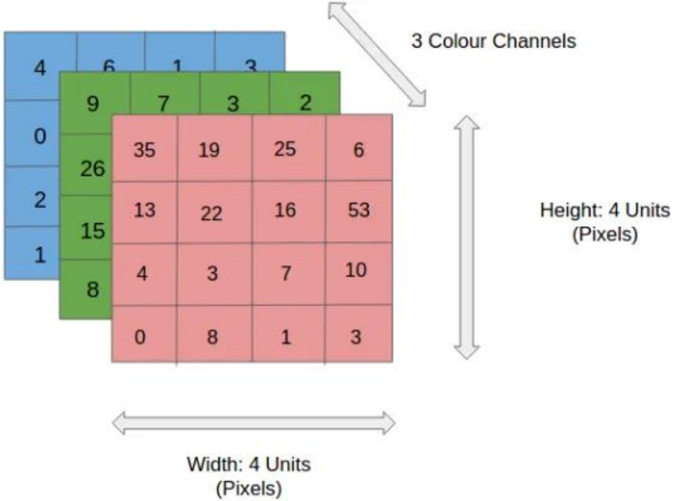
Computer Vision

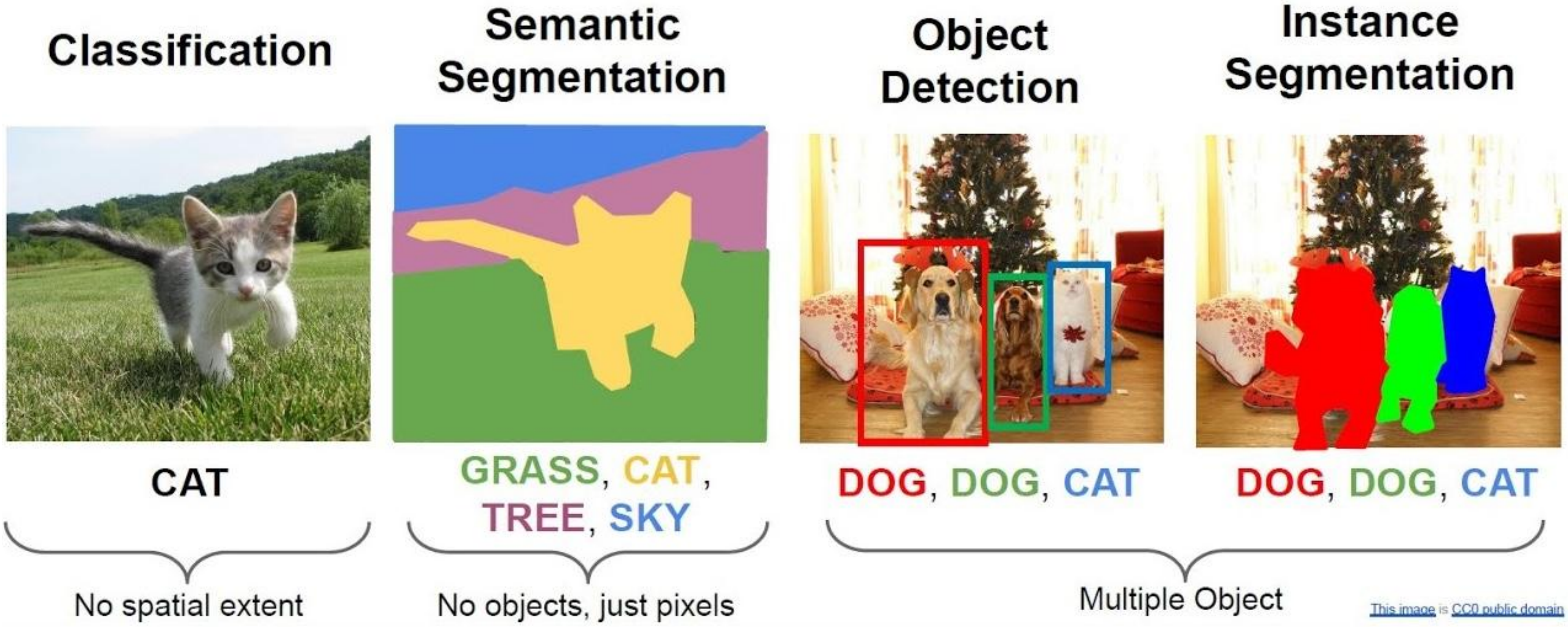
Image and Video Analysis

Introduction to Computer Vision



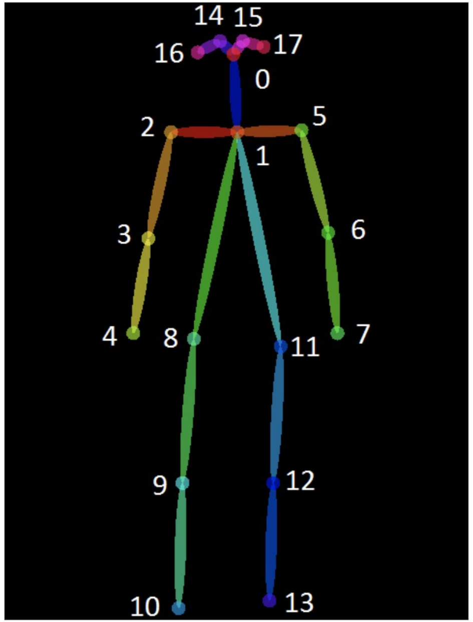
- Automate the derivation of useful information from digital images
- Types of image data:
    - RGB images
    - Multiple images from different cameras
    - Video sequences
    - 3D point clouds (LiDaR)

Computer Vision Tasks

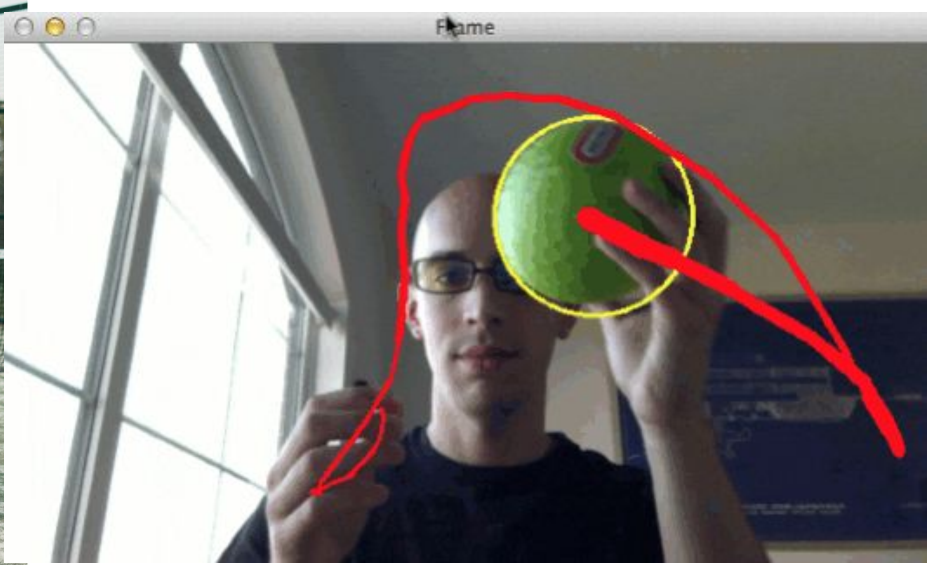


This image is CC0 public domain

Computer Vision Tasks



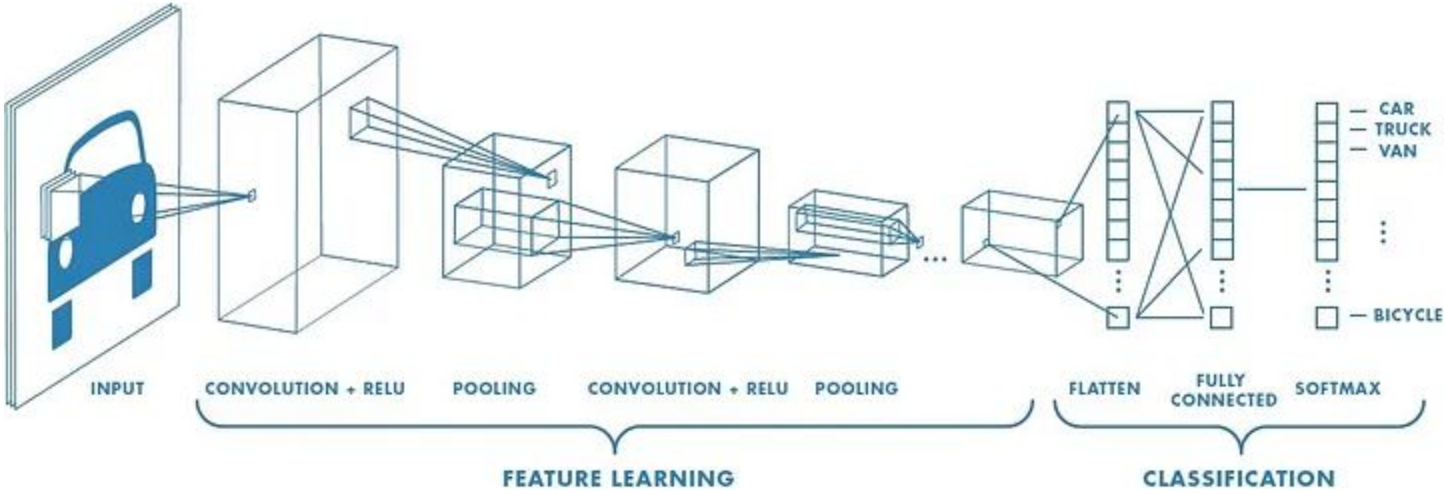
Pose estimation
Activity recognition



Object tracking

Convolutional Neural Networks

Architecture Overview



Why do we need a specialized architecture for images?

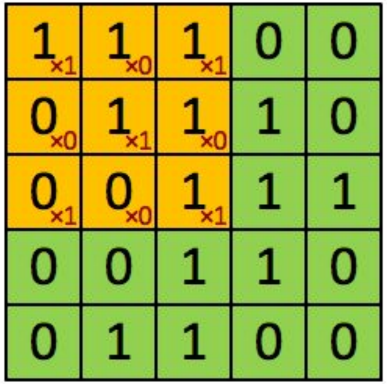
- Fully connected layers do not scale well for images (WxHx3)
    - Leads to massive numbers of parameters
    - AlexNet example
        - 227x227x3 image -> 55x55x96 feature space (11x11x3 kernels)
            - 364 parameters/kernel * 96 kernels = 34,944 parameters
        - FC network would connect each of these neurons: ~44 billion connections
- Fully connected layers do not exhibit translation invariance
    - FC feature detection depends on the location within an image, due to the way the neurons are wired
    - Want to capture features regardless of the location in the image

Convolution

- “Sweep” a kernel over an image to detect features

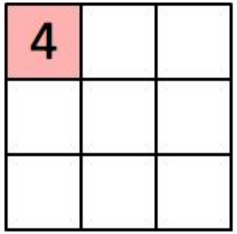
$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Kernel / Filter



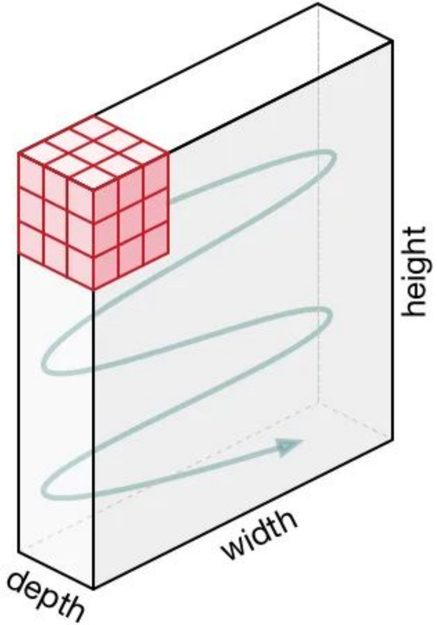
$1_{\times 1}$	$1_{\times 0}$	$1_{\times 1}$	0	0
$0_{\times 0}$	$1_{\times 1}$	$1_{\times 0}$	1	0
$0_{\times 1}$	$0_{\times 0}$	$1_{\times 1}$	1	1
0	0	1	1	0
0	1	1	0	0

Image

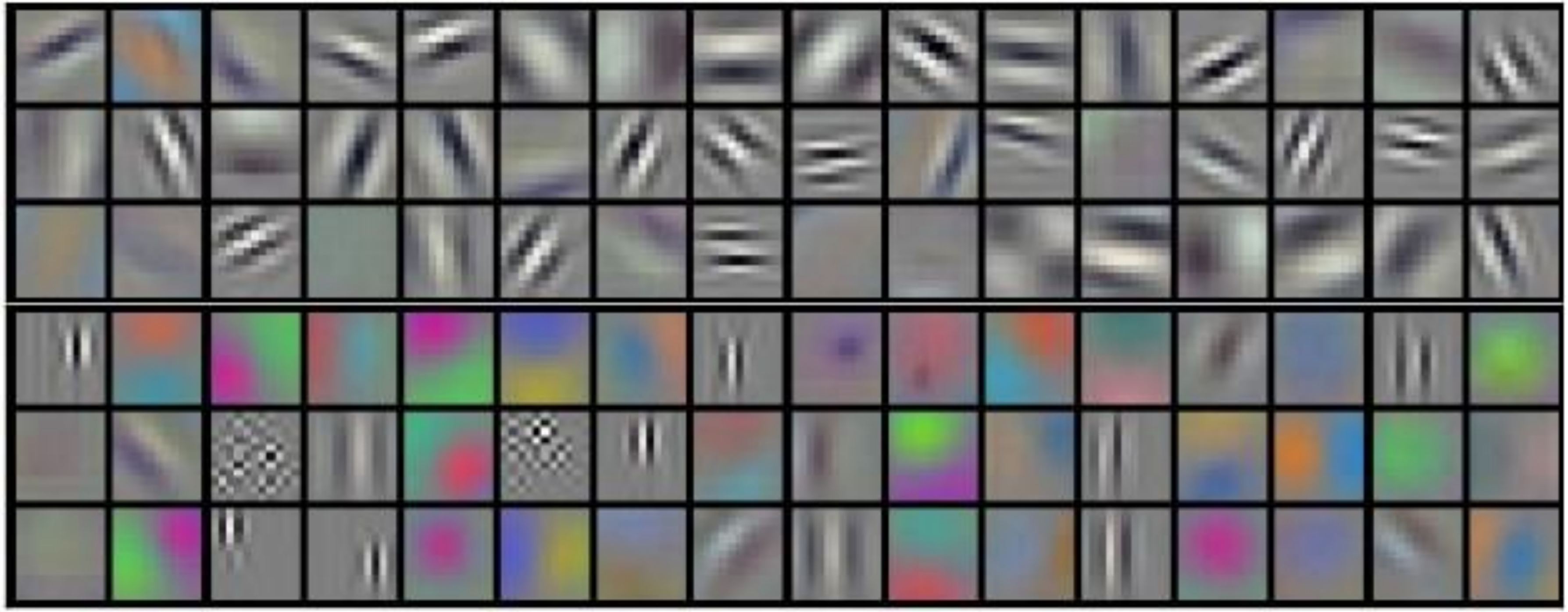


4		

Convolved Feature

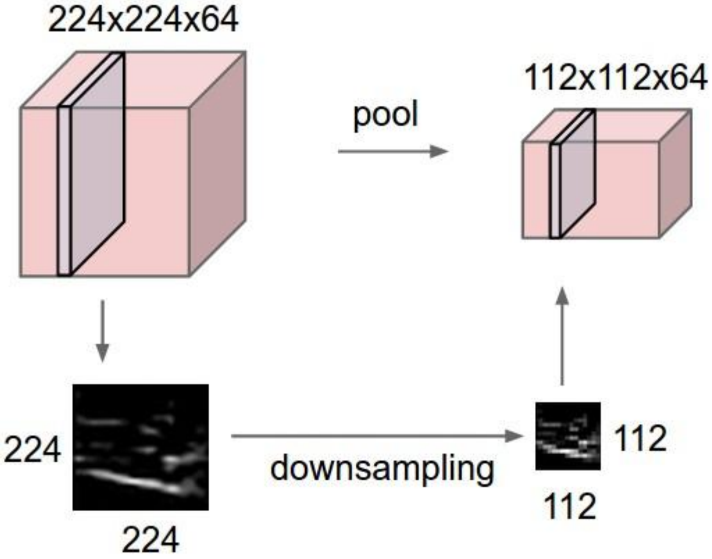


Learned Filter Bank, ImageNet Krizhevsky et al.



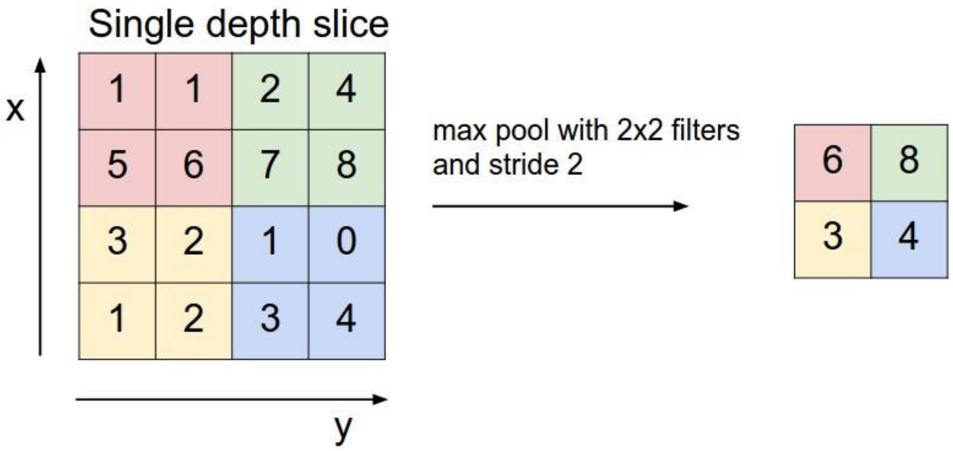
Pooling Operations

- Feature maps have very high dimension
- Pooling operations reduce the dimensionality before further convolutional/FC layers



Pooling Operations

- Max pool
  - Take the maximum value within the filter
- Average pool
  - Take the average value within the filter



Single depth slice

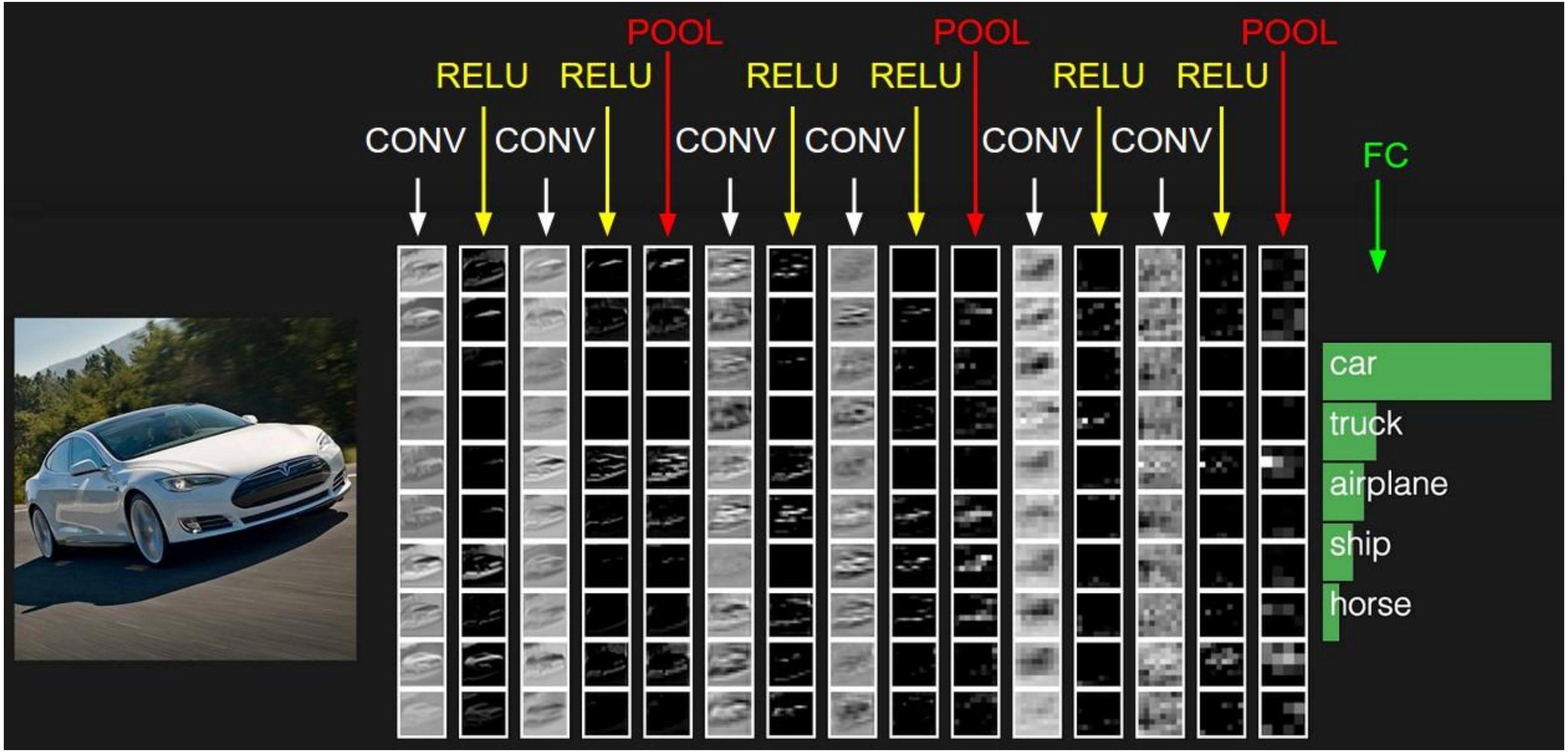
1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

max pool with 2x2 filters and stride 2

6	8
3	4

CNN Intuition

- As we move forward in the CNN, activations go from low level features to high level semantic information
  - The first convolutional layer may learn filters that represent points and lines
    - High spatial resolution, low semantic content
  - At later stages, lower spatial resolution, high semantic content
- The final pooling layer leads into a fully-connected neural network for the final prediction

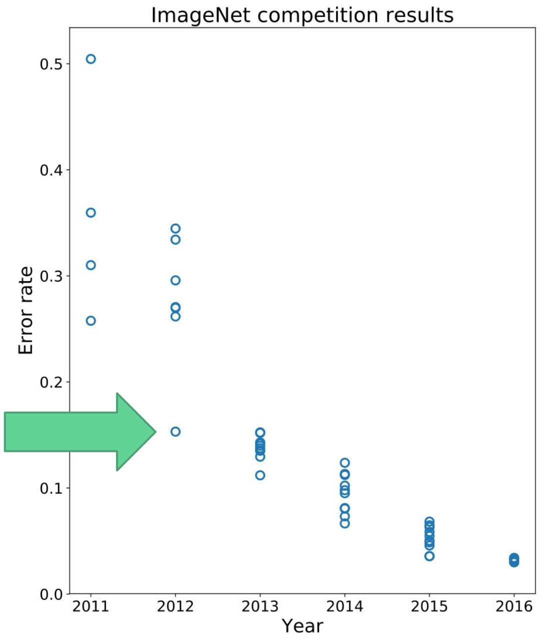


High spatial resolution
at early layers

Reduced dimensionality, learned
a compressed representation

ImageNet Challenge

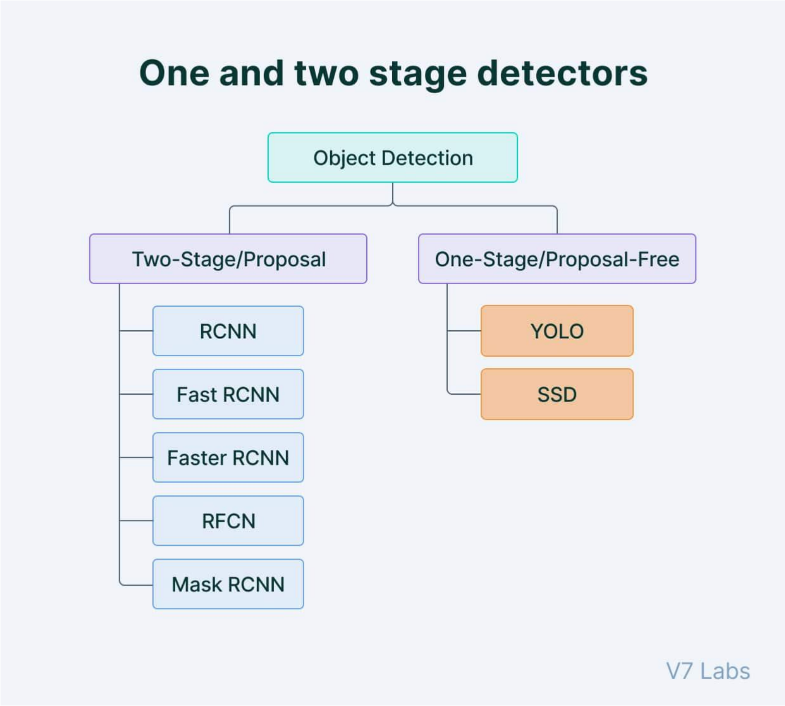
- Compete to obtain best top-5 test error rate when classifying an image into 1 of 1000 categories
- AlexNet, 2012
    - Convolutional layers
    - ReLU activation function



AlexNet

Some Other CV DNNs

Object Detection



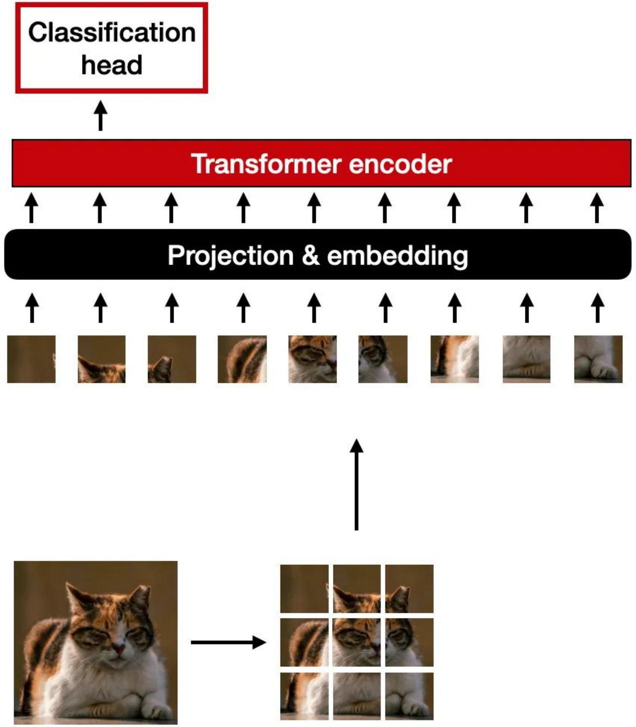
Single-shot object detection:

- YOLO - You Only Look Once
- Single pass of input image to predict the location and classification of objects
- Real-time processing, low latency applications

Two-stage detection:

- Region Proposal Network suggests bounding boxes, then a CNN refines and classifies
- High accuracy

Vision Transformer (ViT)



- Taking a leaf from NLP, ViTs use a transformer architecture
- Tokenize an image
- Self-attention mechanisms model dependencies between tokens