Deep Learning

Dr. Xiaowei Huang

https://cgi.csc.liv.ac.uk/~xiaowei/

Topics

- Deep Learning in
  - Computer Vision
  - Natural Language Processing (NLP)
  - Speech
  - Robotics and AI
  - Music and the arts!
- A brief history of Neural Networks

Deep Learning in Computer Vision

- Object and activity recognition
  - https://www.youtube.com/watch?v=qrzQ_AB1DZk
- Object detection and segmentation
  - https://www.youtube.com/watch?v=CxanE_W46ts
- Image captioning and Q&A
  - https://www.youtube.com/watch?v=8BFzu9m52sc

Why should we be impressed?

- Vision is ultra challenging!
  - For a small 256x256 resolution and for 256 pixel values
  - a total $2^{524,288}$ of possible images
  - In comparison there are about 10^{24} stars in the universe
- Visual object variations
  - Different viewpoints, scales, deformations, occlusions
- Semantic object variations
  - Intra-class variation
  - Inter-class overlaps

Deep Learning in Robotics

- Self-driving cars
  - https://www.youtube.com/watch?v=-96BEoXJMs0
- Drones and robots
  - https://www.youtube.com/watch?v=2hGngG64dNM
- Game AI
  - https://www.youtube.com/watch?v=V1eYniJ0Rnk

Why should we be impressed?

- Typically robotics are considered in controlled environments
  - Laboratory settings, Predictable positions, Standardized tasks (like in factory robots)
- What about real life situations?
  - Environments constantly change, new tasks need to be learnt without guidance, unexpected factors must be dealt with
- Game AI
  - At least $10^{10^{48}}$ possible GO games. Where do we even start?

Deep Learning in NLP and Speech

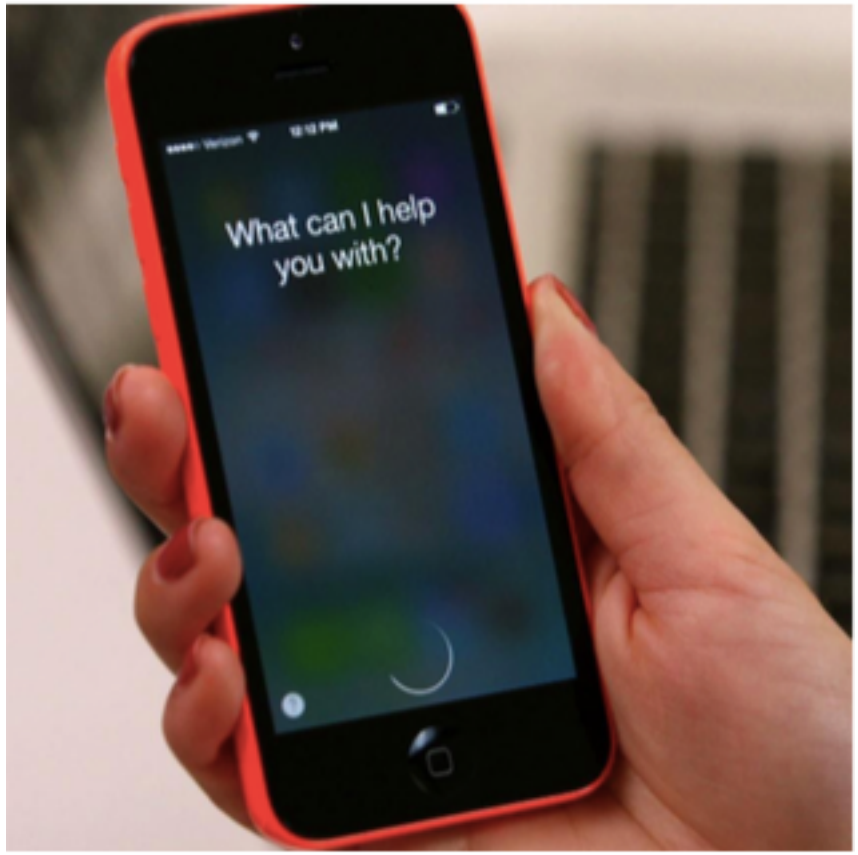
- Word and sentence representations

Newspapers			
New York	New York Times	Baltimore	Baltimore Sun
San Jose	San Jose Mercury News	Cincinnati	Cincinnati Enquirer
NHL Teams			
Boston	Boston Bruins	Montreal	Montreal Canadiens
Phoenix	Phoenix Coyotes	Nashville	Nashville Predators
NBA Teams			
Detroit	Detroit Pistons	Toronto	Toronto Raptors
Oakland	Golden State Warriors	Memphis	Memphis Grizzlies
Airlines			
Austria	Austrian Airlines	Spain	Spainair
Belgium	Brussels Airlines	Greece	Aegean Airlines
Company executives			
Steve Ballmer	Microsoft	Larry Page	Google
Samuel J. Palmisano	IBM	Werner Vogels	Amazon

Table 2: Examples of the analogical reasoning task for phrases (the full test set has 3218 examples). The goal is to compute the fourth phrase using the first three. Our best model achieved an accuracy of 72% on this dataset.

Deep Learning in NLP and Speech

- Speech recognition and Machine translation

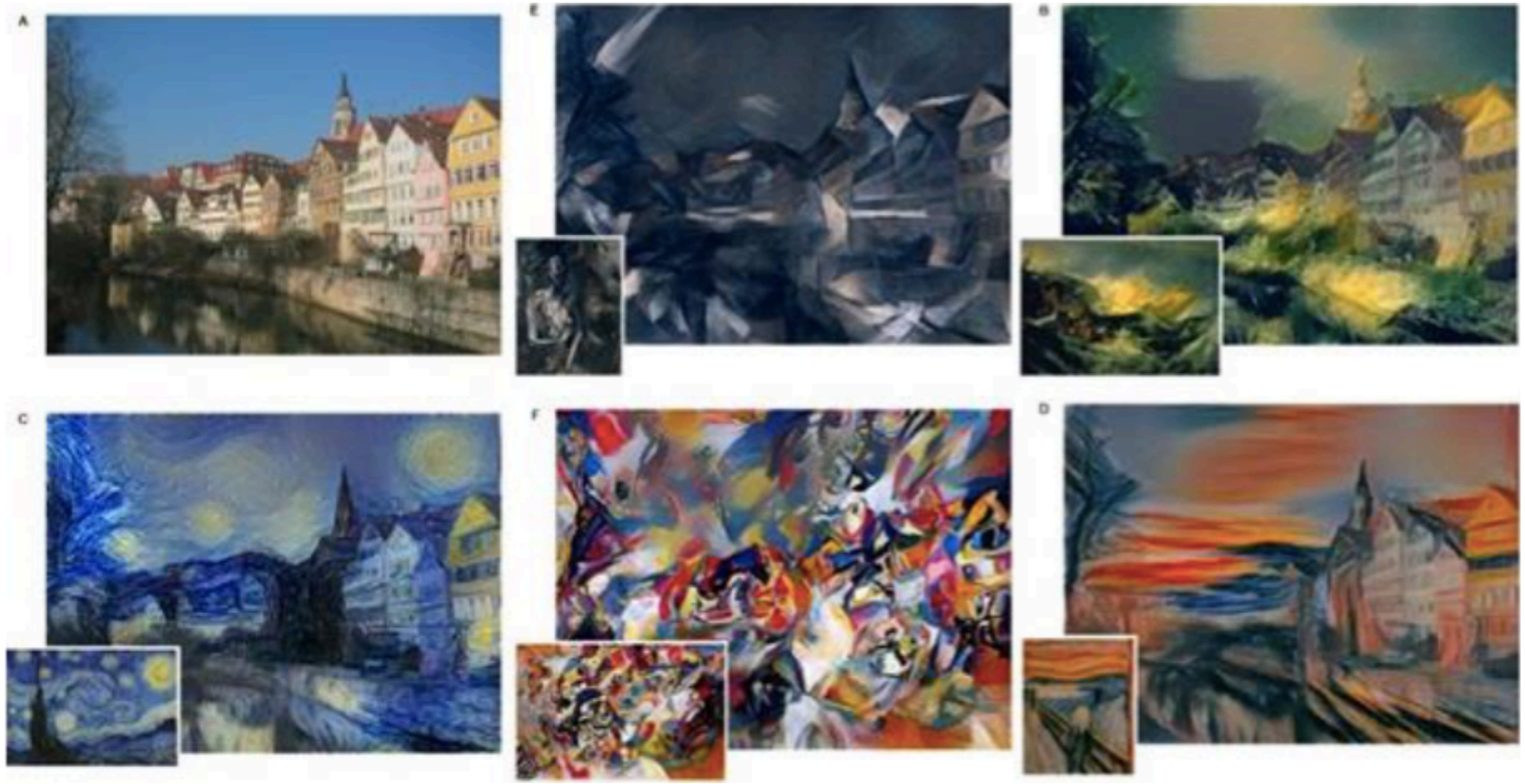


Why should we be impressed?

- NLP is an extremely complex task
  - synonymy (“chair”, “stool” or “beautiful”, “handsome”)
  - ambiguity (“I made her duck”, “Cut to the chase”)
- NLP is very high dimensional
  - assuming 150K english words, we need to learn 150K classifiers
  - with quite sparse data for most of them
- Beating NLP feels the closest to achieving true AI
  - although true AI probably goes beyond NLP, Vision, Robotics, ... alone

Deep Learning in the arts

- Imitating famous painters



Handwriting

- http://www.cs.toronto.edu/~graves/handwriting.html

Why should we be impressed?

- Music, painting, etc. are tasks that are uniquely human
  - Difficult to model
  - Even more difficult to evaluate (if not impossible)
- If machines can generate novel pieces even remotely resembling art, they must have understood something about “beauty”, “harmony”, etc.
- Have they really learned to generate new art, however?
  - Or do they just fool us with their tricks?

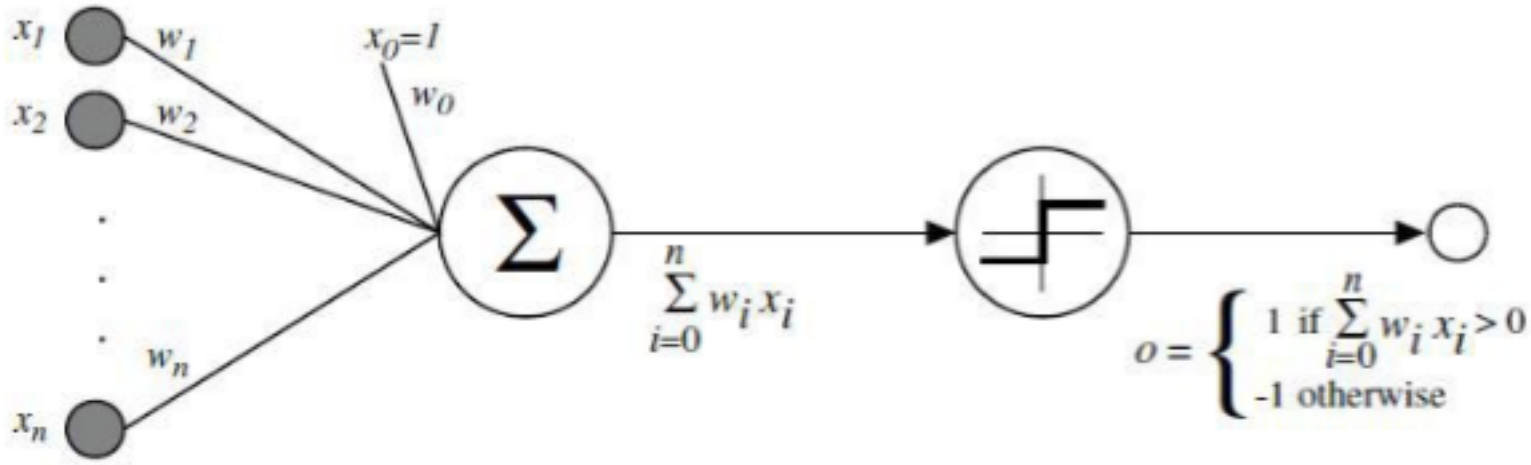
A brief history of Neural Networks & Deep Learning

First Appearance (Roughly)

Perceptrons

- Rosenblatt proposed a machine for binary classifications
- 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 0

$w^T x + b?$



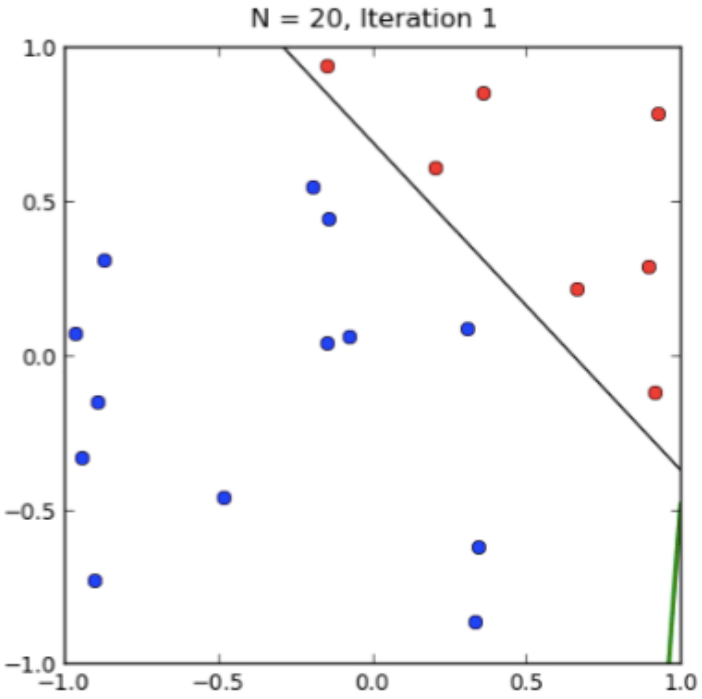
Training a perceptron

- Rosenblatt's innovation was mainly the learning algorithm for perceptrons
- 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

$$\Delta w_i = \eta(y - \hat{y})x_i \qquad w_i \leftarrow w_i + \Delta w_i$$

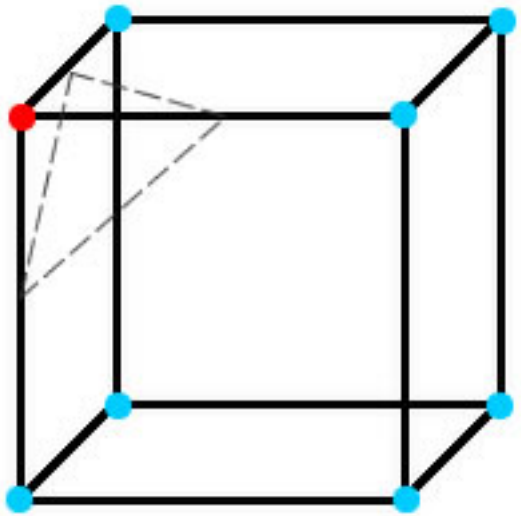
Repeat until no errors are made

η is *learning rate*;
set to value << 1



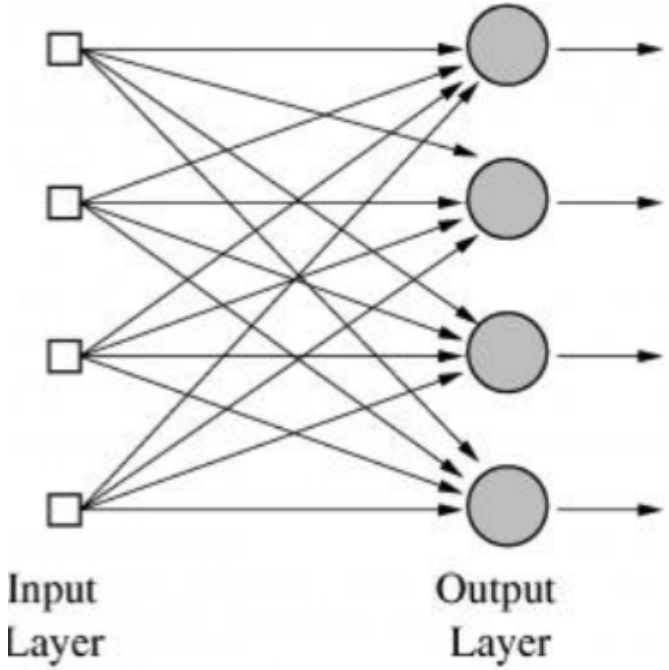
Representational power of perceptrons

- in previous example, feature space was 2D so decision boundary was a line
- in higher dimensions, decision boundary is a hyperplane



From a perceptron to a neural network

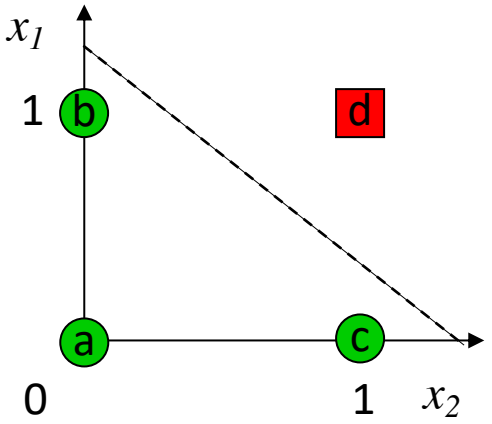
- One perceptron = one decision
- What about multiple decisions?
  - E.g. digit classification
- Stack as many outputs as the possible outcomes into a layer
  - Neural network
- Use one layer as input to the next layer
  - Multi-layer perceptron (MLP)



Some linearly separable functions

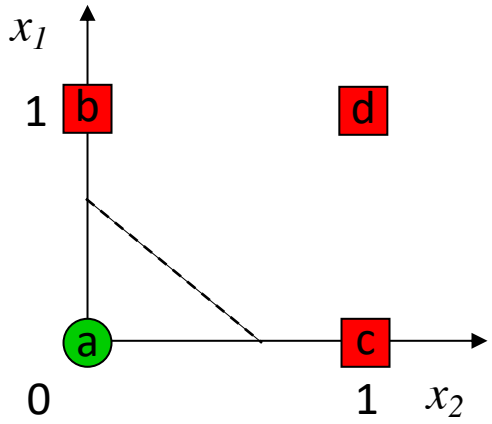
AND

	x_1	x_2	y
a	0	0	0
b	0	1	0
c	1	0	0
d	1	1	1



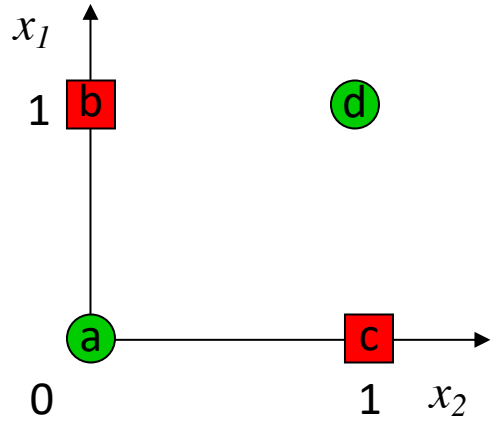
OR

	x_1	x_2	y
a	0	0	0
b	0	1	1
c	1	0	1
d	1	1	1



XOR is not linearly separable

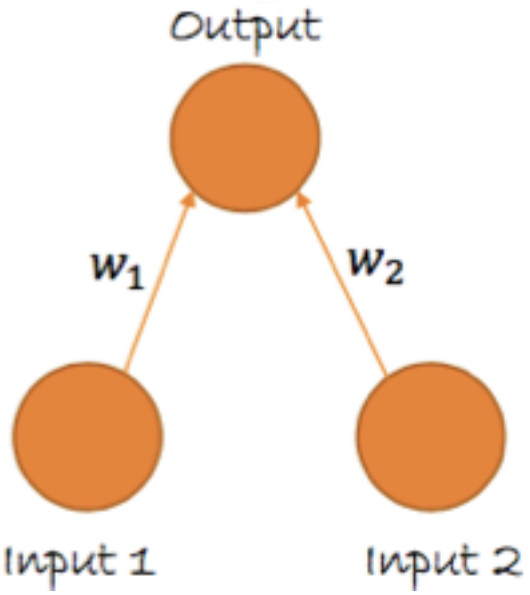
	x_1	x_2	y
a	0	0	0
b	0	1	1
c	1	0	1
d	1	1	0



XOR & Multi-layer Perceptrons

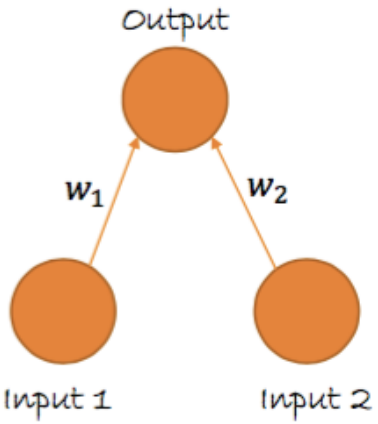
- However, the exclusive or (XOR) cannot be solved by perceptrons
  - [Minsky and Papert, “Perceptrons”, 1969]

Input 1	Input 2	Output
1	1	0
1	0	1
0	1	1
0	0	0



XOR & Multi-layer Perceptrons

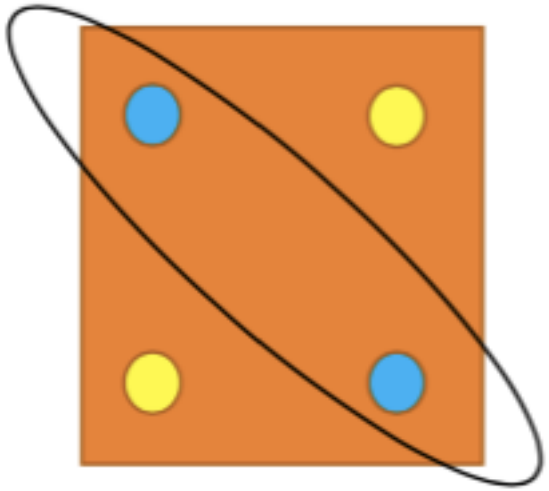
- However, the exclusive or (XOR) cannot be solved by perceptrons
  - [Minsky and Papert, “Perceptrons”, 1969]



Input 1	Input 2	Output
1	1	0
1	0	1
0	1	1
0	0	0

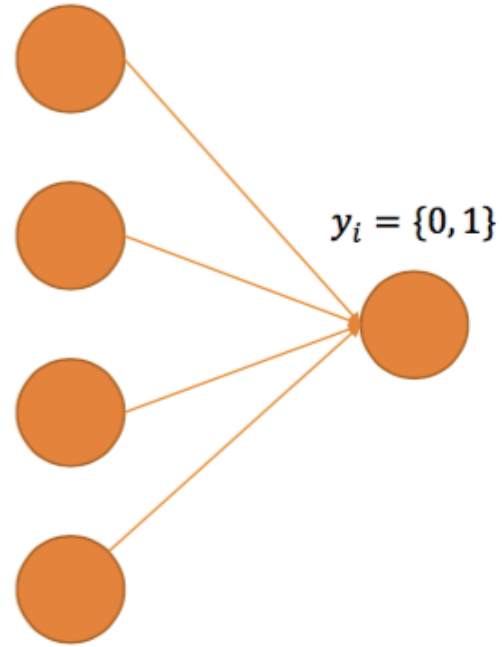
- $0\,w_1 + 0w_2 < \theta \rightarrow 0 < \theta$
- $0\,w_1 + 1w_2 > \theta \rightarrow w_2 > \theta$
- $1\,w_1 + 0w_2 > \theta \rightarrow w_1 > \theta$
- $1\,w_1 + 1w_2 < \theta \rightarrow w_1 + w_2 < \theta$

Inconsistent!!

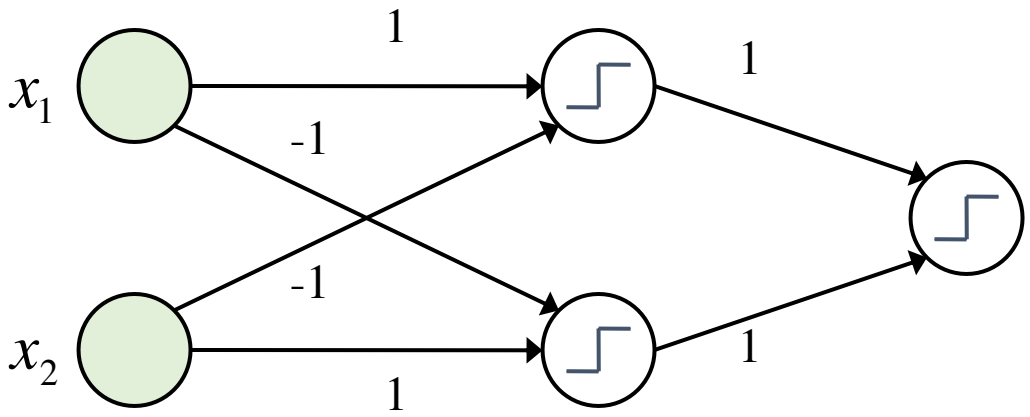


Minsky & Multi-layer perceptrons

- Interestingly, Minksy never said XOR cannot be solved by neural networks
  - Only that XOR cannot be solved with 1 layer perceptrons
- Multi-layer perceptrons can solve XOR
  - 9 years earlier Minsky built such a multi-layer perceptron



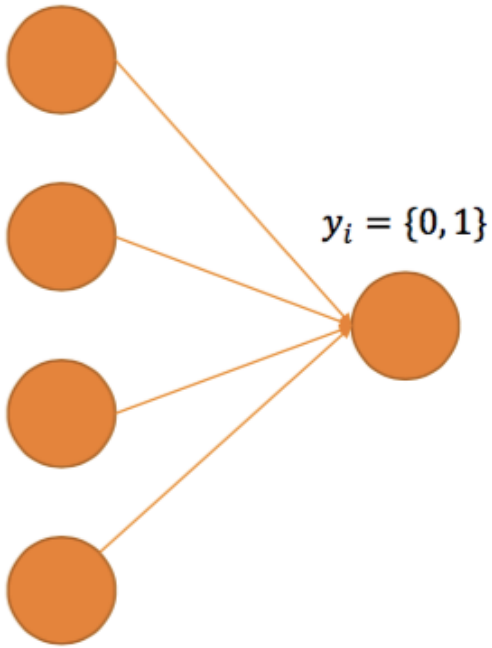
a multilayer perceptron can represent XOR



assume $w_0 = 0$ for all nodes

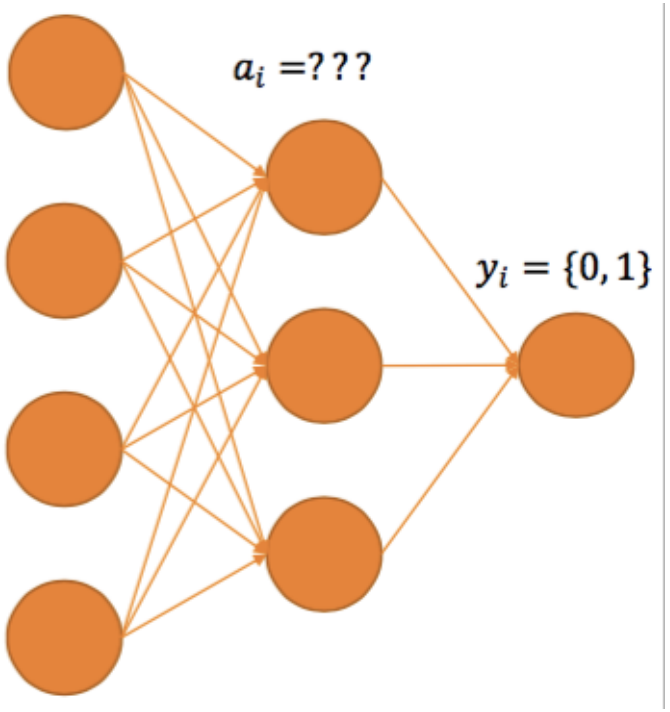
Minsky & Multi-layer perceptrons

- Interestingly, Minksy never said XOR cannot be solved by neural networks
  - Only that XOR cannot be solved with 1 layer perceptrons
- Multi-layer perceptrons can solve XOR
  - 9 years earlier Minsky built such a multi-layer perceptron
- However, how to train a multi-layer perceptron?
- Rosenblatt's algorithm not applicable, as it expects to know the desired target
  - For hidden layers we cannot know the desired target

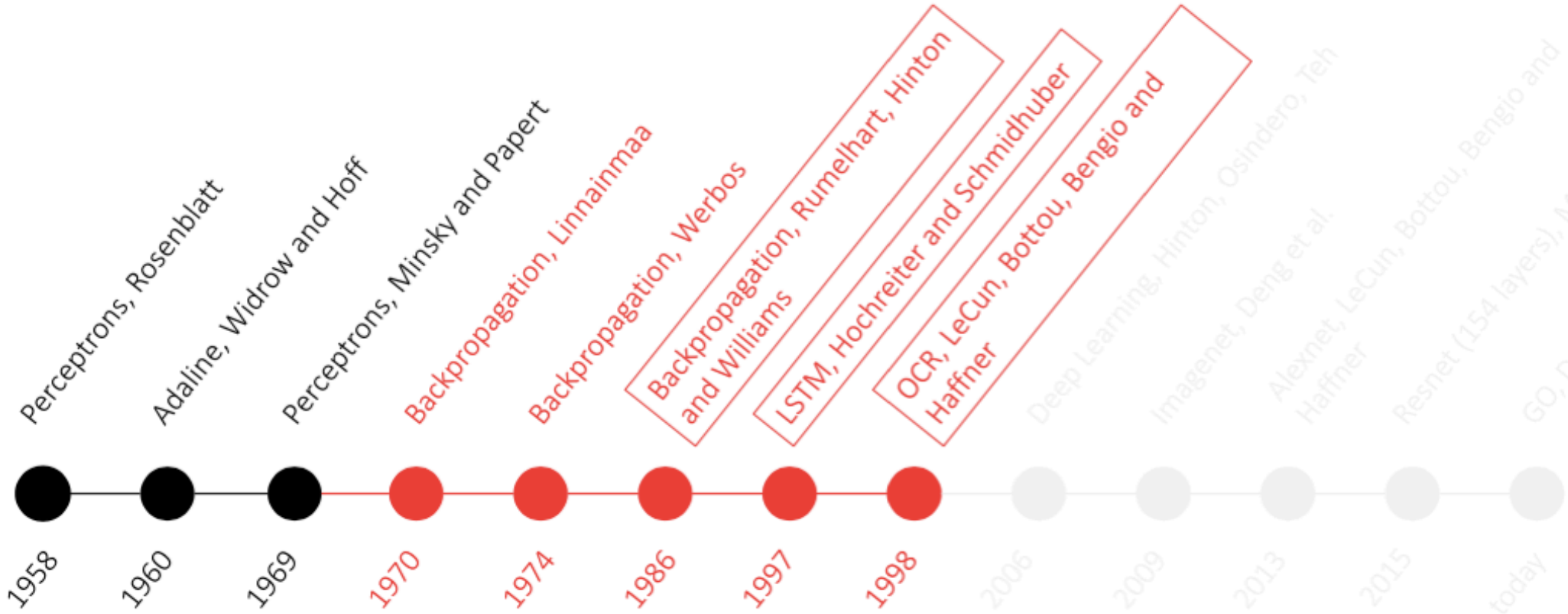


Minsky & Multi-layer perceptrons

- Interestingly, Minksy never said XOR cannot be solved by neural networks
  - Only that XOR cannot be solved with 1 layer perceptrons
- Multi-layer perceptrons can solve XOR
  - 9 years earlier Minsky built such a multi-layer perceptron
- However, how to train a multi-layer perceptron?
- Rosenblatt's algorithm not applicable, as it expects to know the desired target
  - For hidden layers we cannot know the desired target



The “AI winter” despite notable successes



The first “AI winter”

- What everybody thought: “If a perceptron cannot even solve XOR, why bother? “
  - Also, the exaggeration did not help (walking, talking robots were promised in the 60s)
- As results were never delivered, further funding was slushed, neural networks were damned and AI in general got discredited
- “The AI winter is coming”
- Still, a few people persisted
- Significant discoveries were made, that laid down the road for today’s achievements

Backpropagation

- Learning multi-layer perceptrons now possible
  - XOR and more complicated functions can be solved
- Efficient algorithm
  - Process hundreds of example without a sweat
  - Allowed for complicated neural network architectures
- Backpropagation still is the backbone of neural network training today
- Digit recognition in cheques (OCR) solved before the 2000

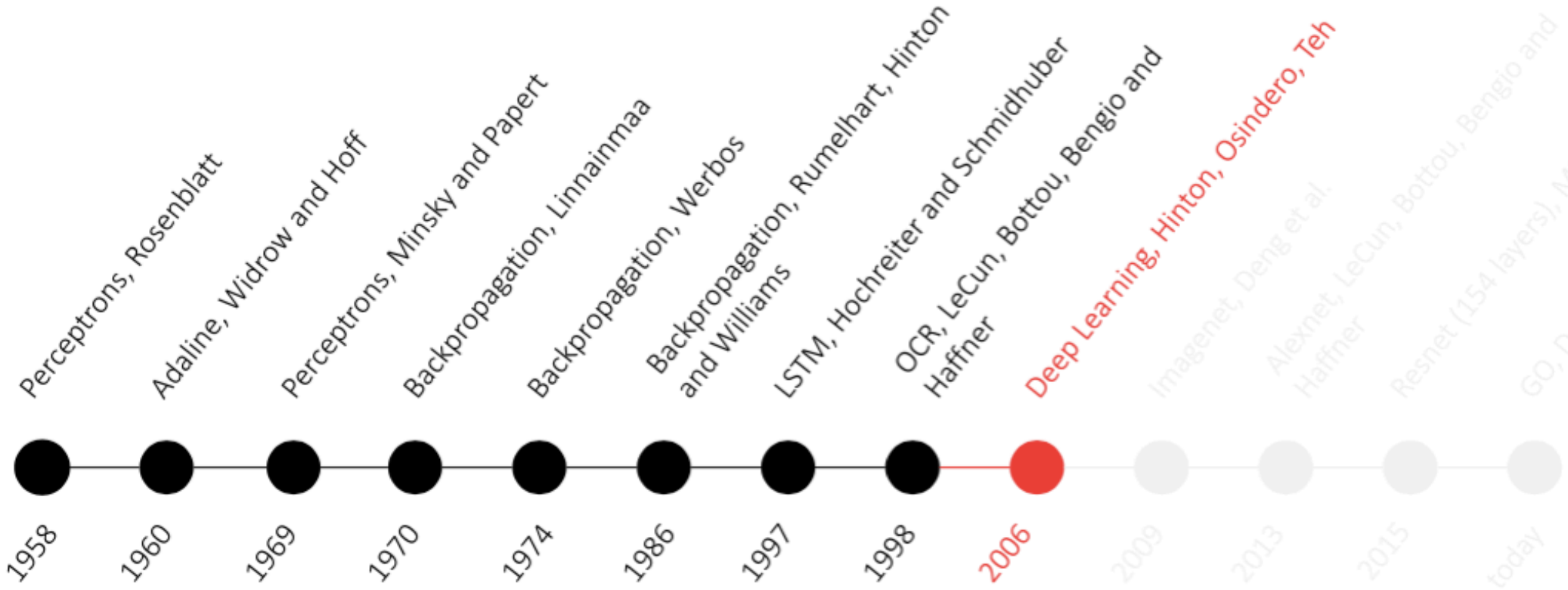
Recurrent networks

- Traditional networks are “too plain”
  - Static Input => Processing => Static Output
- What about dynamic input
  - Temporal data, Language, Sequences
- Memory is needed to “remember” state changes
  - Recurrent feedback connections
- What kind of memory
  - Long, Short?
  - Both! Long-short term memory networks (LSTM), Schmidhuber 1997

The second “AI winter”

- Until 1998 some nice algorithms and methods were proposed
  - Backpropagation
  - Recurrent Long-Short Term Memory Networks
  - OCR with Convolutional Neural Networks
- However, at the same time Kernel Machines (SVM etc.) suddenly become very popular
  - Similar accuracies in the same tasks
  - Neural networks could not improve beyond a few layers
  - Kernel Machines included much fewer heuristics & nice proofs on generalization
- As a result, once again the AI community turns away from Neural Networks

The thaw of the “AI winter”



Neural Network and Deep Learning problems

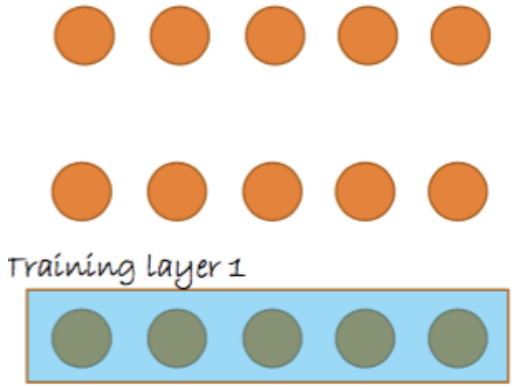
- Lack of processing power
  - No GPUs at the time
- Lack of data
  - No big, annotated datasets at the time
- Overfitting
  - Because of the above, models could not generalize all that well
- Vanishing gradient
  - While learning with NN, you need to multiply several numbers $a1 \cdot a2 \cdot \cdots \cdot an$
  - If all are equal to 0.1, for n = 10 the result is 0.0000000001, too small for any learning

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

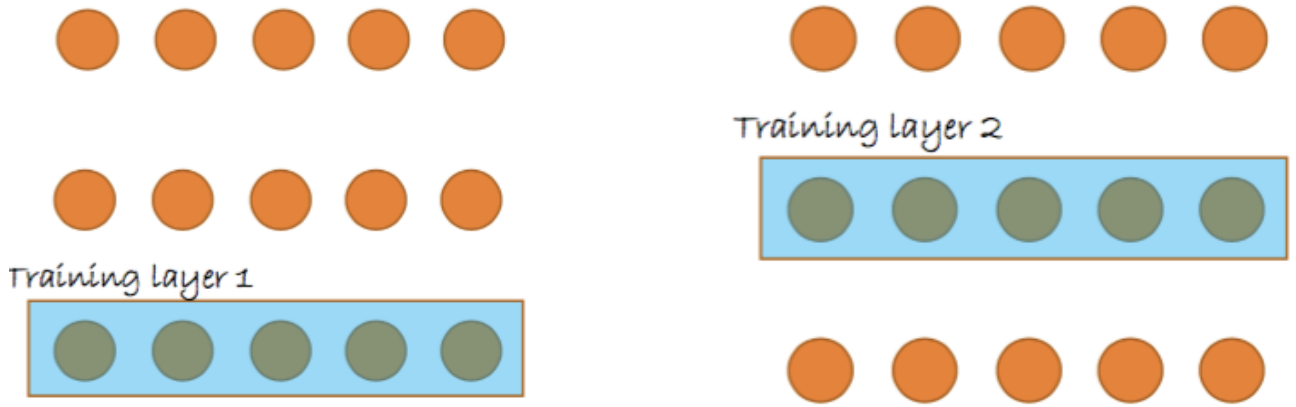
Deep Learning arrives

- Layer-by-layer training
  - The training of each layer individually is an easier undertaking
- Training multi-layered neural networks became easier
- Per-layer trained parameters initialize further training using contrastive divergence



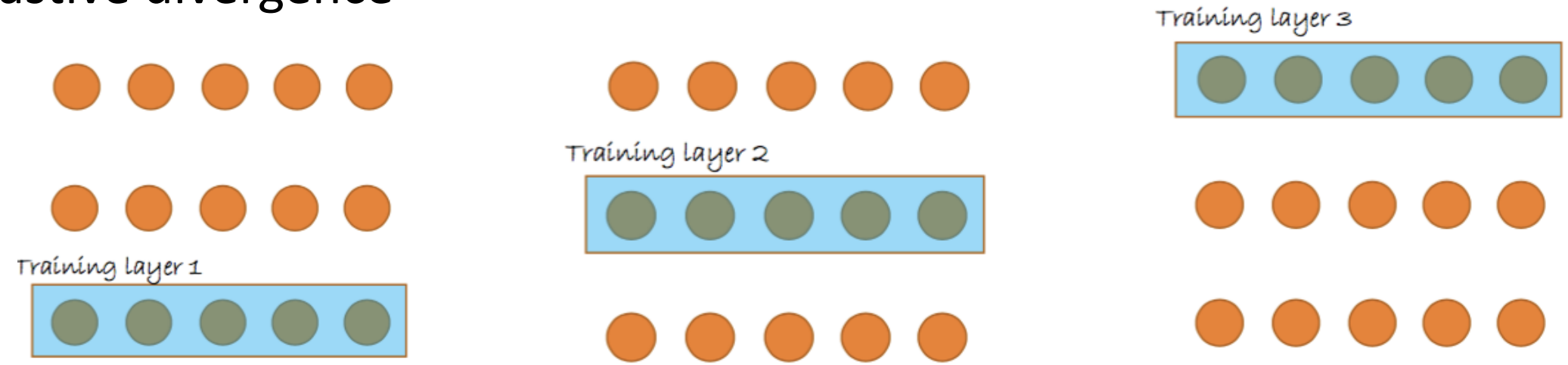
Deep Learning arrives

- Layer-by-layer training
  - The training of each layer individually is an easier undertaking
- Training multi-layered neural networks became easier
- Per-layer trained parameters initialize further training using contrastive divergence

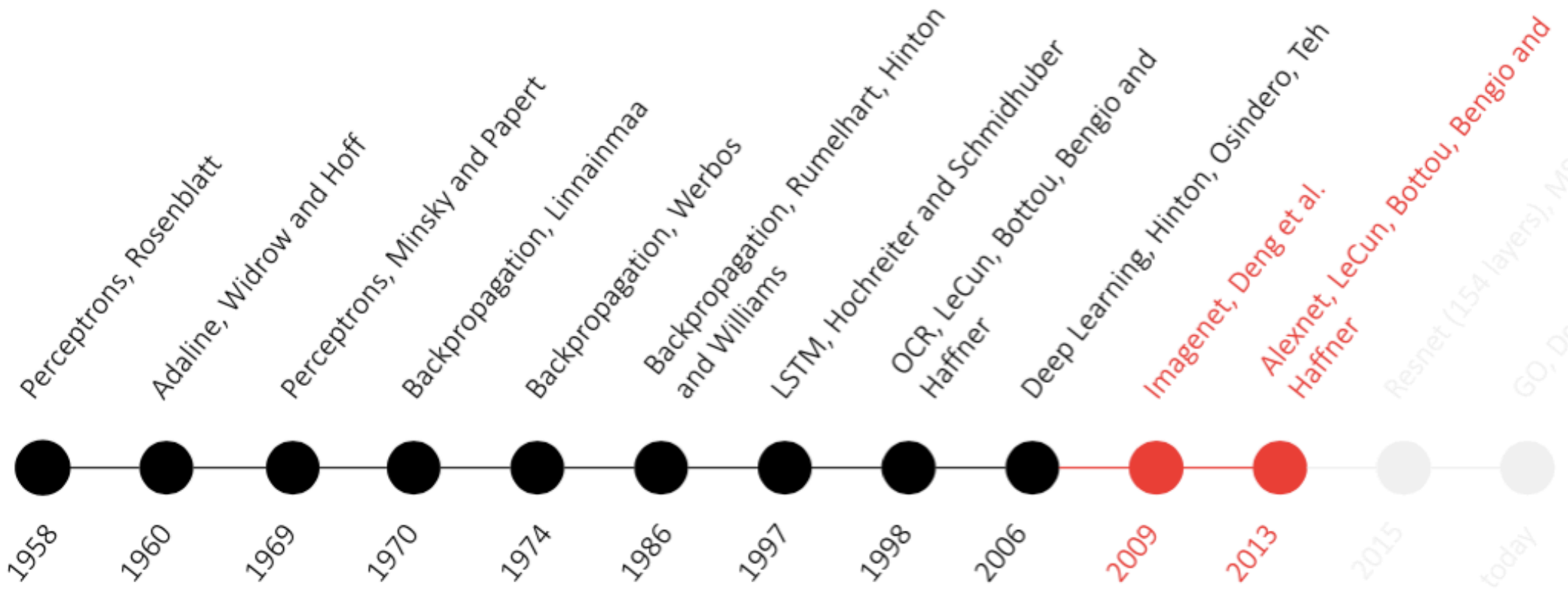


Deep Learning arrives

- Layer-by-layer training
  - The training of each layer individually is an easier undertaking
- Training multi-layered neural networks became easier
- Per-layer trained parameters initialize further training using contrastive divergence



Deep Learning Renaissance



More data, more …

- In 2009 the Imagenet dataset was published [Deng et al., 2009]
  - Collected images for each term of Wordnet (100,000 classes)
  - Tree of concepts organized hierarchically
    - “Ambulance”, “Dalmatian dog”, “Egyptian cat”, ...
  - About 16 million images annotated by humans
- Imagenet Large Scale Visual Recognition Challenge (ILSVRC)
  - 1 million images
  - 1,000 classes
  - Top-5 and top-1 error measured

Alexnet

- In 2013 Krizhevsky, Sutskever and Hinton re-implemented [Krizhevsky2013] a convolutional neural network [LeCun1998]
  - Trained on Imagenet, Two GPUs were used for the implementation
- Further theoretical improvements
  - Rectified Linear Units (ReLU) instead of sigmoid or tanh
  - Dropout
  - Data augmentation
- In the 2013 Imagenet Workshop a legendary turmoil
  - Blasted competitors by an impressive 16% top-5 error, Second best around 26%
  - Most didn’t even think of NN as remotely competitive
- At the same time similar results in the speech recognition community
  - One of G. Hinton students collaboration with Microsoft Research, improving state-of-the-art by an impressive amount after years of incremental improvements [Hinton2012]

Alexnet architecture

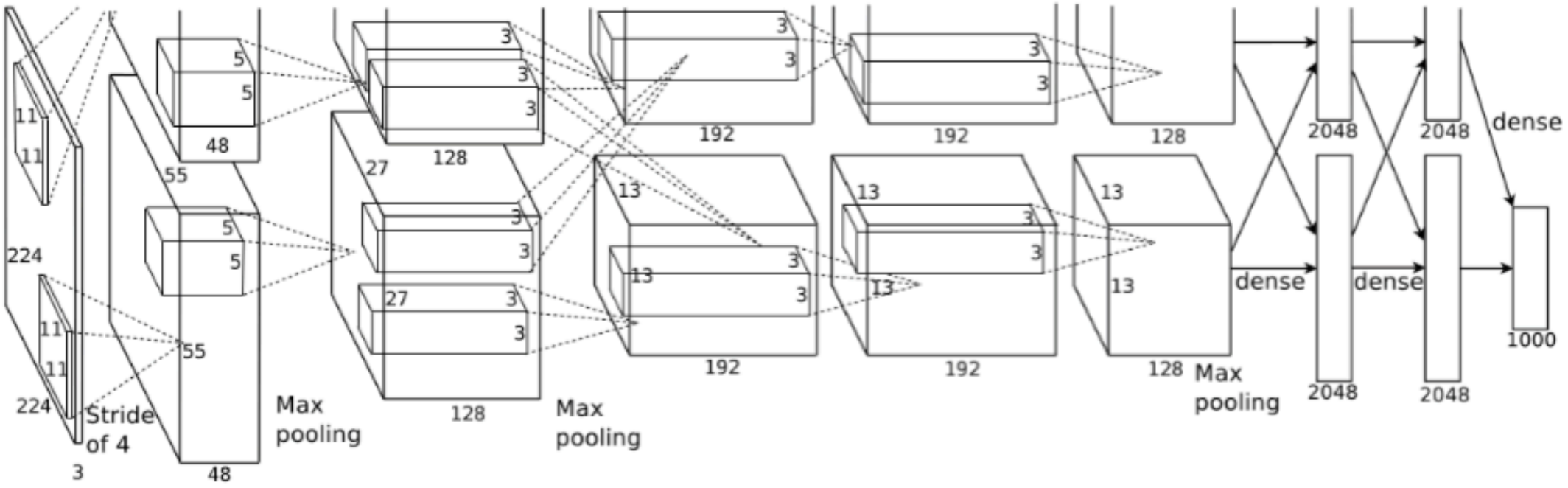
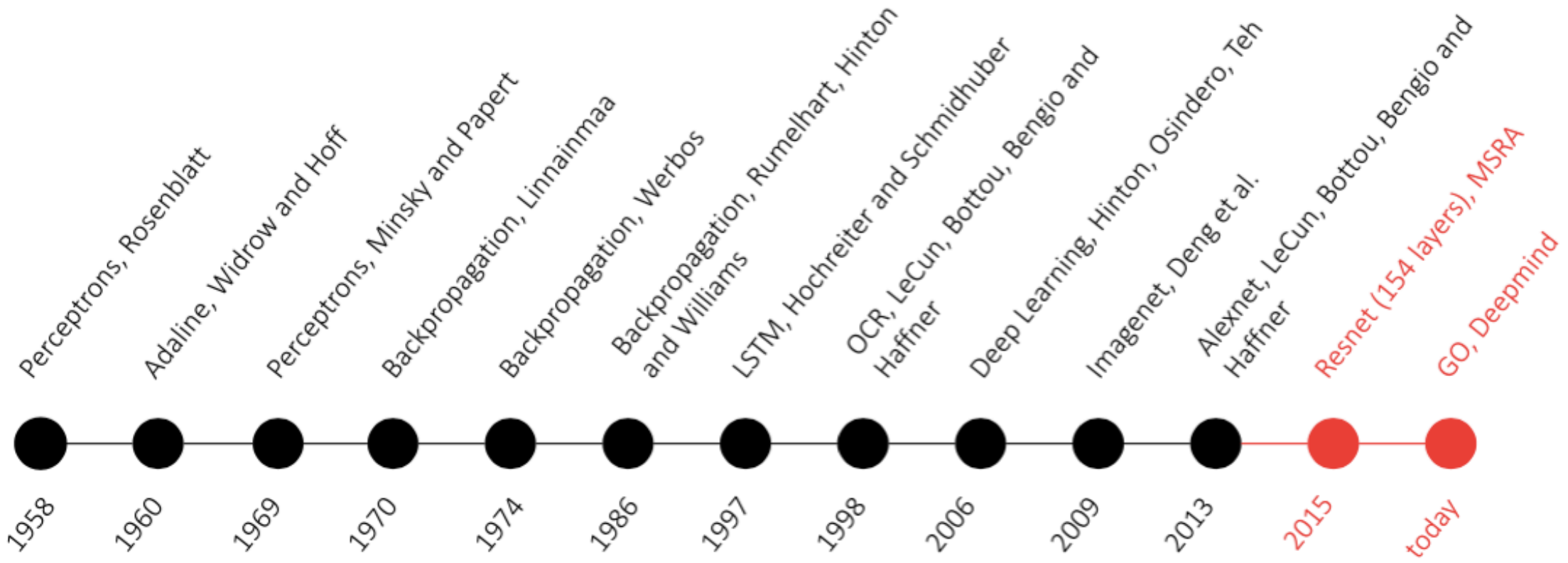


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input 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.

Deep Learning Golden Era



The today

- Deep Learning is almost everywhere
  - Object classification
  - Object detection, segmentation, pose estimation
  - Image captioning, question answering
  - Machine translation
  - Speech recognition
  - Robotics
- Some strongholds
  - Action classification, action detection
  - Object retrieval
  - Object tracking

The ILSVC Challenge

CNN based, non-CNN based

2012 Teams	%error
Supervision (Toronto)	15.3
ISI (Tokyo)	26.1
VGG (Oxford)	26.9
XRCE/INRIA	27.0
UvA (Amsterdam)	29.6
INRIA/LEAR	33.4

2013 Teams	%error
Clarifai (NYU spinoff)	11.7
NUS (singapore)	12.9
Zeiler-Fergus (NYU)	13.5
A. Howard	13.5
OverFeat (NYU)	14.1
UvA (Amsterdam)	14.2
Adobe	15.2
VGG (Oxford)	15.2
VGG (Oxford)	23.0

2014 Teams	%error
GoogLeNet	6.6
VGG (Oxford)	7.3
MSRA	8.0
A. Howard	8.1
DeeperVision	9.5
NUS-BST	9.7
TTIC-ECP	10.2
XYZ	11.2
UvA	12.1

Figures taken from Y. LeCun's CVPR 2015 plenary talk

2015 ILSVRC Challenge

- Microsoft Research Asia won the competition with a legendary 150-layered network
- Almost superhuman accuracy: 3.5% error
  - In 2016 <3% error
- In comparison in 2014 GoogLeNet had 22 layers

So, why now?

- Better hardware
- Bigger data
- Better regularization methods, such as dropout
- Better optimization methods, such as Adam, batch normalization