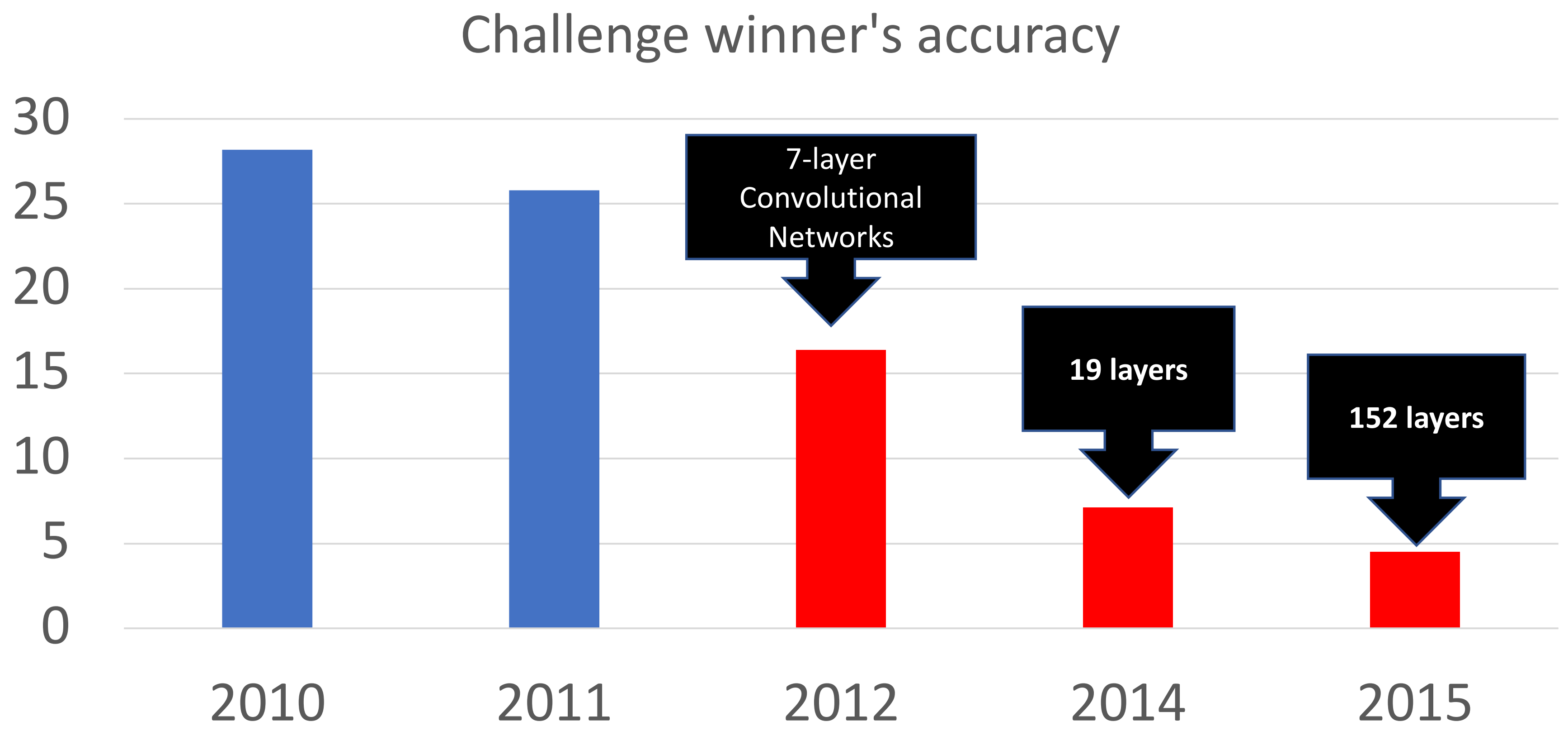
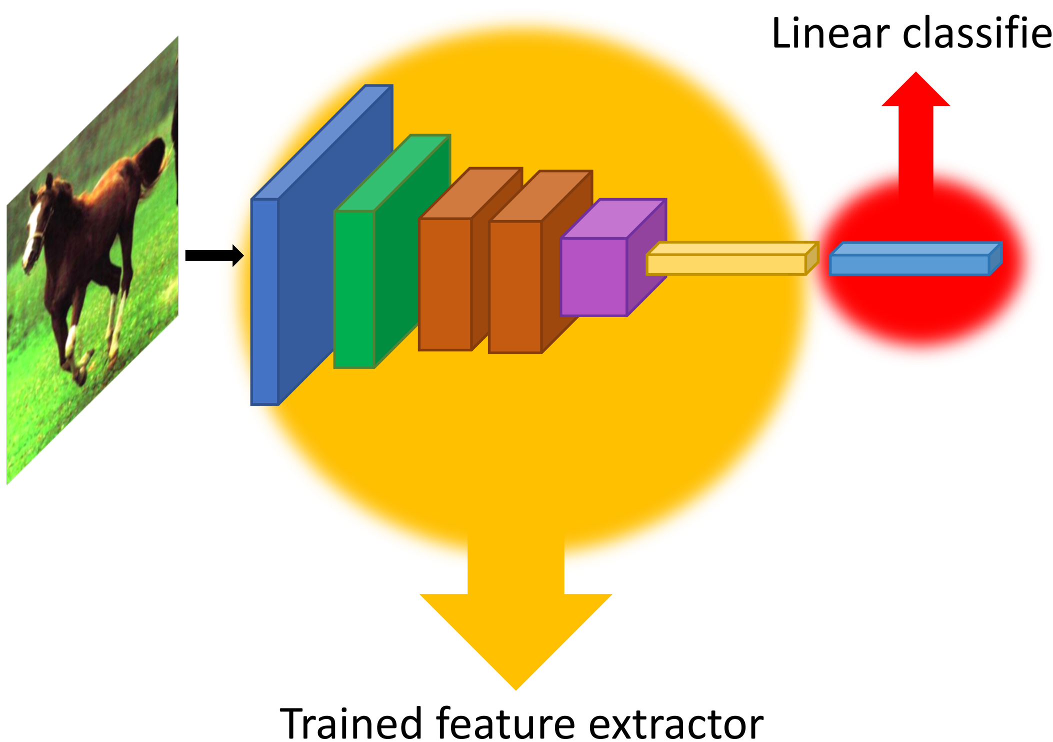
Transfer learning with convolutional networks


Challenge winner's accuracy
30
25
20
15
10
5
0
7-layer Convolutional Networks
19 layers
152 layers
2010
2011
2012
2014
2015

Transfer learning with convolutional networks

- What do we do for a new image classification problem?
- Key idea:
  - *Freeze* parameters in feature extractor
  - *Retrain* classifier



Transfer learning with convolutional networks

Dataset	Non-Convnet Method	Non-Convnet perf	Pretrained convnet + classifier	Improvement
Caltech 101	MKL	84.3	87.7	+3.4
VOC 2007	SIFT+FK	61.7	79.7	+18
CUB 200	SIFT+FK	18.8	61.0	+42.2
Aircraft	SIFT+FK	61.0	45.0	-16
Cars	SIFT+FK	59.2	36.5	-22.7

Why transfer learning?

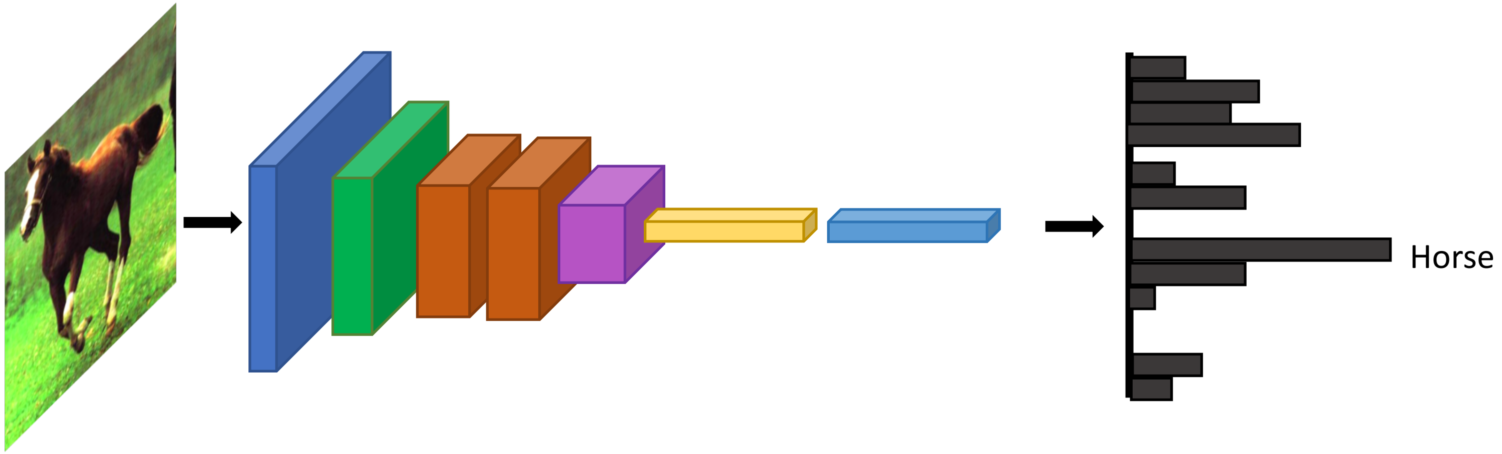
- Availability of training data

- Computational cost

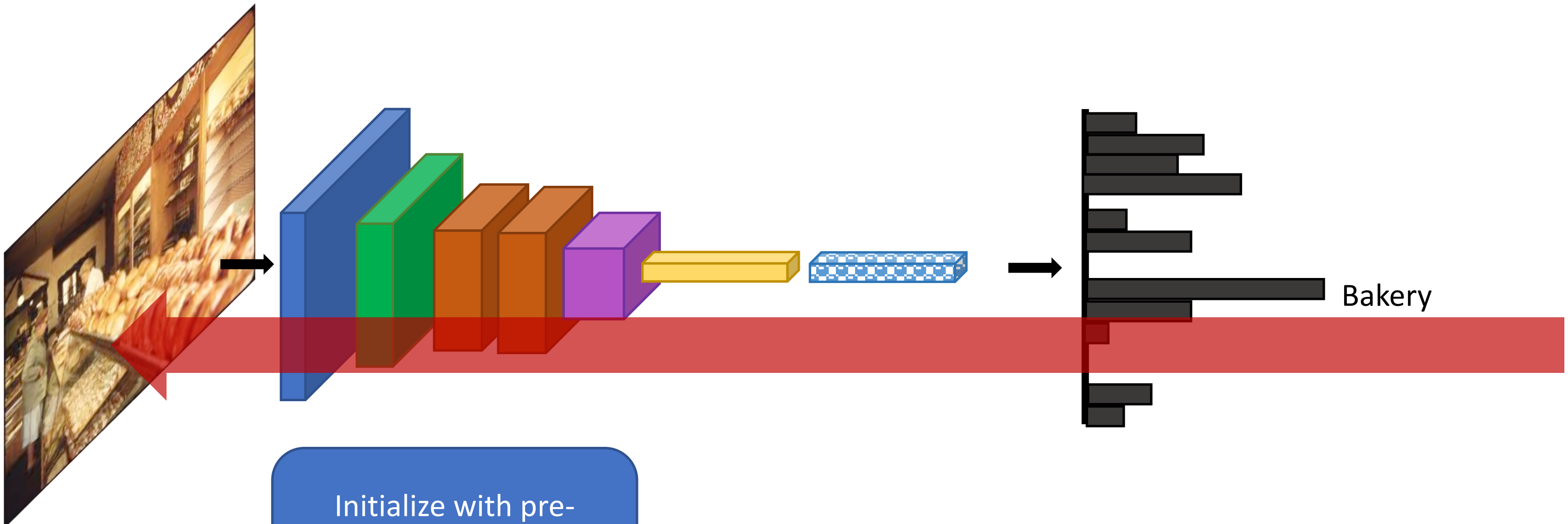
- Ability to pre-compute feature vectors and use for multiple tasks

- *Con: NO end-to-end learning*

Finetuning



Finetuning



Initialize with pre-trained, then train with low learning rate

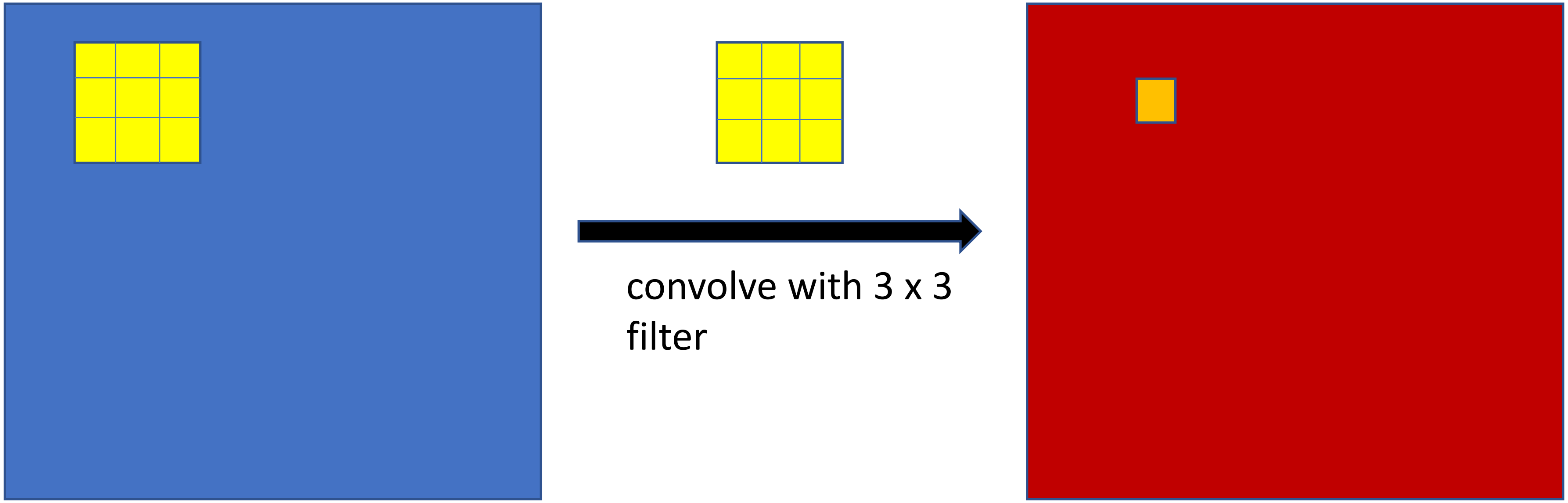
Finetuning

Dataset	Non-Convnet Method	Non-Convnet perf	Pretrained convnet + classifier	Finetuned convnet	Improvement
Caltech 101	MKL	84.3	87.7	88.4	+4.1
VOC 2007	SIFT+FK	61.7	79.7	82.4	+20.7
CUB 200	SIFT+FK	18.8	61.0	70.4	+51.6
Aircraft	SIFT+FK	61.0	45.0	74.1	+13.1
Cars	SIFT+FK	59.2	36.5	79.8	+20.6

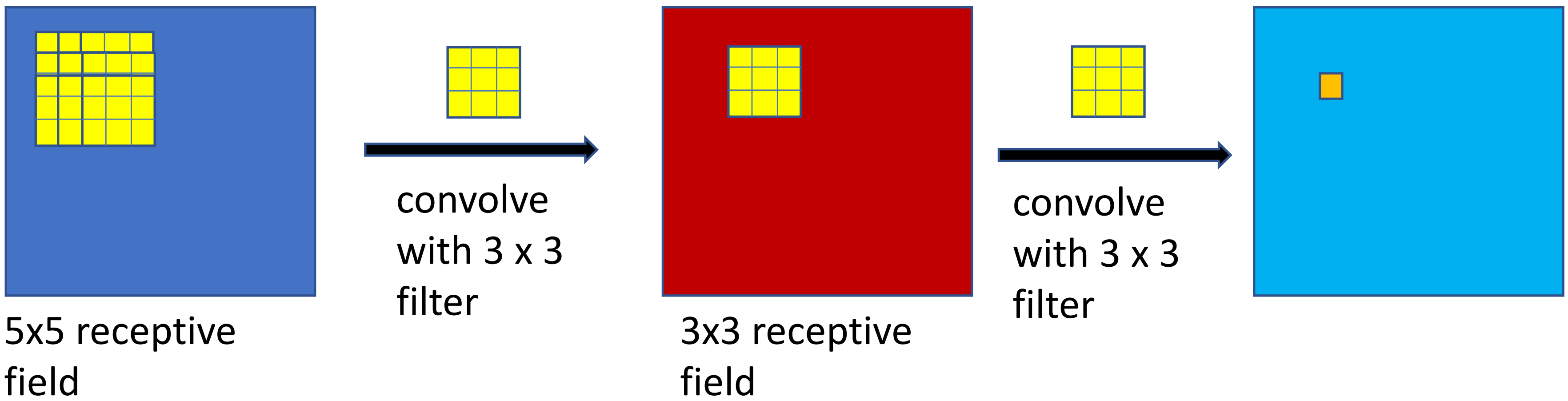
Visualizing convolutional networks

Receptive field

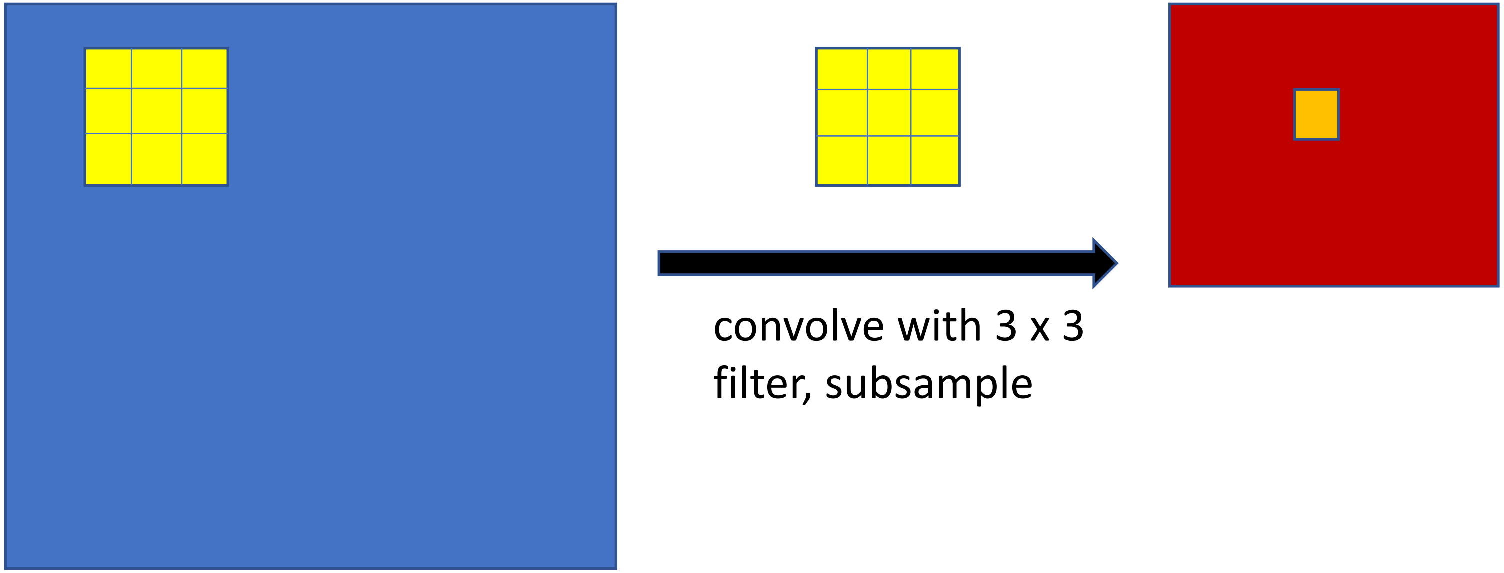
- Which input pixels does a particular unit in a feature map depends on



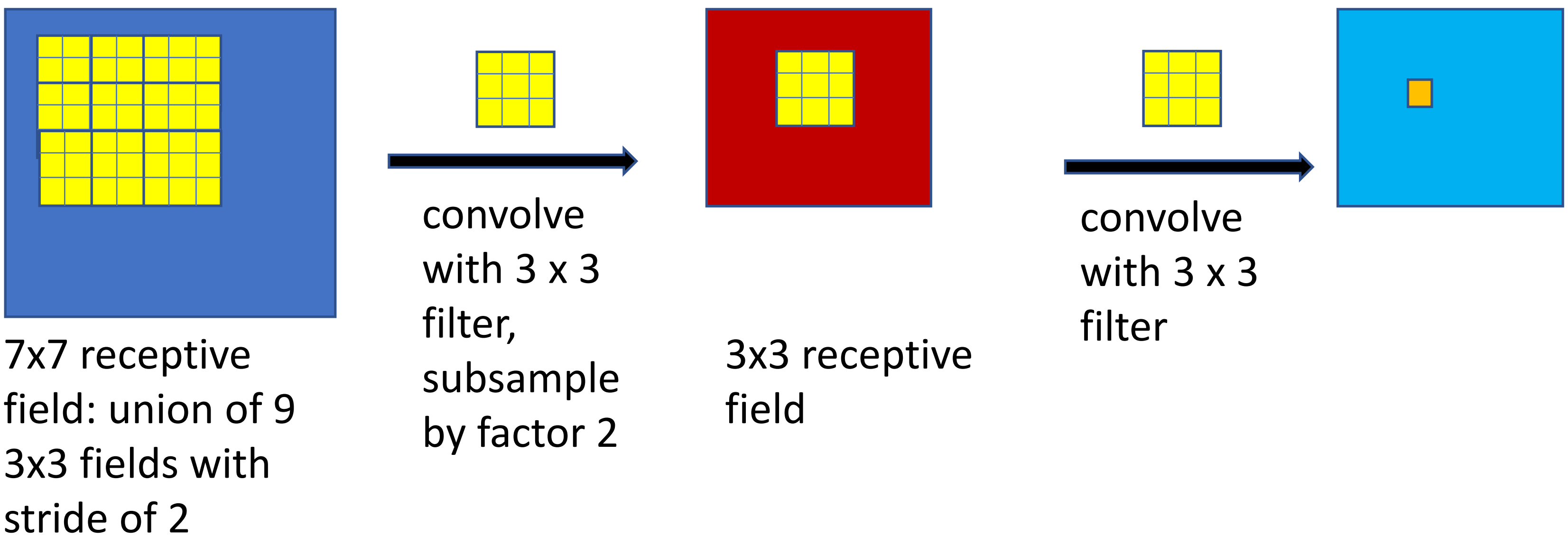
Receptive field



Receptive field

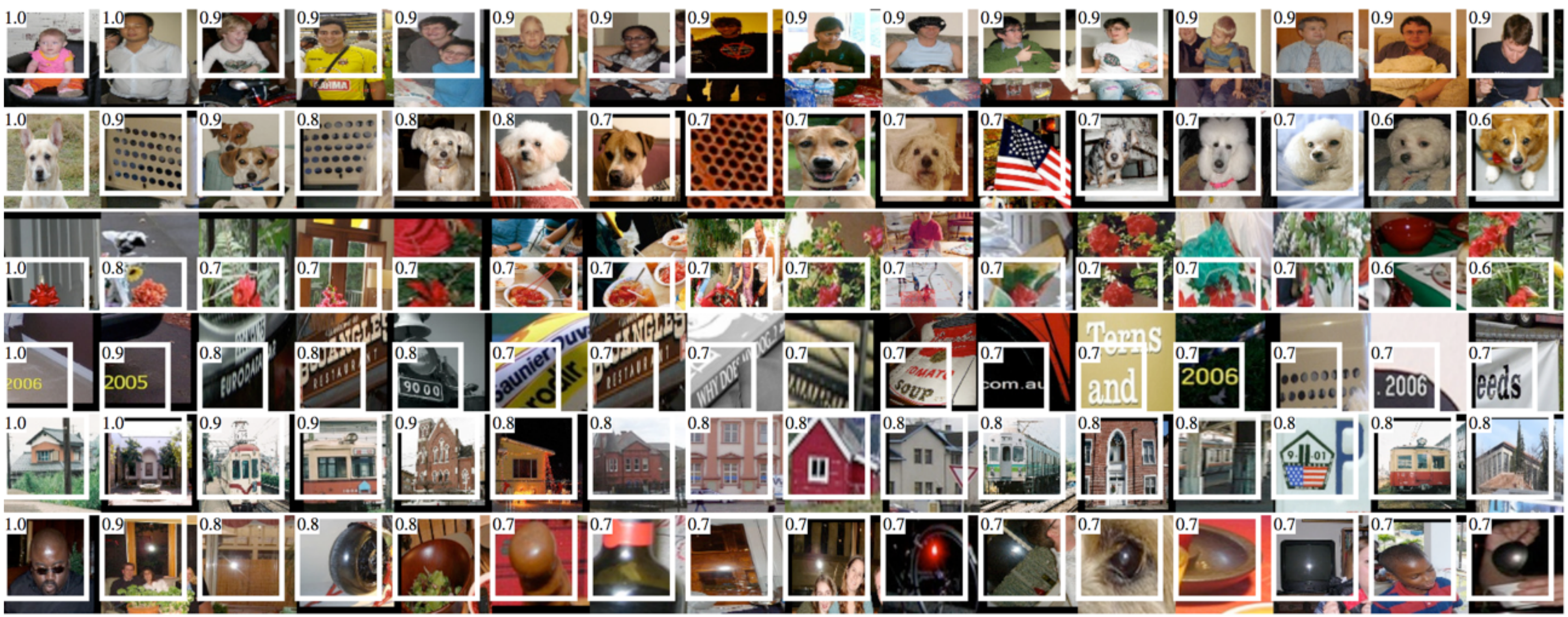


Receptive field



Visualizing convolutional networks

- Take images for which a given unit in a feature map scores high
- Identify the receptive field for each.



Rich feature hierarchies for accurate object detection and semantic segmentation. R. Girshick, J. Donahue, T. Darrell, J. Malik. In *CVPR,* 2014.

Visualizing convolutional networks II

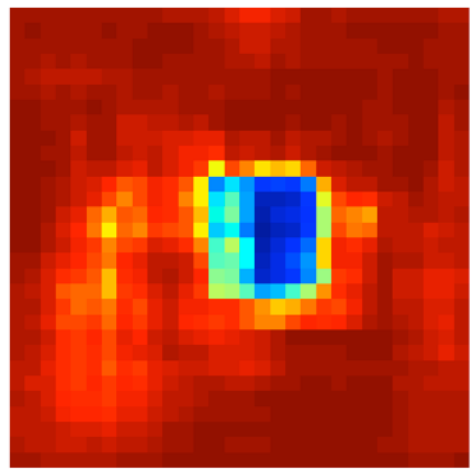
- Block regions of the image and classify

Visualizing and Understanding Convolutional Networks. M. Zeiler and R. Fergus. In *ECCV 2014.*

Visualizing convolutional networks II

- Image pixels important for classification = pixels when blocked cause misclassification

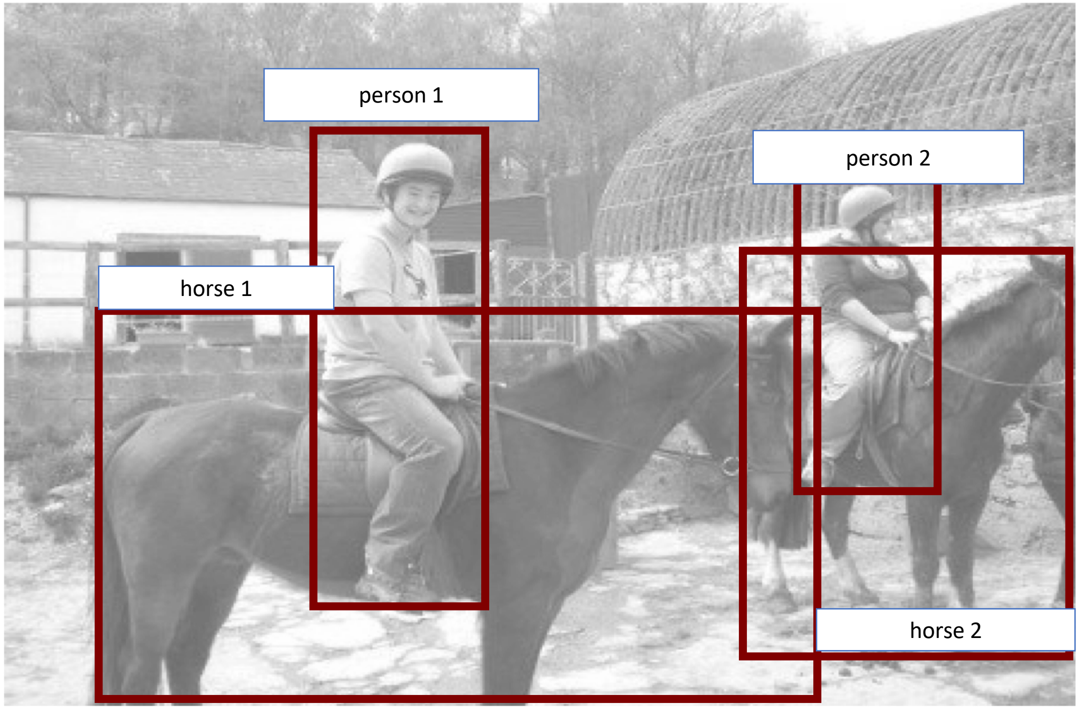
(d) Classifier, probability of correct class



Visualizing and Understanding Convolutional Networks. M. Zeiler and R. Fergus. In *ECCV 2014.*

Object detection

The Task



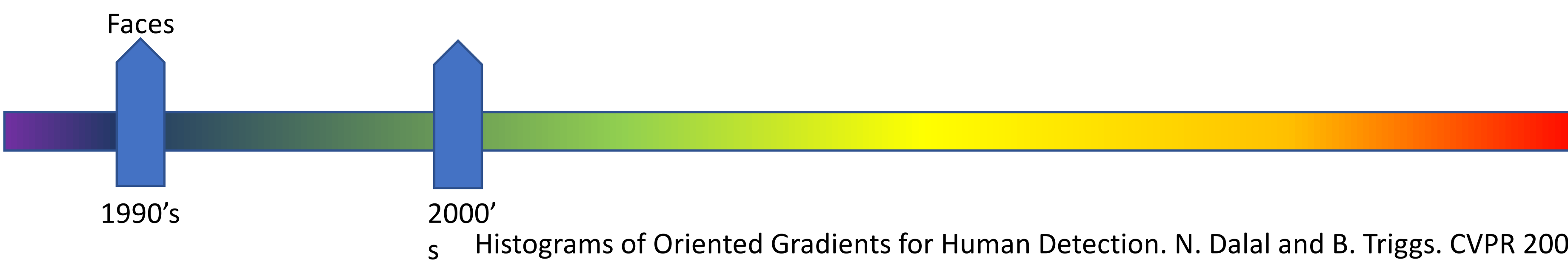
Datasets

- Face detection
- One category: face
- Frontal faces
- Fairly rigid, unoccluded

Human Face Detection in Visual Scenes. H. Rowley, S. Baluja, T. Kanade. 1995.

Pedestrians

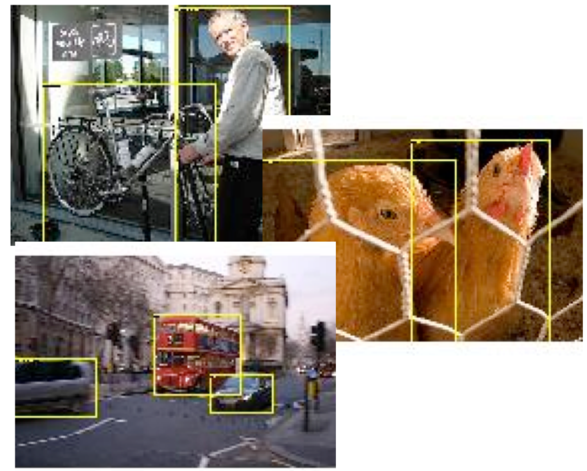
- One category: pedestrians
- Slight pose variations and small distortions
- Partial occlusions

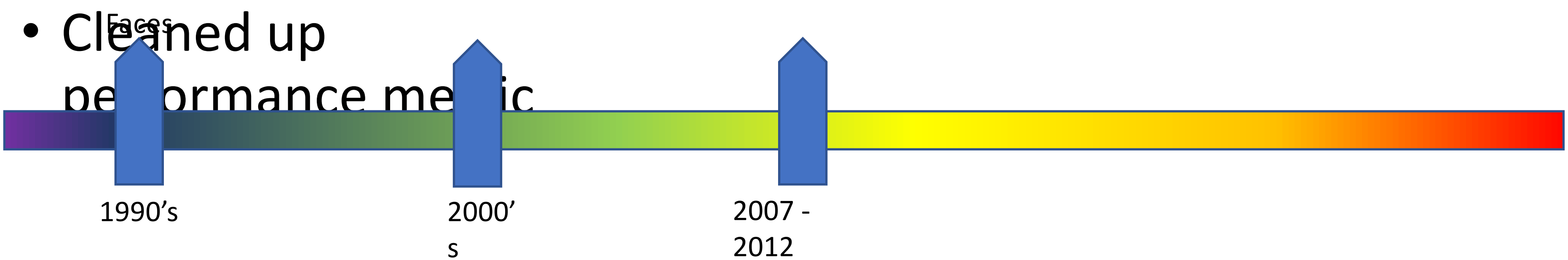


Histograms of Oriented Gradients for Human Detection. N. Dalal and B. Triggs. CVPR 2005

PASCAL VOC

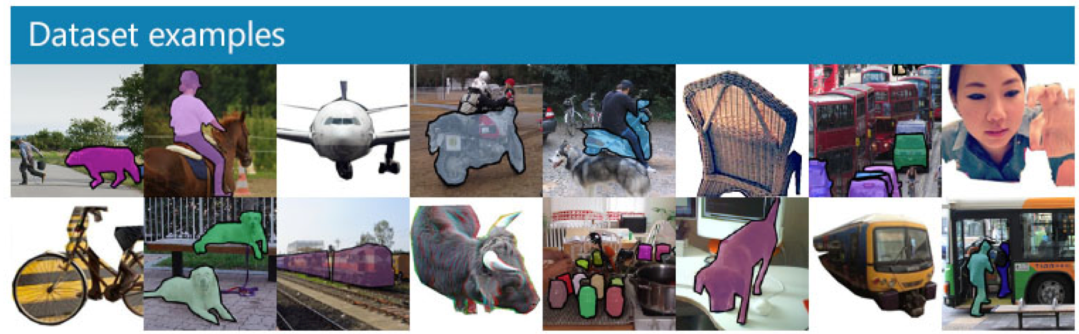
- 20 categories
- 10K images
- Large pose variations, heavy occlusions
- Generic scenes
- Cleaned up performance metric

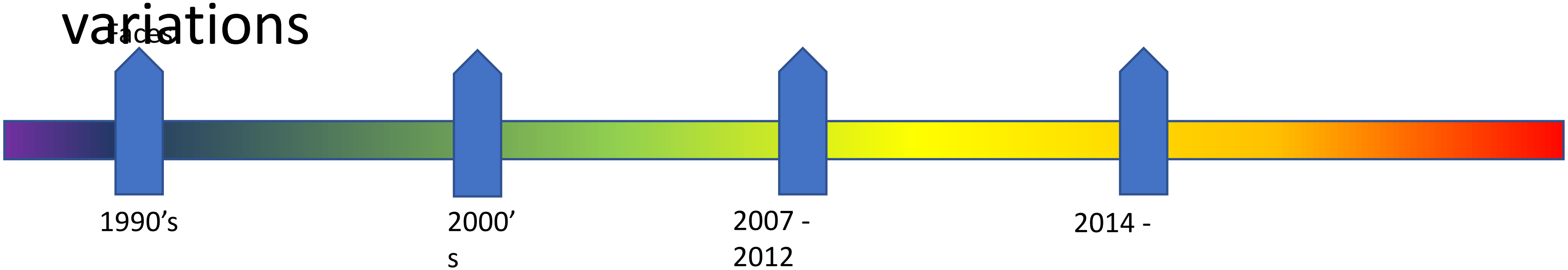




Coco

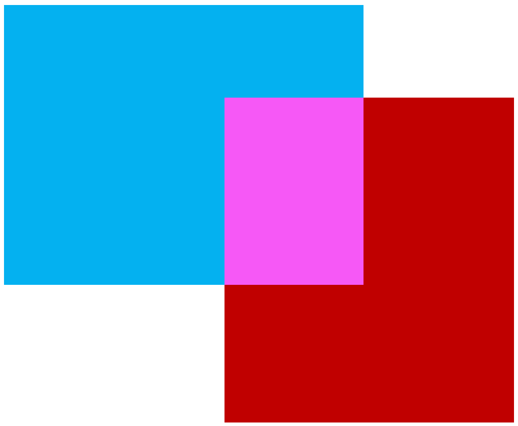
- 80 diverse categories
- 100K images
- Heavy occlusions, many objects per image, large scale variations





Evaluation metric

Matching detections to ground truth



$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

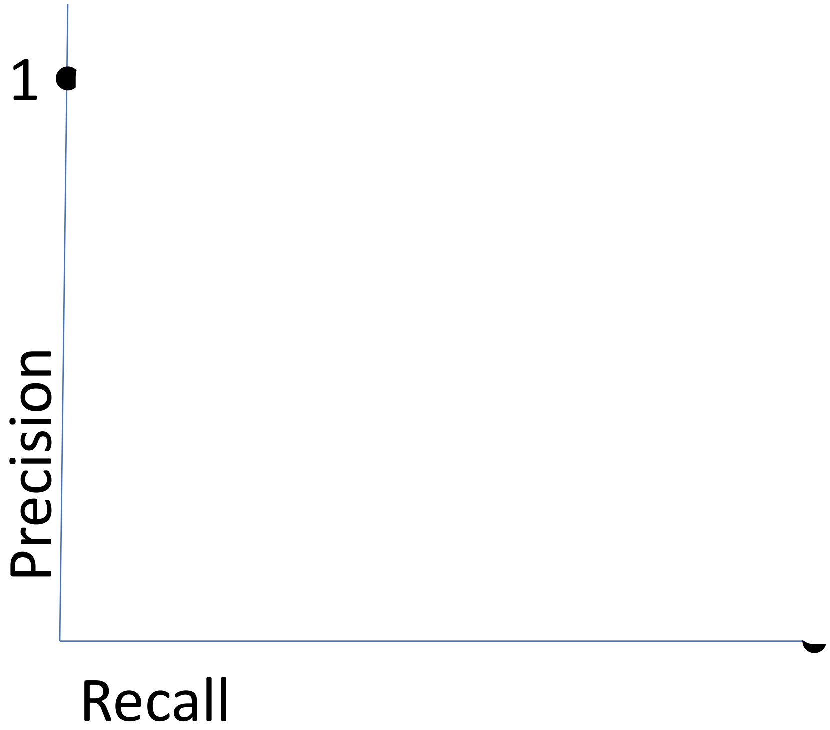
Matching detections to ground truth

- Match detection to most similar ground truth
  - highest IoU
- If IoU > 50%, mark as correct
- If multiple detections map to same ground truth, mark only one as correct
- **Precision** = #correct detections / total detections
- **Recall** = #ground truth with matched detections / total ground truth

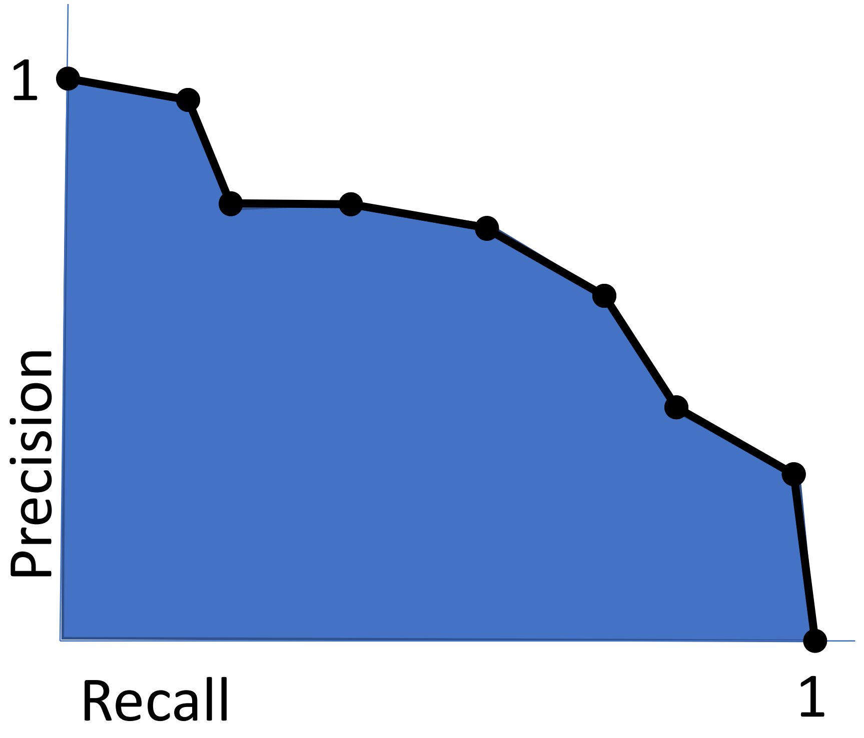
Tradeoff between precision and recall

- ML usually gives scores or probabilities, so threshold
- Too low threshold → too many detections → low precision, high recall
- Too high threshold → too few detections → high precision, low recall
- Right tradeoff depends on application
  - Detecting cancer cells in tissue: need high recall
  - Detecting edible mushrooms in forest: need high precision

Average precision



Average precision



Average average precision

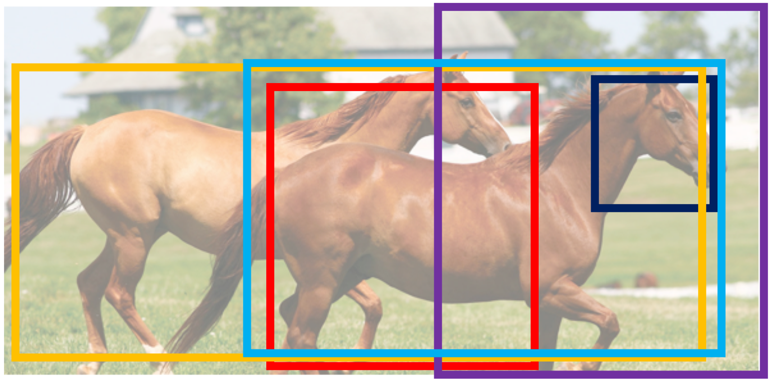
- AP marks detections with overlap > 50% as correct
- But may need better localization
- *Average* AP across multiple overlap thresholds
- Confusingly, still called average precision
- Introduced in COCO

Mean and category-wise AP

- Every category evaluated independently
- Typically report mean AP averaged over all categories
- Confusingly called “mean Average Precision”, or “mAP”

Why is detection hard(er)?

- Precise localization



Why is detection hard(er)?

- Much larger impact of pose

Why is detection hard(er)?

- Occlusion makes localization difficult

Why is detection hard(er)?

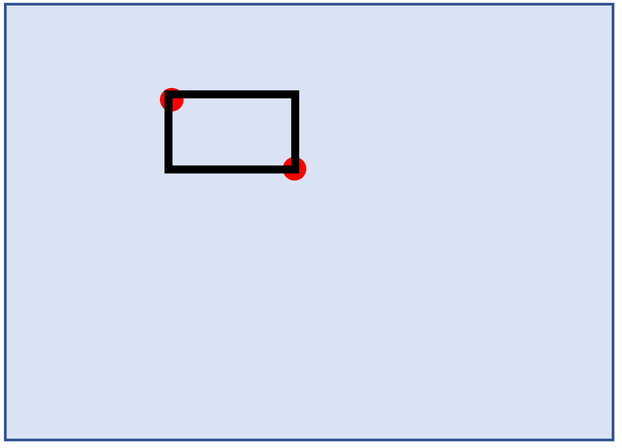
- Counting

Why is detection hard(er)?

- Small objects

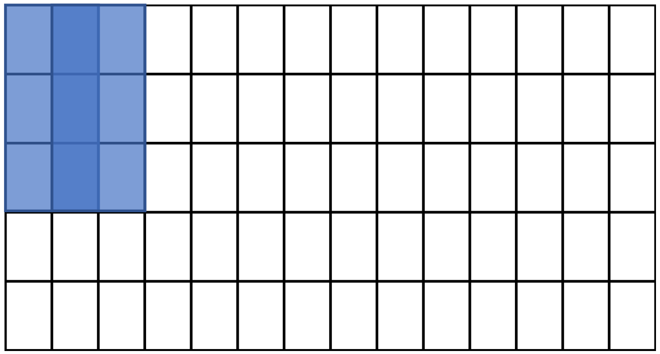
Detection as classification

- Run through every possible box and classify
- How many boxes?
  - Every pair of pixels = 1 box
  - $\binom{N}{2} = O(N^2)$
  - For 300 x 500 image, N
  - 2.25×10^{10} boxes!



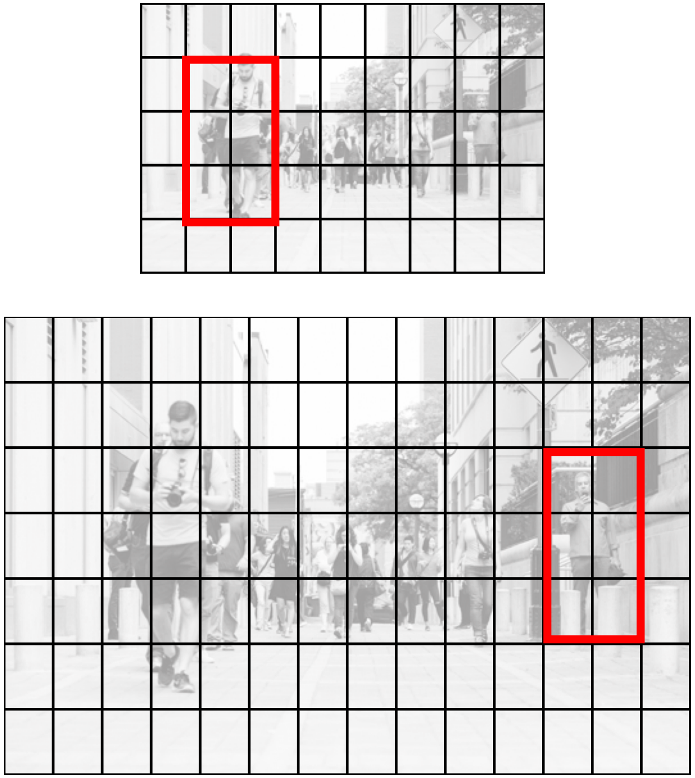
Idea 1: scanning window

- Fix size
  - Can take a few different sizes
- Fixed stride
- Convolution with a filter
  - Classic: compute HOG features over entire image



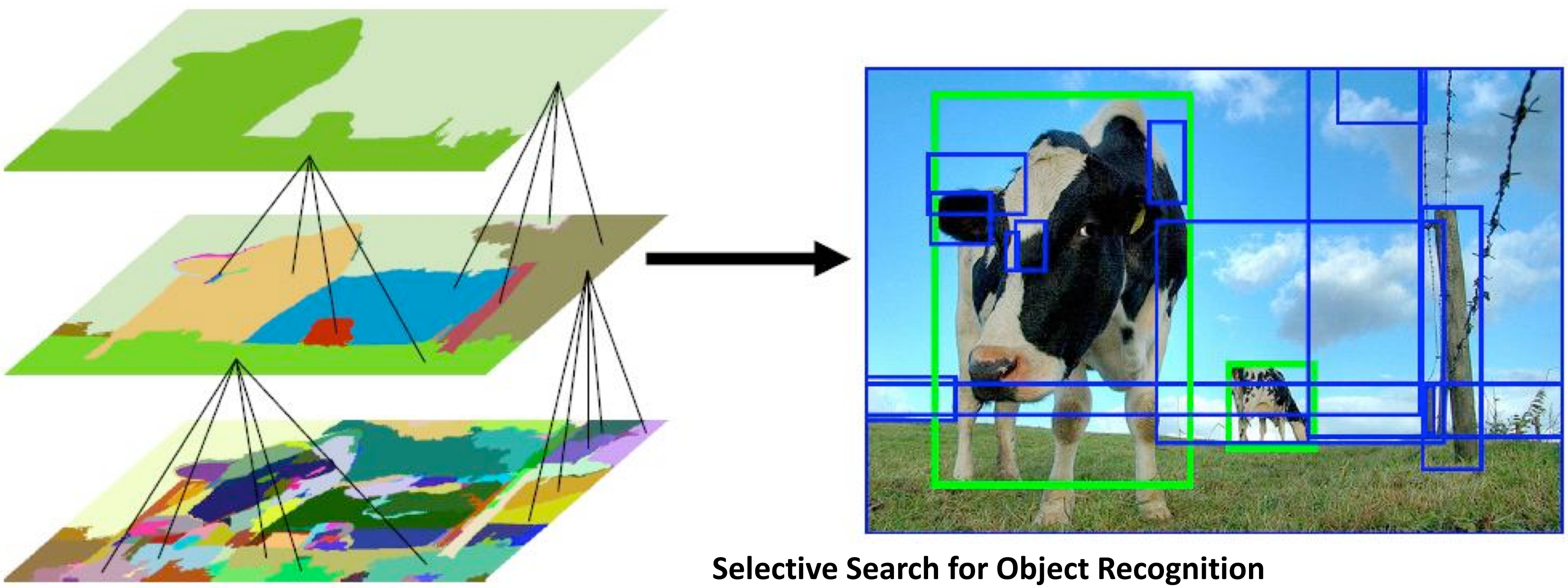
Dealing with scale

Dealing with scale



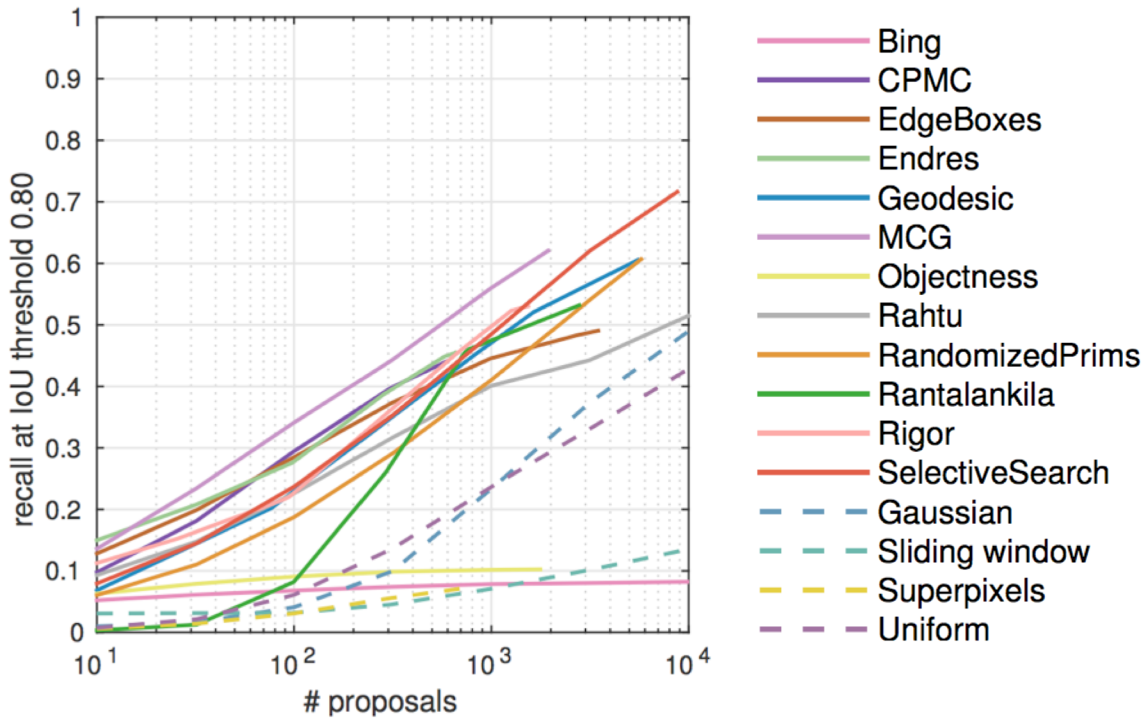
Idea 2: Object proposals

- Use segmentation to produce ~5K candidates



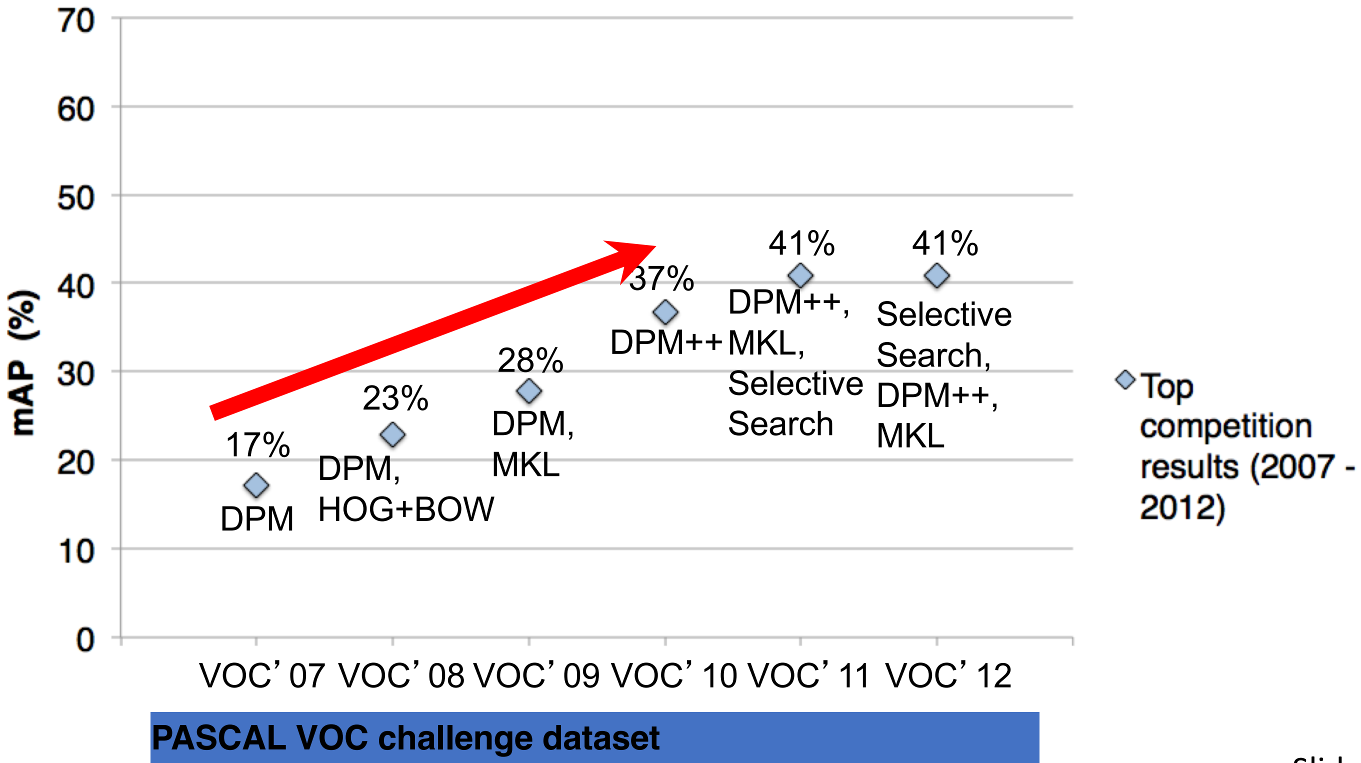
Selective Search for Object Recognition
J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders
In International Journal of Computer Vision 2013.

Idea 2: Object proposals



What makes for effective detection proposals? J. Hosang, R. Benenson, P. Dollar, B. Schiele. In TPAMI

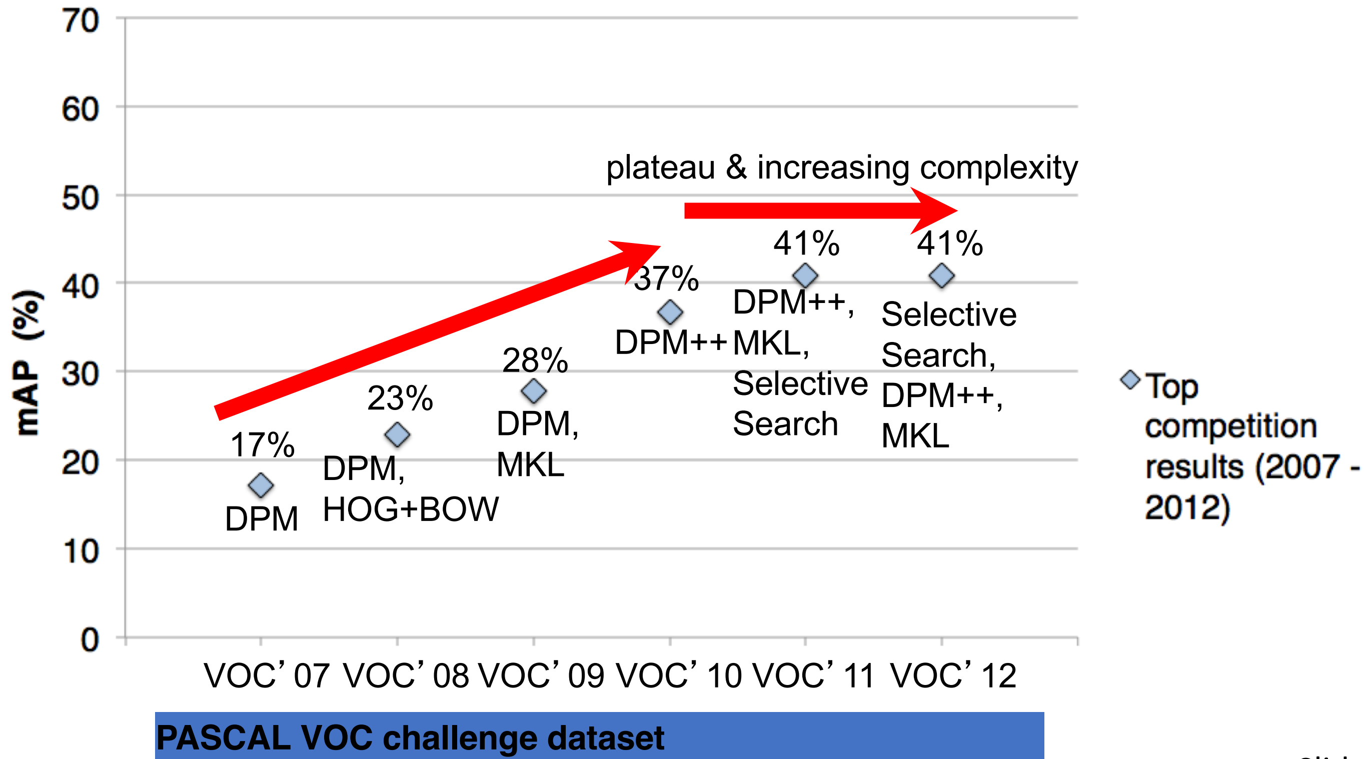
A rapid rise in performance



[Source: http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc20{07,08,09,10,11,12}/results/index.html]

Slide credit : Ross Girshick

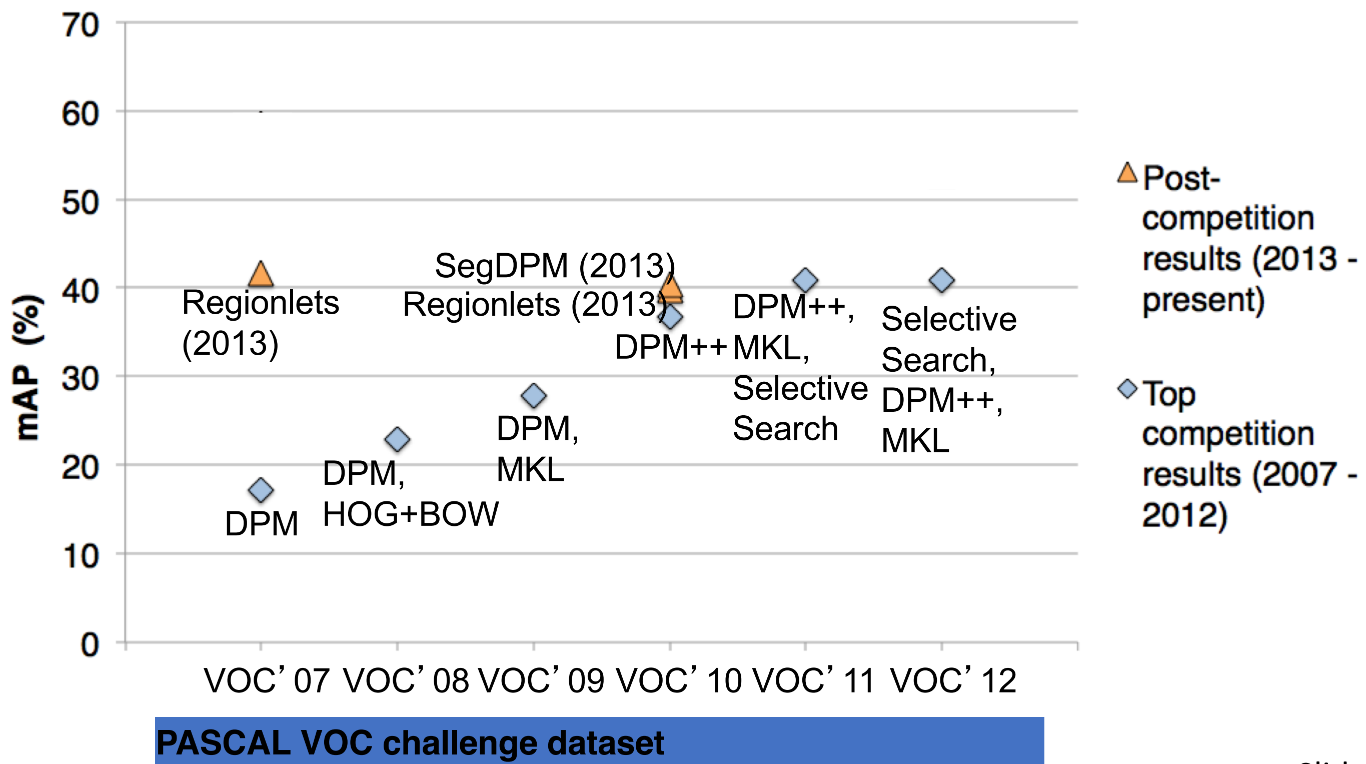
Complexity and the plateau



[Source: http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc20{07,08,09,10,11,12}/results/index.html]

Slide credit : Ross Girshick

SIFT, HOG, LBP, ...

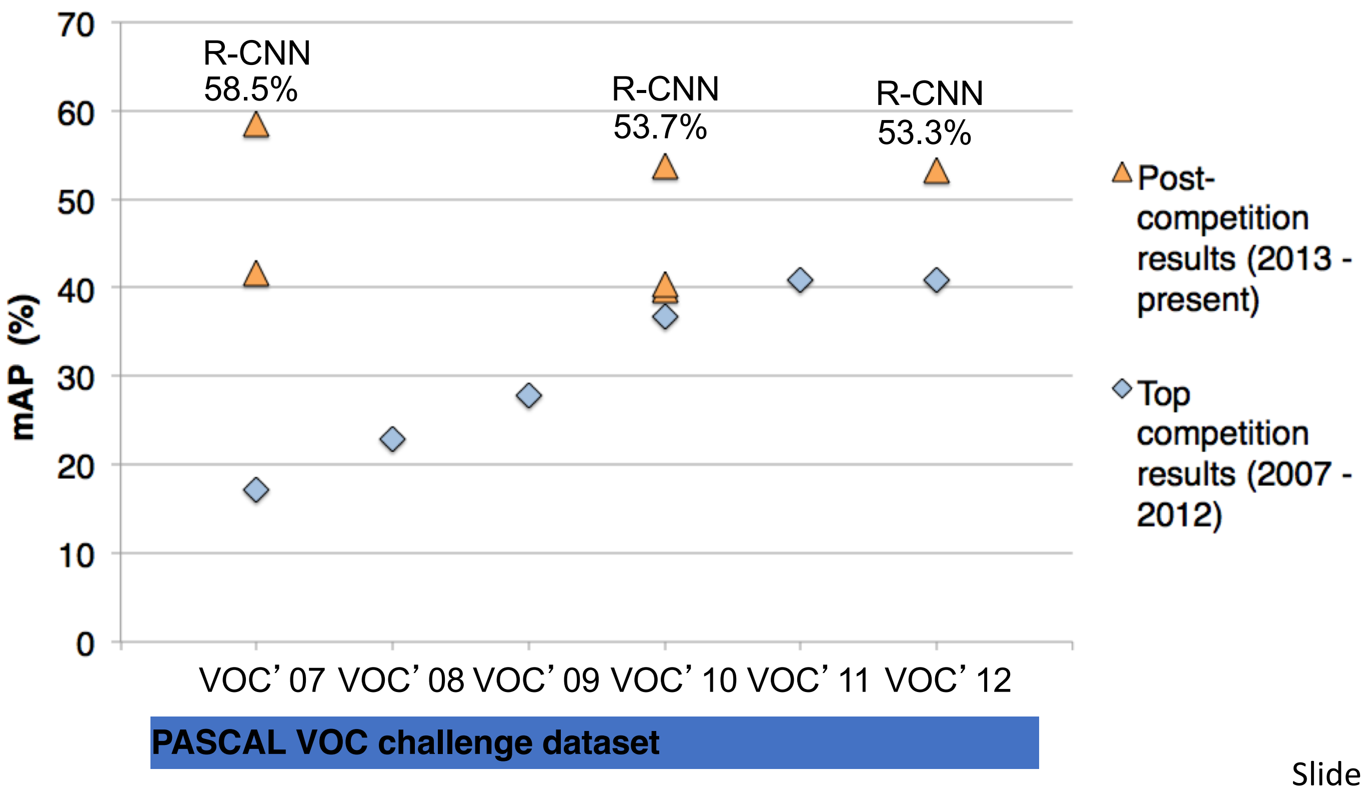


[Regionlets. Wang et al. ICCV' 13]

[SegDPM. Fidler et al. CVPR' 13]

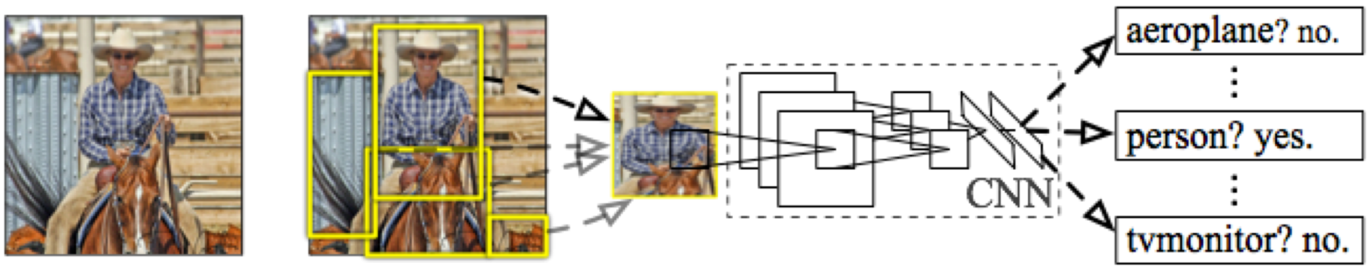
Slide credit : Ross Girshick

R-CNN: Regions with CNN features



Slide credit : Ross Girshick

R-CNN: Regions with CNN features



Input image

Extract region proposals (~2k / image)

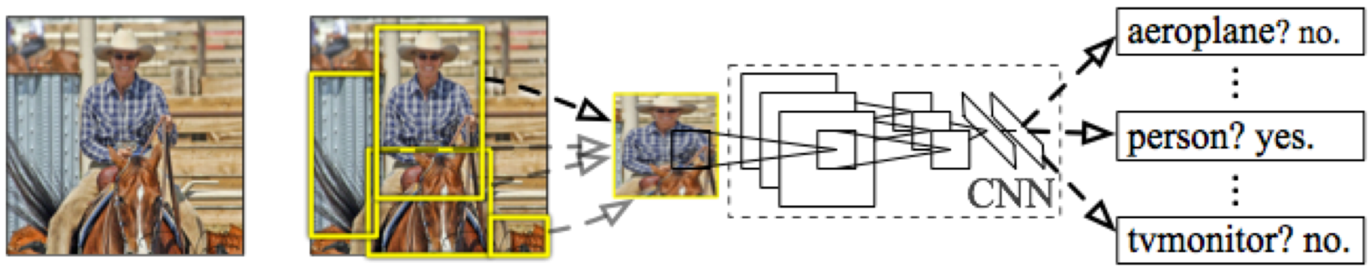
Compute CNN features

Classify regions (linear SVM)

Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation
R. Girshick, J. Donahue, T. Darrell, J. Malik
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014

Slide credit : Ross Girshick

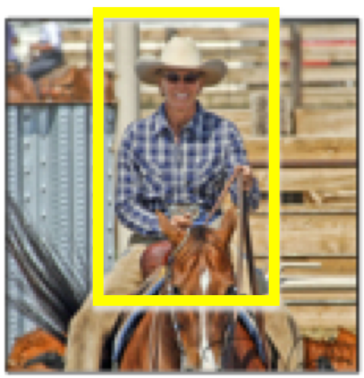
R-CNN at test time: Step 2



Input image

Extract region proposals (~2k / image)

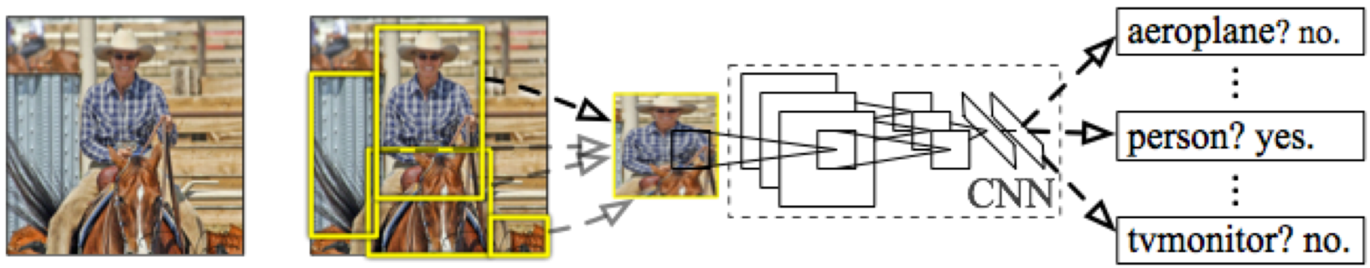
Compute CNN features



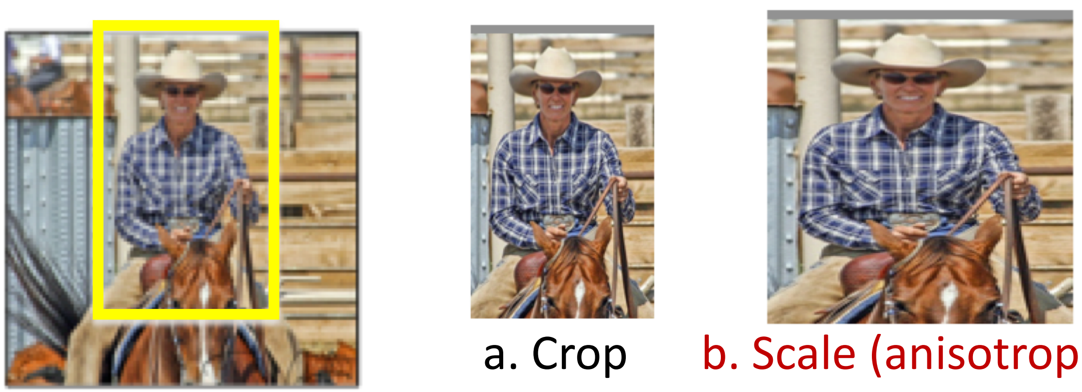
a. Crop

Slide credit : Ross Girshick

R-CNN at test time: Step 2



Input image | Extract region proposals (~2k / image) → Compute CNN features

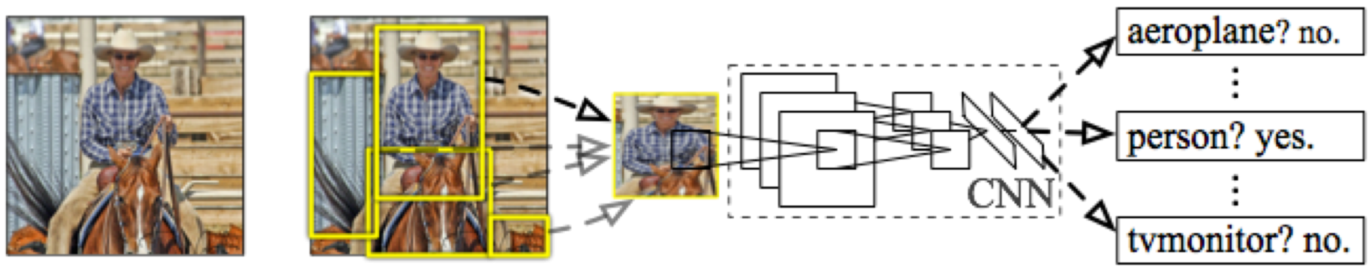


227 x 227

a. Crop | b. Scale (anisotropic)

Slide credit : Ross Girshick

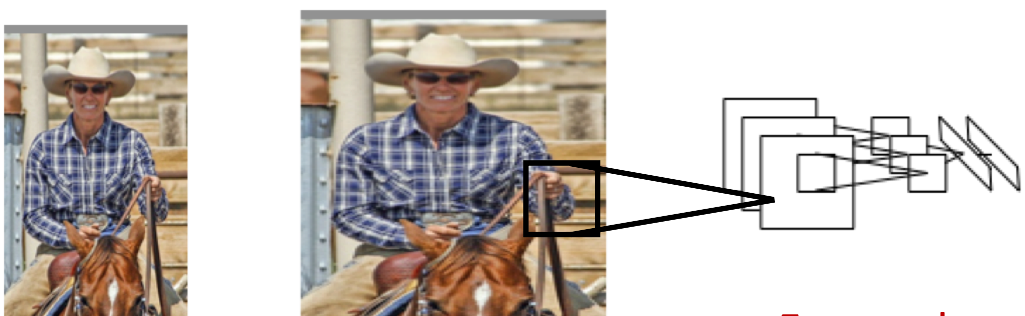
R-CNN at test time: Step 2



Input image

Extract region proposals (~2k / image)

Compute CNN features



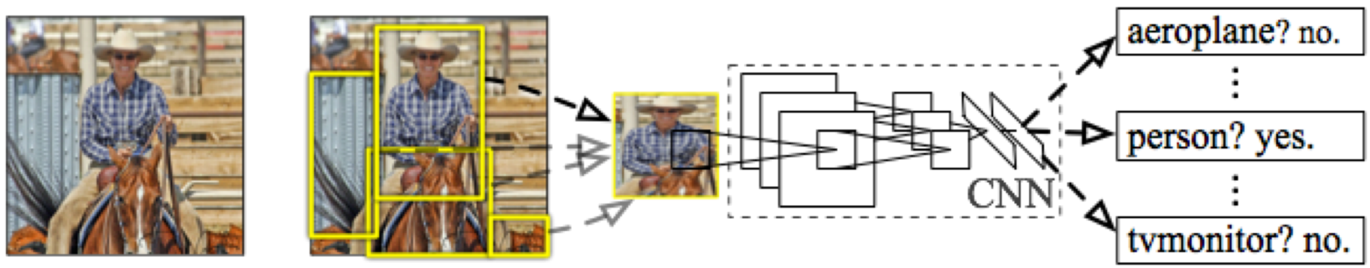
1. Crop

b. Scale (anisotropic)

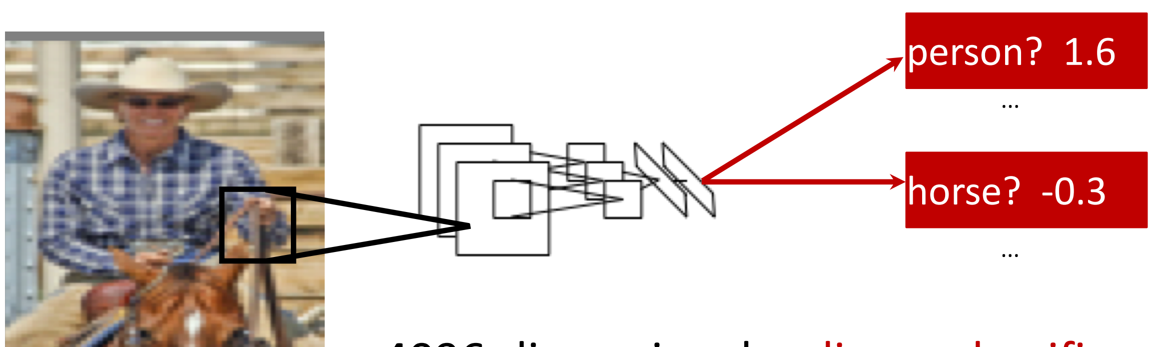
c. Forward propagate
Output: “fc_7” features

Slide credit : Ross Girshick

R-CNN at test time: Step 3



Input image | Extract region proposals (~2k / image) | Compute CNN features → Classify regions



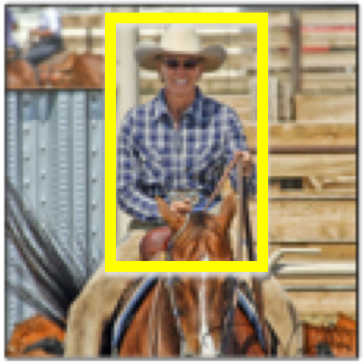
Warped proposal

4096-dimensional fc_7 feature vector

linear classifiers (SVM or softmax)

Slide credit : Ross Girshick

Step 4: Object proposal refinement



Linear regression

on CNN features

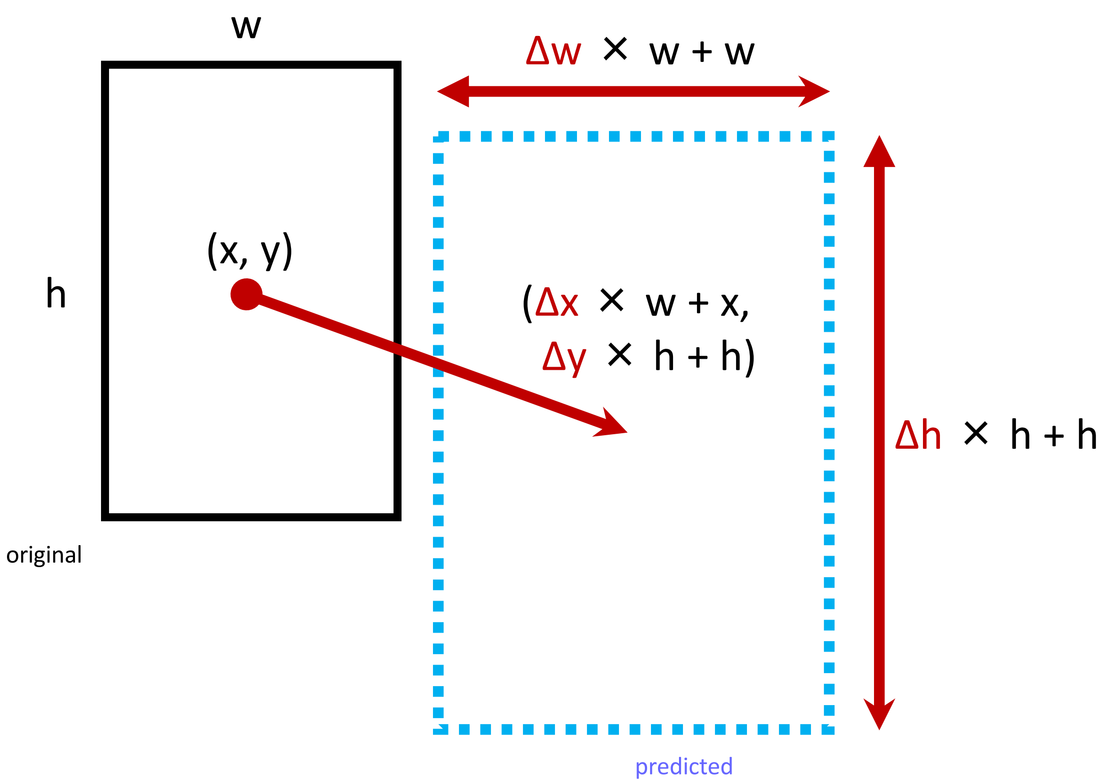
Original
proposal

Predicted
object bounding box

Bounding-box regression

Slide credit : Ross Girshick

Bounding-box regression



Slide credit : Ross Girshick

R-CNN results on PASCAL

	VOC 2007	VOC 2010
DPM v5 (Girshick et al. 2011)	33.7%	29.6%
UVA sel. search (Uijlings et al. 2013)		35.1%
Regionlets (Wang et al. 2013)	41.7%	39.7%
SegDPM (Fidler et al. 2013)		40.4%

Reference systems

metric: mean average precision (higher is better)

Slide credit : Ross Girshick

R-CNN results on PASCAL

	VOC 2007	VOC 2010
DPM v5 (Girshick et al. 2011)	33.7%	29.6%
UVA sel. search (Uijlings et al. 2013)		35.1%
Regionlets (Wang et al. 2013)	41.7%	39.7%
SegDPM (Fidler et al. 2013)		40.4%
R-CNN	54.2%	50.2%
R-CNN + bbox regression	58.5%	53.7%

Slide credit : Ross Girshick

Training R-CNN

- Train convolutional network on ImageNet classification
- *Finetune* on detection
  - Classification problem!
  - Proposals with IoU > 50% are positives
  - Sample fixed proportion of positives in each batch because of imbalance

Other details - Non-max suppression

How do we deal with multiple detections on the same object?

Other details - Non-max suppression

- Go down the list of detections starting from highest scoring
- Eliminate any detection that overlaps highly with a higher scoring detection
- Separate, heuristic step