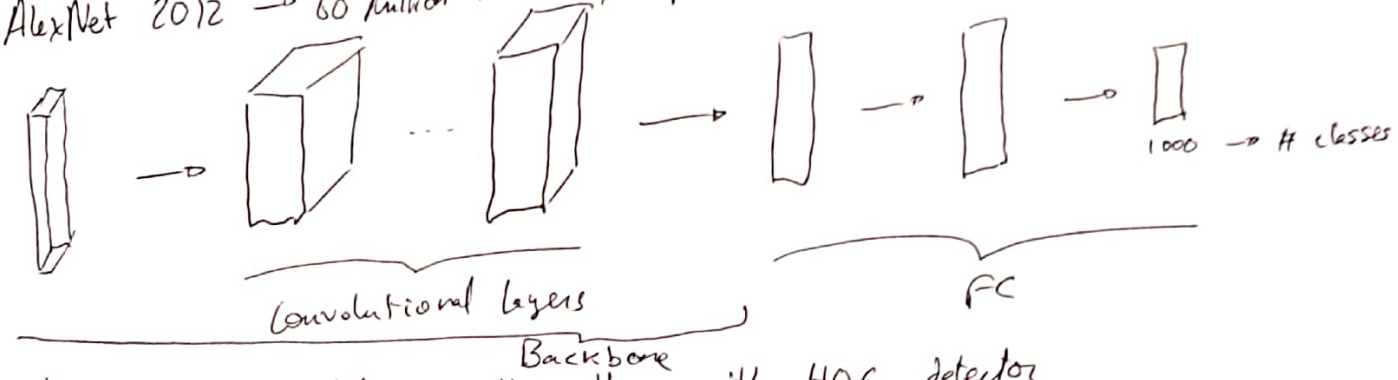
# Lecture 6 Deep Networks for object cate

Features are not hand crafted but learned autonomously

- Two-stage and one-stage NN
- State-of-the-art

AlexNet 2012 → 60 Million Parameters / Top-5 error: 16%




You need more data on this than with HOG detector

ImageNet classification challenge

- 1000 categories
- 1000 images from each category (~1 M images)
- 100 k images for testing
- The backbone is pretrained on this data set

Top-5 error (considers)
- many other stuff on the image which is not labelled
- the 5 highest probability output answers must match the expected answer

Top-1 error (conventional accuracy) → the model answer is the one with highest probability which must be exactly the expected answer

A picture of a cat is shown, and these are the outputs of the NN

- Tiger: 0.4
- Dog: 0.3
- Cat: 0.1
- Lynx: 0.09
- Lion: 0.08
- Bird: 0.02
- Bear: 0.01

From Top-1 approach output would be wrong since it is a tiger and not a cat

Top-5 → is cat in this top 5? → output correct in Top-5 error

VGG-16 $\xrightarrow{2014}$ 138 M parameters and Top-5 error: 7%

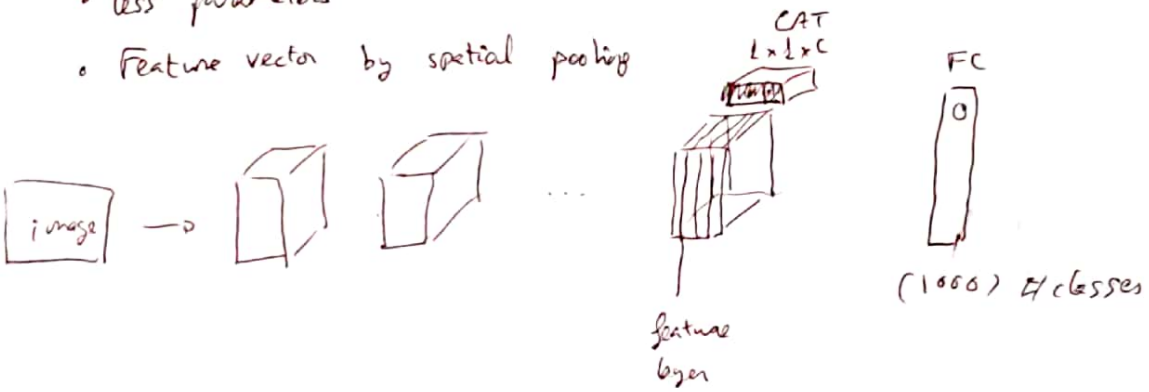
So much deeper in comparison

ResNet $\xrightarrow{2015}$ 60 M parameters (152 layers) Top-5 error: 4%

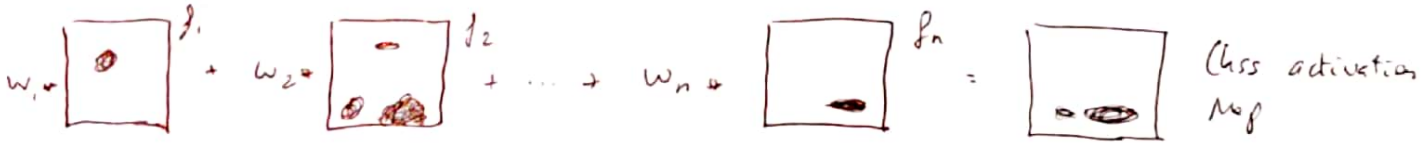
Squeeze & excitation $\xrightarrow{2017}$ 152 layers Top-5 error: 2.3%

CNNs for detection - intuition 1

- Modern classification architectures (ResNet or Inception) use convolutional layers throughout
  - No fully connected layers
  - less parameters
  - Feature vector by spatial pooling




Class Activation Mapping



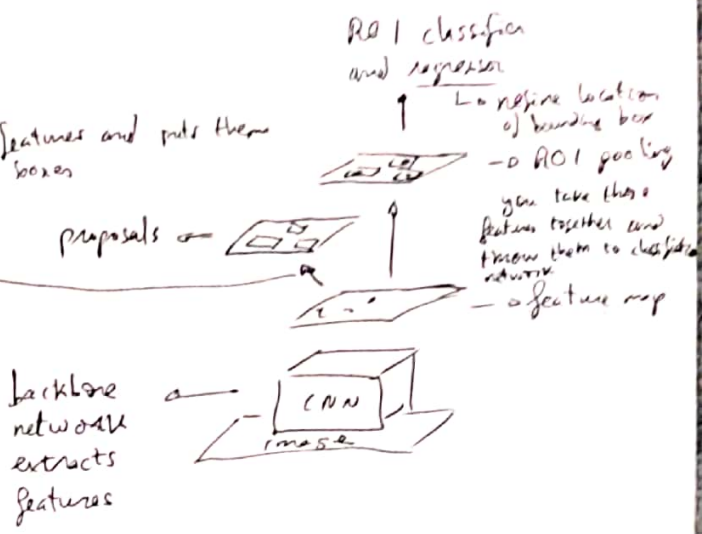

Class activation Map for a certain class

Classical object detectors

- Two stage procedure
  - Propose class agnostic regions in this image (sliding window or proposal)
  - Classify regions into object classes or background
- Can this be captured in a deep network

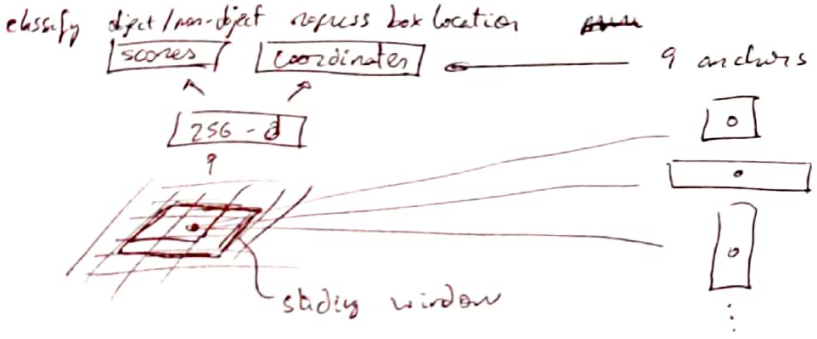
# Faster R-CNN

- Two stage system
  - Region proposal network (RPN) ← it does recognize features and puts them into bounding boxes
  - Classification / regression network
- Base network VGG16




## Region Proposal Network

- Slide a small window on feature map
- Window position provides localization with reference to the image
- Box regression provides a finer localization with reference to the window




Anchors predefine candidate regions

- Multi-scale/size anchors are used at each position: 3 scales × 3 aspect ratio yields 9 anchors
- Each anchor has its own prediction function
- Single-scale features, multi-scale predictions