香港中文大學
The Chinese University of Hong Kong

CNN Applications in Computer Vision

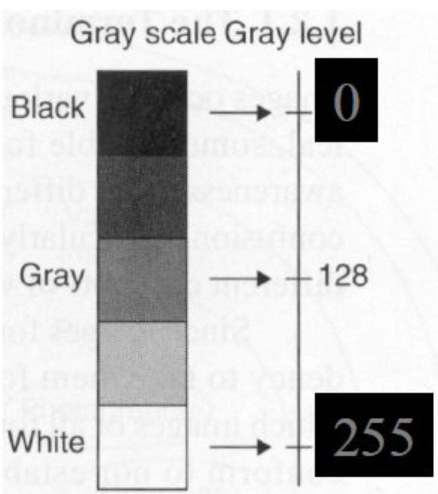
ELEG 5491 Tutorial

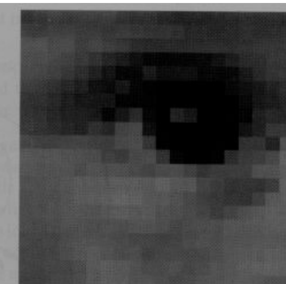
Xihui Liu

Table of Contents

Image Representation

- Grayscale image
  - Can be represented by 2D matrices
  - By default, we use 8 bits per pixel



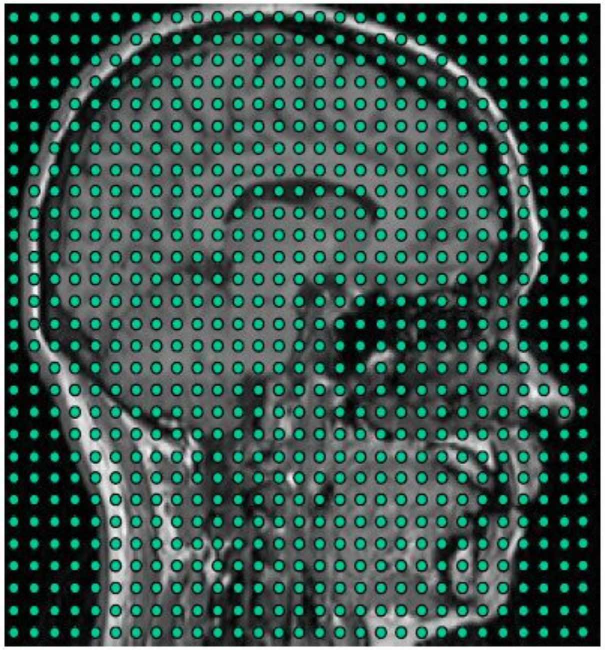


117	125	133	127	130	130	133	121	116	115	100	91	93	94	99	103	112	105	109	106
134	133	138	138	132	134	130	133	128	123	121	113	106	102	99	106	113	109	109	113
146	147	138	140	125	134	124	115	102	96	93	94	99	96	99	100	103	110	109	110
144	141	136	130	120	108	88	74	53	37	31	37	35	39	53	79	93	100	109	116
139	136	129	119	102	85	58	31	41	77	51	53	53	33	37	41	69	94	105	108
132	127	117	102	87	57	49	77	42	28	17	15	13	13	17	41	53	69	88	100
124	120	108	94	72	74	72	31	35	31	15	13	15	11	15	13	46	75	83	96
125	115	102	93	88	82	42	79	113	41	19	100	82	11	11	17	31	91	99	100
124	116	109	99	91	113	99	140	144	57	20	20	15	11	15	17	63	87	119	124
136	133	133	135	138	133	132	144	150	120	24	17	15	15	17	20	115	113	88	150
158	157	157	154	149	145	133	127	146	150	116	35	20	19	28	105	124	128	141	171
155	154	156	155	146	155	154	154	147	139	148	150	138	120	128	129	130	151	156	165
150	151	154	162	166	167	169	174	172	167	177	166	164	140	134	120	121	120	127	172
145	149	151	157	165	169	173	179	176	166	166	157	145	136	129	124	120	136	163	168
144	148	153	160	159	158	165	172	165	169	157	151	149	141	130	140	151	162	169	167
144	141	147	155	154	149	156	151	157	157	151	144	147	147	149	159	158	159	166	165
139	140	140	150	153	151	150	146	140	139	138	140	145	151	149	156	156	162	162	161
136	134	138	146	156	164	153	146	145	136	139	139	140	141	149	157	159	161	169	166
136	133	136	135	144	159	168	159	151	142	141	145	139	146	153	156	164	167	172	168
133	129	140	142	146	159	167	165	154	151	146	141	147	154	156	160	161	157	153	154

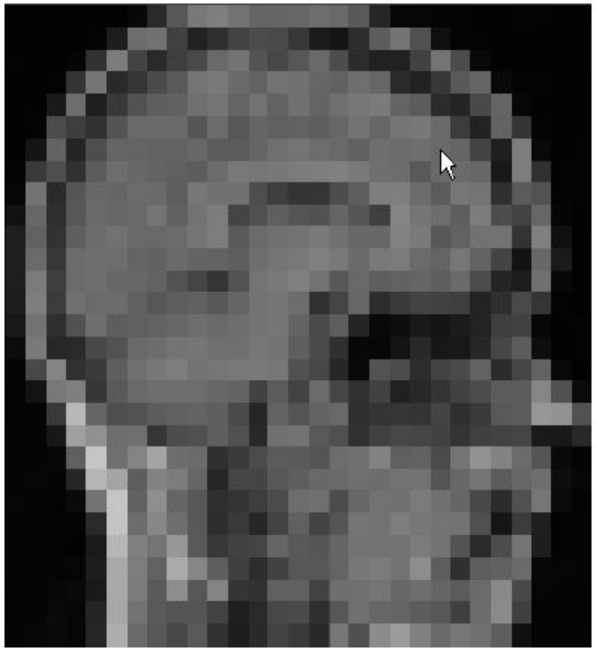
Image Representation

- Image is a 2D array of pixels (picture element) with FIXED Number of samples : N x M

N x M = 256 x 256



N x M = 30 x 30

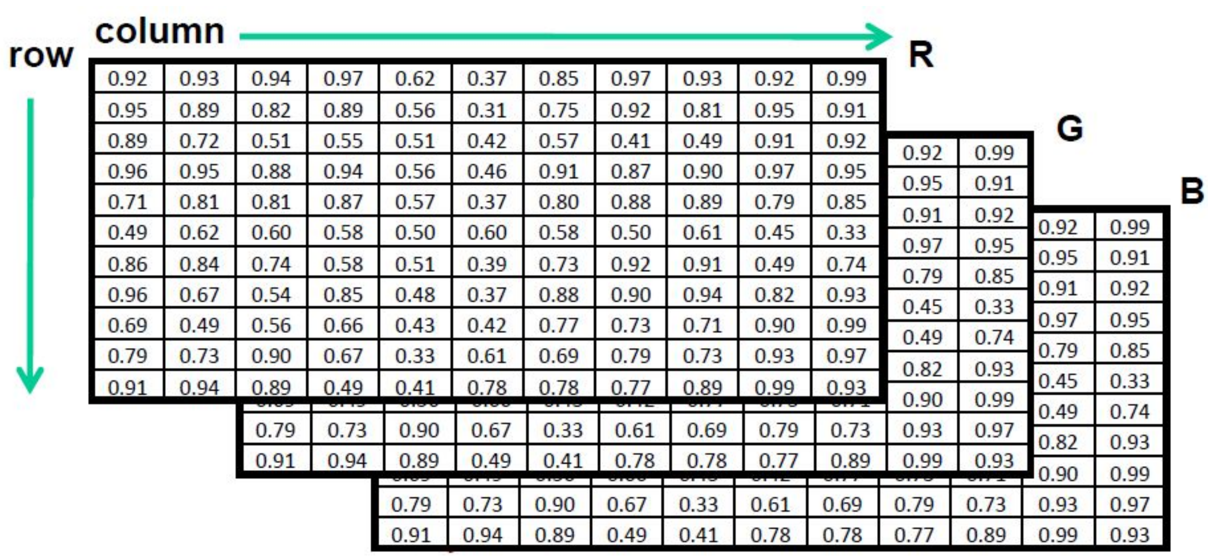


Color Image Representation

- Color image
  - Each pixel is specified by three values, (R, G, B) in the range of [0,255] (8-bit integers)

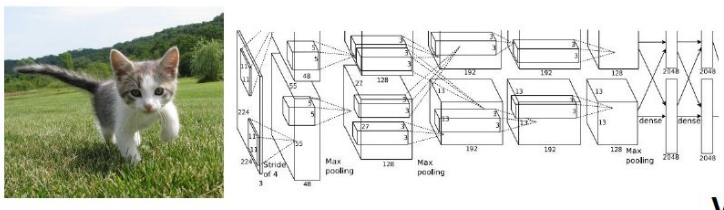
Color Image Representation

- Color image
  - Color images are stored in a 3 x M x N tensor
  - [0,255] is usually mapped to [0.0,1.0] in PyTorch (a deep learning library)



CNN Applications in Computer Vision

- Image Classification
  - Given an input image, classify it into a predefined class



Vector:
4096

Fully-Connected:
4096 to 1000

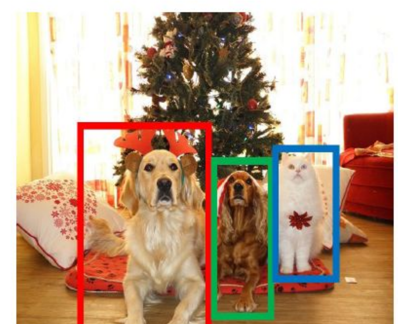
Class Scores
Cat: 0.9
Dog: 0.05
Car: 0.01
...

- Other computer vision tasks

Semantic Segmentation

GRASS, CAT, TREE, SKY

Object Detection



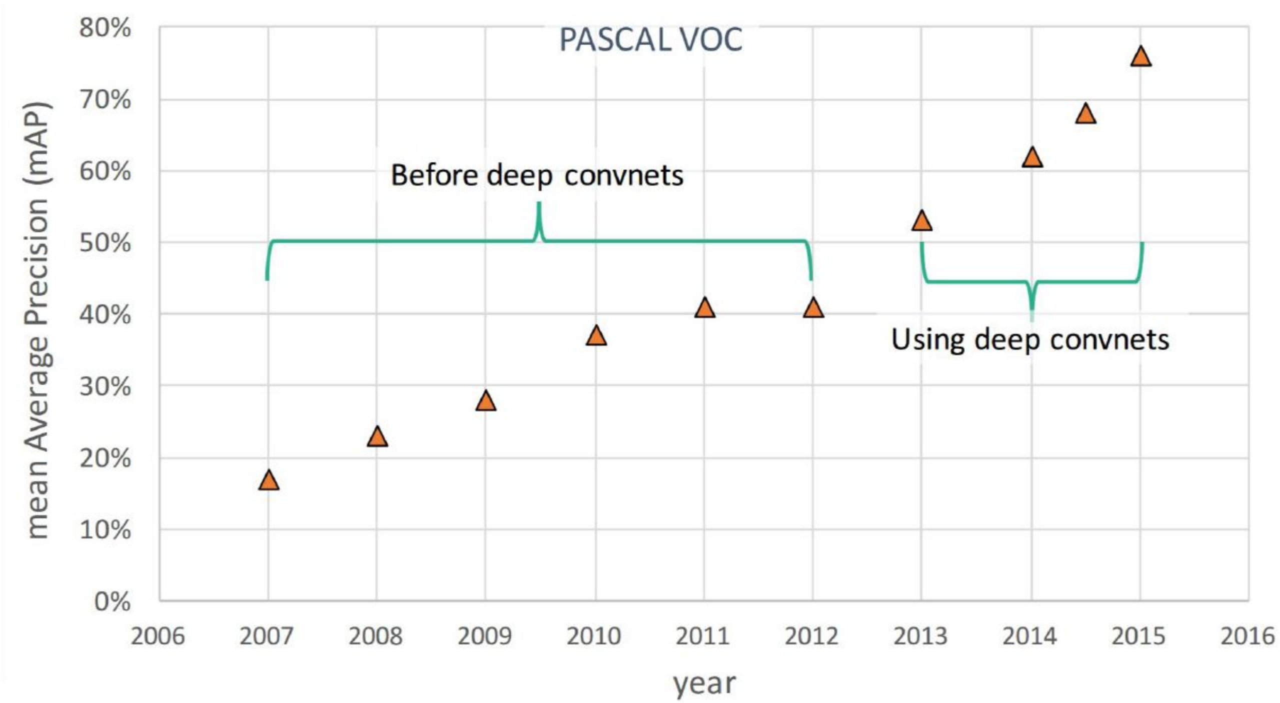
DOG, DOG, CAT

Table of Contents

- Image Representation & Pre-processing
- Object detection
- Semantic Segmentation
- Instance Segmentation

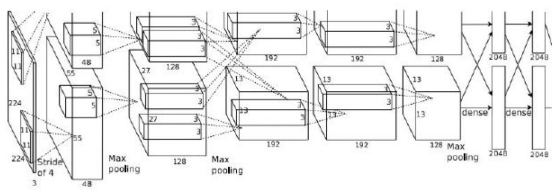
Object Detection: Impact of Deep Learning

- PASCAL VOC is a classical object detection benchmark



Object Detection as Classification: Sliding Window

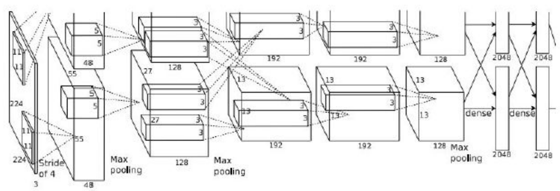
- Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? NO
Cat? NO
Background? YES

Object Detection as Classification: Sliding Window

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

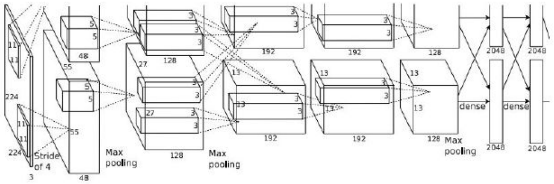


Dog? YES
Cat? NO
Background? NO

香港中文大學 The Chinese University of Hong Kong

Object Detection as Classification: Sliding Window

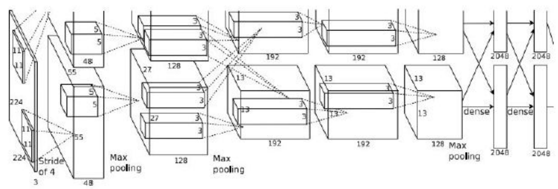
- Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? YES
Cat? NO
Background? NO

Object Detection as Classification: Sliding Window

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

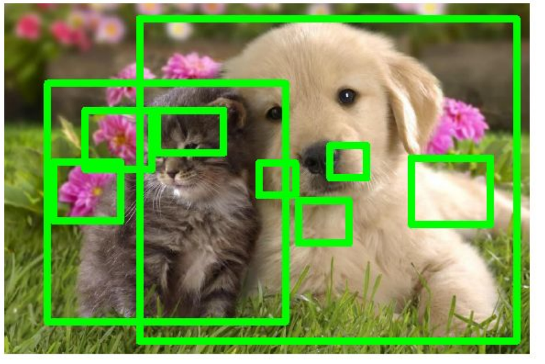


Dog? NO
Cat? YES
Background? NO

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

Region Proposals

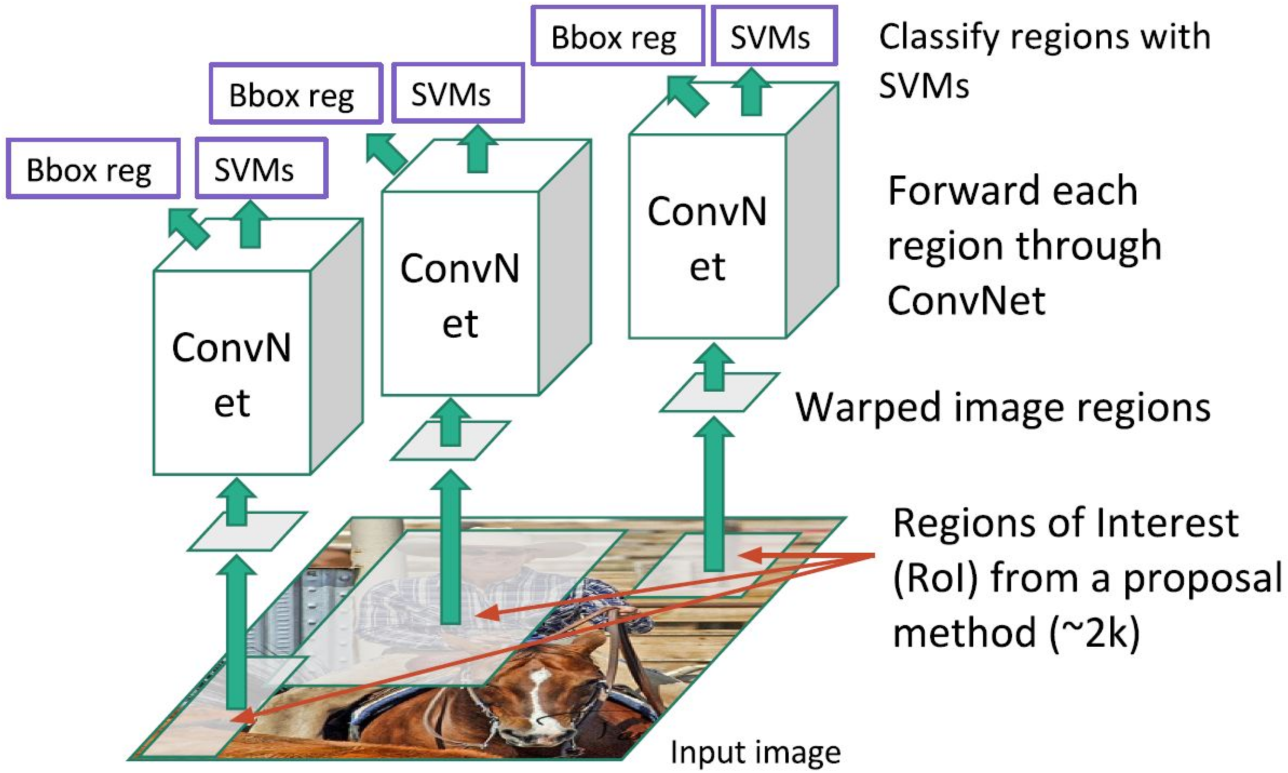
- Find plausible image regions that are likely to contain objects
- Relatively fast to run; e.g. Selective Search gives 1000 region proposals in a few seconds on CPU



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

香港中文大學 The Chinese University of Hong Kong

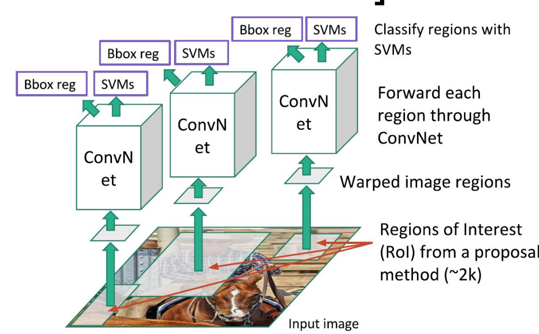
R-CNN



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

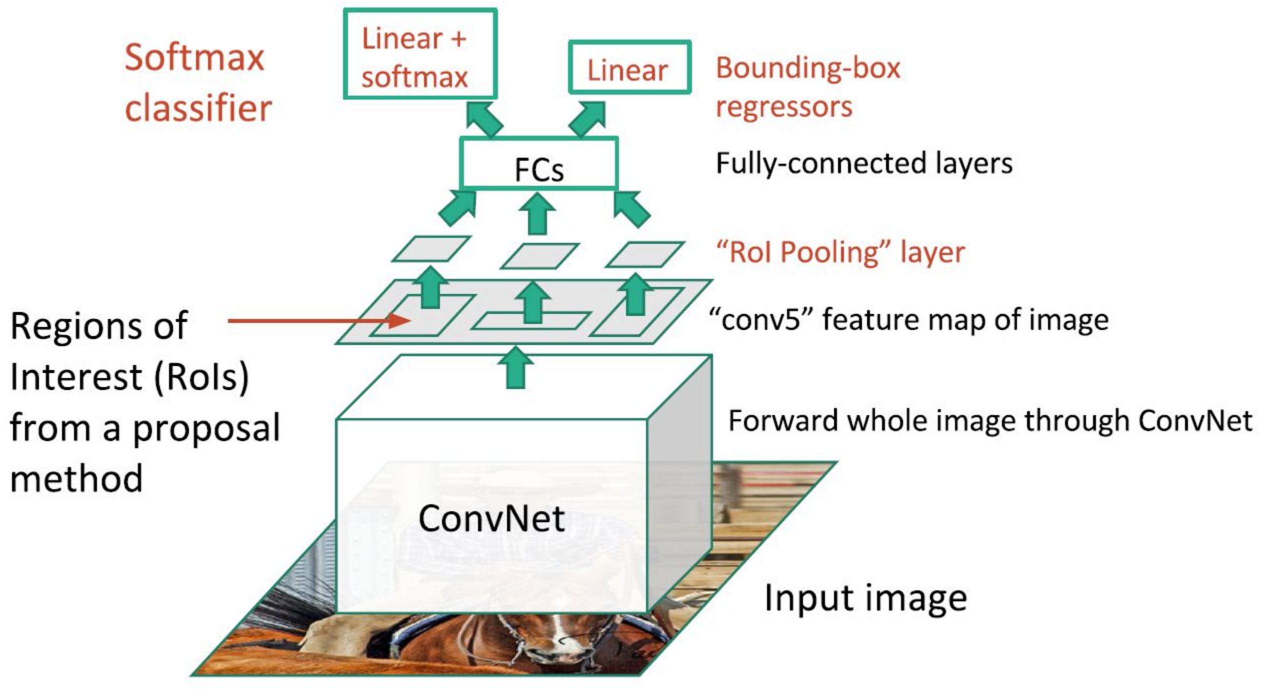
R-CNN: Problems

- Ad hoc training objectives
  - Fine-tune network with softmax classifier (log loss)
  - Train post-hoc linear SVMs (hinge loss)
  - Train post-hoc bounding-box regressions (least squares)
- Training is slow (84h), takes a lot of disk space
- Inference (detection) is slow
  - 47s / image with VGG16 [Simonyan & Zisserman. ICLR15]
  - Fixed by SPP-net [He et al. ECCV14]



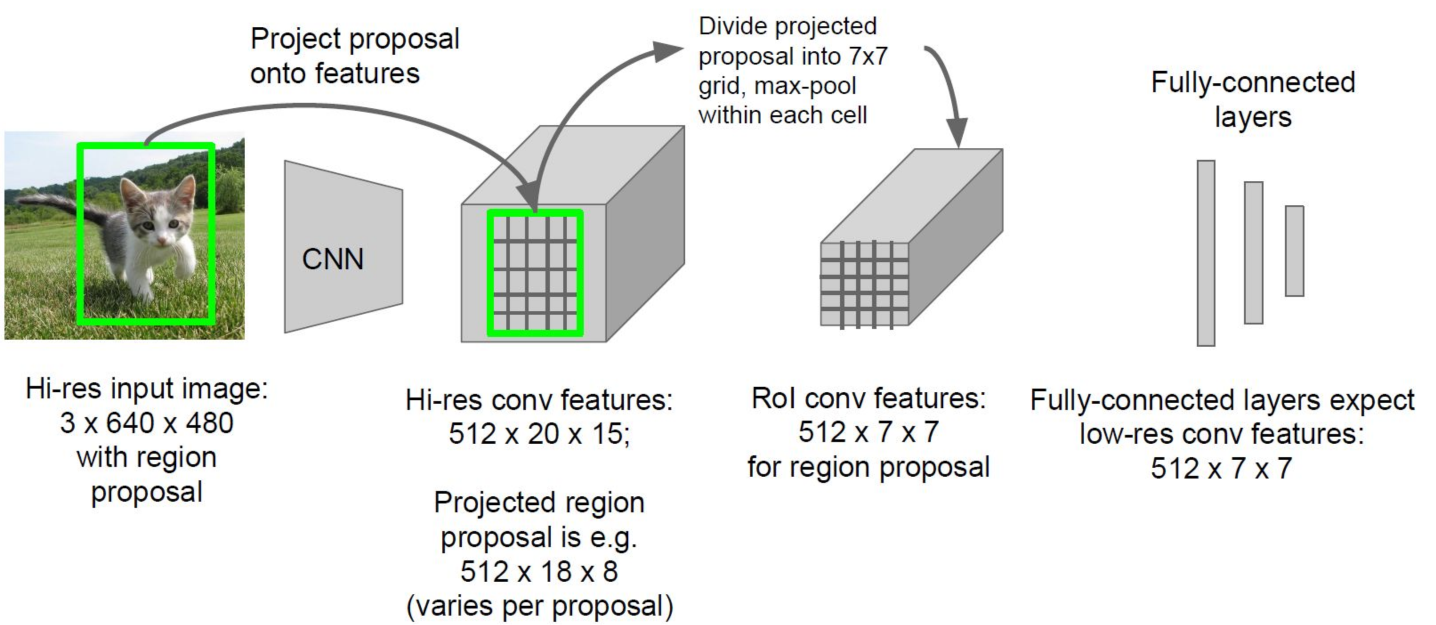
Girshick et al, “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR 2014.

Fast R-CNN



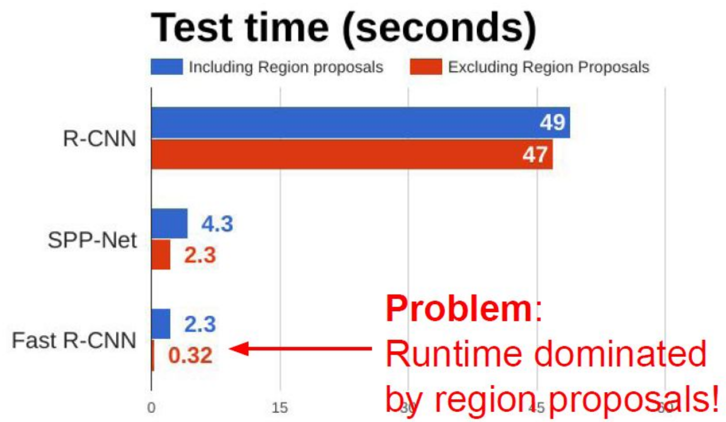
Girshick et al, "Fast R-CNN", ICCV 2015.

Fast R-CNN: ROI Pooling



Girshick et al, “Fast R-CNN”, ICCV 2015.

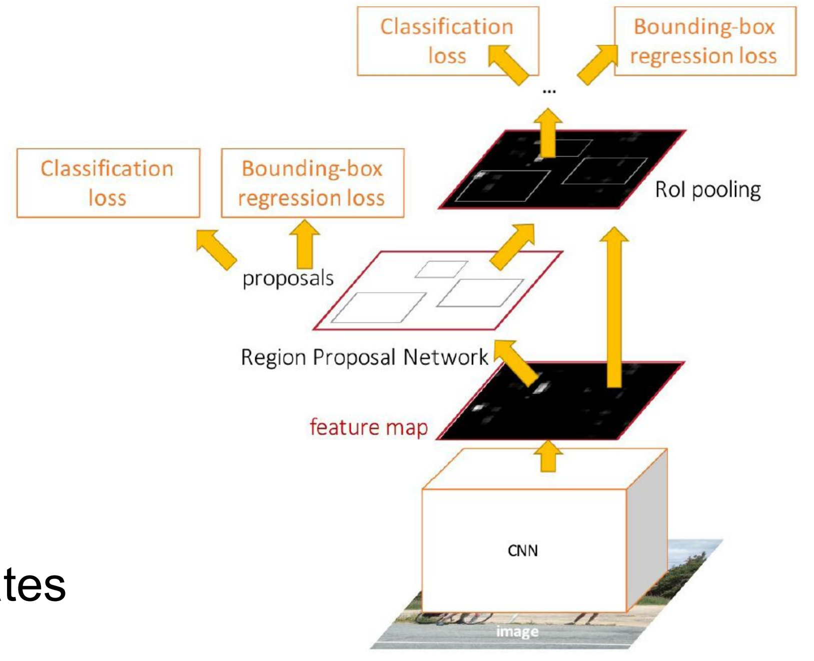
R-CNN vs SPP vs Fast R-CNN



He et al, "Spatial pyramid pooling in deep convolutional networks for visual recognition", ECCV 2014
Girshick et al, "Fast R-CNN", ICCV 2015.

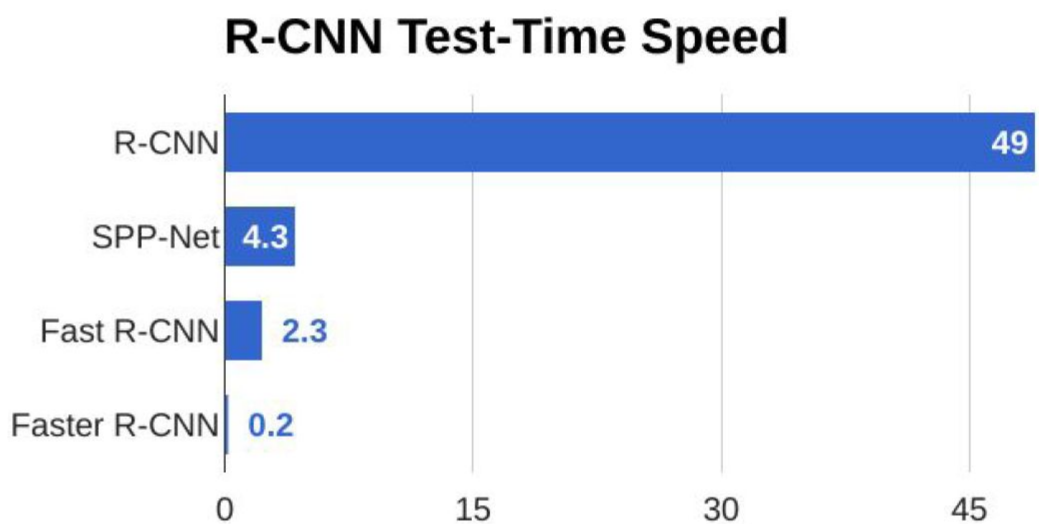
Faster R-CNN

- Make CNN do proposals!
- Insert **Region Proposal Network (RPN) to predict** proposals from features
- Jointly train with 4 losses:
  - RPN classify object / not object
  - RPN regress box coordinates
  - Final classification score (object classes)
  - Final box coordinates



Ren et al, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NIPS 2015

Faster R-CNN

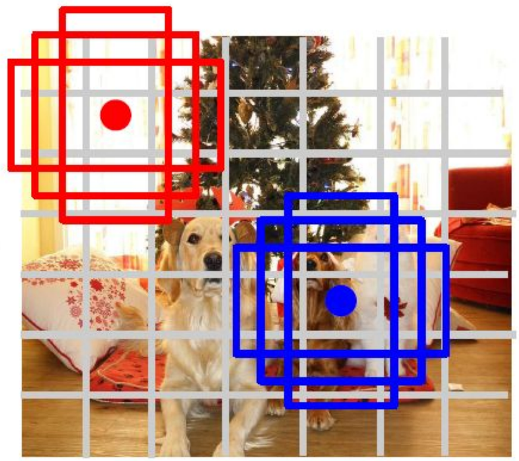


Ren et al, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NIPS 2015

One-stage Methods without Proposals: YOLO / SSD

Go from input image to tensor of scores with one big convolutional network!

Input image
3 x H x W



Divide image into grid
7 x 7

Image a set of **base boxes** centered at each grid cell
Here B = 3

Within each grid cell:

- Regress from each of the B base boxes to a final box with 5 numbers:
  (dx, dy, dh, dw, confidence)
- Predict scores for each of C classes (including background as a class)

Output:
7 x 7 x (5 * B + C)

Redmon et al, “You Only Look Once: Unified, Real-Time Object Detection”, CVPR 2016
Liu et al, “SSD: Single-Shot MultiBox Detector”, ECCV 2016

Object Detection: Lots of variables ...

Base Network
VGG16
ResNet-101
Inception V2
Inception V3
Inception
ResNet
MobileNet

Object Detection architecture
Faster R-CNN
R-FCN
SSD

Image Size
 # Region Proposals
….

Takeaways
Faster R-CNN is slower but more Accurate

SSD is much faster but not as accurate

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

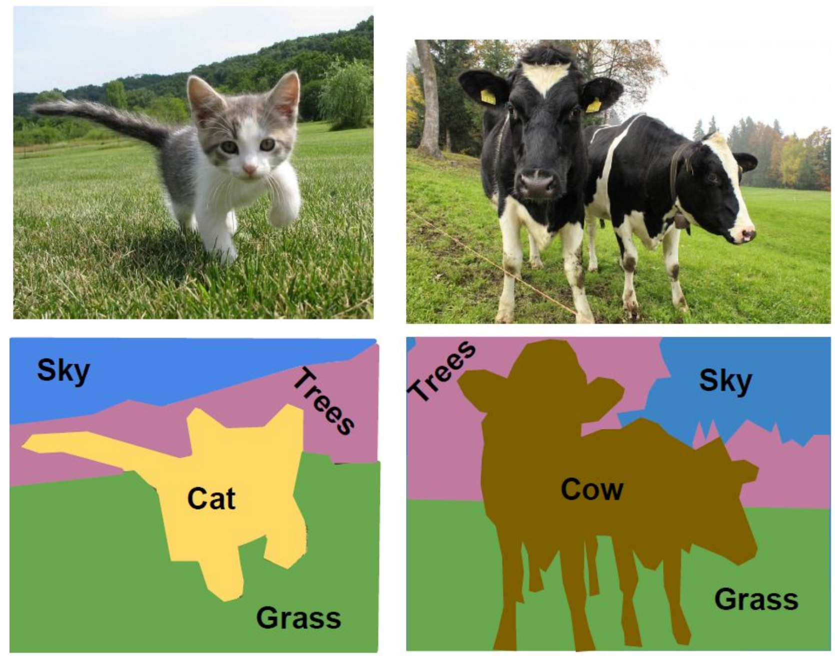
R-FCN: Dai et al, "R-FCN: Object Detection via Region-based Fully Convolutional Networks", NIPS 2016
Inception-V2: Ioffe and Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", ICML 2015
Inception V3: Szegedy et al, "Rethinking the Inception Architecture for Computer Vision", arXiv 2016
Inception ResNet: Szegedy et al, "Inception-V4, Inception-ResNet and the Impact of Residual Connections on Learning", arXiv 2016
MobileNet: Howard et al, "Efficient Convolutional Neural Networks for Mobile Vision Applications", arXiv 2017

Table of Contents

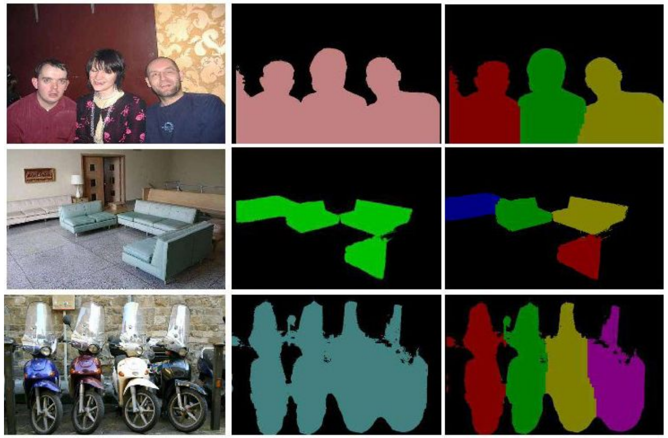
- Image Representation & Pre-processing
- Object detection
- Semantic Segmentation
- Instance Segmentation

Semantic Segmentation

- Classical Computer Vision problem
- Label each pixel in the image with a class label
- Does not differentiate instance, only care about pixels



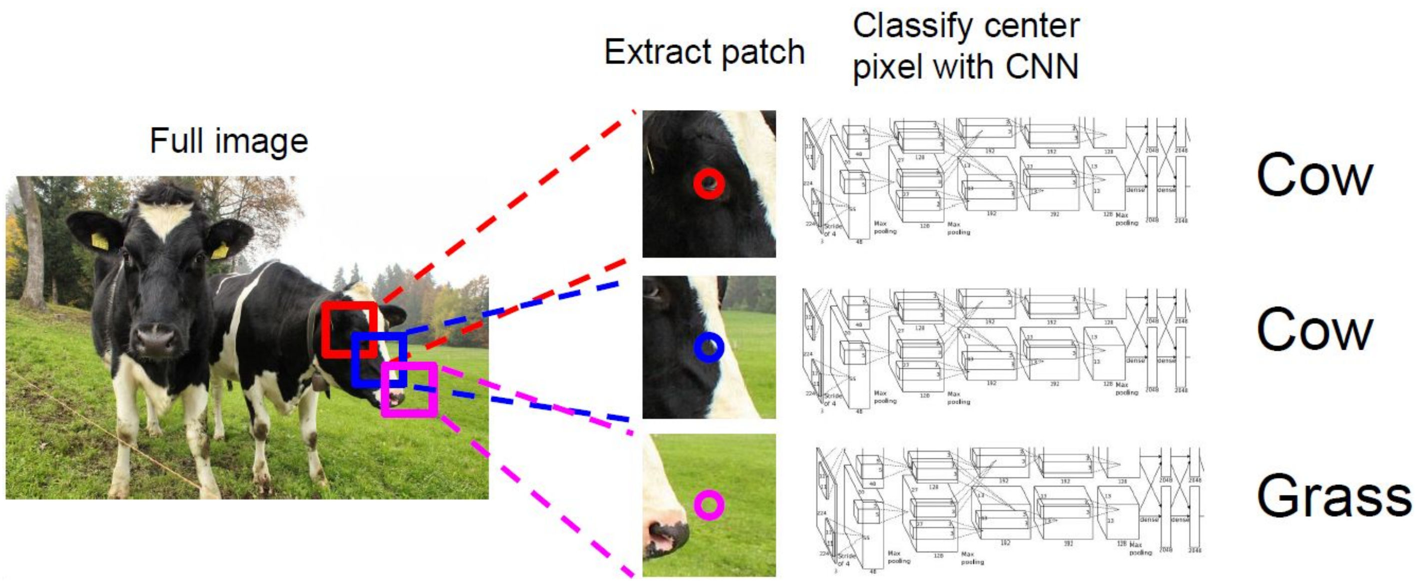
Some Public Semantic Segmentation Datasets



Pascal Visual Object Classes
20 Classes
~ 5.000 images

Microsoft COCO
80 Classes
~ 300.000 images

Semantic Segmentation Idea: Sliding Window

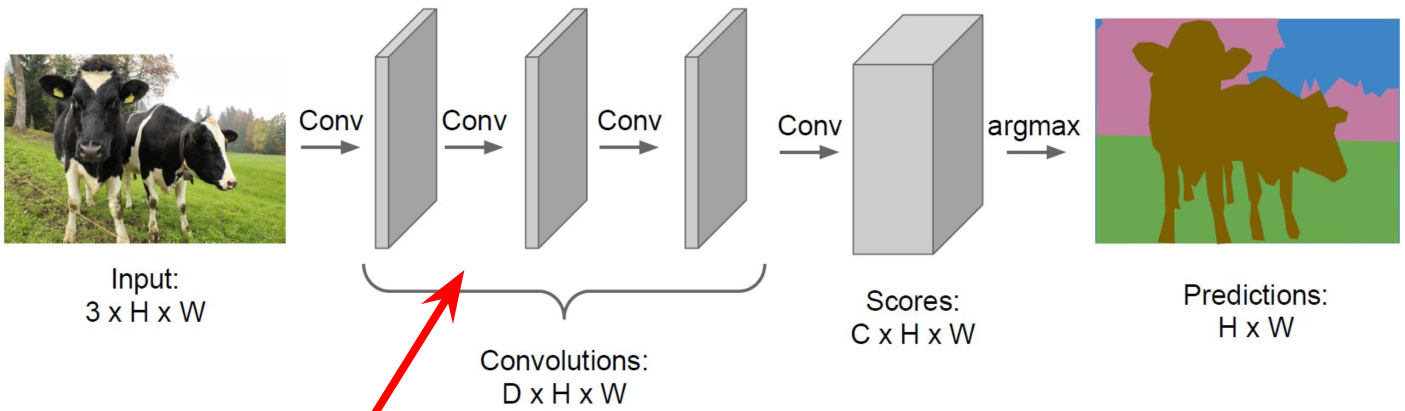


Problem: Very inefficient! Not reusing shared features between overlapping patches

Farabet et al, “Learning Hierarchical Features for Scene Labeling,” TPAMI 2013
Pinheiro and Collobert, “Recurrent Convolutional Neural Networks for Scene Labeling”, ICML 2014

Semantic Segmentation Idea: Fully Convolutional

Design a network as a bunch of convolutional layers to make predictions for pixels all at once!



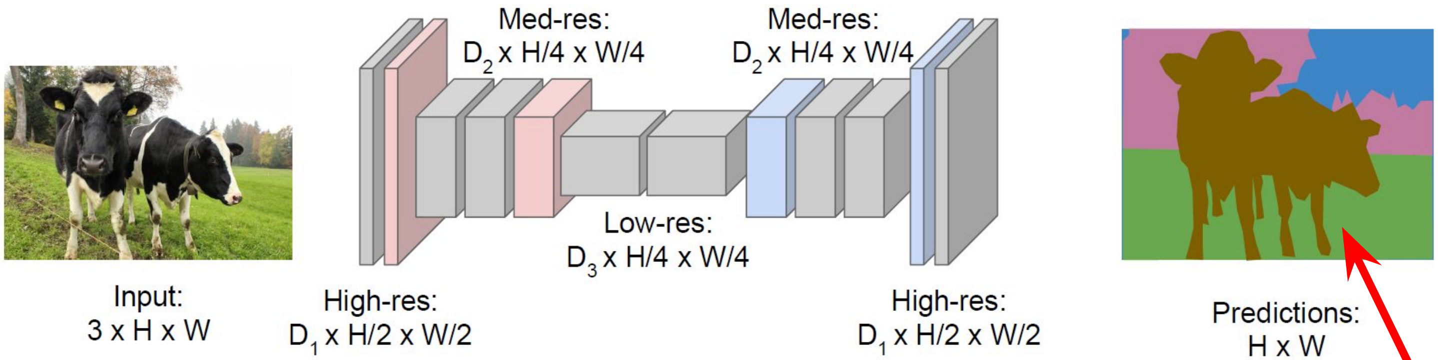
Problem: convolutions at original image resolution will be very expensive ...

Semantic Segmentation Idea: Fully Convolutional

Downsampling:
Pooling, strided convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

Upsampling:
???

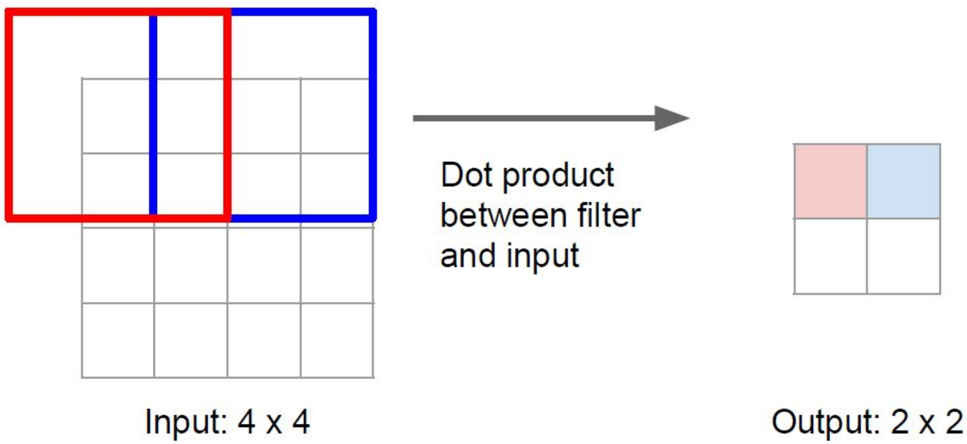


Apply cross-entropy loss at every pixel of the predicted label map

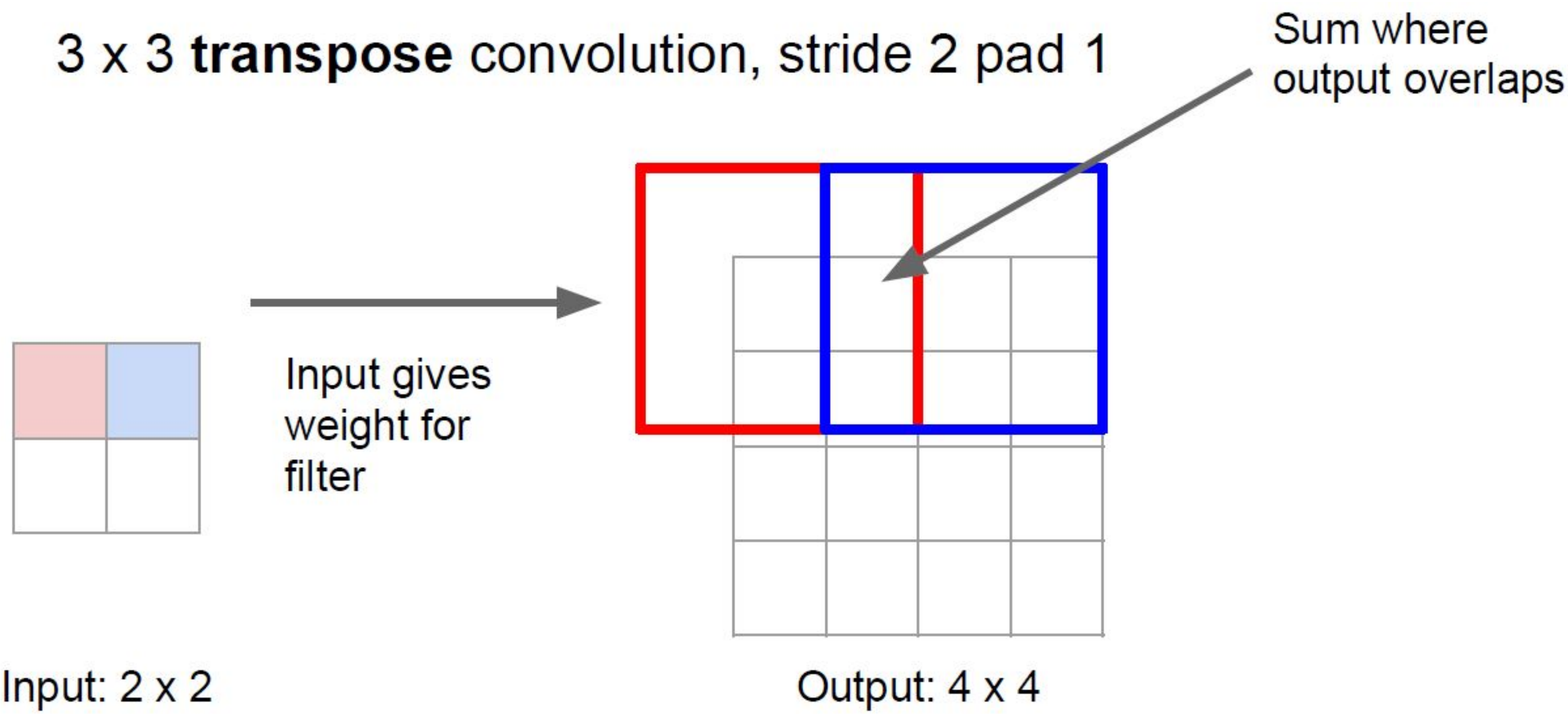
Long, Shelhamer, and Darrell, “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015
Noh et al, “Learning Deconvolution Network for Semantic Segmentation”, ICCV 2015

Convolution Layer

Typical 3 x 3 convolution, stride 2 pad 1



"Deconvolution" Layer for Upsampling

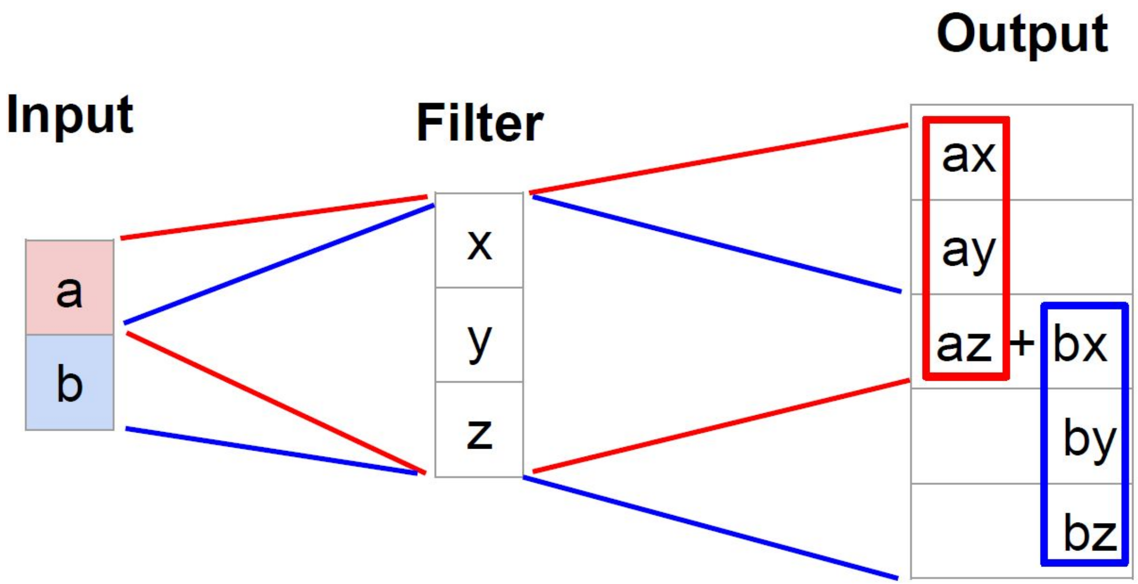


Other names:
-Deconvolution (bad)
-Upconvolution
-Fractionally strided convolution
-Backward strided convolution

Filter moves 2 pixels in the output for every one pixel in the input

Stride gives ratio between movement in output and input

Transpose Convolution: 1D Example



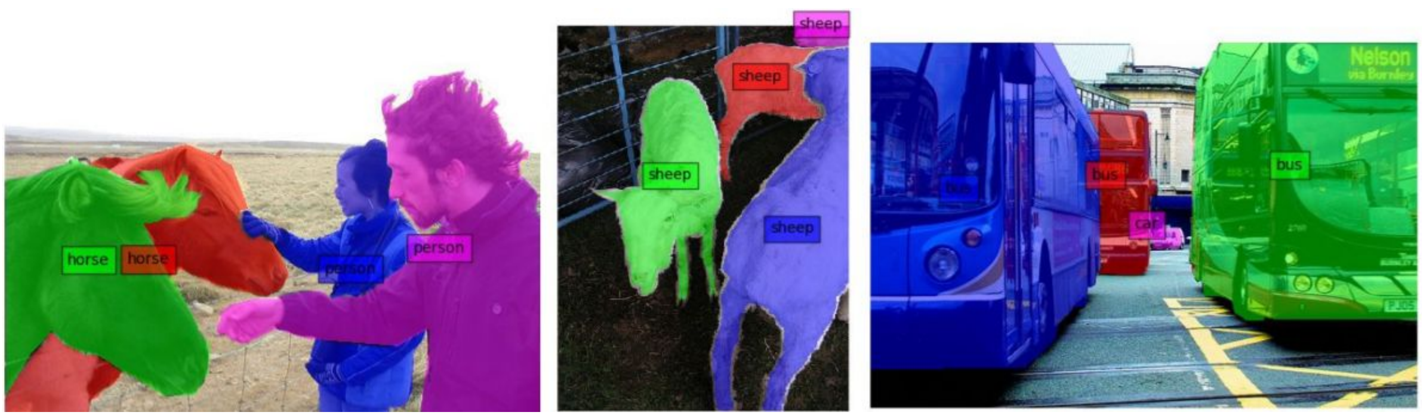
Output contains copies of the filter weighted by the input, summing at where at overlaps in the output

Need to crop one pixel from output to make output exactly 2x input

Table of Contents

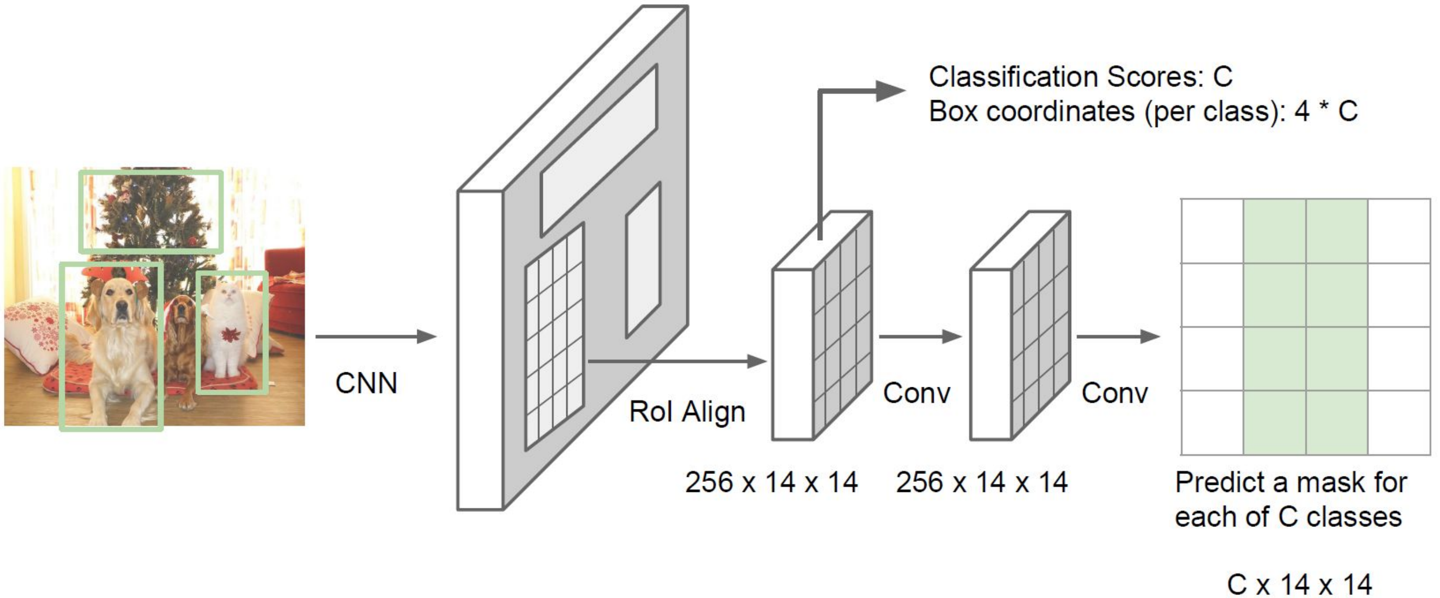
Instance Segmentation

- Not only to segment each pixel but differentiate different instances of the same class
- Idea: combining object detection and semantic segmentation for instance segmentation



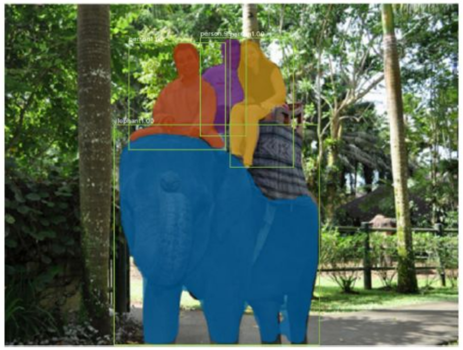
Mask R-CNN

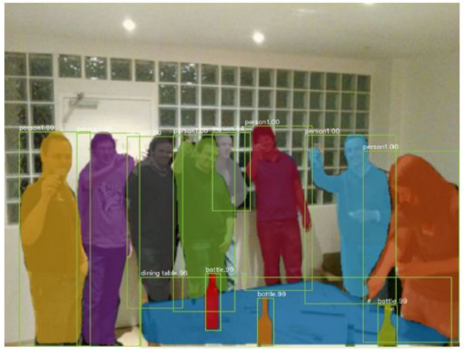
- Idea: combining object detection and semantic segmentation for instance segmentation



He et al, “Mask R-CNN”, ICCV 2017

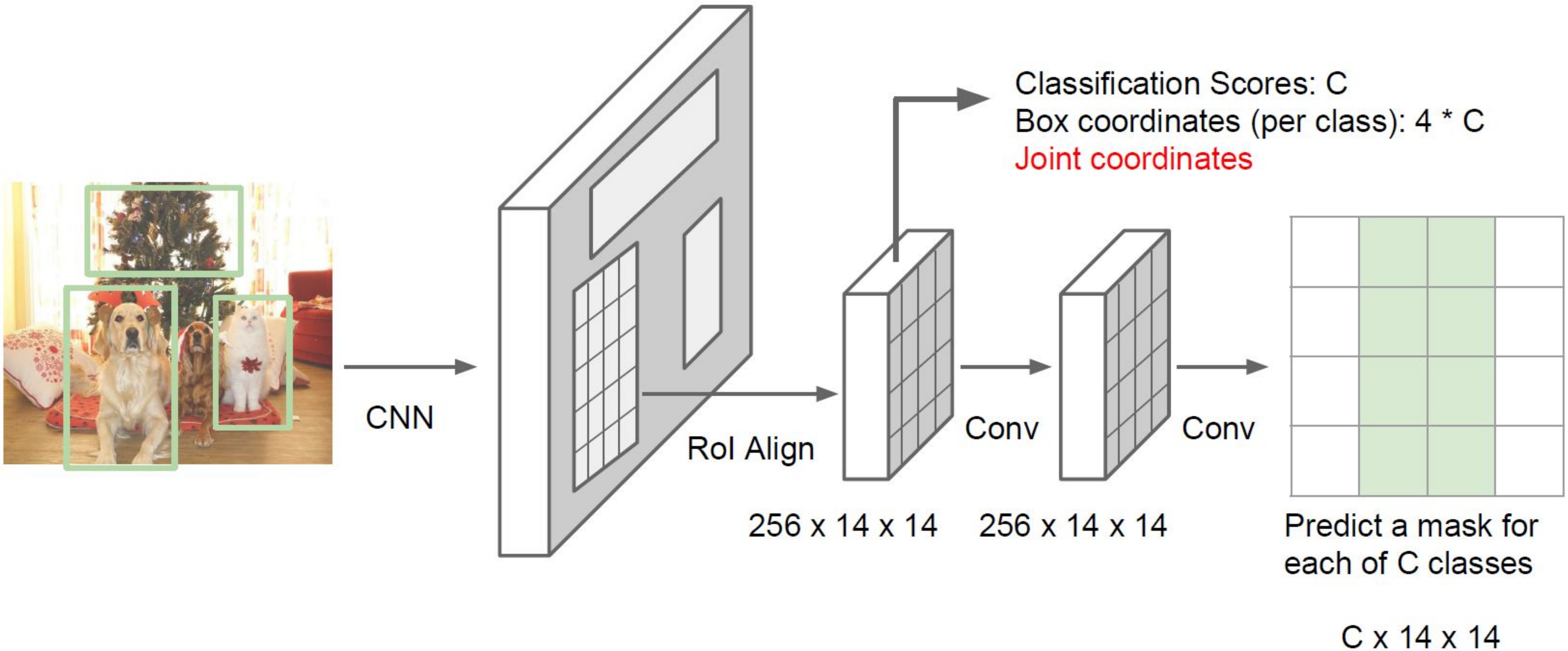
Mask R-CNN: Very Good Results





He et al, “Mask R-CNN”, ICCV 2017

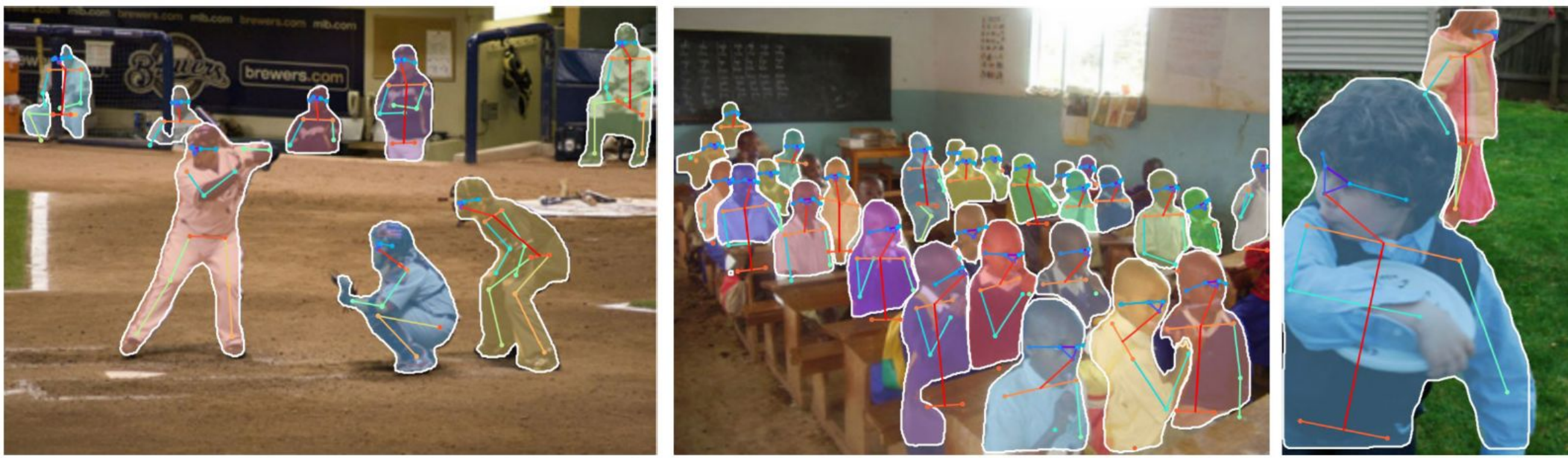
Mask R-CNN: Also Can Estimate Human Poses



He et al, “Mask R-CNN”, ICCV 2017

香港中文大學
The Chinese University of Hong Kong

Mask R-CNN: Also Can Estimate Human Poses



He et al, "Mask R-CNN", ICCV 2017

香港中文大學
The Chinese University of Hong Kong

Thanks!

ELEG 5491 Tutorial
Xihui Liu