Deep Learning for Segmentation and Detection

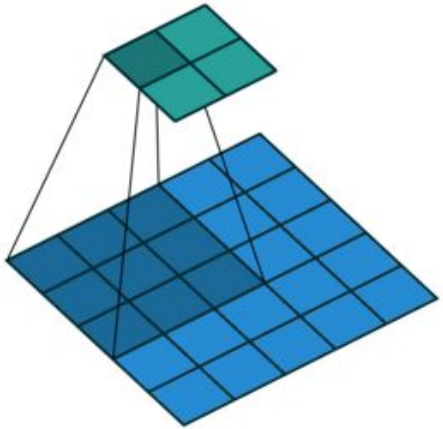
Ayush Thakur

Chair, IEEE EDS Student Branch Chapter
CTO, The Code Foundation

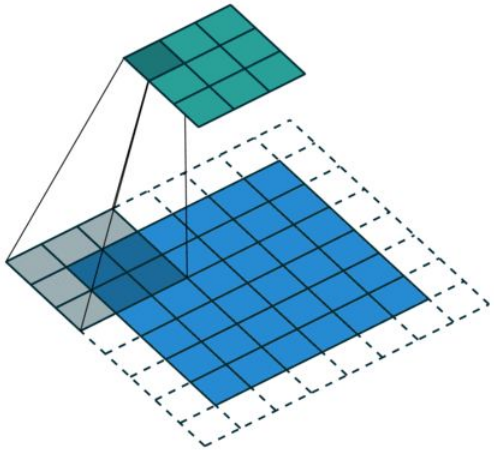
Talk Overview

1) Overview on Standard CNN
2) Image Classification Task
3) Beyond Classification Task
4) Localization Task
5) Segmentation Task
6) Fully Convolutional Network
7) Detection Task

Overview on Standard CNN



- Input Volume = 5
- Padding = 0
- Stride = 2
- Kernel size = 3



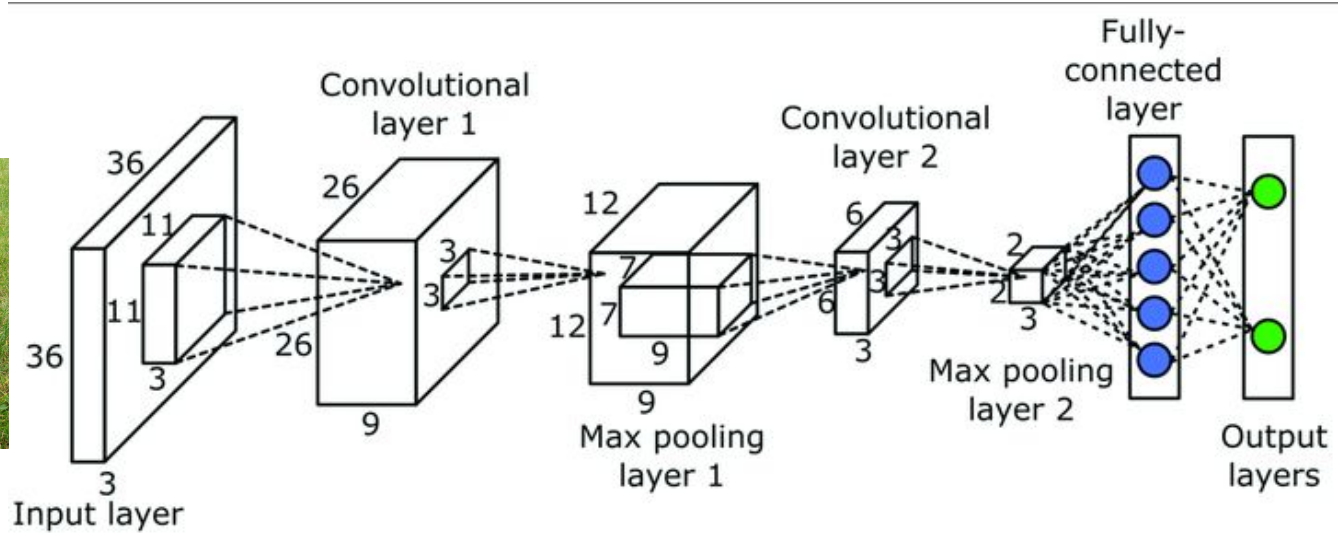
- Input Volume = 6
- Padding = 1
- Stride = 2
- Kernel size = 4

$$\left(\frac{n+2p-f}{s}+1, \frac{n+2p-f}{s}+1, n_c\right)$$

Where,

n = Input Volume
p = Padding
s = Stride
f = Kernel Size
n_c = Number of channels in input volume

Image Classification



Dog = 0.96

Cat = 0.04

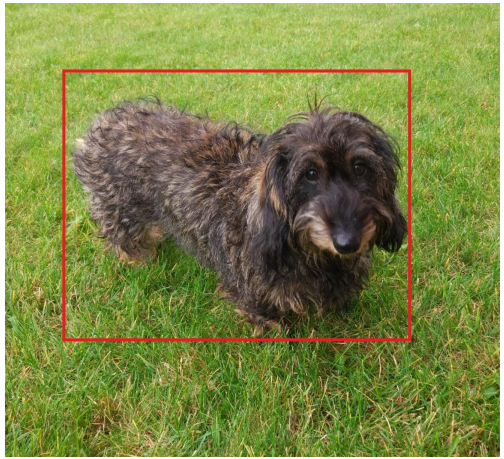
Beyond Image Classification

Limitations:

1) Mostly on centered images.
2) Not enough for real world scenarios.
3) Not getting much insights needed for autonomous systems.
4) Not utilizing the real power of convolution.

So what's more can we achieve?

Beyond Image Classification Contd.

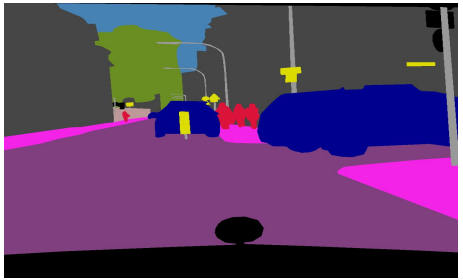
	Classification	Classification + Localization
Single Object		

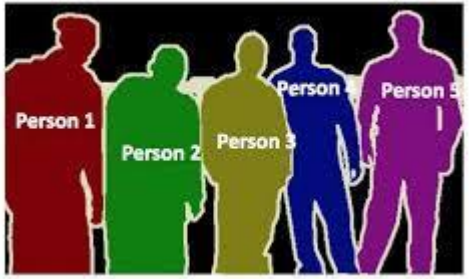
Beyond Image Classification Contd.

Multiple Objects

Object Detection

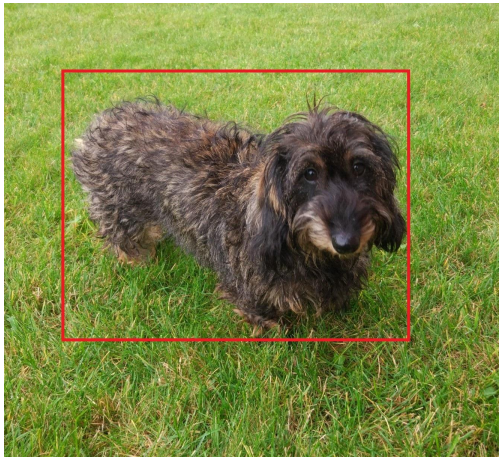
Semantic Segmentation





Instance Segmentation

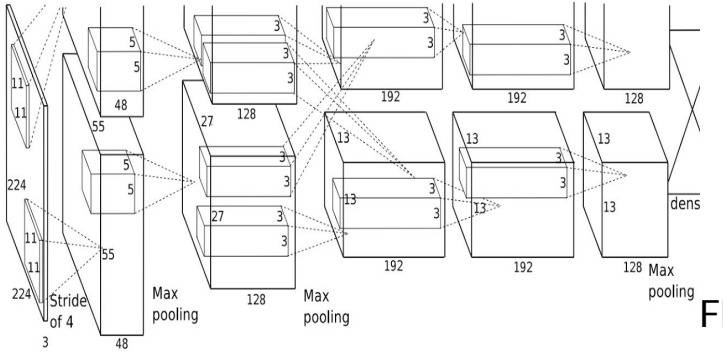
Localisation



- Single object per image
- Predict bounding box (x, y , h, w)
- Evaluate via IoU
- Treat localisation as a Regression problem

Localisation

224x224x3



Flatten

FC: 4096-1000

Class Scores:
Dog: 0.94
Cat: 0.05
Car: 0.01

Softmax Loss

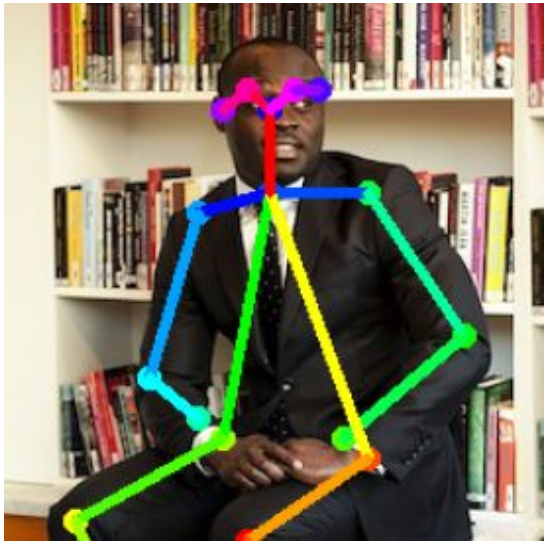
FC: 4096-4

Box Coordinates:

(x, y, h, w)

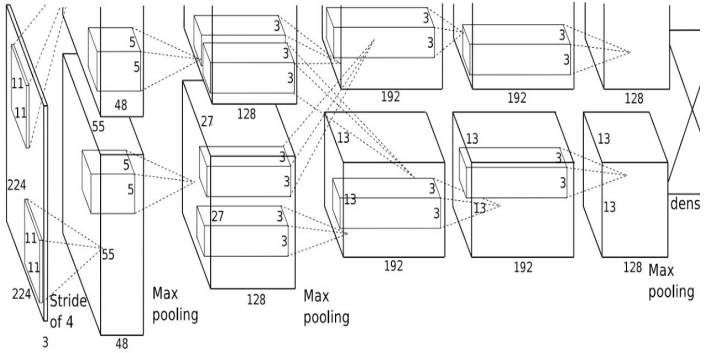
L2 Loss

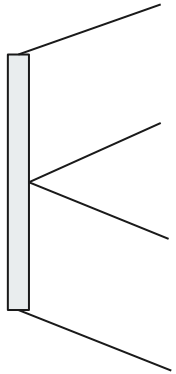
Human Pose Estimation



- Human body can be represented with 14 joints.
- Instead of four outputs of the bounding box.
- The regression layer will output 14 coordinates.
- L2 loss

Human Pose Estimation





L2 Loss

Left Foot: (x1, y1)

……

……

Right upper arm: (x6, y6)

……

……

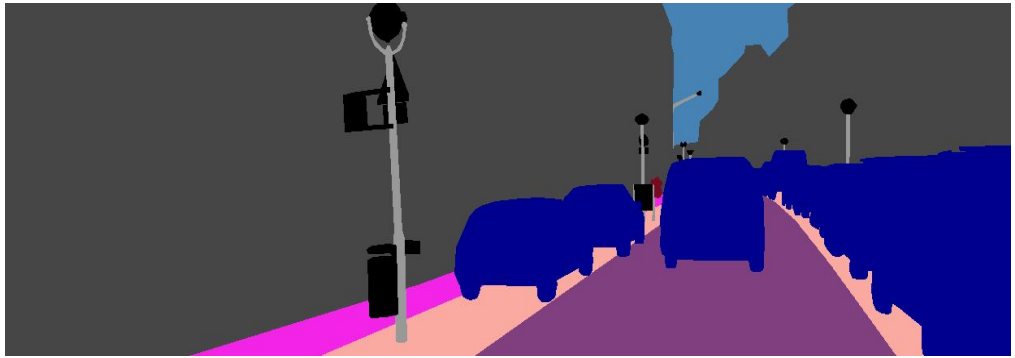
Left Shoulder : (x10, y10)

……

……

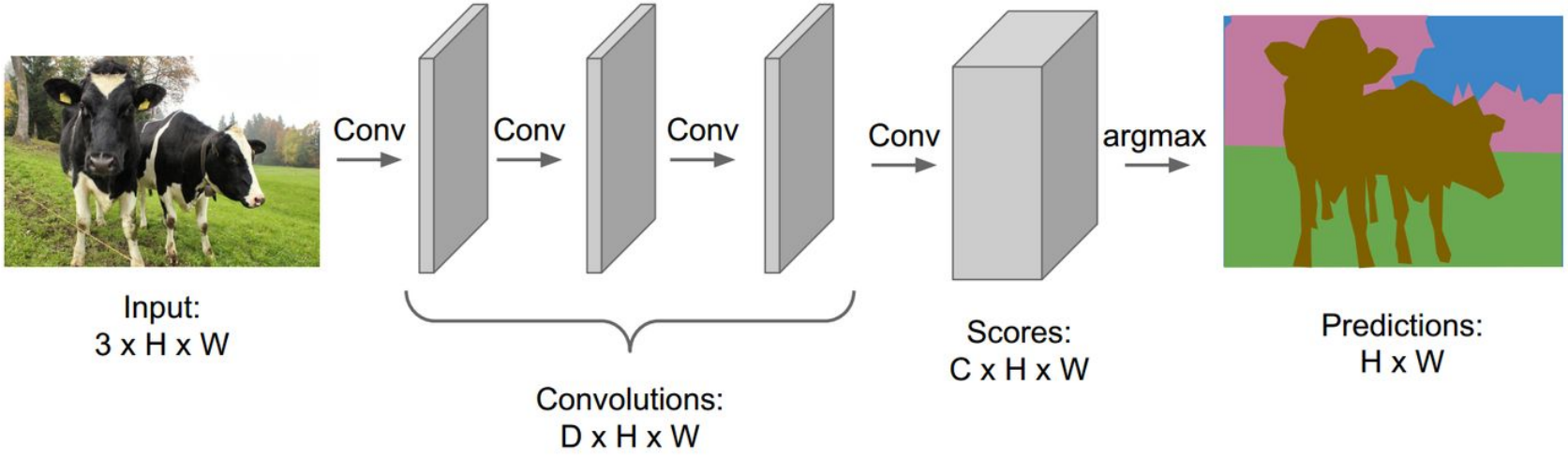
Head: (x14, y14)

Semantic Segmentation

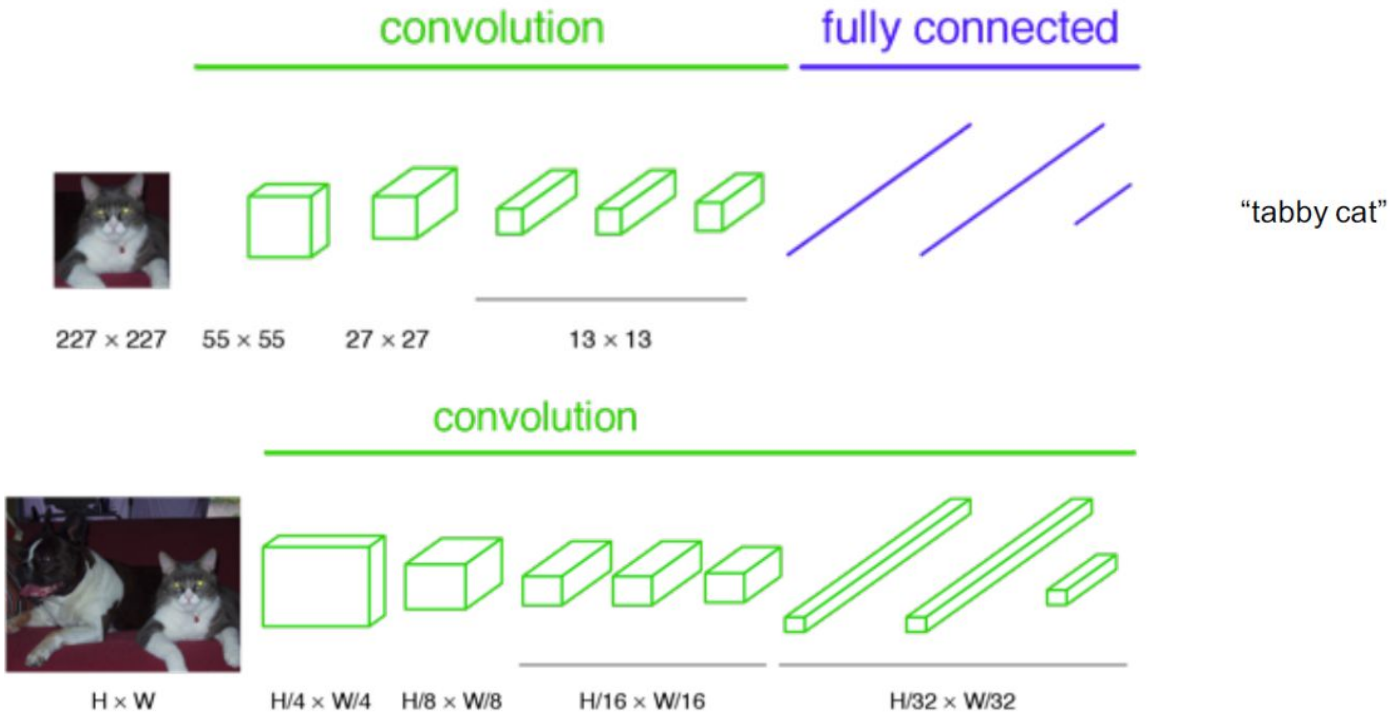


- Label each pixel in the image with a category label.
- Don’t differentiate instances, but care about pixels.
- Labels: Building, vehicles, lamp posts, road, pavement, etc.
- Output a class map for each pixels.

Semantic Segmentation

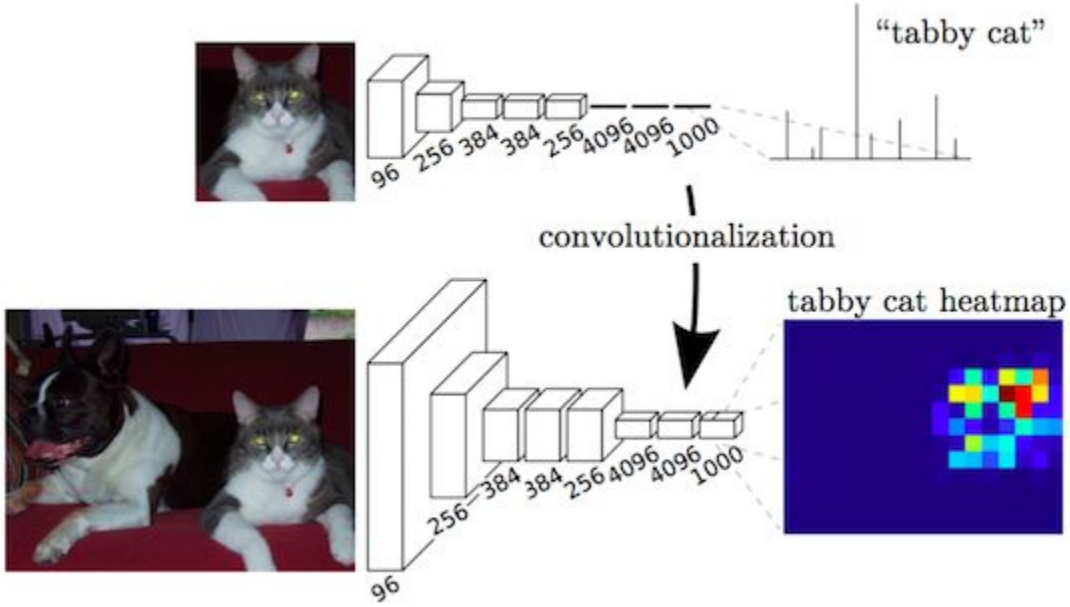


Fully Convolutional Network

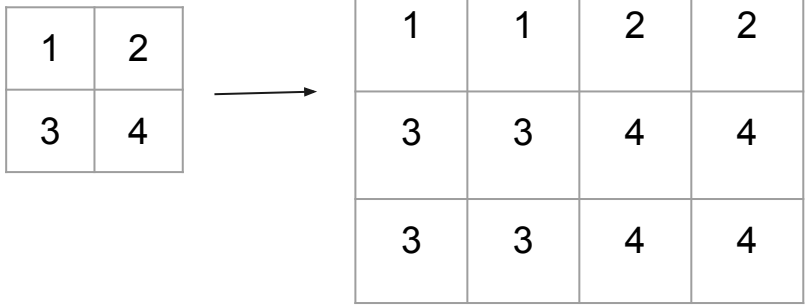


"tabby cat"

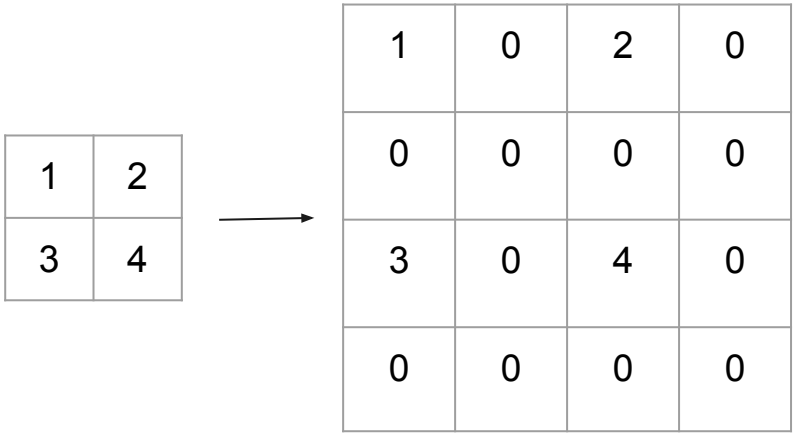
FCN Contd.


"tabby cat"
96
256
384
384
256
4096
4096
1000
convolutionalization
tabby cat heatmap
96
256
384
384
256
4096
4096
1000

Upsampling/Unpooling



Nearest Neighbour



Bell of Nails

Max Unpooling

Max Pooling

1	2	4	2
5	3	1	0
2	4	2	0
3	1	3	1

→

5	4
4	3

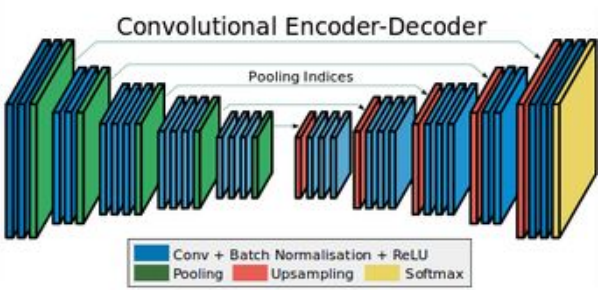
..Rest of the N/W..

1	2
3	4

→

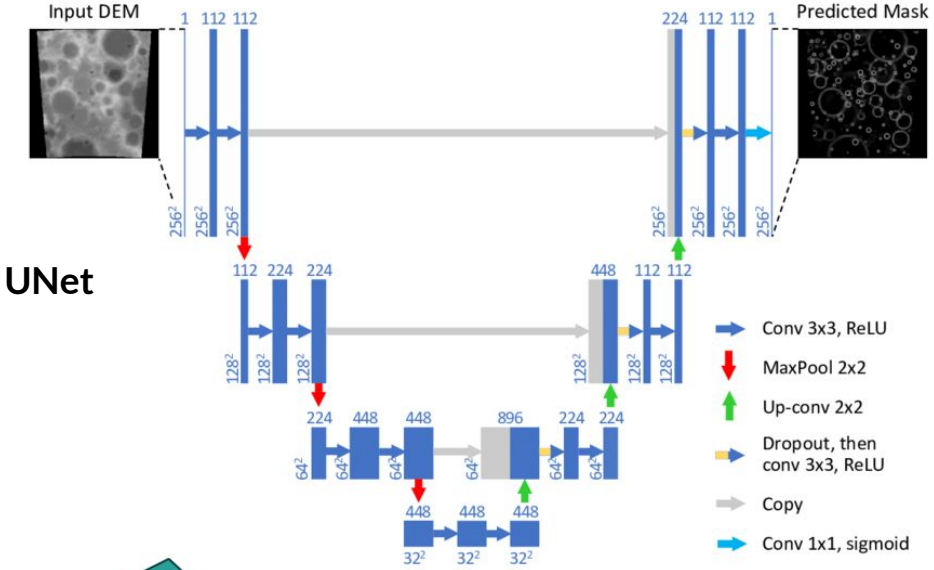
Max Unpooling

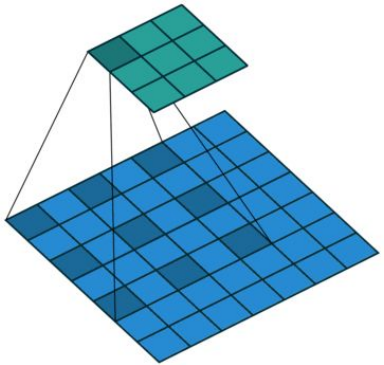
0	0	2	0
1	0	0	0
0	3	0	0
0	0	4	0



Various Architecture

1. FCN
2. SegNet
3. Dilated Convolutions
4. DeepLab (v1 & v2)
5. RefineNet
6. PSPNet
7. Large Kernel Matters
8. DeepLab v3



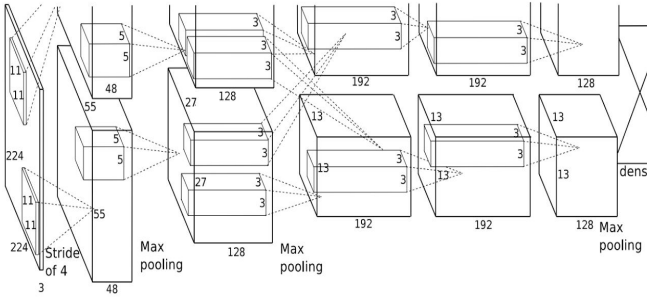


Dilated Convolution

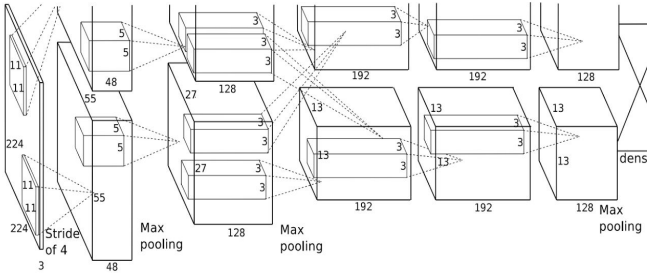
Object Detection

- One of the core problem in computer vision.
- Classical methods used something like Haar Cascade.
- Task is to draw a bounding box to every category which appears in an input image.
- Unlike classification + Localisation, have no idea about the number of objects in an image.

Object Detection as Regression?



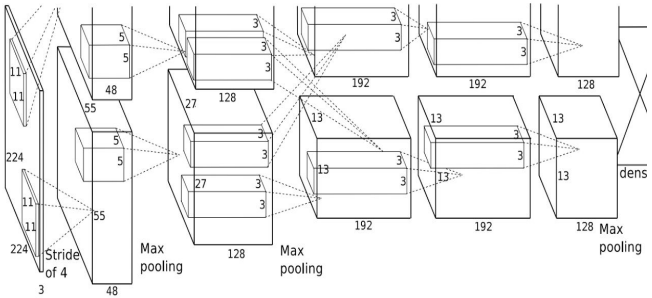
DOG: (x, y, h, w)



Car1: (x1, y1, h1, w1)
Car2: (x2, y2, h2, w2)
.
.
Truck1: (x', y', h', w')

Object Detection as Sliding Window

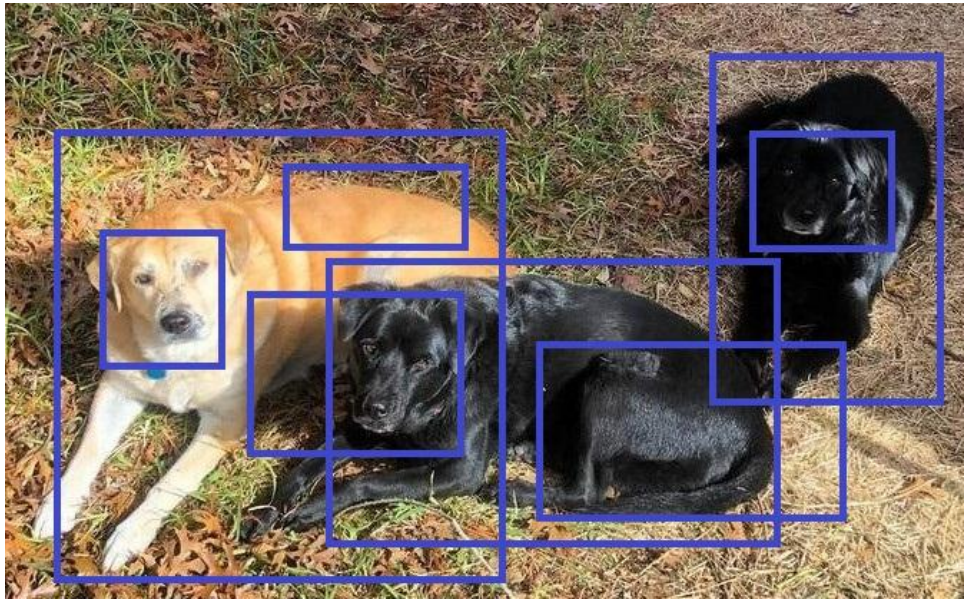
Apply CNN to different crops of image.



DOG ? Yes
Background? No

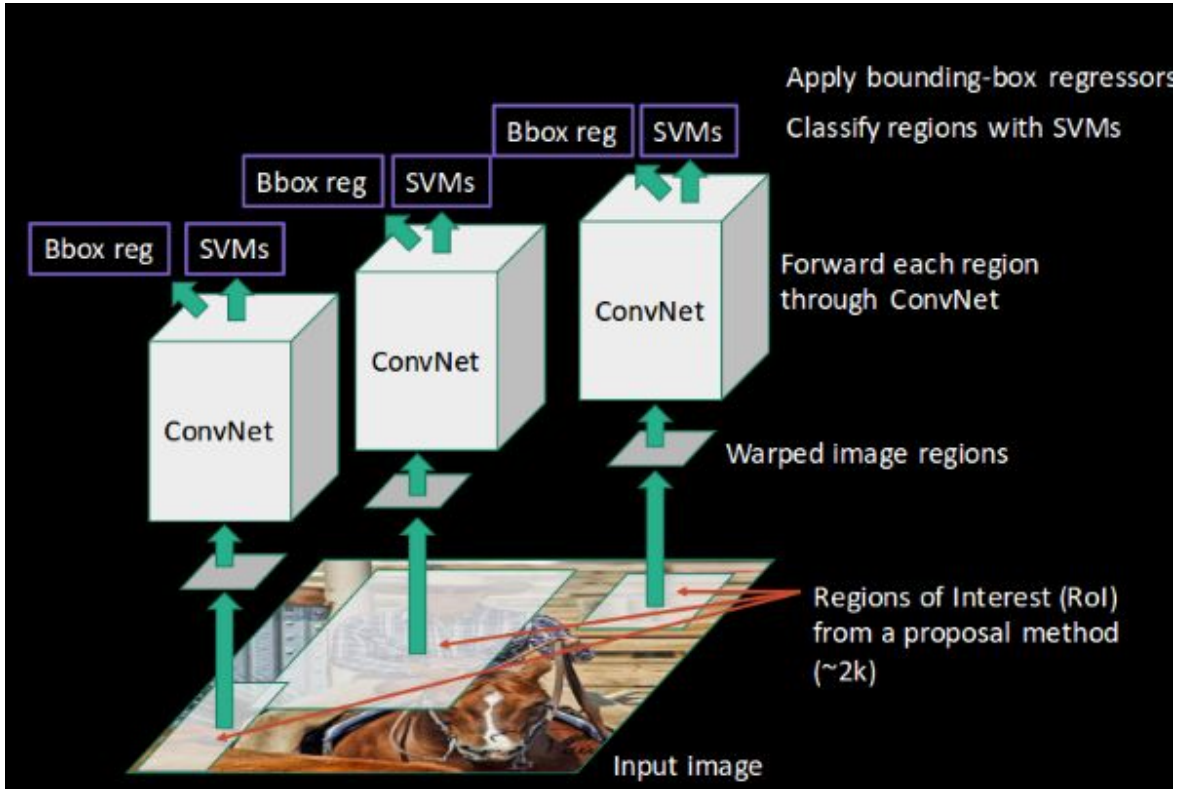
- So many crops.
- Computationally expensive
- What will be the size of crop?

Region Proposals



- Find image regions that are likely to contain objects.
- Number of regions as output can be set as a parameter.
- Classical image processing techniques used.
- Faster than sliding window approach.

R-CNN


Apply bounding-box regressors
Classify regions with SVMs
Bbox reg
SVMs
Bbox reg
SVMs
Bbox reg
SVMs
ConvNet
ConvNet
ConvNet
Forward each region through ConvNet
Warped image regions
Regions of Interest (RoI) from a proposal method (~2k)
Input image

Going Ahead

- **Mask R-CNN**
- **Fast R-CNN**
- **Faster R-CNN**
- **YOLO - You Only Look Once**
- **SSD - Single Shot Detector**

- Faster R-CNN is slower but accurate.
- SSD is insanely fast but not so accurate.

Thank You

Mail to: mein2work@gmail.com

Connect: Linkedin

Find slide on github: ayulockin/talks