Convolutional Neural Networks II

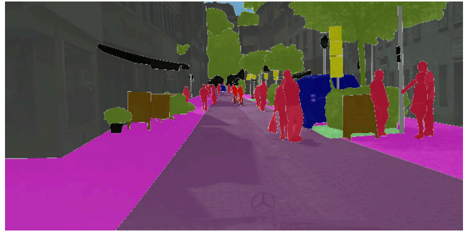
Lecture 11

Automatic Image Analysis

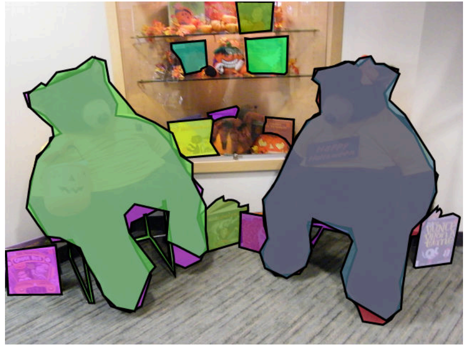
July 19, 2021

Technische Universität Berlin

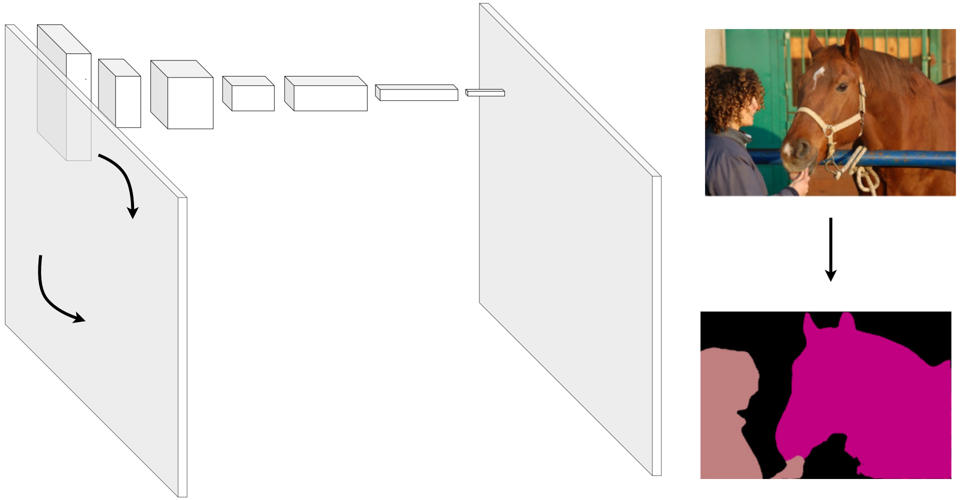
- Semantic Segmentation is the task of classifying every pixel of an image with an object class.
- Often including a background class.



- 30 classes
- 5000 annotated images with fine annotation
- 20000 annotated images with coarse annotations

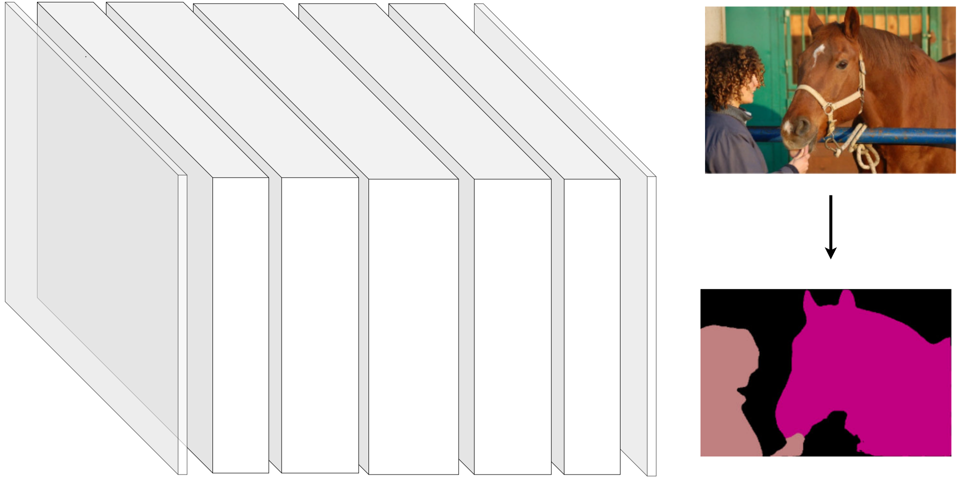


- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 330K images (>200K labeled)

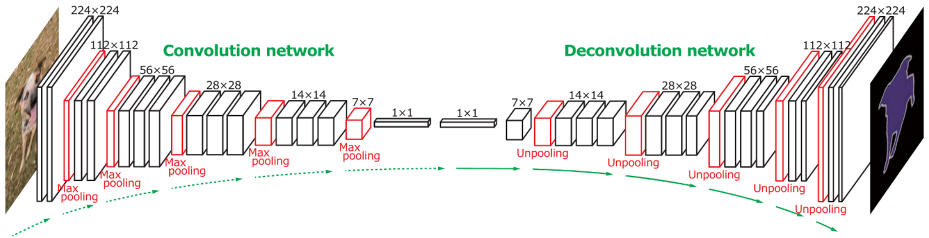


- One forward pass per pixel.

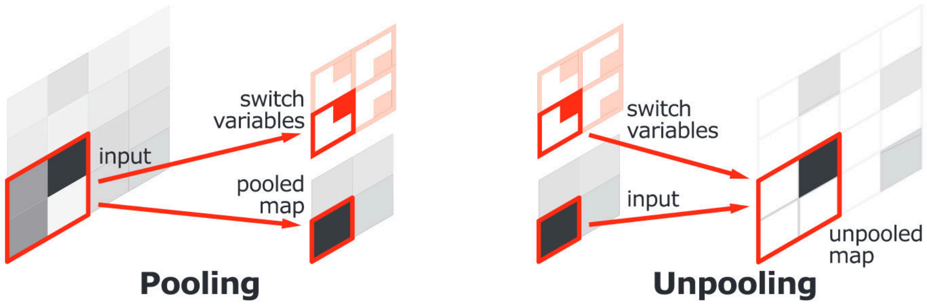
Semantic Segmentation: without downsampling?



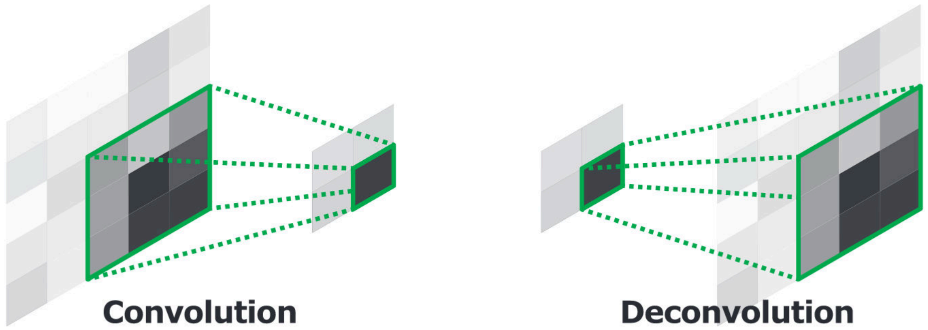
- Huge memory demands.



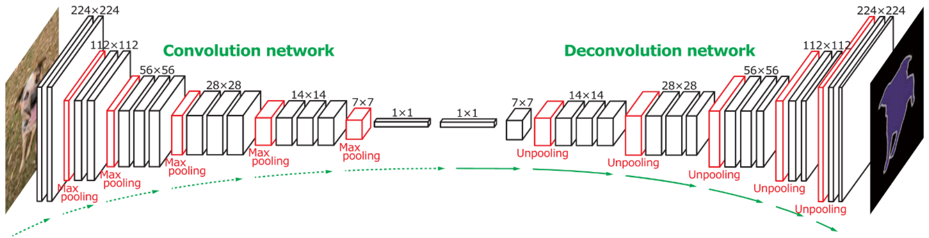
- Image from Learning Deconvolution Network for Semantic Segmentation, Noh et al, ICCV 2015



- Other unpooling methods: nearest neighbour or bed of nails.
- Image from Learning Deconvolution Network for Semantic Segmentation, Noh et al, ICCV 2015

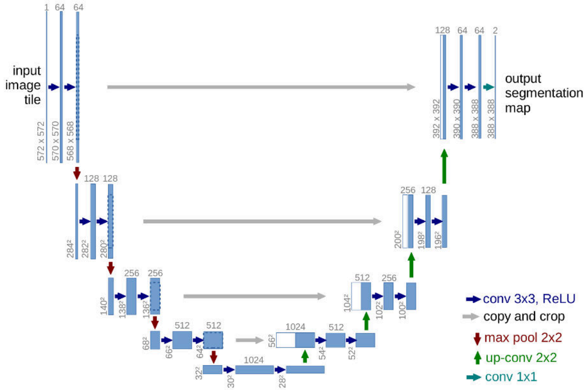


- Transpose convolution, deconvolution
- stride 2, pad 1, the other way
- Image from Learning Deconvolution Network for Semantic Segmentation, Noh et al, ICCV 2015



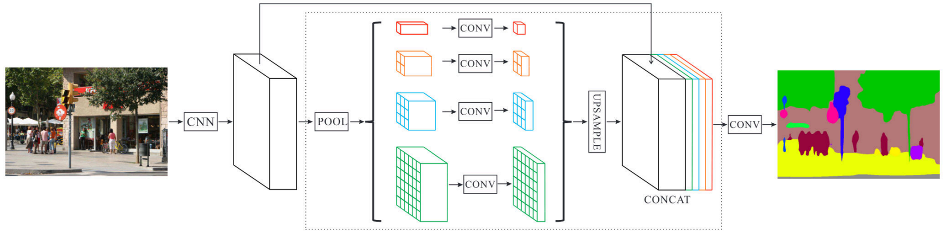
- Problem: the coarse features (encoding in the middle) is supposed to be abstract and to not contain detailed geometrical information.
- Image from Learning Deconvolution Network for Semantic Segmentation, Noh et al, ICCV 2015

UNet/Segnet



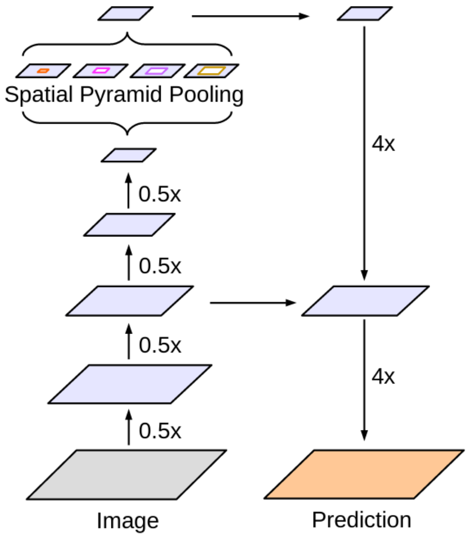
- Solution: skip connection.
- Image from U-Net: Convolutional Networks for Biomedical Image Segmentation, Ronnenberger et al, MICCAI 2015
- SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, Badrinarayanan et al, TPAMI 2017

Pyramid Pooling

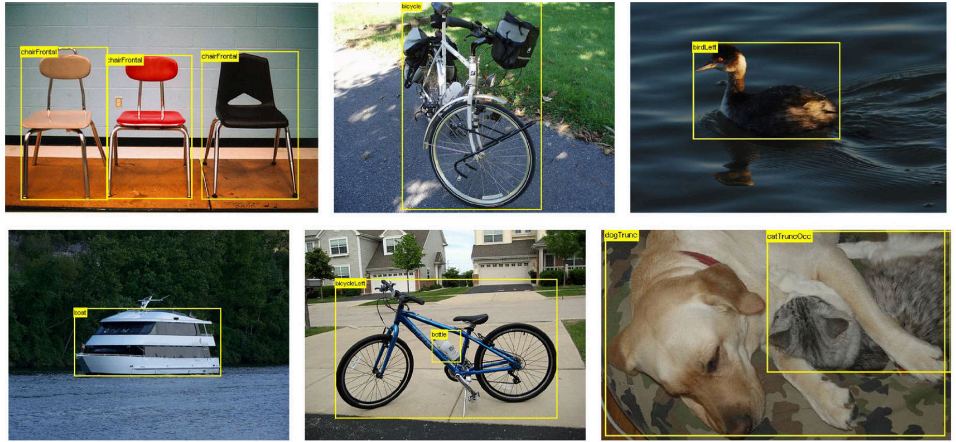


(a) Input Image (b) Feature Map (c) Pyramid Pooling Module (d) Final Prediction

- Global context prior: to allow the network to process the image on different scales improves results.
- Improves the models ability to learn spatial semantics (spatial class co-occurence and spatial coherence).
- Improves recognition of very small object and stuff classes that exceed receptive fields.
- Image from Pyramid Scene Parsing Network, Zhao et al, CVPR 2017

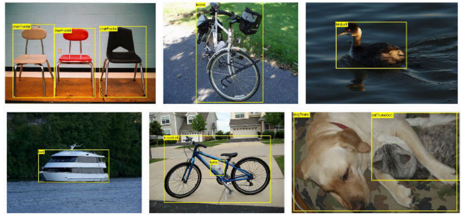


- Case study of a SOTA semantic segmentation network: uses pretrained encoder network plus spatial pyramid pooling and skip connections.
- Image from Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, Chen et al, ECCV 2018



- Image from The PASCAL Visual Object Classes Challenge: A Retrospective, Everingham et al, IJCV 2014

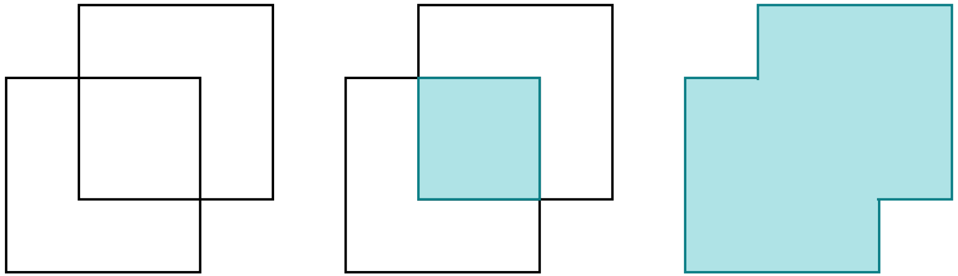
- Pascal VOC (DPM 33.6%)



- 20 classes
- 11k annotated images
- 27k annotated objects

Detection is correct if

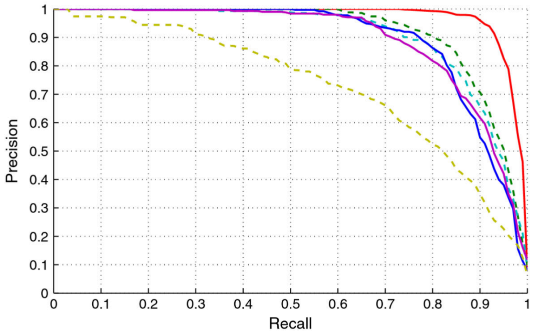
$$intersection/union > threshold$$



- Default threshold was 0.5 for a long time but is now often higher.

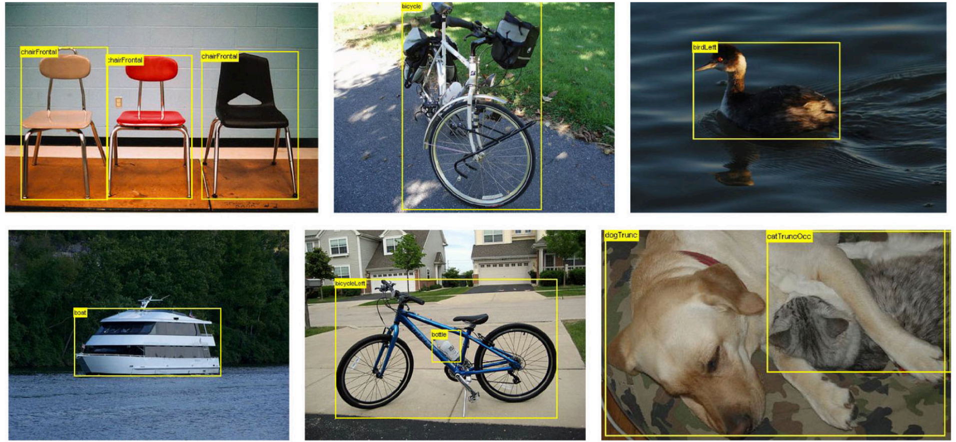
$$precision = \#(correct\ \ detections) / \#(all\ \ detections)$$
$$recall = \#(correct\ \ detections) / \#(all\ \ objects)$$

Average Precision: area under PR curve for specific class
mean Average Precision: AP averaged over all classes

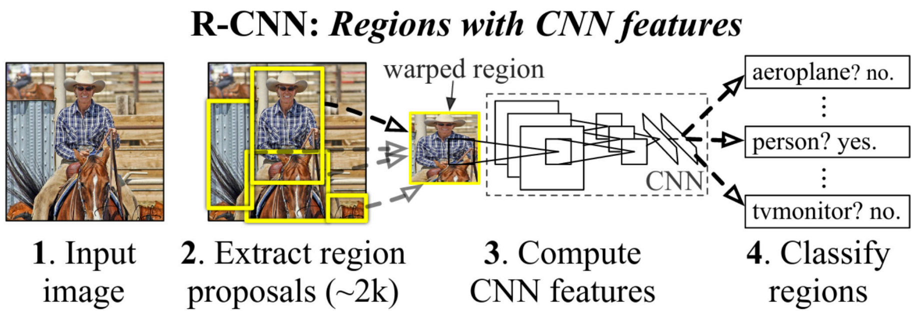


- Image from The PASCAL Visual Object Classes Challenge: A Retrospective, Everingham et al, IJCV 2014

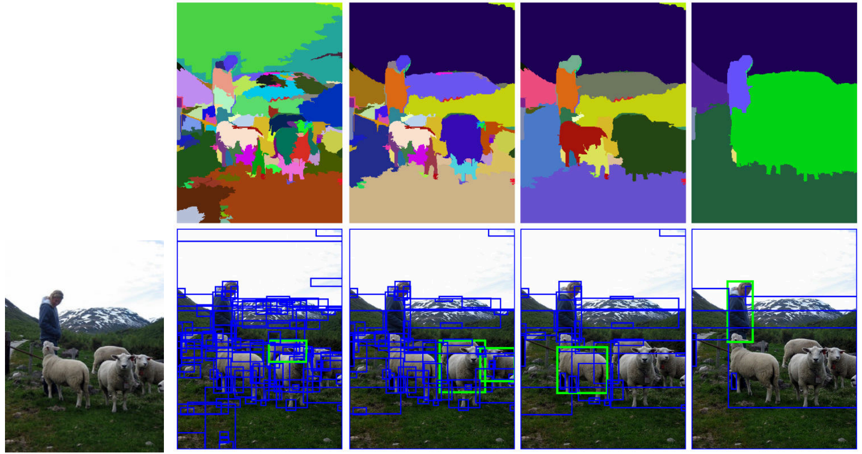
Object Detection: output dimensionality?



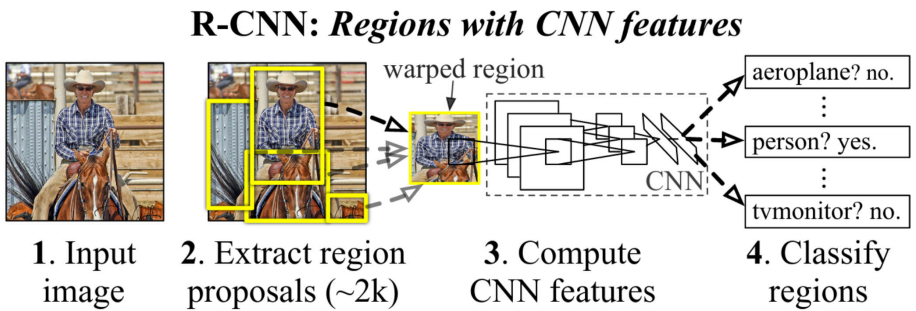
- How would the head of this network look like?
- Image from The PASCAL Visual Object Classes Challenge: A Retrospective, Everingham et al, IJCV 2014



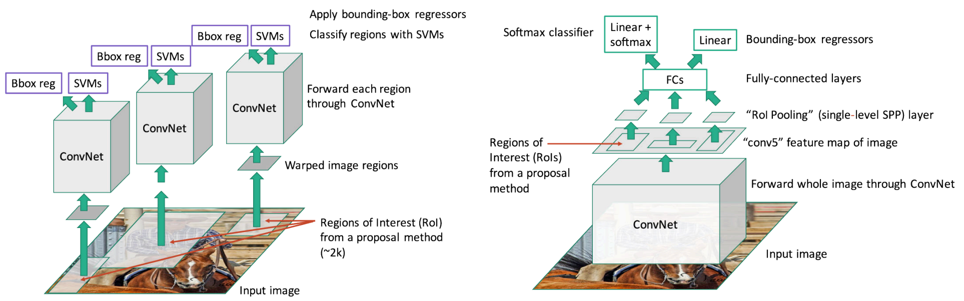
- Same author as DPM.
- Sliding window as in DPM. But NN much slower as SVM, therefore they used region proposals (2k).
- Image from Rich feature hierarchies for accurate object detection and semantic segmentation, Girshick et al, CVPR 2014



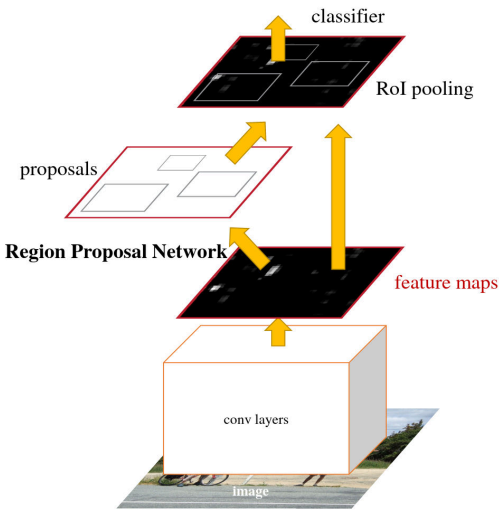
- Image from Selective Search for Object Recognition, Uijlings et al, IJCV 2013



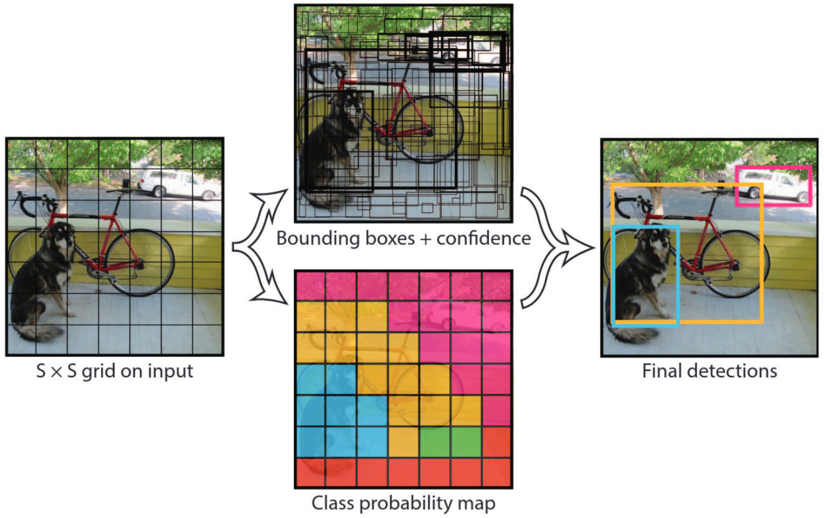
- Network also needs to predict bounding box parameters (size and offset from patch center).
- Non maximum suppression in prediction space.
- Often some high level reasoning (coherence in object relations).
- mAP for Pascal VOC improved to 53% with AlexNet as ConvNet and 62% with VGG (from 33% DPM)
- Image from Rich feature hierarchies for accurate object detection and semantic segmentation, Girshick et al, CVPR 2014



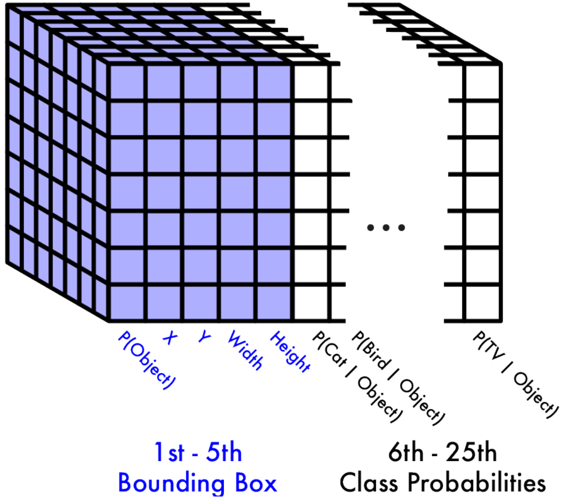
- Moves the cropping of proposed regions to the feature map, saving the many forward passes through the convolutional block.
- Image from Talk at ICCV 2015 by Ross Girshick https://dl.dropboxusercontent.com/s/vlyrkgd8nz8gy5l/fast-rcnn.pdf?dl=0



- Region proposal is now the expensive step in Fast-RNN.
- Faster-RCNN does bounding box regression with a neural network based on the same image features the classifier uses, removing the region proposal step completely.
- Solution: Do region proposal in feature map.



- Image from You Only Look Once:Unified, Real-Time Object Detection, Redmon et al, CVPR 2016

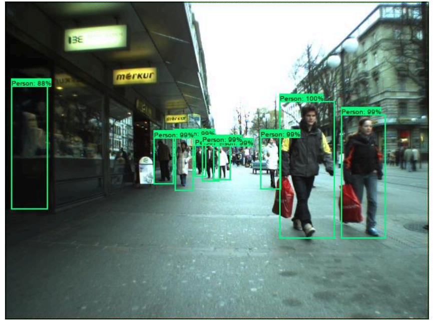


- Newer versions of YOLO have multiple detections per cell for different object sizes.
- Image from Ancient Secrets of Computer Vision Lecture 18, Joseph Redmon

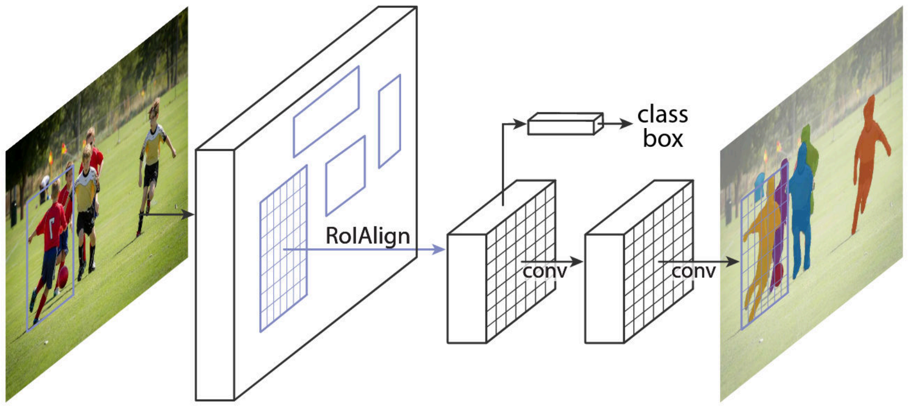
$$
\begin{aligned}
\mathcal{L} &= \alpha_1 \mathcal{L}_{localization} + \alpha_2 \mathcal{L}_{object\ confidence} + \alpha_3 \mathcal{L}_{classification} \\
\mathcal{L}_{localization} &: root\ mean\ squared\ error \\
\mathcal{L}_{object\ confidence} &: binary\ cross\ entropy \\
\mathcal{L}_{classification} &: multi-class\ cross\ entropy
\end{aligned}
$$

- weighted loss, binary and multi-class cross entropy, MSE
- What would happen without conditional probability?

Why not both? Instance Segmentation

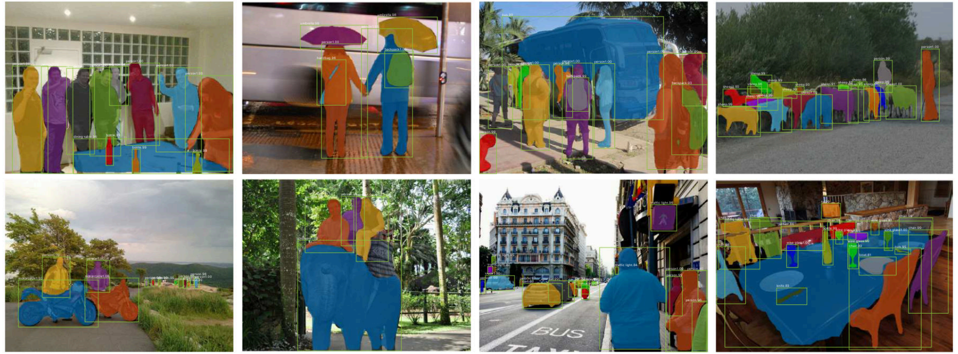


- Pixel level classification with instance boundaries.



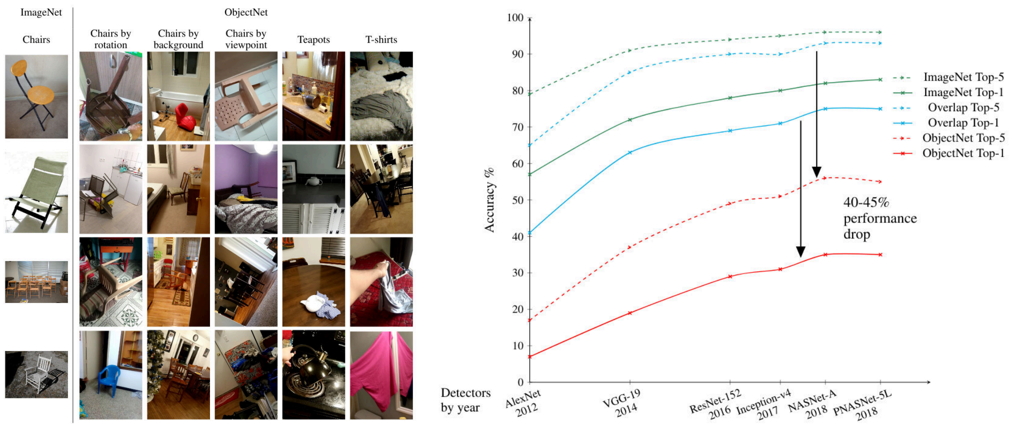
- Faster R-CNN with segmentation network.
- Image from Mask R-CNN, He et al, ICCV 2017

Mask R-CNN

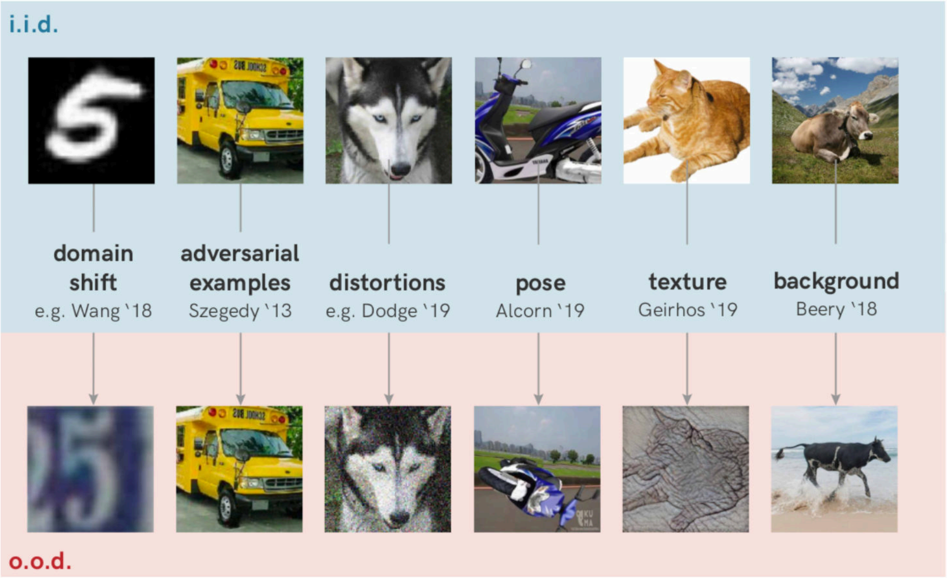


- Results for Mask R-CNN.
- Image from Mask R-CNN, He et al, ICCV 2017

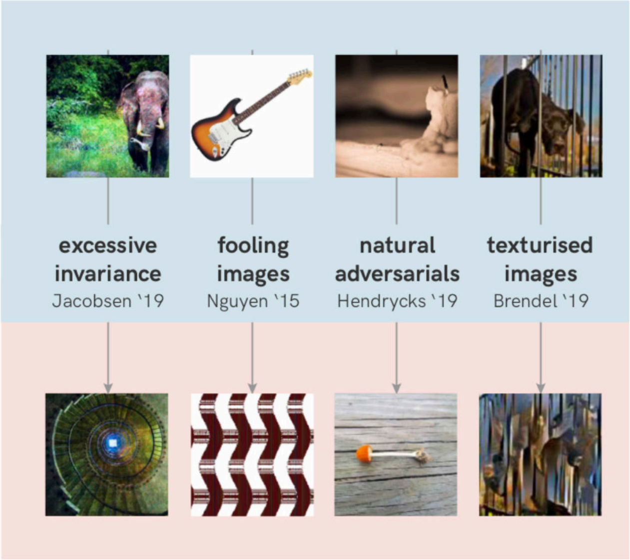
- Mask R-CNN can also learn skeletons.
- Image from Mask R-CNN, He et al, ICCV 2017



- Neural Networks trained and evaluated on ImageNet do not generalize to o.o.d. data.
- Image from ObjectNet: A large-scale bias-controlled dataset forpushing the limits of object recognition models, Barbu et al, NeurIPS 2019

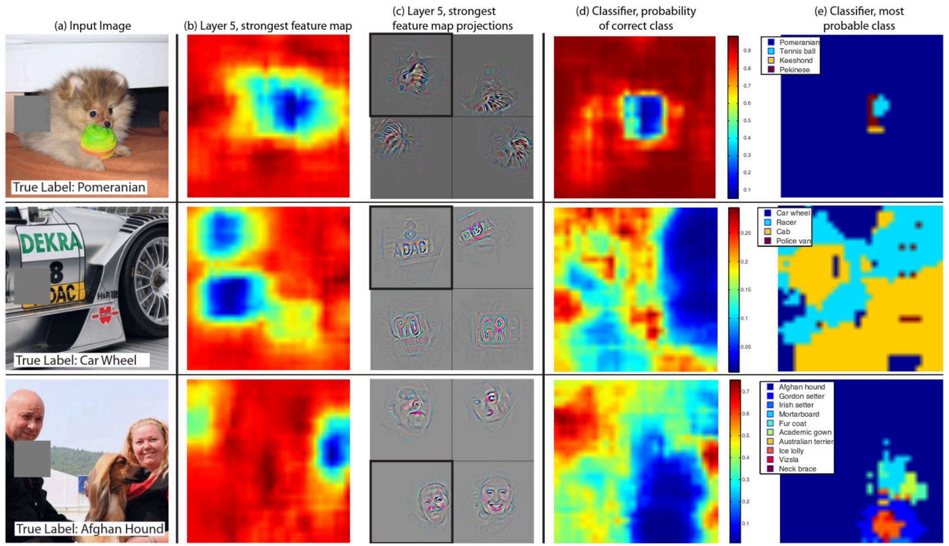


- They learn shortcuts if we let them.
- Image from Shortcut Learning in Deep Neural Networks, Geirhos et al, Nature Machine Intelligence 2020



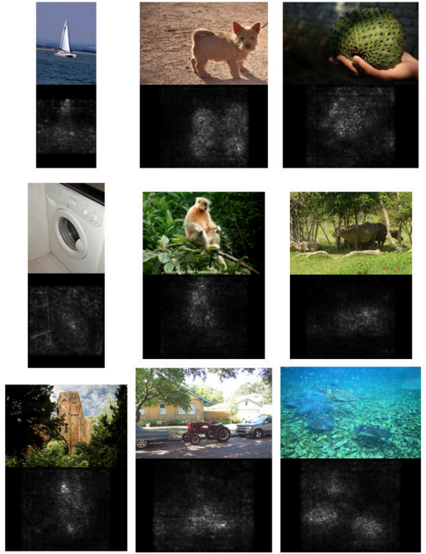
- Image from Shortcut Learning in Deep Neural Networks, Geirhos et al, Nature Machine Intelligence 2020

Investigate decisions: partial occlusion



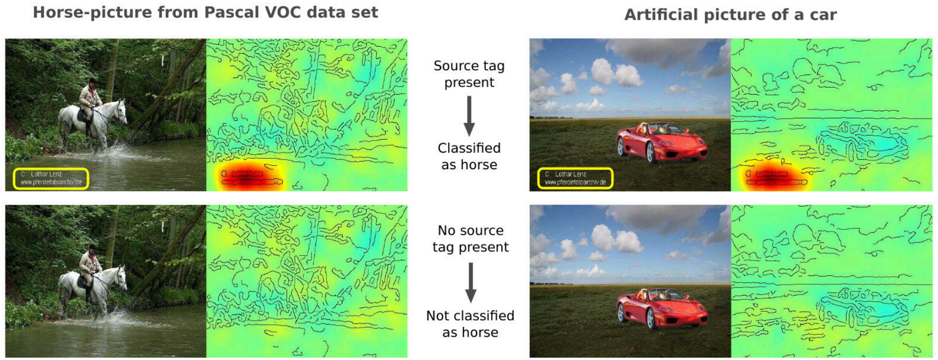
- An easy way for an visual sanity check is occluding parts of the image while watching the accuracy.
- Image from Visualizing and Understanding Convolutional Networks, Zeiler & Fergus, ECCV 2014

Investigate decisions: image gradient



- Looking at the pixel gradient of the network gives some insights too.
- Image from Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, Simonyan et al, 2013

Investigate decisions: relevance propagation



- Explain the output, not the local variation.
- Image from Unmasking Clever Hans Predictors and Assessing What Machines Really Learn, Lapuschkin et al, Nature Communications 2019