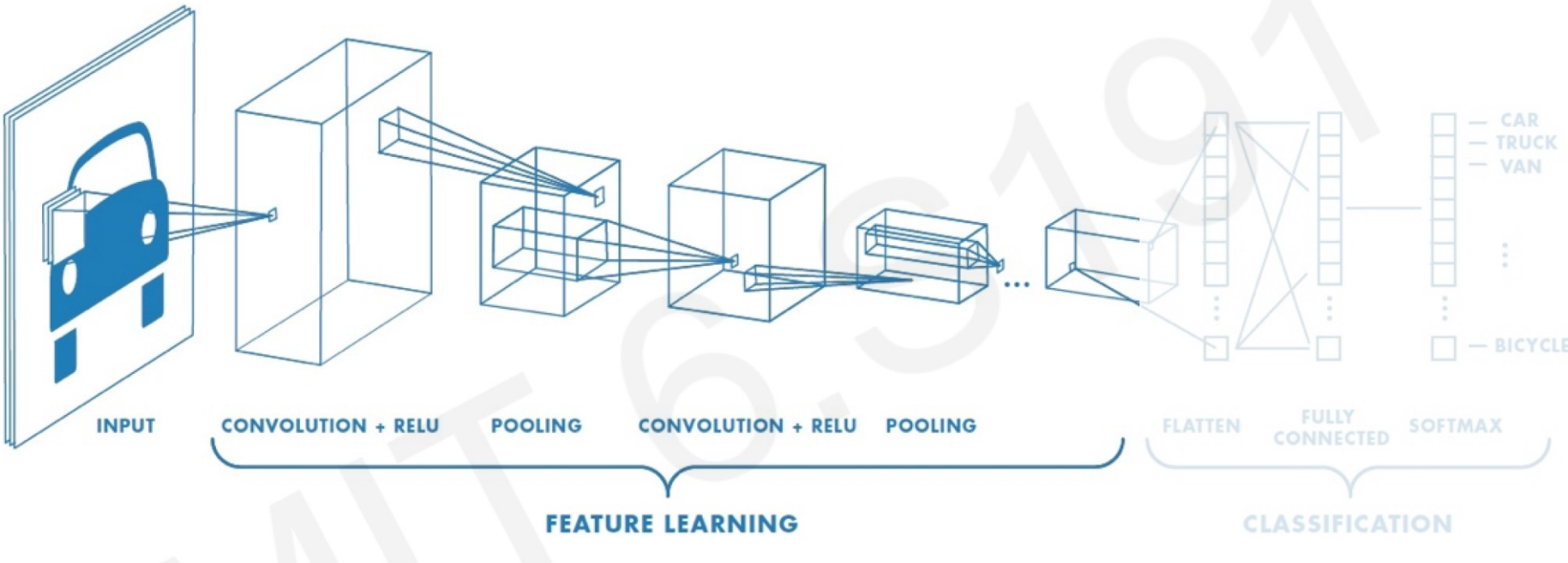
Mid-level features



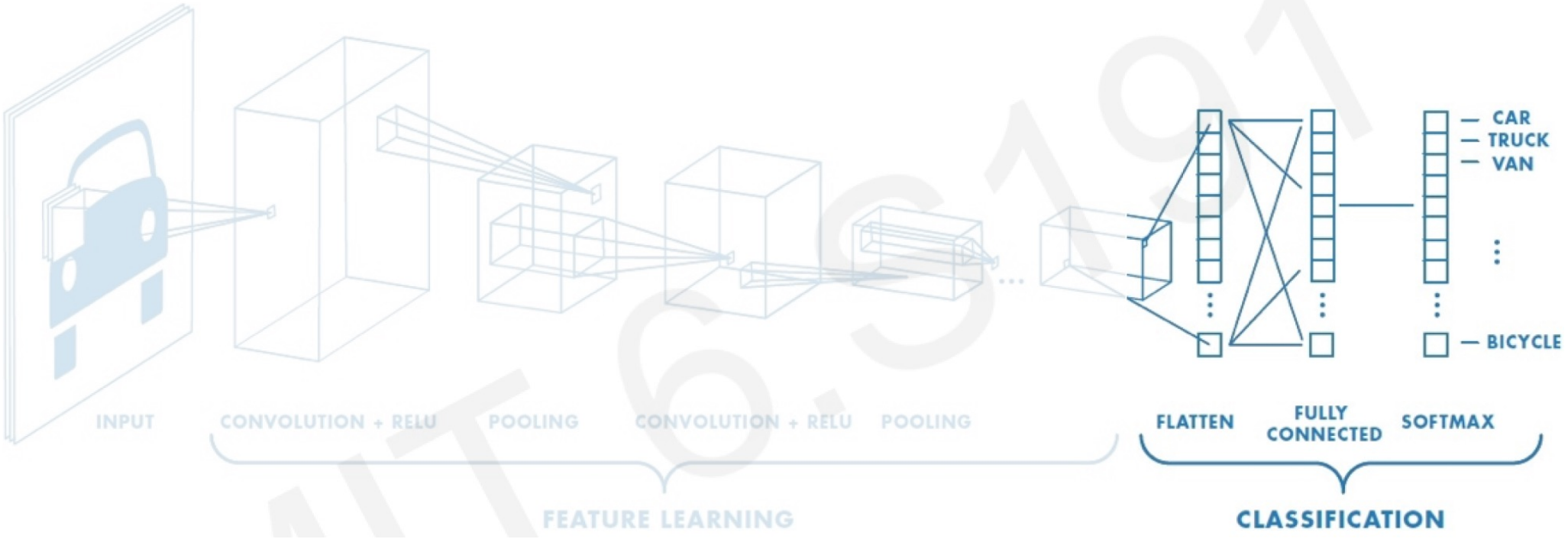
High-level features

- From features engineering to features learning

A typical CNN

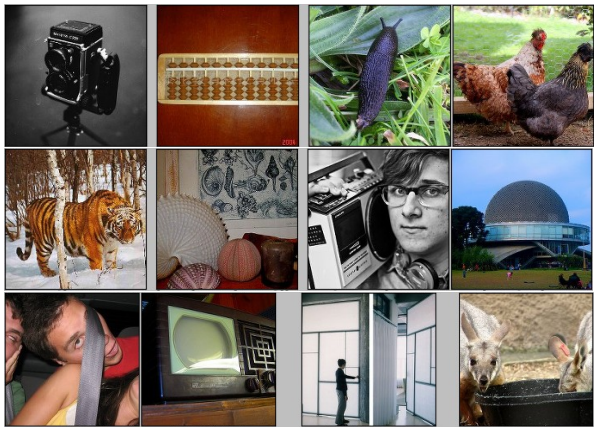


A typical CNN



$$softmax(y_i) = \frac{e^{y_i}}{\sum_j e^{y_i}}$$

The key of success of CNN (ImageNet)

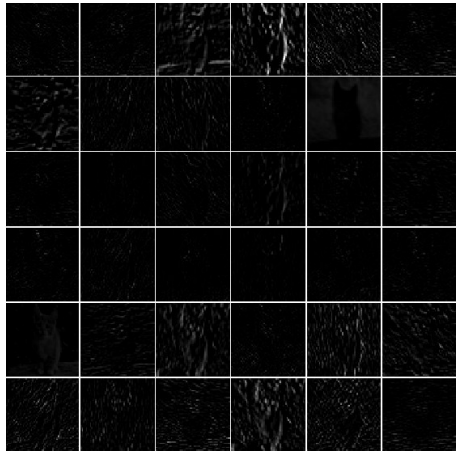


[Deng et al. CVPR 2009]

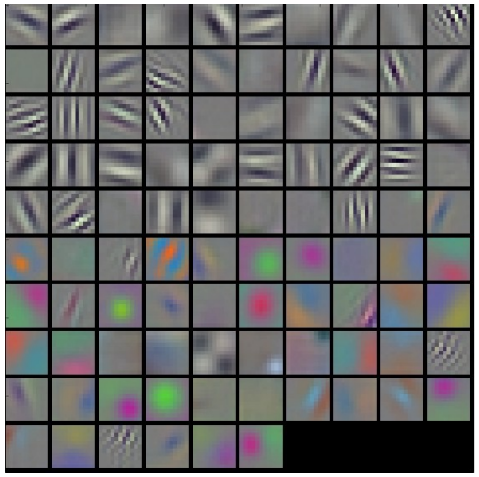
- …. the ImageNet Challenge (2012)
- ~14 million labeled images, 20k classes
- Images gathered from Internet
- Human labels via Amazon Turk (huge scale manual on-line manual labelling)
- ImageNet Object Recognition Challenge: 1.2 million training images, 1000 classes

Visualizing what CNN learn

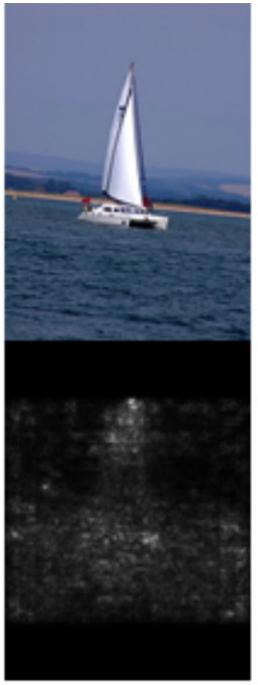
- Visualizing the layer activations



Visualizing the learnt filters



- Images maximizing the class score.

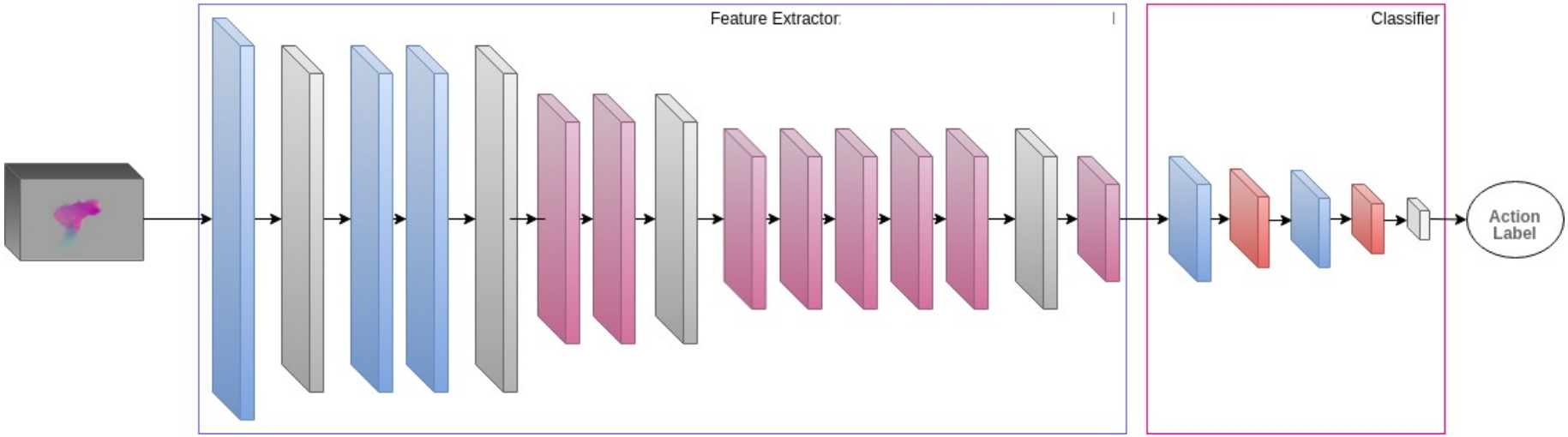


Class saliency map

UniGe | MaLGa

- More infos here: https://cs231n.github.io/understanding-cnn/

Pre-trained features and transfer learning

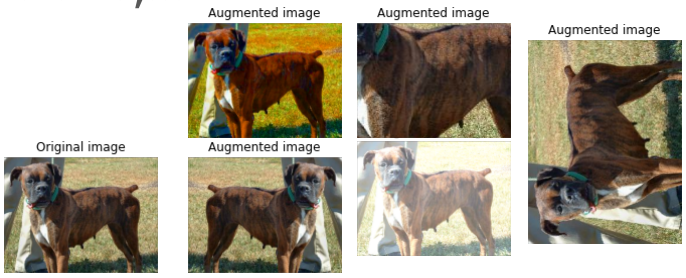


It refers to the possibility of exploiting knowledge in terms of pre-trained models that can be used on different data and tasks (with some constraint)

- CNN can be used as a feature extractor
- Fine-tuning the CNN: (some of) the weights are adapted to the new problem/data starting from the pre-trained model by continuing the backpropagation
- Several models pre-trained on ImageNet are available

CNN training images

- Large architectures are very data hungry
- One of the trick for coping with lack of data is data augmentation
  - The idea is to generate more data by applying some transformation to the image
  - Examples: rotations, scaling, lighting conditions,...



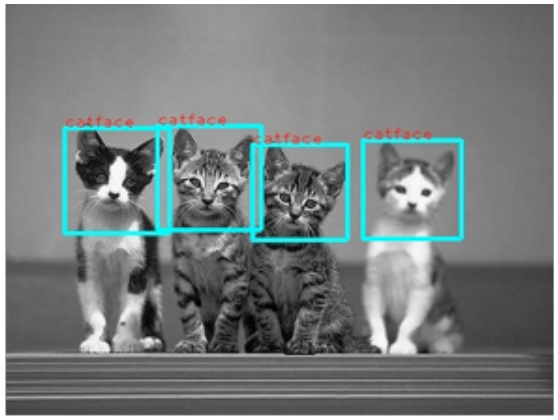
- A recent strategy is to resort to synthetic data, but fine tuning on real data might be necessary to reduce the semantic gap between synthetic and real

Besides image classification

Object detection in its classical formulation

object detection is in essence a classification problem

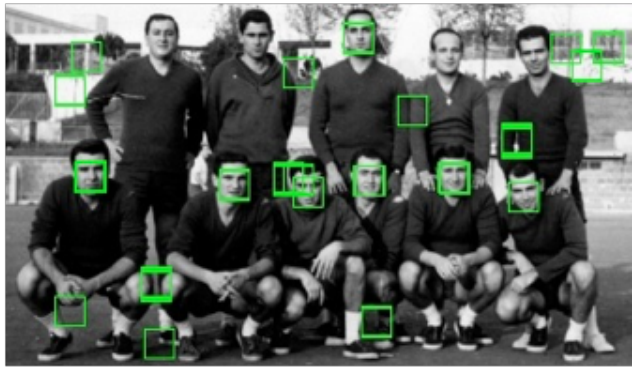
image regions of variable size are classified: *is it an instance of the object or not?*



A special case occurs when we have only a class of interest

unbalanced classes

in this 380x220 px image we perform ~6.5×10^5 tests and we should find only 11 positives



the training set contains

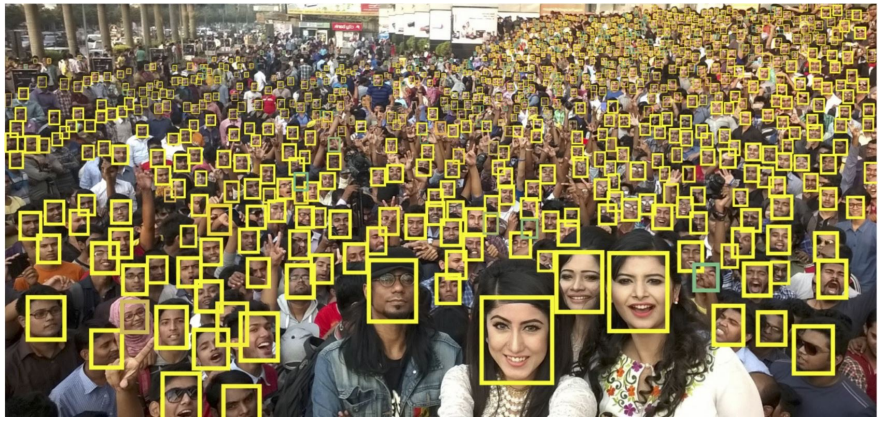
images of positive examples (the object)

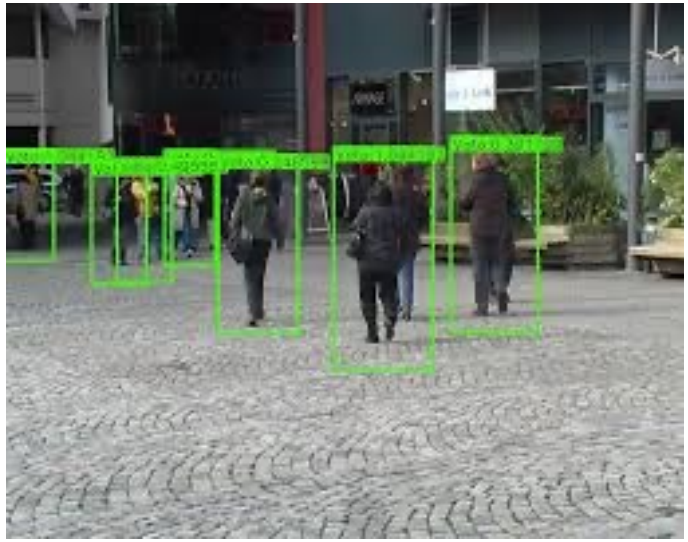
negative examples (background)

1 Class Object detection

Given **a class** of objects of interest (Face, Pedestrian,..) and given an input image I which contains N instances of the object, locate all the instances

that is, find $\{(x_i, y_i, w_i, h_i)\}; i = 1..N$ the center and dimensions of boxes that best localize the objects in image I

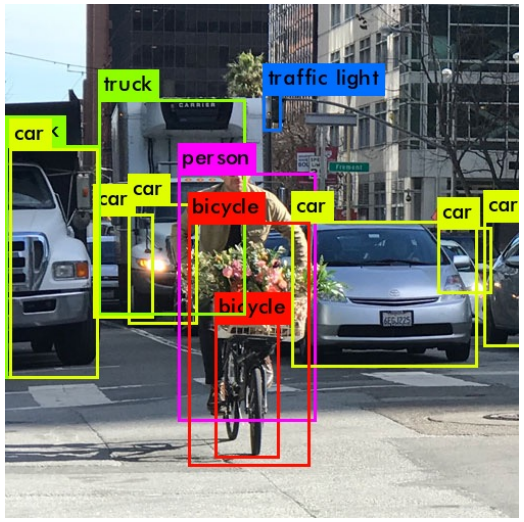




Multi-class object detection

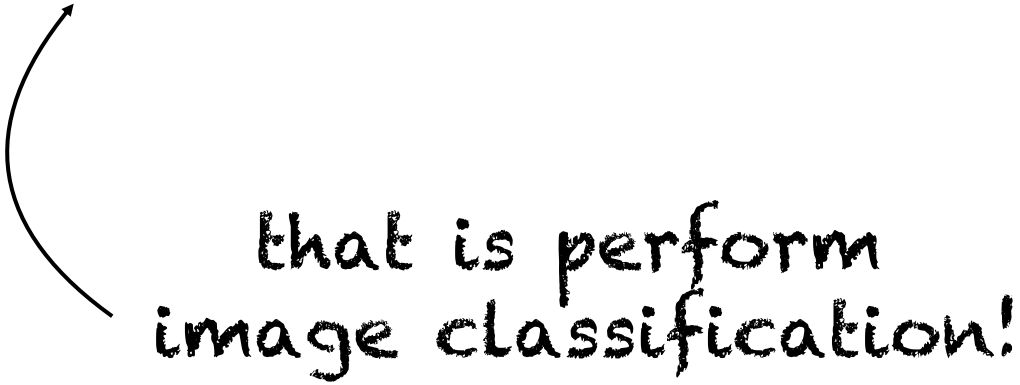
Given a **set of classes of interest** C (e.g. Street objects, office objects,..) and given an input image I that contains N object instances: find the center, the class and dimensions of boxes (aligned with the coordinates system) that best localize the objects in image I, formally:

$$\{(c_i, x_i, y_i, w_i, h_i)\}\,; c_i \in C, i = 1..N$$



Object detection: sliding window basic idea

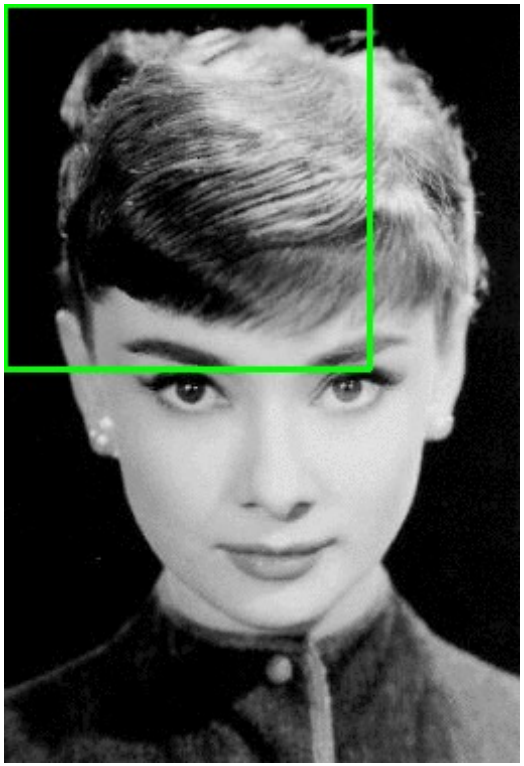
Slide a window across image and evaluate an object model at every location



is it a face?

YES

NO



Object detection: sliding window basic idea

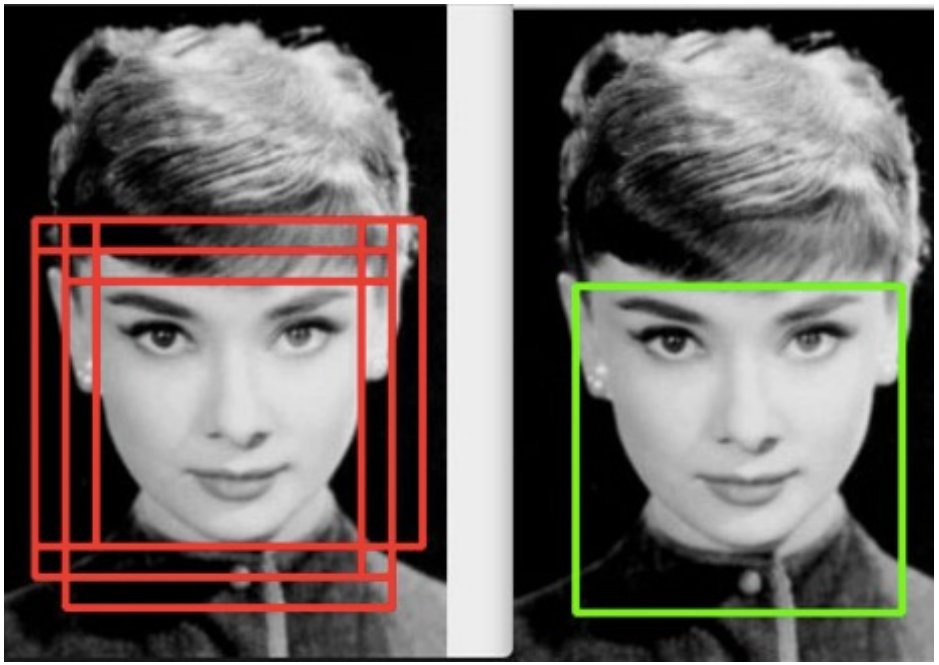
Slide a window across image and evaluate an object model at every location

that is perform image classification!

is it a face?

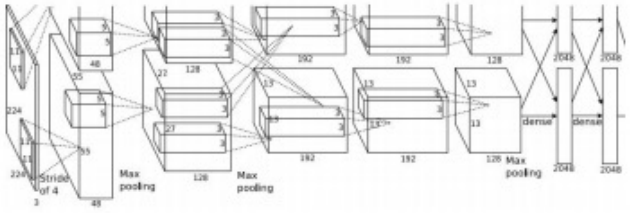
YES

non maxima suppression



Are sliding windows applicable with CNN ?

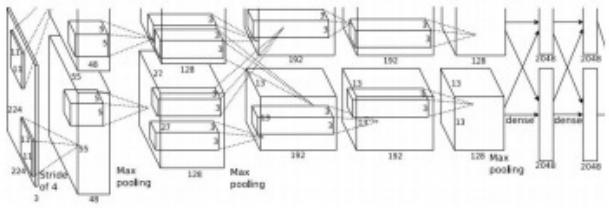
Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? NO
Cat? NO
Background? YES

Are sliding windows applicable with CNN ?

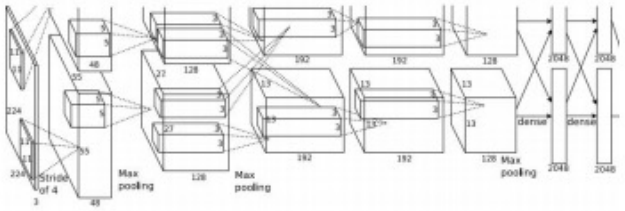
Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? YES
Cat? NO
Background? NO

Are sliding windows applicable with CNN ?

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 and scales, very computationally expensive!

Object detection approaches

Here we mention only supervised methods

Two stage CNN detectors

- for instance Region-based CNN (R-CNN family)

One stage CNN detectors

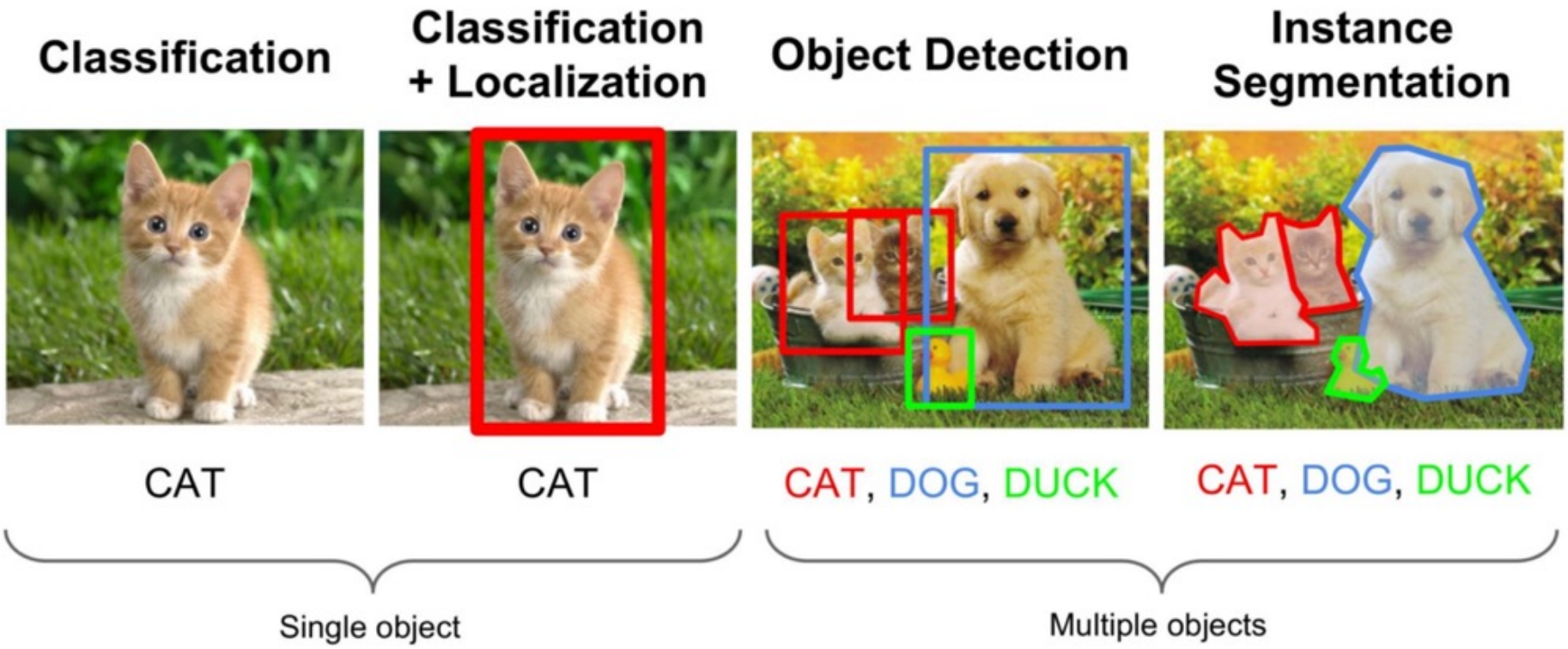
- For instance YOLO family

Transformer based methods (eg DETR)

...

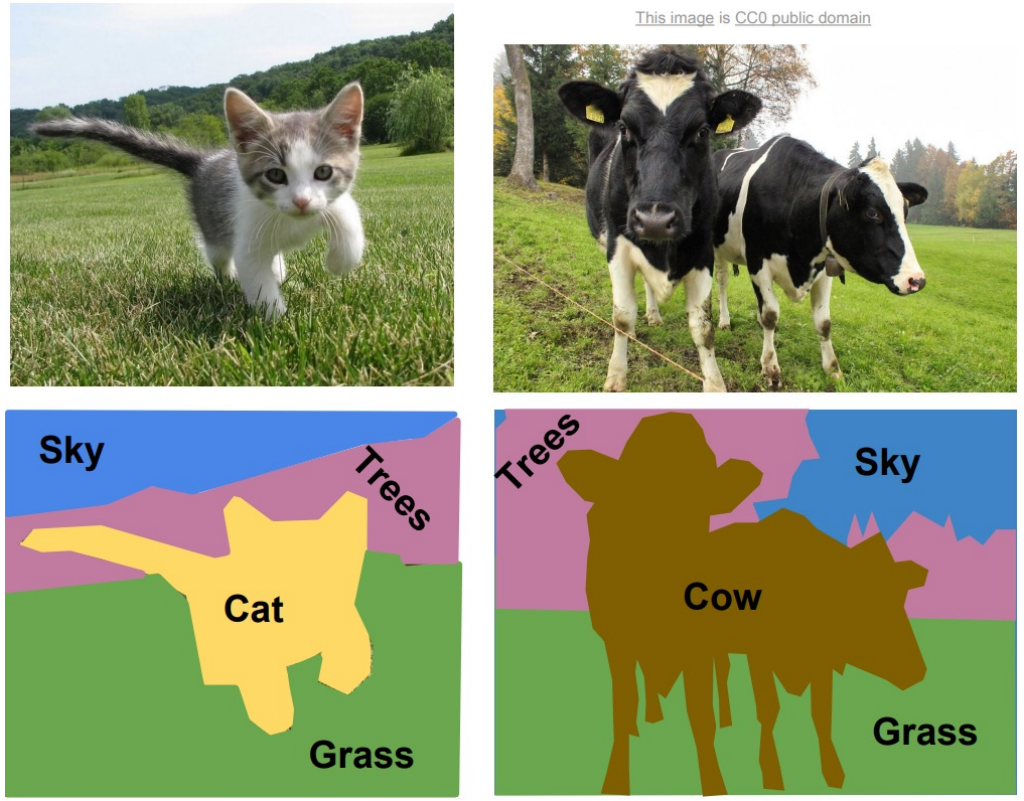
Object detection

Different nuances



Notice the difference

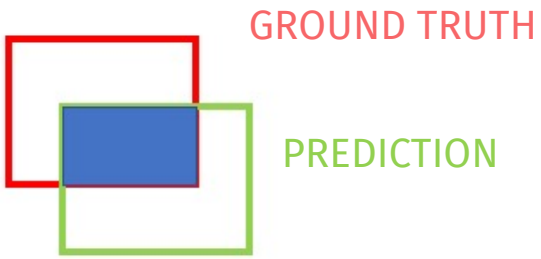
With (semantic) segmentation the goal is to associate a semantic attribute to **every** pixel in the image



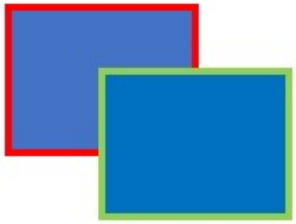
Object detection evaluation

How are object detectors evaluated?

IOU or JACCARD METRIC



Intersection over Union= ——————

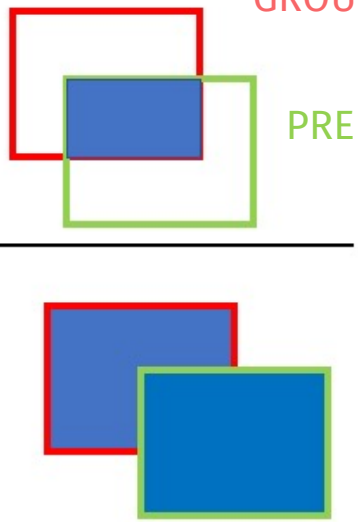


How are object detectors evaluated?

BINARY CASE

IOU METRIC

Intersection over Union= $\frac{\text{GROUND TRUTH} \cap \text{PREDICTION}}{\text{GROUND TRUTH} \cup \text{PREDICTION}}$



TRUE POSITIVE
if IOU >= THRESHOLD

FALSE POSITIVE
if IOU < THRESHOLD

FALSE NEGATIVE
A ground truth region with no associated prediction

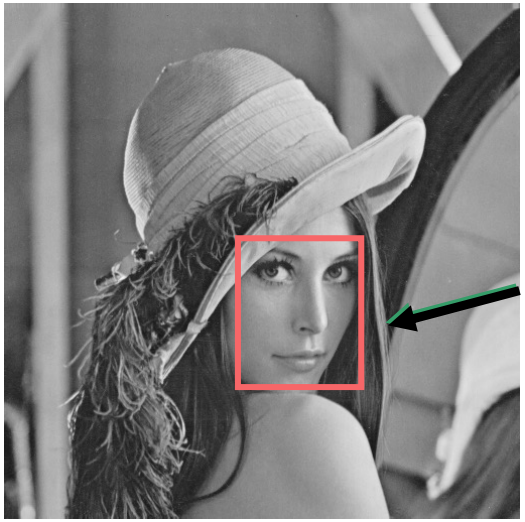
TRUE NEGATIVE
All the other regions

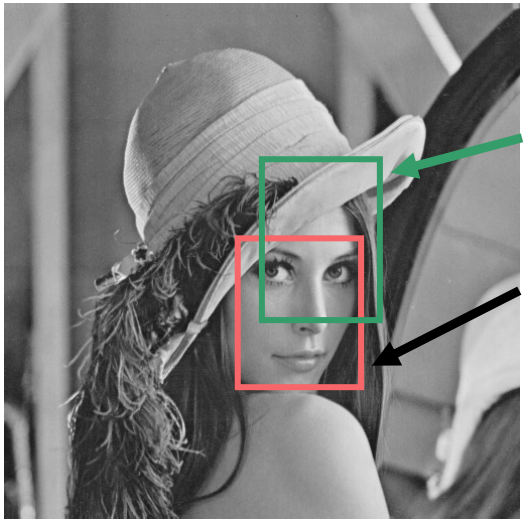
Threshold rule of thumb: greater than 0.5

EXAMPLES

GROUND TRUTH

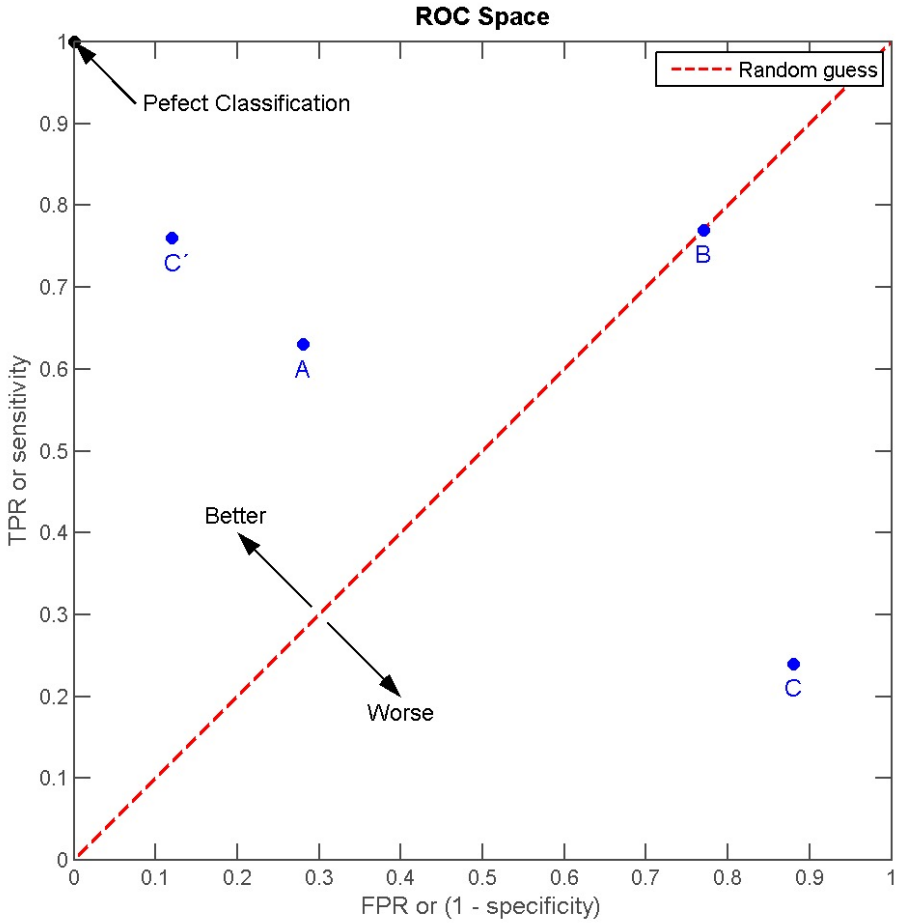
PREDICTION / DETECTION / ESTIMATION

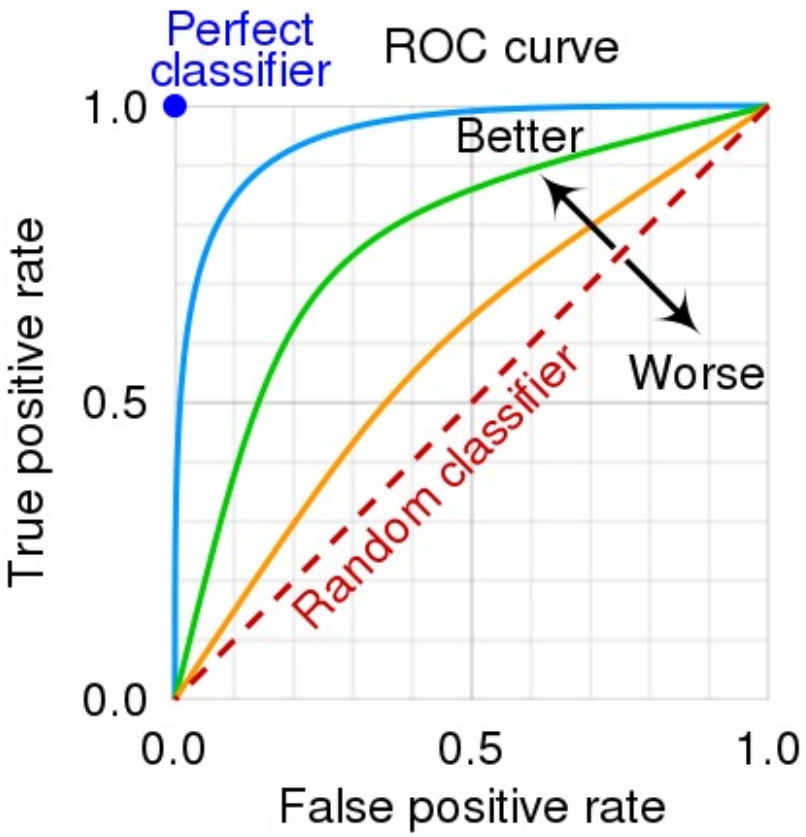




Receiver Operating Characteristic Curve

Plot to illustrate the performance of a binary classifier





https://en.wikipedia.org/wiki/Receiver_operating_characteristic

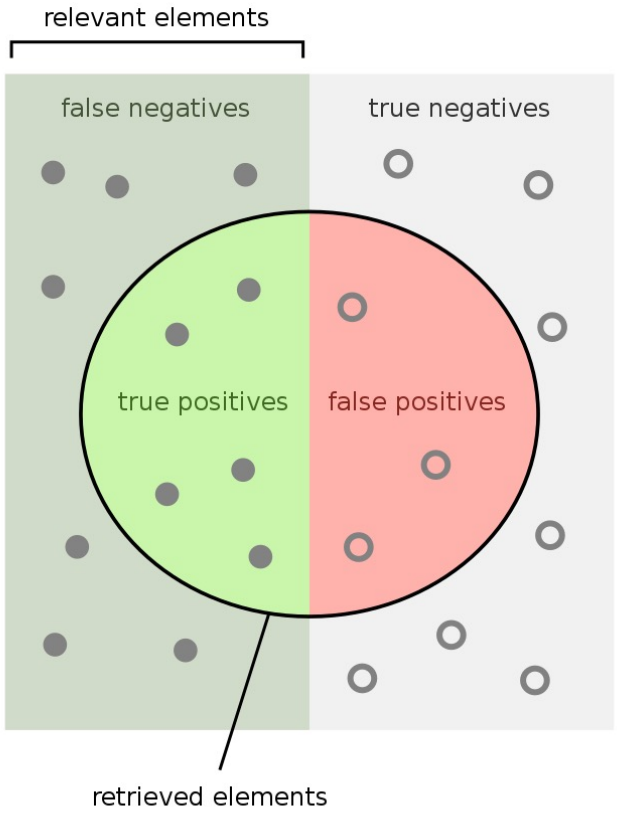
Precision-Recall and F1 score

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

Estimated positives

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

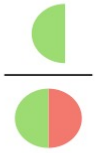
Real positives



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

How many retrieved items are relevant?

Precision =



How many relevant items are retrieved?

Recall =

Precision-Recall curve

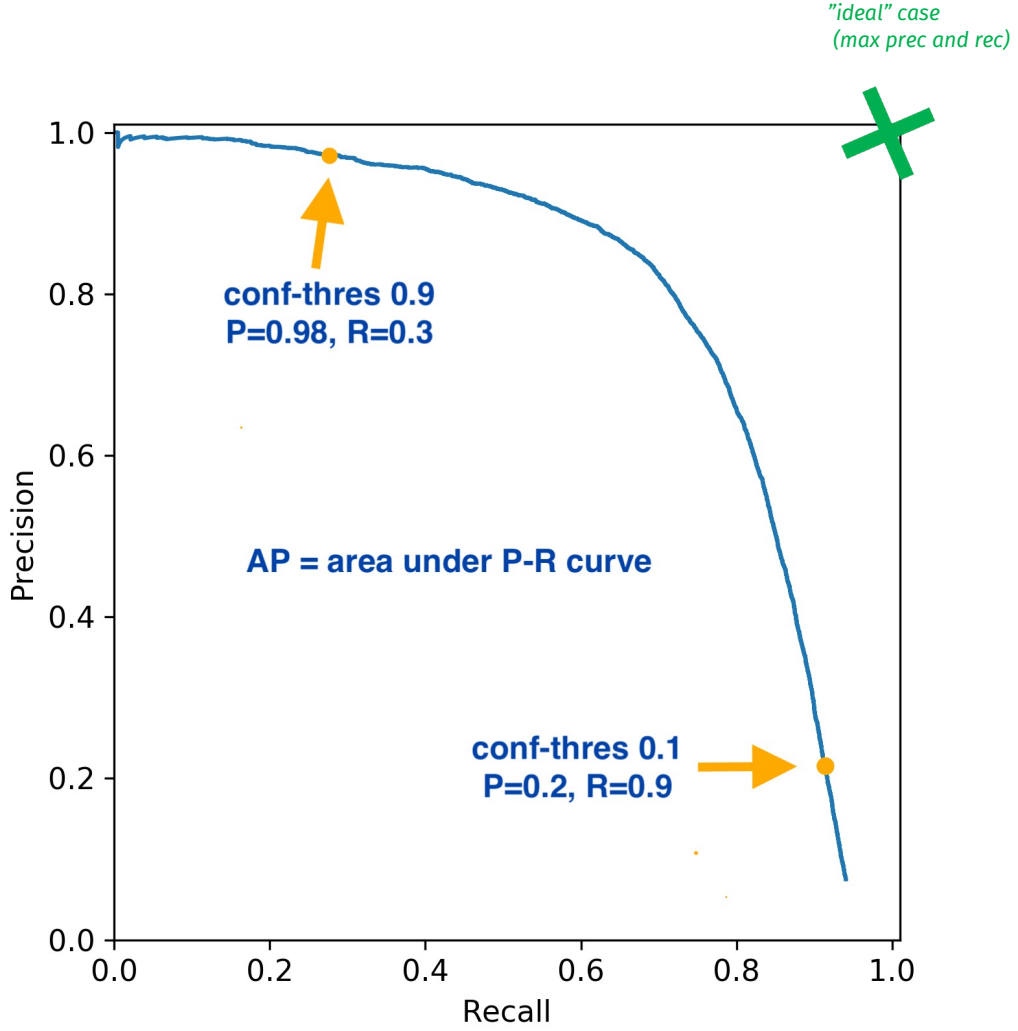
$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

Estimated positives

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

Real positives

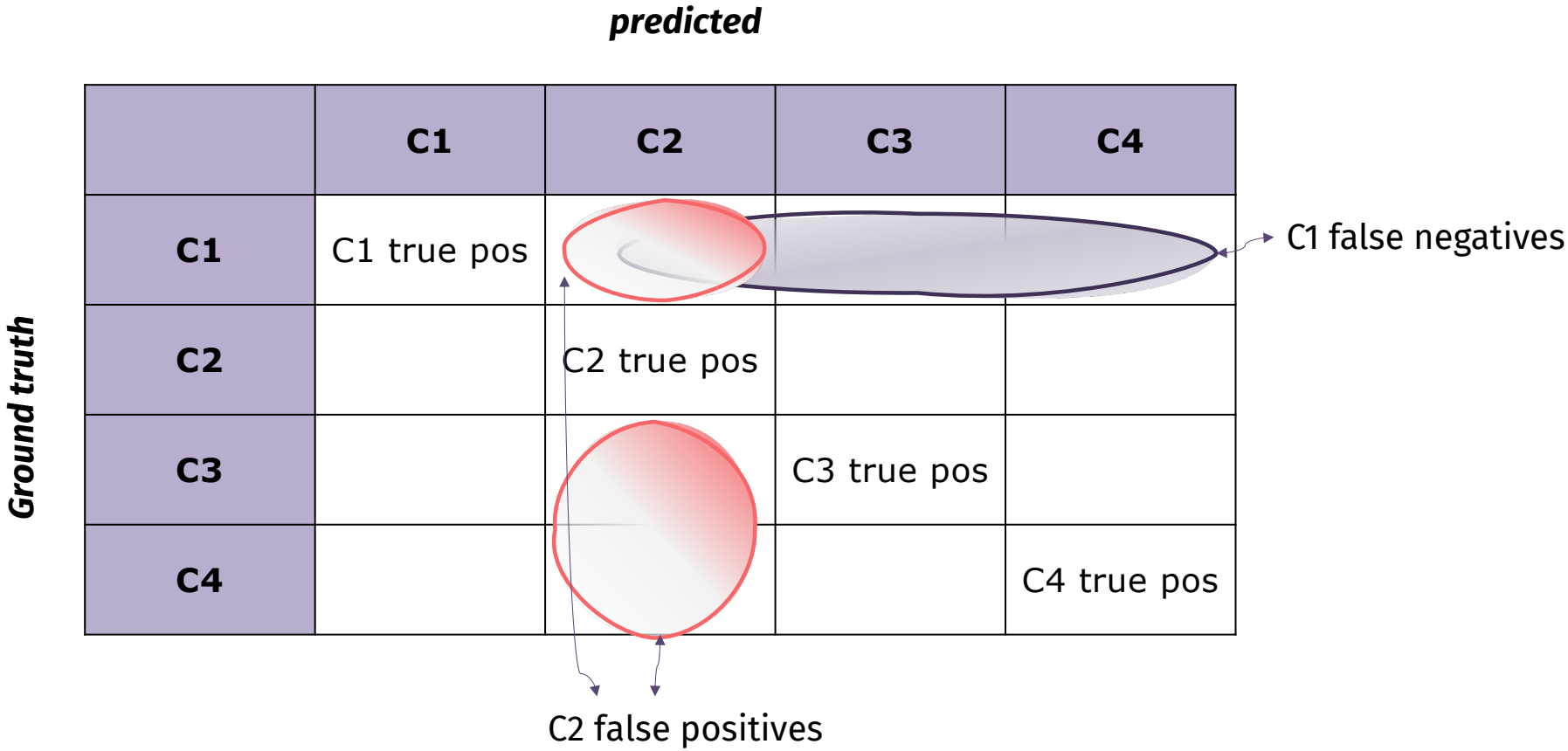
AP = area under P-R curve (for one class)
mAP = mean AP for several classes



Precision-recall curve

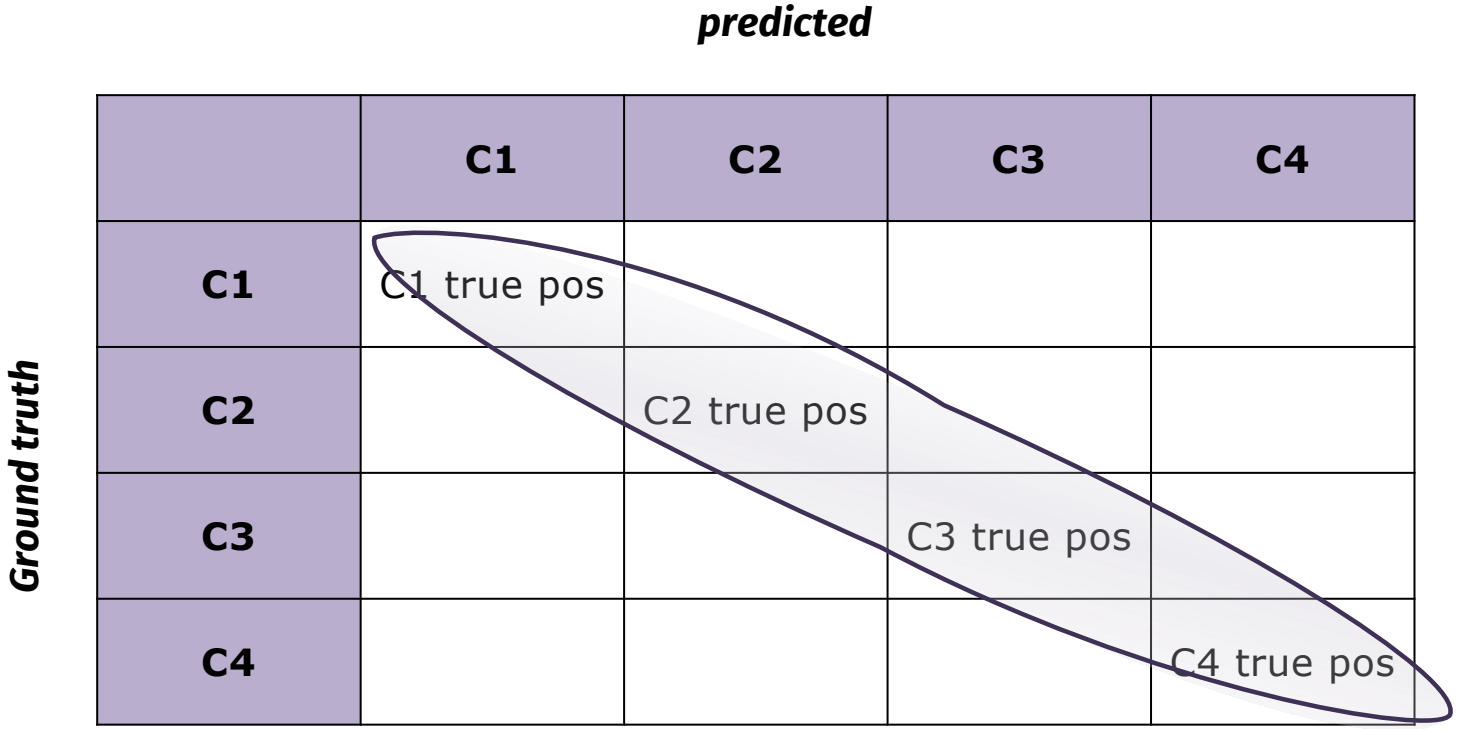
Multi-class evaluation

Confusion matrices



Multi-class evaluation

Confusion matrices



$$Acc_{overall} = \sum_{i} C(i,i)$$

UniGe