

An (incomplete) History of Deep Learning

1943 m1 Thresholded Logic Unit, W McCulloch, W Pitts
(U. Illinois)

- a mathematical model of the human brain
- Paper: "A logical calculus of the ideas immanent in nervous activity", Bulletin of Mathematical Biophysics
- "activity of the neuron is an all-or-none process"

1958 m2 Perceptron, F Rosenblatt (Cornell)

- Paper: "The perceptron: A probabilistic model for information storage and organization in the brain", Psychological Review
- Proposed a 3 "layer" linear classifier
 - One layer contained learnable parameters
 - 0-1 threshold function
- Mark 1 Perceptron is at Smithsonian
- "the embryo of an electronic computer that it [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence" (NYT, 1958)

1965 **M3** Title: "Cybernetic Predicting Devices",

→ A.G. Ivakhnenko, V. G. Lapa (Kyiv Polytechnic Instit, USSR)

→ multilayer network with non-linear activation functions (polynomial) and statistical learning rule (not backprop.)

→ "Cybernetics and forecasting techniques", 1967, Ivakhnenko, Lapa

→ "The Group Method of Data Handling - A Rival Method of Stochastic Approximation", 1968, Ivakhnenko

1967 **B1** Introduce the first multi-layer perceptron

trained with stochastic gradient descent

Shun-Ichi Amari, M. Saito (student) (Kyushu University)

Saito's masters thesis: "Implementation of Amari's

1967 stochastic gradient descent method for multilayer perceptrons"

1969 **M4** Title: "Perceptrons"

→ M Minsky (MIT), S Papert (MIT)

→ argued against the utility of single-layer perceptrons show that, for example, these models could not learn the XOR function

XOR function :

x_1	x_2	Y
0	0	1
0	1	1
1	0	1
1	1	0

[1970] [B2] Seppo Linnainmaa introduced the backpropagation algorithm in his masters thesis "The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors" (U. Helsinki)

[1970s] First AI Winter

[1979] [M5] "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position", K. Fukushima (NHK Labs)
→ first introduced convolutional neural networks with pooling layers

[1982] [B3] "Applications of advances in nonlinear sensitivity analysis", P.J. Werbos (U.S. DoE)
→ applied backpropagation to neural networks

1986 **B4** "Learning representations by backpropagating errors", D. Rumelhart, G. Hinton, R. Williams (Toronto)

→ applied backpropagation to neural networks

1989 **M6** Universal Approximation Theorem for neural networks

→ a multi-layer perceptron with one hidden layer can approximate any bounded continuous function with arbitrary accuracy

→ with two hidden layers, an MLP can approximate any function

→ "Approximation by Superpositions of a Sigmoidal Function",
G. Cybenko (U Illinois)

→ "Multilayer Feedforward Networks are Universal Approximators", Hornik (Techn. Univ. Wein),
Stinchcombe (UCSD), White (UCSD)

1989 **B5** "Backpropagation Applied to Handwritten Zip Code Recognition", Y. LeCun, et al. (Bell Labs)

→ First practical application of convolutional neural networks trained with backpropagation for digit recognition

[1990s] 2nd AI Winter

[1994] [G1] "GPU" is coined by Sony to refer to the graphics card developed for its PlayStation console

[1995] [M7] "Support Vector Networks", C. Cortes, V. Vapnik (Bell Labs)
→ introduced support vector machines

[1997] [M8] "Long short-term memory", S. Hochreiter (Tech. Uni. Munich), J. Schmidhuber (IDSIA)
→ introduced LSTM architecture for recurrent neural networks

[2004] [G2] "GPU implementation of neural networks" K. Oh, K. Jung (Soongsil University)
→ parallelized matrix operation in multilayer perceptron using GPU


[2006] [G3] "High Performance convolutional neural networks for document processing" K. Chellapilla, S. Puri, P. Smard (Microsoft Research)
→ implemented GPU speedup of CNN architecture


[2007] [G4] NVIDIA's CUDA API for GPU computing released


[2009] [G5] "Large-scale deep unsupervised learning using (stanford) graphics processors" R. Raina, A. Madhavan, A. Ng
→ 100M parameter model trained in 1 day using GPU

- 2009** ☐ ImageNet (ILSVRC 2010) dataset is released
- Fei Fei Li (at Princeton at the time, later Stanford)
 - magnet classification contest begins
 - ImageNet-1K (1K classes, 1.2M train, 50K val, 100K test)

2010-2017 ImageNet Large Scale Visual Recognition Challenge
Top-5 Classification Error

- 2010 : 28.2% SVM
- 2011 : 25.2% SVM
-  → 2012 : 16.4% AlexNet (CNN)
- 2013 : 11.7% Cifarai (CNN)
- 2014 : 6.66% GoogLeNet (CNN)
- 2015 : 3.57% ResNet (CNN)
- ⋮
- 2017 : <5% for 29/38 competing teams

2012  "ImageNet Classification with Deep Convolutional Neural Networks" A. Krizhevsky, I. Sutskever, G. Hinton (U. Toronto)

2017  "Attention is all you need" A. Vaswani et al. (Google Brain)

2019 ☐ Turing Award : LeCun, Bengio, Hinton