Lecture 02

A Brief Summary of the History of Neural Networks and Deep Learning

STAT 479: Deep Learning, Spring 2019

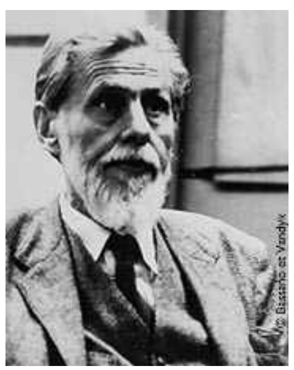
Sebastian Raschka

http://stat.wisc.edu/~sraschka/teaching/stat479-ss2019/

Neural Networks and Deep Learning -- A timeline

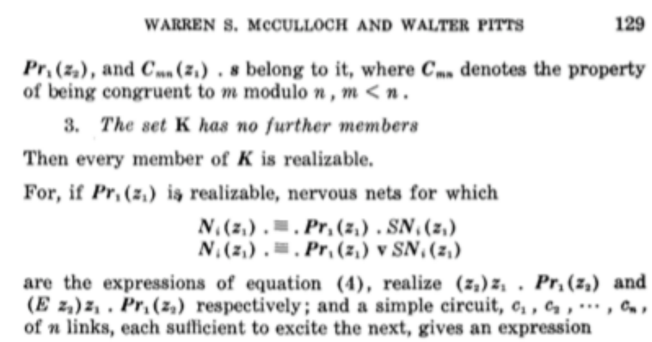
McCulloch & Pitt's neuron model (1943)

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.



Warren McCulloch

Walter Pitts



WARREN S. McCULLOCH AND WALTER PITTS 129

$Pr_1(z_2)$, and $C_{mn}(z_1) . s$ belong to it, where C_{mn} denotes the property of being congruent to m modulo n, $m < n$.

3. *The set* K *has no further members*

Then every member of K is realizable.

For, if $Pr_1(z_1)$ is realizable, nervous nets for which

$$N_i(z_1) . \equiv . Pr_1(z_1) . SN_i(z_1)$$
$$N_i(z_1) . \equiv . Pr_1(z_1) \vee SN_i(z_1)$$

are the expressions of equation (4), realize $(z_2)z_1 . Pr_1(z_2)$ and $(E\ z_2)z_1 . Pr_1(z_2)$ respectively; and a simple circuit, $c_1, c_2, \cdots, c_n$, of n links, each sufficient to excite the next, gives an expression

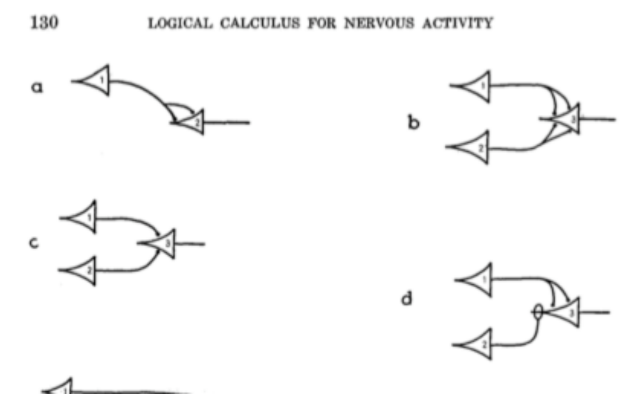
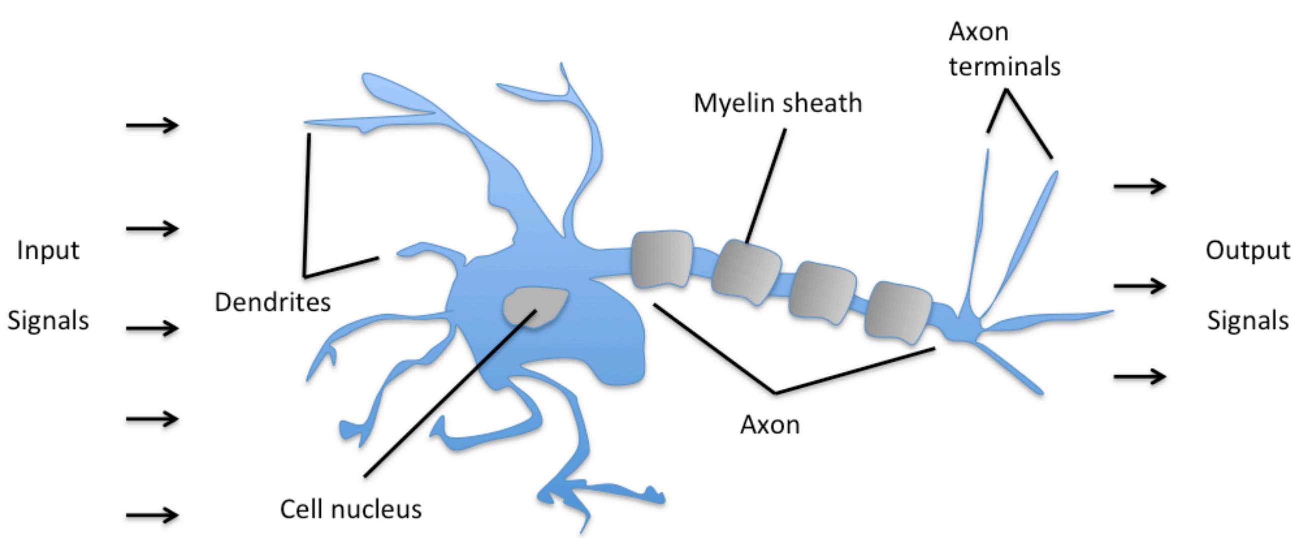


Image Source: https://www.i-programmer.info/babbages-bag/325-mcculloch-pitts-neural-networks.html



Schematic of a biological neuron.

Mathematical formulation of a biological neuron,
could solve AND, OR, NOT problems

Neural Networks and Deep Learning -- A timeline

Frank Rosenblatt's Perceptron (1957)

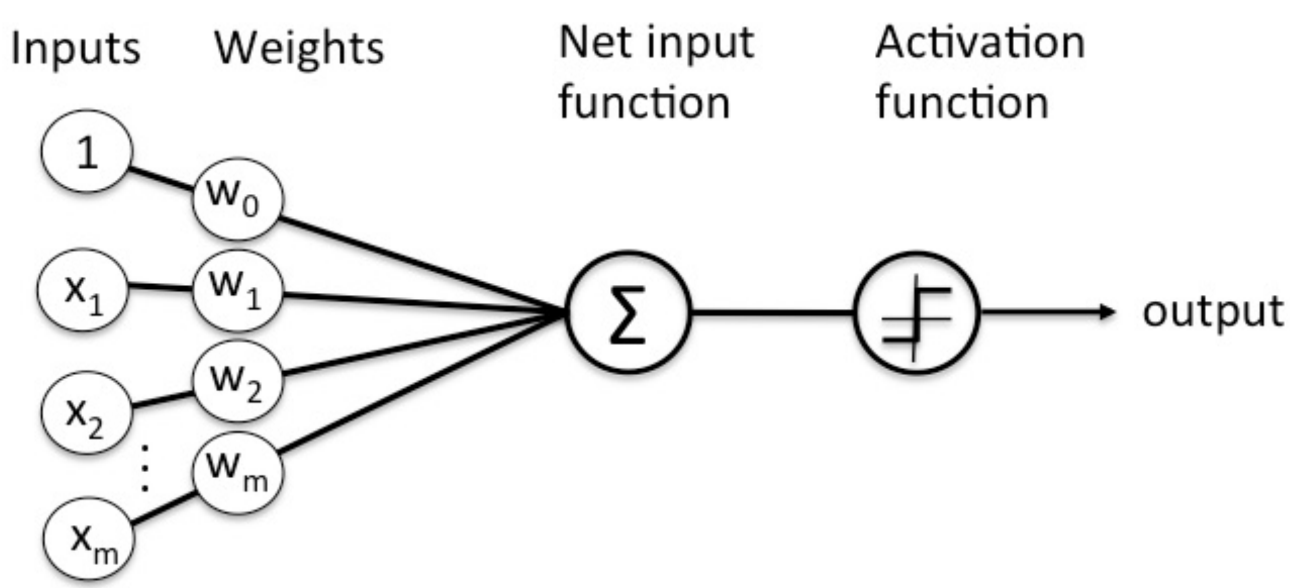
A learning algorithm for the neuron model

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Inspired by

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.

Hebb, D. O. (1949). The organization of behavior. A neuropsychological theory.



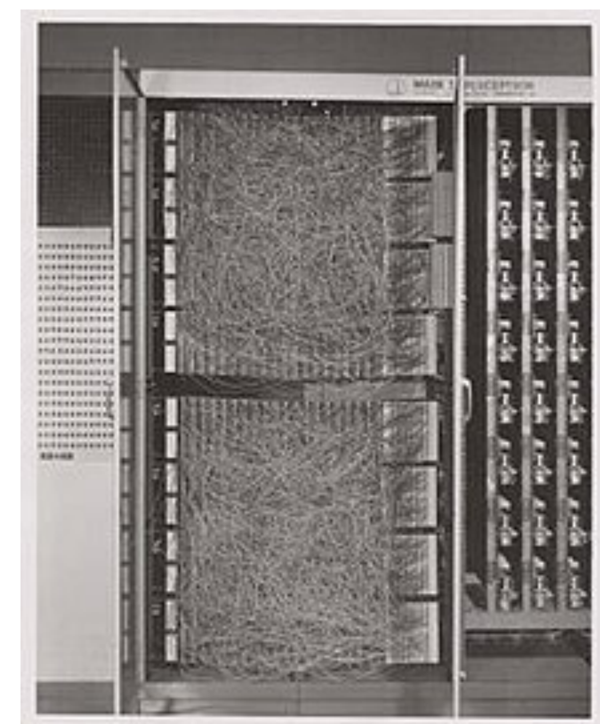


Image source: https://en.wikipedia.org/wiki/Perceptron#/media/File:Mark_I_perceptron.jpeg

Neural Networks and Deep Learning -- A timeline

Widrow and Hoff's ADALINE (1960)

A nicely differentiable neuron model

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits* (No. TR-1553-1). Stanford Univ Ca Stanford Electronics Labs.

Widrow, B. (1960). *Adaptive" adaline" Neuron Using Chemical" memistors."*.

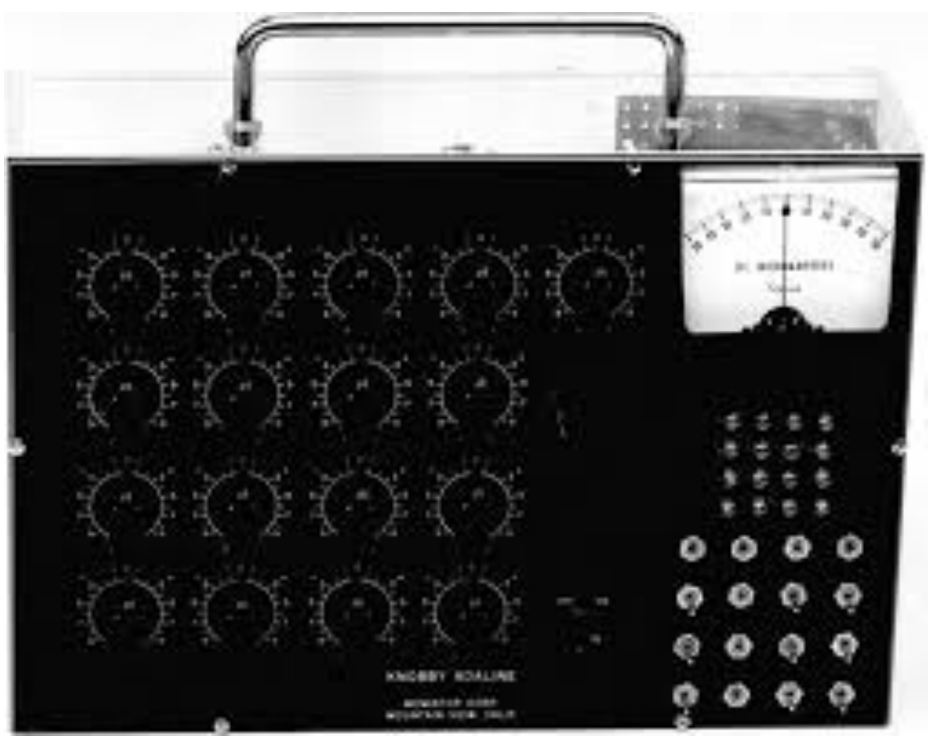
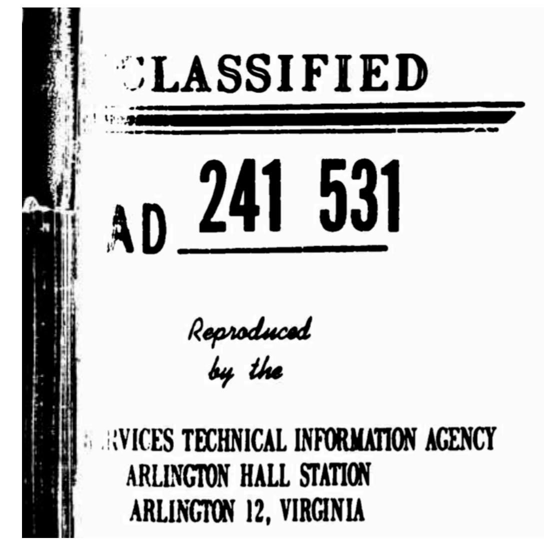


Image source: https://www.researchgate.net/profile/Alexander_Magoun2/publication/265789430/figure/fig2/AS:392335251787780@1470551421849/ADALINE-An-adaptive-linear-neuron-Manually-adapted-synapses-Designed-and-built-by-Ted.png

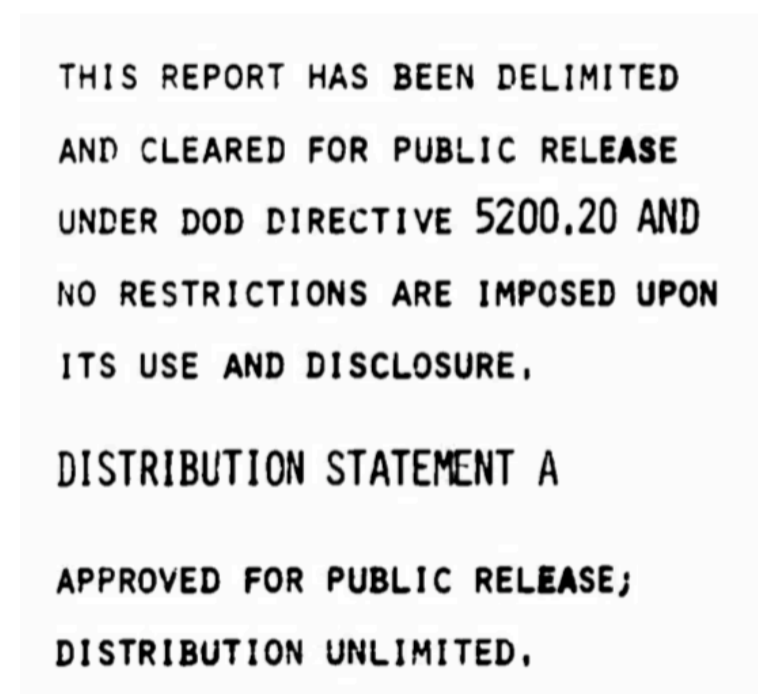


CLASSIFIED

AD 241 531

Reproduced by the

RVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA



THIS REPORT HAS BEEN DELIMITED AND CLEARED FOR PUBLIC RELEASE UNDER DOD DIRECTIVE 5200.20 AND NO RESTRICTIONS ARE IMPOSED UPON ITS USE AND DISCLOSURE.

DISTRIBUTION STATEMENT A

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

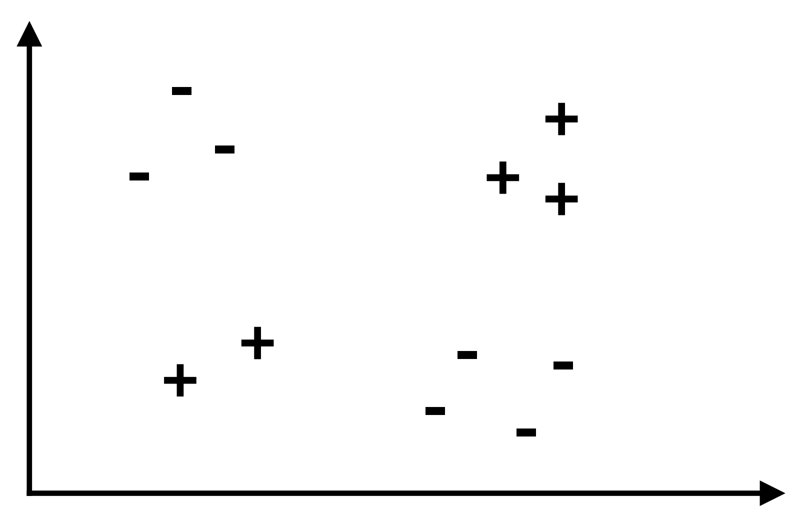
Neural Networks and Deep Learning -- A timeline

Minsky and Pappert (1969): "Perceptrons" book

=> Problem: Perceptrons (and ADALINEs) could not solve XOR problems!

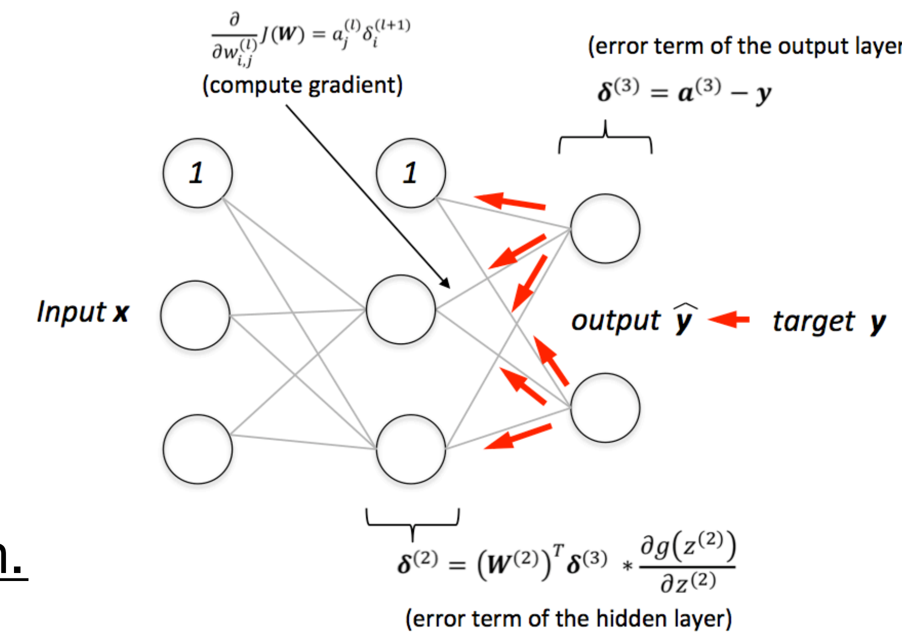
=> Neurons, a dead end? Start of the first "AI Winter"

Minski, M. L., & Papert, S. A. (1969). Perceptrons: an introduction to computational geometry. *MA: MIT Press, Cambridge*.



Neural Networks and Deep Learning -- A timeline

- Solution to the XOR problem: hidden layers and non-linear activation functions
- New problem: Hard to train
- Solution: Backpropagation



Note that backpropagation has been independently formulated many times ...

http://people.idsia.ch/~juergen/who-invented-backpropagation.

Rumelhart and Hinton (1986) formulated it independently and then showed that it really works (and formed the basis of all consequent neural network and DL progress):

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533.

Also, it was later shown that "neural nets" are universal function approximators

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359-366.

Neural Networks and Deep Learning -- A timeline

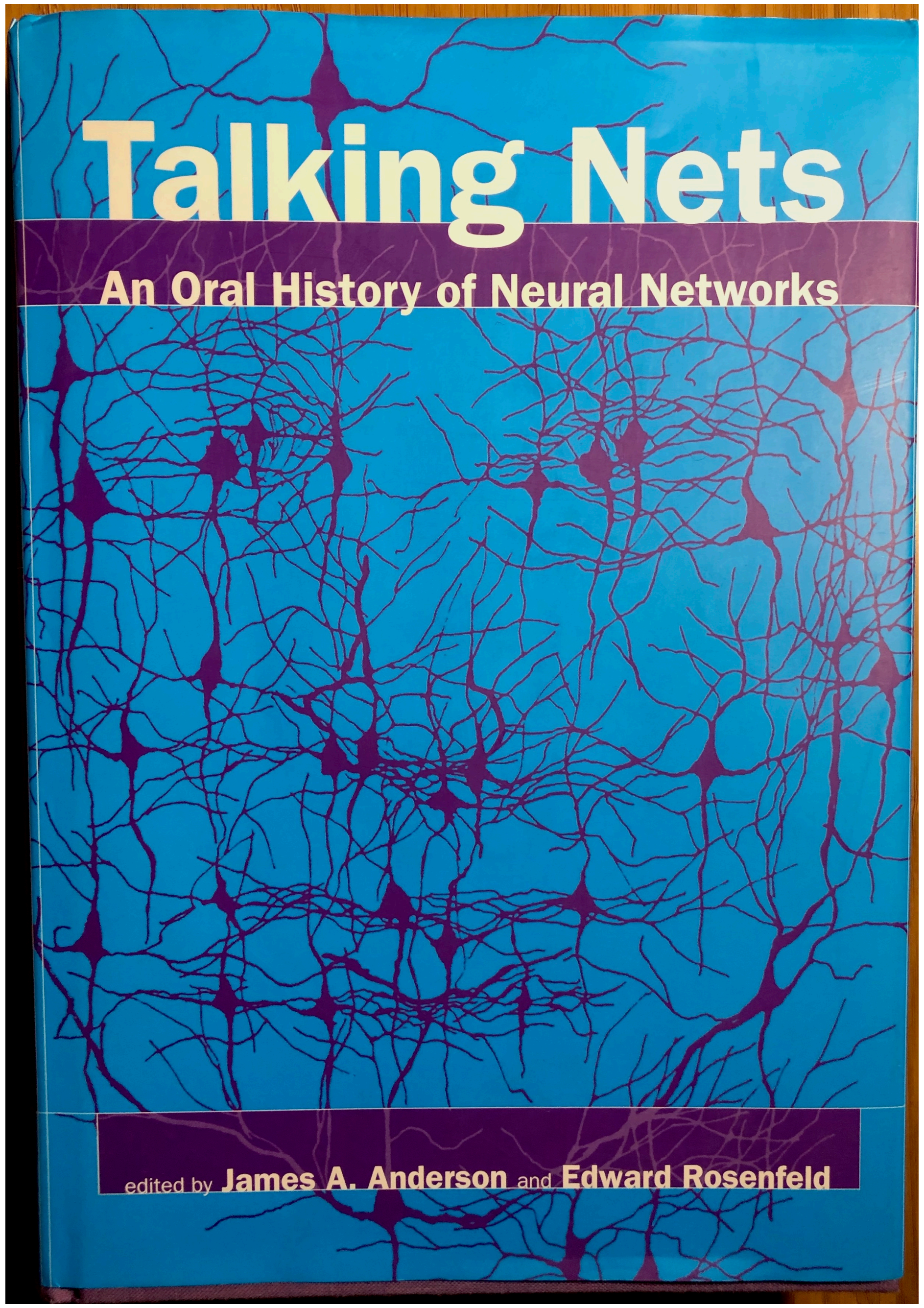
Rumelhart and Hinton (1986) formulated it independently and then showed that it really works (and formed the basis of all consequent neural network and DL progress):

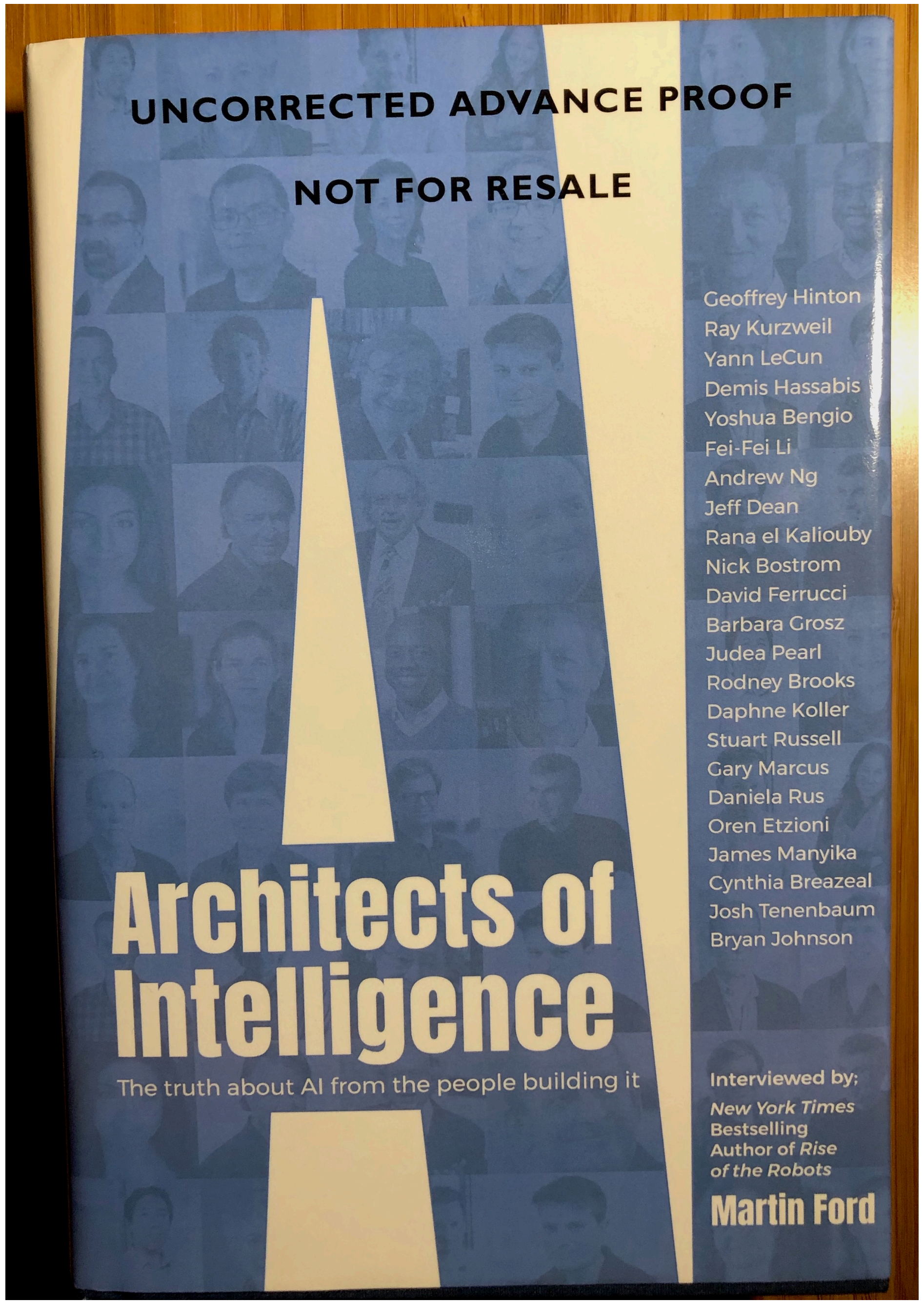
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533.

> You just get the output layer to reproduce the input layer, and then you don't need a separate teaching signal. Then the hidden units are representing some code for the input.
>
> In late 1985, I actually had a deal with Dave Rumelhart that I would write a short paper about backpropagation, which was his idea, and he would write a short paper about autoencoders, which was my idea. It was always better to have someone who didn't come up with the idea write the paper because he could say more clearly what was important.
>
> So I wrote the short paper about backpropagation, which was the *Nature* paper that came out in 1986, but Dave still hasn't written the short paper about autoencoders. I'm still waiting.
>
> What he did do was tell Dave Zipser about the idea of autoencoders and

– Geoffrey Hinton in Talking Nets - An Oral History of Neural Networks, pg. 380

Suggestions for pleasure reading





Neural Networks and Deep Learning -- A timeline

Shortly after followed a breakthrough in image recognition using some clever enhancements to

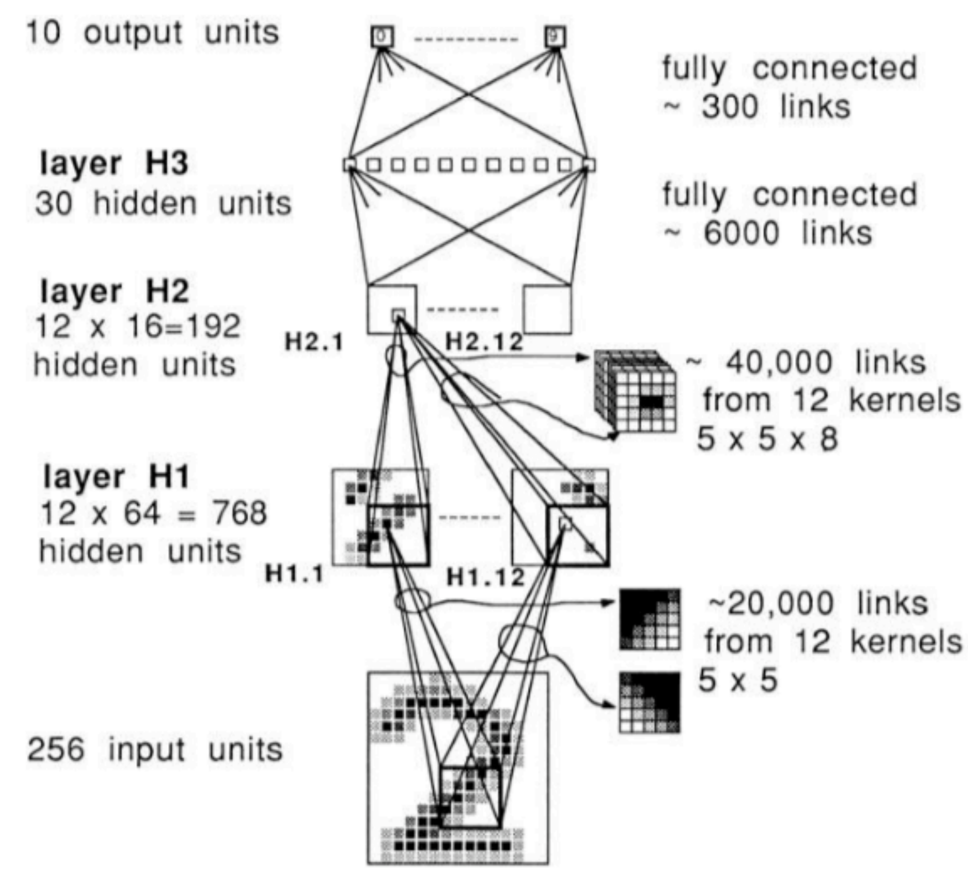
a) make training more efficient
b) extract local features (and better capture feature dependency)

by

- Weight sharing
- Pooling

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541-551.

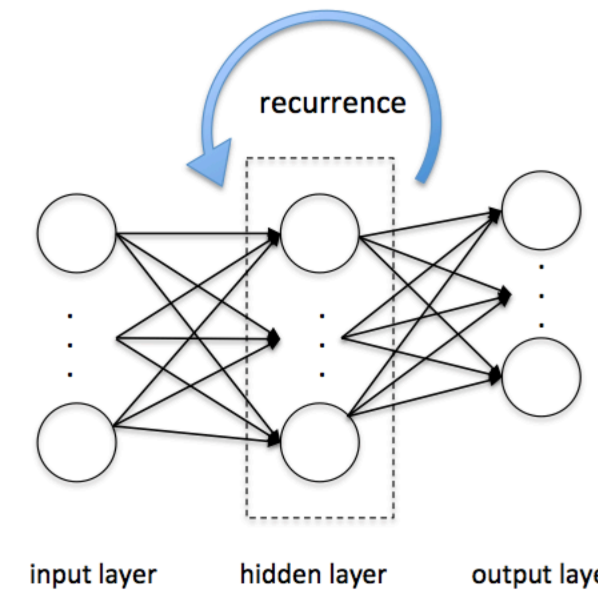
(the origin of Convolutional Neural Networks and MNIST)



Neural Networks and Deep Learning -- A timeline

Recurrent Neural Networks and Backpropagation through time

Some time created in the 1980's based on Rumelhart's work



New problems: vanishing and exploding gradients!

Schmidhuber, Jürgen (1993). *Habilitation thesis: System modeling and optimization*. Page 150 ff demonstrates credit assignment across the equivalent of 1,200 layers in an unfolded RNN.

Solution: LSTMs (still popular and commonly used)

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Neural Networks and Deep Learning -- A timeline

- 2nd "AI Winter" in the late 1990's and 2000's
- Probably due to popularity of
  Support Vector Machines and Random Forests
- Also, neural networks were still expensive to train, until GPUs came into play

Oh, K. S., & Jung, K. (2004). GPU implementation of neural networks. *Pattern Recognition*, *37*(6), 1311-1314.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Image source: https://www.amax.com/blog/?p=907

Specifications	Intel® Core™ i7-6900K Processor Extreme Ed.	NVIDIA GeForce® GTX™ 1080 Ti
Base Clock Frequency	3.2 GHz	< 1.5 GHz
Cores	8	3584
Memory Bandwidth	64 GB/s	484 GB/s
Floating-Point Calculations	409 GFLOPS	11300 GFLOPS
Cost	~ $1000.00	~ $700.00

(data from ~2017)

When did Deep Learning Become Really Popular?

That was the view of people in computer vision until 2012. Most people in computer vision thought this stuff was crazy, even though Yann LeCun sometimes got systems working better than the best computer vision systems, they still thought this stuff was crazy, it wasn't the right way to do vision. They even rejected papers by Yann, even though they worked better than the best computer vision systems on particular problems, because the referees thought it was the wrong way to do things. That's a lovely example of scientists saying, "We've already decided what the answer has to look like, and anything that doesn't look like the answer we believe in is of no interest."

In the end, science won out, and two of my students won a big public competition, and they won it dramatically. They got almost half the error rate of the best computer vision systems, and they were using mainly techniques developed in Yann LeCun's lab but mixed in with a few of our own techniques as well.

MARTIN FORD: This was the ImageNet competition?

GEOFFREY HINTON: Yes, and what happened then was what should happen in science. One method that people used to think of as complete nonsense had now worked much better than the method they believed in, and within two years, they all switched. So, for things like object classification, nobody would dream of trying to do it without using a neural network now.

(Excerpt from "Architects of Intelligence")

Neural Networks and Deep Learning -- A timeline

- Many enhancements where developed to make neural networks perform better and solve new problems

Rectified Linear Units

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807-814).

BatchNorm

Ioffe, S., & Szegedy, C. (2015, June). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning* (pp. 448-456).

Dropout

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929-1958.

GANs

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).

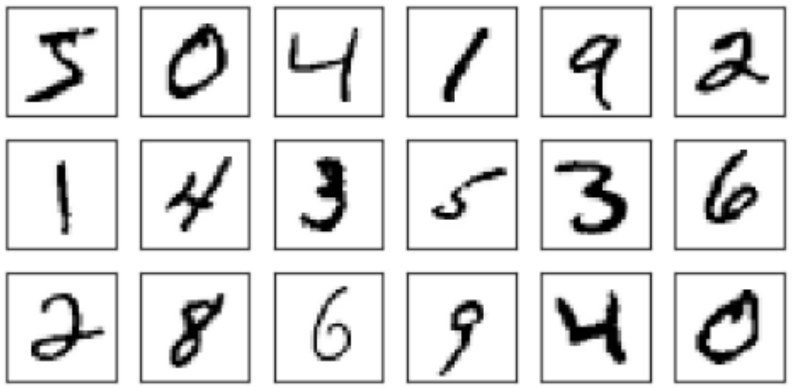
& many more

About the term "Deep Learning" ...

Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep-learning methods are representation-learning methods with multiple levels of representation [...]

-- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436.

Evolution of Benchmark Datasets for Computer Vision

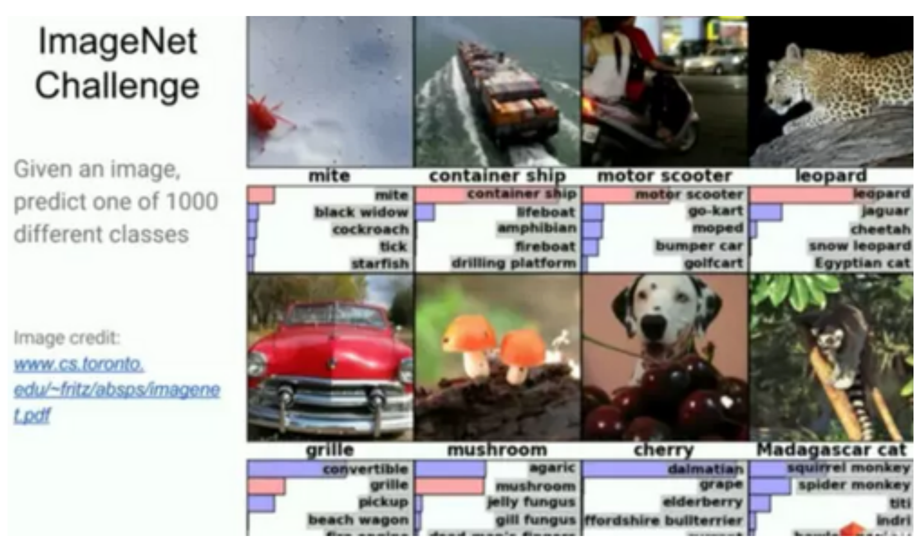


MNIST (1998)

- 60,000 examples, 10 classes
- features: 28x28x1
- http://yann.lecun.com/exdb/mnist/

CIFAR-10/CIFAR-100 (2009)

- 60,000 examples, 10 or 100 classes
- features: 32x32x3,
- https://www.cs.toronto.edu/~kriz/cifar.html



ImageNet (~2010)

- ~14 million images
- features: full resolution
- http://www.image-net.org

Publishing Culture

In contrast to most other fields:

- Mostly Open Access
- Mostly Conferences (double-blind peer review, competitive ~25% acceptance rate)
- Usually preprints online

Publishing Culture

> *The machine learning research community has a long history of open exchange of papers, software, and data sets. There may be trade secrets (obviously I wouldn't know about them), but one key to the recent rapid advances in machine learning has been the willingness of researchers at major companies (especially Google, Facebook, and Microsoft) to share their software and data sets. For example, the Microsoft COCO image data set has been very important for advancing computer vision and machine learning research. The TensorFlow framework developed at Google has made it possible for even high school students to experiment with training deep neural networks.*

> *This is particularly important for students and faculty in poor countries where ther is no money to purchase subscriptions to expensive journals. Keep in mind that a large fraction of machine learning research takes place in universities and is funded by tax payers. The tax payers should have free access to the resulting publications.*

-- Tom Dieterrich, Professor at Oregon State University

https://thenextweb.com/artificial-intelligence/2018/05/01/heres-why-thousands-of-ai-researchers-will-boycott-a-science-journal/

Top Conferences for Machine Learning & Arti. Intelligence

*Ranking is based on **Conference H5-index>=12** provided by Google Scholar Metrics*

Show **Due** only | All Categories | All Countries | Search by keyword

	Hindex	Publisher	Conference Details
1	158	IEEE	**CVPR : IEEE Conference on Computer Vision and Pattern Recognition, CVPR** Jun 15, 2019 - Jun 21, 2019 - Long Beach , **United States** http://cvpr2019.thecvf.com/
2	101	Neural Information Processing Systems Foundation	**NIPS : Neural Information Processing Systems (NIPS)** Dec 3, 2018 - Dec 6, 2018 - Palais des Congrès de Montréal , **Canada** https://nips.cc/
3	98	Springer	**ECCV : European Conference on Computer Vision** Sep 8, 2018 - Sep 14, 2018 - Munich , **Germany** https://eccv2018.org/
4	91	IMLS	**ICML : International Conference on Machine Learning (ICML)** Jul 10, 2018 - Jul 15, 2018 - Stockholm , **Sweden** https://icml.cc/
5	89	IEEE	**ICCV : IEEE International Conference on Computer Vision** Oct 27, 2019 - Nov 3, 2019 - Seoul , **South Korea** http://iccv2019.thecvf.com/ **Deadline : Mon 22 Apr 2019**
10	73	Association for Computing Machinery	**SIGKDD : ACM SIGKDD International Conference on Knowledge discovery and data mining** Aug 19, 2018 - Aug 23, 2018 - London , **United Kingdom** http://www.kdd.org/kdd2018/
16	67		**ACL : Meeting of the Association for Computational Linguistics (ACL)** Aug 28, 2019 - Sep 2, 2019 - Florence , **Italy** http://www.acl2019.org **Deadline : Thu 04 Apr 2019**
22	59	Association for Computing Machinery	**SIGMOD : ACM SIGMOD International Conference on Management of Data** Jun 30, 2019 - Jul 15, 2019 - Amsterdam , **Netherlands**

Source: http://www.guide2research.com/topconf/machine-learning

Journal of Machine Learning Research

Home Page
Papers
Submissions
News
Editorial Board
Announcements
Proceedings
Open Source Software
Search

The Journal of Machine Learning Research (JMLR) provides an international forum for the electronic and paper publication of high-quality scholarly articles in all areas of machine learning. All published papers are freely available online.

JMLR has a commitment to rigorous yet rapid reviewing. Final versions are published electronically (ISSN 1533-7928) immediately upon receipt. Until the end of 2004, paper volumes (ISSN 1532-4435) were published 8 times annually and sold to libraries and individuals by the MIT Press. Paper volumes (ISSN 1532-4435) are now published and sold by Microtome Publishing.

News

- *2019.01.20*: Volume 19 completed; Volume 20 began.
- *2018.08.28*: Volume 18 completed; Volume 19 began.
- *2018.04.16*: Changes in JMLR leadership team.
- *2016.12.01*: Special topic on Learning from Electronic Health Data
- *2015.09.01*: Special issue in Memory of Alexey Chervonenkis.

© JMLR 2019.

http://www.jmlr.org

JAIR Journal of Artificial Intelligence Research

Register Login

CURRENT ARCHIVES ABOUT SPECIAL TRACKS

Search

NEWS

Comments Welcome on New Site

Help Support JAIR

New Special Track on Deep Learning, Knowledge Representation, and Reasoning

New Special Track on AI & Society

Recent Articles

https://www.jair.org/

Login

Search

ICML | 2019

Thirty-sixth International Conference on Machine Learning

Dates ~~Schedule~~ Calls Long Beach

Year (2019)

Help

ICML 2019 Call for Papers

The 36th International Conference on Machine Learning (ICML 2019) will be held in Long Beach, CA, USA from June 10th to June 15th, 2019. The conference will consist of one day of tutorials (June 10), followed by three days of main conference sessions (June 11-13), followed by two days of workshops (June 14-15). We invite submissions of papers on all topics related to machine learning for the conference proceedings, and proposals for tutorials and workshops.

policy, and the organizers have the right to reject such submissions, and remove them from the proceedings.

There are several exceptions to this rule:

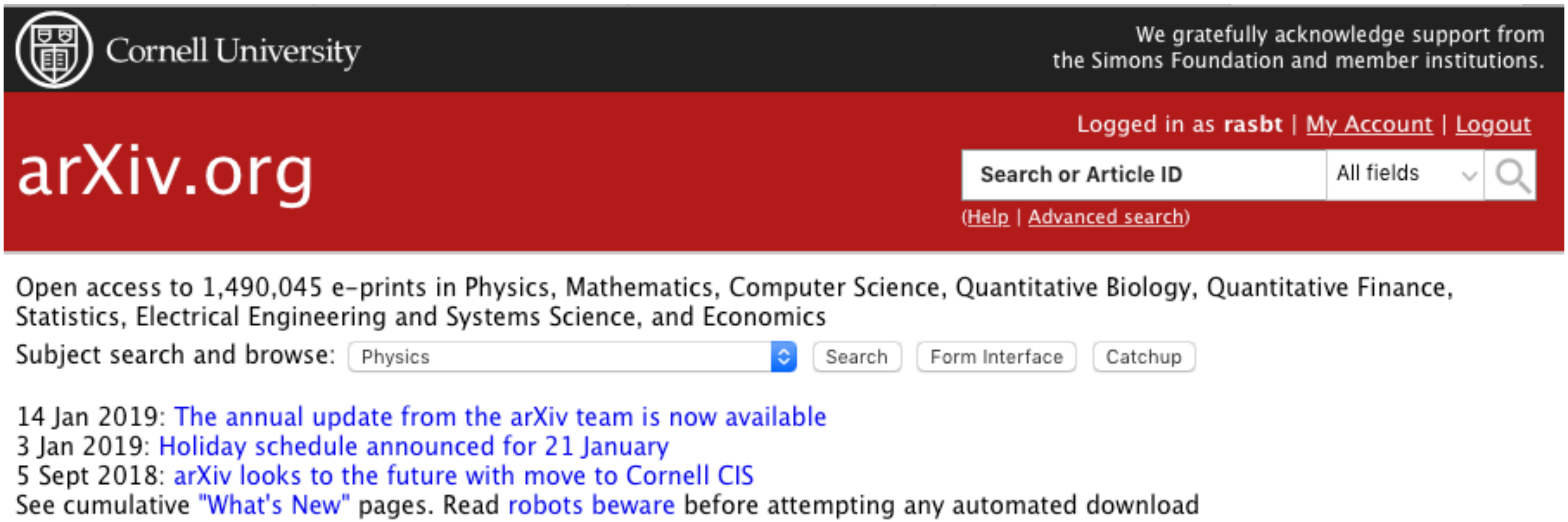
1. Submission is permitted of a short version of a paper that has been submitted to a journal, but will not be published in that journal on or before June 2019. Authors must declare such dual-submissions either through the CMT submission form, or via email to the program chairs (icml2019pc@gmail.com). It is the author's responsibility to make sure that the journal in question allows dual concurrent submissions to conferences.
2. Submission is permitted for papers presented or to be presented at conferences or workshops without proceedings (e.g., ICML or NIPS workshops), or with only abstracts published.
3. Submission is permitted for papers that are available as a technical report (or similar, e.g., in arXiv). In this case we suggest the authors not cite the report, so as to preserve anonymity.

Finally, note that previously published papers with substantial overlap written by the authors must be cited in such a way so as to preserve author anonymity. Differences relative to these earlier papers must be explained in the text of the submission. For example, (This work develops [our earlier work], which showed that).

Reviewing Criteria

Accepted papers must contain significant novel results. Results can be either theoretical or empirical. Results will be judged on the degree to which they have been objectively established and/or their potential for scientific and technological impact. Reproducibility of results and easy availability of code will be taken into account in the decision-making process.

Most ML/DL papers are also available as preprints



https://arxiv.org

(careful & critical though, note that not all preprints were later peer-reviewed / were accepted for publication)

Nice Preprint Recommender System by Andrej Karpathy

http://www.arxiv-sanity.com

Current Trends

- Applications across fields and in industry
- Engineering new tricks
- Developing specialized hardware
- Developing theory and understanding

Applications

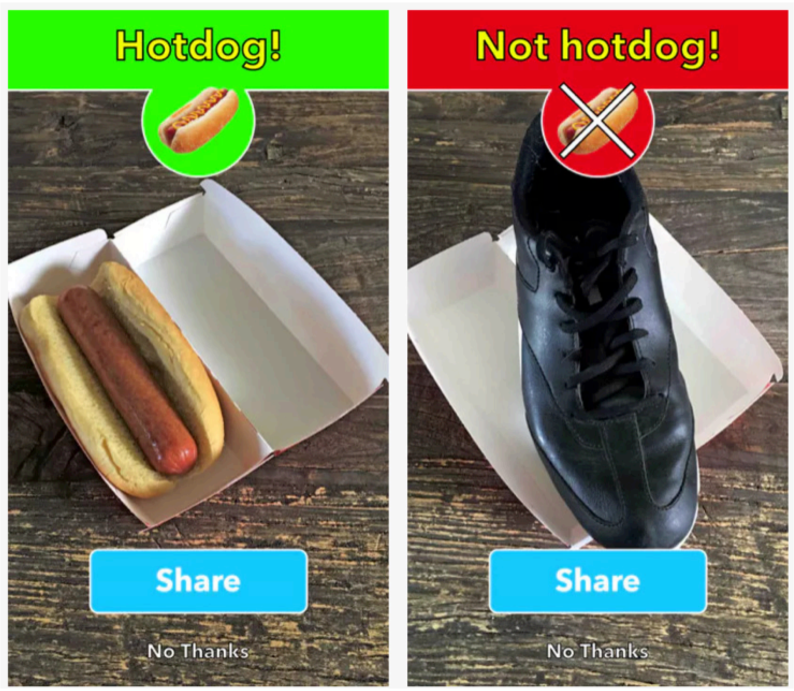
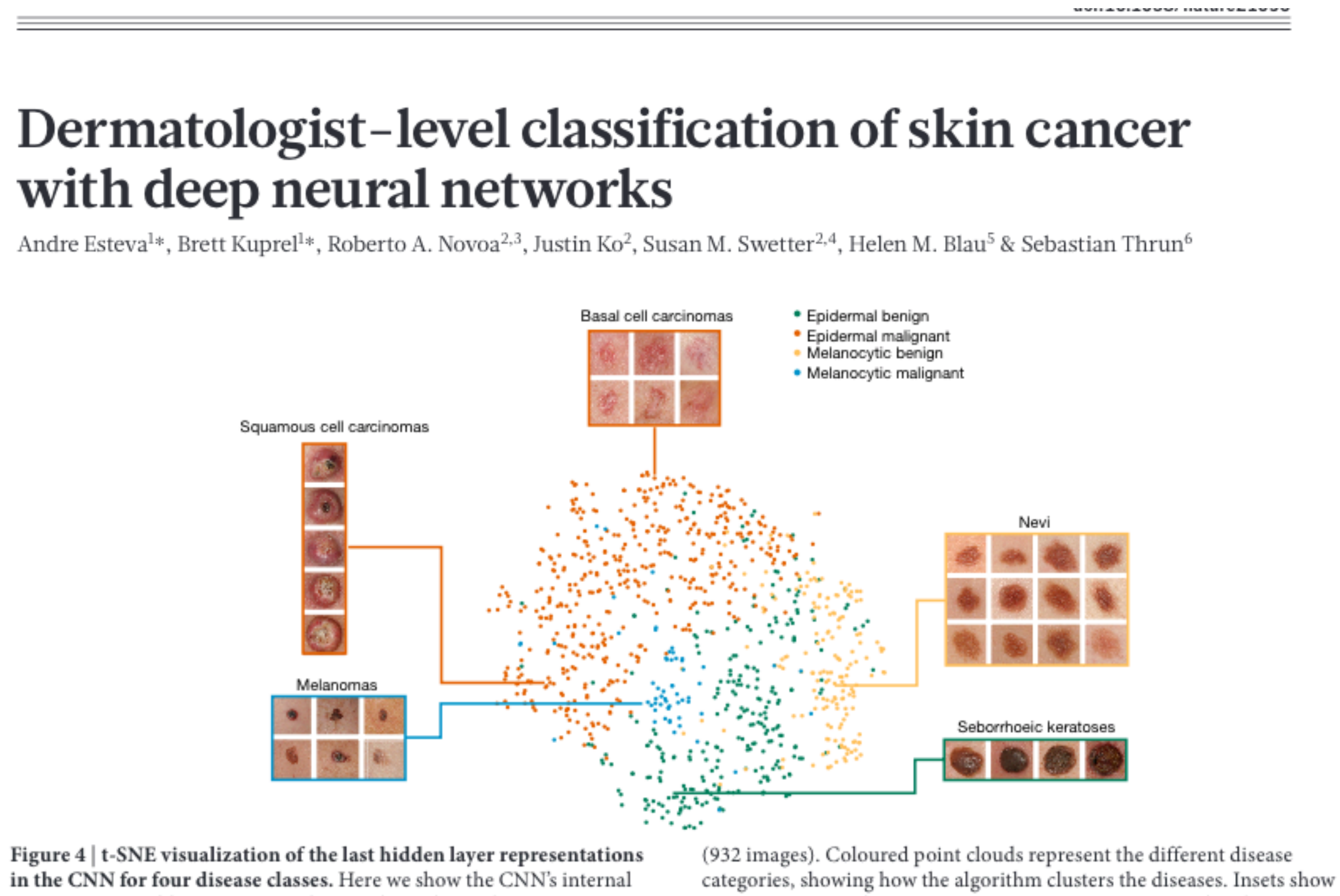


Photo: SeeFood Technologies, Inc.

https://www.theverge.com/tldr/2017/5/14/15639784/hbo-silicon-valley-not-hotdog-app-download



Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1*], Brett Kuprel[1*], Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

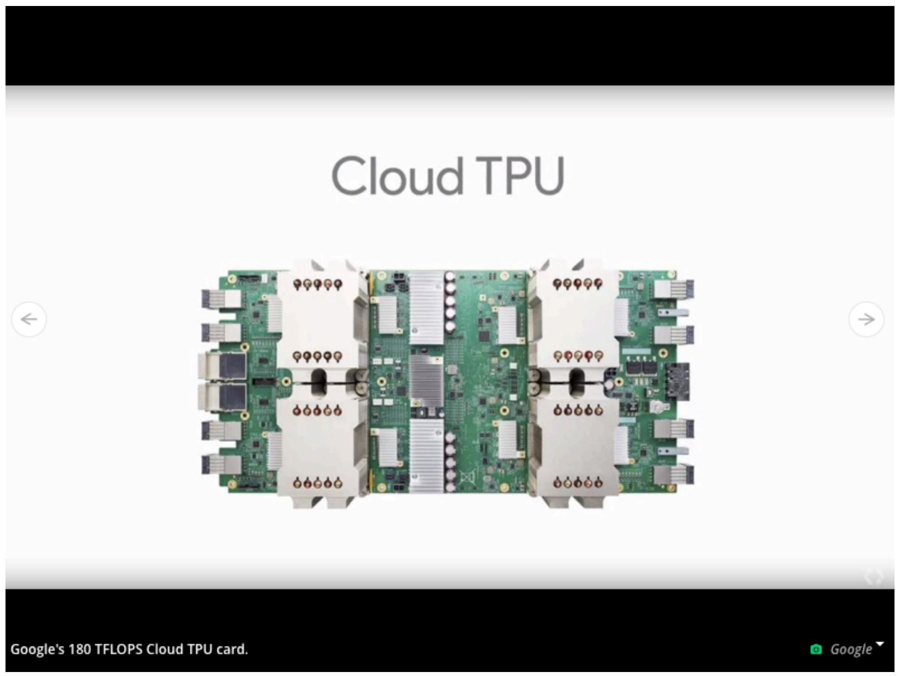
Figure 4 | t-SNE visualization of the last hidden layer representations in the CNN for four disease classes. Here we show the CNN's internal (932 images). Coloured point clouds represent the different disease categories, showing how the algorithm clusters the diseases. Insets show

https://www.nature.com/articles/nature21056.epdf

. . .

https://ai.googleblog.com/2015/07/how-google-translate-squeezes-deep.html

Developing Specialized Hardware

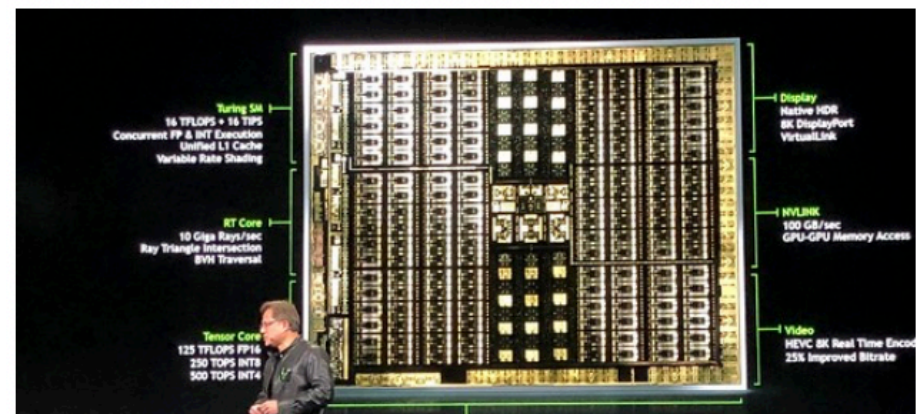


https://arstechnica.com/gadgets/2018/07/the-ai-revolution-has-spawned-a-new-chips-arms-race/

Opinion: New Nvidia chip extends the company's lead in graphics, artificial intelligence

By Ryan Shrout
Published: Aug 14, 2018 2:35 p.m. ET

The only question that remains: How big is Nvidia's advantage over its rivals?



https://www.marketwatch.com/story/new-nvidia-chip-extends-the-companys-lead-in-graphics-artificial-intelligence-2018-08-14

https://developer.arm.com/products/processors/machine-learning/arm-ml-processor

TECHNOLOGY NEWS NOVEMBER 28, 2018 / 2:59 PM / 2 MONTHS AGO

Amazon launches machine learning chip, taking on Nvidia, Intel

https://www.reuters.com/article/us-amazon-com-nvidia/amazon-launches-machine-learning-chip-taking-on-nvidia-intel-idUSKCN1NX2PY

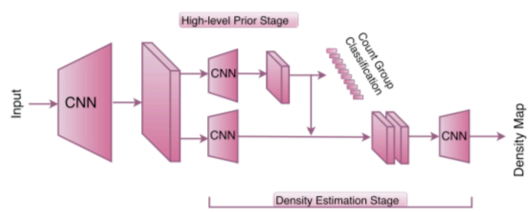
Engineering New Tricks

CNN-based Cascaded Multi-task Learning of High-level Prior and Density Estimation for Crowd Counting

Vishwanath A. Sindagi   Vishal M. Patel
Department of Electrical and Computer Engineering, Rutgers University
94 Brett Road, Piscataway, NJ, 08854, USA
vishwanath.sindagi@rutgers.edu, vishal.m.patel@rutgers.edu

Abstract

Estimating crowd count in densely crowded scenes is an ...emely challenging task due to non-uniform scale vari...ons. In this paper, we propose a novel end-to-end cas...ed network of CNNs to jointly learn crowd count clas...ation and density map estimation. Classifying crowd ...nt into various groups is tantamount to coarsely es-



https://arxiv.org/pdf/1707.09605.pdf

Group Normalization

Yuxin Wu   Kaiming He
Facebook AI Research (FAIR)
{yuxinwu,kaiminghe}@fb.com

Abstract

Batch Normalization (BN) is a milestone technique in the ...elopment of deep learning, enabling various networks ...rain. However, normalizing along the batch dimension ...oduces problems — BN's error increases rapidly when batch size becomes smaller, caused by inaccurate batch ...istics estimation. This limits BN's usage for training ...er models and transferring features to computer vision ...s including detection, segmentation, and video, which ...uire small batches constrained by memory consumption. ...his paper, we present Group Normalization (GN) as ...imple alternative to BN. GN divides the channels into

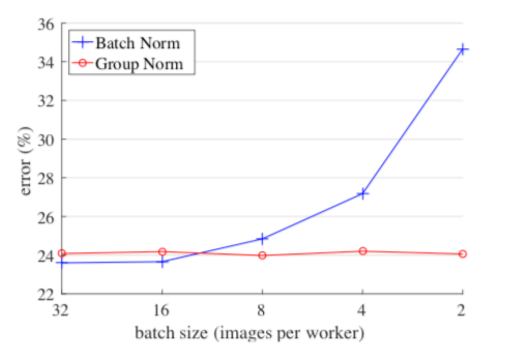


Figure 1. **ImageNet classification error *vs.* batch sizes.** This is

https://arxiv.org/pdf/1803.08494.pdf

Cyclical Learning Rates for Training Neural Networks

Leslie N. Smith
U.S. Naval Research Laboratory, Code 5514
4555 Overlook Ave., SW., Washington, D.C. 20375
leslie.smith@nrl.navy.mil

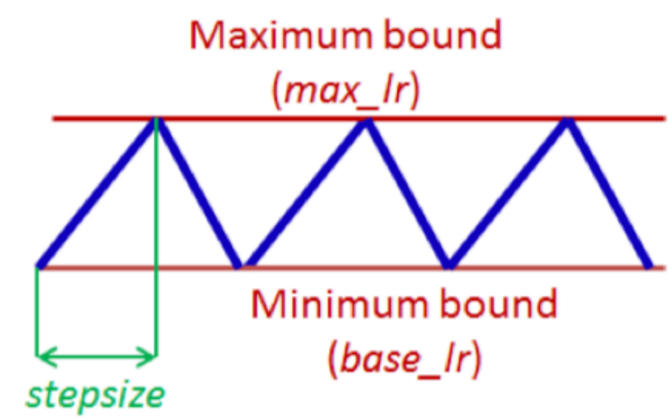


Figure 2. Triangular learning rate policy. The blue lines represent learning rate values changing between bounds. The input parameter *stepsize* is the number of iterations in half a cycle.

https://arxiv.org/pdf/1506.01186.pdf

. . .

Developing Theory and Understanding

Opening the Black Box of Deep Neural Networks via Information

Ravid Shwartz-Ziv, Naftali Tishby

(Submitted on 2 Mar 2017 (v1), last revised 29 Apr 2017 (this version, v3))

Despite their great success, there is still no comprehensive theoretical understanding of learning with Deep Neural Networks (DNNs) or their inner organization. Previous work proposed to analyze DNNs in the \textit{Information Plane}; i.e., the plane of the Mutual Information values that each layer preserves on the input and output variables. They suggested that the goal of the network is to optimize the Information Bottleneck (IB) tradeoff between compression and prediction, successively, for each layer.
In this work we follow up on this idea and demonstrate the effectiveness of the Information-Plane visualization of DNNs. Our main results are: (i) most of the training epochs in standard DL are spent on {\emph compression} of the input to efficient representation and not on fitting the training labels. (ii) The representation compression phase begins when the training errors becomes small and the Stochastic Gradient Decent (SGD) epochs change from a fast drift to smaller training error into a stochastic relaxation, or random diffusion, constrained by the training error value. (iii) The converged layers lie on or very close to the Information Bottleneck (IB) theoretical bound, and the maps from the input to any hidden layer and from this hidden layer to the output satisfy the IB self-consistent equations. This generalization through noise mechanism is unique to Deep Neural Networks and absent in one layer networks. (iv) The training time is dramatically reduced when adding more hidden layers. Thus the main advantage of the hidden layers is computational. This can be explained by the reduced relaxation time, as this it scales super-linearly (exponentially for simple diffusion) with the information compression from the previous layer.

Geometric Understanding of Deep Learning

Na Lei, Zhongxuan Luo, Shing-Tung Yau, David Xianfeng Gu

(Submitted on 26 May 2018 (v1), last revised 31 May 2018 (this version, v2))

Deep learning is the mainstream technique for many machine learning tasks, including image recognition, machine translation, speech recognition, and so on. It has outperformed conventional methods in various fields and achieved great successes. Unfortunately, the understanding on how it works remains unclear. It has the central importance to lay down the theoretic foundation for deep learning.
In this work, we give a geometric view to understand deep learning: we show that the fundamental principle attributing to the success is the manifold structure in data, namely natural high dimensional data concentrates close to a low-dimensional manifold, deep learning learns the manifold and the probability distribution on it.
We further introduce the concepts of rectified linear complexity for deep neural network measuring its learning capability, rectified linear complexity of an embedding manifold describing the difficulty to be learned. Then we show for any deep neural network with fixed architecture, there exists a manifold that cannot be learned by the network.
Finally, we propose to apply optimal mass transportation theory to control the probability distribution in the latent space.

...

Next Lecture:

The Perceptron