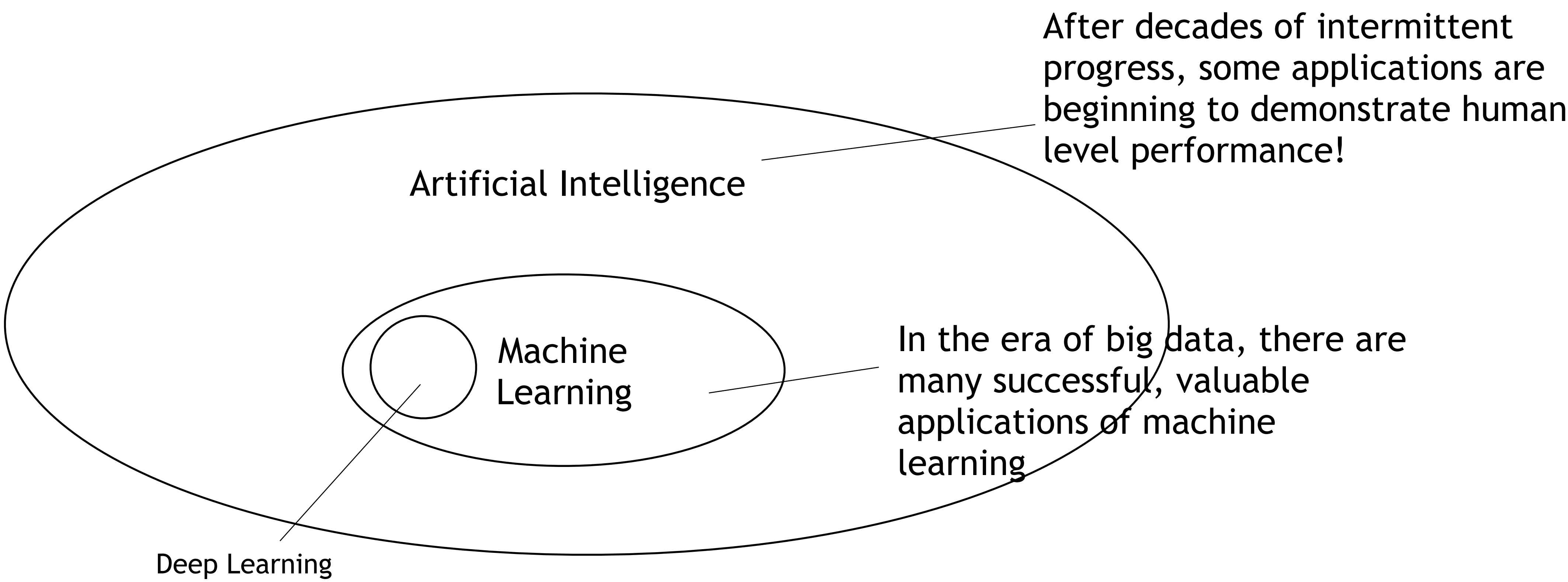
An Introduction to Human-Aided Deep Learning

James K Baker, Bhiksha Raj, Rita Singh

Opportunities in Machine Learning

- Great advances are being made in machine learning



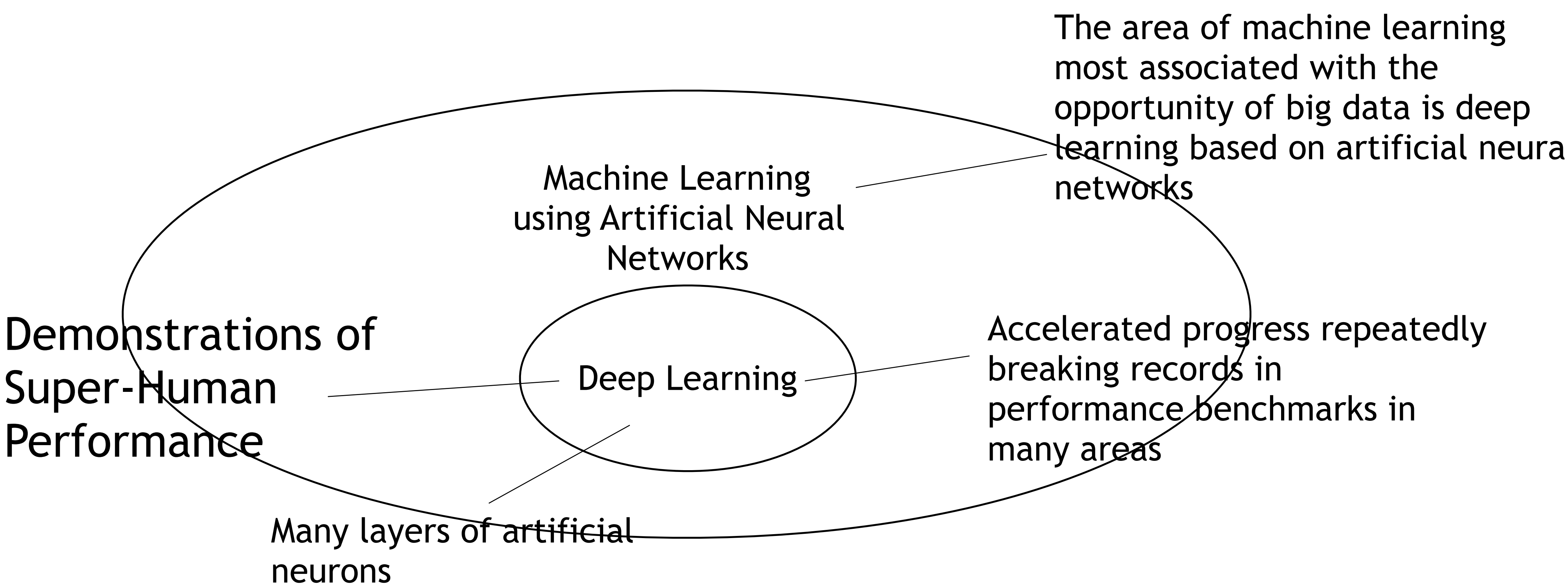
Machine Learning is All Around You, Every Day

- Machine Learning (From A. Ng, Coursera course)
  - Grew out of AI
  - New capabilities for computers
  - You probably use machine learning many times a day without even realizing it (Google search, product recommendations, advertising)
- Examples:
  - Database mining
    - Large data from growth of automation/web
    - E.g. : Web click data, medical records, biology, engineering
  - Applications we can't program by hand
    - E.g.: Autonomous vehicles, handwriting rec, speech rec, NLP, computer vision
  - Self-customizing programs
    - E.g.: Netflix, Amazon product recommendations
  - Understanding human learning
    - E.g.: Modeling the human brain, real AI
  - Deep learning – approaching or exceeding human performance

Opportunities in Machine Learning with Artificial Neural Networks

- **Great advances are being made in deep learning**

Artificial neural networks are networks of simple representations of neurons.



Some of the Recent Successes of Deep Learning

- Super-human performance reading street signs
- Beating a top human player in the game of Go
- Beating previous performance by training an image recognition network with over 100 layers
- Human parity in recognizing conversational speech
- End-to-end training of state-of-the-art question answering in natural language
- Substantial improvement in naturalness of speech synthesis
- Approaching the accuracy of average human translators on some datasets

Deep learning is beginning to meet the grand challenge of AI: Demonstrate human-level performance on tasks that require intelligence when done by humans.

It is important to do it right!

CMU in the news

The New York Times

New Research Center to Explore Ethics of Artificial Intelligence

By JOHN MARKOFF NOV. 1, 2016

The Chimp robot, built by a Carnegie Mellon team, took third place in a competition held by DARPA last year. The school is starting a research center focused on the ethics of artificial intelligence.
Chip Somodevilla/Getty Images

Deep learning raises particular issues because it is very difficult to interpret, much less control, what the millions of inner layer nodes represent or what they are doing.

More on this subject later.

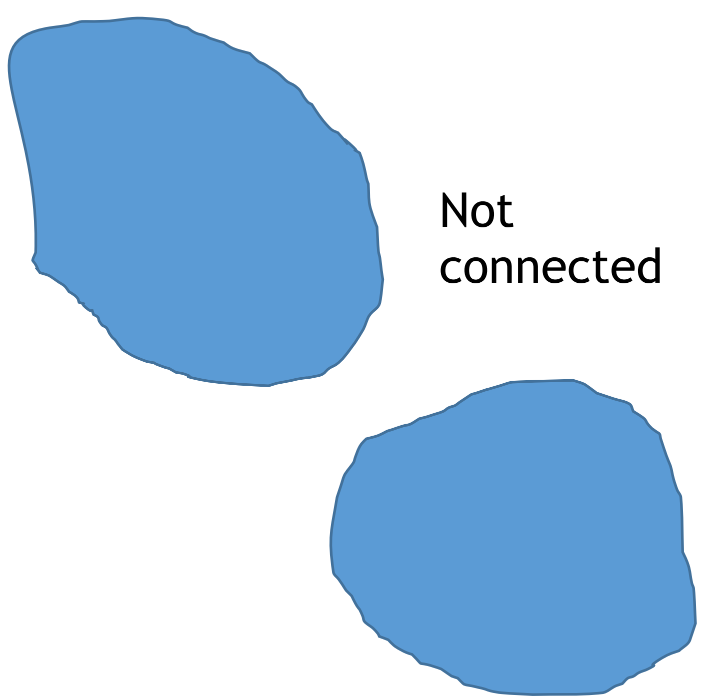
Brief History of Pattern Recognition with Artificial Neural Networks

- 1950s Single neurons (Perceptron) Rosenblatt, *Principles of Neurodynamics*
    - Adaptive learning
- 1960s Single layer of neurons
    - Stochastic gradient descent (perceptron convergence theorem)

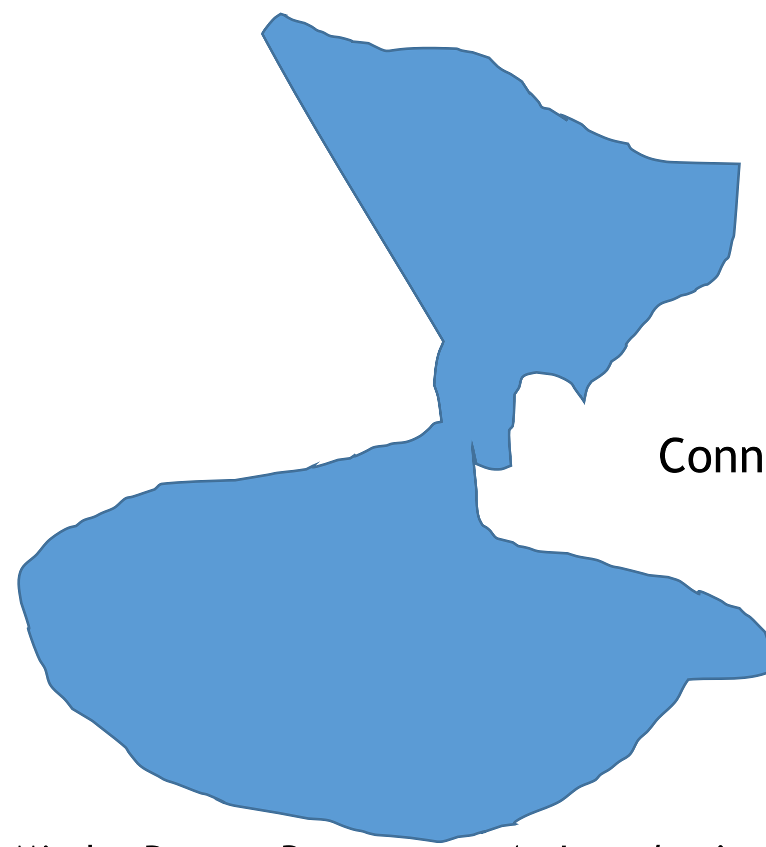
      Minsky, Papert, *Perceptrons: An Introduction to Computational Geometry*
    - Negative result: some things can never be learned with a single layer, no matter how big (e.g. millions in retina)
    - Multiple layers is a hard integer programming problem
- Gap in progress …
- … 1980s and later (continued on a later slide)

Why was there a gap in progress

- Sometimes problems that seem very easy can actually be very hard
- It seems easy to tell at a glance whether two regions are connected



Not connected



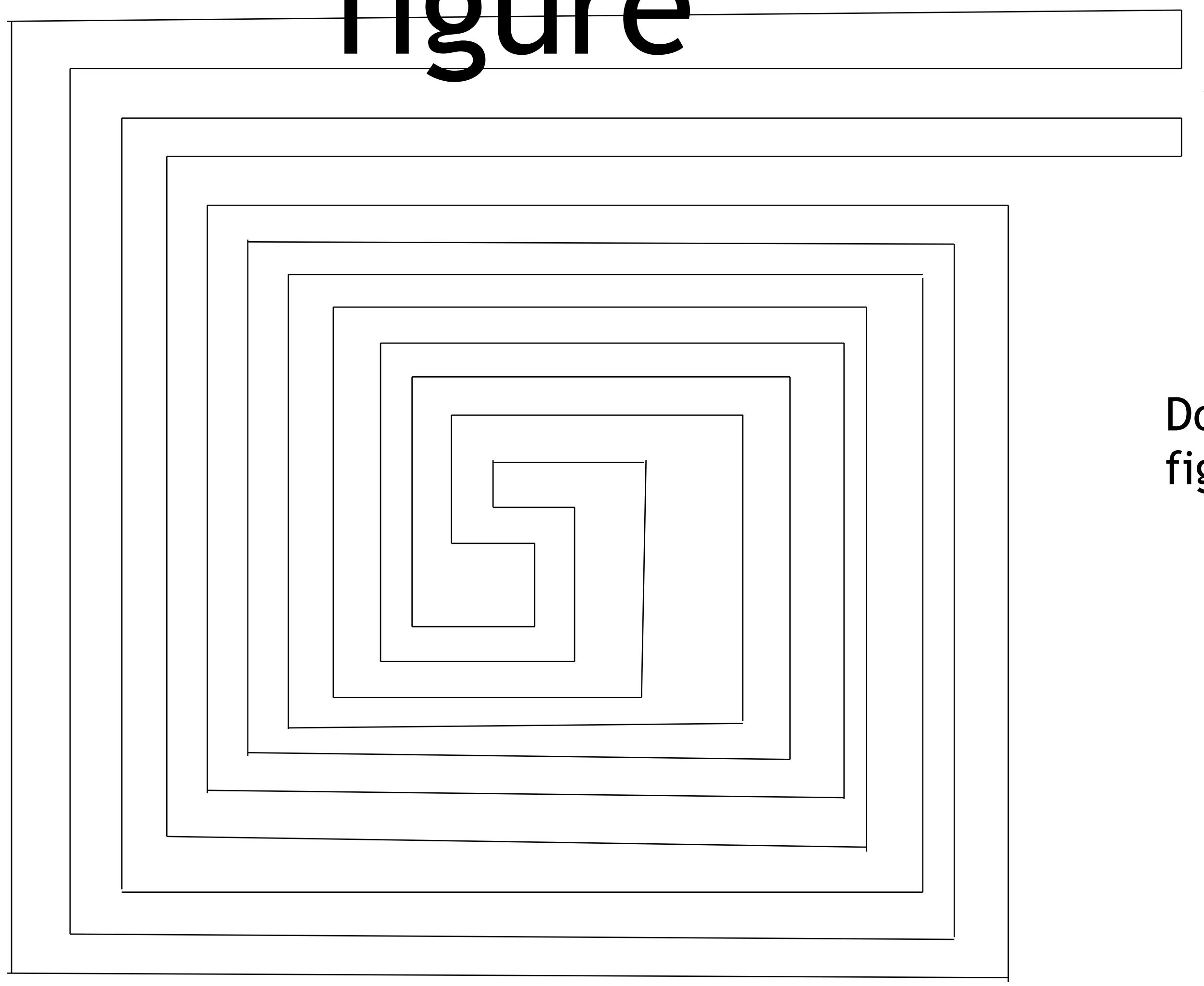
Connected

It is impossible to tell in general, even with an unlimited number of neurons!

Minsky, Papert, *Perceptrons: An Introduction to Computational Geometry*

It looks easy to tell if a region is connected.
Just glance at the figure on the next slide.

Just glance at this figure



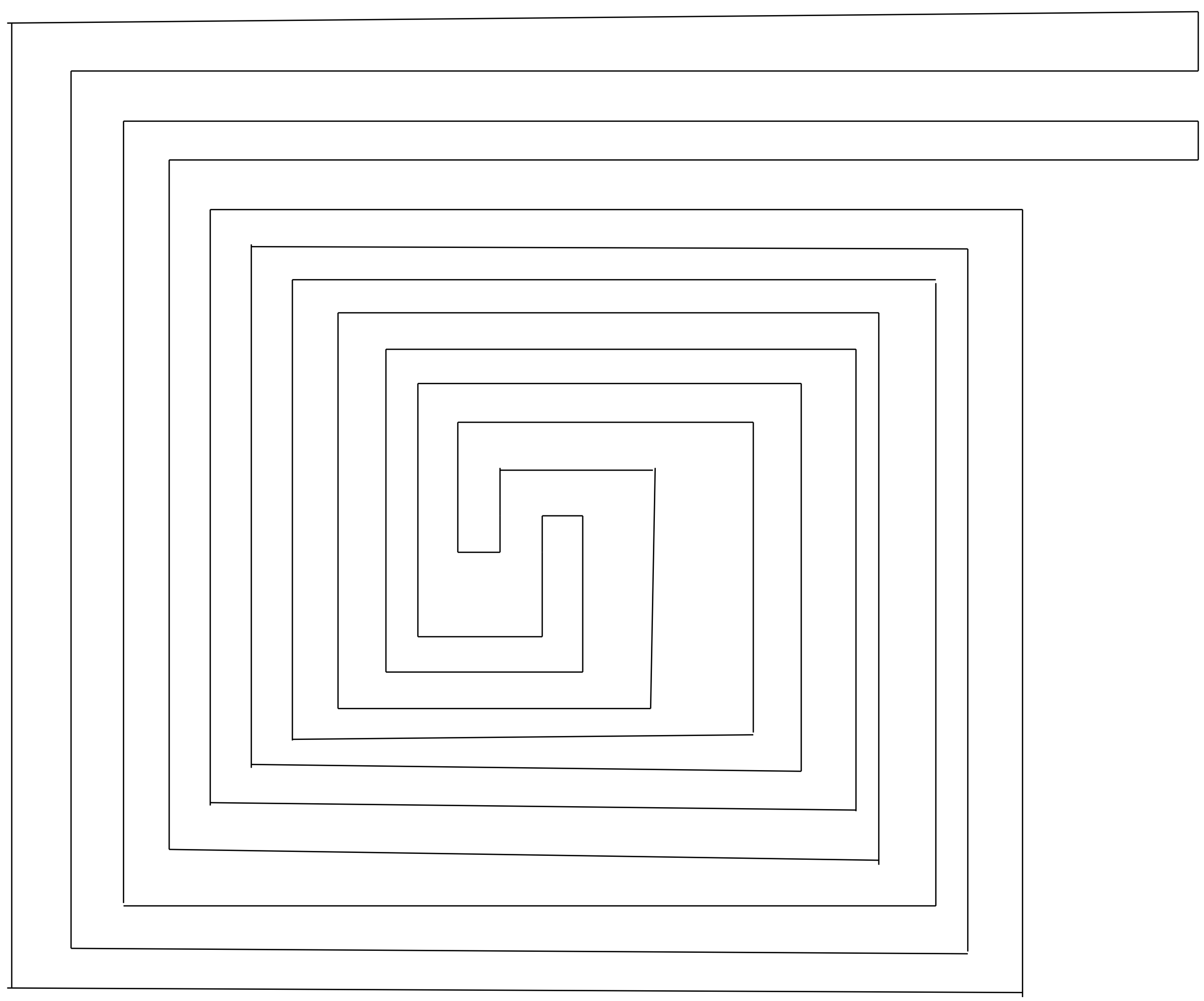
Look here first

Don't try to study the figure.

Was that one snake or two?

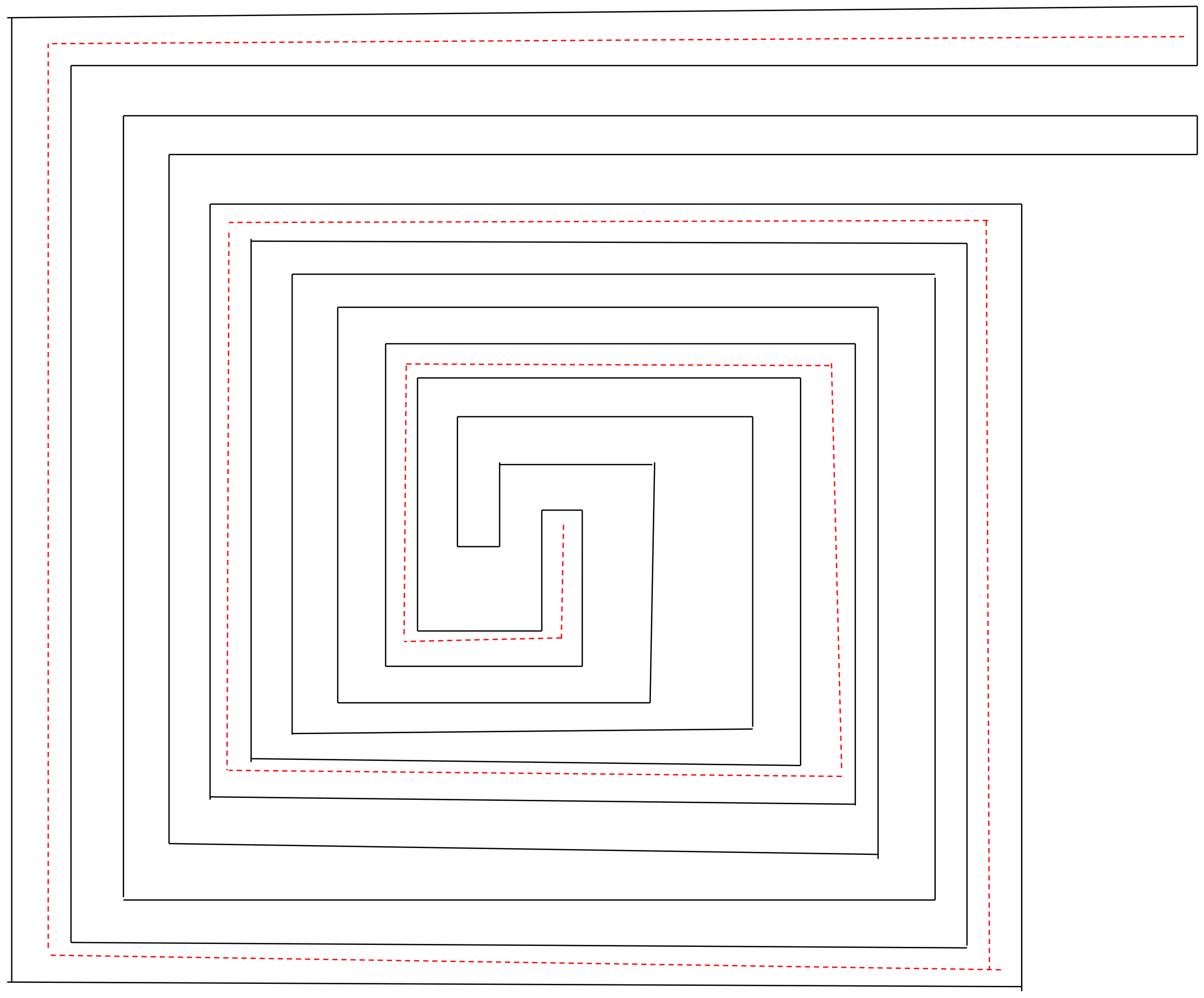
- Are you sure?

One snake or two?



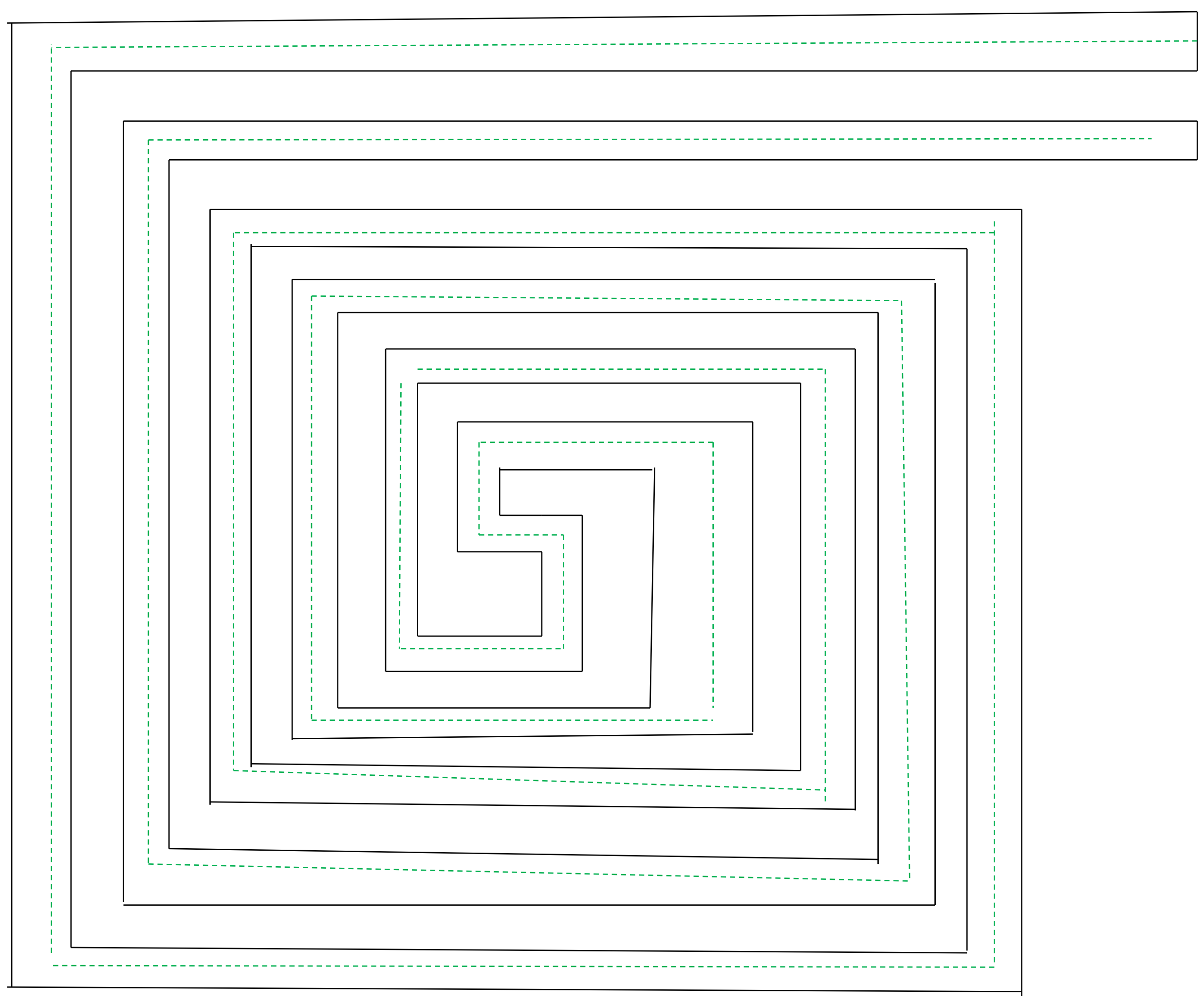
Were you sure?
Take more time.
Did your answer change?

One snake or two?



Did your answer change?
Did you get it right in the first diagram?

But that wasn't the original diagram, this is



This example is in the spirit of the Minsky, Papert book, which was about computer vision. A much simpler example is that the Boolean function Xor cannot be represented with a single layer of perceptrons.

Do you still think you got the original problem correct?
One snake or two?

Brief History of Pattern Recognition with Artificial Neural Networks

- 1950s Single neurons (perceptron)
- 1960s Single layer of neurons
- Gap in progress
- 1982: New interest (Hopfield network)
- 1986: Breakthrough: Error backpropagation algorithm
    - Allows an extra layer (a "hidden" layer between input and output)
    - Key insight: Use a differentiable threshold function (sometimes problems that seem hard are easy)

J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", *Proceedings of the National Academy of Sciences of the USA*, vol. 79 no. 8 pp. 2554–2558, April 1982.

Rumelhart, D., Hinton, G., Williams, R., Learning representations by back propagating errors, Nature vol. 323, 9 October 1986.

Brief History of Pattern Recognition with Artificial Neural Networks

- 1960s Single layer of neurons
- 1986: Backprop: One hidden layer
    - Many successes, but it was difficult to train more than one hidden layer
    - Other machine learning algorithms eventually beat benchmarks set by ANNs
- 1990s – 2006: Research continued, but progress slowed

LeCun,Y.etal. Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation 1:(4)-541-551, 1989.
(Convolutional neural networks)

Hochreiter, S. & Schmidhuber, J. Long short-term memory. Neural Comput. 9, 1735–1780 (1997). (Recurrent neural networks and LSTM)

LeCun, Buttou, et al, Effiicient Backpropagation, in Orr and Muller, *Neural Networks: Tricks of the Trade*, 1998
(Various tricks, including how to initialize the weights)

- 2006: Breakthrough: Efficiently training multiple hidden layers

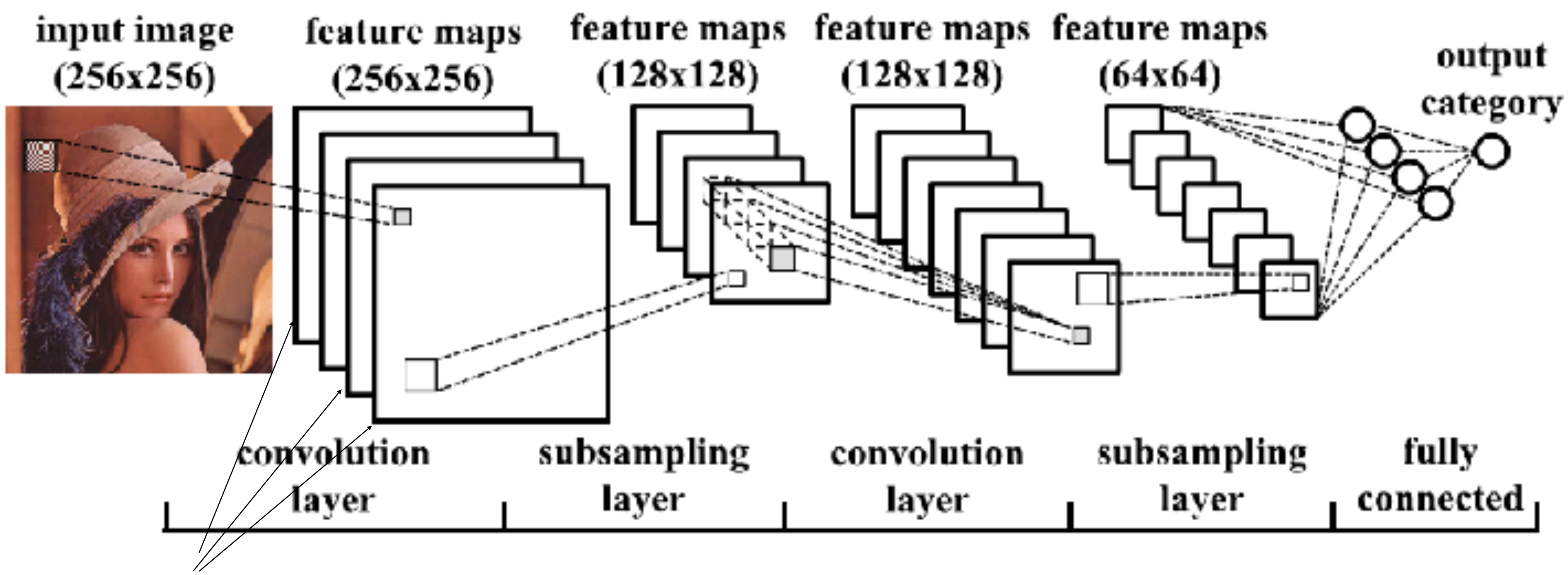
Hinton,G.E.,Osindero,S.&Teh,Y.-W.A fast learning algorithm for deep belief nets. Neural Comp. 18, 1527–1554 (2006).

Artificial Neural Networks 1986 - 2006

- ANNs set several new pattern recognition benchmarks
- Innovations continued (convolutional neural nets, recurrent networks, LSTM)
- But, new methods (SVMs, random forests) began having higher performance than ANNs)
- Although backpropagation can be done with multiple hidden layers, there was little success applying it (slow convergence, problems with local minima, overfitting)
- Progress slowed, but didn’t stop

Architecture of Convolutional Neural Network

Suggested by processing in real eyes and brains.
Greatly reduces the amount of computation required
to train very large networks.



The same weights are used, shifted in position. Thus the output is the input convolved with the weights.

Because the weights are shared, there is more data per update estimate. Also, there is less memory required, so a larger network fits into RAM. There is also somewhat less computation.

The subsampling reduces both the number of nodes and the number of weights.

Fast Training for Deep Belief Nets - 2006

Game changing result: Launched the era of deep learning

- Unsupervised training – one layer at a time
  - Unsupervised training allows one layer at a time
- Requires special architecture
  - Top two layers form an undirected associative memory
- Efficiently trained nets with many layers and millions of nodes
- After unsupervised training of all layers, do an up-down pass of supervised training
- Achieved 20 year goal of efficient multi-layer training for large networks

Hinton,G.E.,Osindero,S.&Teh,Y.-W.A fast learning algorithm for deep belief nets. Neural Comp. 18, 1527–1554 (2006).

Training Deep Learning Nets – 2006+

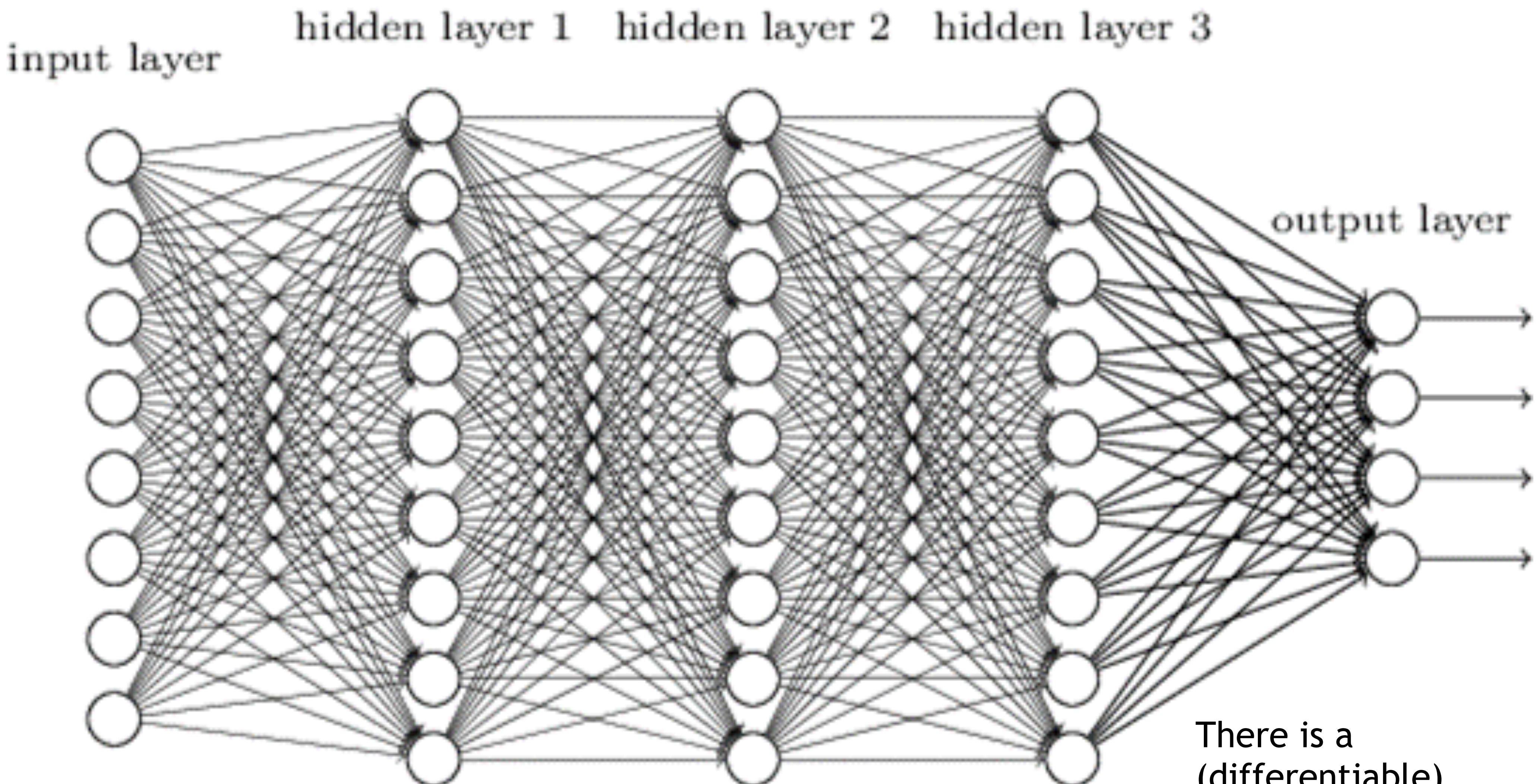
- First, it turned out that the special architecture was not required
  - Other methods of unsupervised training to get the initial weights for multi-layered feedforward nets, followed by supervised training with backprop were also successful
- Gradually, it became clear that even the initial unsupervised training was not essential, other fairly simple ways were found to get adequate initial weights

LeCun, Bottou, et al, Efficient Backpropagation, in Orr and Muller, *Neural Networks: Tricks of the Trade*, 1998
(Various tricks, including how to initialize the weights)

Glorot, Bengio, Understanding the difficulty of training deep feedforward neural networks, AISTATS, 2010

- What did make the difference?
  - Large networks, very large amounts of data, very large amount of computation
  - 1980-90s computers were not fast enough and did not have enough memory

Deep neural network



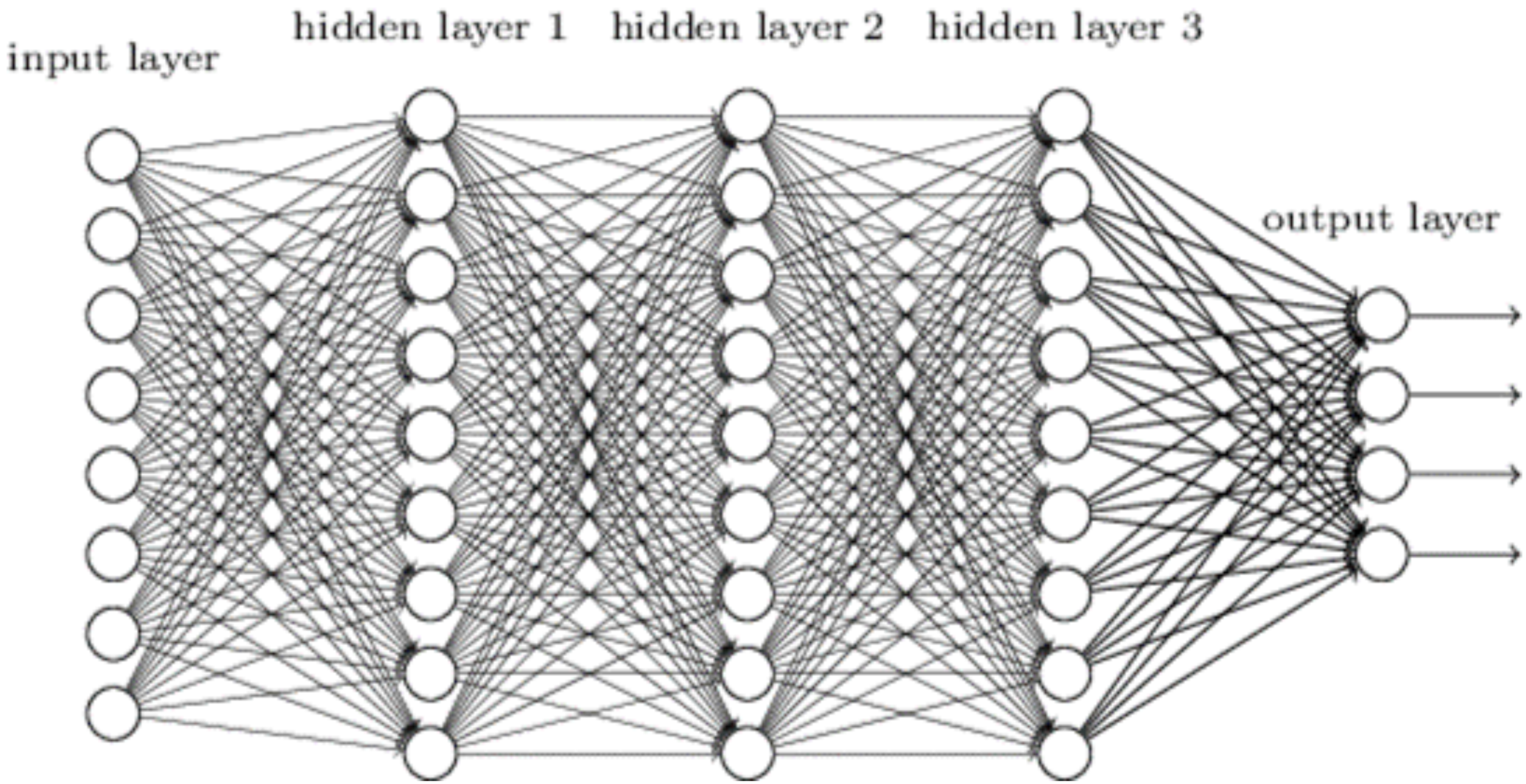
There is a (differentiable) function that measures the discrepancy of the actual output from the desired output.

Each arc has a weight w,
which is multiplied by its input.

Each node generates an output that is a differentiable function of the sum of its inputs.

Forward computation: The computation of the output of each layer of nodes proceeds from left to right.

Deep neural network



Backpropagation: The computation of the derivative of the error function with respect to the weights proceeds backwards. (This is just the ordinary chain rule of elementary calculus.) Make an incremental update to each weight proportional to minus the derivative.

Rumelhart, D., Hinton, G., Williams, R., Learning representations by back propagating errors, Nature vol. 323, 9 October 1986.

Training a deep neural network

- It is (almost) as easy as it looks (I have left out some details)
  - Just do the {feedforward, backprop, update weights} computation for each item of training data (an epoch), and then repeat epochs until convergence
- But, it requires a lot of computation
  - Millions of nodes, billions weights, thousands of epochs, and as many data items per epoch as possible (sometimes millions)
- Fortunately, it is easy to implement for parallel computation
  - Implementation on GPUs typically speeds up the computation by two orders of magnitude

Other Issues (with some solutions)

- With the very large number of parameters, there is always a danger of overfitting the training data
  - Several things can reduce the amount of overfitting
  - One of the best is dropout; For each data item, randomly pick some of the nodes to “dropout” and not participate
- Some large problems still require too much computation for general purpose networks (e.g. computer vision, speech recognition)
  - But they have a repetitive specialized structure: use convolutional neural nets
- Some problems require learning sequences (number grows exponentially with length)
  - Use recurrent neural nets to track the sequences (with LSTM)

- There are other issues which remain as problems; they will be discussed later
  - Vanishing gradient, overfitting, degradation with more layers, non-interpretablility, knowledge not explicit, non-use of domain-specific knowledge

Some of the Recent Successes of Deep Learning (Short List)

- Super-human performance reading street signs
- Beating a top human player in the game of Go
- Beating previous performance by training an image recognition network with over 100 layers
- Human parity in recognizing conversational speech
- Substantial improvement in naturalness of speech synthesis
- Distilling the knowledge of a large number of networks into a single network of the same size

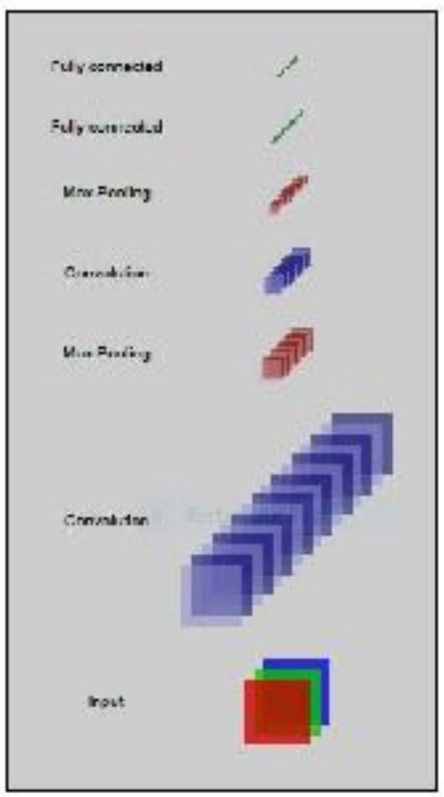
Deep learning is beginning to meet the grand challenge of AI: Demonstrate human-level performance on tasks that require intelligence when done by humans.

Multi-Column Architecture

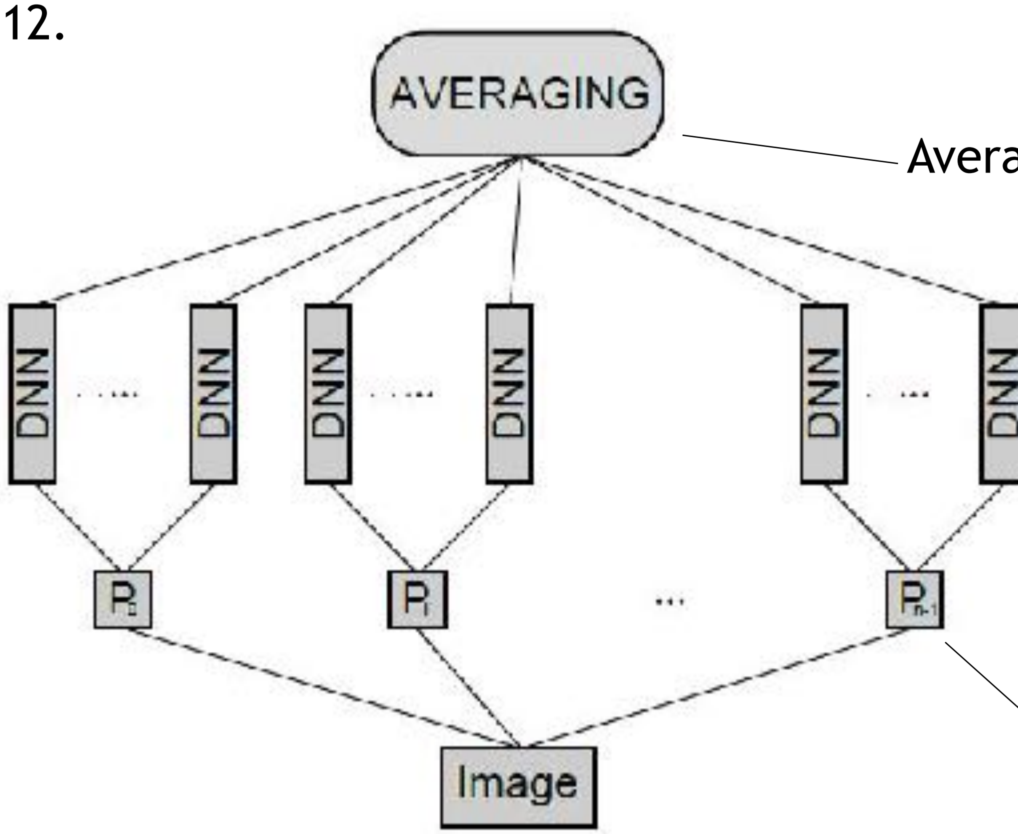
(On traffic signs, outperforms humans by factor of two)

Ciresan, Meier, Masci, Schmidhuber; Multi-column deep neural network for traffic sign classification; 2012.

One DNN from the Multi-Column



(a)

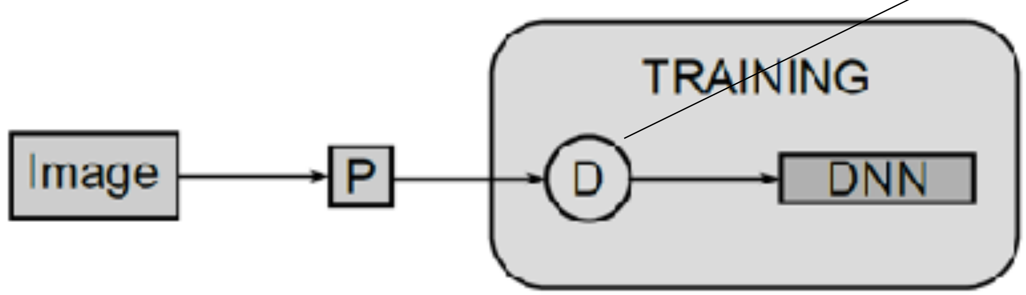


(b)

Averaging an ensemble

What's new?
Multiple preprocessing methods and distortions; Averaging an ensemble.

Various forms of pre-processing and distortions



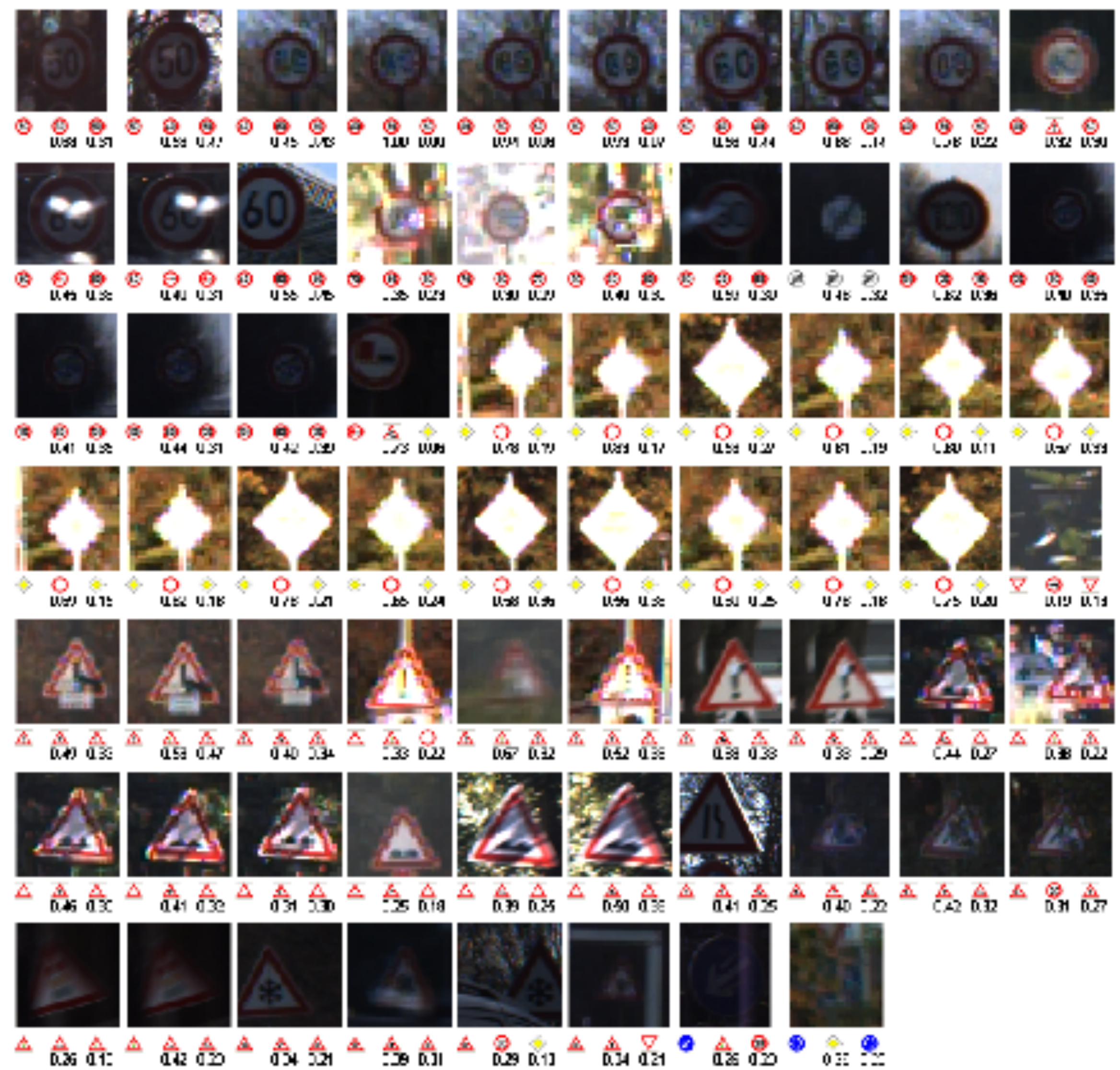
(c)

Traffic Signs Dataset

Ciresan, Meier, Masci, Schmidhuber; Multi-column deep neural network for traffic sign classification; 2012.

- 12569 images, only 68 errors
  - Human error rate
    twice as high
  - Second best algorith
    3 times as many errors

Here are the 68 images it missed



Machine Learning and Games

- Perfect information games (like checkers, chess, Othello, and Go) can be represented as a tree, with a node for each possible position and a branch for each possible move from that position
- If the tree is too large for exhaustive search:
  - Define a *policy function* setting the probability distribution among the possible moves (cuts done on the effective number of branches at each node)
  - Define a *value function* to give an estimated value for each node when the search is terminated before end
- Most successful game algorithms use Monte Carlo tree search (MCTS)
  - The computer plays games against itself and keeps a tree representing all games played
  - As more games are played, the value function becomes more accurate
  - The policy also improves by selected children with higher values
  - Further enhanced by policies that attempt to match human experts
  - Prior art: Shallow policies or value functions based on linear combination of input features

Alpha Go

Silver, Huang, et al, Hassabis, mastering the game of Go with deep neural networks and tree search, Nature VOL 529, 28 January 2016

- Three deep neural networks: SL policy, RL policy, RL value
  - The architecture of each of the networks is a 2-d convolutional neural network based on the 19x19 grid of the Go board
- Stage one: The SL policy network is trained to imitate the play of professional Go players, using supervised learning
- Stage two (MCST): The algorithm plays games against itself and trains the RL policy and RL value networks using reinforcement learning

Probabilistic sampling of the search tree based on the value function.

Using deep learning to train the value function rather than simple linear combination of features.

The New York Times

ASIA PACIFIC

Google's Computer Program Beats Lee Se-dol in Go Tournament

By CHOE SANG-HUN MARCH 15, 2016

Lee Se-dol with his daughter Lee Hye-lim on his way to the last Go match with Google's AlphaGo artificial intelligence program in Seoul, South Korea. Kim Hong-Ji/Reuters

News about AlphaGo, (Silver, Huang, et al, Hassabis, mastering the game of Go with deep neural networks and tree search, Nature VOL 529, 28 January 2016)

Milestone: Achieving Human Parity in Conversational Speech Recognition

Xiong, et al, Zweig, Achieving Human Parity in Conversational Speech Recognition, Microsoft Technical Report MSR-TR-2016-71

Table 7. Word error rates (%) on the NIST 2000 CTS test set with different acoustic models. Unless otherwise noted, models are trained on the full 2000 hours of data and have 9k senones. Our automated system makes about a dozen fewer errors than people on the SWB set, not visible below due to rounding.

Model	N-gram LM		RNN-LM		LSTM-LM	
	CH	SWB	CH	SWB	CH	SWB
300h ResNet	19.2	10.0	17.7	8.2	17.0	7.7
ResNet GMM alignment	15.3	8.8	13.7	7.3	12.8	6.9
ResNet	14.8	8.6	13.2	6.9	12.5	6.6
VGG + ResNet	14.5	8.4	13.0	6.9	12.2	6.4
VGG	15.7	9.1	14.1	7.6	13.2	7.1
LACE	14.8	8.3	13.5	7.1	12.7	6.7
BLSTM	16.6	8.9	15.1	7.4	14.4	7.0
BLSTM 27k senones	16.2	8.7	14.6	7.5	13.6	7.0
BLSTM 27k, spatial smoothing	14.9	8.3	13.7	7.0	13.0	6.7
Final ASR System	13.3	7.4	12.0	6.2	**11.1**	**5.9**
Human Performance	-	-	-	-	11.3	5.9

Each system is a carefully engineered combination of previously successful system components with a few innovations.

Ensemble performance

Matches human performance!

Conclusion: Ensembles win benchmarks

Can you get better learning just by adding more layers?

- Problem: Vanishing gradient
  - After back propagating many layers, the gradient is close to 0
  - This problem was eventually solved (intermediate normalization layers)
- Another problem: With additional layers, accuracy saturates and then rapidly degrades (Why?)
  - Not due to overfitting: performance on training data also degrades
  - (See a solution in next paper)

Deep Residual Learning for Image Recognition

https://arxiv.org/abs/.03385 (He, Zhang, Ren, Sun, Deep residual learning for image recognition, 2015; Building DNNs with many more layers; Winner of ISVRC & COCO 2015 competitions)

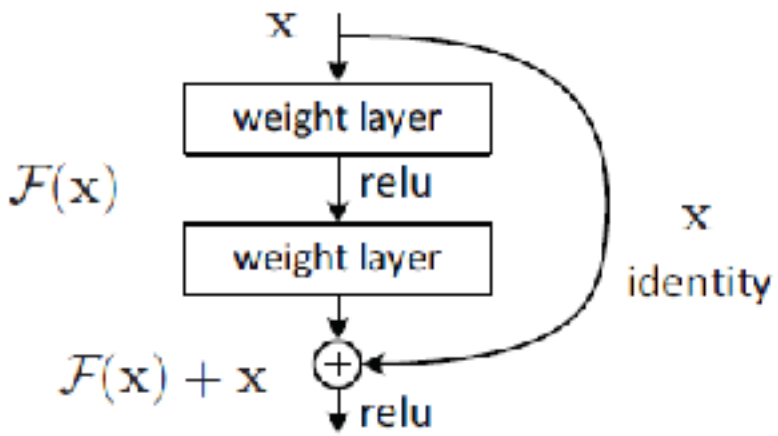
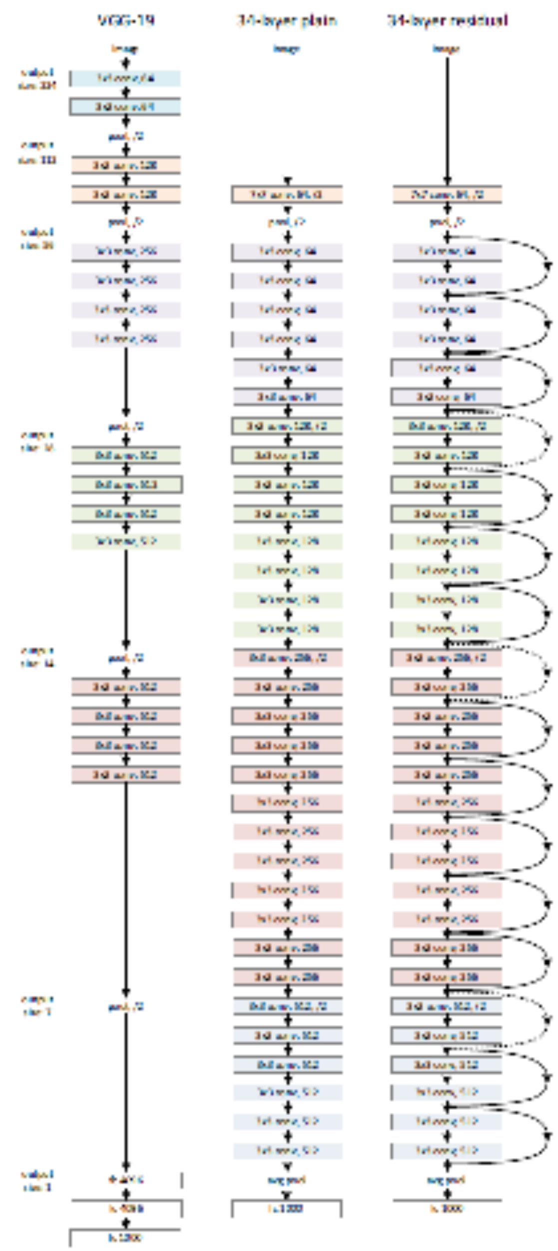


Figure 2. Residual learning: a building block.

Deep Residual Learning for Image Recognition

https://arxiv.org/abs/.03385 (He, Zhang, Ren, Sun, Deep residual learning for image recognition, 2015; Building DNNs with many more layers; Winner of ISVRC & COCO 2015 competitions)

Comparison of three systems, each with many layers.



Deep residual learning allows so many layers that it is difficult to show them on a slide.

Deep Residual Learning for Image Recognition

Deep residual learning wins the 2015 competition.

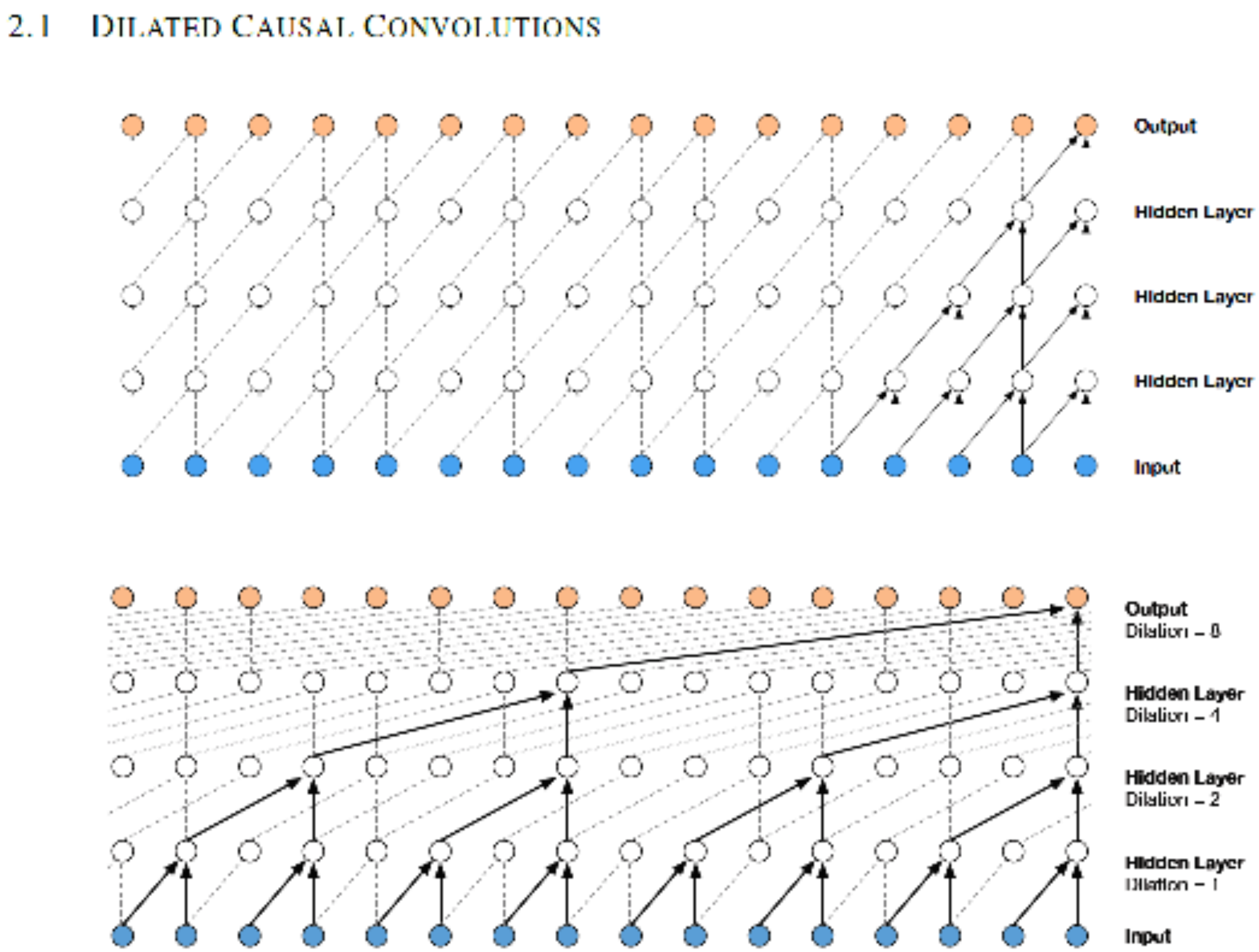
method	top-1 err.	top-5 err.
VGG [41] (ILSVRC'14)	-	8.43†
GoogLeNet [44] (ILSVRC'14)	-	7.89
VGG [41] (v5)	24.4	7.1
PReLU-net [13]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	**19.38**	**4.49**

Table 4. Error rates (%) of **single-model** results on the ImageNet validation set (except † reported on the test set).

Residual learning successfully trained 152 layers.

https://arxiv.org/abs/.03385 (He, Zhang, Ren, Sun, Deep residual learning for image recognition, 2015; Building DNNs with many more layers; Winner of ISVRC & COCO 2015 competitions)

WaveNet: A Generative Model for Raw Audio



Causal convolution

Dilated causal convolution

Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

https://regmedia.co.uk/2016/09/09/wavenet.pdf (van den Oord, et al, WaveNet: A Generative Model for Raw Audio, DeepMind, 2016)

WaveNet: A Generative Model for Raw Audio

https://regmedia.co.uk/2016/09/09/wavenet.pdf (van den Oord, et al, WaveNet: A Generative Model for Raw Audio, DeepMind, 2016)

2.4 RESIDUAL AND SKIP CONNECTIONS

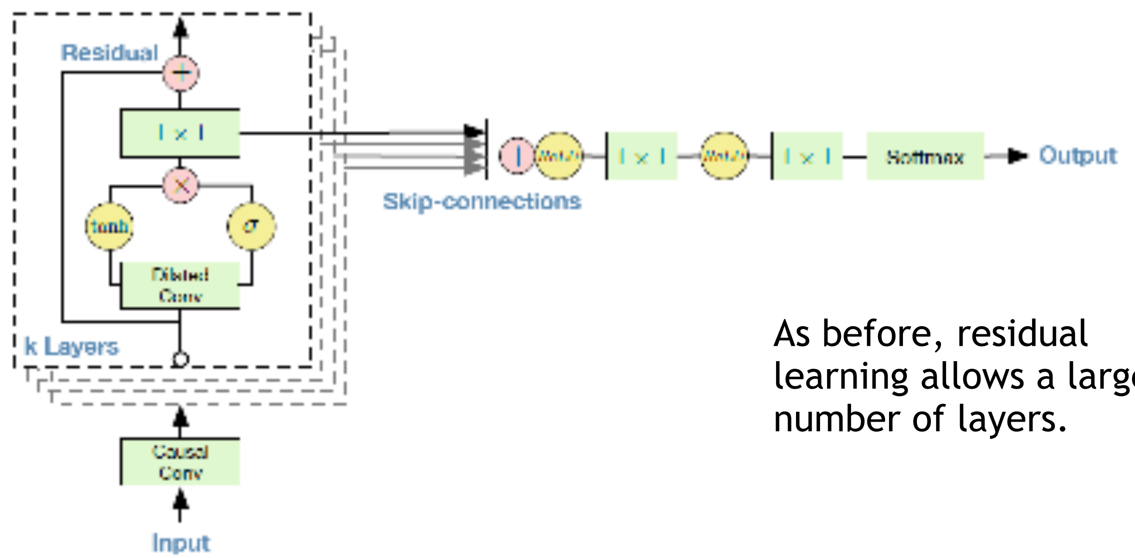


Figure 4: Overview of the residual block and the entire architecture.

As before, residual learning allows a large number of layers.

With dilated causal convolution, using residual learning to enable training many layers, WaveNet is able to produce synthetic speech that sounds much more natural than any previous systems.

Dilated causal convolution and residual learning.

A Sample of Handwritten Digits (MNIST)

Soft Decisions (Hinton, 2015)

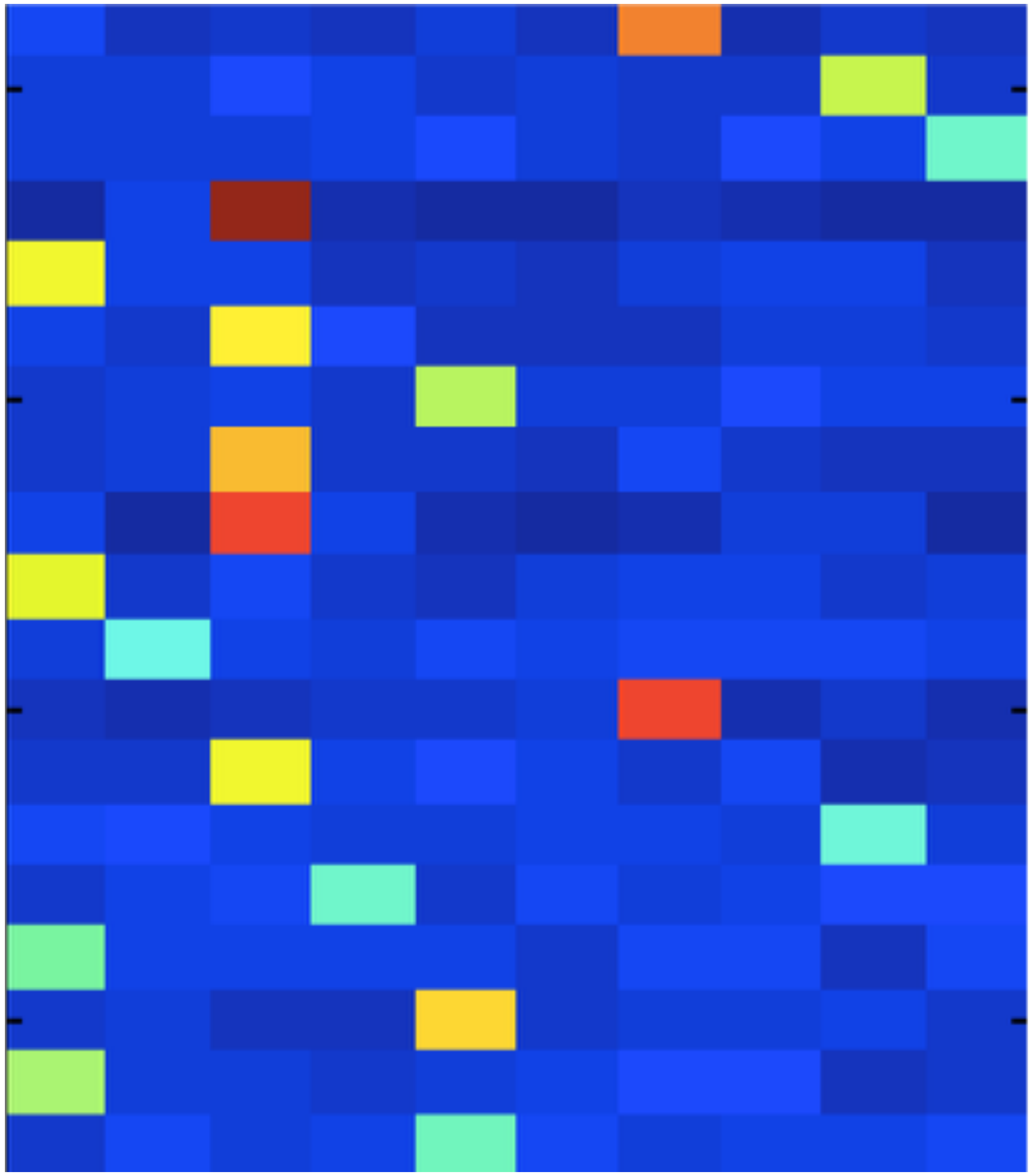
The “dark knowledge” is the knowledge available from the 2nd best and other scores. However, the scores need to be “softened” because the ensemble is too confident in the right answer.

The blue regions all look black with normal “hard” scoring. The extra knowledge is in these “dark” regions.

this 2 resembles a 1 and nothing much else

this 2 resembles 0, 3, 7, 8

this 2 resembles 4 and 7



https://arxiv.org/abs/1503.02531
(Distilling the Knowledge of a Neural Network, Hinton, 2015, Uses MNIST as an example)

Distillation of Knowledge from an Ensemble to a Single Network

- Train the ensemble
- Average a soften version of the output of each member of the ensemble
- Use this average as the objective for training a single network

Uses output of ensemble as supervision for a single network; Softens the output before averaging.

- Result: The single network is much closer to the performance of the ensemble than to the performance of a conventionally trained single network

https://arxiv.org/abs/1503.02531 (Distilling the Knowledge of a Neural Network, Hinton, 2015, Uses MNIST as an example)

About the Course: An Introduction to Human-Aided Deep Learning

- This is a reading and research course
  - That means that you will read state-of-the-art papers like those I have summarized and present them to your fellow students
  - We will begin with simpler papers providing background in the techniques
  - You will also have projects implementing these techniques
  - You will eventually implement a state-of-the-art benchmark (up to the capacity of our computing facility)
  - You will also have an opportunity to go beyond

Some Remaining Problems

- Deep learning systems lack the wisdom of Socrates
  - "The only thing I know is that I don't know anything."
- They are mysterious
  - It is difficult or impossible to know what the nodes and weights of inner layers represent
- End-to-end training with no supplied expert knowledge is a major AI milestone
  - But it is also a major weakness and limitation
- Ethical issue
  - How can we control systems if we don't know what they are doing, and they don't take advice or guidance?

Over confidence.

Non-transparency

The Missing Ingredient: Human Knowledge

- Dilemma: How can we give advice or control deep neural nets if we can't what the node activations and connections weights mean?

- Idea: deep learning networks are good at learning many different things. Why not use a deep learning network to learn how to communicate with deep learning networks?

- Introducing the concept of a Socratic coach: A Socratic coach is a second deep learning system associated with a primary deep learning system. However, rather than studying the primary data, the Socratic coach studies the primary deep learning system itself.
- This concept changes the game.

How the Objective of the Game Changes

- Being able to learn things on their own is one of the major achievements of deep learning systems. Does assistance from humans undercut that achievement?
- In my opinion, if we can use machine learning to facilitate communication with end-to-end trained machines, we will have added to the achievement. The objective becomes the performance of the combined system.
- The Socratic coach automates the task of a machine learning researcher. That is, it does a task requiring intelligence better than a human can do it.

How the Tactics of the Game Change

- Using an outside expect, the Socratic coach, that acquires knowledge about the primary machine learning system greatly facilitates development of improvements in the primary system.
    - The Socratic coach learns to understand the primary system in ways that the primary system itself can't even represent.
    - The Socratic coach can automate development testing, doing many more experiments than could be done by hand.

How the Game Changes for You

- Everything that you learn about deep learning can be applied to designing and developing Socratic coaches, which can in turn be used to help develop better primary systems.
- You will immediately be working at the cutting edge of new developments.

- There may be an opportunity to put this to practice in a follow-on course or as an intern for a start-up.

Take Action

- If you are excited about deep learning or about this opportunity to be at the cutting edge, please take the course 11-364: Introduction to Human-Aided Deep Learning

- In any case, best wishes to you and thank you for your attention.

S-17 --11-364:An Introduction to Human-Aided Deep Learning and Socratic Coaches

- You will read papers like those that I have discussed
  - You will present summaries of these papers to your peers
  - These papers are not be easy to read.
    - They assume a lot of prior knowledge from other papers.
- You will implement at least one of the systems
- You will replicate state-of-the-art results

This will be a challenging course, but you will learn a lot. I hope you will learn more than from any normal course. How much you learn will be up to you.

- Potential follow-on: Implementing ideas never tried before and/or interning with a start-up