CSE 575: Statistical Machine Learning

Deep Learning

Jingrui He
CIDSE, ASU

Acknowledgement

- Part of Slides are from
  - Lei Li (Baidu USA)
  - Yu Cheng (IBM Watson)

Outline

- Why Deep Learning?
- What is Deep Learning?
- Where is Deep Learning Heading to?

Success of Deep Learning

The New York Times

Scientists See Promise in Deep-Learning Programs

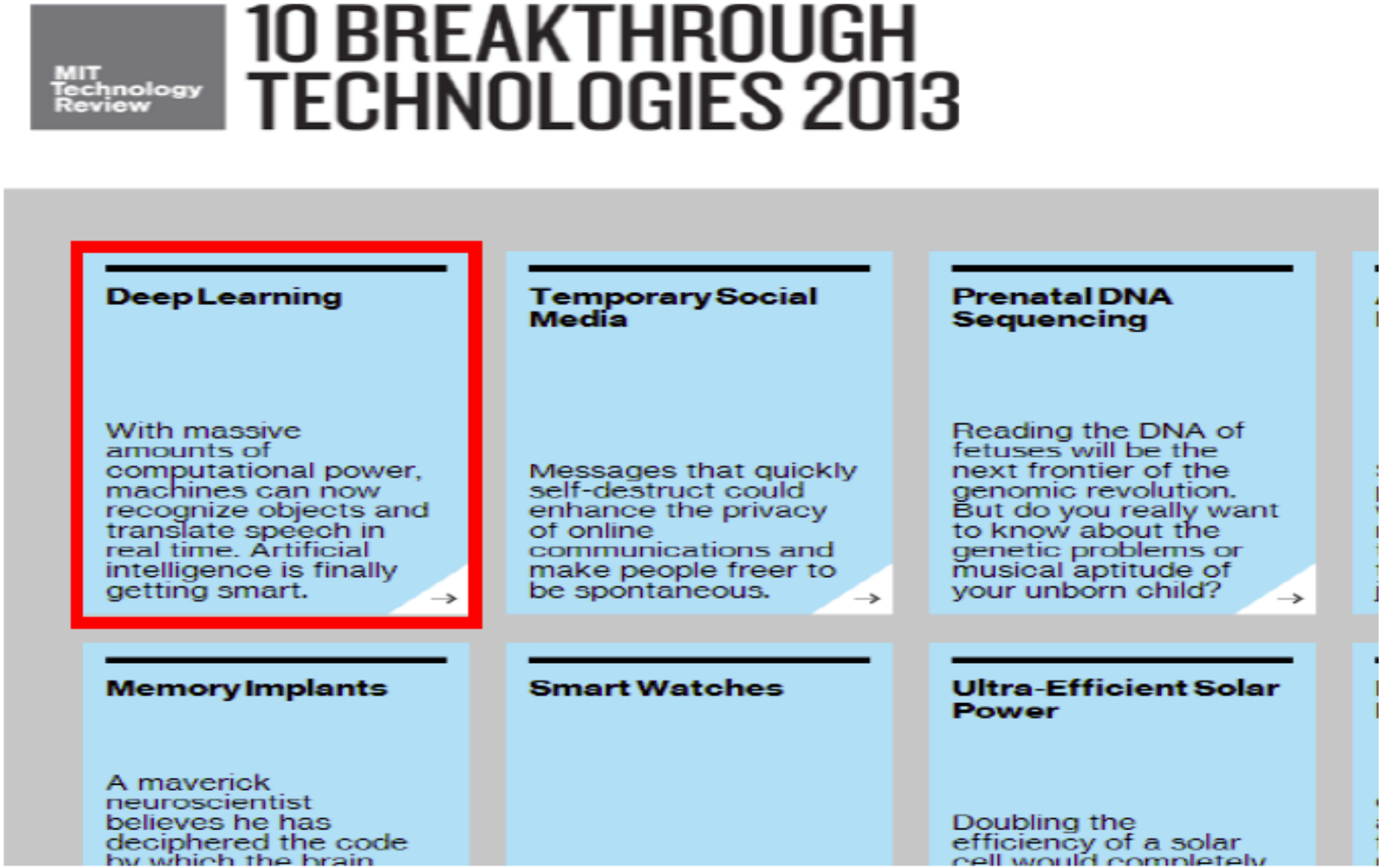
Hao Zhang/The New York Times

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese.

By JOHN MARKOFF
Published: November 23, 2012

Using an artificial intelligence technique inspired by theories about how the brain recognizes patterns, technology companies are reporting startling gains in fields as diverse as computer vision, speech recognition and the identification of promising new molecules for designing drugs.

Success of Deep Learning



Success of Deep Learning

MIT Technology Review

Facebook Launches Advanced AI Effort to Find Meaning in Your Posts

A technique called deep learning could help Facebook understand its users and their data better.

By Tom Simonite on September 20, 2013

Facebook is set to get an even better understanding of the 700 million people who use the social network to share details of their personal lives each day.

A new research group within the company is working on an emerging and powerful approach to artificial intelligence known as deep learning, which uses simulated networks of brain cells to process data. Applying this method to data shared on Facebook could allow for novel features and perhaps boost the company's ad targeting.

Deep learning has shown potential as the basis for software that could work out the emotions or events described in text even if they aren't explicitly referenced, recognize objects in photos, and make sophisticated predictions about people's likely future behavior.

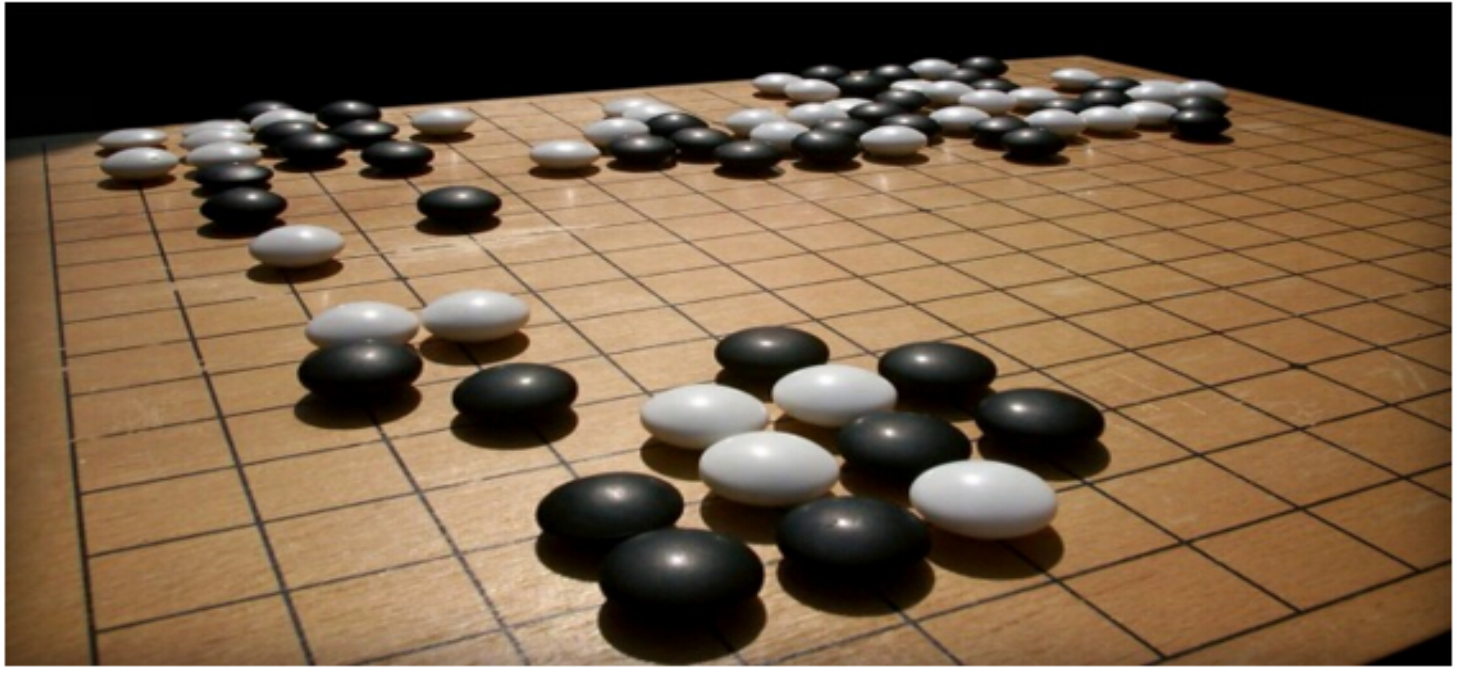
WHY IT MATTERS

Facebook's piles of data on people's lives could allow it to push the boundaries of what can be done with the emerging AI technique

Success of Deep Learning

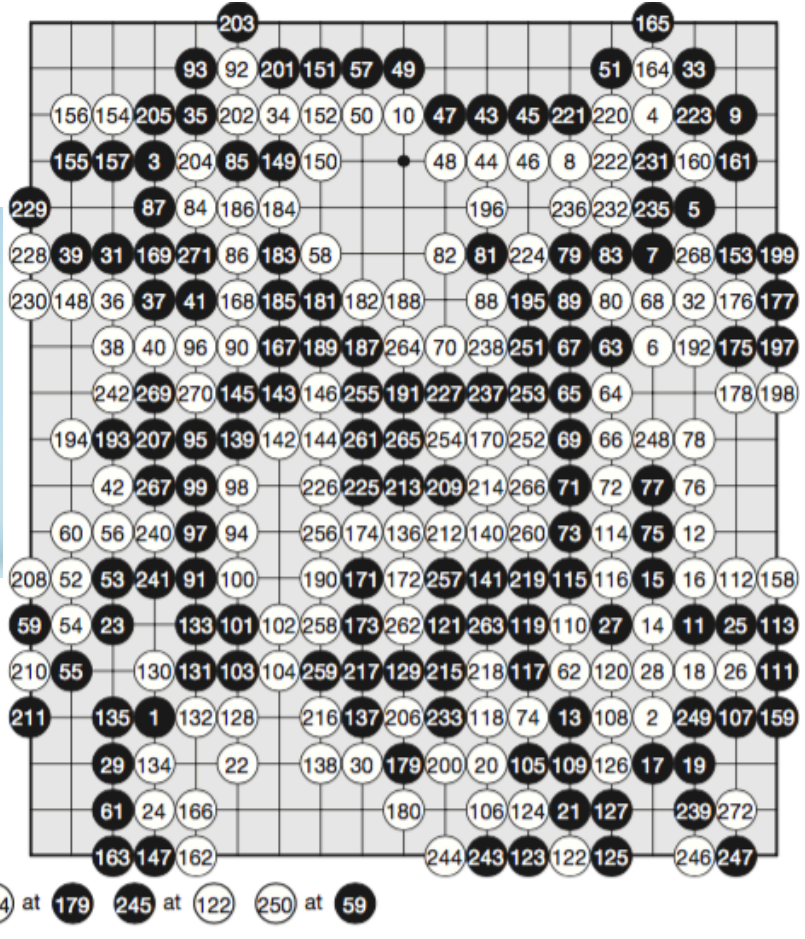
Google's DeepMind AI beats humans at the massively complex game Go

By Ryan Whitwam on January 27, 2016 at 4:00 pm | 11 Comments



Success of Deep Learning

4



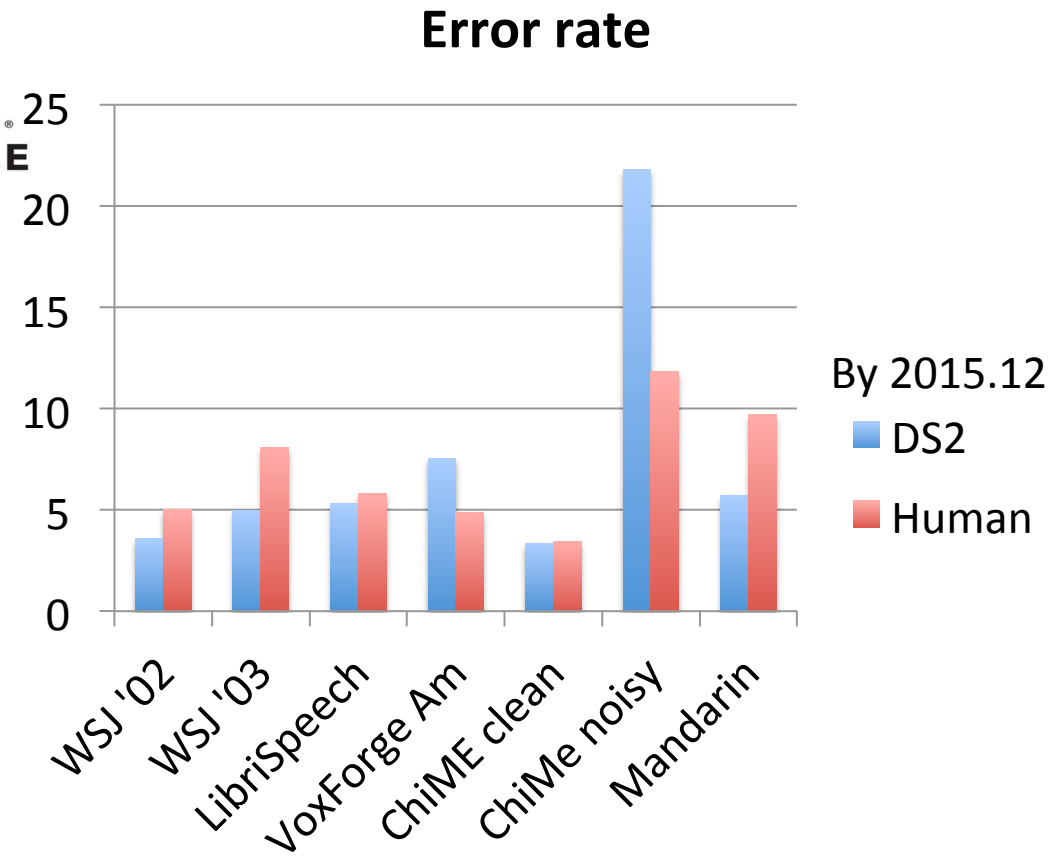
1

Towards human level in speech recognition

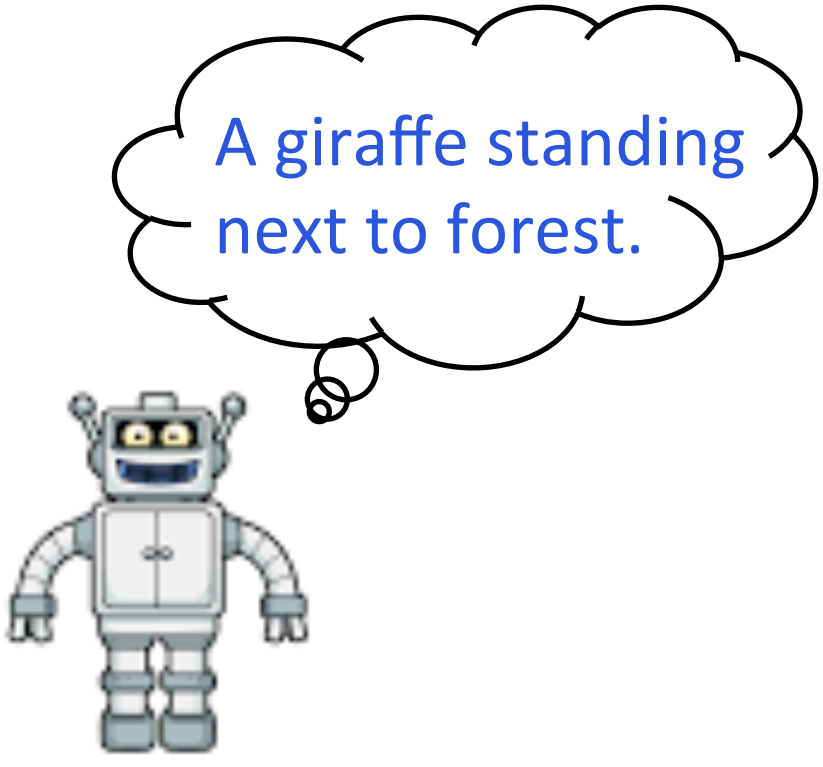
NUANCE

iFLY

Baidu 百度



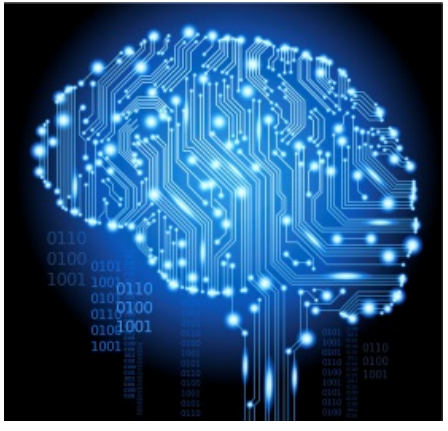
Telling stories in images



In Dec. 2015, Baidu's driverless car completed 18.6-mile route around Beijing that included side streets as well as highways

New power source for AI

Big(ger) data



Better model
& (learning) alg.

High performance
computing

Outline

- Why Deep Learning?
- What is Deep Learning?
- Where is Deep Learning Heading to?

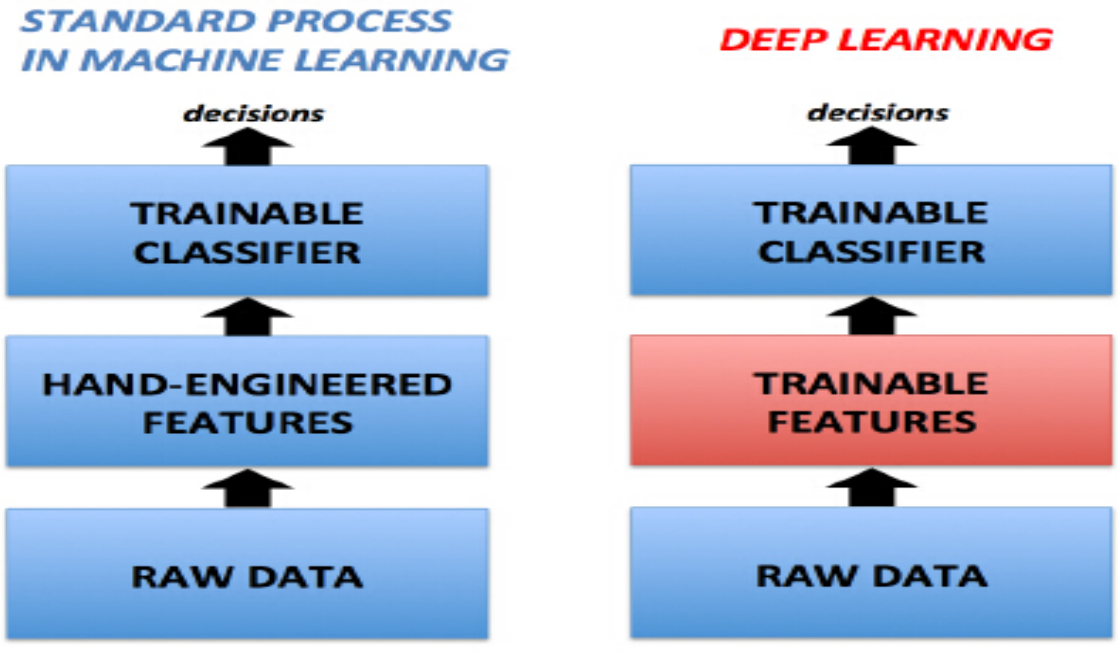
Classification: handwriting recognition

0
1
2
3
4
5
6
7
8
9

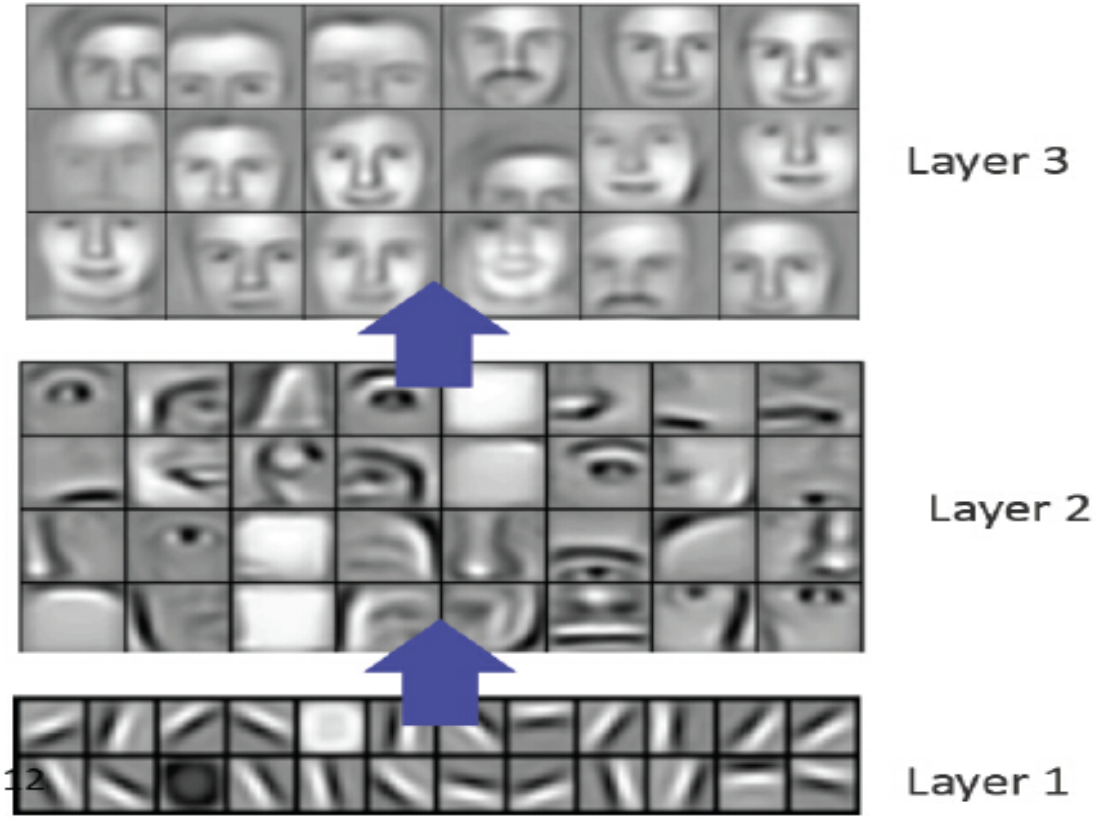
What is Deep Learning?

Definition

A family of methods that uses deep architectures to learn high-level feature representations.



Example of trainable features

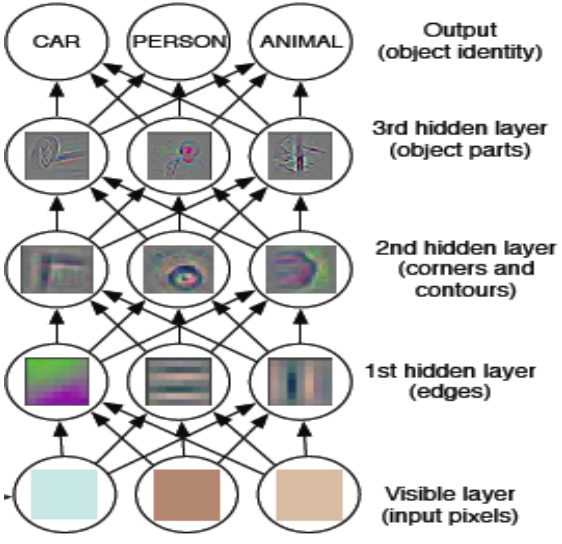


Why Deep Learning Works?

Why Deep Model?

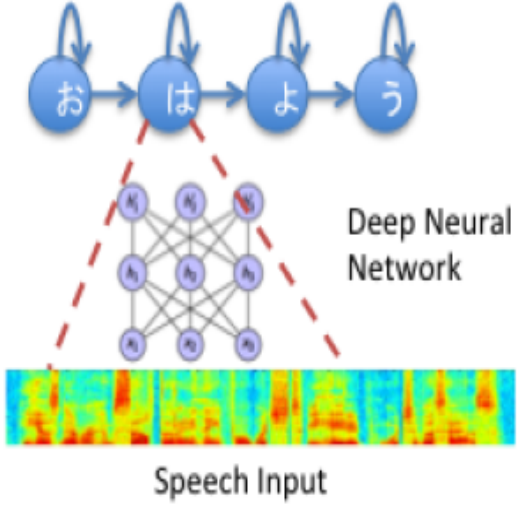
1) Fill in the gap between low-level feature and semantic meaning
2) Learn useful high-level abstraction with less variant/noise

object



pixels

sentence

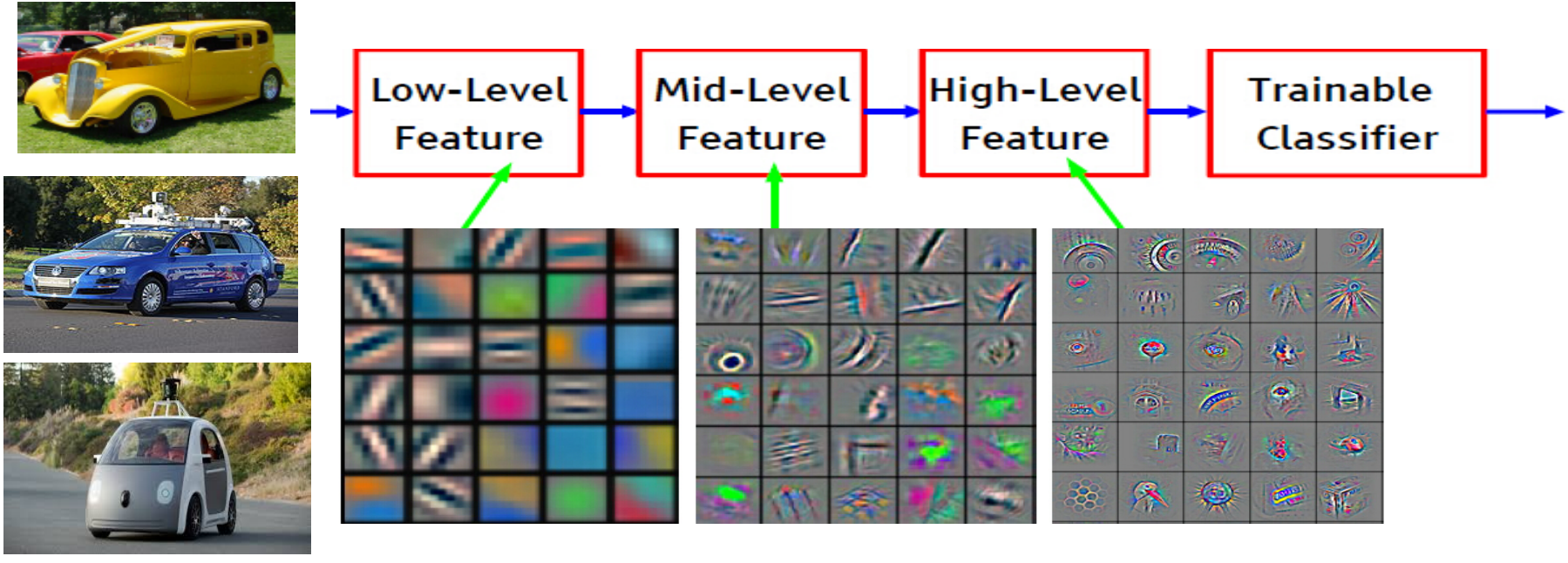


signals

Why Deep Learning Works?

Why Deep Model?

1) Fill in the gap between low-level feature and semantic meaning
2) Learn useful high-level abstractions with less variant/noise



Inspired by a biological neuron

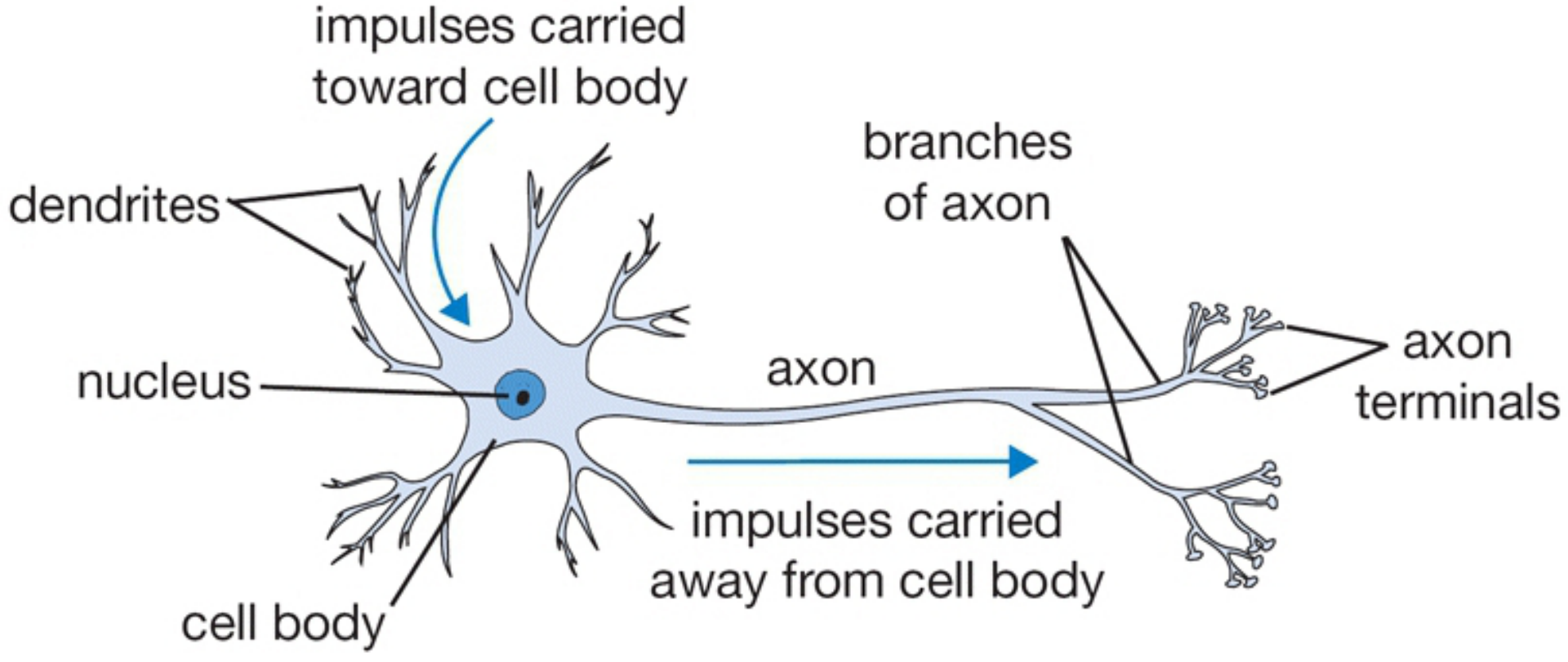
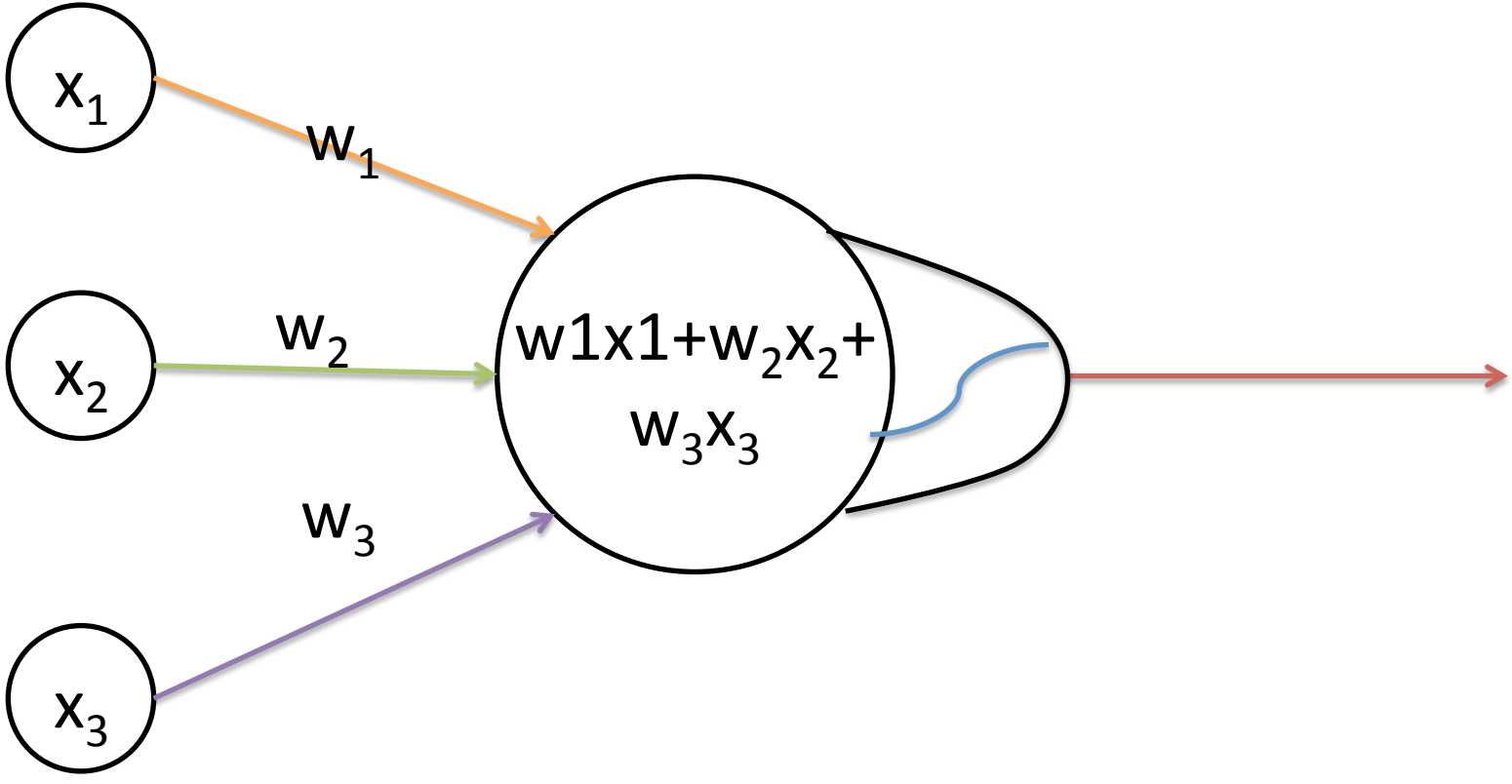


Image credit:
http://cs231n.github.io/neural-networks-1/

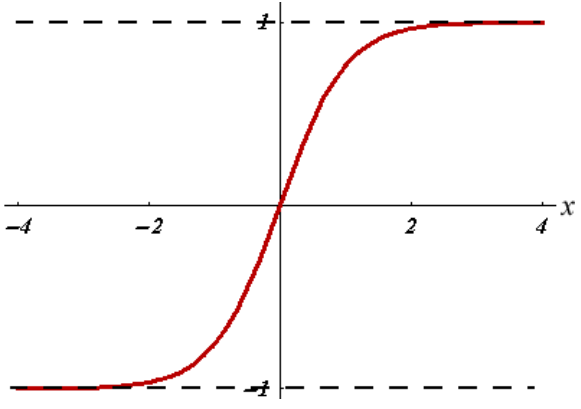
How to model a single neuron?

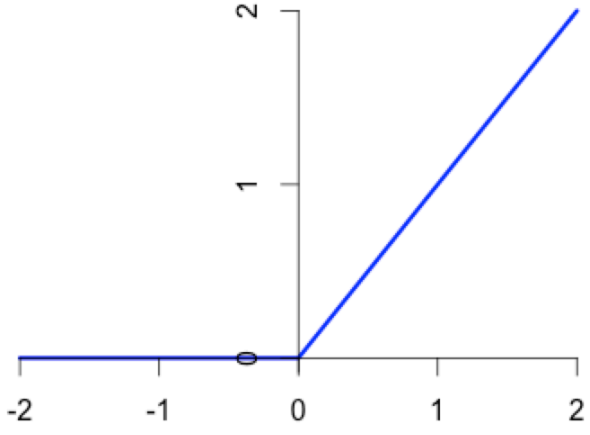


Activation function (σ)

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$$

$$\text{relu}(x) = \max(0, x)$$

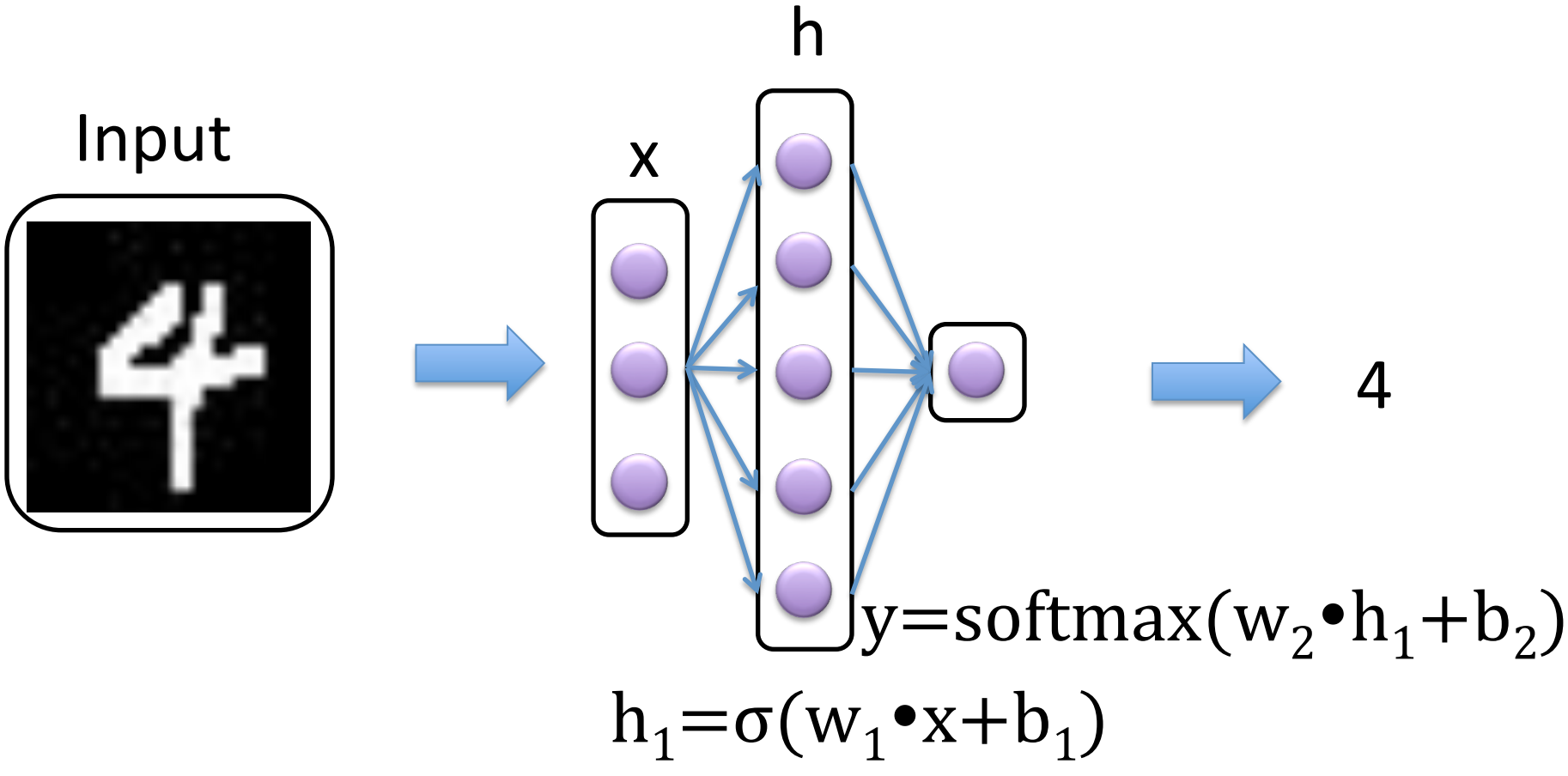




$$\text{softmax}(x)_i = \frac{e^{x_i}}{\sum e^{x_i}}$$

Useful for modeling probability (in classification task)

Supervised Learning with Neural Nets



Numerical Example

Input

x

0
0
1
0
1
0
…

784 x 1

h

0.3
0.2
0.8
-.2
-.7
.5
.3
.1
-.2
…

.1
0
.1
.8
0
…

Loss(w) =Σy'logf(x)

Other loss:
sq loss, hinge-loss

4

$y=\text{softmax}(w_2 \bullet h_1 + b_2)$

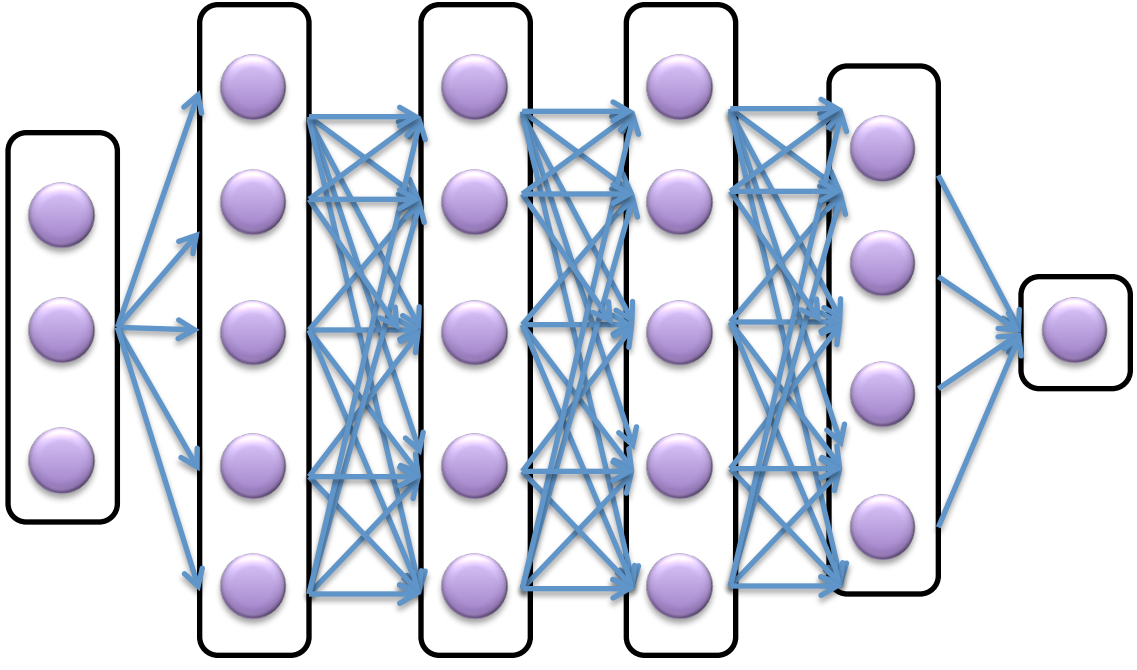
w_2 : 256 x 1

$h_1=\sigma(w_1 \bullet x + b_1)$

w_1 : 256 x 784

Deep Neural Nets

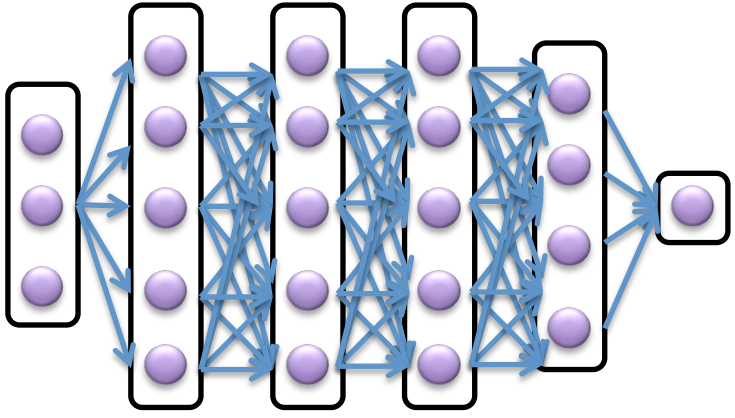
Input



$$h_1 = \sigma_1(w_1 \bullet x + b_1)$$

$$h_2 = \sigma_2(w_2 \bullet h_1 + b_2)$$

Supervised Learning



Given: N data points
$(x_1, y_1) \ldots (x_N, y_N)$

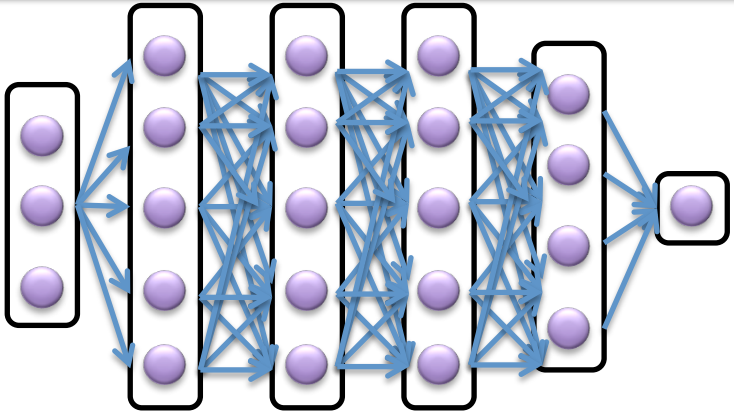
Goal: find the best model parameter w, to minimize cost

$$L(w) = \sum_{i=1}^{N} l(f(x_i, w), y_i)$$

Q1: How to Learn the Model Parameters?

Q2: How to Design a Deep Neural Nets (DNN)?

Training deep neural nets



Stochastic gradient descent algorithm

for iteration 1 to N (or until convergence)

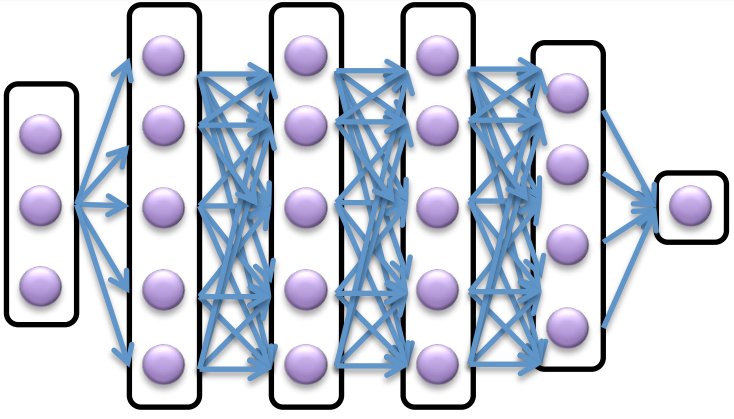
compute g = $\partial / \partial w_j$

w = w – a • g

To improve efficiency: Mini-Batch

Advanced alg: Momentum, Adagrad, Adam, …

Training deep neural nets



Chain rule

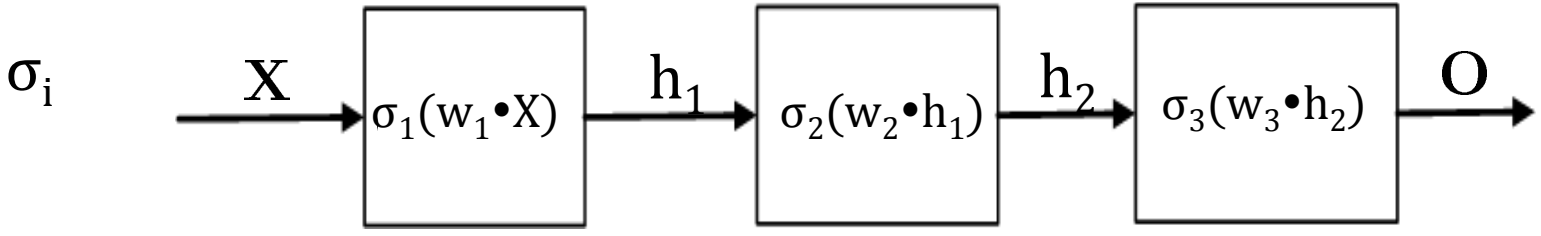
forward pass: computing network prediction

$$h_i = \sigma_i(w_i \bullet h_{i-1})$$

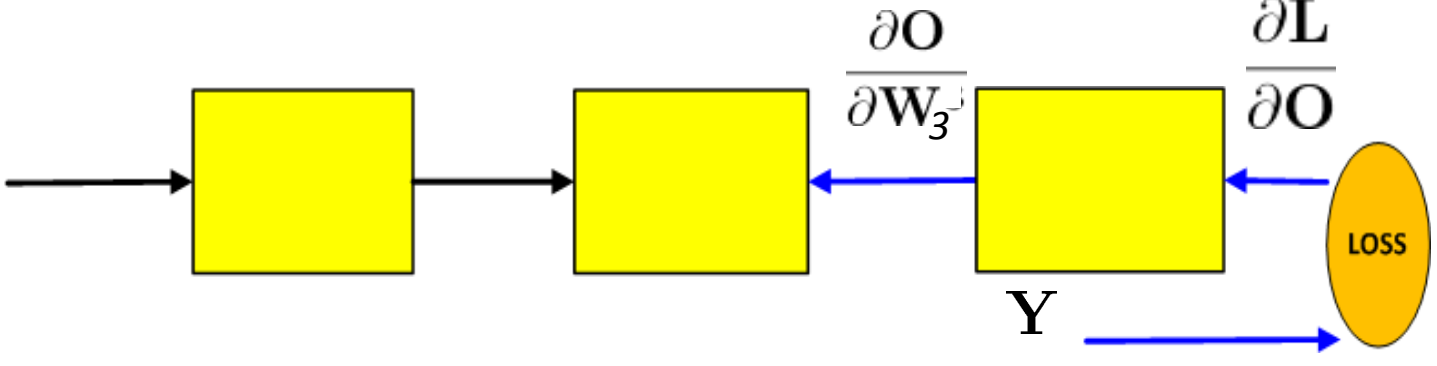
backward prop: computing gradient from layer-wise error

$$\delta_{i-1} = w_i^T \bullet (\delta_i \odot \sigma_{i-1}') \qquad \partial / \partial w_{i-1} = h_{i-1} \bullet \delta_i^T$$

Forward-backward prop: Illustration

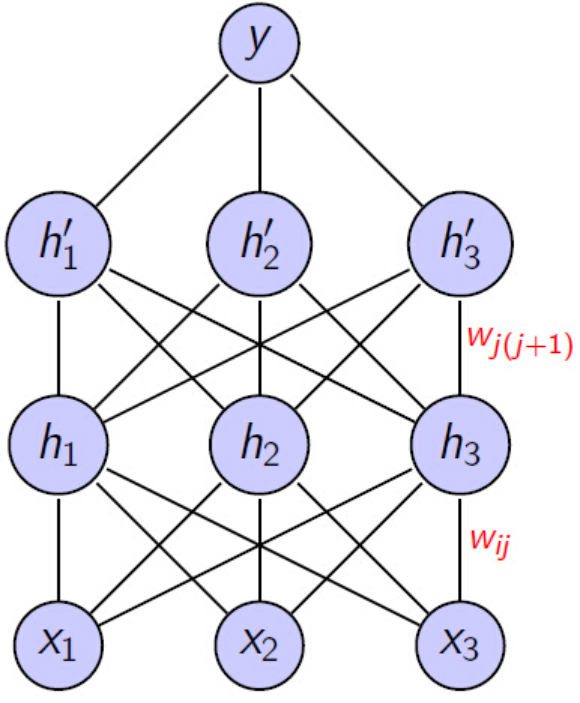


Forward propagation: $h_i = \sigma_i(w_i \bullet h_{i-1})$



Backward propagation: $\frac{\partial L}{\partial W_3} = \frac{\partial L}{\partial O}\frac{\partial O}{\partial W_3}$ **(Chain Rule)**

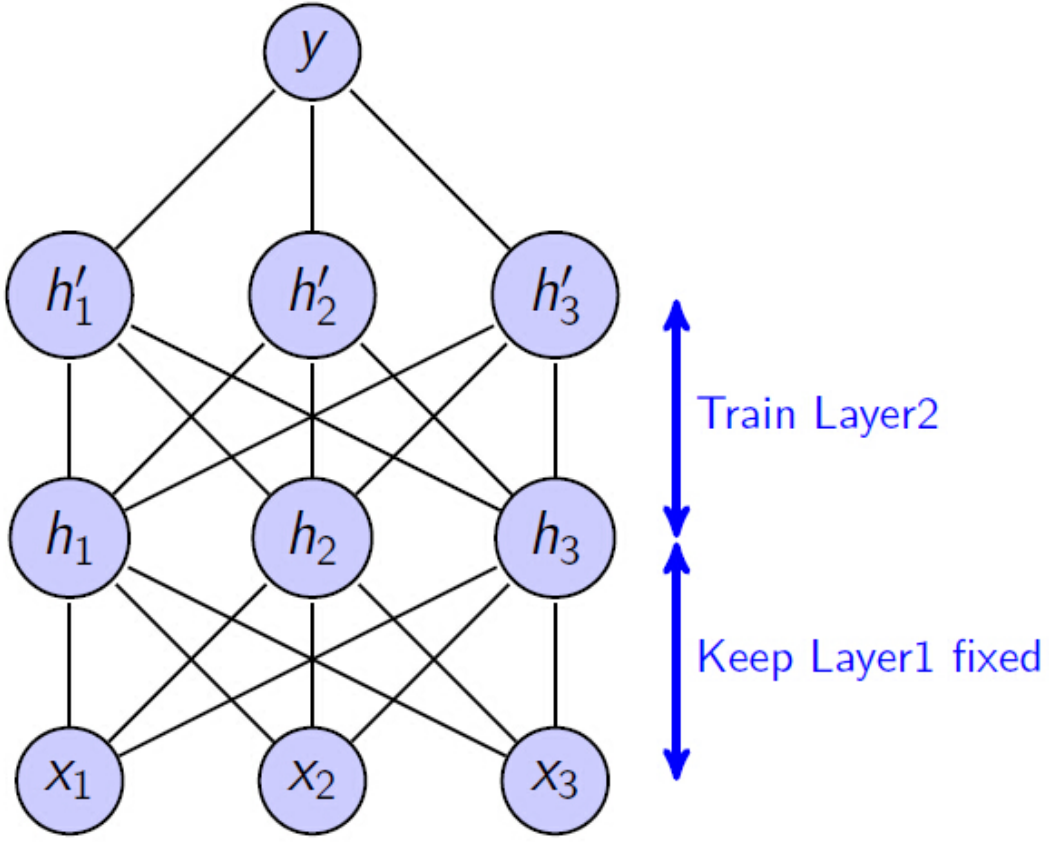
Why deep architectures hard to train?



- Vanishing gradient problem (BP)
- Insufficient training data
- Weak computational source

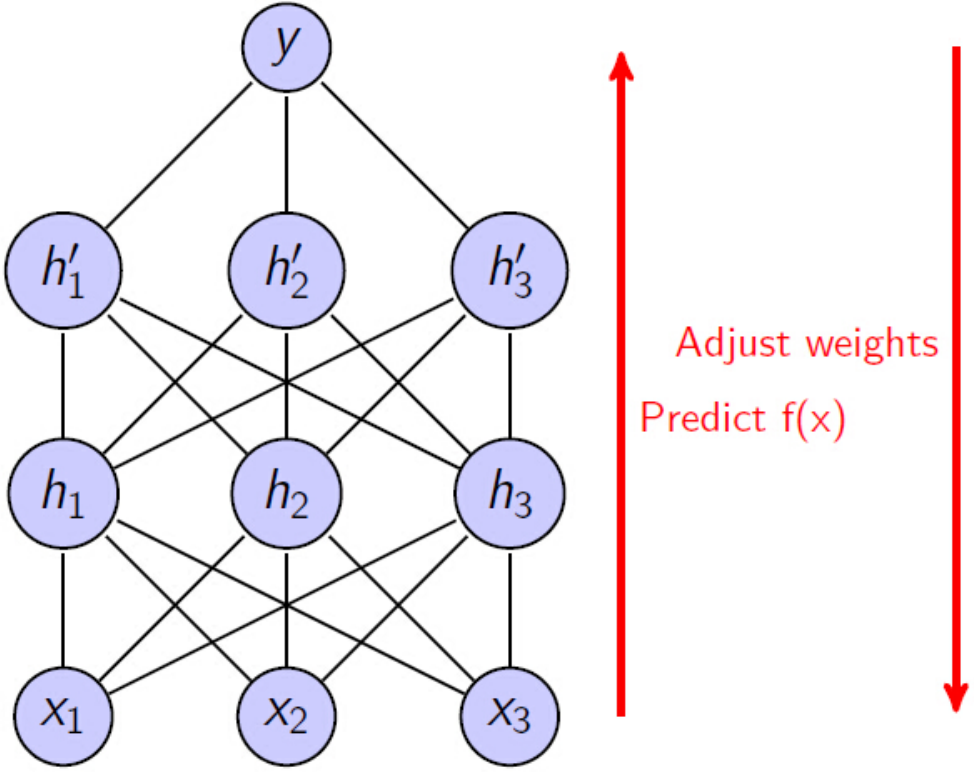
Layer-wise pre-training [Hinton et al. 2006]

First, train one layer at a time, optimizing data-likelihood objective $P(x)$



Layer-wise pre-training [Hinton et al. 2006]

Finally, fine-tune labeled objective $P(y|x)$ by Backpropagation

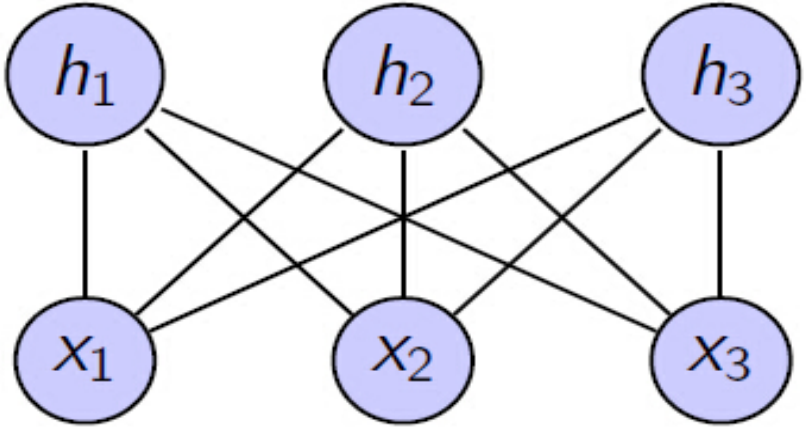


Typical Structures of DNN

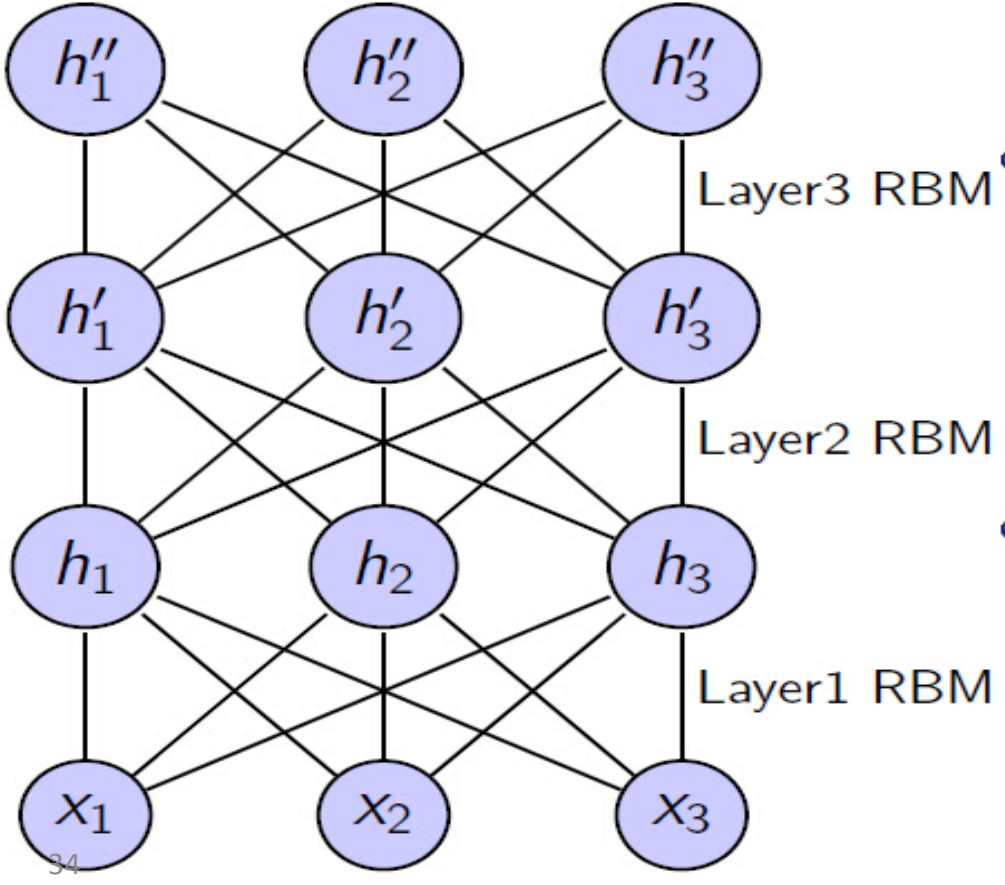
- DBN: Deep Belief Nets
- CNN: Convolutional Neural Nets
- RNN: Recurrent Neural Nets

Restricted Boltzmann Machine (RBM)

- RBM is a simple energy-based model: $p(x, h) = \frac{1}{Z_\theta} \exp\left(-E_\theta(x, h)\right)$
  - with only h-x interactions: $E_\theta(x, h) = -x^T W h - b^T x - d^T h$
  - here, we assume h_j and x_i are binary variables
  - normalizer: $Z_\theta = \sum_{(x,h)} \exp(-E_\theta(x, h))$ is called partition function

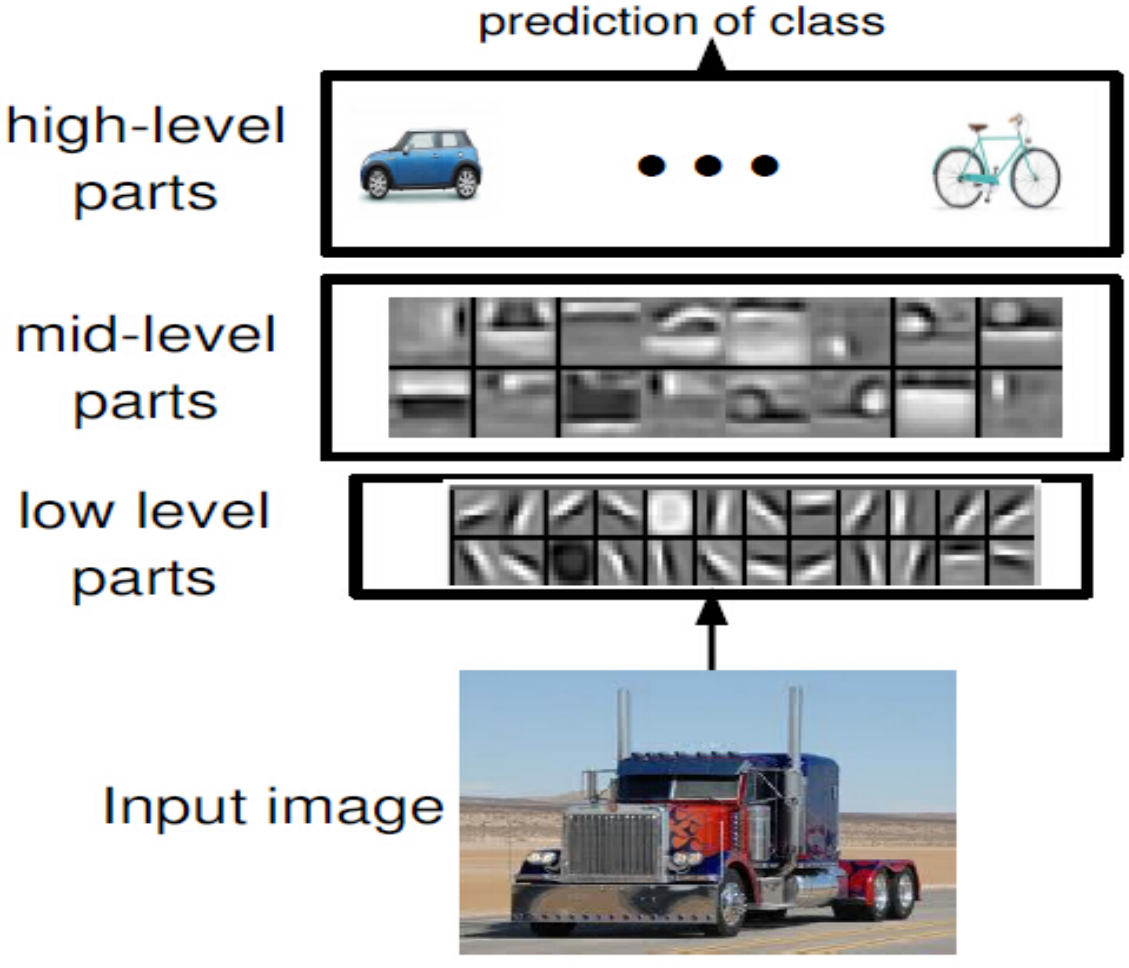


Deep Belief Nets (DBN) = Stacked RBM

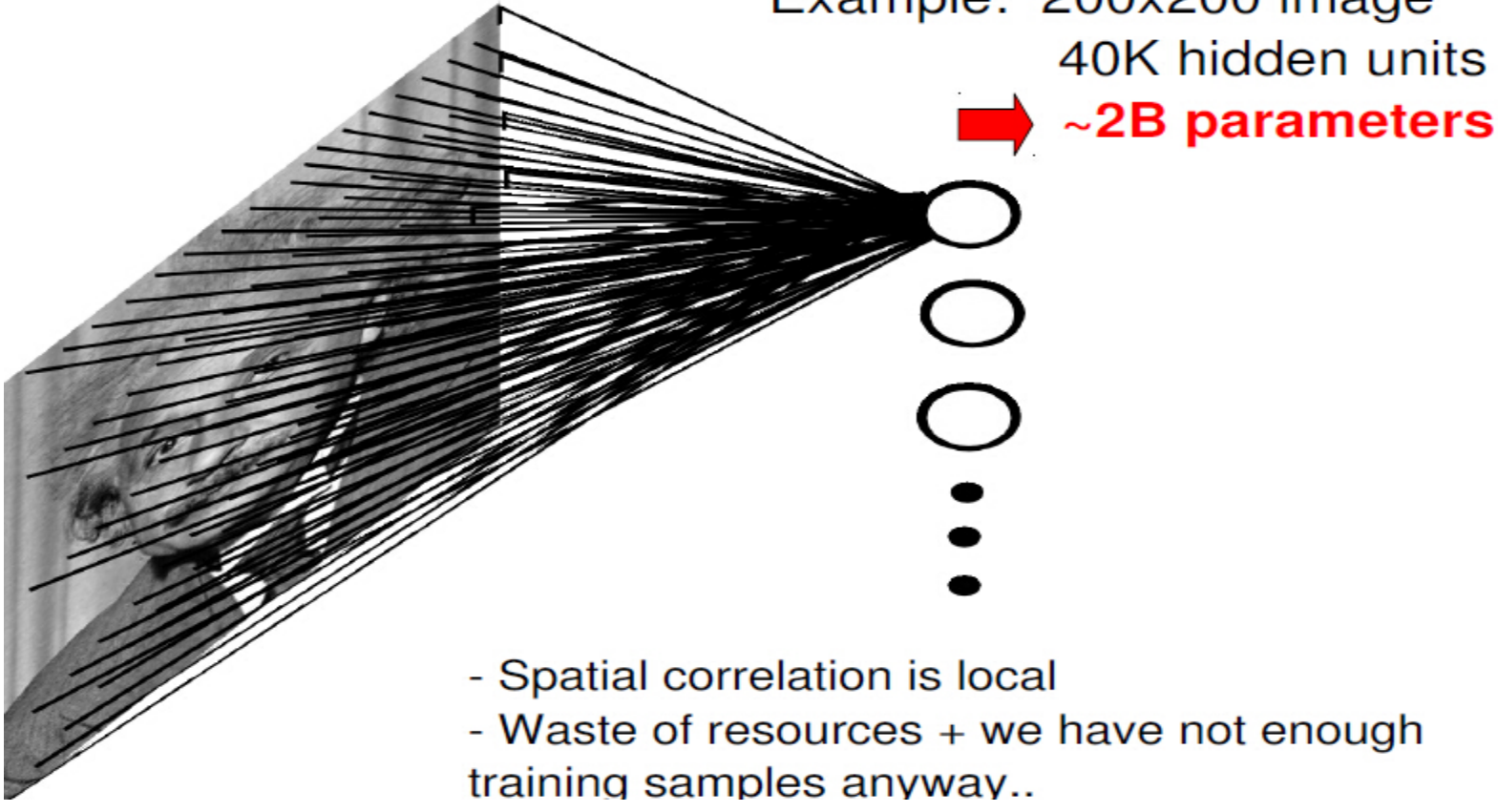


- DBN defines a probabilistic generative model $p(x) = \sum_{h,h',h''} p(x|h)p(h|h')p(h', h'')$ (top 2 layers is interpreted as a RBM; lower layers are directed sigmoids)
- Stacked RBMs can also be used to initialize a Deep Neural Network (DNN)

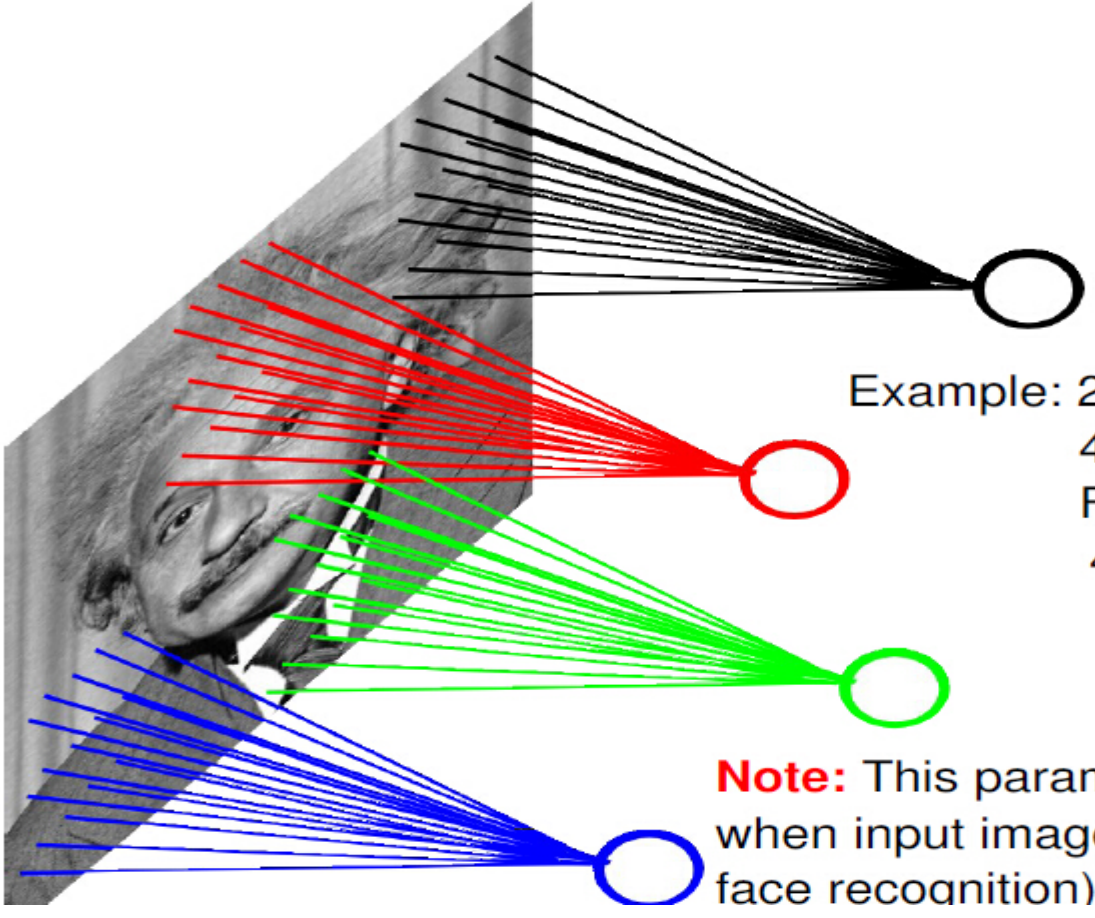
Convolutional Neural Networks



#1: Fully Connected Layer



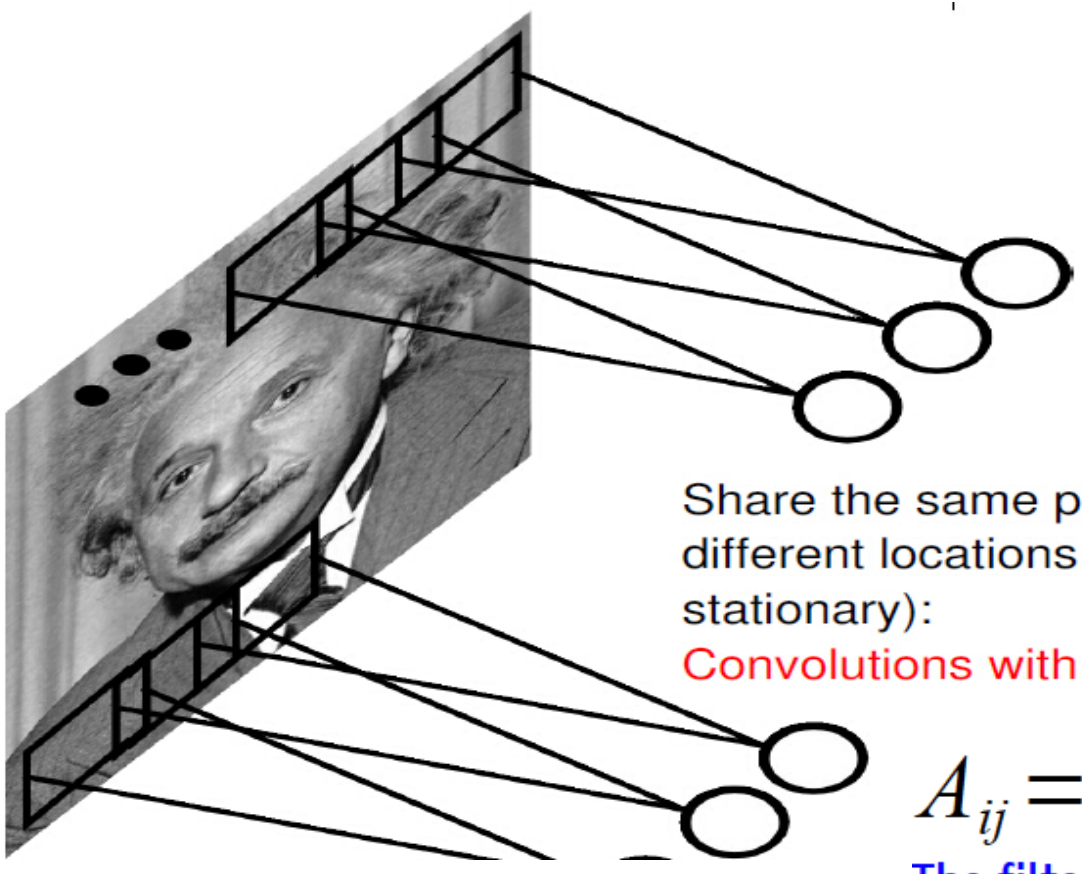
#2: Locally Connected Layer



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).

#3: Convolutional Layer (the key!)



Share the same parameters across different locations (assuming input is stationary):

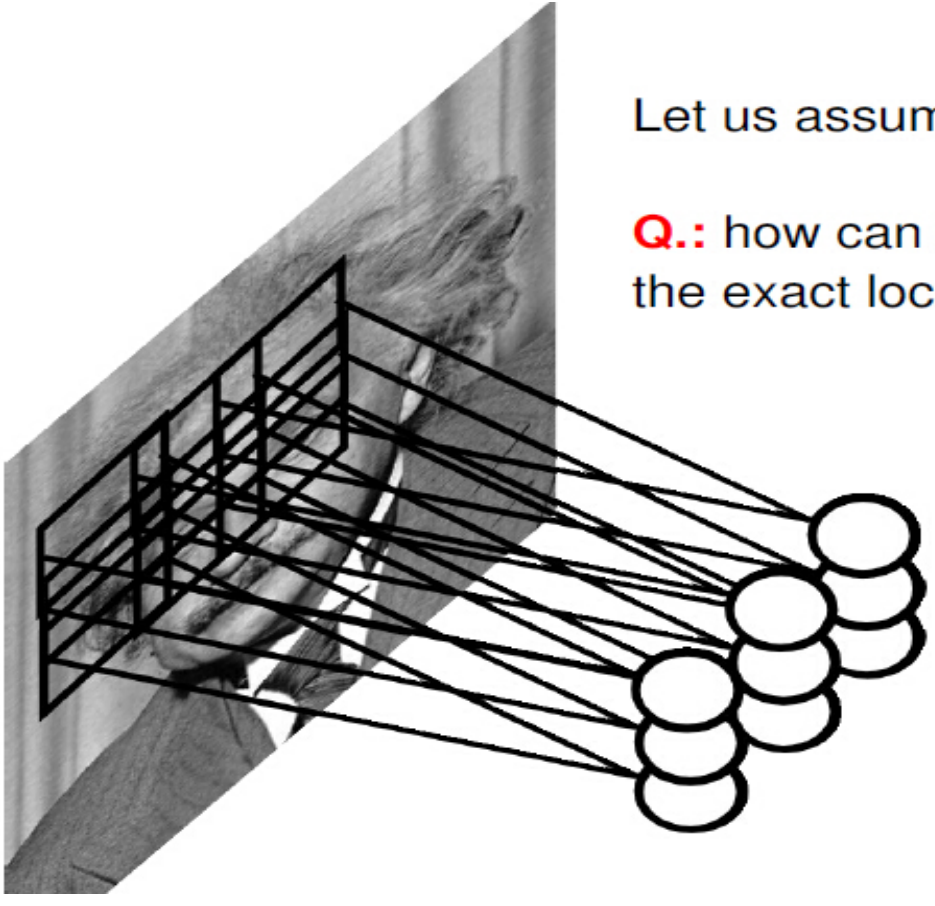
Convolutions with learned kernels

$$A_{ij} = \sum_{kl} W_{kl} X_{i+j.\,k+l}$$

The filtered "image" Z is called a feature map

$$Z_{ij} = max\left(0, A_{ij}\right)$$

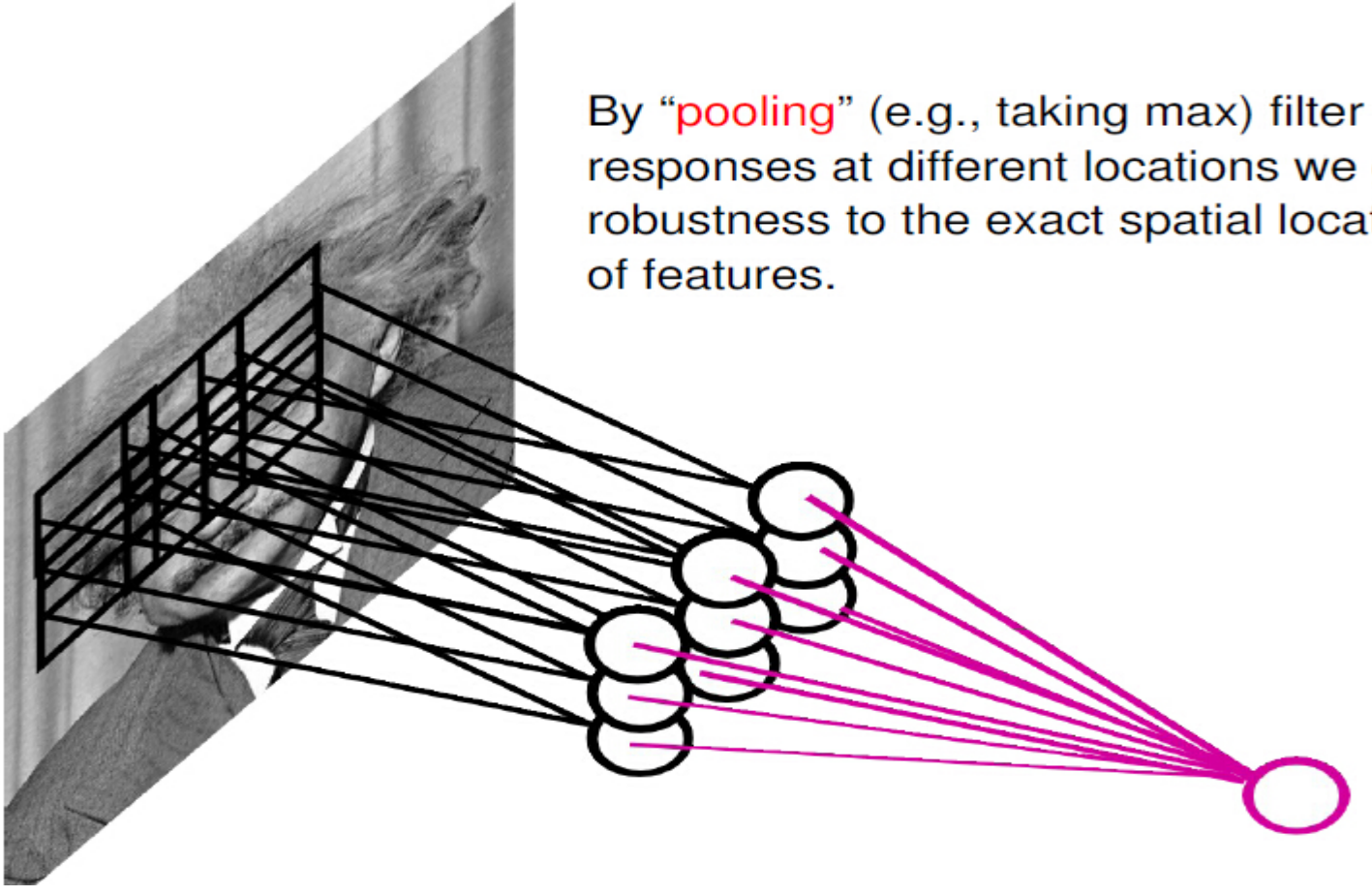
#4: Pooling Layer (subsampling)



Let us assume filter is an “eye” detector.

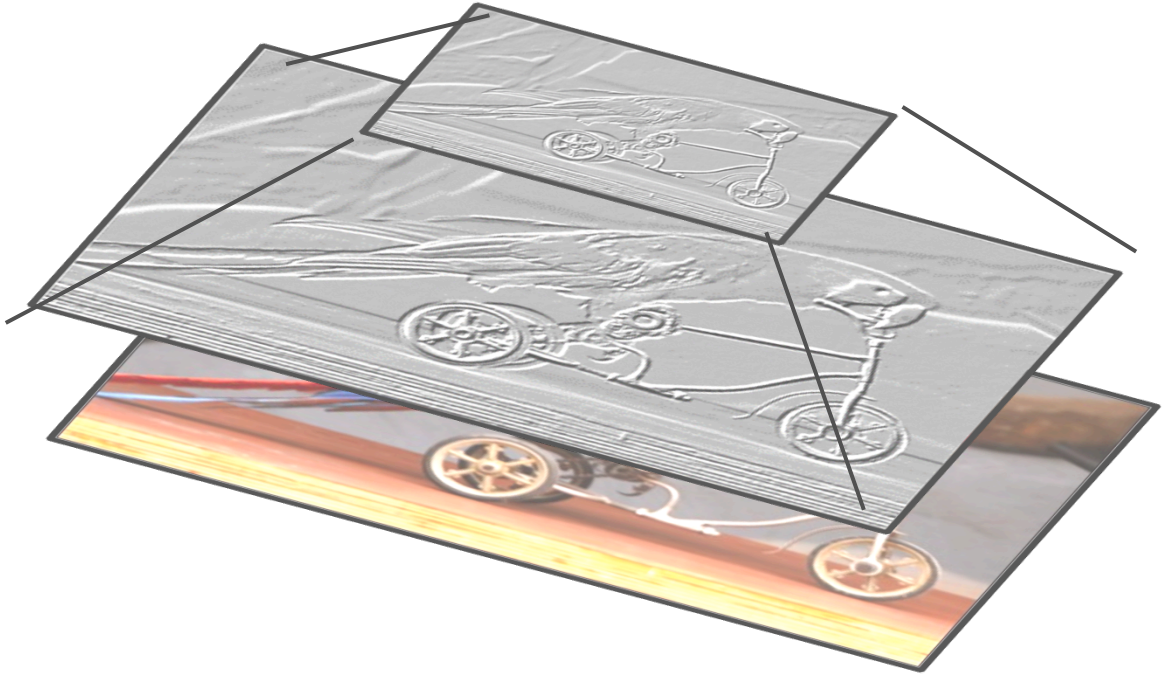
Q.: how can we make the detection robust to the exact location of the eye?

Pooling Layer (subsampling)



By "pooling" (e.g., taking max) filter responses at different locations we gain robustness to the exact spatial location of features.

An example: Convolution + Pooling

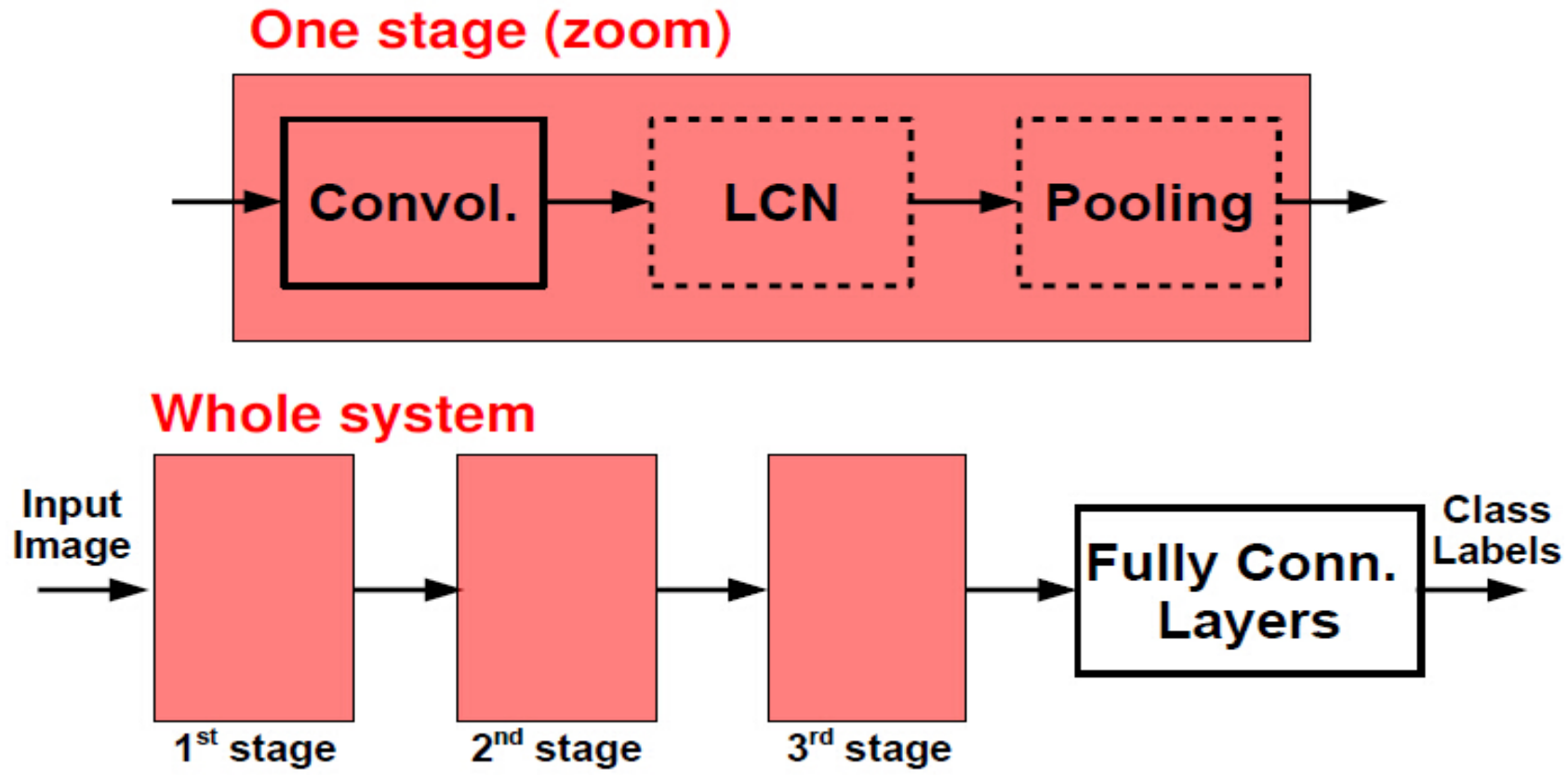


Pooling

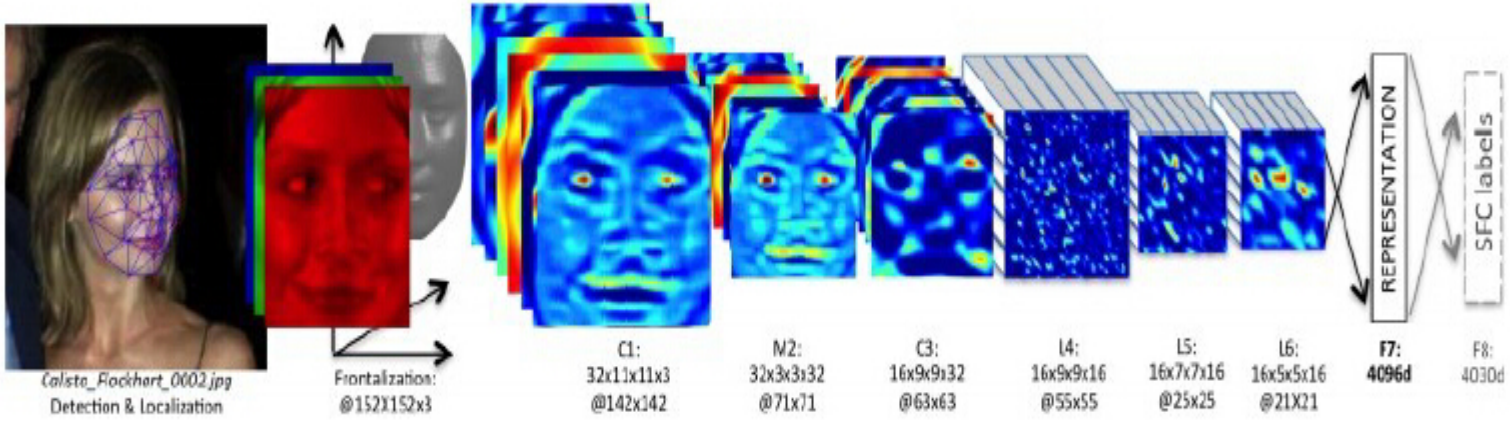
Convolution

Image

ConvNets: Typical Architecture



Examples – DeepFace

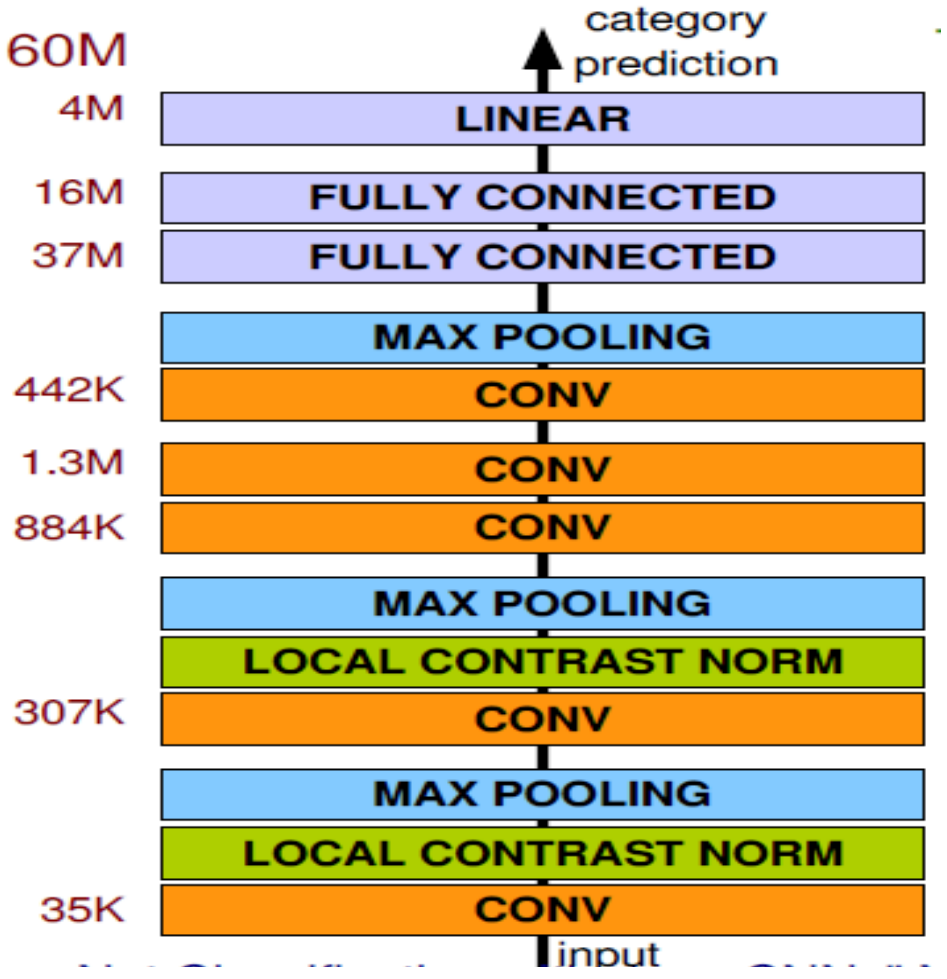


Source of the image: DeepFace: Closing the Gap to Human-Level Performance in Face Verification: CVPR 2014

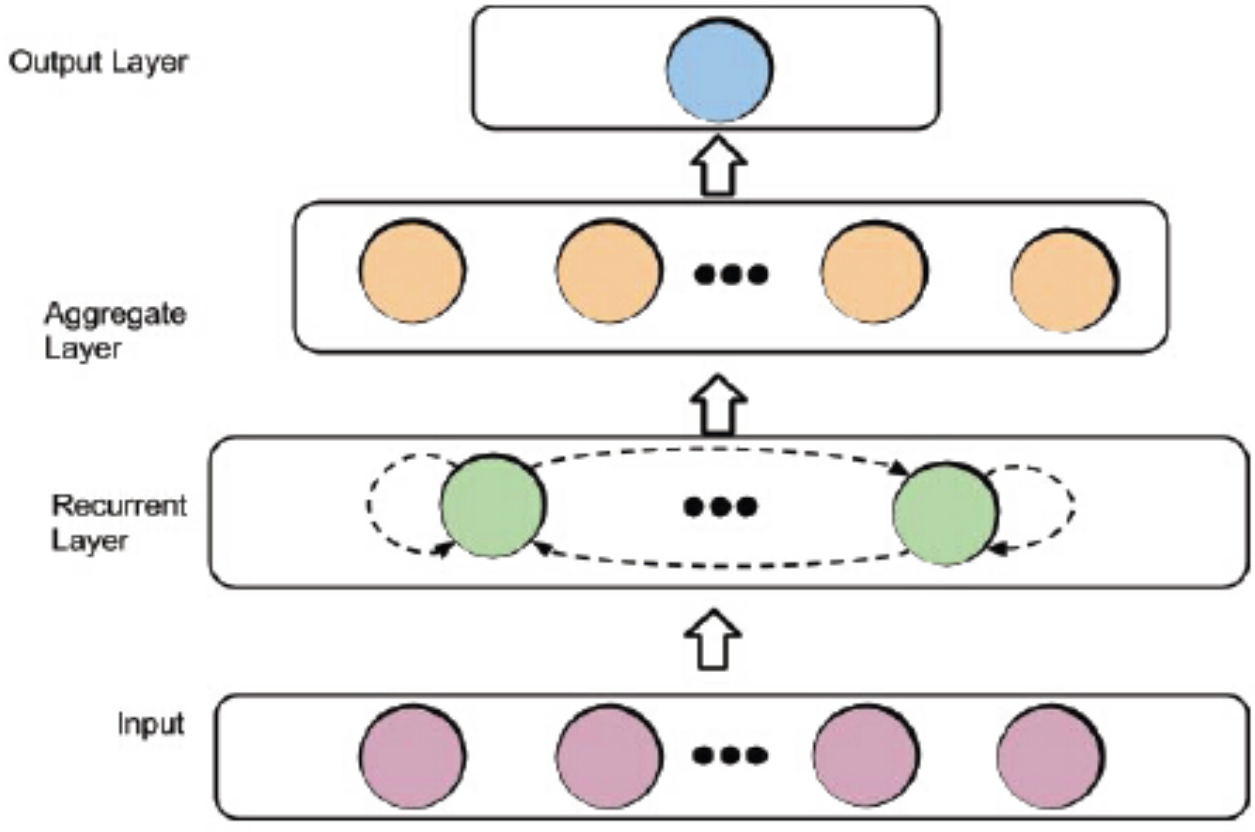
Examples – Alex CNN

7 hidden layers
650,000 neurons
60,000,000 parameters

Trained on 2 GPUs
for a week!

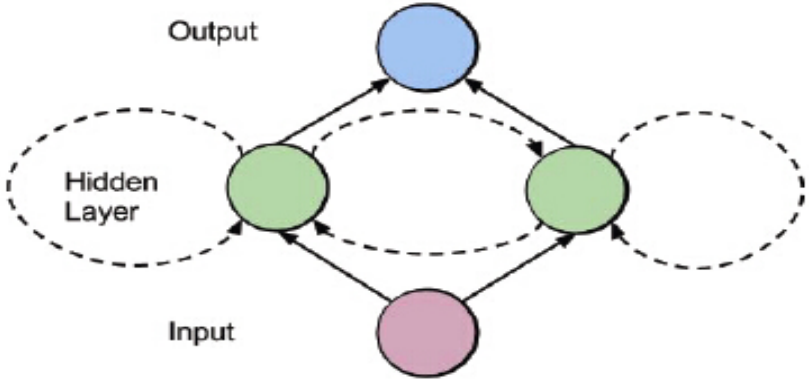


Recurrent Neural Nets



Input: sequential data/time series

Simple Recurrent Neural Net



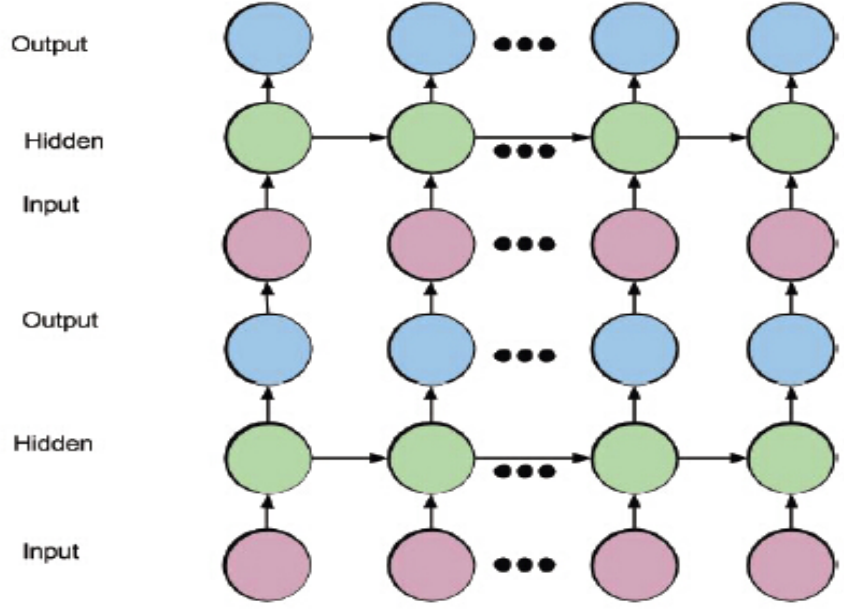
$$\mathbf{h}_t = \sigma(\mathbf{W}_h \mathbf{x}_t + \mathbf{U}_h \mathbf{h}_{t-1} + \mathbf{b}_h)$$

$$\mathbf{y}_t = \text{softmax}(\mathbf{W}_y \mathbf{h}_t + \mathbf{b}_y)$$

- Recurrent edges may form cycles, including self-connections
- Optimization is especially difficult due to long-range dependencies
- Vanishing and exploding gradients occur when propagating errors across many time steps[Lipton, 2015]

Stacking Recurrent Layer: deep RNNs

- Common RNN do not accommodate temporal hierarchy [Hermans and Schrauwen, 2013]
- A deep hidden-to-output function can be useful to disentangle the factors of variations in the hidden state



Towards human level capability

Language

Image

Video

Semantic parsing in general

- Named entity recognition

  In **April 1775** fighting broke out between **Massachusetts** militia units and **British** regulars at **Lexington** and **Concord** .

  State-of-the-art result: F1 74.39% on ontonotes-5.0 18-class data. [Lu, Li, Xu, 2015]

- Semantic role labeling

  The excess supply pushed gasoline prices down in that period .

  subject verb object

  State-of-the-art result: F1 81.27% on CoNLL-2012 test data. [Zhou & Xu, ACL 2015]

- Question Answering: subject parsing

  Who created Harry Potter ?

Neural casual chatting machine


今天午饭好好吃好开心！

So happy to have delicious lunch today!


我也要吃！

I want to eat too!


土豪我们做朋友吧

Let us befriend, rich guy


我不是土豪

I am not rich


你喜欢一见钟情还是日久生情

Would you prefer falling in love at first sight or developing love over time?


一见钟情吧

Falling in love at first sight probably


星球大战好看吗？

Is Star Wars worth watching?


不是很好看

Not very much

[Fei Jiang & Lei Li 2015]

What about longer utterance?

I once let the truest love slip away
from before my eyes,
Only to find myself regretting when it
was too late,
No pain in the world comes near to
this,
If only God would give me another
chance,
I would say to the girl, I love you!
If there had to be a limit of time,
I pray it's ten thousand years.

Bless you.

Describing images through m-RNN

a person standing on a beach holding an umbrella

a clock on the side of a building

a red and yellow train traveling down train tracks

Room to improve

a double decker bus on a city street

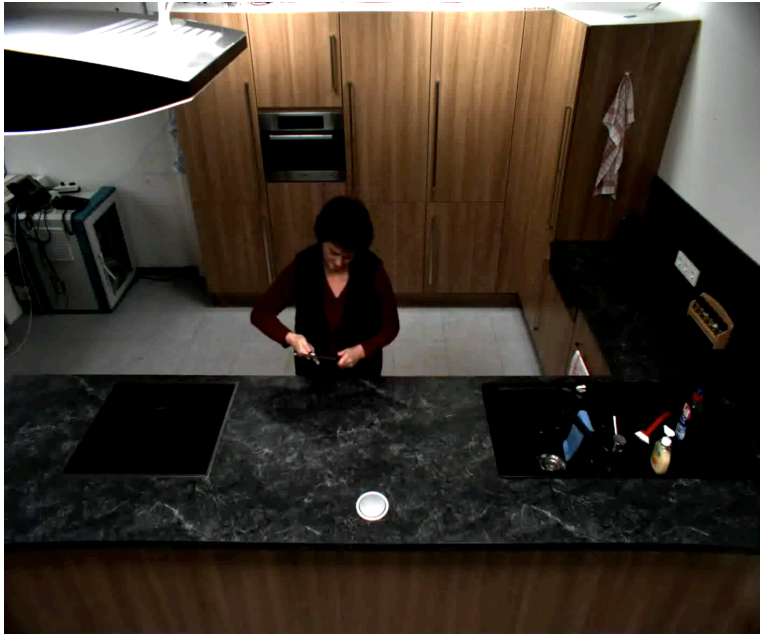
a person riding a horse in a field

a little girl brushing her teeth with a toothbrush

Can machines describe a video

A dog is playing in a bowl.

Hierachical RNN
[Haonan Yu, Jiang Wang et al, CVPR 2016]



The person entered the kitchen.
The person took out a knife and a sharpener.
The person sharpened the knife.

Answering questions about image

Q: what is the color of the bus?
A: yellow

Q: what are there hanging up?
A: umbrellas

Q: What is the color of the cake?
A: red

ABC-CNN
[Chen, Wang et al 2015]

Face recognition

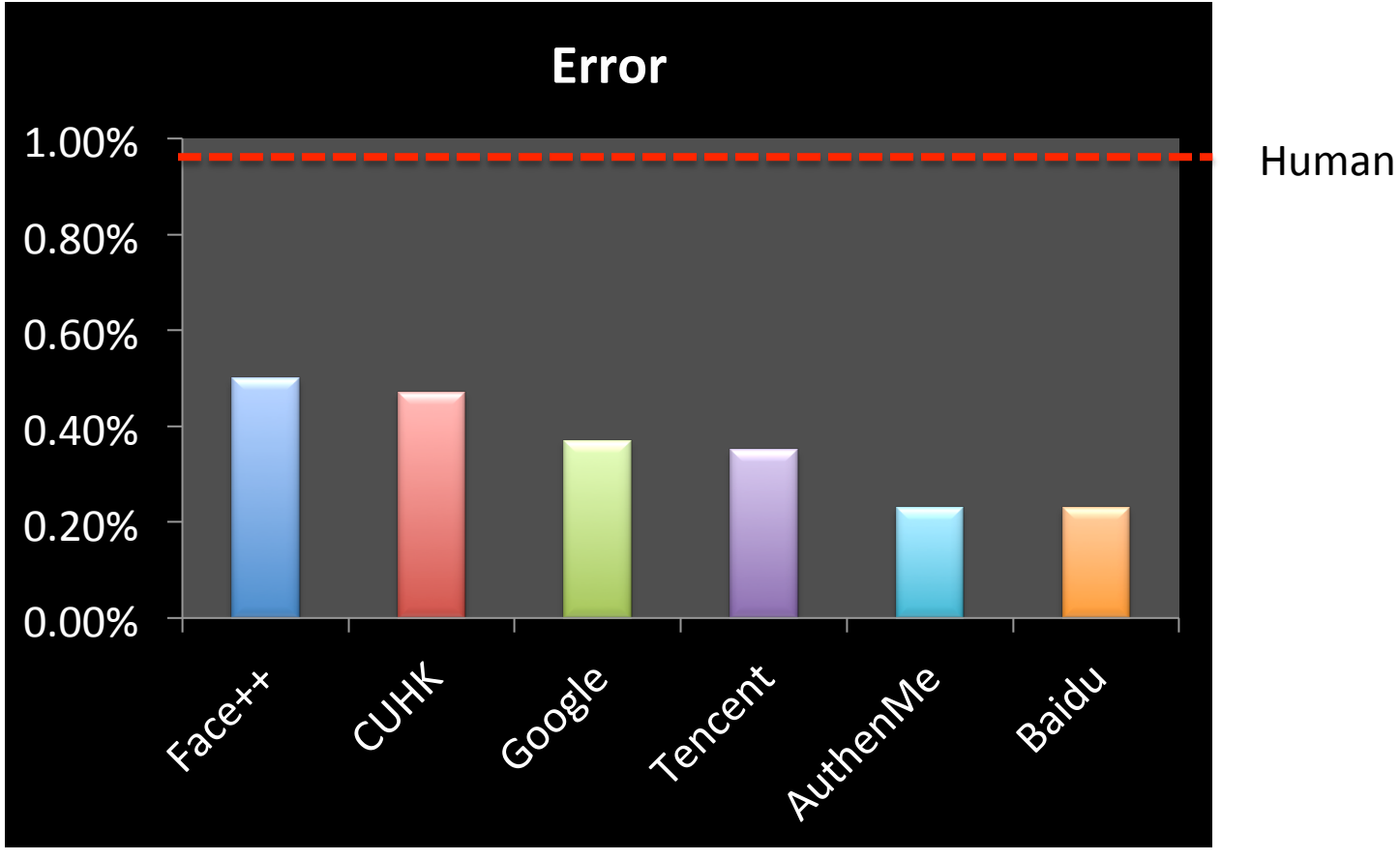
Are they the same person?

Same

Different

LFW Dataset: 13k images of 6k persons

Deep learning to recognize face



Visual Search (image-to-image)

Visual search

Query

Baidu

Other search engine

Summary

- Deep Learning made big success in solving industry real problems
- New paradigm of AI: Big data + better models + scalable computing
- Machines perform better than or close to human on multiple tasks

Year 1955

- "We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer."
- (McCarthy, John; Minsky, Marvin; Rochester, Nathan; Shannon, Claude (1955), A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence)

Looking into the future

Capabilities for computers to acquire

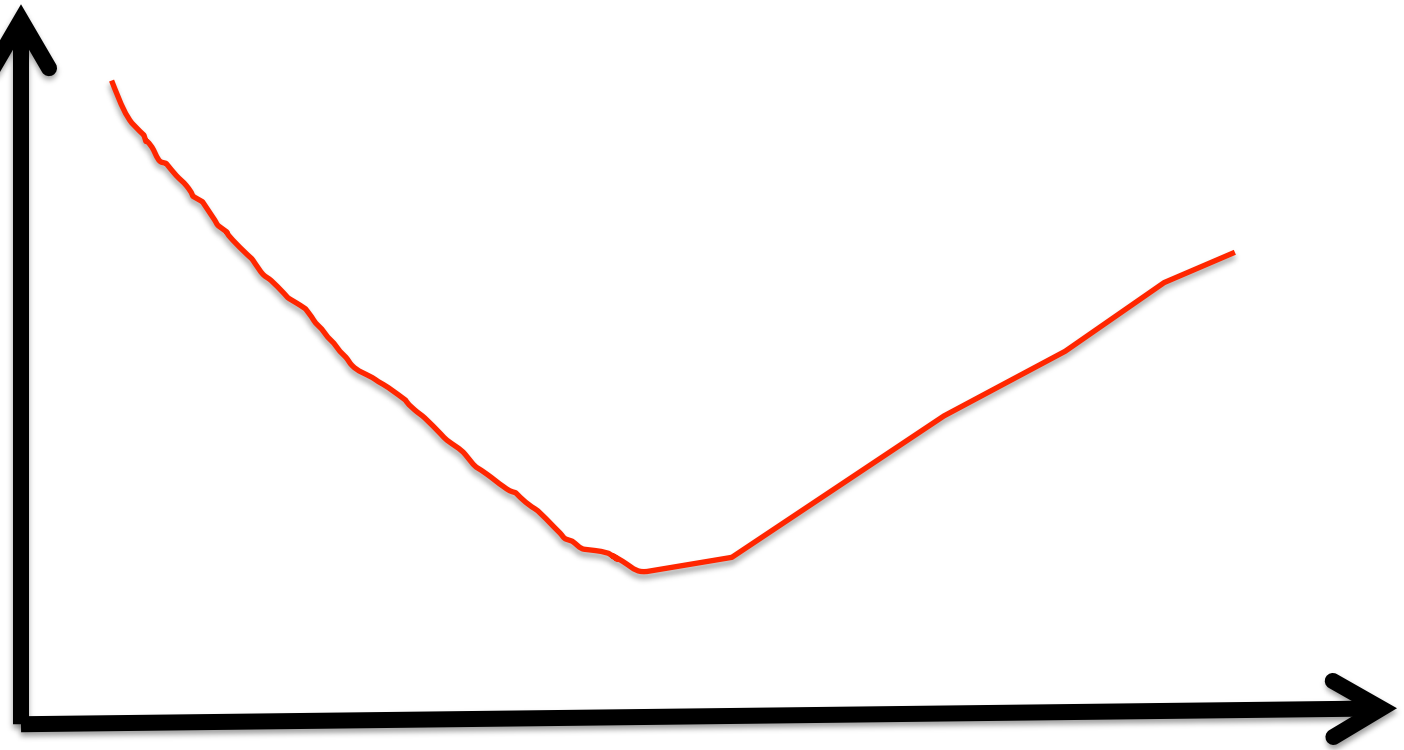
- Learning w/ unlabeled data
- Learning w/ small data
- Reasoning about objects and environment

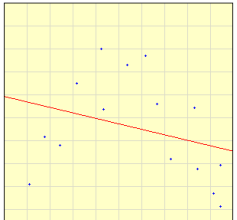
To Be, or Not To Be Deep?

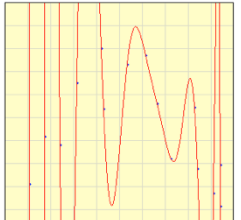
To Be, or Not To Be Deep?

$$error_{train}(\mathbf{w}) = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

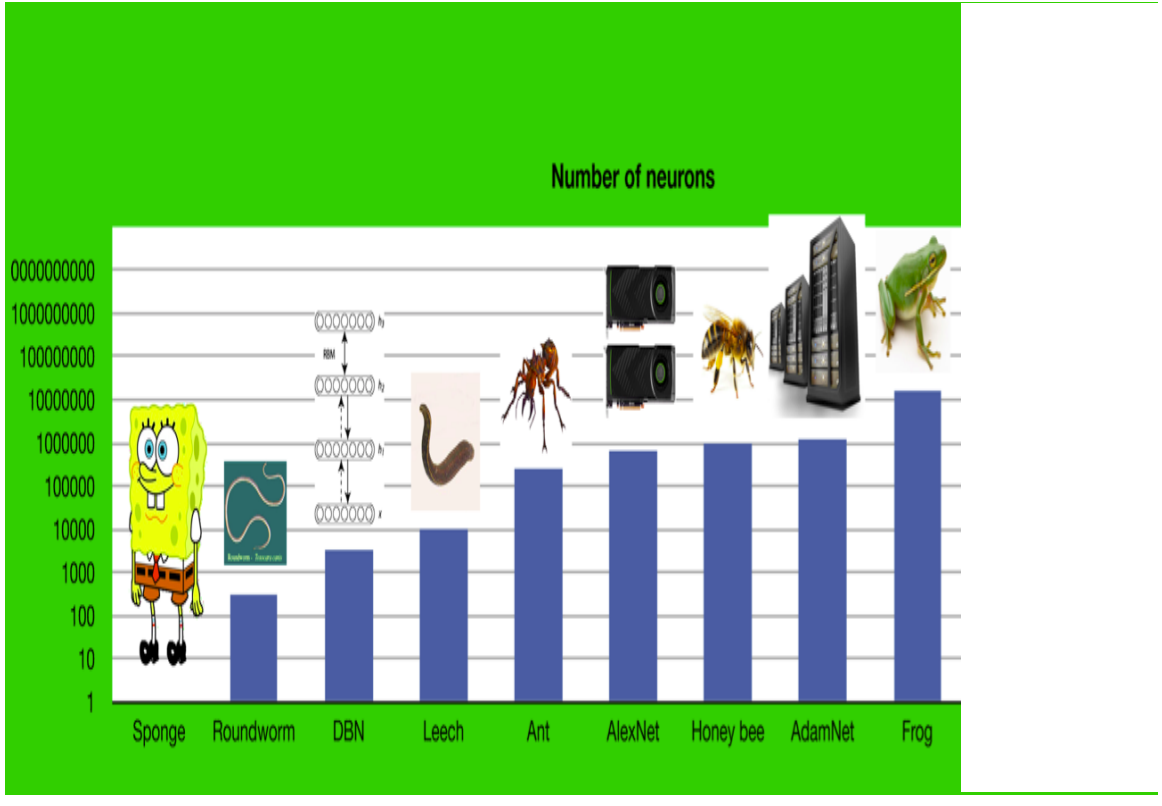
$$error_{true}(\mathbf{w}) = \int_{\mathbf{x}} \left(t(\mathbf{x}) - \sum_i w_i h_i(\mathbf{x}) \right)^2 p(\mathbf{x}) d\mathbf{x}$$







An Optimistic Hope

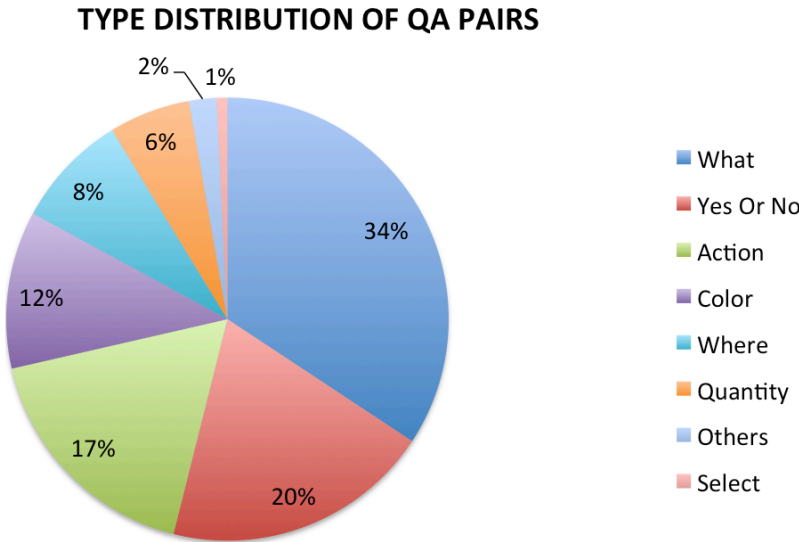


Open source Deep Learning platforms

- Caffe: http://caffe.berkeleyvision.org/
- Torch: http:// www.torch.ch
- Theano: http://deeplearning.net/software/theano/
- TensorFlow: https://www.tensorflow.org/
- DeepLearnToolbox: https://github.com/rasmusbergpalm/DeepLearnToolbox
- Pylearn2: https://github.com/lisa-lab/pylearn2
- Deep Learning book: http://www.deeplearningbook.org/

FM-IQA dataset for public research

- 120,000 images, 250,000 freestyle QA pairs.
- Multilingual
- Complex answers
- http://idl.baidu.com/FM-IQA.html



[Gao, Mao, Zhou, Huang, Wang, Xu, NIPS 2015]

Reference

Seq-to-Seq

- Sutskever et al, Sequence to sequence learning with Neural Networks.
- Bahdanau, et al, Neural machine translation by jointly learning to align and translate.
- Shang et al, Neural responding machine for short-text conversation.
- Vinyals & Le, A neural conversational model.

Reference

Parsing & Sequence labeling

- Collobert et al, Natural language processing almost from scratch.
- Lu et al, Twisted recurrent network for named entity recognition.
- Huang et al, Bidirectional LSTM-CRF models for sequence tagging.
- Zhou et al, End-to-end learning of semantic role labeling using recurrent neural networks.

Reference

Image Captioning

- Mao et al, Explain images with multimodal recurrent neural networks.
- Karpathy et al, Deep visual-semantic alignments for generating image descriptions.
- Kiros et al, Unifying viusal-semantic embeddings with multimodal neural language models.
- Vinyals et al, show and tell.
- Chen & Zitnick, Learning a recurrent visual representation for image caption generation.

Reference

Video captioning

- Venugopalan et al, Translating videos to natural language using deep recurrent neural networks.
- Donahue et al, Long-term recurrent convolutional networks for visual recognition and description.
- Venugopalan et al, Sequence to sequence – video to text.
- Xu et al, A multi-scale multiple instance video description network.
- Yao et al, describing videos by exploiting temporal structure.

Reference

Image QA

- Chen et al, ABC-CNN: An attention based convolutional neural network for visual question answering.
- Gao et al, Are you talking to a machine? dataset and methods for multlingual image question answering.
- Malinowski et al, A neural-based approach to answering questions about images.
- Ren et al, exploring models and data for image question answering.