SaturnLab

iCenter
清华大学iCenter

Artificial Intelligence，Machine Learning and Deep Learning
人工智能，机器学习和深度学习

智能系统实验室
清华大学基础工业训练中心

目录

- 人工智能
- 机器学习
- 深度学习

机器智能
/人工智能

人工/机器智能-简述

- 人工智能或机器智能，通俗的说，就是用机器（如计算机）完成人类需要用脑子完成的任务，代替人脑的工作，比如下棋、开车、阅读理解等等工作。

- 人工智能或机器智能，是指计算机体现的智能的能力，如听说读写到搜索、推理、决策和回答问题等。

- 人工智能或机器智能，是指如何设计实现计算机系统和软件，使其具有智能的行为。

人工/机器智能

机器认知
(Cognition)

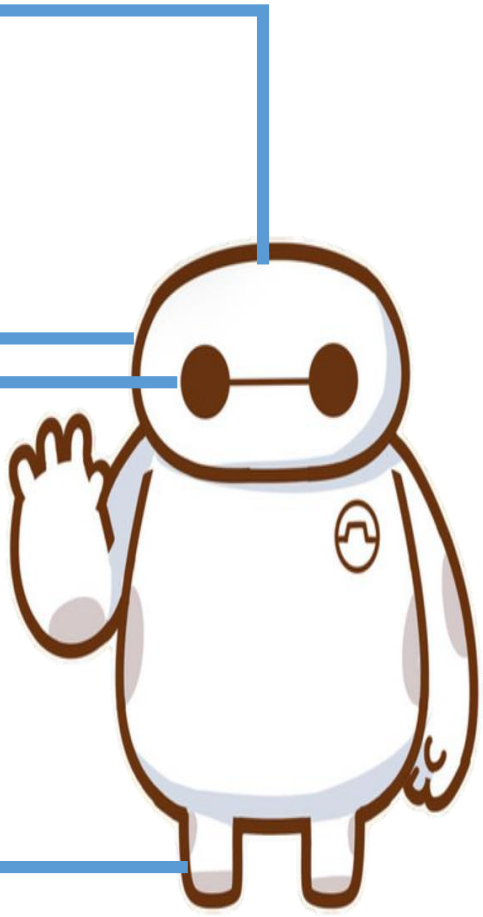
让机器像人一样思考
——机器学习
——自动推理
——人工意识
——知识表示

机器感知
(Perception)

让机器像人一样听懂
——语音识别

让机器像人一样看懂
——视觉识别

让机器像人一样运动
——运动控制



机器智能

机器认知
(Cognition)

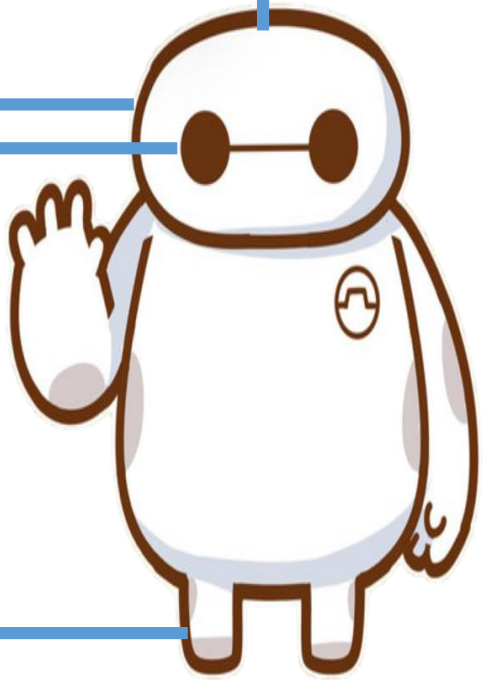
让机器像人一样思考
——机器学习
——自动推理
——人工意识
——知识表示

机器感知
(Perception)

让机器像人一样听懂
——语音识别

让机器像人一样看懂
——视觉识别

让机器像人一样运动
——运动控制



人工智能

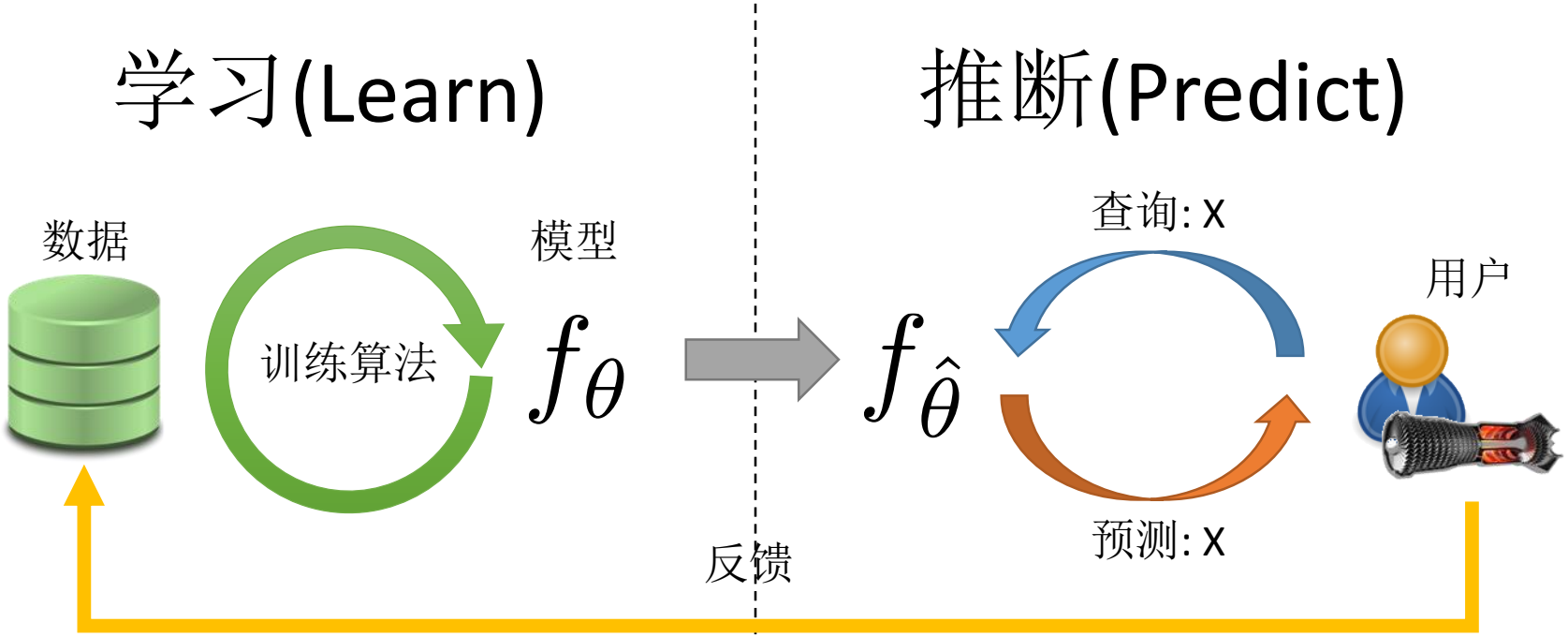
机器学习

机器学习的定义

- 机器学习(Machine Learning)的一个通俗的定义：
  - 给定一个计算机任务T和一个对任务T的性能度量P。
  - 在给出经验集E的前提下，计算机任务T在性能度量P上有提升。
  - 利用经验集E提升任务T的性能P的方法就是机器学习。

- 机器学习是人工智能的一个研究领域
- 深度学习是机器学习的一个分支。

机器学习的过程



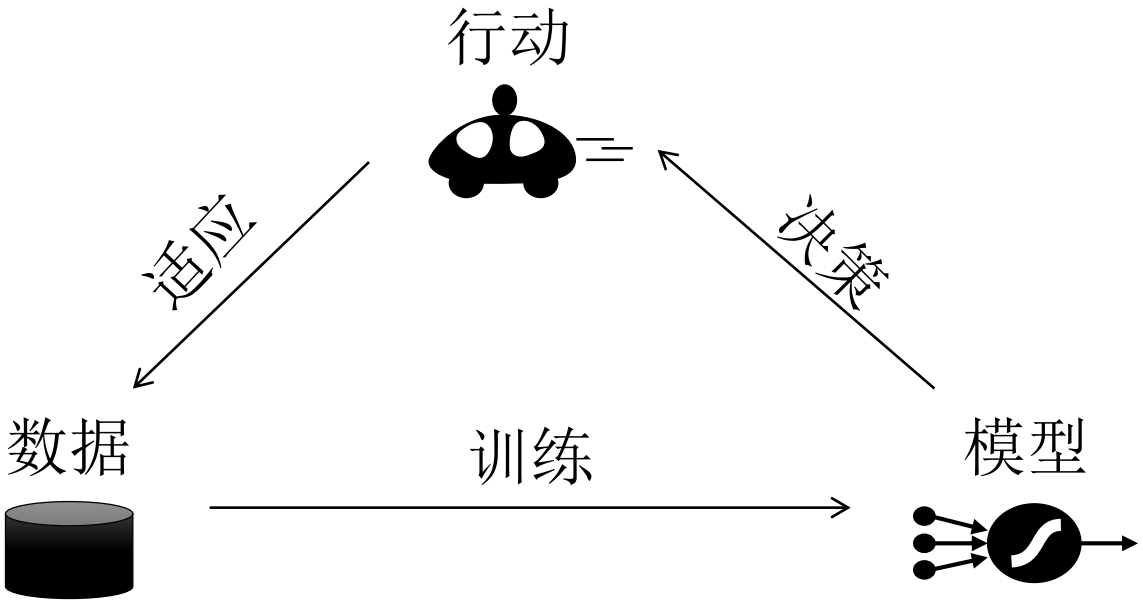
- 训练过程耗时，迭代批处理过程
- 进行快速和鲁棒的预测
  - 检测与追踪反馈
  - 快速推断

机器学习的要素

- 机器学习 = 数据+模型+算法

  训练：建立统计模型，运用数据，用算法去求解模型参数。

  预测：使用训练得到模型参数，预测新的观察事件。



机器学习的任务

- 预测：通过数据集，预测新的数值
- 分类：通过数据集，对新数值进行集合分类
- 房价预测：机器学习==数学建模？
  - 数据：以往房价数据
  - 模型：（假设）线性模型y=a*x + b
  - 算法：求出a, b（最小二乘法）

机器学习的推广能力

- 推广（Generalization）：机器学习通过训练数据上的学习，然后能够推广到新的数据集上的能力。 也称为泛化能力。
- 推广误差（Generalization error）：推广后与正确的分类结果产生的误差。一般用数学公式表示为：GE= AE + EE + OE
  - 逼近误差AE（Approximation error），是指由于模型规模方面而产生误差，要想减少这部分误差，需要扩大模型规模。（模型误差）
  - 估计误差EE（ Estimation error ），是指由数据集规模而产生的误差，要想减少这部分误差，需要增加可用数据的规模。（数据误差）
  - 优化误差OE（Optimization error），是指算法设计而产生的误差，要降低这部分误差，需要设计更优的算法。（算法误差）

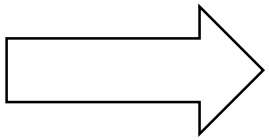
机器学习的本质

- 前提：数据集（经验集）

- 核心两点：
  - （1）机器的学习是可教的
  - （2）机器的学习是可推广的。

机器学习

算法模型



机器智能

智能服务

机器智能服务

To JSON

HTTP Req.

Feature Trans.

机器学习

Encode Prediction

HTTP Response

4

大数据与机器智能

- 假设：机器学习的模型具有好的推广性。
- 更大的数据集，意味着更大的模型和更高的准确度；
- 实用案例：深度学习模型（深度神经网络）

大数据 *训练* 大模型

人工智能

深度学习

深度学习Deep Learning

深度学习的三个领军人物

- Geoffrey Hinton , Yoshua Bengio, Yann LeCun

- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *Nature* 521(7553), pp: 436-444, 2015.

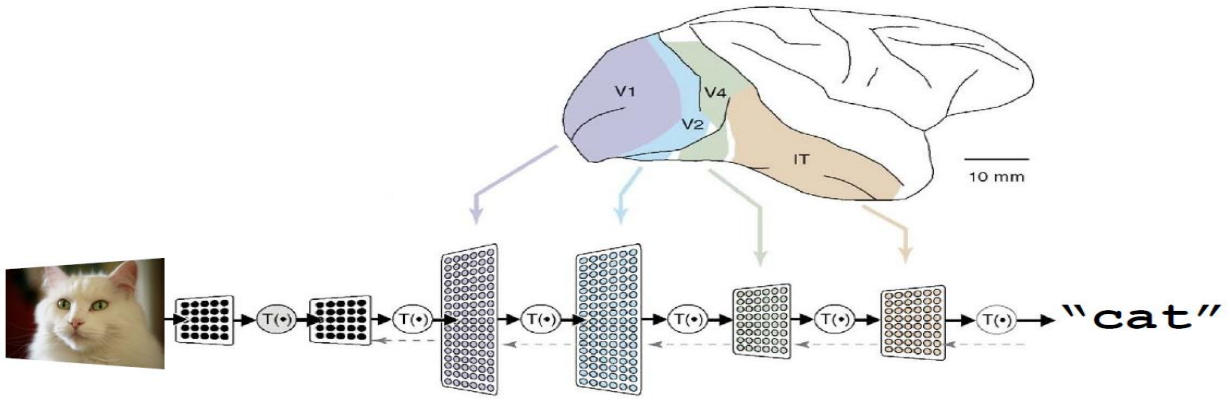


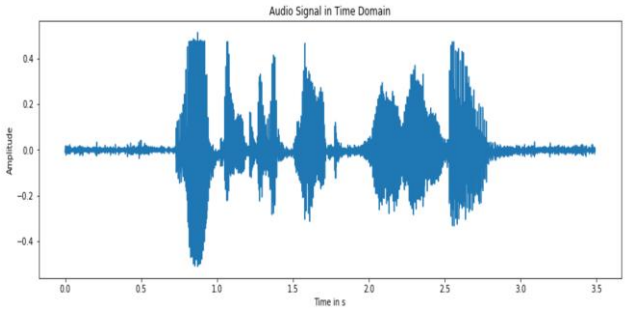
标志性论文

- 深度信仰网络（ deep belief nets ）（2006）
  - Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18.7 (2006): 1527-1554.
- AlexNet（2012）
  - [x] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks." NIPS 2012.

人脑？

- 人脑有正确识别图片的视觉处理能力。
- 人脑有正确识别语音的听觉能力。
- 原因：人脑中存在大量神经元组成的神经网络



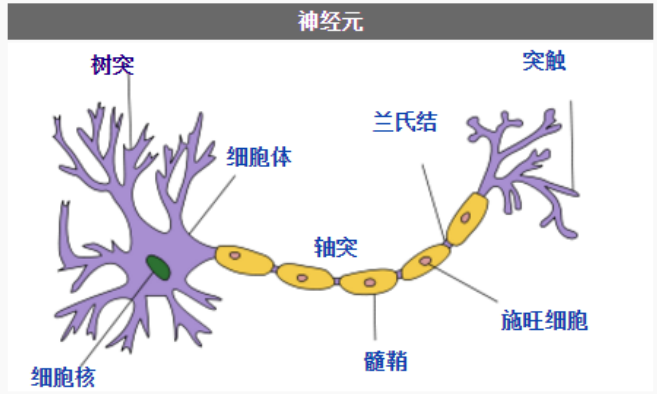


拨打电话

神经元

- 单个神经元（Neuron）的活动原理，目前已经有了比较深入的研究。

- 不论何种神经元，皆可分成：
  - 接收区（receptive zone）
  - 触发区（trigger zone）
  - 传导区（conducting zone）
  - 输出区（output zone）



典型神经元的结构

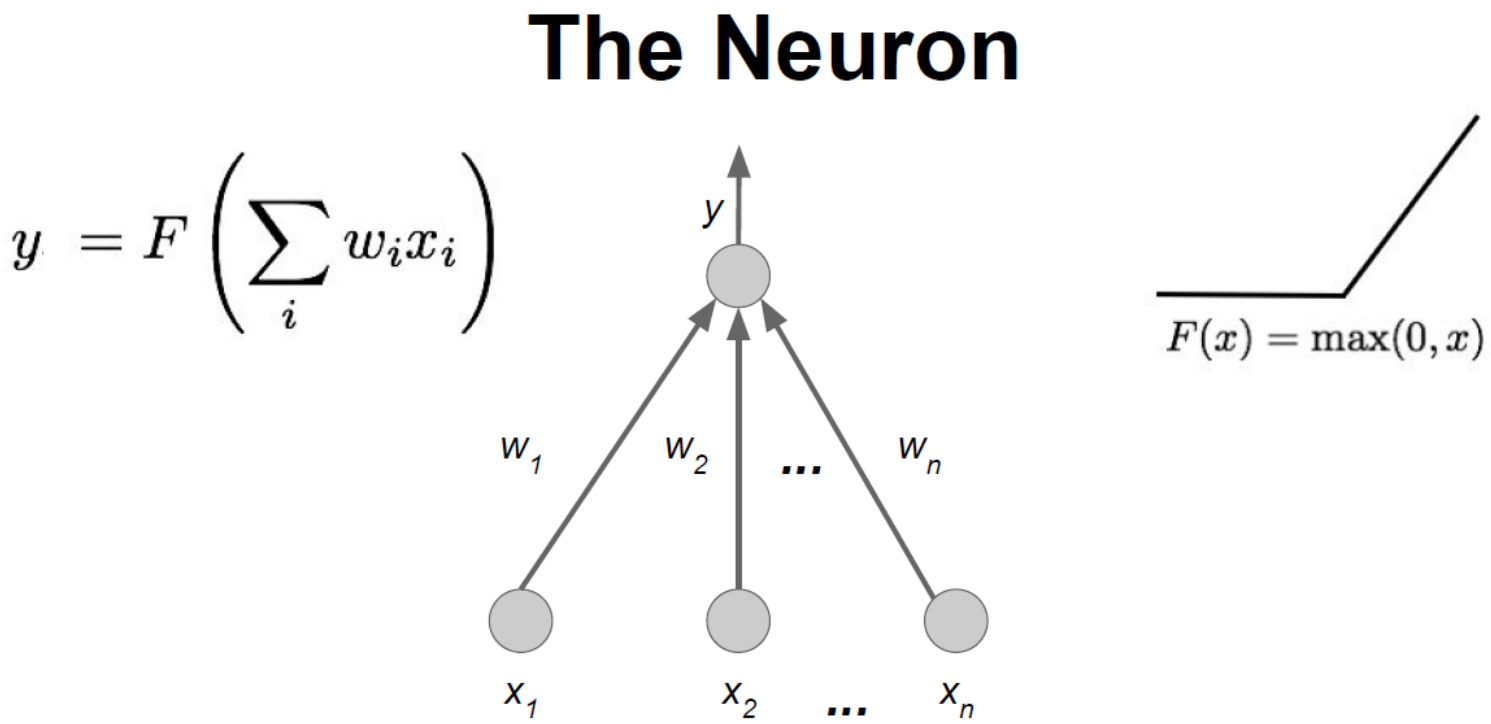
神经元→人工神经元

- 神经元的功能可以数学建模出来，形成人工神经元（Artificial Neuron）。
- 神经元的模型包括：接收的线性部分和触发的非线性部分。
- 历史：
  - McCulloch-Pitts神经元：是生物神经元的一种数学抽象模型（1943）
  - “B型图灵机”：Alan Turing描述的一种智能机器（1948）

[x] W. S. McCulloch and W. Pitts, ‘A Logical Calculus of the Ideas Immanent in Nervous Activity’, Bulletin of Mathematical Biophysics, 5 (1943), 115–33.
[x] A. M. Turing, Intelligent Machinery: A Report for National Physical Laboratory, 1948.

具体建模过程

- 神经元被建模为一个函数F(w, x)，其中w是权重，x是输入
  - 输入的线性加权叠加
  - 一个非线性函数F作用，进行输出，F称为激活函数
  - 激活函数模拟神经元的触发激活特性



激活函数（activation function）

- 激活函数有sigmod函数，tanh函数，ReLU函数等
- Sigmoid函数，又称为逻辑斯提函数(logistic function)或S形函数。数学表达式为：

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$$

- tanh函数或th函数，双曲正切函数，其数学表达式如下：

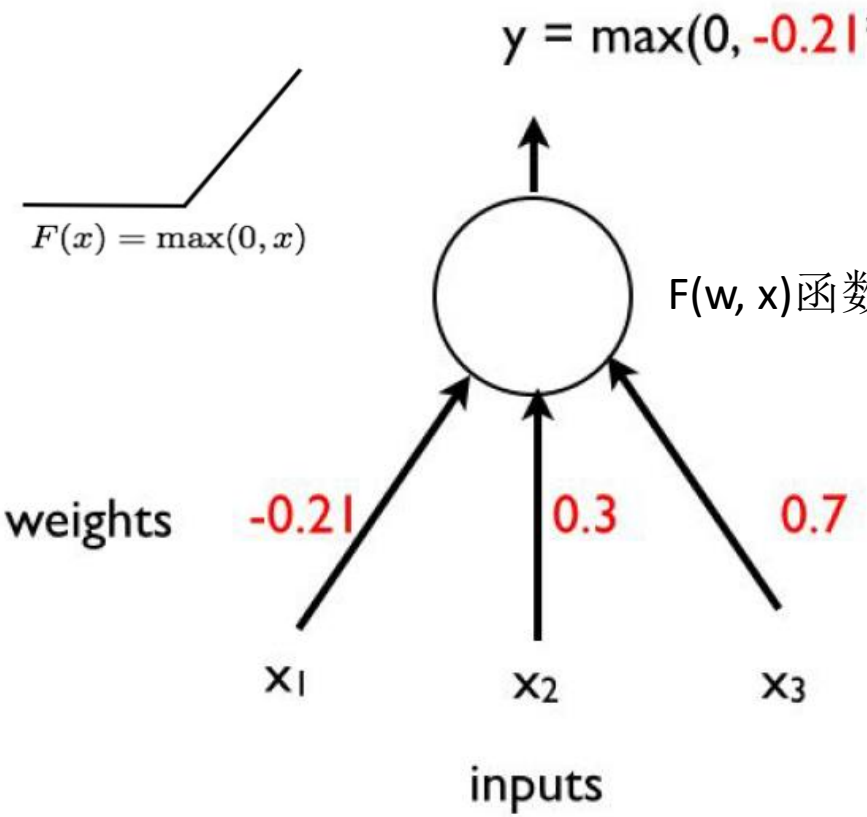
$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

- ReLU函数（rectified linear units），又称为整流线性单元，数学表达式如下：

$$ReLU(x) = max(x, 0)$$

参见python-tensorflow-basics

人工神经元1-ReLU单元（整流线性单元）

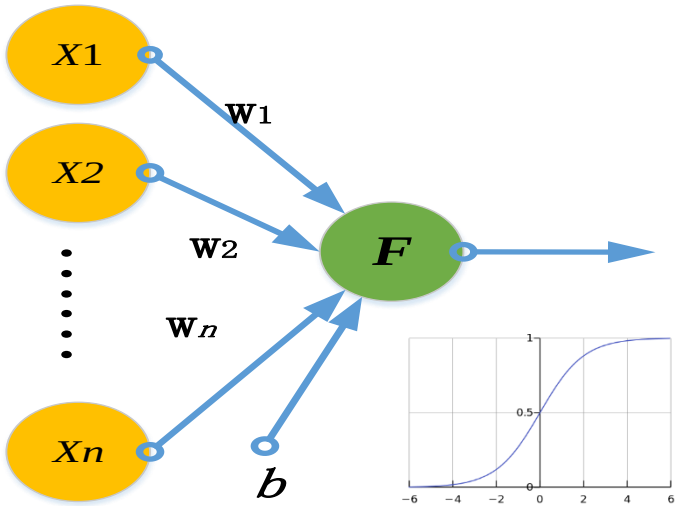


F(w, x)函数通过调节权重w，改变输出结果

- 所有输入加权叠加
- 经过一个非线性函数（激活函数）

人工神经元2-逻辑斯提回归单元

所有输入线性加权叠加，再经过一个非线性函数（激活函数）



$$Y = F\left(\sum_{i=1}^{N} W_i X_i + b\right)$$

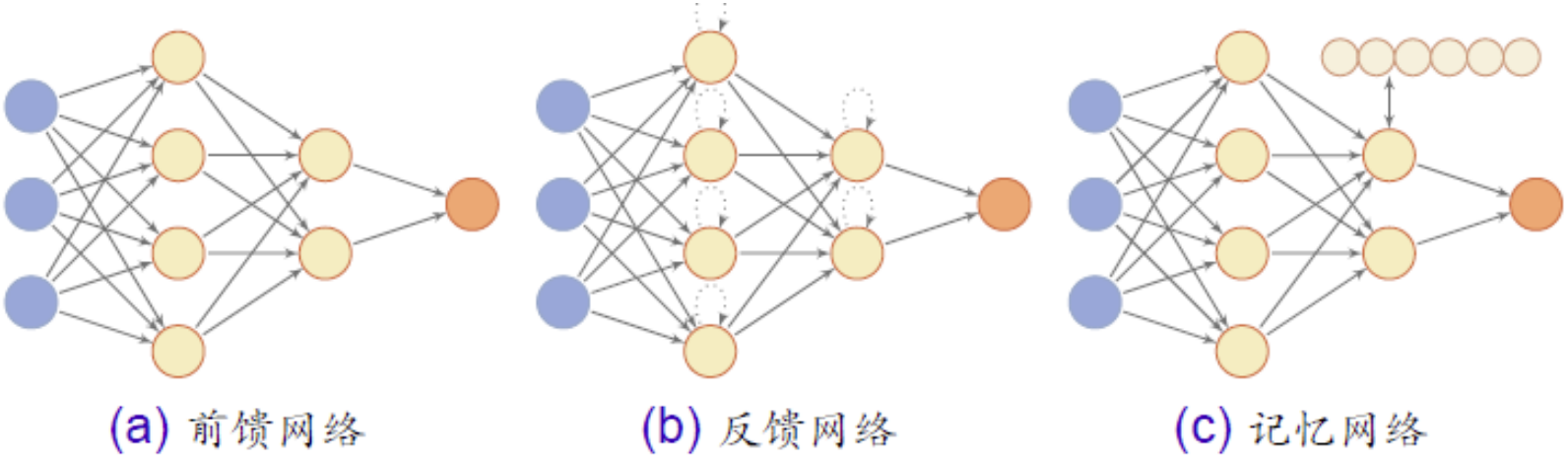
- 逻辑斯提回归单元(Logistic Regression Unit)是最简单人工神经元结构之一。
- 逻辑斯提回归单元的激活函数采用sigmoid函数或逻辑斯提函数

连接主义（Connectionism）方法

- 连接主义Connectionism：将大量的计算单元连接在一起的网络，可以实现复杂的智能行为。
- 神经元的功能可以数学建模出来，称为人工神经元（Artificial Neuron）。
- 将大量人工的神经单元连接在一起的网络，可以实现复杂的智能行为。

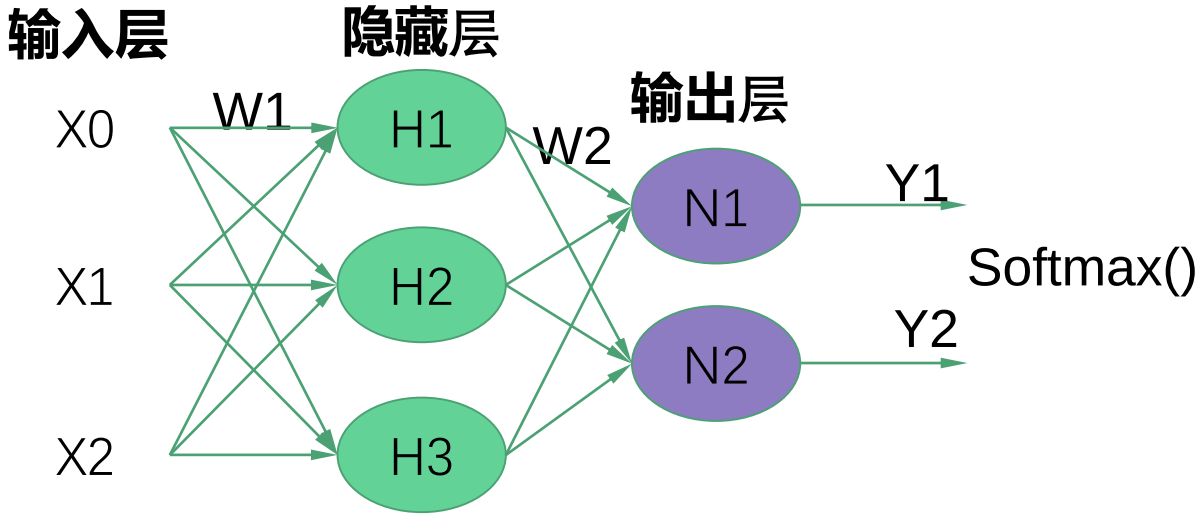
人工神经网络

- 按照连接主义观点：人工神经网络由大量的神经元以及它们之间的有向连接构成。
- 网络的拓扑结构
  - 不同神经元之间的连接关系。
  - 前馈网络（feedforward）、反馈网络（feedback）和记忆网络（memory network)



多层前馈网络（Multilayer feedforward networks）

- 前馈网络结构（feedforward）是一种计算图（Computational Graph）
- 别名：多层全连接网络（FCN）、多层感知机（MLP）
- 神经网络有3个输入，2个输出，中间有1个隐藏层，有1个输出层，共有5个神经元。



logit

- 分对数（logit）

$$\sigma(x) = \frac{1}{1 + \exp(-x)}.$$

$$\forall x \in (0, 1), \sigma^{-1}(x) = \log\left(\frac{x}{1 - x}\right)$$

Softmax处理

- 输出层的Softmax 处理，计算出一个概率分布:

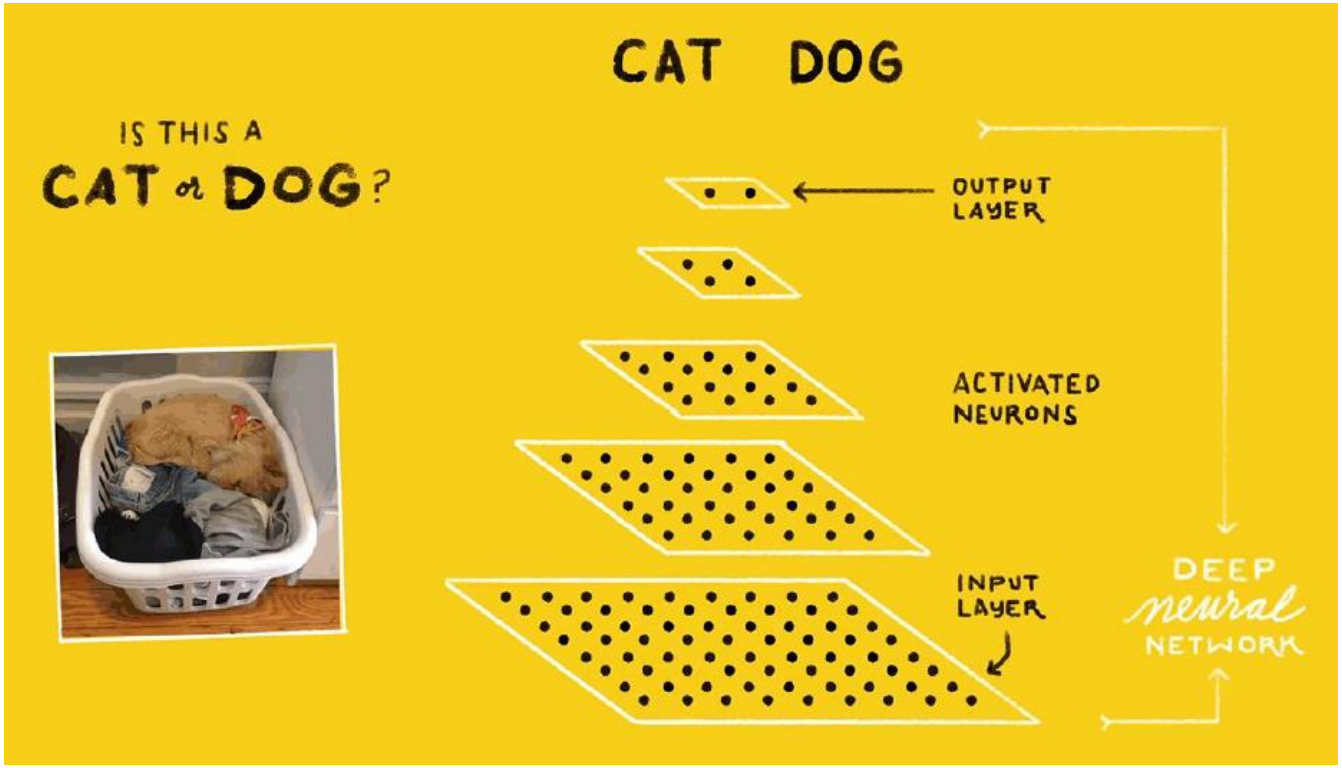
$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}$$

- 比如，输入为[1,2,3,4,5,6,7,8,9]，则softmax输出为[0.0,0.001,0.002,0.004,0.012,0.031,0.086,0.233,0.632]
- 所有分量之和为 1,所有输出的数值是正的。

参见python-tensorflow-basics

神经网络识别示例

- Cat or Dog
- https://www.kaggle.com/c/dogs-vs-cats/data



拉普拉斯定律

- 法国数学家皮埃尔·西蒙·拉普拉斯（Pierre Simon laplace）
- 1774年，在完全不知道贝叶斯的工作（Bayes' theorem）的情况下，拉普拉斯发表了一篇名为“事件原因的概率论”论文。
- 如果买n 张彩票共w 张中奖，那么中奖率就是中奖数加1，除以所购买的数目加2，即（w+1）/（n+2）。
- 这种令人难以置信的简单方法，估计概率的简单方法被称为拉普拉斯定律，它很容易就能适用于任何你需要通过历史事件来评估概率的情况。
- 相信太阳明天会升起是有道理的，这句话告诉我们：地球已经连续看到太阳上升约1.6 万亿天（45亿年），在下一次的“尝试”中看见太阳不升起来的机会，几乎没有可能。

[1] Christian, Brian, and Tom Griffiths. Algorithms to live by: The computer science of human decisions. Macmillan, 2016.

https://en.wikipedia.org/wiki/Rule_of_succession

深度学习Deep Learning

数学理论基础

深度学习的本质（数学）

- 深度学习的数学基础：微积分，线性代数和概率论
- 基本数学概念：
  - 映射（ Mapping ）：两个集合之间的关系，如满射、单射、一一对应即双射（满射和单射）
  - 函数（Function）： 定义x, y 变量，对于每个x数值，按照一定法则总有确定的数值y和它对应，则称y为x的函数，x叫自变量，y叫因变量
- 函数逼近问题：泛函分析

万能近似器(Universal Approximators)

- 万能近似定理（universal approximation theorem）
  - 多层前馈网络提供了一种万能近似框架
  - 如果具有线性输出层和至少一层具有任何一种‘‘挤压’’性质的激活函数（例如logistic sigmoid激活函数）的隐藏层，只要给予网络足够数量的隐藏单元
  - 给定一个函数，存在一个前馈网络能够近似该函数
- 以任意的精度来近似任何从一个有限维空间到另一个有限维空间的任意连续函数（ Borel 可测函数）
  - 前馈网络的导数也可以任意好地来近似函数的导数。
  - 可以近似从任何有限维离散空间映射到另一个的任意函数

理论与实际中存在的问题

- 万能近似定理
  - 无论试图学习什么函数，存在一个大的MLP 一定能够表示这个函数。
- 理论上
  - 存在一个足够大的网络能够达到我们所希望的任意精度，但是定理并没有说这个网络有多大。
- 实际中
  - 训练算法并不能保证能够学得这个函数。即使MLP能够表示该函数，学习也可能因两个不同的原因而失败。
  - 用于训练的优化算法可能找不到用于期望函数的参数值。
  - 训练算法可能由于过拟合而选择了错误的函数。

神经网络训练-原理

- 采用带标签的训练样本对神经网络进行学习，确定网络的权重参数
  - 数据集：(x1,y1), (x2,y2), (x3,y3), …,
- 神经网络输出
  - 实际输出：(x1, y1’), (x2, y2’), (x3, y3’), …,
  - 其中：Y’为实际输出
- 预期输出与实际输出的差异：
  - Y为预期输出（标签）
  - 标签与预测集：(y1, y1’), (y2, y2’), (y3, y3’), …,
- 差异最小化：
  - 度量函数：绝对值求和，平方和，交叉熵等
  - 目标优化：min{(|y1-y1’|+ |y2-y2’|+ |y3-y3’|+…)}
- 训练的本质，就是找到相应的内部权重，使得在训练数据（样本）输入到网络后，网络的实际输出与预期输出（即标签）之间差异最小。

神经网络训练-损失函数

- 神经网络的预期输出结果会与所对应的实际标签之间的存在差别。
- 用度量函数来表示这种差异，称为损失函数(Loss)或成本函数(Cost)。
- 损失函数的具体方式：
  - 对于回归任务，通过均方误差的公式，来计算损失。
  - 对于分类任务，通过交叉熵的公式，来计算损失(Loss)。
- 训练过程是损失函数 $H_{y'}(y) = -\sum_i y'_i \log(y_i)$ 学习的问题转化为最优化问题。

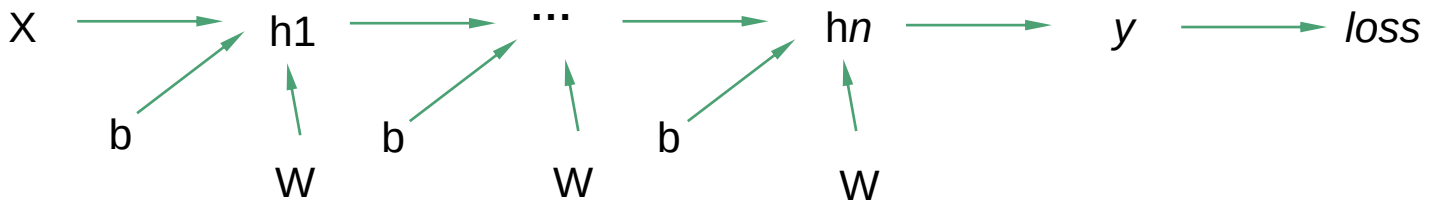
交叉熵

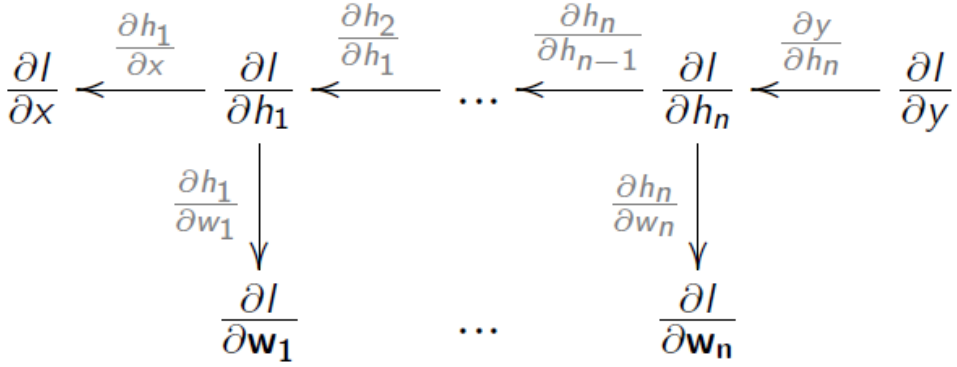
- 交叉熵损失函数表示如下：其中y’表示训练样本对应的标签，y表示神经网络的输出。
- 交叉熵是负对数似然损失函数，主要用于分类任务
- 比如，MNIST手写字体识别，0的向量[1,0,0,0,0,0,0,0,0,0]，网络输出的概率向量是[0.9, 0.5,0.5,0,0,0,0,0,0,0]

$$H_{y'}(y) = -\sum_i y'_i \log(y_i)$$

反向传播算法（backprop）

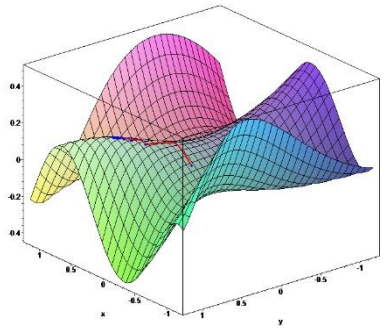
- 根据损失函数的性质以及链式求导法则，反向逐层计算损失函数对权重的梯度（各个偏导数），这个过程称为反向传播算法。
- 多层神经网络的一种更加抽象表示如下图所示，这种表示抽象了每层基本特征。

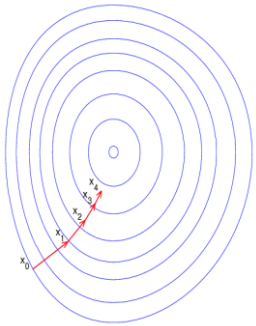




梯度下降法（Gradient descent）

- 梯度下降法也称为最速下降法，是一个最优化算法。
- 梯度的定义：
  - 函数F(x)在某个点上增长变化率最大的方向，就是导数的方向。
- 梯度下降法的原理：
  - 如果实值函数F(x)在点 a 处可微且有定义，那么函数F(x) 在a 点沿着梯度相反的方向下降最快。
- 梯度下降法的过程：
  - 找到一个函数的局部极小值，必须向函数上当前点对应梯度的反方向，按照规定步长距离点，进行迭代搜索。





梯度下降法-参数值更新（梯度算子）

- 在原值θ的基础上，沿着梯度方向的步长，这样就得到了一个新的θ数值，使得损失函数变化的最快
- 当不断调整θ数值，Loss(θ)是一个随着θ取值的序列series，这个序列的数值不再有大的变化，就说明收敛了。

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta).$$

神经网络的实际训练过程

- 神经网络的训练将样本数据“分批训练”，
- 最简单的批量选择是使用整个数据集，又称为批量训练（batch training）。
- 优点：
  - 批量训练的收敛性是有保证的。
  - 共轭梯度法和L-BFGS加速技术在批量训练中效果显著
- 缺点：
  - 是模型参数的更新前需要遍历整个数据集

随机梯度下降法

- 随机梯度下降方法(stochastic gradient descent, SGD)是最常用的权重调节方法，通过权重的调整，最小化损失函数。
- 随机梯度下降法步骤如下：
  - 步骤 1. 随机初始化每个神经元输入权重和偏差；
  - 步骤 2. 选取一个随机样本；
  - 步骤 3. 根据网络的输出结果，从最后一层开始后，逐层计算每层权重的偏导数；
  - 步骤 4. 逐层调整每层的权重，产生新的权重值。
  - 返回到步骤2，继续随机选取下一个样本。
- 随机梯度下降法的核心是一个“随机样本”

神经网络的实际训练过程

- 小批量训练是批量训练和随机梯度下降法的一个折中
  - 整个训练集称为一个批次（batch），先将整个训练集分成多个大小相同的子集，每个子集称为一个迷你批次（mini-batch）。子集的大小由参数迷你批次大小（mini-batch_size）控制。
  - 每个批次的数据被依次送入网络进行训练。训练完一个迷你批次， 被称为一次迭代（iteration）。
  - 一个时代（epoch）是指训练集中所有训练样本都被送入网络，完成一次训练的过程。
- 其中每一步的模型参数的更新使用
  - 使用不止一个样本，这称为迷你批次（minibatch）。
  - 每次迭代所用的样本数目称为迷你批次大小（minibatch size）。
  - 当迷你批次大小为1时，就退化为普通的随机梯度下降。

深度学习Deep Learning

数值计算方法实现

数值计算-数值量化

- 数值分析（Numerical analysis）是研究各种数学问题求解的数值计算方法
  - 数值逼近、数值微分和积分、非线性方程求解等
- 计算机是数值计算的基本工具，数值分析也叫计算方法
  - 计算机字长是固定，无法表示如无理数之类的数值，需要对数值进行量化（ Quantization ）
  - 单精度single，双精度double，浮点数float，64位整型int64，32位整型int32
- [x]教材：李庆扬等，数值分析，华中理工大学出版社，1982.

数值计算的误差来源

- 模型误差：
  - 数学模型与实际问题之间出现的这种误差叫模型误差
- 观测误差：
  - 数学模型中往往会有一些观察的物理量，这些参量也包含误差，这种有观测产生的误差叫观测误差
- 方法误差或截断误差：
  - 数学模型不能得到精确解时，通常用近似解，近似解和精确解之间的误差称为方法误差
- 舍入误差：
  - 计算机字长有限，原始数据在计算机上表现会产生误差，计算机运算又会产生新误差

数值计算的基本原则

- 数值稳定：
  - 运算过程舍入误差不增长的计算公式称为数值稳定的。
- 数值计算的基本原则
  - 要避免除数绝对值远远小于被除数绝对值的除法
  - 要避免两相近数相减
  - 要防止大数“吃掉”小数
  - 简化计算步骤，减小运算次数
- 举例：softmax 函数
  - 当接近零的数被四舍五入为零时，发生下溢（underflow）。
  - 当大量级的数被近似为无穷大 或负无穷大时，发生上溢（overflow）。
  - 必须对上溢和下溢进行数值稳定的
- 参见python-tensorflow-basics

求导运算

- 数值微分(Numerical differentiation)
  - 根据导数定义来计算：
  - $f'(x)=\lim_{\Delta x\to 0}(f(x+\Delta x)-f(x))/\Delta x$
  - 从两侧逼近：$f'(x)=\lim_{\Delta x\to 0}(f(x+\Delta x)-f(x-\Delta x))/2\Delta x$
- 符号微分(Symbolic differentiation)
  - 符号微分是精确，其误差为0
  - 符号微分返回一个公式（计算图）
- *‘自动微分’ (Automatic differentiation)
  - ‘自动微分’直接返回f(x)的导数值f’(x)
  - ‘自动微分’不对运算进行符号化
  - 狭义地讲，‘自动微分’可以指使用二元数（dual number）进行自动导数运算的自动求导方式。

符号微分原理

- 从表达式出发，将表达式本身看做符号的运算，直接得到求导结果的表达式。
- 1. 统一定义基本运算Op
- 2. 统一定义初等函数的导数运算
  - 线性运算
  - 三角函数运算
  - 指数对数
  - …
- 3. 预定义四则运算求导：
  - d(u/v)=(udv−vdu)/v2d(u/v)=(udv−vdu)/v2
  - d(u∗v)=udv+vdu
  - d(u+v)=du+dv …
- 4. 预定义导数的链式法则：
  - dy/dx=dy/du∗du/dx

深度学习Deep Learning

基本网络结构

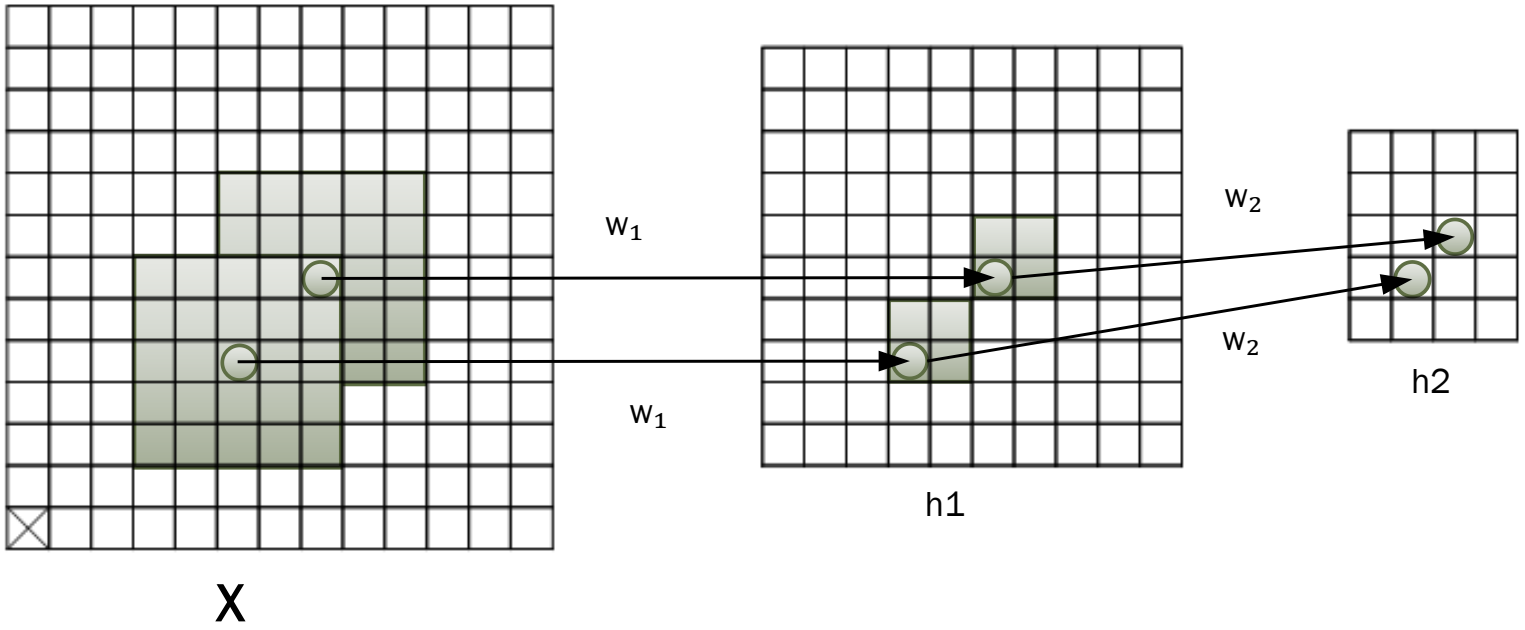
典型常用的网络结构

- 卷积网络CNN
- （Convolutional neural network，简称CNN）

- 循环网络RNN
- （Recurrent neural network，简称RNN）

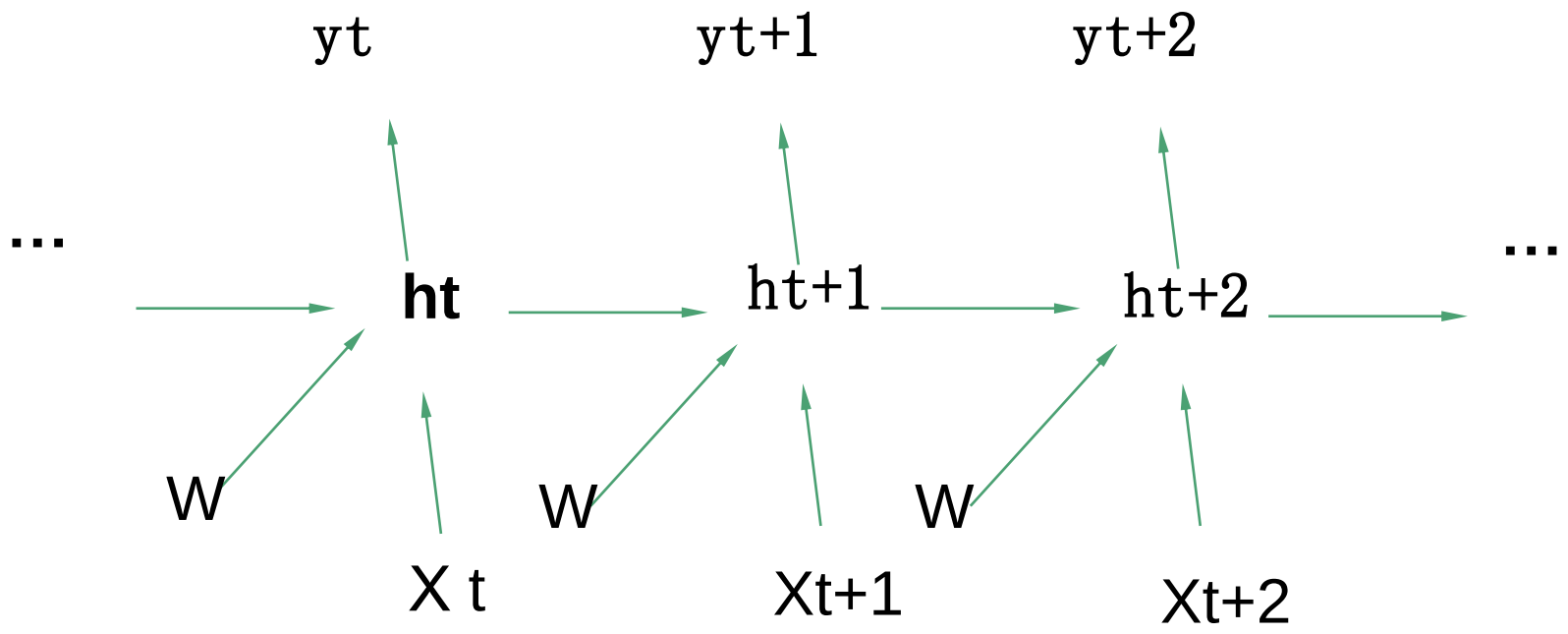
卷积网络

- 卷积网络（Convolutional neural network，简称CNN）
- 特点：局部区域的权重W共用（weight sharing）（空间维度）
- 每一个卷积层后通常紧跟着一个下采样层，比如采用max-pooling 方法完成下采样。



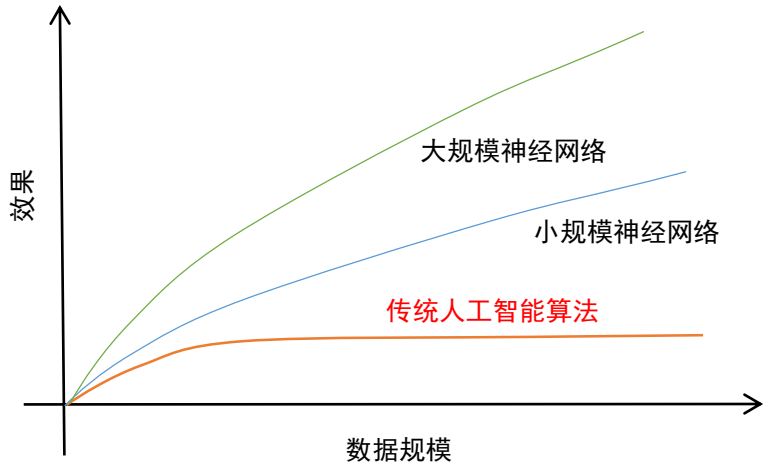
循环网络-RNN

- 循环网络（Recurrent neural network，简称RNN）的每一个时间步处理时，权重共享。（时间维度）
- 用于时间序列的预测

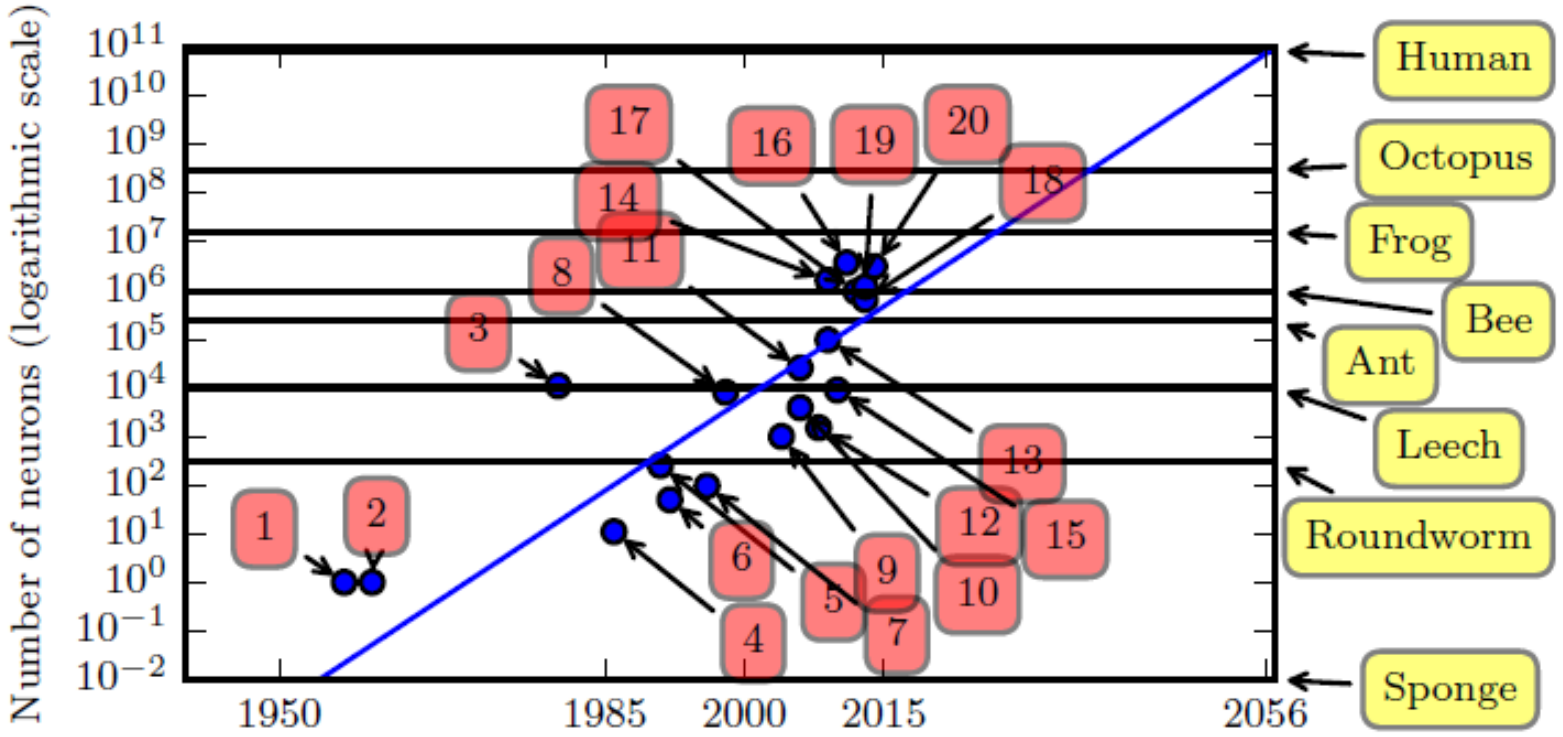


大数据是深度学习基础

- 传统机器学习方法主要涉及数据、模型和算法三个方面。
- 手工或人为的特征选取，随着训练数据规模的提高，传统机器学习方法提升效果并不明显
- 以深度学习为代表的方法，随着训练数据规模的提高，提升效果显著，大大超过了传统机器学习的方法。



人工神经网络规模



- 自从引入隐藏单元，人工神经网络的大小大约每2.4 年翻一倍。

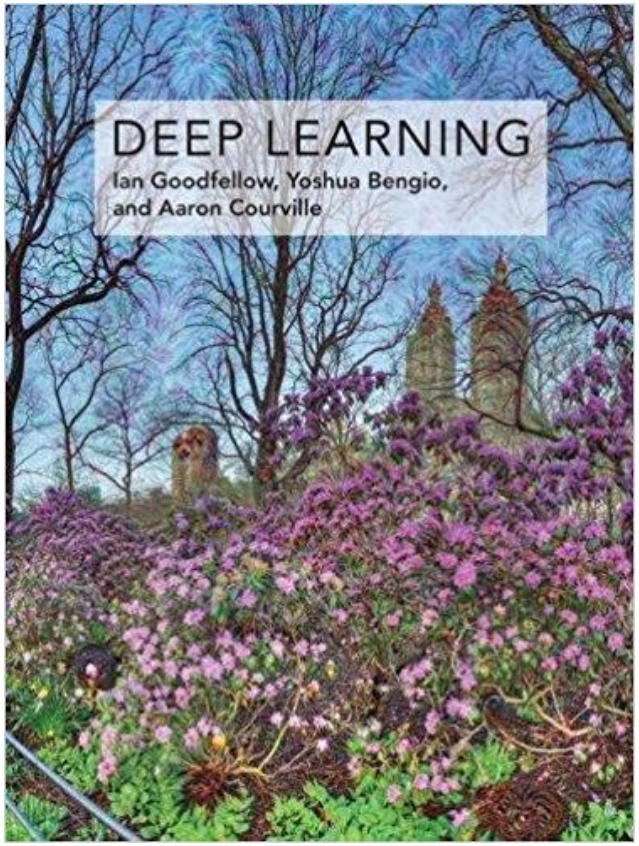
[x] Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, MIT Press, 2016.

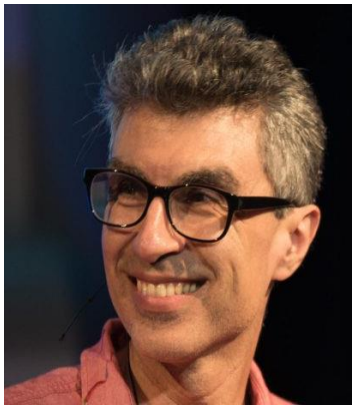
总结

- 机器智能技术是大数据平台+机器学习算法

- 深度学习引发了新的机器智能潮流

参考书





- [1] Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT press; 2016 Nov 18.
- [1] 中文版：深度学习 [deep learning]，人民邮电出版社，2017年8月.

谢谢！

zhenchen@Tsinghua.edu.cn