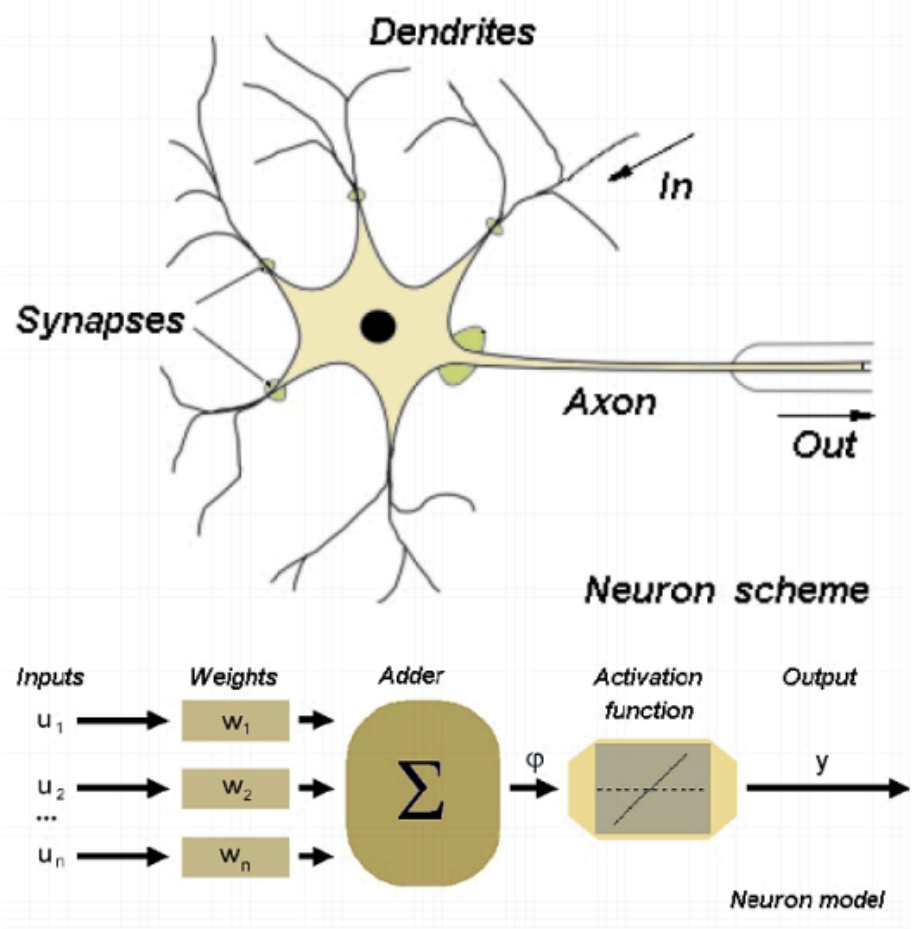
History

Suleyman Demirel University

CSS634: Deep Learning

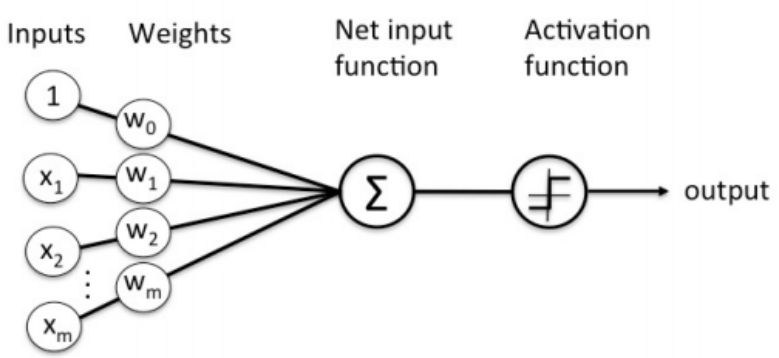
PhD Abay Nussipbekov

First Generation Neural Networks: McCullogh Pitts (1943)

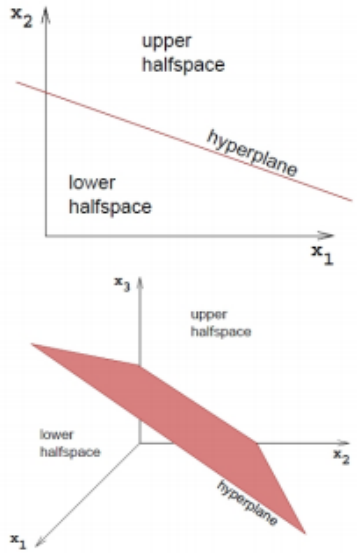


Frank Rosenblatt's Perceptron (1957)

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b \\ 0 & \text{otherwise} \end{cases}$$



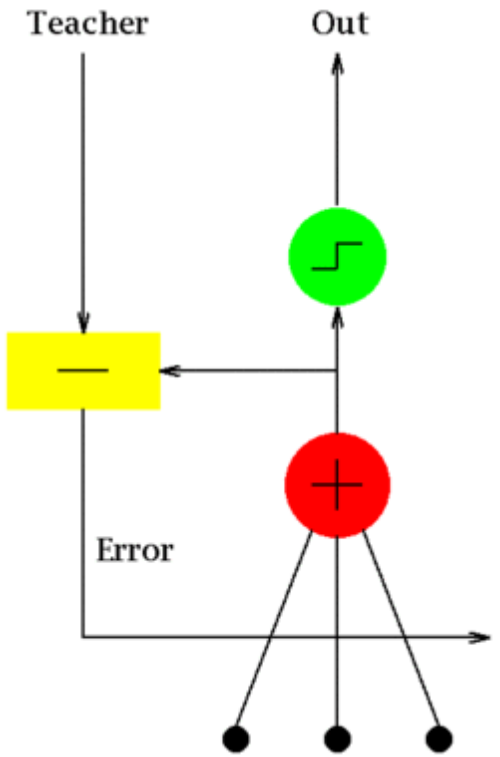
- Assumes data are linearly separable
- A perceptron represents a decision surface in a *d* dimensional space as a hyperplane
- Many boolean functions can be represented by a perceptron: AND, OR, NAND,NOR



Widrow and Hoff's ADALINE (1960)

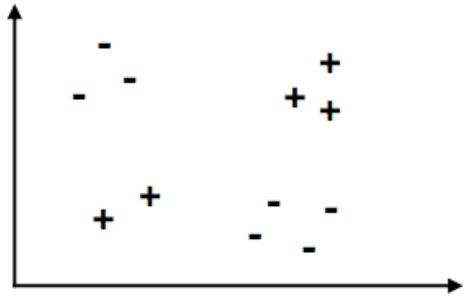
- A nicely differentiable neuron model
- Adaline is a single layer neural network with multiple nodes where each node accepts multiple inputs and generates one output.

$$y = \sum_{j=1}^{n} x_j w_j + c$$

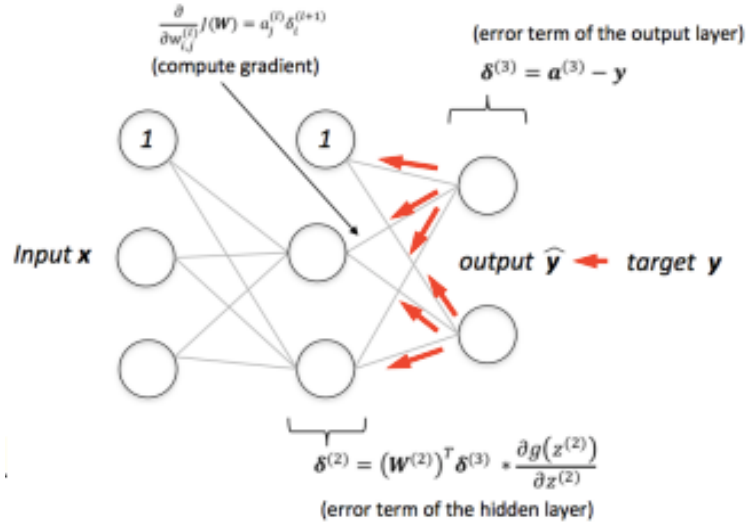


Problems

- Perceptrons (and ADALINEs) could not solve XOR problems
- Neurons, a dead end? Start of the first "AI Winter”



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



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

Improvements

- 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.

Other Early Works

- 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.

New Problems

- 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

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")

New Enhancements

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

“Deep Learning” term

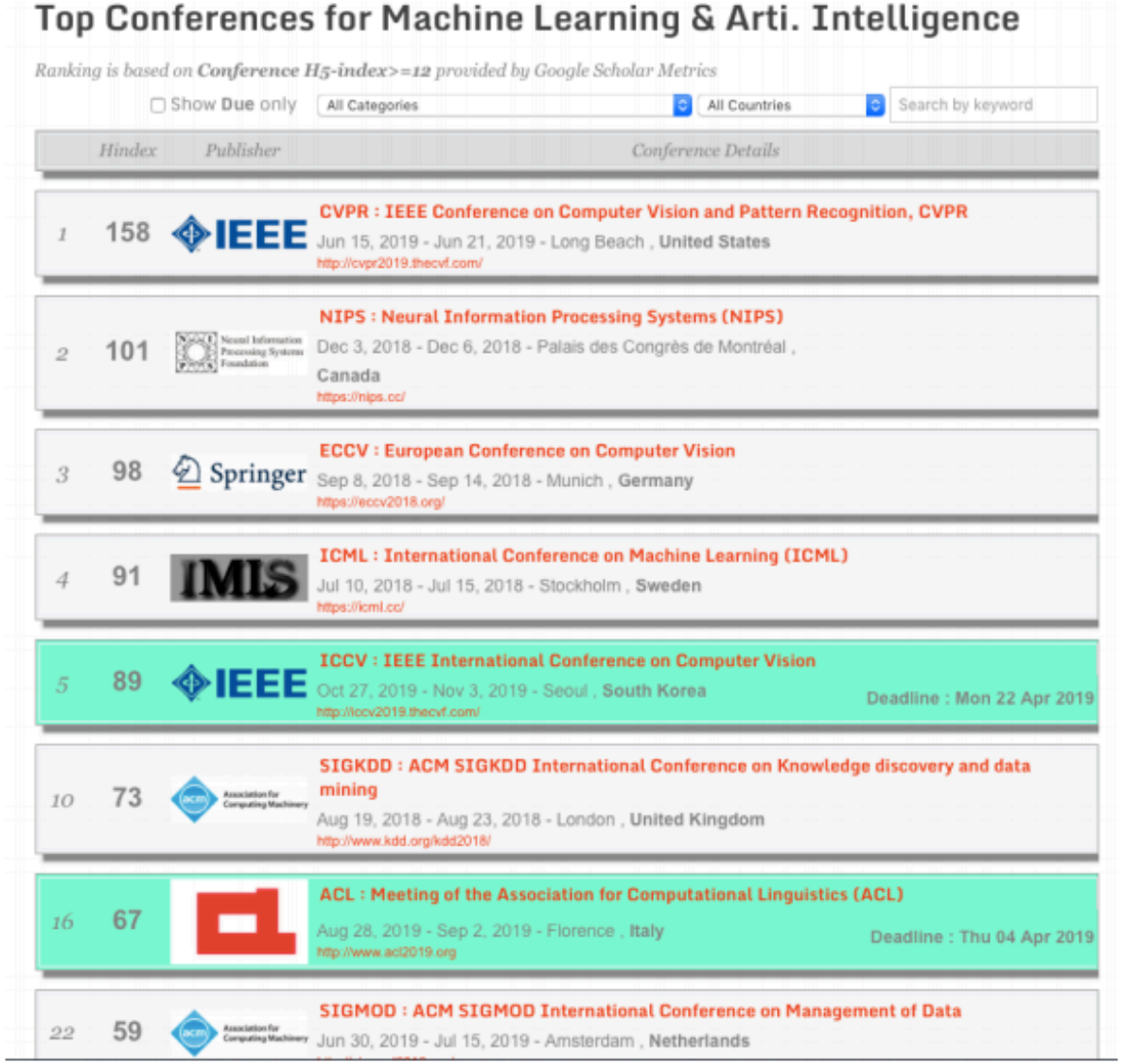
Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep-learning methods are representation-learning methods with multiple levels of representation [...]

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

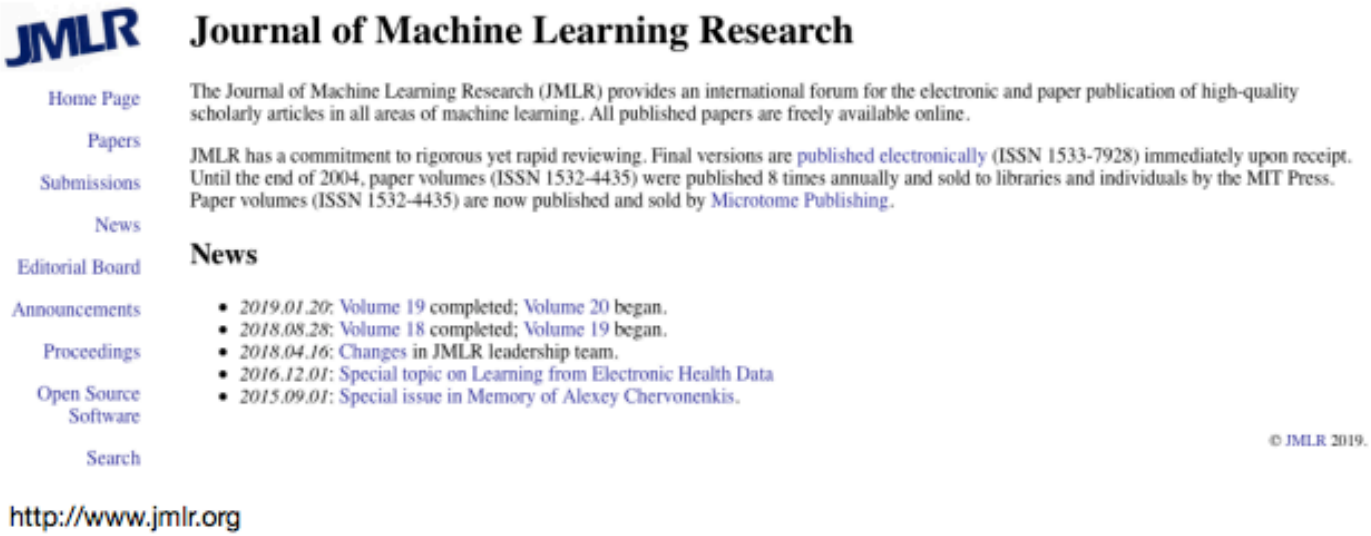
Publishing Culture

In contrast to most other fields:

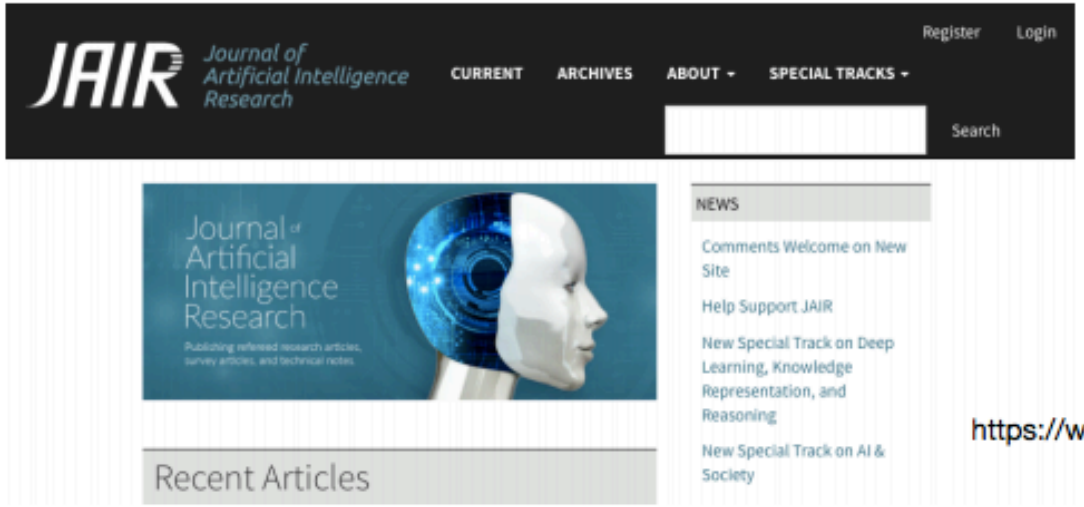
- Mostly Open Access
- Mostly Conferences (double-blind peer review, competitive ~25% acceptance rate)
- Usually preprints online



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



http://www.jmlr.org



https://www.jair.org/

ICML | 2019

Thirty-sixth International Conference on Machine Learning

Year (2019)

Help

Dates Schedule Calls Long Beach

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

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

There are several exceptions to this rule:

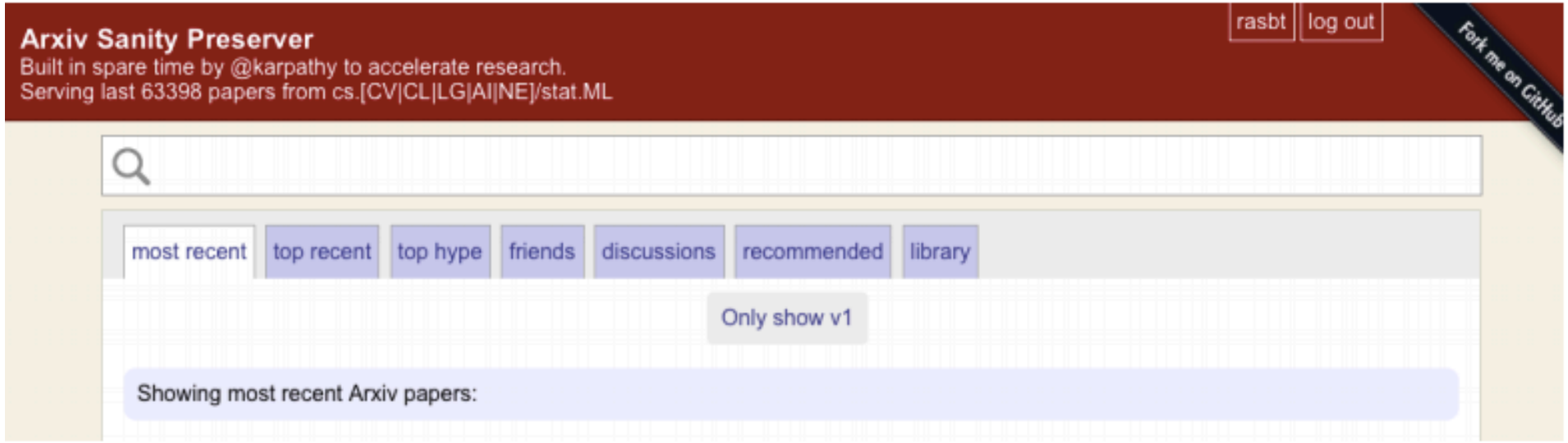
1. Submission is permitted of a short version of a paper that has been submitted to a journal, but will not be published in that journal on or before June 2019. Authors must declare such dual-submissions either through the CMT submission form, or via email to the program chairs (icml2019pc@gmail.com). 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 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.

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

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]

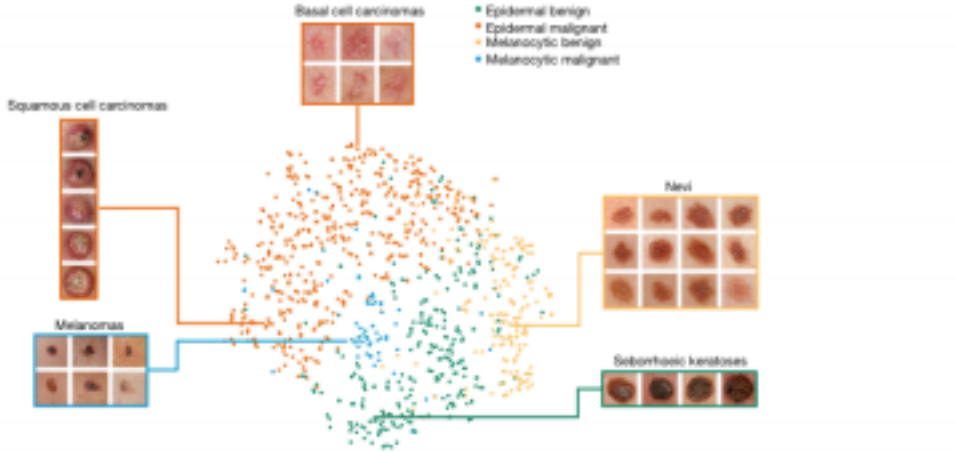
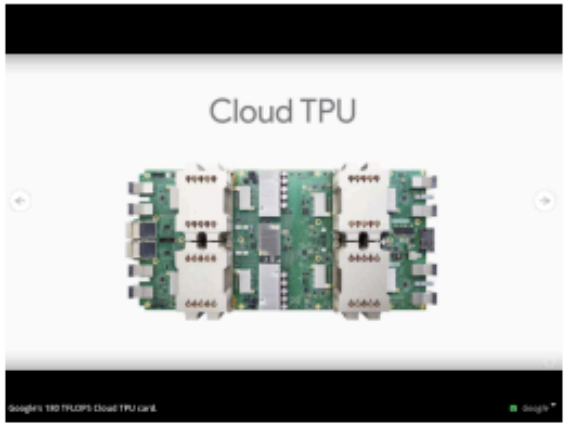


Figure 4 | t-SNE visualization of the last hidden layer representations in the CNN for four disease classes. Here we show the CNN's internal (932 images). Coloured point clouds represent the different disease categories, showing how the algorithm clusters the diseases. Insets show

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

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

Developing Specialized Hardware



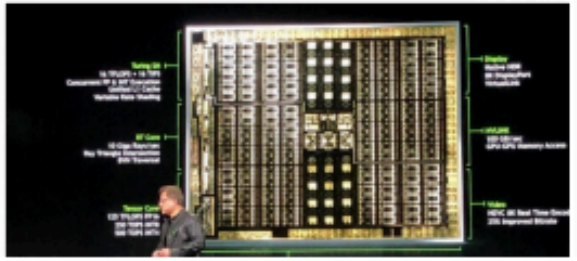
https://arstechnica.com/gadgets/2018/07/the-ai-revolution-has-spawned-a-new-chips-arms-race/

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 extremely challenging task due to non-uniform scale variation. In this paper, we propose a novel end-to-end cascaded network of CNNs to jointly learn crowd count classification and density map estimation. Classifying crowd count 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 development of deep learning, enabling various networks to train. However, normalizing along the batch dimension introduces problems — BN's error increases rapidly when batch size becomes smaller, caused by inaccurate batch statistics estimation. This limits BN's usage for training larger models and transferring features to computer vision tasks including detection, segmentation, and video, which require small batches constrained by memory consumption. In this paper, we present Group Normalization (GN) as a simple alternative to BN. GN divides the channels into

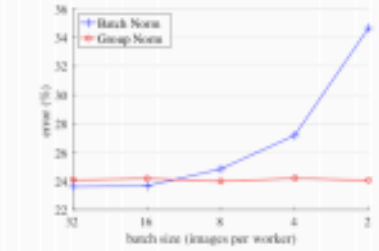


Figure 1. **ImageNet classification error vs. batch sizes**. This is

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

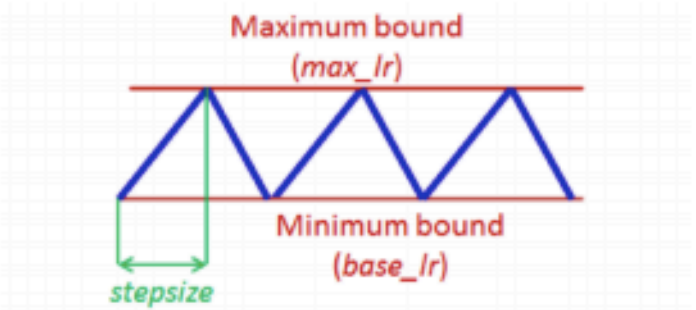


Figure 2. Triangular learning rate policy. The blue lines represent learning rate values changing between bounds. The input parameter *stepsize* is the number of iterations in half a cycle.

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.

. . .

Resources Used

- STAT 479: Deep Learning by Sebastian Raschka
- CMSC 35246 Deep Learning by Shubhendu Trivedi and Risi Kondor