Convolutional Neural Networks

Classical and State-of-the-art designs

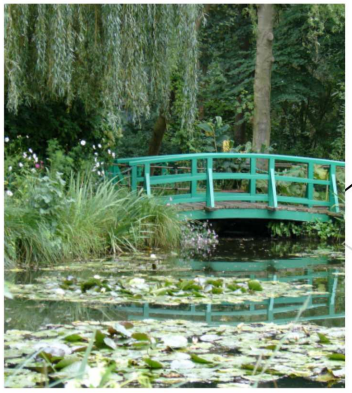
Vamsi K. Ithapu
CS 540 Lecture

December 8, 2017

Contents

What is (computer) vision?

Image (or video)



Sensing device

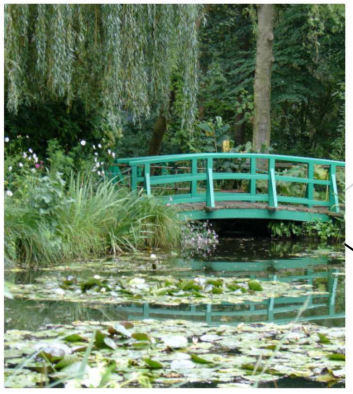
Interpreting device

Interpretations

garden, spring, bridge, water, trees, flower, green, etc.

What is (computer) vision?

Image (or video)



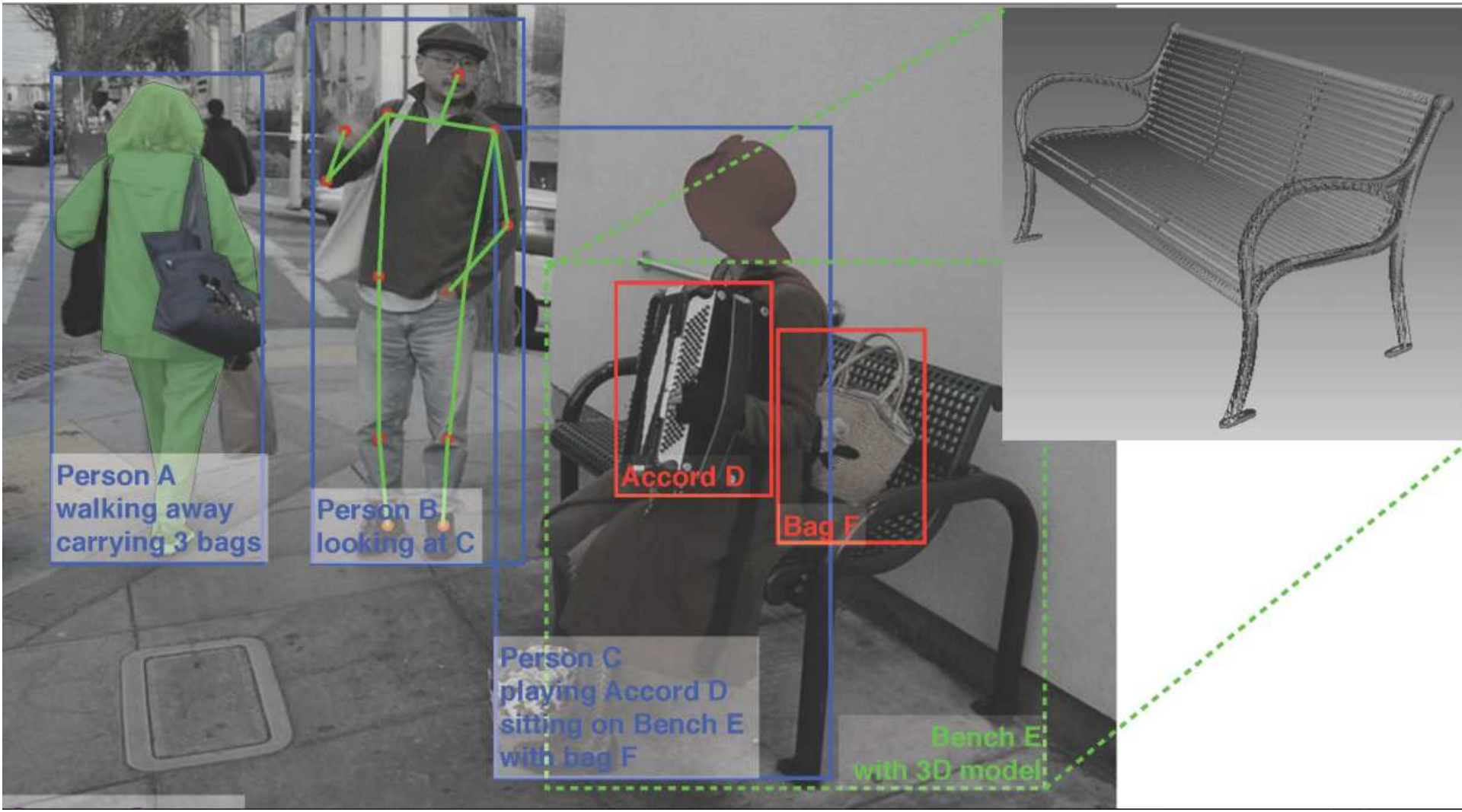
Sensing device

Interpreting device

Interpretations

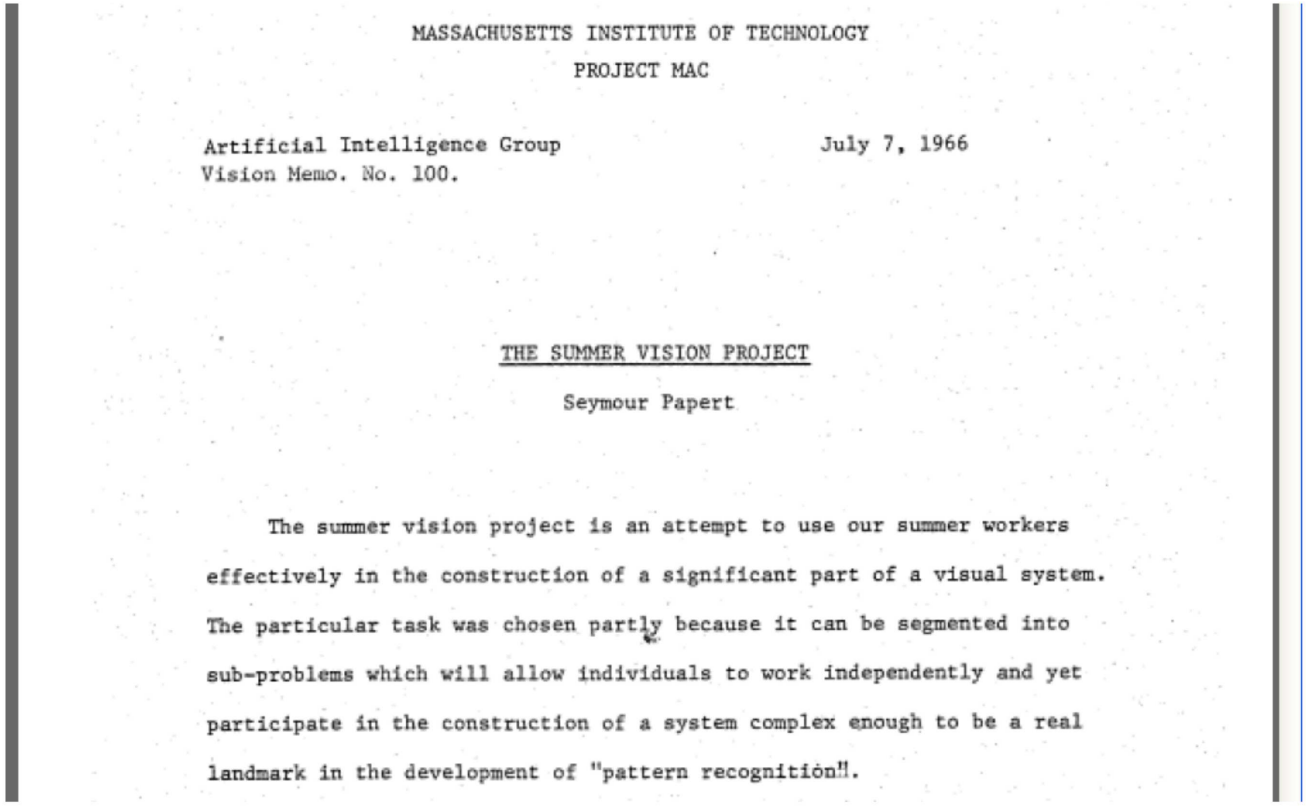
garden, spring, bridge, water, trees, flower, green, etc.

What we would like to infer…



Will person B put some money into Person C's tip bag?

Origins of computer vision: an MIT undergraduate summer project



MASSACHUSETTS INSTITUTE OF TECHNOLOGY
PROJECT MAC

Artificial Intelligence Group
Vision Memo. No. 100.

July 7, 1966

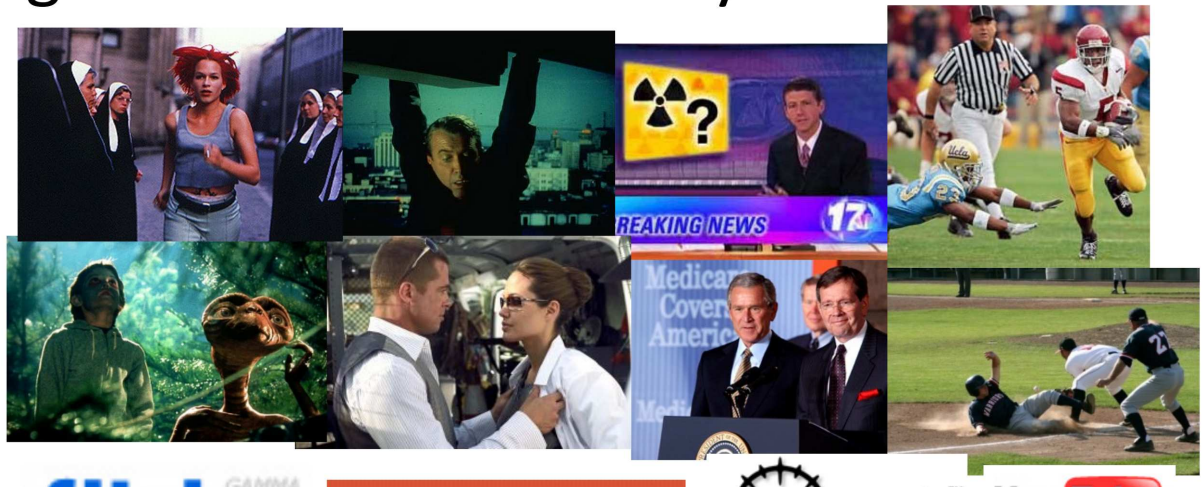
THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

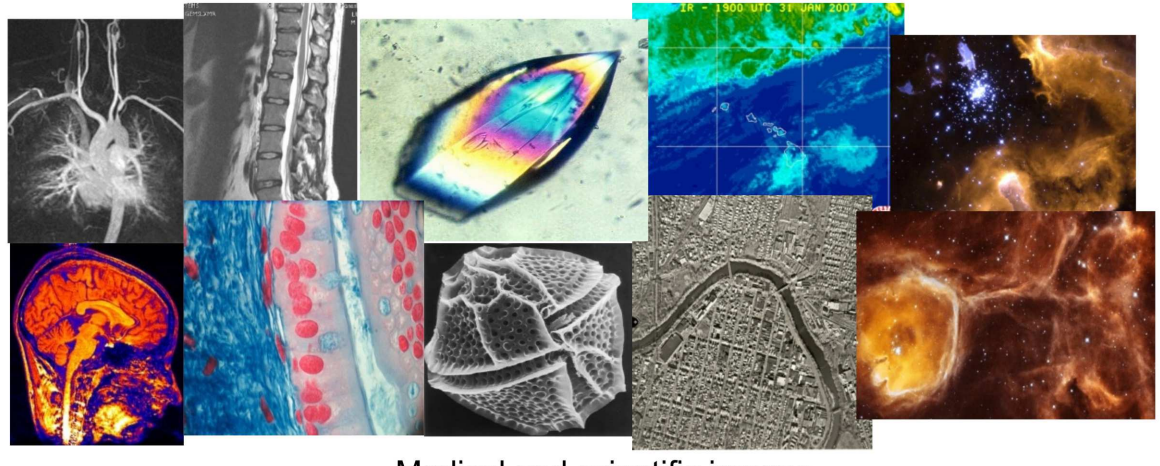
Why study computer vision?

- Vision is useful: Images and video are everywhere!



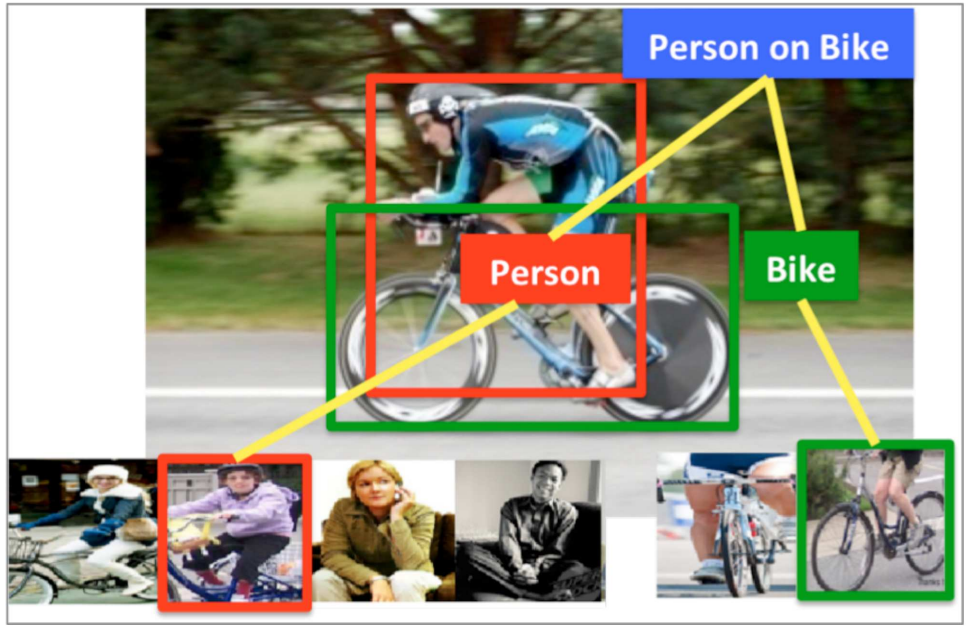
Google Image Search Picasa™ flickr™ webshots picsearch™ YouTube Broadcast Yourself™

Surveillance and security



Medical and scientific images

- Object detection
- Action classification
- Image captioning
- …



Face detection

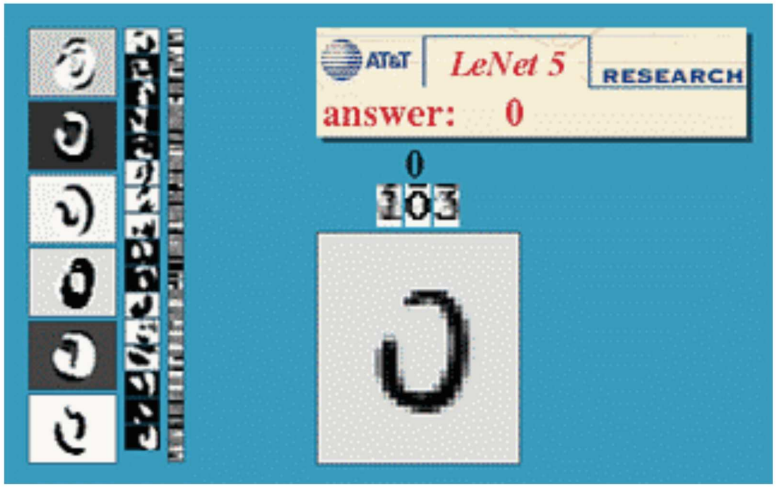
- Many new digital cameras now detect faces
  - Canon, Sony, Fuji, …

Source: S. Seitz

Optical character recognition (OCR)

Technology to convert scanned docs to text

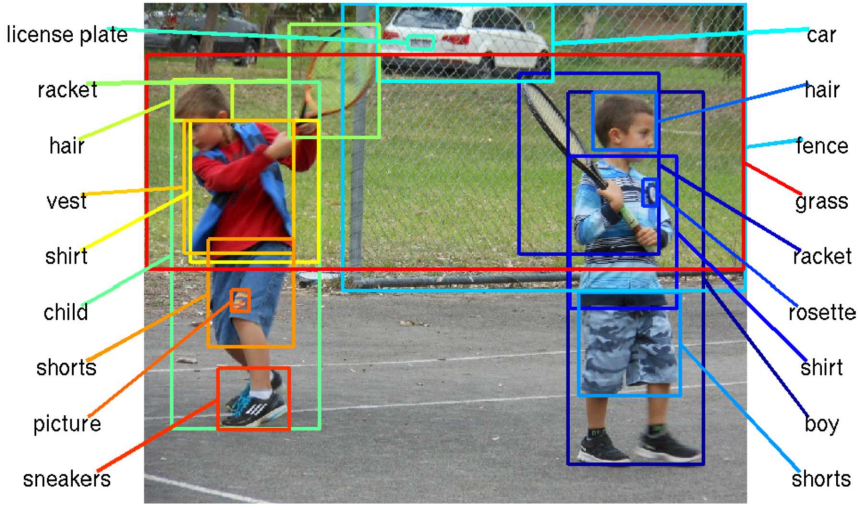
- If you have a scanner, it probably came with OCR software

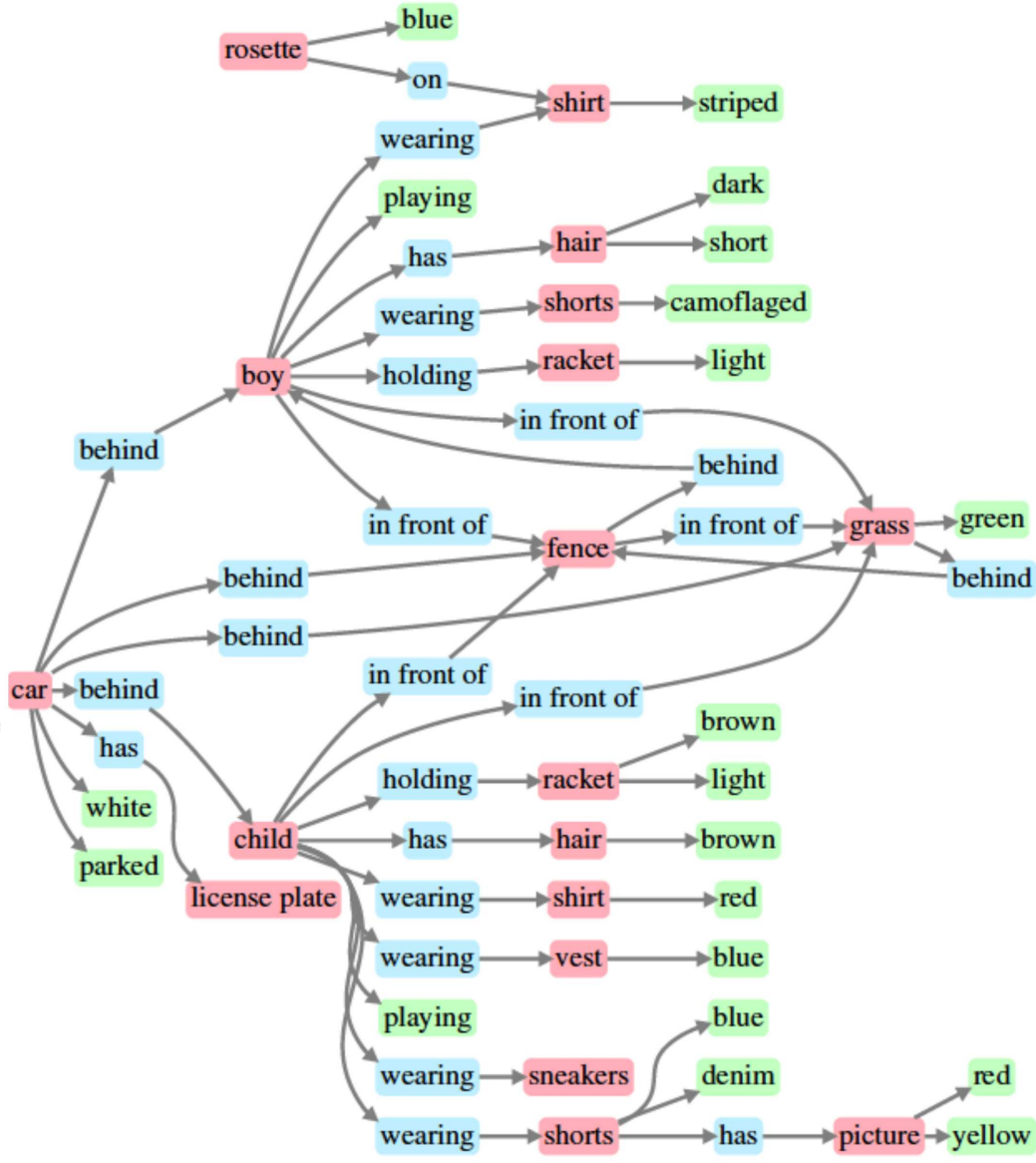


Digit recognition, AT&T labs

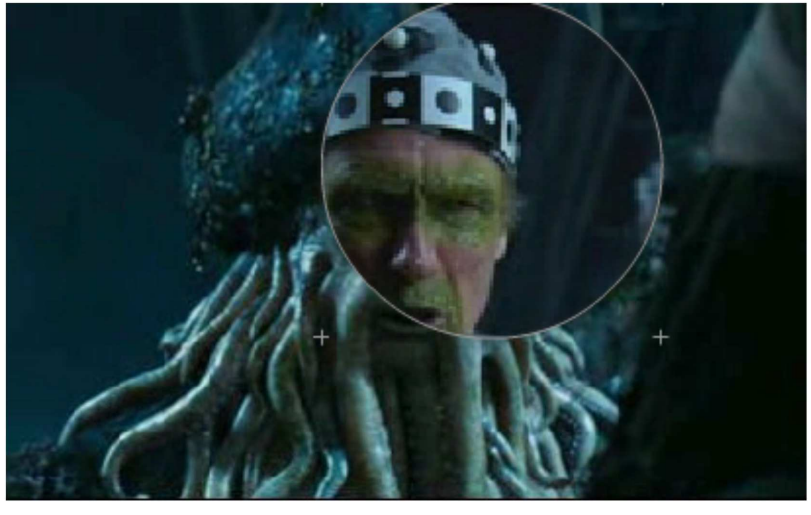
License plate readers
http://en.wikipedia.org/wiki/Automatic_number_plate_recognition

Source: S. Seitz


license plate
racket
hair
vest
shirt
child
shorts
picture
sneakers
car
hair
fence
grass
racket
rosette
shirt
boy
shorts

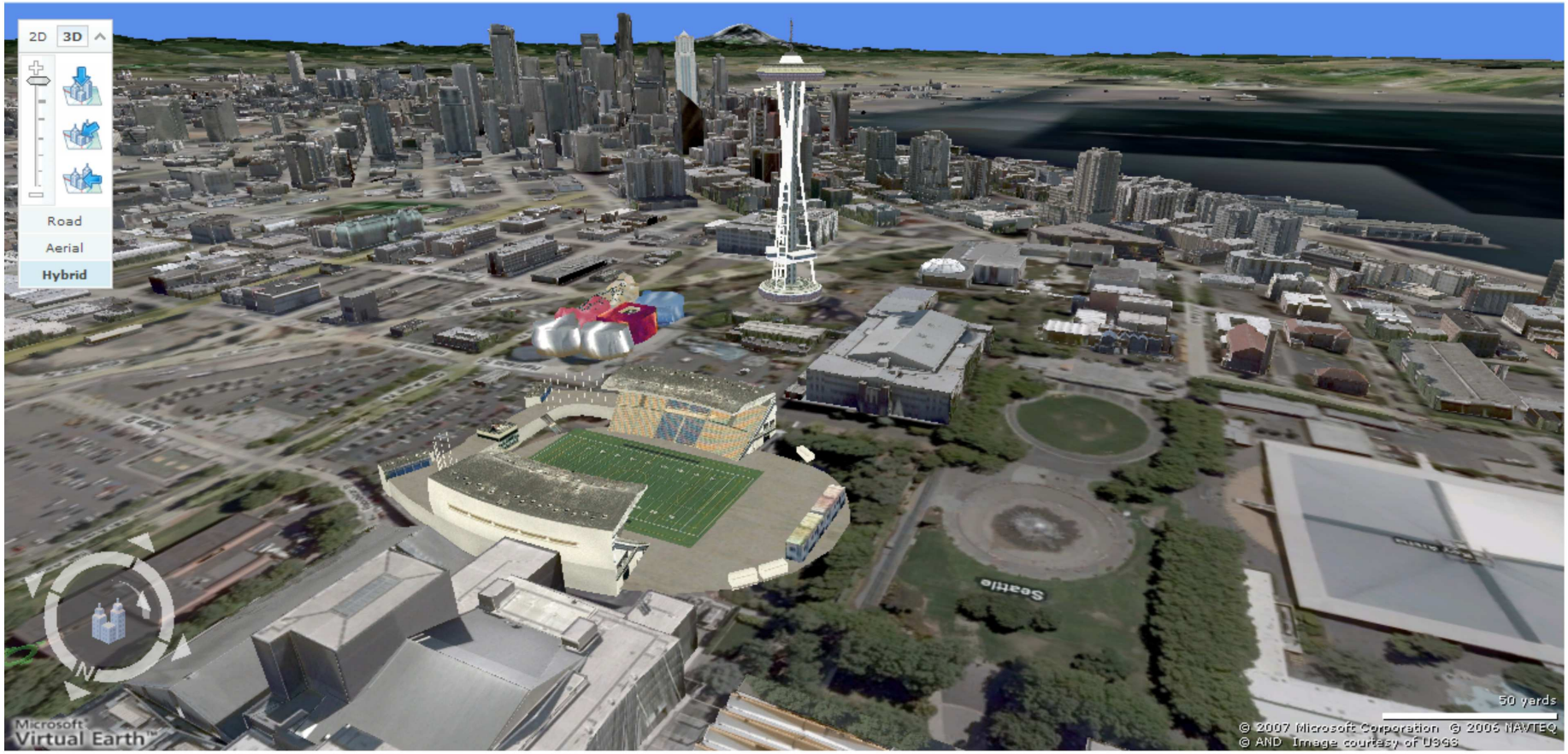

rosette
blue
on
shirt
striped
wearing
playing
dark
has
hair
short
wearing
shorts
camoflaged
boy
holding
racket
light
in front of
behind
behind
in front of
fence
in front of
grass
green
behind
behind
behind
car
behind
in front of
in front of
has
white
parked
brown
holding
racket
light
child
has
hair
brown
license plate
wearing
shirt
red
wearing
vest
blue
playing
blue
wearing
sneakers
denim
red
wearing
shorts
has
picture
yellow

Special effects: shape and motion capture



Source: S. Seitz

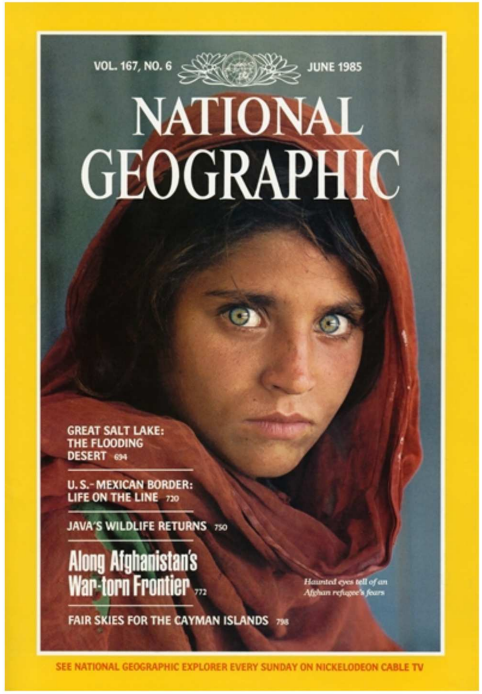
3D urban modeling



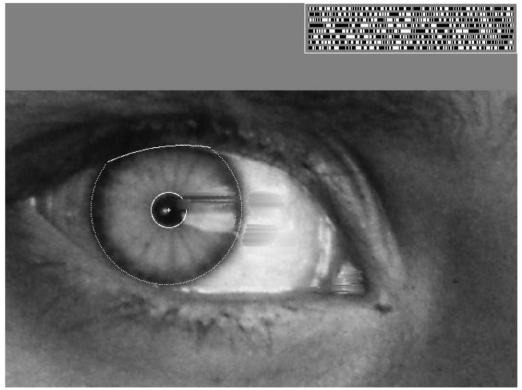
Bing maps, Google Streetview

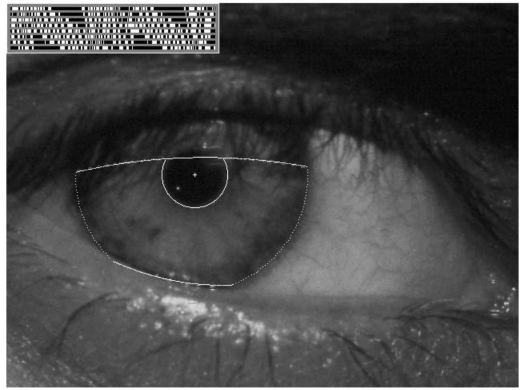
Source: S. Seitz

Biometrics



[How the Afghan Girl was Identified by Her Iris Patterns]





Source: S. Seitz

Toys and Robots

Mobile visual search: Google Goggles

Google Goggles in Action

Click the icons below to see the different ways Google Goggles can be used.

Google goggles labs

Landmark | Book | Contact Info. | Artwork | Places | Wine | Logo

Automotive safety

- Mobileye: Vision systems in high-end BMW, GM, Volvo models
  - "In mid 2010 Mobileye will launch a world's first application of full emergency braking for collision mitigation for pedestrians where vision is the key technology for detecting pedestrians."

Source: A. Shashua, S. Seitz

Vision in supermarkets

[LaneHawk by EvolutionRobotics]

"A smart camera is flush-mounted in the checkout lane, continuously watching for items. When an item is detected and recognized, the cashier verifies the quantity of items that were found under the basket, and continues to close the transaction. The item can remain under the basket, and with LaneHawk, you are assured to get paid for it… "

Source: S. Seitz

Vision-based interaction (and games)

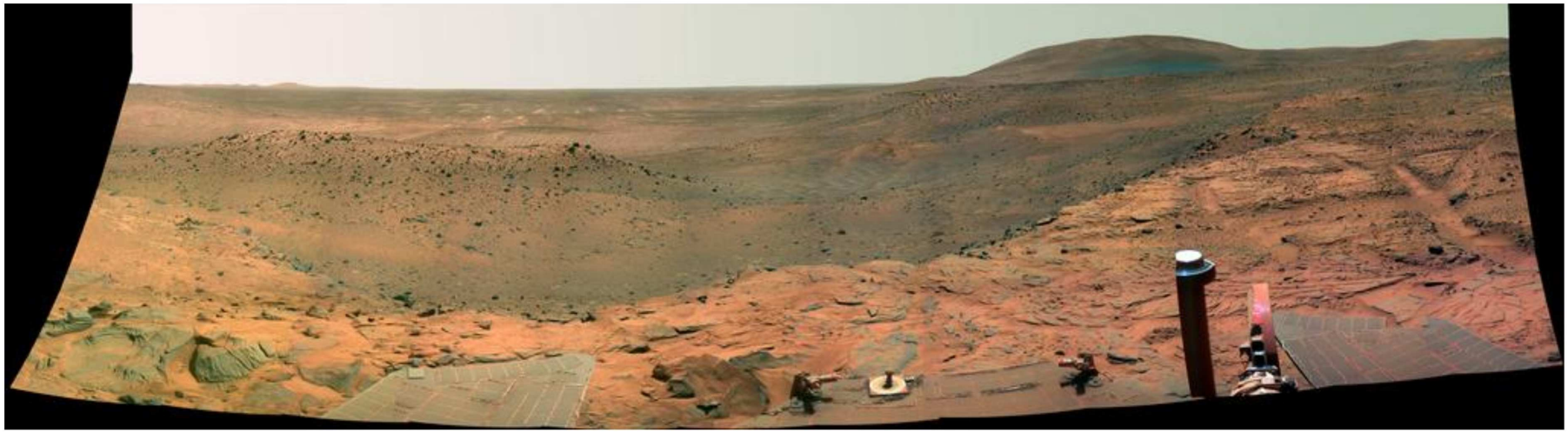
Microsoft's Kinect

Sony EyeToy

Assistive technologies

Source: S. Seitz

Vision for robotics, space exploration



NASA'S Mars Exploration Rover Spirit captured this westward view from atop a low plateau where Spirit spent the closing months of 2007.

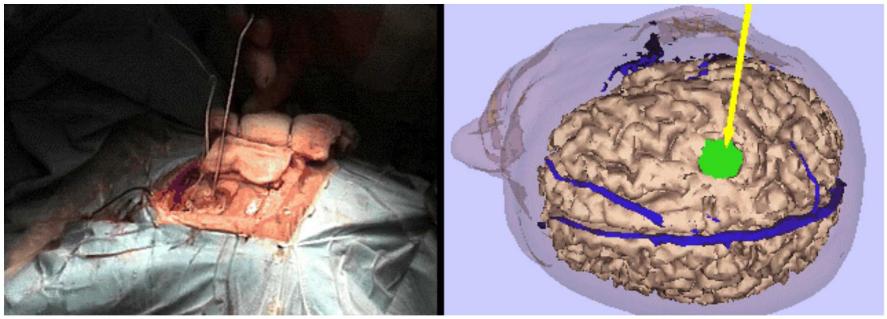
Vision systems (JPL) used for several tasks

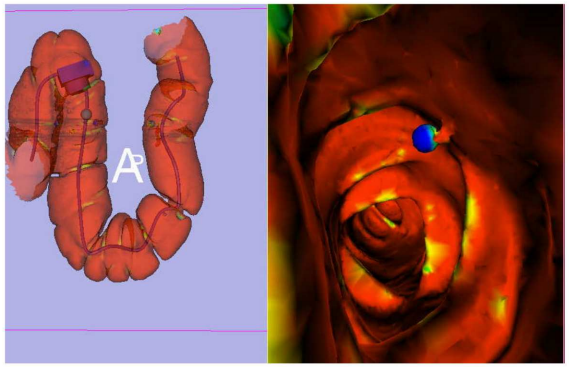
- Panorama stitching
- 3D terrain modeling
- Obstacle detection, position tracking
- For more, read "Computer Vision on Mars" by Matthies et al.

Source: S. Seitz

Medical Vision

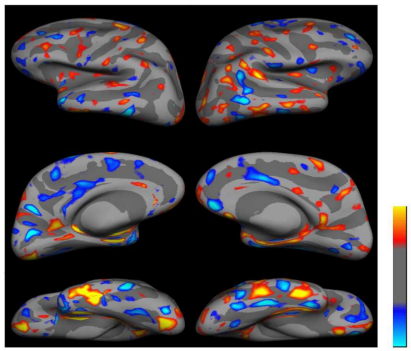
Surgical planning and navigation





Simulation

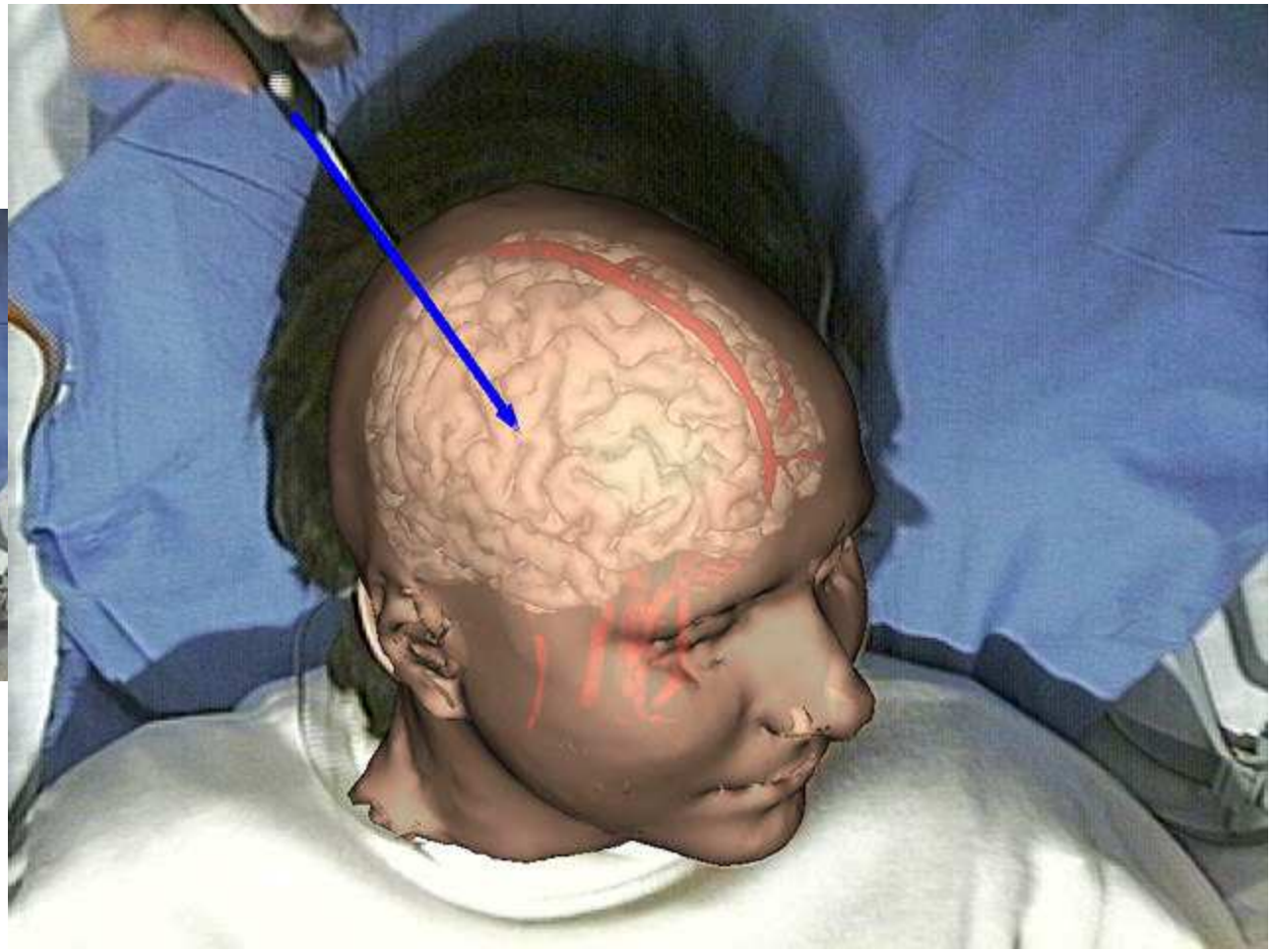
Population modeling



SOURCE

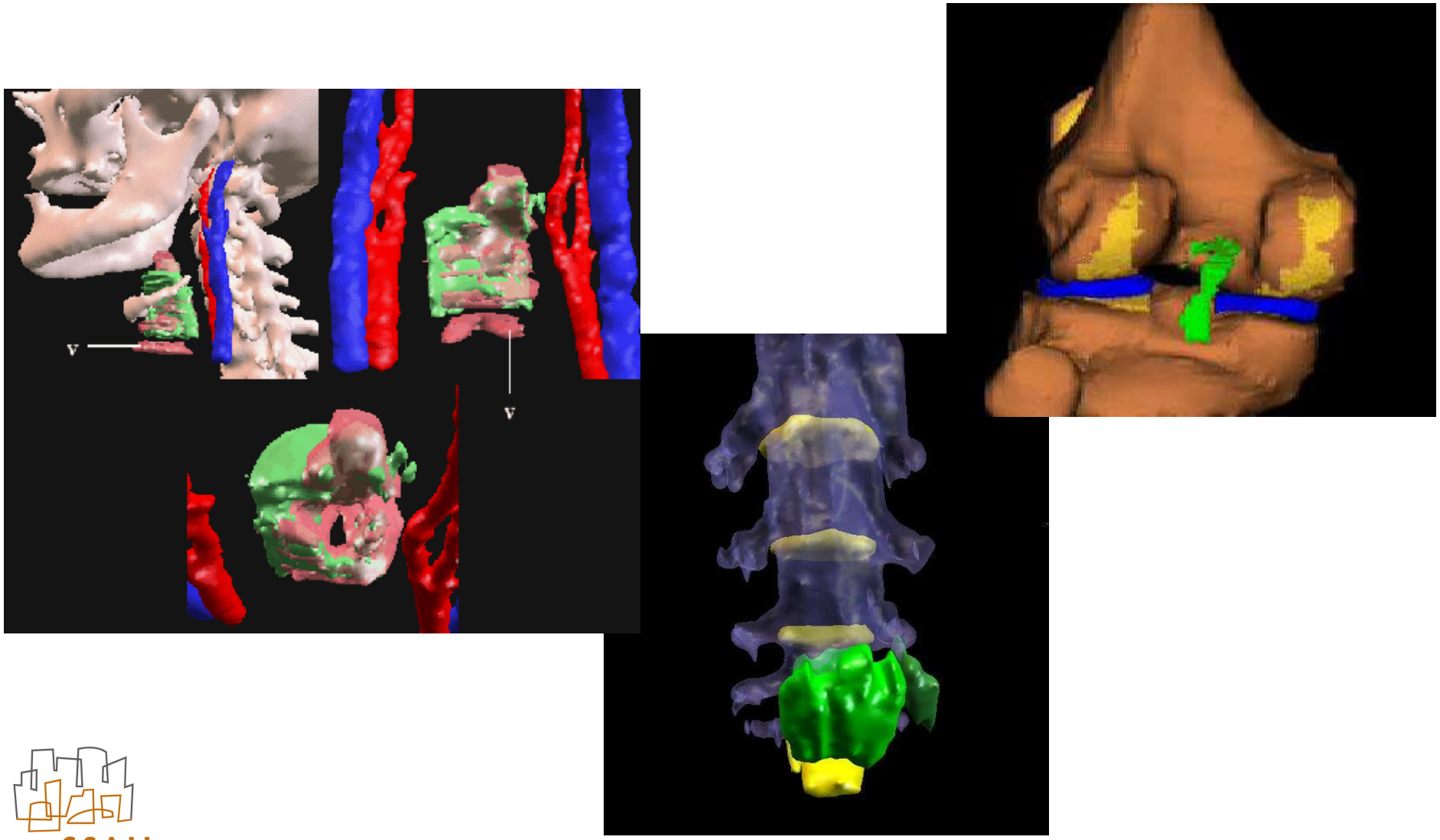
CSAIL

Image Guided Surgery: See under the surface



CSAIL

Example patient specific models



CSAIL

Big Breakthroughs!

Convolutional Neural Networks have been applied to all of these problems

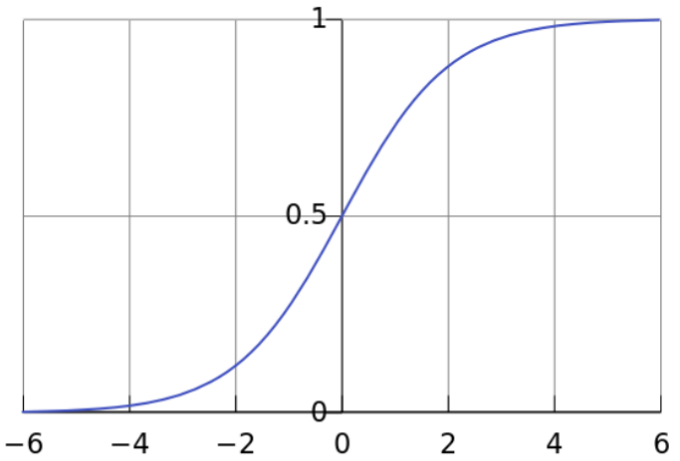
Big Breakthroughs!

Convolutional Neural Networks have
solved many of these problems

Contents

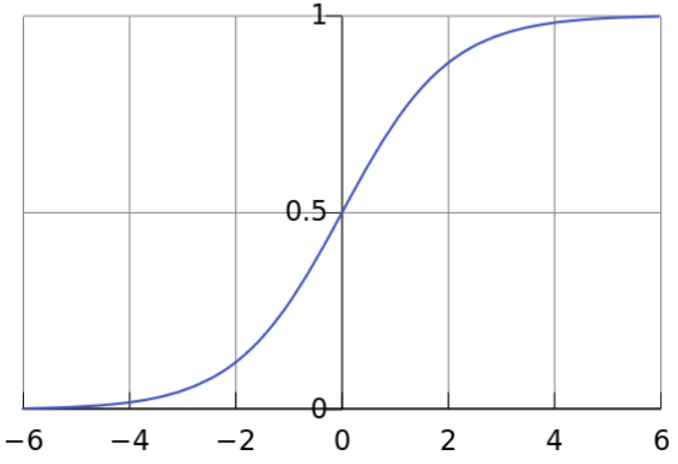
1-Layer Nets (Logistic Regression)

- Function model: $f(x) = \sigma(w^T \cdot x + b)$
  - Parameters: vector $w \in R^d$, b is scalar bias term
  - σ is a non-linearity, e.g. sigmoid: $\sigma(z) = 1/(1 + \exp(-z))$
  - For simplicity, sometimes write $f(x) = \sigma(w^T x)$ where $w = [w; b]$ and $x = [x; 1]$



1-Layer Nets (Logistic Regression)

- Function model: $f(x) = \sigma(w^T \cdot x + b)$
  - Parameters: vector $w \in R^d$, b is scalar bias term
  - σ is a non-linearity, e.g. sigmoid: $\sigma(z) = 1/(1 + \exp(-z))$
  - For simplicity, sometimes write $f(x) = \sigma(w^T x)$ where $w = [w; b]$ and $x = [x; 1]$



- Non-linearity will be important in expressiveness multi-layer nets. Other non-linearities, e.g., $tanh(z) = (e^z - e^{-z})/(e^z + e^{-z})$

A bit of history...

The **Mark I Perceptron** machine was the first implementation of the perceptron algorithm.

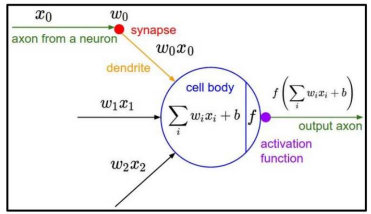
The machine was connected to a camera that used 20×20 cadmium sulfide photocells to produce a 400-pixel image.

recognized
letters of the alphabet

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

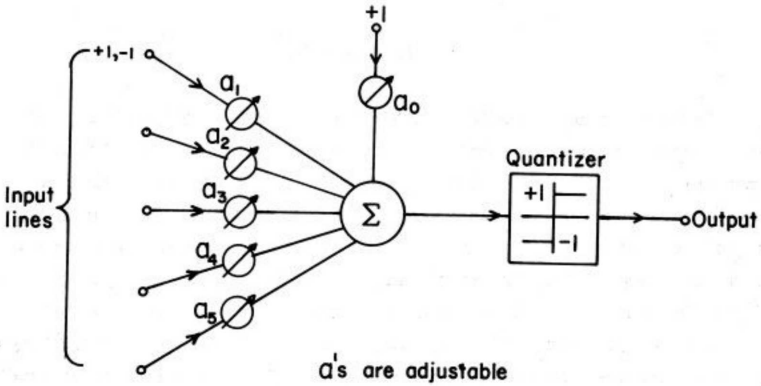
update rule:

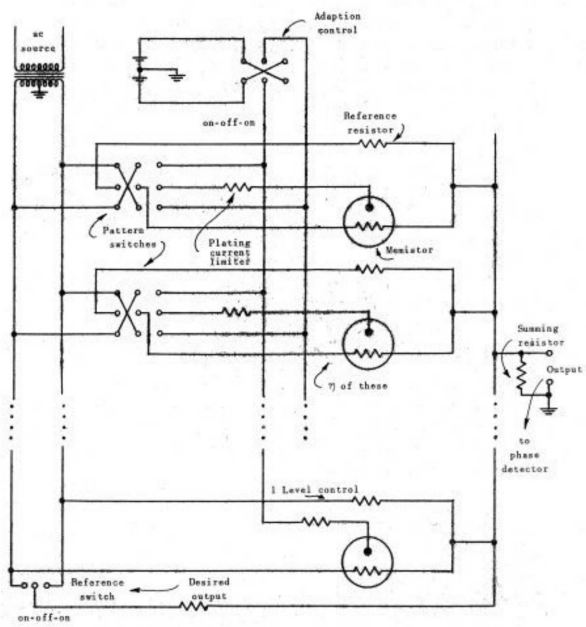
$$w_i(t+1) = w_i(t) + \alpha(d_j - y_j(t))x_{j,i},$$



Frank Rosenblatt, ~1957: Perceptron

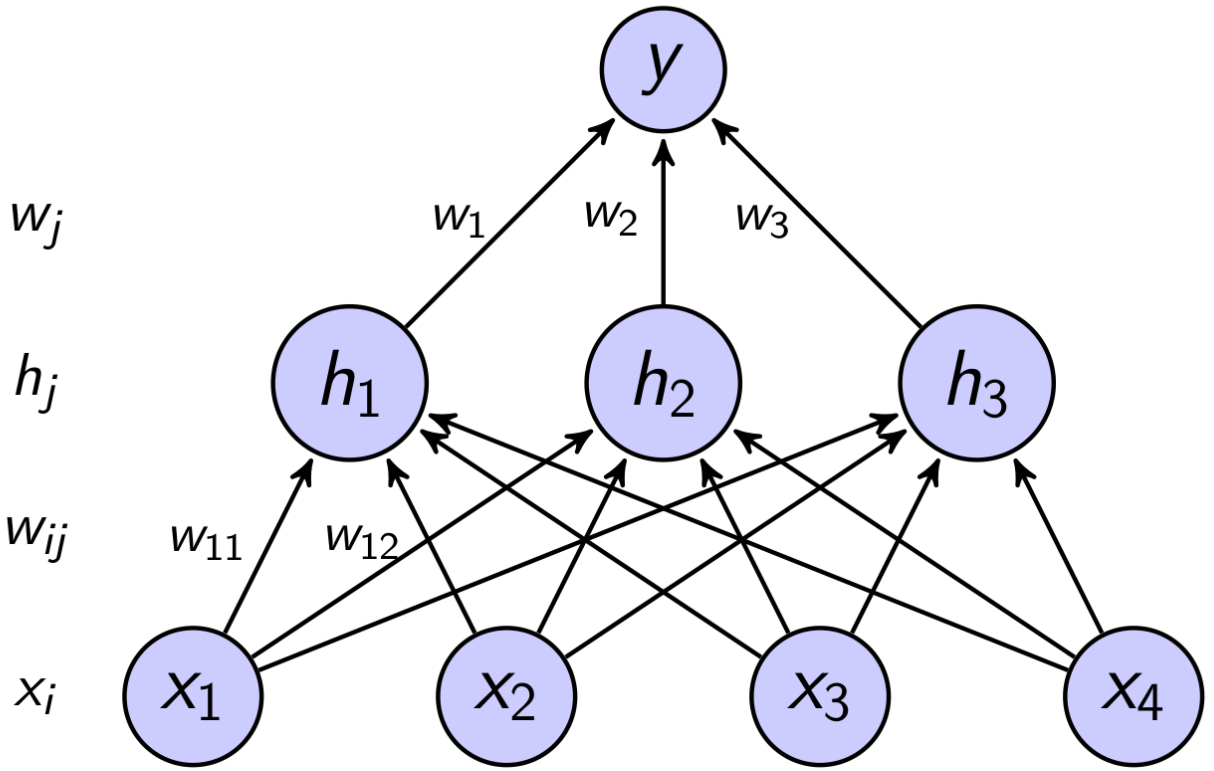
A bit of history...





Widrow and Hoff, ~1960: Adaline/Madaline

2-Layer Neural Networks

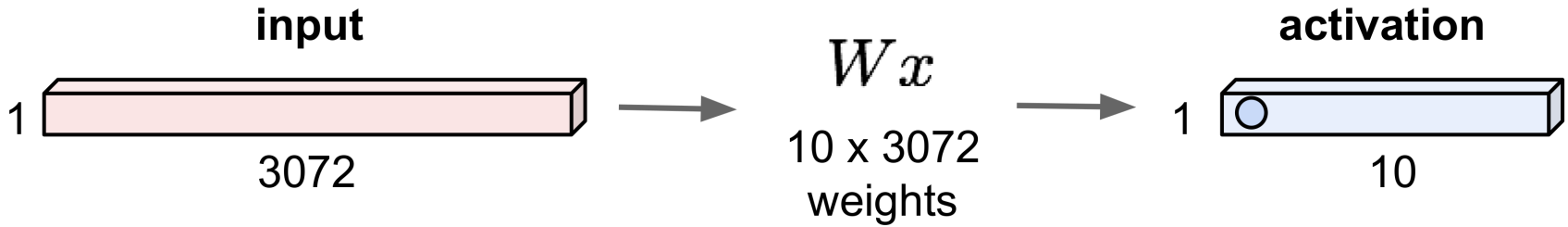


$f(x) = \sigma(\sum_j w_j \cdot h_j) = \sigma(\sum_j w_j \cdot \sigma(\sum_i w_{ij} x_i))$

Hidden units h_j's can be viewed as new "features" from combining x_i's

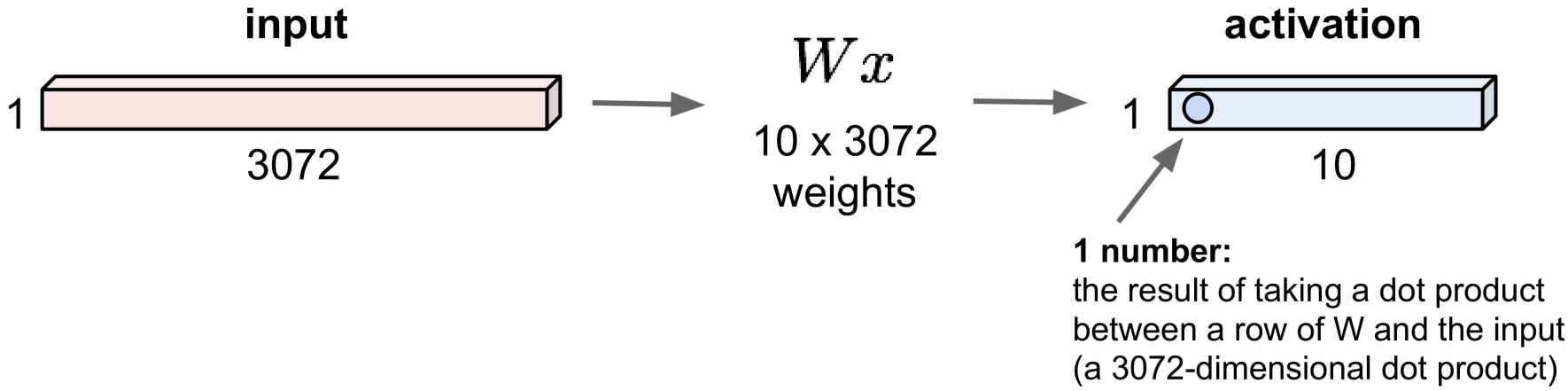
Fully Connected Layer

32x32x3 image -> stretch to 3072 x 1



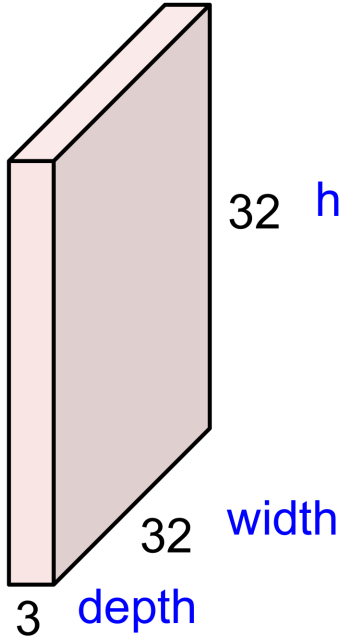
Fully Connected Layer

32x32x3 image -> stretch to 3072 x 1



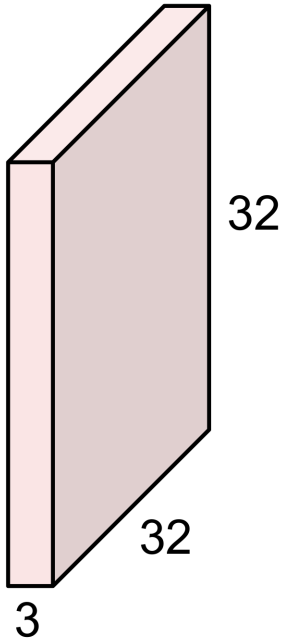
Convolution Layer

32x32x3 image -> preserve spatial structure

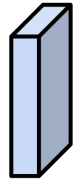


Convolution Layer

32x32x3 image



5x5x3 filter

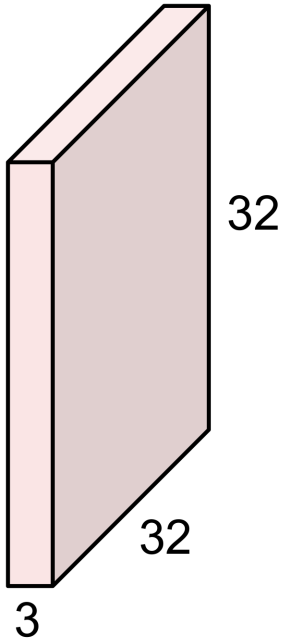


Convolve the filter with the image
i.e. “slide over the image spatially,
computing dot products”

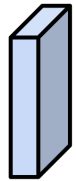
Convolution Layer

Filters always extend the full depth of the input volume

32x32x3 image

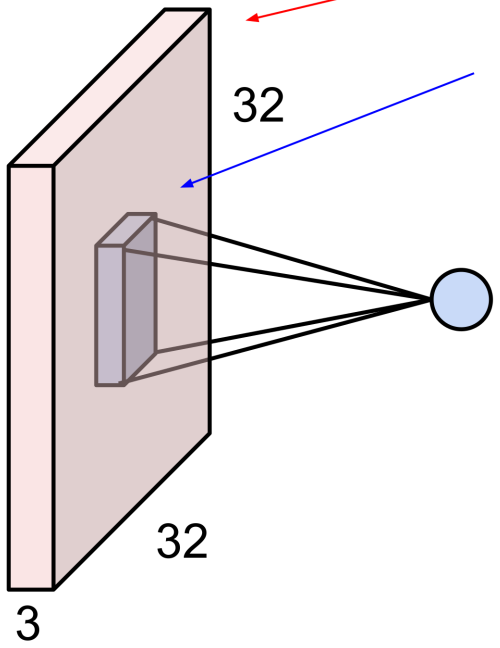


5x5x3 filter



Convolve the filter with the image i.e. “slide over the image spatially, computing dot products”

Convolution Layer



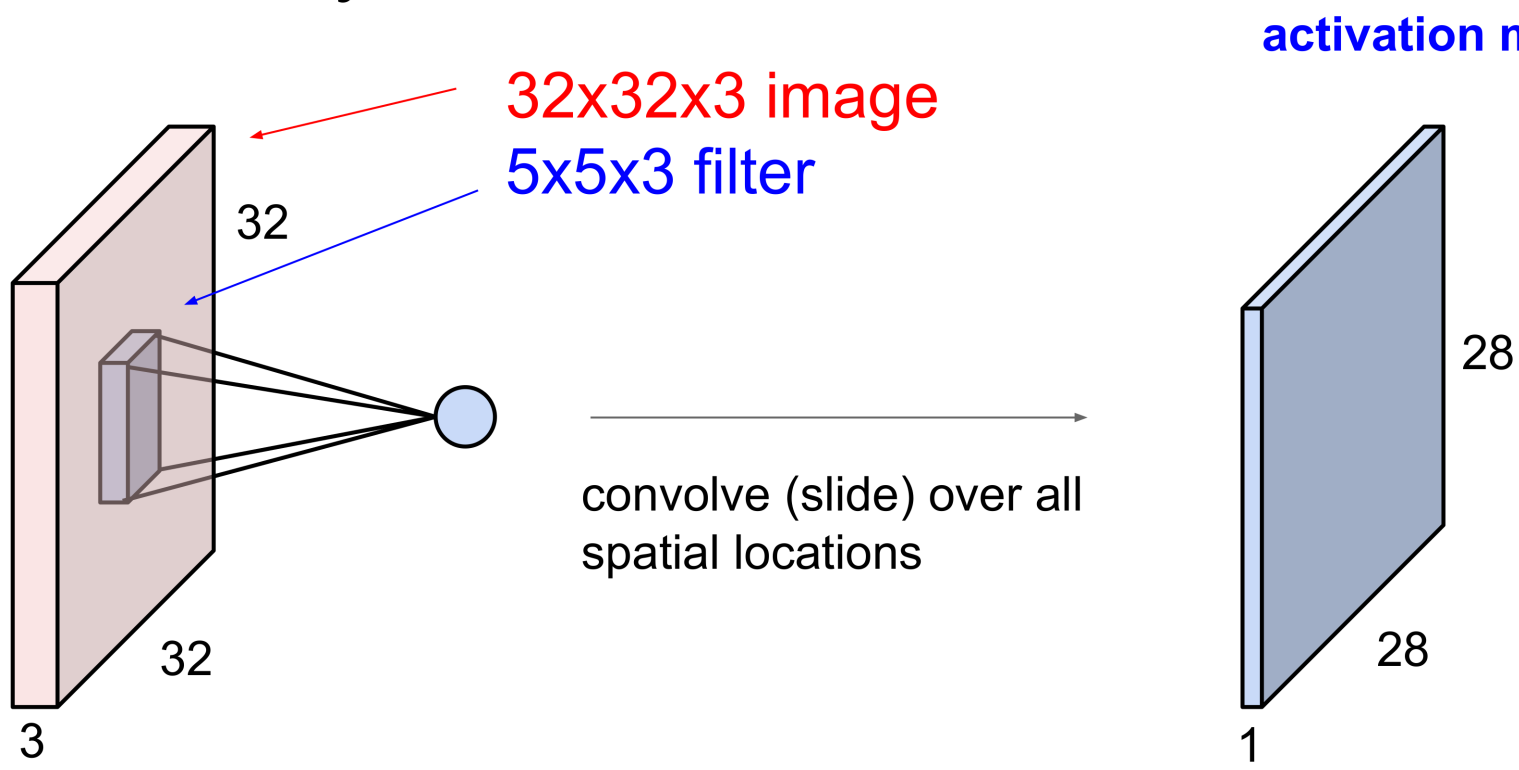
32x32x3 image

5x5x3 filter w

1 number:
the result of taking a dot product between the filter and a small 5x5x3 chunk of the image
(i.e. 5*5*3 = 75-dimensional dot product + bias)

$$w^T x + b$$

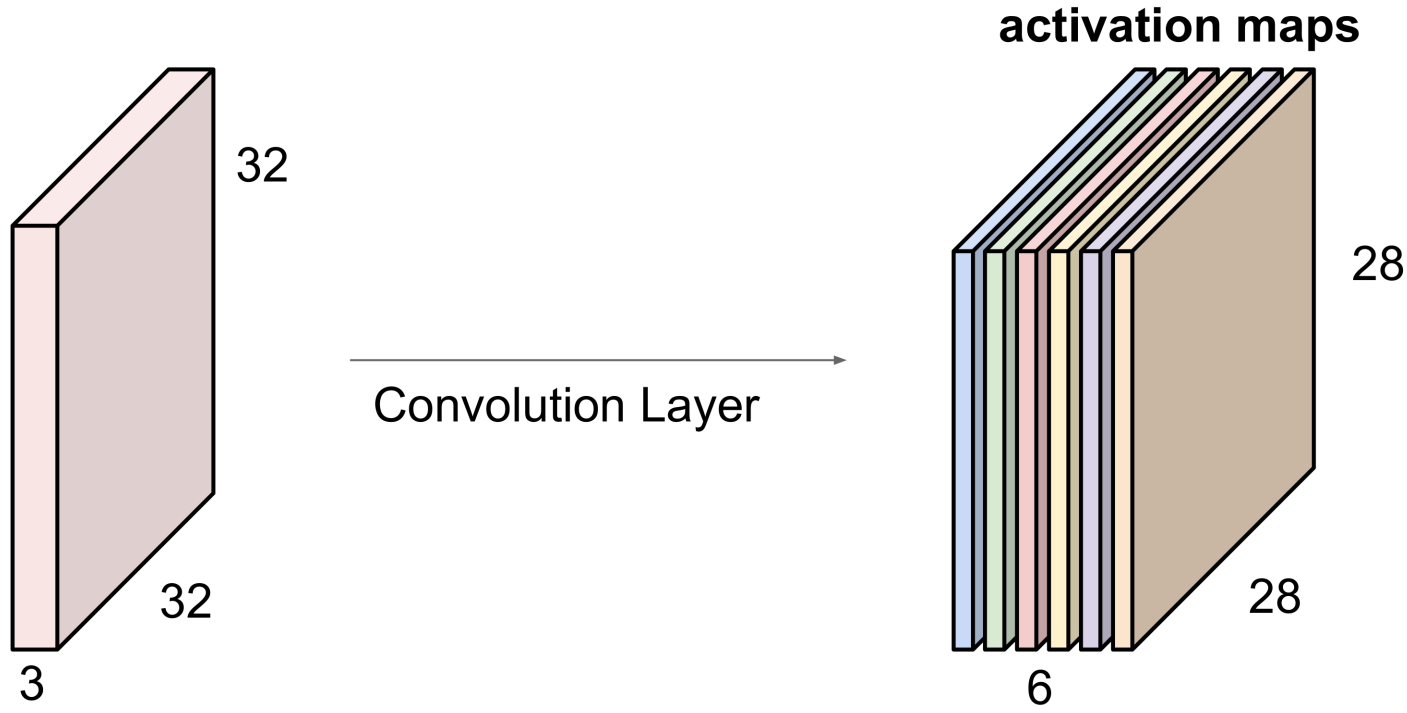
Convolution Layer



Convolution Layer

consider a second, green filter

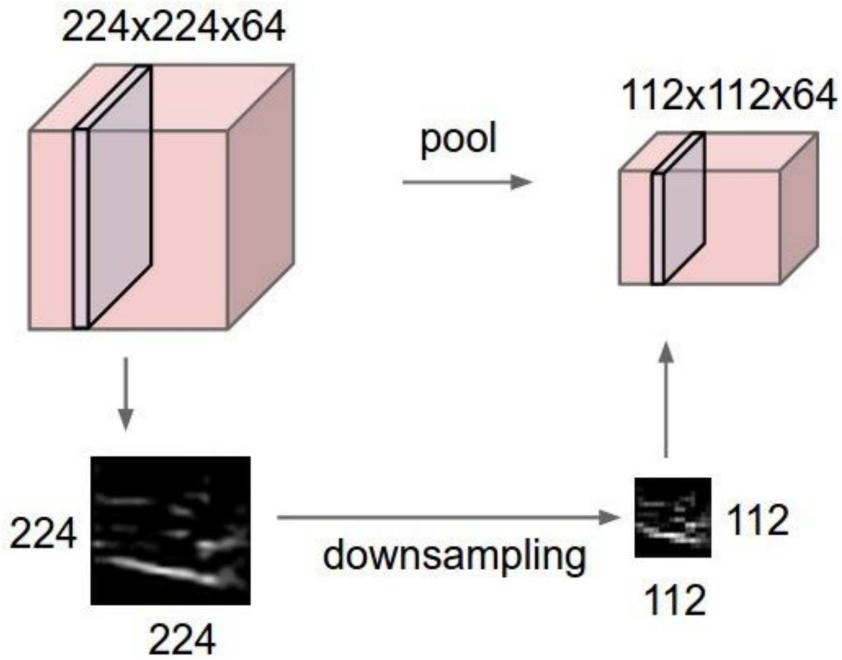
For example, if we had 6 5x5 filters, we'll get 6 separate activation maps:



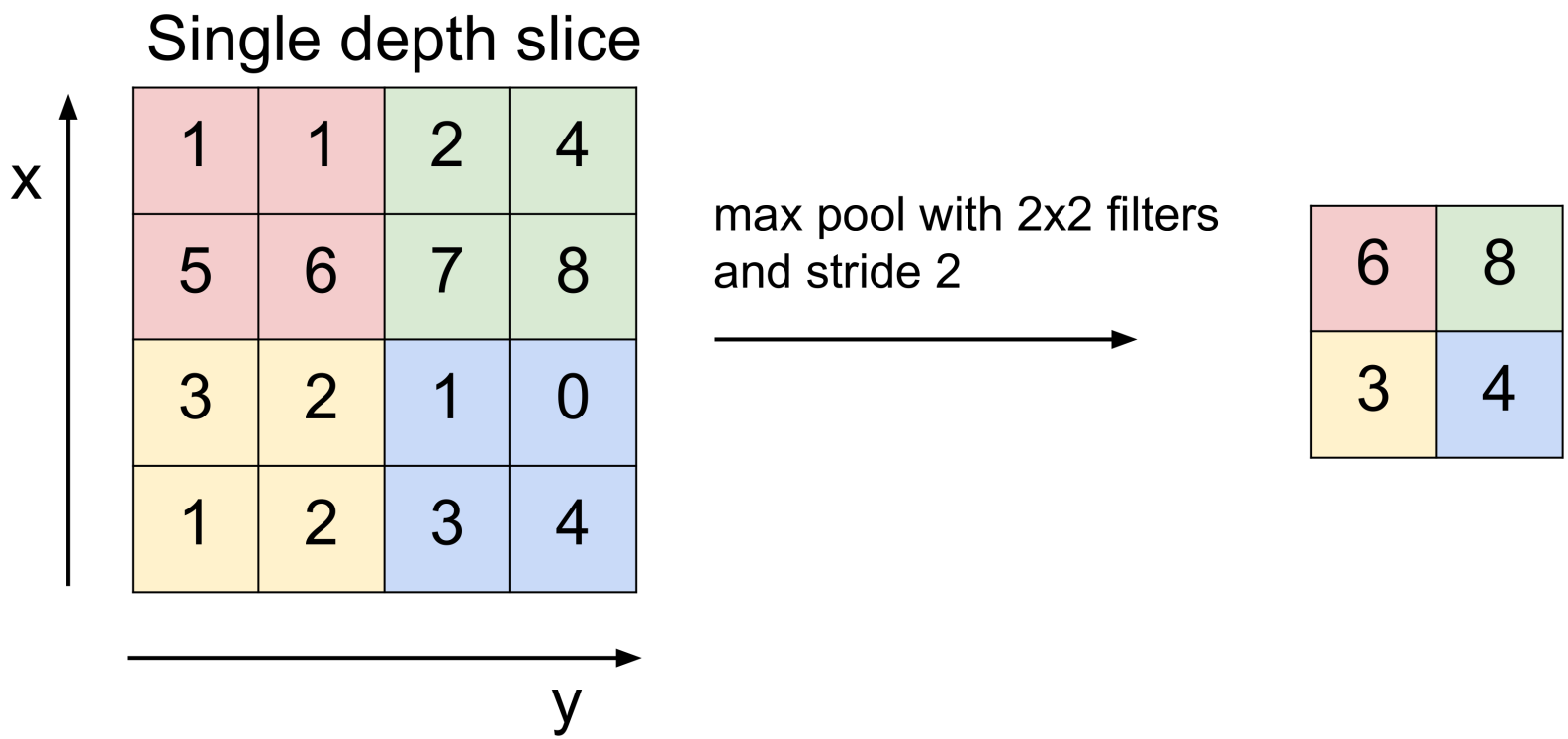
We stack these up to get a "new image" of size 28x28x6!

Pooling layer

- makes the representations smaller and more manageable
- operates over each activation map independently:

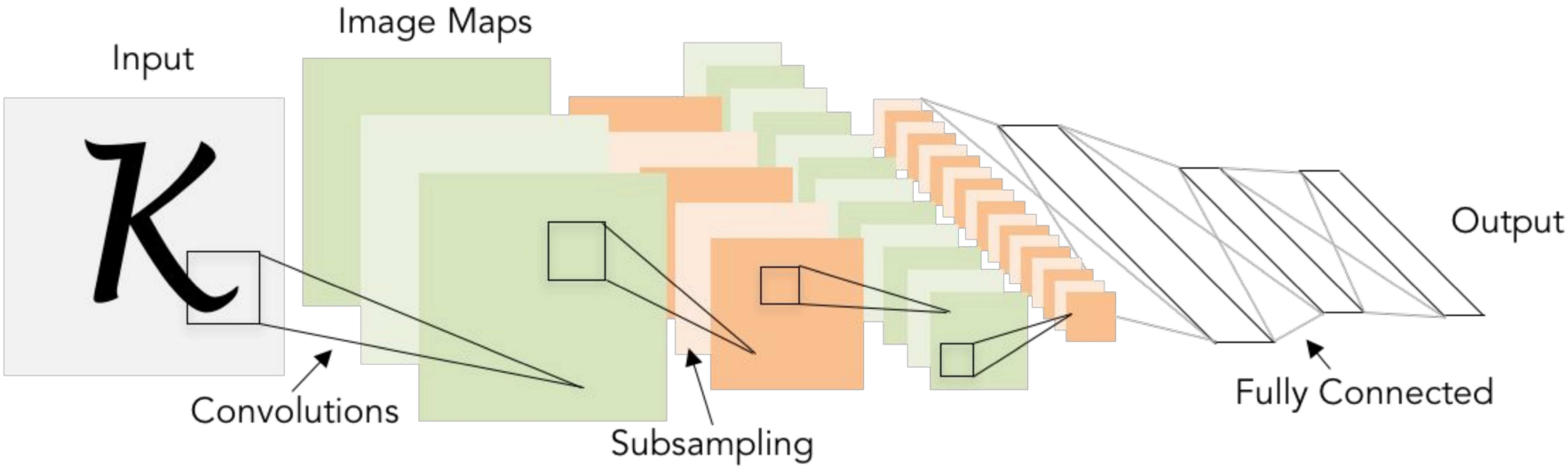


MAX POOLING


Single depth slice
x
y
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

LeNet-5

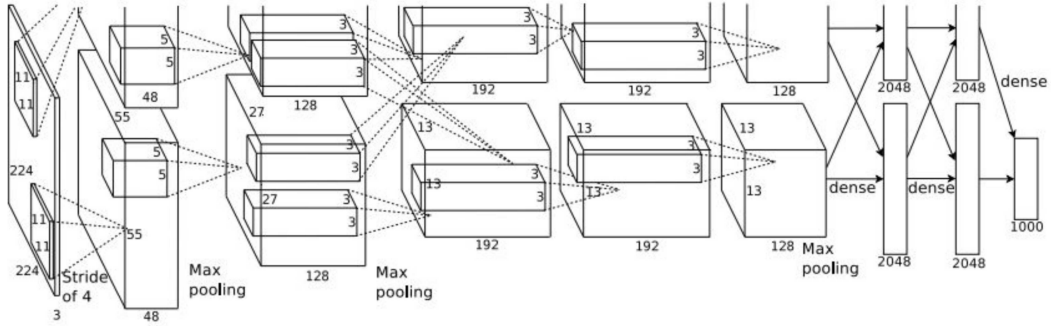
[LeCun et al., 1998]



Conv filters were 5x5, applied at stride 1
Subsampling (Pooling) layers were 2x2 applied at stride 2
i.e. architecture is [CONV-POOL-CONV-POOL-FC-FC]

AlexNet

[Krizhevsky et al. 2012]

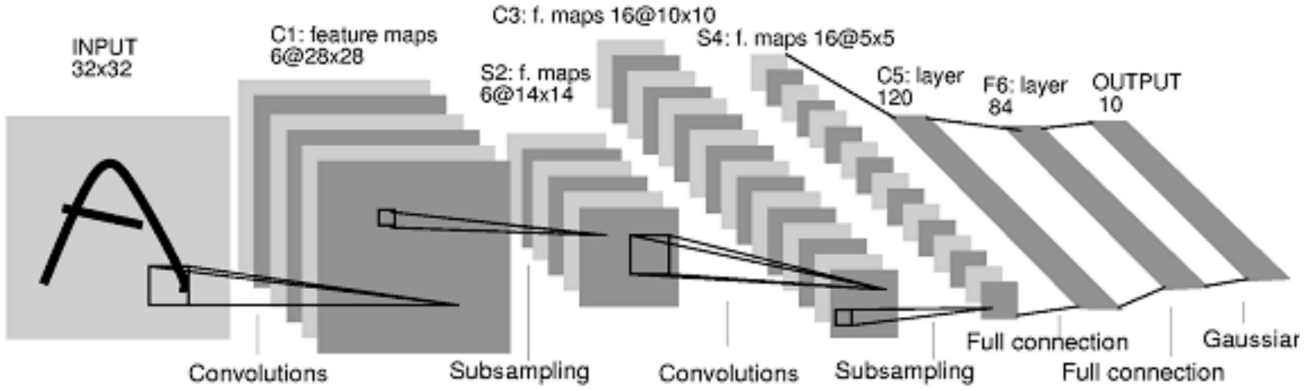


Architecture:

CONV1
MAX POOL1
NORM1
CONV2
MAX POOL2
NORM2
CONV3
CONV4
CONV5
Max POOL3
FC6
FC7
FC8

1998

LeCun et al.



# of transistors

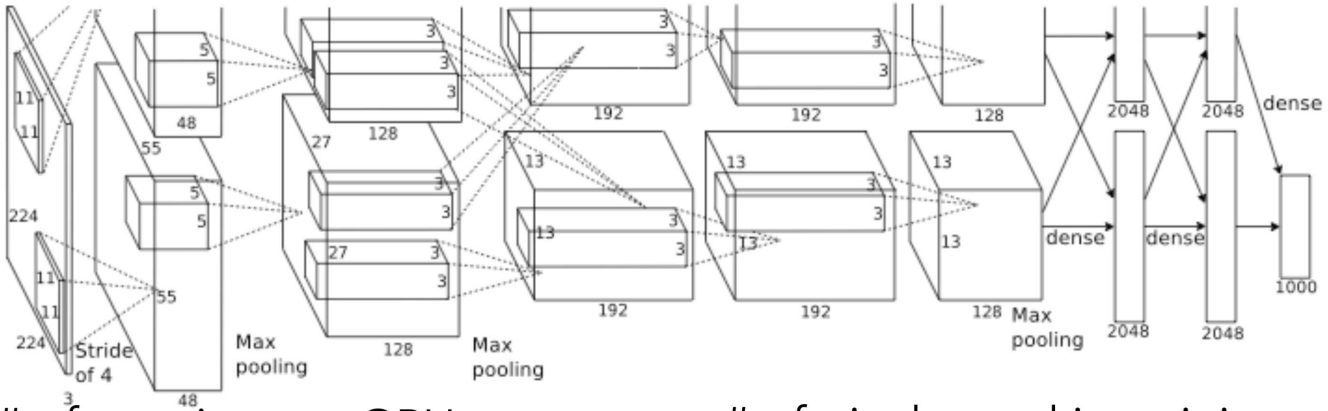
10^6

# of pixels used in training

10^7 NIST

2012

Krizhevsky et al.



# of transistors

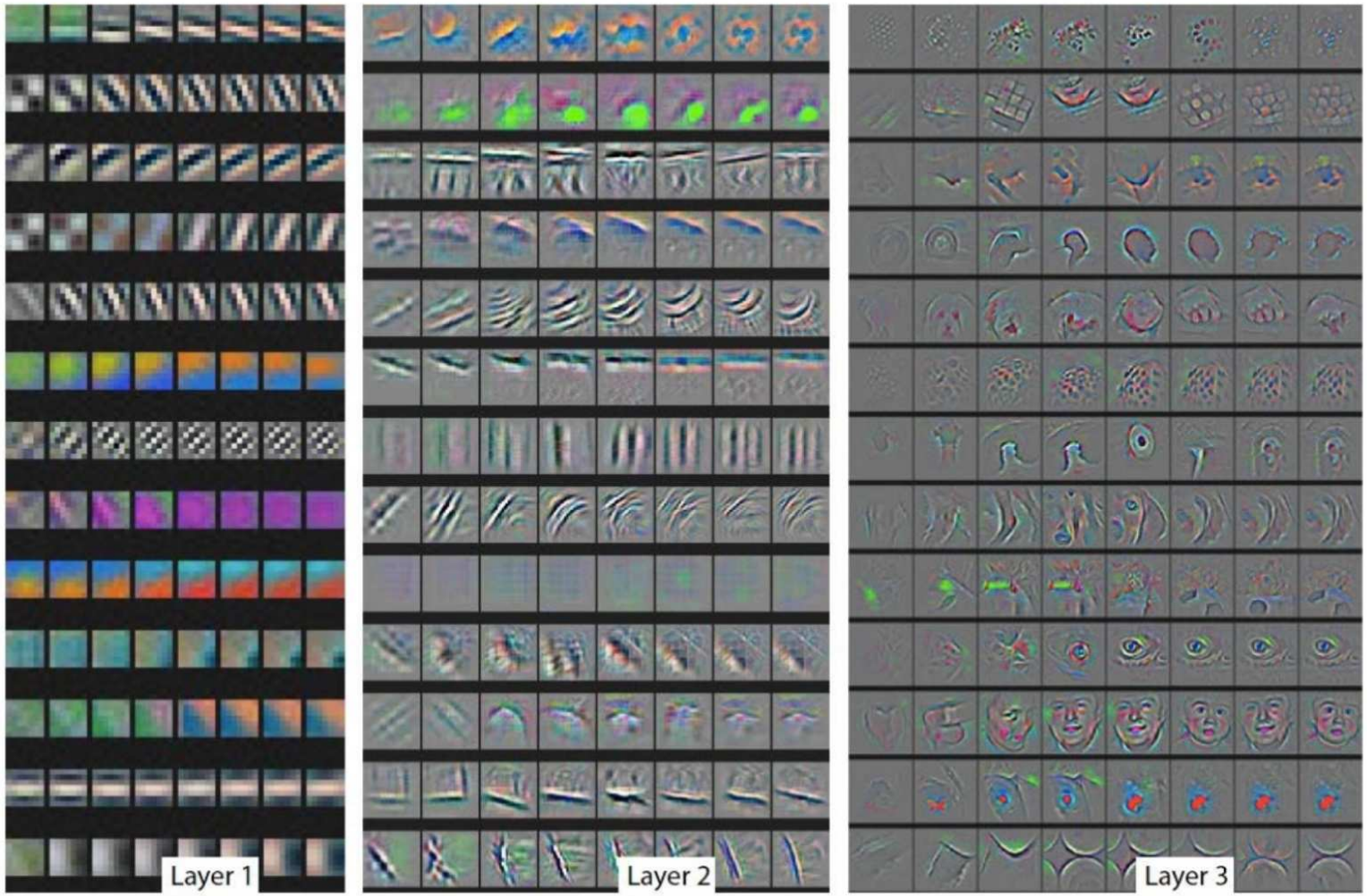
10^9

GPUs

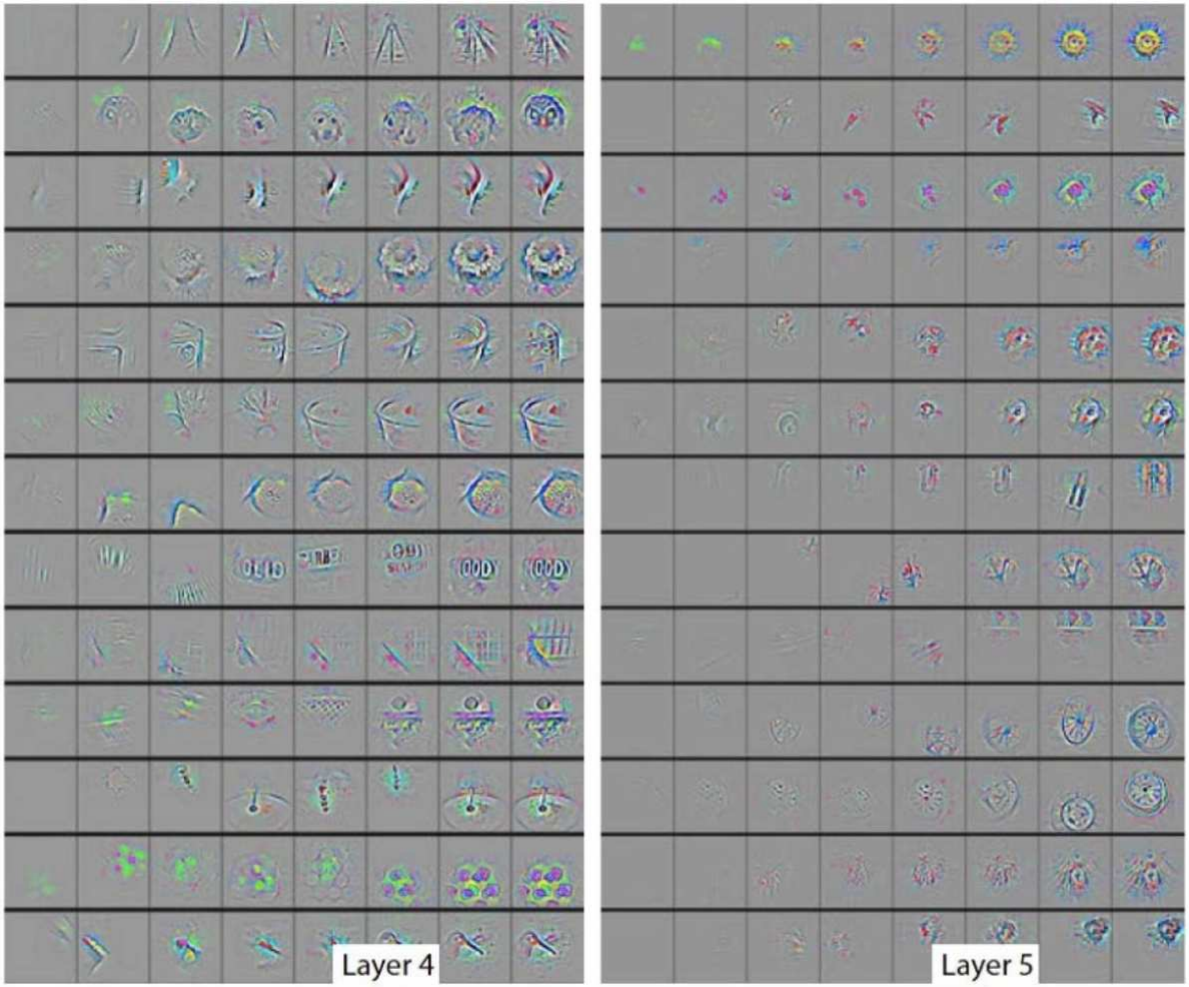
# of pixels used in training

10^{14} IMAGENET

Evolution of Filters



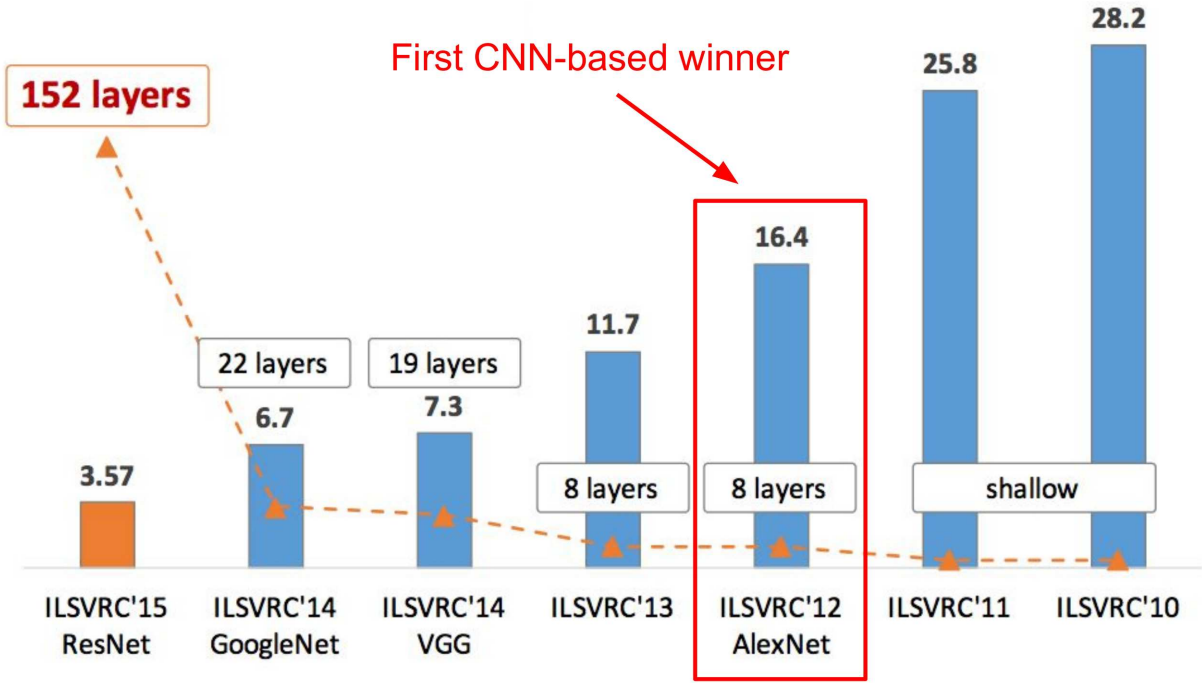
Evolution of Filters



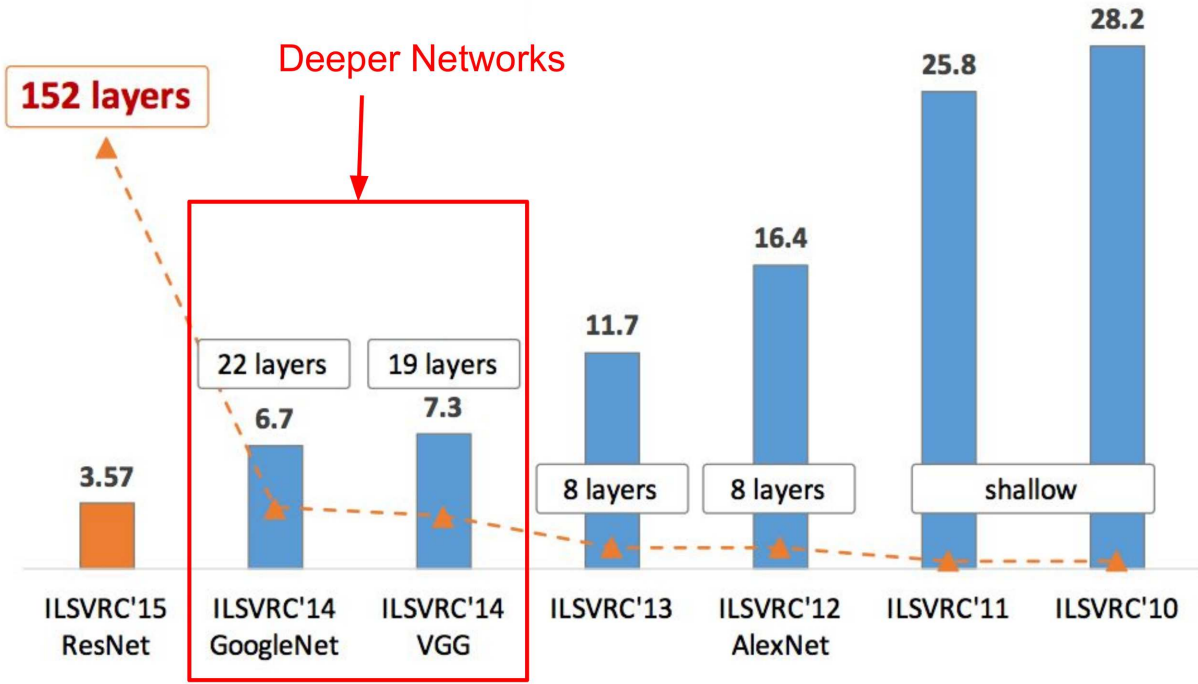
Caveat?

Contents

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners



ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners

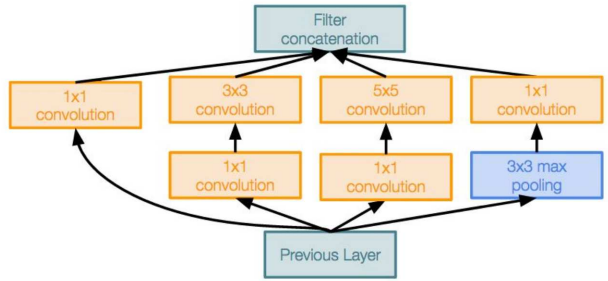


Case Study: GoogLeNet

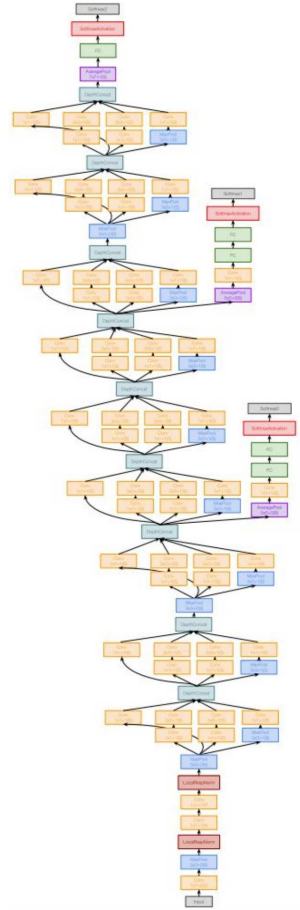
[Szegedy et al., 2014]

Deeper networks, with computational efficiency

- 22 layers
- Efficient “Inception” module
- No FC layers
- Only 5 million parameters!
  12x less than AlexNet
- ILSVRC’14 classification winner
  (6.7% top 5 error)



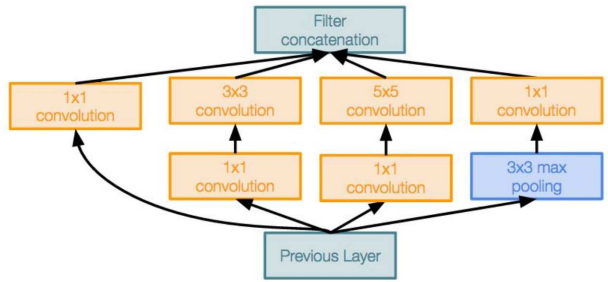
Inception module



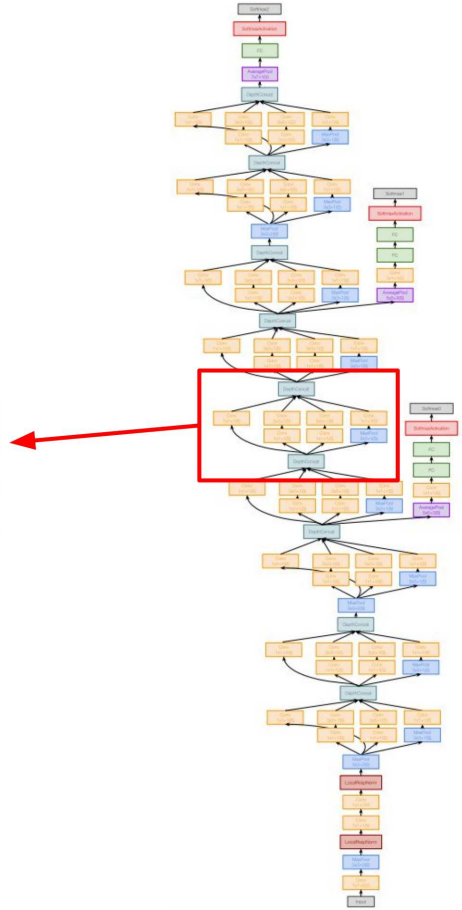
Case Study: GoogLeNet

[Szegedy et al., 2014]

"Inception module": design a good local network topology (network within a network) and then stack these modules on top of each other



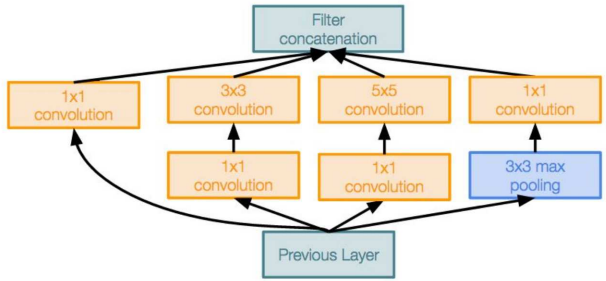
Inception module



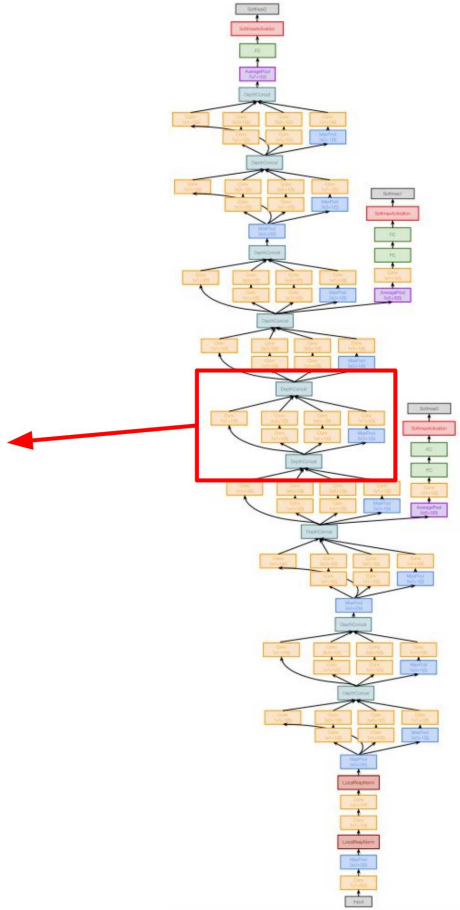
Case Study: GoogLeNet

[Szegedy et al., 2014]

Stack Inception modules with dimension reduction on top of each other



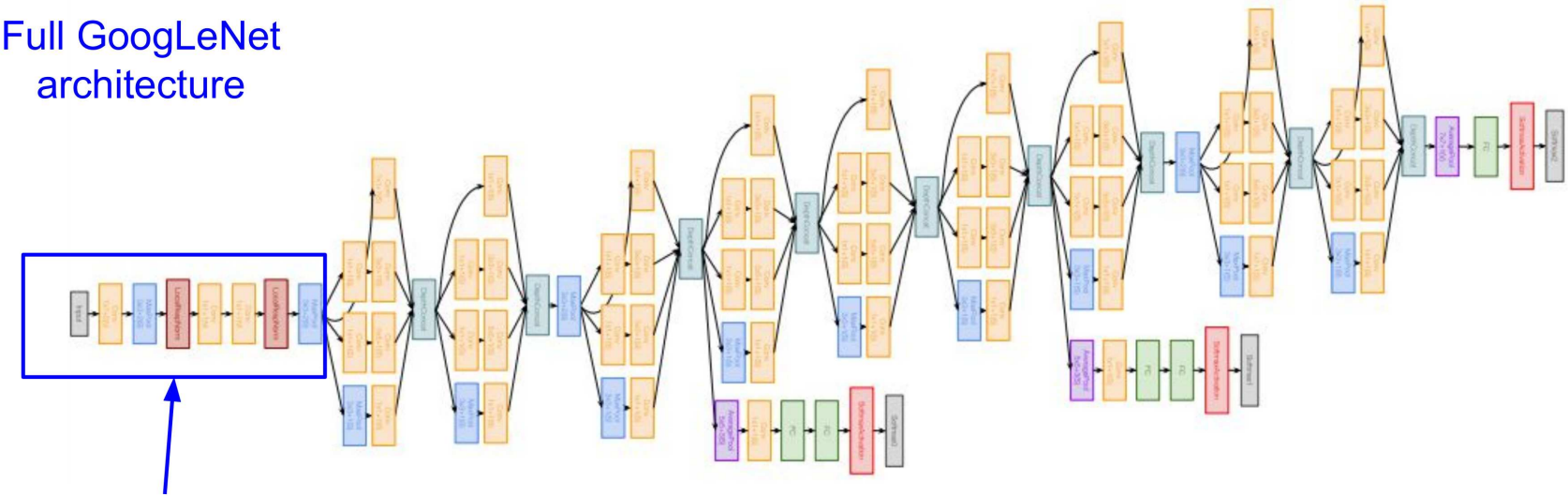
Inception module



Case Study: GoogLeNet

[Szegedy et al., 2014]

Full GoogLeNet architecture

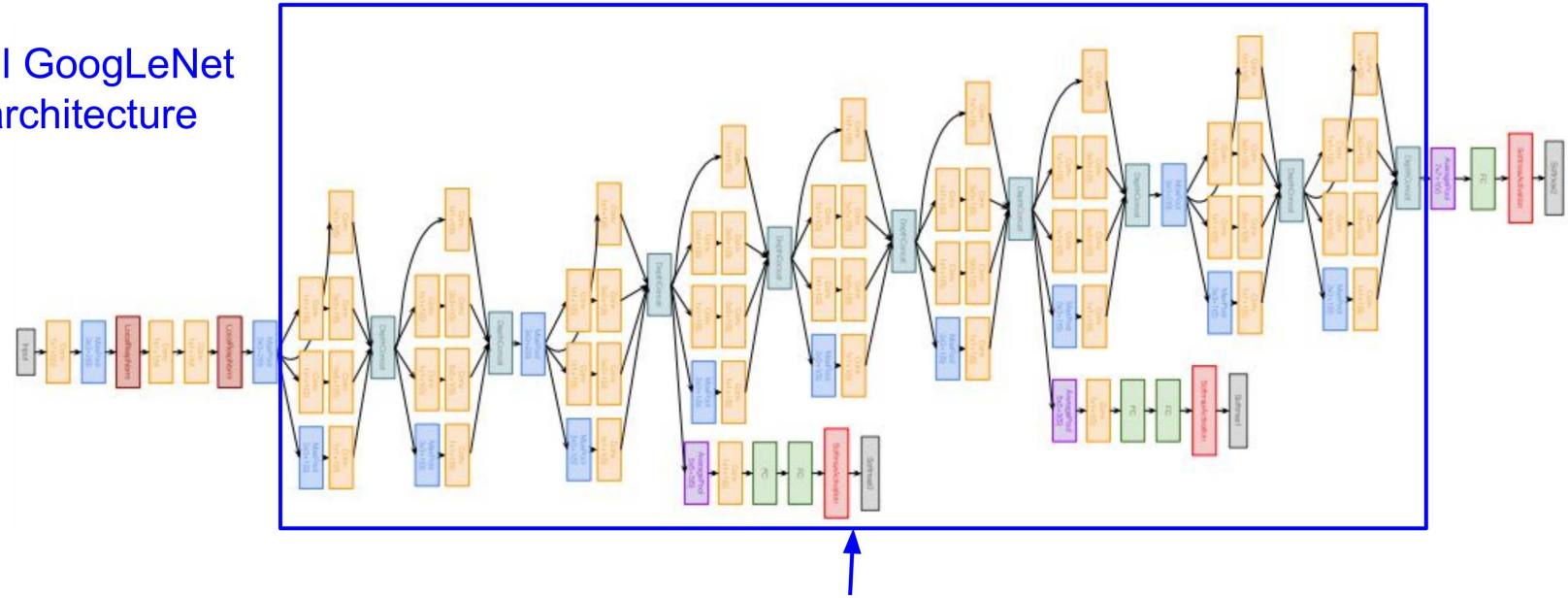


Stem Network: Conv-Pool- 2x Conv-Pool

Case Study: GoogLeNet

[Szegedy et al., 2014]

Full GoogLeNet architecture

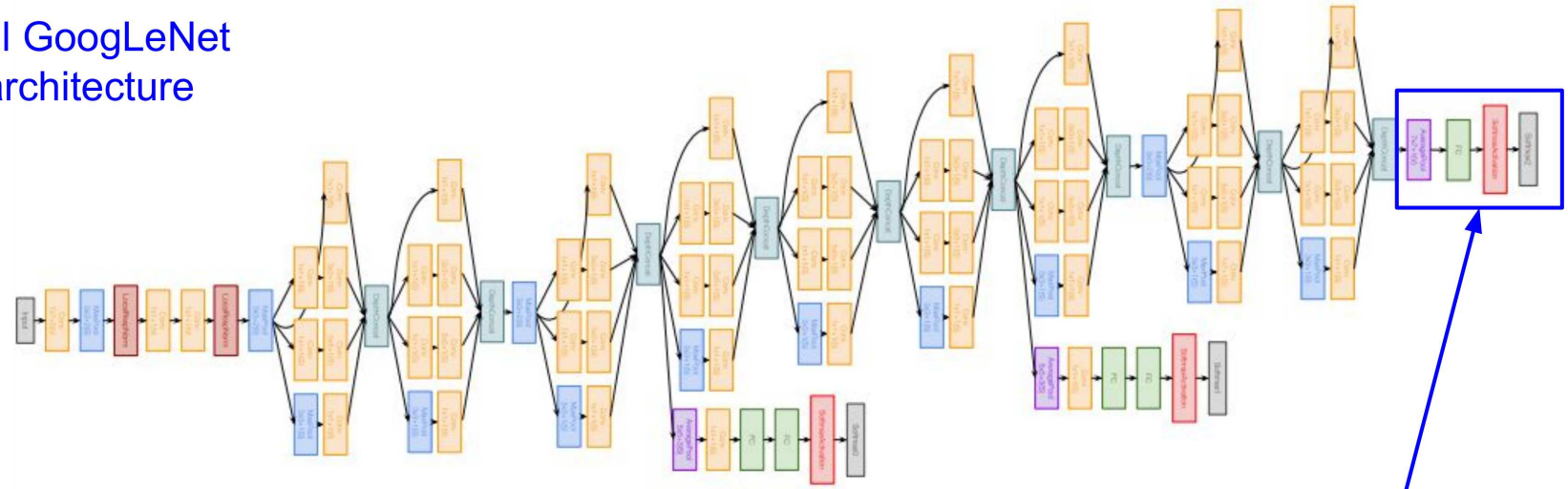


Stacked Inception Modules

Case Study: GoogLeNet

[Szegedy et al., 2014]

Full GoogLeNet architecture

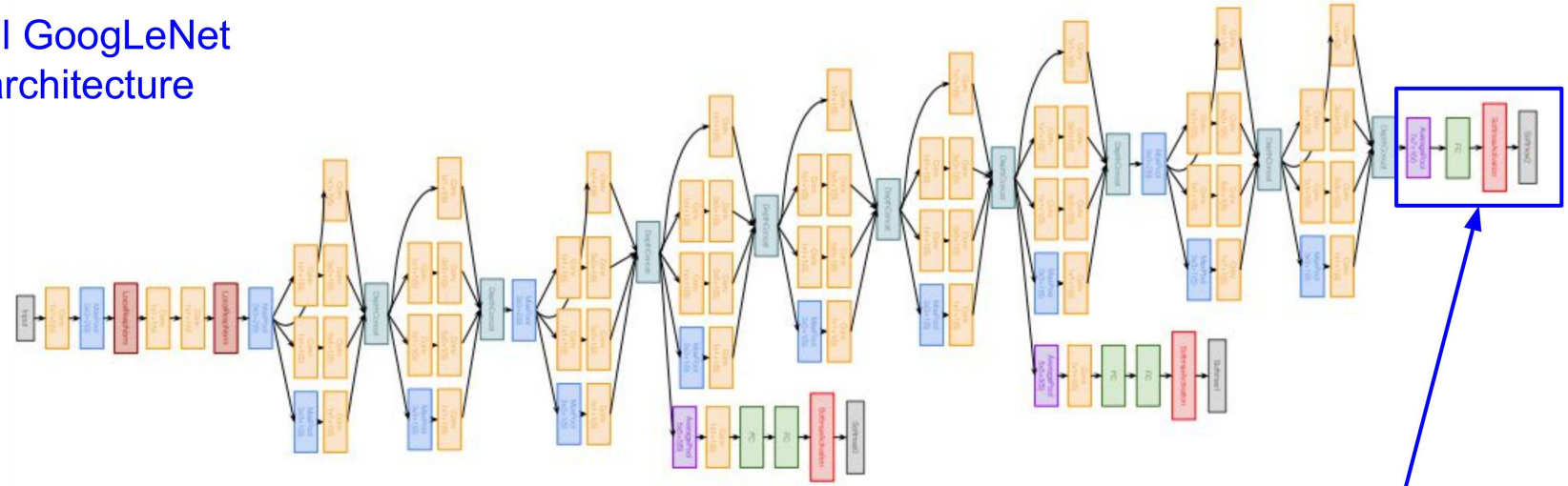


Classifier output

Case Study: GoogLeNet

[Szegedy et al., 2014]

Full GoogLeNet architecture

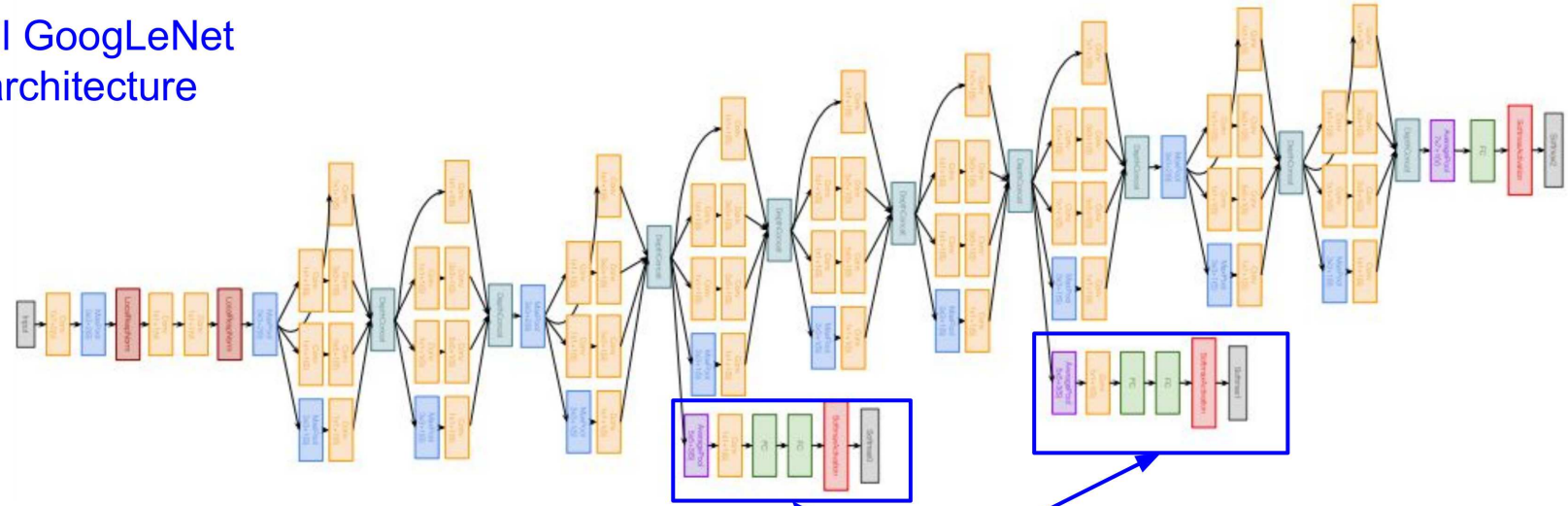


Classifier output
(removed expensive FC layers!)

Case Study: GoogLeNet

[Szegedy et al., 2014]

Full GoogLeNet architecture

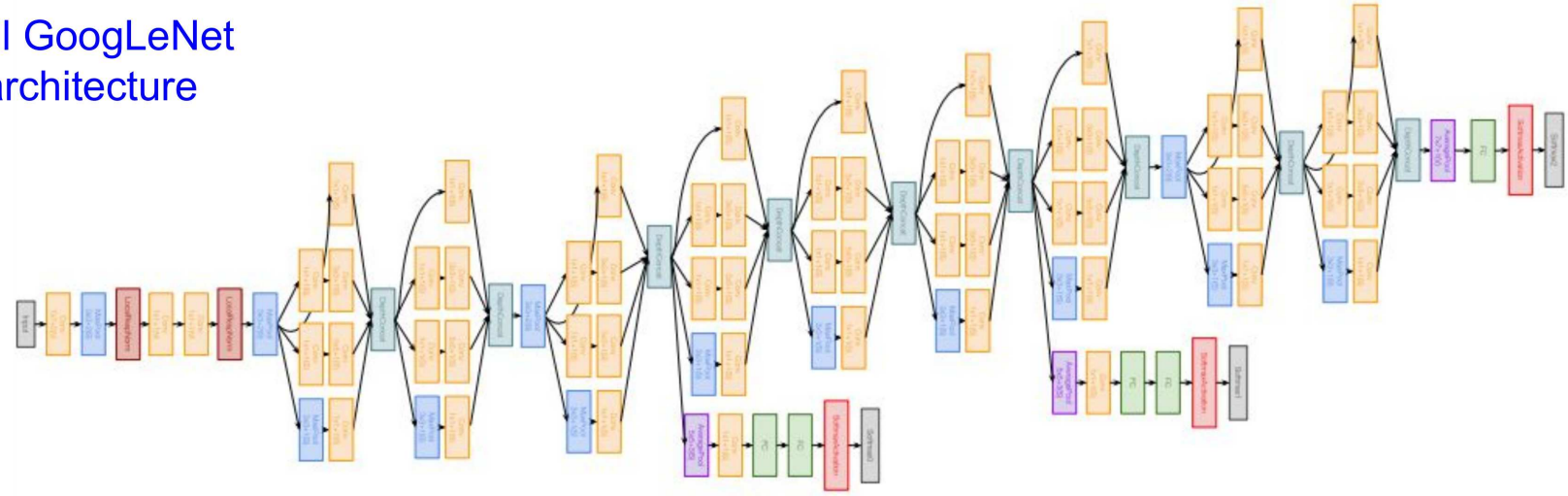


Auxiliary classification outputs to inject additional gradient at lower layers
(AvgPool-1x1Conv-FC-FC-Softmax)

Case Study: GoogLeNet

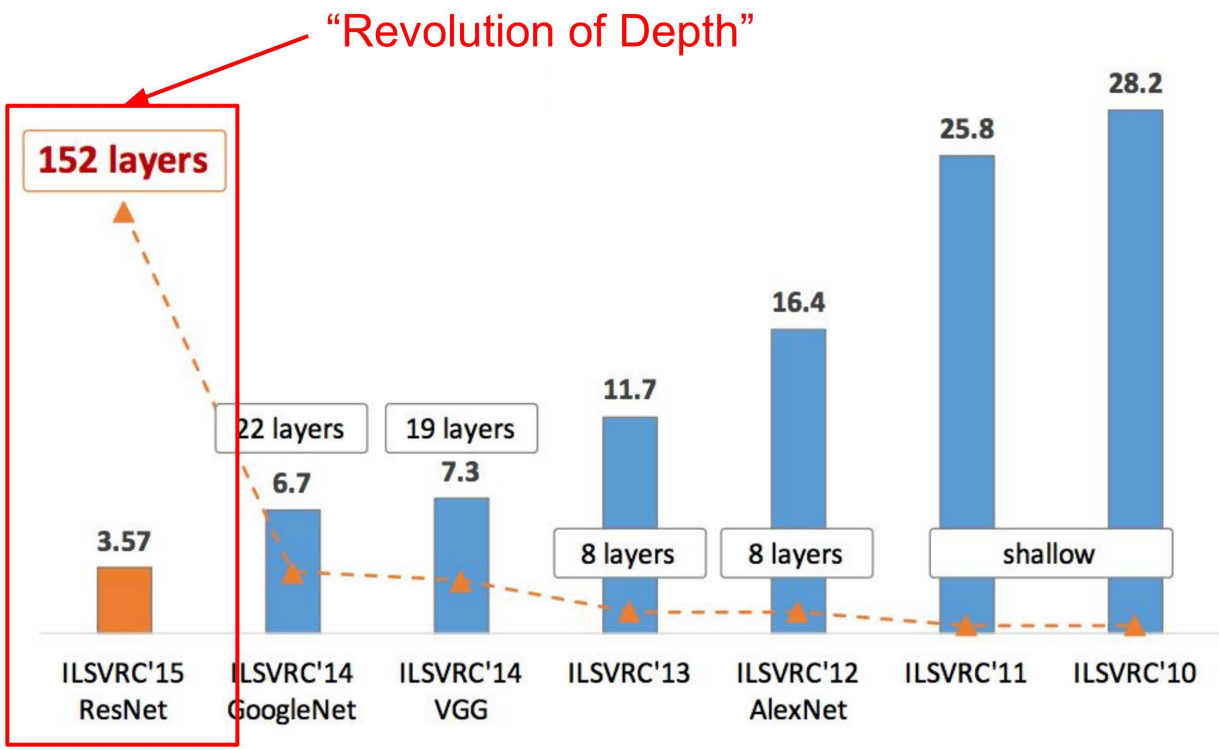
[Szegedy et al., 2014]

Full GoogLeNet architecture



22 total layers with weights (including each parallel layer in an Inception module)

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners

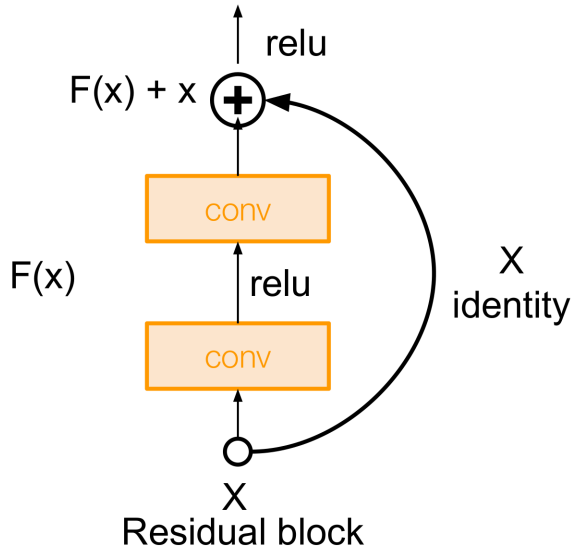


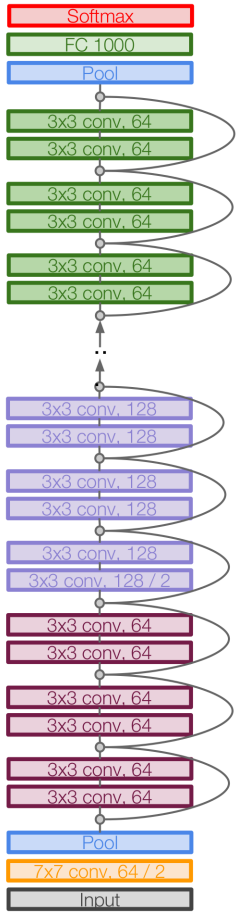
Case Study: ResNet

[He et al., 2015]

Very deep networks using residual connections

- 152-layer model for ImageNet
- ILSVRC'15 classification winner (3.57% top 5 error)
- Swept all classification and detection competitions in ILSVRC'15 and COCO'15!





Case Study: ResNet

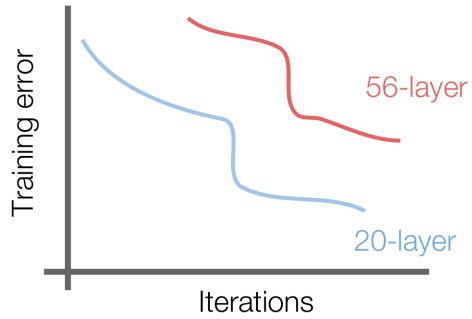
[He et al., 2015]

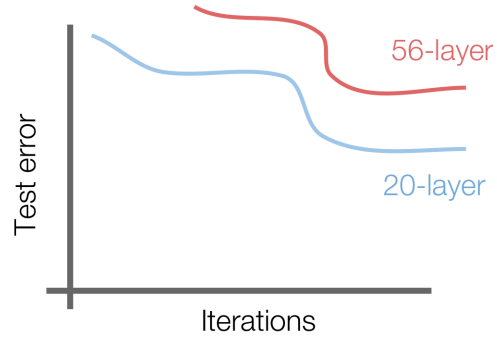
What happens when we continue stacking deeper layers on a “plain” convolutional neural network?

Case Study: ResNet

[He et al., 2015]

What happens when we continue stacking deeper layers on a “plain” convolutional neural network?



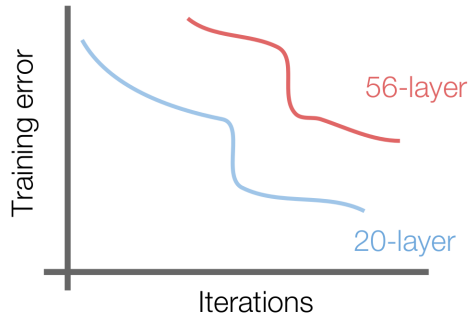


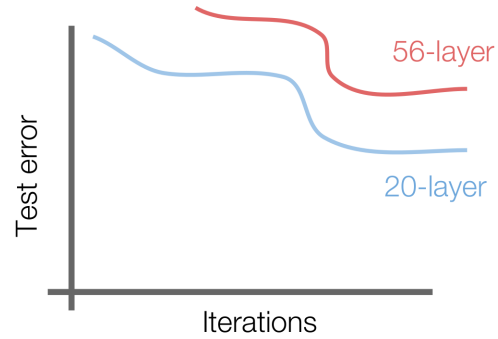
Q: What’s strange about these training and test curves?
[Hint: look at the order of the curves]

Case Study: ResNet

[He et al., 2015]

What happens when we continue stacking deeper layers on a “plain” convolutional neural network?





56-layer model performs worse on both training and test error
-> The deeper model performs worse, but it’s not caused by overfitting!

Case Study: ResNet

[He et al., 2015]

Hypothesis: the problem is an *optimization* problem, deeper models are harder to optimize

Case Study: ResNet

[He et al., 2015]

Hypothesis: the problem is an *optimization* problem, deeper models are harder to optimize

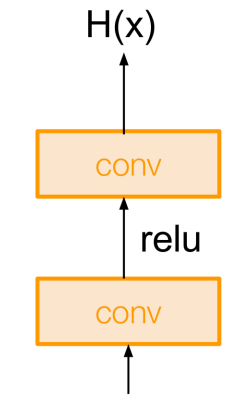
The deeper model should be able to perform at least as well as the shallower model.

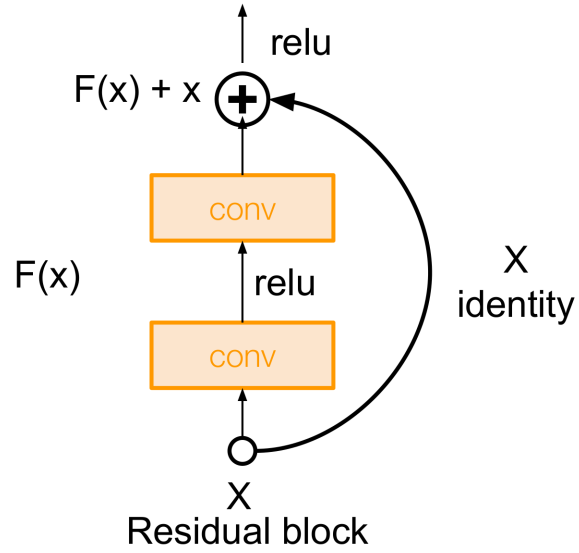
A solution by construction is copying the learned layers from the shallower model and setting additional layers to identity mapping.

Case Study: ResNet

[He et al., 2015]

Solution: Use network layers to fit a residual mapping instead of directly trying to fit a desired underlying mapping

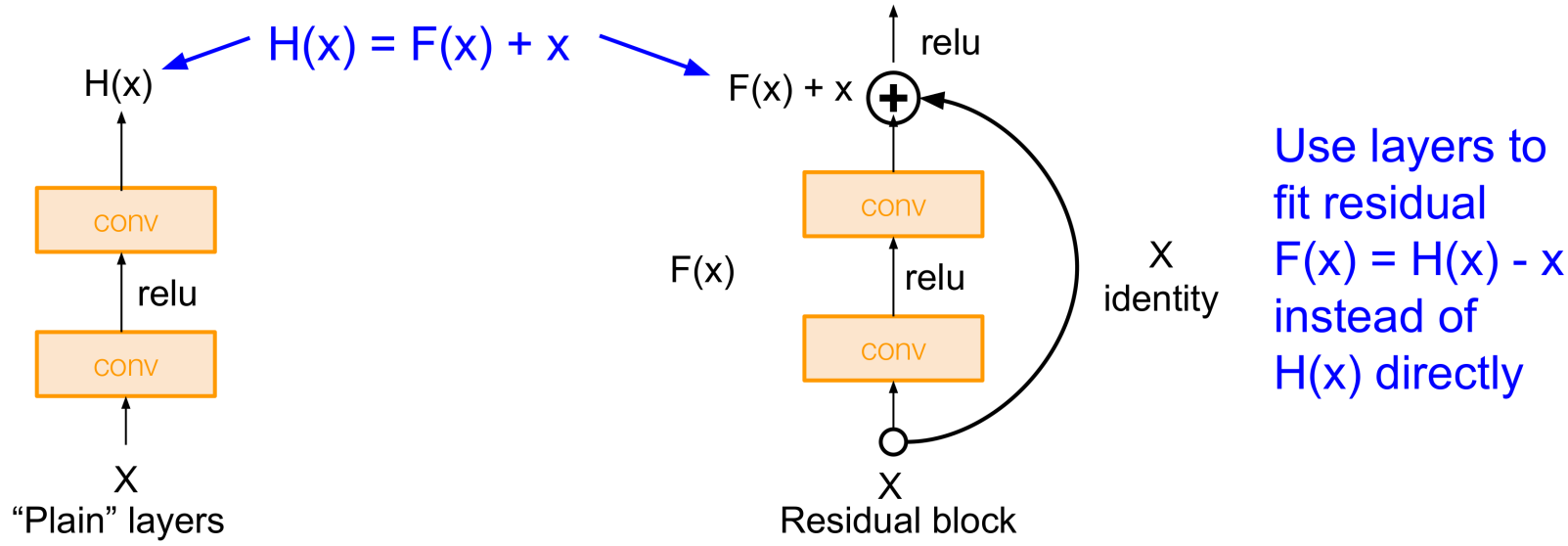




Case Study: ResNet

[He et al., 2015]

Solution: Use network layers to fit a residual mapping instead of directly trying to fit a desired underlying mapping

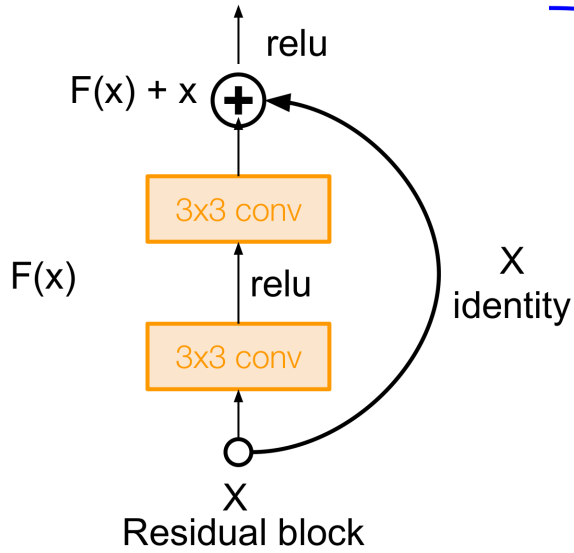


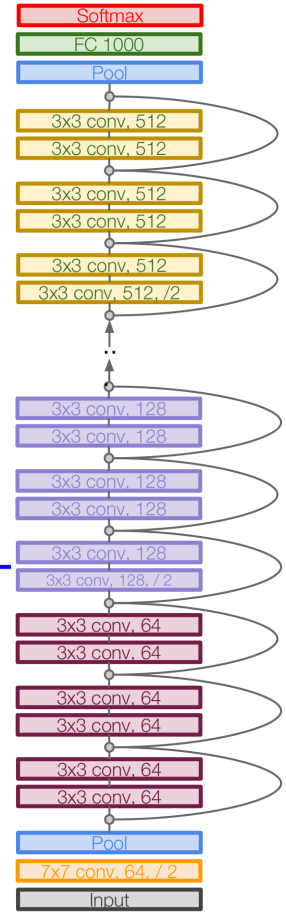
Case Study: ResNet

[He et al., 2015]

Full ResNet architecture:

- Stack residual blocks
- Every residual block has two 3x3 conv layers



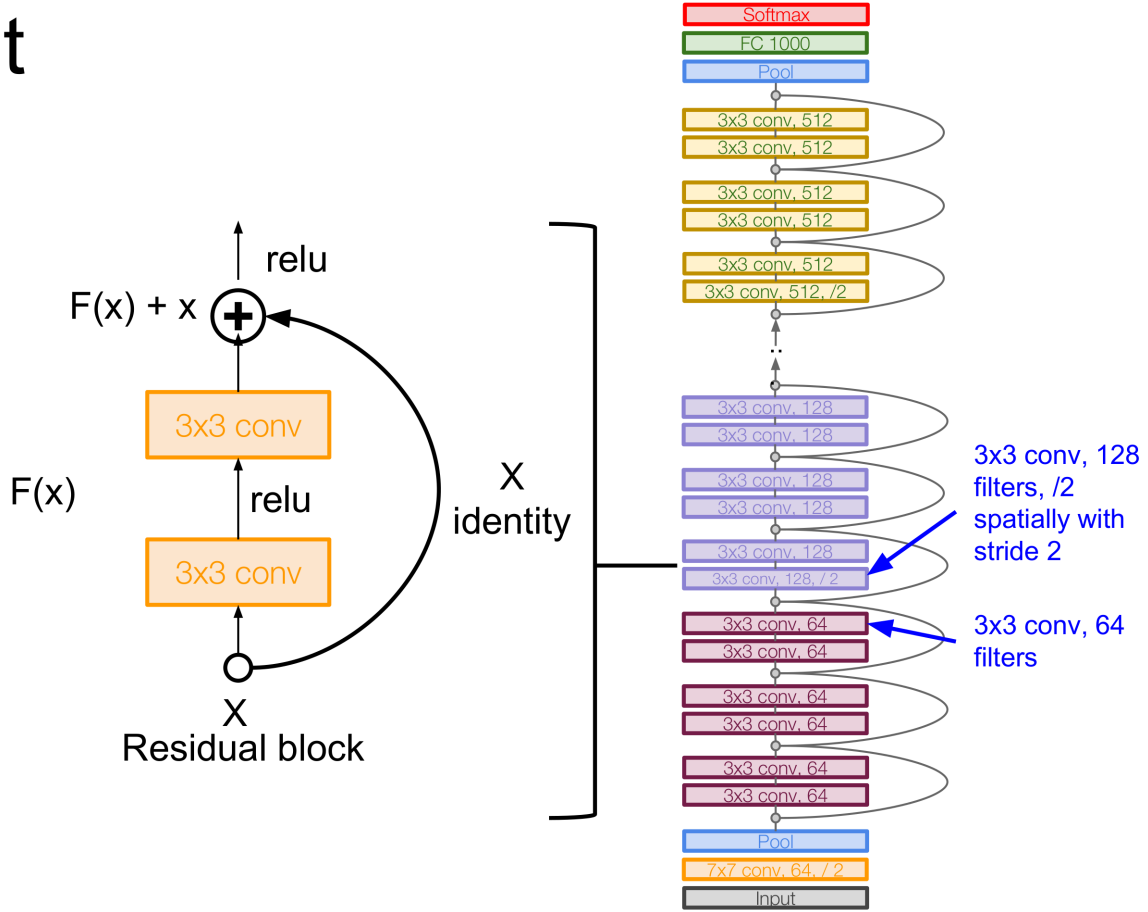


Case Study: ResNet

[He et al., 2015]

Full ResNet architecture:

- Stack residual blocks
- Every residual block has two 3x3 conv layers
- Periodically, double # of filters and downsample spatially using stride 2 (/2 in each dimension)

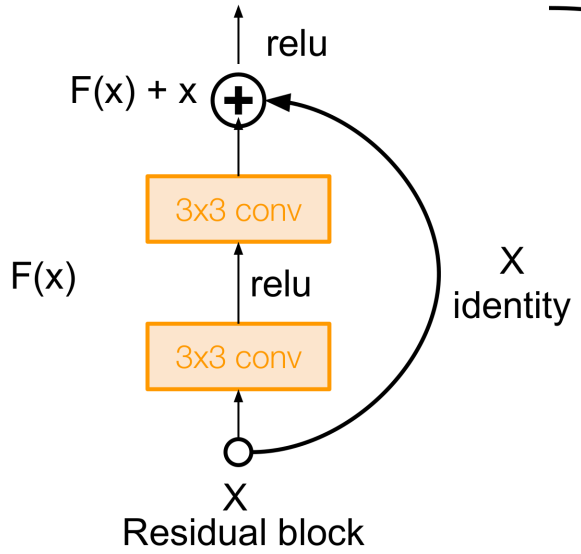


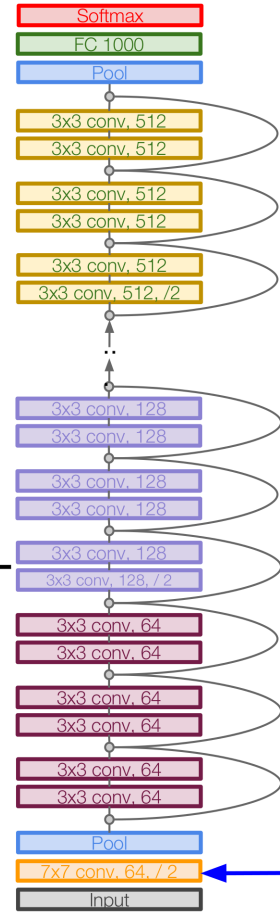
Case Study: ResNet

[He et al., 2015]

Full ResNet architecture:

- Stack residual blocks
- Every residual block has two 3x3 conv layers
- Periodically, double # of filters and downsample spatially using stride 2 (/2 in each dimension)
- Additional conv layer at the beginning





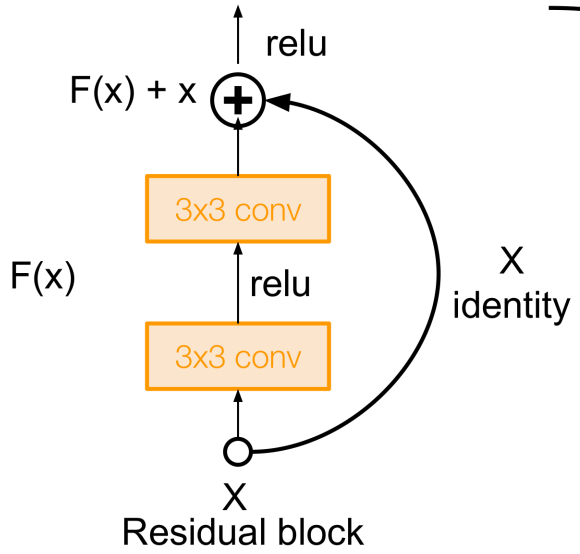
Beginning conv layer

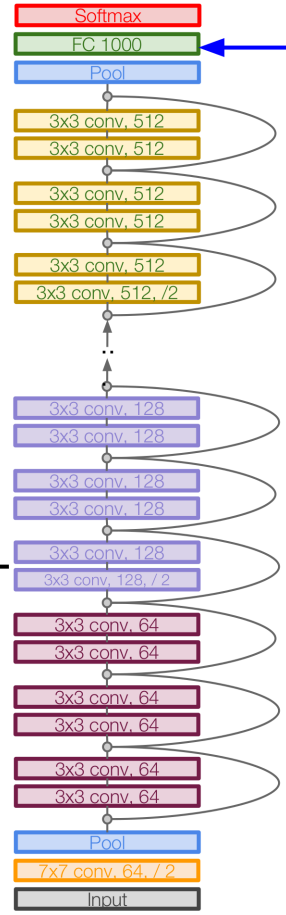
Case Study: ResNet

[He et al., 2015]

Full ResNet architecture:

- Stack residual blocks
- Every residual block has two 3x3 conv layers
- Periodically, double # of filters and downsample spatially using stride 2 (/2 in each dimension)
- Additional conv layer at the beginning
- No FC layers at the end (only FC 1000 to output classes)

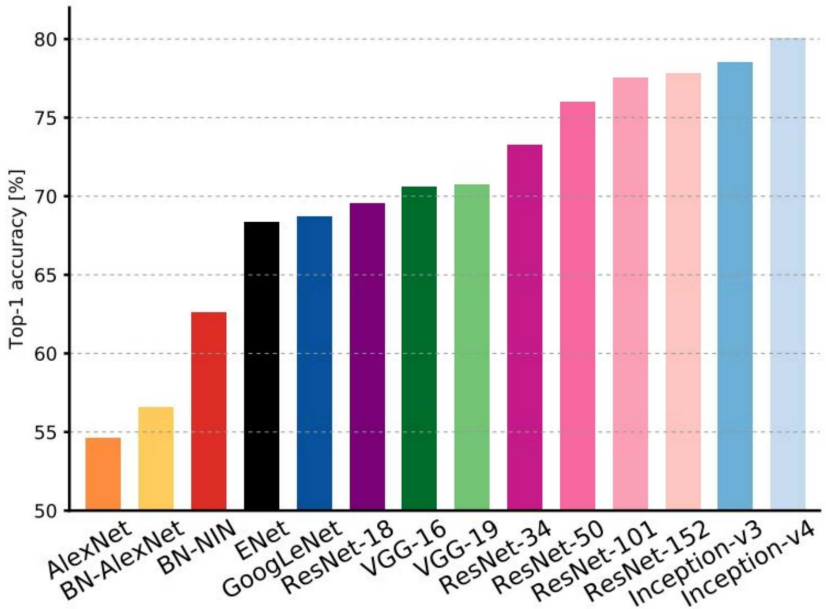


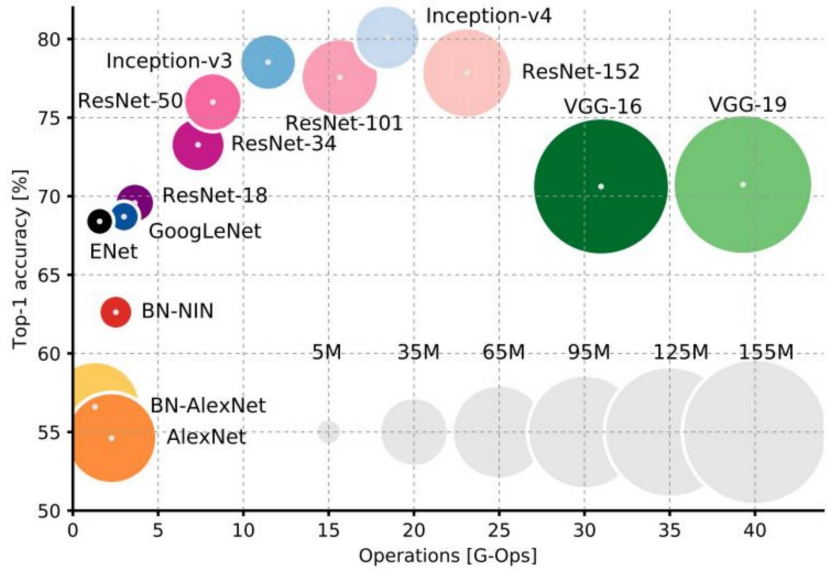


No FC layers besides FC 1000 to output classes

Global average pooling layer after last conv layer

Comparing complexity...

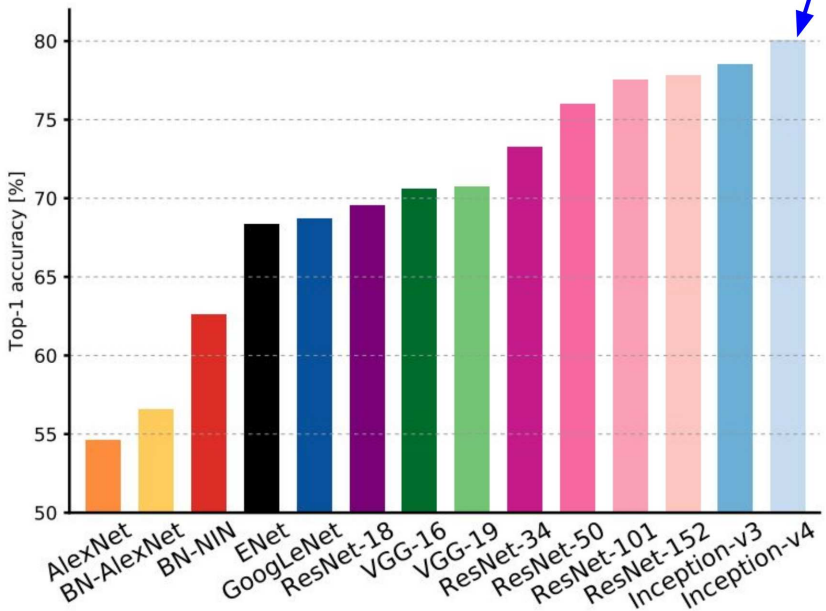


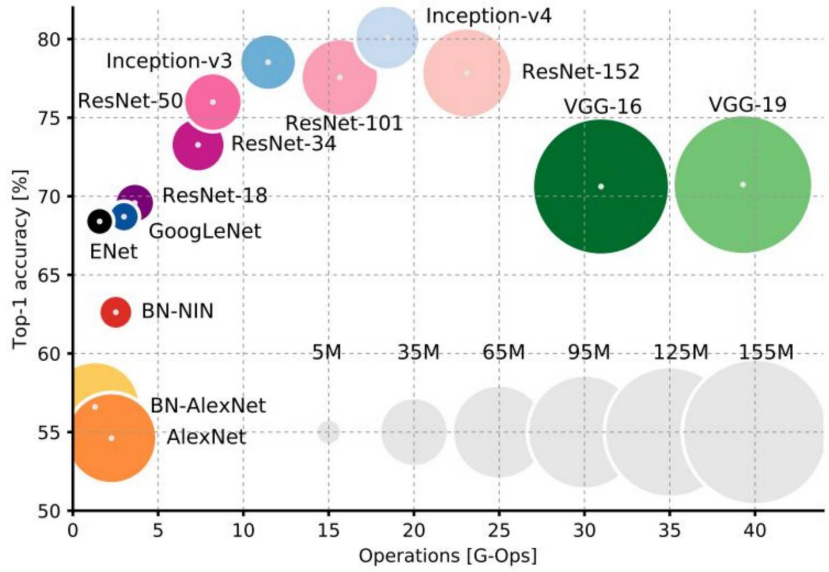


An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Comparing complexity...

Inception-v4: Resnet + Inception!

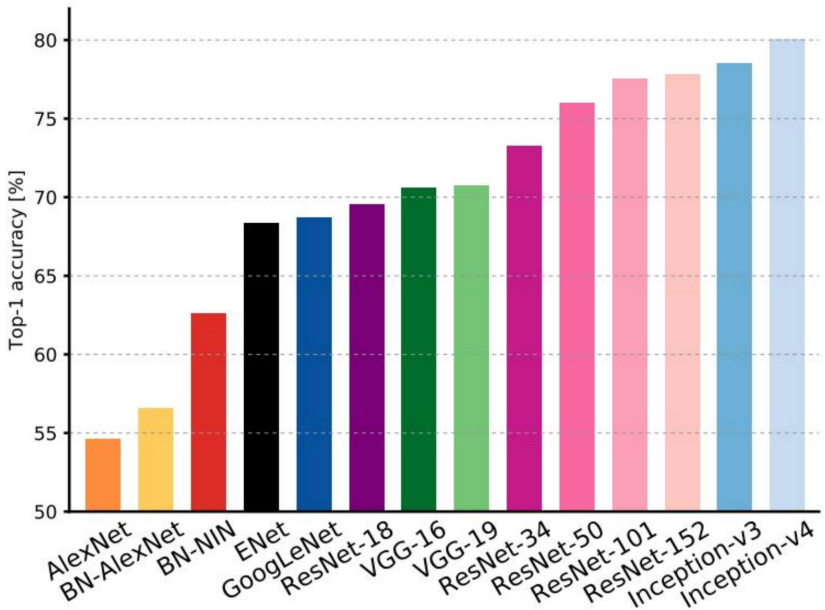


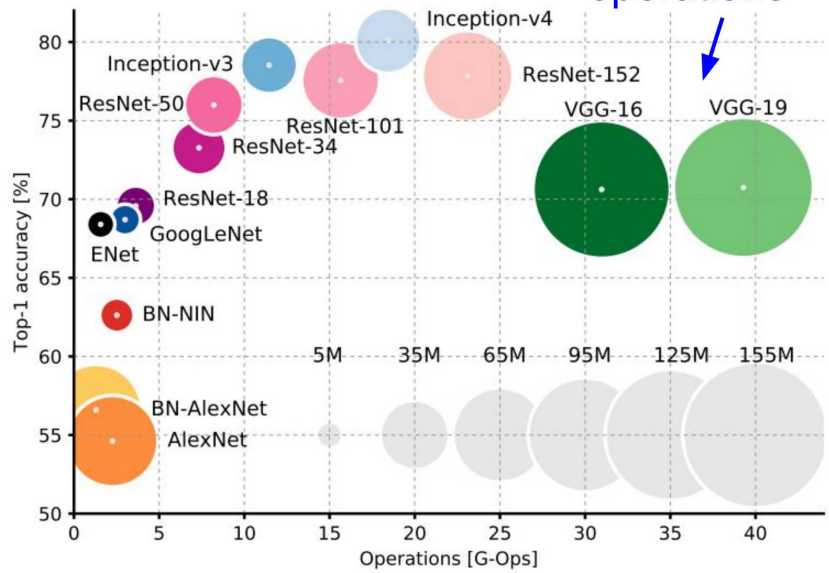


An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Comparing complexity...

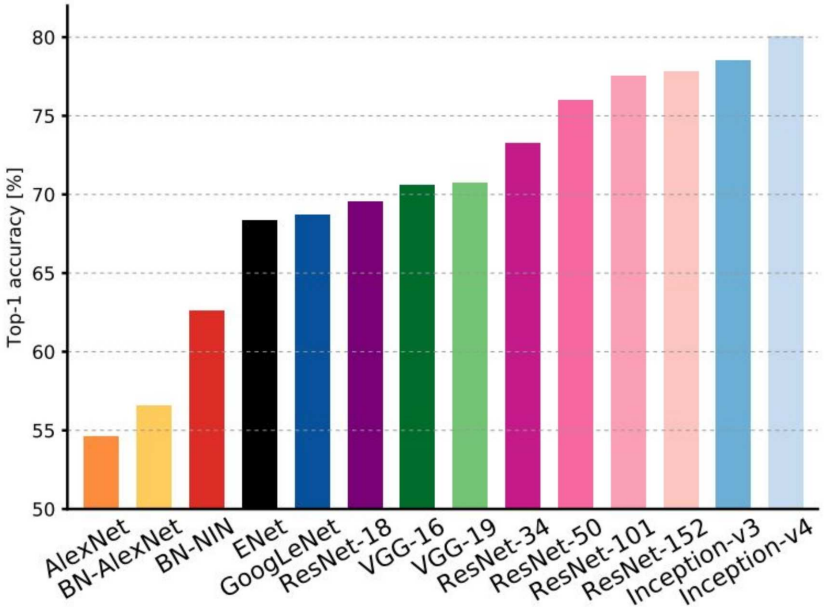
VGG: Highest memory, most operations



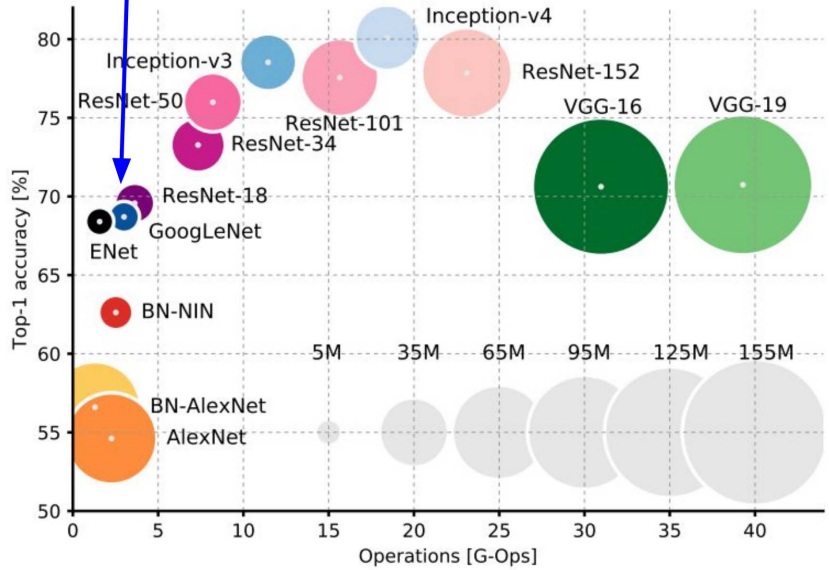


An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Comparing complexity...



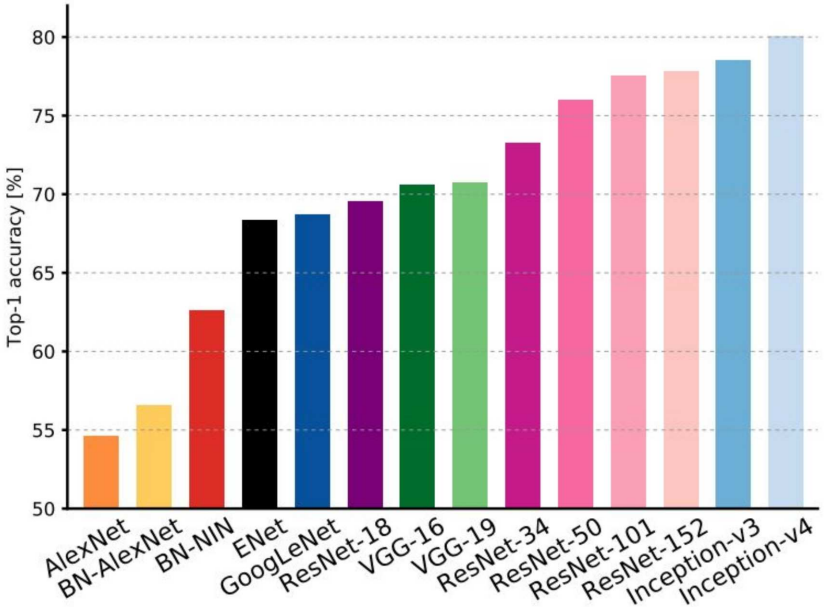
GoogLeNet:
most efficient

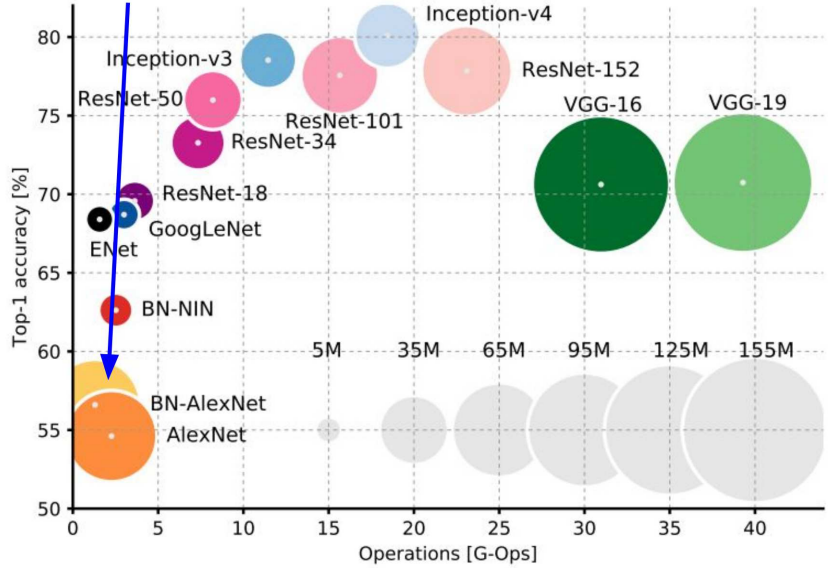


An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Comparing complexity...

AlexNet:
Smaller compute, still memory heavy, lower accuracy

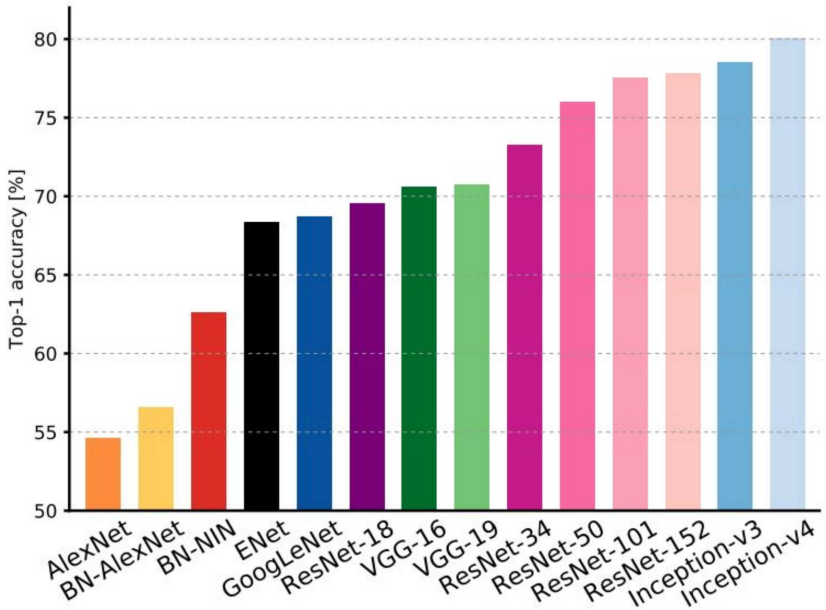


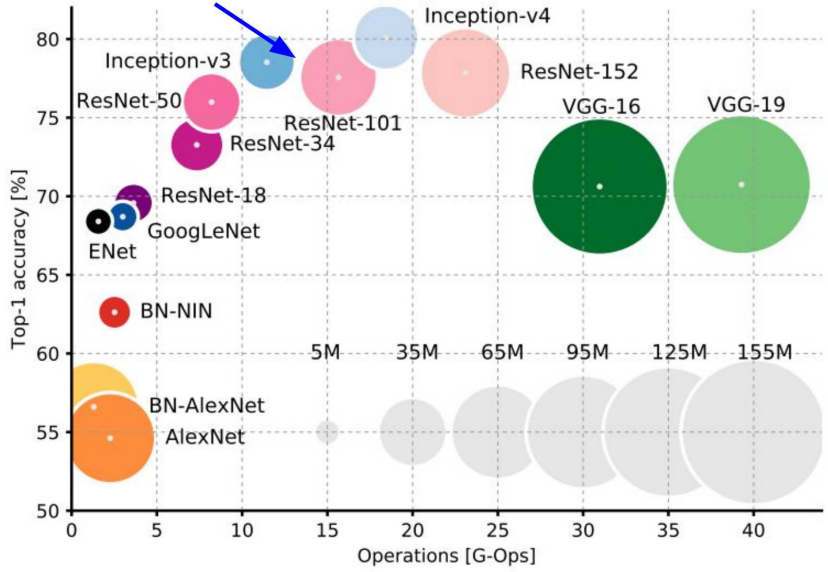


An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Comparing complexity...

ResNet:
Moderate efficiency depending on model, highest accuracy





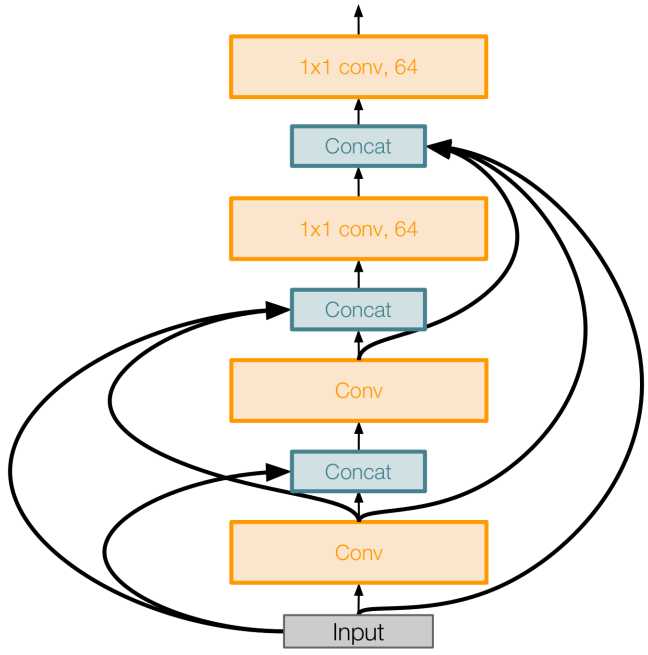
An Analysis of Deep Neural Network Models for Practical Applications, 2017.

Beyond ResNets...

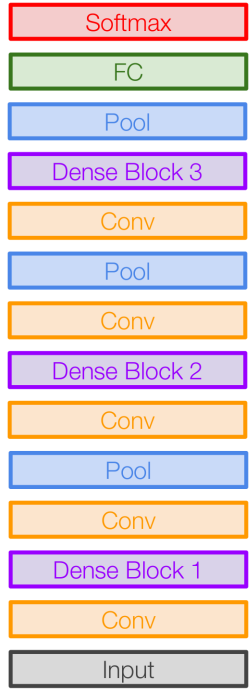
Densely Connected Convolutional Networks

[Huang et al. 2017]

- Dense blocks where each layer is connected to every other layer in feedforward fashion
- Alleviates vanishing gradient, strengthens feature propagation, encourages feature reuse



Dense Block



Contents

This year ...

Caffe
(UC Berkeley)

Caffe2
(Facebook)

Torch
(NYU / Facebook)

PyTorch
(Facebook)

Theano
(U Montreal)

TensorFlow
(Google)

Paddle
(Baidu)

CNTK
(Microsoft)

MXNet
(Amazon)

Developed by U Washington, CMU, MIT, Hong Kong U, etc but main framework of choice at AWS

And others...

Today

Caffe
(UC Berkeley)

Caffe2
(Facebook)

Torch
(NYU / Facebook)

PyTorch
(Facebook)

Theano
(U Montreal)

TensorFlow
(Google)

Mostly these

Paddle
(Baidu)

CNTK
(Microsoft)

MXNet
(Amazon)

Developed by U Washington, CMU, MIT, Hong Kong U, etc but main framework of choice at AWS

And others...

Thank You – Questions?

References:

Deep Learning, Ian Goodfellow, Yoshua Bengio

Deep Learning, Nature, Yann LeCun, Yoshua Bengio, Geoffrey Hinton

Deep learning in neural networks: An overview, J. Schmidhuber

Going Deeper with Convolutions, C. Szegedy et. al.

Deep Residual Learning for Image Recognition, K. He

GIT: awesome-deep-learning-papers