简介

一、 介绍

1. 深度学习：计算机从经验中学习，以层次化的概念（`concept`）来理解世界。
   - 从经验中学习：避免了人工指定计算机学习所需的所有知识。
   - 层次化的概念：计算机通过从简单的概念来构建、学习更复杂的概念。

   如果绘制一张图来展示这些概念的关系，那么这张图是一个深度的层次结构，因此称这种方法为深度学习。

1.1 知识表达

1. 计算机需要获取大量的`"常识"`才能以人工智能的方式行动，如：`树叶是绿色的、乒乓球比足球小`。

   这些`"常识"`大部分是主观的和直观的，因此很难以正式的方式表达。一个问题是：如何将这些`"常识"`传给计算机。

   有两种方式将知识传递给计算机：

   - 知识库（`knowledge base`）：通过形式化语言硬编码关于真实世界的知识，计算机使用逻辑推理自动推理这些硬编码的知识。

     最出名的知识库项目是`Cyc`，但这些项目都没有取得重大成功。

   - 机器学习：`AI`系统通过从原始数据中提取模式来获得知识，并作出看起来`"智能"`的决策。

     如：通过朴素贝叶斯算法分离正常的电子邮件和垃圾邮件。

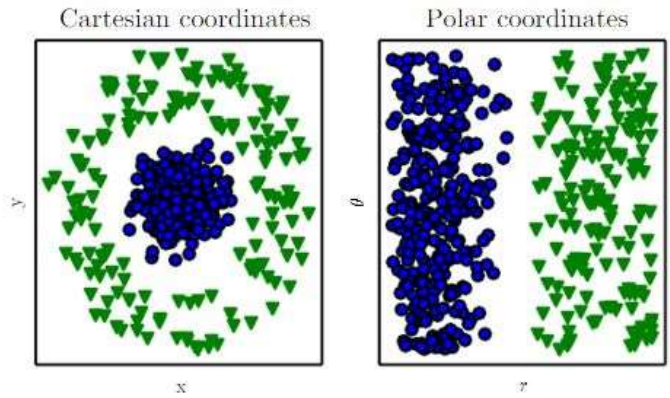
2. 传统机器学习算法严重依赖于数据的表达方式（`representation`）。如：对病人的诊断中，`AI`系统并不是直接接触病人，而是由医生告诉`AI`系统关于病人的一些信息（如身高、体重等）。这些信息称作特征（`feature`）。

   传统机器学习算法在三个层面上严重依赖于数据的表达方式：

   - 传统机器学习算法无法确定需要哪些特征。如：是直接给身高和体重，还是给出肥胖系数？
   - 传统机器学习算法也无法确定这些特征的方式。如：是否将特征离散化？
   - 某些信息，传统的机器学习算法无法学习。如：给出一份核磁共振的影像，由于每一个像素点与诊断结果相关性非常微小，因此传统机器学习算法无法学习。

3. 这种对数据的表达方式的依赖是计算机科学甚至生活中的一般现象。
   - 如：生活中人们很容易对阿拉伯数字进行算术运算，但是对于罗马数字的算术运算更费时间。
   - 下图的线性分类任务中：左图采用笛卡尔坐标系，右图采用极坐标系。

     可以看到数据的不同表达方式（坐标系的不同）导致左图难以线性分类，右图可以容易的线性分类。



4. 在传统的机器学习应用中，通常针对特定的任务来设计一套专用的、有效的特征集合，然后采集这些特征描述下的数据。如：语音识别中， 一个有效的特征就是讲话者的声音的声道（`vocal tract`）。

   但是大多数任务中，很难给出有效的特征是哪些。如：从图片中检查汽车的任务，可以使用是否有轮子作为一个特征。但是很难根据像素点来准确描述轮子。因为可能由于阴影、光照条件、观察角度等导致轮子的像素集合非常复杂。

5. 特征设计的一个解决方案是：通过机器学习来发现特征。即：**不仅学习`representation`到输出的映射（即模型），也学习`representation`本身**。这称作表达学习（`reprensentation learning`）。

   其优点有：

   - 往往比人为设计的特征的性能要好得多。
   - 允许`AI`系统快速适应新任务，用最少的人工干预。
   - 对于简单任务它可以在几分钟内学到一组好的特征，对于复杂任务可以在几小时到几个月的时间内学到一组好的特征。

     在复杂任务中，人工设计特征需要消耗大量的人力和时间。

1.2 特征的组合

1. 在设计特征或者学习特征时，一个良好的准则是：将能够解释数据的那些变化因子分离。

   通常这些因子不是直接观察到的量，而是影响那些能够直接观察到的量。如：在语音识别中，变化因子就是：讲话者的年龄、性别、口音、讲话的单词等。在汽车相关的图片识别中，变化因子就是：汽车的位置、汽车的颜色、观察角度等。

2. 在变化因子分离的过程中，有两个问题：

   - 大多数因子仅仅影响观察到的数据的某个部分，因此需要分解这些影响数据的因子，提取我们关心的因子。
   - 从原始数据中提取某些高级的、抽象的特征可能非常困难，这使得提取这种特征几乎和解决原始问题一样难。

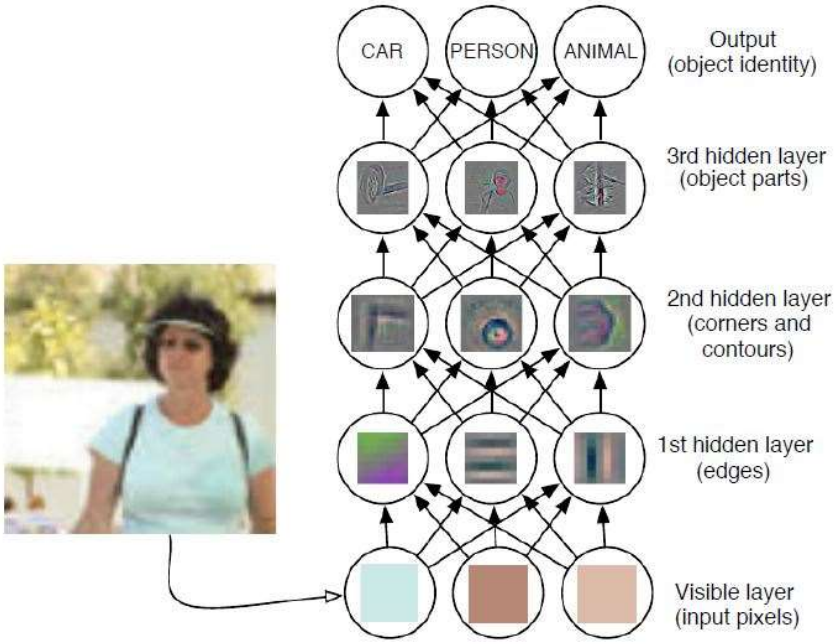
     如：说话者的口音只能用接近人类的、抽象的概念来表达。

3. 深度学习在学习特征时采用的解决方案是：高级特征以低级特征来表示。即：**通过组合简单的概念（`concept`）来构建复杂的概念。**

   如下所示的图片识别任务中，如果直接学习从一组像素到物体的映射很困难。深度学习通过将所需的复杂映射分解成一系列嵌套的简单映射来解决该问题。每个映射由模型的不同层来描述：

   - 可见层为输入，因为它包含了能够观察到的变量。

- 第一层隐层：描述了边(`edge`)的概念。通过比较相邻像素的亮度，则容易地识别边缘。
- 第二层隐层：描述了角(`corner`)和轮廓(`contour`)的概念。通过识别边的集合，则容易识别角。
- 第三层隐层：描述了特定物体整体（如：人物）的概念（物体由特定的角/等高线集合组成）。通过识别轮廓和角的特点集合，则容易识别物体整体。



4. 深度学习的一个经典案例是多层感知机(`multilayer perceptron:MLP`)。

   一个多层感知机就是一个函数：它将一组输入值映射到输出值，而这个函数由许多更简单的函数组成。可以认为每个函数都给出了输入的一个新的`representation`。

5. 深度学习的两个观点：

   - 一个观点是：深度学习就是学习数据正确的`representaion`，正如多层感知机所描述的。
   - 另一个观点是：深度学习让计算机学习一个多步计算程序：
     - 每一层的`representaion`被认为是在并行执行一组指令之后，计算机的存储器的状态。
     - 更深层的网络可以按顺序地执行更多的指令。
     - 序列越深的指令功能越强大，因为序列后面的指令可以参考序列前期指令的结果。

     根据这种观点：每一层的`representaion`中，并非所有的信息都对应了输入数据的特征信息，它还存储了辅助多步计算程序执行的状态信息：类似于计数器或者指针，它与输入的内容无关，但是有助于模型的组织处理过程。

1.3 深度

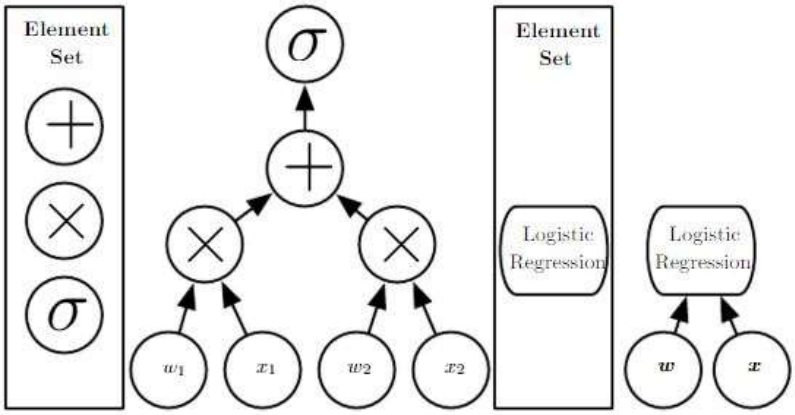
1. 深度学习的"深度"有两种度量方式。

   - 第一种度量方式为：框架中必须执行的顺序指令的数量。

     可以视为通过流程图的最长路径的长度，该流程图描述了如何根据输入来计算模型的输出。但是提供不同的函数单元，同一个模型可能具有不同的深度。

     下图给出的是`logistic regression`模型的深度。其中输出 $\hat{y}=\sigma(\vec{\mathbf{w}}^T\vec{\mathbf{x}})$， $\sigma(z)=\frac{1}{1+\exp(-z)}$ 为`sigmoid`函数。

- 左图中：将加法、乘法、`sigmoid`函数作为基本运算单元，则模型深度为 3。
- 右图中：将`logistic regression`模型本身作为基本运算单元，则模型深度为 1 。



- 第二种度量方式为：概念`concept`图的深度。

  流程图的深度可能远远大于概念图本身，因为如果给定了复杂概念，则简单的概念可以得到更好的理解。

  如：一个面部识别应用中，如果一只眼睛在阴影中，那么`AI`最初只能看到一只眼睛。在检测到面部的存在后，`AI`可以推断出第二只眼睛很可能存在。

  此时概念图只有两层：眼睛为第一层、面部为第二层。但是此时计算概念图的流程图可能为 $2n$ 层，其中 n 是对每个概念进行修正的次数。

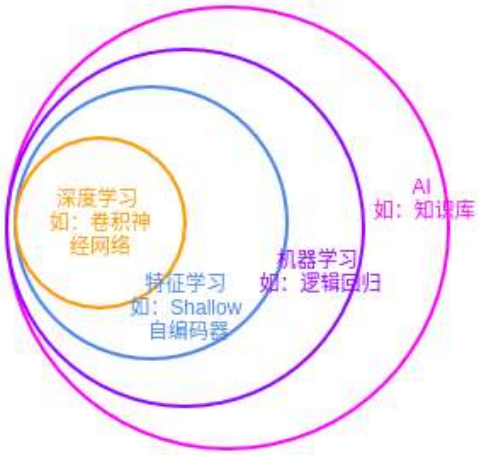
2. 对于模型的深度并没有一个标准的值，也没有说哪种度量方式是合适的。
   - 究竟模型的深度值为多少才能称作“深”，也没有标准答案。
   - 通常深度学习被认为是涉及大量的概念`concept`的模型的学习。

1.4 深度学习与 AI

1. 深度是机器学习的一种，它是一种特定类型的机器学习，通过学习将世界表示为层次化嵌套的概念。

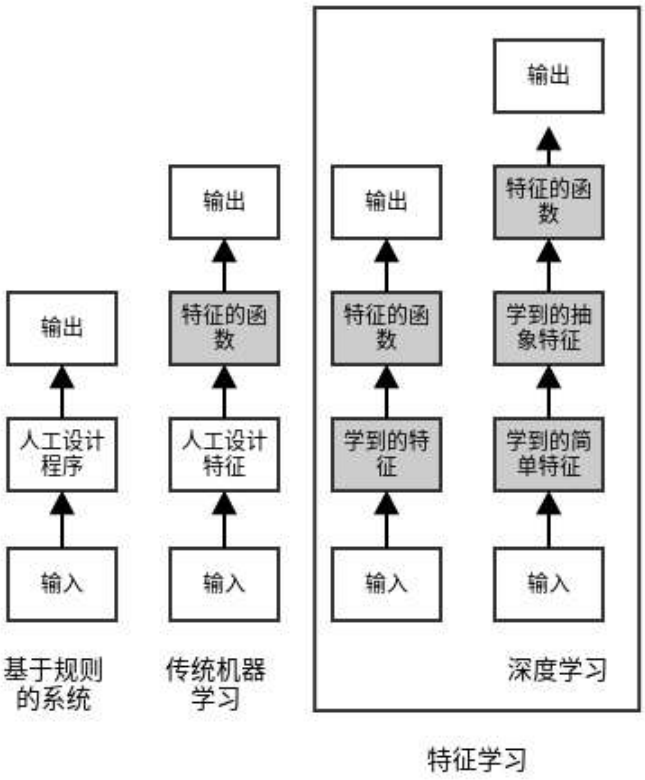
   每个概念都由一些更简单的概念定义，更抽象的特征由不那么抽象的特征来计算。

2. 下图给出了深度学习的隶属关系：深度学习 < 特征学习 < 机器学习 < 人工智能`AI` 。



3. 下图给出了不同`AI`系统中，不同部分的关联。阴影方框表示计算机从数据中学习获得的部分。

- 规则学习：硬编码知识。计算机所以无法、也不需要从数据中学习知识。
- 经典的机器学习：人工设计特征。计算机从数据中学习到了 “特征 --> `label` ” 之间的映射。
- 特征学习：机器从数据中自动学习到特征，然后学习到了 “特征 --> `label` ” 之间的映射。
- 深度学习：机器从数据中自动学习到了多层特征（深层特征由浅层特征来表达），然后学习到了 “特征 --> `label` ” 之间的映射。

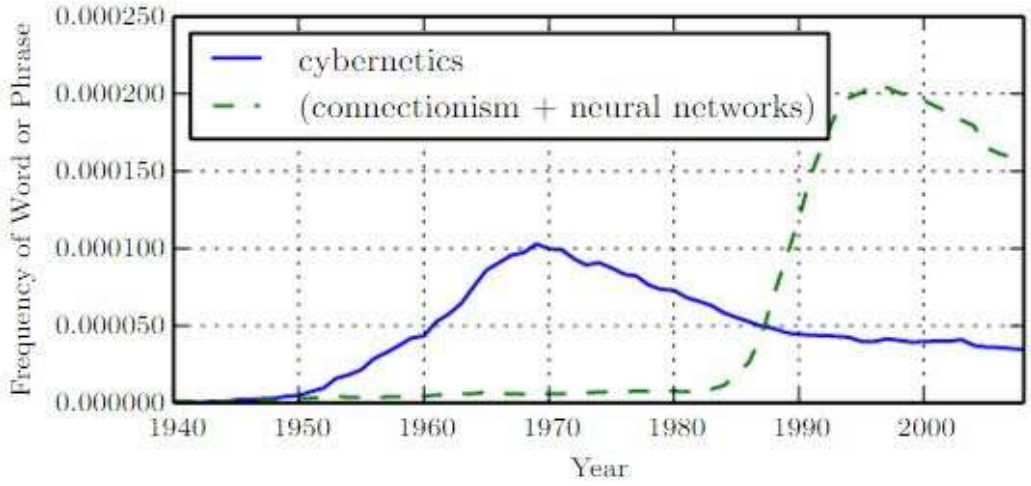


二、历史

2.1 历史简介

1. 深度学习分为三个时期：
   - `1940s-1960s`：这时它被称作控制论`cybernetics`。
   - `1980s-1990s`：这时它被称作连接机制`connectionism`。
   - `2006--`：这时被称作`deep learning`。

   深度学习在历史上的不同名字反映了不同的哲学观点。
2. 下图展示了神经网络研究的三个历史浪潮中的两个（因为第三波太近了）。
   - 第一波随着生物学习理论的发展和第一个模型的出现（如感知机神经元）。
   - 第二波用反向传播训练一层或者两层隐层神经网络。



3. 第一波浪潮：早期的人工智能算法模拟生物的学习过程：对大脑的学习过程建模。
    - 此时的深度学习被称作人工神经网络`artificial neural networks:ANNs`， 深度学习模型因此被认为是受生物大脑启发的工程系统。
    - 现代的深度学习超越了神经科学的观点：它是一种多层次学习的、通用的机器学习框架，而不必是从神经科学中获取灵感。
4. 第二波浪潮：连接机制的中心思想是：大量简单的计算单元在连接时可以实现智能行为。
    - 这种观点适用于生物神经系统中的神经元，以及深度网络模型中的隐层神经元。
    - 在连接机制期间，有一些核心思想仍然影响了后续的神经网络：
        - 分布式表达`distributed representation`：系统的每个输入应该由许多特征表示，每个特征描述了输入的一个部分。

          如一个视觉识别系统可以识别：汽车、卡车、鸟，这些对象可以为红色、蓝色、绿色。
            - 表达这些输入的一种方式为：9个神经元分别表示红色卡车、红色汽车、红色鸟、绿色汽车...等等。
            - 如果使用分布式表达，则使用6个神经元：3 个神经元来表示卡车、汽车、鸟，另外3个神经元来表示红色、绿色、蓝色。
    - 反向传播算法`back-propagation`：它是当前主要的训练深度模型的算法。
    - `long short-term memory:LSTM`网络：它是一种序列模型，解决了许多自然语言处理任务。
5. 第三波浪潮从2006年开始突破。

   `Geoffrey Hinton`给出了一种称作深度信念网络（`deep belief network`），该网络可以使用`greedy layer-wise pre-training`策略来有效地训练。

2.2 目前状况

1. 深度学习的成功的关键有两个：
    - 训练集大小的增长。
    - 硬件和软件的发展。包括更快的`CPU`、通用的`GUP`发展、更大内存、更快的网络连接、更好的软件基础。

2.2.1 训练集大小

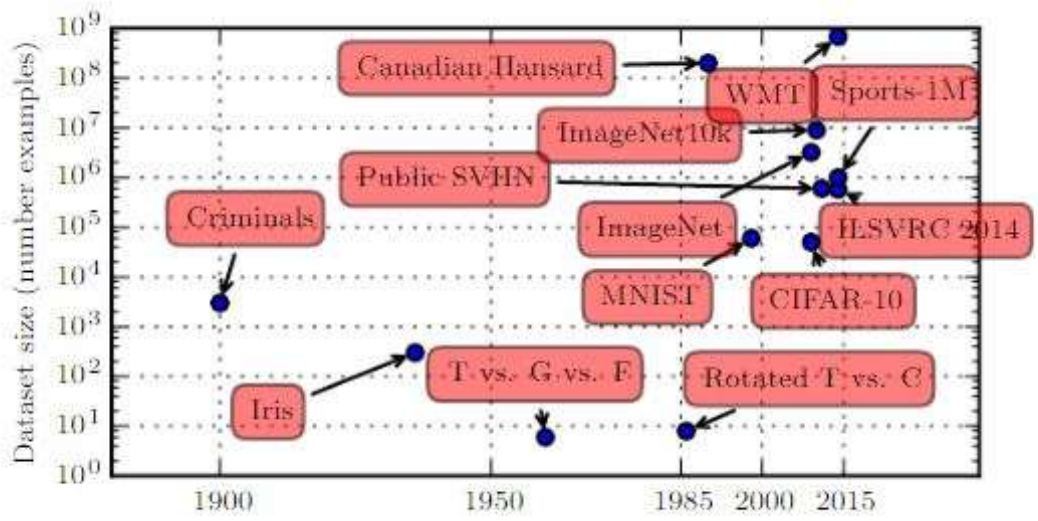
1. 早期的深度学习需要一些`trick`才能从算法中获得良好的性能。随着训练数据量的增加，所需的`trick`在降低。

   截止2016年，一个粗略的经验法则是：

   - 在深度学习的监督学习中要想获取可接受的性能，那么每个分类集合需要大约5000个标记样本。
   - 要想匹配甚至超越人类的性能，则训练集至少包含 1000 万个标记样本。
   - 对于小于这个数量的数据集，如何获取良好的性能是个重要的研究领域。尤其关注如何使用未标记样本（通过无监督学习或者半监督学习）。

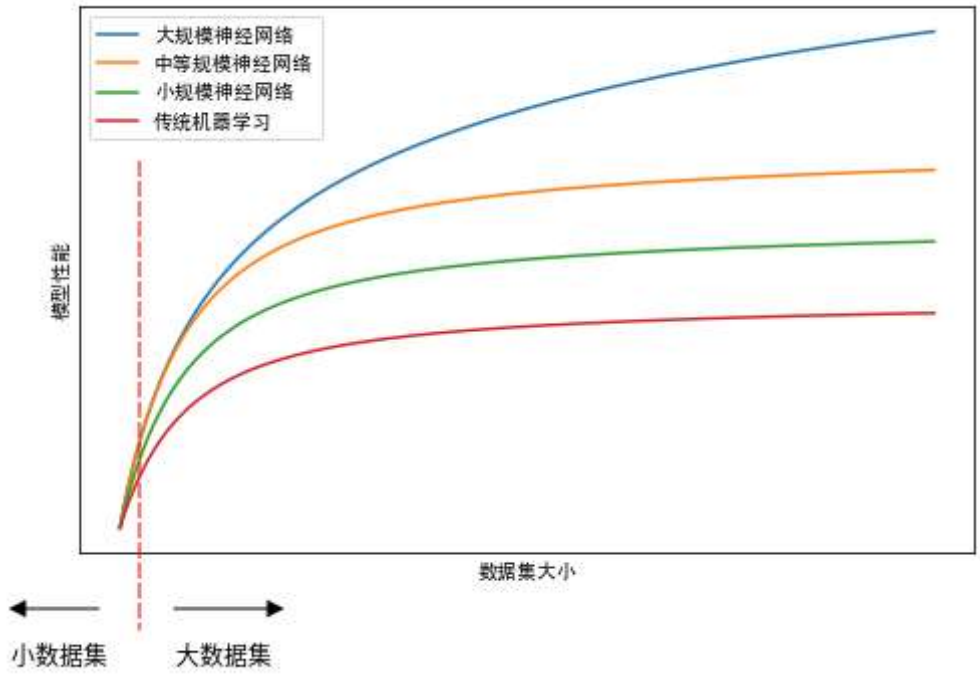
2. 下图显示了`benchmark`数据集大小随时间的变化，这种变化趋势是由于整个社会的数字化的推动。

   - 20 世纪初，统计学家使用数百或者数千的人工制作的数据来研究 。
   - 20世纪50年代到80年代，AI 专家采用较小的、合成的数据集（如低分辨率的位图的字母），从而证明神经网络能够完成特定任务。
   - 20世纪80年代到90年代，机器学习本质上更具有统计性。人们开始利用包含成千上万个例子的大数据集，如`MNIST`数据集。
   - 21世纪前十年，继续产生了同样大小的更复杂的数据集，如`CIFAR-10`数据集。
   - 2010年以来，更大的数据集包含了数十万到数千万的样本，深刻地改变了深度学习的可行性。



3. 神经网络的表现与它的规模和数据量有关。如今在神经网络上获取更好的性能的最可靠的方法就是：训练一个更大的神经网络，以及投入更多的数据。

   - 在大规模数据集上：更大的神经网络表现得更好。它们都优于传统算法。
   - 在小规模的数据集中，各种算法的性能排名事实上不是很确定。
     - 如果没有大量的训练集，则最终效果取决于你的特征工程能力，以及在算法细节上的处理上。
     - 只有在超大规模的训练集上，神经网络才能占领统治地位。

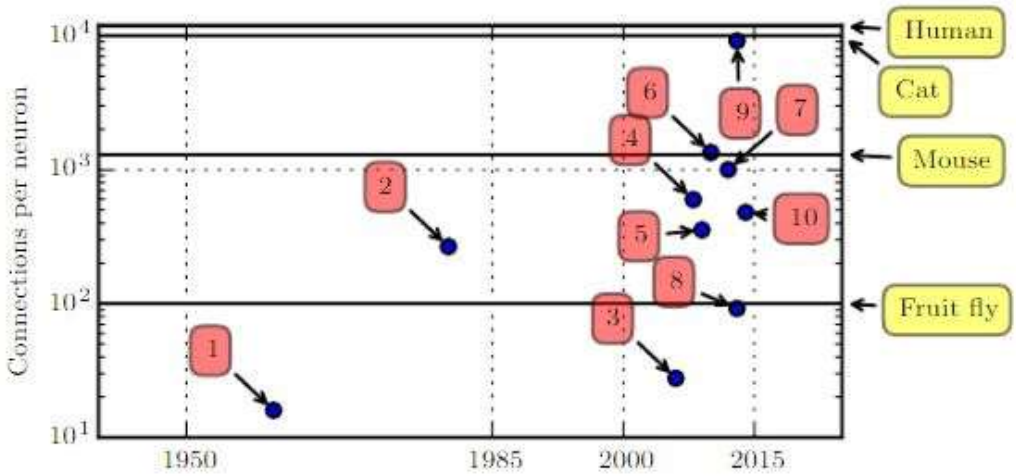


2.2.2 计算资源

1. 神经网络最近成功的另一个关键原因是：有充足的计算资源来运行更大的模型。
2. 连接主义的主要观点是：单个神经元或者少量的神经元集合没什么用处，只有许多神经元一起工作（更大的模型）才产生智能。

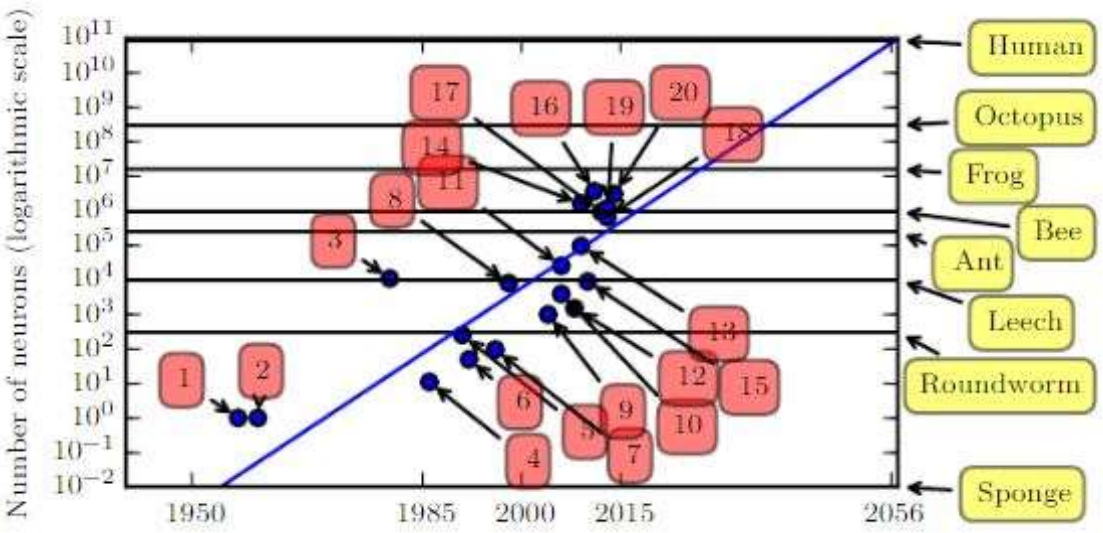
   更大的模型有两层含义：单个神经元的连接数量更多、模型的神经元的数量更多。
3. 生物的神经元的连接并不是非常密集。下图给出了一些动物和人的大脑中，每个神经元的连接数量。

   早期人工神经元之间的连接数受到硬件能力的限制，今天神经元之间的连接数量多数出于设计考虑。目前一些人工神经网络的单个神经元的连接数量已经和猫相同。



4. 在神经元总量上，早期的神经网络非常的小，直到最近才有较强的改观。
   - 自从引入了隐层以来，人工神经网络的大小大约每隔 2.4年翻一番。这种增长是由更大内存、更快的计算机（更快的`CPU`、通用`GPU`的出现、更快的网络连接、更好的软件基础）和更大的数据集来驱动的。
   - 更大的神经网络能够在更复杂的任务上实现更高的精度。

- 按照目前的趋势，大约2050年人工神经网络将具有与人类大脑相同数量的神经元。但是生物神经元可能拥有比人工神经元更复杂的功能。



2.2.3 算法优化

1. 在最近几年，许多算法方面的创新也推动了神经网络的发展。这些算法主要使得神经网络运行的更快。
   - 如：激活函数从 `sigmoid` 转换到 `ReLU` 函数。它使得基于梯度下降的算法运行的更快。
   - 快速计算的一个重要作用是：迭代网络的效率更高。