Computer Vision with CNN

Behnia - Heydari

Amirkabir University of Technology

May 21, 2019

Overview

1. Motivation
2. ImageNet Challenge
3. From Cat's Brain to ResNet

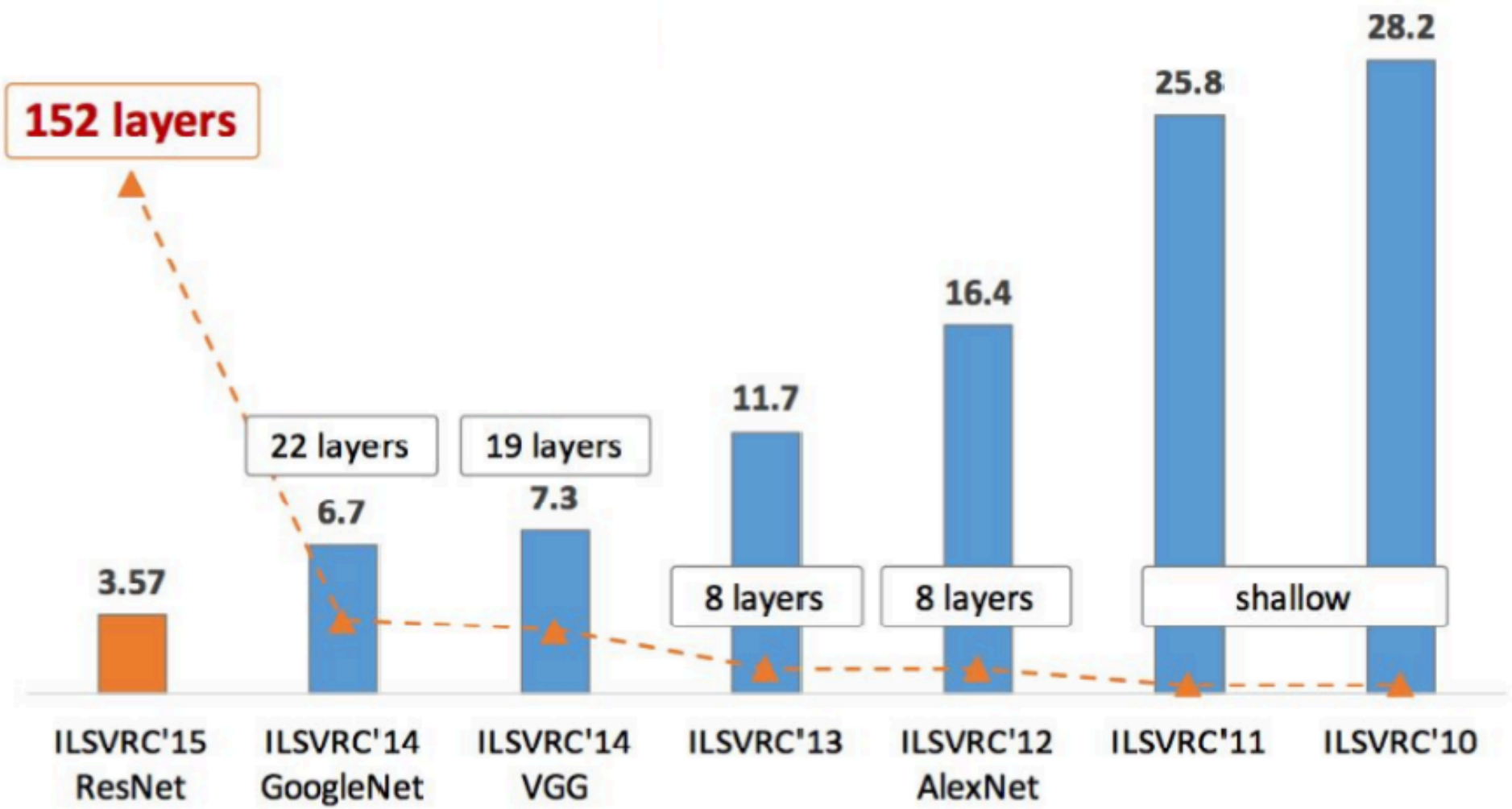
Cameras Everywhere

Smartphones

- Exploding number of sensors vs. humans

Neat Dataset

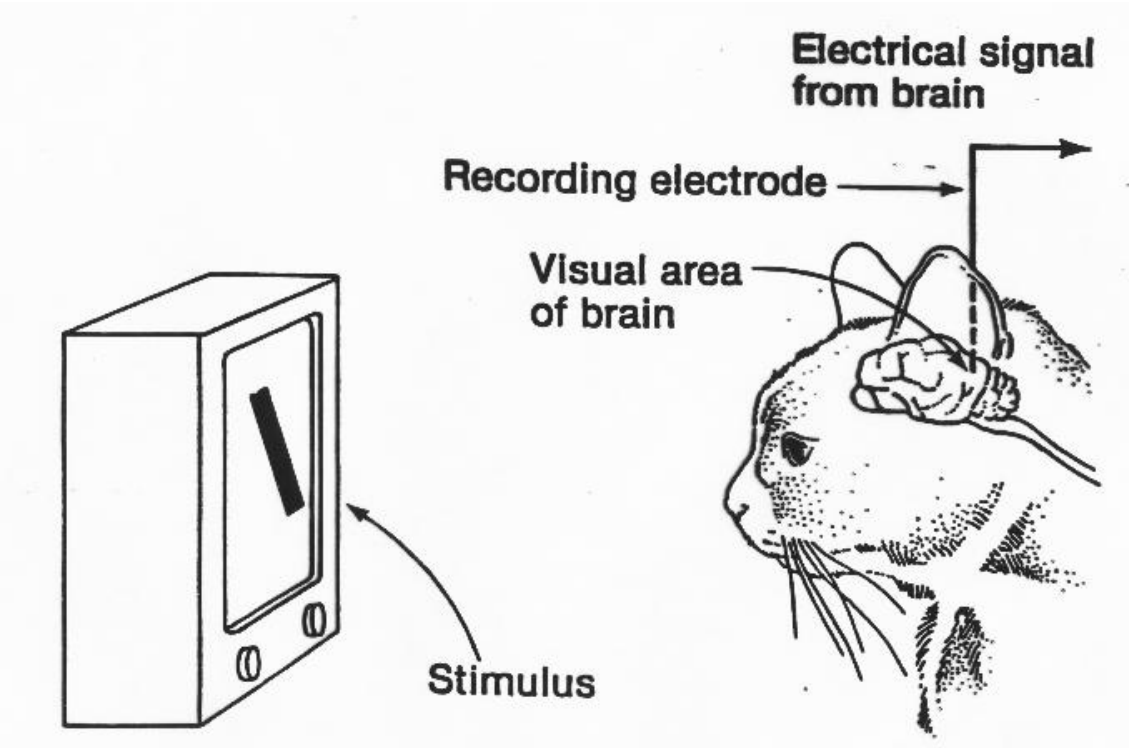
Role of CNN


152 layers
22 layers
19 layers
8 layers
8 layers
shallow
3.57
6.7
7.3
11.7
16.4
25.8
28.2
ILSVRC'15 ResNet
ILSVRC'14 GoogleNet
ILSVRC'14 VGG
ILSVRC'13
ILSVRC'12 AlexNet
ILSVRC'11
ILSVRC'10

Cat's Brain

Types of cells:

- simple cells
- complex cells
- hypercomplex cells



AlexNet

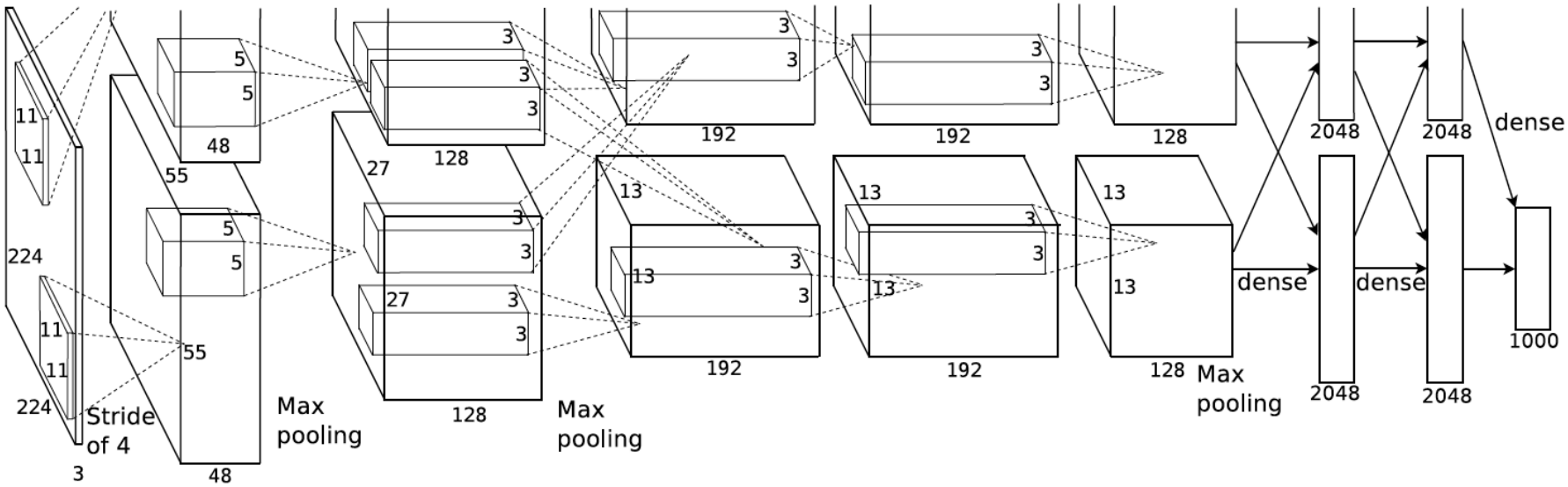


Figure: AlexNet: [Krizhevsky, Sutskever, Hinton] 2012

Convolution

Figure: Convolution Layer

Complexity of Features

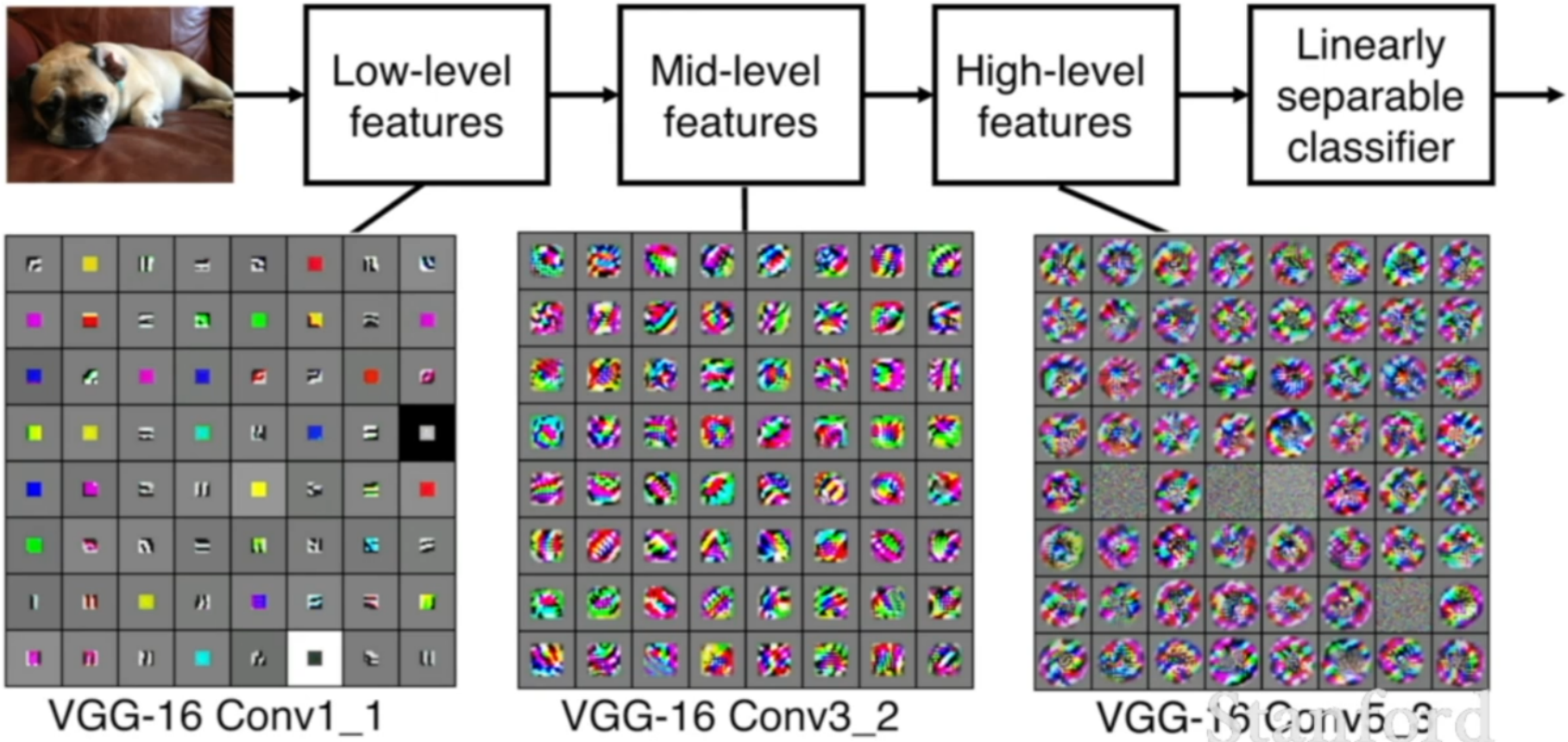


Figure: complexity of features in each layer

Sliding a Filter

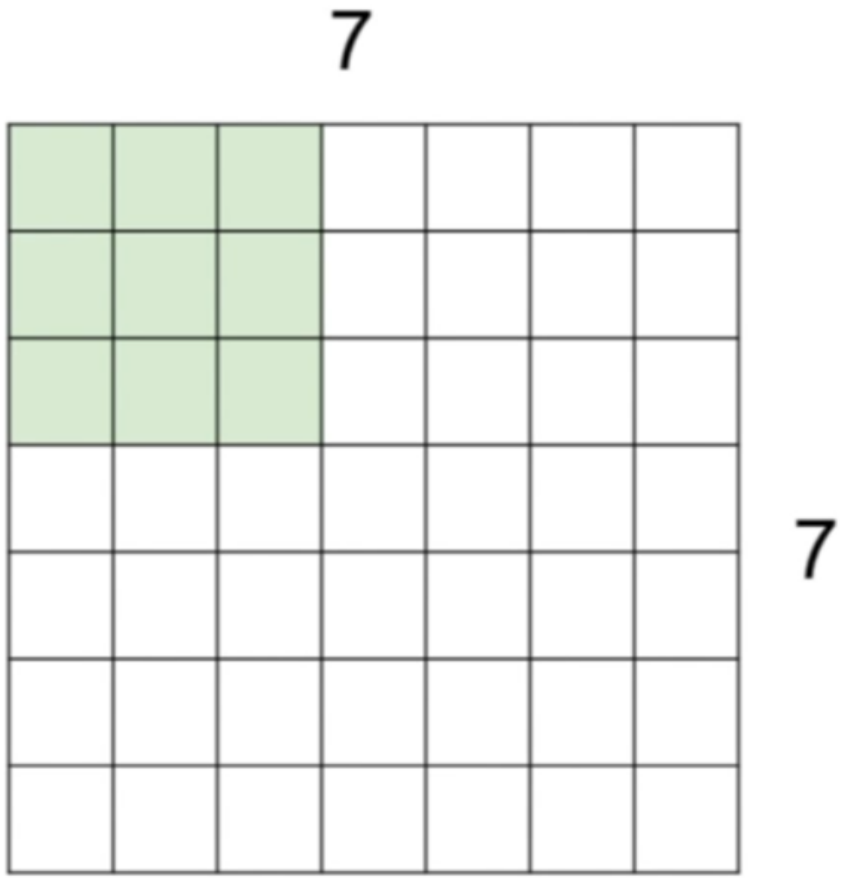


Figure: filtering example

Sliding a Filter (Cont.)

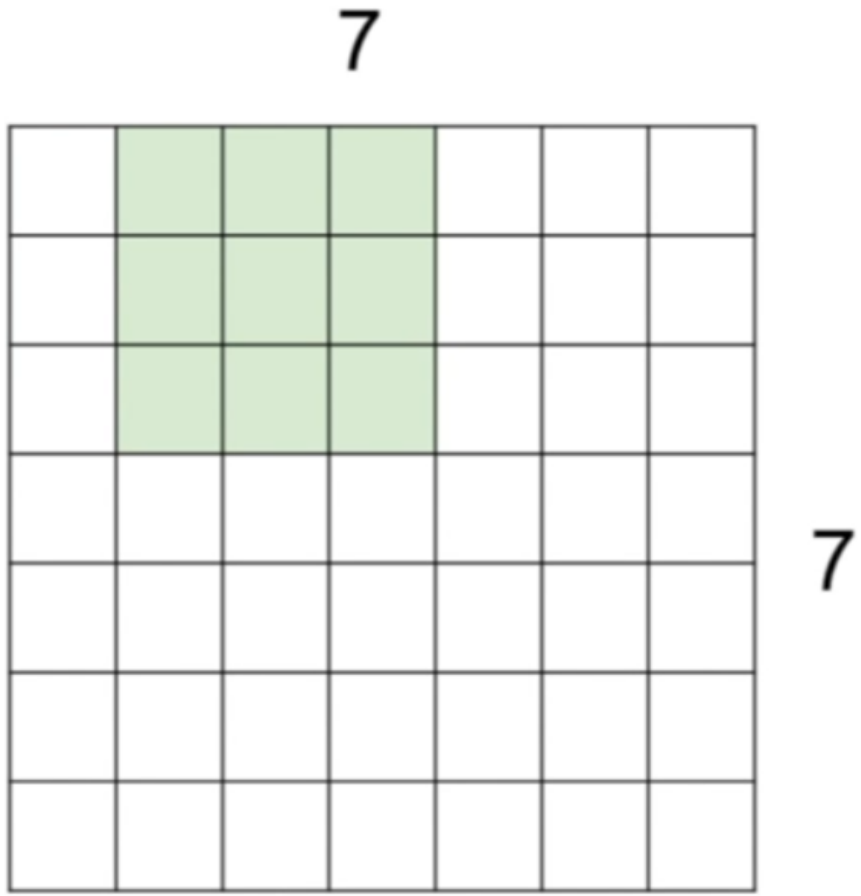


Figure: filtering example

Sliding a Filter (Cont.)

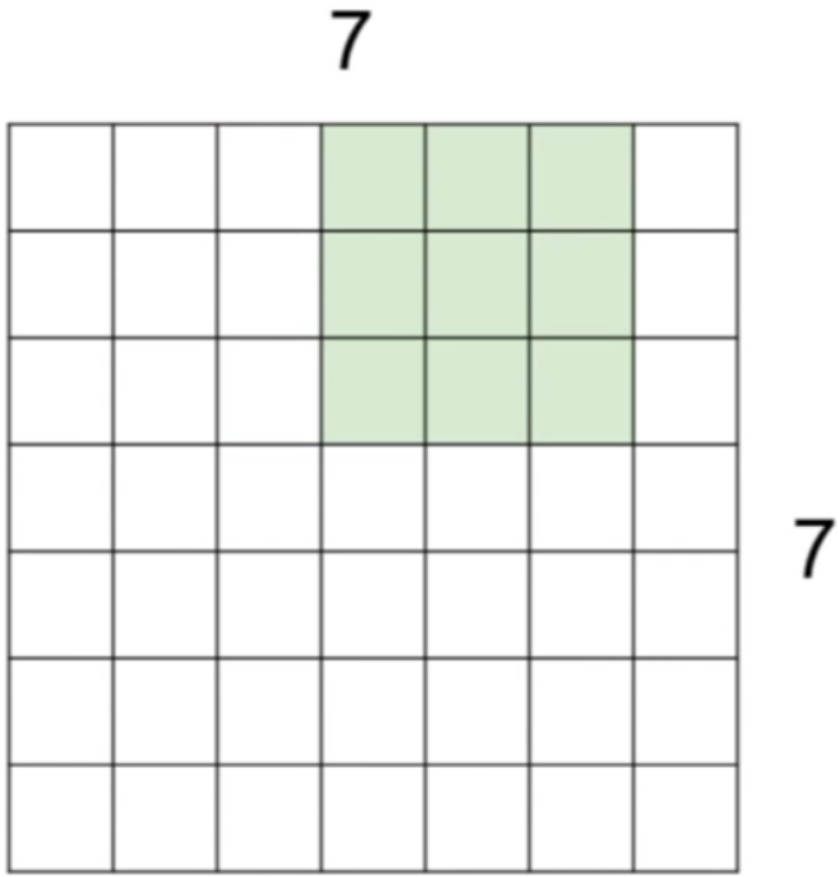


Figure: filtering example

Sliding a Filter (Cont.)

What if using a 3*3 filter with stride 3 ?!

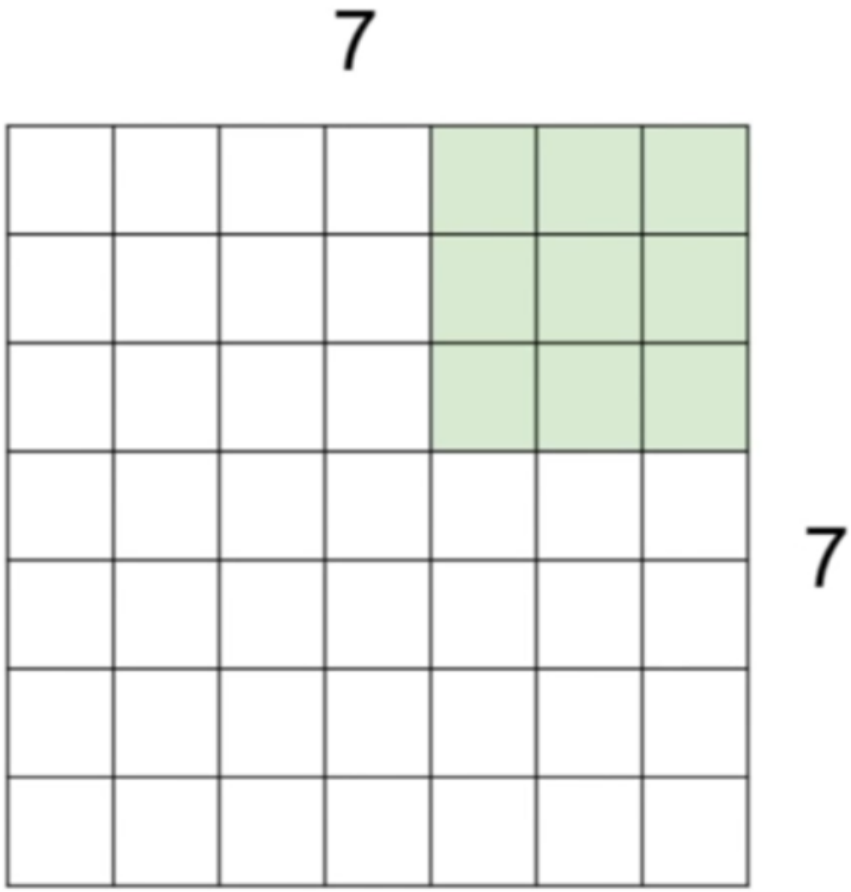


Figure: filtering example

Padding

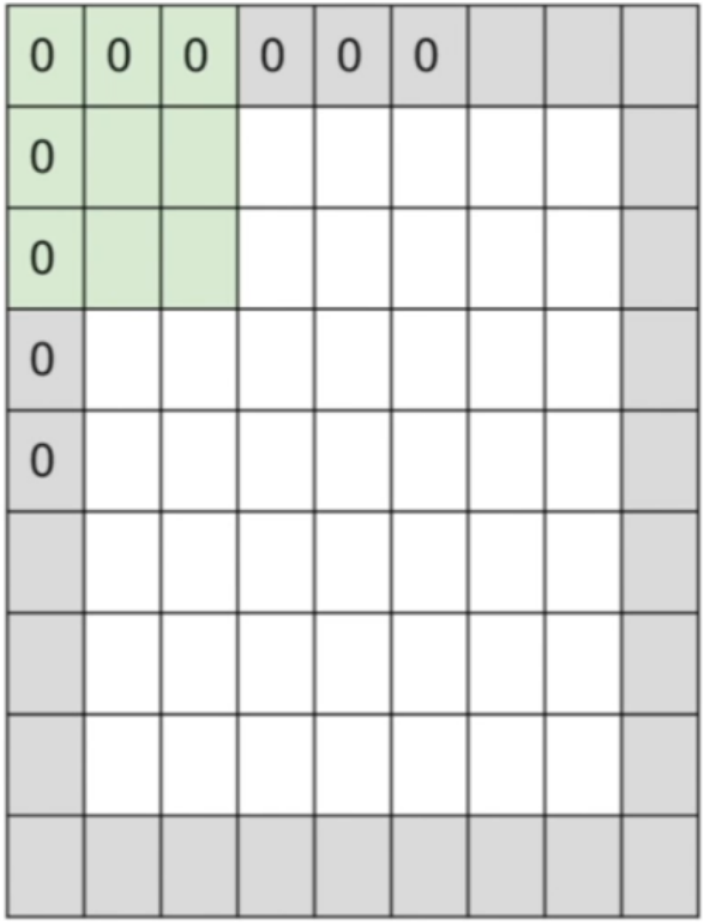


Figure: Padding: control output size

Summary

- Accept a volume of size N x N
- requires 4 hyper parameters:
  - Filter's spatial extent F
  - Filter's Stride S
  - Amount of zero padding P
- Produces a volume of size M X M
  - M $= \frac{(N-F+2p)}{S} + 1$
  - Number of parameters : $(F \times F \times D + b) \times k$

Pooling

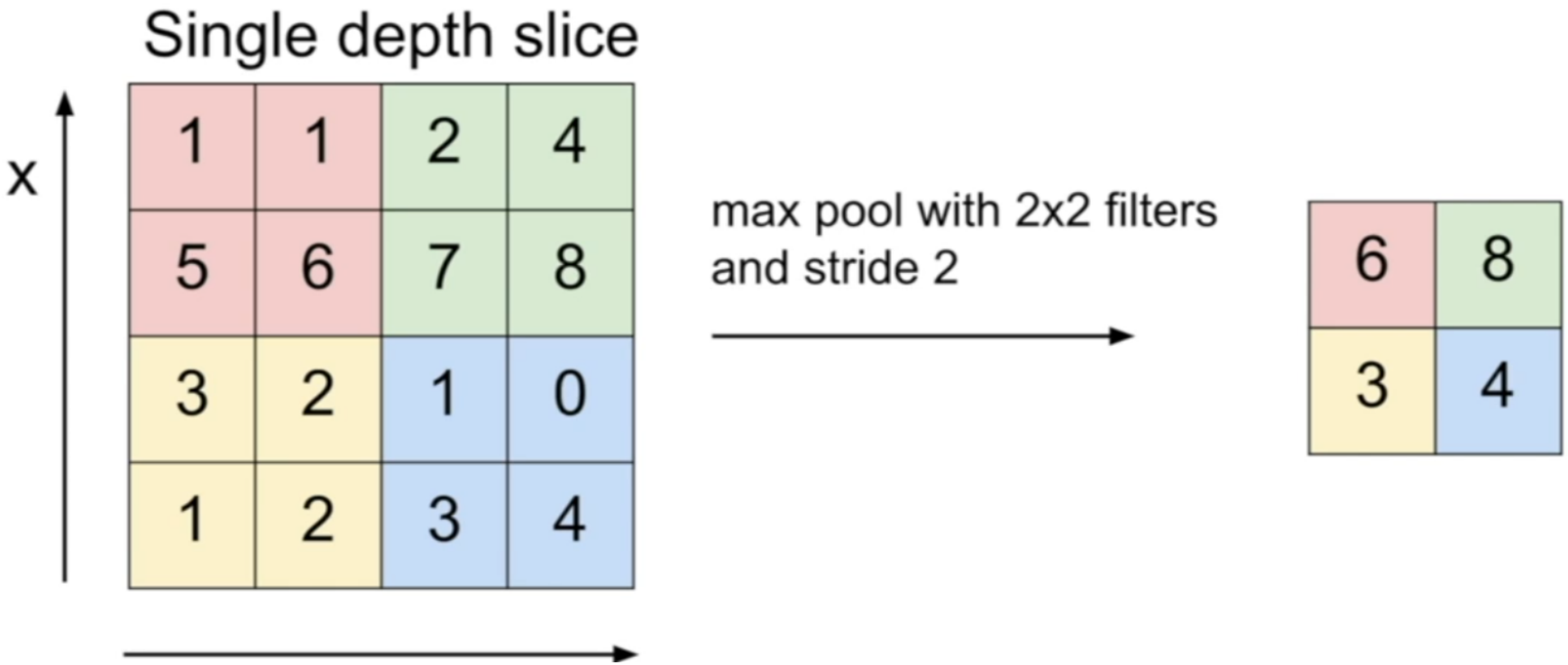


Figure: Max-pooling

Fully Connected Layer

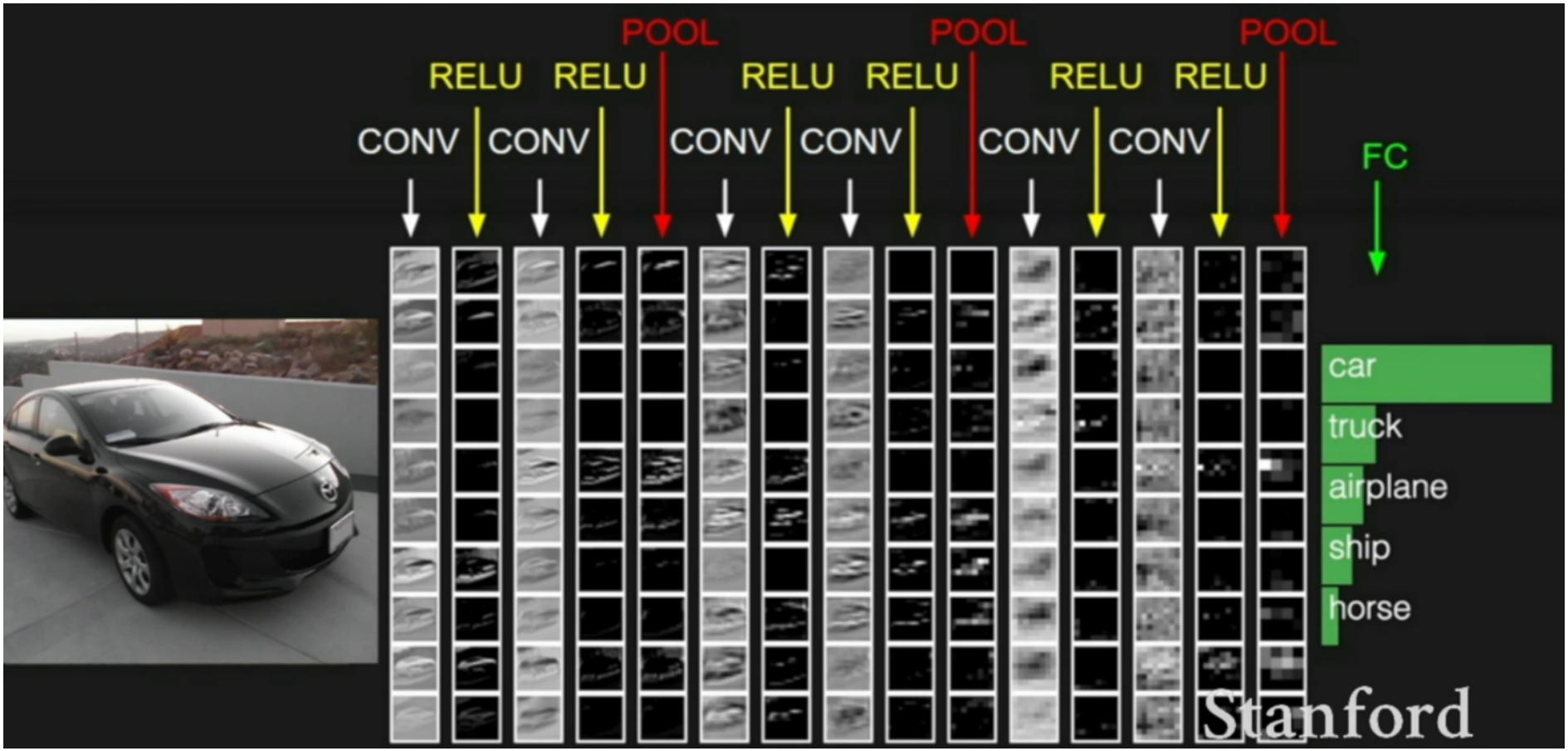


Figure: Fully Connected Layer

AlexNet

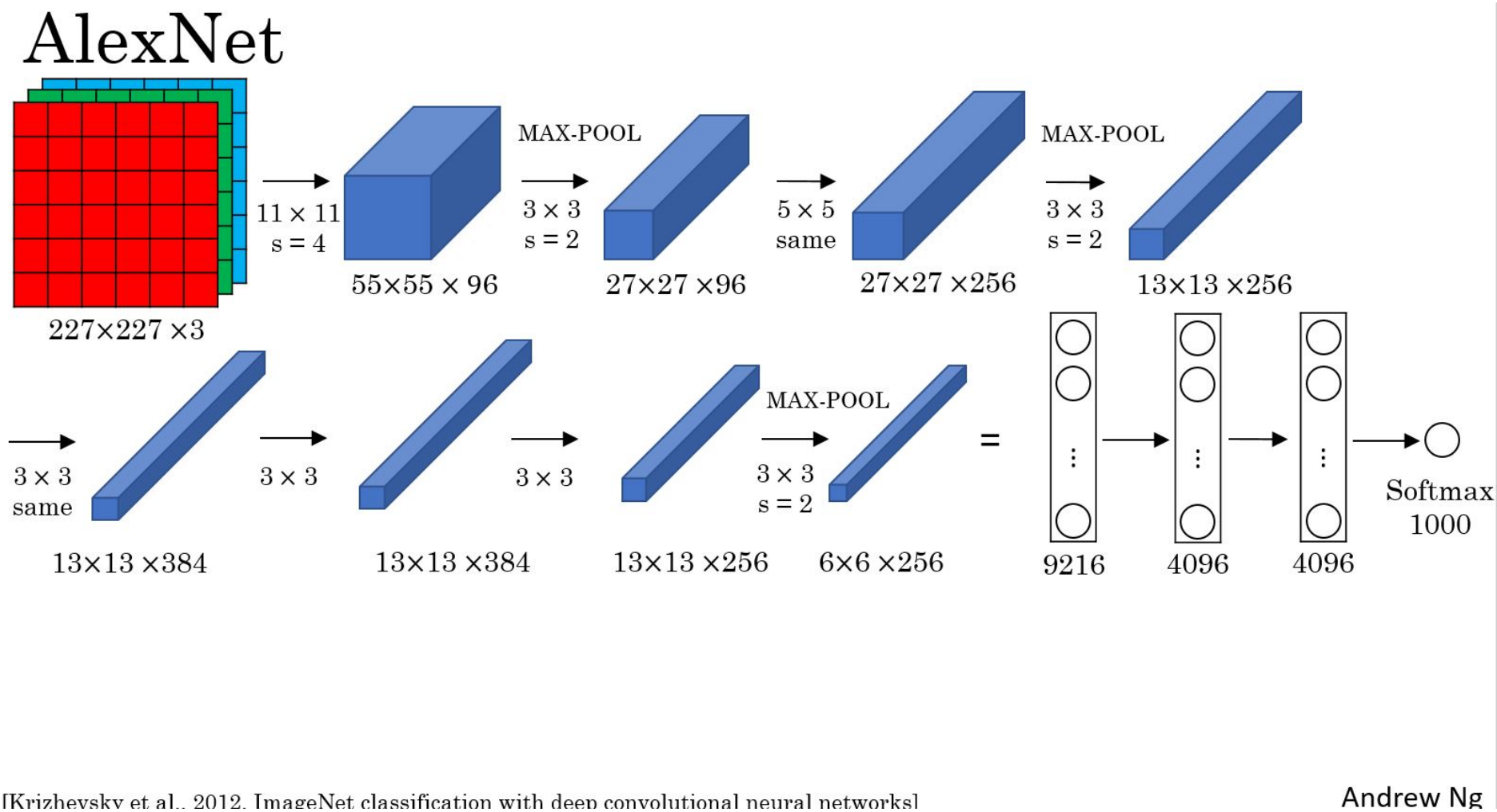


Figure: AlexNet architecture

Comparing different Structures

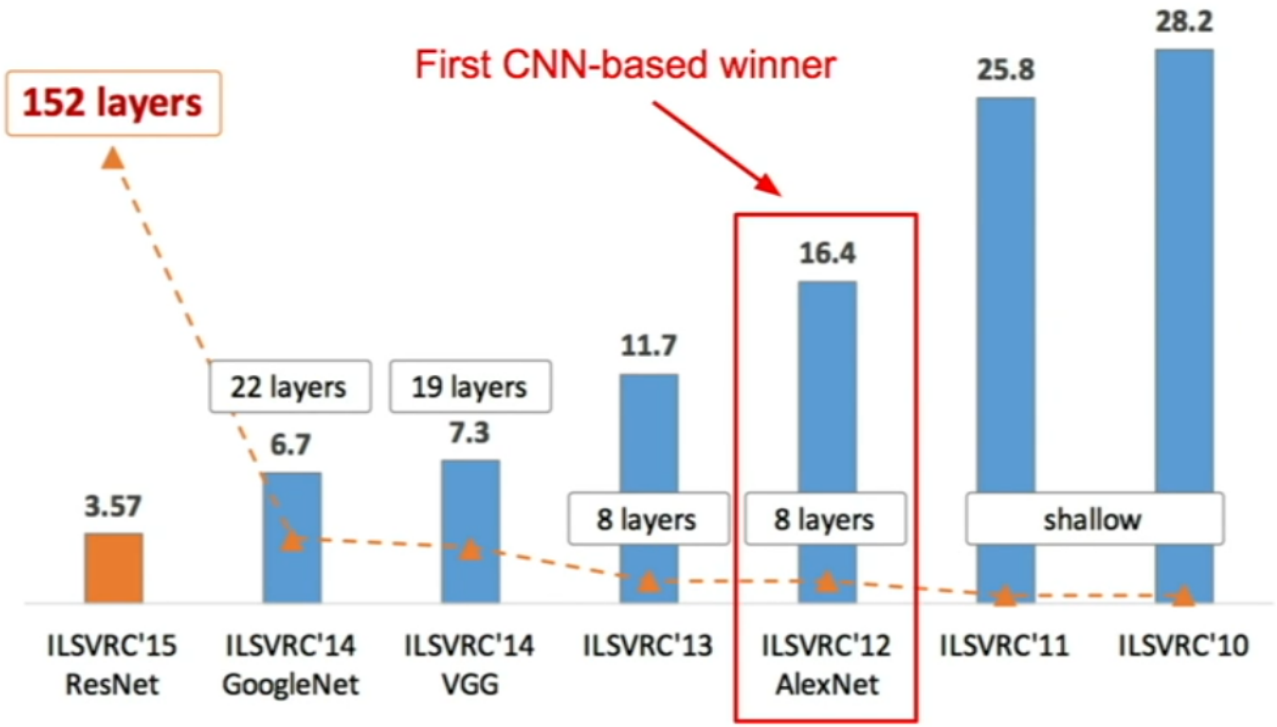
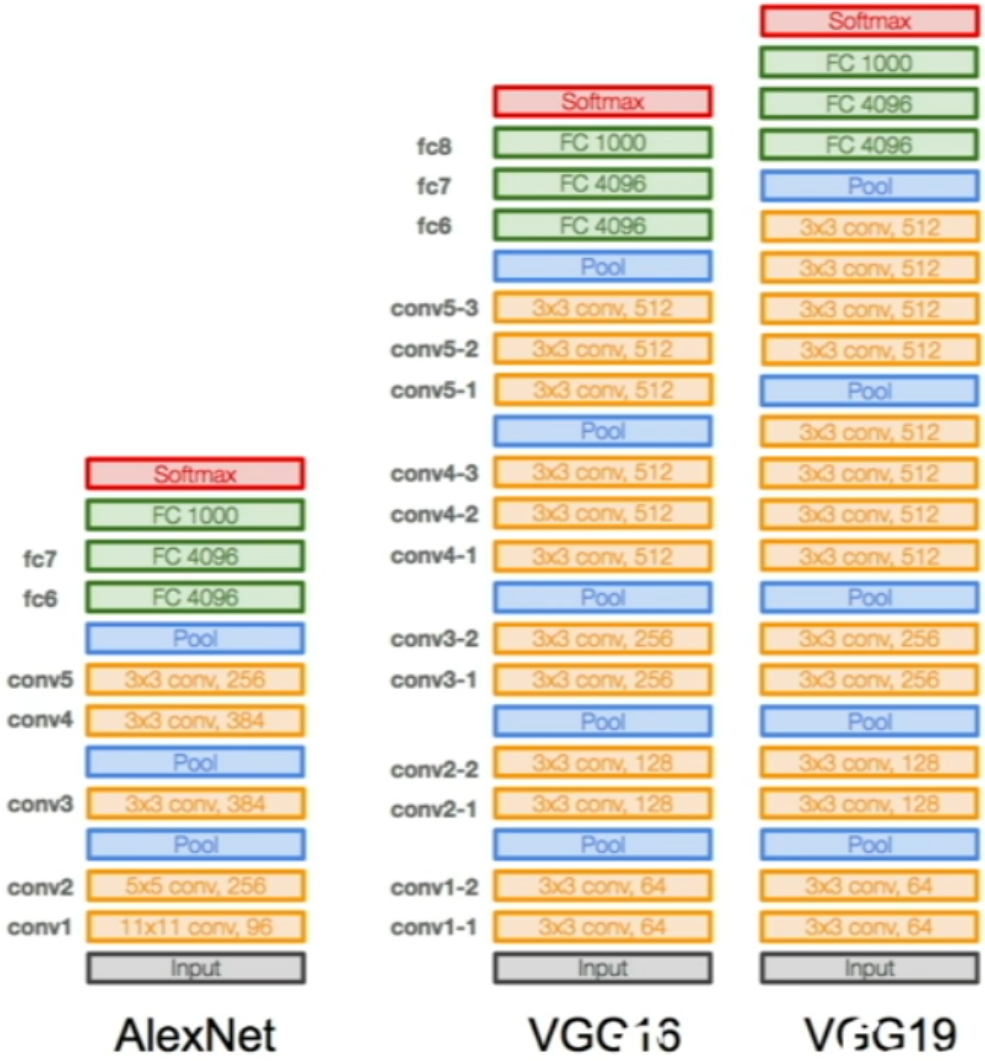


Figure: Comparing Structures

Deeper Networks, Smaller Filters



Comparing

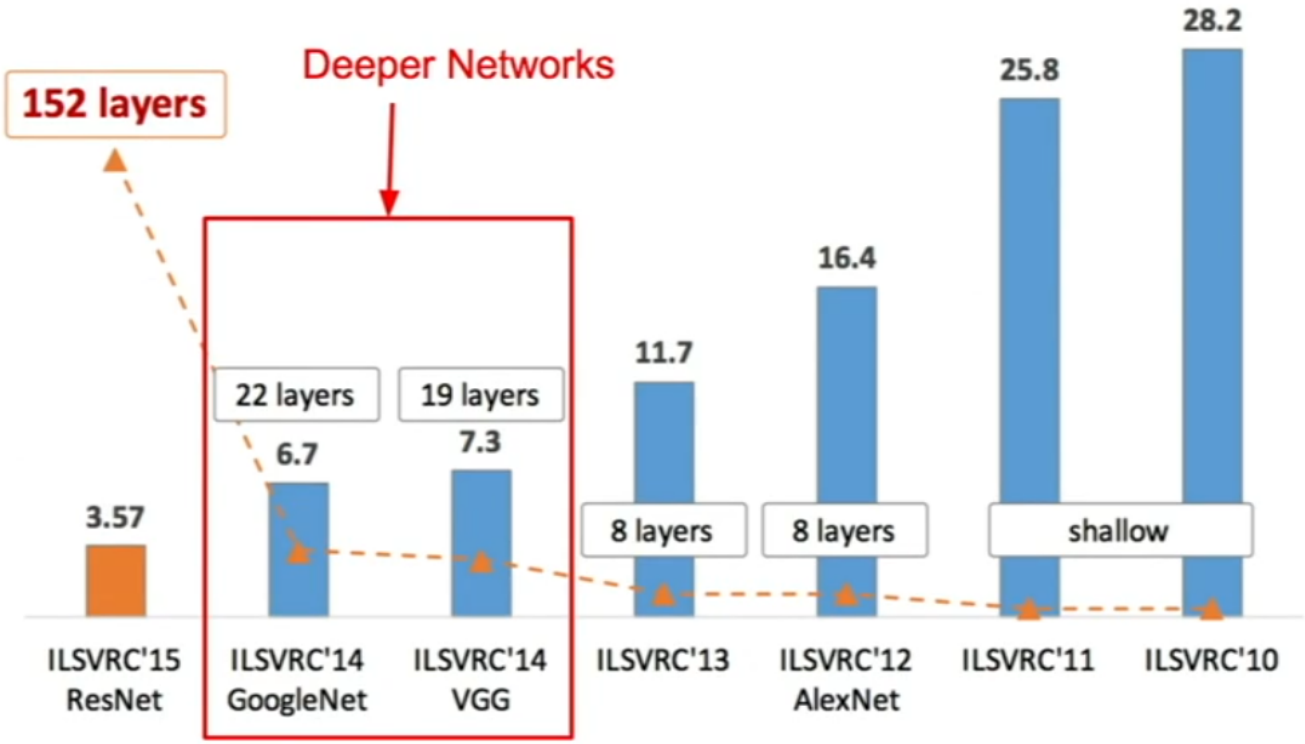


Figure: Comparing Structures

Google Net

Deeper Networks, with computationally inexpensive

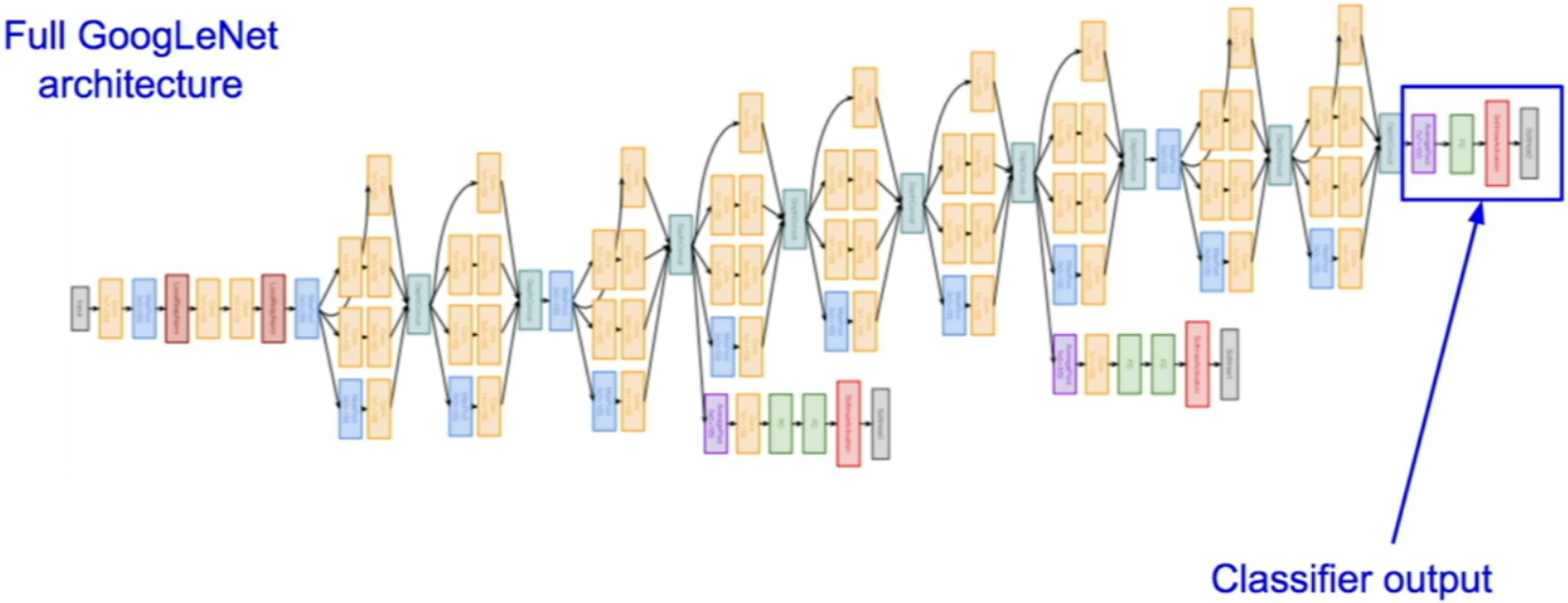


Figure: Googlenet Structures

Comparing

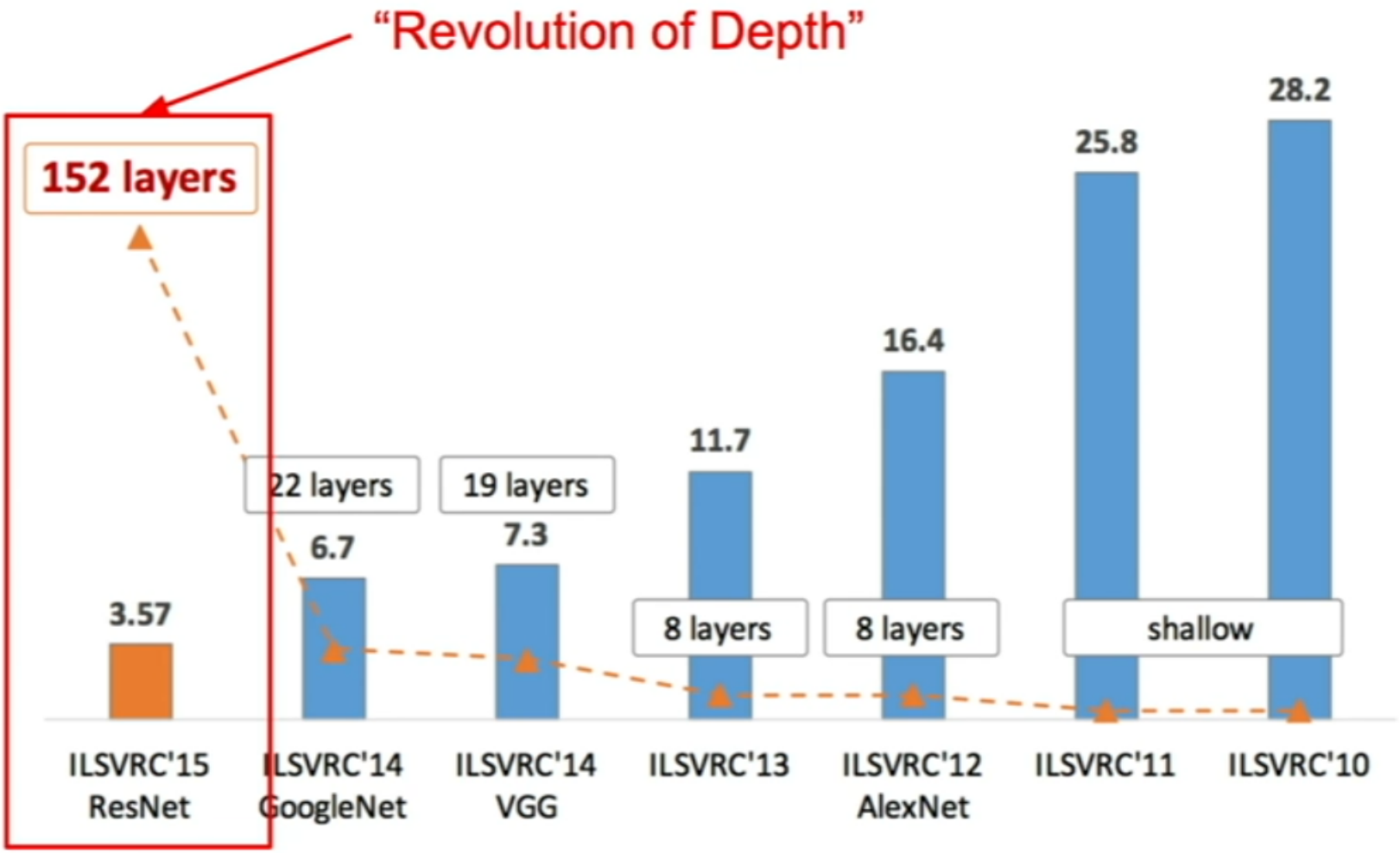


Figure: Comparing Structures

ResNet

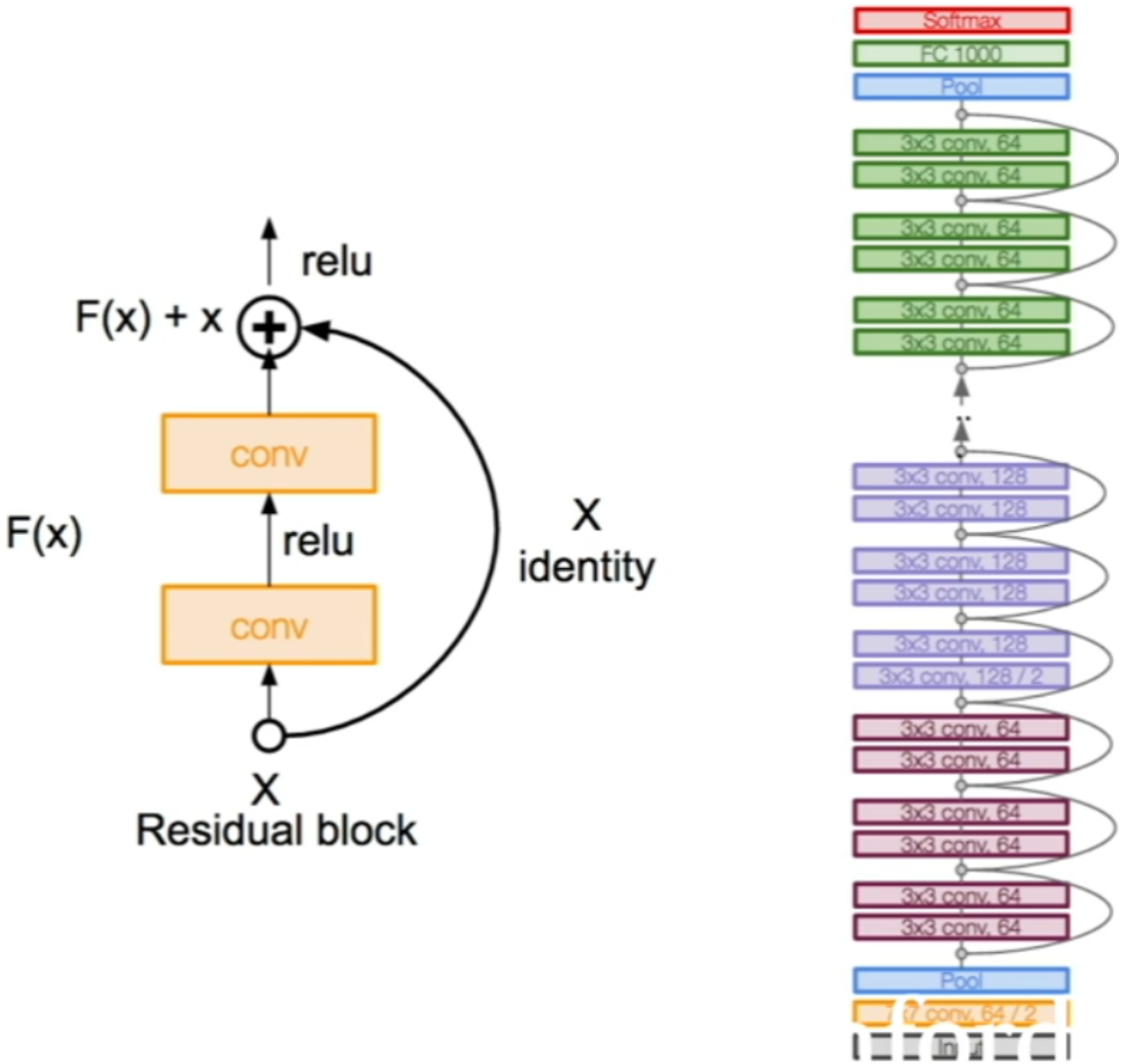


Figure: ResNet Structures

Comparing Structures

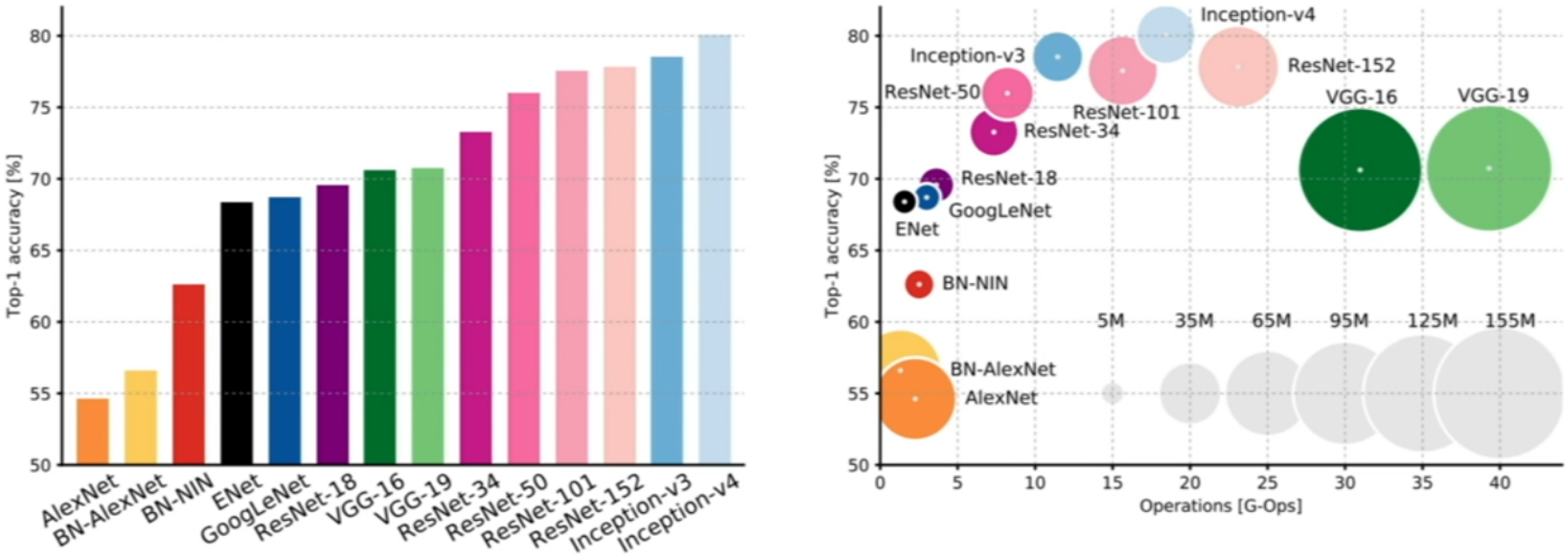


Figure: Comparing Structures

ResNet

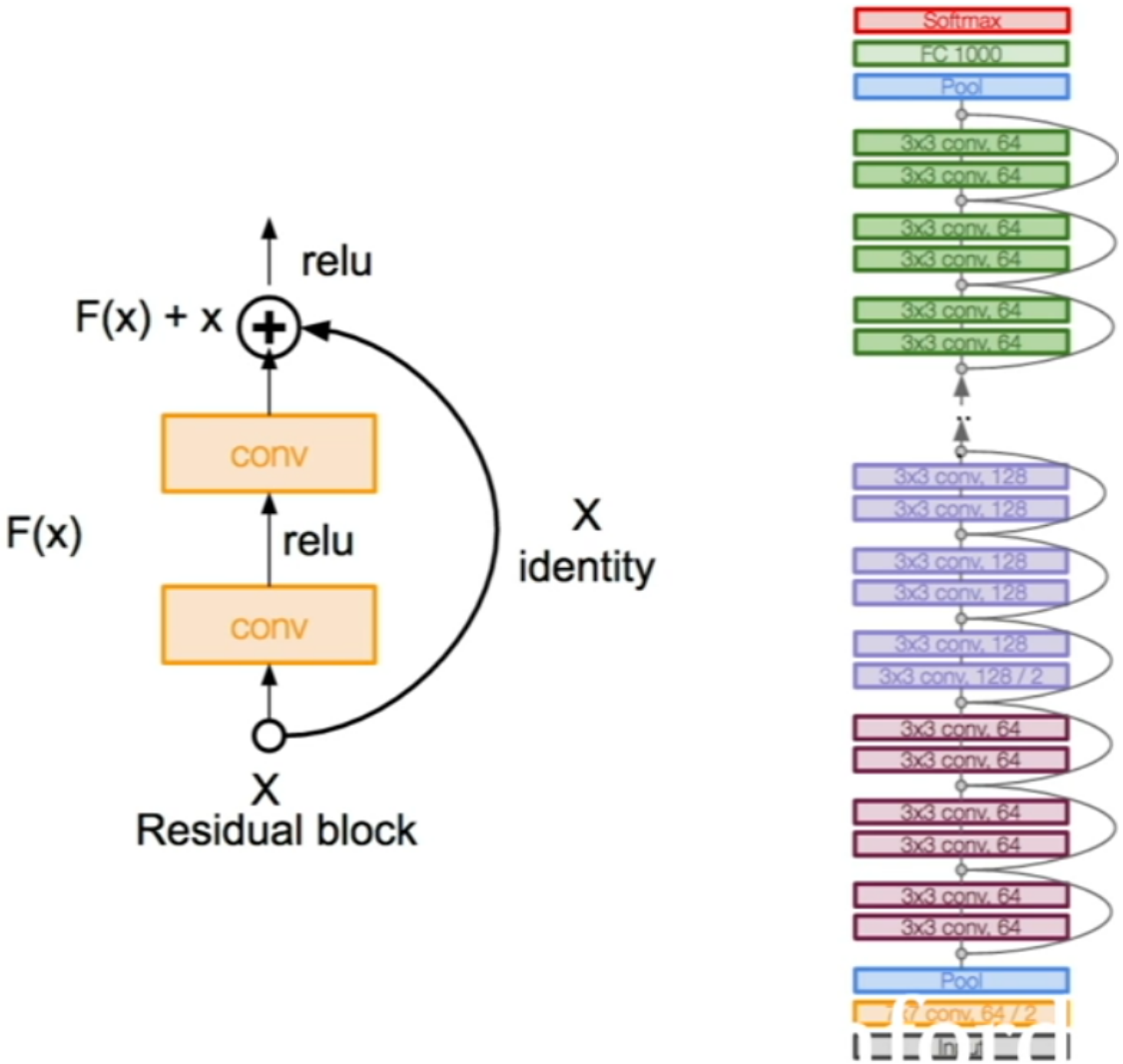
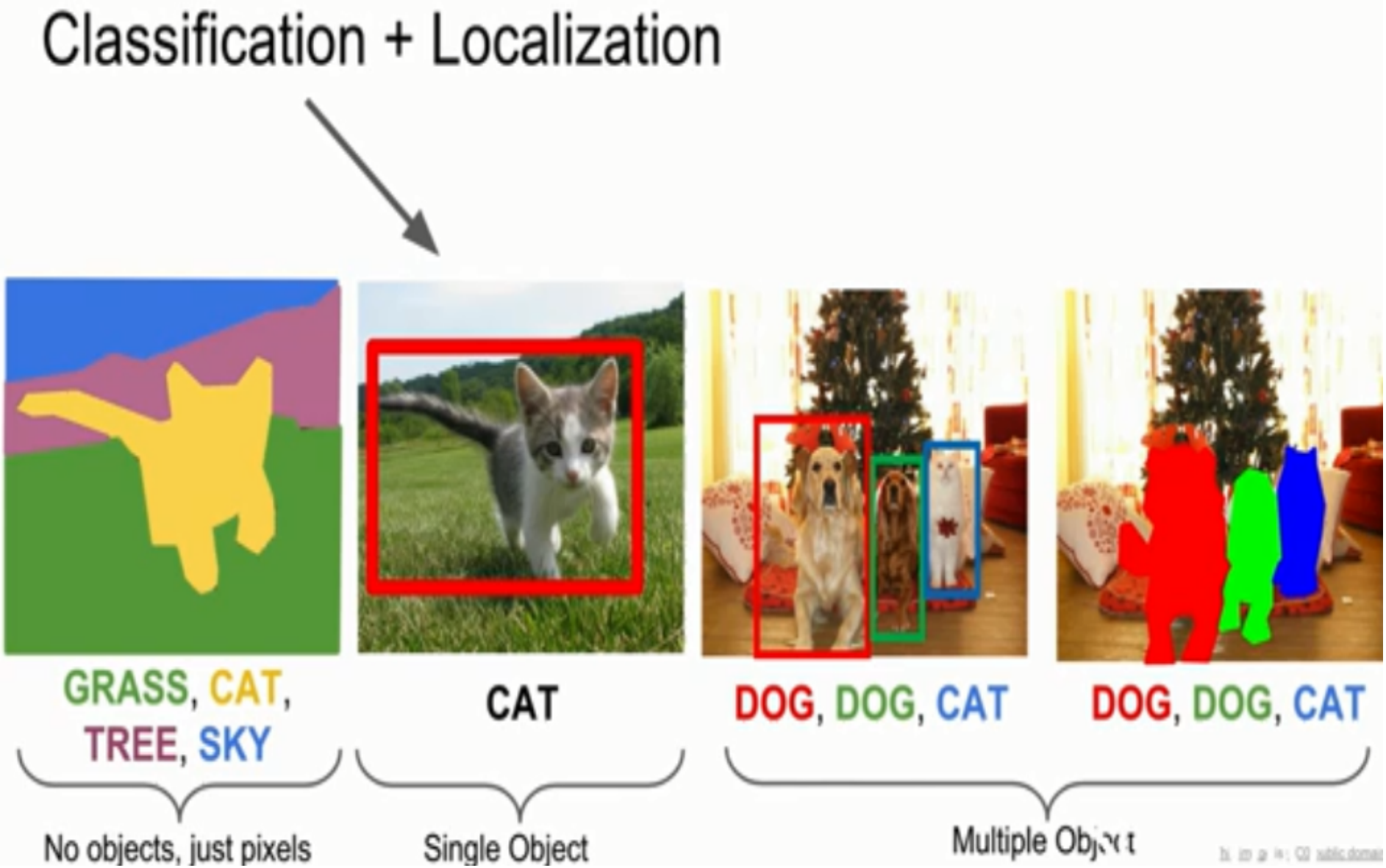
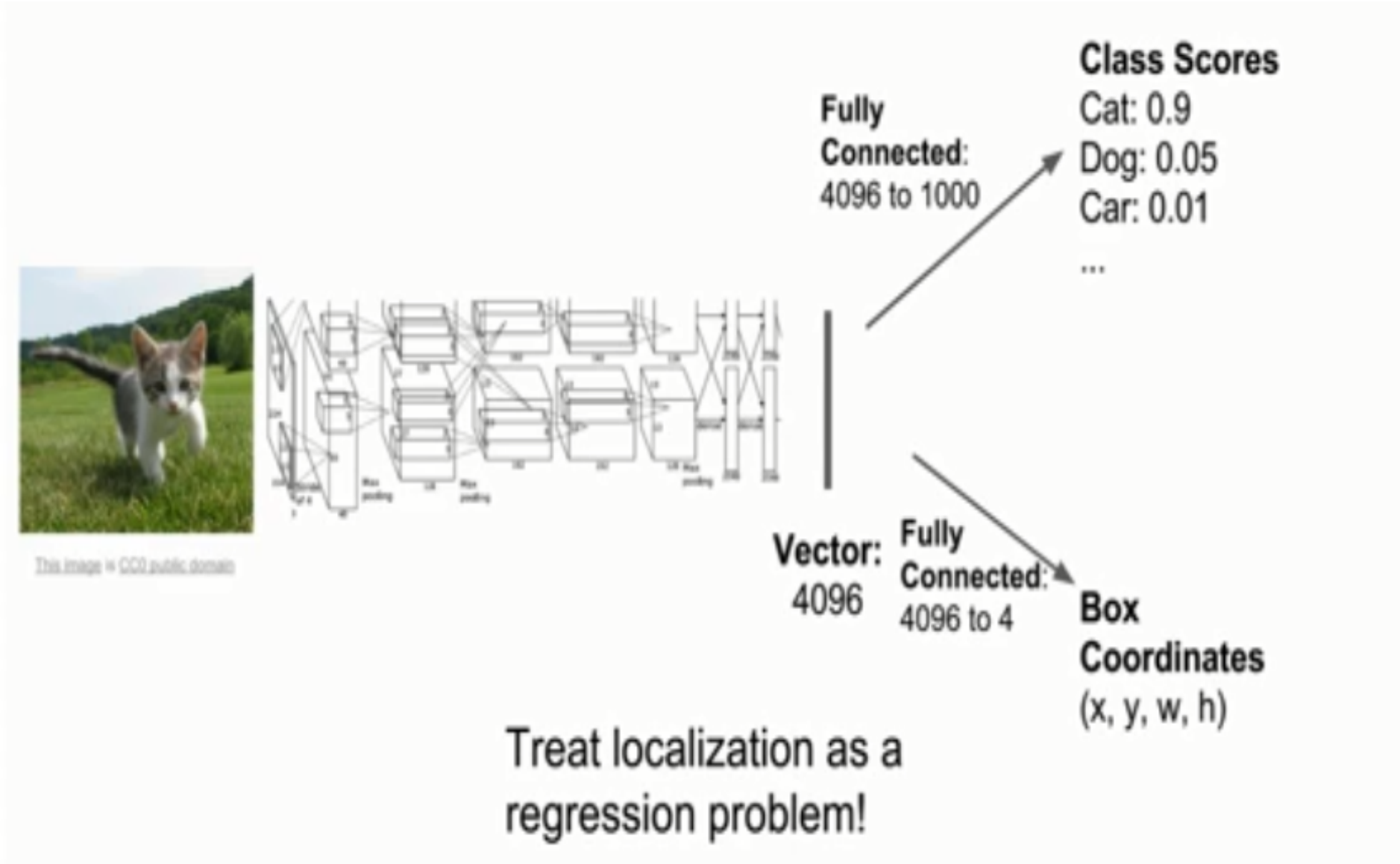


Figure: ResNet Structures

Class + Location



Class + Location

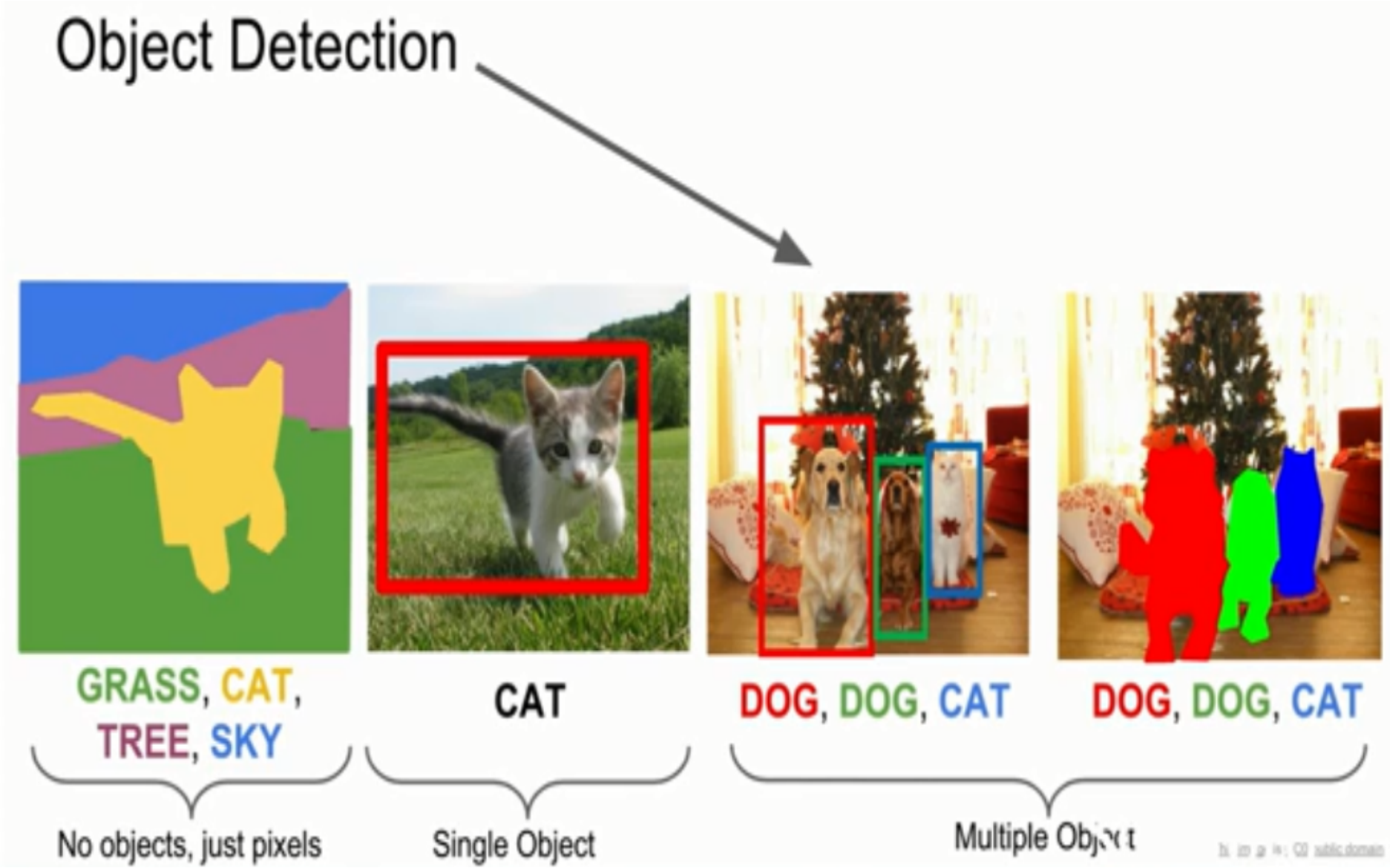


Example

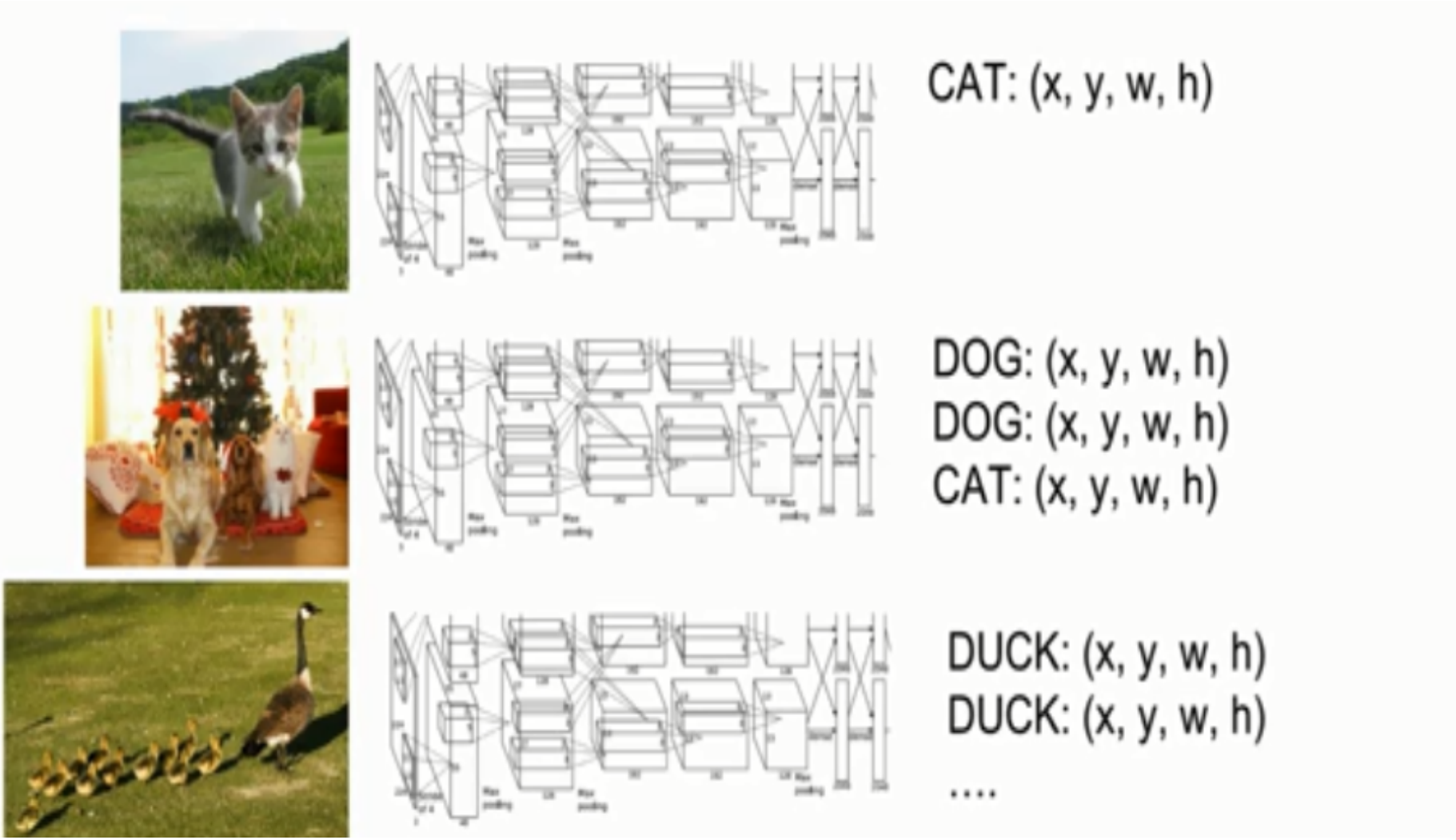
Represent pose as a set of 14 joint positions:

Left / right foot
Left / right knee
Left / right hip
Left / right shoulder
Left / right elbow
Left / right hand
Neck
Head top

Object detection

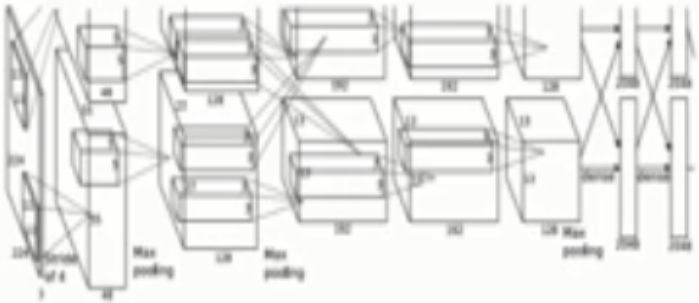


Object detection



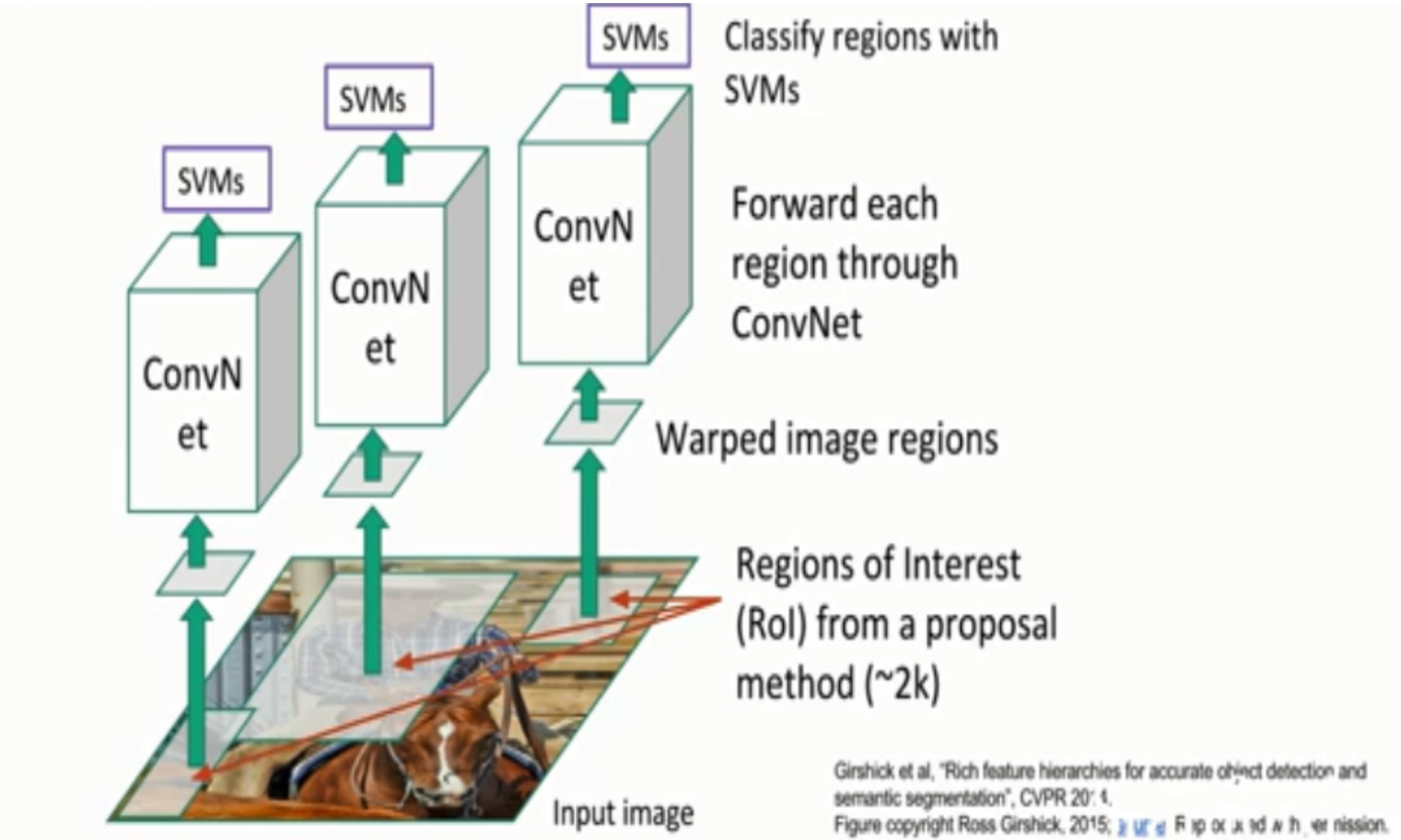
Object detection

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



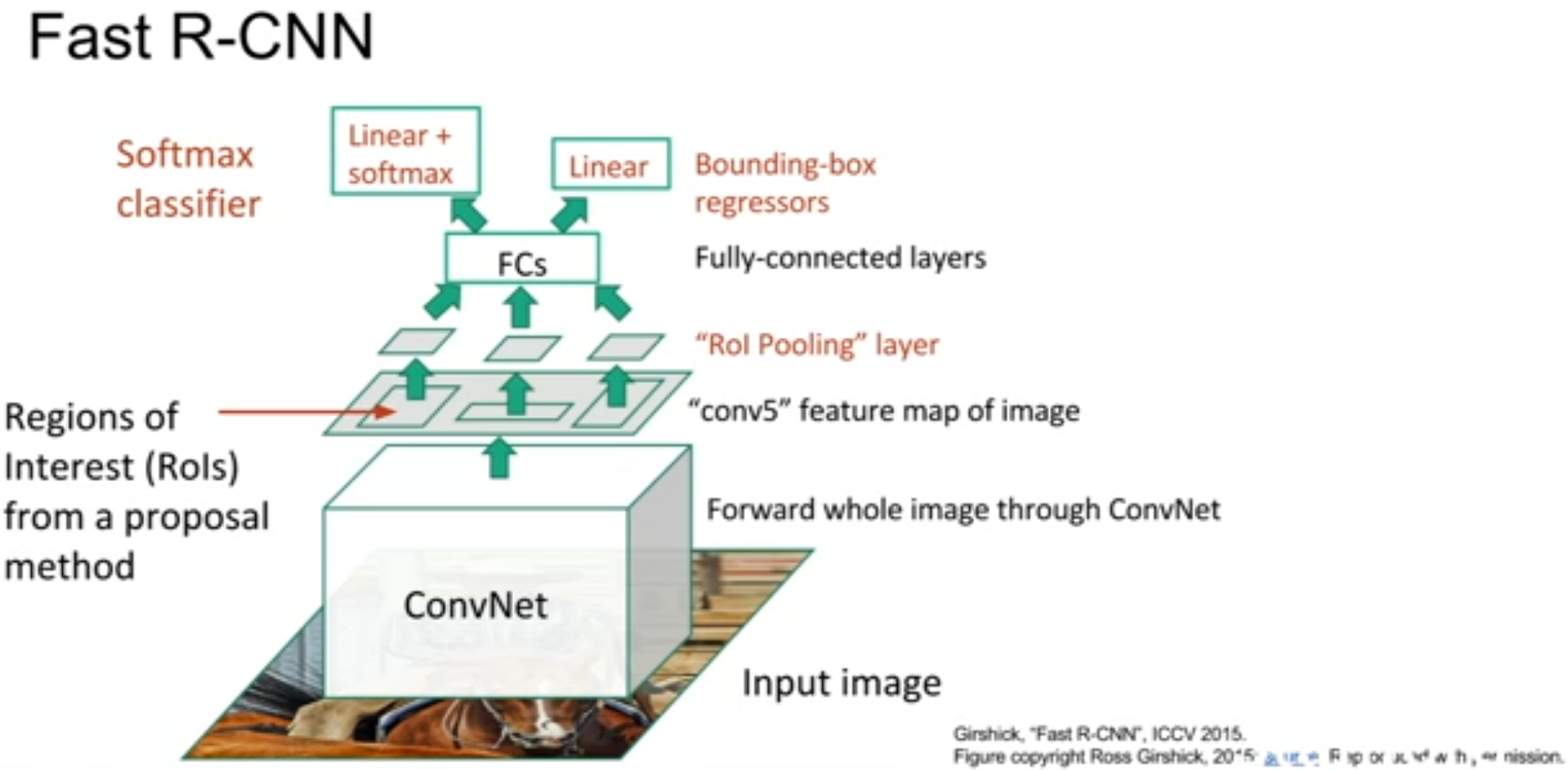
Dog? NO
Cat? NO
Background? YES

RCNN

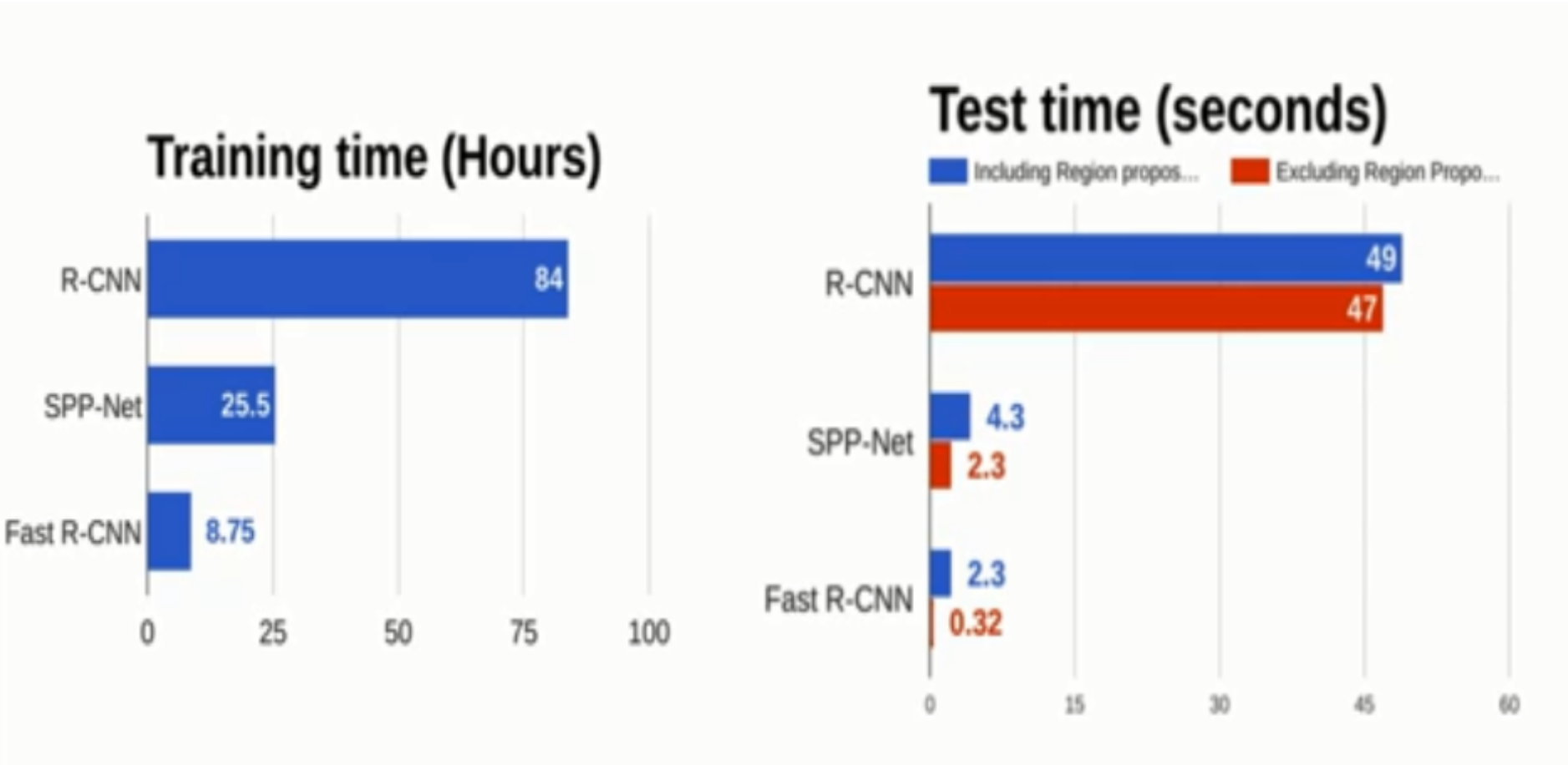


Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 20'. t.
Figure copyright Ross Girshick, 2015; [illegible] R ep ce d w h er nission.

Fast RCNN


Fast R-CNN
Softmax classifier
Linear + softmax
Linear
Bounding-box regressors
FCs
Fully-connected layers
"RoI Pooling" layer
Regions of Interest (RoIs) from a proposal method
"conv5" feature map of image
Forward whole image through ConvNet
ConvNet
Input image
Girshick, "Fast R-CNN", ICCV 2015.

Comparing


Training time (Hours)
R-CNN
84
SPP-Net
25.5
Fast R-CNN
8.75
0
25
50
75
100
Test time (seconds)
Including Region propos...
Excluding Region Propo...
R-CNN
49
47
SPP-Net
4.3
2.3
Fast R-CNN
2.3
0.32
0
15
30
45
60

Even Faster

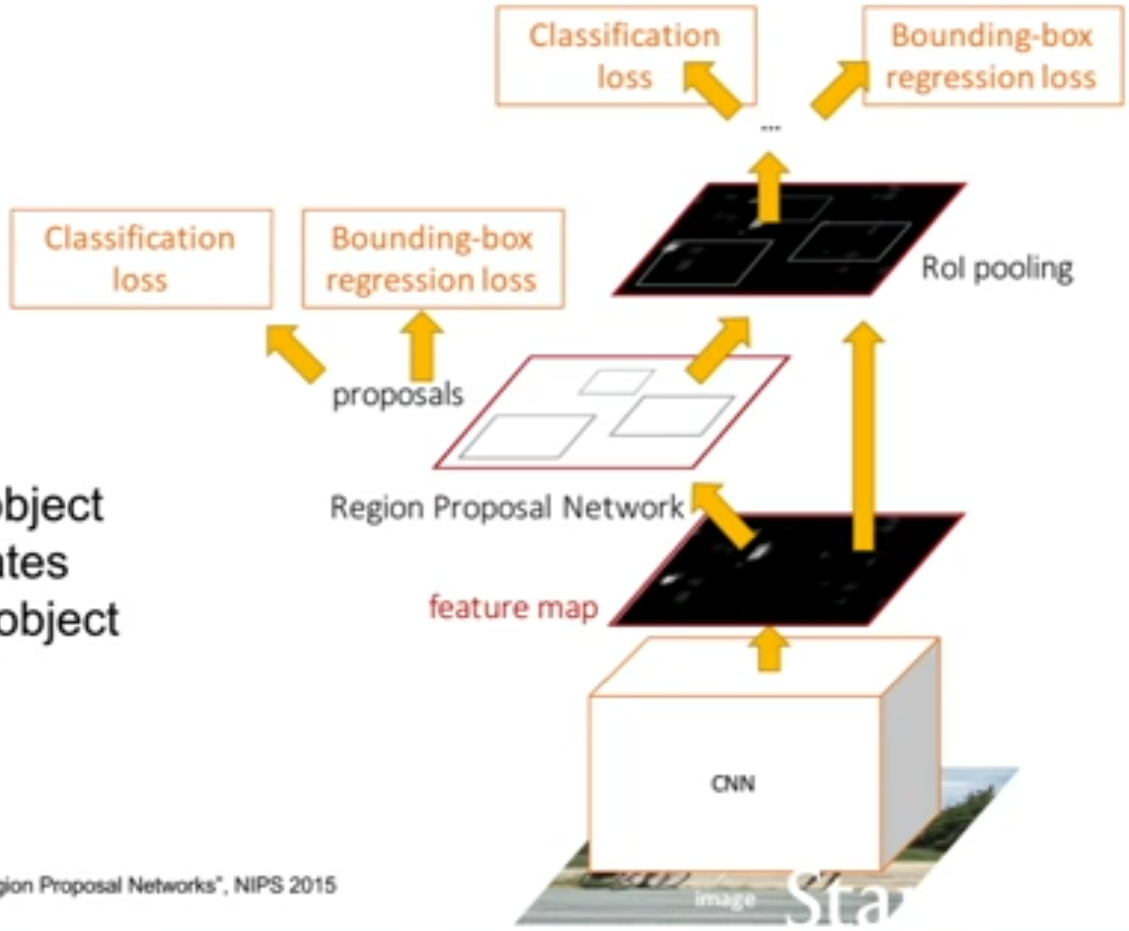
Faster R-CNN:

Make CNN do proposals!

Insert **Region Proposal Network (RPN)** to predict proposals from features

Jointly train with 4 losses:

1. RPN classify object / not object
2. RPN regress box coordinates
3. Final classification score (object classes)
4. Final box coordinates

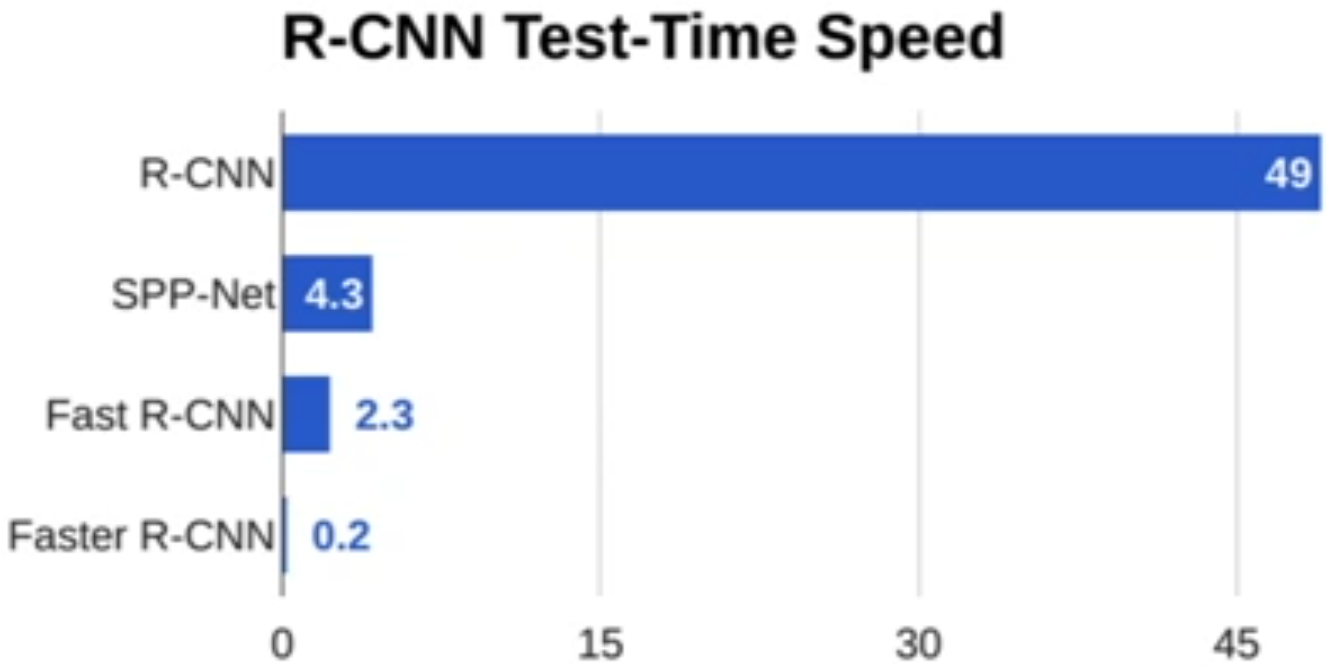


Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015
Figure copyright 2015, Ross Girshick; reproduced with permission

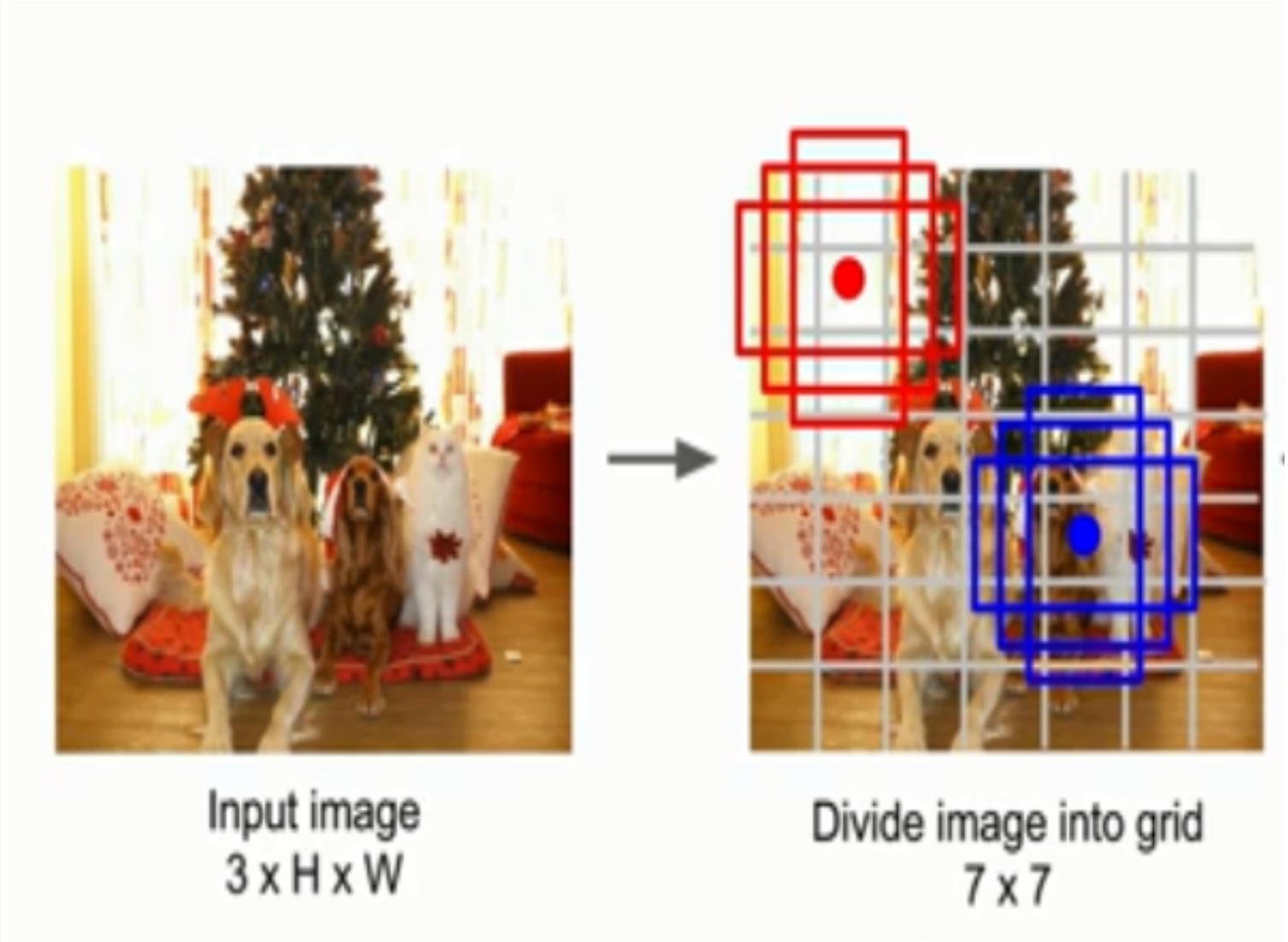
Even Faster

Faster R-CNN:

Make CNN do proposals!



Yolo / SSD



References:

- Stanford CS231n

Any Question?