1. 导言

WU Xiaokun 吴晓堃

xkun.wu [at] gmail

2021/03/01

本章内容

人工智能概论。机器学习简史。深度学习兴起的原因。深度学习工作站的配置。测试深度学习的简单示例：MNIST手写数字分类。

重点：基本概念，机器学习的编程范式和深度学习算法的工作流程；

难点：深度学习工作站的配置，测试深度学习的简单示例。

学习目标

- 理解基本概念：人工智能，学习，机器学习，深度学习；
- 了解深度学习相关学科的历史发展过程，相关技术的更新迭代过程；
- 理解机器学习的编程范式和深度学习算法的工作流程；
- 理解深度学习兴起的技术前提（硬件、数据、算法）和有望长盛不衰的决定因素（简单、可扩展、可复用、可迁移）；
- 掌握深度学习工作站的配置方法，并能够独立测试深度学习的简单示例。

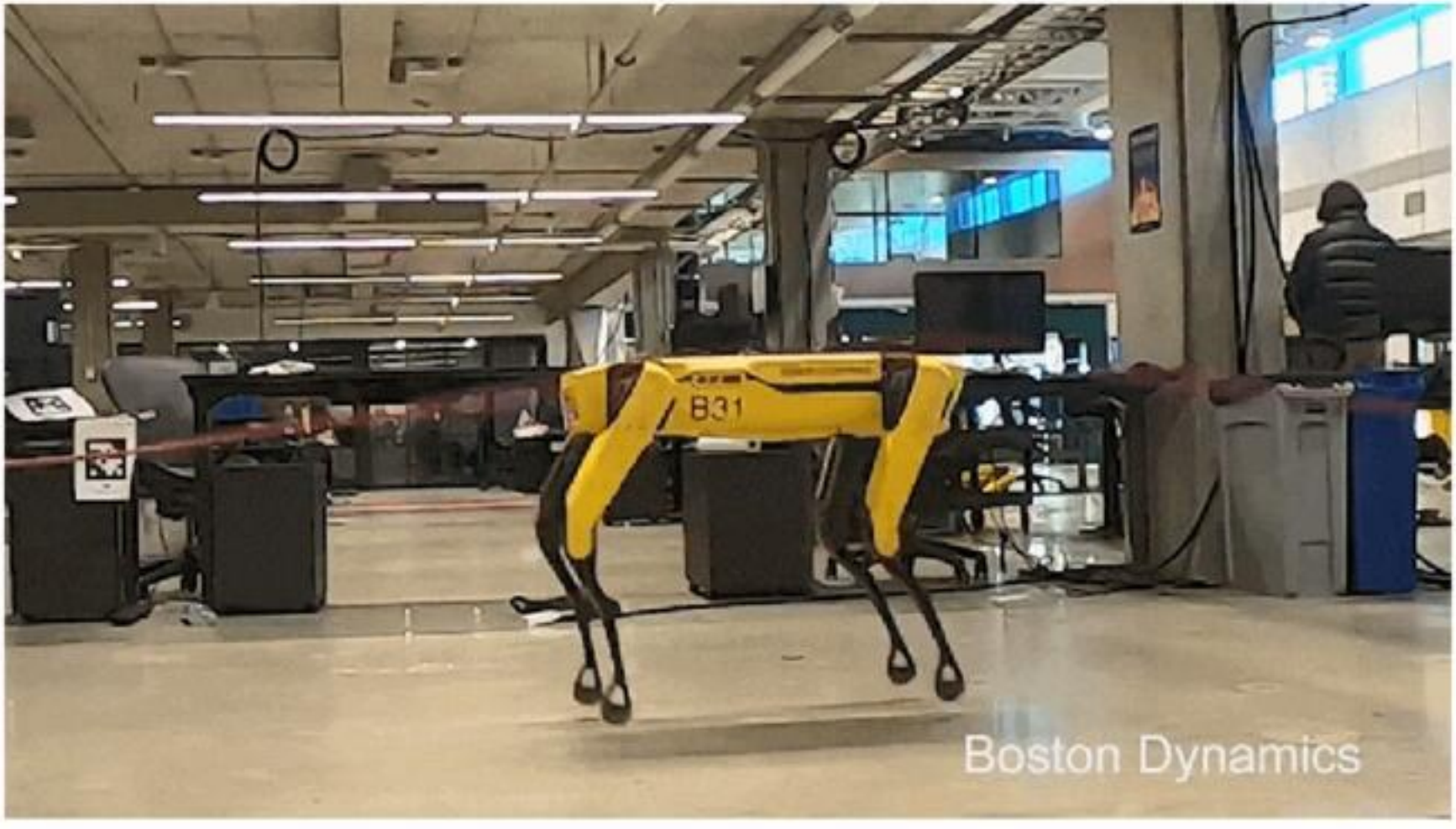
AI：科幻与现实

AI无所不在

底特律：化身为人

波士顿动力 - 拾取

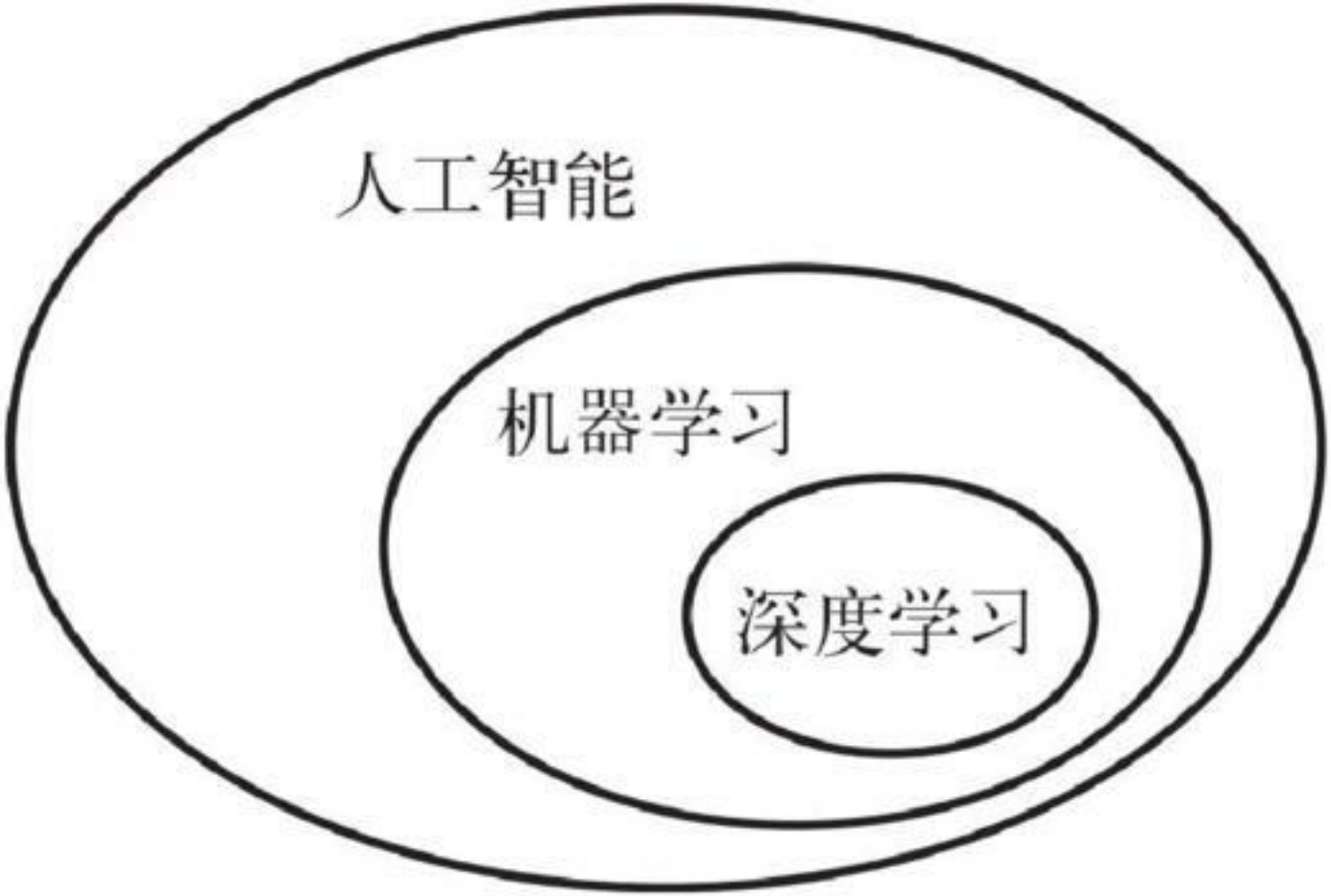
波士顿动力 - 跳绳



波士顿动力：双足

人工智能、机器学习与深度学习

三者关系



人工智能

人工智能诞生于20世纪50年代。

人工智能的简洁定义

将通常由人类完成的智力任务自动化。

人工智能

人工智能诞生于20世纪50年代。

人工智能的简洁定义

将通常由人类完成的智力任务自动化。

> *在相当长的时间内，许多专家相信，只要程序员**精心编写足够多的明确规则**来处理知识，就可以实现与人类水平相当的人工智能。*

- 符号主义人工智能（symbolic AI）：从 20 世纪 50 年代到 80 年代末
- 专家系统（expert system）：20 世纪 80 年代

符号主义人工智能

符号主义人工智能适合用来解决定义明确的逻辑问题，比如下象棋。

但难以给出明确的规则来解决复杂、模糊的现实问题

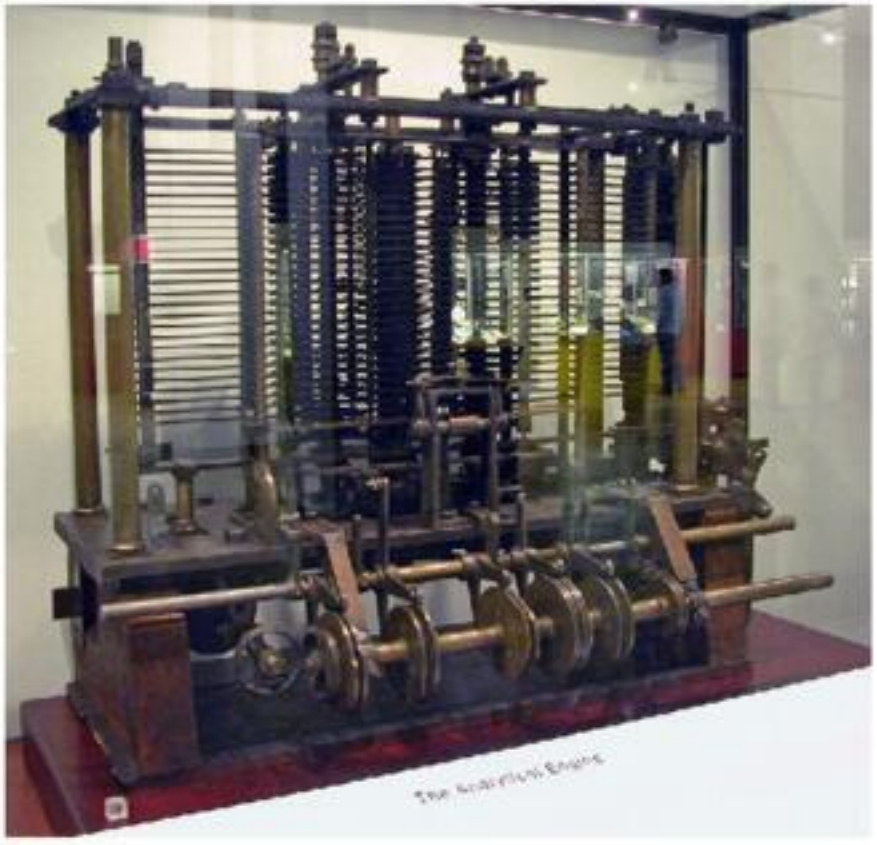
比如图像分类、语音识别和语言翻译。

于是人们研究出了一种新的方法来替代符号主义人工智能，这就是机器学习（machine learning）。

机器学习 - 史前时代

在维多利亚时代的英格兰，Charles Babbage发明了分析机（Analytical Engine），即第一台通用的机械式计算机。

> *“分析机谈不上能创造什么东西。它只能完成我们命令它做的任何事情……它的职责是帮助我们去实现我们已知的事情。”*
>
> *– Lovelace伯爵夫人Ada*



人工智能的里程碑：Turing测试

人工智能先驱Alan Turing在 1950 年发表论文《计算机器和智能》。

Turing测试

测试者与被测试者（一个人和一台机器）隔开的情况下，通过一些装置（如键盘）向被测试者随意提问。进行多次测试后，如果机器让平均每个参与者做出超过30%的误判，那么这台机器就通过了测试，并被认为具有人类智能。

机器学习与Turing

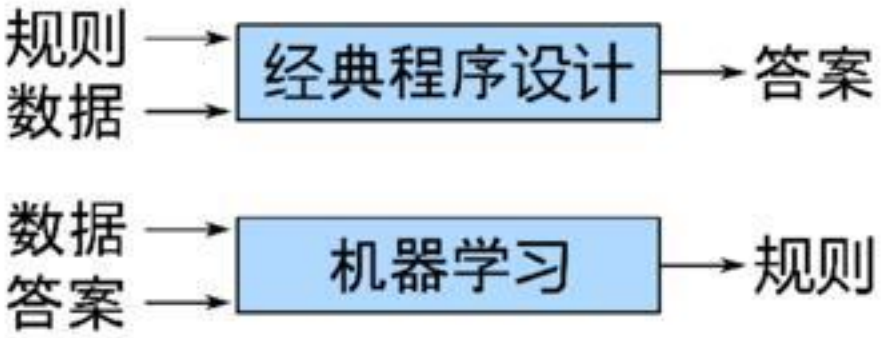
机器学习的概念来自于Turing引述 Lovelace伯爵夫人 Ada 的问题后的进一步思考：

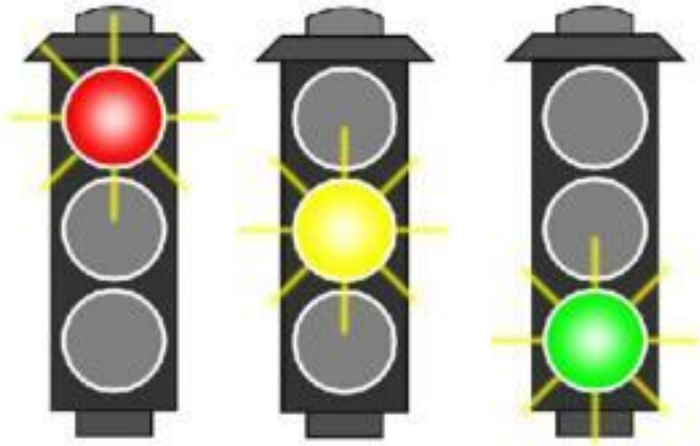
通用计算机是否能够学习与创新?

对于计算机而言，除了“我们命令它做的任何事情”之外，它能否自我学习执行特定任务的方法?

如果没有程序员精心编写的数据处理规则，计算机能否通过观察数据自动学会这些规则?

新的编程范式





注意：机器学习系统是训练出来的，而不是明确地用程序编写出来的。

举例：原始人穿越来到现在，如何学会交通规则？观察与试错。

机器学习

机器学习在 20 世纪 90 年代才开始蓬勃发展。

- 驱动力：速度更快的硬件与更大的数据集。
- 以工程为导向：想法更多地是靠实践来证明，而不是靠理论推导。

对比：经典的统计分析方法（比如贝叶斯分析）。

机器学习算法三要素

给定包含预期结果的示例，机器学习将会发现执行一项数据处理任务的规则。

- **输入数据点。**
- **预期输出的示例。**
- **衡量算法效果好坏的方法。**

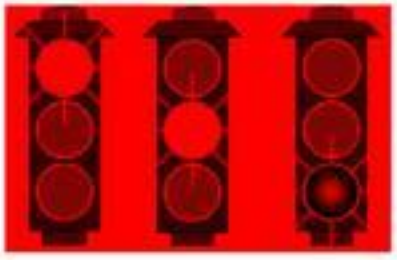
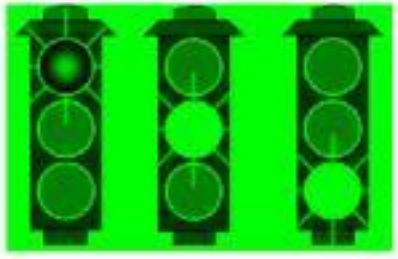
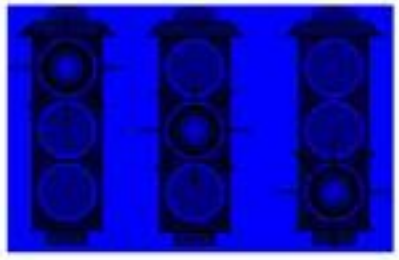
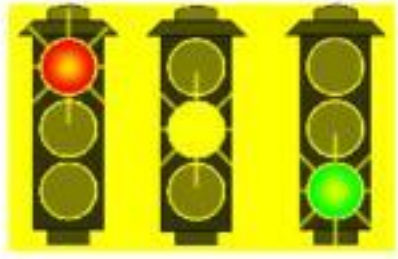
“学习”的简单理解

衡量结果是一种反馈信号，用于调节算法的工作方式。这个调节步骤就是我们所说的**学习**。

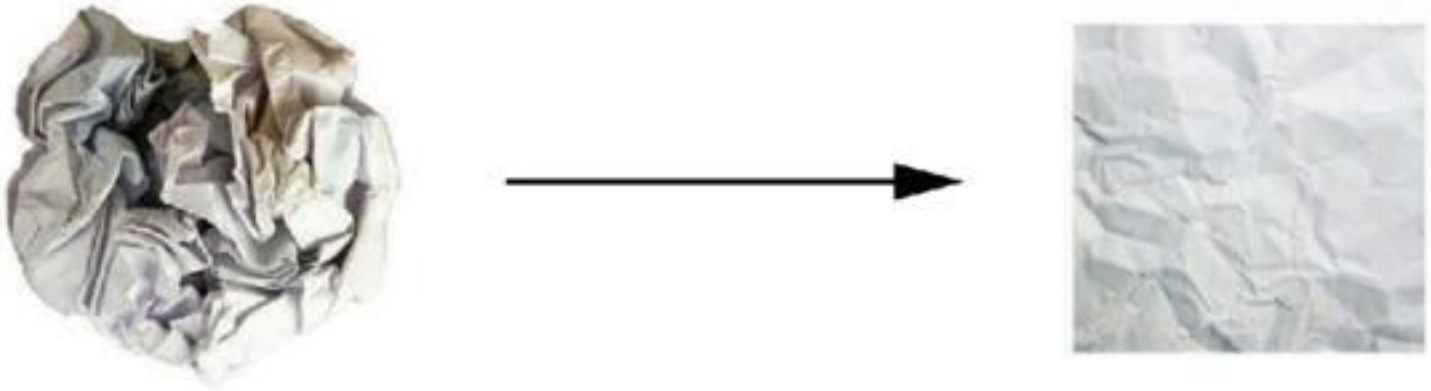
机器学习算法的核心问题

机器学习模型将输入数据变换为有意义的输出，这是一个从已知的输入和输出示例中进行“学习”的过程。

机器学习和深度学习的核心问题在于有意义地变换数据，换句话说，在于学习输入数据的有用表示（representation）——这种表示可以让数据更接近预期输出。

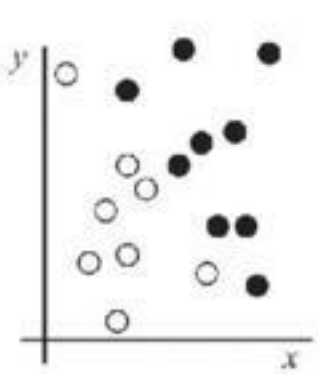





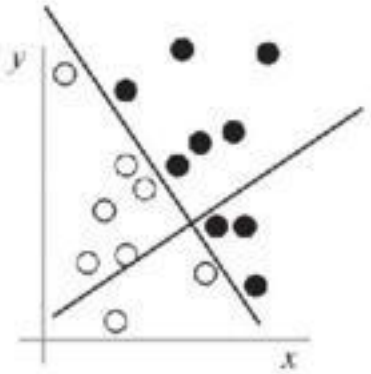
数据变换

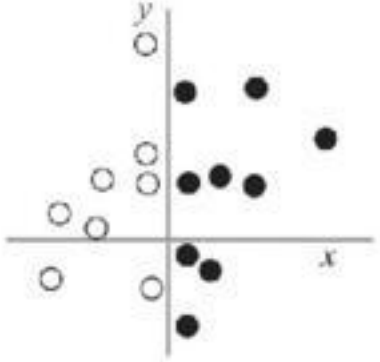


- 两张纸代表了不同的类别；
- 解开纸团的过程就是学习的过程：
  - 展开纸张的操作就是数据变换。
  - 寻找更好的展开状态就是寻找更好的数据表示。

黑白点划分问题







- 输入是点的坐标；
- 预期输出是点的颜色；
- 衡量算法效果好坏的一种方法是，正确分类的点所占的百分比。

問題：我们人为定义了坐标变换。

机器学习 - 学习与假设空间

学习

寻找更好数据表示的自动搜索过程。

假设空间

机器学习算法在解决问题时通常没有什么创造性，而仅仅是遍历一组预先定义好的操作集合，这个操作集合叫作假设空间（hypothesis space）。

机器学习 - 技术定义

在预先定义好的可能性空间中，利用反馈信号的指引来寻找输入数据的有用表示。

深度学习之“深度”

深度学习

它是从数据中学习表示的一种新方法，强调从连续的层（layer）中进行学习，这些层对应于越来越有意义的表示。

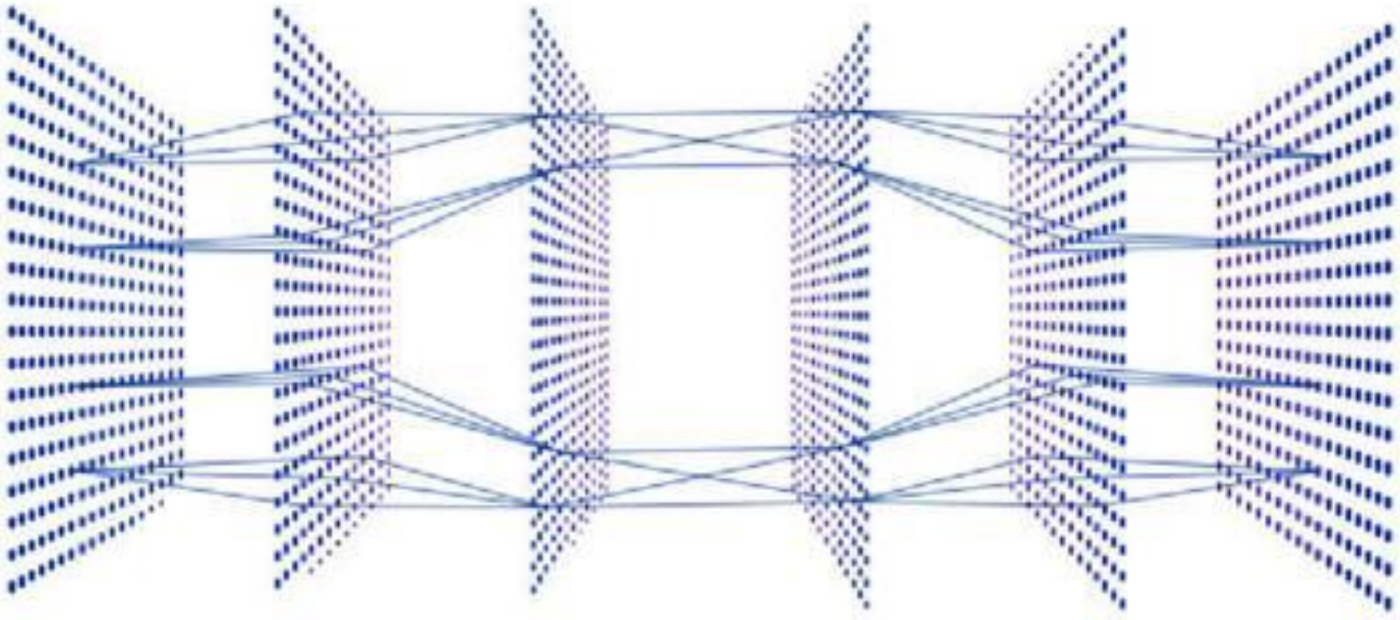
注意

“深度学习”中的“深度”指的并不是利用这种方法所获取的更深层次的理解，而是指一系列连续的表示层。

深度学习 - 层、深度

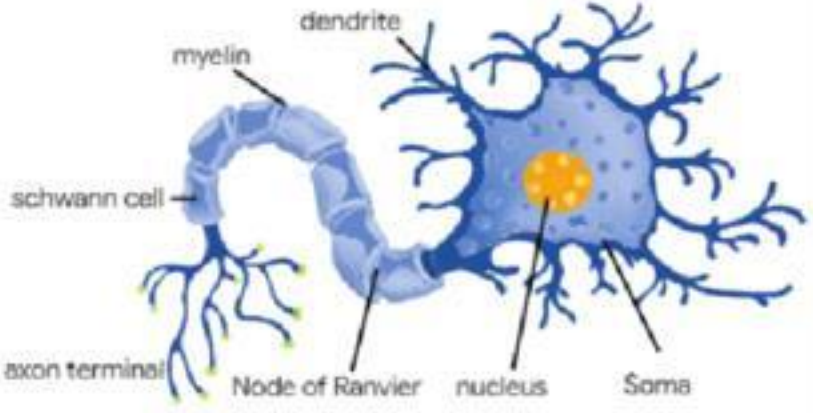
层、深度

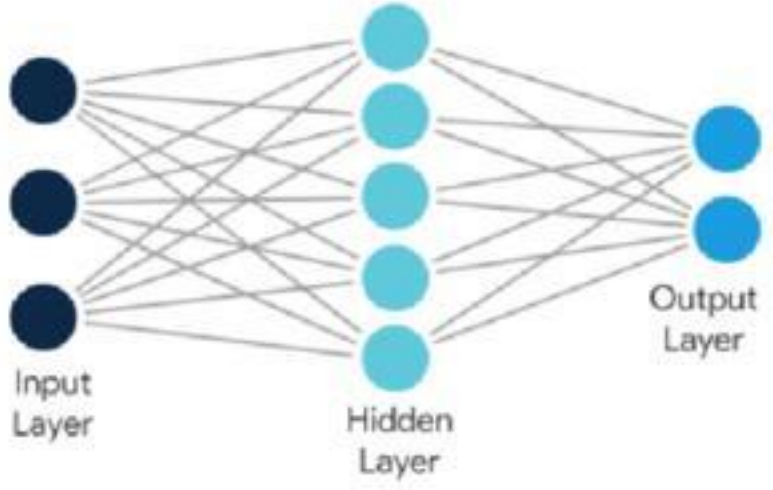
数据模型中包含多少层，这被称为模型的深度（depth）。



深度学习 - 神经网络（neural network）

神经网络这一术语来自于神经生物学，然而，虽然深度学习的一些核心概念是从人们对大脑的理解中汲取部分灵感而形成的，但没有证据表明大脑的学习机制与现代深度学习模型所使用的相同。

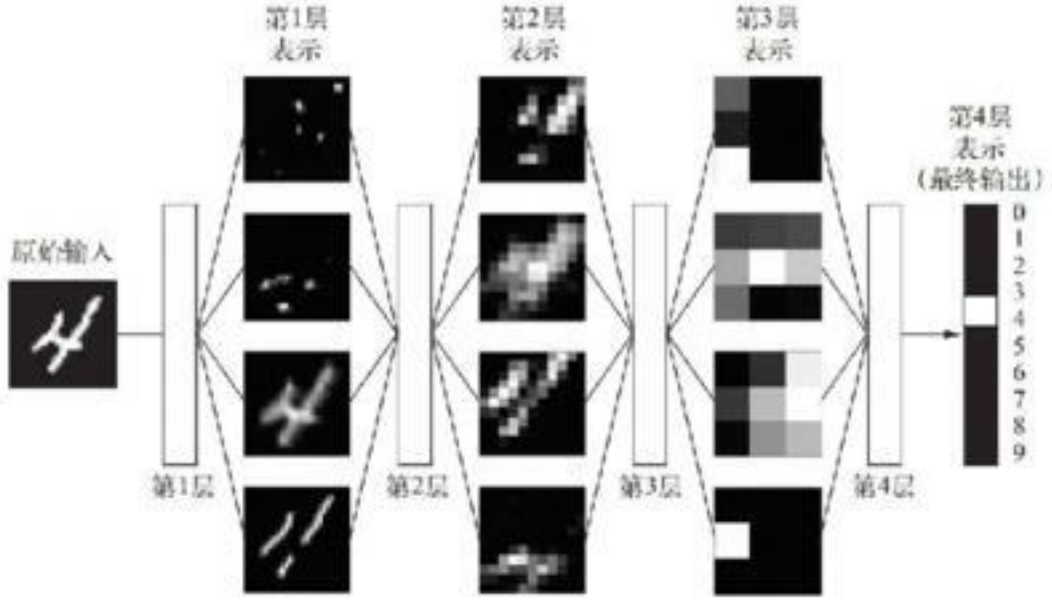




就表现形式而言，深度学习是从数据中学习表示的一种数学框架。

深度学习算法学到的表示是什么样的?

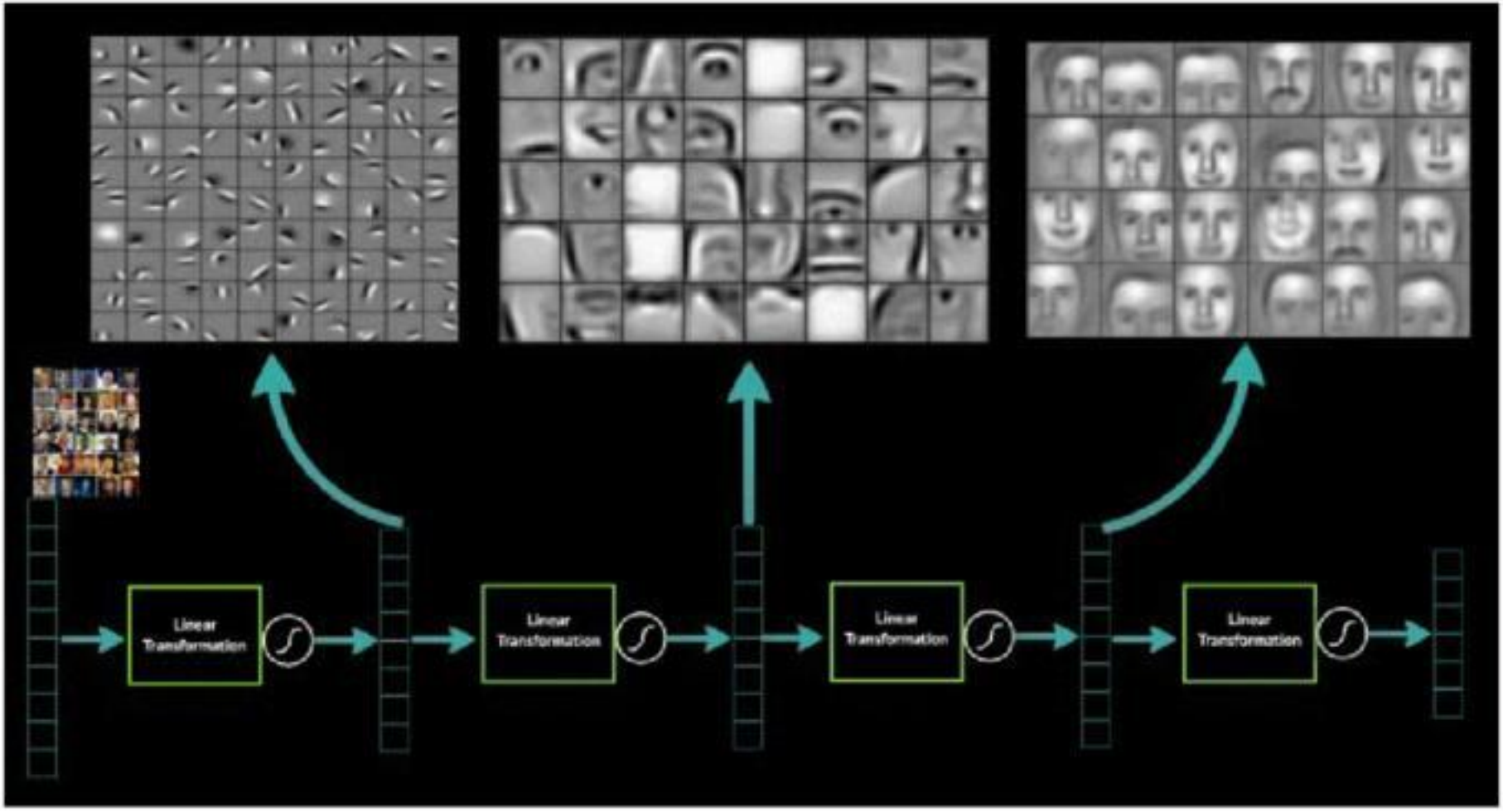
你可以将深度网络看作多级信息蒸馏操作：信息穿过连续的过滤器，其纯度越来越高（即对任务的帮助越来越大）。



深度学习 - 技术定义

学习数据表示的多级方法。

CNN 示意图


Linear Transformation
Linear Transformation
Linear Transformation
Linear Transformation

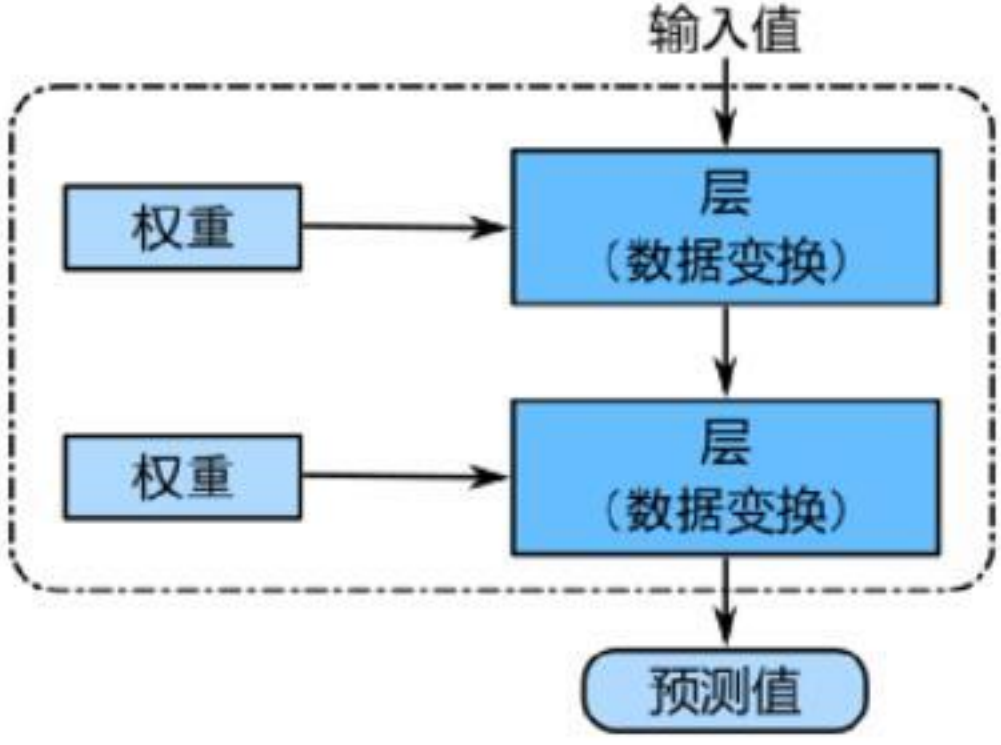
小结

- 机器学习：是将输入（比如图像）映射到目标（比如标签“猫”），这一过程是通过观察许多输入和目标的示例来完成的。
- 深度神经网络通过一系列简单的数据变换（层）来实现这种输入到目标的映射，而这些数据变换都是通过观察示例学习到的。

下面用三张图来具体看一下深度学习的过程是如何发生的。

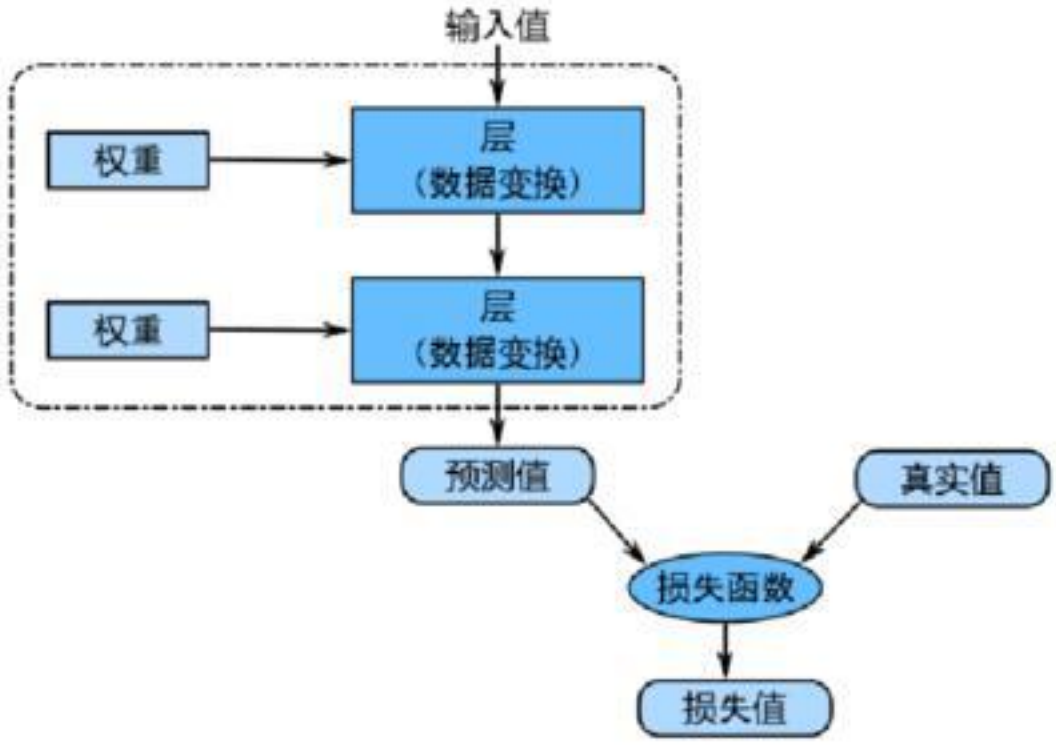
神经网络是由其权重来参数化

神经网络中每层对输入数据所做的具体操作保存在该层的权重（weight）中，其本质是一串数字，有时也被称为该层的参数（parameter）。



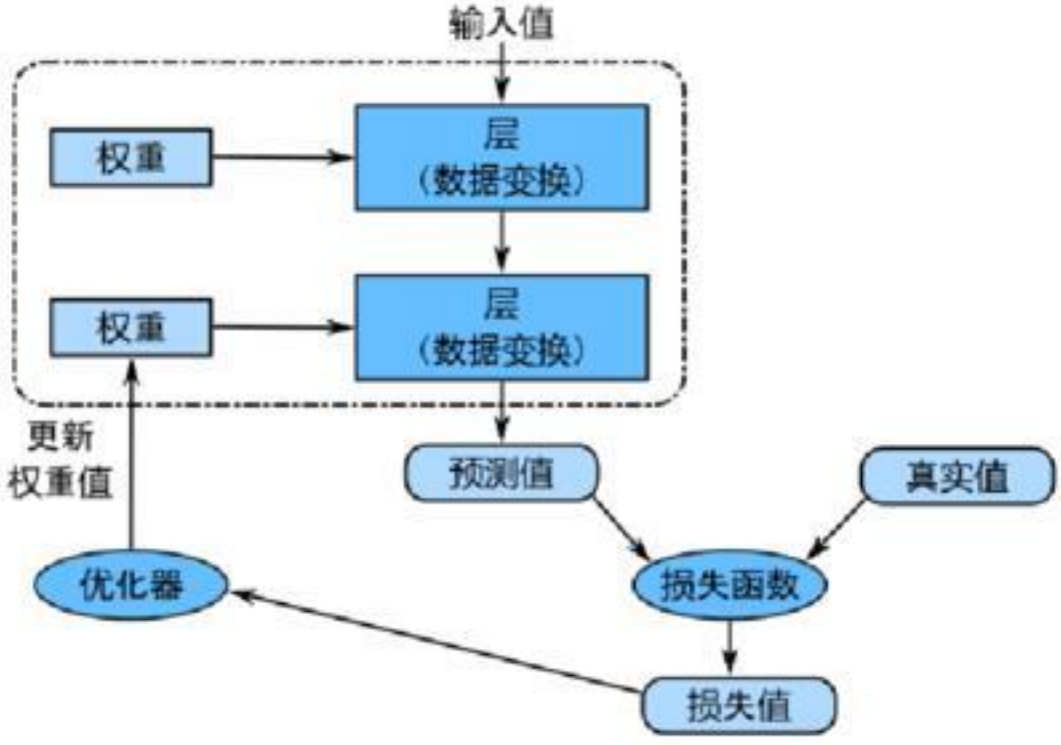
损失函数用来衡量网络输出结果的质量

想要控制神经网络的输出，就需要能够衡量该输出与预期值之间的距离。 这是神经网络损失函数（loss function）的任务，该函数也叫目标函数（objective function）。



将损失值作为反馈信号来调节权重

深度学习的基本技巧是利用这个距离值作为反馈信号来对权重值进行微调，以降低当前示例对应的损失值。这种调节由优化器（optimizer）来完成，它实现了所谓的反向传播（backpropagation）算法，这是深度学习的核心算法。



量变导致质变

一开始对神经网络的权重随机赋值，因此网络只是实现了一系列随机变换。 其输出结果自然也和理想值相去甚远，相应地，损失值也很高。

但随着网络处理的示例越来越多，权重值也在向正确的方向逐步微调，损失值也逐渐降低。 这就是训练循环（training loop），将这种循环重复足够多的次数（通常对数千个示例进行数十次迭代），得到的权重值可以使损失函数最小。

深度学习已经取得的进展

- 接近（或超过）人类水平的图像分类
- 接近人类水平的语音识别
- 接近人类水平的手写文字转录
- 更好的机器翻译
- 更好的文本到语音转换
- 数字助理，比如谷歌即时（Google Now）和亚马逊 Alexa
- 接近人类水平的自动驾驶
- 更好的广告定向投放，Google、百度、必应都在使用
- 更好的网络搜索结果
- 能够回答用自然语言提出的问题
- 在围棋上战胜人类

不要相信短期炒作

我们尤其不应该把达到人类水平的通用智能（human-level general intelligence）的讨论太当回事。 在短期内期望过高的风险是，一旦技术上没有实现，那么研究投资将会停止，而这会导致在很长一段时间内进展缓慢。

两次人工智能冬天（AI winter）：

- 20 世纪 60 年代的符号主义人工智能
- 20 世纪 80 年代的符号主义人工智能——专家系统（expert system）

我们可能正在见证人工智能炒作与让人失望的第三次循环，而且我们仍处于极度乐观的阶段。 最好的做法是降低我们的短期期望，确保对这一技术领域不太了解的人能够清楚地知道深度学习能做什么、不能做什么。

人工智能的未来

过去五年里，人工智能研究一直在以惊人的速度发展，这在很大程度上是由于人工智能短短的历史中前所未见的资金投入，但到目前为止，这些进展却很少能够转化为改变世界的产品和流程。

但不要怀疑：人工智能的时代即将到来。人工智能最终将应用到我们社会和日常生活的几乎所有方面，正如今天的互联网一样。

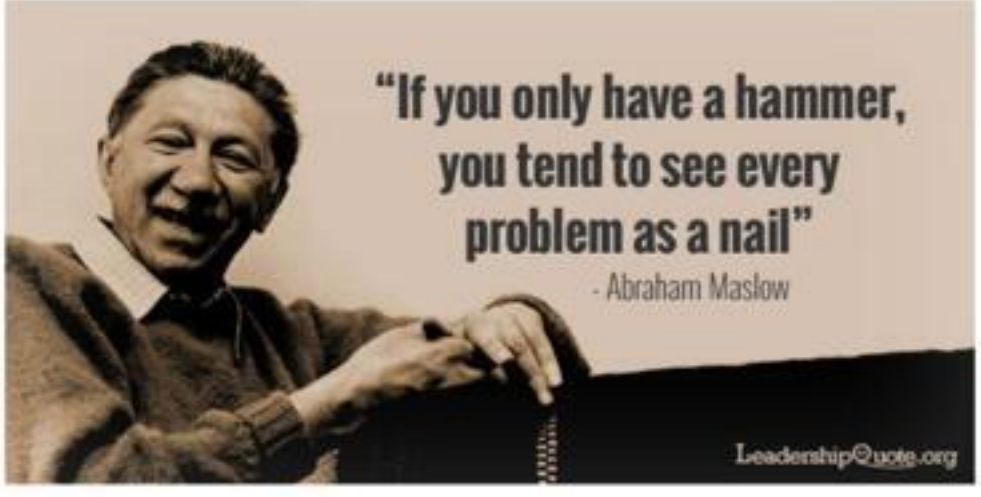
深度学习之前：机器学习简史

为什么要了解经典机器学习方法?

深度学习并不是机器学习的第一次成功。 可以这样说，当前工业界所使用的绝大部分机器学习算法都不是深度学习算法。

锤子 - 钉子

如果你第一次接触的机器学习就是深度学习，那你可能会发现手中握着一把深度学习“锤子”，而所有机器学习问题看起来都像是“钉子”。



概率建模（probabilistic modeling）

概率建模是统计学原理在数据分析中的应用，在统计学习课程中专门讲授。 它是最早的机器学习形式之一，至今仍在广泛使用。

- 其中最有名的算法之一就是朴素贝叶斯（Naive Bayes）算法，比计算机出现得还要早，在其第一次被计算机实现（很可能追溯到 20 世纪 50 年代）的几十年前就已经靠人工计算来应用了。
- 另一个密切相关的模型是 logistic 回归（logistic regression，简称 logreg），它有时被认为是现代机器学习的“hello world”。 不要被它的名称所误导——logreg 是一种分类算法，而不是回归算法。

早期神经网络

神经网络早期的迭代方法已经完全被本章所介绍的现代方法所取代。

- 人们早在 20 世纪 50 年代就将神经网络作为玩具项目，但在很长一段时间内，一直没有训练大型神经网络的有效方法。
- 在 20 世纪 80 年代中期很多人都独立地重新发现了反向传播算法，并开始将其应用于神经网络。

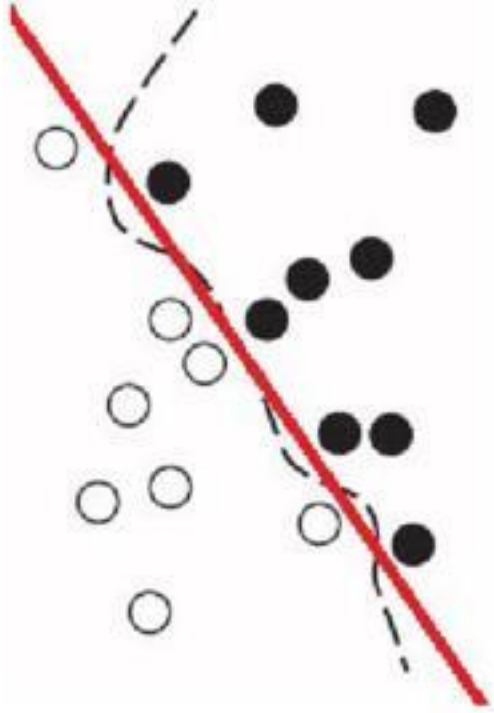
贝尔实验室于 1989 年第一次成功实现了神经网络的实践应用，当时 Yann LeCun 将卷积神经网络的早期思想与反向传播算法相结合，并将其应用于手写数字分类问题，由此得到名为LeNet 的网络，在 20 世纪 90 年代被美国邮政署采用，用于自动读取信封上的邮政编码。

核方法（kernel method）

SVM 刚刚出现时，在简单的分类问题上表现出了最好的性能，很快就使人们将神经网络抛诸脑后。

SVM 的目标是通过找到良好决策边界（decision boundary）来解决分类问题，这个过程分为两步：

1. 将数据映射到高维表示，这时决策边界可以用超平面来表示。
2. 尽量让超平面与每个类别最近的数据点之间的距离最大化，这一步叫作间隔最大化（maximizing the margin）。



核方法 - 核技巧（kernel trick）

将数据映射到高维表示从而使分类问题简化？在实践中通常是难以计算的。

核技巧（kernel trick）

要想在新的表示空间中找到良好的决策超平面，你不需要在新空间中直接计算点的坐标，只需要在新空间中计算点对之间的距离。

核函数（kernel function）

实现将原始空间中的任意两点映射为这两点在目标表示空间中的距离的函数。

注意

核函数通常是人为选择的，而不是从数据中学到的——对于 SVM 来说，只有分割超平面是通过学习得到的。

核方法 - 评价

优点

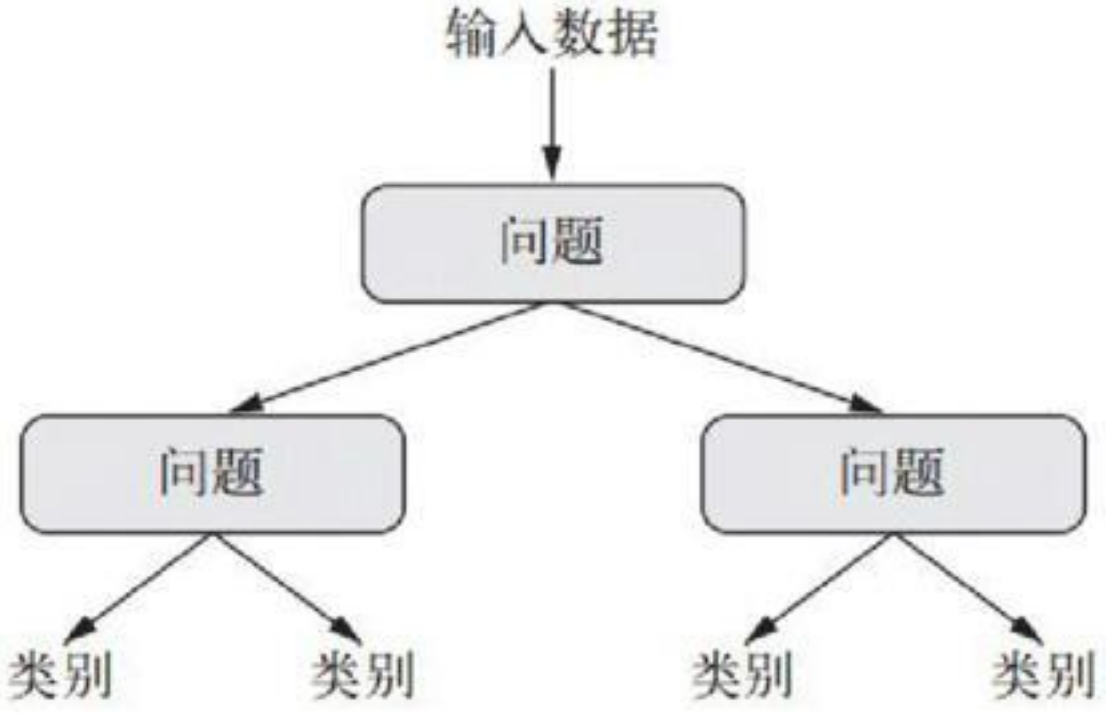
- SVM 刚刚出现时，在简单的分类问题上表现出了最好的性能。
- 大量的理论支持，适合用于严肃的数学分析，因而非常易于理解和解释。

缺点

- 很难扩展到大型数据集，并且在图像分类等感知问题上的效果也不好。
- SVM是一种比较浅层的方法，因此要想将其应用于感知问题，首先需要手动提取出有用的表示（这叫作特征工程），这一步骤很难，而且不稳定。

决策树、随机森林与梯度提升机

决策树（decision tree）是类似于流程图的结构，可视化和解释都很简单。 到了 2010 年，决策树经常比核方法更受欢迎。



决策树、随机森林与梯度提升机

随机森林（random forest）算法构建多个决策树，然后将它们的输出集成在一起。

- 随机森林适用于各种各样的问题——对于任何浅层的机器学习任务来说，它几乎总是第二好的算法。
- 广受欢迎的机器学习竞赛网站Kaggle在 2010 年上线后，随机森林迅速成为平台上人们的最爱，直到 2014 年才被梯度提升机所取代。

梯度提升机（gradient boosting machine）通过迭代地训练新模型来专门解决之前模型的弱点。

- 将梯度提升技术应用于决策树时，得到的模型与随机森林具有相似的性质，但在绝大多数情况下效果都比随机森林要好。
- 它可能是目前处理非感知数据最好的算法之一（如果非要加个“之一”的话）。和深度学习一样，它也是 Kaggle 竞赛中最常用的技术之一。

Kaggle 竞赛 - 向优秀的人学习

要想了解机器学习算法和工具的现状，一个好方法是看一下 Kaggle 上的机器学习竞赛：哪种算法能够可靠地赢得竞赛呢？顶级参赛者都使用哪些工具？

在 2017 年左右，Kaggle 上主要有两大方法：梯度提升机和深度学习。具体而言：

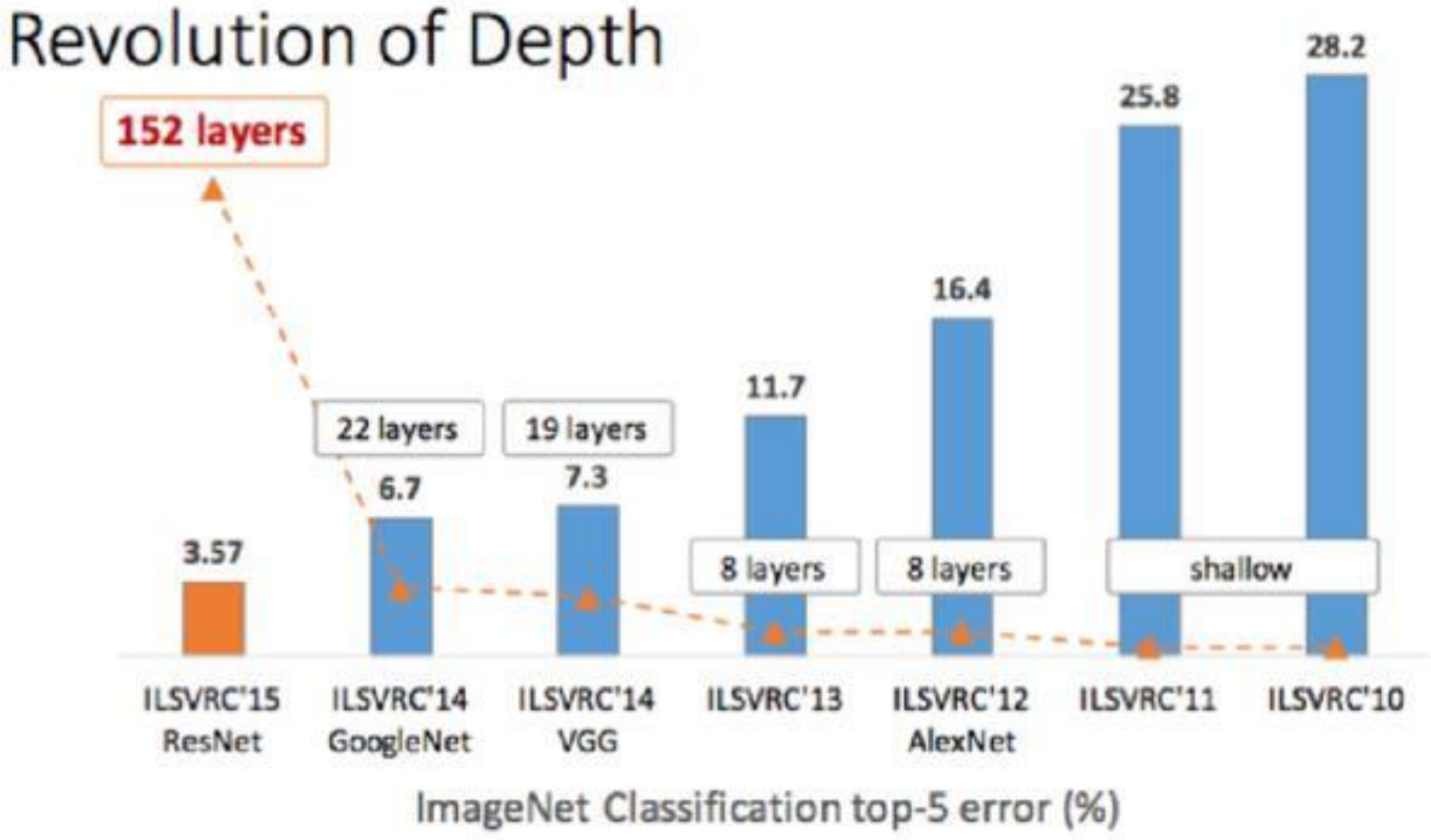
- 梯度提升机用于处理结构化数据的问题，大多使用 XGBoost 库。
- 深度学习则用于图像分类等感知问题，大多使用 Keras 库。

回到神经网络

Turing Award 2018

ImageNet

NN depth milestones



神经网络的巅峰

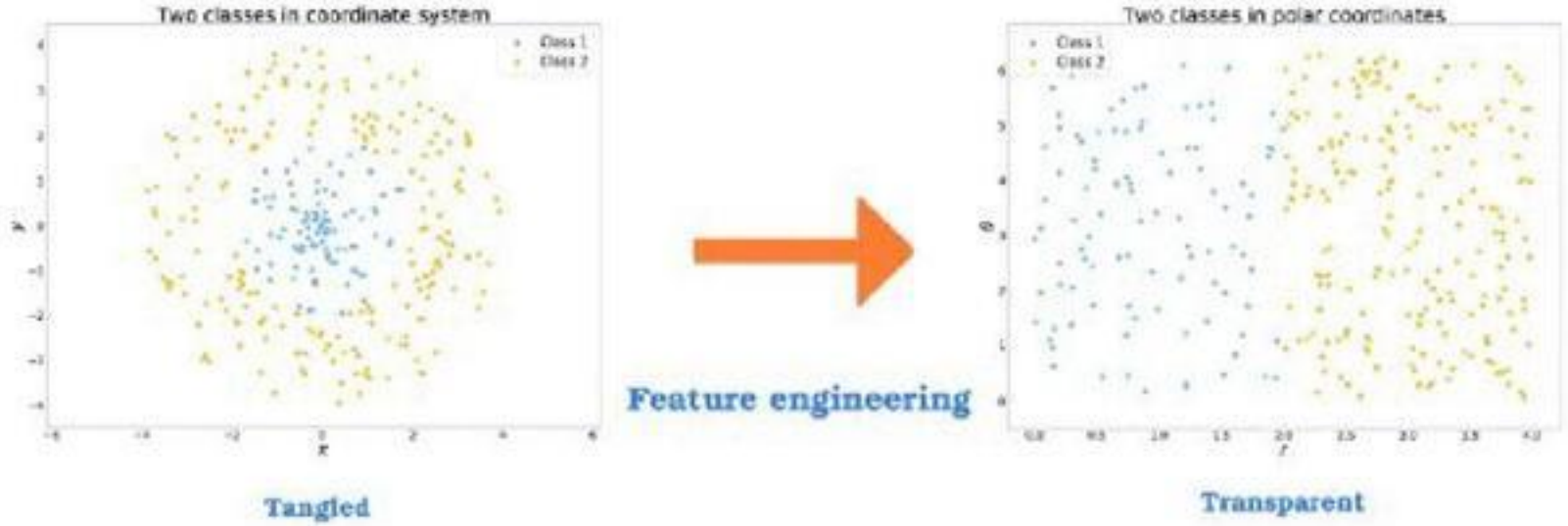
- 自 2012 年以来，深度卷积神经网络（convnet）已成为所有计算机视觉任务的首选算法。
- 更一般地说，它在所有感知任务上都有效。
- 在 2015 年和 2016 年的主要计算机视觉会议上，几乎所有演讲都与 convnet 有关。
- 与此同时，深度学习也在许多其他类型的问题上得到应用，比如自然语言处理。
- 它已经在大量应用中完全取代了 SVM 与决策树。

深度学习有何不同?

深度学习发展得如此迅速，主要原因在于它在很多问题上都表现出更好的性能。

深度学习还让解决问题变得更加简单，因为它将特征工程完全自动化，而这曾经是机器学习工作流程中最关键的一步。

什么是特征工程?



- 先前的机器学习技术（浅层学习）需要手动设计表示空间之间的数据变换操作。
- 深度学习完全将这个步骤自动化：利用深度学习，你可以一次性学习所有特征。

深度学习不是浅层学习的简单叠加

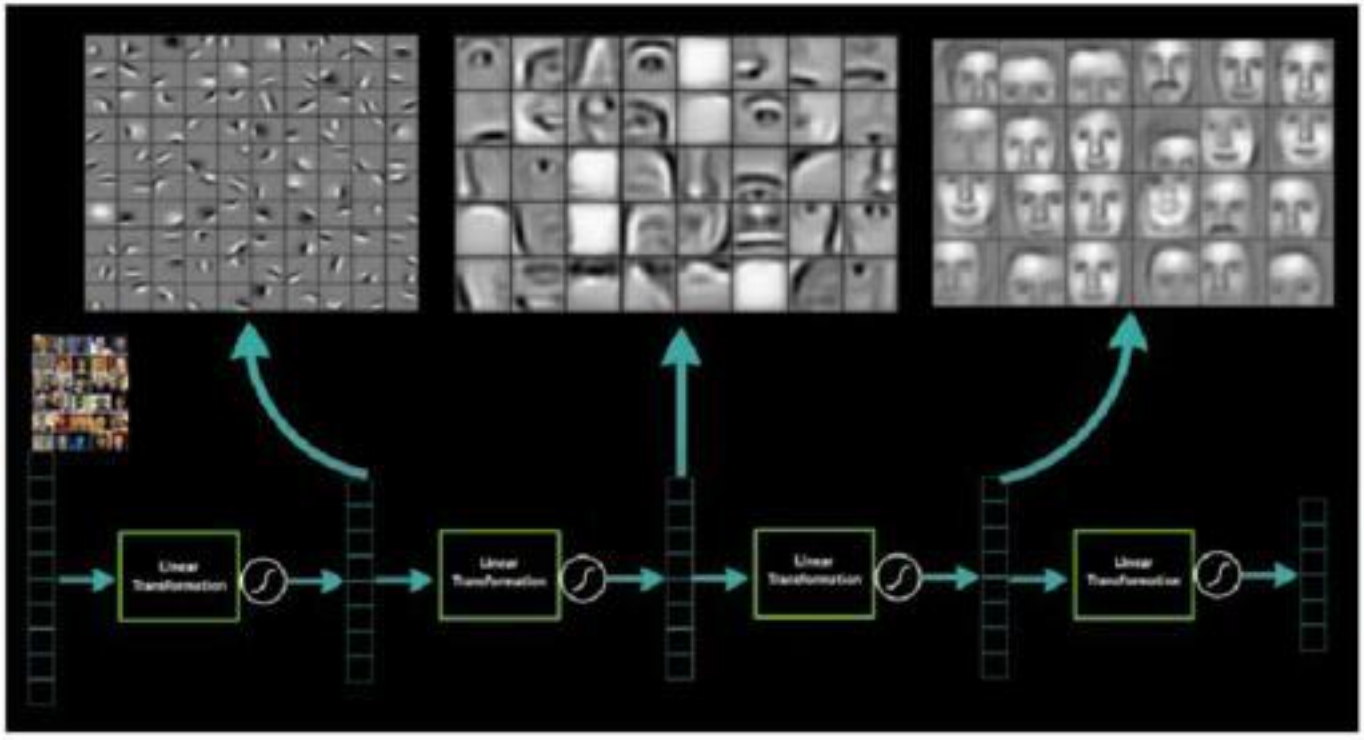
在实践中，如果连续应用浅层学习方法，其收益会随着层数增加迅速降低，因为三层模型中最优的第一表示层并不是单层或双层模型中最优的第一表示层。

深度学习的变革性在于，模型可以在同一时间共同学习所有表示层，一切都由单一反馈信号来监督。

罗恩的选择

深度学习更加成功秘诀

- 将特征工程完全自动化。
- 渐进、逐层地形成越来越复杂的表示。
- 所有层同时趋向全局最优解：每一层的变化都需要同时考虑上下两层的需要。



为什么是深度学习，为什么是现在

从技术成熟到广泛应用

深度学习用于计算机视觉的两个关键思想，即卷积神经网络和反向传播，在1989 年就已经为人们所知。

长短期记忆（LSTM，long short-term memory）算法是深度学习处理时间序列的基础，它在 1997 年就被开发出来了，而且此后几乎没有发生变化。

那么为什么深度学习在2012 年之后才开始取得成功？这二十年间发生了什么变化？

总的来说，三种技术力量在推动着机器学习的进步：

- 硬件。
- 数据集和测试基准。
- 算法上的改进。

机器学习是一门工程科学

只有当合适的数据和硬件可用于尝试新想法时（或者将旧想法的规模扩大，工程实践往往如此），才可能出现算法上的改进。

在 20 世纪 90 年代和 21 世纪前十年，真正的瓶颈在于数据和硬件。 但在这段时间内发生了下面这些事情：互联网高速发展，并且针对游戏市场的需求开发出了高性能图形芯片。

硬件

从 1990 年到 2010 年，非定制 CPU 的速度提高了约 5000 倍。

在 20 世纪前十年里，NVIDIA 和 AMD 等公司投资数十亿美元来开发快速的大规模并行芯片（图形处理器，GPU），以便为越来越逼真的视频游戏提供图形显示支持。

2007 年，NVIDIA 推出了 CUDA，作为其 GPU 系列的编程接口。

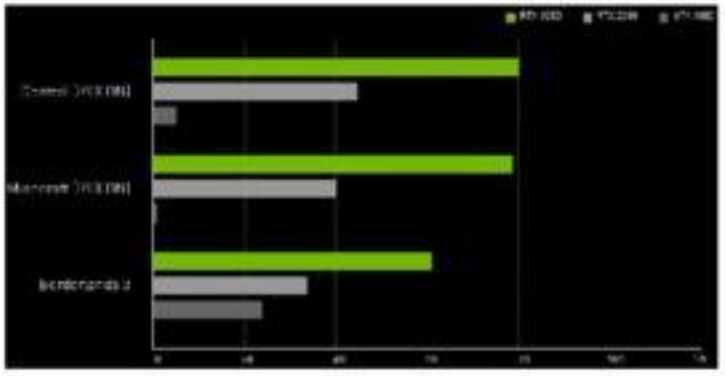
- 游戏中实时渲染复杂的3D场景需要大量的并行化矩阵计算。
- 非常巧合，深度神经网络也需要大量的并行化矩阵计算。

游戏玩家的力量

游戏市场推动了人工智能应用必需的超级计算能力

NVIDIA TITAN X 比一台现代笔记本电脑的速度要快约 350 倍。

使用一块 TITAN X 显卡，只需几天就可以训练出几年前赢得 ILSVRC 竞赛的 ImageNet 模型。与此同时，大公司还在包含数百个 GPU 的集群上训练深度学习模型，这种类型的 GPU 是专门针对深度学习的需求开发的，比如 NVIDIA Tesla K80。



2016 年，Google 展示的张量处理器（TPU）比最好的 GPU 还要快 10 倍。

数据

人工智能有时被称为新的工业革命。 如果深度学习是这场革命的蒸汽机，那么数据就是煤炭，即驱动智能机器的原材料，没有煤炭一切皆不可能。

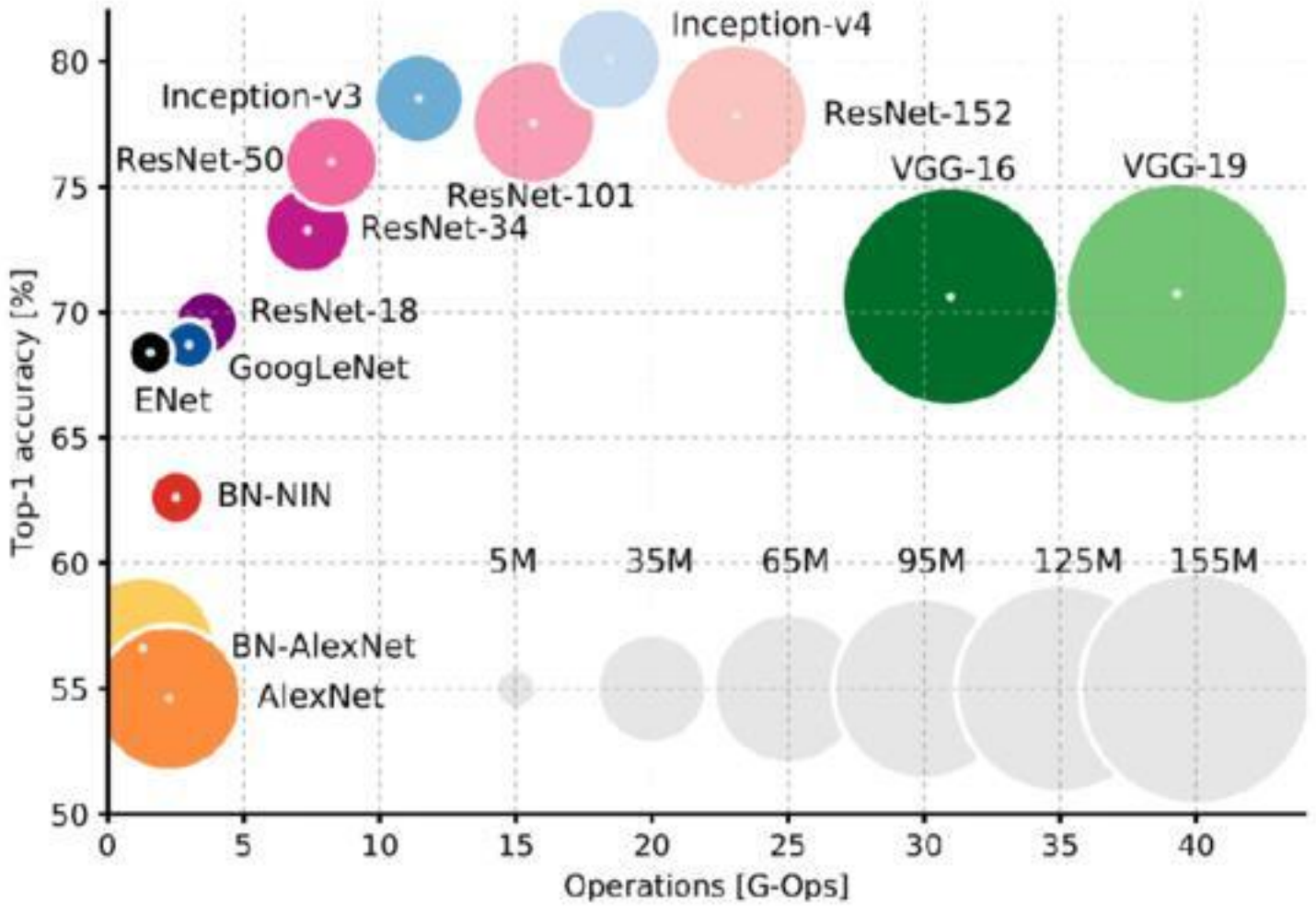
就数据而言，除了过去 20 年里存储硬件的指数级增长（遵循摩尔定律），最大的变革来自于互联网的兴起，它使得收集与分发用于机器学习的超大型数据集变得可行。

测试基准

深度学习兴起的催化剂：ImageNet 数据集。

ImageNet 的特殊之处不仅在于数量之大，还在于与它相关的年度竞赛：ILSVRC。

复杂度比较


Inception-v4
Inception-v3
ResNet-152
ResNet-50
VGG-16
VGG-19
ResNet-101
ResNet-34
ResNet-18
GoogLeNet
ENet
BN-NIN
5M
35M
65M
95M
125M
155M
BN-AlexNet
AlexNet
Top-1 accuracy [%]
80
75
70
65
60
55
50
0
5
10
15
20
25
30
35
40
Operations [G-Ops]

新的投资热潮

- 2011 年在人工智能方面的风险投资总额大约为 1900 万美元，几乎全都投给了浅层机器学习方法的实际应用。
- 到了 2014 年，这一数字已经涨到了 3.94 亿美元。
- 这三年里创办了数十家创业公司，试图从深度学习炒作中获利。
- Google、Facebook、百度、微软等大型科技公司已经在内部研究部门进行投资，其金额很可能已经超过了风险投资的现金流。
- 2013 年，Google 收购了深度学习创业公司 DeepMind，报道称收购价格为 5 亿美元。
- 2014年，百度在硅谷启动深度学习研究中心，并投资 3 亿美元。

> *“机器学习这一具有变革意义的核心技术将促使我们重新思考做所有事情的方式。 我们用心将其应用于所有产品，无论是搜索、广告、YouTube 还是 Google Play。 我们尚处于早期阶段，但你将会看到我们系统性地将机器学习应用于所有这些领域。”*
>
> *– Google 首席执行官 Sundar Pichai*

深度学习的大众化

技术层面

- 在早期，从事深度学习需要精通 C++ 和 CUDA，而它们只有少数人才能掌握。
- Keras 等用户友好型库则使深度学习变得像操纵乐高积木一样简单。

教育层面

- 我国正在试图在初等教育中推广人工智能和Python编程课程。
- 从2020年起大量高校开设人工智能学院或专业。

这种趋势会持续吗?

> *深度神经网络成为企业投资和研究人员纷纷选择的方法，是否只是难以持续的昙花一现?*

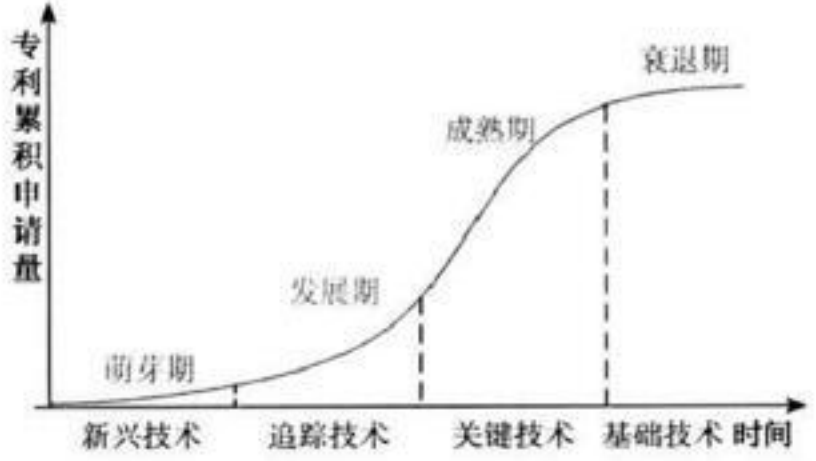
深度学习有几个重要的性质，证明了它确实是人工智能的革命，并且能长盛不衰。

- 简单：不需要特征工程，用户友好型库。
- 可扩展：高度并行的特性可以充分利用摩尔定律，小批量数据迭代训练的特性可以在任意大小的数据集上进行训练。
- 可复用：可用于连续在线学习，
- 可迁移：可以将训练好的模型迁移到不同领域的问题。

展望 20 年后

20 年后我们可能不再使用神经网络，但我们那时所使用的工具都是直接来自于现代深度学习及其核心概念。

在一次科学革命之后，科学发展的速度通常会遵循一条 S 形曲线：首先是一个快速发展时期，接着随着研究人员受到严重限制而逐渐稳定下来，然后进一步的改进又逐渐增多。 深度学习仍然处于这条 S 形曲线的前半部分，在未来几年将会取得更多进展。



The end

概念

人工智能

将通常由人类完成的智力任务自动化。

学习

数据观点：寻找对目标任务而言最佳数据表示的自动搜索过程。

函数观点：寻找逼近目标任务的最佳显函数的自动搜索过程。

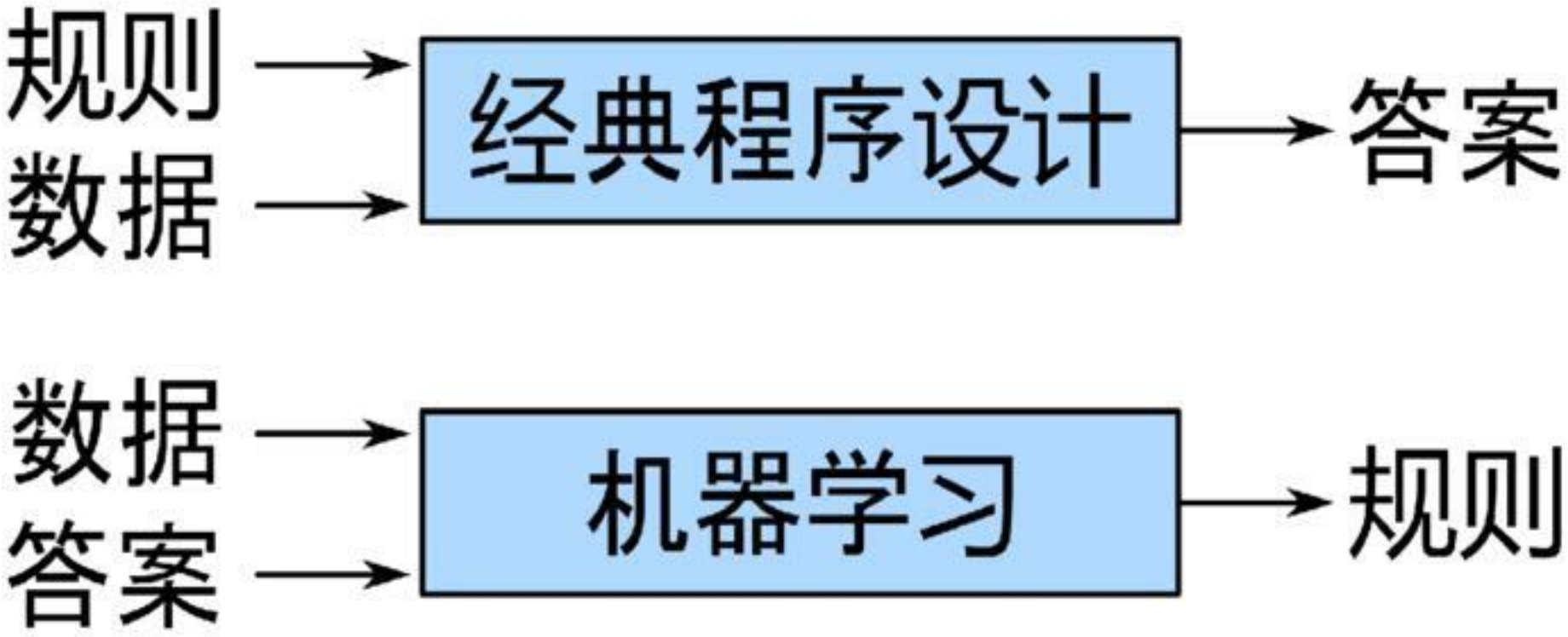
机器学习

在预先定义好的假设空间中，利用反馈信号的指引来自动寻找一个最优的显函数。这个函数能够将输入数据映射为解决目标任务的最佳数据表示。

深度学习

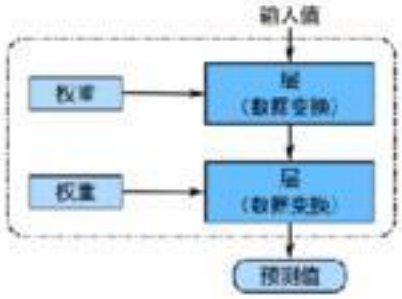
学习最佳数据表示的多层方法；其中层是数据处理单元。

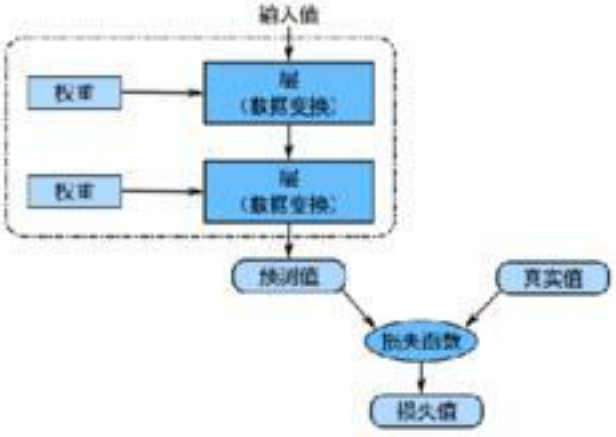
理解机器学习的编程范式

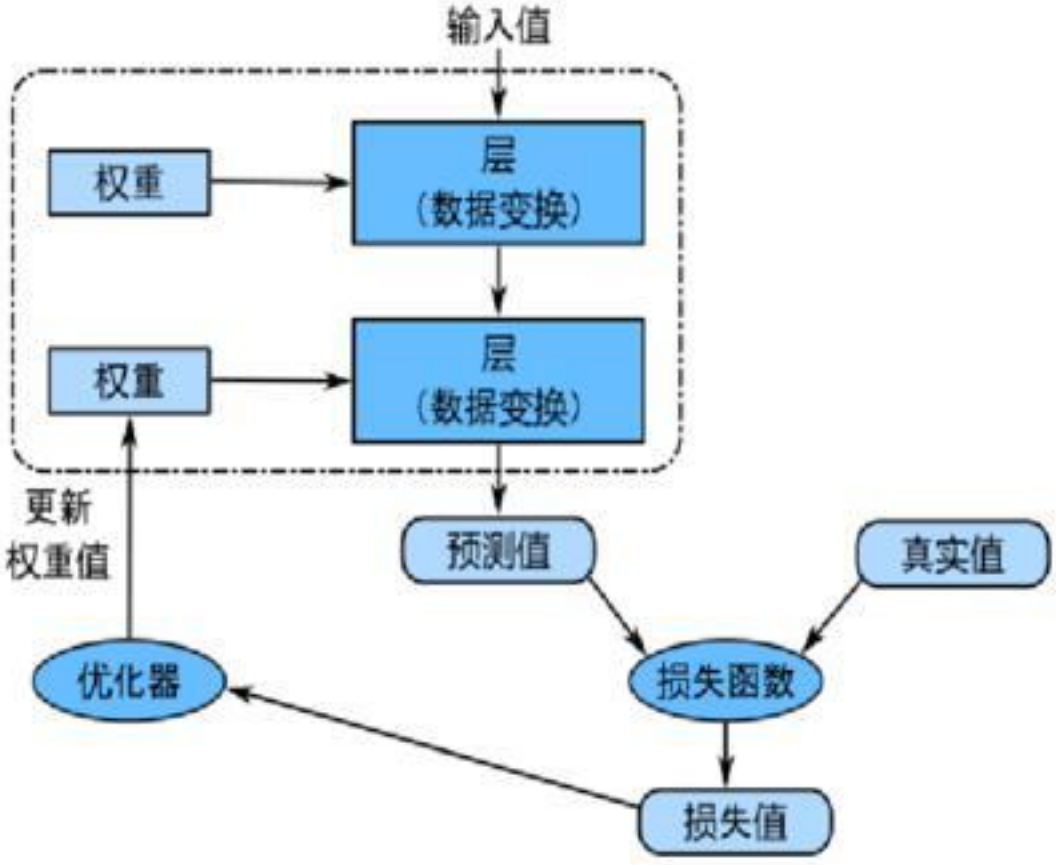


注意：机器学习系统（即模型）是训练出来的，而不是人为编写出来的。

理解深度学习算法的工作流程







机器学习方法

浅层方法（了解）

- 概率建模：常用于基准。
  - 朴素贝叶斯、logistic 回归
- 核方法：理论完美，人工特征工程。
  - 支持向量机（SVM）
- 图算法：简单、有效，适用于非感知数据。
  - 决策树、随机森林与梯度提升机（错题本）

深度学习的基本特征

- 将特征工程完全自动化。
- 渐进、逐层地形成越来越复杂的表示。
- 所有层同时趋向全局最优解。

深度学习的现在与未来

三种技术力量

- 硬件。
- 数据集和基准。
- 算法上的改进。

深度学习有希望能长盛不衰的几个性质

- 简单：不需要特征工程，易于使用的库。
- 可扩展：高度并行，小批量数据迭代训练。
- 可复用：可用于连续在线学习，
- 可迁移：可以迁移训练好的模型。

深度学习具有技术革命的基本特征

- 现阶段主要得益于在资源和人力上的指数式投资。
- 未来很光明，尽管短期期望有些过于乐观。

深度学习工作站的简单配置

安装Anaconda/Miniconda

打开命令行终端Anaconda Powershell Prompt，并输入以下代码：

```
conda create -n dl tensorflow-gpu keras-gpu
```

MNIST 手写数字分类：简单全连接

典型的 Keras 工作流程

1. 定义训练数据：输入张量和目标张量。
2. 定义层组成的网络（或模型），将输入映射到目标。
3. 配置学习过程：选择损失函数、优化器和需要监控的指标。
4. 调用模型的 fit 方法在训练数据上进行迭代。

实际应用训练好的模型进行预测

- 调整输入数据的格式：预测输入要与训练输入格式一致。