第一讲：深度学习发展历史

History of Deep Learning

张盛平

s.zhang@hit.edu.cn

计算学部
哈尔滨工业大学

2021 年秋季学期

致谢

- 讲稿中很多资料或素材来源于网络，包括国外一些大学的相关课程、一些博客、维基百科等。后面不一一列举来源，在此一并表示感谢
- 引用网络资源时，由于本人的理解能力，可能存在一些偏差
- 本讲稿会经常更新

§1: Biological Neurons

Reticular Theory（网状理论）

早在 18 世纪初，科学家就提出了“所有生物组织都是由细胞组成”的假设。然而，神经组织一直是个例外，因为人们始终无法找到神经细胞

1871 年，德国解剖学家 Joseph von Gerlach 认为神经系统是一个单一的互联网状结构，中间不存在任何断点，也没有所谓「独立神经细胞」

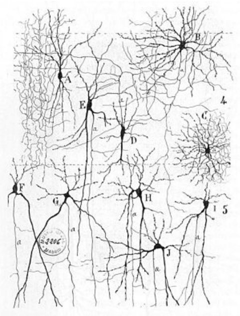
Staining Technique（染色技术）

1873 年, 意大利生理学家 Camillo Golgi (卡米洛 · 高尔基) 发明了铬酸银染色法, 可以清楚观察神经纤维走向

他是网状理论的支持者

Neuron Doctrine（神经元学说）

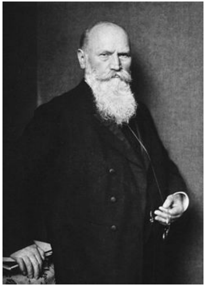
1891 年，西班牙组织学家 Santiago Ramón y Cajal (圣地亚哥 · 拉蒙 · 卡哈尔) 使用高尔基的染色技术发现神经系统是由单一神经元相连而成，神经元包含胞体、树突及轴突，神经之间通过突触彼此联结



The Term Neuron

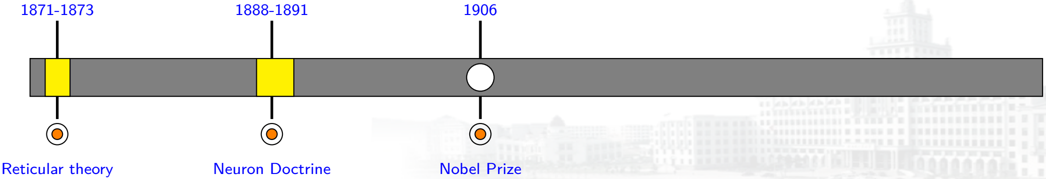
神经元一词是由 Heinrich Wilhelm Gottfried von Waldeyer-Hartz 于 1891 年左右创造的

他进一步巩固了神经元学说



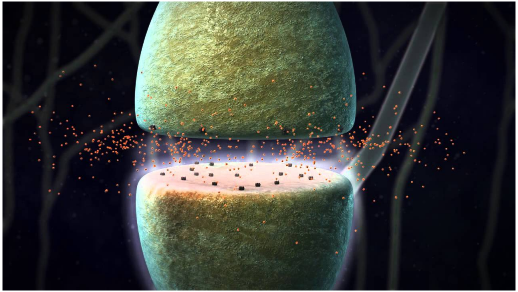
Nobel Prize

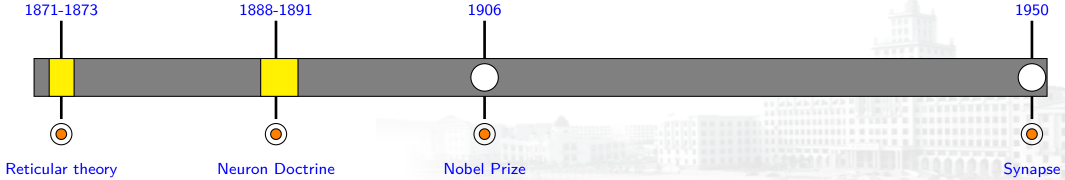
1906 年第六届诺贝尔生理学 / 医学奖，颁给了意大利生理学家卡米洛 ▪ 高尔基 (Camillo Golgi）和西班牙组织学家圣地亚哥 ▪ 拉蒙-卡哈尔（Santiago Ramón y Cajal)，这也是诺贝尔生理学医学奖第一次同时颁给两位获奖人。



The Final Word

20 世纪 50 年代，电子显微镜观察到单个神经细胞通过触突相互连接，进一步证实了神经元学说。

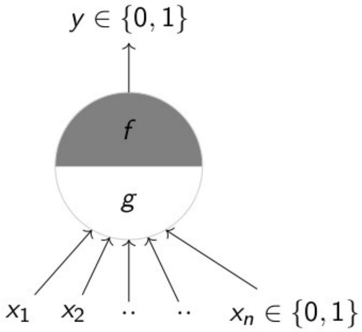


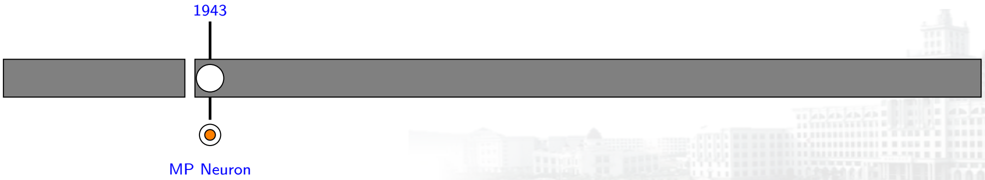


§2: From Spring to Winter of Neural Network

M-P 神经元

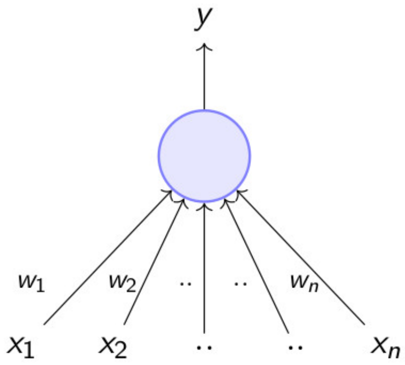
1943 年, McCulloch (神经科学家) and Pitts (逻辑学家) 提出了一个高度简化版的神经元模型，称为 M-P 神经元 [1]，从而开创了人工神经网络研究的时代

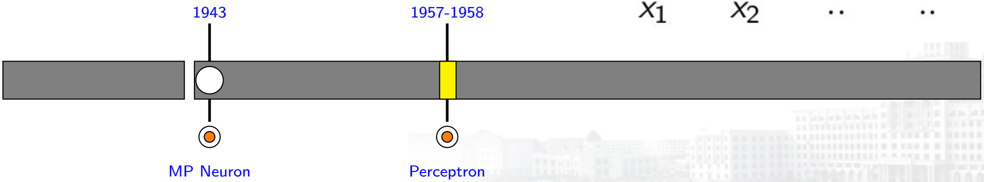




感知机（Perceptron）

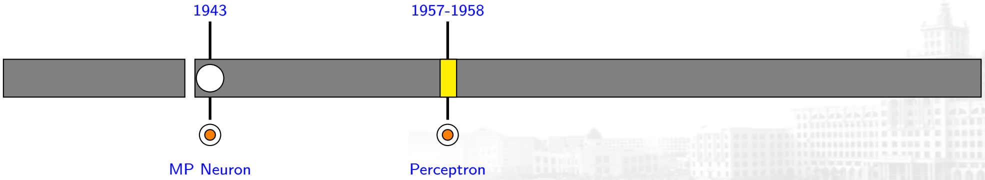
Frank Rosenblatt（弗兰克·罗森布拉特）提出了可以模拟人类感知能力的机器，并称之为『感知机』，它可以被视为一种最简单形式的前馈神经网络





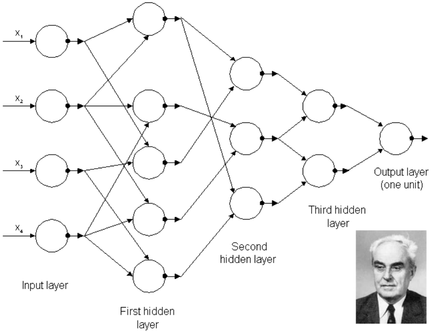
感知机学习算法

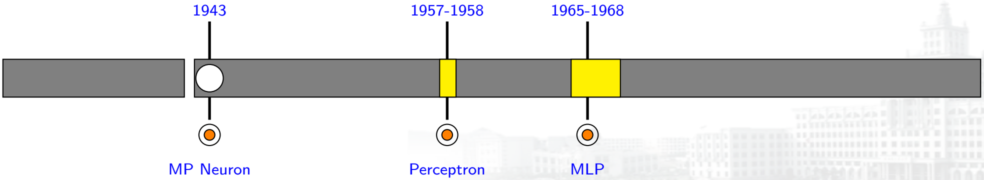
为了‘教导’感知机识别图像，弗兰克·罗森布拉特在 Hebb 学习法则的基础上，发展了一种迭代、试错、类似于人类学习过程的学习算法——感知机学习算法



多层感知机(Multilayer Perceptrons)

1965 年，Ivakhnenko 等人提出了多层感知机 [2]，能够解决线性不可分的问题

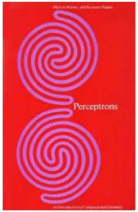


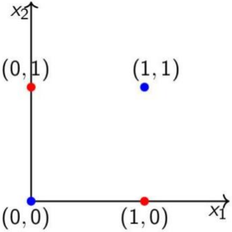


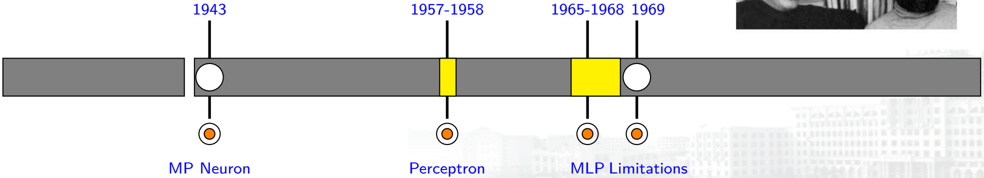
感知机的不足

虽然最初被认为有着良好的发展潜能，但感知机最终被证明不能处理诸多的模式识别问题

1969 年，马文·明斯基和西摩尔·派普特在《Perceptrons》书中，仔细分析了以感知机为代表的单层神经网络系统的功能及局限，证明感知机不能解决简单的异或(XOR)等线性不可分问题 [3]

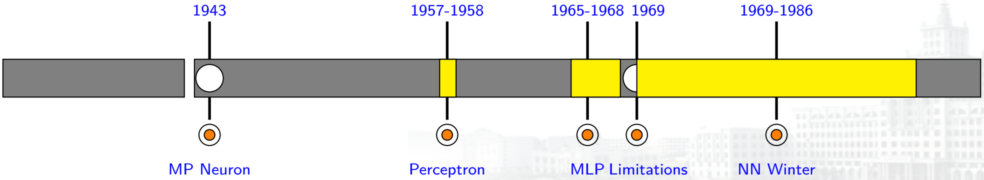






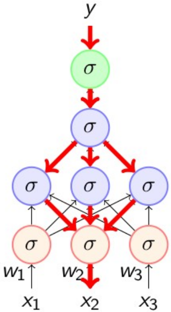
人工神经网络研究的低潮

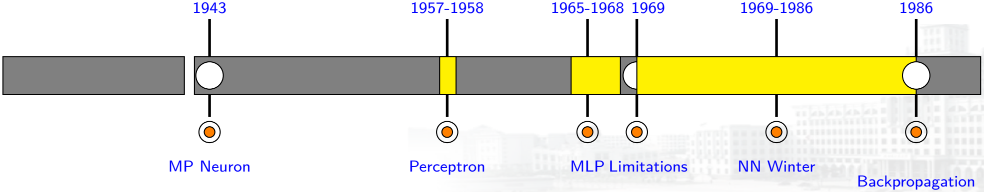
由于弗兰克·罗森布拉特等人没能够及时推广感知机学习算法到多层神经网络上，又由于《Perceptrons》在研究领域中的巨大影响，及人们对书中论点的误解，造成了人工神经领域发展的长年停滞及低潮



反向传播算法

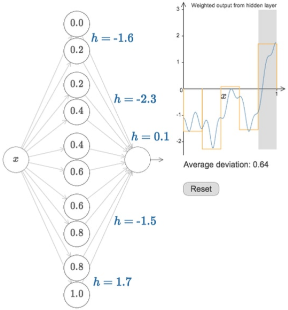
- 在 1960’s 到 1970’s 期间，发明了反向传播算法。
- 1982 年，Werbos [4] 首次将反向传播算法用于人工神经网络
- 1986 年，Rumelhart 等人的工作极大地让反向传播算法被大家认识 [5]

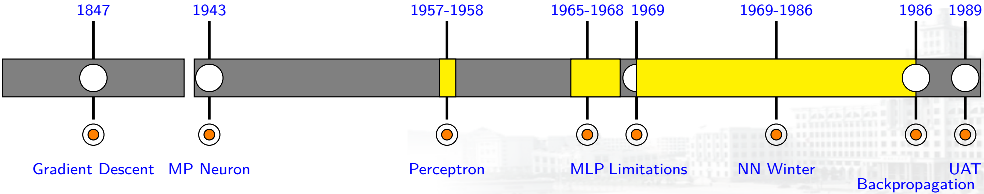




通用逼近定理（Universal Approximation Theorem）

一个包含单一隐含层的多层神经元网络能够以任何期望的精度近似任何连续函数[6]

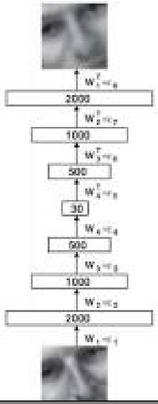




§3: The Deep Revival

无监督预训练（Unsupervised Pre-Training）

Hinton and Salakhutdinov 提出了一个有效的权重初始化方法，允许深度自编码网络学习数据的低维表示 [7]

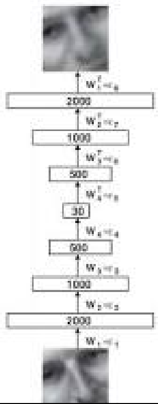


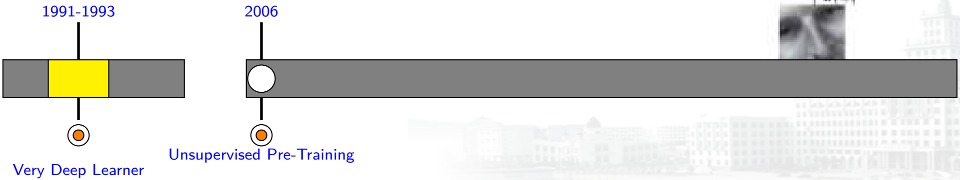
2006

Unsupervised Pre-Training

无监督预训练

无监督预训练的思想可以追溯到 1991-1993 年 J. Schmidhuber 使用这一思想来训练一 “Very Deep Learner”





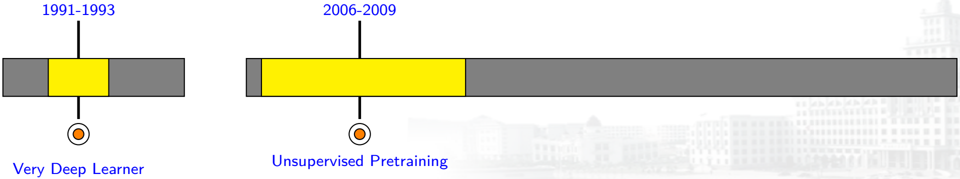
进一步的研究 (2007-2009)

对无监督预训练有效性的进一步的研究

Greedy Layer-Wise Training of Deep Networks

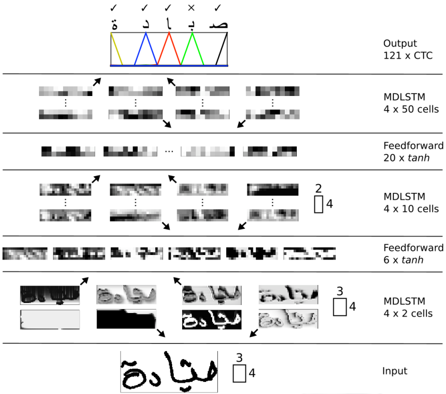
Why Does Unsupervised Pre-training Help Deep Learning?

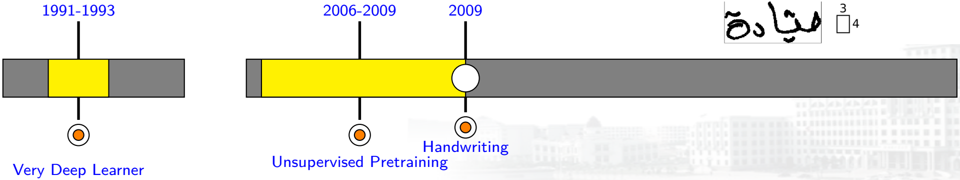
Exploring Strategies for Training Deep Neural Networks



成功应用于手写字识别

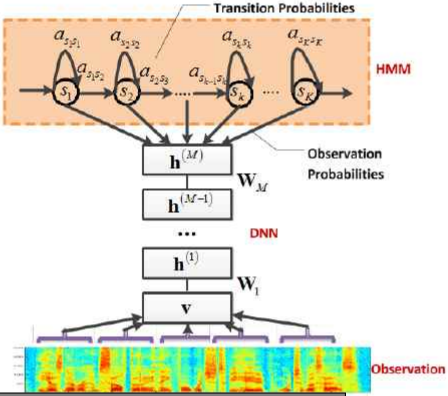
Graves et. al. outperformed all entries in an international Arabic handwriting recognition competition [8]

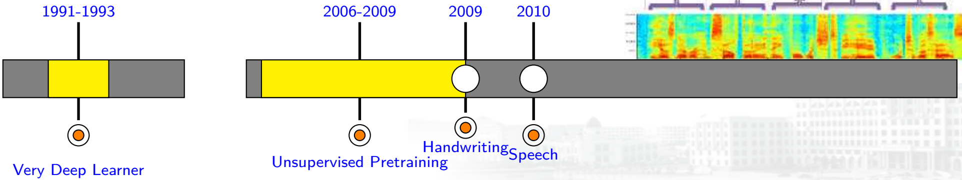




成功应用于语音识别

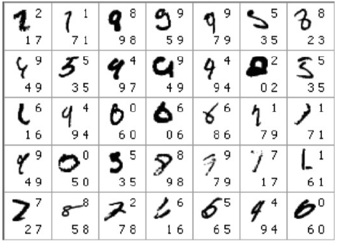
Dahl et. al. showed relative error reduction of 16.0% and 23.2% over a state of the art system [9]

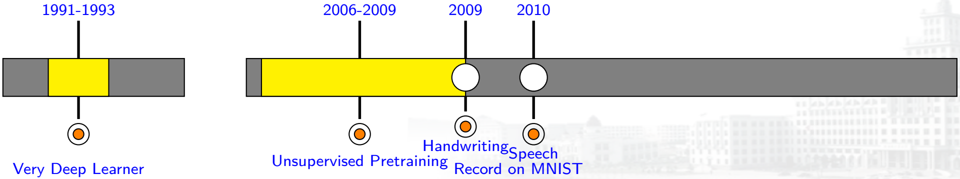




MNIST 上新的记录

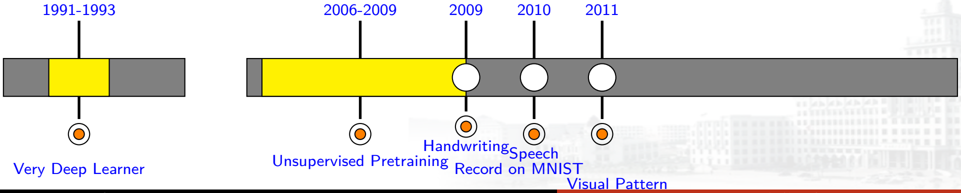
Ciresan et. al. set a new record on the MNIST dataset using good old backpropagation on GPUs (GPUs enter the scene)[10]





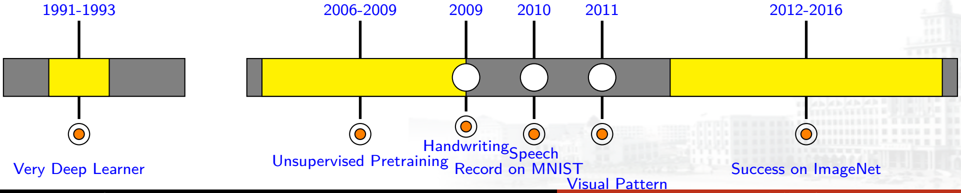
第一个超级视觉模式识别算法

D. C. Ciresan et. al. achieved 0.56% error rate in the IJCNN Traffic Sign Recognition Competition [11]



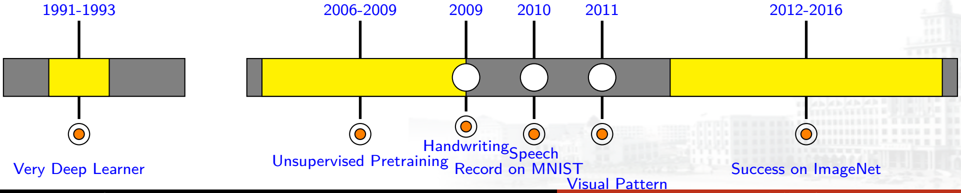
ImageNet 视觉识别挑战赛

Network	Error	Layers
AlexNet [12]	16.0%	8



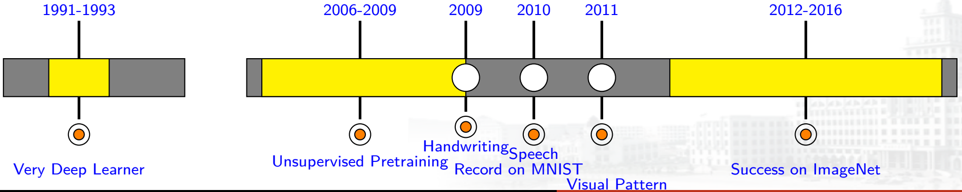
ImageNet 视觉识别挑战赛

Network	Error	Layers
AlexNet [12]	16.0%	8
ZFNet [13]	11.2%	8



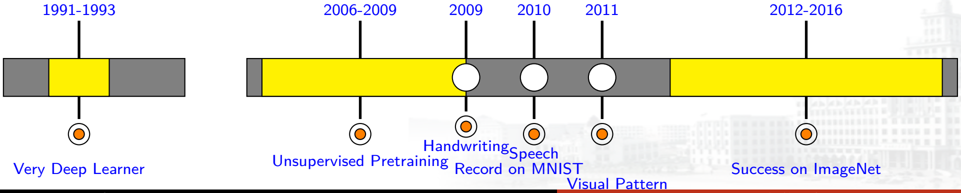
ImageNet 视觉识别挑战赛

Network	Error	Layers
AlexNet [12]	16.0%	8
ZFNet [13]	11.2%	8
VGGNet [14]	7.3%	19



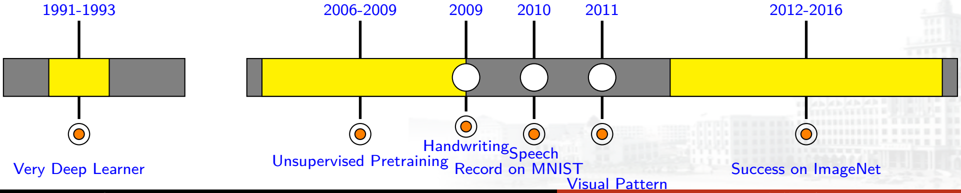
ImageNet 视觉识别挑战赛

Network	Error	Layers
AlexNet [12]	16.0%	8
ZFNet [13]	11.2%	8
VGGNet [14]	7.3%	19
GoogLeNet [15]	6.7%	22



ImageNet 视觉识别挑战赛

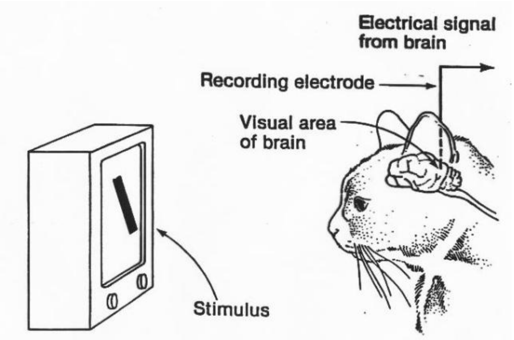
Network	Error	Layers
AlexNet [12]	16.0%	8
ZFNet [13]	11.2%	8
VGGNet [14]	7.3%	19
GoogLeNet [15]	6.7%	22
MS ResNet [16]	3.6%	152!!

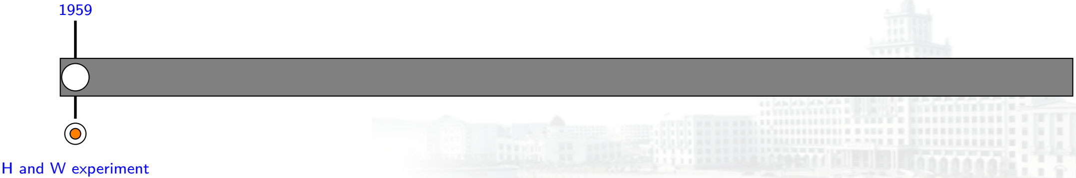


§4: From Cats to Convolutional Neural Networks

Hubel and Wiesel 实验

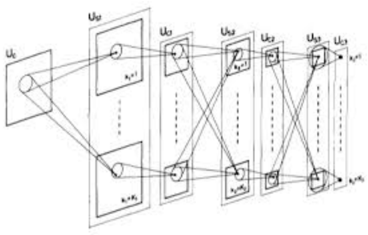
1959 年，Hubel and Wiesel 的实验表明每个神经元有一个固定的感受野 — 一个神经元只对一个特定区域内的视觉激励『开火』[17]





Neocognitron

1980 年,Fukushima 等人提出 Neocognitron 用于手写字符识别和模式识别 [18]



1959

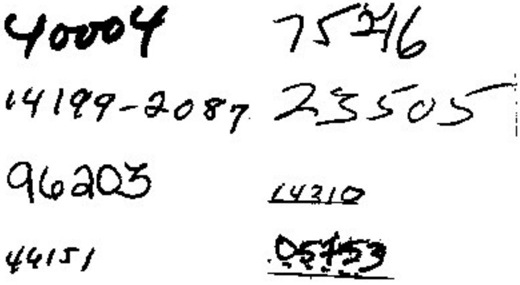
1980

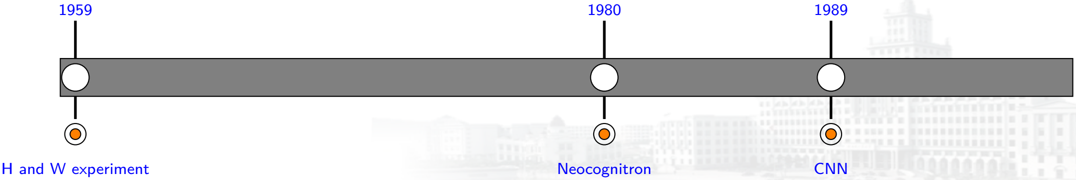
H and W experiment

Neocognitron

Convolutional Neural Network

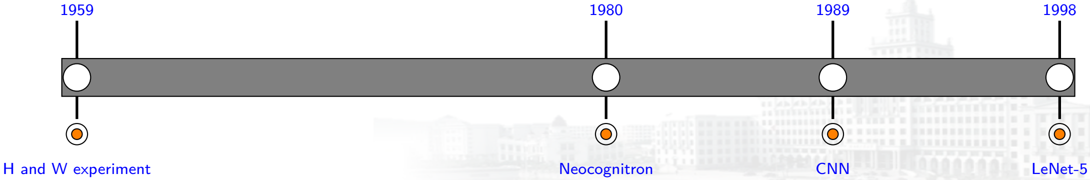
1989 年，LeCun 等人提出卷积神经网络用于手写数字识别 (LeCun et. al.) [19]





LeNet-5

1998 年，LeCun 等人提出 LeNet-5 模型用于 MNIST 数据集上的手写数字识别 [20]

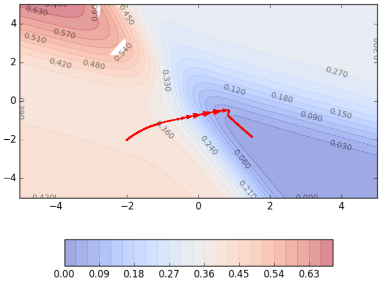


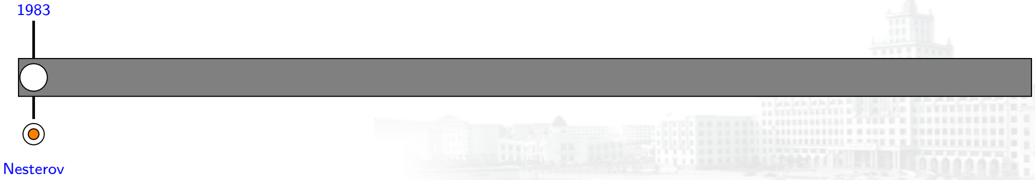
受猫实验启发得到算法正被用来检测视频中的猫:-)

§5: Faster, higher, stronger

更好的优化方法

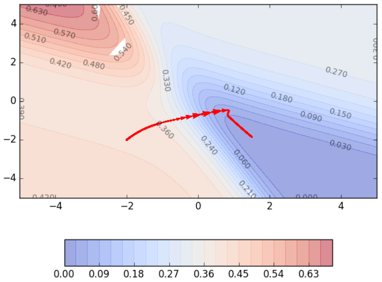
Faster convergence, better accuracies





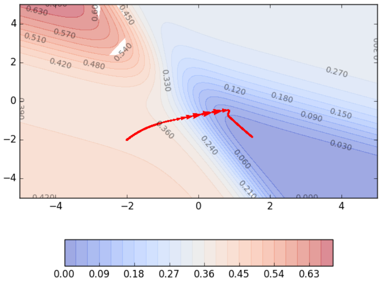
更好的优化方法

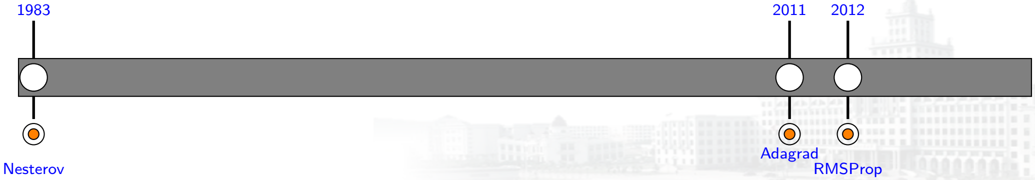
Faster convergence, better accuracies



更好的优化方法

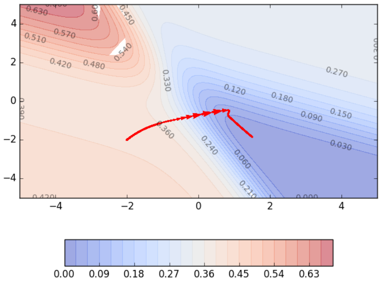
Faster convergence, better accuracies

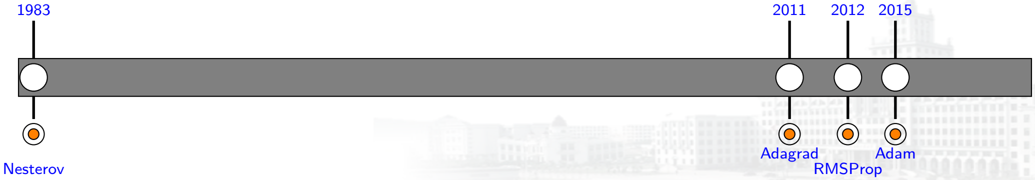




更好的优化方法

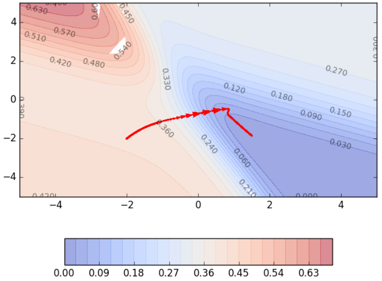
Faster convergence, better accuracies

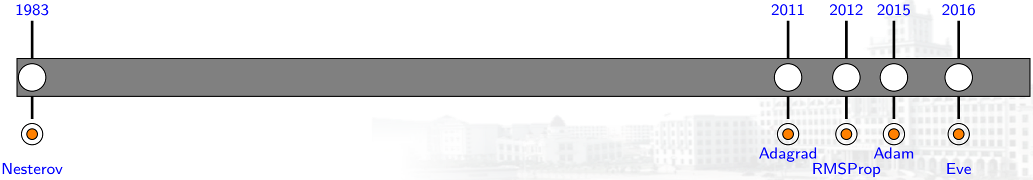




更好的优化方法

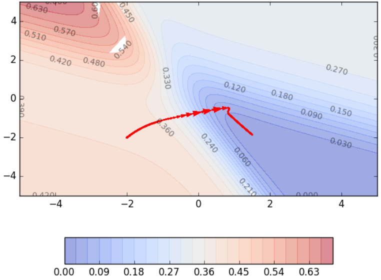
Faster convergence, better accuracies

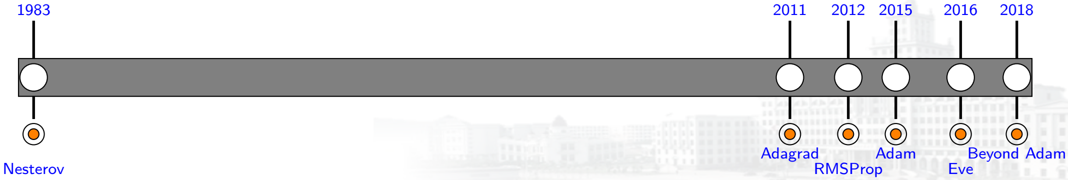




更好的优化方法

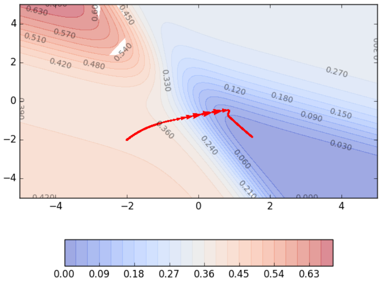
Faster convergence, better accuracies

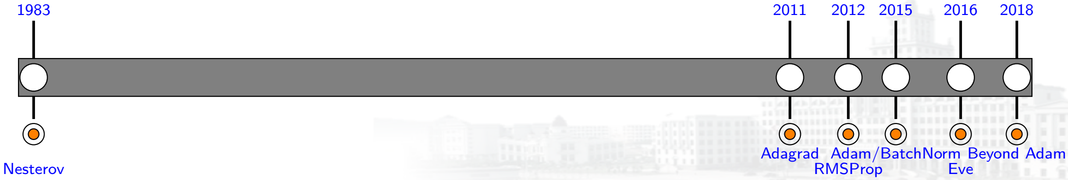




更好的优化方法

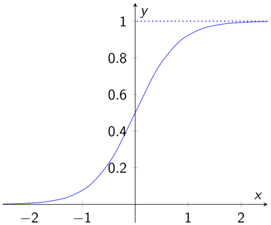
Faster convergence, better accuracies





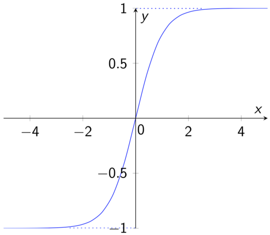
更好的激活函数

The **logistic 函数**是 80's 最常用的激活函数



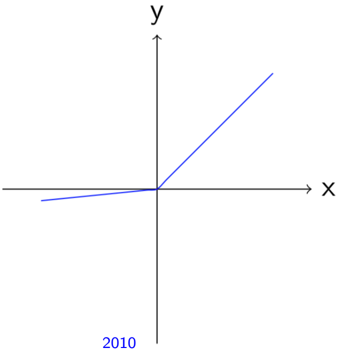
更好的激活函数

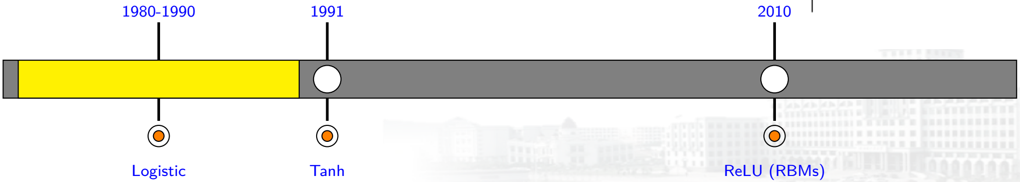
The **tanh 函数**是零中心化的，能够导致更好的收敛 [21]



更好的激活函数

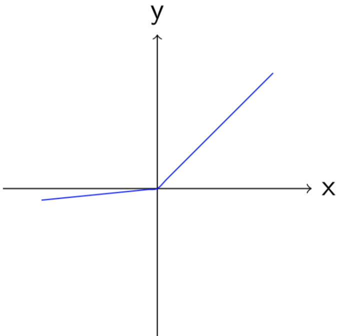
最近，**Rectified Linear Units (ReLUs)** 和它的变体得到了更好的性能 [22], [23],[24]





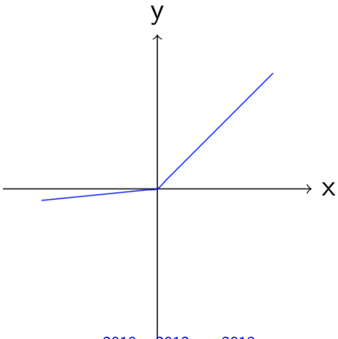
更好的激活函数

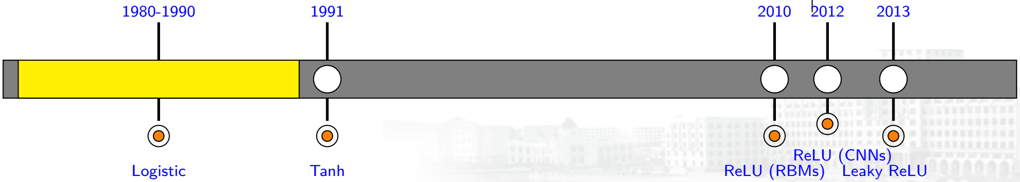
最近，**Rectified Linear Units (ReLUs)** 和它的变体得到了更好的性能 [22], [23],[24]



更好的激活函数

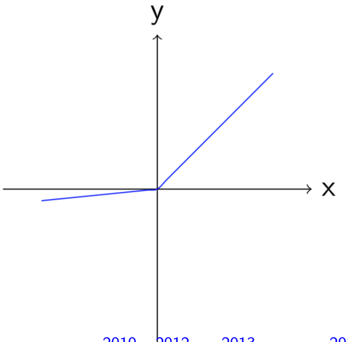
最近，**Rectified Linear Units (ReLUs)** 和它的变体得到了更好的性能 [22], [23],[24]





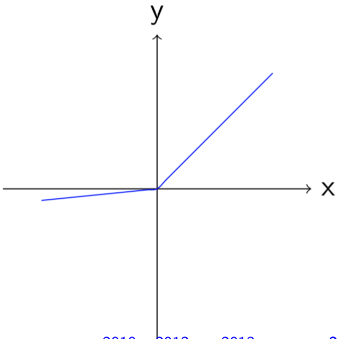
更好的激活函数

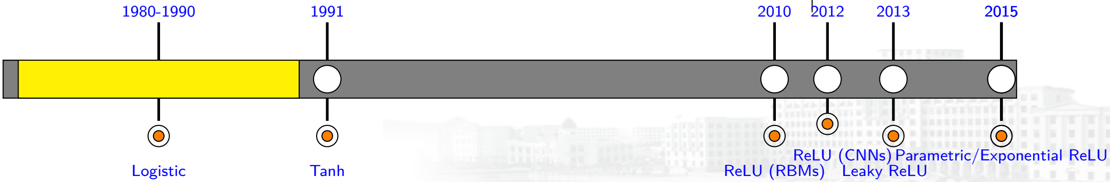
最近，**Rectified Linear Units (ReLUs)** 和它的变体得到了更好的性能 [22], [23],[24]



更好的激活函数

最近，**Rectified Linear Units (ReLUs)** 和它的变体得到了更好的性能 [22], [23],[24]





References I

[1] W.S.McCulloch and W.Pitts.
A logival calculus of the ideas imminent in nervous activity.
1943.

[2] A.G. Ivakhnenko and V.G. Lapa.
Cybernetic predicting devices.
1965.

[3] M.Minsky and S.Papert.
Perceptrons.
1969.

[4] P. J. Werbos.
Applications of advances in nonlinear sensitivity analysis.
In *Proceedings of the 10th IFIP Conference, 31.8 - 4.9, NYC*, pages 762–770, 1981.

[5] D. E. Rumelhart, G. E. Hinton, and R. J. Williams.
Learning internal representations by error propagation.
In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, pages 318–362. MIT Press, 1986.

[6] Kurt Hornik, Maxwell Stinchcombe, and Halbert White.
Multilayer feedforward networks are universal approximators.
Neural Networks, 2(5):359–366, 1989.

References II

[7] Geoffrey E. Hinton and Ruslan Salakhutdinov.
Reducing the dimensionality of data with neural networks.
Science, 313(5786):504–507, 2006.

[8] Alex Graves and Jürgen Schmidhuber.
Offline handwriting recognition with multidimensional recurrent neural networks.
In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 545–552. Curran Associates, Inc., 2009.

[9] G. E. Dahl, Dong Yu, Li Deng, and A. Acero.
Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition.
Trans. Audio, Speech and Lang. Proc., 20(1):30–42, January 2012.

[10] Dan Claudiu Ciresan, Ueli Meier, Luca Maria Gambardella, and Jürgen Schmidhuber.
Deep big simple neural nets excel on handwritten digit recognition.
CoRR, abs/1003.0358, 2010.

[11] Dan C. Ciresan, Ueli Meier, and Jürgen Schmidhuber.
Multi-column deep neural networks for image classification.
CoRR, abs/1202.2745, 2012.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton.
Imagenet classification with deep convolutional neural networks.
In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

References III

[13] Matthew D. Zeiler and Rob Fergus.
Visualizing and understanding convolutional networks.
CoRR, abs/1311.2901, 2013.

[14] Karen Simonyan and Andrew Zisserman.
Very deep convolutional networks for large-scale image recognition.
CoRR, abs/1409.1556, 2014.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich.
Going deeper with convolutions.
CoRR, abs/1409.4842, 2014.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
Deep residual learning for image recognition.
CoRR, abs/1512.03385, 2015.

[17] D. H. Wiesel and T. N. Hubel.
Receptive fields of single neurones in the cat's striate cortex.
J. Physiol., 148:574–591, 1959.

[18] K. Fukushima.
Neocognitron: A self-organizing neural network for a mechanism of pattern recognition unaffected by shift in position.
Biological Cybernetics, 36(4):193–202, 1980.

References IV

[19] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel.
Back-propagation applied to handwritten zip code recognition.
Neural Computation, 1(4):541–551, 1989.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner.
Gradient-based learning applied to document recognition.
Proceedings of the IEEE, 86(11):2278–2324, November 1998.

[21] Y. LeCun, I. Kanter, and S. A. Solla.
Second order properties of error surfaces: Learning time and generalization.
In D. S. Lippman, J. E. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 918–924. Morgan Kaufmann, 1991.

[22] Vinod Nair and Geoffrey E. Hinton.
Rectified linear units improve restricted Boltzmann machines.
In *International Conference on Machine Learning (ICML)*, 2010.

[23] Alex Krizhevsky, I Sutskever, and G. E Hinton.
Imagenet classification with deep convolutional neural networks.
In *Advances in Neural Information Processing Systems (NIPS 2012)*, page 4, 2012.

[24] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng.
Rectifier nonlinearities improve neural network acoustic models.
In *International Conference on Machine Learning (ICML)*, 2013.