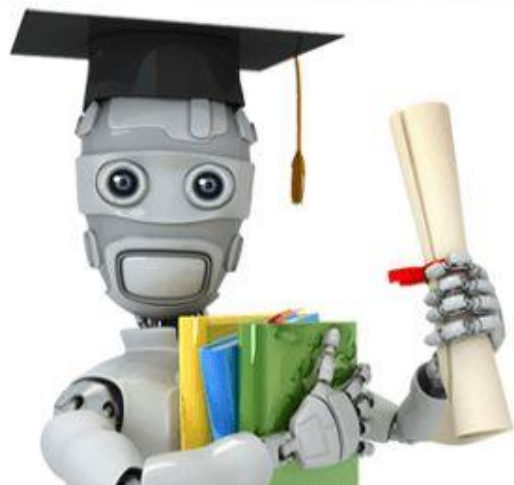




机器学习概述

Introduction to Machine Learning



杨 丹
2018年9月



1、什么是机器学习

2、人工智能发展简介

3、机器学习的分类





1、什么是机器学习

机器学习在身边 –阿尔法狗

阿尔法狗 (AlphaGo) 是第一个击败人类职业围棋选手第一个战胜围棋世界冠军的人工智能程序。由谷歌 (Google) 旗下DeepMind公司戴密斯·哈萨比斯领衔的团队开发。

2016年3月，阿尔法围棋与围棋世界冠军、职业九段棋手李世石进行围棋人机大战，以4比1的总比分获胜；

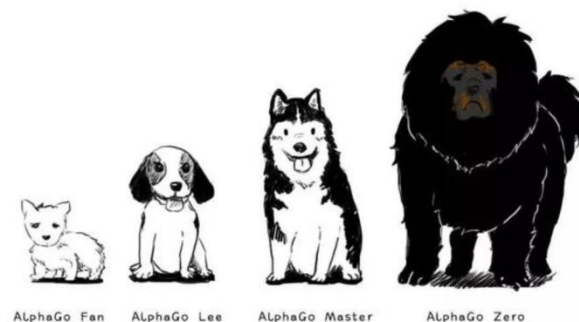
2016年末2017年初，该程序在中国棋类网站上以“大师” (Master) 为注册帐号与中日韩数十位围棋高手进行快棋对决，连续60局无一败绩。

2017年5月，在中国乌镇围棋峰会上，它与排名世界第一的世界围棋冠军柯洁对战，以3比0的总比分获胜。

围棋界公认阿尔法围棋的棋力已经超过人类职业围棋顶尖水平。

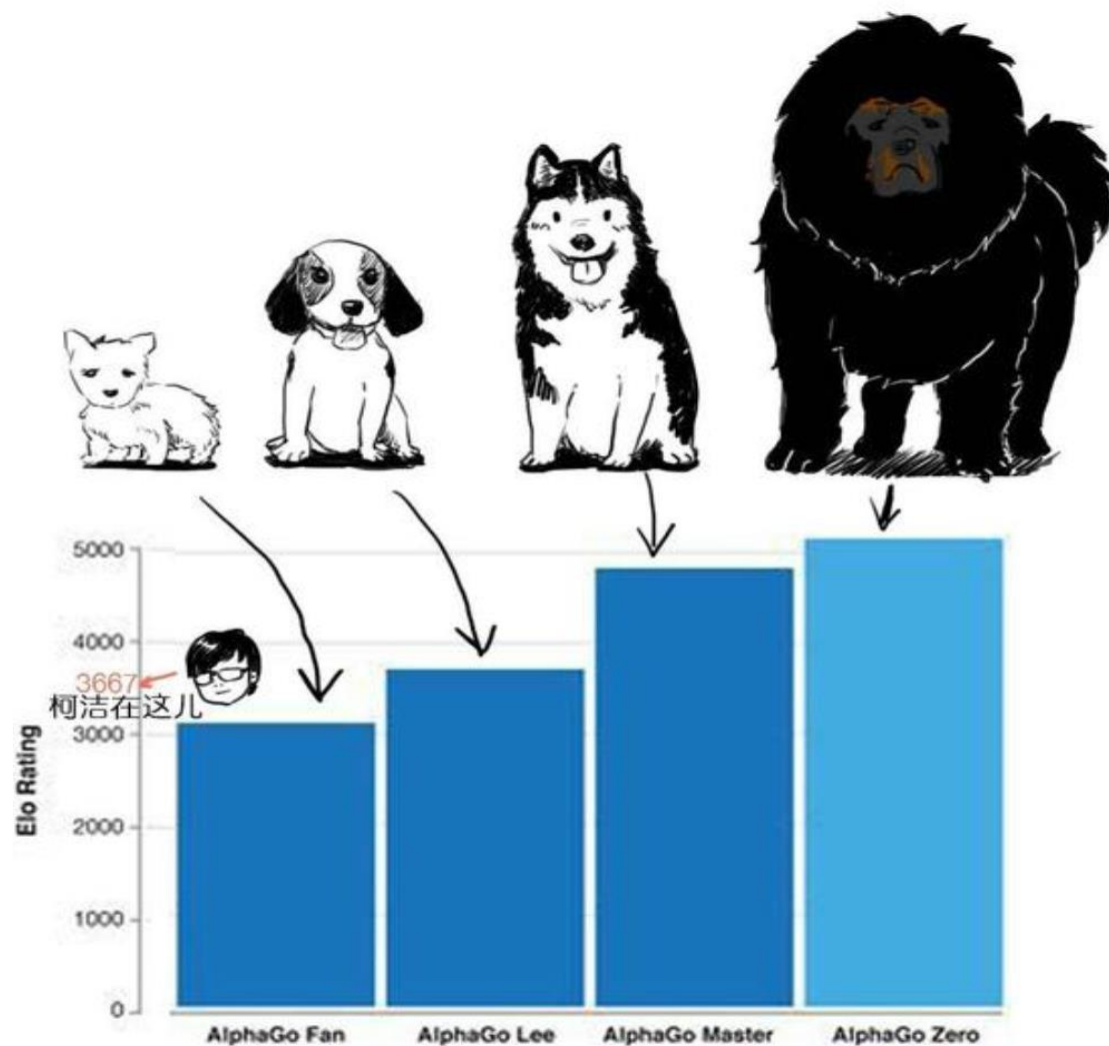
2017年5月27日，在柯洁与阿尔法围棋的人机大战之后，阿尔法围棋团队宣布阿尔法围棋将不再参加围棋比赛。

你可能已经听说了
AlphaGo家族又添新狗了



埃罗排名 (Elo Ratings) 用以计量个体在对决类比赛中相对技能的算法系统。

最早应用于国际象棋和围棋，目前已广泛用于足球等多人参与的对决性比赛。在网游WOW, DOTA中使用到了ELO算法。





计算机下棋史 -- 下棋一直就是人类智能的挑战，自然也成了人工智能的标志之一。

跳棋

- 1951年，图灵的朋友克里斯特拉切（Christopher Strachey）在曼彻斯特Mark-1上写了第一款跳棋程序。图灵在1952年曾与之对弈一局，轻松取胜。
- 1956年IBM的塞缪尔（Arthur Samuel，人工智能里程碑达特茅斯会议的参加者之一）写了第二个跳棋程序，这款程序的特点是自学习，这也是最早的机器学习程序之一，后来不断改进，曾经赢过盲人跳棋大师。
- 20世纪80年代末，最强的跳棋程序是加拿大阿尔伯塔大学的Chinook。数学家丁斯利（Marion Tinsley）自50年代起就一直是跳棋的人类冠军。1992年丁斯利大战Chinook并取胜，1994年再战，但丁斯利不久病逝。目前最好棋手的最好结局是打成平手。

国际象棋

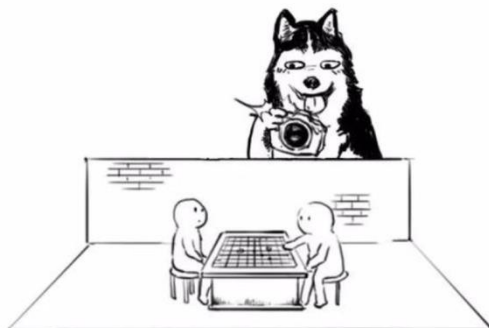
- 第一个可以走全局的下棋程序是IBM的工程师伯恩斯坦（Alex Bernstein）1958年在一台IBM 704上做的。
- 1970年开始，美国计算机学会（ACM）的年会都在晚餐时举行计算机象棋比赛，作为娱乐节目。
- 1996年，IBM开发“深蓝”（Deep Blue）机器棋手，对阵当时的世界冠军俄罗斯特级大师卡斯帕罗夫，卡4:2赢得比赛。但1997年5月11日，卡认输，“深蓝”成了第一位战胜当时世界冠军的机器。

中国象棋

- 至2006年，中国象棋程序开始击败特级大师级别的人类棋手。

...because chess requires intelligence.（下棋需要智能。）——Alan Turing（图灵）

第一步，
先让零基础的AlphaGo观摩海量的
人类棋谱



真正到了实战环节
在狗每一次举棋不定的时候
它身边会有两位导师帮它拿主意

老湿，
下这儿行不行？



2

但是只有人类的棋谱
样本量还是远远不够的
工程师于是就决定让AlphaGo
自己跟自己下，
又创造出海量的棋谱

4

其中一位就是

导师1号

策略网络

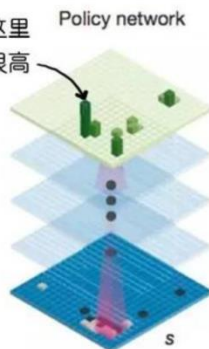
(Policy network)



在学习了大量棋谱的套路之后
AlphaGo形成了超强的**预测能力**

5

棋子落在这里
的概率会很高



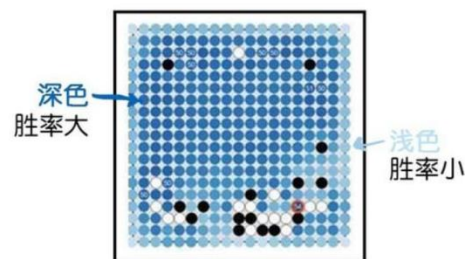
这时它发现了几个
出现概率较高、看上去很有潜力的点
它们就需要第二位导师的重点关注了

6



导师2号
价值网络
(Value network)

他解决的问题是
“怎么下能赢？”



7

整个过程还需要用到一种叫作
蒙特卡洛树搜索(MCTS)
的神器



然后经过导师们的商议



8

这只狗只需了解围棋最基本的获胜规则

剩下的就是让两只狗不停地对弈，
随着局数的积累，逐渐总结经验

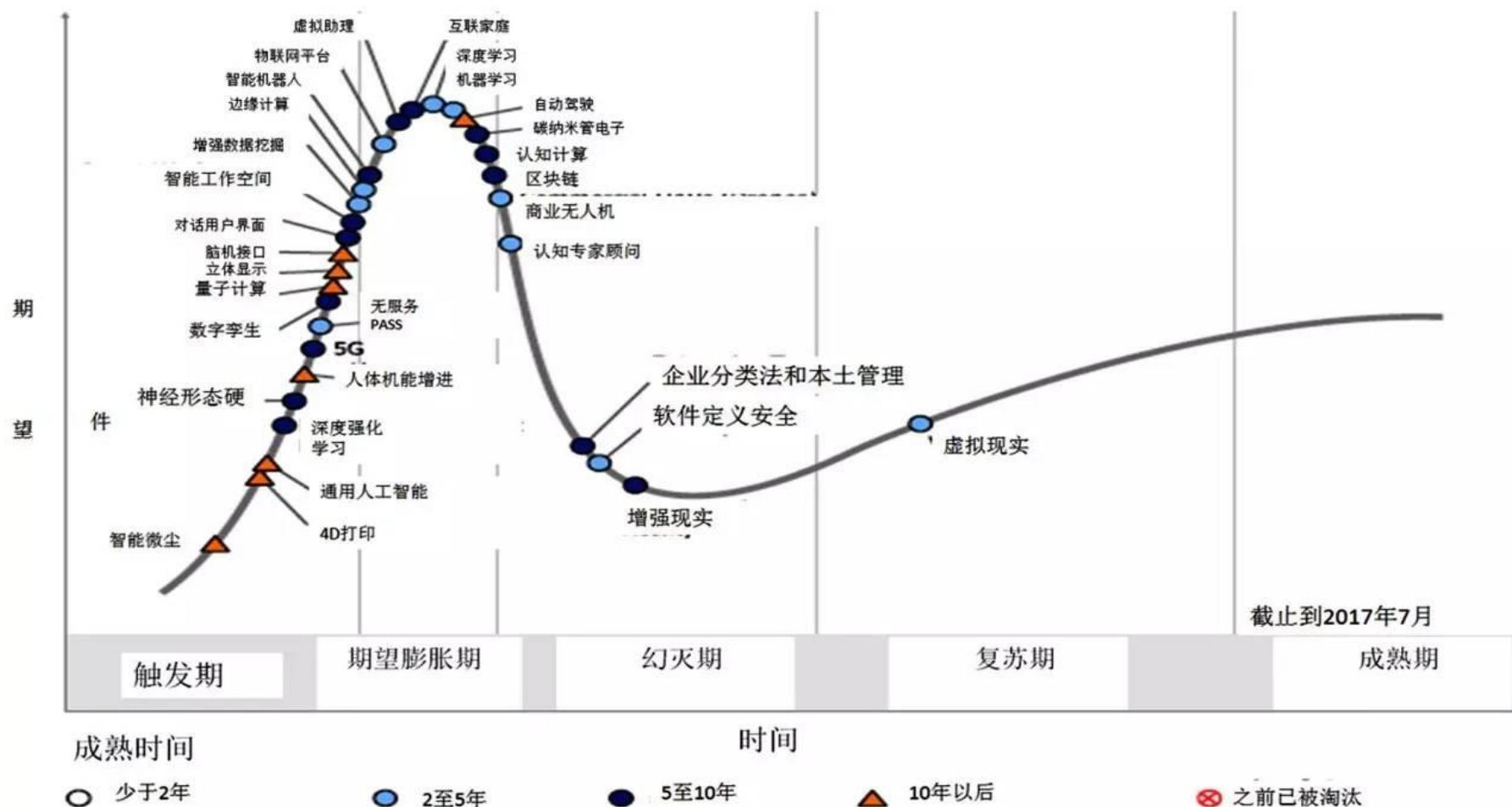


这种学习方式就是“无监督学习”

阿尔法狗
无监督学习

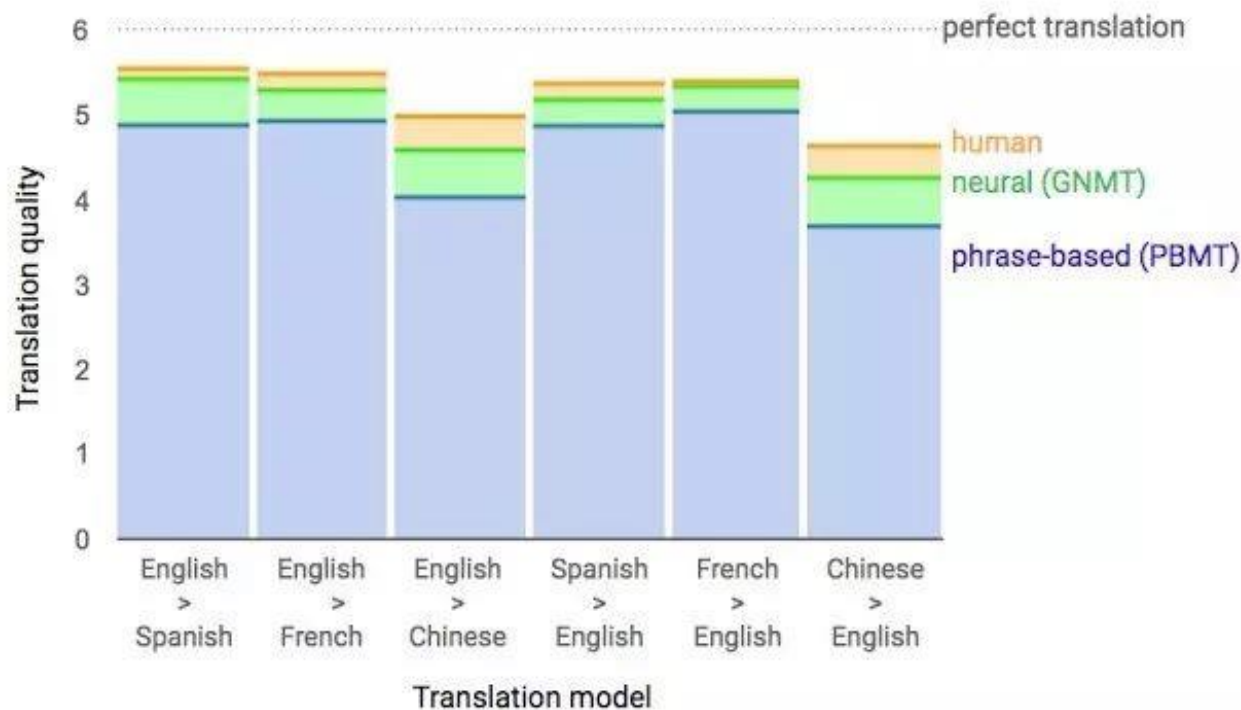
2017新兴技术成熟度曲线（来源：Gartner 2017年7月）

Gartner认为，2017年技术成熟度曲线揭示了未来5-10年的三方面技术趋势，一是无处不在的人工智能、二是身临其境的体验、三是数字化平台。



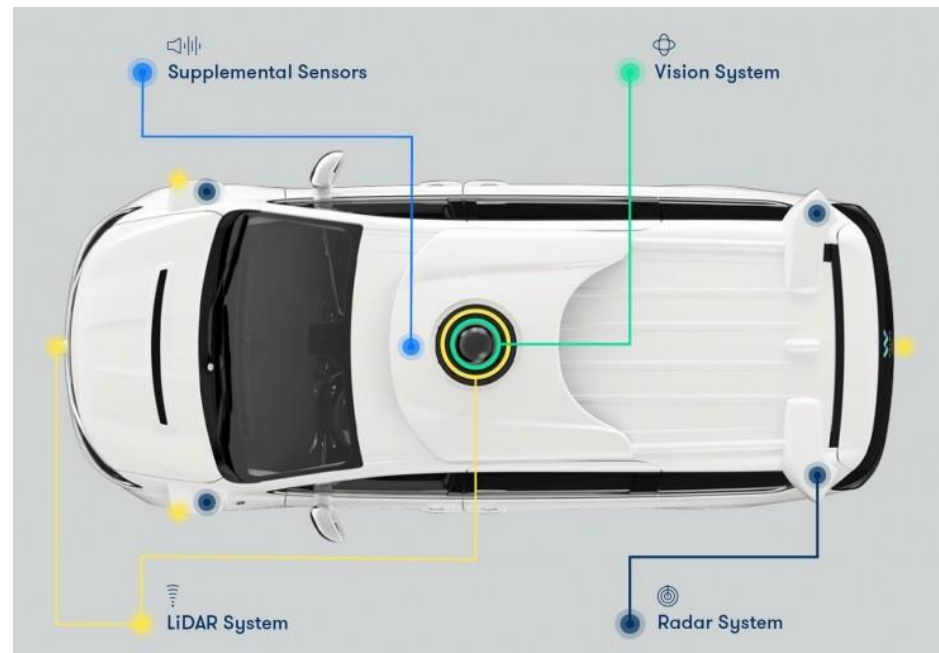
标志性进展

- **谷歌神经机器翻译**：关键结果，与人类翻译准确率的差距缩小了 55-85%（研究者使用 6 个语言对的评估结果）。但是该模型如果没有谷歌的大型数据集，则很难复现这么优秀的结果。



谷歌 自动驾驶系统Waymo

过去 8 年，Waymo 在美国 4 个州进行了测试，在 20 多个城市里实现了自动驾驶——从晴朗的凤凰城，一直到多雨的柯克兰，整个过程中积累了超过 350 万英里的自动驾驶数据。



Level 4级别可在任何系统故障的情况下，给车辆安全制停的能力，而无需人类驾驶员接管。

2017年11月17日，特斯拉在美国正式发布了Tesla Semi。

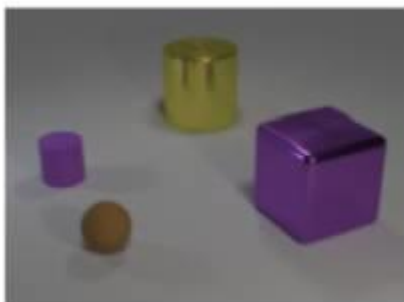
2017年8月，特斯拉为卡车申请自动驾驶路测。特斯拉负责人在写给内华达州机动车管理局官员的一封信中指出：路测首要目标，是在内华达和加利福尼亚州连续对原型卡车进行测试，使它们在无人干预的状态下以编队或自动驾驶方式行驶。



视觉推理

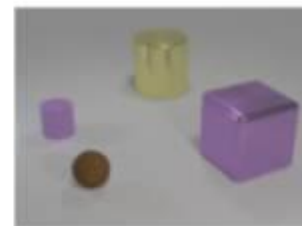
视觉推理指的是让神经网络回答根据照片提出的问题。例如，「照片中有和黄色的金属圆柱的尺寸相同的橡胶物体吗？」

Original Image:



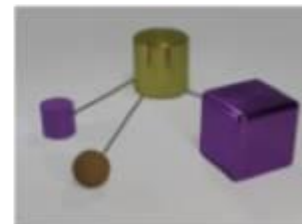
Non-relational question:

What is the size of the brown sphere?



Relational question:

Are there any rubber things that have the same size as the yellow metallic cylinder?



目前的机器学习系统在 CLEVR 上标准问题架构上的回答成功率为 68.5%，而人类的准确率为 92.5%。但是使用了 RN 增强的神经网络，**DeepMind 展示了超越人类表现的 95.5% 的准确率。**



AI在身边 – 例子

- 智能助理：苹果公司的IOS语音助理Siri；
- 新闻推荐和新闻撰稿:2014年，Automated Insights的人功能智能程序已撰写出超过10亿篇的新闻稿
- 机器视觉:广义上的机器视觉既包括人脸识别，也包括图像、视频中的各种物体识别、场景识别、地点识别乃至语义理解。
- AI艺术: “2016年夏天，一款名为Prisma的手机绘画程序在大家的朋友圈里流行开来。Prisma并不是用程序凭空作画，而是根据用户指定的一张照片，将照片变成特定风格的画作。”
- 新一代搜索引擎:谷歌搜索已经是一个绝大部分由人工智能技术支撑的新一代搜索引擎了。
- 机器翻译:2016年9月，谷歌宣布已经在谷歌翻译的中译英的模型中应用了深度学习的一种最新算法，并大幅提高了中文到英文的翻译准确率。
- 机器人:根据2015年第三季度末的统计数据，亚马逊在13个仓储中心使用着超过3万个Kiva机器人。.....根据2015年第三季度末的统计数据，亚马逊在13个仓储中心使用着超过3万个Kiva机器人。.....使用了橙黄色机器人的仓储中心比普通仓储中心可以多存放50%的货物，运营成本也由此降低了20%



Tom M. Mitchell (1997) 提供了一个简洁的定义：对于某类任务T 和性能度量P，一个计算机程序被认为可以从经验E 中学习是指，通过经验E 改进后，它在任务T 上由性能度量P 衡量的性能有所提升。

Machine learning is a category of research and algorithms focused on finding patterns in data and using those patterns to make predictions. Machine learning falls within the artificial intelligence (AI) umbrella, which in turn intersects with the broader field of knowledge discovery and data mining.



机器学习领域是由Arthur Samuel在1959年创造的，他说：“研究的领域使计算机能够在不被明确编程的情况下学习”。

周志华：机器学习所研究的主要内容，是关于在计算机上从数据中产生“模型”的算法，即“学习算法”，因此可以说机器学习史研究关于“学习算法”的学问。



算法 (Algorithm) 是一系列求解问题的清晰指令，它对一定规范的输入，在有限时间内获得所要求的输出。

通常算法：数据输入计算机，算法会利用数据完成接下来的事，然后结果就出来了。

机器学习：输入数据和想要的结果，输出的则是算法。即把数据转换为算法。

语音识别：从一个用户的话语，确定用户提出的具体要求。这样的模型，可以帮助程序能够并尝试自动填充用户需求。带有Siri系统的iPhone就有这种功能。



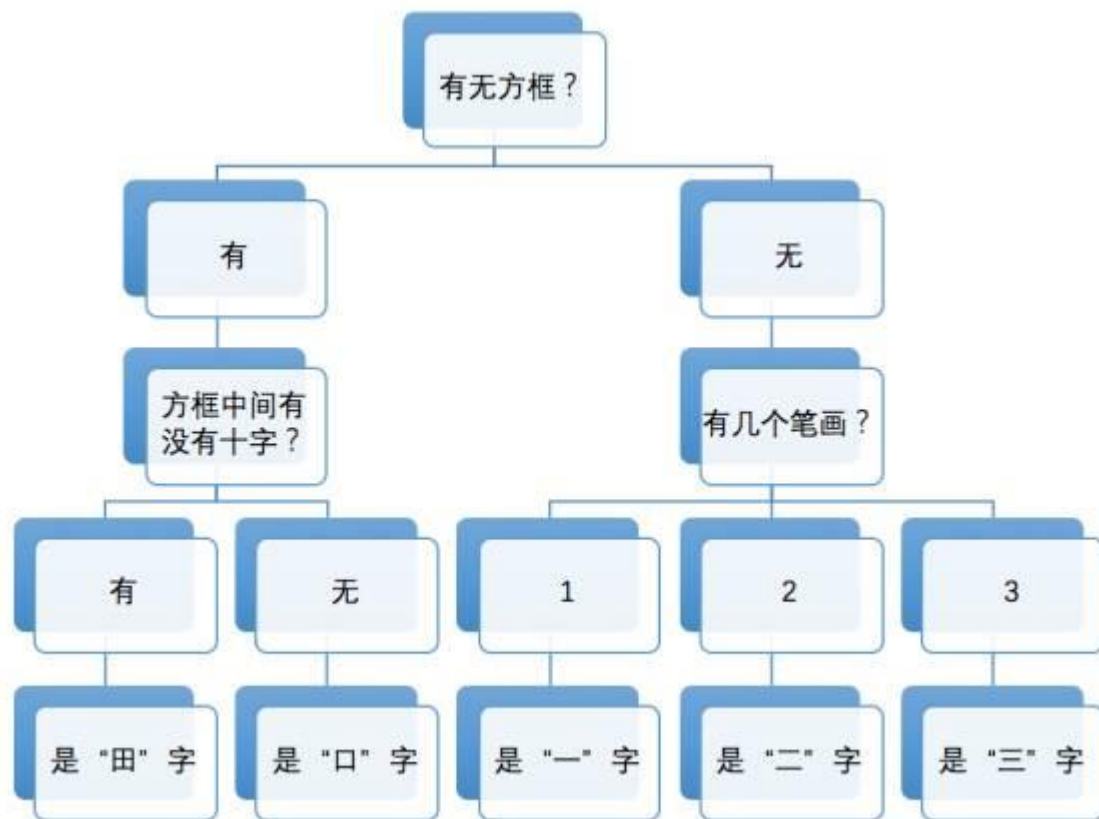


卡片识字：人类小朋友是如何学习的？

从简单到复杂的顺序，小朋友反复看每个汉字的各种写法，看得多了，自然就记住了。下次再见到同一个字，就很容易能认出来。

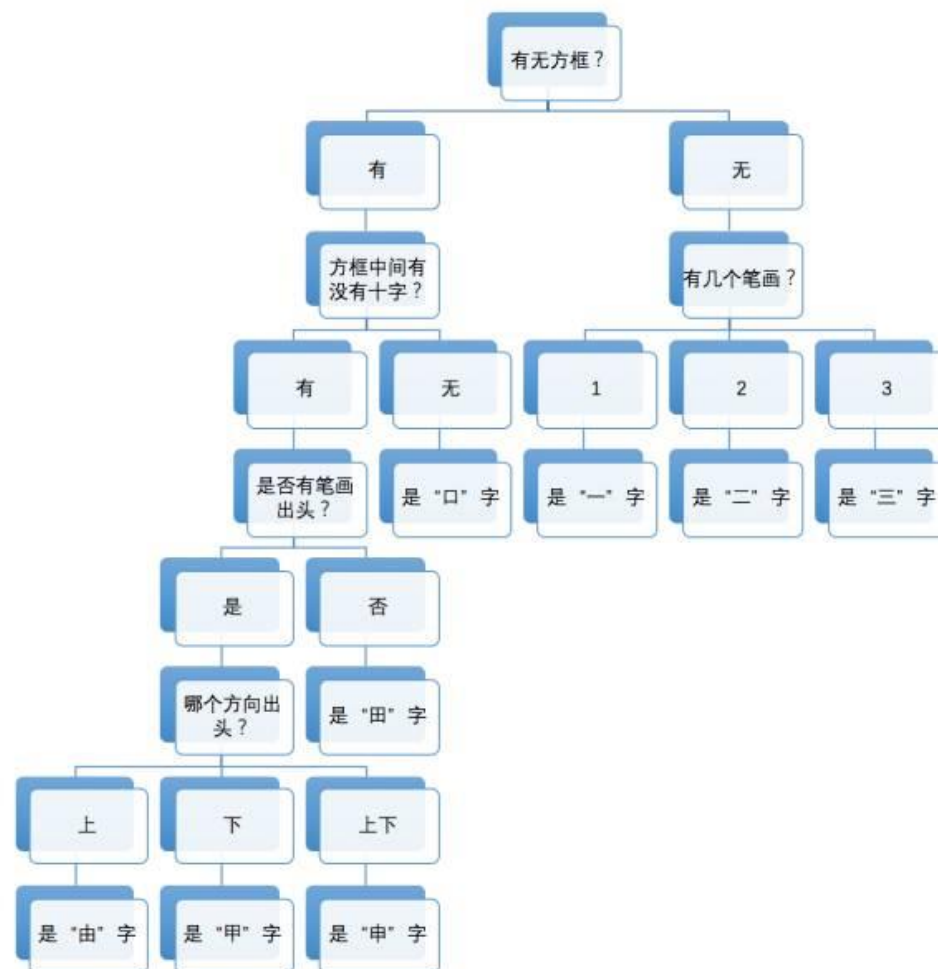
机器学习

计算机看图片（“训练数据集”）；“训练数据集”中，一类数据区别于另一类数据的不同方面的属性或特质，叫做“特征”；计算机在“大脑”中总结规律的过程，叫“建模”；计算机在“大脑”中总结出的规律，就是我们常说的“模型”；而计算机通过反复看图，总结出规律，然后学会认字的过程，就叫“机器学习”。

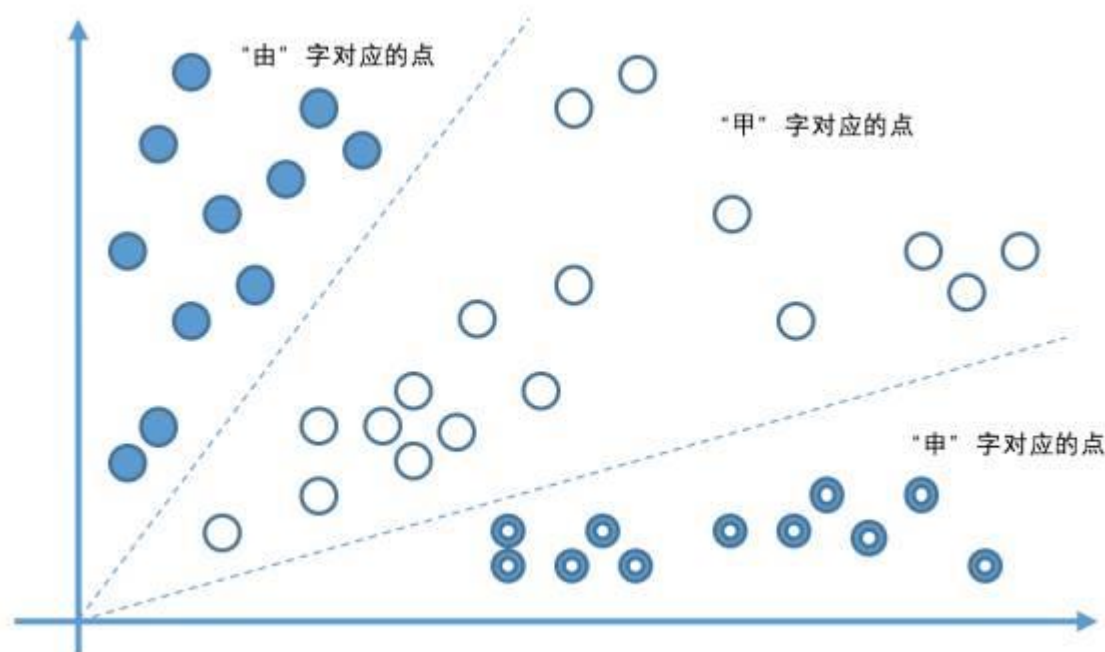


机器学习（决策树）：分辨
“一”、“二”、“三”
、“田”、“口”

计算机学习“由”、“甲”、“申”这三个新汉字前后，计算机内部的决策树的不同。



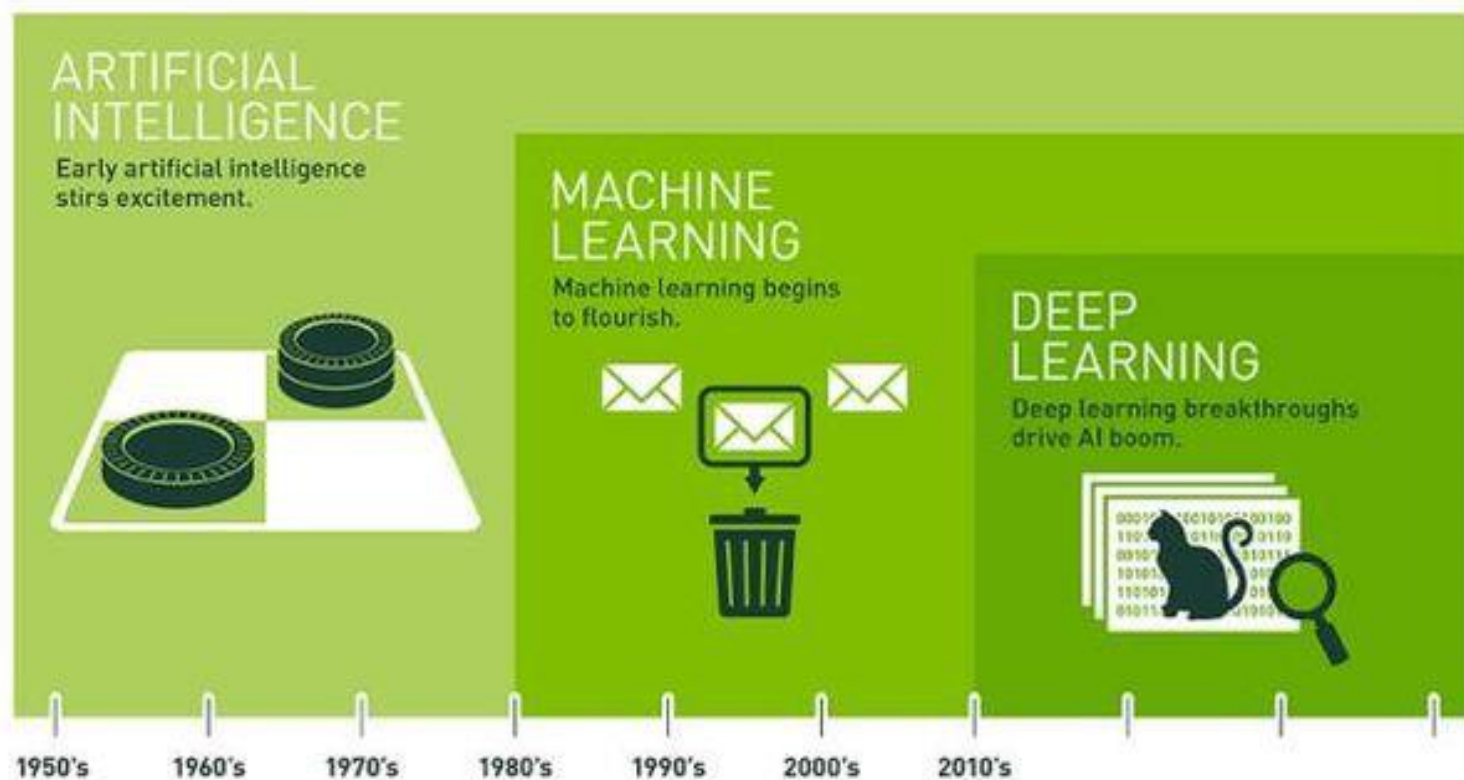
把空间分割成几个相互独立的区域，尽量使得训练数据集中每个字对应的点都位于同一个区域内。如果这种分割是可行的，就说明计算机“学”到了这些字在空间中的分布规律，为这些字建立了模型。



训练数据集中，这三个字的大量不同写法，在计算机看来就变成了空间中的一大堆点。



2、人工智能发展简介

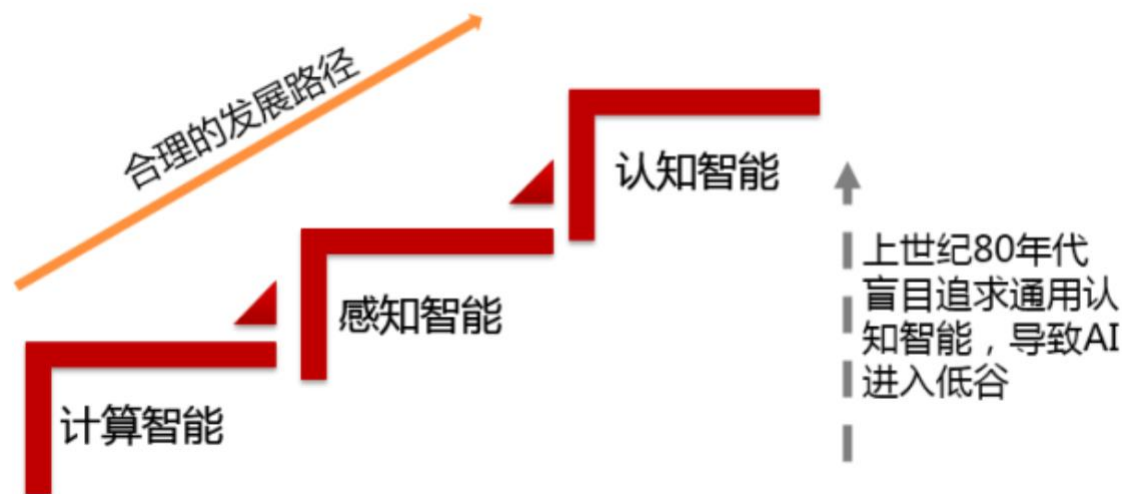


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.



人工智能 (Artificial Intelligence) 是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的技术科学。它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器，该领域的研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等。

人工智能可以对人的意识、思维的信息过程的模拟。人工智能不是人的智能，但能像人那样思考、也可能超过人的智能。



图灵 (Alan Turing)

计算机之父、人工智能之父

- 提出计算机的概念：1950年制出了“自动计算机”ACE样机,1958年制成大型ACE机。
- 把可计算函数定义为图灵机可计算函数.：1937年,图灵在他的“可计算性与 λ 可定义性”一文中证明了图灵机可计算函数与 λ 可定义函数是等价的
- 开创“自动机”这一学科分支,促进了电子计算机的研制工作
- 提出了通用图灵机的概念：开启了后来计算机科学中计算复杂性理论的先河
- 解决了著名的希尔伯特判定问题：狭谓词演算公式的可满足性的判定问题，成为后来人们证明一阶逻辑的公式类的不可判定性的主要方法之一。
- 图灵测试：1946年,图灵发表论文阐述存储程序计算机的设计。
- 人工智能：1950年他发表论文《计算机器与智能》，为后来的人工智能科学提供了开创性的构思。



1950年，阿兰·图灵创造了图灵测试来判定计算机是否智能。如果一台机器能够与人类展开对话（通过电传设备）而不能被辨别出其机器身份，那么称这台机器具有智能。



2014年6月8日，一台计算机（计算机尤金·古斯特曼是一个聊天机器人，一个电脑程序）成功让人类相信它是一个13岁的男孩，成为有史以来首台通过图灵测试的计算机。这被认为是人工智能发展的一个里程碑事件。

图灵奖—计算机领域的诺贝尔奖

- ◆ 美国计算机协会(ACM)于1966年设立图灵奖纪念这位科学家。
- ◆ 共计有67名科学家获此殊荣,有8位科学家是做人工智能的，Minsky Marvin（连接主义）、McCarthy、Newll、Simon、Feigenbaum（符号主义）、Reddy（语音识别）、Valiant（机器学习理论）、Pearl（概率计算和因果推理）

达特茅斯会议与人工智能的缘起

- 麦卡锡 (John McCarthy) :达特茅斯学院的数学系助理教授
- 明斯基 (Marvin L Minsky) :非线性规划和博弈论，获1969年图灵奖
- 塞弗里奇 (Oliver Selfridge) :模式识别奠基人，写出第一个可工作的AI程序
- 克劳德·香农 (Claude Shannon) : 信息论的创始人
- 纽厄尔(Allen Newell) : 获1975年度的图灵奖
- 司马贺 (赫伯特·西蒙 (Herbert A. Simon) : 卡内基理工学院 (卡内基梅隆大学的前身) 1975年获图灵奖，1978年获得诺贝尔经济学奖
- 所罗门诺夫 : MIT教授，发明了“归纳推理机”
-

1956年暑期的达特茅斯会议 (人工智能夏季研讨会 , Summer Research Project on Artificial Intelligence) , 10人 , 历时两个月

人工智能：使一部机器的反应方式 就像一个人在行动时所依据的智能



2006年，会议五十年后，当事人重聚达特茅斯。左起：摩尔，麦卡锡，明斯基，赛弗里奇，所罗门诺夫



人工智能第一次浪潮（1956-1976）

主要是符号主义、推理、专家系统等领域发展很快

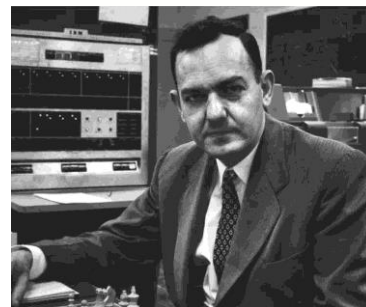
人工智能中符号派的思想源头和理论基础就是定理证明

1958年，西蒙和纽维尔预言10年内

- 计算机将成为国际象棋冠军
- 计算机将发现和证明有意义的数学定理
- 计算机将能谱写优美的乐曲
- 计算机将能实现大多数的心理学定理

第一次人工智能浪潮——推理

- 20世纪50年代，西蒙、纽厄尔、约翰·肖开发了世界上最早的启发式程序“逻辑理论家”(Logic Theorist)。逻辑理论家证明了数学名著《数学原理》一书第二章52个定理中的38个定理，被认为是图灵关于机器可以具有智能这一论断的第一个实际的证明。开创了机器定理证明(mechanical theorem proving)这一新的学科领域。
- 1972年，正式宣布创立Prolog编程语言，在早期的机器智能研究领域，Prolog曾经是主要的开发工具。
- 华人数理逻辑学家王浩1958年夏天在IBM704实现一个完全的命题逻辑程序和一阶逻辑程序，9分钟证明了《数学原理》中一阶逻辑的全部150个定理中的120个。王浩毕业于西南联大数学系，曾和杨振宁同屋。

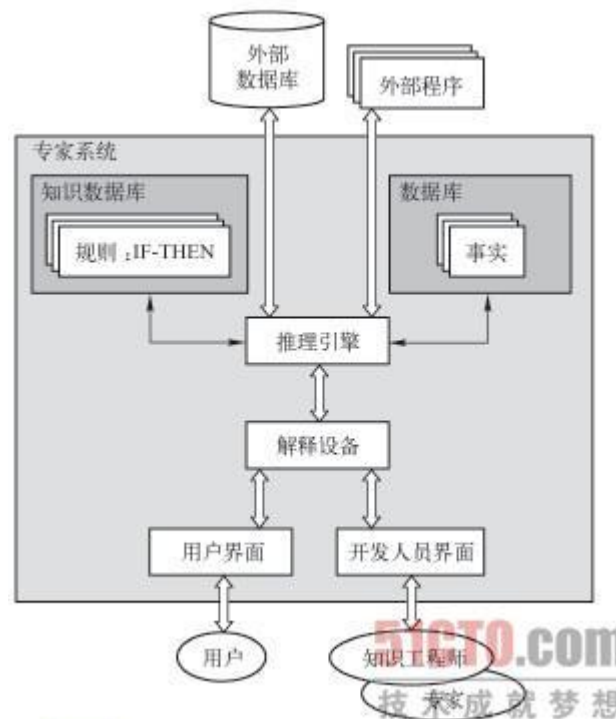
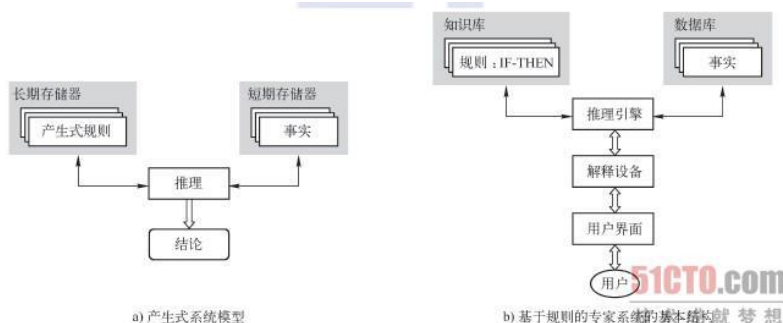


西蒙：获图灵奖、诺贝尔经济学奖



吴文俊：2000年度国家最高科学技术奖

产生式系统模型和基于规则的专家系统的基本结构



基于规则的专家系统由5个部分组成：知识库、数据库、推理引擎、解释设备和用户界面。

专家系统时代最成功的案例是DEC的专家配置系统XCON。1980年，卡内基梅隆大学为数字设备公司设计了一套名为XCON的“专家系统”。



1961年，明斯基写道：“在我们的有生之年，机器可能会在一般智力上超越我们”。

- 第一次人工智能浪潮催生的子领域：基于规则的系统、机器学习、单层和多层感知器网络、自然语言处理(NLP)、语音识别和语音处理、图像处理和计算机视觉、聊天机器人、机器人

萧条期和人工智能冬季 到1975年，人工智能项目基本上限于解决基本问题。

- 1976年，世界上最快的超级计算机(它的成本超过了500万美元)只能每秒执行大约1亿次指令。相对人类有大约860亿个神经元和1万亿突触，创建一个这种规模的感知器网络将花费1.6万亿美元，在1974年消耗整个美国的GDP。
- 1969年，Minsky和Papert出版了《感知器》，不能解决XOR（异或）问题
- 20世纪50年代的大肆宣传使人们对这种大胆的高度产生了预期，当1973年的结果没有实现时，美国 and 英国政府撤回了对AI的研究经费。
- 日本政府在1980年暂时提供了额外的资金，1981年10月，开始研制第五代计算机。但在1980年代后期，它很快就幻灭了，并再次收回投资。



人工智能第二次浪潮（1976-2006）-- 连接主义

80年代Hopfield神经网络和误差反向传播算法（BP算法）的提出，使得人工智能再次兴起，出现了语音识别、语音翻译计划，以及日本提出的第五代计算机。

- 1982年，霍普菲尔德（Hopfield）提出了一种新的神经网络，可以解决一大类模式识别问题，还可以给出一类组合优化问题的近似解。这种神经网络模型后被称为Hopfield神经网络。
- 1982年，David Parker重新发现了BP神经网络学习算法（1974年，Paul Werbos就发明了影响深远的著名BP神经网络学习算法。但没有引起重视）。
- 1983年，Hinton, G. E. 和 Sejnowski, T. J. 设计了玻尔兹曼机，首次提出了“隐单元”的概念。在全连接的反馈神经网络中，包含了可见层和一个隐层，这就是玻尔兹曼机。
- 1986年，David E. Rumelhart, Geoffrey E. Hinton 和 Ronald J. Williams发表文章《Learning representations by back-propagating errors》，重新报道这一方法，BP神经网络学习算法才受到重视。BP算法引入了可微分非线性神经元或者sigmoid函数神经元，克服了早期神经元的弱点，为多层神经网络的学习训练与实现提供了一种切实可行的解决途径。

第二次人工智能浪潮——知识

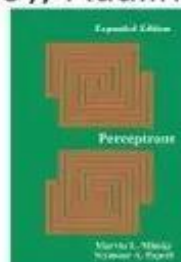
浅层学习模型包括：传统隐马尔可夫模型（HMM）、条件随机场（CRFs）、最大熵模型（MaxEnt）、boosting、支持向量机（SVM）、核回归及仅含单隐层的多层感知器（MLP）等。

神经网络发展回顾

1940年代-萌芽期：M-P模型（1943），Hebb 学习规则（1945）

1958左右-1969左右~繁荣期：感知机（1958），Adaline（1960），...

1969年：Minsky & Papert "Perceptrons"



冰河期

1985左右 -1995左右~繁荣期：Hopfield（1983），BP（1986），...

1995年左右：SVM 及 统计学习 兴起

沉寂期

2010左右-至今~繁荣期：深度学习

交替模式：
热十（年）
冷十五（年）



统计学习方法的春天（1986~2006）

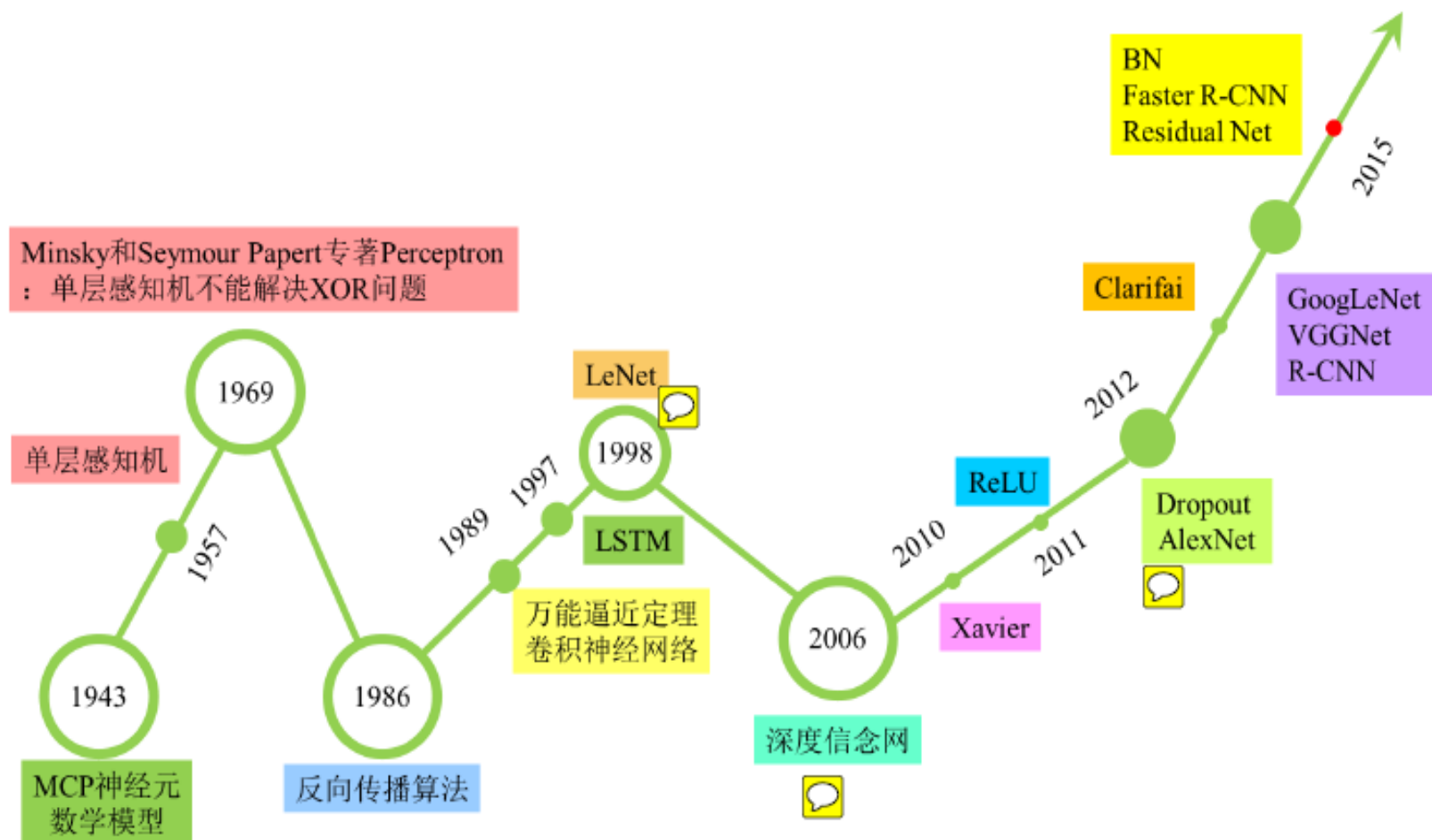
- 1986年提出决策树方法，很快ID3，ID4，CART等改进的决策树方法相继出现，到目前仍然是非常常用的一种机器学习方法。该方法也是符号学习方法的代表。
- 1995年，线性SVM被统计学家Vapnik提出。该方法的特点有两个：由非常完美的数学理论推导而来（统计学与凸优化等），符合人的直观感受（最大间隔）。不过，最重要的还是该方法在线性分类的问题上取得了当时最好的成绩。
- 1997年，AdaBoost被提出，该方法是PAC（Probably Approximately Correct）理论在机器学习实践上的代表，也催生了集成方法这一类。该方法通过一系列的弱分类器集成，达到强分类器的效果。
- 2000年，KernelSVM被提出，核化的SVM通过一种巧妙的方式将原空间线性不可分的问题，通过Kernel映射成高维空间的线性可分问题，成功解决了非线性分类的问题，且分类效果非常好。至此也更加终结了NN时代。
- 2001年，随机森林被提出，这是集成方法的另一代表，该方法的理论扎实，比AdaBoost更好的抑制过拟合问题，实际效果也非常不错。
- 2001年，一种新的统一框架-图模型被提出，该方法试图统一机器学习混乱的方法，如朴素贝叶斯，SVM，隐马尔可夫模型等，为各种学习方法提供一个统一的描述框架。

人工智能第三次浪潮（2006-）-- 基于大数据的深度学习

2006年，计算机处理速度和存储能力大大提高，为深度学习的提出铺平了道路。G. E. Hinton 和他的学生 R. R. Salakhutdinov 在《科学》杂志上发表题为《Reducing the Dimensionality of Data with Neural Networks》的文章，掀起了深度学习在学术界和工业界的研究热潮。文章摘要阐述了两个重要观点：一是多隐层的神经网络可以学习到能刻画数据本质属性的特征，对数据可视化和分类等任务有很大帮助；二是可以借助于无监督的“逐层初始化”策略来有效克服深层神经网络在训练上存在的难度。

这篇文章是一个分水岭，拉开了深度学习大幕，标志着深度学习的诞生。从此，历史这样写就：从感知机提出，到BP算法应用以及2006年以前的历史被称为浅层学习，以后的历史被称为深度学习。

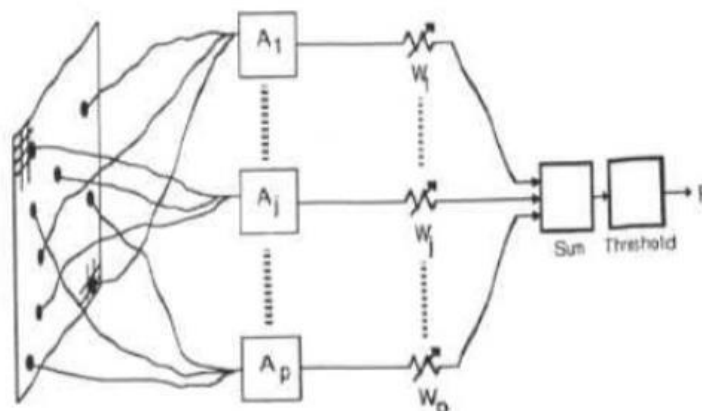
第三次人工智能浪潮——学习



人工神经网络发展历程

1957年, Frank Rosenblatt在《New York Times》上发表文章《Electronic 'Brain' Teaches Itself》, 首次提出了可以模拟人类感知能力的机器, 并称之为感知机 (Perceptron)

Perceptron (1957)



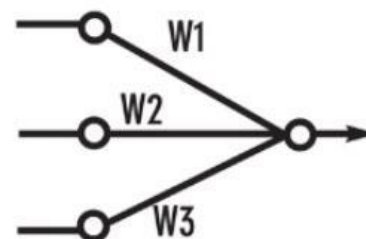
Original Perceptron

(From *Perceptrons* by M. L. Minsky and S. Papert, 1969, Cambridge, MA: MIT Press. Copyright 1969 by MIT Press.)



Frank Rosenblatt
(1928-1971)

Simplified model:



感知机的数学模型（是阈值）：

$$Y = f\left(\sum_{m=0}^{\infty} W_m X_m - \theta\right)$$

其中， $f(\cdot)$ 是阶跃函数，并且有：

$$f(u) = 1, u = \sum_{m=0}^{\infty} W_m X_m - \theta > 0$$

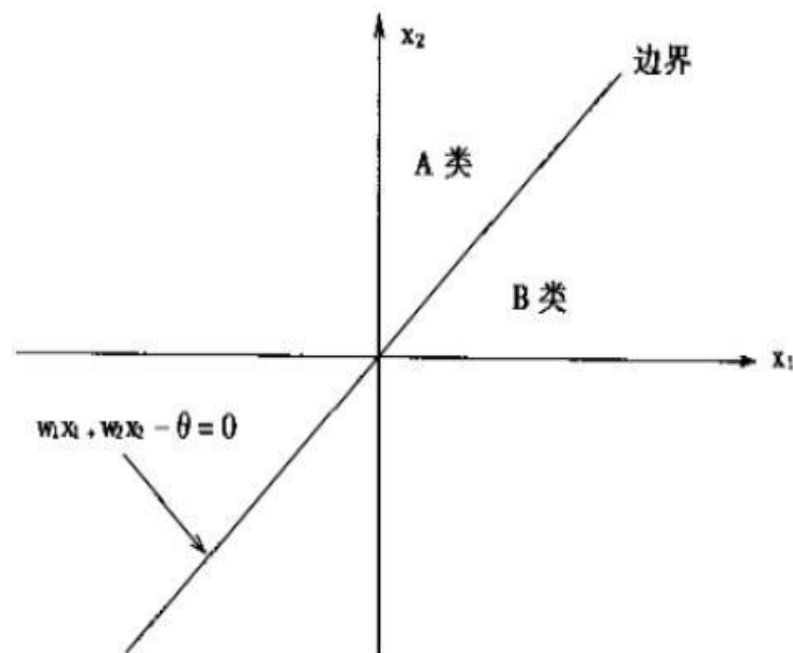
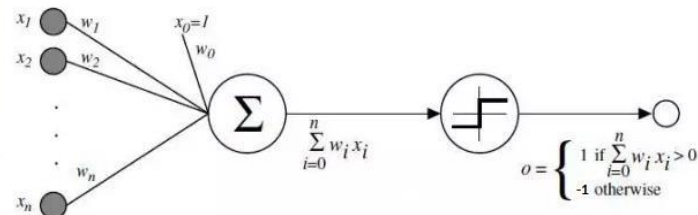
$$f(u) = -1, u = \sum_{m=0}^{\infty} W_m X_m - \theta \leq 0$$

感知器对输入的样本分类，故它可以作为分类器

在输入样本只有两个分量 x_1 和 x_2 时，则分类边界

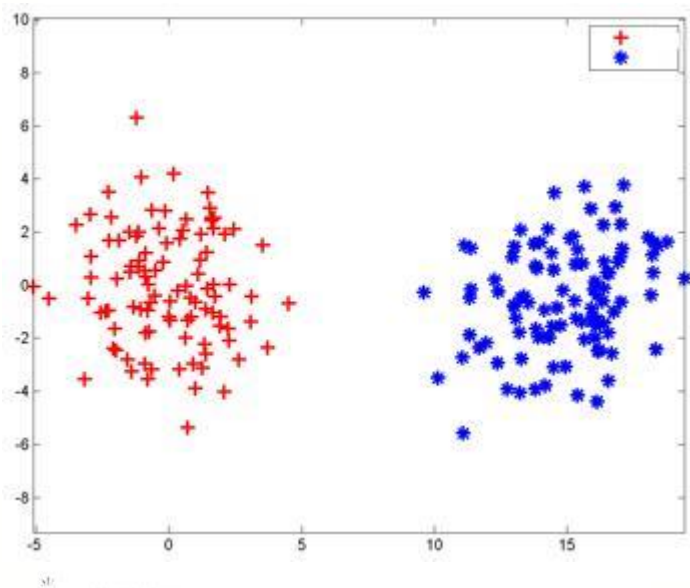
条件： $W_1 X_1 + W_2 X_2 - \theta = 0$

感知机的学习算法：目的在于计算出恰当的权系数（ w_1, w_2, \dots, w_n ），使系统对一个特定的样本（ x_1, x_2, \dots, x_n ）能产生期望值 d 。

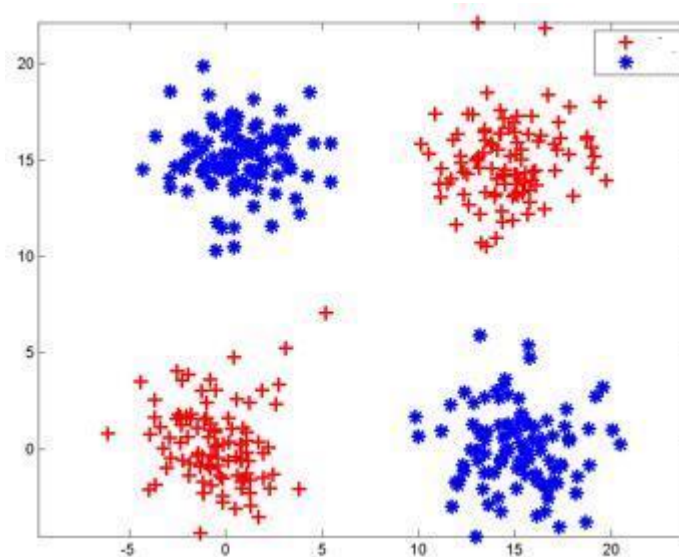


单层感知机的局限性

单层感知机仅对线性问题具有分类能力，即仅用一条直线可分的图形

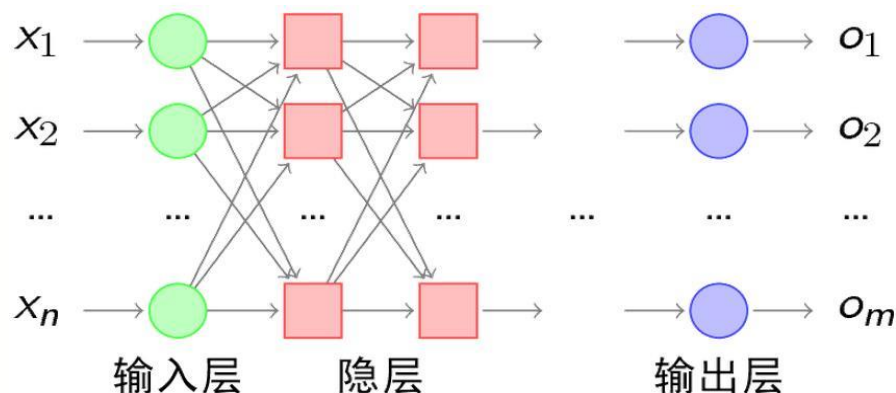


线性可分问题



非线性不可分问题

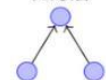

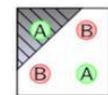
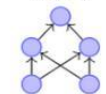

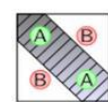
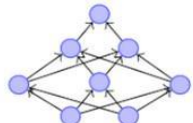

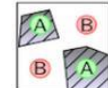
在单层感知机的输入层和输出层之间加入隐藏层，就构成了多层感知机，目的是通过凸域能够正确分类样本。



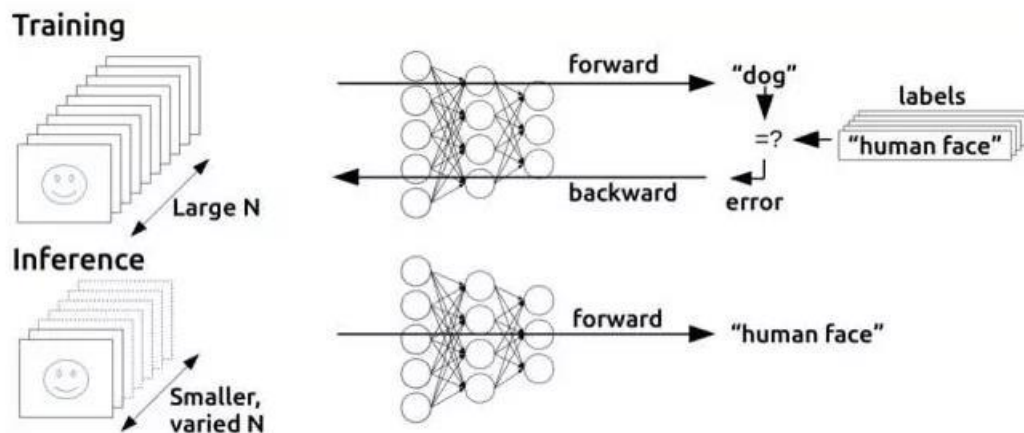
单层感知机和多层感知机
分类能力进行比较

多层感知机

问题：隐藏层的权值怎么训练？对于各隐层的节点来说，它们并不存在期望输出，所以也无法通过感知机的学习规则来训练多层感知机。

结构	决策区域类型	区域形状	异或问题
无隐层 	由一超平面分成两个		
单隐层 	开凸区域或闭凸区域		
双隐层 	任意形状（其复杂度由单元数目确定）		

多层感知器的误差反向传播(Error Back Propagation)算法:通过输出层得到输出结果和期望输出的误差来间接调整隐层的权值



思想：学习过程由信号的正向传播与误差的反向传播两个过程组成。

- 1) 正向传播时，输入样本从输入层传入,经各隐层逐层处理后,传向输出层。若输出层的实际输出与期望的输出不符,则转入误差的反向传播阶段。
- 2) 反向传播时，将输出以某种形式通过隐层向输入层逐层反传,并将误差分摊给各层的所有单元，从而获得各层单元的误差信号,此误差信号即作为修正各单元权值的依据。

BP神经网络模型的问题：基于局部梯度下降对权值进行调整容易出现梯度弥散（Gradient Diffusion）现象，根源在于非凸目标代价函数导致求解陷入局部最优，而不是全局最优。



浅层学习

自80年代末期，人们利用BP算法的人工神经网络模型从大量训练样本中学习统计规律（做预测）。这种基于统计的机器学习方法比起过去基于人工规则的系统凸显优越性。也被称多层感知机（Multi-layer Perceptron），但实际是种只含有一层隐层节点的浅层模型。

浅层机器学习模型相继被提出，例如支撑向量机（SVM，Support Vector Machines）、Boosting、最大熵方法（如LR，Logistic Regression）等。这些模型的结构基本上可以看成带有一层隐层节点（如SVM、Boosting），或没有隐层节点（如LR）。

缺点：理论分析的难度大，训练方法又需要很多经验和技巧

优点：

- (1) 神经网络具备拟合任意复杂函数的特点，例如实现复杂的非线性映射。
- (2) 由于深度神经网络的参数很多，得到的假设空间维度非常高，故有很强大的表征能力。
- (3) 可以自动获取特征，而且提取出来的特征要比人为设定的特定具有更强的泛化性能。
- (4) 具有较大的灵活性，通过设计不同的结构，可以改变网络提取特征的方法，同样也可以对拟合目标函数达到不同的效果。

缺点：

- (1) 由于假设空间大，易陷入局部最优。
- (2) 模型泛化能力较差，易出现过拟合。
- (3) 由于参数量大，训练速度较慢。
- (4) 在网络层数多的情况下，会出现梯度消失，且收敛速度较慢。
- (5) 一般只能用有标签的数据来训练。



深度学习 — 神经网络之后的又一突破

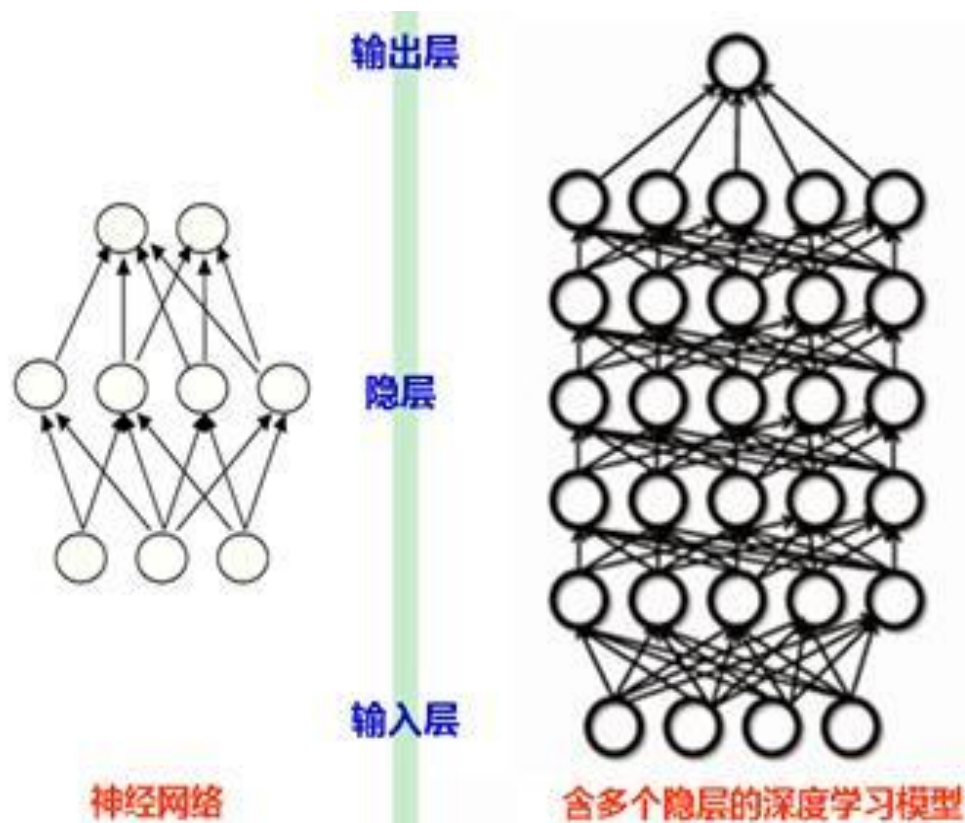
2006年，加拿大多伦多大学教授Geoffrey Hinton在《科学》上的一篇论文中提出：（1）多层人工神经网络模型有很强的特征学习能力，深度学习模型学习得到的特征数据对原始数据有更本质的代表性，这将大大便于分类和可视化问题；（2）对于深度神经网络很难训练达到最优的问题，可以采用逐层训练方法解决。

他提出深度置信网（Deep Belief Net：DBN），其由一系列受限波尔兹曼机（Restricted Boltzmann Machine：RBM）组成，并提出了非监督贪心逐层训练（Layerwise Pre-Training）算法，应用效果取得突破性进展。

深度学习的本质是对观察数据进行分层特征表示，实现将低级特征进一步抽象成高级特征表示。

深度学习分为：（1）生成型深度结构：学习模型有自编码器、受限玻尔兹曼机、深度置信网络等。（2）判别型深度结构：学习模型主要有卷积神经网络和深凸网络等。（3）混合型深度结构：例如通过深度置信网络进行预训练后的深度神经网络。

深度学习可通过学习一种深层非线性网络结构，表征输入数据，实现复杂函数逼近，并展现了强大的从少数样本集中学习数据集本质特征的能力。



深度学习采用了与神经网络相似的分层结构：系统是一个包括输入层、隐层（可单层、可多层）、输出层的多层网络，只有相邻层的节点之间有连接，而同一层以及跨层节点之间相互无连接。

相比传统的神经网络，深度神经网络作出了重大的改进，在训练上的难度（如梯度弥散问题）可以通过“逐层预训练”来有效降低。

通过机器学习解决图像识别、语音识别、自然语言理解、天气预测、基因表达、内容推荐等问题的思路

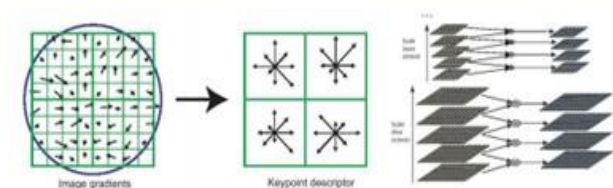
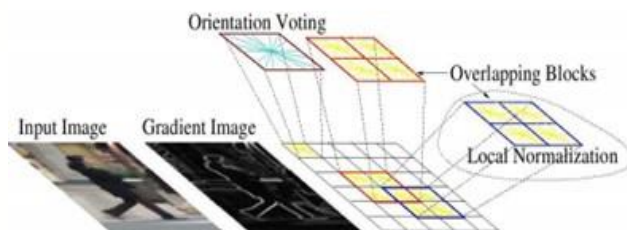
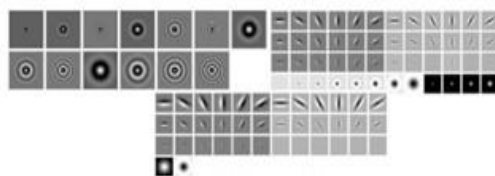
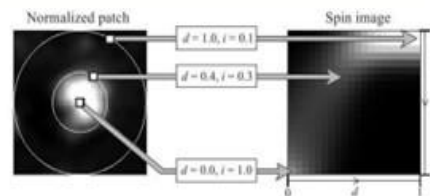
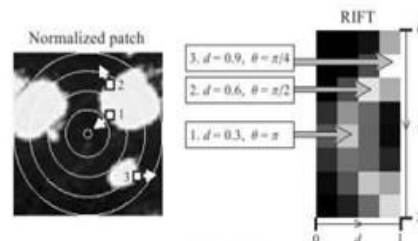
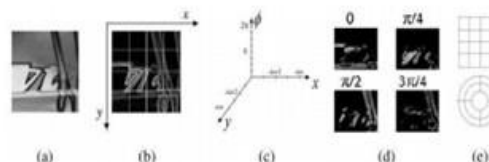


关键是中间三个环节，概括起来就是特征表达。良好的特征表达，对最终算法的准确性起了非常关键的作用，一般都是人工完成的，靠人工提取特征。

手工地选取特征是一件非常费力、启发式（需要专业知识）的方法，能不能选取好很大程度上靠经验和运气，而且它的调节需要大量的时间。

既然手工选取特征不太好，那么能不能自动地学习一些特征呢？答案是能！**Deep Learning**就是用来干这个事情的，看它的一个别名Unsupervised Feature Learning，就可以顾名思义了，Unsupervised的意思就是不要人参与特征的选取过程。

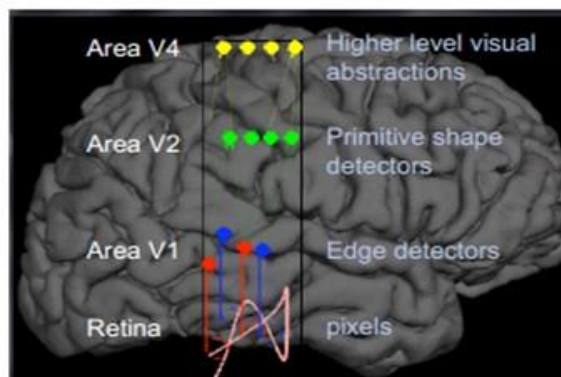
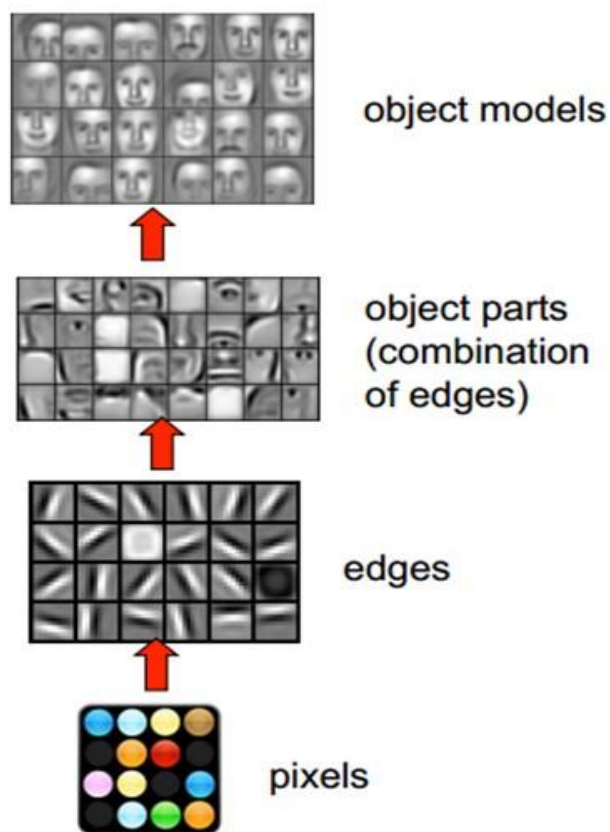


**SIFT****HoG****Textons****Spin image****RIFT****GLOH**

- SIFT (Scale invariant feature transform , Lowe, 2004.) : 特征不变尺度变换
- HoG (Histogram of Oriented Gradient , Dalal , 2005) : 梯度直方图
- Textons : 边缘检测及纹理分析
- Spin Image : 三维计算机视觉中基于点云空间分布的特征描述方法
- RIFT (Rotation-Invariant Feature Transform) : 三维计算机视觉中一种局部特征描述法
- GLOH (Gradient location-orientation histogram , Mikolajczyk and Schmid 2005) : 梯度位置方向直方图

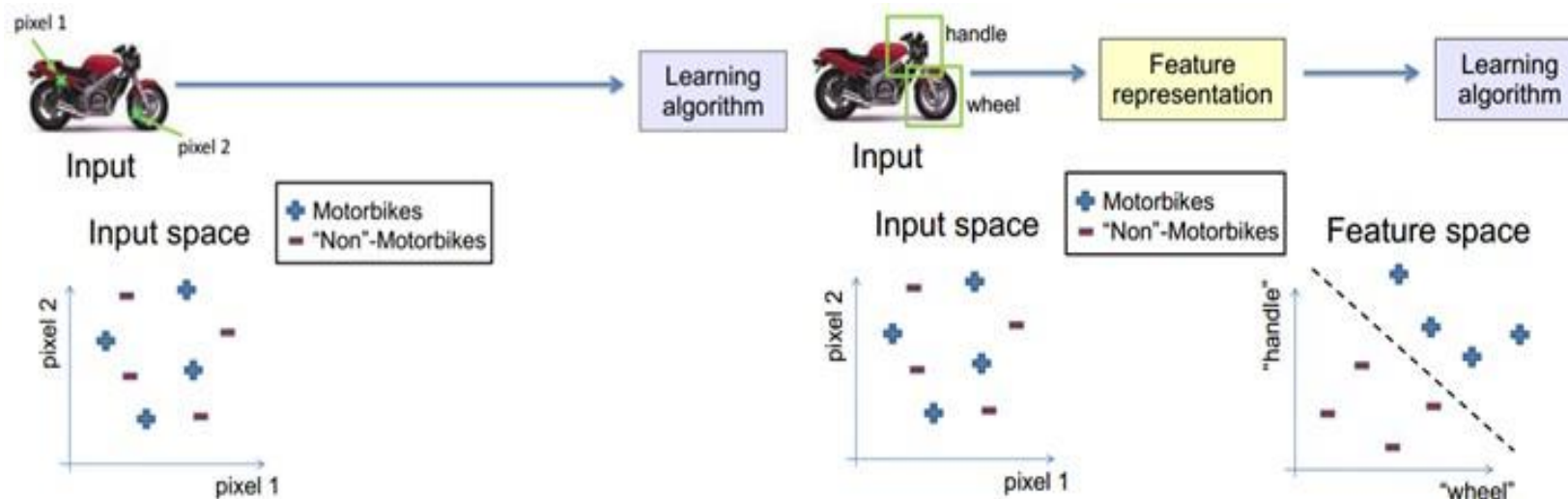
典型特征提取方法

1981 年的诺贝尔医学奖获得者David Hubel和Torsten Wiesel发现了视觉系统的信息处理：**可视皮层是分级的**。即从原始信号，做低级抽象，逐渐向高级抽象迭代。

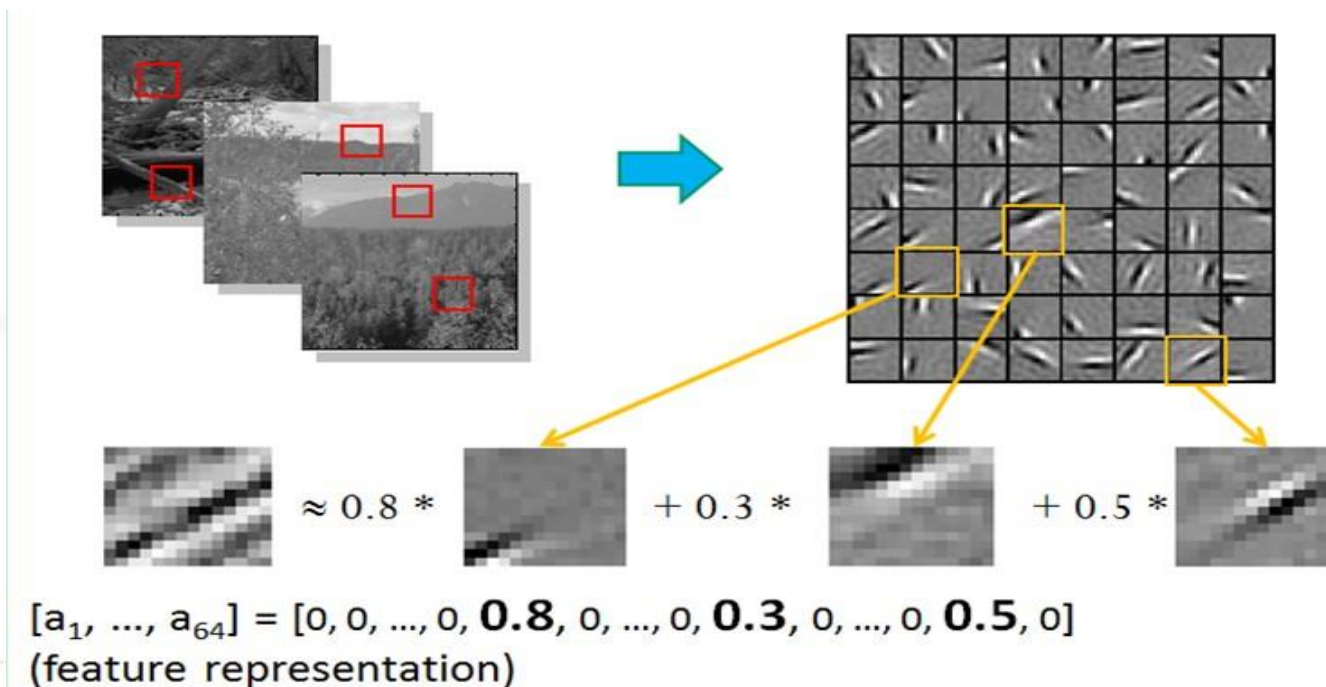


人的视觉系统的信息处理是分级的。从低级的V1区提取边缘特征，再到V2区的形状或者目标的部分等，再到更高层，整个目标、目标的行为等。也就是说高层的特征是低层特征的组合，从低层到高层的特征表示越来越抽象，越来越能表现语义或者意图

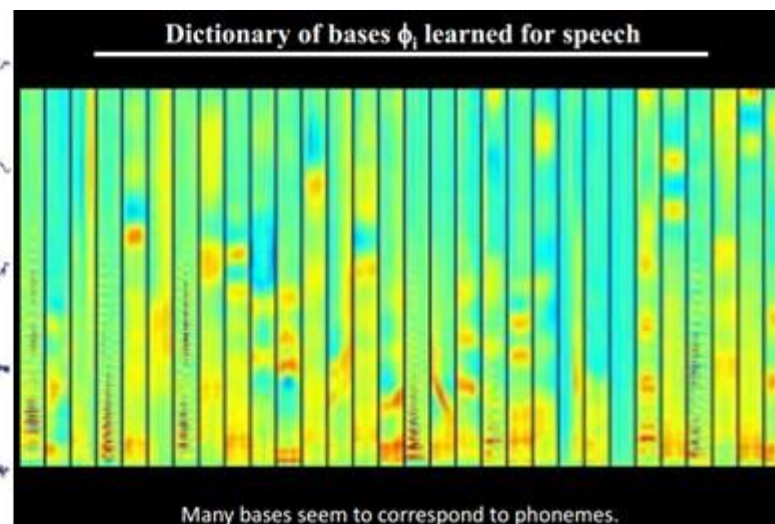
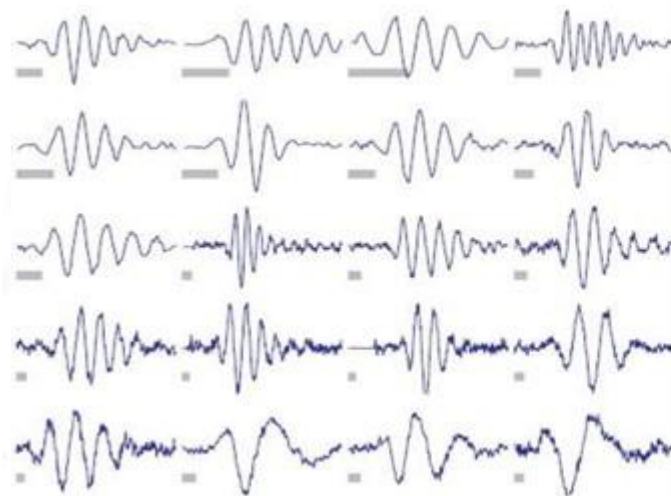
学习算法在一个什么粒度上的特征表示，才有能发挥作用？就一个图片来说，像素级的特征根本没有价值。例如下面的摩托车，从像素级别，根本得不到任何信息，其无法进行摩托车和非摩托车的区分。而如果特征是一个具有结构性（或者说有含义）的时候，比如是否具有车把手（handle），是否具有车轮（wheel），就很容易把摩托车和非摩托车区分，学习算法才能发挥作用。



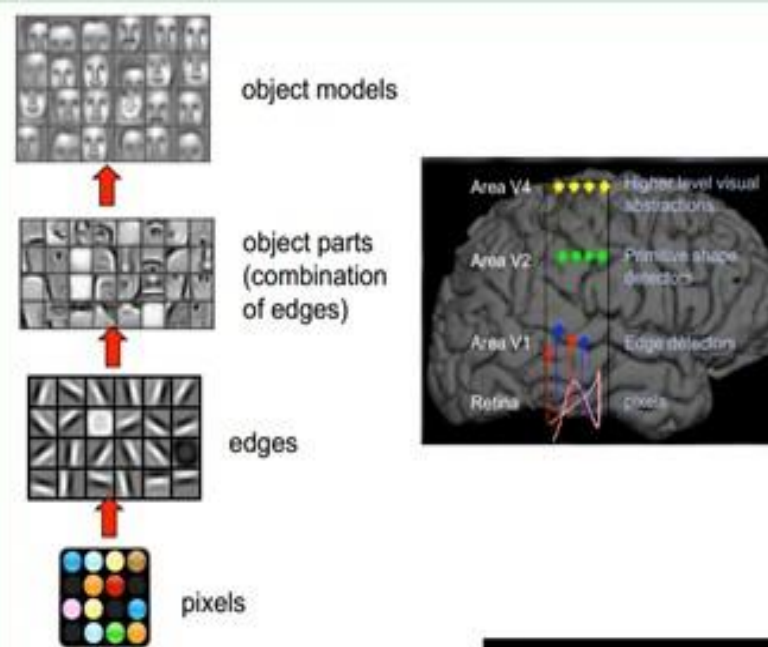
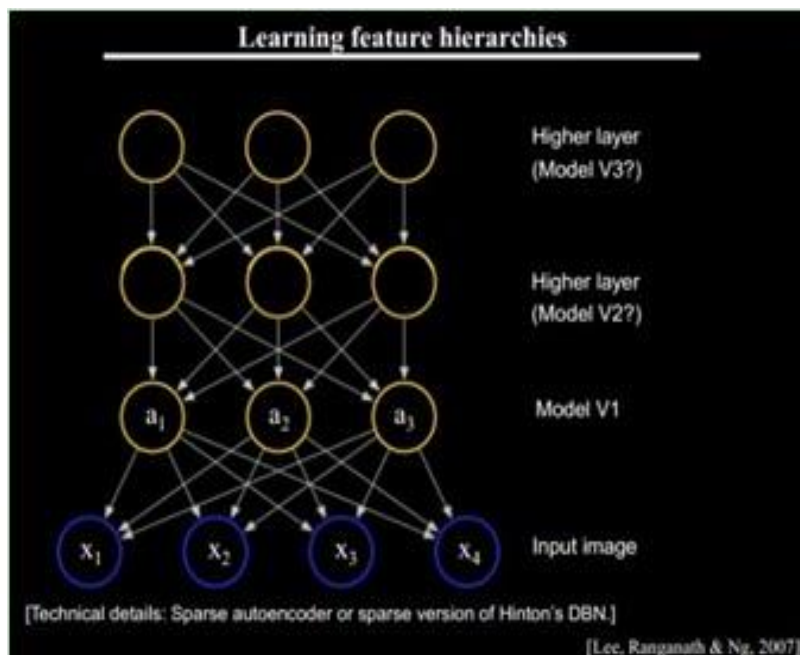
1995 年前后，Bruno Olshausen和 David Field从很多黑白风景照片，提取出400个小碎片，每个照片碎片的尺寸均为 16x16 像素。通过分析发现：复杂图形，往往由一些基本结构组成。比如下图：一个图可以通过用64种正交的edges（可以理解成正交的基本结构）来线性表示。比如样例的x可以用1-64个edges中的三个按照0.8,0.3,0.5的权重调和而成。而其他基本edge没有贡献，因此均为0。

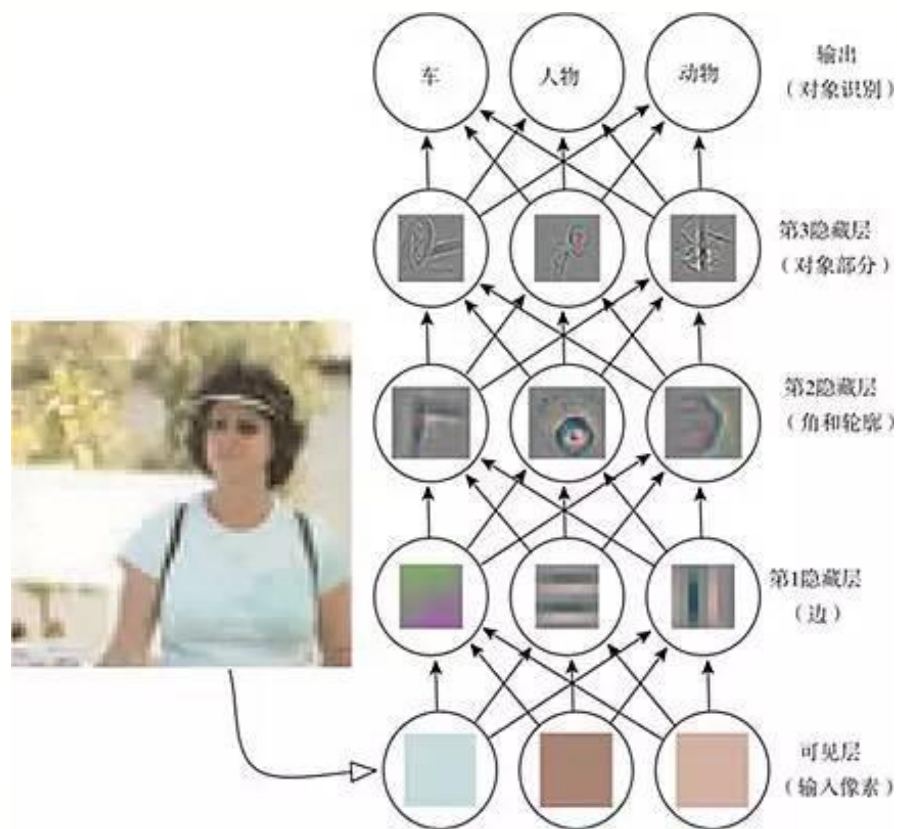


此外，还发现不仅图像存在这个规律，声音也存在。他们从未标注的声音中发现了20种基本的声音结构，其余的声音可以由这20种基本结构合成。



小块的图形可以由基本edge构成，更结构化，更复杂的，具有概念性的图形如何表示呢？这就需要更高层次的特征表示，比如V2，V4。因此V1看像素级是像素级。V2看V1是像素级，这个是层次递进的，高层表达由底层表达的组合而成。专业点说就是基basis。V1取提出的basis是边缘，然后V2层是V1层这些basis的组合，这时候V2区得到的又是高一层的basis。即上一层的basis组合的结果，上上层又是上一层的组合basis.....（所以有大牛说Deep learning就是“搞基”，因为难听，所以美其名曰Deep learning或者Unsupervised Feature Learning）





深度学习将所需的复杂映射分解为一系列嵌套的简单映射（每个由模型的不同层描述）来解决这一难题。输入展示在可见层visible layer。

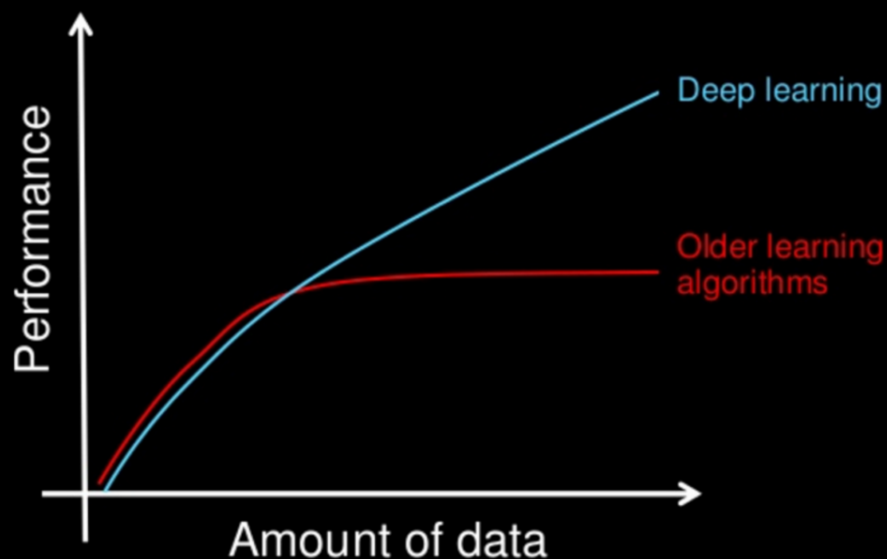
- 一系列从图像中提取越来越多抽象特征的隐藏层hidden layer。
- 给定像素，第 1 层可以轻易地通过比较相邻像素的亮度来识别边缘。
- 有了第 1 隐藏层描述的边缘，第 2 隐藏层可以容易地搜索可识别为角和扩展轮廓的边集合。
- 给定第 2 隐藏层中关于角和轮廓的图像描述，第 3 隐藏层可以找到轮廓和角的特定集合来检测特定对象的整个部分。
- 最后，根据图像描述中包含的对象部分，可以识别图像中存在的对象

Zeiler and Fergus (2014)

深度学习模型的示意图



Why deep learning



How do data science techniques scale with amount of data?

深度学习与传统的机器学习最主要的区别在于随着数据规模的增加其性能也不断增长。当数据很少时，深度学习算法的性能并不好。这是因为深度学习算法需要大量的数据来完美地理解它。另一方面，在这种情况下，传统的机器学习算法使用制定的规则，性能会比较好。



深度学习的优势

- Hinton：对于深度神经网络很难训练达到最优的问题，可以采用逐层训练方法解决。（Layer-wise training）
- 通过逐层特征变换，将样本在原空间的特征表示变换到一个新特征空间，从而使分类或预测更加容易。（Feature transformation）
- 深度学习可通过学习一种深层非线性网络结构，实现复杂函数逼近，表征输入数据分布式表示，并展现了强大的从少数样本集中学习数据集本质特征的能力。（Sufficient model complexity representation）

深度学习的不足

- 训练需要大量数据，导致决策上对数据过度依赖
- 可迁移性差（环境适应性）
- 可理解性差
- 训练方法多为技巧性方法



深度前馈网络 (deep feedforward network) , 也叫作前馈神经网络 (feedforward neural network) 或者多层感知机 (multilayer perceptron, MLP) , 是典型的深度学习模型。

卷积网络 (convolutional network) (LeCun, 1989) , 也叫做卷积神经网络 (convolutional neural network, CNN) , 是一种专门用来处理具有类似网格结构的数据的神经网络。

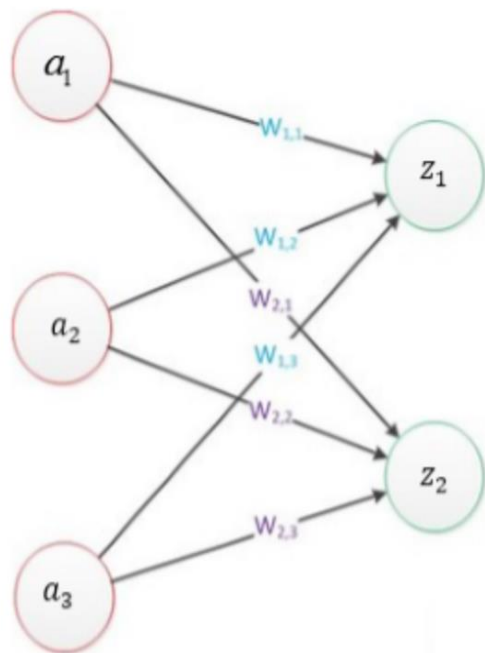
循环神经网络 (recurrent neural network) 或RNN (Rumelhart et al., 1986c) 是一类用于处理序列数据的神经网络。就像卷积网络是专门用于处理网格化数据X (如一个图像) 的神经网络 , 循环神经网络是专门用于处理序列 $x(1); :::: x()$ 的神经网络。

递归神经网络代表循环网络的另一个扩展 , 它被构造为深的树状结构而不是RNN 的链状结构 , 因此是不同类型的计算图。递归神经网络由Pollack (1990) 引入 , 而Bottou (2011) 描述了这类网络的潜在用途——学习推论。递归网络已成功地应用于输入是数据结构的神经网络(Frasconi et al., 1997, 1998) , 如自然语言处理(Socher et al., 2011a,c, 2013a) 和计算机视觉(Socher et al., 2011b)。

深度前馈神经网络也叫作多层感知机，是深度学习中最常用的模型。

单层神经网络

下图显示了带有两个输出单元 z_1, z_2 的单层神经网络：

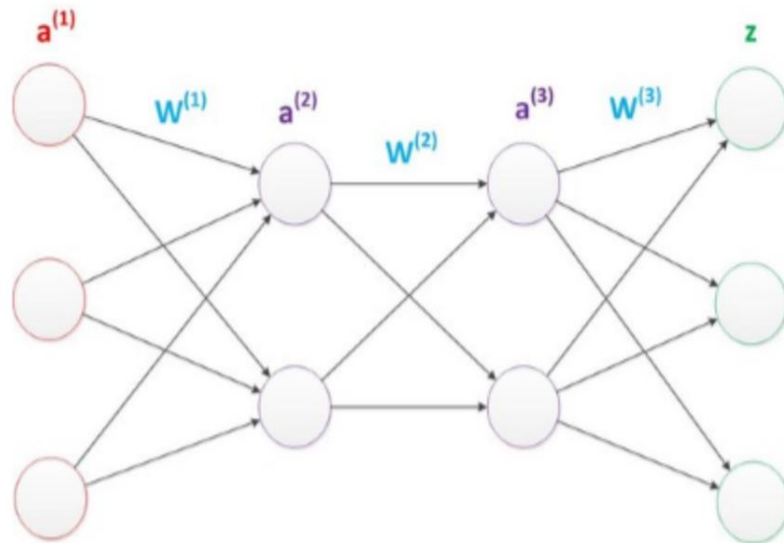


$$z_1 = f(a_1 * w_{1,1} + a_2 * w_{1,2} + a_3 * w_{1,3})$$

$$z_2 = f(a_1 * w_{2,1} + a_2 * w_{2,2} + a_3 * w_{2,3})$$

多层神经网络

多层神经网络的结构：在两层神经网络的输出层后面，继续添加层次。原来的输出层变成中间层，新加的层次成为新的输出层。所以可以得到下图。



已知输入 $a^{(1)}$ ，参数 $W^{(1)}$ ， $W^{(2)}$ ， $W^{(3)}$ 的情况下，输出 z 的推导公式如下：

$$f(W^{(1)} * a^{(1)}) = a^{(2)}$$

$$f(W^{(2)} * a^{(2)}) = a^{(3)}$$

$$f(W^{(3)} * a^{(3)}) = z$$



卷积神经网络 (convolutional neural network, CNN)

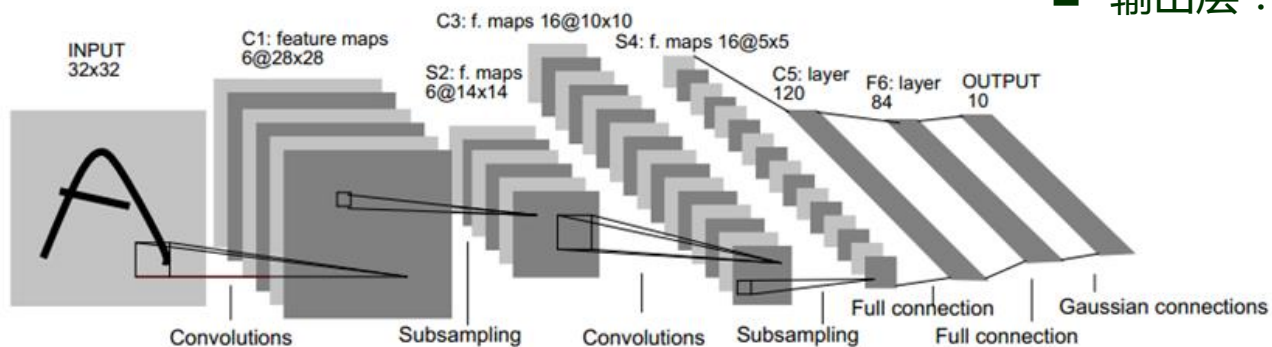
CNN经典网络形式

- **LeNet** : 最早用于数字识别的CNN
- **AlexNet** : 2012年ILSVRC比赛冠军, 远超第二名的CNN, 比LeNet更深, 用**多层小卷积叠加来替换单个的大卷积**
- **ZF Net** : 2013ILSVRC冠军
- **GoogleNet** : 2014ILSVRC冠军
- **VGGNet** : 2014ILSVRC比赛中算法模型, 效果率低于GoogleNet
- **ResNet** : 2015ILSVRC冠军, **结构修正以适应更深层次的CNN训练**

<http://blog.csdn.net/loveliuzz>

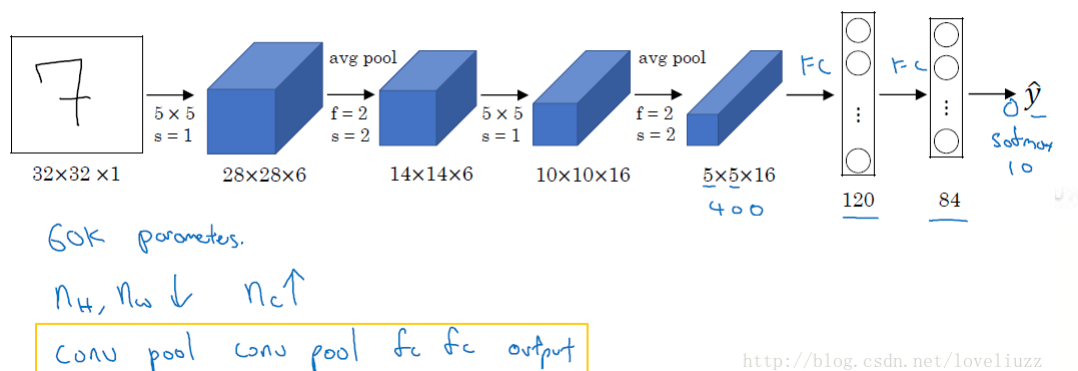
LeNet-5网络

- 输入尺寸：32*32
- 卷积层：2个
- 降采样层(池化层)：2个
- 全连接层：2个
- 输出层：10个类别（数字0-9的概率）



<http://blog.csdn.net/loveliuzz>

LeNet - 5



<http://blog.csdn.net/loveliuzz>

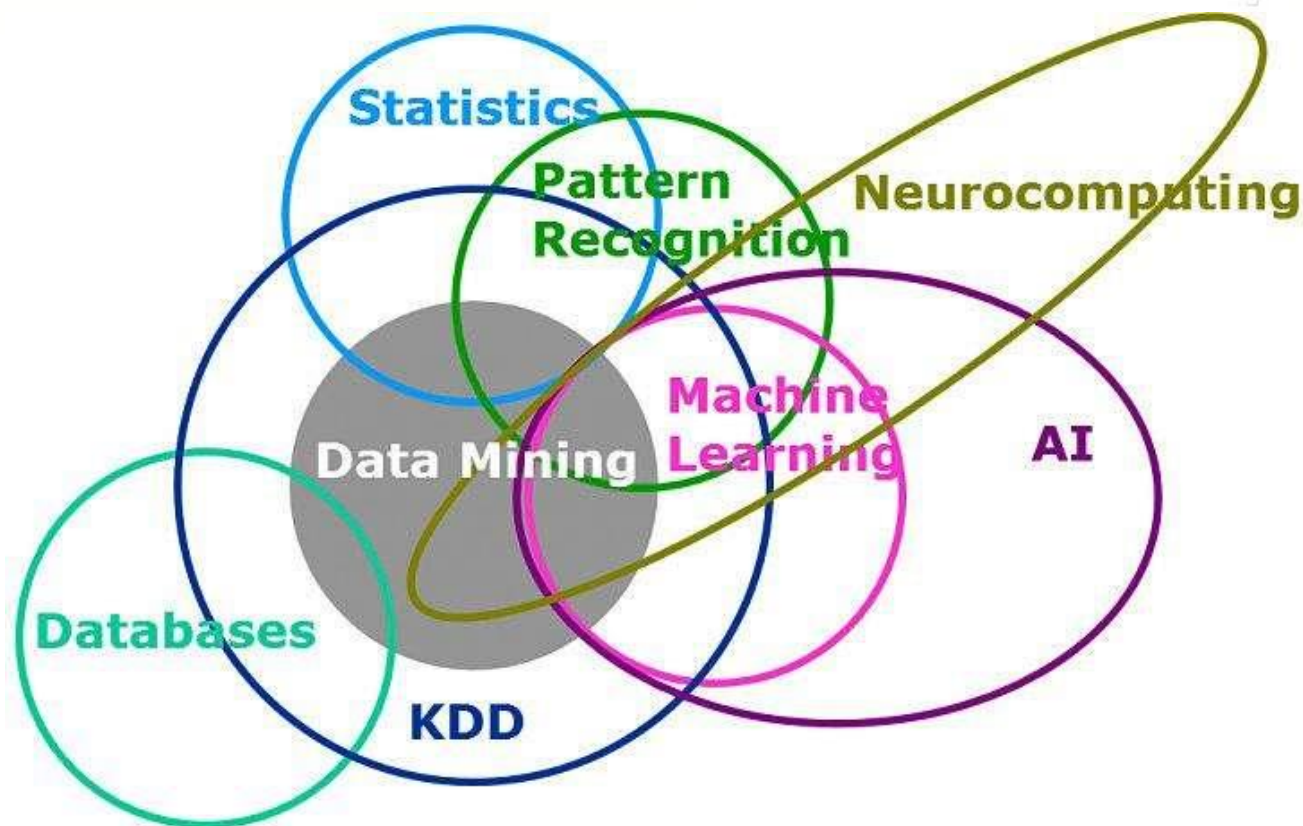
LeNet-5网络是针对灰度图进行训练的，输入图像大小为32*32*1，共有7层，每层都包含可训练参数（连接权重）。注：每个层有多个Feature Map，每个Feature Map通过一种卷积滤波器提取输入的一种特征，然后每个Feature Map有多个神经元。



3、机器学习的分类



- 数据库
- 统计学
- 数据挖掘
- 知识发现
- 神经计算
- 模式识别
- 机器学习
- 深度学习
- 人工智能





符号学派

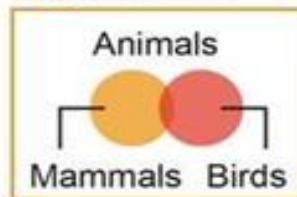
贝叶斯学派

联结学派

进化学派

类推学派

Symbolists

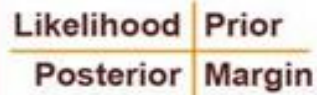


Use symbols, rules, and logic to represent knowledge and draw logical inference

Favored algorithm

Rules and decision trees

Bayesians

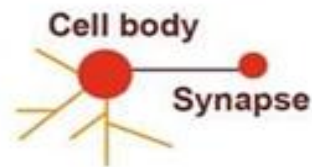


Assess the likelihood of occurrence for probabilistic inference

Favored algorithm

Naive Bayes or Markov

Connectionists



Recognize and generalize patterns dynamically with matrices of probabilistic, weighted neurons

Favored algorithm

Neural networks

Evolutionaries



Generate variations and then assess the fitness of each for a given purpose

Favored algorithm

Genetic programs

Analogizers



Optimize a function in light of constraints ("going as high as you can while staying on the road")

Favored algorithm

Support vectors

符号学派 (Symbolists)：是使用基于规则的符号系统做推理的人。大部分AI都围绕着这种方法。使用Lisp和Prolog的方法属于这一派，使用Semantic Web，RDF和OWL的方法也属于这一派。其中一个最雄心勃勃的尝试是Doug Lenat在80年代开发的Cyc，试图用逻辑规则将我们对这个世界的理解编码。这种方法主要的缺陷在于其脆弱性，因为在边缘情况下，一个僵化的知识库似乎总是不适用。但在现实中存在这种模糊性和不确定性是不可避免的。常用方法：规则和决策树



贝叶斯学派 (Bayesians) : 是使用概率规则及其依赖关系进行推理的一派。概率图模型 (PGM) 是这一派通用的方法 , 主要的计算机制是用于抽样分布的蒙特卡罗方法。这种方法与符号学方法的相似之处在于 , 可以以某种方式得到对结果的解释。这种方法的另一个优点是存在可以在结果中表示的不确定性的量度。常用方法 : 朴素贝叶斯或马尔科夫



联结学派 (Connectionists) : 这一派的研究者相信智能起源于高度互联的简单机制。这种方法的第一个具体形式是出现于1959年的感知器。自那以后, 这种方法消亡又复活了好几次。其最新的形式是深度学习。常用方法: 神经网络

进化学派 (Evolutionists) : 是应用进化的过程 , 例如交叉和突变以达到一种初期的智能行为的一派。在深度学习中 , GA确实有被用来替代梯度下降法 , 所以它不是一种孤立的方法。这个学派的人也研究细胞自动机 (cellular automata) , 例如Conway的 “生命游戏” 和复杂自适应系统 (GAS) 。常用方法 : 遗传算法

类推学派 (The analogizers) : 更多地关注心理学和数学最优化, 通过外推来进行相似性判断。类推学派遵循 “最近邻” 原理进行研究。各种电子商务网站上的产品推荐 (例如亚马逊或 Netflix 的电影评级) 是类推方法最常见的示例。常用方法 : 支持向量机 (SVM)



谢谢！