Advanced Section #4: Semantic Segmentation and Object Detection

Pavlos Protopapas and Robbert Struyven

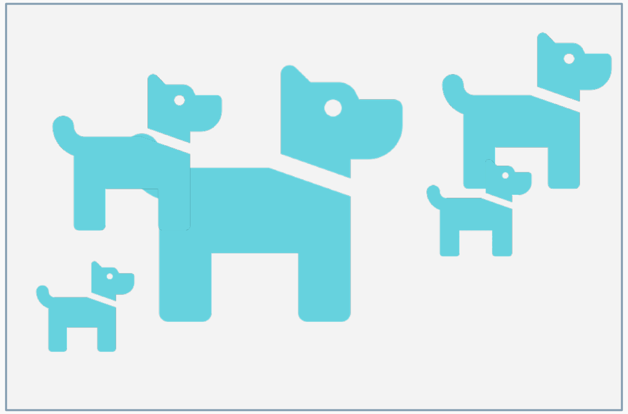
CS109B Advanced Topics in Data Science

Pavlos Protopapas

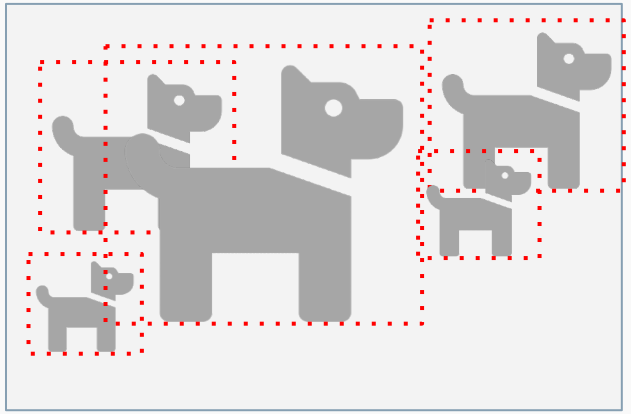
Computer Vision Tasks

Classification

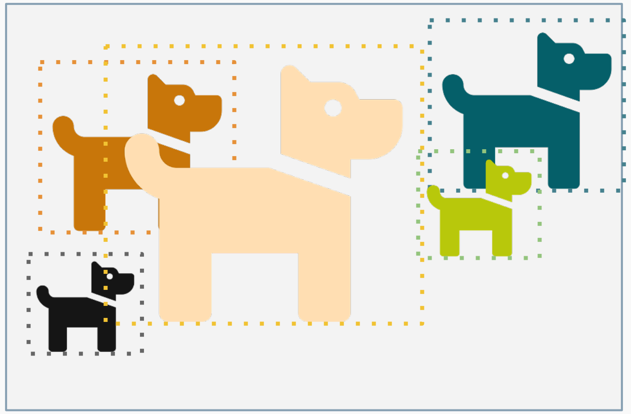
Semantic Segmentation



Object Detection



Instance Segmentation



Object Detection & Semantic Segmentation

Object Detection: let’s classify and locate

- Sliding Window versus Region Proposals
- Two stage detectors: the evolution of R-CNN , Fast R-CNN, Faster R-CNN
- Single stage detectors: detection without Region Proposals: YOLO / SSD

Semantic Segmentation: classify every pixel

- Fully-Convolutional Networks
- SegNet & U-NET
- Faster R-CNN linked to Semantic Segmentation: Mask R-CNN

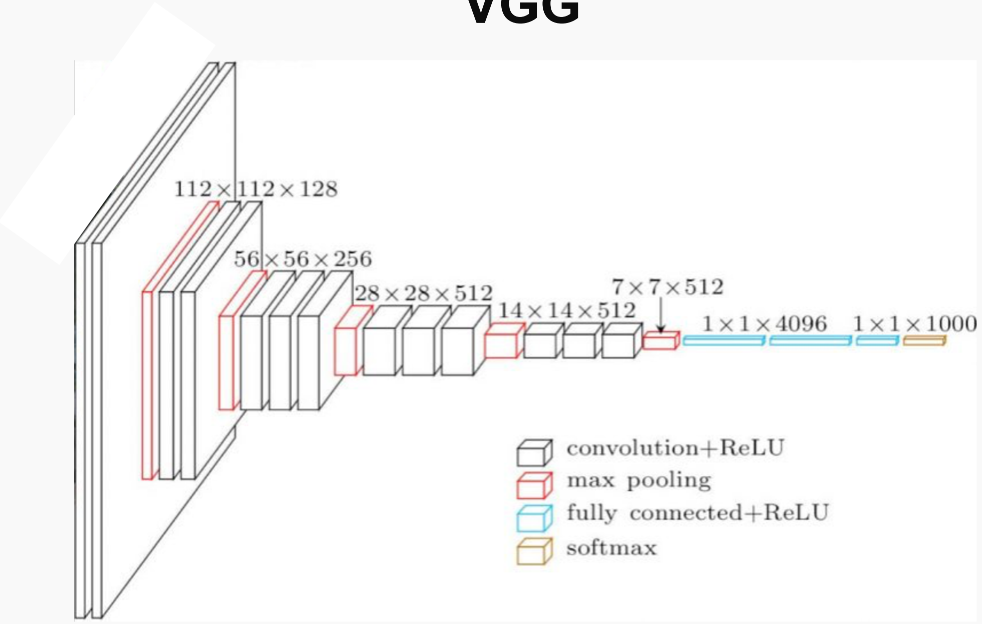
Task: Image Classification using Fully-Connected CNN

- Fundamental to computer vision given a set of labels {dog, cat, human, ...};
- Predict the most likely class.

A SOTA CNN
See a-sec 5 for more

Input

VGG



Output

Classification (C = 1000):

- Dog: 0.95
- Cat: 0.02
- Human: 0.01
- ...

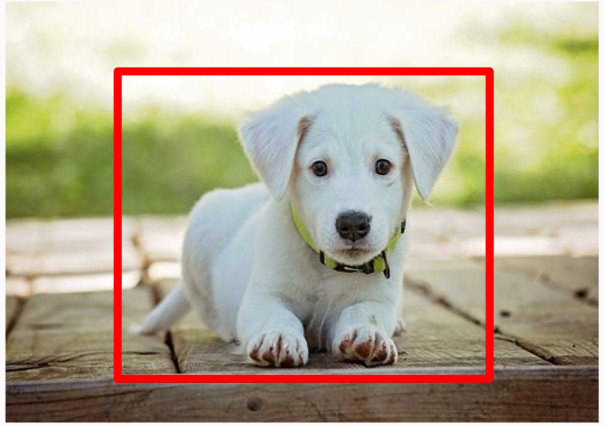
Task: From Classification to Classification + Localization

- Localization demands to compute **where 1 object is present in an image**
- Limitation: only 1 object (also non-overlapping)
- Typically implemented using a bounding box (x, y, w, h)

Predict

Classification Output:
- Dog: 0.95
- Cat: 0.02
- Human: 0.01
- ...

Predict



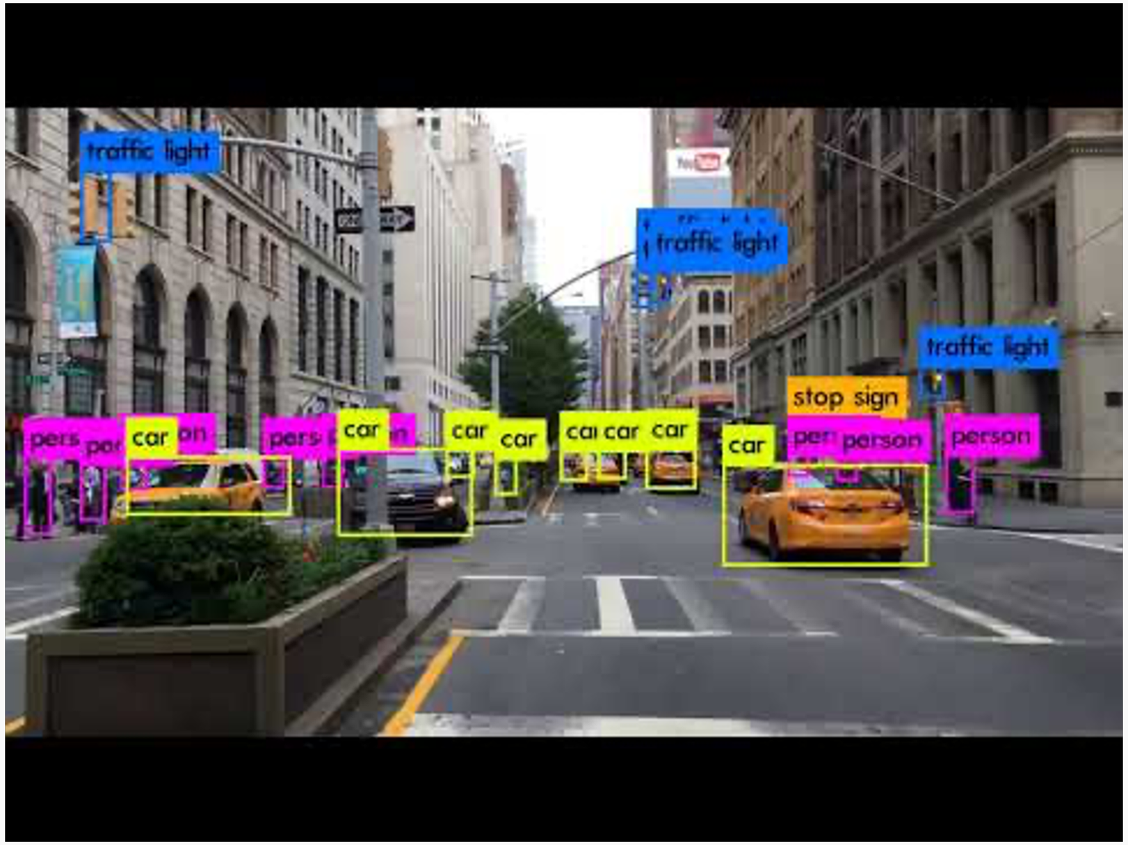
Classification output:
- Dog: 0.95
- Cat: 0.02
- Human: 0.01

Localization output:
- Bounding-Box: (x, y, w, h)

Output: Regular Image Classification

Task: From Classification + Localization to Object Detection

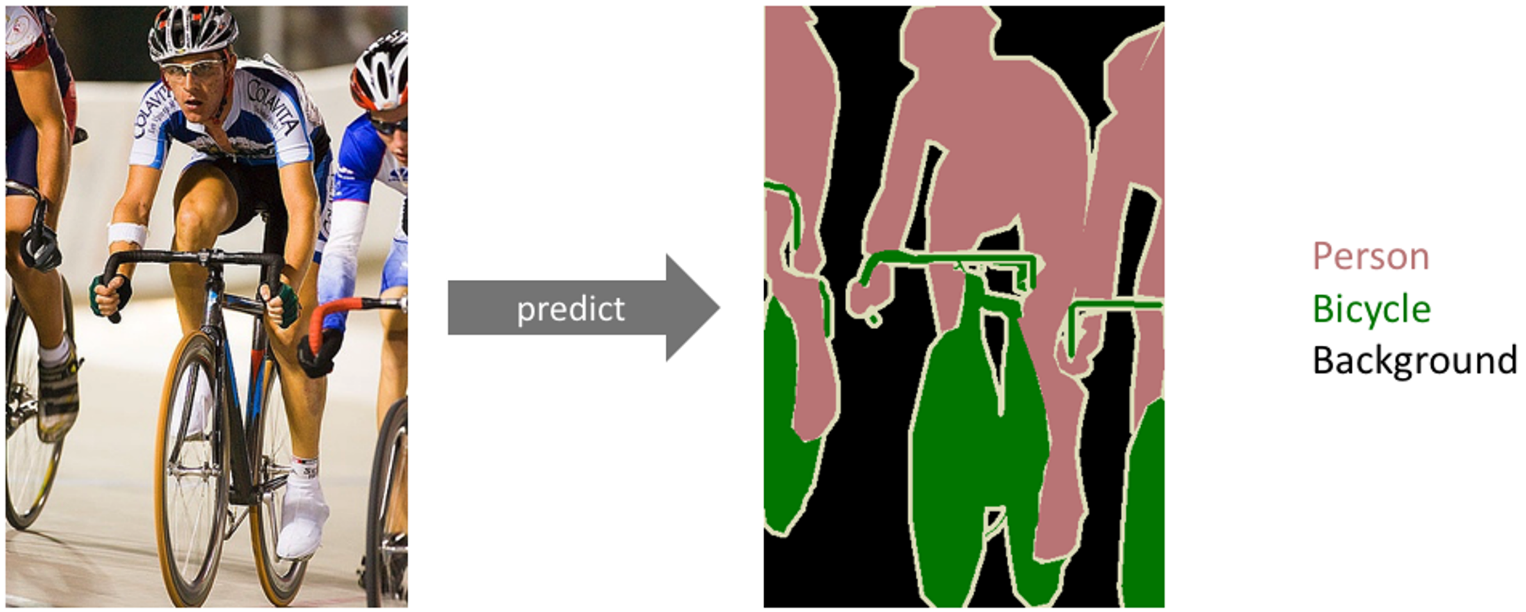
- Classification and Localization extended to multiple objects



Youtube 'YOLO in New York" by Joseph Redmon (creator of YOLO)

Task: From Classification to Semantic Segmentation

- **Image Classification:** assigning a single label to **the entire picture**
- **Semantic Segmentation:** assigning a semantically meaningful label to **every pixel in the image**



Long, Shelhamer et al. “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015 : Cited by 14480

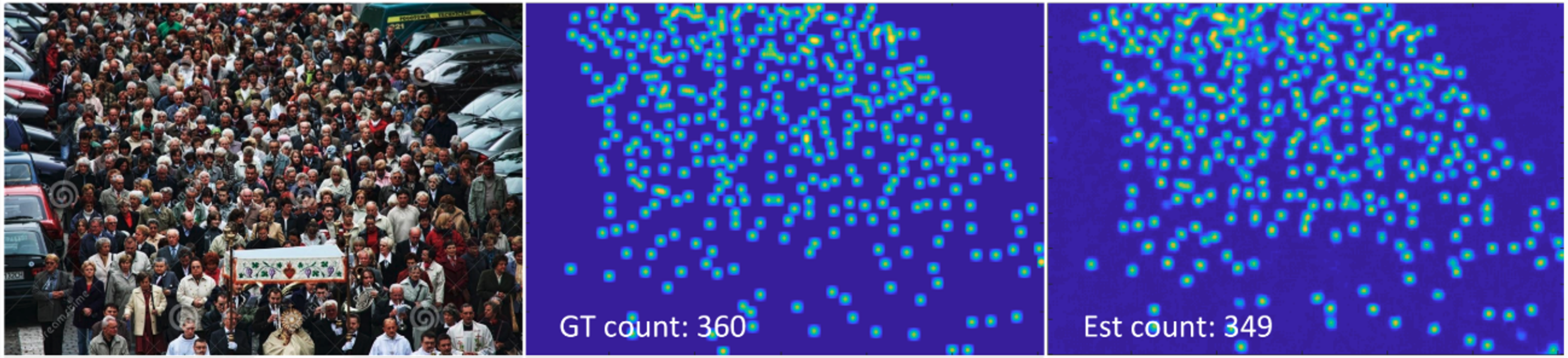
Why Object Detection and Semantic Segmentation

Computer Vision:

- Autonomous vehicles
- Biomedical Imaging detecting cancer, diseases
- Video surveillance:
  - Counting people
  - Tracking people
- Aerial surveillance
- Geo Sensing: tracking wildfire, glaciers, via satellite

Note:

- Efficiency/inference-time is important!
- How many frames/sec. can we predict?
- Must for real-time segmentation & detection.



Why Object Detection and Semantic Segmentation

Youtube: “Tensorflow DeepLab v3 Xception Cityscapes”(link)

How to Measure Quality in Detection and Segmentation?

- **Pixel Accuracy:**
  - Percent of pixels in your image that are classified correctly
  - Our model has 95% accuracy! Great!

Input **Labels** **Predict**

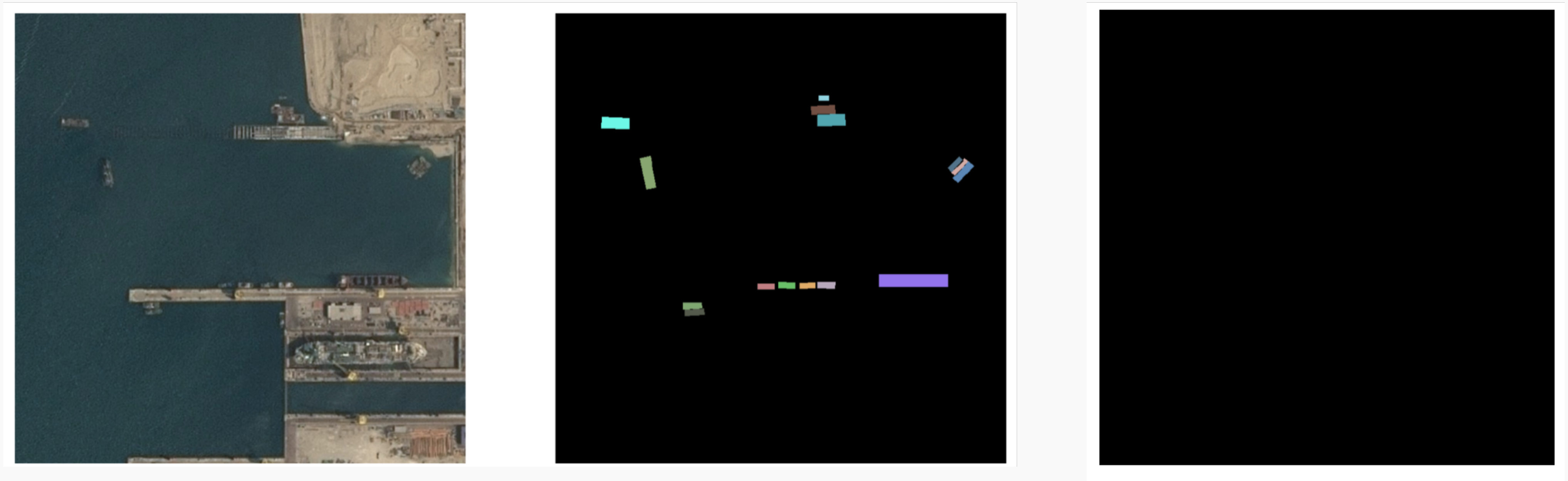


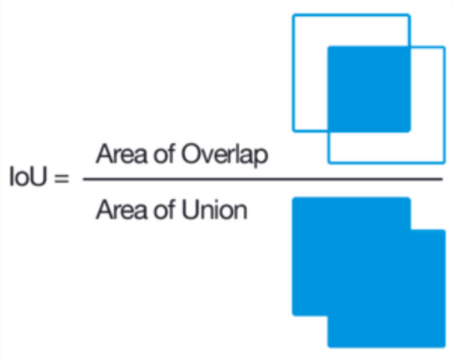
Image from Vlad Shmyhlo in article: Image Segmentation: Kaggle experience in TDS

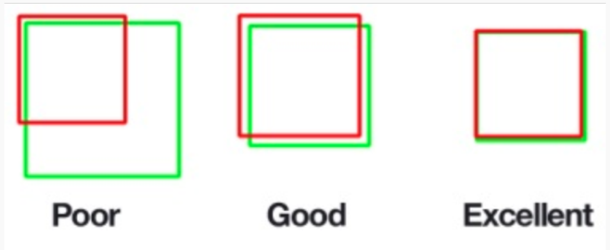
- **Problem with accuracy: unbalanced data!**

How Do We Measure Accuracy?

- **Pixel Accuracy**: Percent of pixels in your image that are classified correctly
- **IOU:** Intersection-Over-Union (Jaccard Index): Overlap / Union
- **mAP:** Mean Average Precision: AUC of Precision-Recall curve standard (0.5 is high)
- **DICE:** Coefficient (F1 Score): 2 x Overlap / Total number of pixels

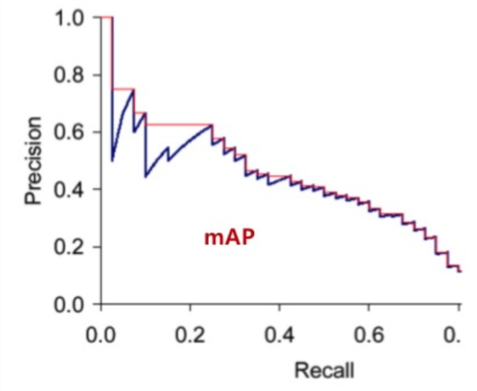
IOU



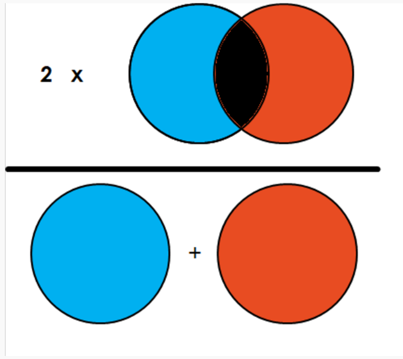


IoU: 0.40 **IoU:** 0.73 **IoU:** 0.92

mAP



DICE



Object Detection: let’s classify and locate

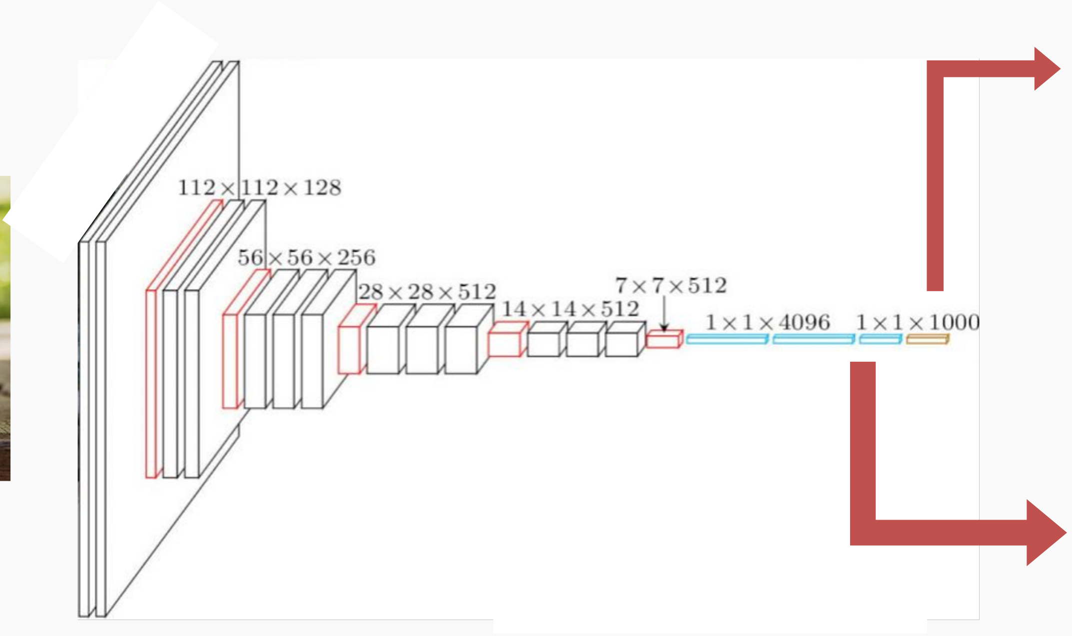
- Sliding Window versus Region Proposals
- Two stage detectors: the evolution of R-CNN , Fast R-CNN, Faster R-CNN
- Single stage detectors: detection without Region Proposals: YOLO / SSD

Task: Object Detection - Let's Classify and Locate

- Object detection is just classification and localization combined:
  - Classification using standard CNN;
  - Localization using regression problem for predicting box coordinates
  - Combining loss from Classification (Softmax) and Regression (L2)

Input

VGG



Classification:
- Dog: 0.95
- Cat: 0.02
- Human: 0.01
- ...

Bounding-Box :
- x
- y
- width (w)
- height (h)

Output

Softmax Loss

Multi-Task Learning

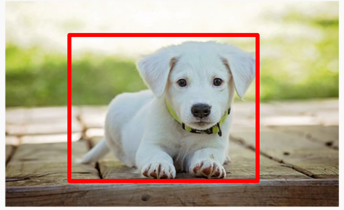
L2-norm Loss

Sliding Windows, from Single to Multiple Objects

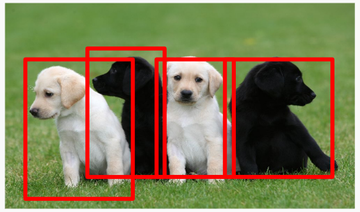
- Might work for single object, but not for multiple objects
- Each image containing "n" objects: needs "n" number of classification and localization outputs
- Solution for multiple objects:
  - Crop the image "in a smart way"
  - Apply the CNN to each crop

- Can we just use sliding windows?
  - Problem: Need for applying CNN to huge number of locations, scales, bbox aspect ratios: very computationally expensive

Solution: Region Proposals methods to find object-like regions



Dog: (x, y, w, h)



Dog: (x, y, w, h)
Dog: (x, y, w, h)
Dog: (x, y, w, h)
Dog: (x, y, w, h)

Adapted from Fei-Fei Li & Justin Johnson & Serena Yeung Stanford CS231n 2019 "Convolutional Neural Networks for Visual Recognition" Lecture 12 Slide 37
Uijlings et al, Selective Search for Object Recognition" IJCV 2013 link

Object Detection: Region Proposal Networks!

- **Problem**: Need for applying CNN to huge number of locations, scales, bbox aspect ratios: very computationally expensive
- **Solution:** Region Proposals methods to find object-like regions
- **Selective Search Algorithm:** returns boxes that are likely to contain objects
    - Use hierarchical segmentation
    - Start with small superpixels
    - Merge based on similarity
- **Output:** Where are object like regions
    - No classification yet

Uijlings et al, Selective Search for Object Recognition" IJCV 2013 [link]

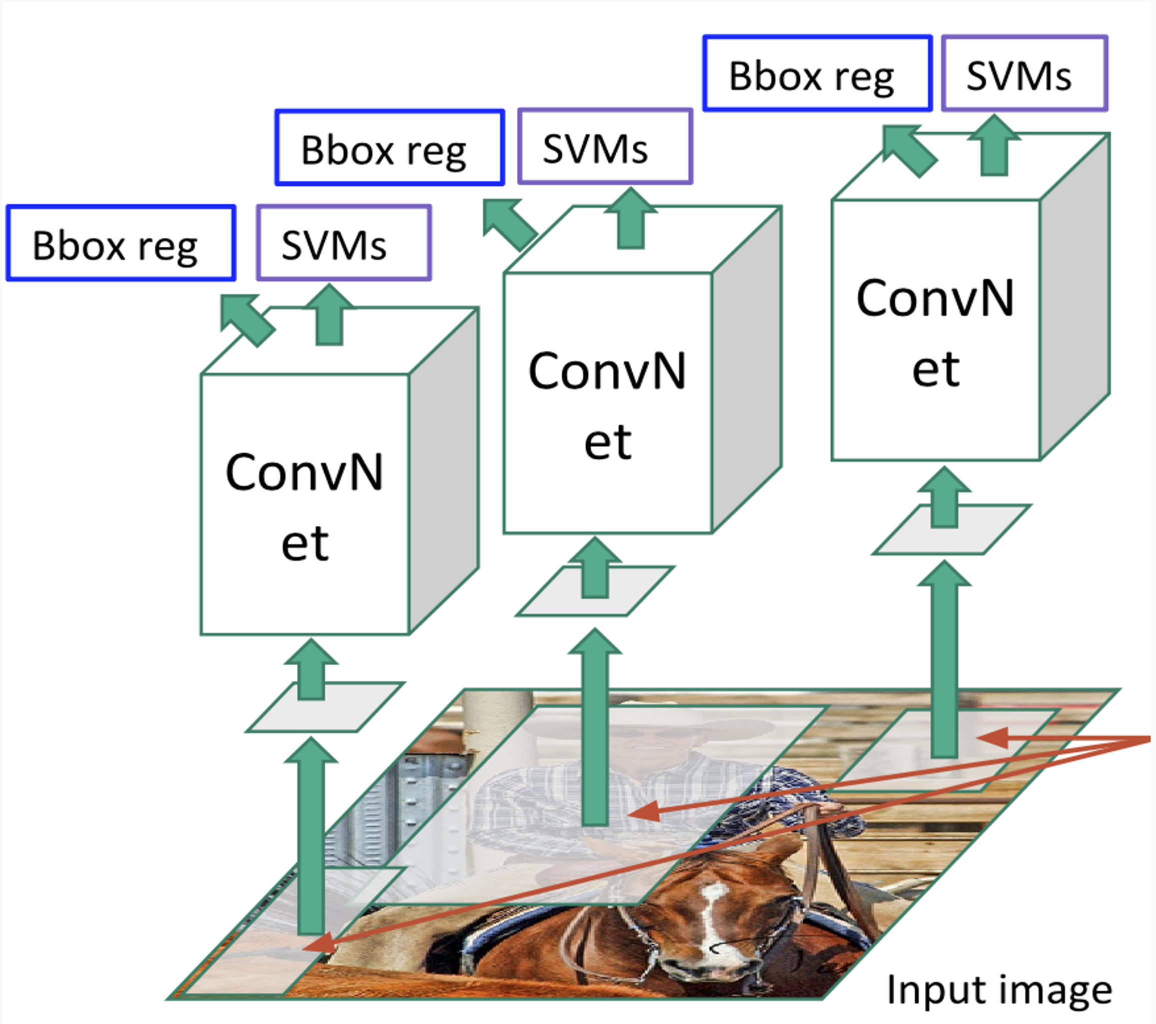
The Evolution of R-CNN: R-CNN, Fast R-CNN, Faster R-CNN

R-CNN = Region-based CNN

- Correct Bbox by Bbox regressor (dx,dy,dw,dh)
- Forward each region through CNN
- Resize proposed RoI (224x224)

Region of Interest (RoI) from selective search region proposal (approx 2k)

Problem: need to do 2k independent forward passes for each image! **('slow' R-CNN)**



Adapted from Fei-Fei Li & Justin Johnson & Serena Yeung Stanford CS231n 2019 "Convolutional Neural Networks for Visual Recognition"
Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation" CVPR2014
Ross Girshick, "Fast R-CNN" Slides 2015

The Evolution of R-CNN: R-CNN, Fast R-CNN, Faster R-CNN

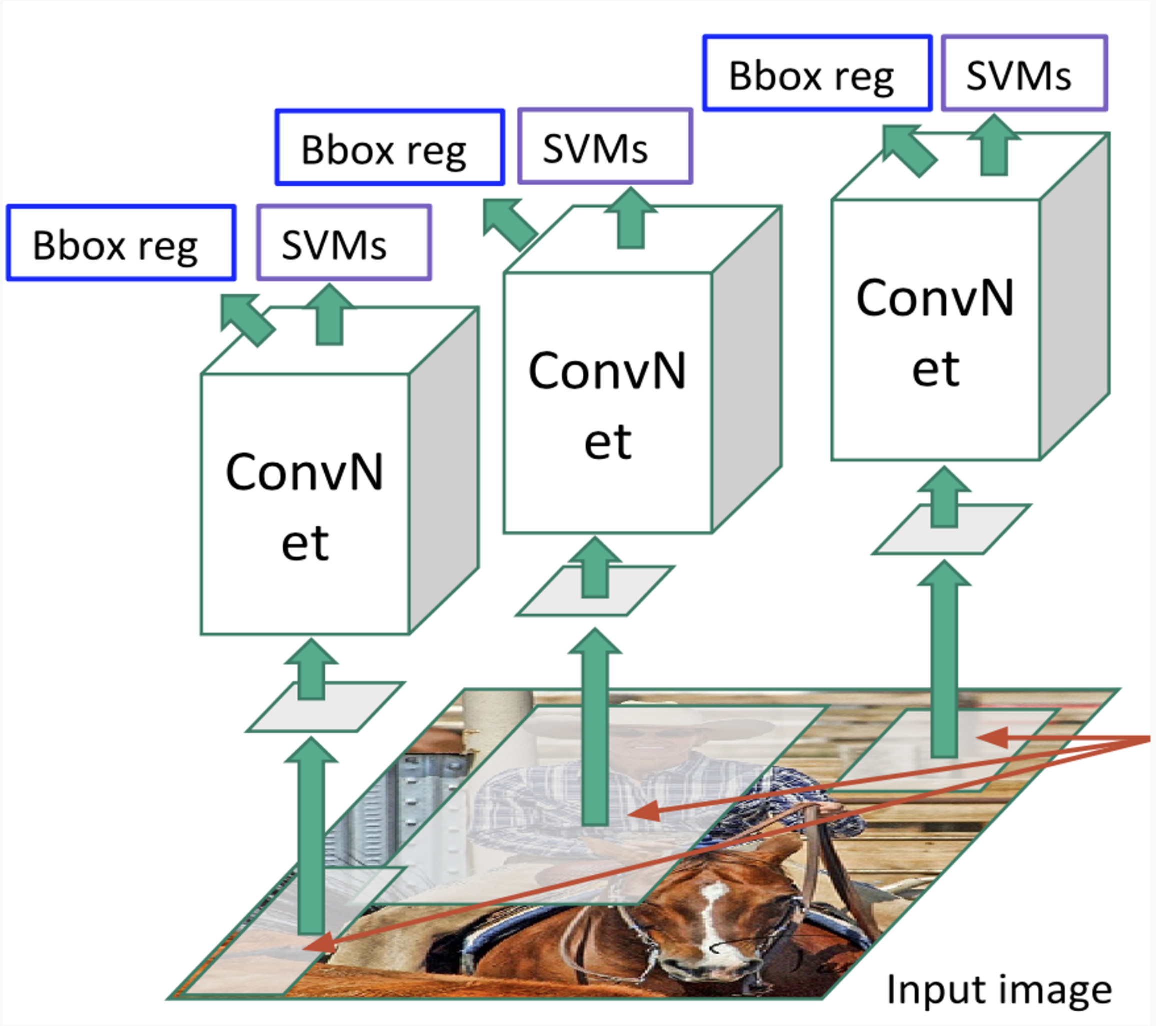
R-CNN = Region-based CNN

- Correct Bbox by Bbox regressor (dx,dy,dw,dh)
- Forward each region through CNN
- Resize proposed RoI (224x224)

Region of Interest (RoI) from selective search region proposal (approx 2k)

Problem: need to do 2k independent forward passes for each image! **('slow' R-CNN)**

Solution: can we process the image before cropping?

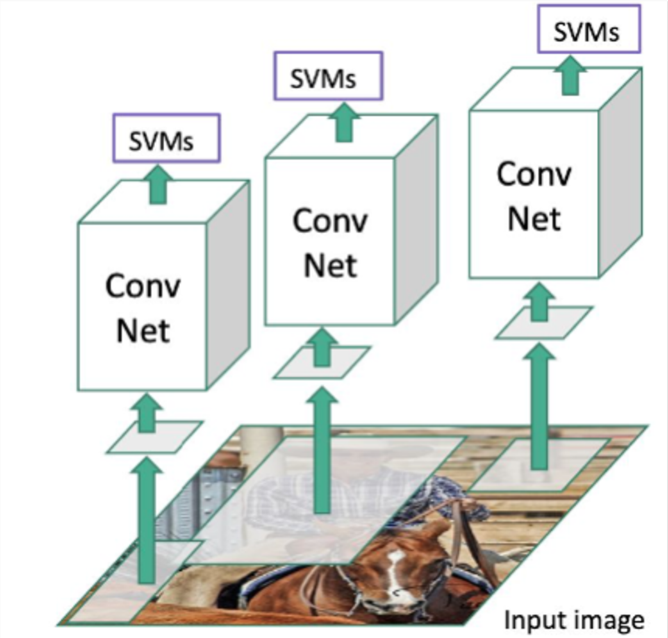


Adapted from Fei-Fei Li & Justin Johnson & Serena Yeung Stanford CS231n 2019 "Convolutional Neural Networks for Visual Recognition"
Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation" CVPR2014
Ross Girshick, "Fast R-CNN" Slides 2015

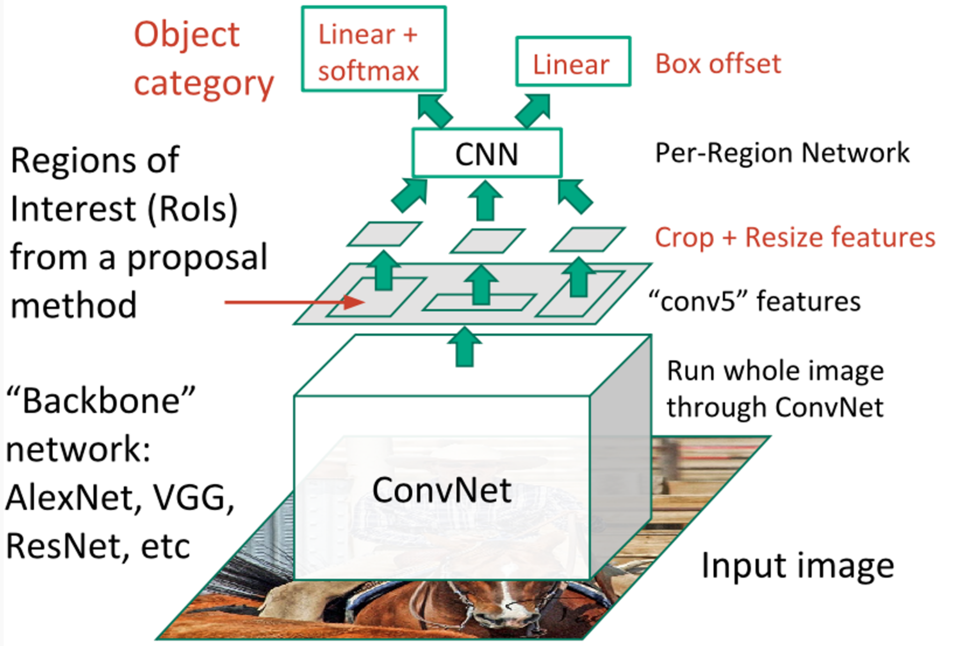
The Evolution of R-CNN: R-CNN, Fast R-CNN, Faster R-CNN

- **Problem:** need to do 2k independent forward passes for each image! **('slow' R-CNN)**
- Even inference is slow: 47s/image with VGG16 [Simonyan & Zisserman, ICLR 15]
- **Solution:** can we process (CNN forward pass) the image before cropping generates 2k regions?

Slow R-CNN



Fast R-CNN

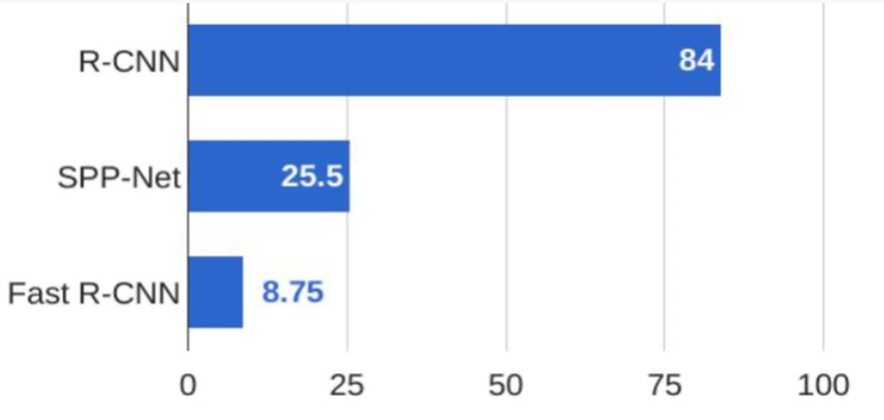


Adapted from Fei-Fei Li & Justin Johnson & Serena Yeung Stanford CS231n 2019 "Convolutional Neural Networks for Visual Recognition"
Ross Girshick, "Fast R-CNN" Slides 2015

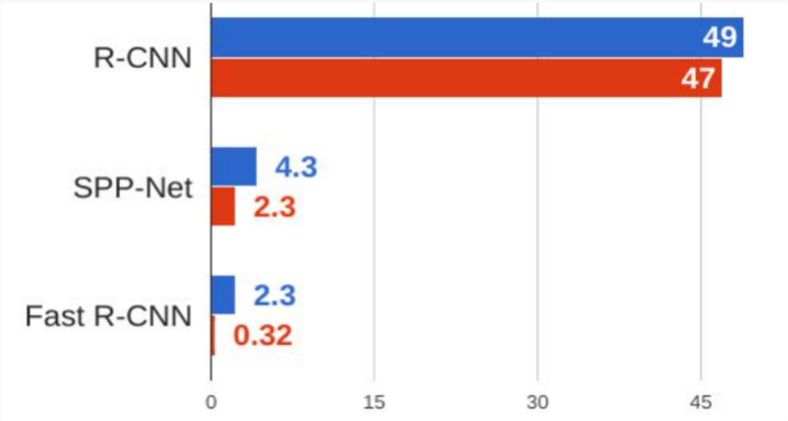
The Evolution of R-CNN: R-CNN, Fast R-CNN, Faster R-CNN

- Fast R-CNN is much faster than R-CNN
- Runtime dominated by region proposals; an iterative method (‘like selective search’);
- **Solution:** Can we make the CNN do proposals?!

Training Time (Hours)



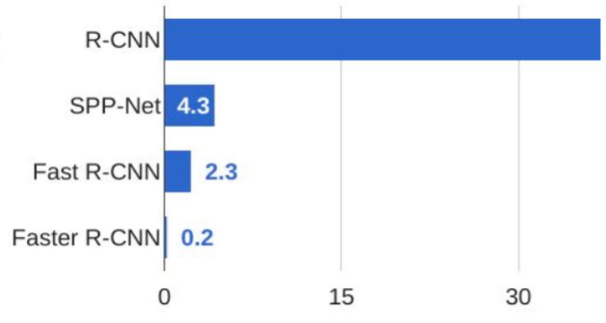
Test Time (Seconds)

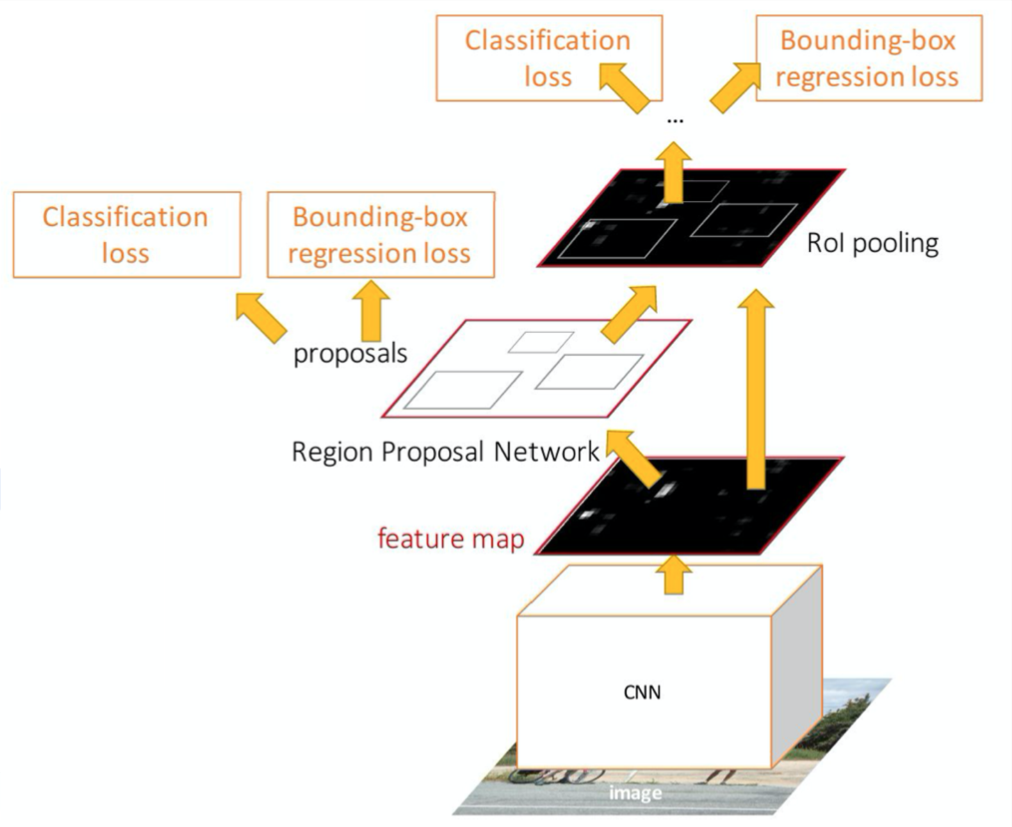


Adapted from Fei-Fei Li & Justin Johnson & Serena Yeung Stanford CS231n 2019 “Convolutional Neural Networks for Visual Recognition”
Ross Girshick, “Fast R-CNN” Slides 2015

The Evolution of R-CNN: R-CNN, Fast R-CNN, Faster R-CNN

- **Faster R-CNN**: Have the CNN make proposals! (single forward, not iterative selective search)
- **CNN Region Proposal Network (RPN)**: Predict region proposals from features
- Otherwise same as Fast R-CNN: crop and classify
- End-to-end quadruple loss:
    - RPN classify object / not object
    - RPN regress box coordinates
    - Final classification score (object classes)
    - Final box coordinates
- Test-time seconds per image

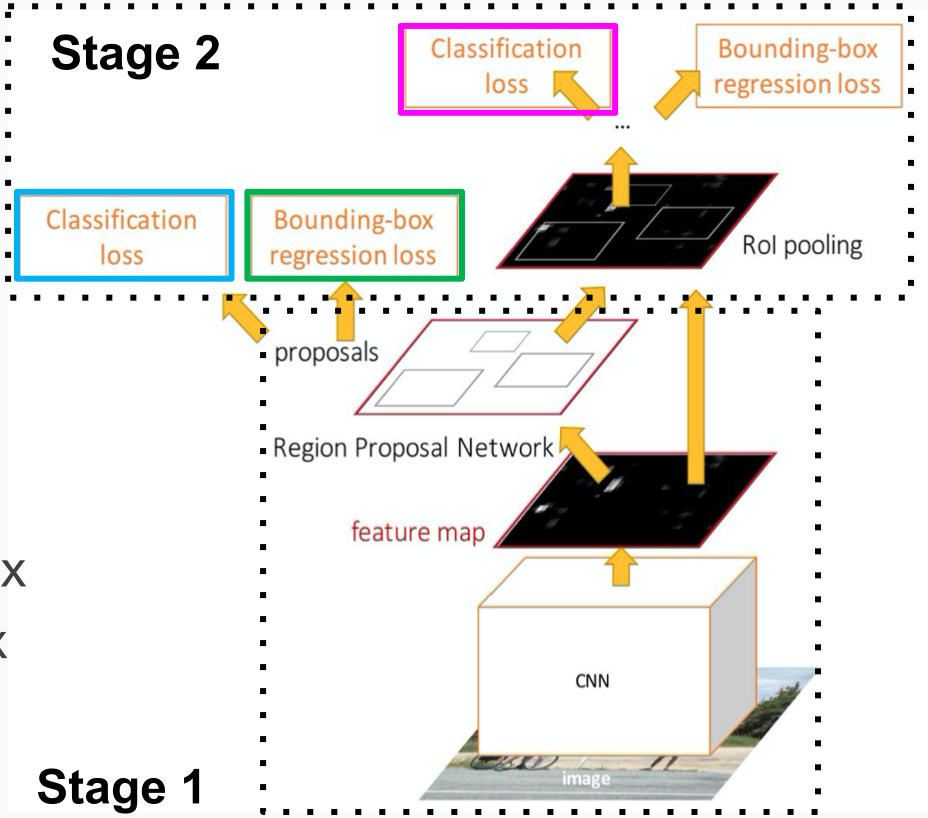




Adapted from Fei-Fei Li & Justin Johnson & Serena Yeung Stanford CS231n 2019 “Convolutional Neural Networks for Visual Recognition”
Ross Girshick, “Fast R-CNN” Slides 2015

The Evolution of R-CNN: R-CNN, Fast R-CNN, Faster R-CNN

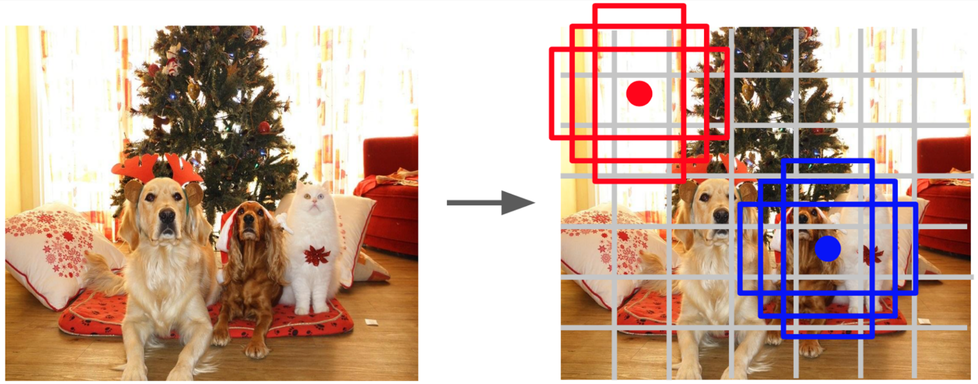
- Previously we said: “Multiple objects? We need Region Proposal Networks!”
- **Faster R-CNN is a two-stage object detector**
  - **Stage 1:** backbone network + RPN (once/image)
  - **Stage 2:** crop - predict object & bbox (once/region)
- What is our RPN again?
- RPN runs prediction on many many anchor boxes:
  - **Loss 1:** Tells is does the anchor bbox contain an object
  - **Loss 2:** For the top 300 boxes its adjusts the box
- What is the difference between our 2 classification losses?
  - one is classifying **object** (i.e. object/not object) – green box
  - one is classifying specific **categories** (e.g. dog) – pink box
- Do we really need two stages?



Adapted from Fei-Fei Li & Justin Johnson & Serena Yeung Stanford CS231n 2019 “Convolutional Neural Networks for Visual Recognition”
Ross Girshick, “Fast R-CNN” Slides 2015

Single-Stage Detection Without Region Proposals: YOLO, SSD

- Within each **NxN** grid, regress over each **B** base boxes, predict: (x,y,h,w, confidence = 5)
- **Predict C** category specific class scores
    - Output : N x N x S (5 B + C)
- YOLOv3 (Joseph Redmon):
    - predicts at 3 scales, S = 3
    - predicts 3 boxes at each scale, B=3
    - Darknet-53 as feature extractor (similar to ResNet 152, and 2x faster!)

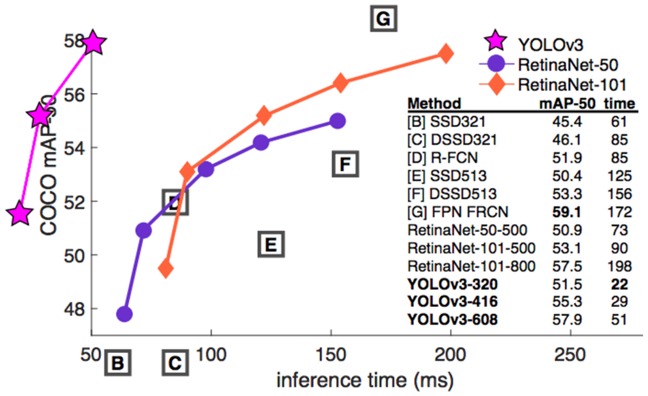


Input image
3 x H x W

Divide image into grid
7 x 7

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

Redmon et al, "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016
Liu et al, "SSD: Single-Shot MultiBox Detector", ECCV 2016
Lin et al, "Focal Loss for Dense Object Detection", ICCV 2017



(YOLO) Redmon, "You Only Look Once: Unified, Real-Time Object Detection" CVPR 2015: Cited by 8057 (link)

Fei-Fei Li & Justin Johnson & Serena Yeung Stanford CS231n 2019 "Convolutional Neural Networks for Visual Recognition"

Semantic Segmentation

Semantic Segmentation: Classify Each Pixel

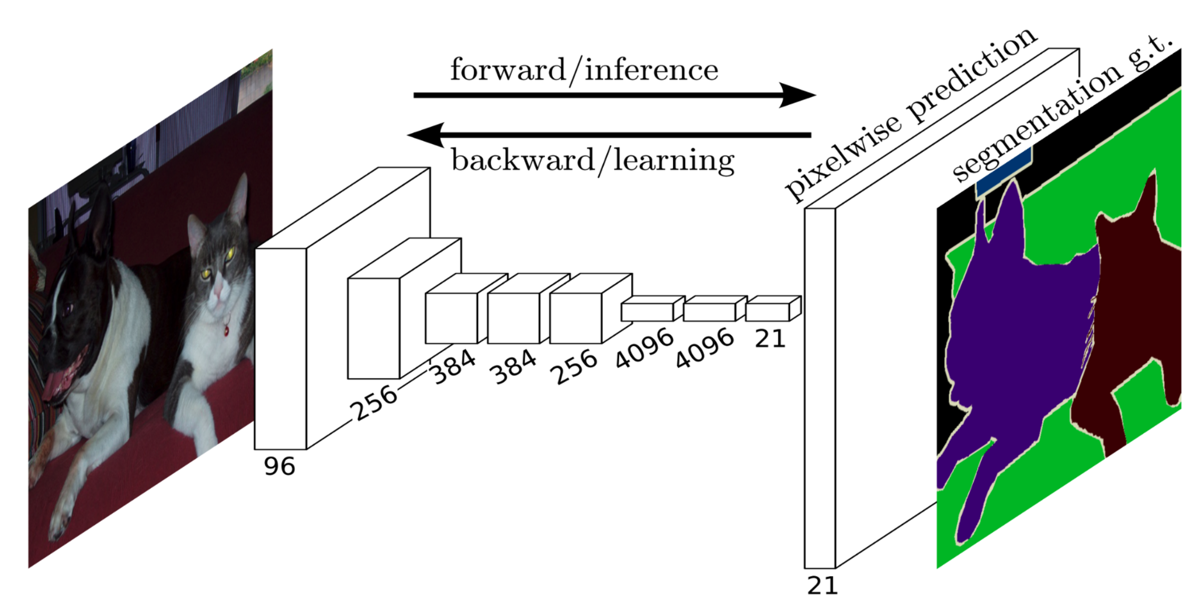
- Fully-Convolutional Networks
- SegNet & U-NET
- Faster R-CNN linked to Semantic Segmentation: Mask R-CNN

Semantic Segmentation: Classify Every Pixel

- **Image Classification:** assigning a single label to **the entire picture**
- **Semantic Segmentation:** assigning a semantically meaningful **label to every pixel in the image**

So, our output shouldn't be a class prediction (C numbers) but a picture (C x *w* x *h*)

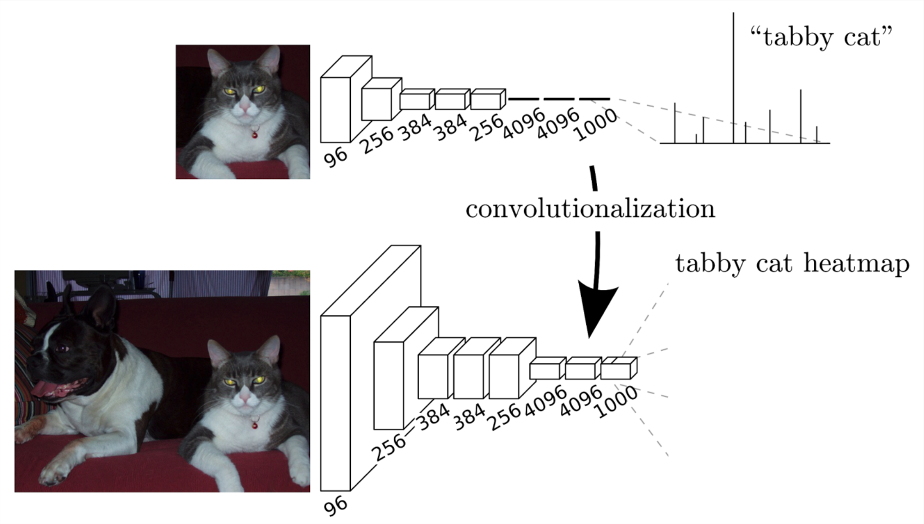
- Can we have a network for each pixel location?
- Sliding window inputs of patches predicting the class of the pixel in the center?
- Many forward passes! Not reusing overlapping patches and features.



(FCN) Long, Shelhamer et al. "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015: Cited by 14480 (link)

Fully-Convolutional Networks

- Semantic segmentation: assigning a semantically meaningful label to every pixel in the image
- So our output shouldn't be a classification prediction (C numbers) but a picture (C x w x h)
  - Maybe we can have a network for each pixel location? Many (w times h) networks!
  - Sliding window inputs of patches predicting the class of the pixel in the center? Many forward passes! Overlapping features not used.
- Solution: FCN = Fully-Convolutional Networks! (not fully-connected)
  - 1 network - 1 prediction would be a lot better
  - Why convolutions? every pixel is very much influenced by its neighborhood



(FCN) Long, Shelhamer et al. "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015: Cited by 14480 (link)

Fig: top, Image Classification (FC), bottom, Image Segmentation (FCN)

26

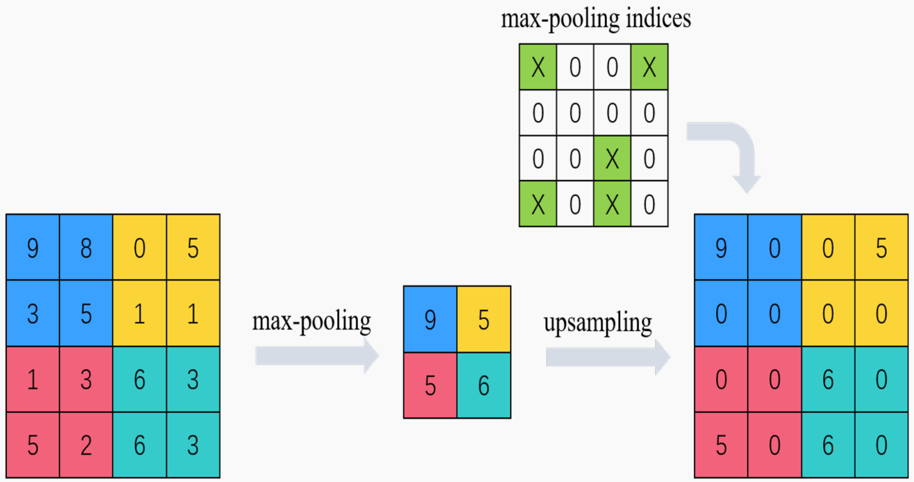
Fully-Convolutional Networks

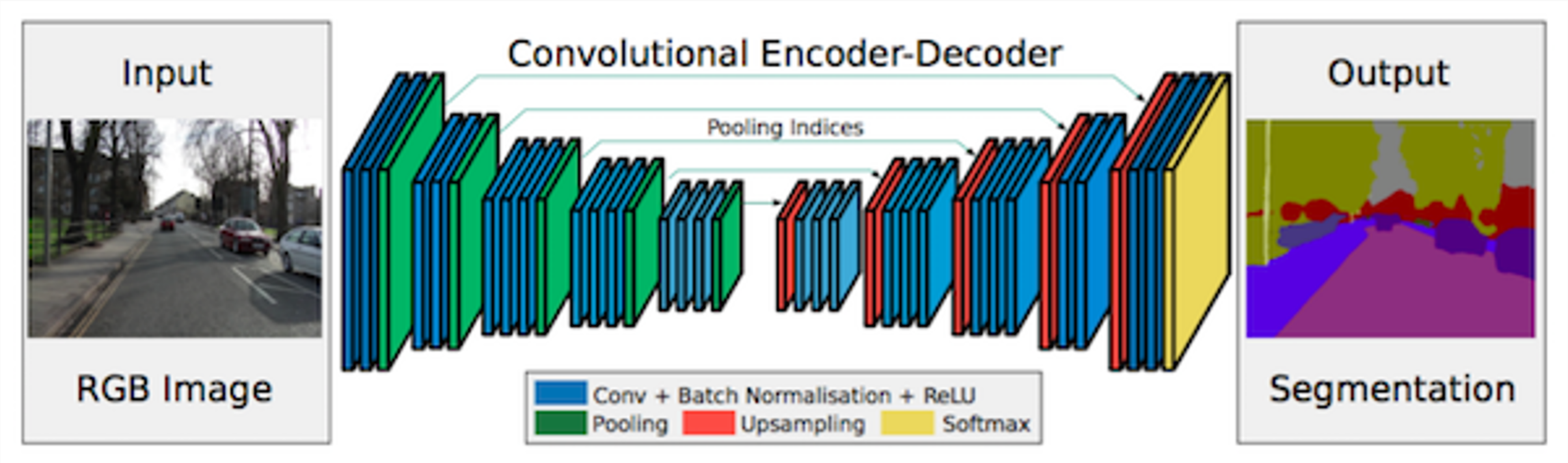
- **FCN:** design a network as a bunch of conv layers to make predictions for all pixels all at once.
  - **Encoder** (= Localization)**:** **downsample** through **convolutions.** Reduces number of params (bottleneck), can make network deeper
  - **Decoder** (= Segmentation): **upsampled** through **transposed convolutions**
  - **Loss:** cross-entropy loss on every pixel.
- **Contribution:**
  - Popularize the use of end-to-end CNNs for semantic segmentation;
  - Re-purpose imagenet pretrained networks for segmentation = Transfer Learning
  - Upsample using transposed layers.
- **Negative:**
  - upsampling = loss of information during pooling;
  - 224x224 image downsampled to 20x20 back upsampled to 224x224.

(FCN) Long, Shelhamer et al. "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015: Cited by 14480 (link)

SegNet

- The indices from max pooling down sampling are transferred to the decoder: **pooling indices**
- Improves fine segmentation resolution, we want “pixel-perfect”;
- More efficient since no transposed convolutions to learn.

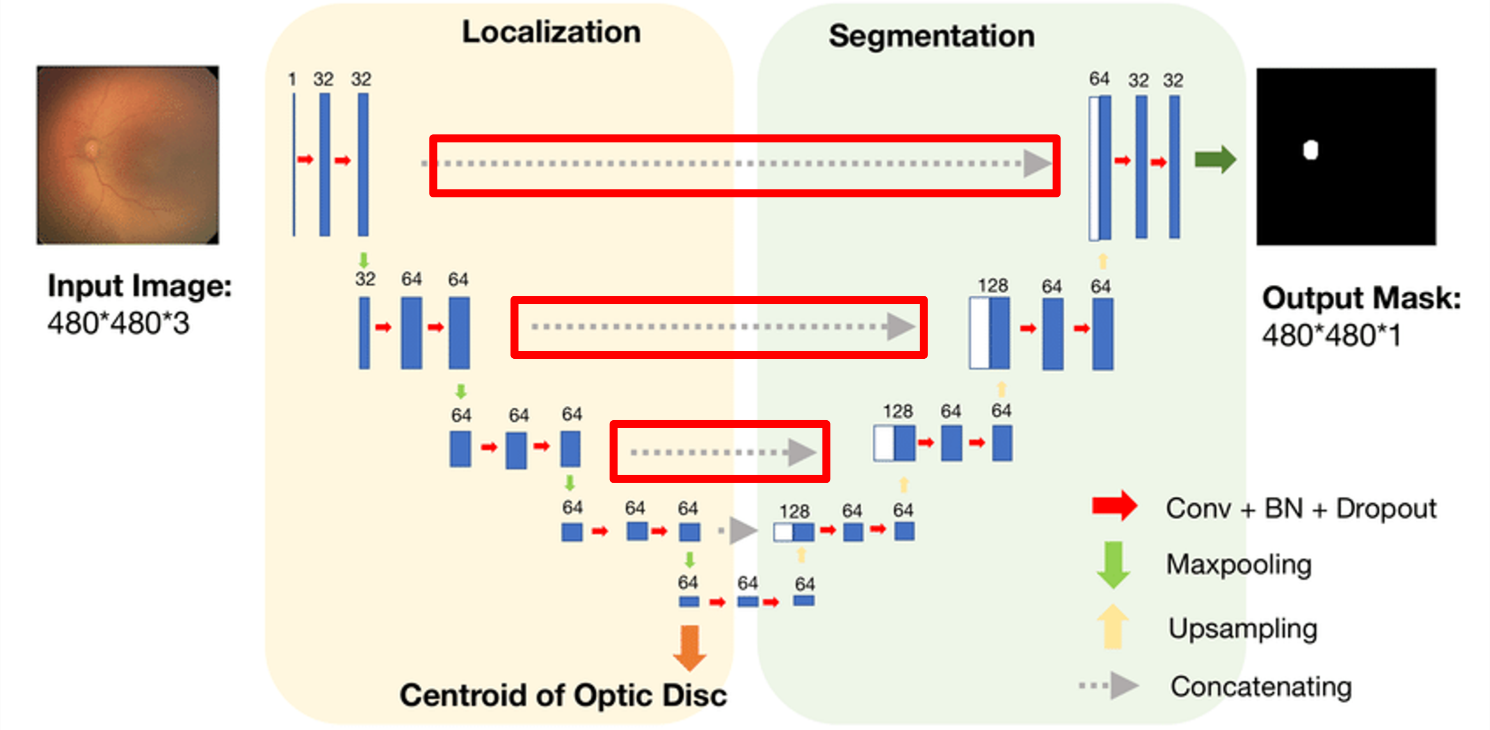




SegNet: A deep Convolutional Encoder-Decoder Architecture for Image Segmentation. ([link]())

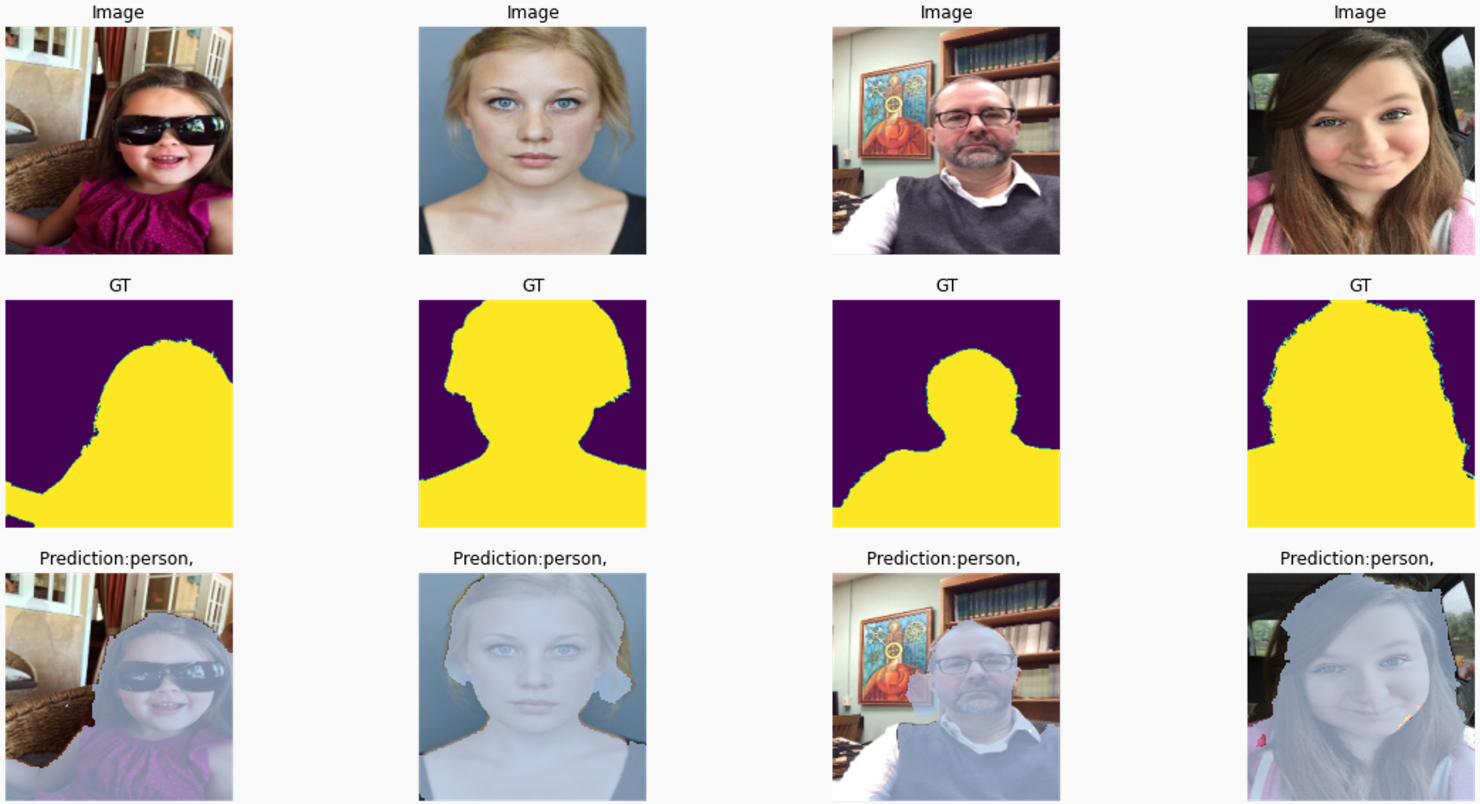
U-NET: Long Skip Connections

- The U-Net is an encoder decoder using:
  - **location information** from the down sampling path of the encoder;
  - **contextual information** in the up sampling path by the "concatenating" long-skip connections.



Tutorial: Using Transfer Learning to train a U-NET

[Colab Notebook]



References

Presentations:

- Fei-Fei Li & Justin Johnson & Serena Yeung Stanford CS231n 2019/2018 “Conv. Neural Networks for Visual Recognition” Lecture 12 !
    - BTW: Great course / youtube series (youtube 2017)
- Ross Girshick, “Fast R-CNN” Slides 2015 (link)

Papers:

- **VGG** Simonyan, Zisserman. “Very Deep CNNs for Large-scale Image Recognition”, ILSVRC 2014: Cited by 34652 (link)
- **Select. Search** Uijlings et al, Selective Search for Object Recognition” IJCV 2013: Cited by 3944 (link)
- **R-CNN** Girshick et al, “Rich feature hierarchies for accurate object detect. & sem. segmentation” CVPR2014: Cited by 12000 (link)
- **Fast-R-CNN** Girshick, ‘Fast R-CNN“ ICCV 2015: Cited by 8791 (link)
- **Faster- R-CNN** Ren et al, “Faster R-CNN: Real-Time Object Det. with Region Proposal Networks” NEURIPS 2015 Cited by 16688 (link)
- **Mask-R-CNN** He et al, “Mask R-CNN” ICCV 2017: Cited by 5297 (link)
- **YOLO** Redmon, “You Only Look Once: Unified, Real-Time Object Detection” CVPR 2015: Cited by 8057 (link)
- **FCN** Long, Shelhamer et al. “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015: Cited by 14480 (link)
- **SegNet** Badrinarayanan et al. “SegNet: A deep Conv Encoder-Decoder Architecture for Image Segmentation”. Cited by 4258 (link)
- **U-Net** Ronneberger et al. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. Cited by 12238 (link)