

Learning

- Learning agents
- Inductive learning
 - Different Learning Scenarios
 - Evaluation
- Neural Networks
 - Perceptrons
 - Multilayer Perceptrons
 - Deep Learning
- Reinforcement Learning
 - Temporal Differences
 - Q-Learning
 - SARSA

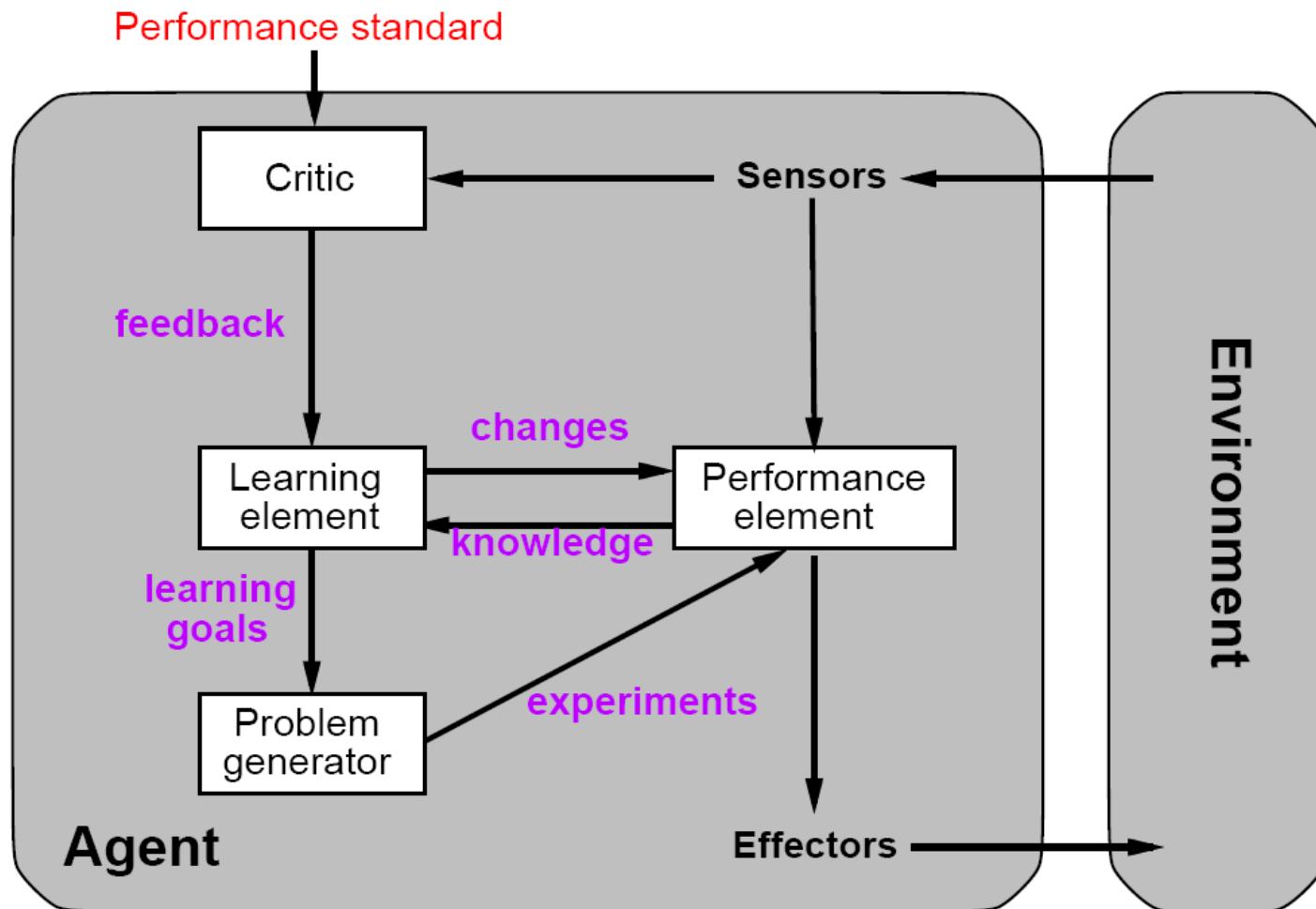
Material from
Russell & Norvig,
chapters 18.1,
18.2, 20.5 and 21

Slides based on Slides
by Russell/Norvig,
Ronald Williams,
Torsten Reil,
Alyosha Efros,
Fei-Fei li,
Viktoria Sharmanyska,
Geoff Hinton,
and many more

Learning

- Learning is essential for **unknown environments**,
 - i.e., when designer lacks omniscience
- Learning is useful as a **system construction method**,
 - i.e., expose the agent to reality rather than trying to write it down
- Learning modifies the agent's decision mechanisms to **improve performance**

Learning Agents



Learning Element

- Design of a learning element is affected by
 - Which components of the performance element are to be learned
 - What feedback is available to learn these components
 - What representation is used for the components
- Type of feedback:
 - Supervised learning:
 - correct answers for each example
 - Unsupervised learning:
 - correct answers not given
 - Reinforcement learning:
 - occasional rewards for good actions

Different Learning Scenarios

Supervised Learning

- A teacher provides the value for the target function for all training examples (labeled examples)
- concept learning, classification, regression

Semi-supervised Learning

- Only a subset of the training examples are labeled

Unsupervised Learning

- There is no information except the training examples
- clustering, subgroup discovery, association rule discovery

Inductive Learning

Simplest form: learn a function from examples

- f is the (unknown) **target function**
- An **example** is a pair $(x, f(x))$
- Problem: find a **hypothesis** h
 - given a **training set** of examples
 - such that $h \approx f$
 - on *all* examples
 - i.e. the hypothesis must **generalize** from the training examples
- This is a **highly simplified** model of real learning:
 - Ignores prior knowledge
 - Assumes examples are given

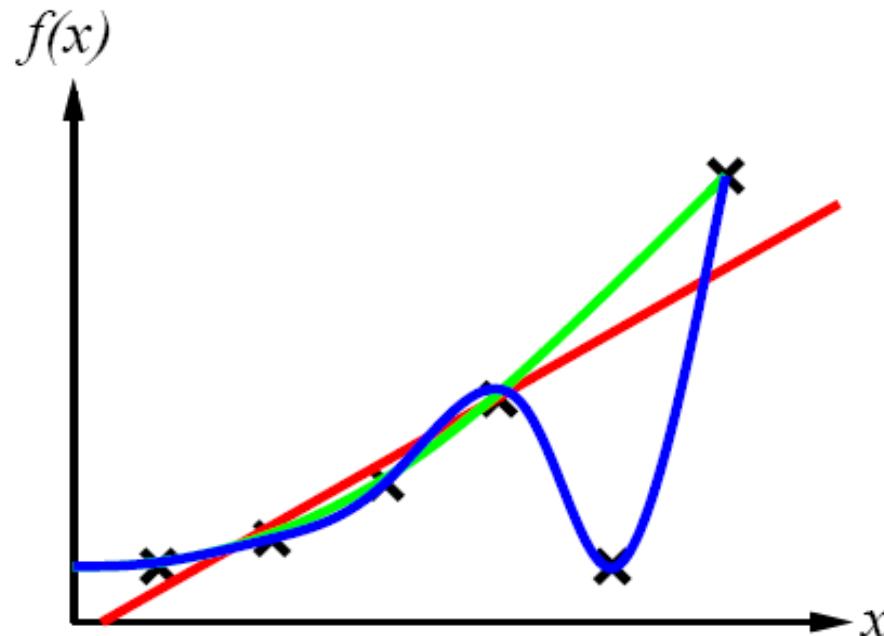
Inductive Learning Method

Construct/adjust h to agree with f on training set

- h is **consistent** if it agrees with f on all examples

Example:

- curve fitting



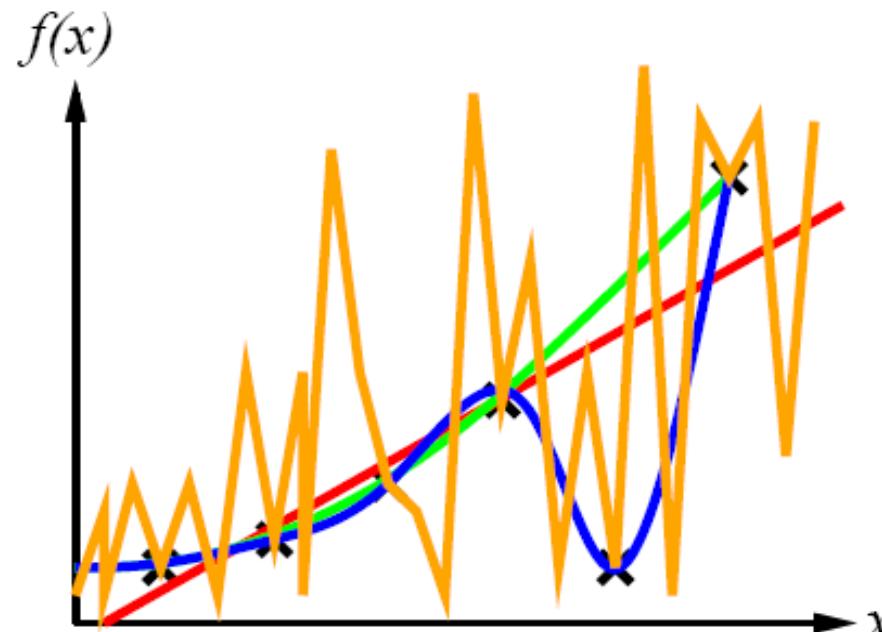
Inductive Learning Method

Construct/adjust h to agree with f on training set

- h is **consistent** if it agrees with f on all examples

Example:

- curve fitting

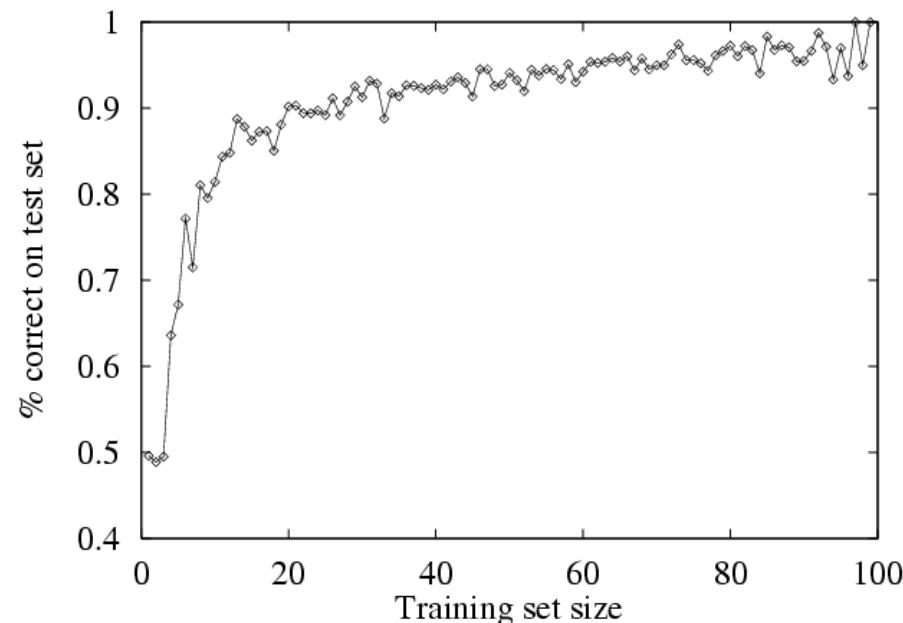


- **Ockham's Razor**
 - The best explanation is the simplest explanation that fits the data
- **Overfitting Avoidance**
 - maximize a combination of consistency and simplicity

Performance Measurement

- How do we know that $h \approx f$?
 - Use theorems of computational/statistical learning theory
 - Or try h on a new **test set** of examples where f is known
(use **same distribution** over example space as training set)

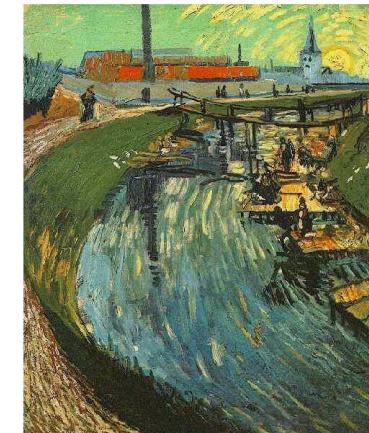
Learning curve = % correct on test set over training set size



Pigeons as Art Experts

Famous experiment (Watanabe *et al.* 1995, 2001)

- Pigeon in Skinner box
- Present paintings of two different artists (e.g. Chagall / Van Gogh)
- Reward for pecking when presented a particular artist



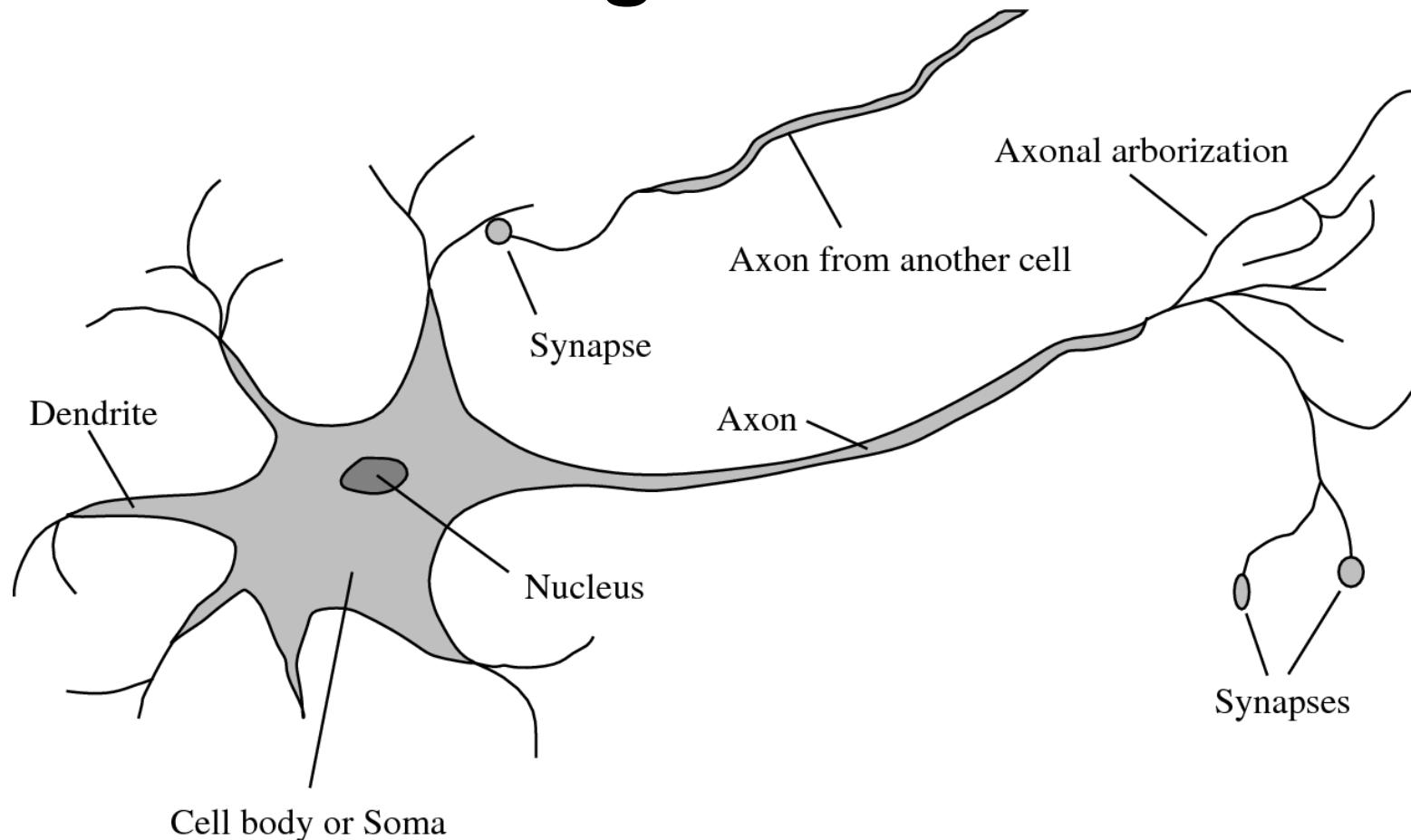
Results

- Pigeons were able to discriminate between Van Gogh and Chagall with 95% accuracy
 - when presented with pictures they had been trained on
 - Discrimination still 85% successful for previously unseen paintings of the artists
- Pigeons do not simply memorise the pictures
 - They can extract and recognise patterns (the ‘style’)
 - They generalise from the already seen to make predictions
- This is what neural networks (biological and artificial) are good at (unlike conventional computer)

What are Neural Networks?

- **(Highly simplified)** models of the brain and nervous system
- Highly parallel
 - Process information much more like the brain than a serial computer
- Learning
- Very simple principles
- Very complex behaviours
- Applications
 - As powerful problem solvers
 - As (approx.) biological models

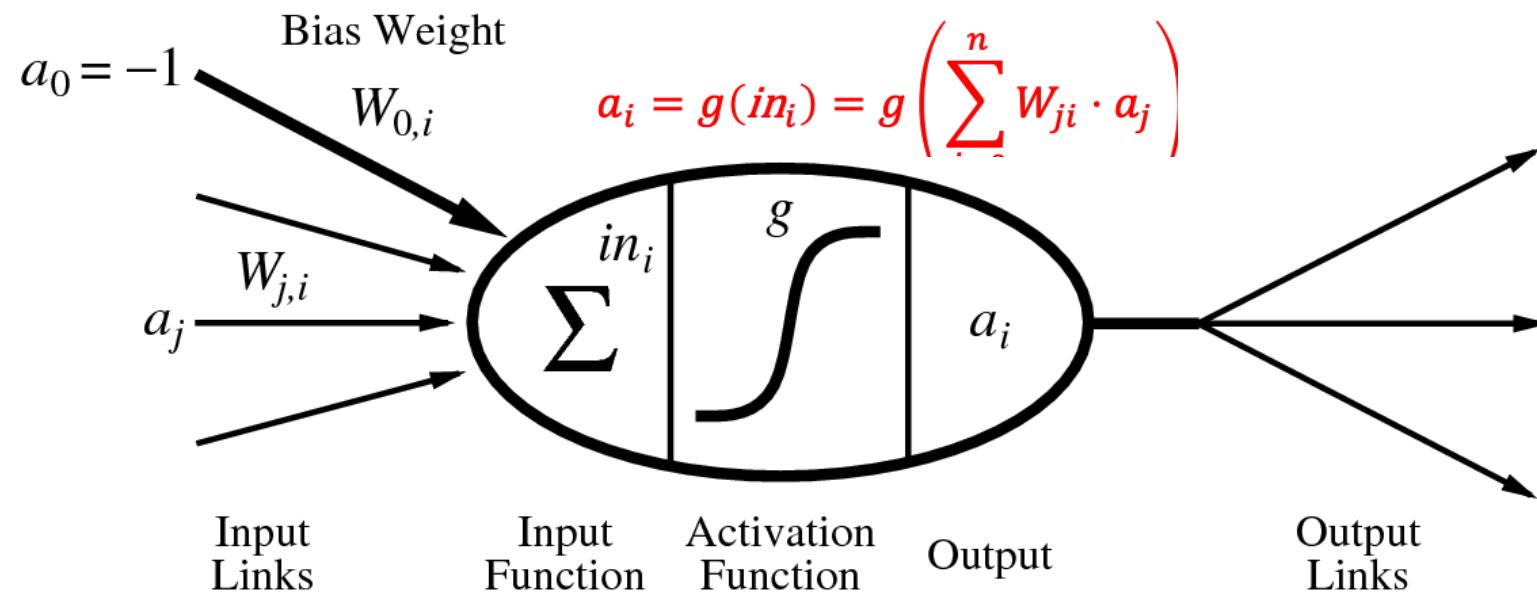
A Biological Neuron



- Neurons are connected to each other via synapses
- If a neuron is activated, it spreads its activation to all connected neurons

An Artificial Neuron

(McCulloch-Pitts, 1943)



- Neurons correspond to nodes or **units**
- A **link** from unit j to unit i propagates activation a_j from j to i
- The **weight** $W_{j,i}$ of the link determines the strength and sign of the connection
- The total **input activation** is the sum of the input activations
- The **output activation** is determined by the activation function g

Perceptron

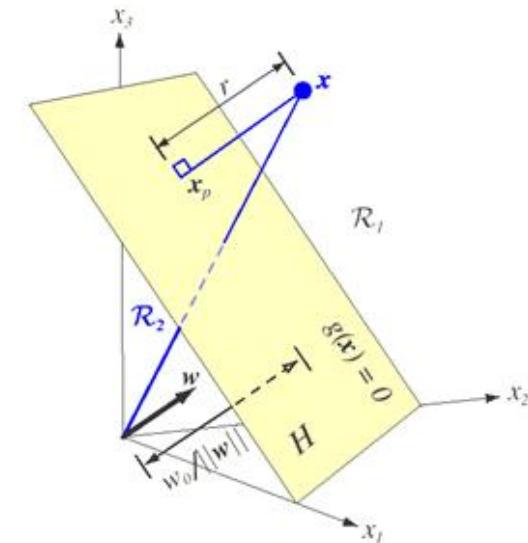
(Rosenblatt 1957, 1960)

- A single node
 - connecting n input signals a_j with one output signal a
 - typically signals are -1 or $+1$

- Activation function
 - A simple threshold function:

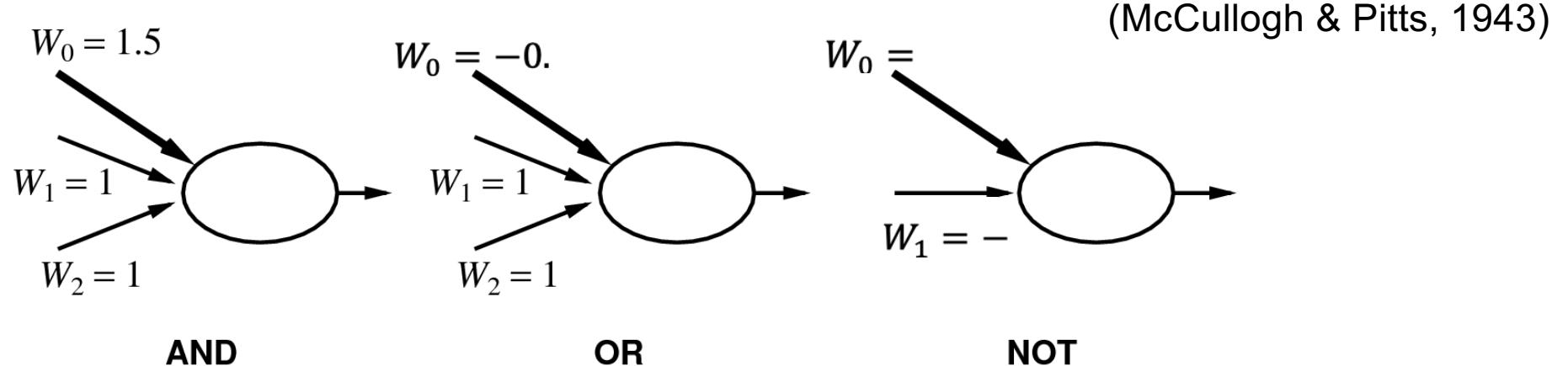
$$a = \begin{cases} -1 & \text{if } \sum_{j=0}^n w_j \cdot a_j \leq \\ & \\ 1 & \text{if } \sum_{i=0}^n w_j \cdot a_j > \end{cases}$$

- Thus it implements a **linear separator**
 - i.e., a hyperplane that divides n -dimensional space into a region with output -1 and a region with output 1

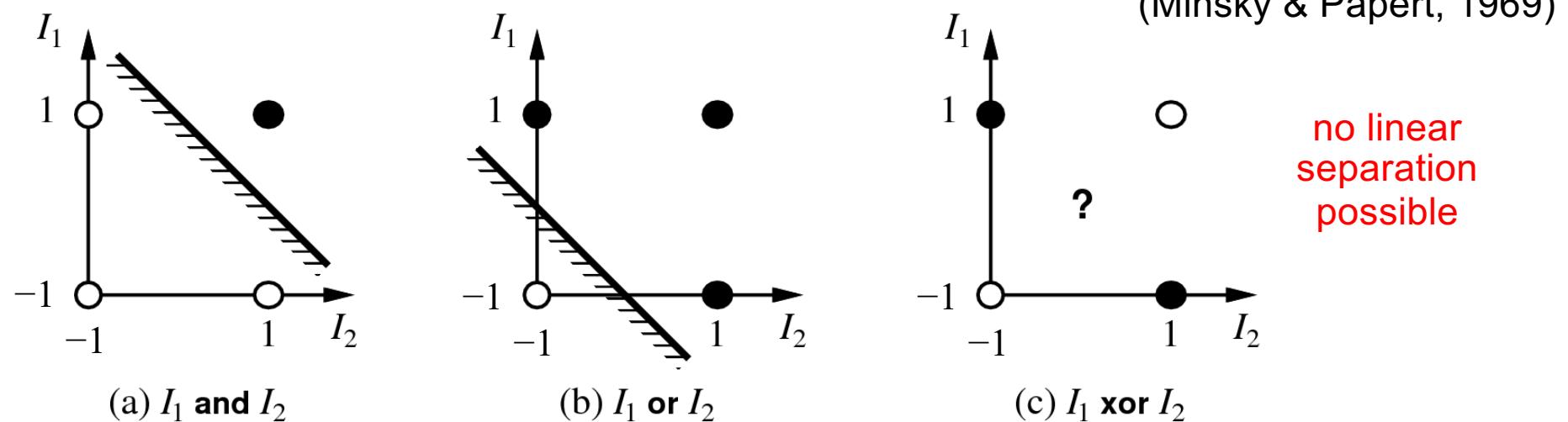


Perceptrons and Boolean Functions

- a Perceptron can implement all elementary logical functions



- more complex functions like XOR cannot be modeled



Perceptron Learning

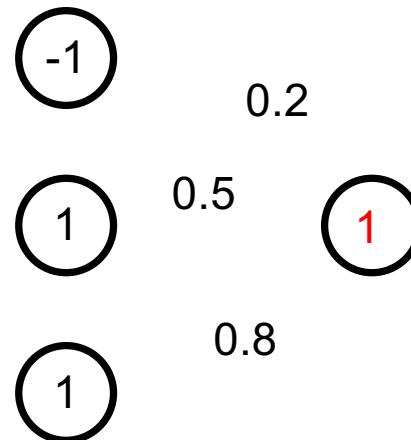
- Perceptron Learning Rule for Supervised Learning

$$W_j \leftarrow W_j + \alpha \cdot (f(\mathbf{x}) - h(\mathbf{x})) \cdot x_j$$

learning rate error

- Example:

Computation of output signal $h(x)$



$$\text{in}(x) = -1 \cdot 0.2 + 1 \cdot 0.5 + 1 \cdot 0.8 =$$

$h(x) = 1$ because $\text{in}(x) > 0$ (activation function)

Assume target value $f(x) = -1$ (and $\alpha = 0.5$)

$$W_0 \leftarrow 0.2 + 0.5 \cdot (-1 - 1) \cdot -1 = 0.2 + 1 :$$

$$W_1 \leftarrow 0.5 + 0.5 \cdot (-1 - 1) \cdot 1 = 0.5 - 1 =$$

$$W_2 \leftarrow 0.8 + 0.5 \cdot (-1 - 1) \cdot 1 = 0.8 - 1 =$$

Measuring the Error of a Network

- The error for one training example \mathbf{x} can be measured by the squared error
 - the squared difference of the output value $h(\mathbf{x})$ and the desired target value $f(\mathbf{x})$

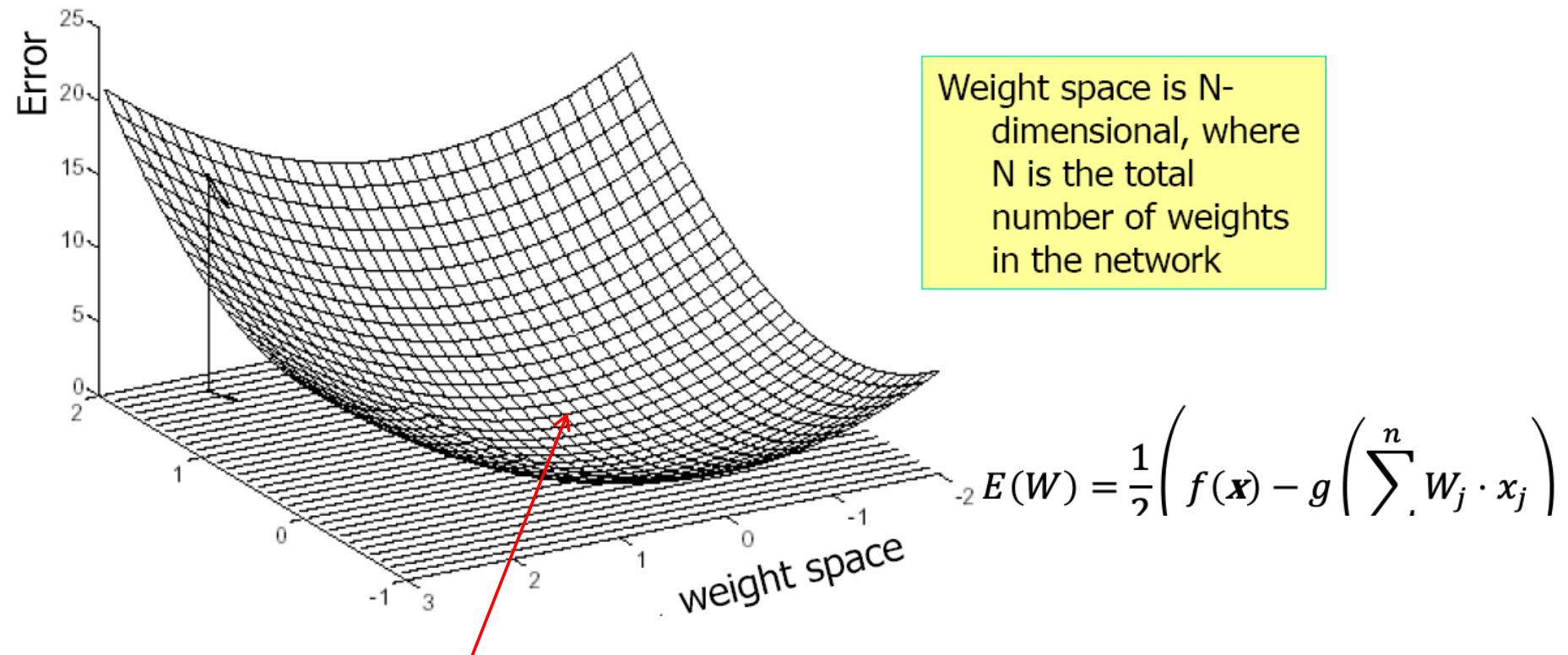
- $$E(\mathbf{x}) = \frac{1}{2} Err^2 = \frac{1}{2} (f(\mathbf{x}) - h(\mathbf{x}))^2 = \frac{1}{2} \left(f(\mathbf{x}) - g\left(\sum_{i=0}^n w_i \cdot x_i\right) \right)^2$$

- For evaluating the performance of a network, we can try the network on a set of datapoints and average the value (= sum of squared errors)

$$E(\text{Network}) = \sum_{i=1}^N E(\mathbf{x}_i)$$

Error Landscape

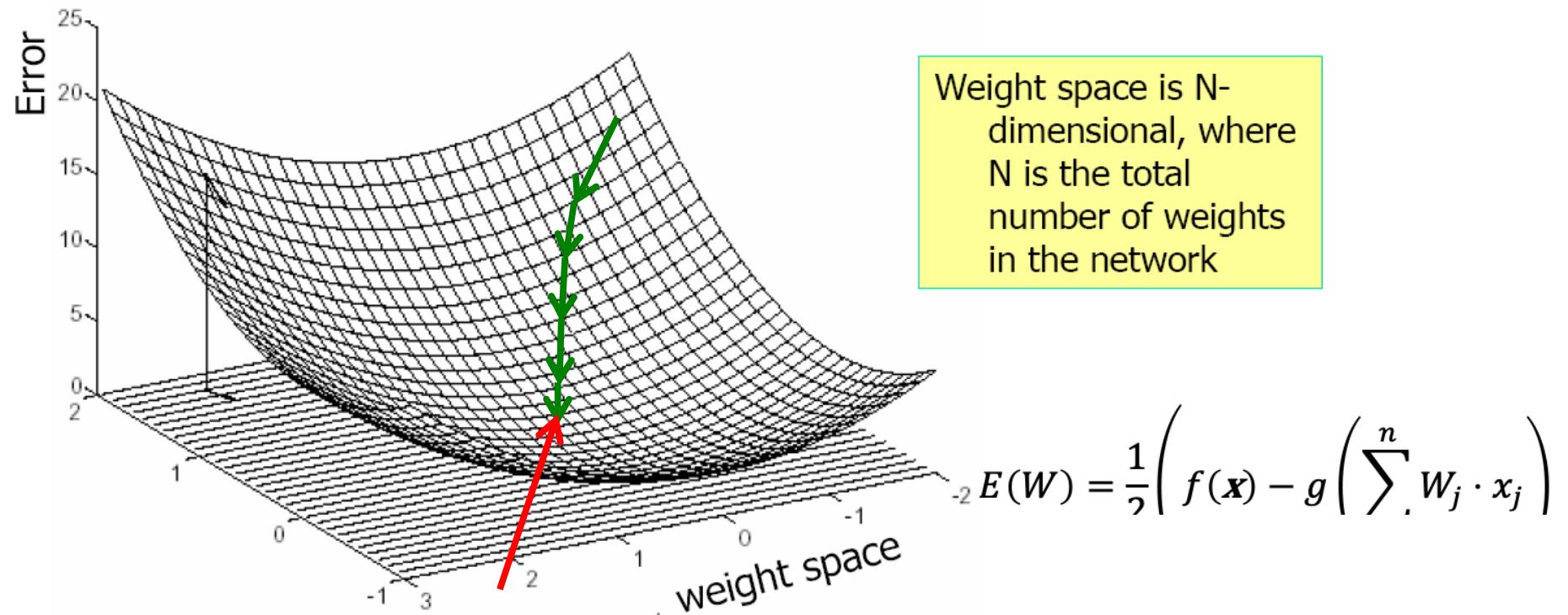
- The error function for one training example may be considered as a function in a multi-dimensional weight space



- The best weight setting for one example is where the error measure for this example is minimal

Error Minimization via Gradient Descent

- In order to find the point with the minimal error:
 - go downhill in the direction where it is steepest



- ... but make small steps, or you might shoot over the target

Error Minimization

- It is easy to derive a perceptron training algorithm that minimizes the squared error

$$E = \frac{1}{2} Err^2 = \frac{1}{2} (f(\mathbf{x}) - h(\mathbf{x}))^2 = \frac{1}{2} \left(f(\mathbf{x}) - g \left(\sum_j w_j \cdot x_j \right) \right)^2$$

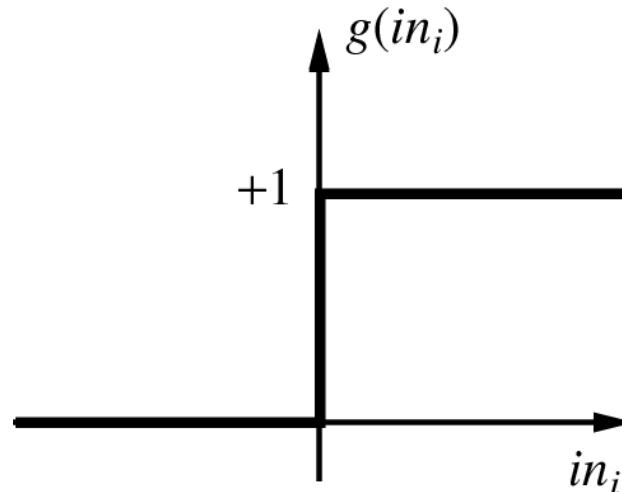
- Change weights into the direction of the steepest descent of the error function

$$\frac{\partial E}{\partial w_j} = Err \cdot \frac{\partial Err}{\partial w_j} = Err \cdot \frac{\partial}{\partial w_j} \left(f(\mathbf{x}) - g \left(\sum_k w_k \cdot x_k \right) \right) = -Err \cdot g'(in) \cdot x_j$$

- To compute this, we need a continuous and differentiable activation function g !
- Weight update with learning rate α : $w_j \leftarrow w_j + \alpha \cdot Err \cdot g'(in) \cdot x_j$
 - positive error → increase network output
 - increase weights of nodes with positive input
 - decrease weights of nodes with negative input

Threshold Activation Function

- The regular threshold activation function is problematic
 - $g'(x) = 0$, therefore $\frac{\partial E}{\partial W_{j,i}} = -Err \cdot g'(in_i) \cdot x_j = 0$
 - → no weight changes!

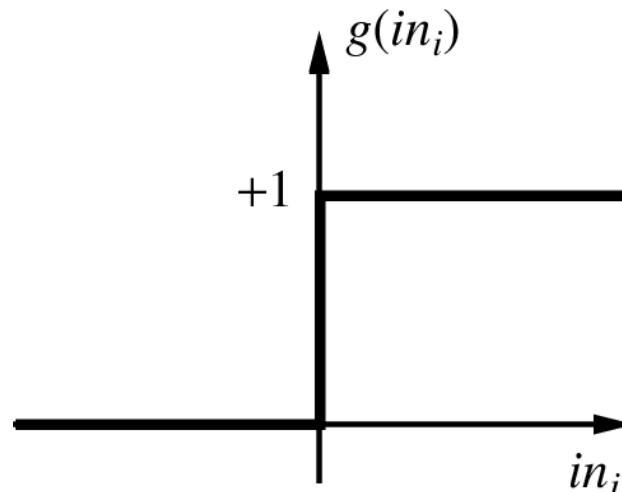


$$g(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

$$g'(x) = 0$$

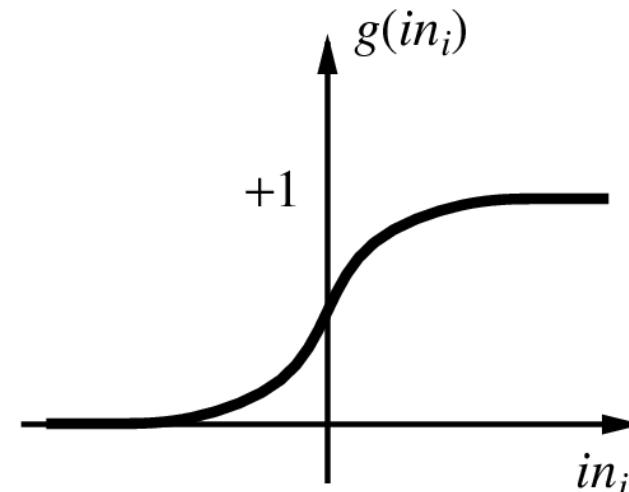
Sigmoid Activation Function

- A commonly used activation function is the sigmoid function
 - similar to the threshold function
 - easy to differentiate
 - non-linear



$$g(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

$$g'(x) = 0$$

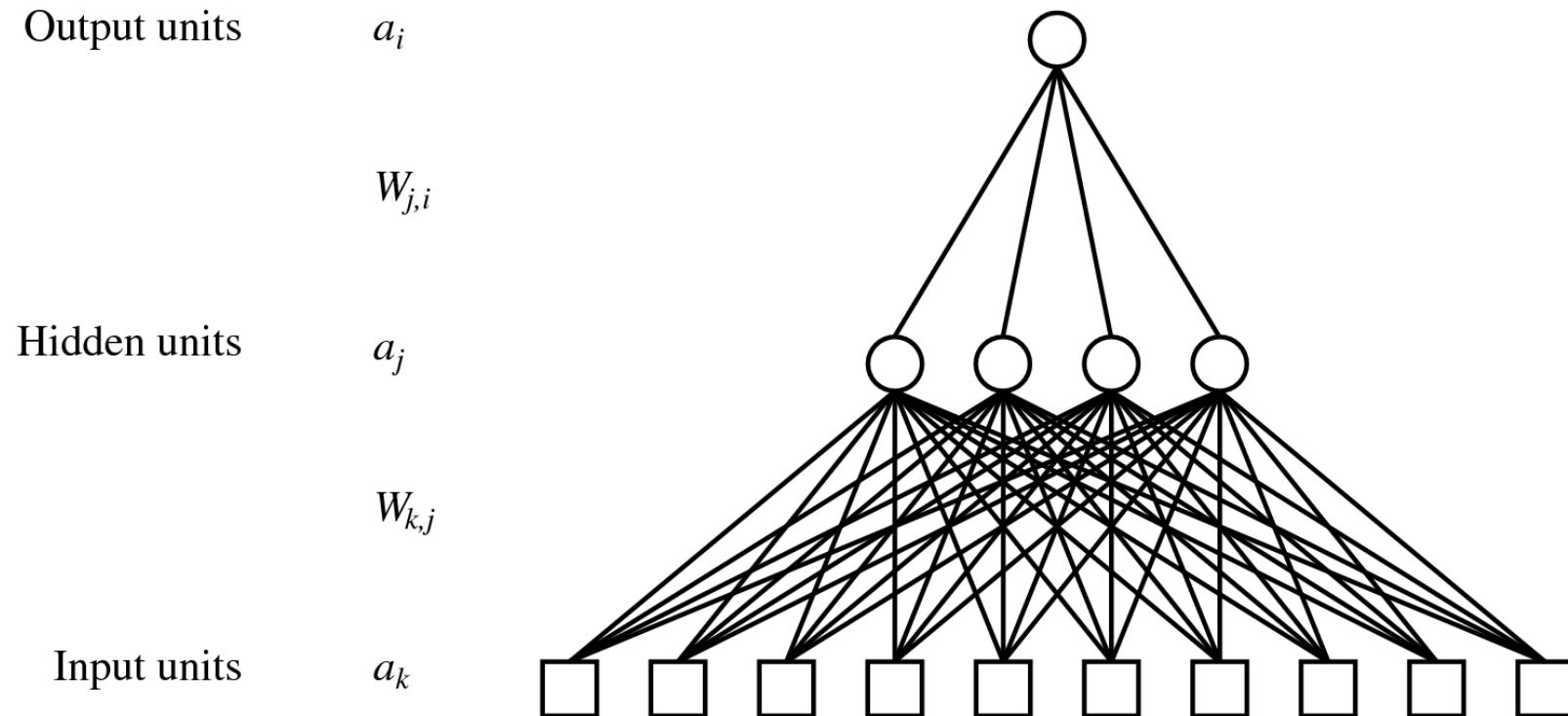


$$g(x) = \frac{1}{1 + e^{-x}}$$

$$g'(x) = g(x)(1 - g(x))$$

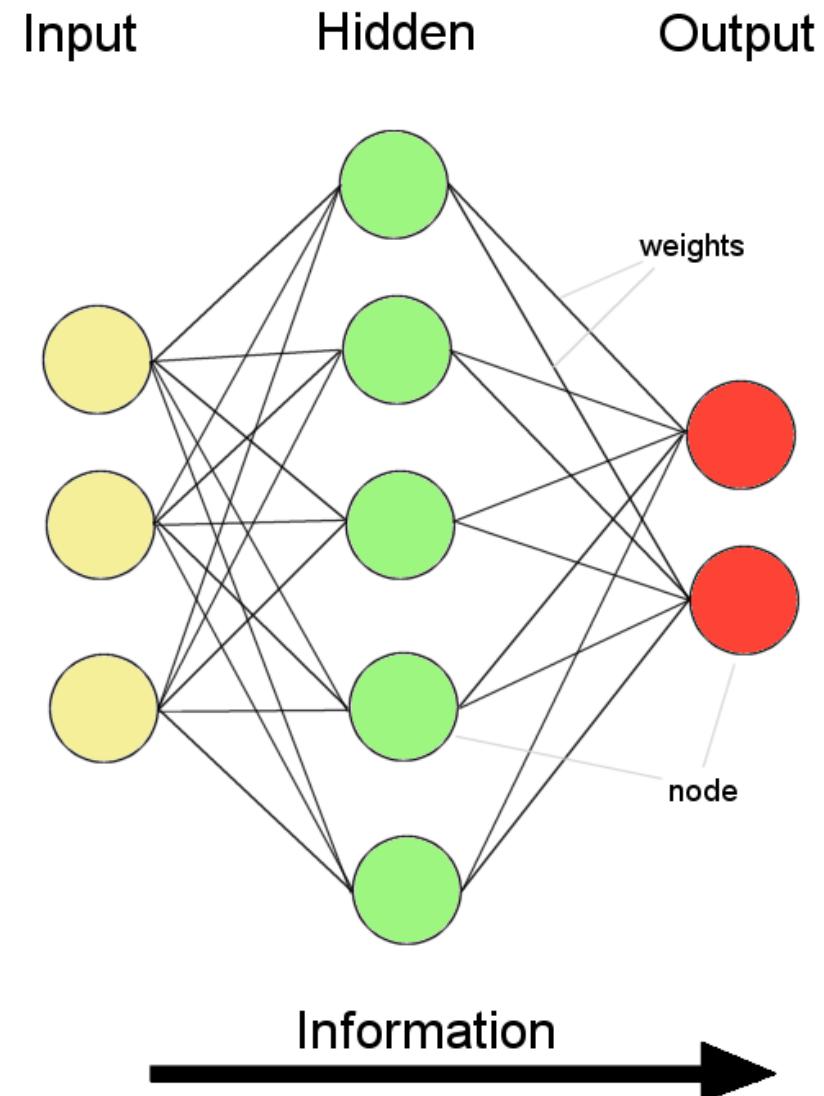
Multilayer Perceptrons

- Perceptrons may have multiple output nodes
 - may be viewed as multiple parallel perceptrons
- The output nodes may be combined with another perceptron
 - which may also have multiple output nodes
- The size of this **hidden layer** is determined manually



Multilayer Perceptrons

- Information flow is unidirectional
- Data is presented to *Input layer*
- Passed on to *Hidden Layer*
- Passed on to *Output layer*
- Information is distributed
- Information processing is parallel



Expressiveness of MLPs

- Every continuous function can be modeled with three layers
 - i.e., with one hidden layer
- Every function can be modeled with four layers
 - i.e., with two hidden layers

Backpropagation Learning

- The **output nodes** are trained like a normal perceptron

$$W_{ji} \leftarrow W_{ji} + \alpha \cdot Err_i \cdot g'(in_i) \cdot x_j = W_{ji} + \alpha \cdot \Delta_i \cdot x_j$$

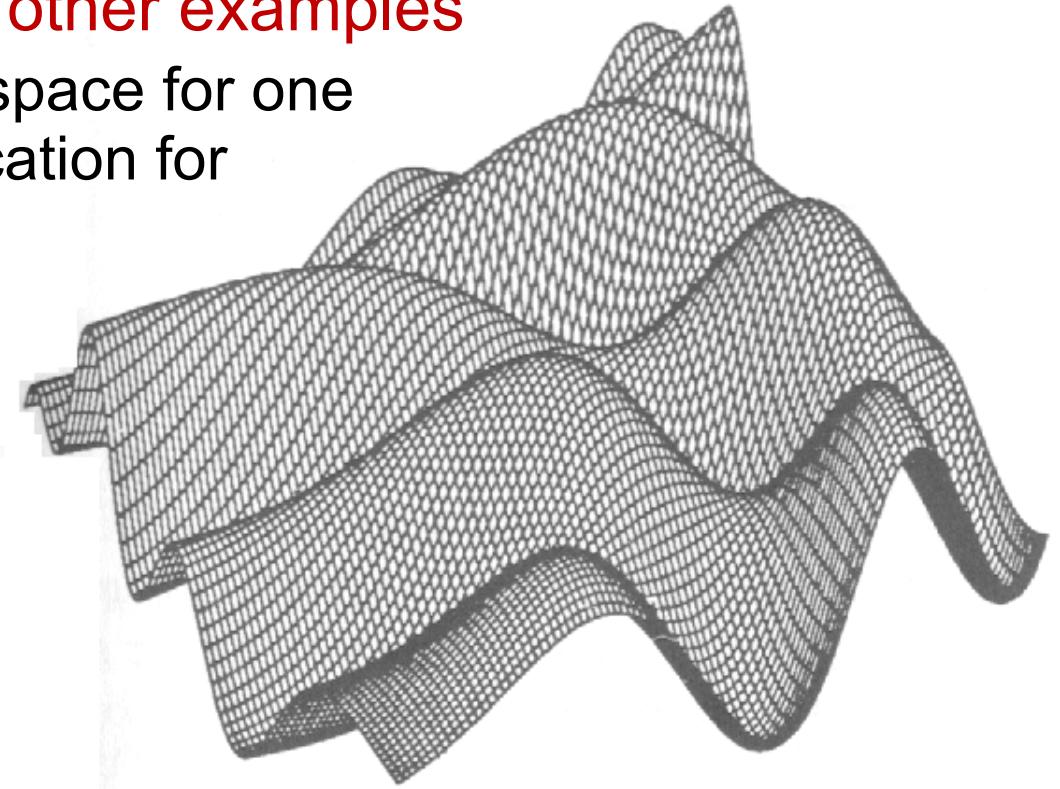
- Δ_i is the error term of output node i times the derivation of its inputs
- the error term Δ_i of the output layers is propagated back to the **hidden layer**

$$\Delta_j = \left(\sum_i W_{ji} \cdot \Delta_i \right) \cdot g'(in_j) \quad W_{kj} \leftarrow W_{kj} + \alpha \cdot \Delta_j \cdot x_k$$

- the training signal of hidden layer node j is the weighted sum of the errors of the output nodes
- Thus the information provided by the **gradient flows backwards** through the network

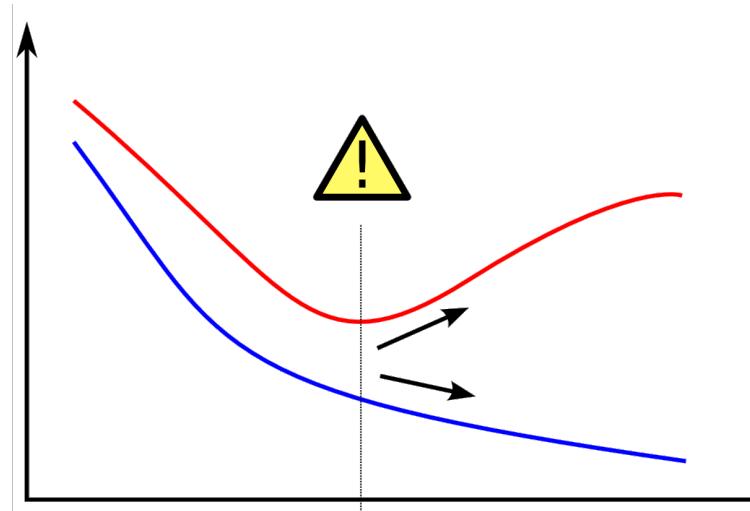
Minimizing the Network Error

- The error landscape for the entire network may be thought of as the sum of the error functions of all examples
 - will yield many local minima → hard to find global minimum
- Minimizing the error for one training example may destroy what has been learned for other examples
 - a good location in weight space for one example may be a bad location for another examples
- Training procedure:
 - try all examples in turn
 - make small adjustments for each example
 - repeat until convergence
- One Epoch = One iteration through all examples



Overfitting

- Training Set Error continues to decrease with increasing number of training examples / number of epochs
 - an epoch is a complete pass through all training examples
- Test Set Error will start to increase because of overfitting



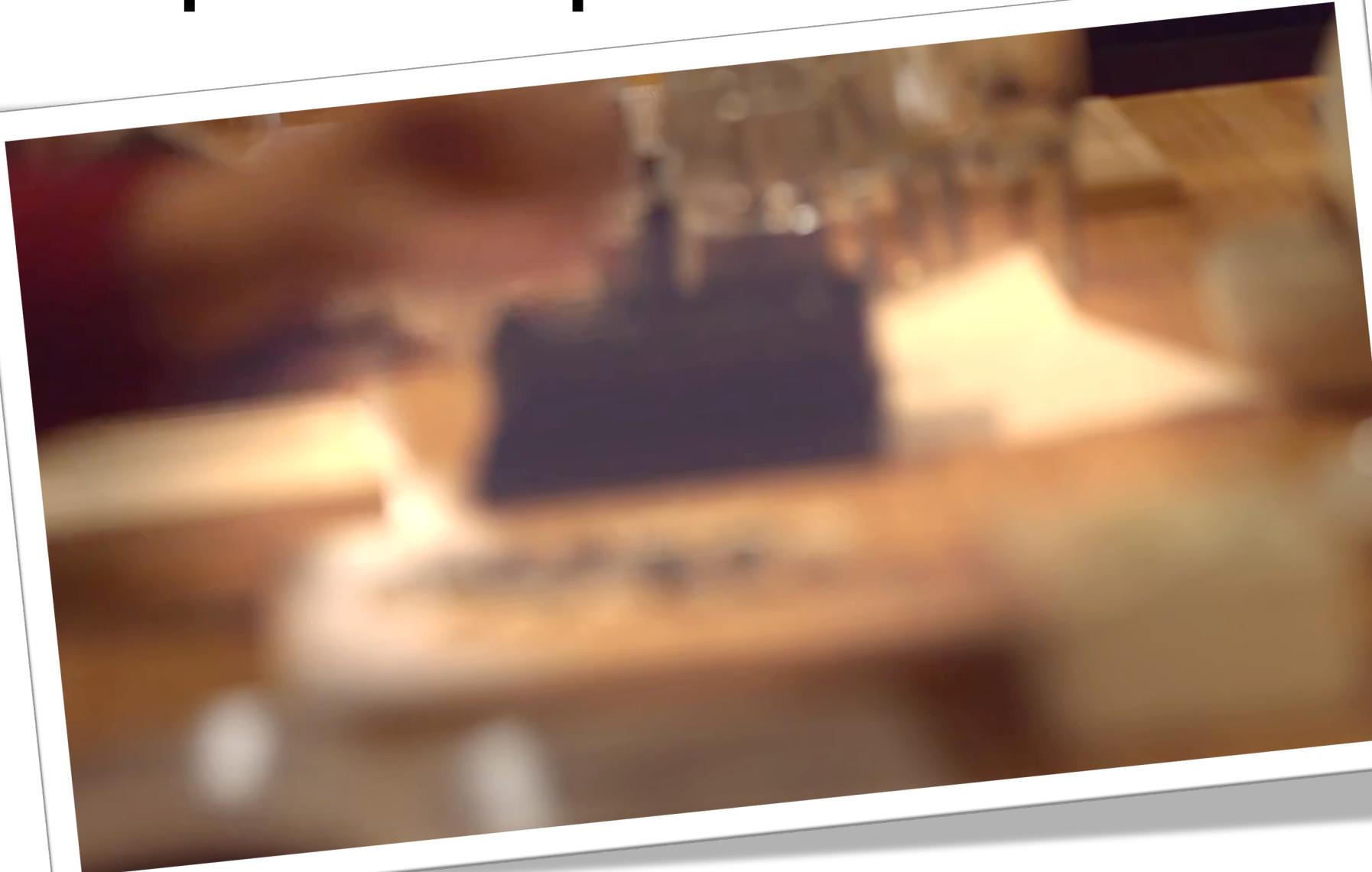
- Simple training protocol:
 - keep a separate validation set to watch the performance
 - validation set is different from training and test sets!
 - stop training if error on validation set gets down

Deep Learning

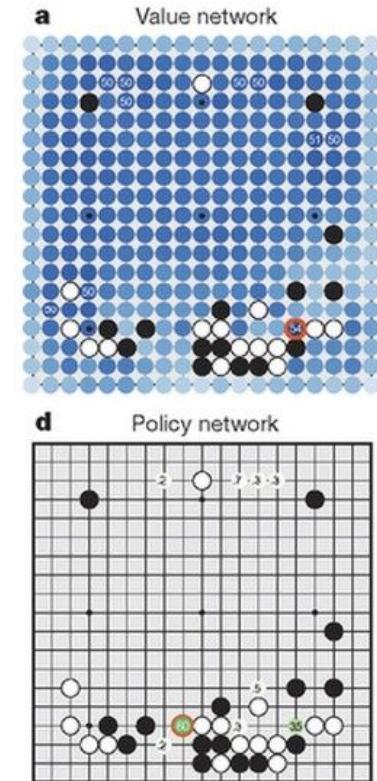
- In the last years, great success has been observed with training „deep“ neural networks
 - Deep networks are networks with multiple layers
- Key ingredients:
 - A lot of training data are needed and available (big data)
 - Fast processing and a few new tricks made fast training for big data possible

Watch NATURE video at <https://www.youtube.com/watch?v=g-dKXOlsf98>

DeepMind's AlphaGo



DeepMind's AlphaGo



Deep policy network is trained to produce probability map of promising moves. The deep value network is used to prune the (mcmc) search tree

Goal of Deep Architectures

To this aim most approaches use (stacked) neural networks

Deep learning methods aim at

- learning feature hierarchies
- where features from higher levels of the hierarchy are formed by lower level features.

High-level semenatical representations

Edges, local shapes, object parts

Low level representation

very high level representation:

MAN SITTING ...

... etc ...

slightly higher level representation

raw input vector representation:

$$\mathcal{V} = [23 \ 19 \ 20 \ \dots \ 18]$$

$x_1 \ x_2 \ x_3 \ \dots \ x_n$



Figure is from Yoshua Bengio

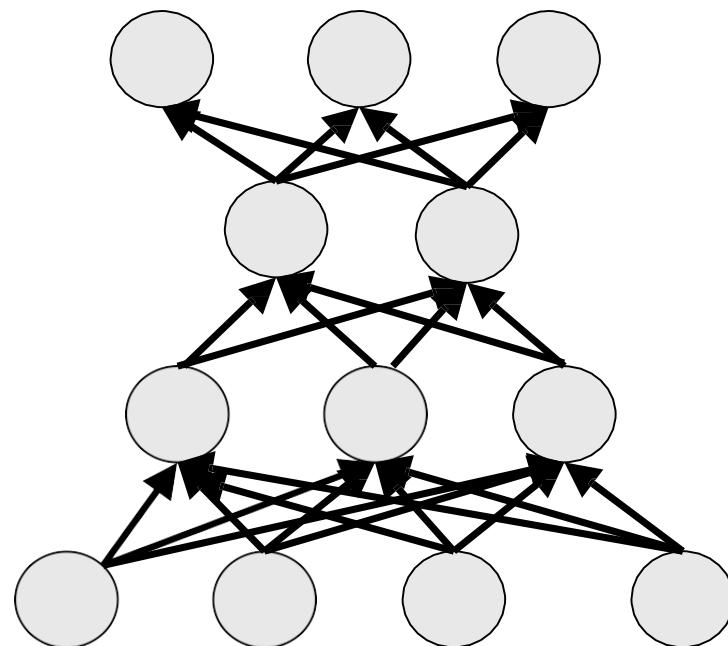
Deep Architectures

Deep architectures are composed of multiple levels of non-linear operations, such as neural nets with many hidden layers.

Output layer

Hidden layers

Input layer



Examples of non-linear activations:

$$\tanh(x)$$

$$\sigma(x) = (1 + e^{-x})^{-1}$$

$$\max(0, x)$$

In practice, NN with multiple hidden layers work better than with a single hidden layer.

For training use backpropagation as discussed before

(Deep) Convolutional Networks CNNs

Compared to standard neural networks with similarly-sized layers,

- CNNs have much fewer connections and parameters
- and so they are easier to train
- and typically have more than five layers (a number of layers which makes fully-connected neural networks almost impossible to train properly when initialized randomly)

LeNet, 1998 LeCun Y, Bottou L, Bengio Y, Haffner P: Gradient-Based Learning Applied to Document Recognition, Proceedings of the IEEE

AlexNet, 2012 Krizhevsky A, Sutskever I, Hinton G: ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012

First breakthrough on image classification using ImageNet Dataset

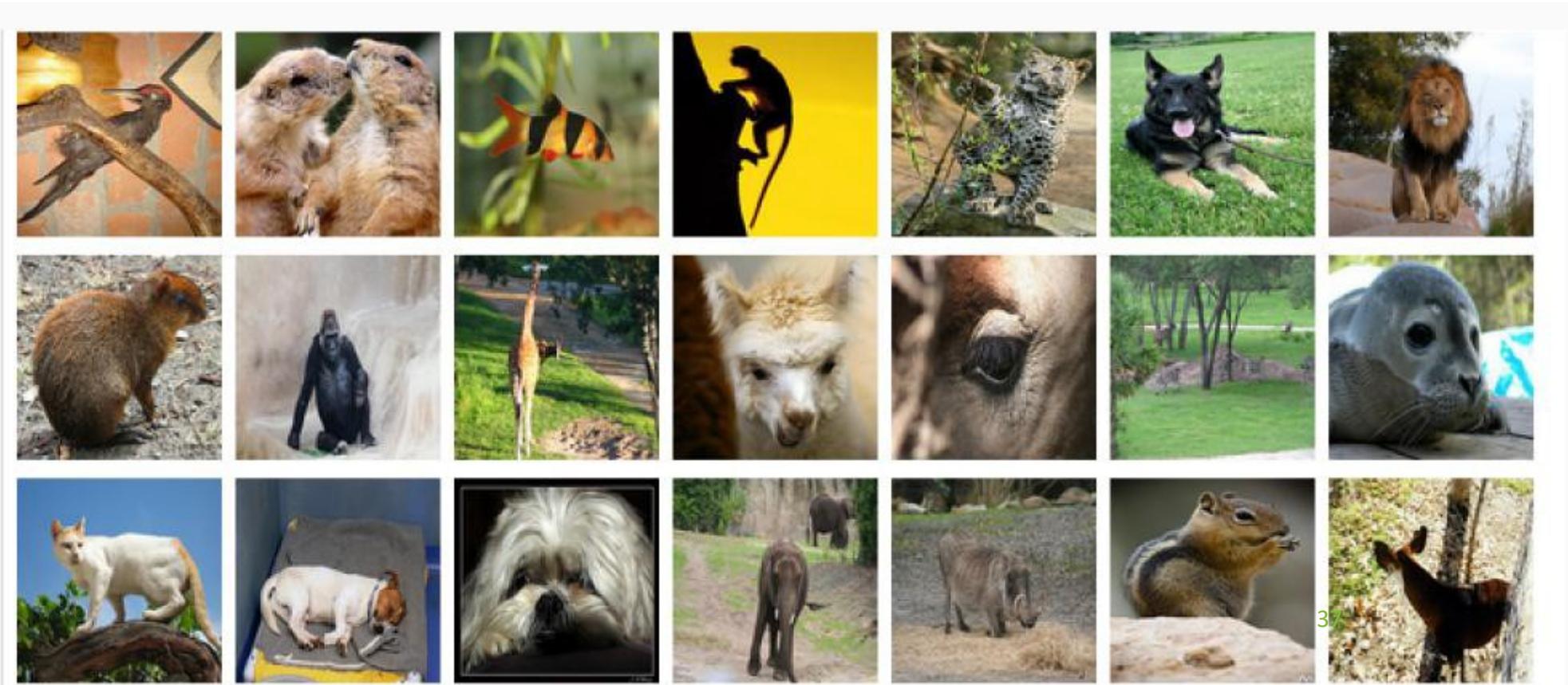
- 15M images
- 22K categories
- Images collected from Web
- Human labelers (Amazon's Mechanical Turk crowd-sourcing)
- ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2010)
 - 1K categories
 - 1.2M training images (~1000 per category)
 - 50,000 validation images
 - 150,000 testing images
- RGB images; mean normalization
- Variable-resolution, but this architecture scales them to 256x256 size

36

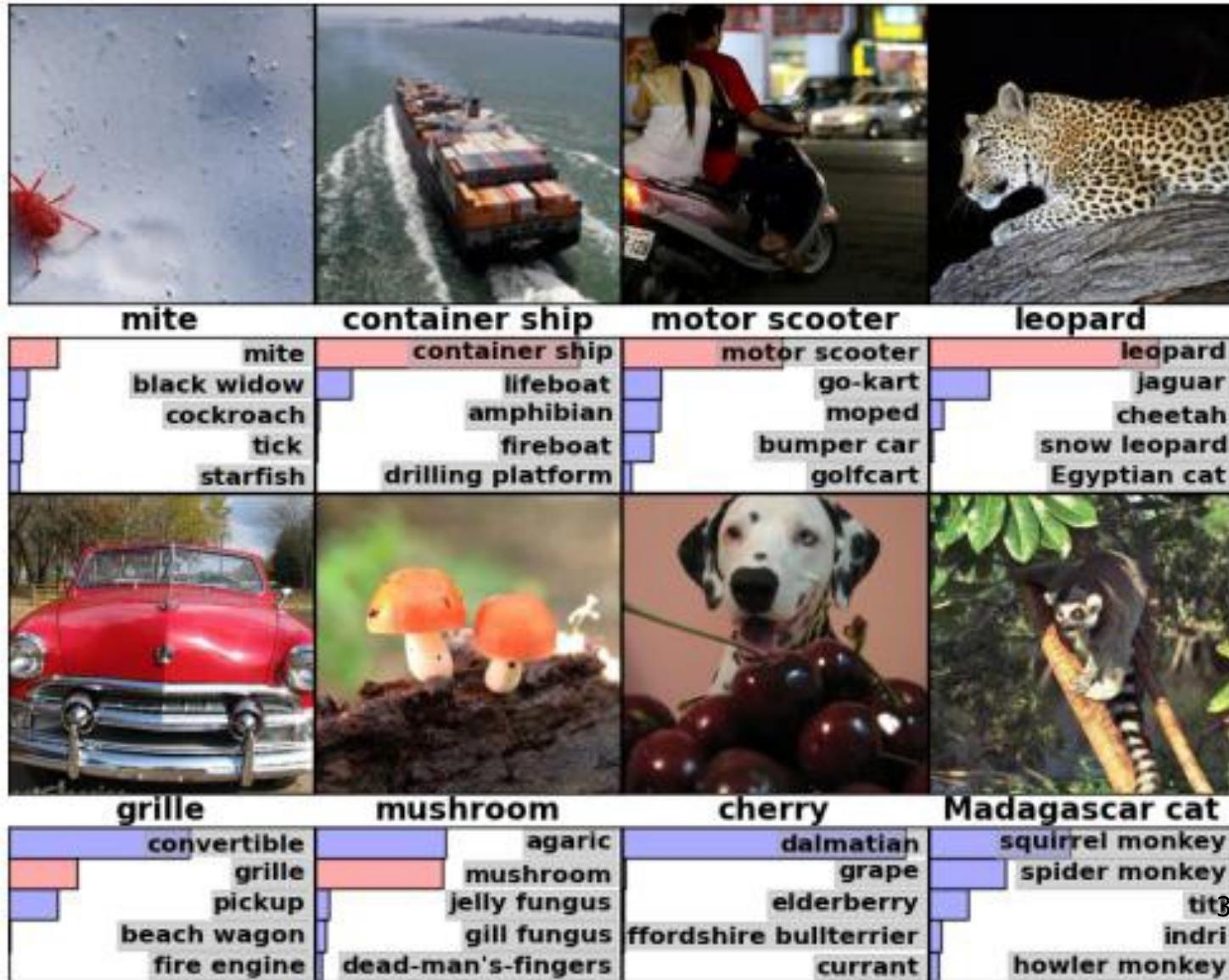
ImageNet

Classification goals:

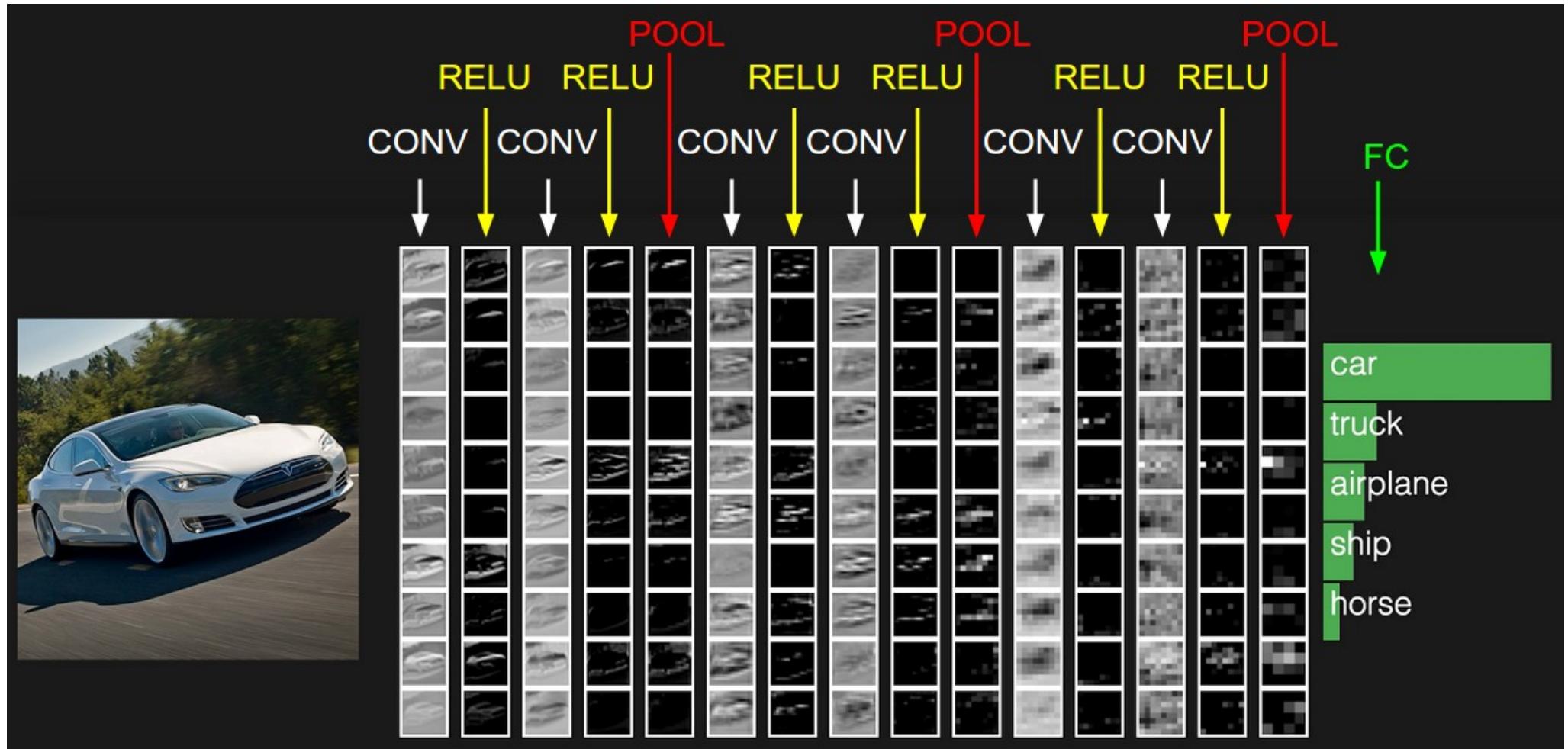
- Make 1 guess about the label (Top-1 error)
- make 5 guesses about the label (Top-5 error)



Results: ImageNet



CNN Architecture Convolutions, RELU, Pooling

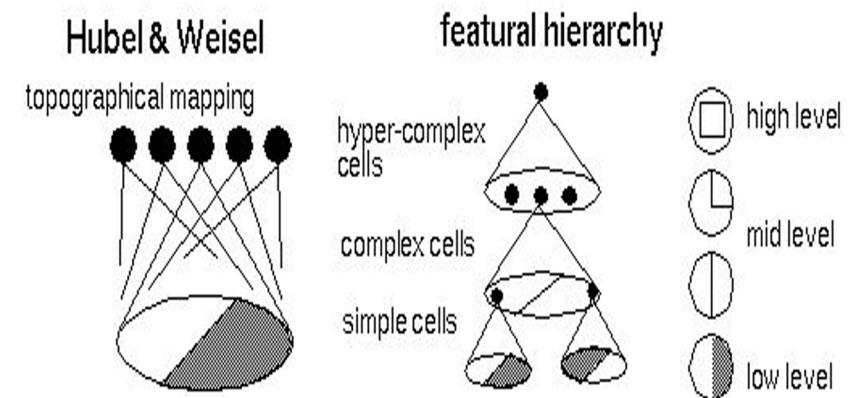


Deep Learning- Kristian Kersting

39

Cortical Receptive Fields

David Hubel & Thorston Wiesel, Nobel Prize 1981



<https://www.youtube.com/watch?v=IOHayh06LJ4>

Convolutional Neural Networks

- Convolution:
 - for each pixel of an image, a new feature is computed using a weighted combination of its $n \times n$ neighborhood

35	40	41	45	50
40	40	42	46	52
42	46	50	55	55
48	52	56	58	60
56	60	65	70	75



5x5 image

0	1	0		
0	0	0		
0	0	0		

3x3 convolution
runs over all
possible 3x3
subimages
of picture



resulting image
only one
pixel shown

Convolution - Blur



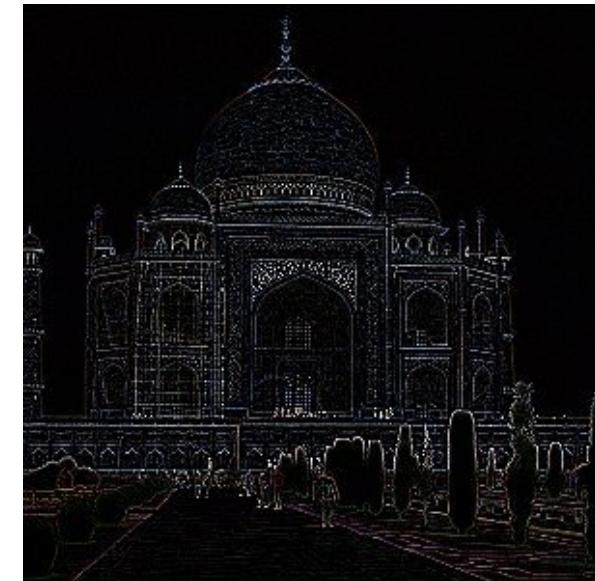
0	0	0	0	0
0	1	1	1	0
0	1	1	1	0
0	1	1	1	0
0	0	0	0	0



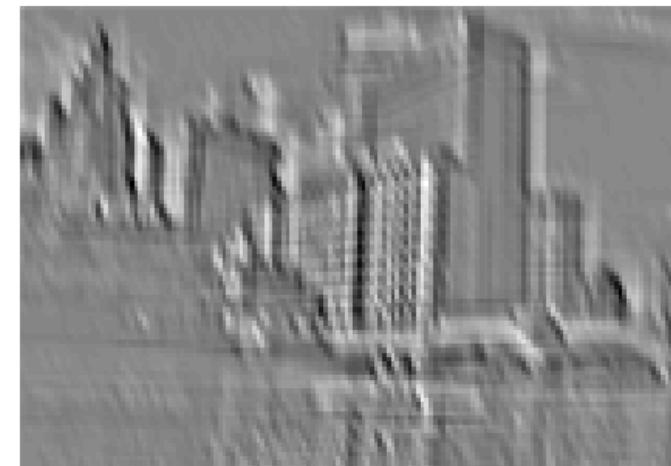
Convolution - Edge detection



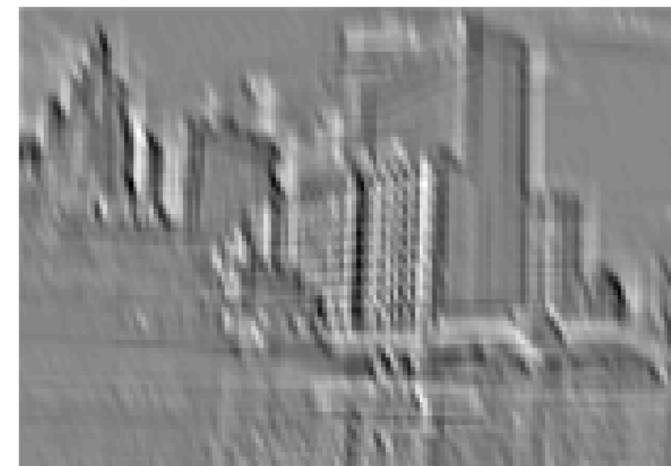
$$\begin{array}{|c|c|c|} \hline & & \\ \hline & 0 & 1 & 0 \\ \hline 1 & -4 & 1 \\ \hline 0 & 1 & 0 \\ \hline \end{array}$$



Outputs of Convolution



Outputs of Convolution

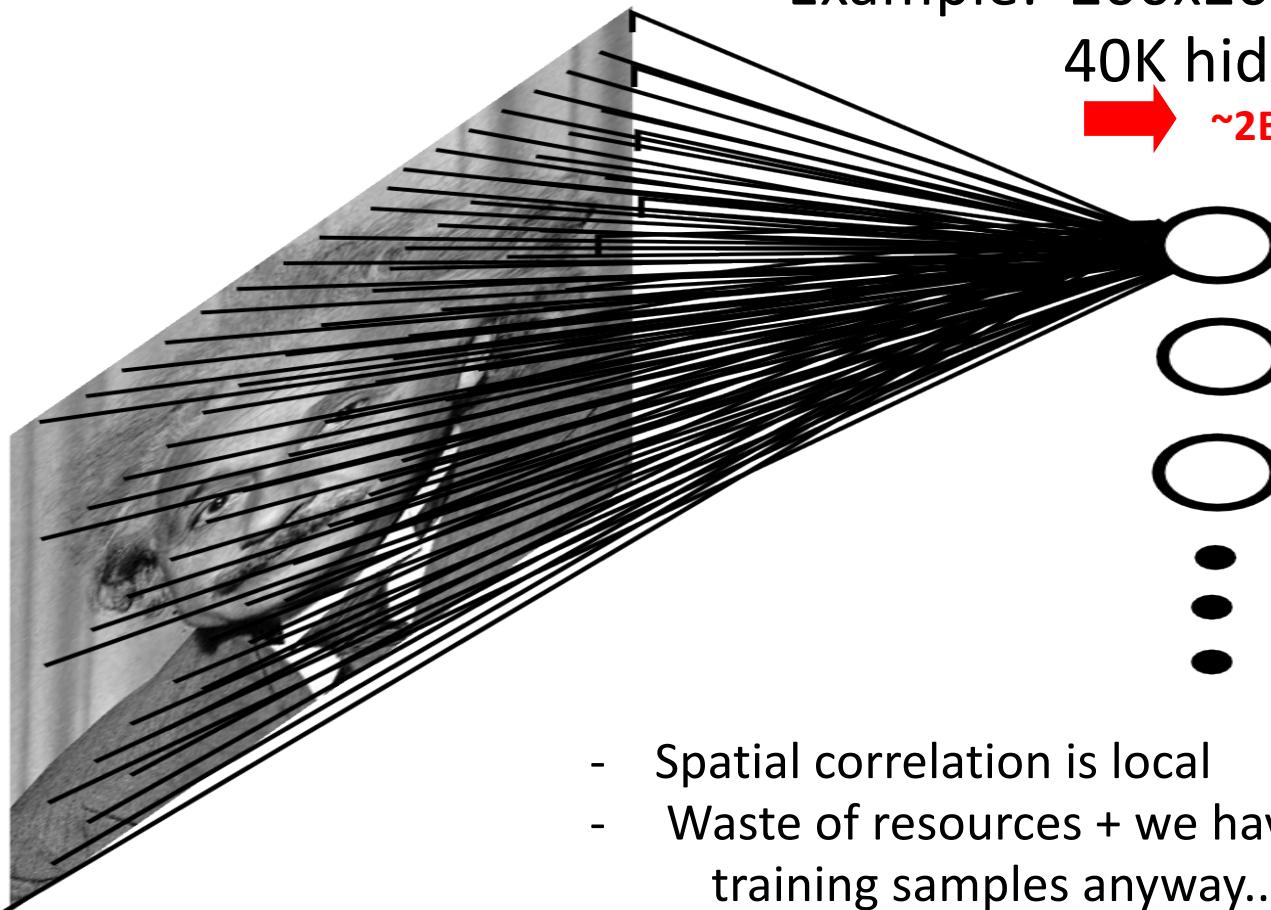


Fully Connected Layer

Example: 200x200 image

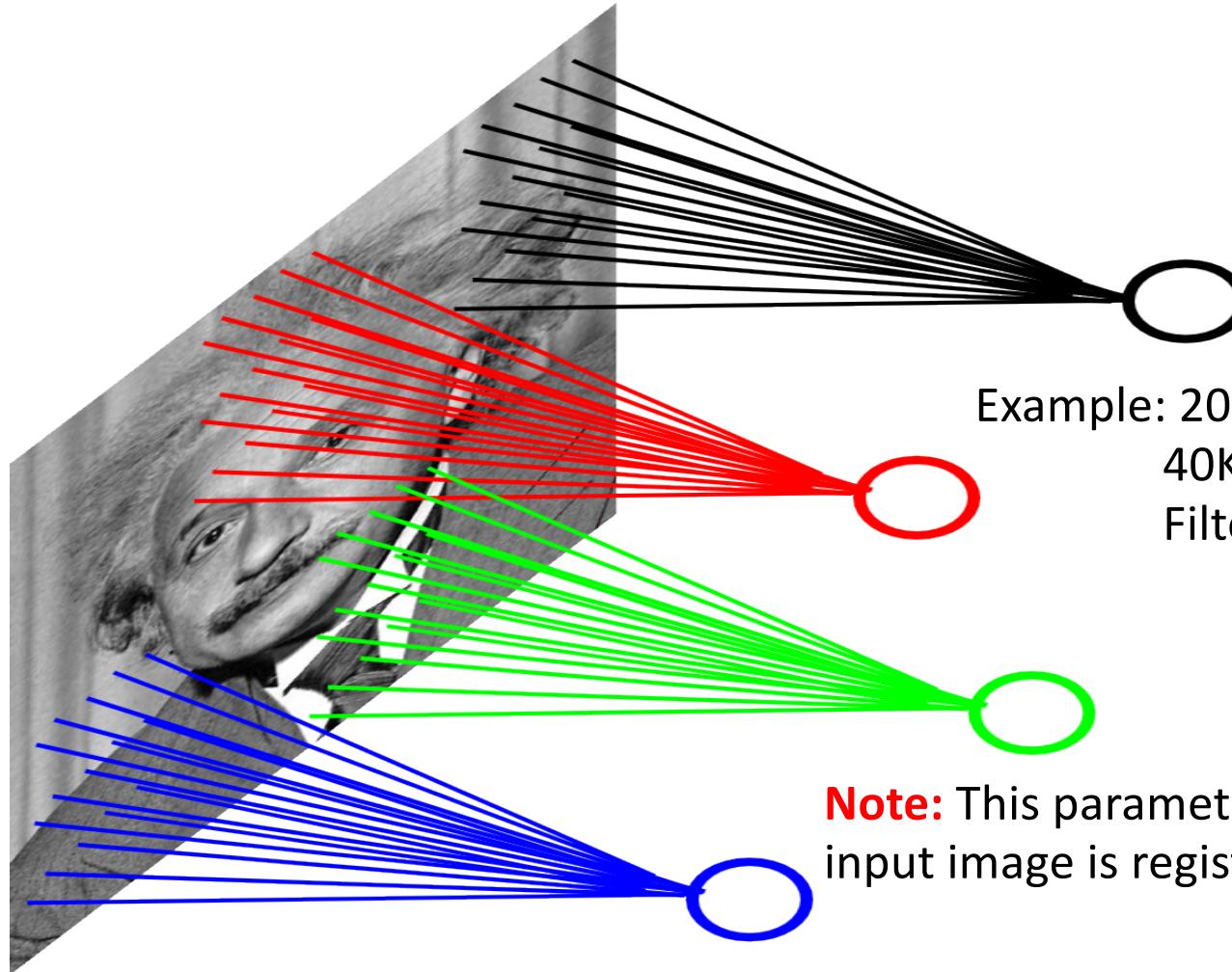
40K hidden units

→ ~2B parameters!!!



- Spatial correlation is local
- Waste of resources + we have not enough training samples anyway..

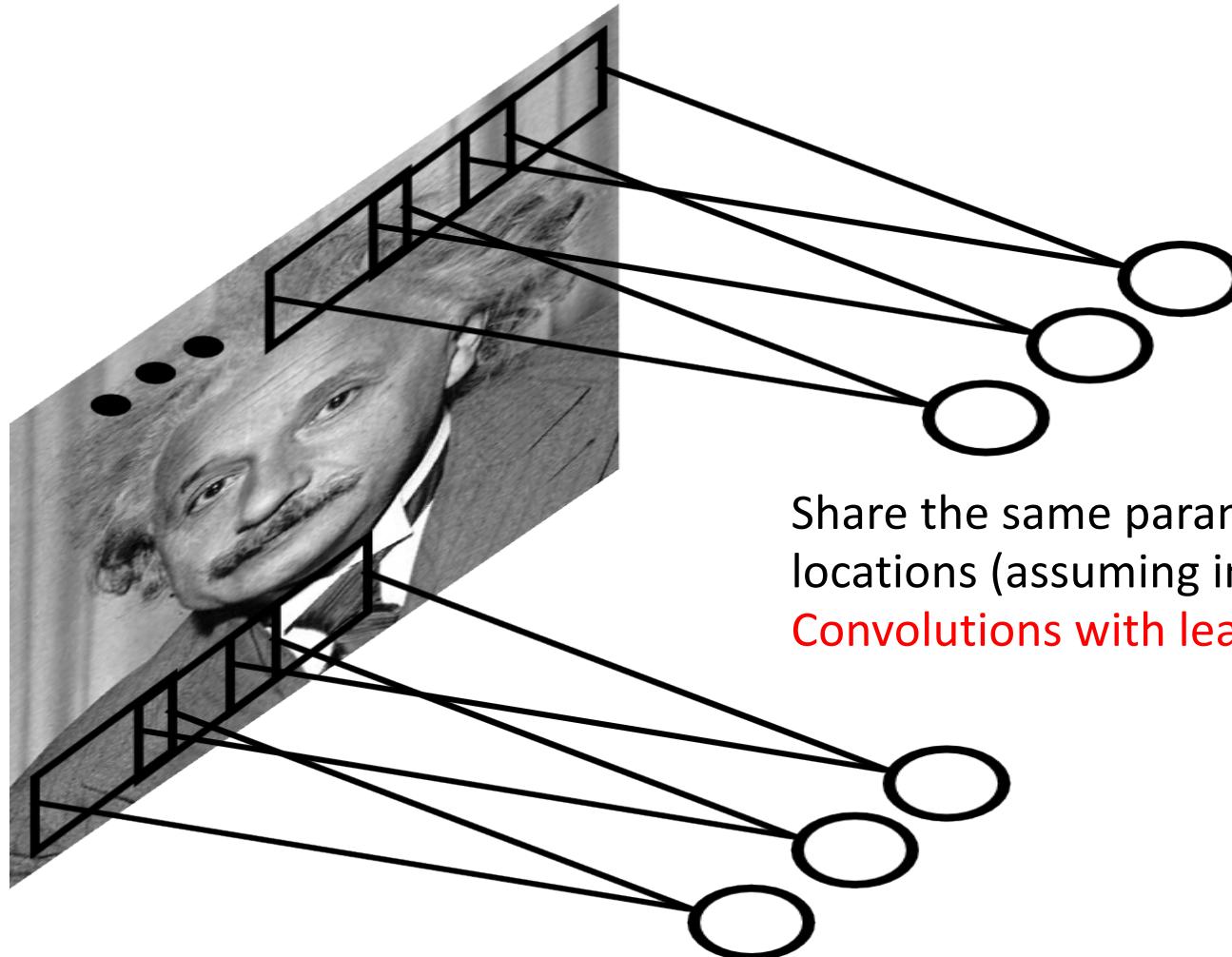
Locally Connected Layer



Example: 200x200 image
40K hidden units
Filter size: 10x10
4M parameters

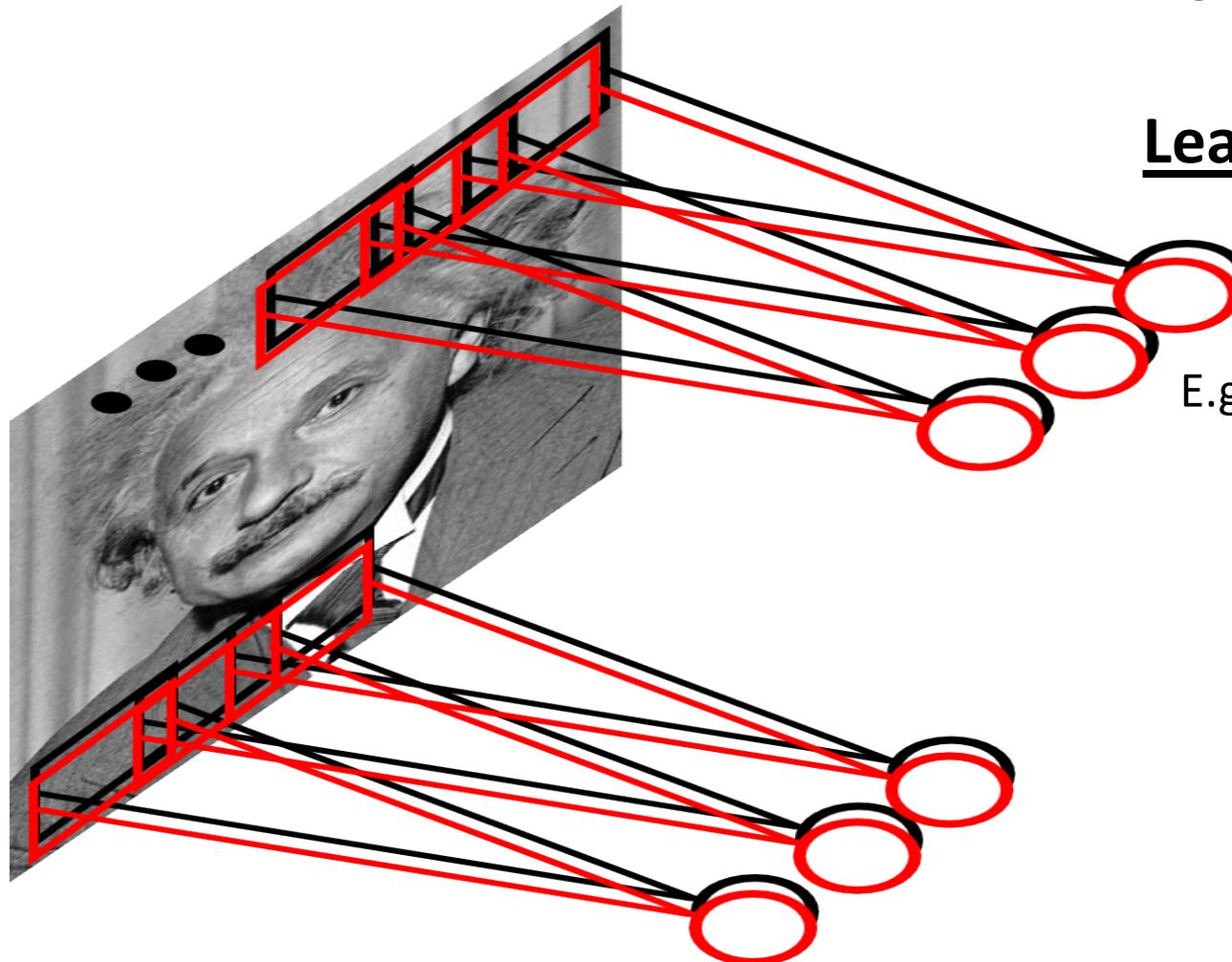
Note: This parameterization is good when input image is registered (e.g., face recognition).

Convolutional Layer



Share the same parameters across different locations (assuming input is stationary):
Convolutions with learned kernels

Convolutional Layer



Learn multiple filters.

E.g.: 200x200 image
100 Filters
Filter size: 10x10
10K parameters



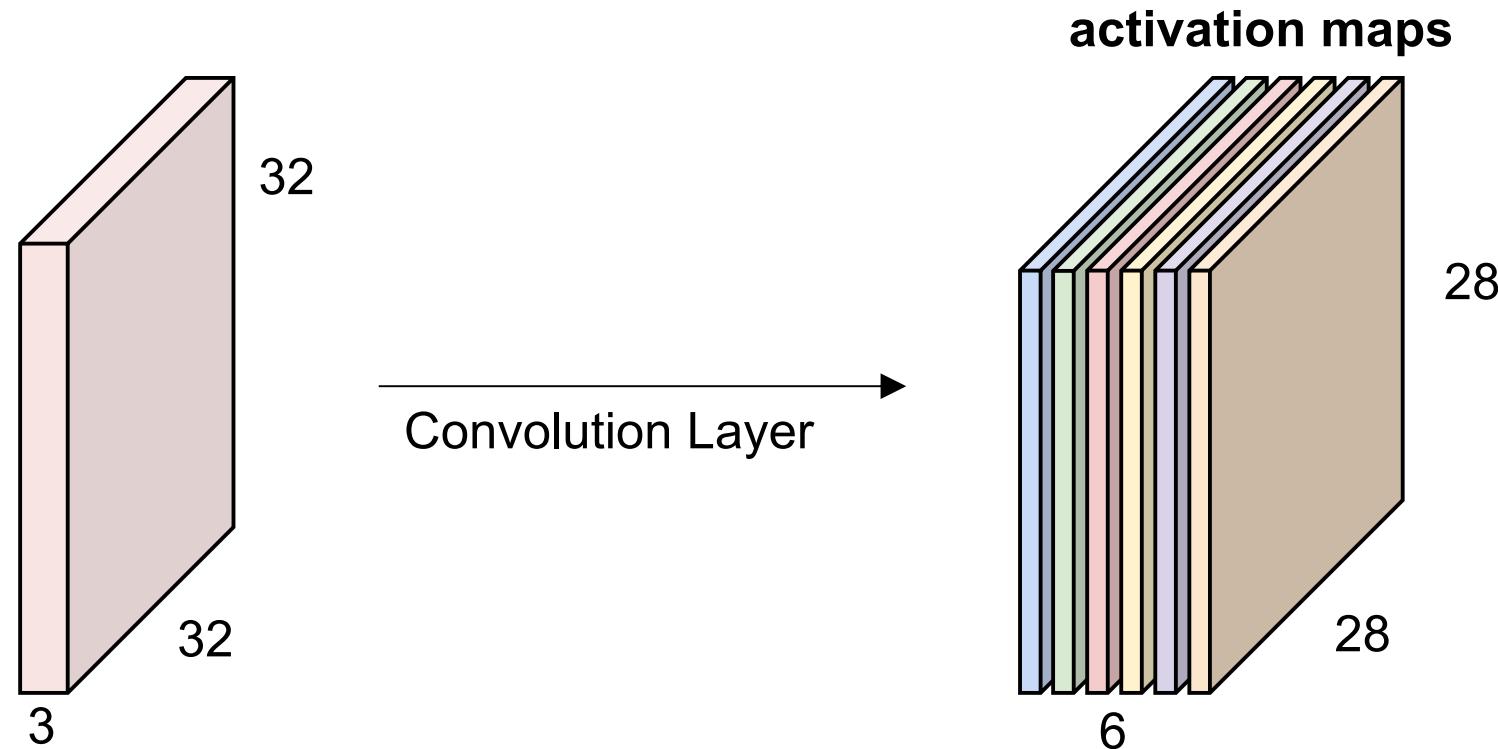
Ranzato 

Convolution Layer

consider a second, green filter

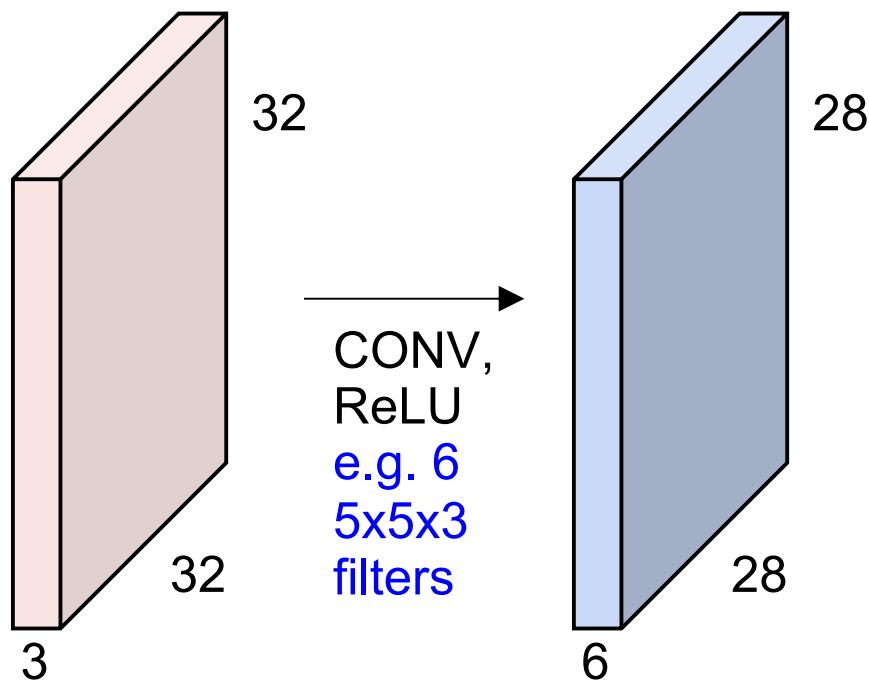


For example, if we had 6 5×5 filters, we'll get 6 separate activation maps:

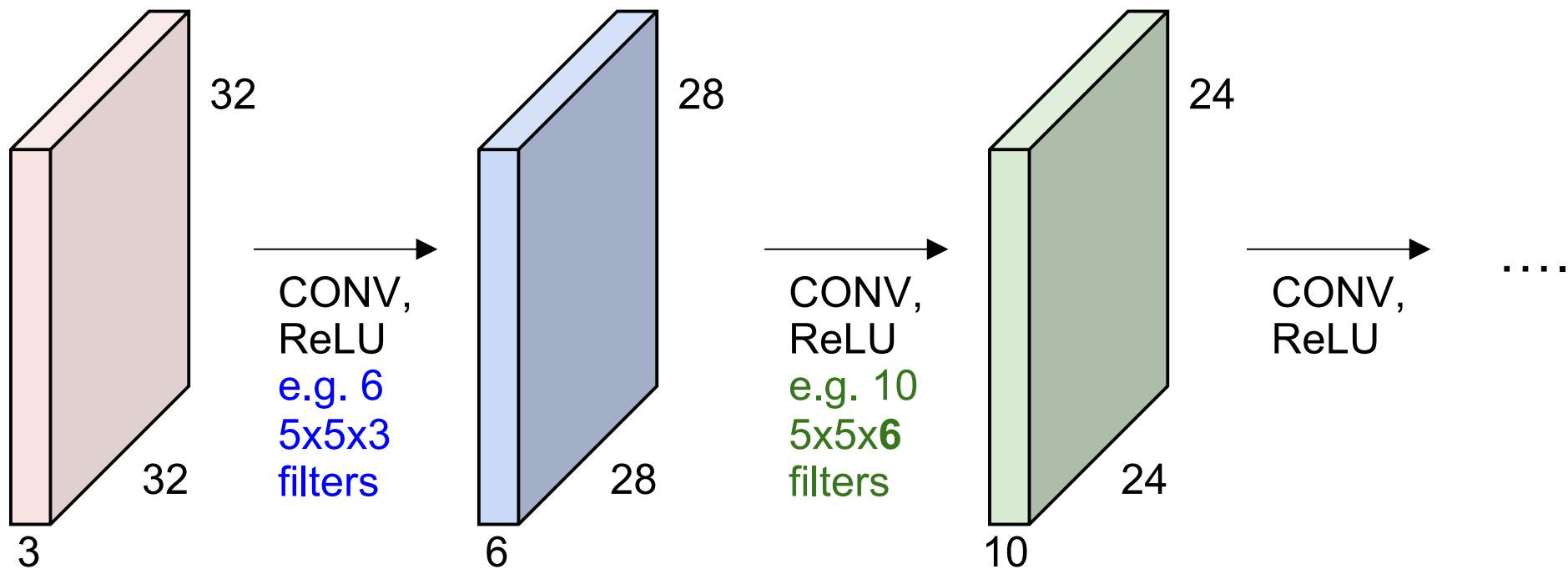


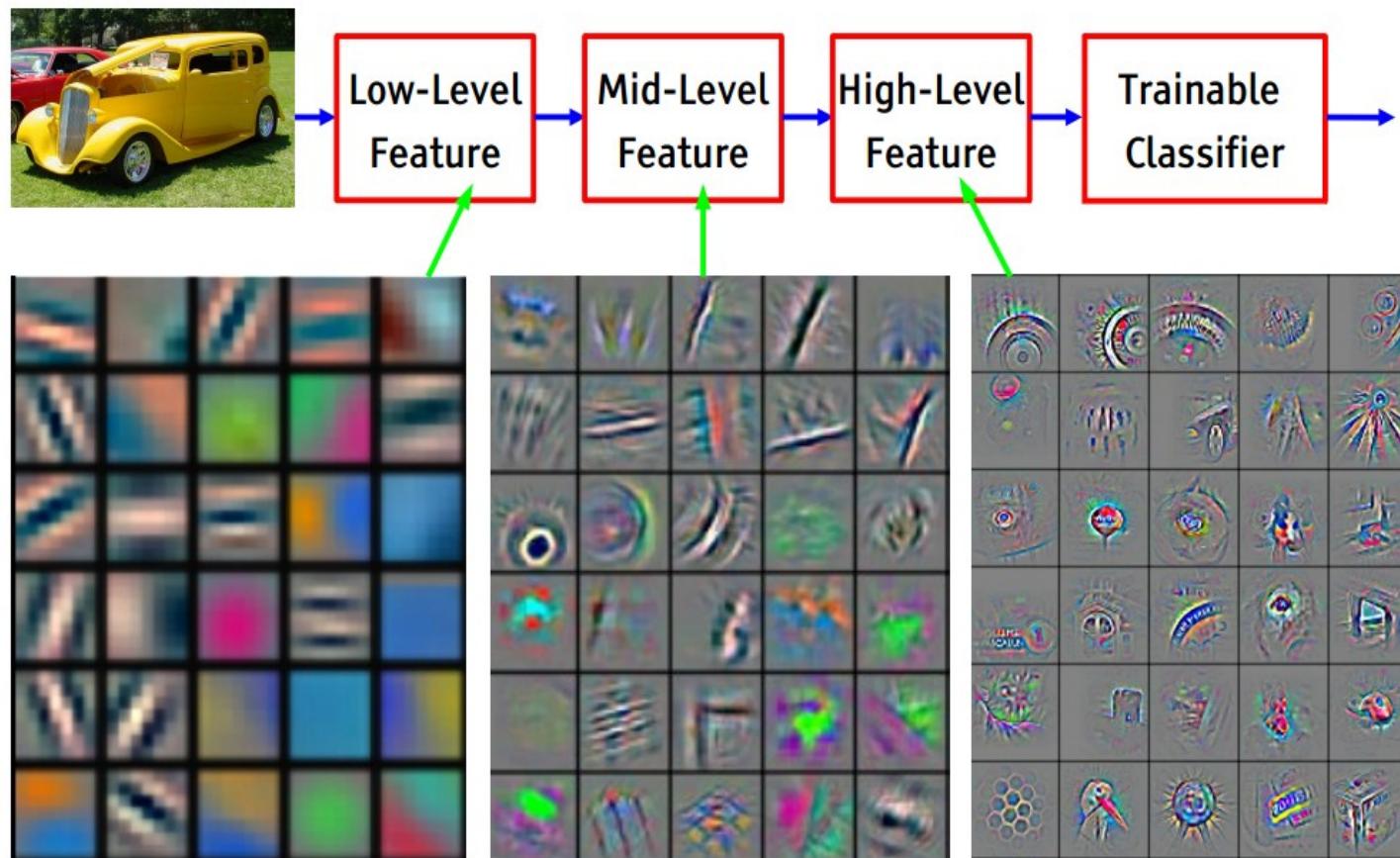
We stack these up to get a “new image” of size $28 \times 28 \times 6$!

ConvNet is a sequence of Convolution Layers, interspersed with activation functions

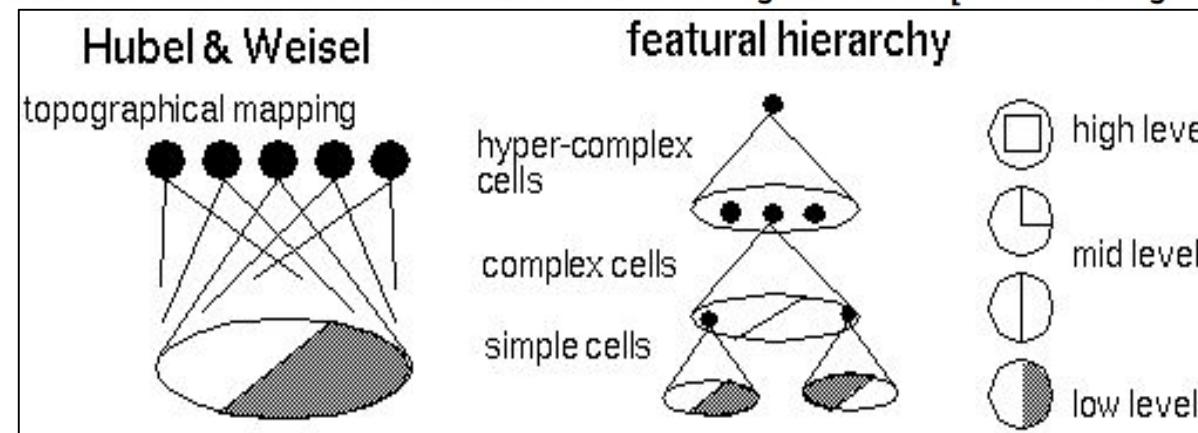


ConvNet is a sequence of Convolution Layers, interspersed with activation functions



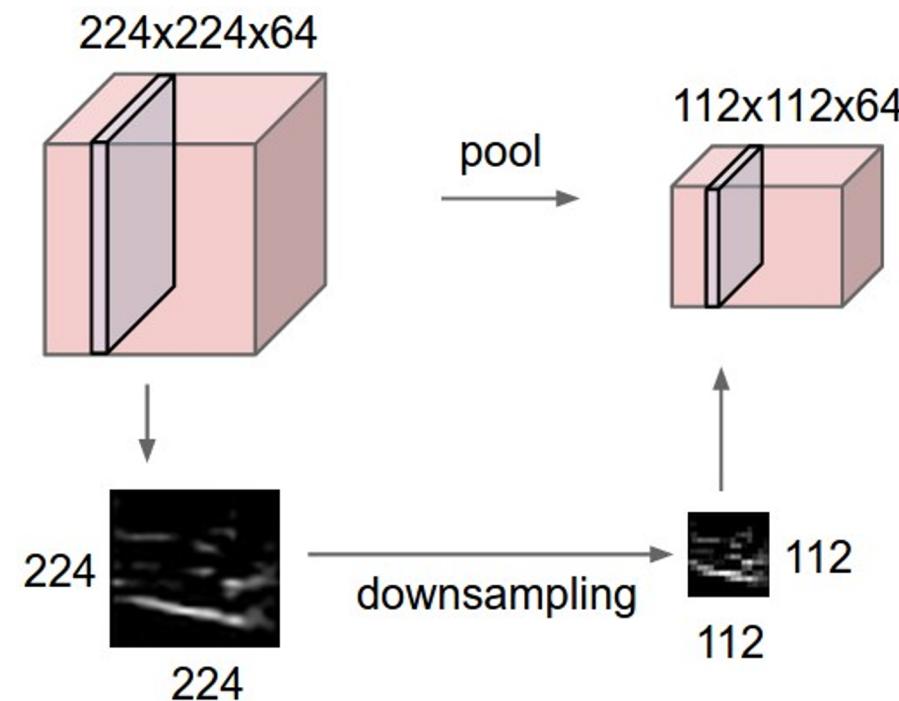


Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

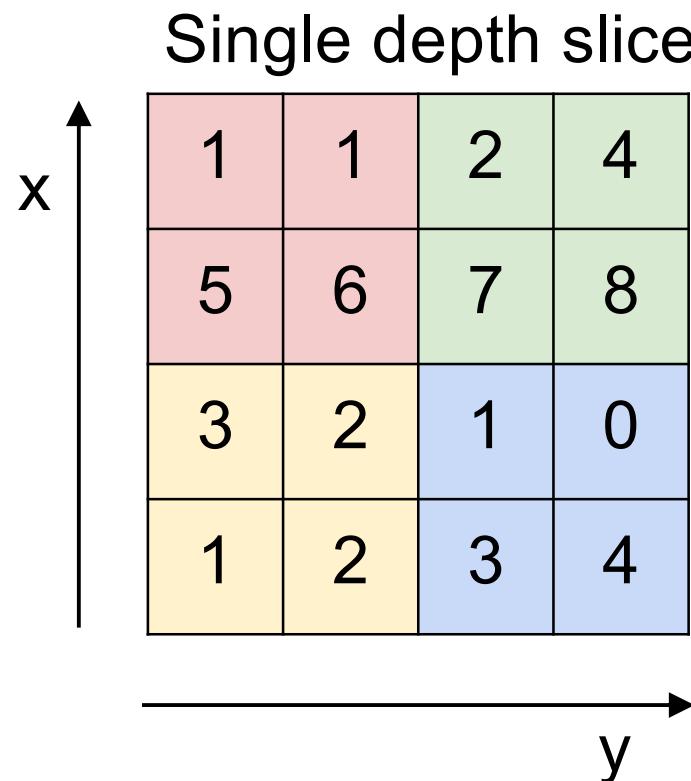


Pooling layer

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



MAX POOLING

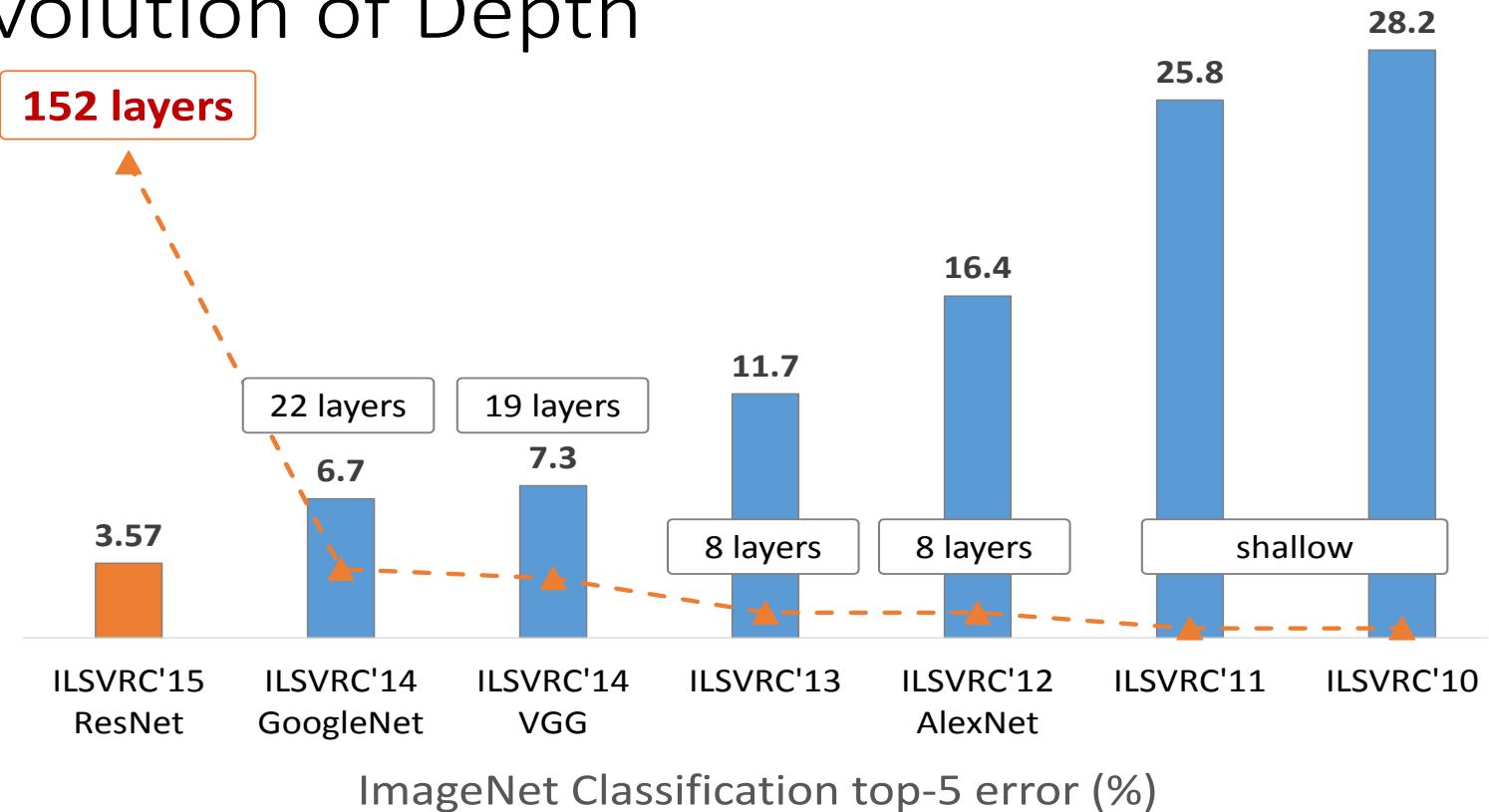


max pool with 2x2 filters
and stride 2

6	8
3	4

Fast-forward to today: Image Classification on ImageNet

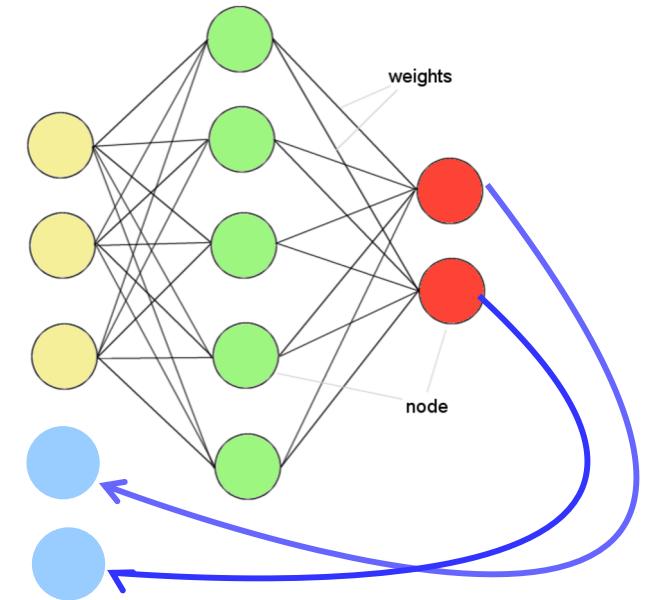
Revolution of Depth



Kaiming He, et al. Deep residual learning for Image Recognition, 2015

There are many more architectures, e.g., Recurrent Neural Networks

- Recurrent Neural Networks (RNN)
 - allow to process sequential data
 - by feeding back the output of the network into the next input
- Long-Short Term Memory (LSTM)
 - add „forgetting“ to RNNs
 - good for mapping sequential input data into sequential output data
 - e.g., text to text, or time series to time series
- Deep Learning often allows „end-to-end learning“
 - e.g., learn a network that does the complete translation of text in one language into another language
 - previously, learning often concentrated on individual components (e.g. word sense disambiguation)



What have we learnt?



Neural networks are inspired by biological brains



Trained using backpropagation (gradient)



Deep Neural Networks aim at learning feature hierachies



We have understood the structure of convolutional neural networks, one of the central DNN architectures



Convolutional layer, ReLU, Max pooling layer, fully connected layer



DNNs are rather large but result in state-of-the-art performance on many tasks