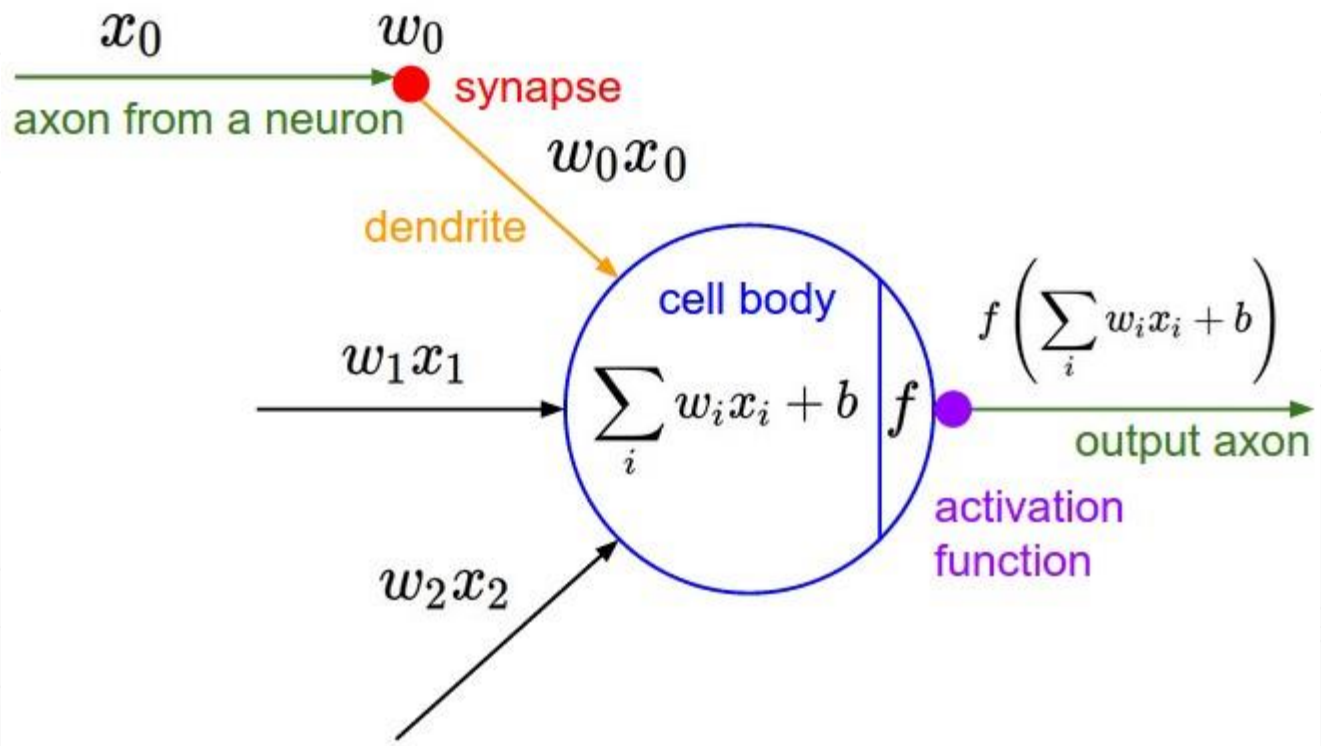
Perceptron

PITTS WITH LETTVIN: Pitts with Jerome Lettvin and one subject of their experiments on visual perception (1959).

Wikipedia



http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

Perceptron: 1958

Psychological Review
Vol. 65, No. 6, 1958

THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN[1]

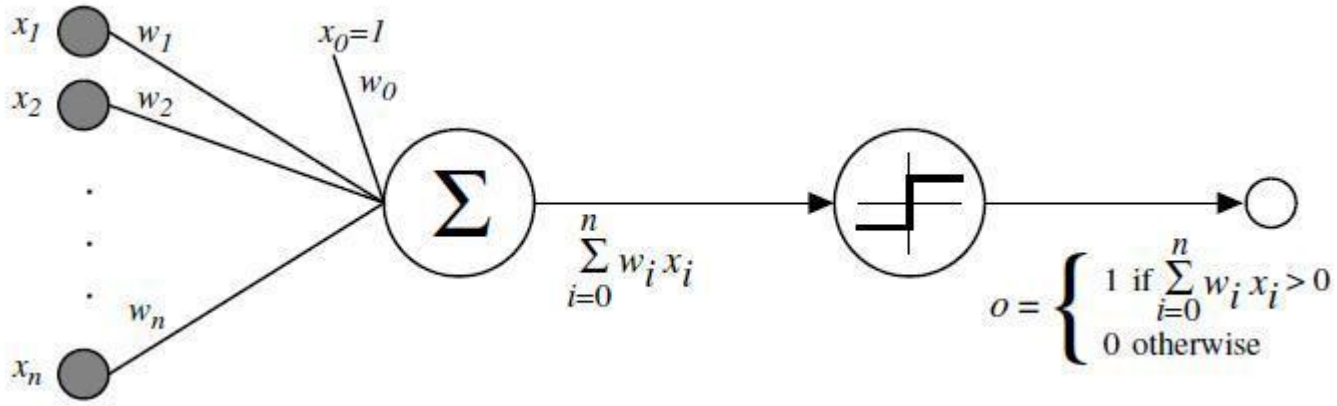
F. ROSENBLATT

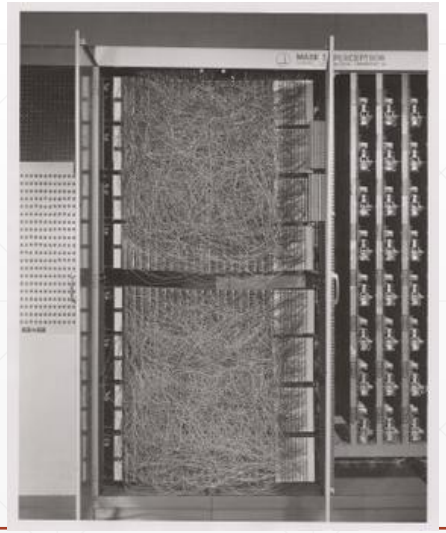
Cornell Aeronautical Laboratory

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.335.3398&rep=rep1&type=pdf

Perceptron

1. Start off with a Perceptron having random weights and a training set
2. For the inputs of an example in the training set, compute the Perceptron's output
3. If the output of the Perceptron does not match the output that is known to be correct for the example: * If the output should have been 0 but was 1, decrease the weights that had an input of 1. * If the output should have been 1 but was 0, increase the weights that had an input of 1.
4. Go to the next example in the training set and repeat steps 2-4 until the Perceptron makes no more mistakes

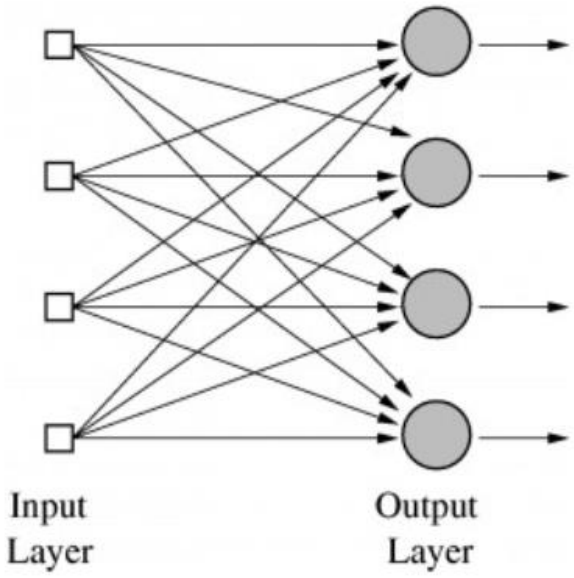




http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

Multi-Output Perceptron

- Extend to cope with multi categories

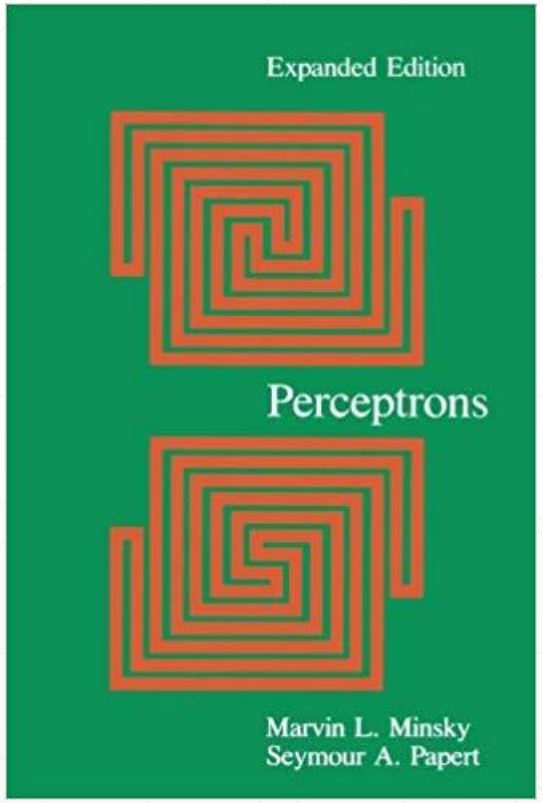


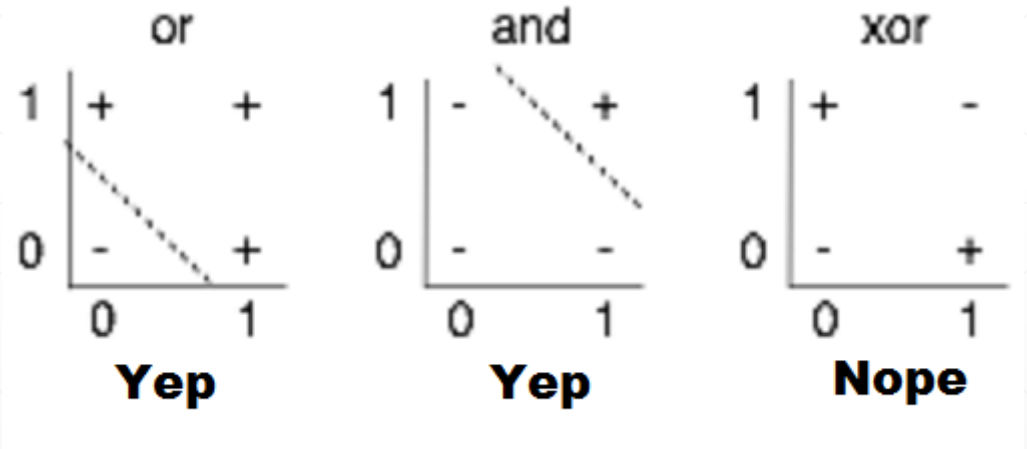
A neural net with multiple outputs.

The stuff promised in this video - still not really around.

https://youtu.be/aygSMgK3BEM

Perceptrons' Limitation: 1969





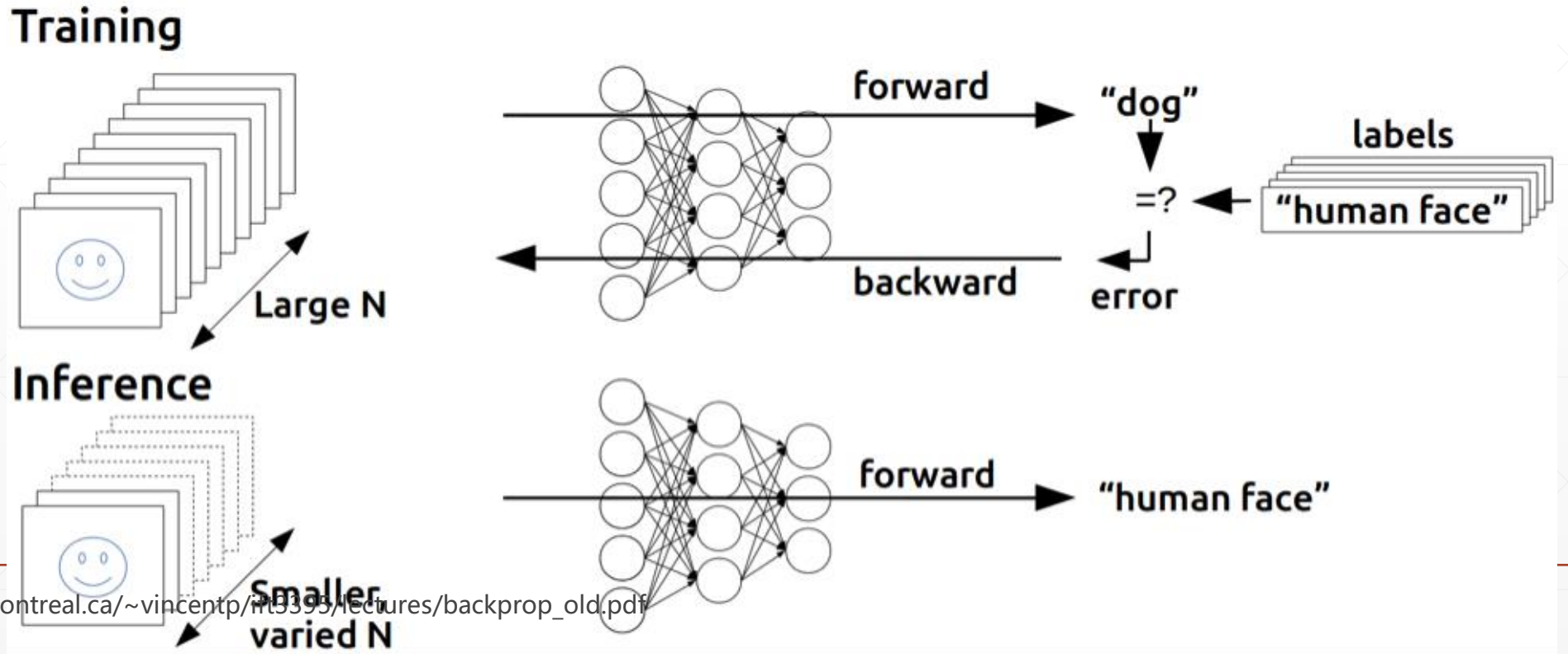
http://science.sciencemag.org/content/165/3895/780

Is it Winter?

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

NO! Multi-Layer Perceptron is coming

- New Issue: How to train MLP
- Chain Rules => Backpropagation



http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

Backpropagation: First Spark

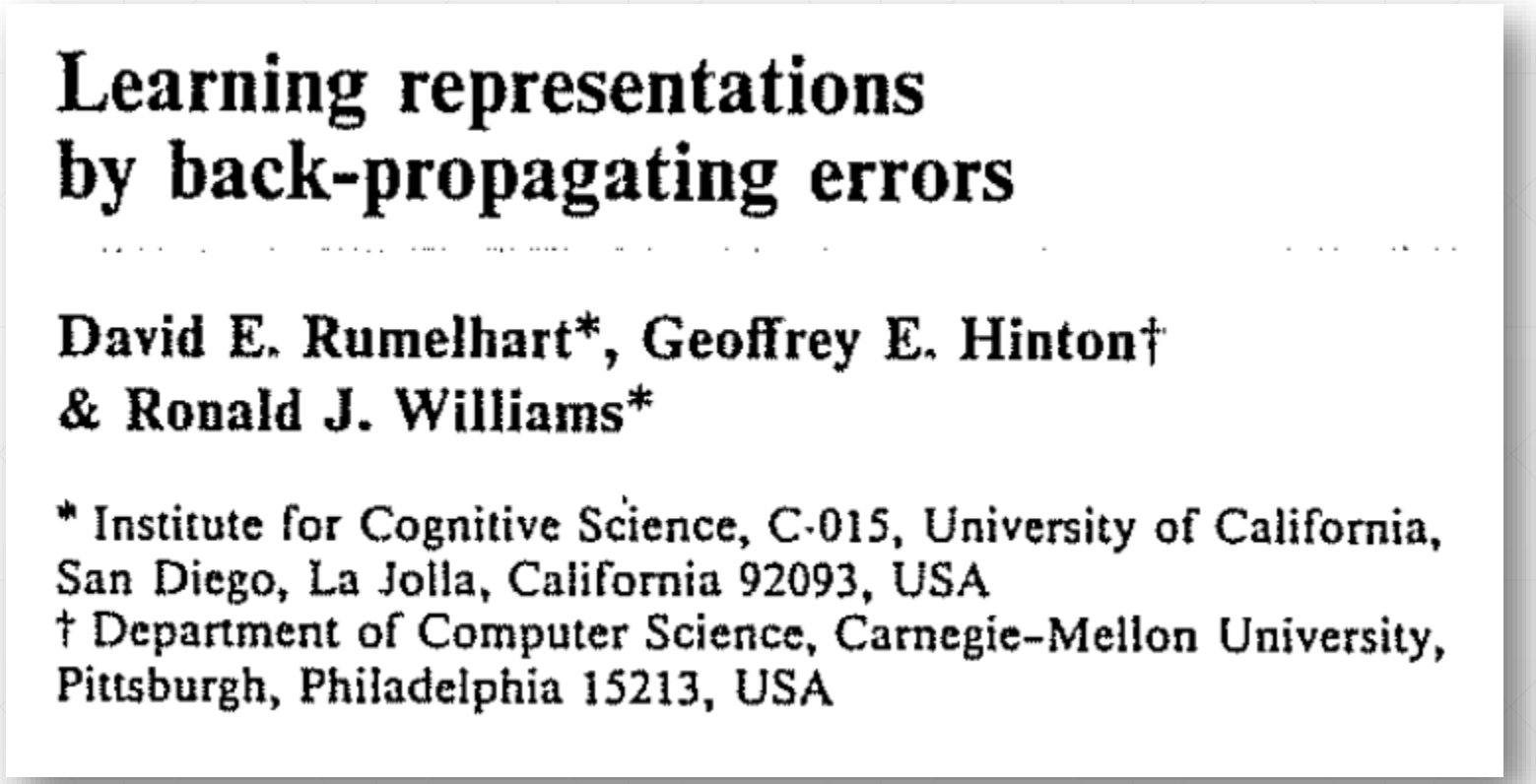
- Derived in early 60’ s
- Run on computer by Seppo Linnainma in 1970
- Introduced to Neural Networks by Paul Werbos in 1974 PhD Thesis
- Published until 1982 due to AI Winter

> *in fact visit Minsky at MIT. I proposed that we do a joint paper showing that MLPs can in fact overcome the earlier problems ... But Minsky was not interested(14). In fact, no one at MIT or Harvard or any place I could find was interested at the time.”*

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

Rediscover! 1986 on Nature

- David Parker and Yann LeCun mentioned.



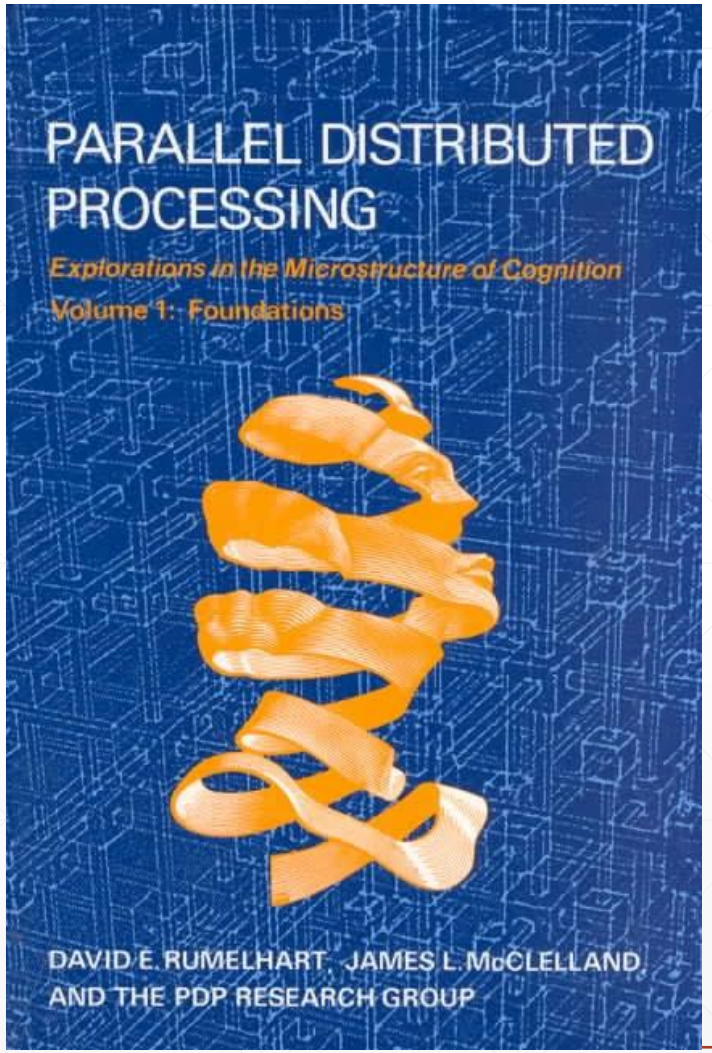
Learning representations
by back-propagating errors

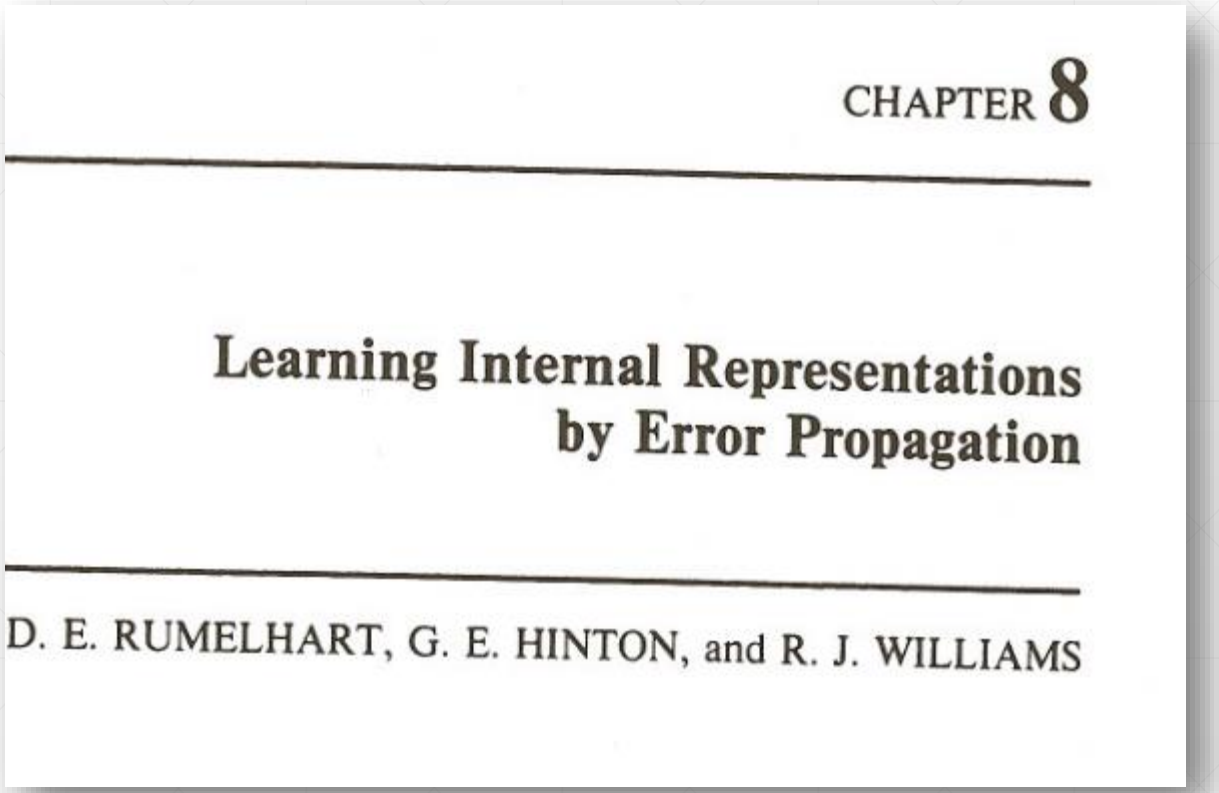
David E. Rumelhart*, Geoffrey E. Hinton†
& Ronald J. Williams*

* Institute for Cognitive Science, C-015, University of California, San Diego, La Jolla, California 92093, USA
† Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Philadelphia 15213, USA

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

1986





CHAPTER 8

Learning Internal Representations by Error Propagation

D. E. RUMELHART, G. E. HINTON, and R. J. WILLIAMS

https://web.stanford.edu/class/psych209a/ReadingsByDate/02_06/PDPVolIChapter8.pdf

Multilayer Feedforward Networks are Universal Approximators

KUR' HORNIK

Technische Universität Wien

MAXWELL STINCHCOMBE AND HALBERT WHITE

University of California, San Diego

(*Received* 16 *September* 1988; *revised and accepted* 9 *March* 1989)

https://www.cs.cmu.edu/~epxing/Class/10715/reading/Kornick_et_al.pdf

CPU?

1985

- National Semiconductor debuts the 32000 32-bit processor. [513]
- In England, Advanced RISC Machines ships a 32-bit ARM processor. Its first application is in an accelerator card for Acorn PCs. [32] [599.15]
- Intel introduces the 80287 math coprocessor. [511.309] (1980 [1064.26])
- Motorola unveils its 68008 CPU chip. [120]
- Intel and IBM sign an agreement allowing IBM to manufacture its own x86 processors and design derivative products, but not to sell them directly on the open market. [979]
- Sun Microsystems begins work on its SPARC processor. [160]
- Intel begins circulating prototype chipsets of the 386 processor. [606.77]

October 16

- Intel introduces the 16 MHz 80386DX microprocessor. It uses 32-bit registers and a 32-bit (16 MHz) data bus, and incorporates 275,000 transistors (1.5 micron width). Initial price is US$299. It can access 4 gigabytes of physical memory, or up to 64 terabytes of virtual memory. Intel spent US$100 million in development costs. [41] [75] [176.74] [177.102] [296] [347.61] [477.125] [540.64] [62] [690.94] [879.116] [900] [940.106] [947.102] [1389.D4] [1635.52] [1897.128]

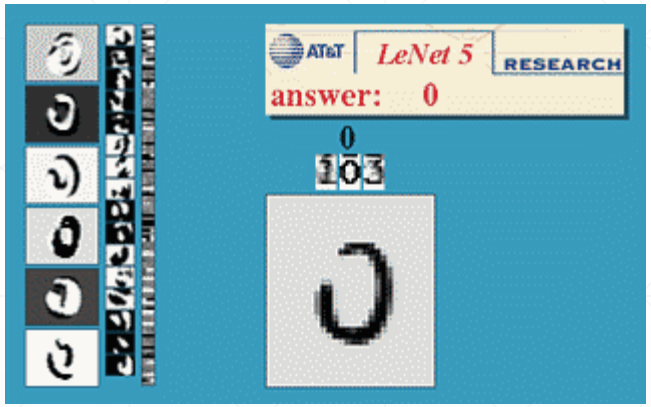
http://processortimeline.info/proc1980.htm

Now theoretically solved:1989

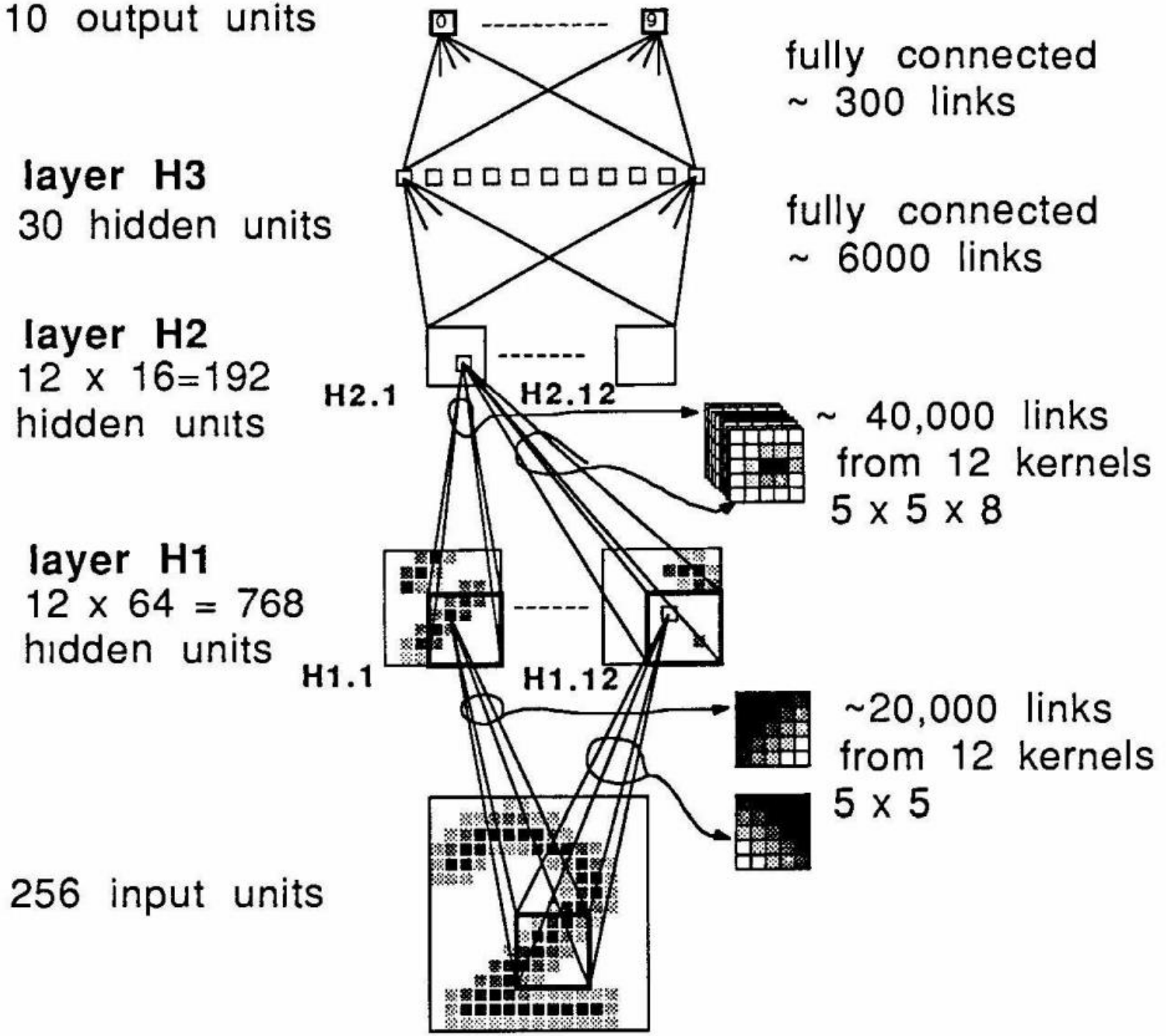
Communicated by Dana Ballard

Backpropagation Applied to Handwritten Zip Code Recognition

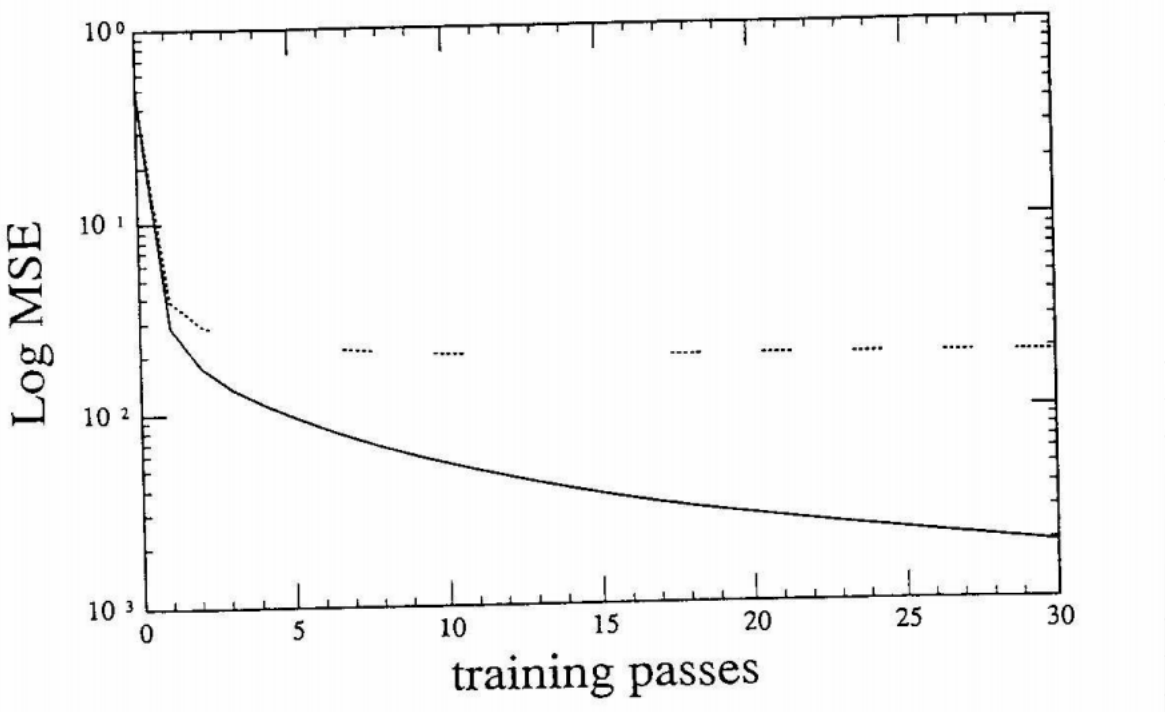
Y. LeCun
B. Boser
J. S. Denker
D. Henderson
R. E. Howard
W. Hubbard
L. D. Jackel
AT&T Bell Laboratories Holmdel, NJ 07733 USA

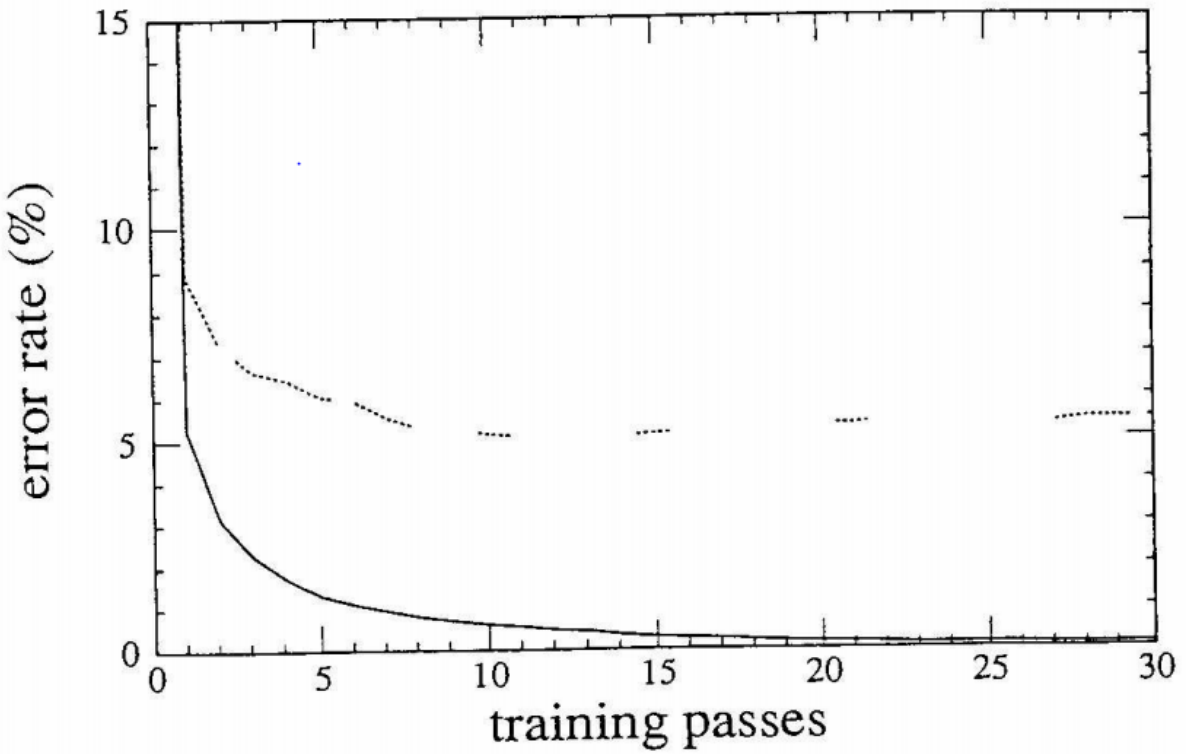


http://yann.lecun.com/exdb/publis/pdf/lecun-89e.pdf


10 output units
0
9
fully connected
~ 300 links
layer H3
30 hidden units
fully connected
~ 6000 links
layer H2
12 x 16=192
hidden units
H2.1
H2.12
~ 40,000 links
from 12 kernels
5 x 5 x 8
layer H1
12 x 64 = 768
hidden units
H1.1
H1.12
~20,000 links
from 12 kernels
5 x 5
256 input units

9760 independent parameters





http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

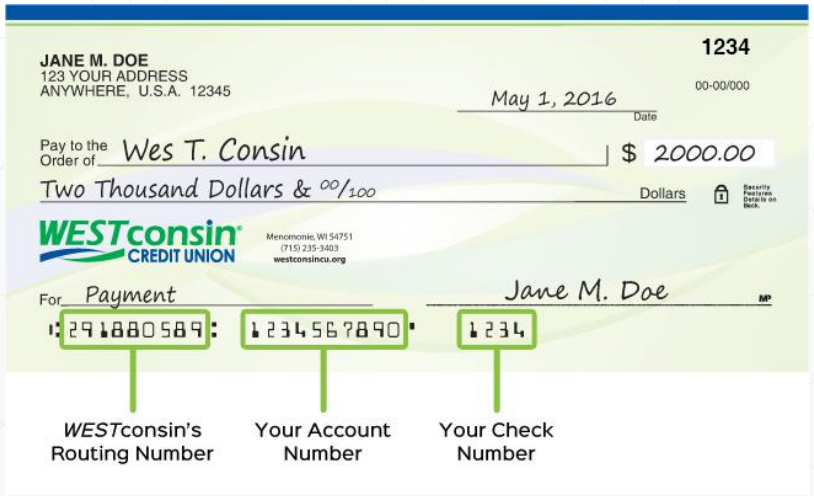
On DSP

5.2 DSP Implementation. During the recognition process, almost all the computation time is spent performing multiply accumulate operations, a task that digital signal processors (DSP) are specifically designed for We used an off-the-shelf board that contains 256 kbytes of local memory and an AT&T DSP-32C general purpose DSP with a peak performance of 12 5 million multiply add operations per second on 32 bit floating point numbers (25 MFLOPS). The DSP operates as a coprocessor; the host is a personal computer (PC), which also contains a video acquisition board connected to a camera.

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

- https://youtu.be/FwFduRA_L6Q

- At some point in the late 1990s, one of these systems was reading 10 to 20% of all the checks in the US.



http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

Another Learning pattern in 1986

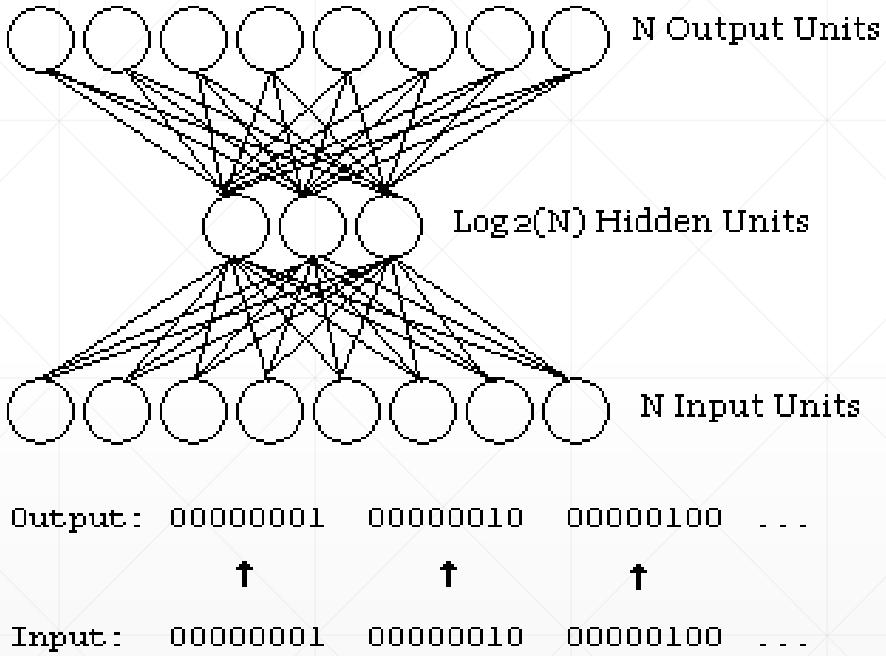
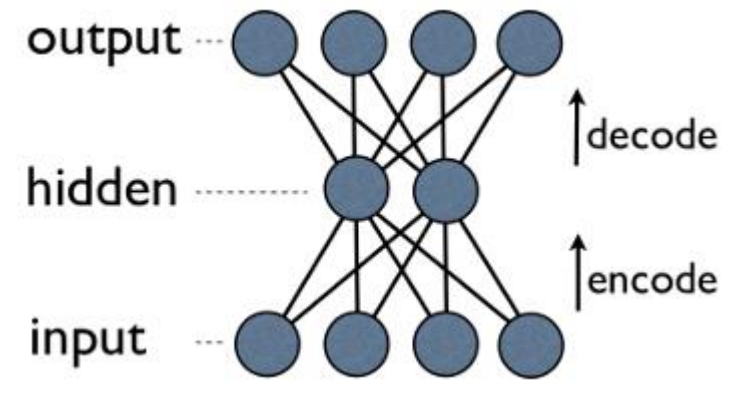


Figure 1. The encoding problem (Rumelhart, Hinton, & Williams, 1986)



probability distributions

http://www.cs.toronto.edu/~fritz/absps/cvq.pdf

Meanwhile: Speech Sequence

- No Memory
- Time delay NN

328 IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

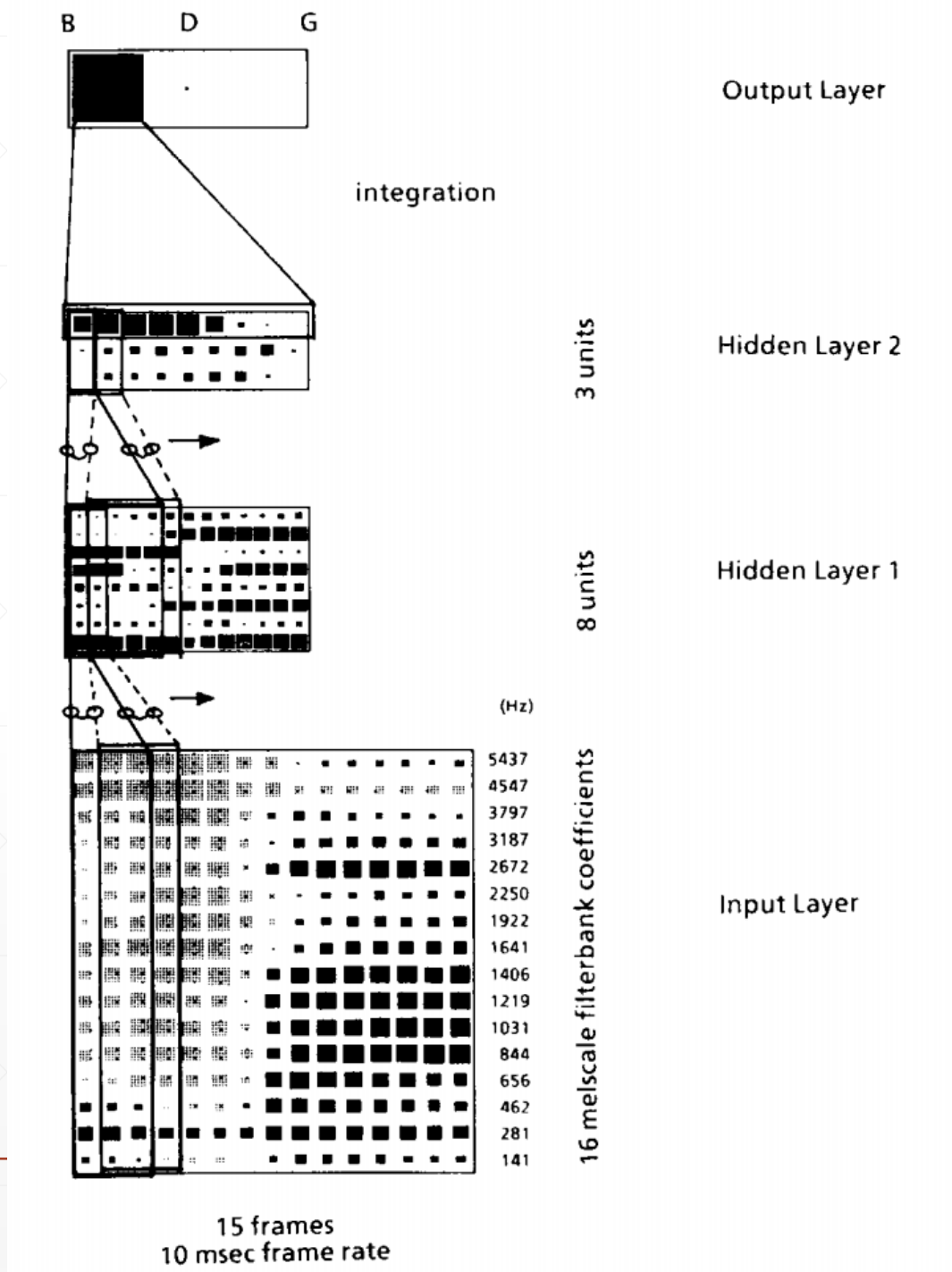
Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON,
KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

http://www.cs.toronto.edu/~fritz/absps/waibelTDNN.pdf

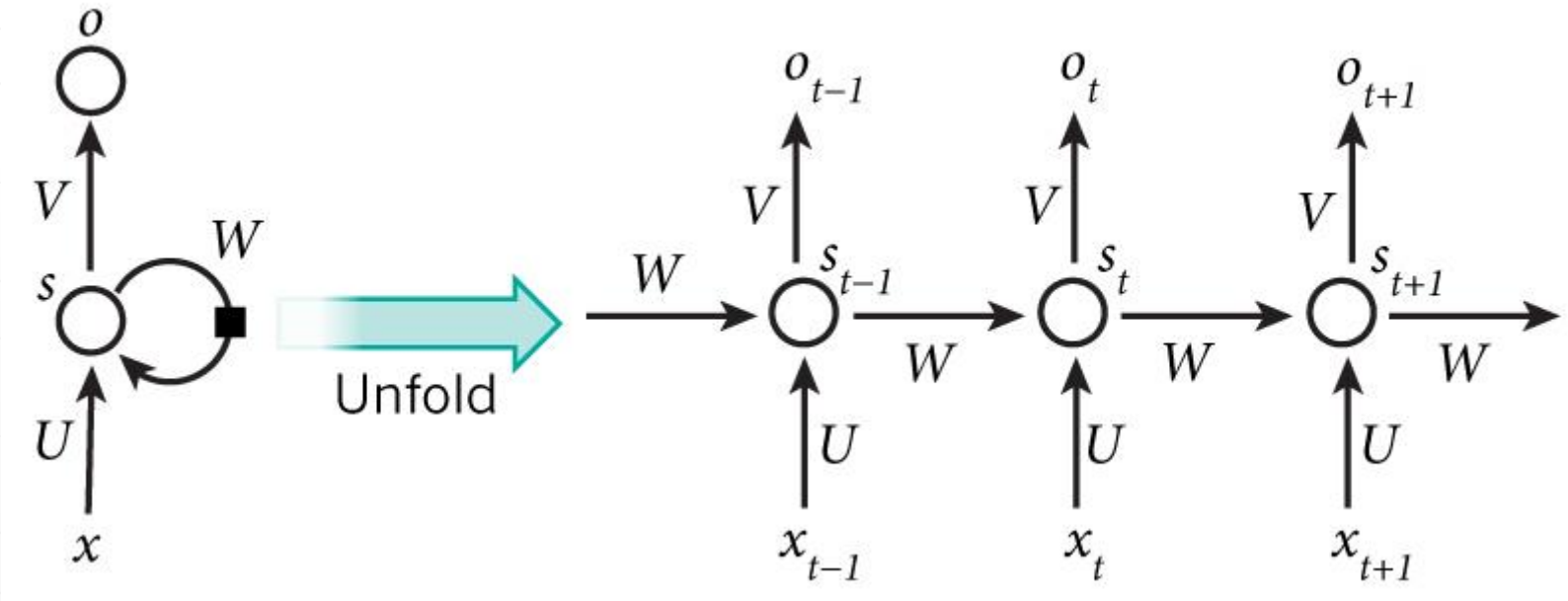
Moving window

- Inspired LeCun



Recurrent Neural Network

- Spatial Local
- Temporal Local



http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

Yoshua Bengio:1993

A Connectionist Approach to Speech Recognition

Yoshua Bengio

AT&T Bell Laboratories,

HO-4G312, Crawfords Corner Rd, Holmdel, NJ 07733.

http://www.iro.umontreal.ca/~lisa/publications2/index.php/attachments/single/161

Yann LeCun & Yoshua Bengio

Convolutional Networks for Images, Speech, and

Time-Series

Yann LeCun

Rm 4G332, AT&T Bell Laboratories

101 Crawfords Corner Road

Yoshua Bengio

Dept. Informatique et Recherche

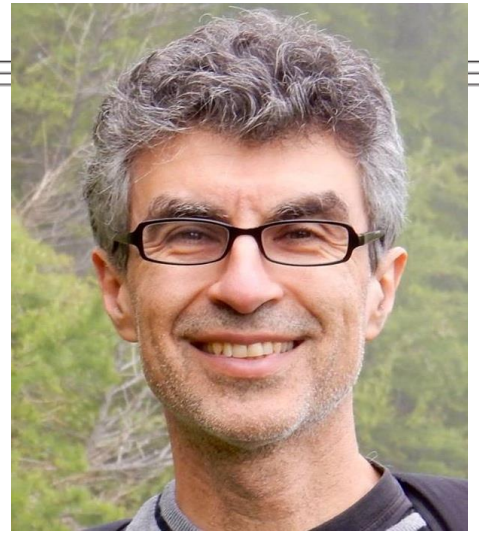
Opérationnelle, Université de Montréal,

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

REVIEW

Deep learning

Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]


Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

https://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf

LSTM: 1997

- Long memory

LONG SHORT-TERM MEMORY

NEURAL COMPUTATION 9(8):1735–1780, 1997

Sepp Hochreiter
Fakultät für Informatik
Technische Universität München
80290 München, Germany
hochreit@informatik.tu-muenchen.de
http://www7.informatik.tu-muenchen.de/~hochreit

Jürgen Schmidhuber
IDSIA
Corso Elvezia 36
6900 Lugano, Switzerland
juergen@idsia.ch
http://www.idsia.ch/~juergen

https://github.com/dzitkowskik/StockPredictionRNN/blob/master/docs/Hochreiter97_lstm.pdf

Method	Delay p	Learning rate	# weights	% Successful trials	Success after
RTRL	4	1.0	36	78	1,043,000
RTRL	4	4.0	36	56	892,000
RTRL	4	10.0	36	22	254,000
RTRL	10	1.0-10.0	144	0	$>$ 5,000,000
RTRL	100	1.0-10.0	10404	0	$>$ 5,000,000
BPTT	100	1.0-10.0	10404	0	$>$ 5,000,000
CH	100	1.0	10506	33	32,400
LSTM	100	1.0	10504	100	5,040

Table 2: *Task 2a: Percentage of successful trials and number of training sequences until success, for "Real-Time Recurrent Learning" (RTRL), "Back-Propagation Through Time" (BPTT), neural sequence chunking (CH), and the new method (LSTM). Table entries refer to means of 18 trials. With 100 time step delays, only CH and LSTM achieve successful trials. Even when we ignore the unsuccessful trials of the other approaches, LSTM learns much faster.*

Predictability Minimization V.S. GAN

https://zhuanlan.zhihu.com/p/27159510

Can deal with simplest scenario

Why

> *"A diploma thesis (Hochreiter, 1991) represented a milestone of explicit DL research. As mentioned in Sec. 5.6, by the late 1980s, experiments had indicated that traditional deep feedforward or recurrent networks are hard to train by backpropagation (BP) (Sec. 5.5). Hochreiter's work formally identified a major reason: Typical deep NNs suffer from the now famous problem of vanishing or exploding gradients. With standard activation functions (Sec. 1), cumulative backpropagated error signals (Sec. 5.5.1) either shrink rapidly, or grow out of bounds. In fact, they decay exponentially in the number of layers or CAP depth (Sec. 3), or they explode. "*

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

New Star: Support Vector Machine: 1992

A Training Algorithm for Optimal Margin Classifiers

Bernhard E. Boser*
EECS Department
University of California
Berkeley, CA 94720
boser@eecs.berkeley.edu

Isabelle M. Guyon
AT&T Bell Laboratories
50 Fremont Street, 6th Floor
San Francisco, CA 94105
isabelle@neural.att.com

Vladimir N. Vapnik
AT&T Bell Laboratories
Crawford Corner Road
Holmdel, NJ 07733
vlad@neural.att.com

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

Dark time

- Paper got rejected
- Hinton moved to CIFAR seeking for funding
- Conspiracy: rebrand “neural network” as “deep learning”

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

Rekindle: 2006

- Weights initialized by pretraining
  - train each layer one by one with unsupervised training

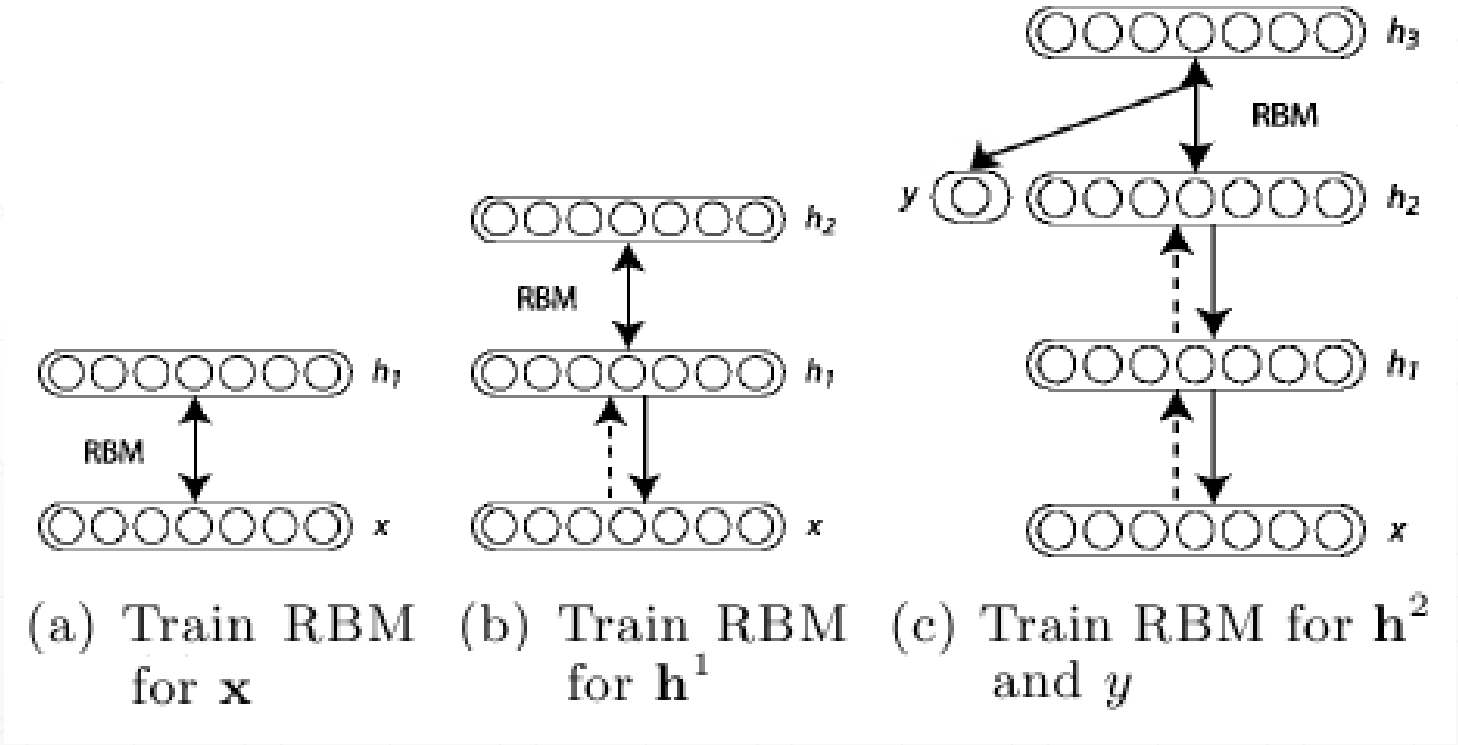
A fast learning algorithm for deep belief nets *

Geoffrey E. Hinton and **Simon Osindero**
Department of Computer Science University of Toronto
10 Kings College Road
Toronto, Canada M5S 3G4
{hinton, osindero}@cs.toronto.edu

Yee-Whye Teh
Department of Computer Science
National University of Singapore
3 Science Drive 3, Singapore, 117543
tehyw@comp.nus.edu.sg

https://www.cs.toronto.edu/~hinton/absps/fastnc.pdf

1.4% => 1.25%



(a) Train RBM for $\mathbf{x}$ (b) Train RBM for $\mathbf{h}^1$ (c) Train RBM for $\mathbf{h}^2$ and y

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

Deep is more efficient: representation learning

Greedy Layer-Wise Training of Deep Networks

Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle
Université de Montréal
Montréal, Québec
{bengioy,lamblinp,popovicd,larocheh}@iro.umontreal.ca

http://papers.nips.cc/paper/3048-greedy-layer-wise-training-of-deep-networks.pdf

Another Hero: NVIDIA

> *"Inspired by one of Hinton's lectures on deep neural networks, Mohamed began applying them to speech - but deep neural networks required too much computing power for conventional computers – so Hinton and Mohamed enlisted Dahl. A student in Hinton's lab, Dahl had discovered how to train and simulate neural networks efficiently using the same high-end graphics cards which make vivid computer games feasible on personal computers.*

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

NVIDIA GPU Timeline

- https://en.wikipedia.org/wiki/GeForce

2009

Large-scale Deep Unsupervised Learning using Graphics Processors

Rajat Raina RAJATR@CS.STANFORD.EDU
Anand Madhavan MANAND@STANFORD.EDU
Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department, Stanford University, Stanford CA 94305 USA

http://www.machinelearning.org/archive/icml2009/papers/218.pdf

Table 2. Average running time in seconds for processing 1 million input examples for learning an RBM, with contrastive divergence updates applied in batches of 192 examples each. The size of the RBM in each column is denoted by the number of visible units × number of hidden units. The GPU speedup is computed w.r.t. the fastest CPU-based result.

Package	Architecture	576x1024	1024x4096	2304x16000	4096x11008
Goto BLAS	Single CPU	563s	3638s	172803s	223741s
Goto BLAS	Dual-core CPU	497s	2987s	93586s	125381s
GPU		38.6s	184s	1376s	1726s
GPU Speedup		**12.9x**	**16.2x**	**68.0x**	**72.6x**

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

MNIST on GPU

Deep Big Simple Neural Nets Excel on Hand-written Digit Recognition

Dan Claudiu Cireşan[1, 2],
Ueli Meier[1, 2],
Luca Maria Gambardella[1, 2],
Jürgen Schmidhuber[1, 2]

[1]IDSIA, Galleria 2, 6928 Manno-Lugano, Switzerland.

[2]University of Lugano & SUPSI, Switzerland.

Keywords: NN (Neural Network) , MLP (Multilayer Perceptron), GPU (Graphics Processing Unit), training set deformations, MNIST [1], BP (back-propagation).

Abstract

Good old on-line back-propagation for plain multi-layer perceptrons yields a very low 0.35% error rate on the famous MNIST handwritten digits benchmark. All we need to achieve this best result so far are many hidden layers, many neurons per layer, numerous deformed training images, and graphics cards to greatly speed up learning.

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

BOOM! 2012

- 60,000,000 parameters
- 5 conv layers
- 26.2% -> 15.3%, top-5 error rate

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

https://papers.nips.cc/paper/4824-i
neural-networks.pdf

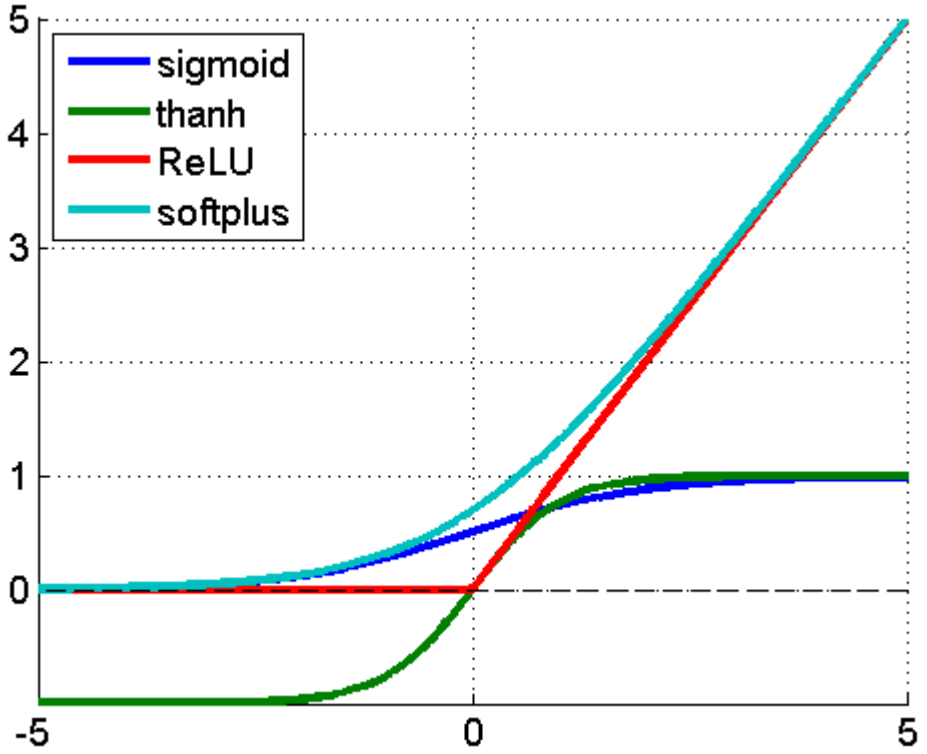
Deep Neural Networks for Acoustic Modeling in Speech Recognition

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

Other Heroes

- Big Data
- ReLU
- BatchNorm
- Xavier Initialization
- Kaiming Initialization
- Dropout



http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

GAN:2014

Generative Adversarial Nets

Ian J. Goodfellow,* **Jean Pouget-Abadie**,† **Mehdi Mirza**, **Bing Xu**, **David Warde-Farley**,
Sherjil Ozair,‡ **Aaron Courville**, **Yoshua Bengio**§
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

BigGAN

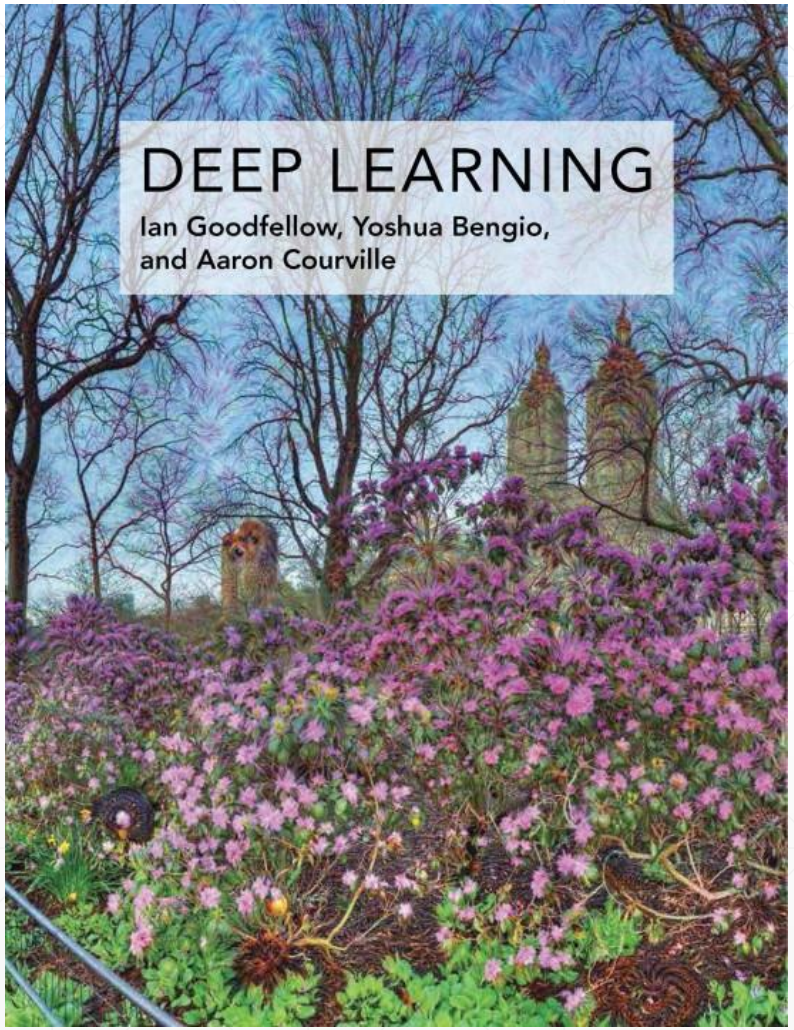
1 INTRODUCTION

Figure 1: Class-conditional samples generated by our model.

https://arxiv.org/pdf/1809.11096.pdf

Ian Goodfellow

- How I fail



https://veronikach.com/how-i-fail/how-i-fail-ian-goodfellow-phd14-computer-science/

Yann LeCun, Director of AI Research at Facebook and Professor at NYU

Answered Jul 29, 2016 · Upvoted by Joaquin Quiñonero Candela, studied Machine Learning and Gokul Krishnan, M.Sc Computer Science & Machine Learning, ETH Zurich (2018)

There are many interesting recent development in deep learning, probably too many for me to describe them all here. But there are a few ideas that caught my attention enough for me to get personally involved in research projects.

The most important one, in my opinion, is adversarial training (also called GAN for Generative Adversarial Networks). This is an idea that was originally proposed by Ian Goodfellow when he was a student with Yoshua Bengio at the University of Montreal (he since moved to Google Brain and recently to OpenAI).

This, and the variations that are now being proposed is the most interesting idea in the last 10 years in ML, in my opinion.

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

LETTER

doi:10.1038/nature14236

Human-level control through deep reinforcement learning

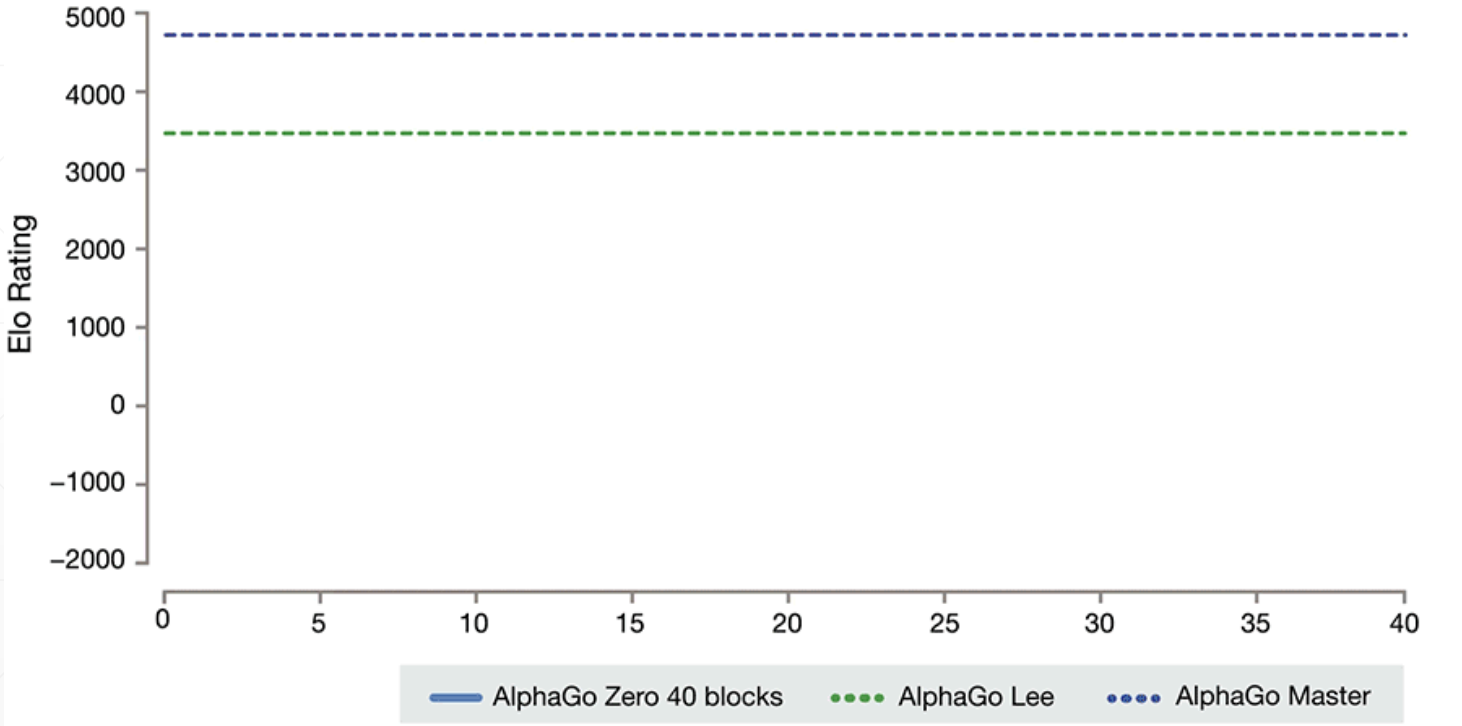
Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

- 2015

strate that the deep Q-network agent, receiving only the pixels and the game score as inputs, was able to surpass the performance of all previous algorithms and achieve a level comparable to that of a professional human games tester across a set of 49 games, using the same algorithm, network architecture and hyperparameters. This work bridges the divide between high-dimensional sensory inputs and actions, resulting in the first artificial agent that is capable of learning to excel at a diverse array of challenging tasks.

https://storage.googleapis.com/deepmind-media/dqn/DQNNaturePaper.pdf

AlphaZero



http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

OpenAI Five

- https://blog.openai.com/openai-five/
- https://youtu.be/UZHTNBMAfAA

Baidu Apollo

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

So many to list

- … …

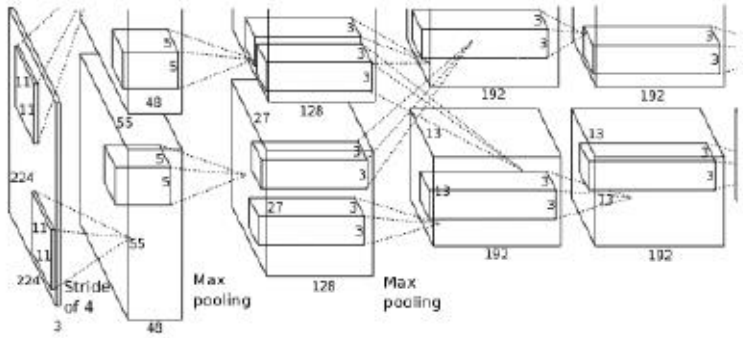
http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

Deep Learning

The Deep Learning "Computer Vision Recipe"

Big Data: ImageNet

+



Deep Convolutional Neural Network

+

Backprop on GPU

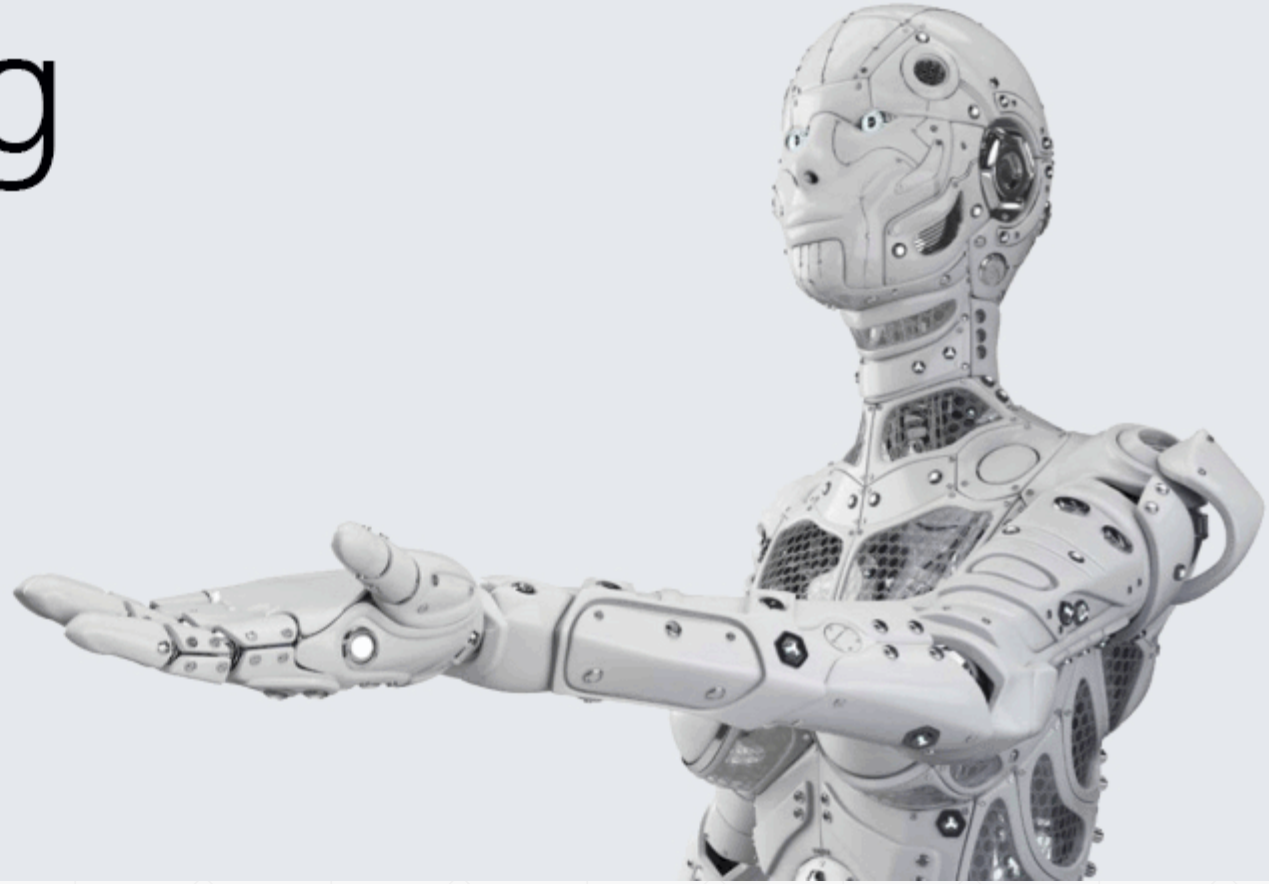
=

Learned Weights

http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

AI's coming of age

The progress into the AGI phase and the beginning of true autonomy.



http://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf

Thank You.