Computer Vision by Learning

Cees Snoek, University of Amsterdam

Efstratios Gavves, University of Amsterdam

With an invited tutorial by: Serge Belongie, University of Copenhagen

http://computervisionbylearning.info

Abstract

Computer vision has been first revolutionized since the year 2000. Learning from examples became leading. Another revolution happened in 2012, with deep learning from examples.

None of the methods for learning in computer vision is older than 3 years. In the course we will discuss methods of computing, invariance, equivariance and learning to distinguish and generate objects, actions and what more.

The course is supplemented with practical work and is completed with an assignment.

Where and When

Monday 9th of May to Thursday 12th of May

Lectures	09:30-12:15	CASA – theater room
Lunch	12:15-13:30	*included*
Lab	13:30-17:00	CASA – 3 lab rooms

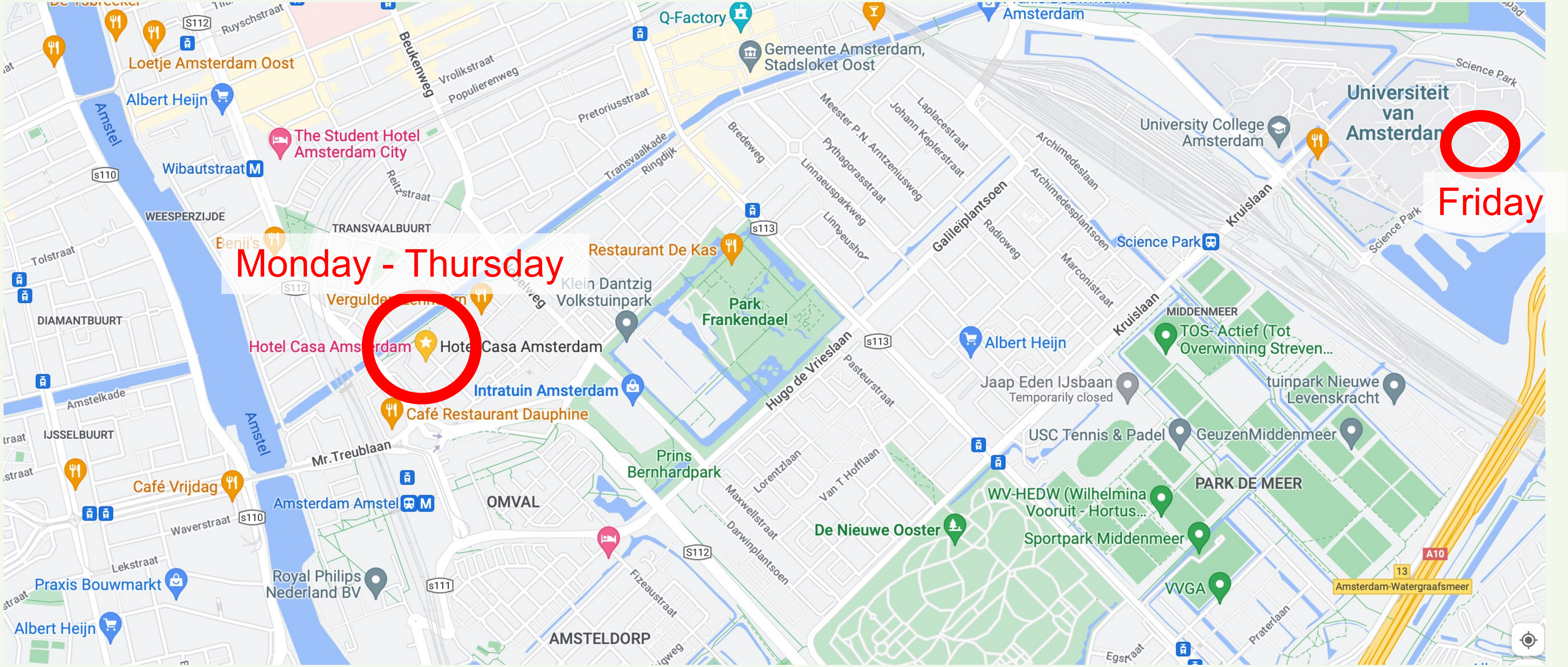
Thursday 12th of May

Borrel	17:00-18:00	CASA

Friday 13th of May

Invited tutorial	09:30-12:15	Startup Village – Venture studio
Closing	12:15-12:30	

Map


Monday - Thursday
Friday
Hotel Casa Amsterdam
Universiteit van Amsterdam
University College Amsterdam
Science Park
Q-Factory
Gemeente Amsterdam, Stadsloket Oost
Loetje Amsterdam Oost
Albert Heijn
The Student Hotel Amsterdam City
Wibautstraat
WEESPERZIJDE
TRANSVAALBUURT
Restaurant De Kas
Klein Dantzig Volkstuinpark
Park Frankendael
DIAMANTBUURT
Intratuin Amsterdam
Café Restaurant Dauphine
Amstel
IJSSELBUURT
Mr.Treublaan
Café Vrijdag
Amsterdam Amstel
OMVAL
Praxis Bouwmarkt
Royal Philips Nederland BV
AMSTELDORP
Prins Bernhardpark
De Nieuwe Ooster
Jaap Eden IJsbaan
Temporarily closed
USC Tennis & Padel
GeuzenMiddenmeer
PARK DE MEER
MIDDENMEER
Sportpark Middenmeer
VVGA
Kruislaan
Hugo de Vrieslaan
Galileiplantsoen
Amsterdam-Watergraafsmeer
A10

Startup village @ Science Park 608

Program

Monday	Fundamentals
Tuesday	Computer vision by learning
Wednesday	Machine learning for computer vision
Thursday	Computer video by learning
Friday	Invited tutorial by Serge Belongie

Serge Belongie

Guest speakers

Subhransu Maji

Martin Oswald

Erik Bekkers

Yuki Asano

Hazel Doughty

Lab

Lab Monday	Vision by multi-layer perceptron and convnet
Lab Tuesday	Vision by transformer
Lab Wednesday	Vision by geometric learning
Lab Thursday	Vision by self-supervised learning

TA team every afternoon available for support.

Each **group of 2 students** submits a report about their findings during the practicals. Your report should have roughly 1 page per practical, with a maximum of 8 pages. See lab assignments for all details on format, questions, PyTorch code etc.

Deadline: **May 31th, 2022**

http://computervisionbylearning.info

Overview

1. **Introduction**, history, tasks, impact.
2. **Invariance**, the need for, color invariants.
3. **Neural networks**, basics, perceptrons, multiple-layers.
4. **Convolutional networks**, local receptive fields, sharing, pooling.
5. **ImageNet classification** with deep convolutional networks.

1. Introduction

AI is not new

A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

August 31, 1955

John McCarthy, Marvin L. Minsky,
Nathaniel Rochester,
and Claude E. Shannon

Computer vision is also not new, how old?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group
Vision Memo. No. 100.

July 7, 1966

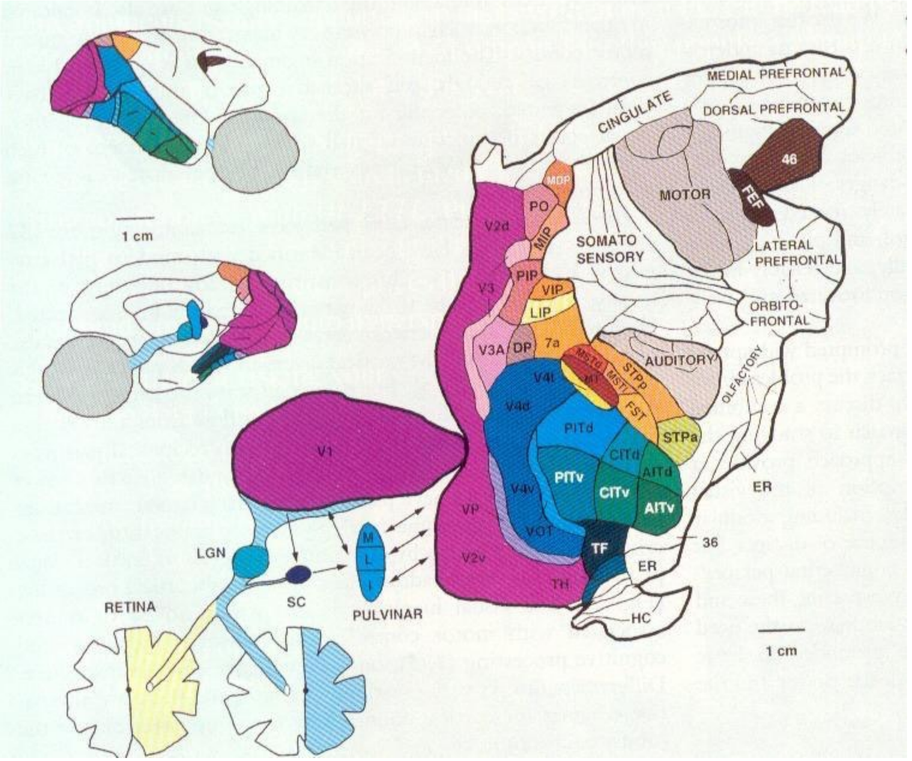
<u>THE SUMMER VISION PROJECT</u>

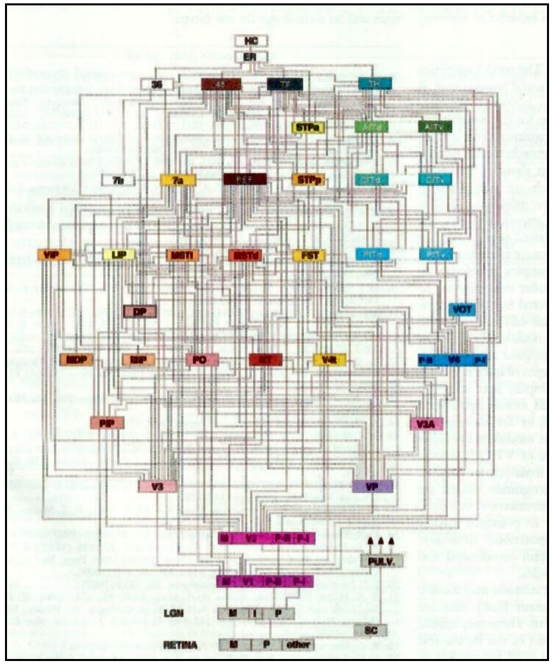
Seymour Papert

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".

How difficult is the problem?

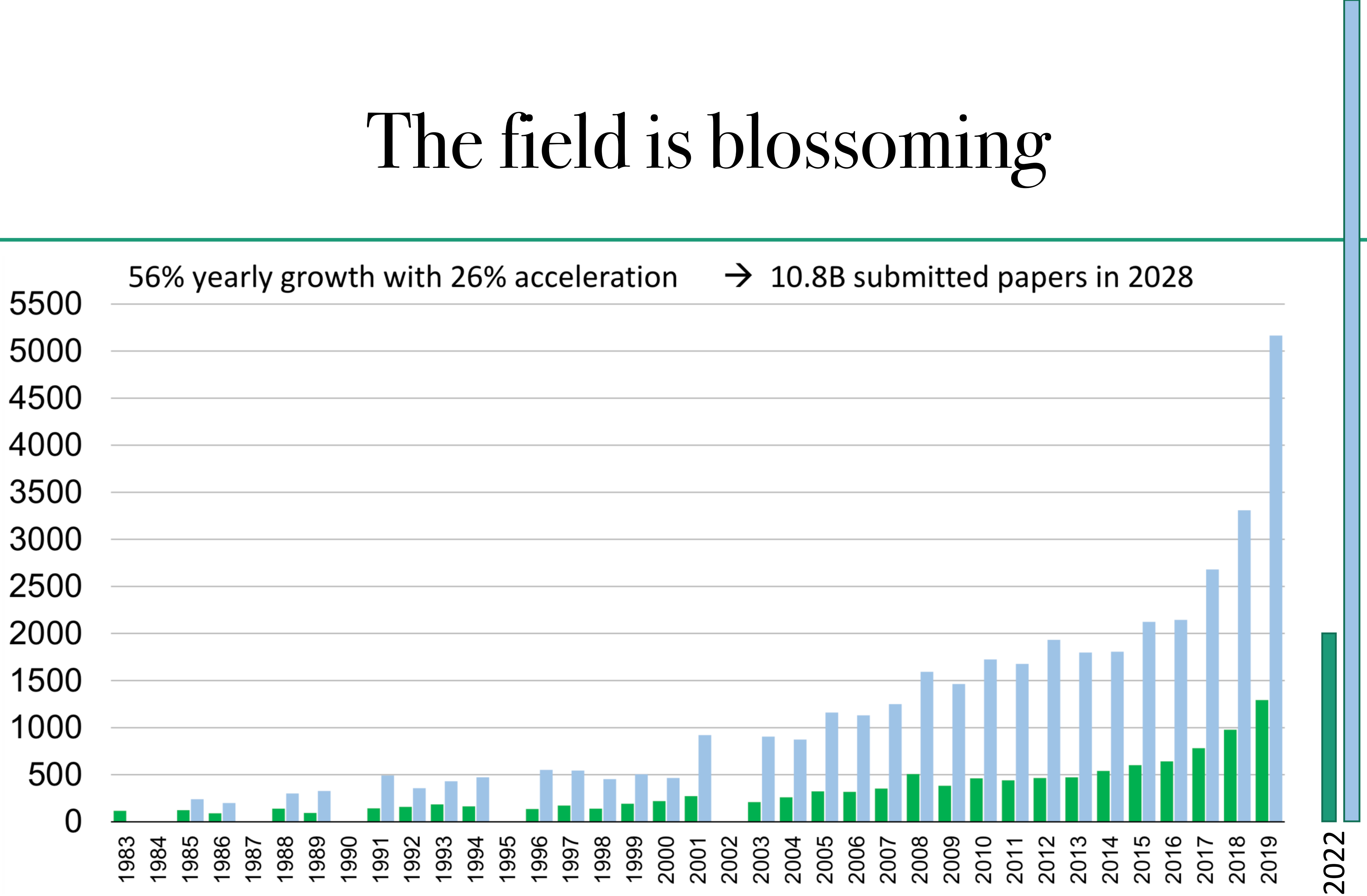
Human vision consumes 50% brain power…





Van Essen, Science 1992

The field is blossoming



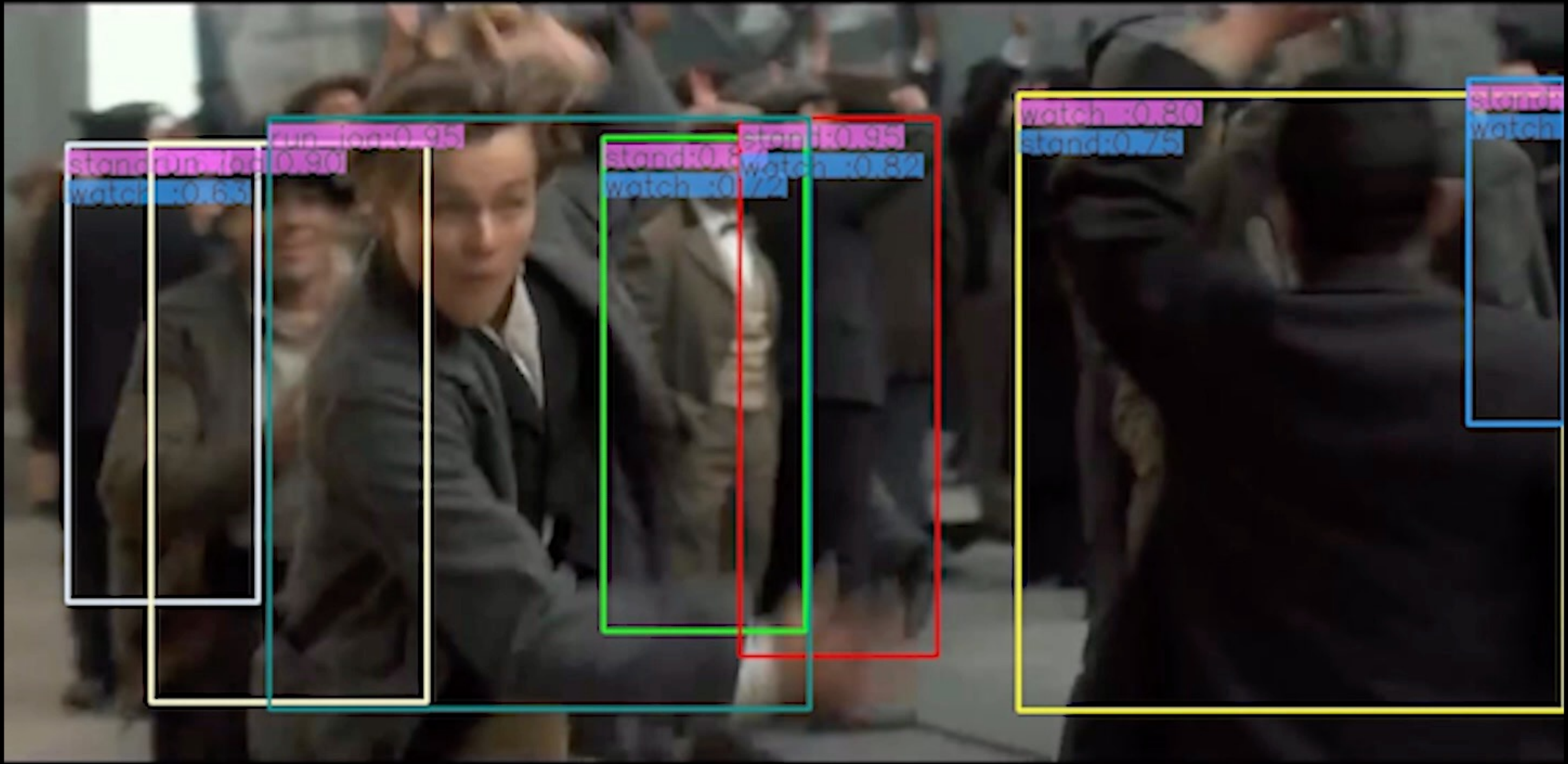
Submitted and accepted papers at CVPR, vision's main research venue

Dalle-2 image generation

"academic researchers before a deadline in the style of Edvard Munch"

src: @ankurhandos

Kien Nguyen *et al.*, CVPR22



Jiaojiao Zhao et al., CVPR 2022

Esteva, Kuprel, *et al.*, Nature 2017

Most cited science in 2016-2020

1. **Deep Residual Learning for Image Recognition**
   CVPR 2016
   82,588 citations

2. **Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China**
   The Lancet 2020
   30,529 citations

3. **Attention is all you need**
   NeurIPS 2017
   23,606 citations

According to Google Scholar, July 2021

2. Invariance

Invariance aims to exclude all irrelevant variations.

Color invariants are powerful for everything related to illumination, and in the end simple.

The quality of invariant features: invariance + discrimination.

The need for invariance

There are a million appearances to one object

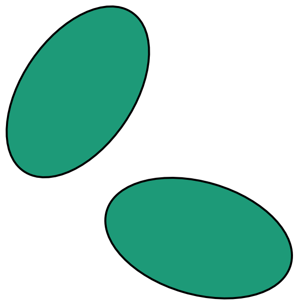
Same part of same shoe does not have same appearance in the image.

Remove unwanted variance from the representation, but when?

The need for invariance

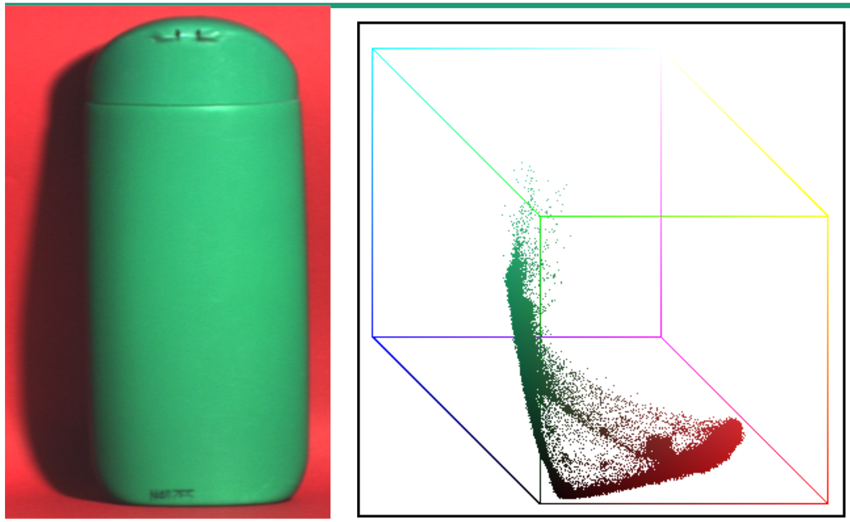
A feature g is invariant under condition (transform) W caused by accidental conditions at the time of recording, iff g observed on equal objects t_1 and t_2 is constant:

$$t_1 \overset{W}{\sim} t_2 \Rightarrow f_{t_1} = f_{t_2}$$

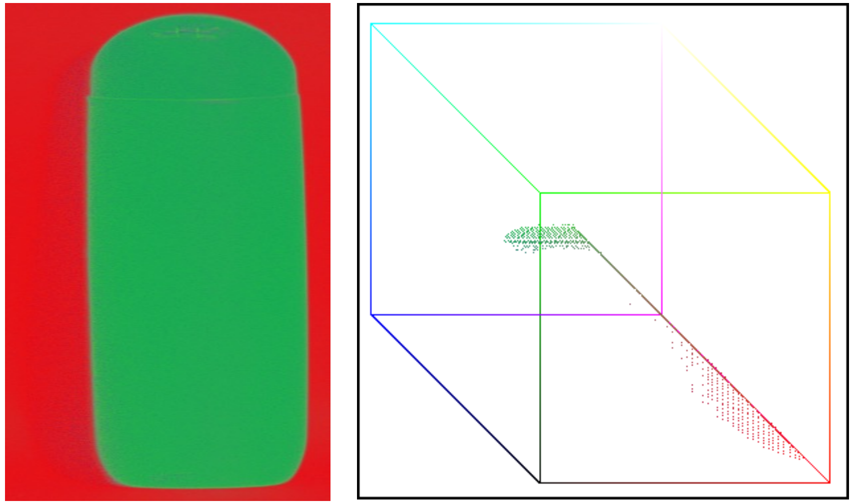


Length of long axis / short axis is independent of scale and rotation.

Simple color space example



RGB space



C space

$$c_1(R,G,B) = \arctan \frac{R}{\max\{G,B\}}$$

$$c_2(R,G,B) = \arctan \frac{G}{\max\{R,B\}}$$

$$c_3(R,G,B) = \arctan \frac{B}{\max\{R,G\}}$$

Gevers TIP 2000

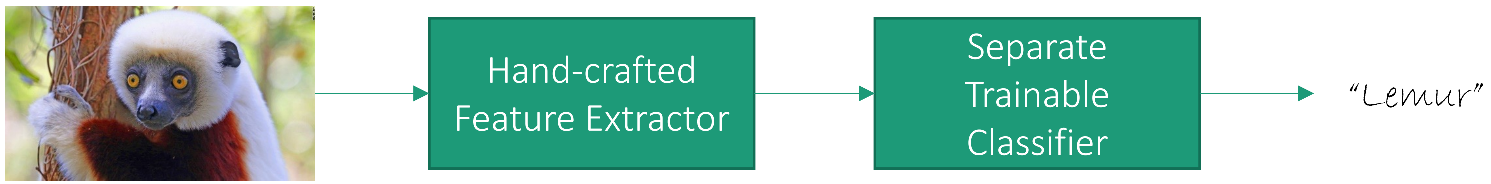
Invariance & Discrimination

The most invariant feature is the value “42”.

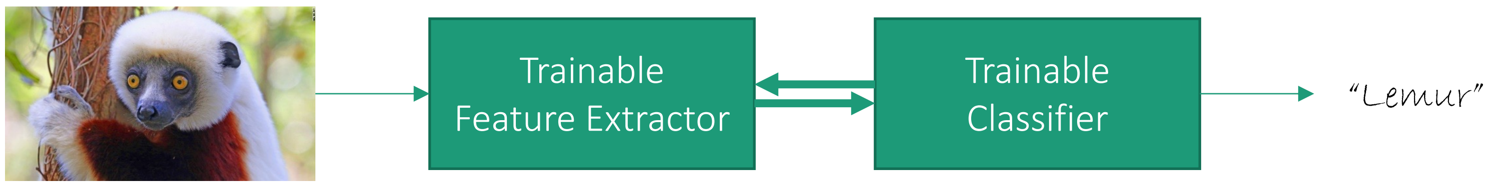
Balance desired invariance with undesired loss of discriminative power.

Modeling or learning invariance?

Traditional computer vision by learning



End-to-end-learning



3. Basics of Neural Networks

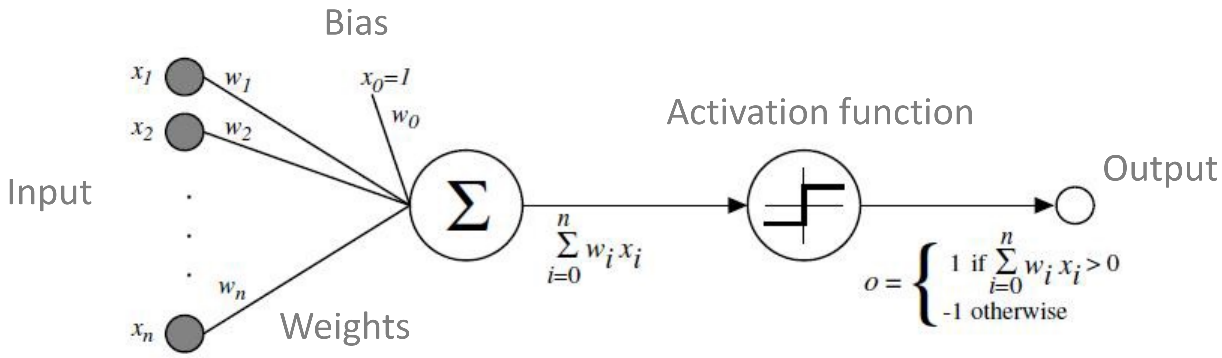
In this chapter we discuss the basics of neural networks. Covering perceptrons, multiple-layers, backpropagation and activation functions.

Perceptrons

Rosenblatt proposed a machine for **binary** classification in 1958

Main idea

- One weight w_i per input x_i
- Multiply weights with respective inputs and add bias x_0 =+1
- If result larger than threshold return 1, otherwise -1



Key innovation: a learning algorithm

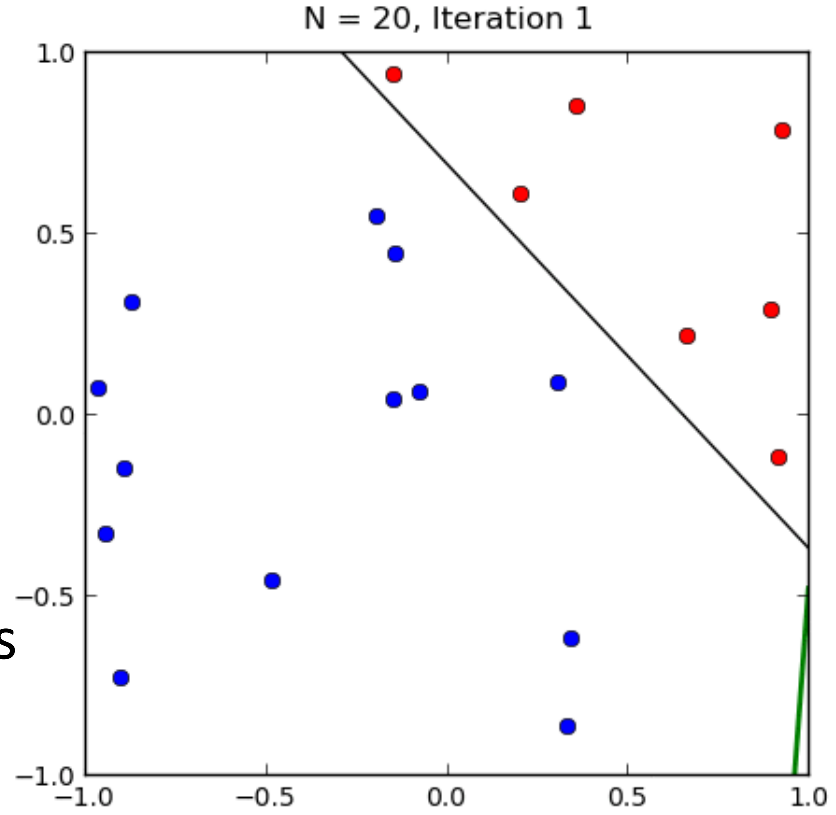
Initialize weights randomly

Take one sample x_i and predict y_i

For erroneous predictions update weights

- If the output was $\hat{y}_i = 0$ and $y_i = 1$, increase weights
- If the output was $\hat{y}_i = 1$ and $y_i = 0$, decrease weights

Repeat until no errors are made



It did not pass unnoticed…

NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI) —The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's $2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's

ings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

Without Human Controls

The Navy said the perceptron would be the first non-living mechanism "capable of receiving, recognizing and identifying its surroundings without any human training or control."

The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch

The New York Times

In today's demonstration, the "704" was fed two cards, one with squares marked on the left side and the other with squares on the right side.

Learns by Doing

In the first fifty trials, the machine made no distinction between them. It then started registering a "Q" for the left squares and "O" for the right squares.

Dr. Rosenblatt said he could explain why the machine learned only in highly technical

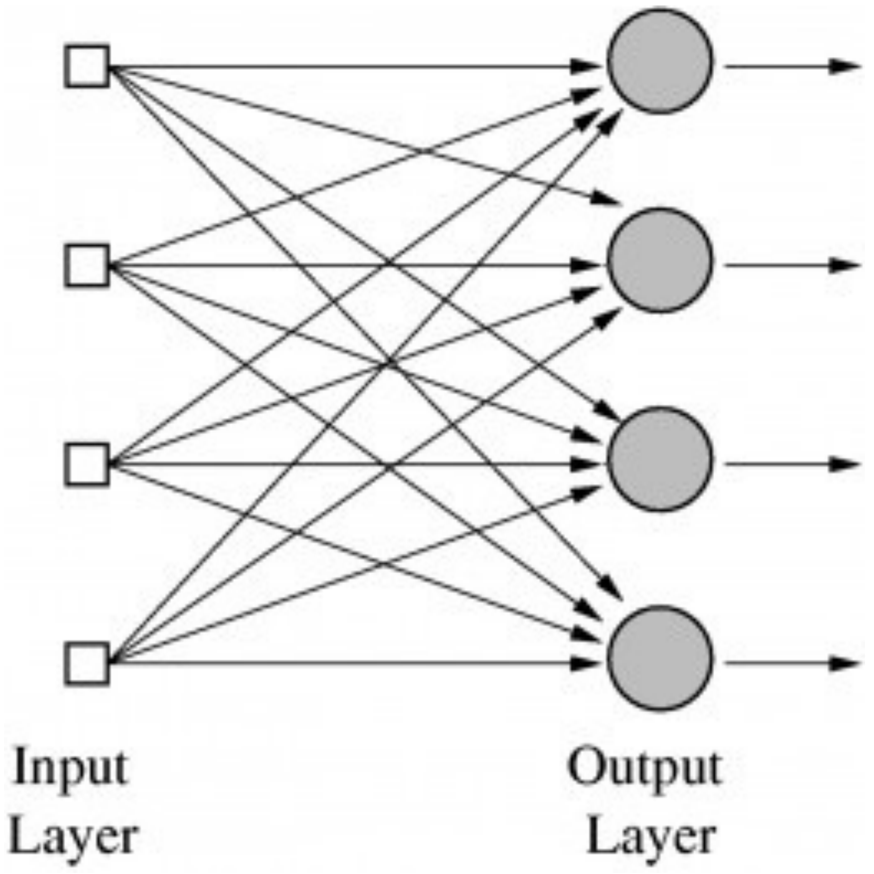
From perceptron to neural network

One perceptron = one decision

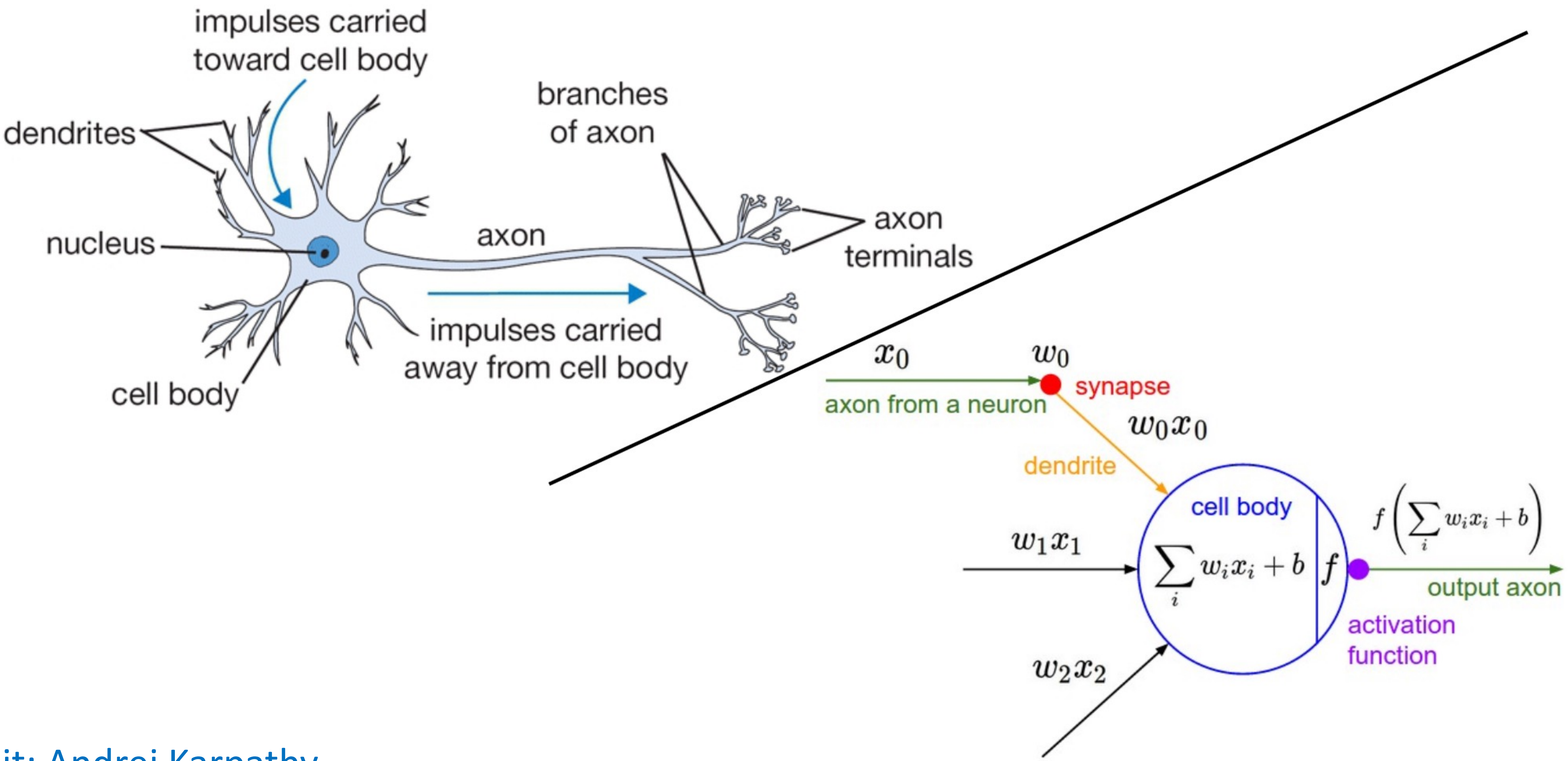
What about multiple decisions?
- E.g. digit classification

Stack outputs into a layer
- Neural network

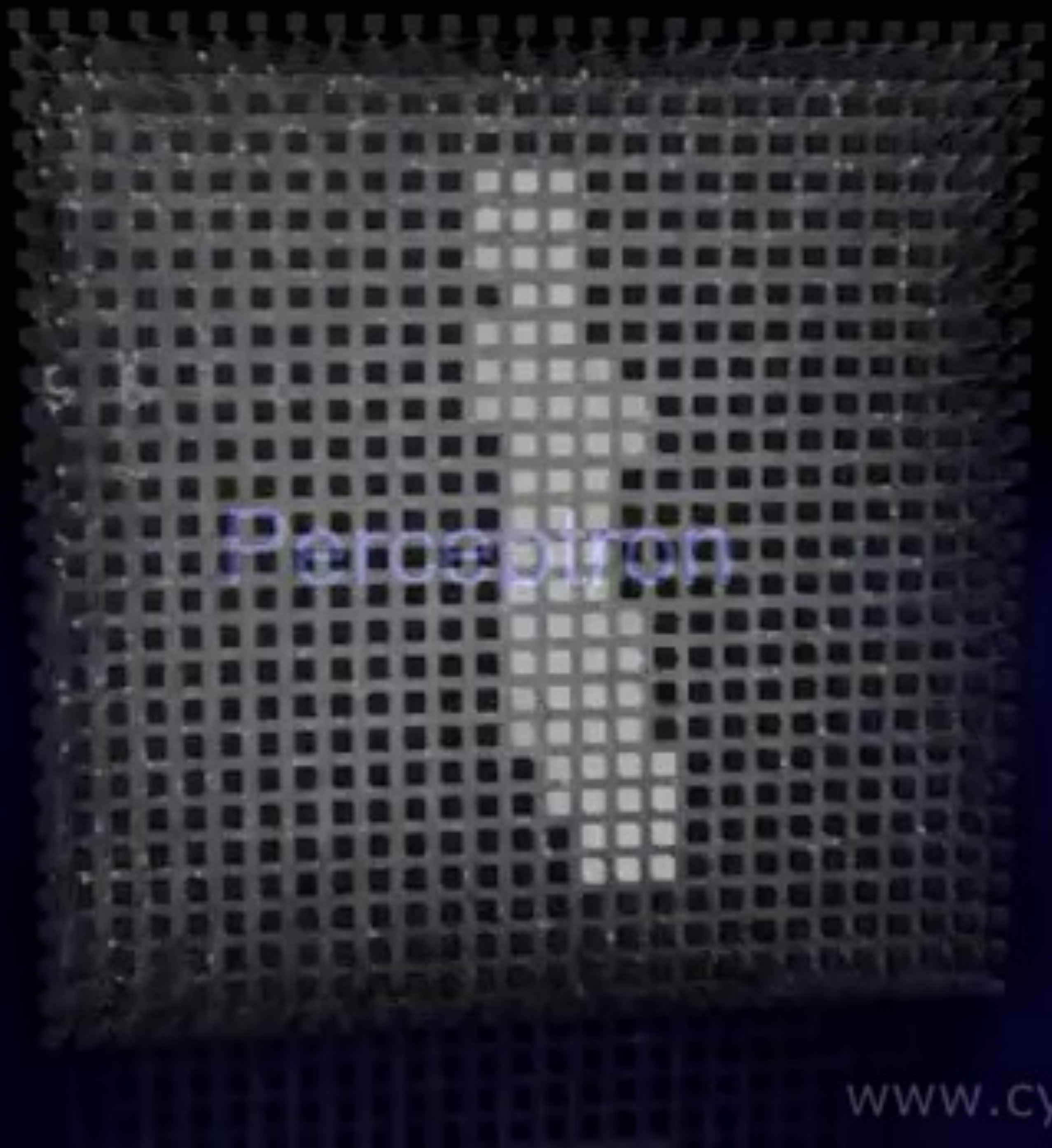
Use one layer as input to the next layer
- Multi-layer perceptron (MLP)



Neuro equivalence



Slide credit: Andrej Karpathy


Perceptron

Backpropagation

Algorithm that looks for minimum of the error function in weight space

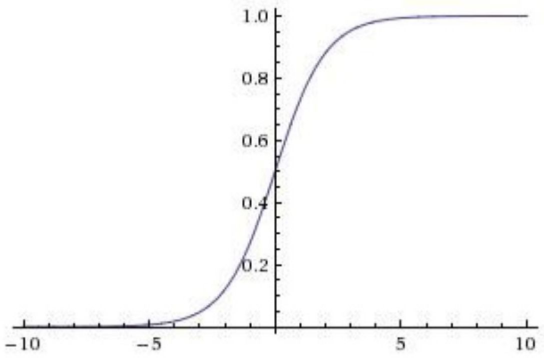
- Uses gradient descent
- Practical application of the chain rule
- Known since early-1960s, not widely understood until mid-1980s

Derivatives can be computed by working backwards from the gradient wrt the output of each layer in a network

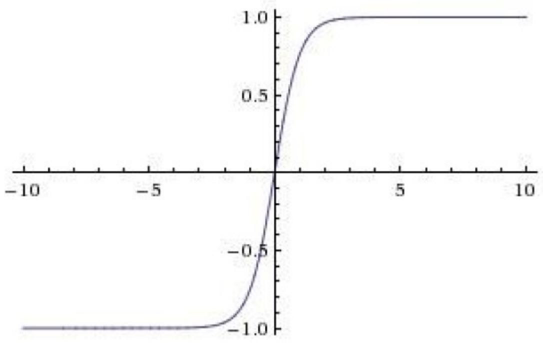
- As long as layer does not use heaviside step activation function, as in perceptron

Rumelhart et al., Nature 1986

Derivative-friendly activation functions



Sigmoid



tanh(x)

+ Squashes numbers to [0,1] range
- Sigmoid outputs not zero-centered
- Saturated neurons kill the gradient

+ Squashes numbers to [-1,1] range
+ Zero-centered
- Saturated neurons kill the gradient

Backpropagation

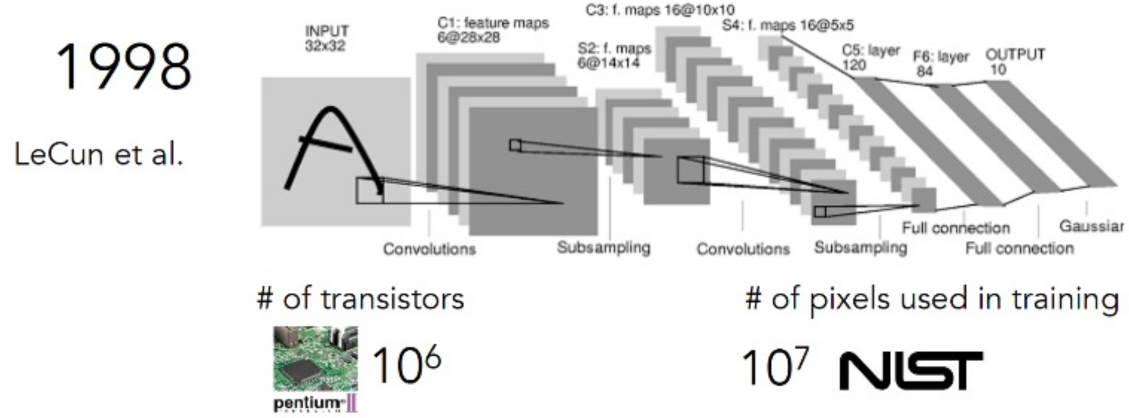
Learning multi-layer perceptrons made possible

- XOR and more complicated functions can be solved

Efficient algorithm

- Process hundreds of example without a sweat
- Allowed for more complicated neural network architectures

Still the engine of neural network training today



4. Convolutional Neural Networks

Convolutional neural networks are a specialized kind of neural networks for processing data that has a grid-like topology, especially image data. We discuss its filters, weight-sharing principles and its pooling operator.

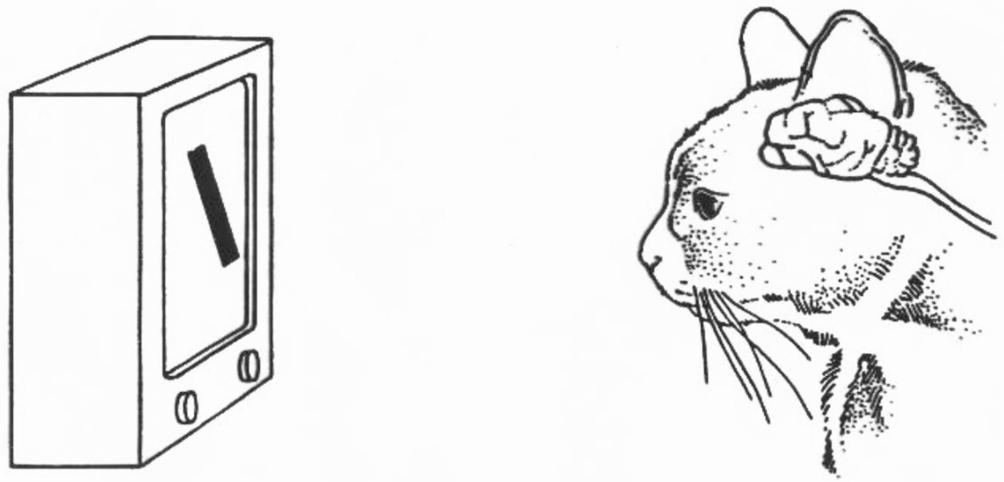
How to classify an image with an MLP?

A **256x256 RGB image** requires 200 000 input values

MLP with a single hidden layer with 500 units already implies
100 million parameters

Clearly we need to incorporate an **inductive bias** into the architecture

Hubel & Wiesel's cat experiments



Neurons are spatially localized

Define topographic feature maps

Provide hierarchical feature processing

Nobel Prize 1981

LeCun et al. PIEEE 1998

Convolutional layers

Force receptive fields of hidden units **to be local** so they capture points, edges and corners and build from there

Elementary feature detectors are useful for the entire image allows to **share their weights**

Sequential implementation corresponds to **convolution operation.**

Convolutional layer

The convolution layer has a set of **filters**. Its output is a set of **feature maps**, each one obtained by convolving the image with a filter

Blur

$$* \begin{array}{|c|c|c|} \hline 0 & 1 & 0 \\ \hline 1 & 4 & 1 \\ \hline 0 & 1 & 0 \\ \hline \end{array} =$$

Edge detect

$$* \begin{array}{|c|c|c|} \hline 0 & -1 & 0 \\ \hline -1 & 4 & -1 \\ \hline 0 & -1 & 0 \\ \hline \end{array} =$$

Why does it work?

Slide credit: Roger Grosse and Jimmy Ba

Convolutional layers are translation equivariant

If the input image is shifted, the feature map output will be shifted with the same amount, but will be unchanged otherwise

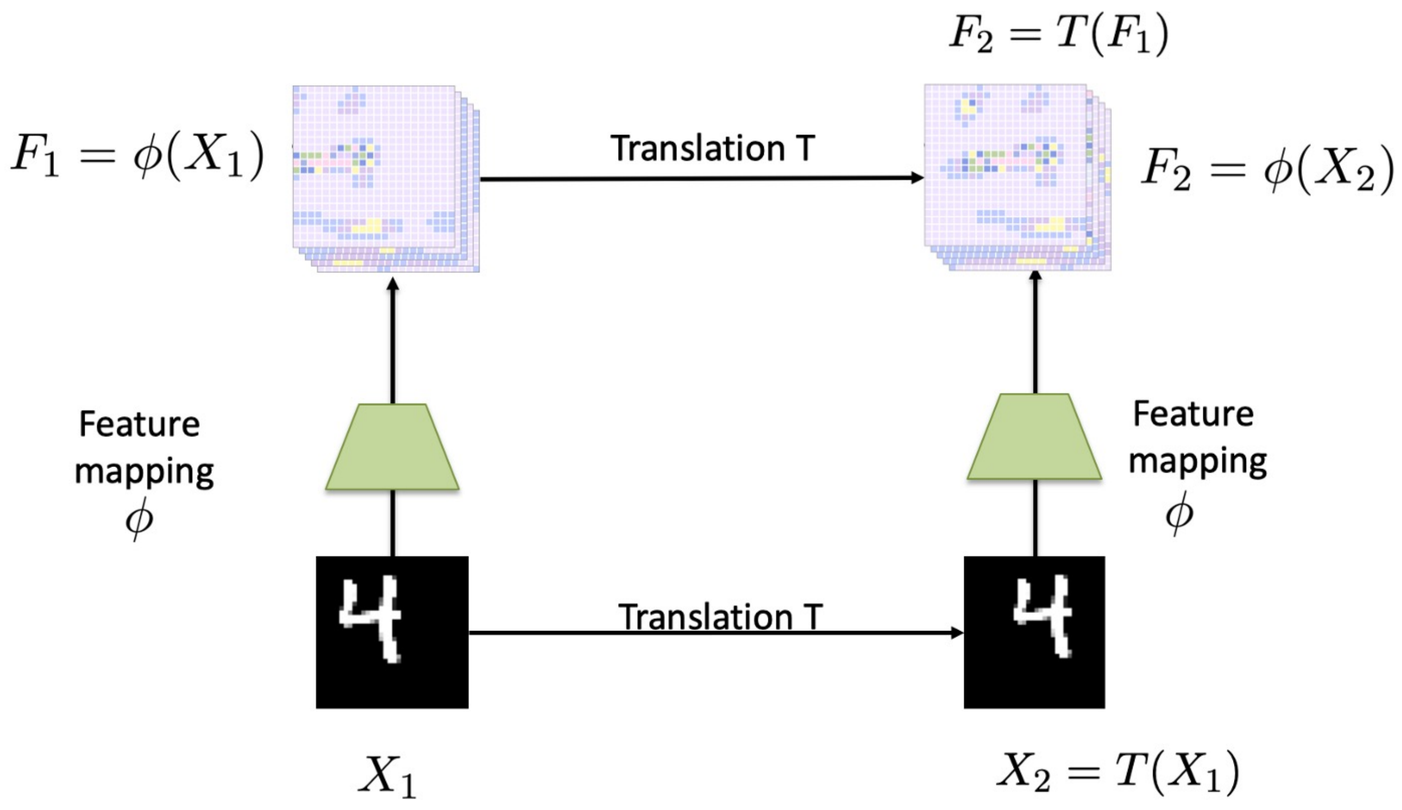
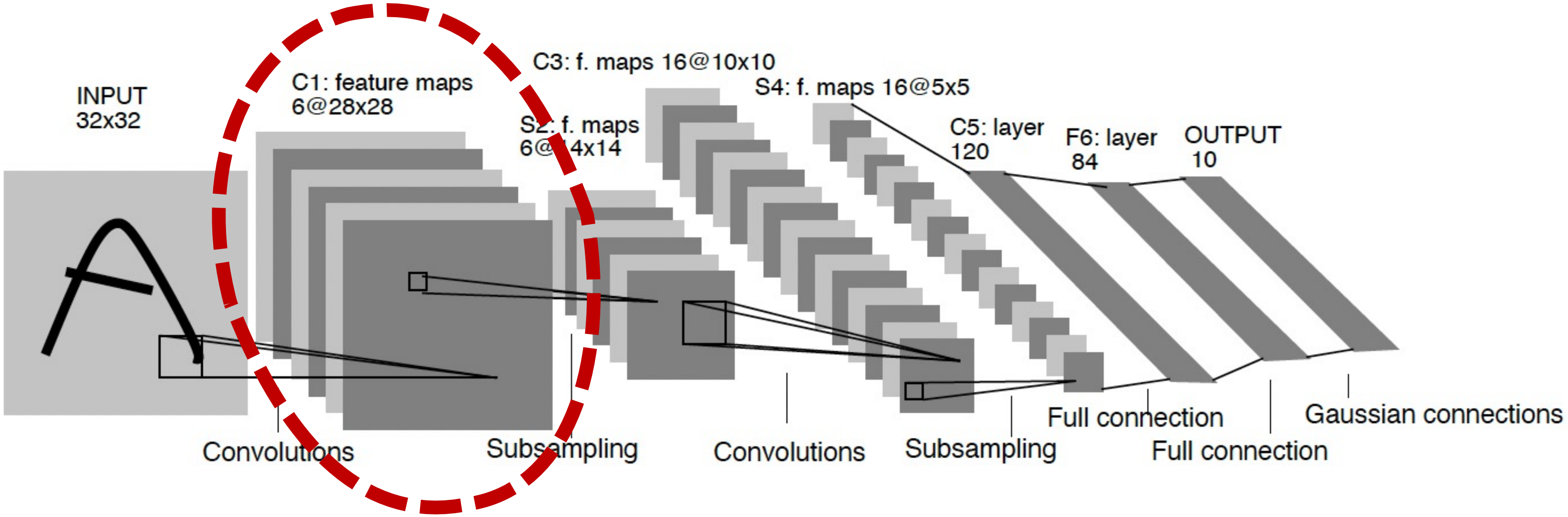


Figure credit: Christian Wolf

Example implementation from LeNet-5

A complete convolutional layer is composed of several feature maps so that multiple fatures can be extracted at multiple places



5x5 receptive field

6 feature maps, each with own set of weights

ConvNets approximate translation invariance

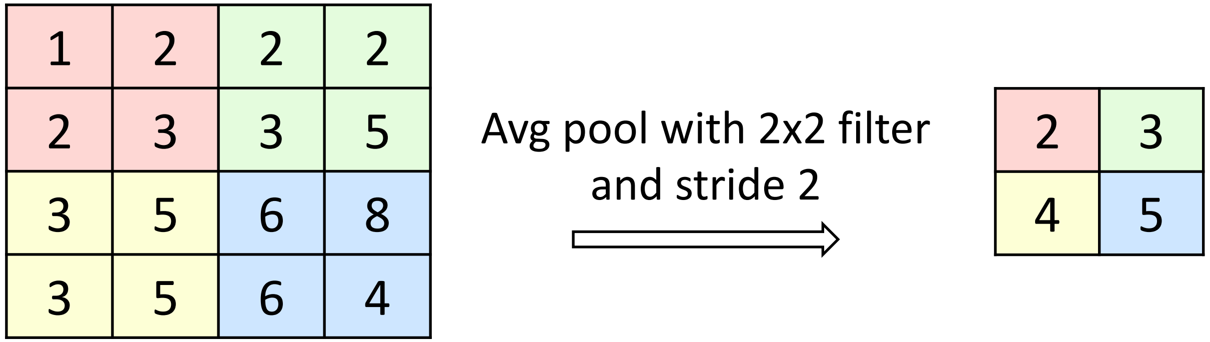
Exact locations are not important, relative locations are key

A simple way to soften position encoding is to reduce spatial resolution

Commonly known as **pooling**

avg pooling used in LeNet5

Also reduces computation



World not ready for ConvNet's yet?

LeNet-5 had good success for recognition of handwriting and machine-printed characters, but not much so beyond these domains.

Training was still slow.

At the same time Kernel Machines (SVM *etc.*) became very popular.

5. ImageNet with deep networks

The breakthrough of deep learning in computer vision happened when the AlexNet won the ImageNet large scale visual recognition challenge. In this chapter we detail the birth of deep learning, its absorption in convolutional neural networks and the record-breaking results in ImageNet.

Despite Backpropagation …

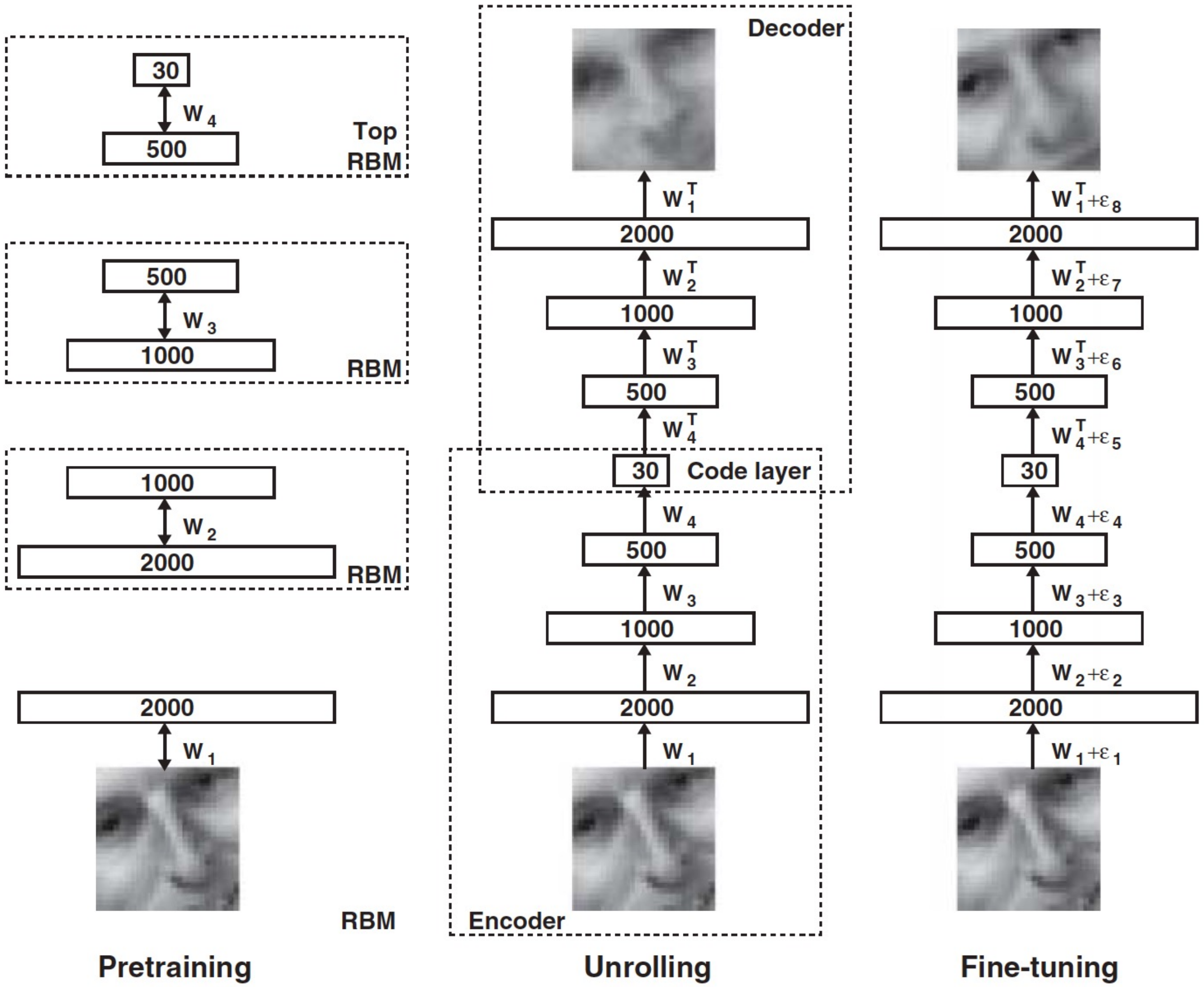
Experimentally, training multi-layer perceptrons was not that useful

- Accuracy didn't improve with more layers

The inevitable question

- Are 1-2 hidden layers the best neural networks can do?
- Or is it that the learning algorithm is not really mature yet

Deep learning arrives

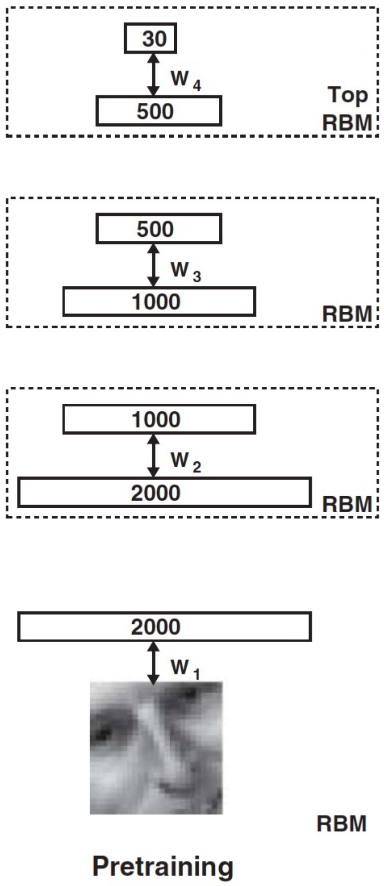


Hinton & Salakhutdinov, Science 2006

Deep learning arrives

Introduction of pretraining

Layer-by-layer learning

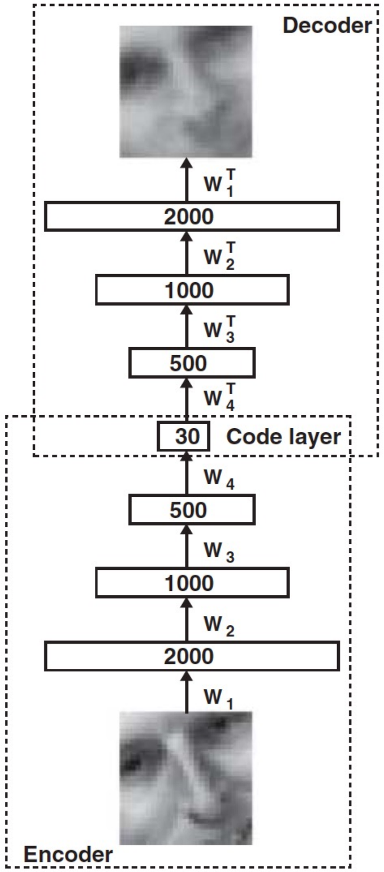


Deep learning arrives

Introduction of pretraining

Layer-by-layer learning

Unroll into encoder and decoder



Unrolling

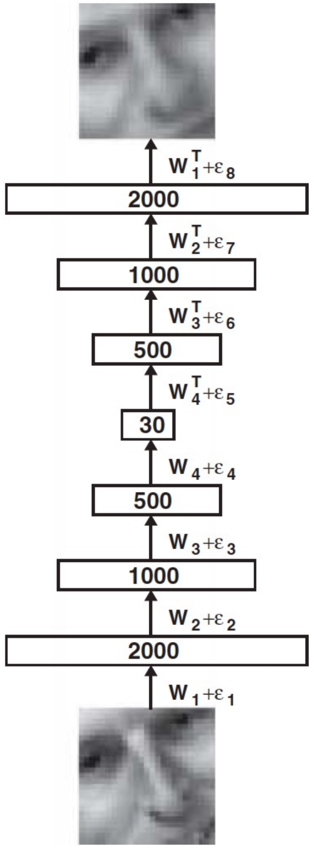
Deep learning arrives

Introduction of pretraining

Layer-by-layer learning

Unroll into encoder and decoder

Finetune entire network



Fine-tuning

Some results

Autoencoder to extract 30d codes for Olivetti face images

Input

Autoencoder

PCA

ImageNet arrives

In 2009 the ImageNet dataset was published

Collected images for each term of Wordnet

Tree of concepts organized hierarchically

“ambulance”, “Dalmatian dog”, “Egyptian cat”, …

Constructing ImageNet

Step 1:
Collect candidate images
via the Internet

Step 2:
Clean up the candidate
Images by humans

What is the downside of ImageNet construction?

Ethical and privacy concerns

- Containing personal information taken without consent,
- Unclear license usage,
- Biases, and,
- In some cases, even problematic image content.

Statistics

July 2008: 0 images

Dec 2008: 3 million images, 6K+ synsets

April 2010: 11 million images, 15K+ synsets

Finally: 14 million images, 21K synsets indexed

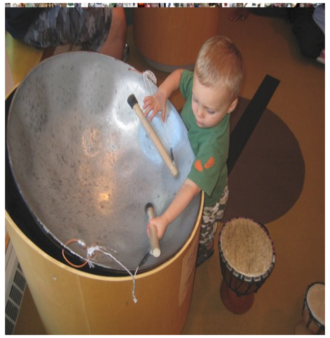
ImageNet Large Scale Visual Recognition

Ran from 2010 to 2017

- Today a Kaggle competition

Main task: image classification

- Automatically label 1.4M images with 1K objects
- Measure top-5 classification error



Output
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

✓

Output
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

ImageNet 2012 winner: AlexNet

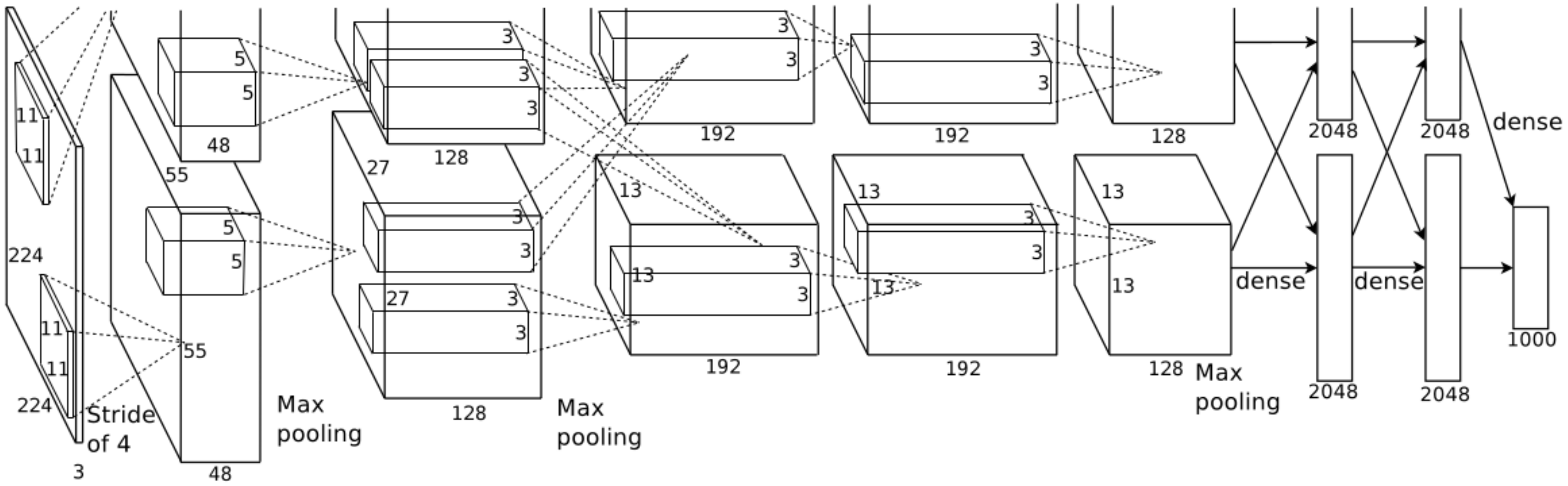


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

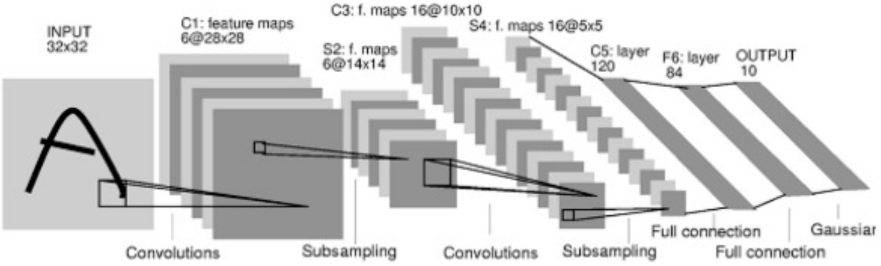
Krizhevsky, Sutskever & Hinton, NIPS 2012

AlexNet is a ConvNet

1998

LeCun et al.

LeNet



\# of transistors

10^6

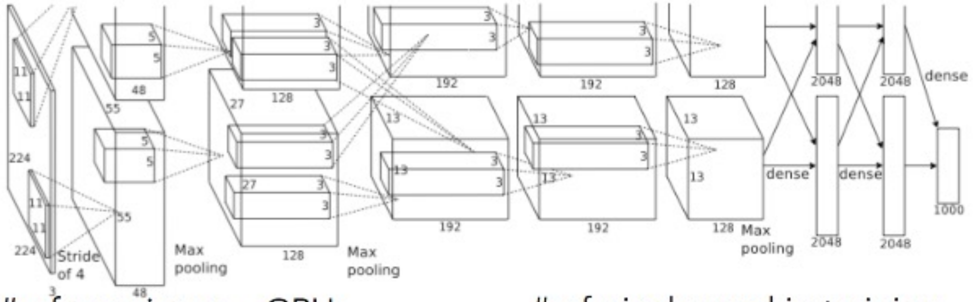
\# of pixels used in training

10^7

2012

Krizhevsky et al.

AlexNet



\# of transistors

10^9

GPUs

\# of pixels used in training

10^{14}

AlexNet introduced a few clever tricks

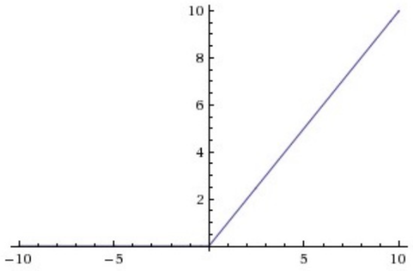
ReLU activation function

Data augmentation

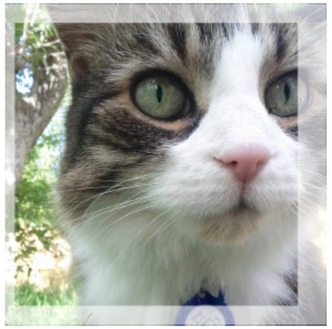
Testing on multiple crops

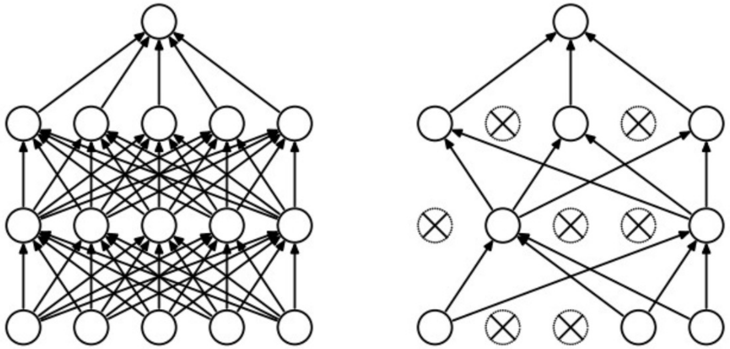
Dropout

GPU's



ReLU

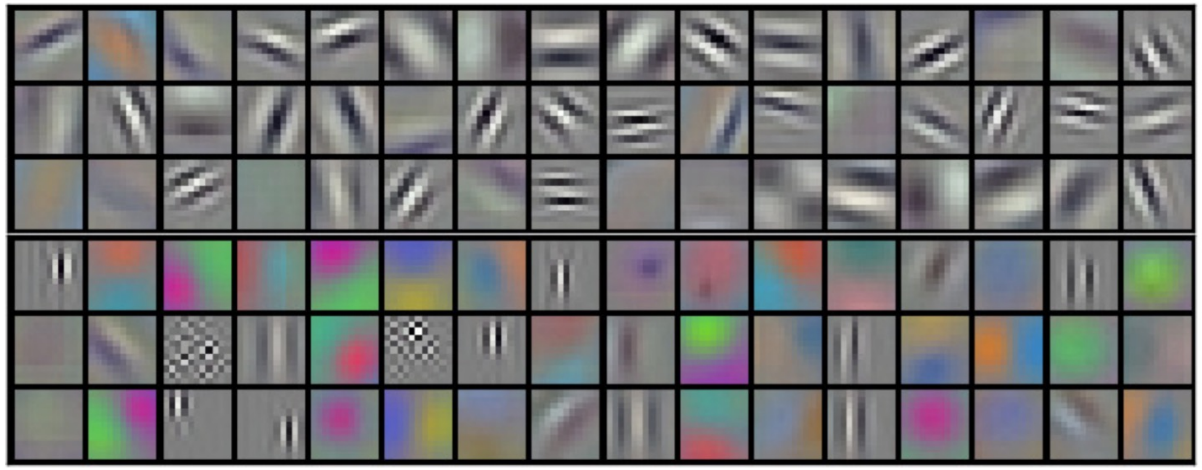




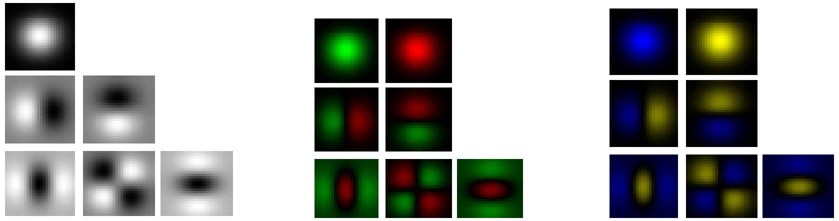
(a) Standard Neural Net

(b) After applying dropout.

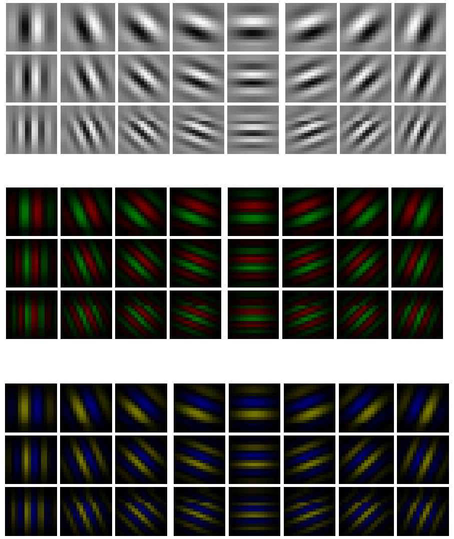
Learning or modeling?



Filters learned by first layer of AlexNet



Gaussian filters



Gabor filters