Introduction to Artificial Neural Network

Minlie Huang
aihuang@tsinghua.edu.cn
Dept. of Computer Science and Technology
Tsinghua University
http://coai.cs.tsinghua.edu.cn/hml/

Content

- Recent Success of Neural Networks
- Origin, History, and Present of Neural Networks
- Basics of Neural Networks
- Characteristics
- Basics of Machine Learning

Great Success

- Speech
  - MSRA, Interspeech 2011
    ⋆ Switchboard, **27.4%->18.5%**
- Vision
  - Swiss AI Lab IDSIA, CVPR 2012
    ⋆ MNIST, NIST SD 19, HW DB 1.0 on, HW DB 1.0 off, CIFAR 10, IJCNN 2011, NORB: **improve at least 26%**
  - Hinton
    ⋆ ImageNet Large Scale Visual Recognition Challenge 2012: **15.3% vs 26.2%**
  - MSRA, ResNet, CVPR 2016
    ⋆ ILSVRC2016 Object Detection: **62.1% mAP vs 37.2%(2014) and 22.5%(2013)**

Computer Vision

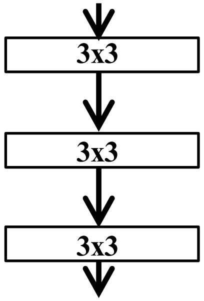
ImageNet

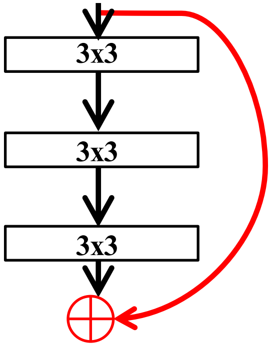
- 1000 categories
- 1.2 M images for training
- 50K for validation
  100K for testing

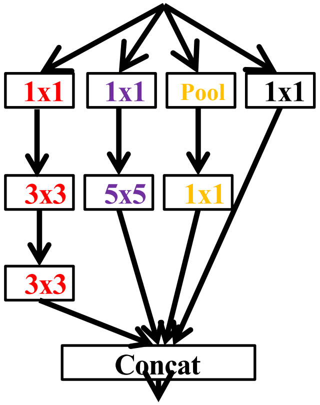
Team	Time	Place	Top-5 error
SuperVision	2012	1	16.42%
ISI	2012	2	26.17%
Clarifai	2013	1	11.74%
NUS	2013	2	12.95%
GoogLeNet	2014	1	6.66%
VGG	2014	2	7.32%
MSRA	2014	3	8.06%
Human	2014	-	5.1%
MSRA PReLU-nets	2015.2	-	4.94%
BN-Inception	2015.2	-	4.82%
ResNet	2015.12	1	3.57 %
Inception-v4	2016.2	-	3.08 %
WRN	2016.10	1	2.99%

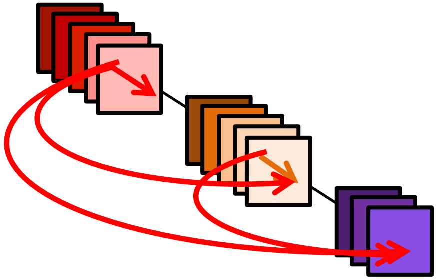
Computer Vision

ImageNet









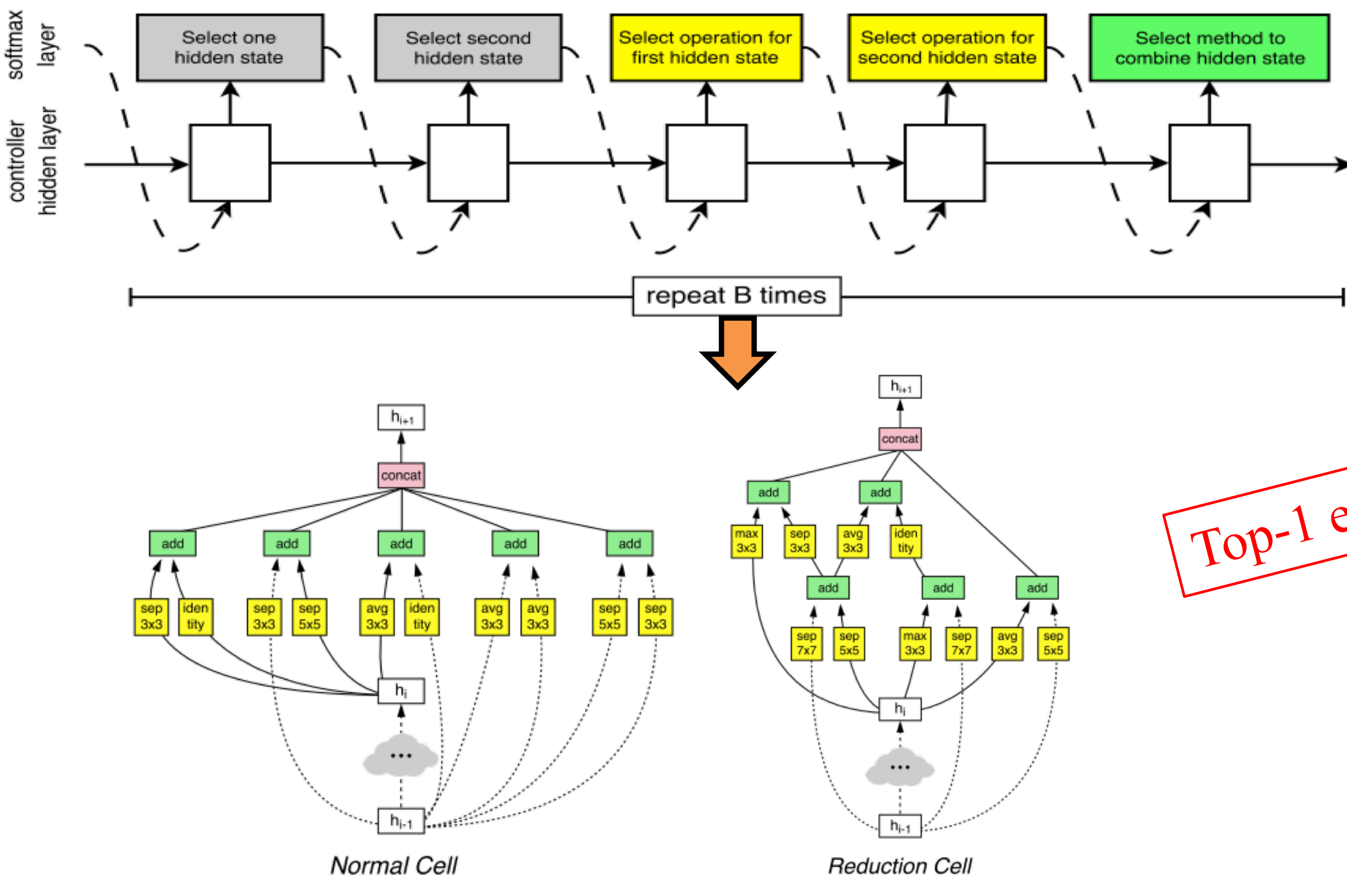
VGG(2015) ResNet(2016) Inception V1-V4 (2015-17) DenseNet(2017)

Top-1 err:

VGG(2015)	ResNet(2016)	Inception V1-V4 (2015-17)	DenseNet(2017)
28.5%	23.2%	31.2% → 19.8%	19.6%

Computer Vision

ImageNet



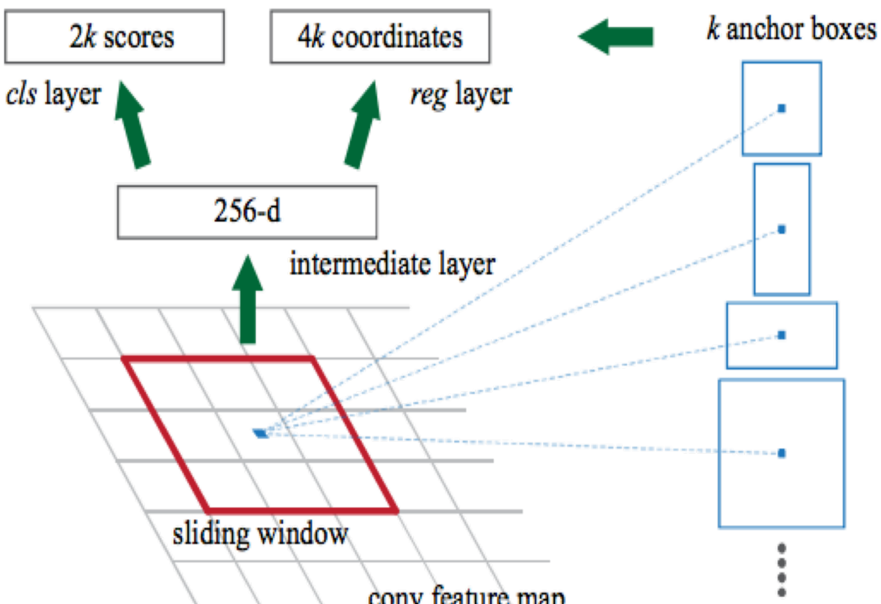
Top-1 err: 17.3%

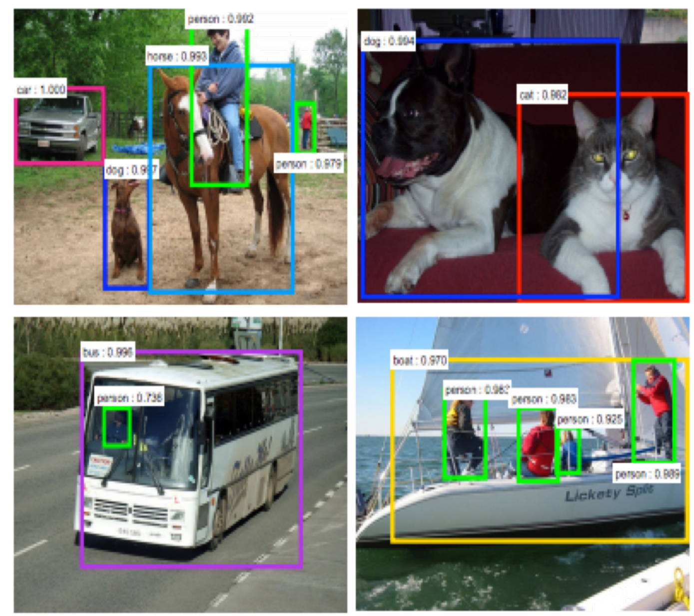
AutoML(2018)

Computer Vision

Detection — VOC & COCO

- 80 object categories
- 1.5 million object instances





Framework of Detection - (Faster R-CNN)

Computer Vision

Detection — VOC 2017

R-CNN

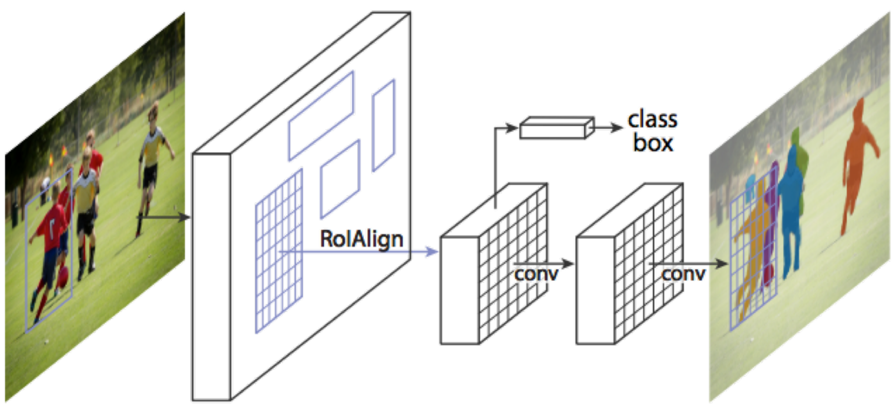
- R-CNN 2014
  - **47s/frame，improve the performance from 34.3%(DPM HSC) to 66%；**
  - Generate proposals by using selective search.
- Fast R-CNN 2015
  - **3s/frame，mAP 70%；**
  - Reduce the time consumption related to the high number of models necessary to analyze all region proposals.
- Faster R-CNN 2015
  - **5FPS，mAP 73.2%；**
  - Introduced Region Proposal Network (RPN) to directly generate region proposals.
- RFCN 2016
  - **5.9FPS，mAP 76.6%；**
  - Propose position-sensitive score maps in order to take into account the spatial localization of a particular object.
- FPN 2017
  - **Base on RFCN or Faster R-CNN, mAP 78.8%；**
  - Use multi-scale features to improve the performance.

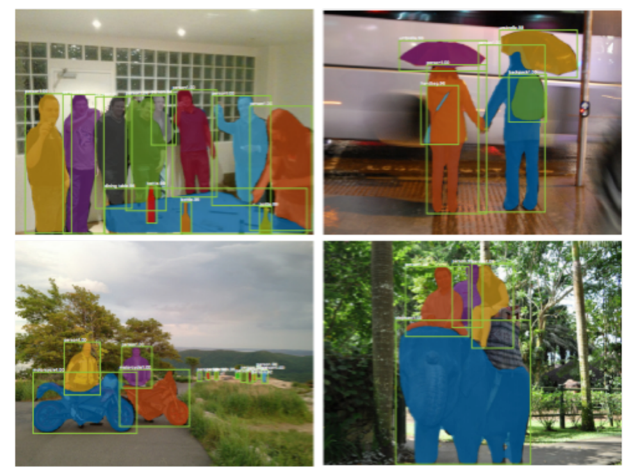
Computer Vision

COCO 2018 Object Detection Task

Detection → Instance Segmentation

- 80 object categories
- 200,000 images





Framework for instance segmentation – (Mask R-CNN)

Natural Language Processing

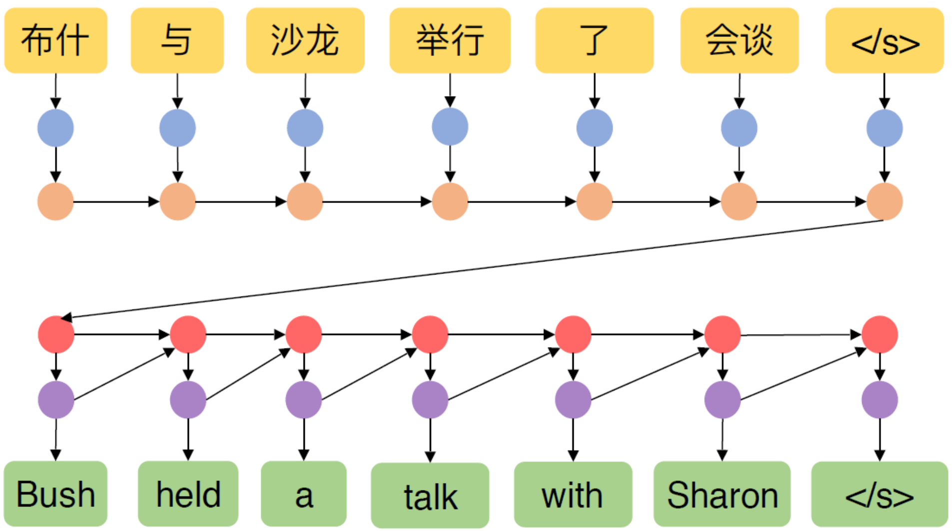
- Sentiment Classification on Stanford Sentiment Treebank
  - **SVM/MNB ~ 40%**
  - **Neural models ~50%+**

State-of-the-art

Method	Fine-grained
SVM [Pang and Lee 2008]	40.7
MNB [Wang and Manning 2012]	41.0
bi-MNB [Wang and Manning 2012]	41.9
RNN [Socher et al. 2011]	43.2
RNTN [Socher et al. 2013]	45.7
MV-RNN [Socher et al. 2012]	44.4
AdaMC-RNN [Dong et al. 2014]	45.8
AdaMC-RNTN [Dong et al. 2014]	46.7
DRNN [Irsoy and Cardie 2014]	49.8
TG-RNN (ours)	46.1(0.3)
TE-RNN (ours)	47.8(0.3)
TE-RNTN (ours)	48.8(0.4)
CNN [Kim 2014]	48.0
DCNN [Kalchbrenner et al. 2014]	48.5
LSTM [Tai et al. 2015]	46.4(1.1)
Bi-directional LSTM [Tai et al. 2015]	49.1(1.0)
Tree-LSTM [Tai et al. 2015]	51.0(0.5)
TW-LSTM (ours)	49.9(0.4)
TW-LSTM+p (ours)	50.6(0.4)
TE-LSTM (ours)	50.3(0.2)
TE-LSTM+p (ours)	51.3(0.4)
TW-LSTM+c (ours)	52.0(0.4)
TW-LSTM+c,p (ours)	52.1(0.4)
TE-LSTM+c (ours)	52.3(0.4)
TE-LSTM+c,p (ours)	52.6(0.6)

Natural Language Processing

- Machine translation develops with sequence-to-sequence (seq2seq) learning and attention mechanism.

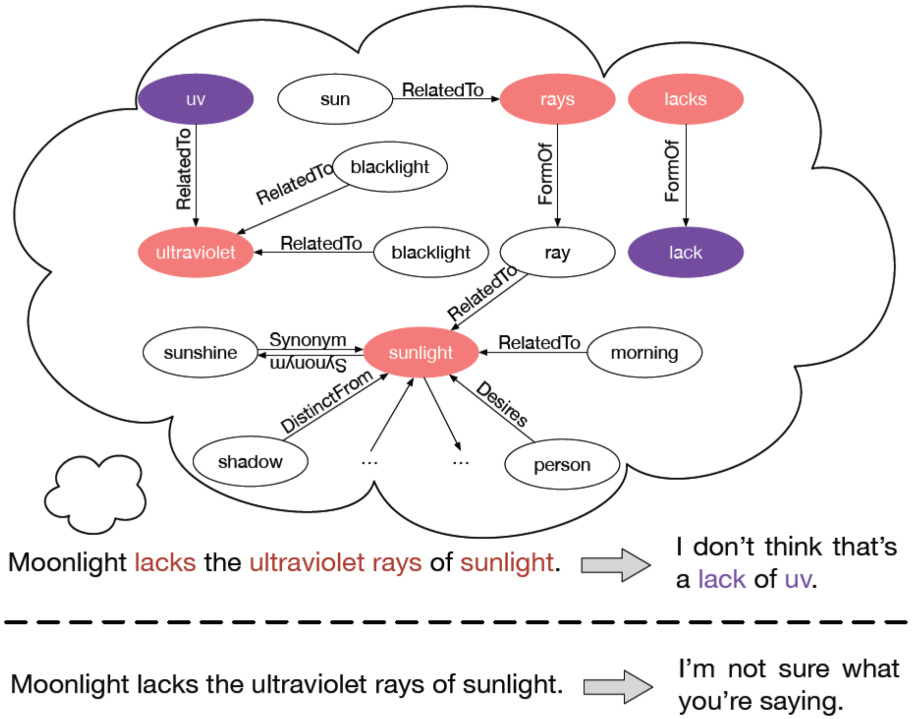


Model	BLEU	
	EN-DE	EN-FR
ByteNet [18]	23.75	
Deep-Att + PosUnk [39]		39.2
GNMT + RL [38]	24.6	39.92
ConvS2S [9]	25.16	40.46
MoE [32]	26.03	40.56
Deep-Att + PosUnk Ensemble [39]		40.4
GNMT + RL Ensemble [38]	26.30	41.16
ConvS2S Ensemble [9]	26.36	**41.29**
Transformer (base model)	27.3	38.1
Transformer (big)	**28.4**	**41.8**

[1] Ilya Sutskever, Oriol Vinyals, Quoc V. Le. Sequence to Sequence Learning with Neural Networks. NIPS 2014.
[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A. N., et al. Attention is all you need. NIPS 2017.

Natural Language Processing

- Knowledge graph augments NLP tasks (e.g. dialogue generation) with understanding.



Post	Why are you so **breakable**?
Knowledge	(**glass**, RelatedTo, **breakable**), (**brittle**, RelatedTo, **breakable**), (**rule**, RelatedTo, **breakable**)
Seq2Seq	I'm not a **OOV** , I'm just a **OOV**.
MemNet	I'm not **OOV**. I'm just a really nice person.
CopyNet	I'm not. I'm just a lurker.
CCM	Because I'm a **brittle** man .

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, Xiaoyan Zhu. Commonsense Knowledge Aware Conversation Generation with Graph Attention. IJCAI-ECAI 2018. **distinguished paper**

Natural Language Processing

- Reading comprehension on SQuAD dataset

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

SQuAD1.1 Leaderboard

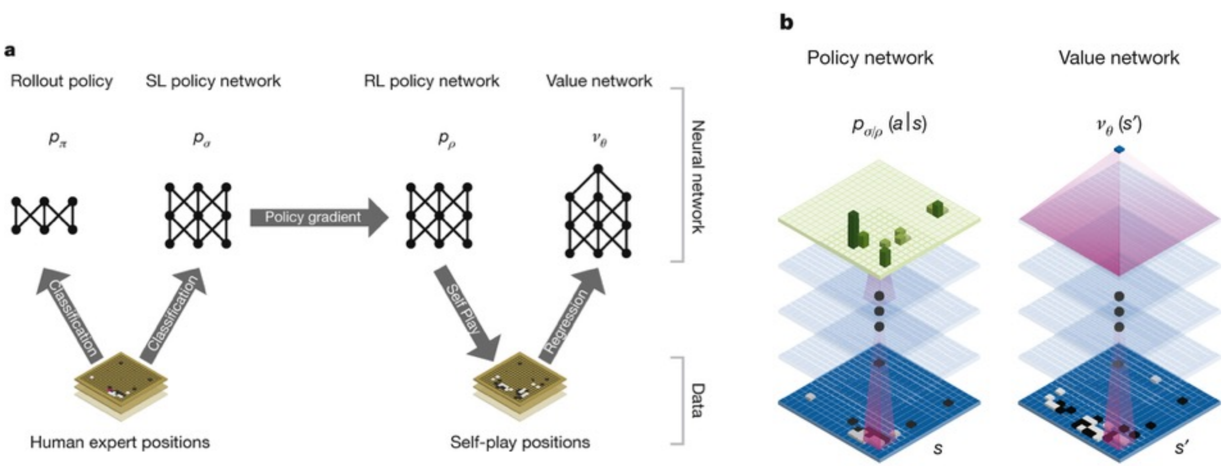
Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	**84.454**	**90.490**
2 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
3 Jun 20, 2018	MARS (ensemble) *YUANFUDAO research NLP*	83.982	89.796
4 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737
5 Jun 21, 2018	MARS (single model) *YUANFUDAO research NLP*	83.122	89.224

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. EMNLP 2016.

Human-Machine Chess

- The first computer program to ever beat a professional player at the game of Go
  - Deep Reinforcement Learning

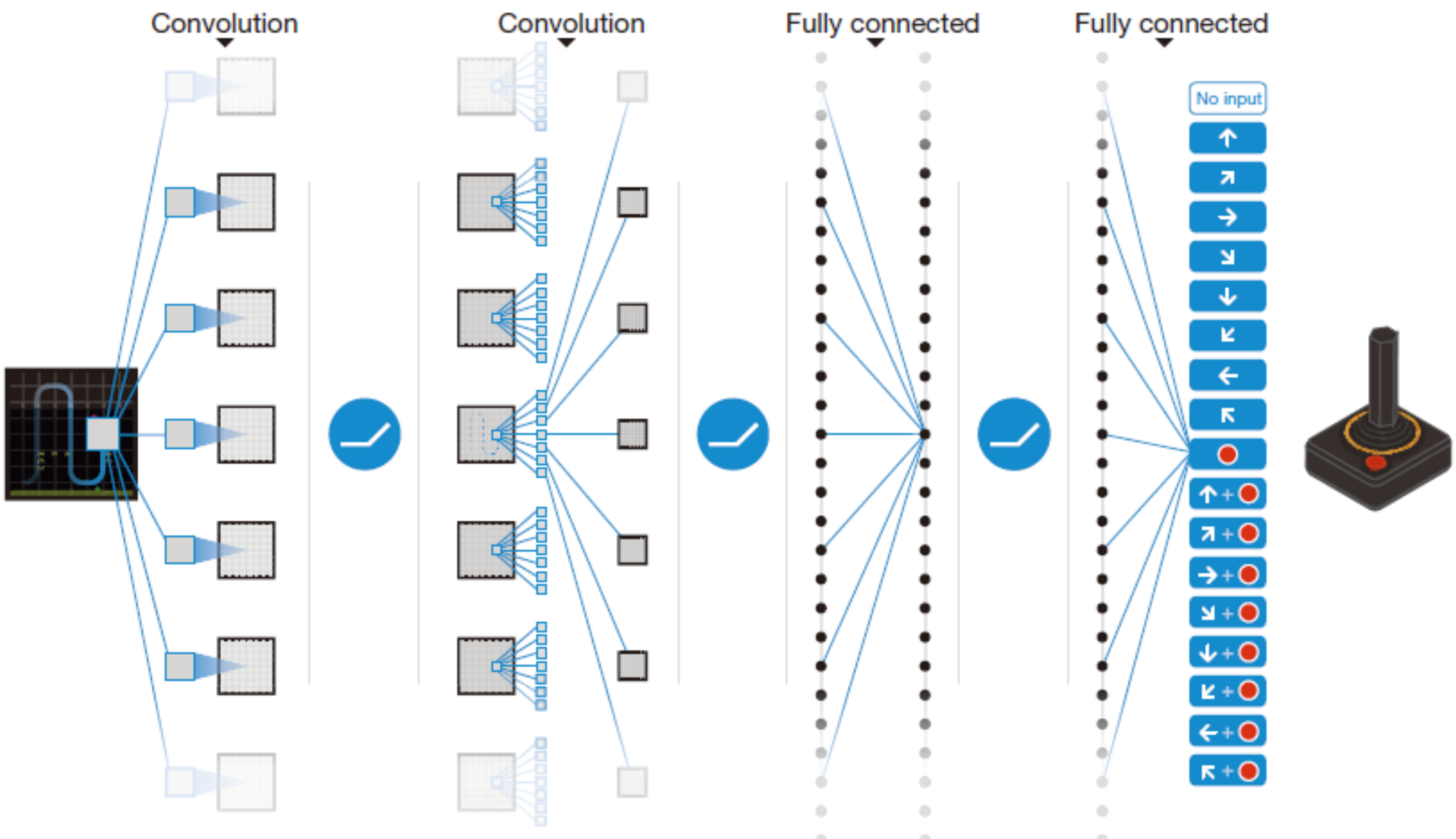


[1] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search.

Game Playing

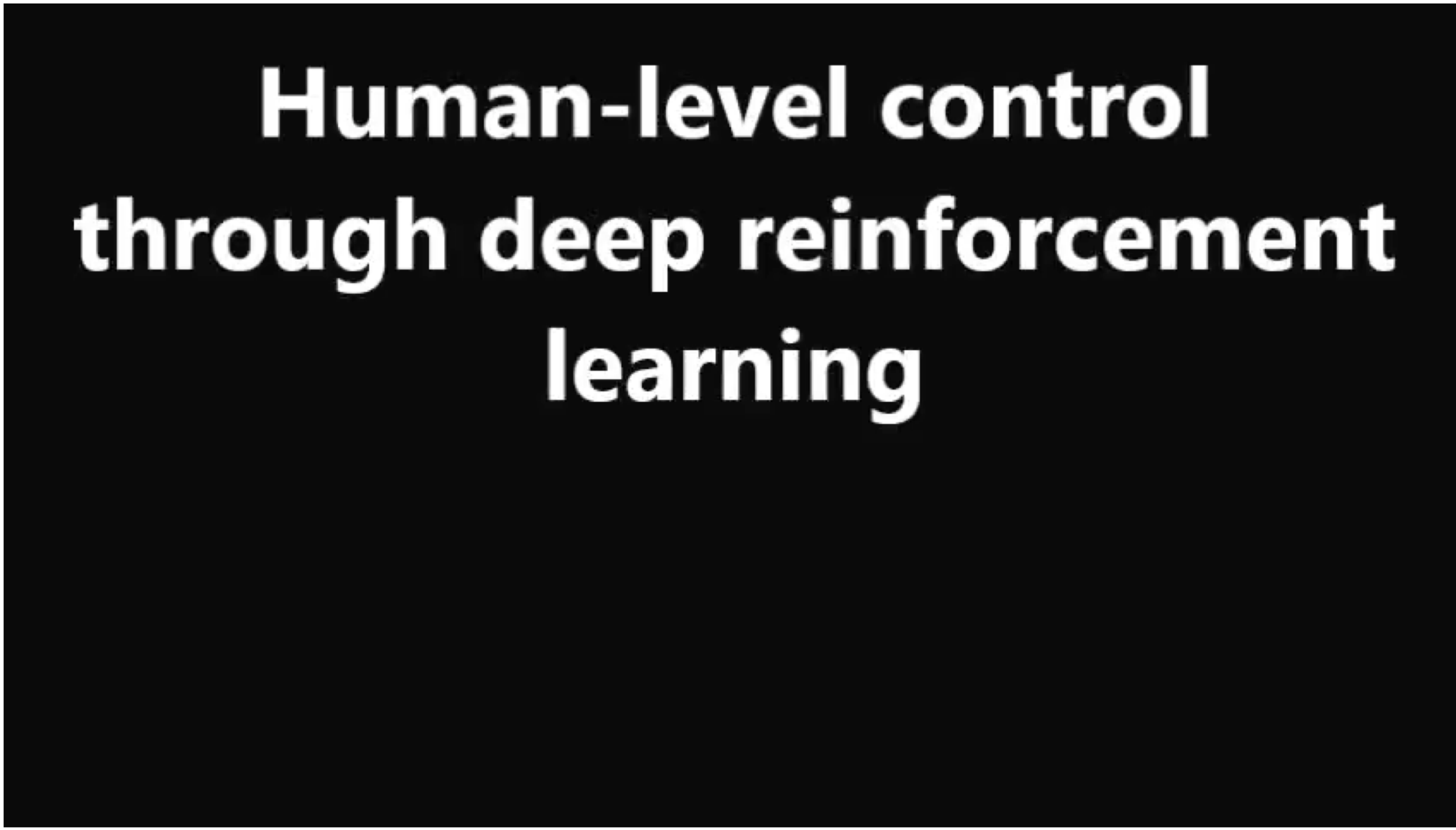
- Atari 2600 platform offers 49 games
- Google’s deep Q-network (DQN) performs the same as or better than the human expert in 29 games
- The same network, same learning algorithms

Game Playing



Deep Q-learning. Mnih et al., 2015

Game Playing



IBM: Use supercomputer to simulate the human brain

- 2012年11月13日，IBM宣布了世界上最大的人脑模拟计划“Compass”，他们使用了世界排名第二的超级计算机Sequoia Blue Gene/Q和一个全新的低功耗计算机架构，终于模拟出了与人脑相当的5300亿个神经元和137万亿个神经突触，向着真正的“人工大脑”迈出了重要的一步。

从猕猴大脑结构中推导出来的神经突触网络

Google Brain

Stanford University professor Andrew Ng started Google's Deep Learning project (which would later acquire the name *Google Brain*) in 2011 as one of the Google X projects.

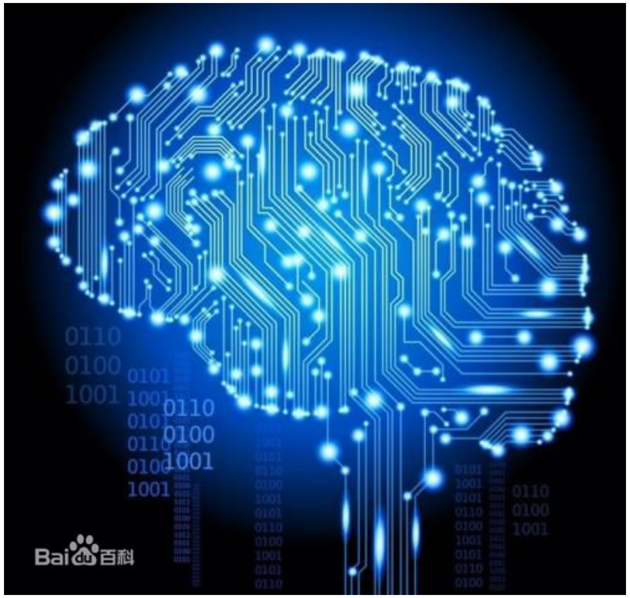
In March 2013, Google hired Geoffrey Hinton, and acquired the company DNNResearch Inc. headed by Hinton.

On 26 January 2014, multiple news outlets stated that Google had purchased DeepMind Technologies.

Baidu Brain

- 百度创始人李彦宏首次对外披露，百度目前正在推进一个名为“百度大脑”的项目，利用计算机技术模拟人脑，已经可以做到2-3岁孩子的智力水平。李彦宏表示，“相信随着硬件成本越来越低，计算能力越来越提升，计算机的能力将非常接近人的能力。”

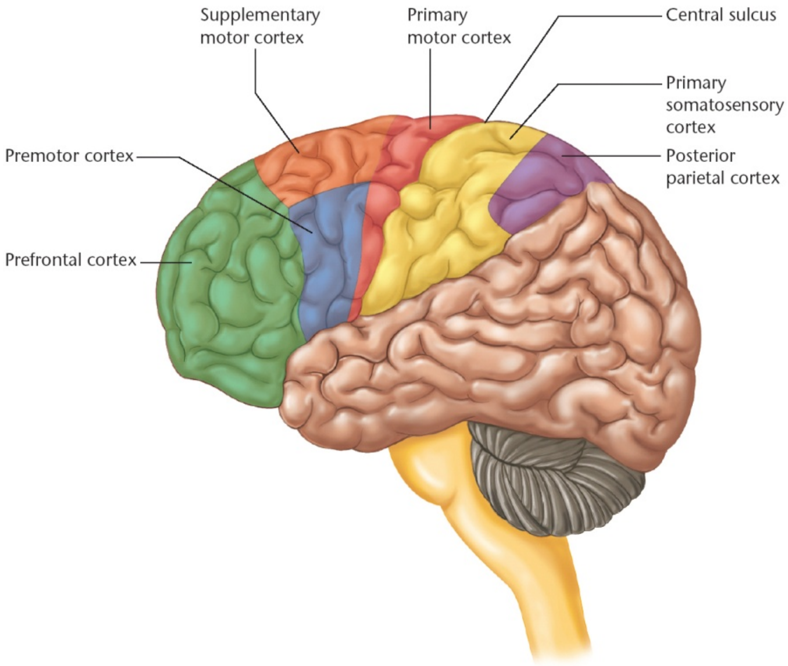
✓ 语音
✓ 图像
✓ 自然语言处理
✓ 用户画像

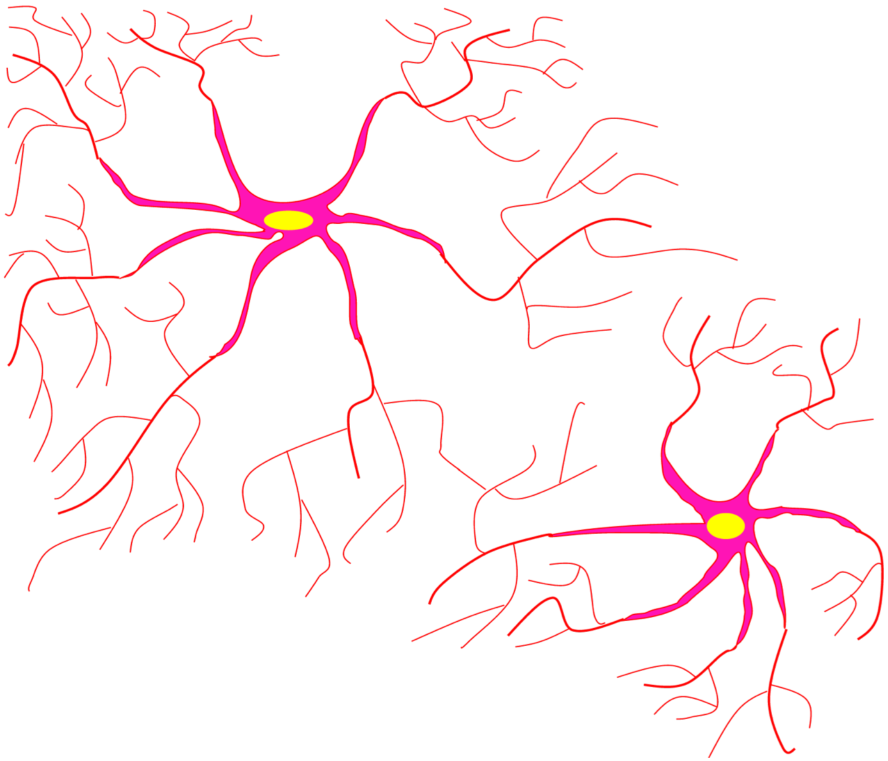


Origin, History, and Present

- Biological Discovery
- Bio-inspired Computational Models

Network of Neurons and Synapses

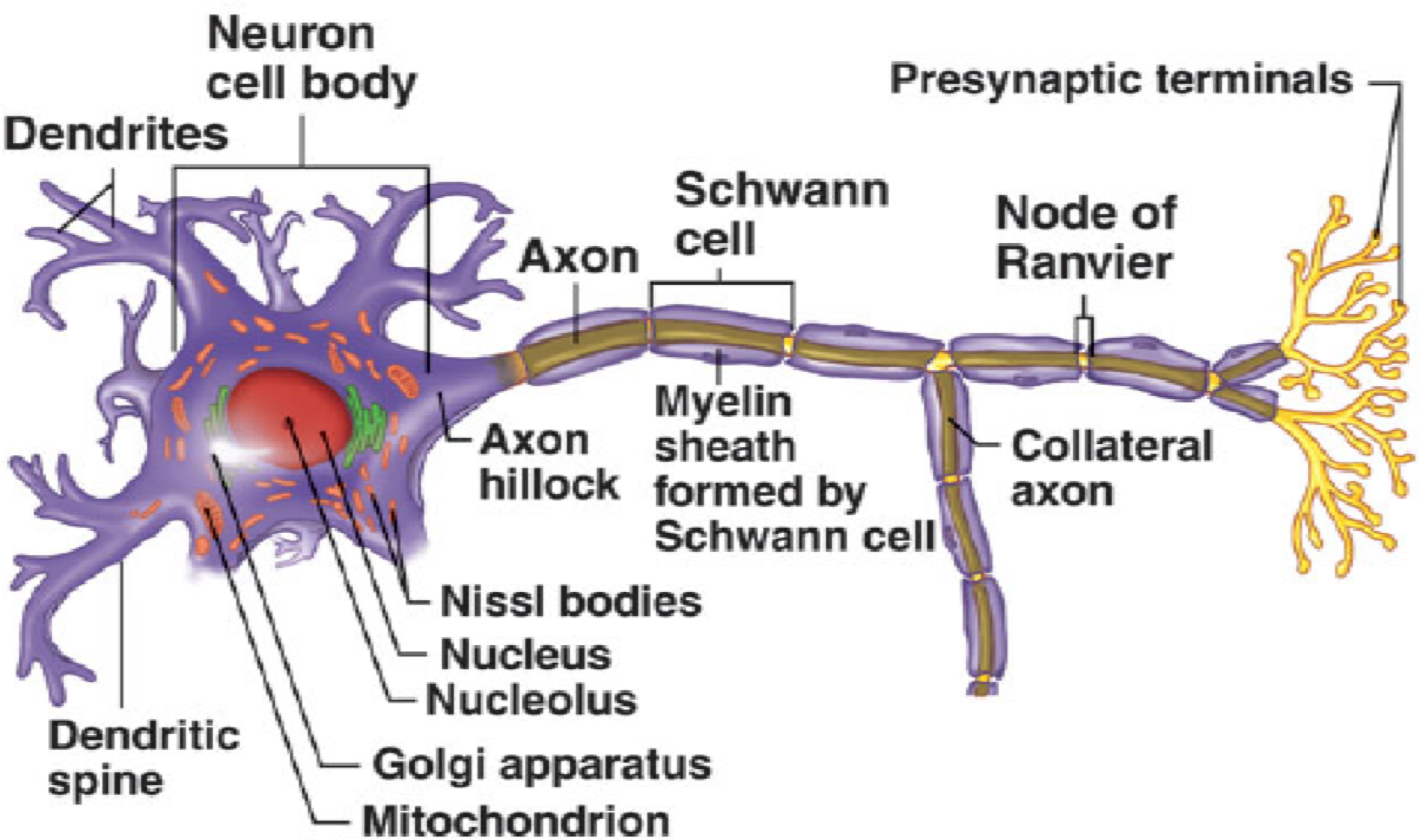




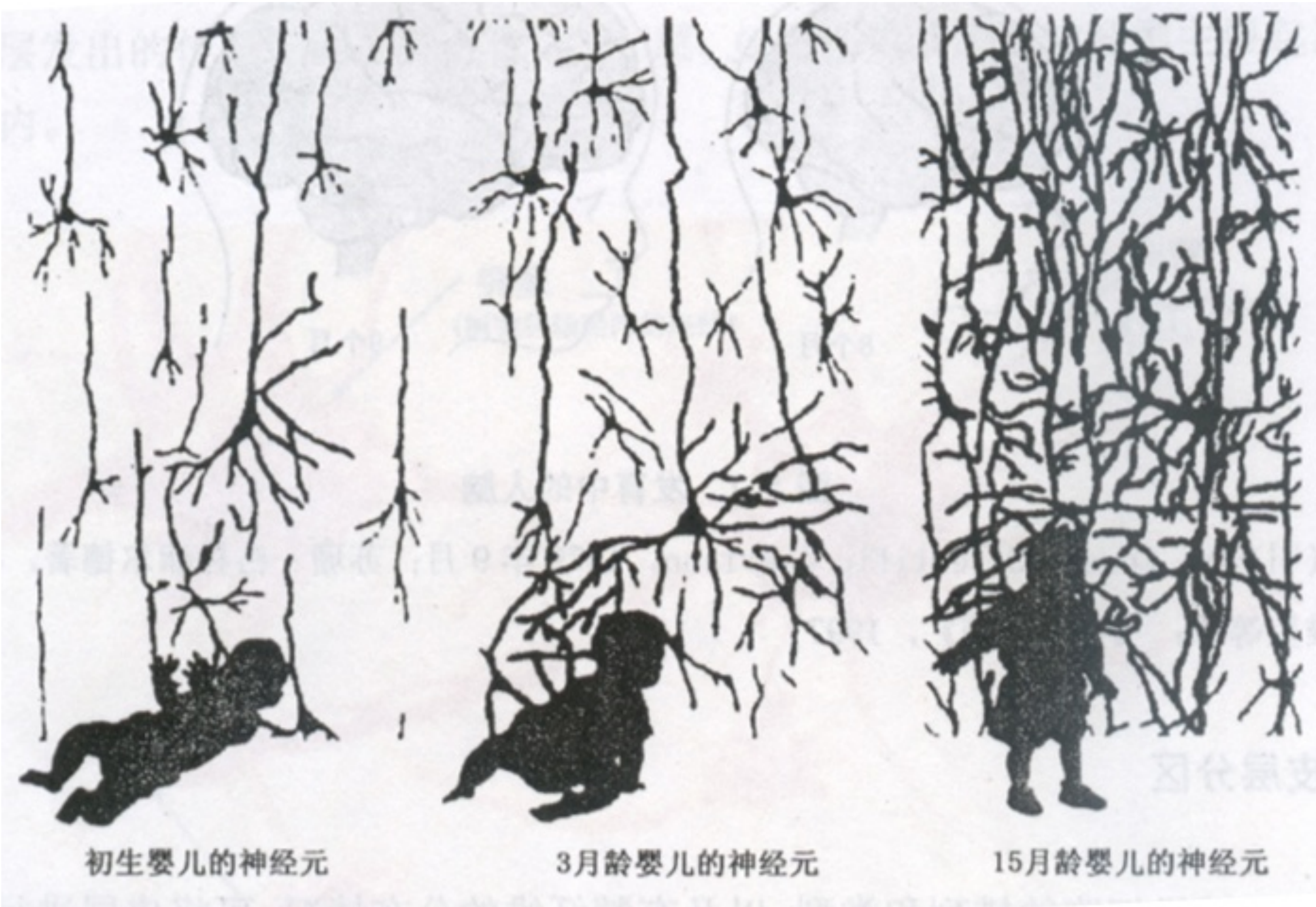
Some Facts

- Human brain contains **~10^{11}** neurons, **~10^{14}** synapses, **~10^{15}** glial cells
- Each neuron connected ~10^4 others
- Some scientists compared the brain with a "complex, nonlinear, parallel computer"
- The largest modern neural networks achieve the complexity comparable to a nervous system of a fly

Cells of Nervous System



Physiology Basis



Neuron and Perceptron

Neuron (biology)

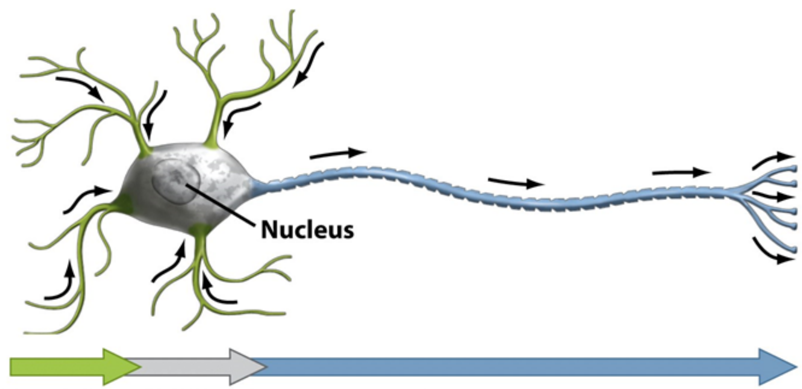
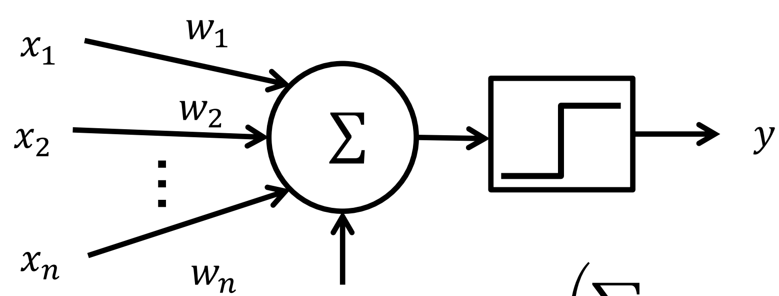


Figure 45-2b Biological Science, 2/e

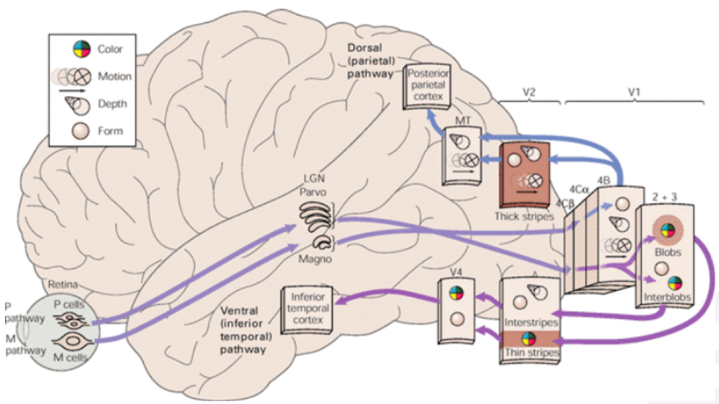
Perceptron (computation)



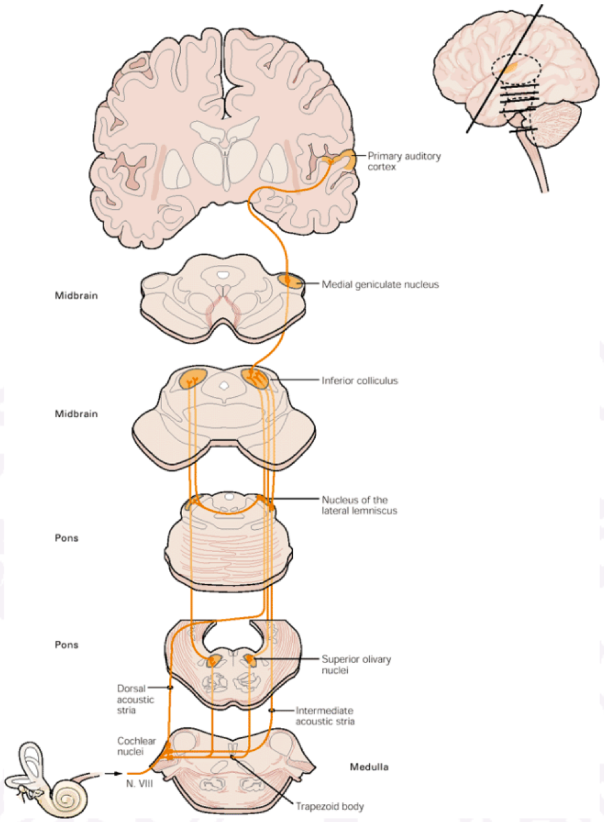
$$y = f\left(\sum_i w_i x_i + \theta\right)$$

Rosenblatt, Frank (1957), The Perceptron--a perceiving and recognizing automaton. Report 85-460-1, Cornell Aeronautical Laboratory.

Beyond Primary Visual Cortex

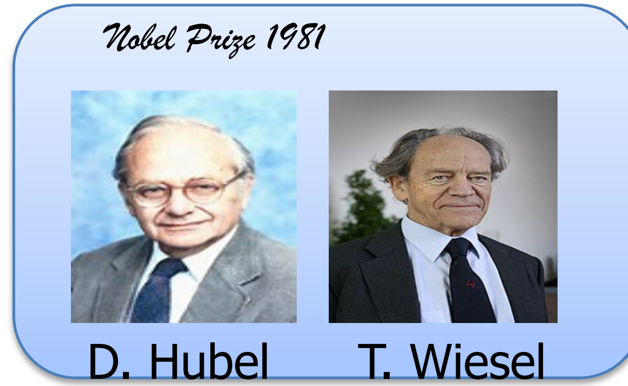


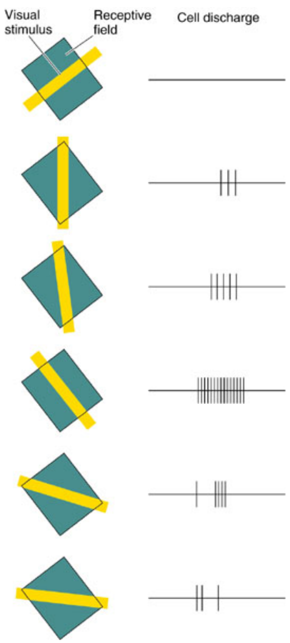
visual pathway



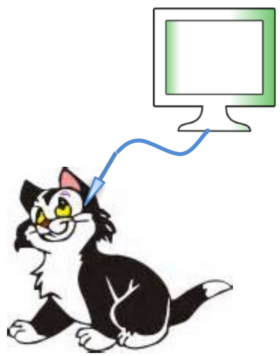
auditory pathway

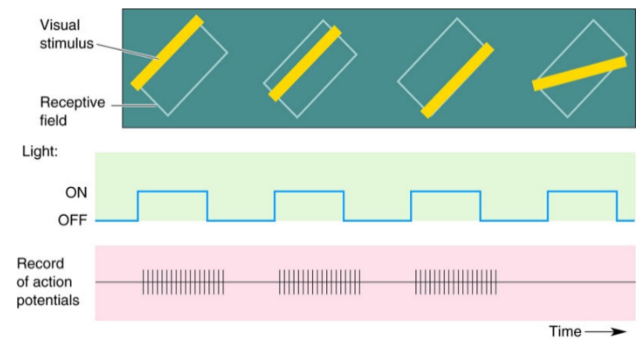
Simple Cell and Complex Cell





Simple cell
(location and orientation)

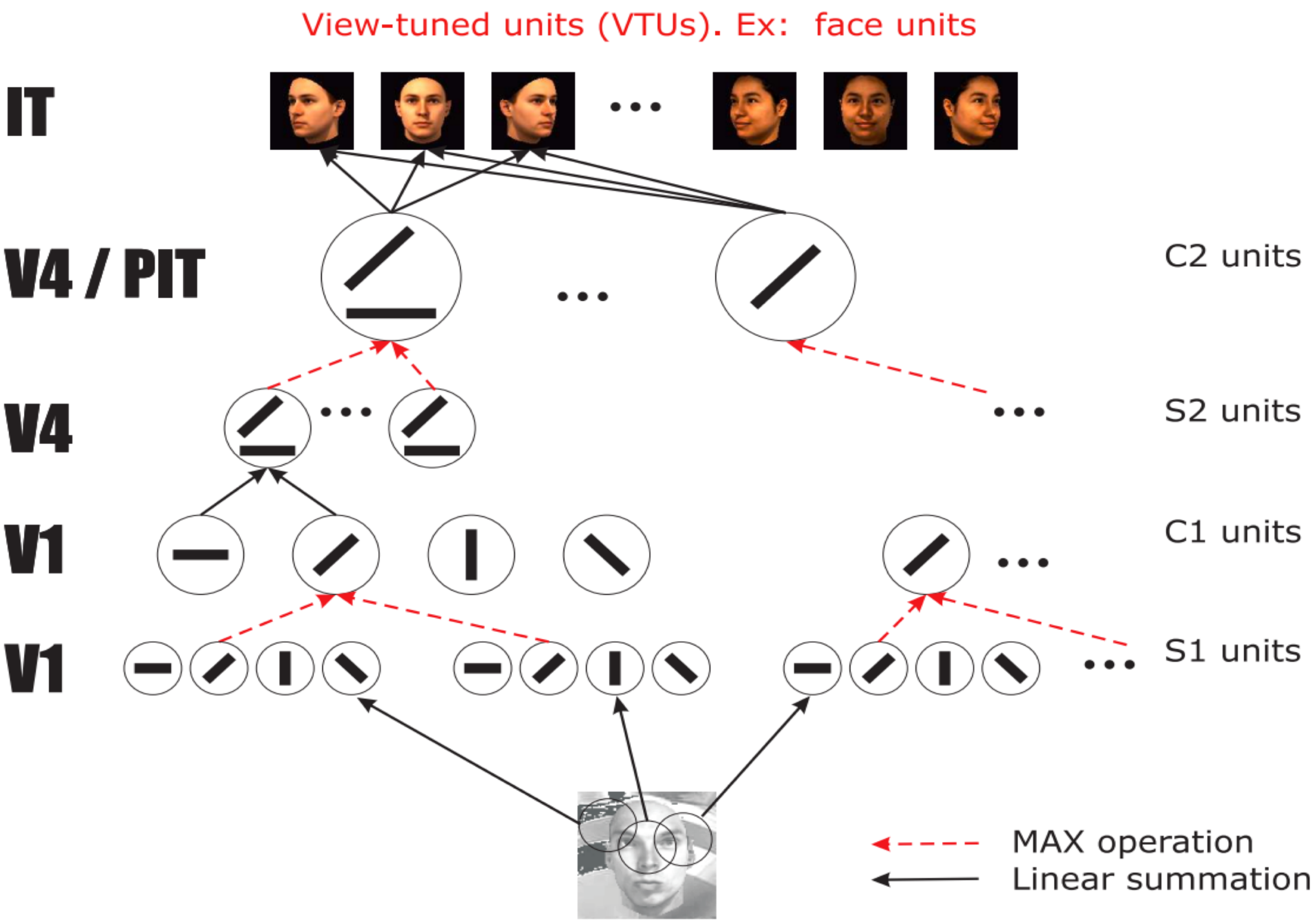




Complex cell
(only orientation)

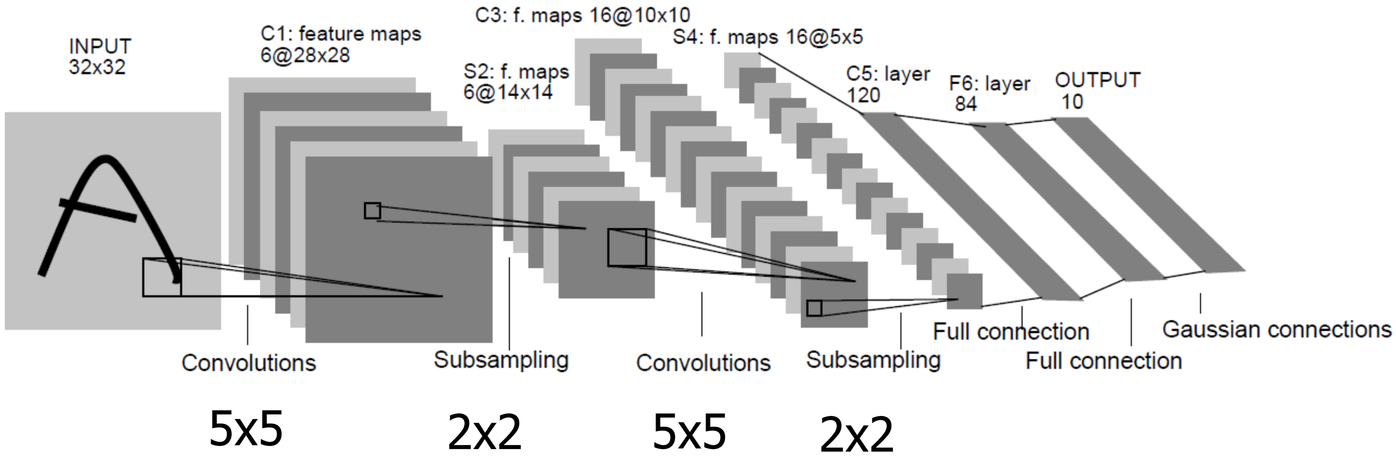
Bear, Connors, Paradiso, 2006

HMAX: Hierarchical Models of Object Recognition in Cortex



Riesenhuber, Poggio, 1999, MIT

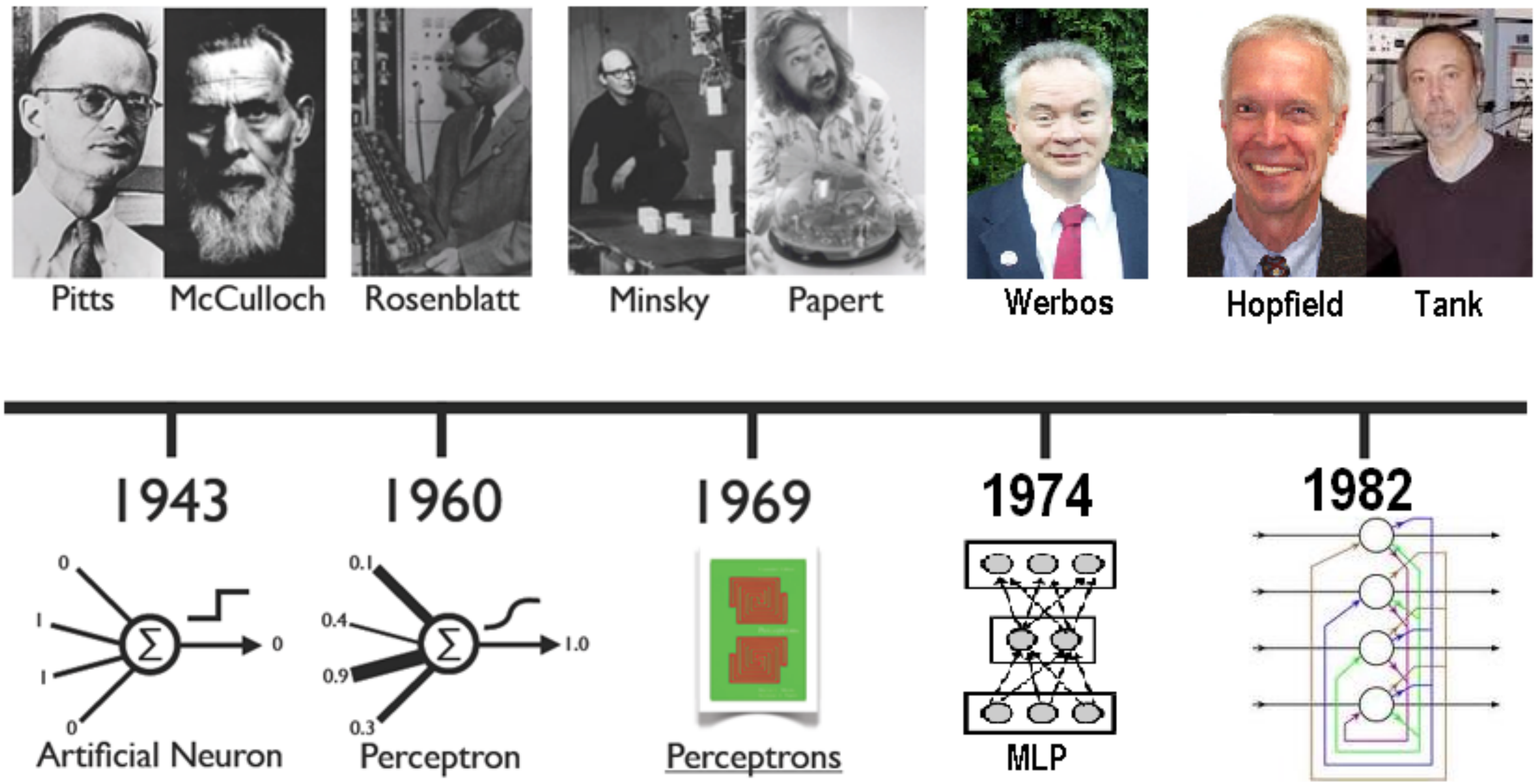
Convolutional Neural Network (CNN)



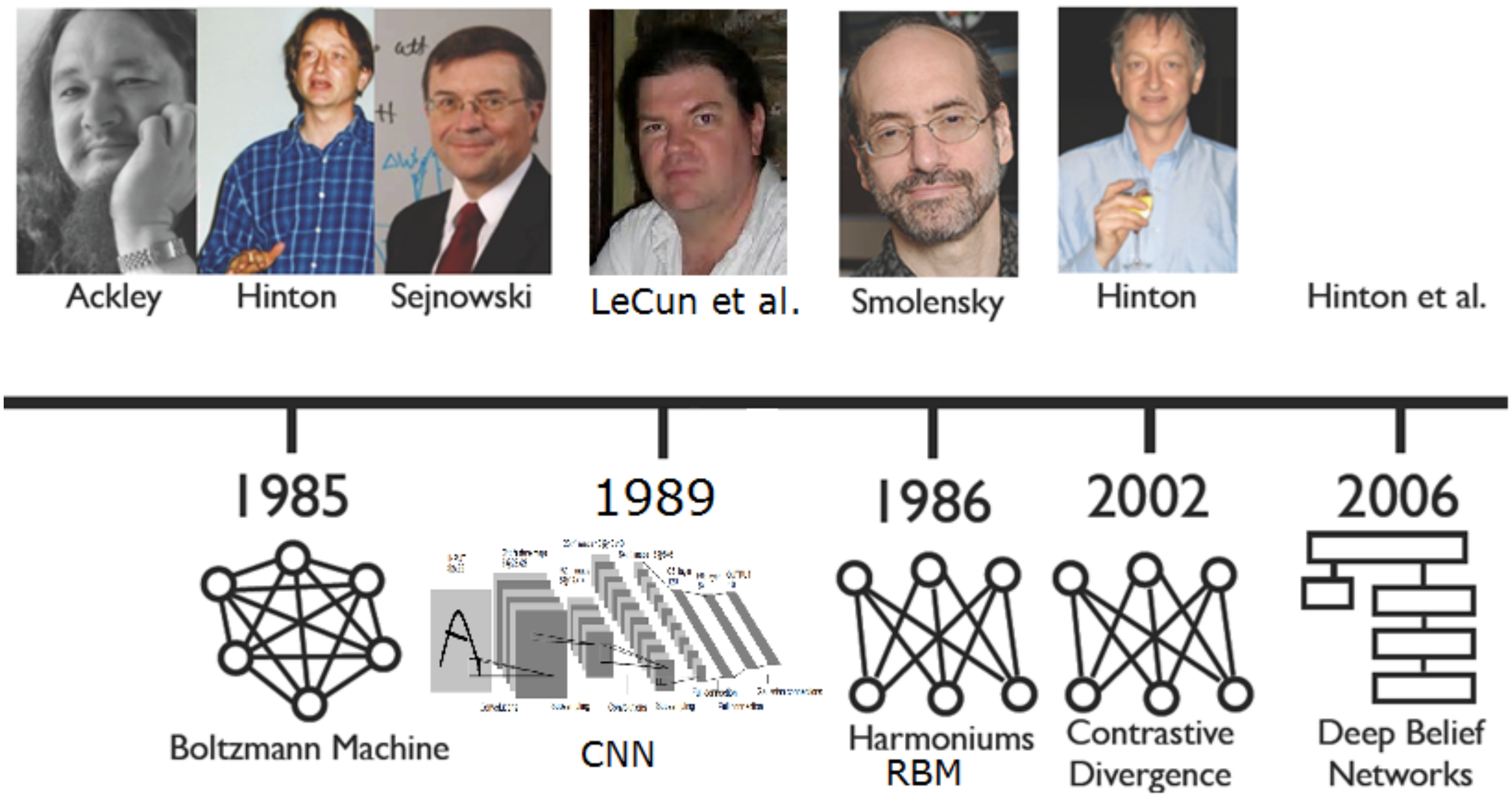
- Local connections and weight sharing
- C layers: convolution (simple cell)
  - Convolution filters (automatically learned)
  - Different feature maps or channels
- S layers: subsampling or pooling (complex cell)

History

Overview of History

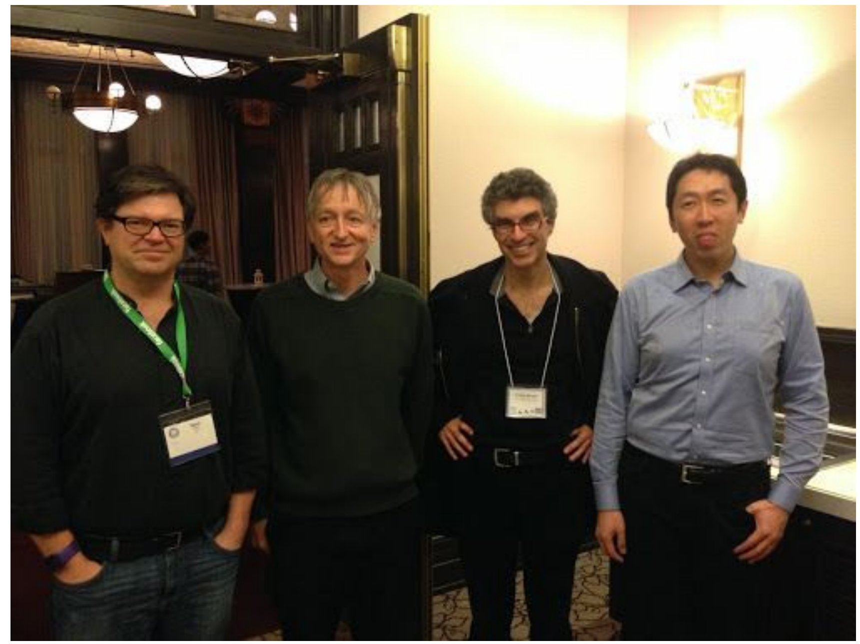


Overview of History (cont.)



Deep Learner

- Yann LeCun
- Geoffrey Hinton
- Yoshua Bengio



nature
International journal of science

Review Article | Published: 27 May 2015

Deep learning

Yann LeCun, Yoshua Bengio & Geoffrey Hinton

Nature **521**, 436–444 (28 May 2015) | Download Citation

Abstract

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object

https://www.nature.com/articles/nature14539

History (cont.)

- In **1957**, **Frank Rosenblatt** defined a neural network structure called Perceptron (感知器), and δ law learning algorithm.
- Contribution
  - Initially take neural network research from theory to application
  - Computer simulations at IBM70 yield evidence that this model was capable to achieve correct classification through tuning the weight learning
  - Bring the thriving period of neural network research

$$y_j = f\left(\sum_{i=1}^{n} w_{ij} x_i - \theta_j\right)$$

$$f(u_j) = \begin{cases} 1 & u_j \geq 0 \\ 0 & u_j < 0 \end{cases}$$

http://en.wikipedia.org/wiki/IEEE_Frank_Rosenblatt_Award

History (cont.)

- In **1969**, M.Minsky and S. Papert wrote a book named "**Perceptrons**", which indicated the perceptron only solved first-order Predicate Logic and linear partition.
- It failed to solve nonlinear or other partition problems, e.g., a **simple XOR problem**.

Marvin Minsky

1927-2016， MIT

Seymour Papert

1928-2016， MIT

History (cont.)

- **In 1982,** **John J. Hopfield** presented full-connected network and discrete neural network model
  - An energy function which is proportional to the active value of each neuron and the connection weight between neurons
  - Reach a local minimum by energy function degression. He proved the two cases (discrete and continuous) that network would reach steady state.

John J. Hopfield 1933-

History (cont.)

- In **1986**, a parallel computing group in US presented **Back Propagation (BP) algorithm** for Feed-forward Network.
- The back propagation algorithm becomes one of the most widely used methods nowadays.

Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. (8 October 1986). "Learning representations by back-propagating errors". *Nature*. **323** (6088): 533–536.

History (cont.)

Vladimir

Vapnik

- When the network becomes deeper, BP algorithm will be trapped in local minima.
- During 1990s, Vladimir Vapnik and his co-workers developed SVM, which is a specialized two-layer neural network.
  - a convex QP problem
- Vapnik received the 2010 Neural Networks Pioneer Award, the 2012 IEEE Frank Rosenblatt Award, etc.
- BP network researches then encounter "Winter" in 1990s.

History (cont.)

Geoffrey Hinton

1947-

Deep Learning

- In 2006, the concept of deep learning was brought to the community by Geoffrey Hinton.
- Essentially it is a new learning paradigm for BP networks. The basic idea is to learn a representation of the data in an unsupervised way, then apply the BP algorithm for fine tuning the weights.
- The unsupervised learning algorithm is based on a stochastic network, restricted Boltzmann machine (RBM), which is in agreement of the structure of the BP network.
- BP network researches have revived since then.

Neural Networks are coming back

Present – Neural Network

Today neural network is more connected with

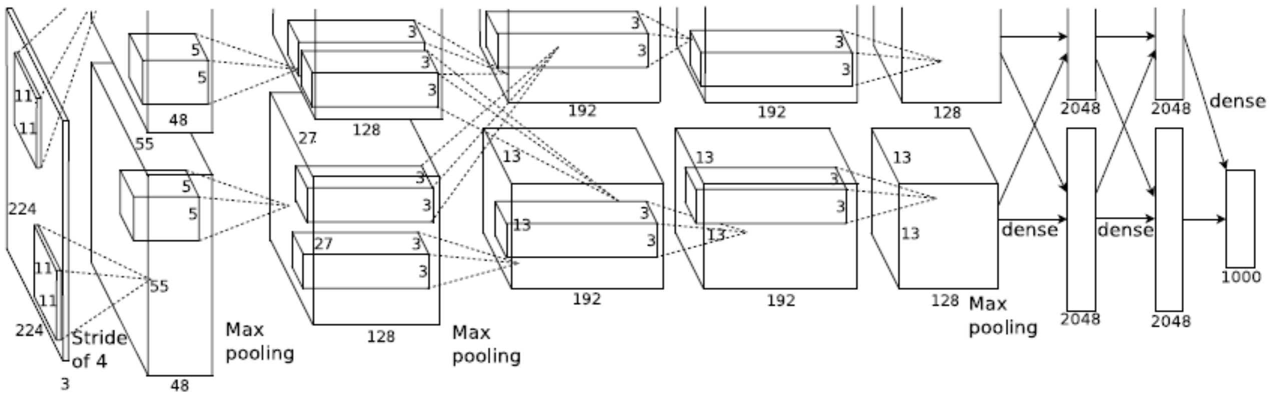
DEEP LEARNING!

- ◆ **Very large networks**
- ◆ **Very deep structures**
- ◆ **Impressive performance**

Present - 1st of 10 breakthrough technologies 2013

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart

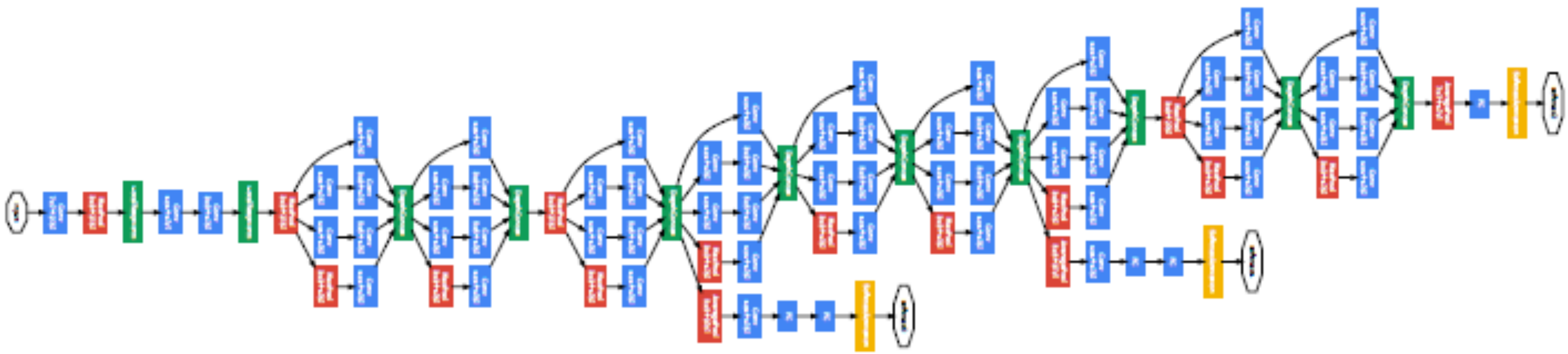
CNN for image classification

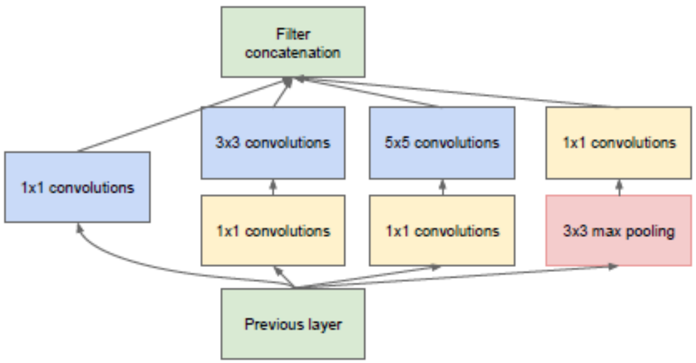


- Network dimension: 150,528(input)-253,440–186,624–64,896–64,896–43,264–4096–4096–1000(output)
- In total: 60 million parameters
- Task: classify 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes
- Results: state-of-the-art accuracy on ImageNet

Krizhevsky, Sutskever and Hinton, NIPS, 2012

GoogLeNet

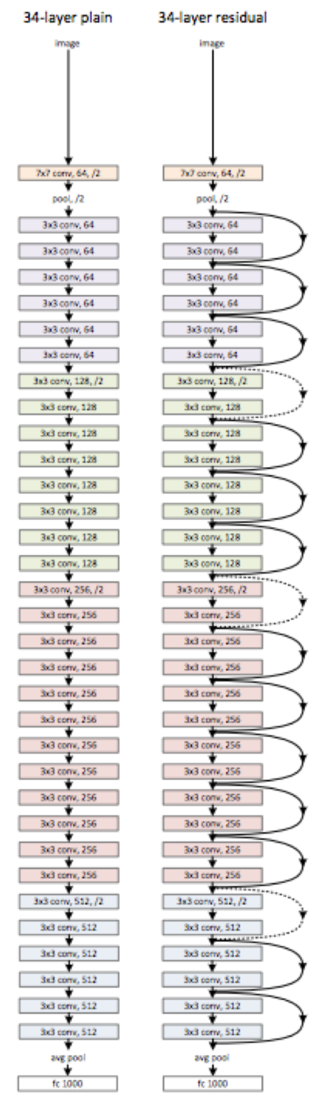




Szegedy, et al., 2014

- 22 weight layers
- Small filters (1x1, 3x3, 5x5)
- Two auxiliary classifiers connected to intermediate layers are used to increase the gradient signal for BP algorithm
- A cpu-based implementation on distributed system

Residual Network



- Won the 1st place on the ILSVRC 2015 **classification**, ImageNet **detection**, ImageNet **localization**, COCO **detection**, and COCO **segmentation** [1]
- *The ImageNet dataset, up to* ***152*** *layers—8× deeper than VGG nets*
- *CIFAR-10 dataset, 100 and 1000 layers*

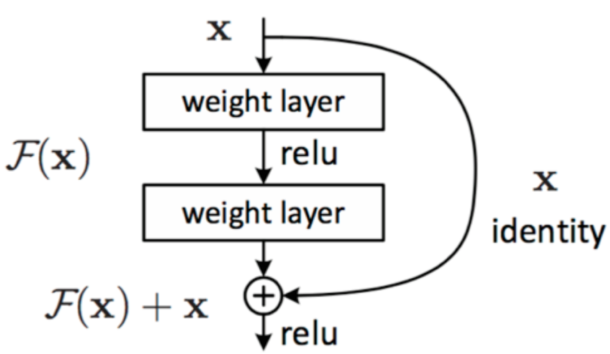
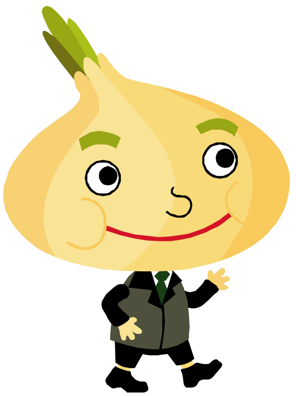


Figure 2. Residual learning: a building block.

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition.



Basics of ANN

Excitation and Inhibition

- The receptors of a neuron are called **synapses**, and they are located on many branches called **dendrites.** There are many types of synapses, but roughly they can be divided into two classes:
  - **Excitatory**: a signal received at this synapse "encourages" the neuron to fire.
  - **Inhibitory**: a signal received at this synapse will try to make the neuron "shut up"

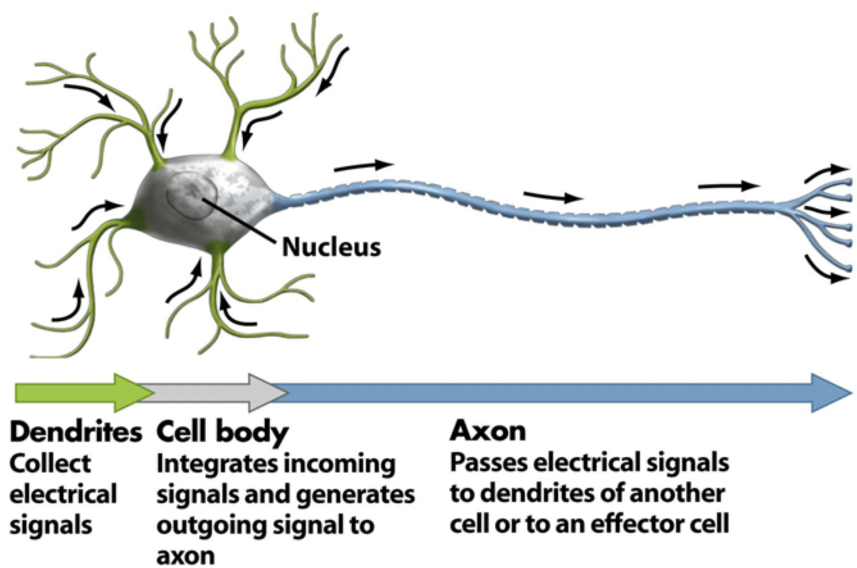
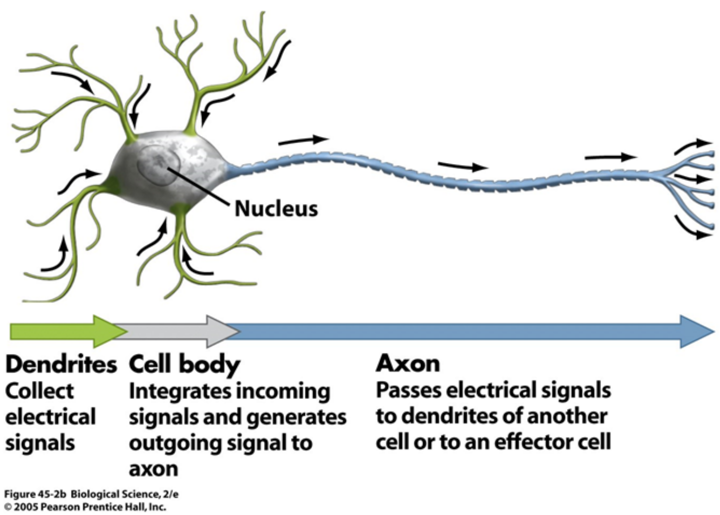


Figure 45-2b Biological Science, 2/e

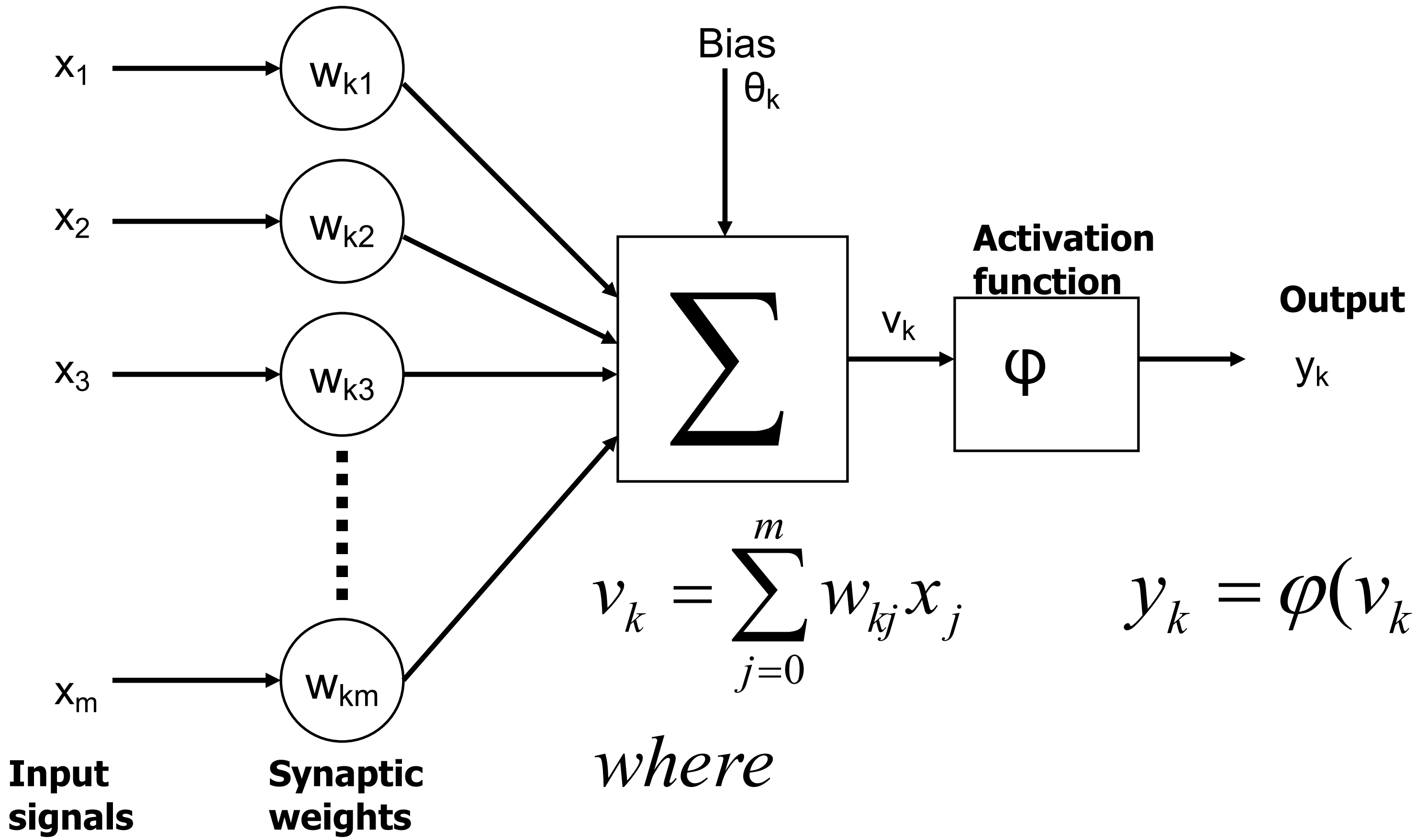
Weighted Input

- Synapses (receptors) of a neuron have weights. $W=(w_1, w_2, \ldots, w_n)$, which can have positive (excitatory) or negative (inhibitory) values. Each incoming signal is multiplied by the weight of the receiving synapse $w_i x_i$. Then all the "weighted" inputs are added together into a weighted sum v:

$$v = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n = \langle w, x \rangle$$



Neuron Model



$$v_k = \sum_{j=0}^{m} w_{kj} x_j \qquad y_k = \varphi(v_k)$$

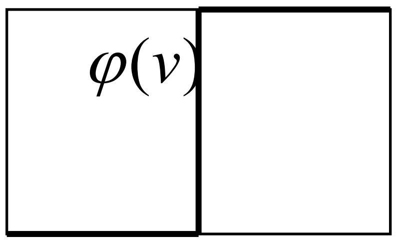
where

$$x_0 = -1$$

$$w_{k0} = \vartheta_k$$

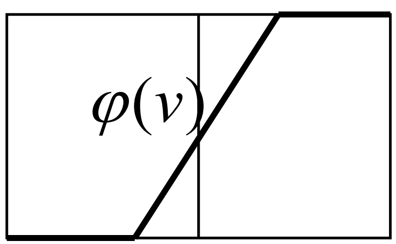
Activation Functions

- Threshold function



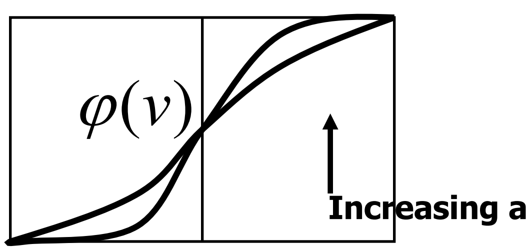
$$\varphi(v) = \begin{cases} 1 & if \quad v \geq 0 \\ 0 & if \quad v < 0 \end{cases}$$

- Piecewise-linear function



$$\phi(v) = \begin{cases} 1 & v >= \frac{1}{2} \\ v + \frac{1}{2} & \frac{1}{2} > v > -\frac{1}{2} \\ 0 & v \leq -\frac{1}{2} \end{cases}$$

- Sigmoid/Hyperbolic function

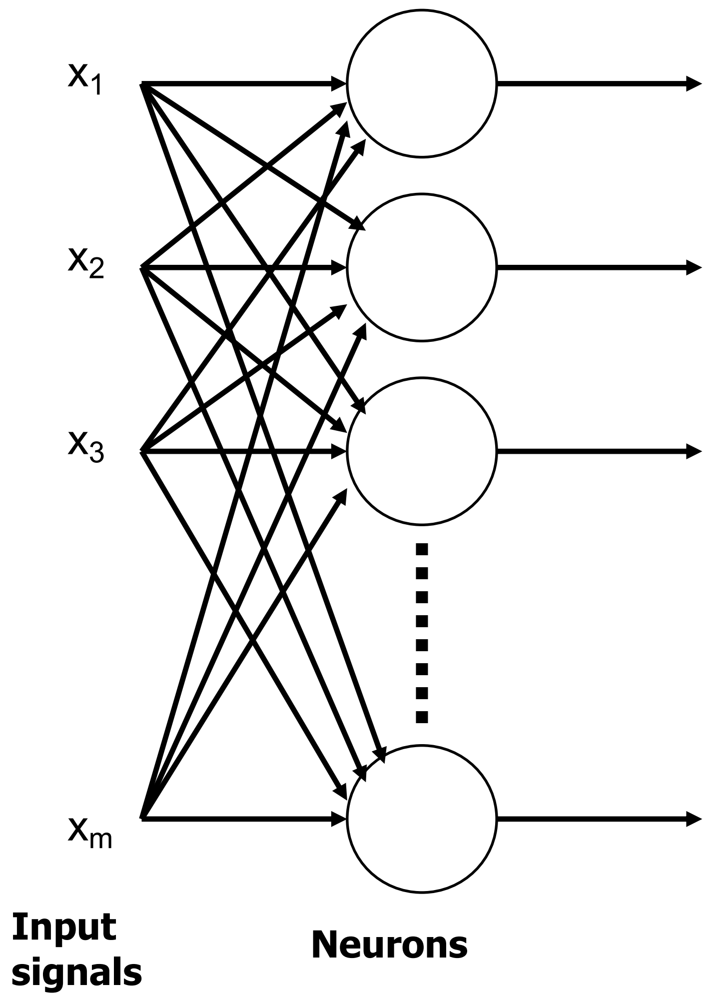


$$\varphi(v) = \frac{1}{1 + \exp(-av)} \qquad \varphi(v) = \tanh(av)$$

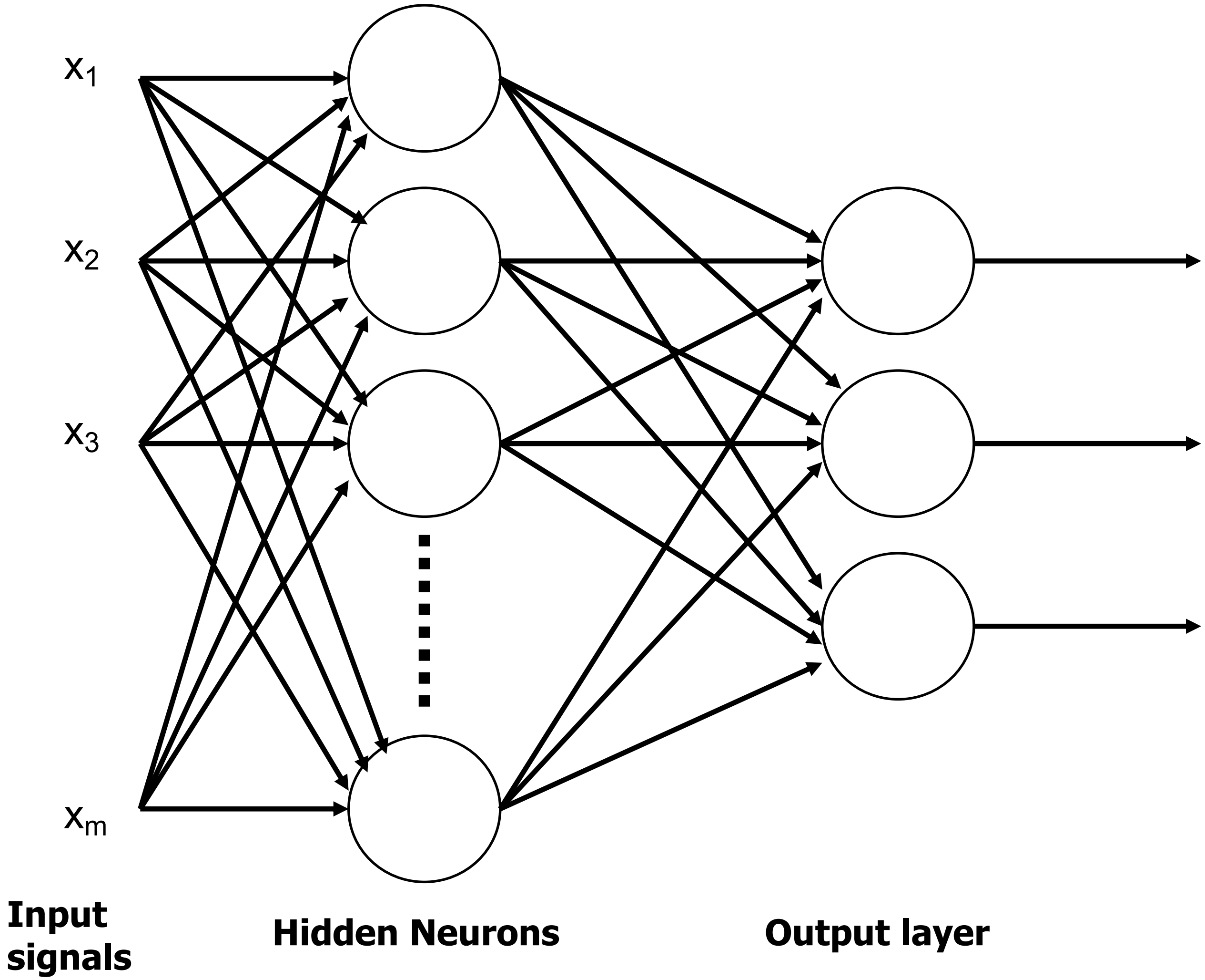
Typical Structures

- Feed-forward Networks
- Feed-back Networks

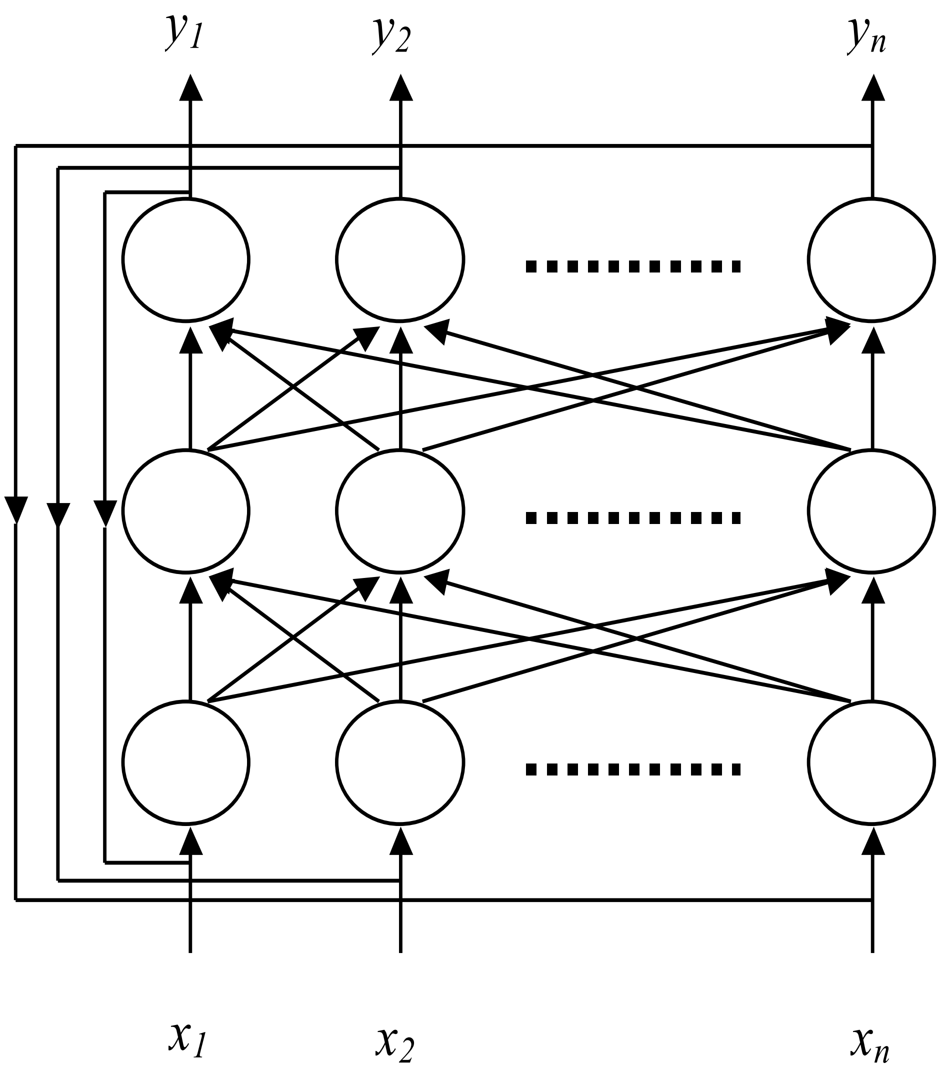
Single-layer Feed-forward Network



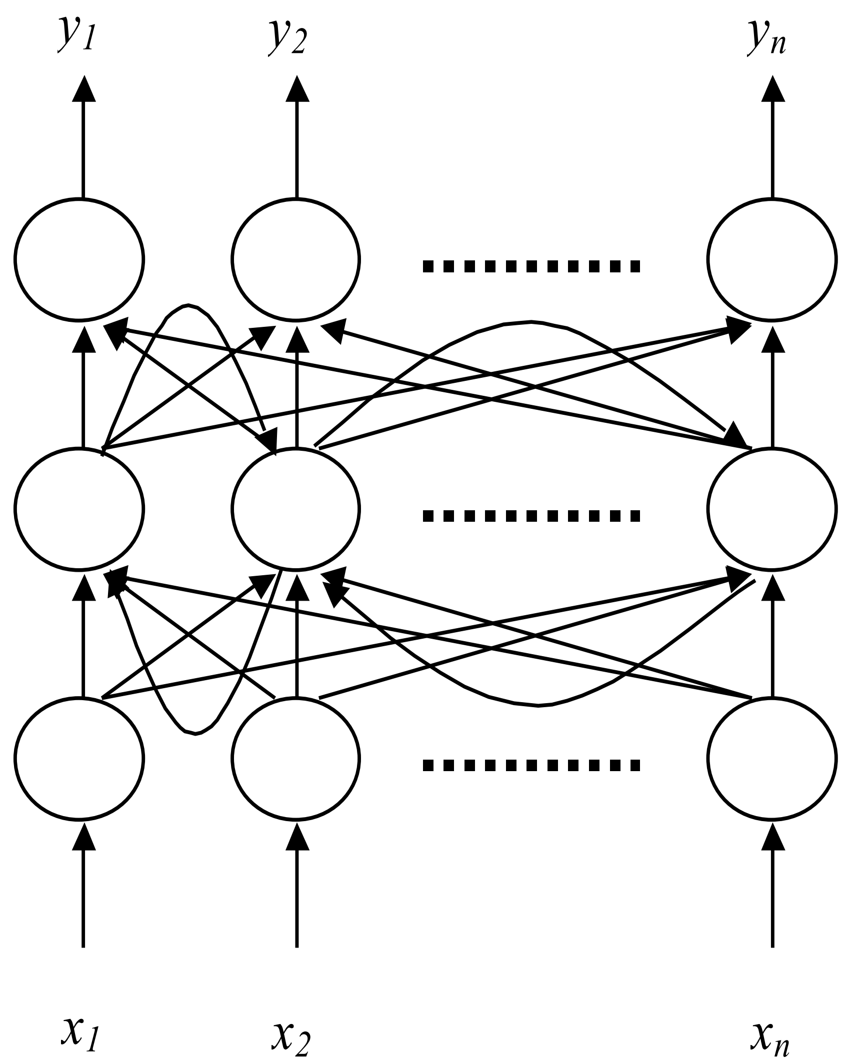
Multilayer Feed-forward Network



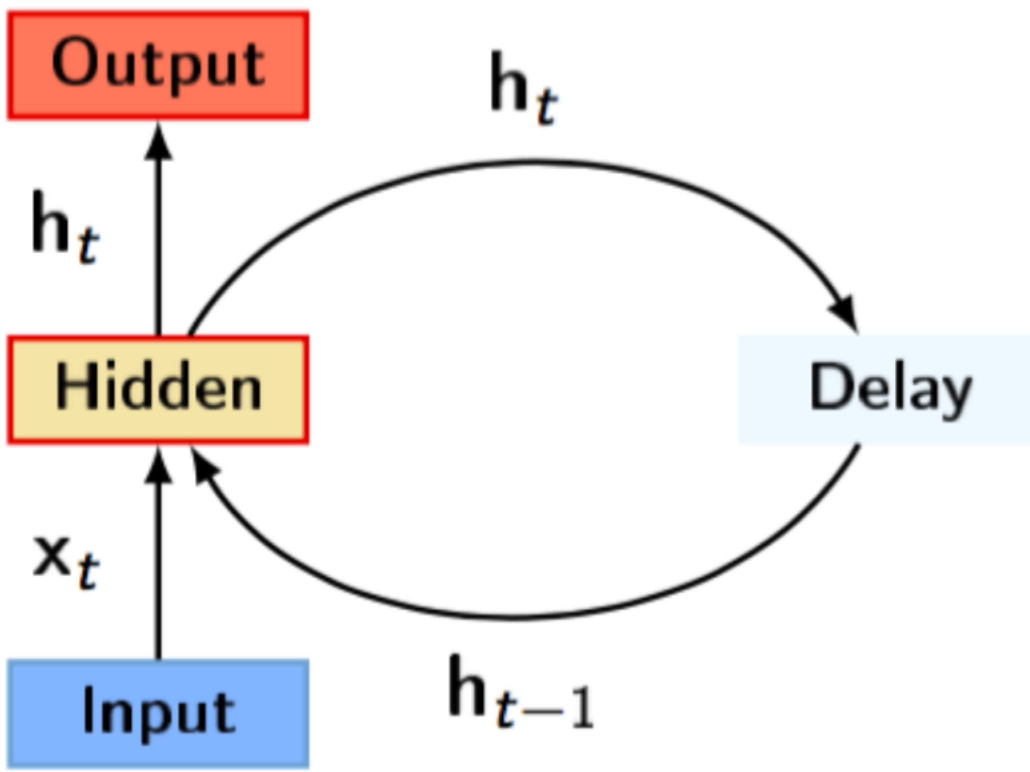
Feed-back Feed-forward Network



Inter-feedback Feed-forward Network



Recurrent Neural Networks



Characteristics

Characteristics

- Non-linearity (非线性)
  - Non-linearity is a ubiquitous feature of nature
- Non-locality (非局域性)
  - The feature of single neuron
  - **Interaction and connection between neurons**

Characteristics (cont.)

- Non-constancy (非定常性)
  - Neural network is capable of self-adaption, self-organization, and self-learning
- Non-convexity (非凸性)
  - the function has more than one extremums, i.e., system has multiple steady states

Basics of Machine Learning

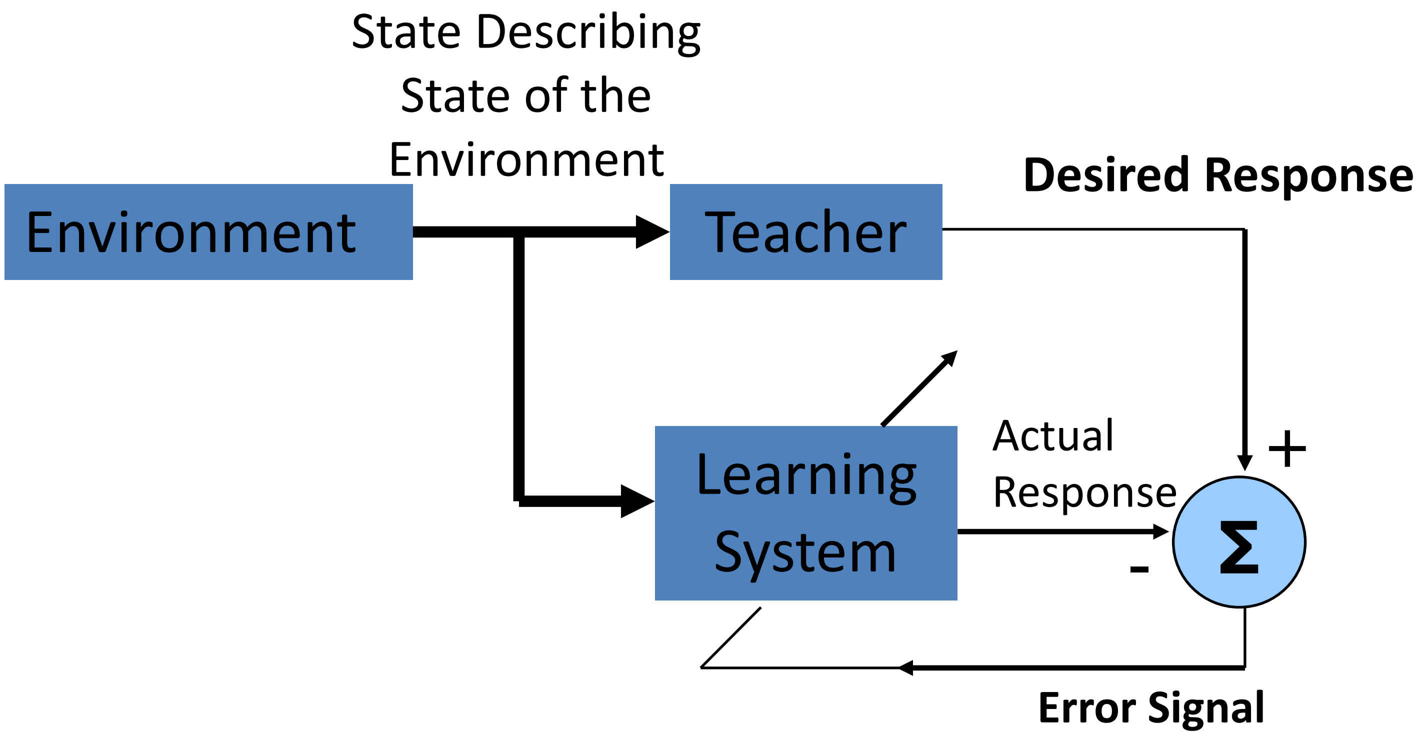
Learning

- Learning is essential for unknown environments,
  - i.e., when designer lacks omniscience
- Learning is useful as a system construction method,
  - i.e., expose the agent to reality rather than trying to write it down
- Learning modifies the agent's decision mechanisms to improve performance

Machine Learning

- X: sample; Y: output
- Training data: $\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$
- Supervised Learning P(Y|X)
  - Most Neural Models (MLP, CNN, RNN)
- Unsupervised learning P(X)
  - Autoencoders
  - Boltzmann Machine

Supervised Learning

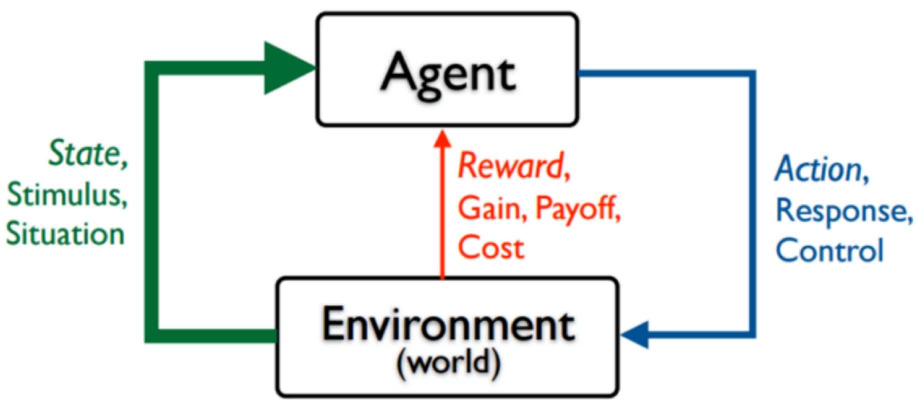


Supervised Learning

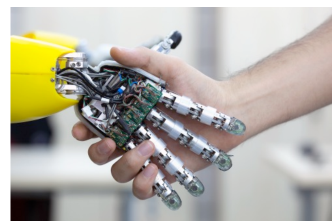
- **Maximum Likelihood Estimation**: maximizing the probability of observing the data
  - **Max** $\boldsymbol{\sum_{i=1}^{n} logP(x_i, y_i)}$
- **Cost Function** defines the cost when an error is given by the learning agent.
  - Mean Square Error: $E_{p(x,y)}[(\hat{y} - y)^2]$
  - Cross Entropy: $H(\hat{y}, y) = E_{p(\hat{y})}[-logp(y)]$
  - Hinge Loss: $l(y) = \max(0, \gamma - \hat{y} * y)$

Deep Reinforcement Learning

Learning through interactions with the environment



Deep learning to represent states, actions, or policy functions



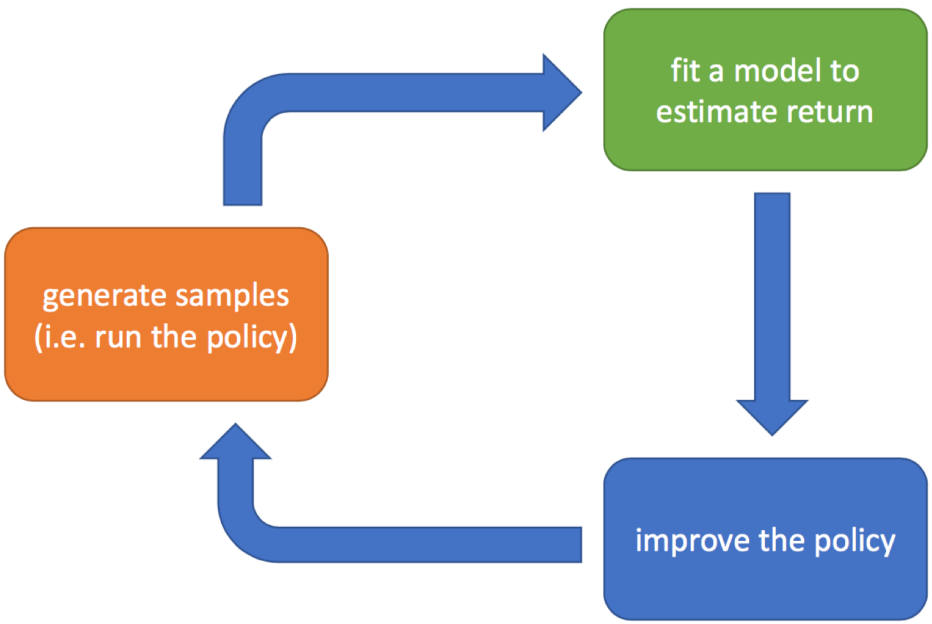
机器人控制

自主驾驶

语言交互

系统运营

Deep Reinforcement Learning



compute $\hat{Q} = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$ (MC policy gradient)

fit $Q_\phi(\mathbf{s}, \mathbf{a})$ (actor-critic, Q-learning)

estimate $p(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ (model-based)

$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$ (policy gradient)

$\pi(\mathbf{s}) = \arg\max Q_\phi(\mathbf{s}, \mathbf{a})$ (Q-learning)

optimize $\pi_\theta(\mathbf{a}|\mathbf{s})$ (model-based)

From ICML Tutorial by Sergey Levine and Chelsea Finn

End