University of Nottingham
UK | CHINA | MALAYSIA

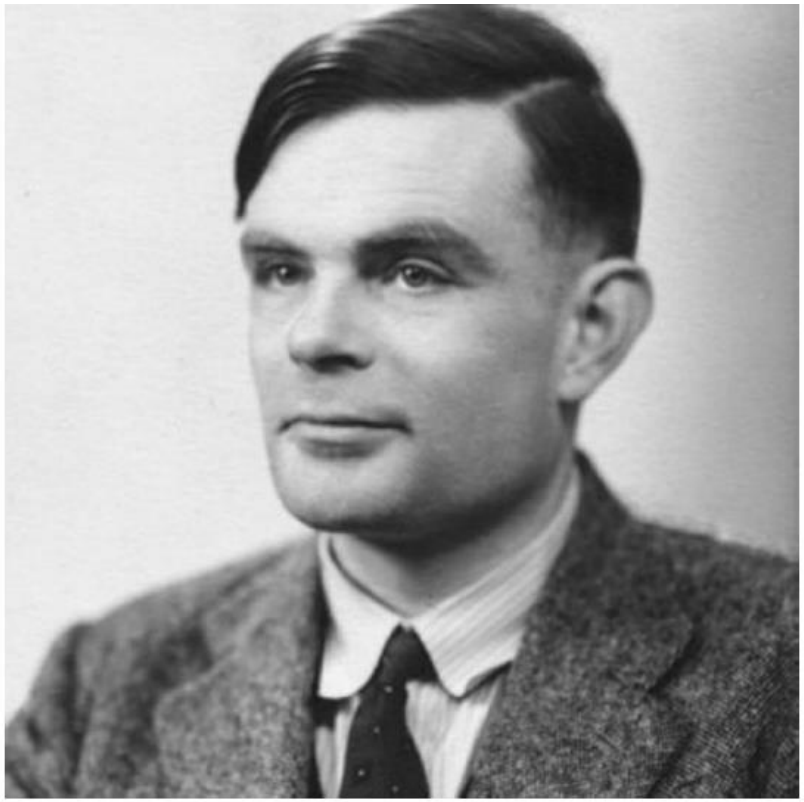
COMP3055 Machine Learning

Topic 14 – Deep Learning - Basics

Zheng Lu

2024 Autumn

History of Deep Learning Ideas



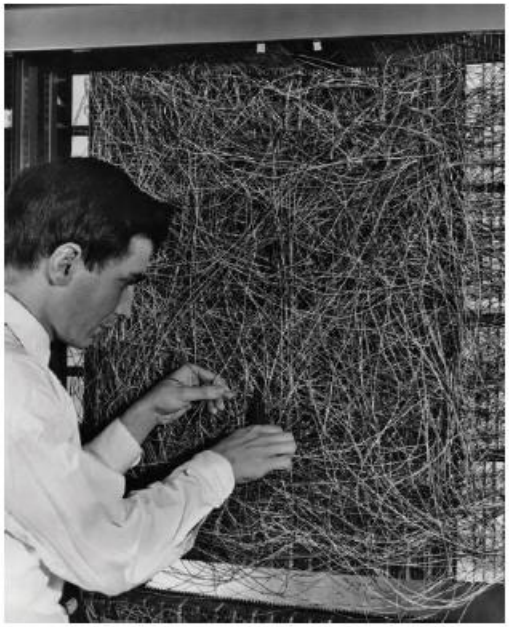
Dreams, mathematical foundations, and engineering in reality.

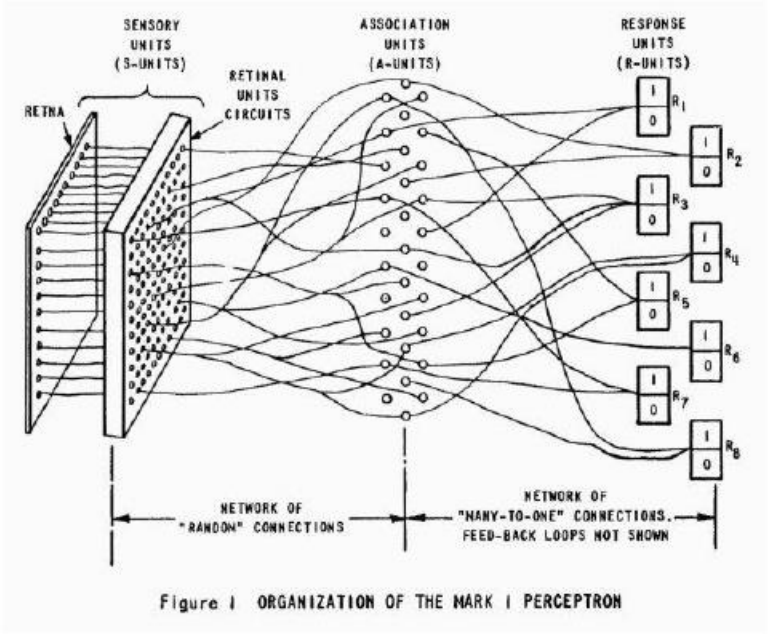
Alan Turing, 1951: "It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. They would be able to converse with each other to sharpen their wits. At some stage therefore, we should have to expect the machines to take control."

Milestones

- 1943: Neural networks
- 1957-62: Perceptron
- 1974-86: Backpropagation, RBM, RNN
- 1989-98: CNN, MNIST, LSTM, Bidirectional RNN
- 2006: “Deep Learning”, DBN
- 2009-12: ImageNet + AlexNet
- 2014: GANs
- 2016-17: AlphaGo, AlphaZero
- 2017-19: BERT, Transformers

Milestones



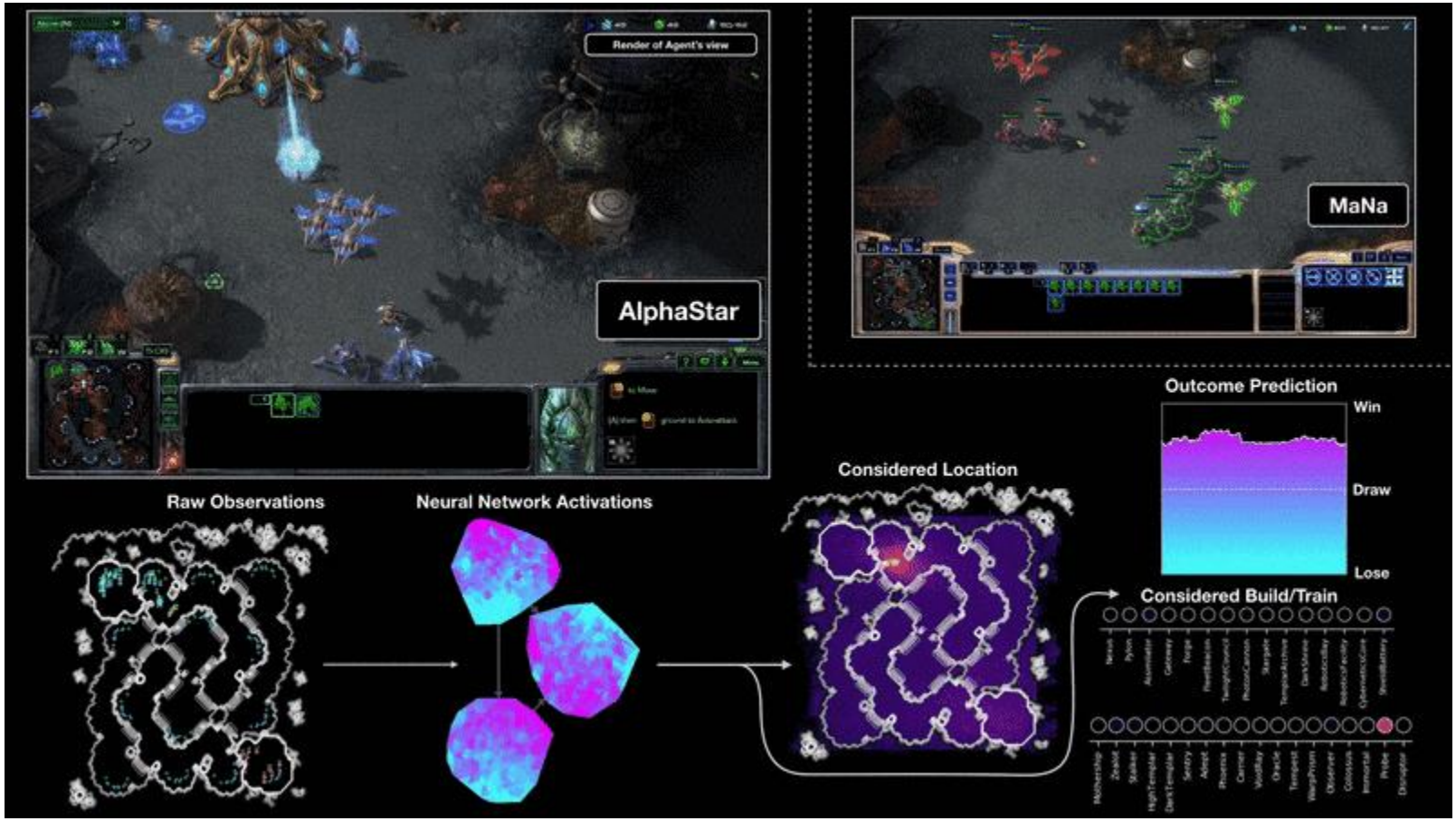


Frank Rosenblatt, Perceptron (1957, 1962): Early description and engineering of single-layer and multi-layer artificial neural networks.

Milestones

Lee Sedol vs AlphaGo, 2016

Milestones



- AlphaStar beats MaNa , one of the world's strongest professional StarCraft players, 5:0.

Turing Award for Deep Learning

- Yann LeCun
- Geoffrey Hinton
- Yoshua Bengio

Turing Award given for:

- "The conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing."

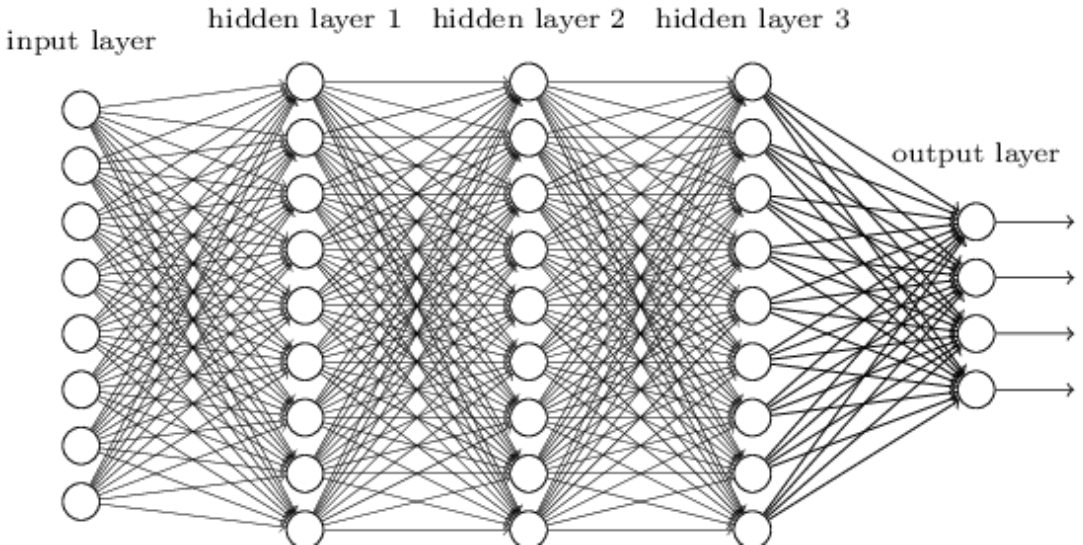
What is Deep Learning?

- Most machine learning methods work well because of human-designed input features or representations
  - Our job is to find the best features to send to learning techniques such as SVM.

- Machine learning becomes just optimizing weights to best make a final prediction.

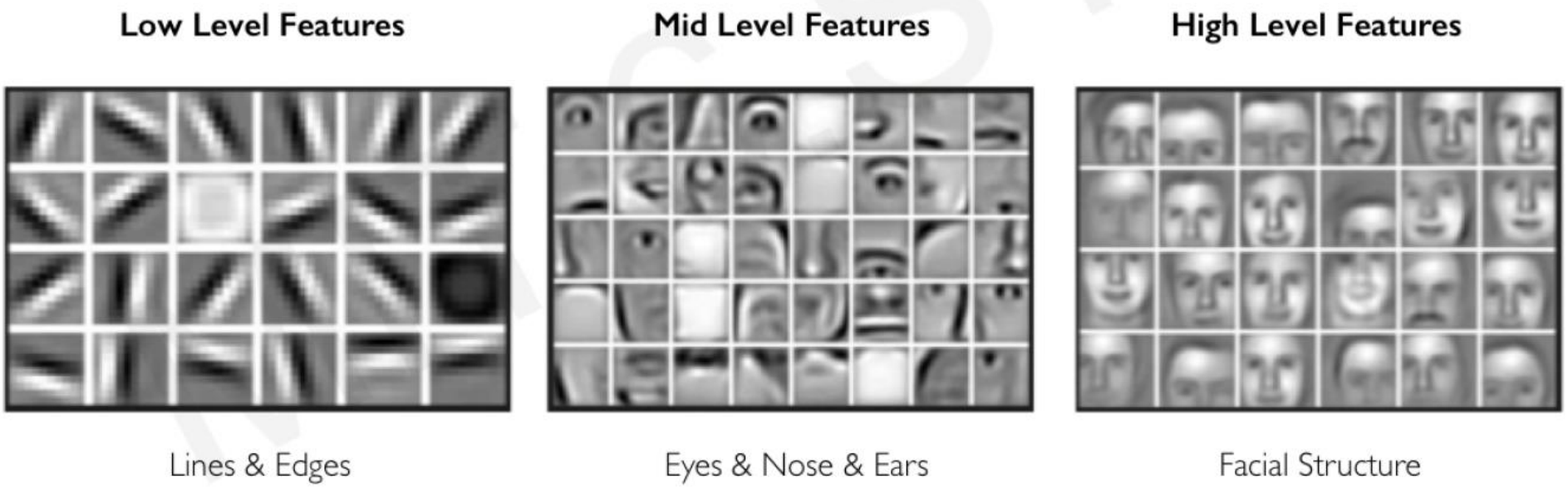
What is Deep Learning?

- In contrast to standard machine learning
- **Representation learning** attempts to automatically learn good features or representations
- **Deep learning algorithms** learn multiple levels of representations (here: hidden layer 1, 2, 3) and an output layer
- From "raw" inputs x (e.g. sound, pixels, characters, or words)
- **Neural networks** are the currently successful method for deep learning
- A.k.a. "**Differentiable Programming**"



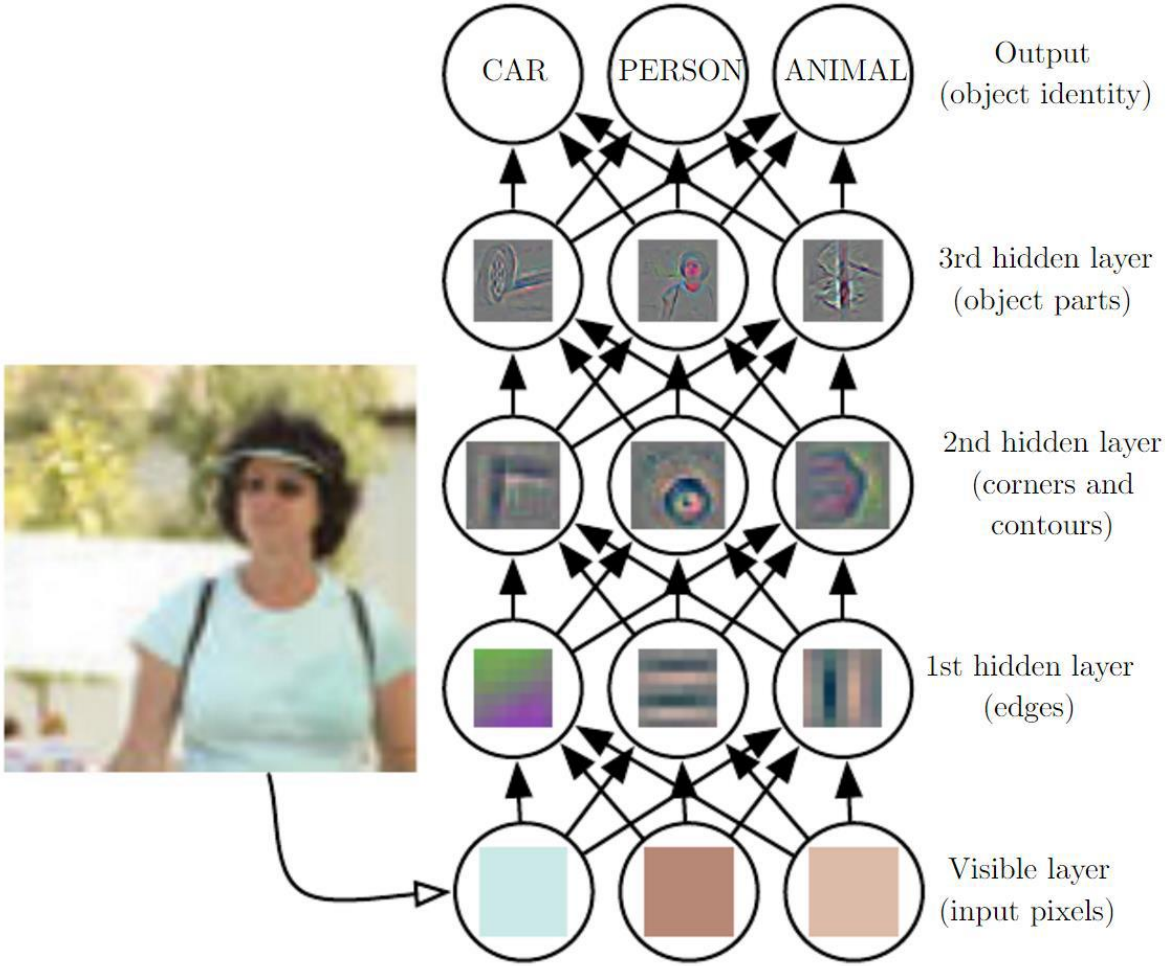
Why Deep Learning?

- To learn the underlying features directly



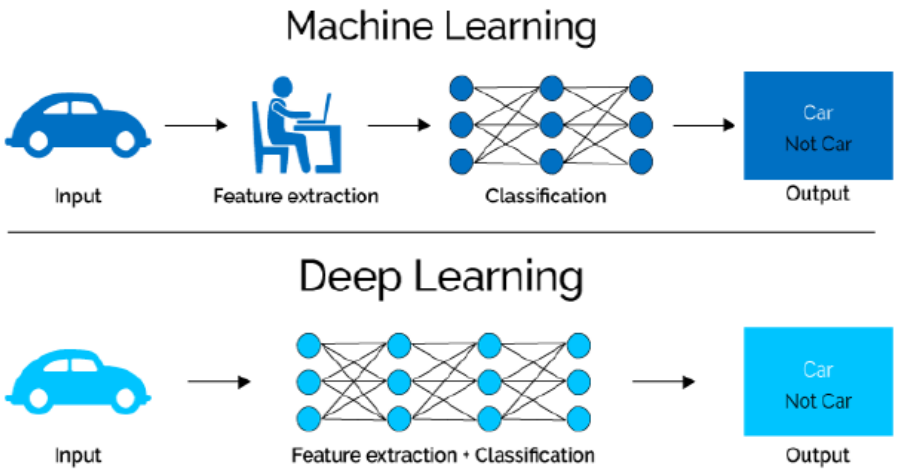
Why Deep Learning?

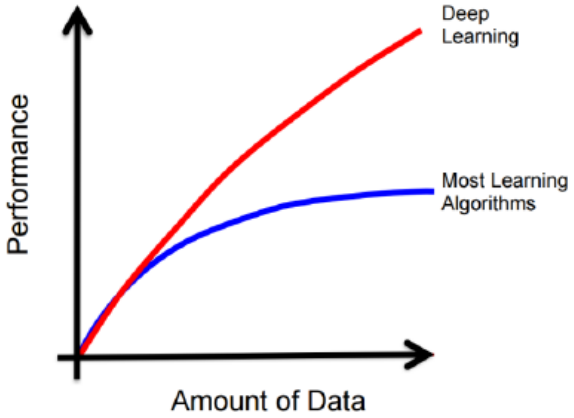
- To learn the underlying features directly



Why Deep Learning?

- Scalable machine learning





Why Now?

- Big data
  - Easier collection and storage
  - Larger datasets (ImageNet, COCO)

- Hardware
  - Graphics Processing Units (GPUs)
  - Massively parallelizable

- Software
  - Toolboxes
  - Powerful framework

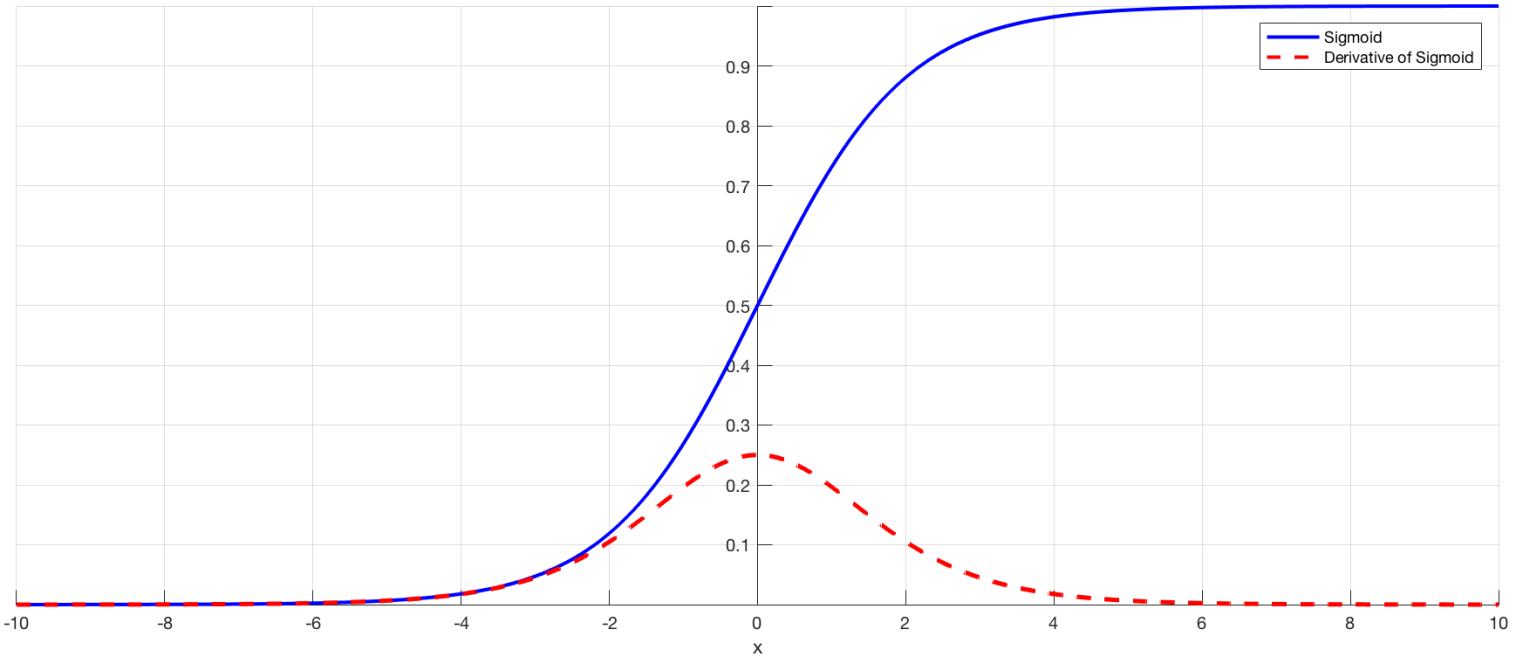
Why Deep

- More sophisticated models
  - Learn more good features (representations)
  - So better classification results

Problems with Going Deeper

- Vanishing gradients
  - As more layers using certain activation functions are added to neural networks, the gradients of the loss function approaches zero, making the network hard to train
  - Certain activation functions, like the sigmoid function, squishes a large input space into a small input space between 0 and 1. Therefore, a large change in the input of the sigmoid function will cause a small change in the output. Hence, the derivative becomes small

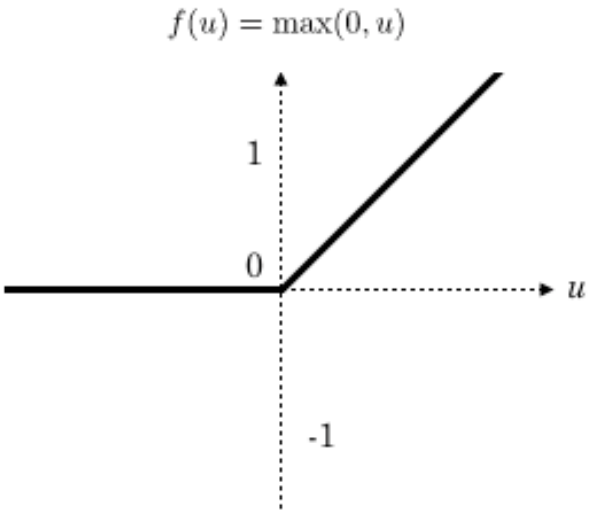
Vanishing Gradients



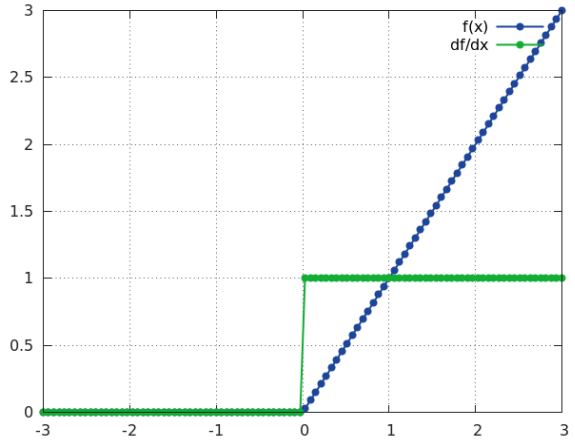
Vanishing Gradients

Use better activation function

- Rectified Linear Units (ramp)
  - $f(x) = \max(0, x)$
  - Derivative: All in or all out (unit step)
    - $f'(x) = 1 \; if \; x > 0 \; else \; 0$



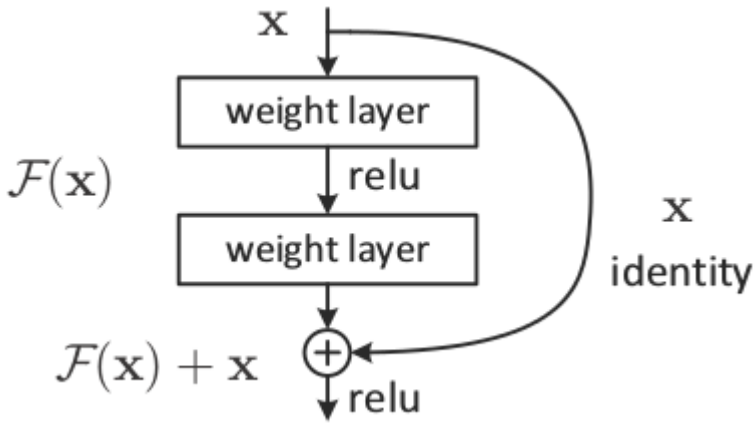
- Dead ReLUs
  - LeakyReLU: $f(x) = \max(x, 0.01x)$
  - PReLU: $f(x) = \max(x, ax)$



Vanishing Gradients

Use better architecture

- Residual networks
  - provide residual connections straight to earlier layers
  - This residual connection doesn't go through activation functions that "squashes" the derivatives, resulting in a higher overall derivative of the block

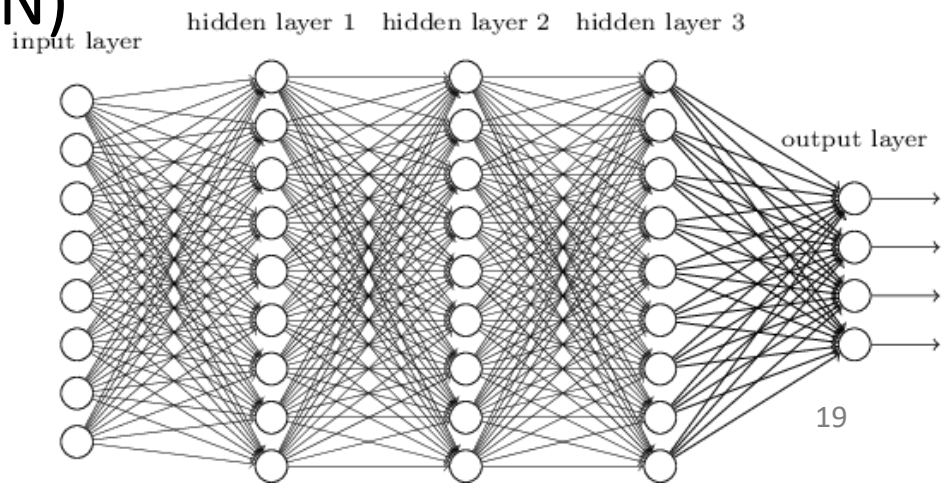


Use normalization

- Batch normalization

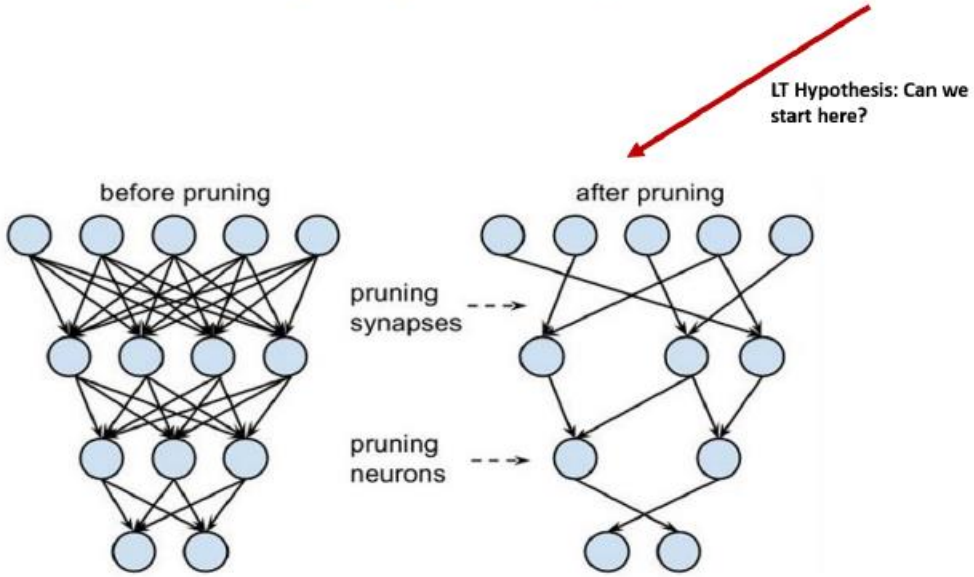
Problems with Going Deeper

- Parameter explosion
  - Too many weights to optimize as we go deeper
  - Search space is much harder to navigate
- Proposal: shared weights
  - Spatially shared (CNN)
  - Temporally shared (RNN)



The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks

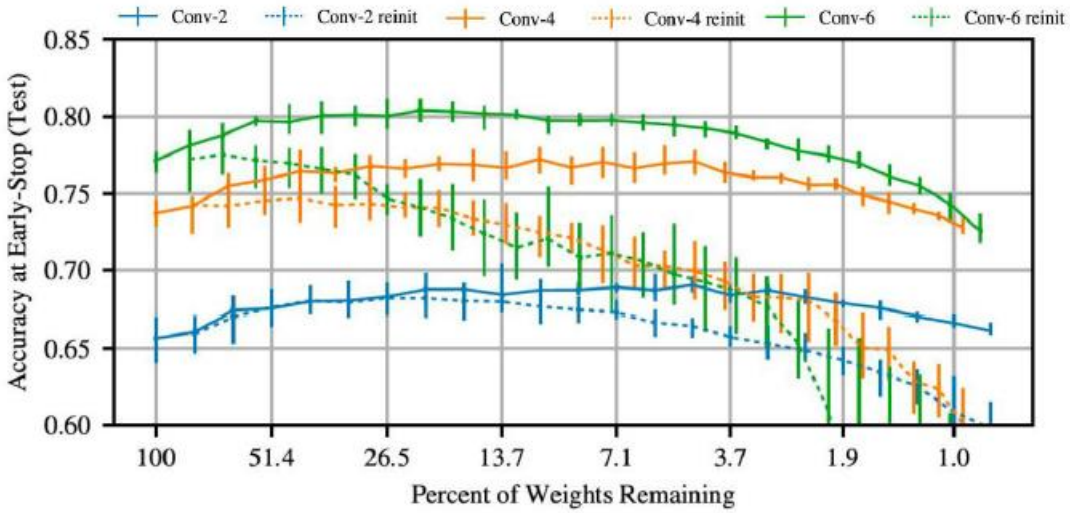
Frankle et al. (MIT) - Best Paper at ICLR 2019



1. Randomly initialize a neural network.
2. Train the network until it converges.
3. Prune a fraction of the network.
4. Reset the weights of the remaining network to initialization values from step 1
5. Train the pruned, untrained network. Observe convergence and accuracy.

The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks

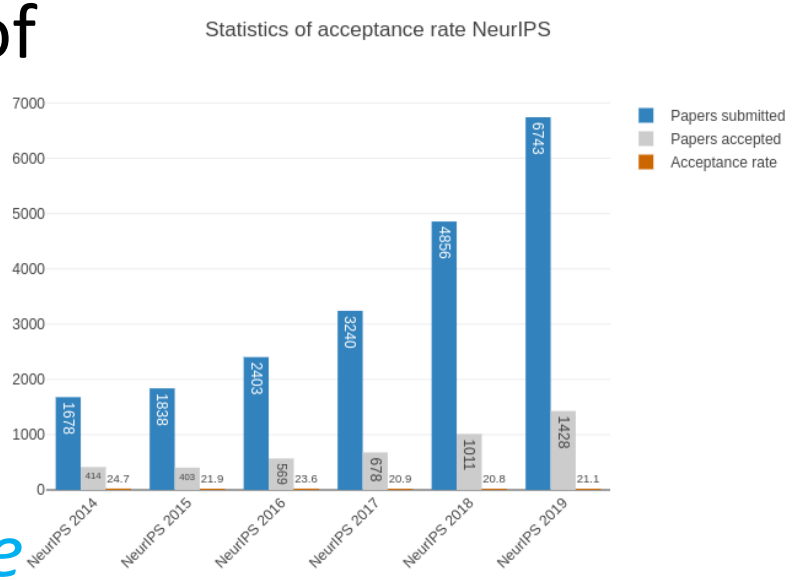
Frankle et al. (MIT) - ICLR 2019 Best Paper



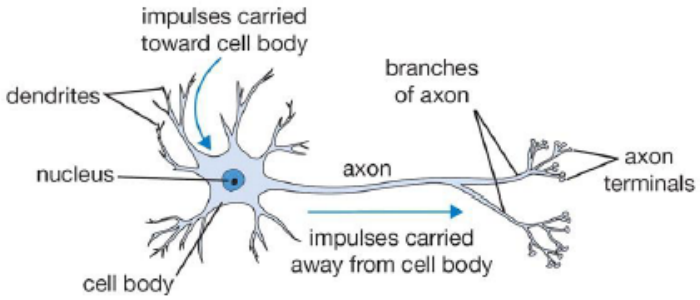
- **Idea:** For every neural network, there is a subnetwork that can achieve the same accuracy in isolation after training.
- **Iterative pruning:** Find this subset subset of nodes by iteratively training network, pruning its smallest-magnitude weights, and re-initializing the remaining connections to their original values. Iterative vs one-shot is key.
- **Inspiring takeaway:** There exist architectures that are much more efficient. Let's find them!

Deep Learning Future

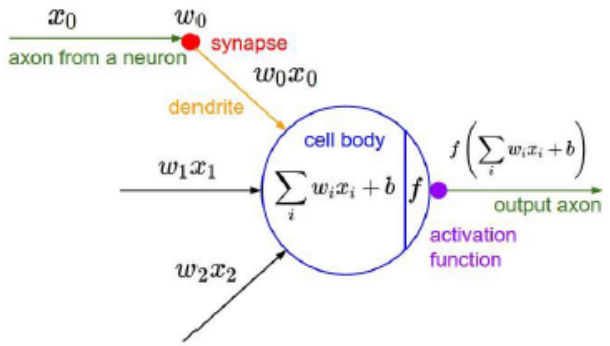
- History of science is a story of both people and ideas
- Many brilliant people contributed to the development of AI
- "*The future depends on some graduate student who is deeply suspicious of everything I have said*." - Geoffrey Hinton



Biological and Artificial Neural Networks



- **Neuron:** computational building block for the brain



- **(Artificial) Neuron:** computational building block for the "neural network"

Key Difference:

- **Parameters:** Human brains have ~10,000,000 times synapses than artificial neural networks.
- **Topology:** Human brains have no "layers". **Async:** The human brain works asynchronously, ANNs work synchronously.
- **Learning algorithm**: ANNs use gradient descent for learning. We don't know what human brains use
- **Power consumption:** Biological neural networks use very little power compared to artificial networks
- **Stages:** Biological networks usually never stop learning. ANNs first train then test.