The Development of Deep Learning

邓仰东
清华大学软件学院

Definitions

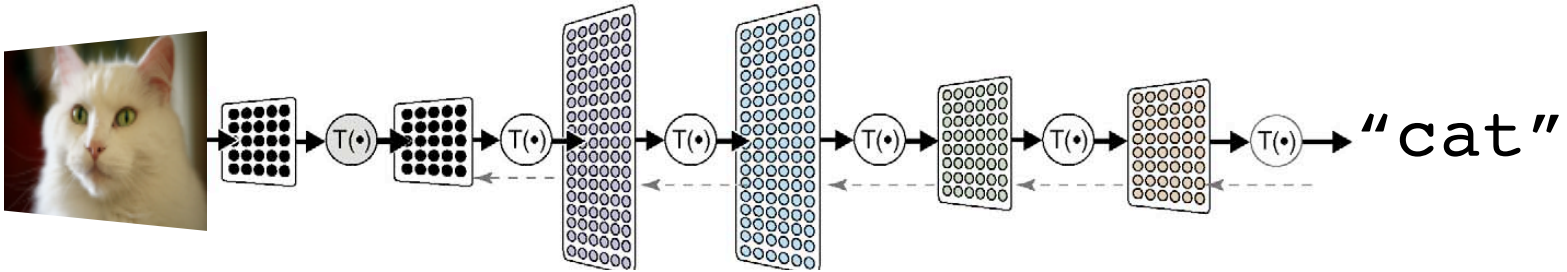
- “Parameter learning” or “training”
  - The calculation of weights of a model (either from closed-form solution, or numerical solution)
- “Inference” or “testing”
  - The process of applying an existing model with weights known onto solving a real-world problem
  - To decode a hidden state of data, e.g. to predict a label
- Overfitting
  - Learning a function that works really well for the training set but badly on the test set

General Machine Learning Approaches

- Supervised learning: Leaning by labeled exampls
  - e.g. An email spam detector
  - Amazingly effective if you have lots of labeled examples
- Unsupervised learning: Discovering patters
  - e.g. Data clustering
  - Difficult in practice, but useful if you lack labeled examples
- Reinforcement Learning: Feedback right/wrong
  - e.g. Learning to play chess by winning or loass
  - Works well in some domains, becoming more important

What Is Deep Learning

- The modern reincarnation of Artificial Neural Networks from the 1980s and 90s.
- A collection of simple trainable mathematical units, which collaborate to compute a complicated function
- Compatible with supervised, unsupervised, and reinforcement learning



Deep Neural Network Can Be Impressive

Both recognized as a
“meal”

Deep Neural Network Can Be Impressive

Human: Three different types of pizza on top of a stove.

Model sample 1: Two pizzas sitting on top of a stove top oven.

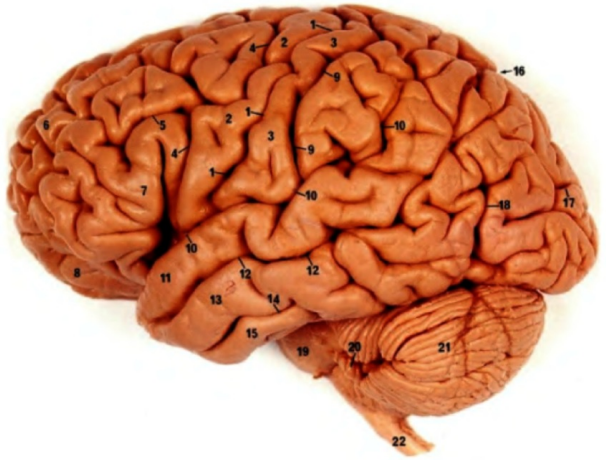
Model sample 2: A pizza sitting on top of a pan on top of a stove.

提纲

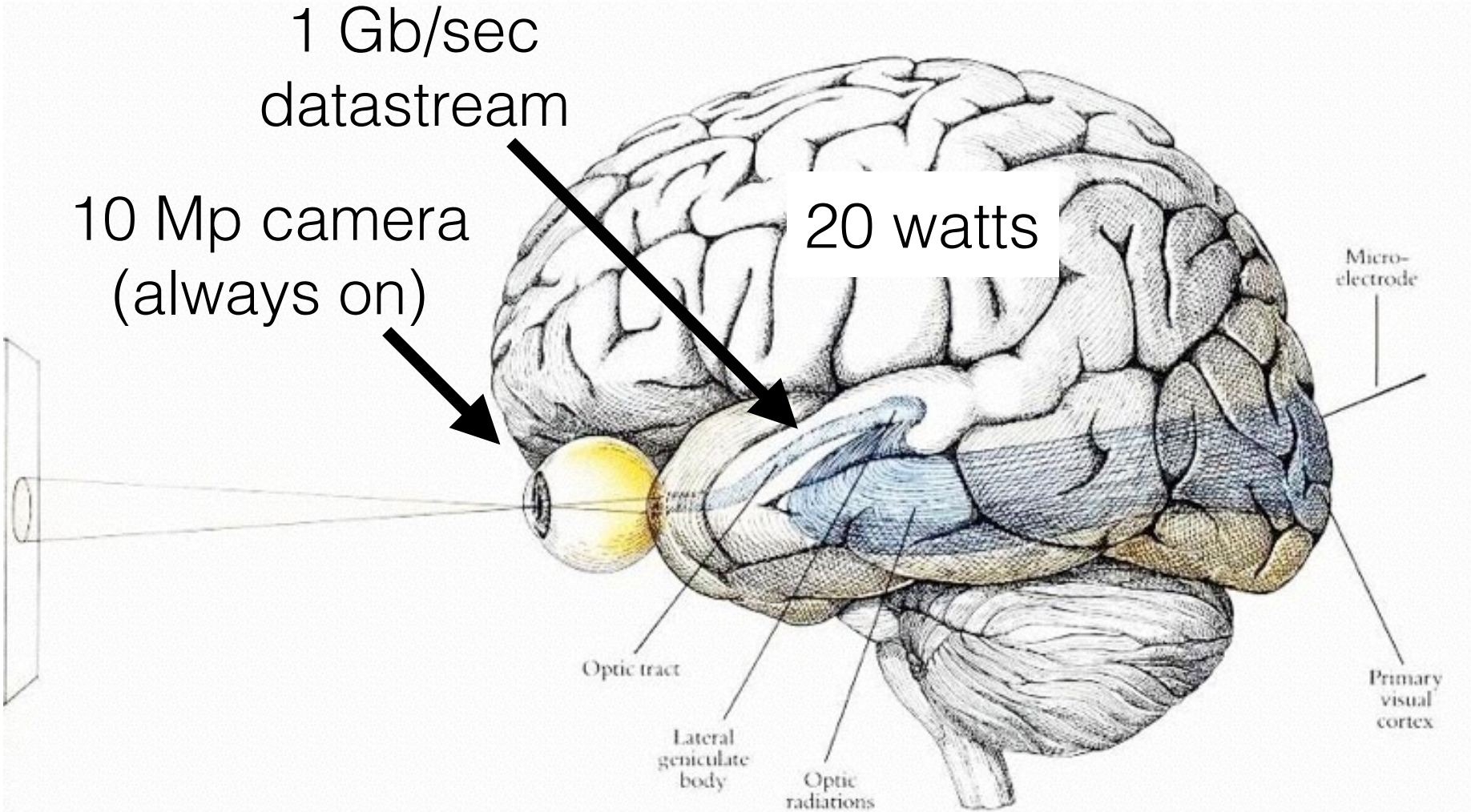
1. Brain Computing
2. The Road to Deep Learning
3. Rising of Deep Learning
4. DNN Applications

Computation in the Brain

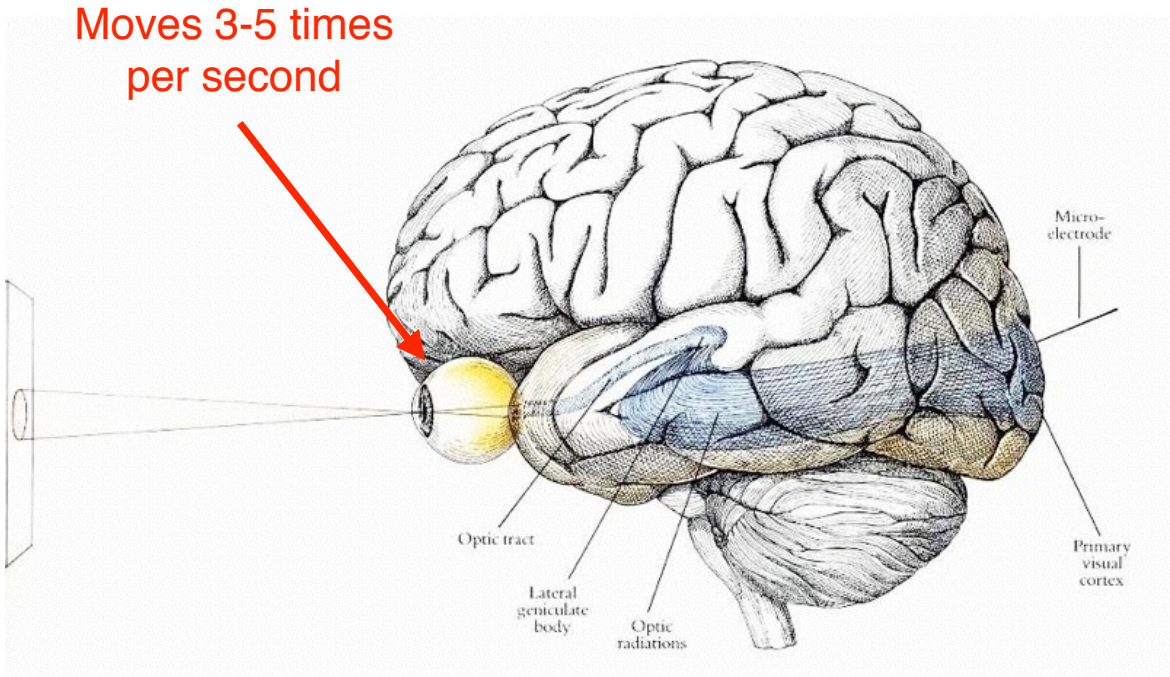
- An engineering perspective
  - Energy efficient (20 watts)
  - 10^{12} Glial cells (power, cooling, support)
  - 10^{11} Neurons (soma + wires)
  - 10^{14} Connections (synapses)
  - Volume = mostly wires
- General computing machine?
  - Slow for mathematical logic, arithmetic, etc.
  - Very fast for vision, speech, language, social interactions, etc.
  - Evolution: vision –> language –> logic.




1 mm^2 of cortex contains 100,000 neurons

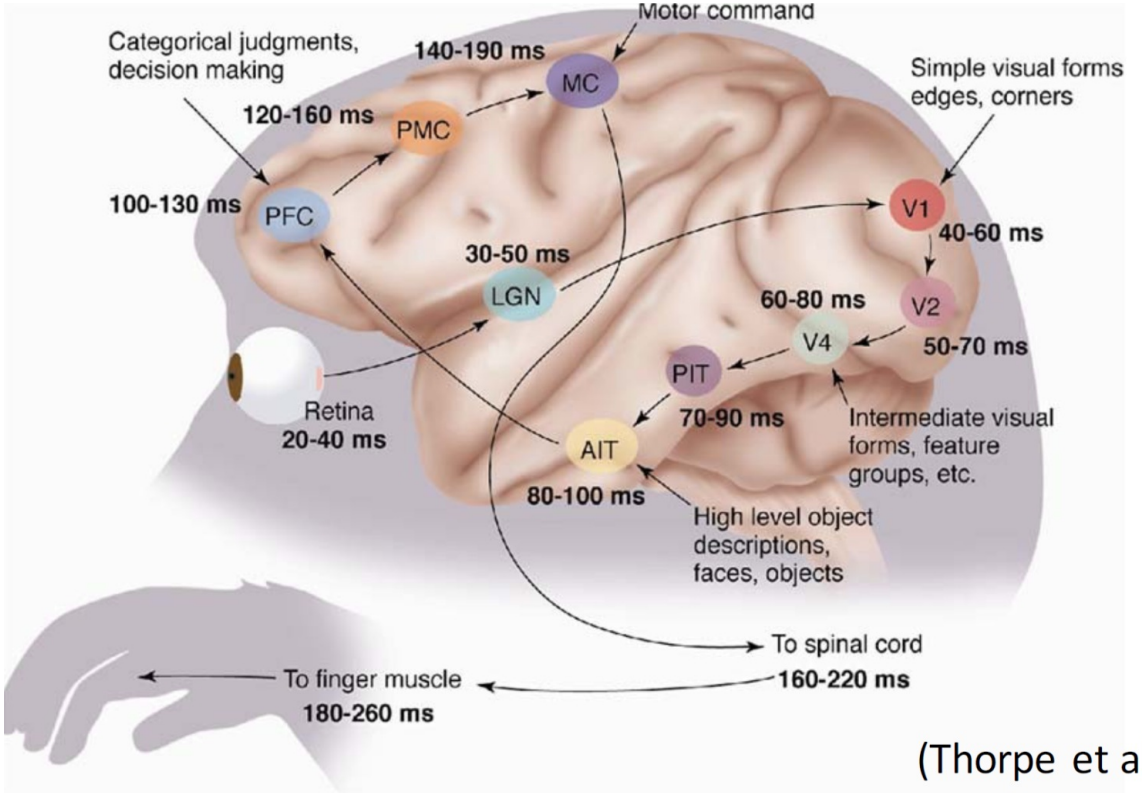

1 Gb/sec
datastream
10 Mp camera
(always on)
20 watts
Micro-
electrode
Optic tract
Lateral
geniculate
body
Optic
radiations
Primary
visual
cortex

Active Perception



Vision Is Fast

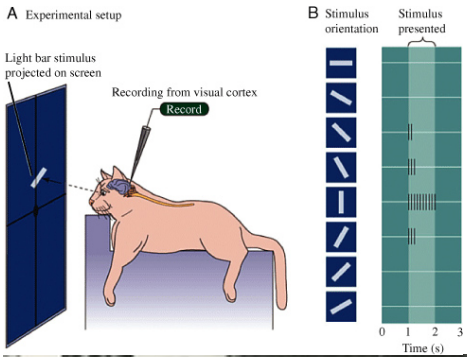
0.1 sec: neurons fire only 10 times!



(Thorpe et al., 1995-…)

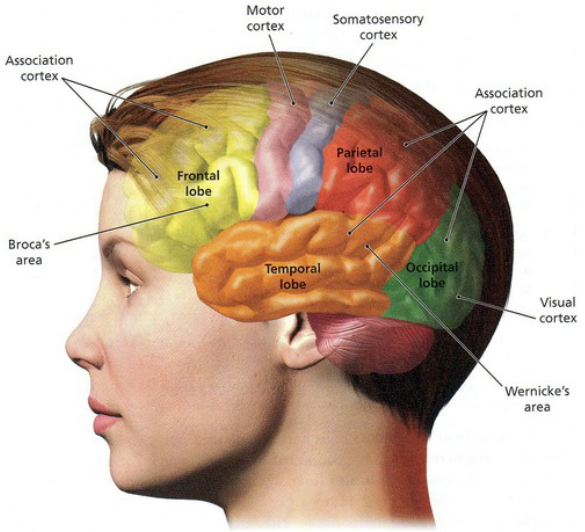
Basics on Visual Cortex (Hubel & Wiesel, 1959)

- A quantum step in the understanding of the visual system
  - 1981 Nobel Prize in Physiology or Medicine
- Insights about image processing in the brain
  - Simple cells detect local features
    - Inspired the SIFT descriptor
  - Complex cells pool local features in a retinotopic neighborhood
  - Some individual neuronal cells in the brain responded (or fired) only in the presence of edges of a certain orientation



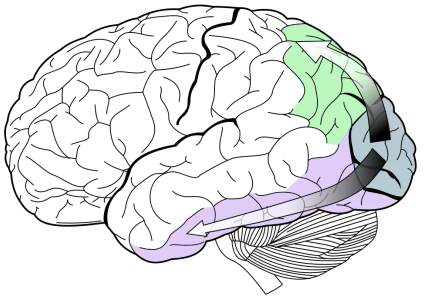
Visual Cortex

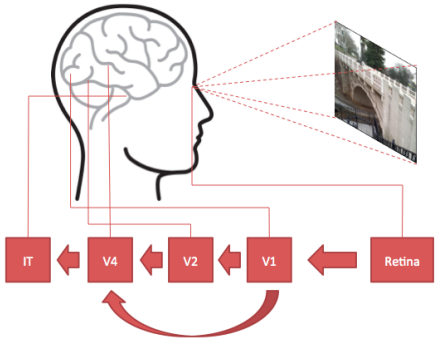
- The visual cortex of the brain, located in the occipital lobe which is located at the back of the skull
- Visual information coming from the eye, goes through a series of brain structures and reaches the visual cortex
- The parts of the visual cortex that receive the sensory inputs is known as the primary visual cortex, also known as area V1
- Visual information is further managed by extrastriate areas, including visual areas two (V2) and four (V4)
- There are also other visual areas (V3, V5, and V6)
- Visual areas that are related to object recognition, which is known as ventral stream and consists of areas V1, V2, V4 and inferior temporal gyrus, which is one of the higher levels of the ventral stream of visual processing, associated with the representation of complex object features, such as global shape, like face perception (Haxby et al., 2000)



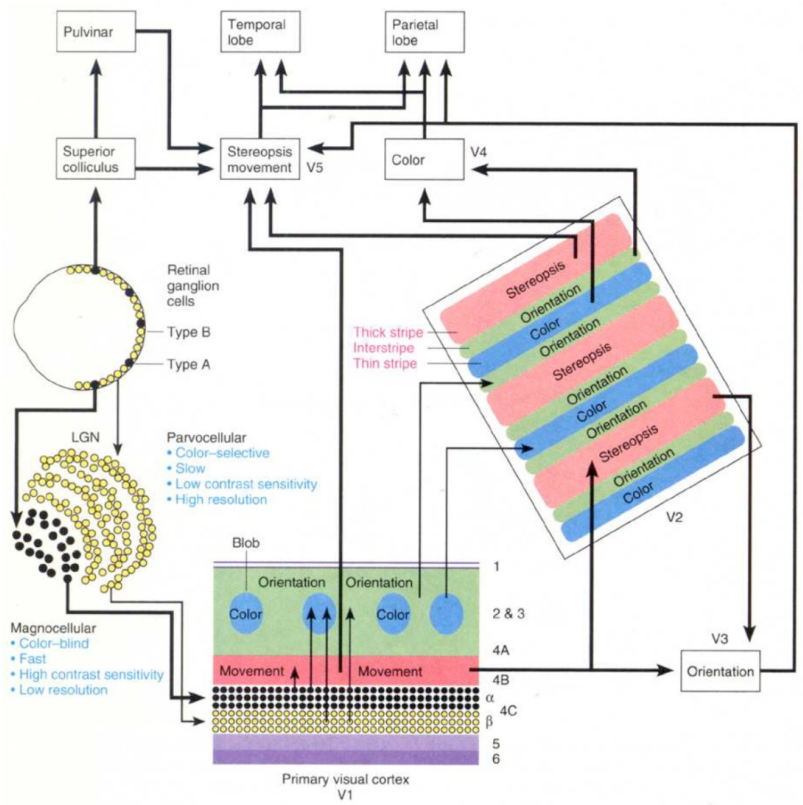
Ventral Stream of the Visual Cortex

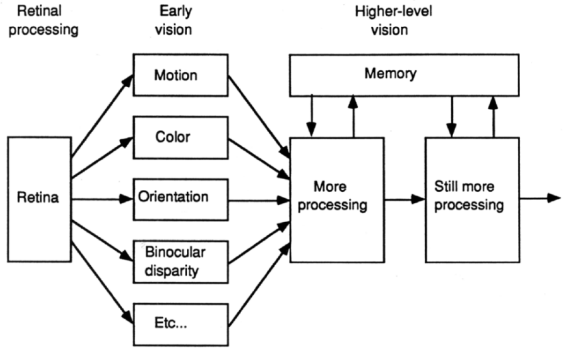
- Retina converts the light energy that comes from the rays bouncing o⸦ of an object into chemical energy. This chemical energy is then converted into action potentials that are transferred onto primary visual cortex. (In fact, there are several other brain structures involved between retina and V1, but we omit these structures for simplicity)
- Primary visual cortex (V1) mainly fulfills the task of edge detection, where an edge is an area with strongest local contrast in the visual signals.
- V2, also known as secondary visual cortex, is the first region within the visual association area. It receives strong feedforward connections from V1 and sends strong connections to later areas. In V2, cells are tuned to extract mainly simple properties of the visual signals such as orientation, spatial frequency, and colour, and a few more complex properties.
- V4 fulfills the functions including detecting object features of intermediate complexity, like simple geometric shapes, in addition to orientation, spatial frequency, and color. V4 is also shown with strong attentional modulation (Moran and Desimone, 1985). V4 also receives direct input from V1.
- Inferior temporal gyrus (TI) is responsible for identifying the object based on the color and form of the object and comparing that processed information to stored memories of objects to identify that object (Kolb et al., 2014). In other words, IT performs the semantic level tasks, like face recognition.

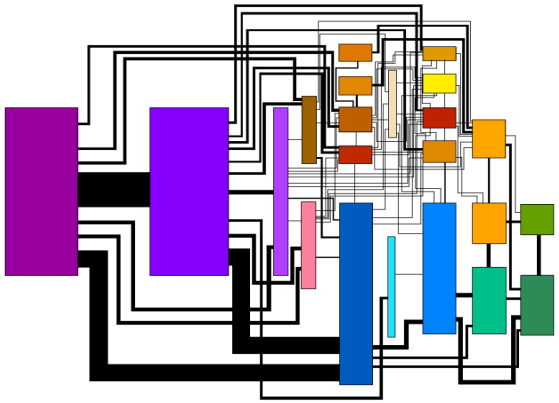




Visual Pathways







提纲

1. Brain Computing

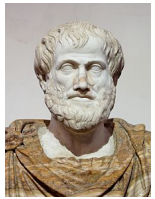
2. The Road to Deep Learning

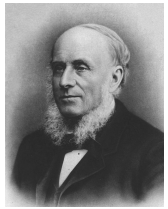
3. Rising of Deep Learning

4. DNN Applications

Deep Learning Milestones

Year	Contributer	Contribution
300 BC	Aristotle	introduced Associationism, started the history of human's attempt to understand brain.
1873	Alexander Bain	introduced Neural Groupings as the earliest models of neural network, inspired Hebbian Learning Rule.
1943	McCulloch & Pitts	introduced MCP Model, which is considered as the ancestor of Artificial Neural Model.
1949	Donald Hebb	considered as the father of neural networks, introduced Hebbian Learning Rule, which lays the foundation of modern neural network.
1958	Frank Rosenblatt	introduced the first perceptron, which highly resembles modern perceptron.
1974	Paul Werbos	introduced Backpropagation
1980	Teuvo Kohonen	introduced Self Organizing Map
	Kunihiko Fukushima	introduced Neocogitron, which inspired Convolutional Neural Network
1982	John Hopfield	introduced Hopfield Network
1985	Hilton & Sejnowski	introduced Boltzmann Machine
1986	Paul Smolensky	introduced Harmonium, which is later known as Restricted Boltzmann Machine
	Michael I. Jordan	defined and introduced Recurrent Neural Network
1990	Yann LeCun	introduced LeNet, showed the possibility of deep neural networks in practice
1997	Schuster & Paliwal	introduced Bidirectional Recurrent Neural Network
	Hochreiter & Schmidhuber	introduced LSTM, solved the problem of vanishing gradient in recurrent neural networks
2006	Geoffrey Hinton	introduced Deep Belief Networks, also introduced layer-wise pretraining technique, opened current deep learning era.
2009	Salakhutdinov & Hinton	introduced Deep Boltzmann Machines
2012	Geoffrey Hinton	introduced Dropout, an efficient way of training neural networks





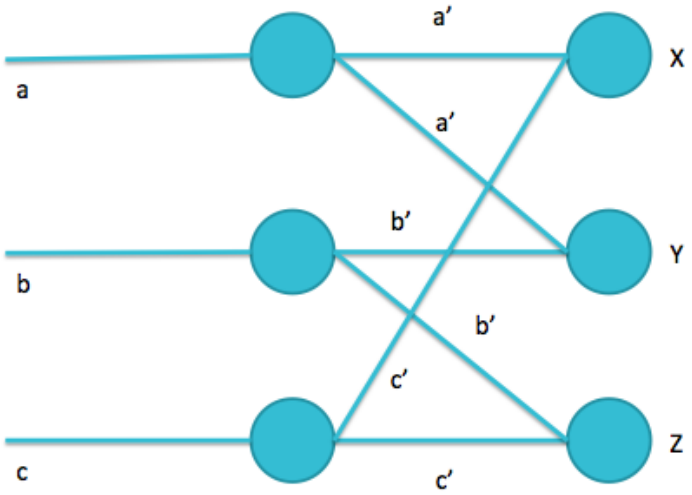
History of Neural Network

Origin

- Associationism
  - A theory states that mind is a set of conceptual elements that are organized as associations between these elements
- Aristotle's four laws of association
  - Contiguity: Things or events with spatial or temporal proximity tend to be associated in the mind
  - Frequency: The number of occurrences of two events is proportional to the strength of association between these two events
  - Similarity: Thought of one event tends to trigger the thought of a similar event
  - Contrast: Thought of one event tends to trigger the thought of an opposite event
- Hartley: Memory could be conceived as smaller scale vibrations in the same regions of the brain as the original sensory experience

Bain's Neural Groupings

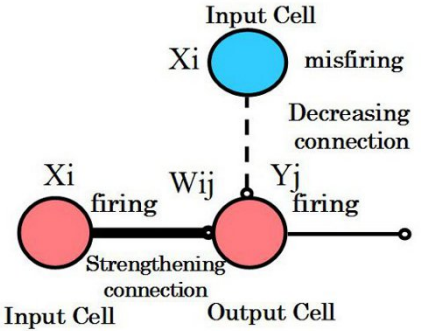
- Alexander Bain (1873) related the processes of associative memory to the distribution of activity of neural groupings (i.e. neural networks)
  - connections are strengthened or weakened through experience via changes of intervening cell-substance



Hebbian Learning Rule

- Donald O. Hebb: The Organization of Behavior (1949)
  - "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that As efficiency, as one of the cells firing B, is increased."
  - Cell assembly: a discrete, strongly interconnected group of active neurons, the 'cell assembly', represents a distinct cognitive entity

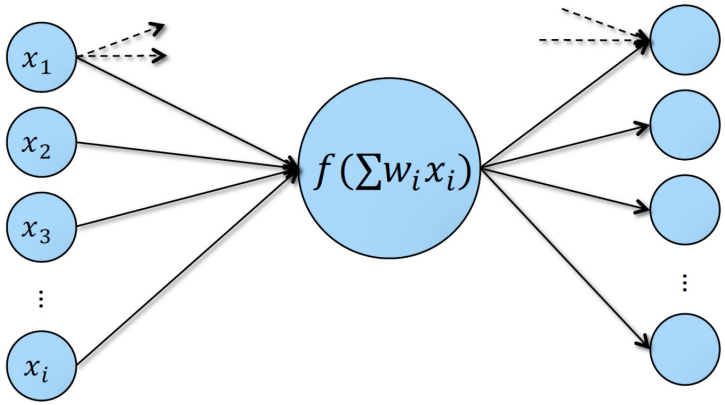
$\Delta Wij = \eta\, Xi \cdot Yj$



ΔWij is the strength of the change in synaptic weight
Xi is the output of the input cell
Yj is the output of the output cell
η is the learning coefficient

MCP Neural Model

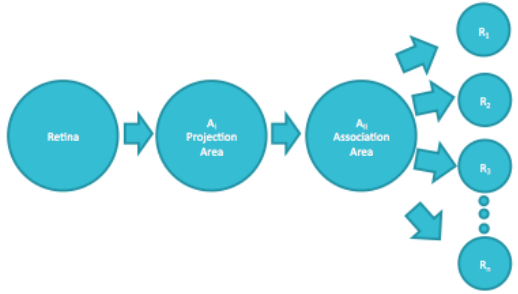
- Proposed by McCulloch & Pitts (1943)
- Initially built as electrical circuits
- The weights of MCP Neural Model wi are fixed, in contrast to the adjustable weights in modern perceptron
  - All the weights must be assigned with manual calculation
- The idea of inhibitory input is quite unconventional even seen today
  - It might be an idea worth further study

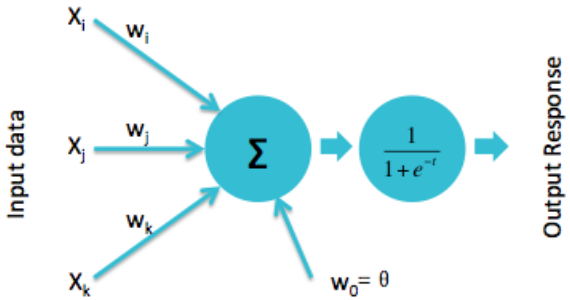


$$y = \begin{cases} 1, \ \sum_i w_i x_i \geq \theta \quad \text{AND} \quad z_j = 0, \forall j \\ 0, \ \text{otherwise} \end{cases}$$

Perceptron

- Rosenblatt (1958)
- Stimuli impact on a retina of the sensory units, which respond in a manner that the pulse amplitude or frequency is proportional to the stimulus intensity.
- Impulses are transmitted to Projection Area (optional)
- Impulses are then transmitted to Association Area through random connections. If the sum of impulse intensities is equal to or greater than the threshold (θ) of this unit, then this unit fires.

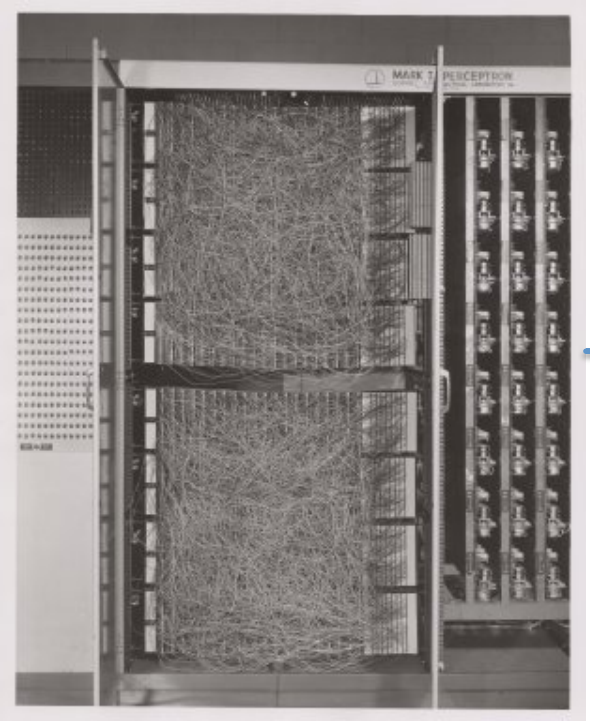


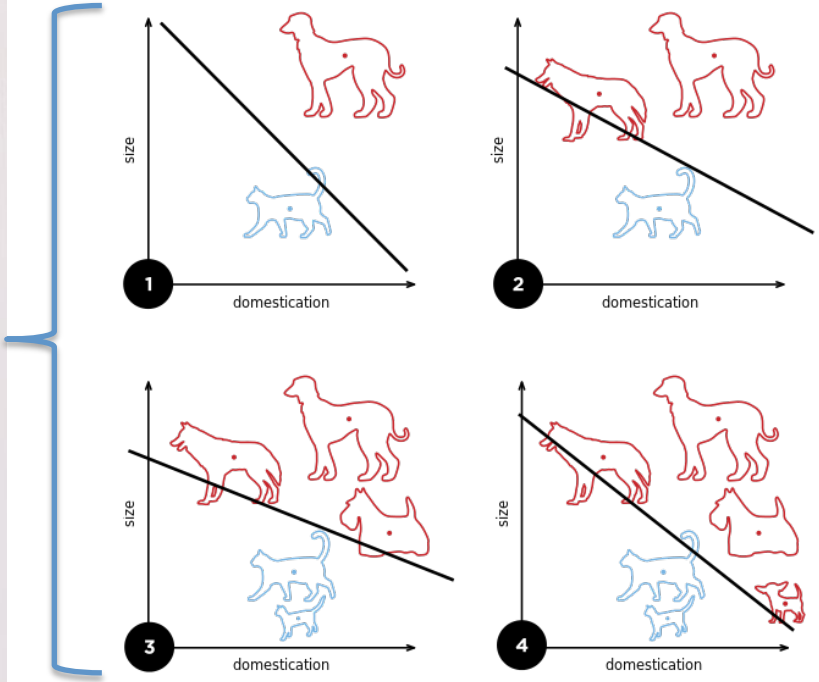


Learning Algorithm for Perceptrons

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

Perceptron

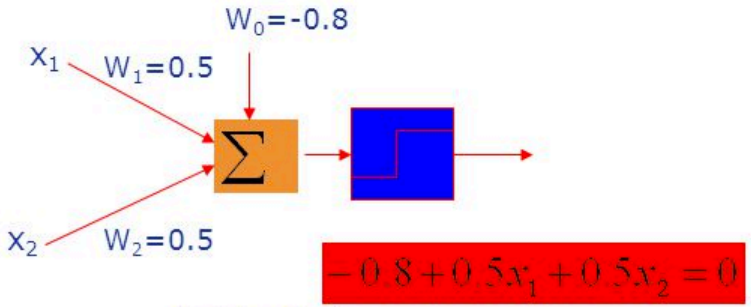

MARK 1 PERCEPTRON


size
domestication
1
size
domestication
2
size
domestication
3
size
domestication
4

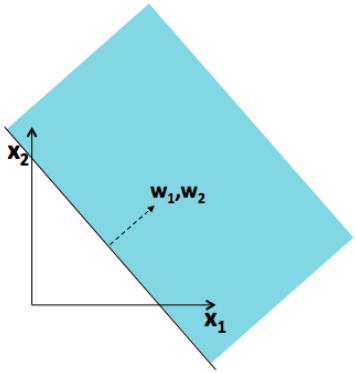
The Linear Representation Power of Perceptron

AND

x1	x2	D
-1	-1	-1
-1	+1	-1
+1	-1	-1
+1	+1	+1

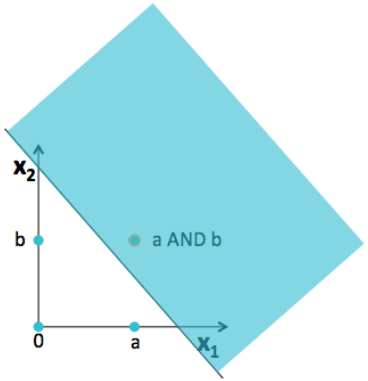


The Linear Representation Power of Perceptron

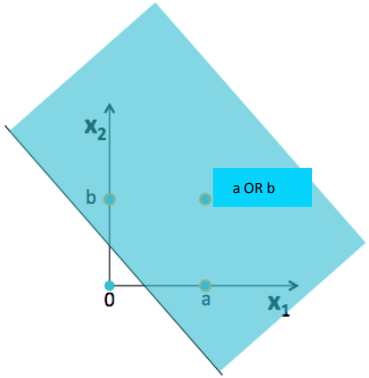


(a)

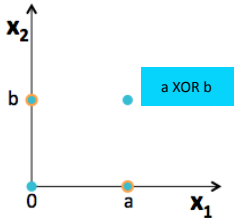
$W_1X_1+W_2X_2$



(b)

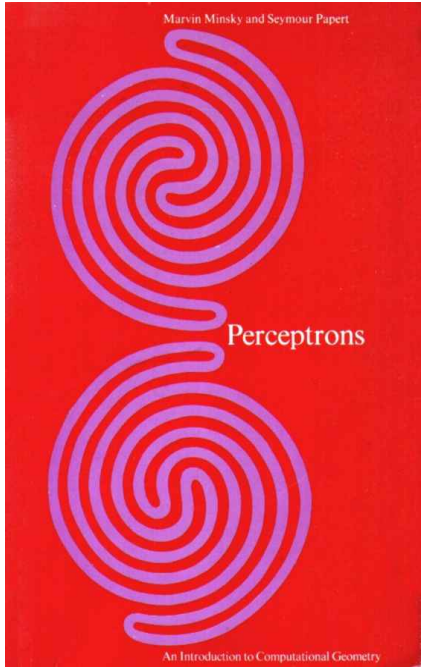


(c)



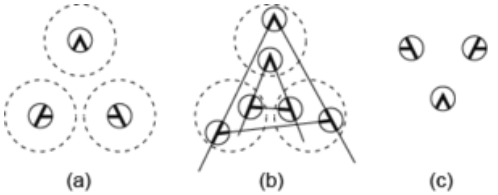
(d)

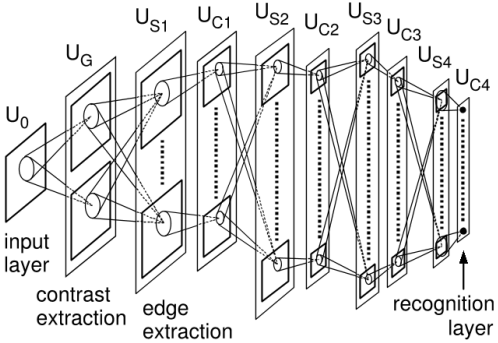
This is the book that killed research into neural networks for 15 years!



The Neocognitron

- Neocogitron, proposed by Fukushima (1980)
- Generally seen as the model that inspires Convolutional Neural Networks on the computation side
- It is a neural network that consists of two di↵erent kinds of layers (S-layer as feature extractor and C-layer as structured connections to organize the extracted features.)



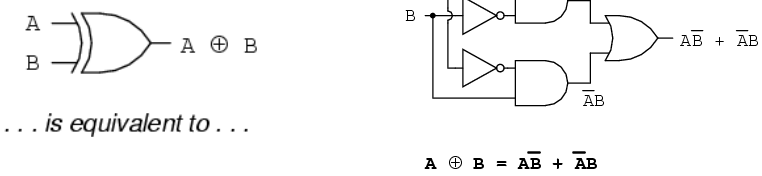


Multi-Layer Perceptrons (MLP)

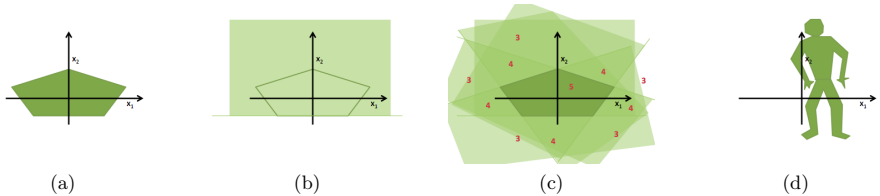
- Kawaguchi (2000)
- Universal Approximation Property
  - Boolean Approximation: an MLP of one hidden layer1 can represent any boolean function exactly
  - Continuous Approximation: an MLP of one hidden layer can approximate any bounded continuous function with arbitrary accuracy
  - Arbitrary Approximation: an MLP of two hidden layers can approximate any function with arbitrary accuracy

The Representation Power of MLP

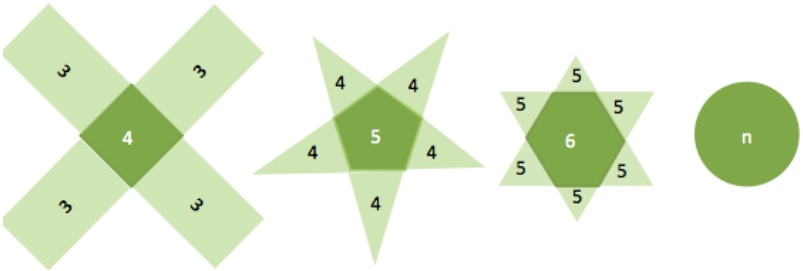
- Boolean Approximation



- Continuous Approximation

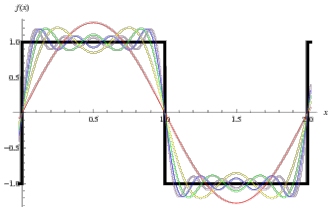


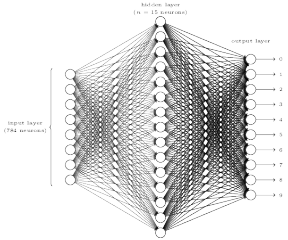
- Arbitrary Approximation

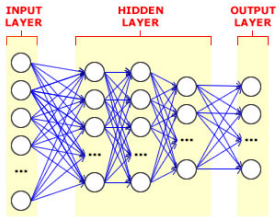


The Representation Power of MLP

- Similarly to how Gibbs phenomenon affects Fourier series approximation, this approximation
- cannot guarantee an exact representation
- The universal approximation properties showed a great potential of shallow neural networks at the price of exponentially many neurons at these layers
- One followed-up question is how to reduce the number of required neurons while maintaining the representation power
  - This question motivates people to proceed to deeper neural networks despite that shallow neural networks already have infinite modeling power
- Another issue worth attention is that, although neural networks can approximate any functions, it is not trivial to find the set of parameters to explain the data







Anything humans can do in 0.1 sec, the right big 10-layer network can do too

Why “Deep”

- Bengio and Delalleau (2011) suggested that it is nature to pursue deeper networks because
  - 1) human neural system is a deep architecture
  - 2) humans tend to represent concepts at one level of abstraction as the composition of concepts at lower levels
- Yao (1985) and Hastad (1986) showed the limitations of shallow circuits functions
  - “There are functions computable in polynomial size and depth k but requires exponential size when depth is restricted to k − 1”.
- A function that could be expressed with O(n) neurons on a network of depth k requiredat least O(2pn) and O((n−1)k) neurons on a two-layer neural network
- Eldan and Shamir (2015) presented a more thorough proof to show that depth of a neural network is exponentially more valuable than the width of a neural network, for a standard MLP with any popular activation functions

Backpropagation

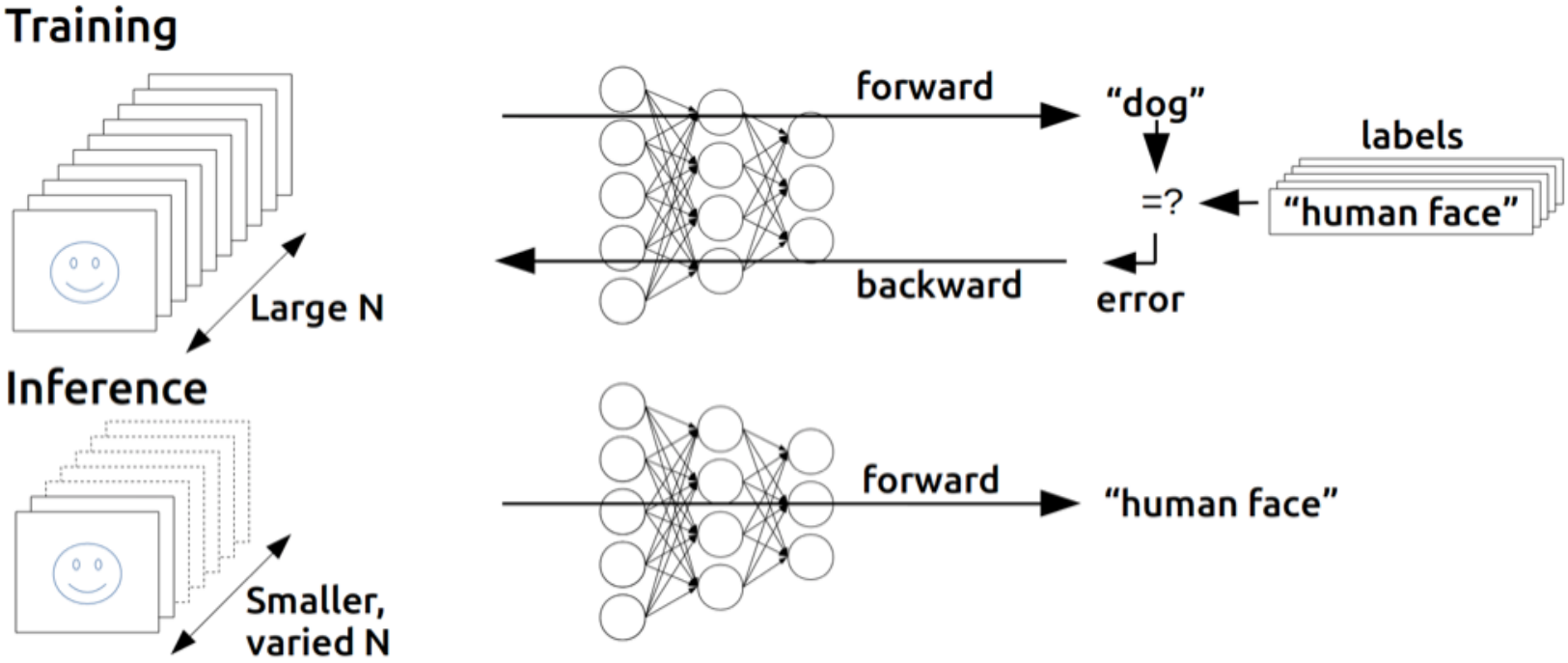
Backpropagation

- Linear Separable Data
  - Gori and Tesi (1992) studied on the problem of local minima in backpropagation and proposed an architecture where global optimal is guaranteed
  - Only a few weak assumptions of the network are needed to reach global optimal, including
    - Pyramidal Architecture: upper layers have fewer neurons
    - Weight matrices are full row rank
    - The number of input neurons cannot smaller than the classes/patterns of data

Backpropagation

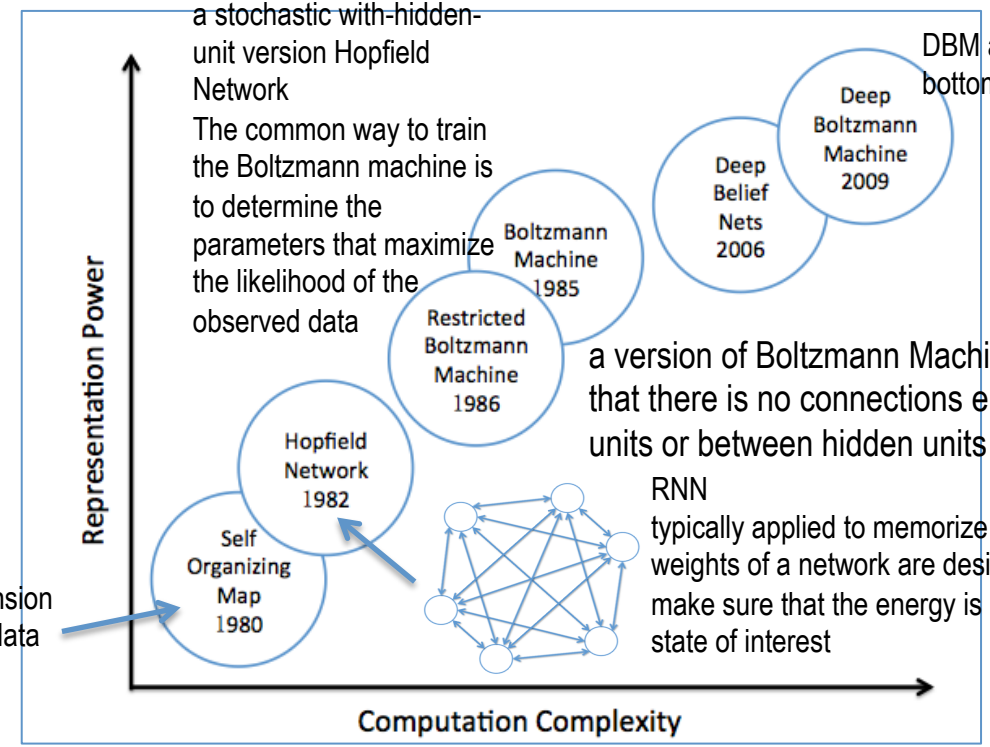
- Linear Separable Data
- Brady et al. (1989) studied the situations when backpropagation fails on linearly separable data sets and showed that there could be situations when the data is linearly separable, but a neural network learned with backpropagation cannot find that boundary.
  - His illustrative examples only hold when the misclassified data samples are significantly less than correctly classified data samples, in other words, the misclassified data samples might be just outliers.

Basic Idea of Backpropagation



Other Neural Networks

Tradeoffs of Representation Power and Computation Complexity of Several NN-Alike Models



learns a two dimension representation of data

a stochastic with-hidden-unit version Hopfield Network

The common way to train the Boltzmann machine is to determine the parameters that maximize the likelihood of the observed data

a version of Boltzmann Machine with a restriction that there is no connections either between visible units or between hidden units

DBM allows bidirectional connections in the bottom layers

RNN

typically applied to memorize the state of data:The weights of a network are designed or learned to make sure that the energy is minimized given the state of interest

Deep Generative Models

- Lake et al. (2015) introduces a Bayesian Program Learning framework that can simulate human learning abilities with large scale visual concepts
- In addition to its performance on one-shot learning classification task, their model passes the visual Turing Test in terms of generating handwritten characters from the worlds alphabets. In other words, the generative performance of their model is indistinguishable from human's behavior
- Being not a deep neural model itself, their model outperforms several concurrent deep neural networks
- Deep neural counterpart of the Bayesian Program Learning framework can be surely expected with even better performance.

提纲

1. Brain Computing

2. The Road to Deep Learning

3. Rising of Deep Learning

4. DNN Applications

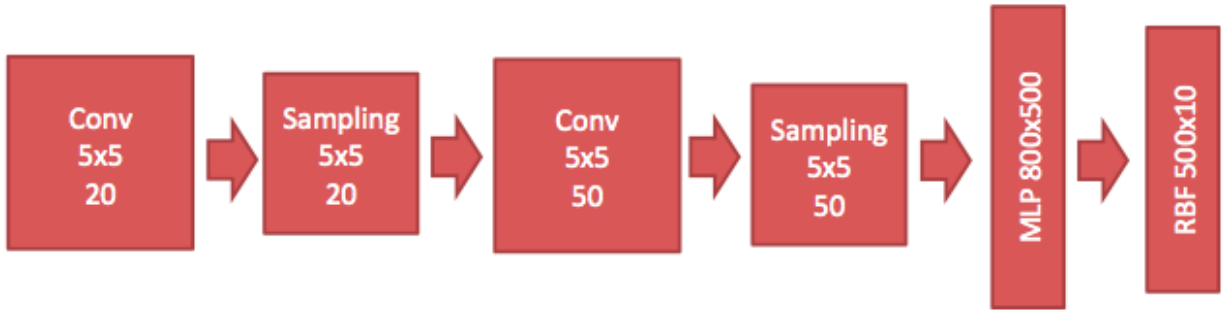
Modern Start of CNN

- The modern Convolutional Neural Networks owe their inception to a well-known 1998 research paper by Yann LeCun and Léon Bottou
- In this highly instructional and detailed paper, the authors propose a neural architecture called LeNet 5 used for recognizing hand-written digits and words that established a new state of the art classification accuracy of 99.2% on the MNIST dataset

- According to the author’s accounts, CNNs are biologically-inspired models. The research investigations carried out by D. H. Hubel and T. N. Wiesel in their paper[6] proposed an explanation for the way in which mammals visually perceive the world around them using a layered architecture of neurons in the brain, and this in turn inspired engineers to attempt to develop similar pattern recognition mechanisms in computer vision

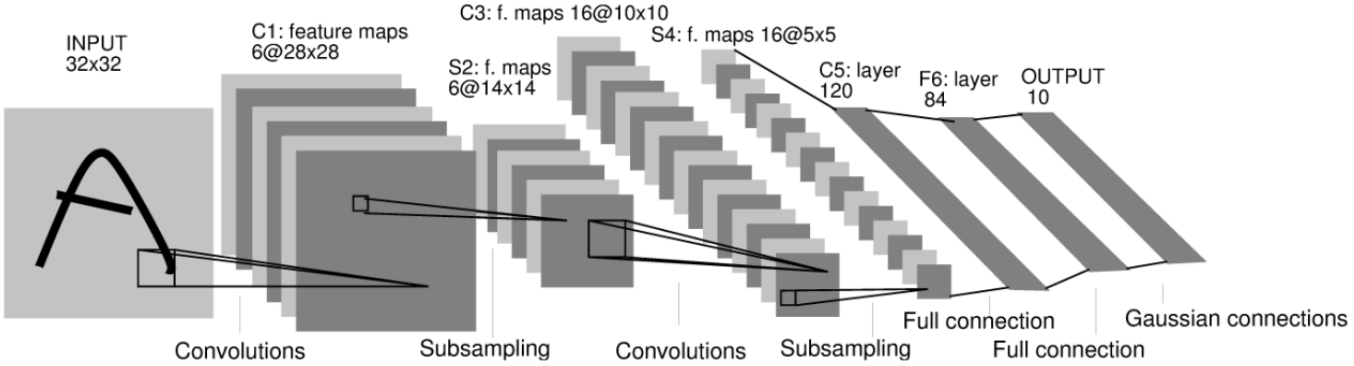
LeNet

- LeNet is known as its ability to classify digits and can handle a variety of different problems of digits including variances in position and scale, rotation and squeezing of digits, and even different stroke width of the digit. Meanwhile, with the introduction of LeNet, LeCun et al. (1998b) also introduces the MNIST database, which later becomes the standard benchmark in digit recognition field.



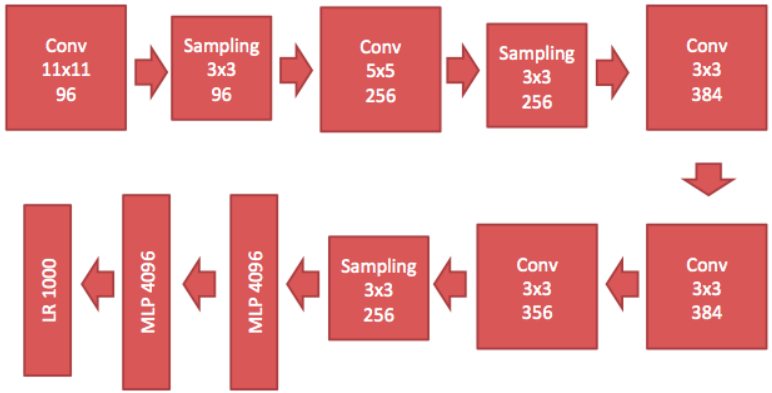
CNN Applications in the 1990s

- 1989 Isolated handwritten character recognition (AT&T Bell Labs)
- 1991 Face recognition. Sonar image analysis. (Neuristique)
- 1993 Vehicle recognition (Onera)
- 1994 Zip code recognition (AT&T Bell Labs)
- 1996 Check reading (AT&T Bell Labs)



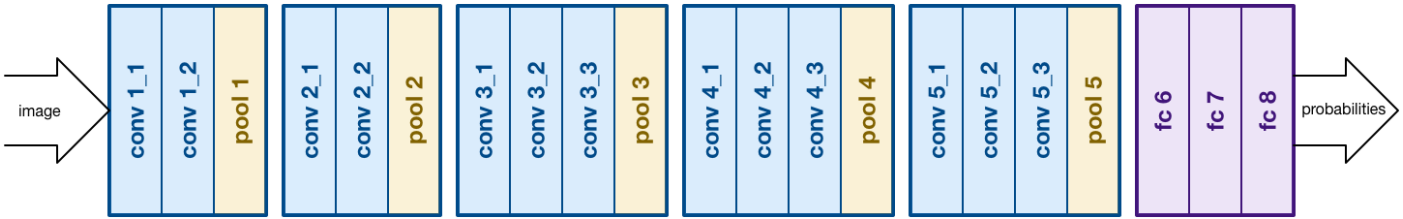
AlexNet

- AlexNet is the first evidence that CNN can perform well on this historically dicult ImageNet dataset and it performs so well that leads the society into a competition of developing CNNs
- The success of AlexNet is not only due to this unique design of architecture but also due to the clever mechanism of training. To avoid the computationally expensive training process, AlexNet has been split into two streams and trained on two GPUs.



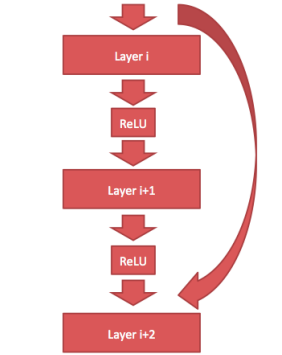
VGG

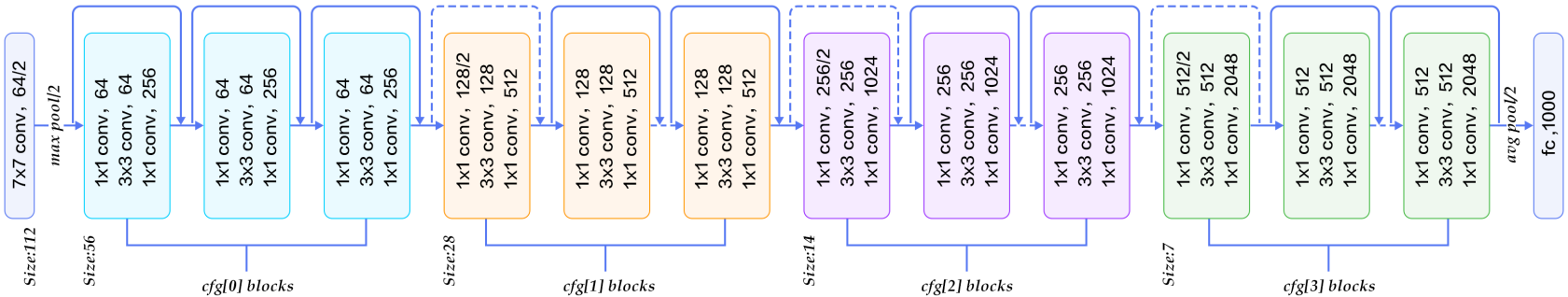
- Although VGG is deeper (19 layer) than other models around that time, the architecture is extremely simplified: all the layers are 3 →| 3 convolutional layer with a 2 →| 2 pooling layer.
- VGG is not the winner of the ImageNet competition of that year (The winner is GoogLeNet invented by Szegedy et al. (2015)). GoogLeNet introduced several important concepts like Inception module and the concept later used by R-CNN (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015), but the arbitrary/creative design of architecture barely contribute more than what VGG does to the society, especially considering that Residual Net, following the path of VGG, won the ImageNet challenge in an unprecedented level.



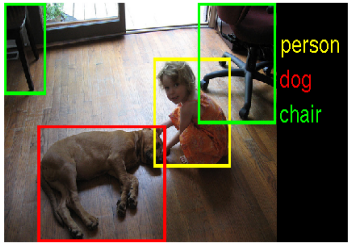
ResNet

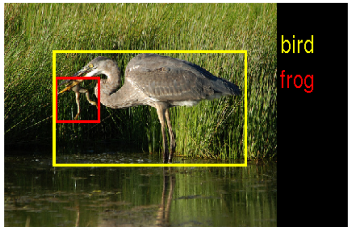
- Residual Net (ResNet) is a 152 layer network, which was ten times deeper than what was usually seen during the time
- The breakthrough ResNet introduces, which allows ResNet to be substantially deeper than previous networks, is called Residual Block.

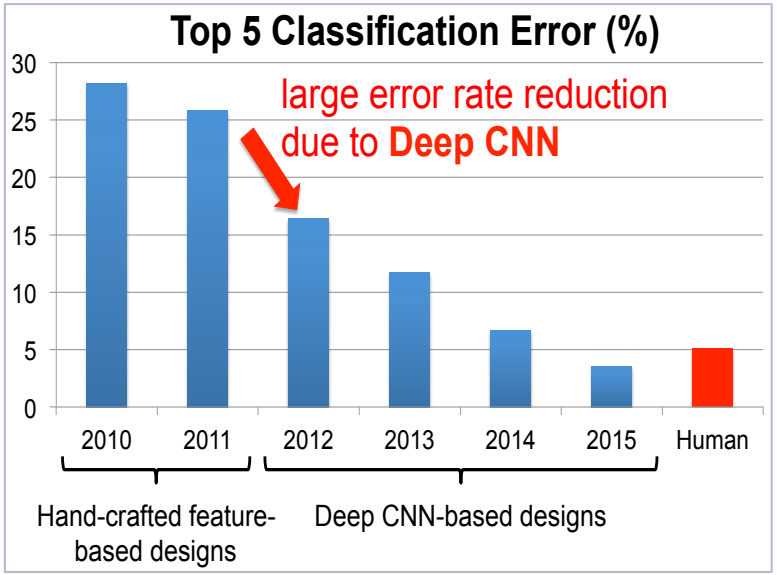




ImageNet Challenge







ImageNet: Image recognition, detection, and location

提纲

1. Brain Computing

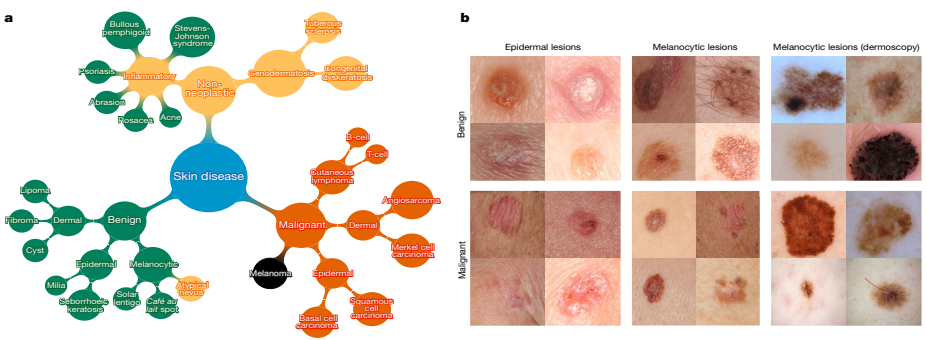
2. The Road to Deep Learning

3. Rising of Deep Learning

4. DNN Applications

深度学习时代

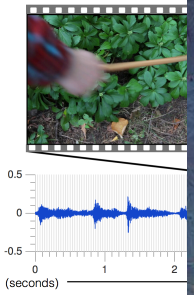
AlphaGo: Deep Reinforcement Learning for Go Game (Nature'16)



Dermatologist-level classification of skin cancer with deep neural networks (Nature'17)

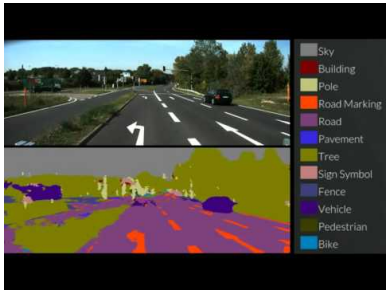
自动染色

自



题

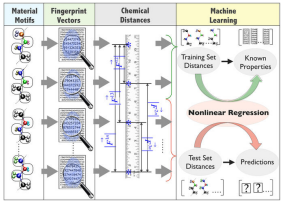
自动驾驶



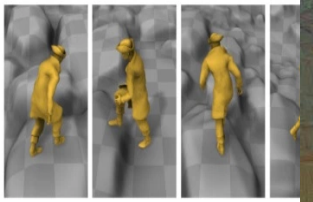
翻译

唇语

新材料发现



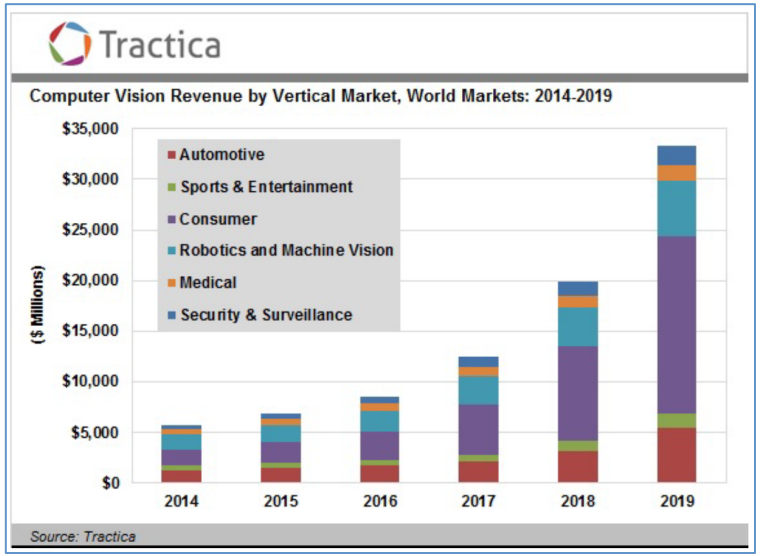
动画人



计算机梦境

未来市场

- 深度学习技术在未来10年将形成5000亿美元的市场！



职业恐惧度

Probability of computerisation of different occupations, 2013
(1 = certain)

Job	Probability
Recreational therapists	0.003
Dentists	0.004
Athletic trainers	0.007
Clergy	0.008
Chemical engineers	0.02
Editors	0.06
Firefighters	0.17
Actors	0.37
Health technologists	0.40
Economists	0.43
Commercial pilots	0.55
Machinists	0.65
Word processors and typists	0.81
Real-estate sales agents	0.86
Technical writers	0.89
Retail salespeople	0.92
Accountants and auditors	0.94
Telemarketers	0.99

Source: "The Future of Employment: How Susceptible are Jobs to Computerisation?", by C. Frey and M. Osborne (2013)

Economist.com

局限

- 需要大量具有标签的样本数据
  - 但是实际应用中更多需要不间断学习
- 成功领域相对有限
  - ImageNet等数据集合的类别频率与真实世界不同
  - 但是较少解决实时、在线、控制类应用
- 与日常生活的融入性不足
  - 但是我们需要“不经意”的深度学习应用

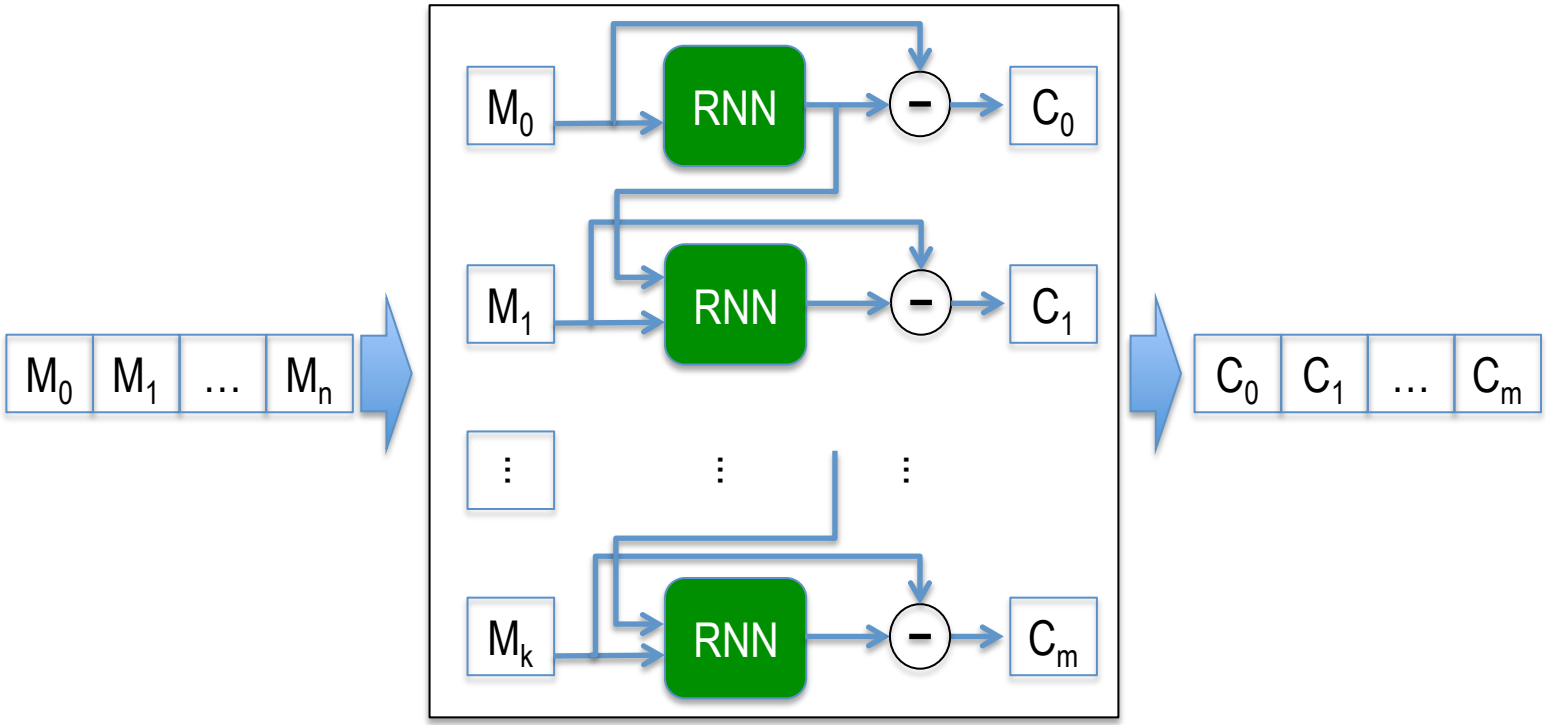
FIRST YOU GET THE DATA, THEN YOU GET THE

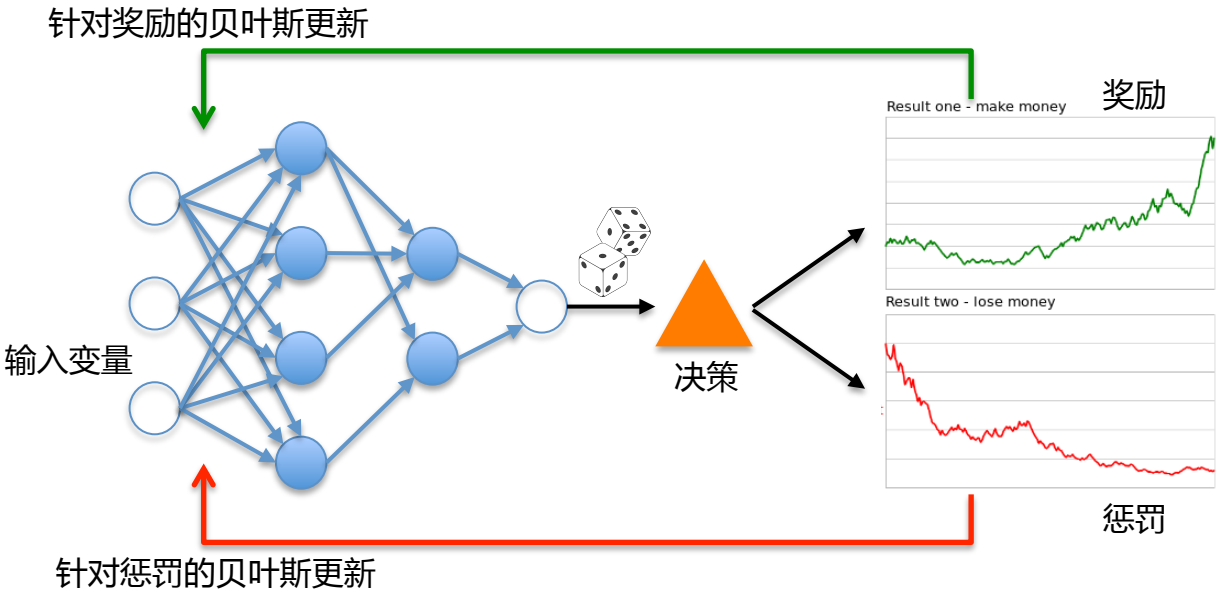
APPLYING DEEP LEARNING TO REAL-WORLD PROBLEMS CAN BE MESSY

DNN Working Not So Impressively

Human: A green monster kite soaring in a sunny sky.

Model: A man flying through the air while riding a skateboard.


M0
M1
…
Mn
M0
RNN
C0
M1
RNN
C1
Mk
RNN
Cm
C0
C1
…
Cm


针对奖励的贝叶斯更新
奖励
Result one - make money
Result two - lose money
输入变量
决策
惩罚
针对惩罚的贝叶斯更新