深度学习：今生前世

2017 年 1 月 25 日

译者按： 深度学习这个术语自 2006 年被正式提出后，在最近 10 年得到了巨大的发展，它使人工智能产生了革命性的技术突破，让我们切实地领略到人工智能改变人类生活的潜力。受人民邮电出版社的邀请，我的几位学生承担了 Goodfellow，Bengio 和 Courville (后续简称他们为 GBC) 撰写的《Deep Learning》一书翻译工作。原著三位作者一直耕耘于机器学习领域的前沿，引领深度学习的发展潮流。

原著第一章关于深度学习的思想、历史发展等的论述深刻、透彻和精辟，也非常耐人回味。我们在阅读该章时启发良多，大有裨益。因此，我们决定把该章的译稿独立成文，并取名为《深度学习：今生前世》。我们期望把它发布出来和大家分享。由于其篇幅较长，为了方便快速抓住其核心思想，这里我们把其中关键的内容摘录出来，同时把我个人一些的体会也一并呈现给大家。

GBC 指出：“人工智能的真正挑战在于解决那些对人来说很容易执行、但很难形式化描述的任务，比如识别人们所说的话或图像中的脸。对于这些问题，我们人类往往可以凭直觉轻易地解决”。为了解决这个挑战，他们提出让计算机从经验中学习，并根据层次化的概念体系来理解世界，而每个概念通过与某些相对简单的概念之间的关系定义。由此，GBC 给出了深度学习的定义。具体地，“层次化的概念让计算机构建较简单的概念来学习复杂概念。如果绘制出这些概念如何建立在彼此之上的图，我们将得到一张‘深’(层次很多) 的图。基于这个原因，我们称这种方法为 AI 深度学习 (deep learning)”。

因此，GBC 认为：“深度学习是通向人工智能的途径之一”。而且他们给出深度学习与机器学习之间的关系。GBC 认为：“深度学习是机器学习的一种，一种能够使计算机系统从经验和数据中得到提高的技术。我们坚信机器学习可以构建出在复杂实际环境下运行的 AI 系统，并且是唯一切实可行的方法。深度学习是一种特定类型的机器学习，具有强大的能力和灵活性，它将大千世界表示为嵌套的层次概念体系 (由较简单概念间的联系定义复杂概念、从一般抽象概括到高级抽象表示)”。

GBC 指出：“一般来说，目前为止深度学习已经经历了三次发展浪潮：20 世纪 40 年代到 60 年代深度学习的雏形出现在控制论 (cybernetics) 中，20 世纪 80 年代到 90 年代深度学习以联结主义 (connectionism) 为代表，并于 2006 年开始，以深度学习之名复兴”。

谈到深度学习与脑科学或者神经科学的关系，GBC 强调：“如今神经科学在深度学习研究中的作用被削弱，主要原因是我们根本没有足够的关于大脑的信息作为指导去使用它。要获得对被大脑实际使用算法的深刻理解，我们需要有能力同时监测 (至少是) 数千相连神经元的活动。我们不能够做到这一点，所以我们甚至连大脑最简单、最深入研究的部分都还远远没有理解”。值得注意的是，我国正致力于把人工智能和脑科学的交叉学科研究被提到战略地位，计划在 “类脑智能” 或 “脑计算” 等方面投入重点资助。且不

论我国是否真有既懂人工智能又懂脑科学或神经科学的学者，我们都应该本着务实、理性的求是态度。唯有如此，我们才有可能在这一波人工智能发展浪潮中有所作为，而不是又成为一群观潮人。

进一步地，GBC 指出：“媒体报道经常强调深度习与大脑的相似性。的确，深度学习研究者比其他机器学习领域 (如核方法或贝叶斯统计) 的研究者更可能地引用大脑作为影响，但大家不应该认为深度学习在尝试模拟大脑。现代深度学习从许多领域获取灵感，特别是应用数学的基本内容如线性代数、概率论、信息论和数值优化。尽管一些深度学习的研究人员引用神经科学作为灵感的重要来源，然而其他学者完全不关心神经科学”。的确，对于广大年青学者和一线的工程师们，我们是可以完全不用因不懂神经 (或脑) 科学而对深度学习、人工智能踯躅不前。数学模型、计算方法和应用驱动才是我们研究人工智能的可行之道。此外，我们诚然可以从哲学层面或角度来欣赏科学问题，但切不能从哲学层面来研究科学问题。

至于谈到人工神经网络在 20 世纪 90 年代中期的衰落，GBC 分析了其原因。“基于神经网络和其他 AI 技术的创业公司开始寻求投资，其做法野心勃勃但不切实际。当 AI 研究不能实现这些不合理的期望时，投资者感到失望。同时，机器学习的其他领域取得了进步。比如，核方法和图模型都在很多重要任务上实现了很好的效果。这两个因素导致了神经网络热潮的第二次衰退，并一直持续到 2007 年”。“其兴也悖焉，其亡也忽焉”。这个教训也同样值得当今基于深度学习的创业界、工业界和学术界等的警醒。

最后，GBC 总结并展望了深度学习：“深度学习是机器学习的一种方法。在过去几十年的发展中，它大量借鉴了我们关于人脑、统计学和应用数学的知识。近年来，得益于更强大的计算机、更大的数据集和能够训练更深网络的技术，深度学习的普及性和实用性都有了极大的发展。未来几年充满了进一步提高深度学习并将它带到新领域的挑战和机遇”。是的，深度学习、人工智能技术不是在我们头顶而是在我们脚下。

中文初译稿是我学生完成的，之后许多同行给出了大量富有建设性的修改意见，我本人也前后做了多次校对。即使这样，由于我们无论是中文还是英文能力都深感有限，译文还是比较生硬，而且我们特别担心未能完整地传达出原作者的真实思想和观点。因此，我们强烈地建议有条件的读者去读英文原著，也非常期待大家去 GitHub 指正我们的译文。

2017 年 1 月 25 日于北大静园六院

远在古希腊时期，发明家就梦想着创造能思考的机器。神话人物皮格马利翁 (Pygmalion)、代达罗斯 (Daedalus) 和赫淮斯托斯 (Hephaestus) 可以被看作传说中的发明家，而加拉蒂亚 (Galatea)、塔洛斯 (Talos) 和潘多拉 (Pandora) 则可以被视为人造生命 (Ovid and Martin, 2004; Sparkes, 1996; Tandy, 1997)。

当人类第一次构思可编程计算机时，就已经在思考计算机能否变得智能（尽管这距造出第一台计算机还有一百多年）(Lovelace, 1842)。如今，**人工智能** (artificial intelligence, AI) 已经成为一个具有众多实际应用和活跃研究课题的领域，并且正在蓬勃发展。我们期望通过智能软件自动地处理常规劳动、理解语音或图像、帮助医学诊断和支持基础科学研究。

在人工智能的早期，那些对人类智力来说非常困难、但对计算机来说相对简单的问题得到迅速解决，比如，那些可以通过一系列形式化的数学规则来描述的问题。人工智能的真正挑战在于解决那些对人来说很容易执行、但很难形式化描述的任务，如识别人们所说的话或图像中的脸。对于这些问题，我们人类往往可以凭直觉轻易地解决。

针对这些比较直观的问题，本书讨论一种解决方案。该方案可以让计算机从经验中学习，并根据层次化的概念体系来理解世界，而每个概念则通过与某些相对简单的概念之间的关系来定义。让计算机从经验获取知识，可以避免由人类来给计算机形式化地指定它需要的所有知识。层次化的概念让计算机构建较简单的概念来学习复杂概念。如果绘制出这些概念如何建立在彼此之上的图，我们将得到一张“深”（层次很多）的图。基于这个原因，我们称这种方法为AI**深度学习** (deep learning)。

AI许多早期的成功发生在相对朴素且形式化的环境中，而且不要求计算机具备很多关于世界的知识。例如，IBM 的深蓝（Deep Blue）国际象棋系统在 1997 年击败了世界冠军Garry Kasparov(Hsu, 2002)。显然国际象棋是一个非常简单的领域，因为它仅含有 64 个位置并只能以严格限制的方式移动 32 个棋子。设计一种成功的国际象棋策略是巨大的成就，但向计算机描述棋子及其允许的走法并不是挑战的困难所在。国际象棋完全可以由一个非常简短的、完全形式化的规则列表来描述，并可以容易地由程序员事先准备好。

讽刺的是，抽象和形式化的任务对人类而言是最困难的脑力任务之一，但对计算机而言却属于最容易的。计算机早就能够打败人类最好的象棋选手，但直到最近计算机才在识别对象或语音任务中达到人类平均水平。一个人的日常生活需要关于世界的巨量知识。很多这方面的知识是主观的、直观的，因此很难通过形式化的方式表达清楚。计算机需要获取同样的知识才能表现出智能。人工智能的一个关键挑战就是如何将这些非形式化的知识传达给计算机。

一些人工智能项目力求将关于世界的知识用形式化的语言进行硬编码 (hard-code)。

计算机可以使用逻辑推理规则来自动地理解这些形式化语言中的申明。这就是众所周知的人工智能的**知识库** (knowledge base) 方法。这些项目没有导致重大的成功。其中最著名的项目是 Cyc (Lenat and Guha, 1989)。Cyc 包括一个推断引擎和一个使用 CycL 语言描述的声明数据库。这些声明是由人类监督者输入的。这是一个笨拙的过程。人们设法设计出足够复杂的形式化规则来精确地描述世界。例如，Cyc 不能理解一个关于名为Fred的人在早上剃须的故事 (Linde, 1992)。它的推理引擎检测到故事中的不一致性：它知道人没有电气零件，但由于Fred正拿着一个电动剃须刀，它认为实体 "正在剃须的Fred"("FredWhileShaving") 含有电气部件。因此它产生了这样的疑问——Fred在刮胡子的时候是否仍然是一个人。

依靠硬编码的知识体系面对的困难表明，AI系统需要具备自己获取知识的能力，即从原始数据中提取模式的能力。这种能力被称为**机器学习** (machine learning)。引入机器学习使计算机能够解决涉及现实世界知识的问题，并能作出看似主观的决策。比如，一个被称为**逻辑回归** (logistic regression) 的简单机器学习算法可以决定是否建议剖腹产 (Mor-Yosef *et al.*, 1990)。而同样是简单机器学习算法的**朴素贝叶斯** (naive Bayes) 则可以区分垃圾电子邮件和合法电子邮件。

这些简单的机器学习算法的性能在很大程度上依赖于给定数据的**表示** (representation)。例如，当逻辑回归被用于推荐剖腹产时，AI系统不直接检查患者。相反，医生需要告诉系统几条相关的信息，诸如子宫疤痕是否存在。表示患者的每条信息被称为一个特征。逻辑回归学习病人的这些特征如何与各种结果相关联。然而，它丝毫不能影响该特征定义的方式。如果将病人的 MRI 扫描作为逻辑回归的输入，而不是医生正式的报告，它将无法作出有用的预测。MRI 扫描的单一像素与分娩过程中并发症之间的相关性微乎其微。

在整个计算机科学乃至日常生活中，对表示的依赖都是一个普遍现象。在计算机科学中，如果数据集合被精巧地结构化并被智能地索引，那么诸如搜索之类的操作的处理速度就可以成指数级地加快。人们可以很容易地在阿拉伯数字的表示下进行算术运算，但在罗马数字的表示下运算会比较耗时。因此，毫不奇怪，表示的选择会对机器学习算法的性能产生巨大的影响。图1展示了一个简单的可视化例子。

许多人工智能任务都可以通过以下方式解决：先提取一个合适的特征集，然后将这些特征提供给简单的机器学习算法。例如，对于通过声音鉴别说话者的任务来说，一个有用的特征是对其声道大小的估计。这个特征为判断说话者是男性、女性还是儿童提供了有力线索。

然而，对于许多任务来说，我们很难知道应该提取哪些特征。例如，假设我们想编写一个程序来检测照片中的车。我们知道，汽车有轮子，所以我们可能会想用车轮的存

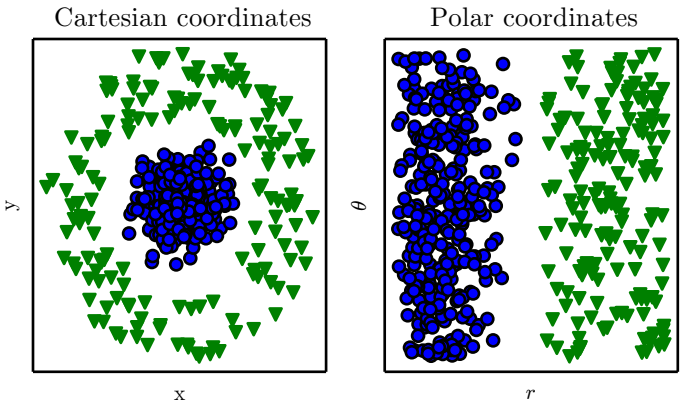

图 1: 不同表示的例子：假设我们想在散点图中画一条线来分隔两类数据。在左图，我们使用笛卡尔坐标表示数据，这个任务是不可能的。右图中，我们用极坐标表示数据，可以用垂直线简单地解决这个任务。(与 David Warde-Farley 合作画出此图。)

在与否作为特征。不幸的是，我们难以准确地根据像素值来描述车轮看上去像什么。虽然车轮具有简单的几何形状，但它的图像可能会因场景而异，如落在车轮上的阴影、太阳照亮的车轮的金属零件、汽车的挡泥板或者遮挡的车轮一部分的前景物体等等。

解决这个问题的途径之一是使用机器学习来发掘表示本身，而不仅仅把表示映射到输出。这种方法我们称之为**表示学习** (representation learning)。学习到的表示往往比手动设计的表示表现得更好。并且它们只需最少的人工干预，就能让AI系统迅速适应新的任务。表示学习算法只需几分钟就可以为简单的任务发现一个很好的特征集，对于复杂任务则需要几小时到几个月。手动为一个复杂的任务设计特征需要耗费大量的人工时间和精力；甚至需要花费整个社群研究人员几十年的时间。

表示学习算法的典型例子是**自编码器** (autoencoder)。自编码器由一个**编码器** (encoder) 函数和一个**解码器** (decoder) 函数组合而成。编码器函数将输入数据转换为一种不同的表示，而解码器函数则将这个新的表示转换到原来的形式。我们期望当输入数据经过编码器和解码器之后尽可能多地保留信息，同时希望新的表示有各种好的特性，这也是自编码器的训练目标。为了实现不同的特性，我们可以设计不同形式的自编码器。

当设计特征或设计用于学习特征的算法时，我们的目标通常是分离出能解释观察数据的**变差因素** (factors of variation)。在此背景下，“因素” 这个词仅指代影响的不同来源；因素通常不是乘性组合。这些因素通常是不能被直接观察到的量。相反，它们可能是现实世界中观察不到的物体或者不可观测的力，但会影响可观测的量。为了对观察到的数据提供有用的简化解释或推断其原因，它们还可能以概念的形式存在于人类的思维中。

它们可以被看作数据的概念或者抽象，帮助我们了解这些数据的丰富多样性。当分析语音记录时，变差因素包括说话者的年龄、性别、他们的口音和他们正在说的词语。当分析汽车的图像时，变差因素包括汽车的位置、它的颜色、太阳的角度和亮度。

在许多现实的人工智能应用中，困难主要源于很多变差因素影响着我们能够观察到的每一个数据。比如，在一张包含红色汽车的图片中，其单个像素在夜间可能会非常接近黑色。汽车轮廓的形状取决于视角。大多数应用需要我们理清变差因素并忽略我们不关心的因素。

显然，从原始数据中提取如此高层次、抽象的特征是非常困难的。许多诸如说话口音这样的变差因素，只能通过对数据进行复杂的、接近人类水平的理解来辨识。这几乎与获得原问题的表示一样困难，因此，乍一看，表示学习似乎并不能帮助我们。

深度学习 (deep learning) 通过其他较简单的表示来表达复杂表示，解决了表示学习中的核心问题。

深度学习让计算机通过较简单概念构建复杂的概念。图2展示了深度学习系统如何通过组合较简单的概念（例如转角和轮廓，它们转而由边线定义）来表示图像中人的概念。深度学习模型的典型例子是前馈深度网络或**多层感知机** (multilayer perceptron, MLP)。多层感知机仅仅是一个将一组输入值映射到输出值的数学函数。该函数由许多较简单的函数复合而成。我们可以认为不同数学函数的每一次应用都为输入提供了新的表示。

学习数据的正确表示的想法是解释深度学习的一个视角。另一个视角是深度促使计算机学习一个多步骤的计算机程序。每一层表示都可以被认为是并行执行另一组指令之后计算机的存储器状态。更深的网络可以按顺序执行更多的指令。顺序指令提供了极大的能力，因为后面的指令可以参考早期指令的结果。从这个角度上看，在某层激活函数里，并非所有信息都蕴涵着解释输入的变差因素。表示还存储着状态信息，用于帮助程序理解输入。这里的状态信息类似于传统计算机程序中的计数器或指针。它与具体的输入内容无关，但有助于模型组织其处理过程。

目前主要有两种度量模型深度的方式。第一个观点是基于评估架构所需执行的顺序指令的数目。假设我们将模型表示为给定输入后，计算对应输出的流程图，则可以将这张流程图中的最长路径视为模型的深度。正如两个使用不同语言编写的等价程序将具有不同的长度；相同的函数可以被绘制为具有不同深度的流程图，其深度取决于我们可以用来作为一个步骤的函数。图3说明了语言的选择如何给相同的架构两个不同的衡量。

另一种是在深度概率模型中使用的方法，它不是将计算图的深度视为模型深度，而是将描述概念彼此如何关联的图的深度视为模型深度。在这种情况下，计算每个概念表示的计算流程图的深度可能比概念本身的图更深。这是因为系统对较简单概念的理解在

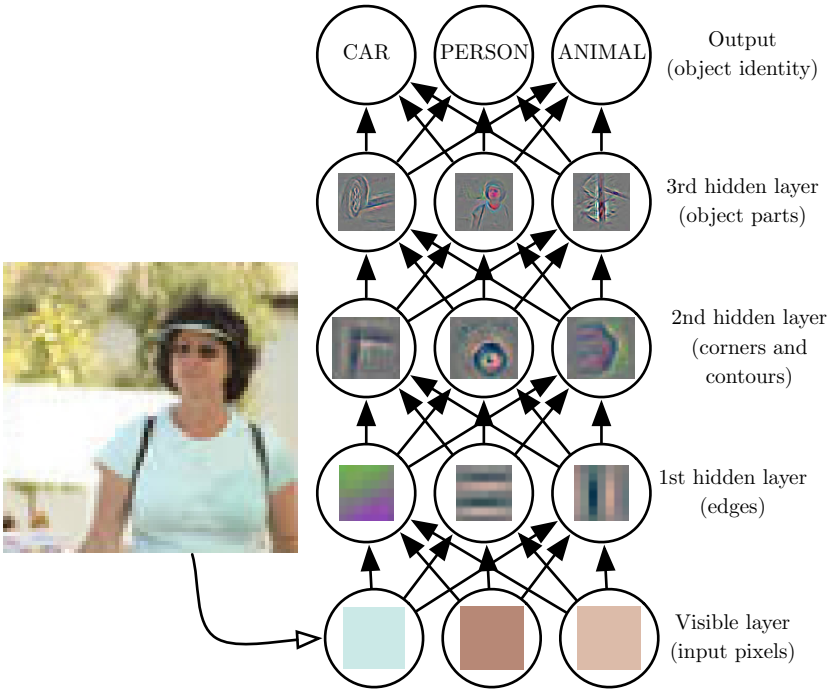

图 2: 深度学习模型的示意图。计算机难以理解原始感观输入数据的含义，如表示为像素值集合的图像。将一组像素映射到对象标识的函数非常复杂。如果直接处理，学习或评估此映射似乎是不可能的。深度学习将所需的复杂映射分解为一系列嵌套的简单映射（每个由模型的不同层描述）来解决这一难题。输入展示在**可见层** (visible layer)，这样命名的原因是因为它包含我们能观察到的变量。然后是一系列从图像中提取越来越多抽象特征的**隐藏层** (hidden layer)。因为它们的值不在数据中给出，所以将这些层称为 "隐藏"; 模型必须确定哪些概念有利于解释观察数据中的关系。这里的图像是每个隐藏单元表示的特征的可视化。给定像素，第一层可以轻易地通过比较相邻像素的亮度来识别边缘。有了第一隐藏层描述的边缘，第二隐藏层可以容易地搜索可识别为角和扩展轮廓的边集合。给定第二隐藏层中关于角和轮廓的图像描述，第三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。经Zeiler and Fergus (2014) 许可转载此图。

给出更复杂概念的信息后可以进一步精细化。例如，一个AI系统观察其中一只眼睛在阴影中的脸部图像时，它最初可能只看到一只眼睛。但当检测到脸部的存在后，系统可以推断第二只眼睛也可能是存在的。在这种情况下，概念的图仅包括两层（关于眼睛的层和关于脸的层），但如果我们根据每个概念给出的其他 n 次估计进行细化，计算的图将包括 $2n$ 层。

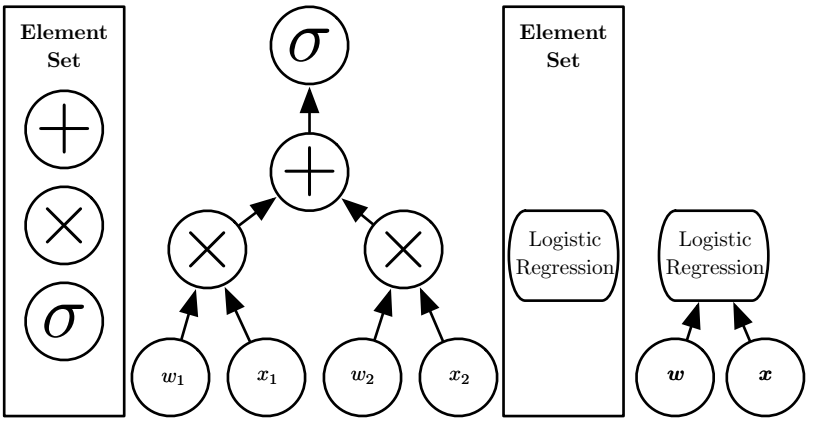

图 3: 将输入映射到输出的计算图表的示意图，其中每个节点执行一个操作。深度是从输入到输出的最长路径的长度，但这取决于可能的计算步骤的定义。这些图中所示的计算是逻辑回归模型的输出，$\sigma(\boldsymbol{w}^T\boldsymbol{x})$，其中 σ 是logistic sigmoid函数。如果我们使用加法、乘法和logistic sigmoid作为我们计算机语言的元素，那么这个模型深度为三。如果我们将逻辑回归视为元素本身，那么这个模型深度为一。

由于并不总是清楚计算图的深度或概率模型图的深度哪一个是最有意义的，并且由于不同的人选择不同的最小元素集来构建相应的图，因此就像计算机程序的长度不存在单一的正确值一样，架构的深度也不存在单一的正确值。另外，也不存在模型多么深才能被修饰为 ‘‘深’’ 的共识。但相比传统机器学习，深度学习研究的模型涉及更多学到功能或学到概念的组合，这点毋庸置疑。

总之，这本书的主题——深度学习是通向人工智能的途径之一。具体来说，它是机器学习的一种，一种能够使计算机系统从经验和数据中得到提高的技术。我们坚信机器学习可以构建出在复杂实际环境下运行的AI系统，并且是唯一切实可行的方法。深度学习是一种特定类型的机器学习，具有强大的能力和灵活性，它将大千世界表示为嵌套的层次概念体系（由较简单概念间的联系定义复杂概念、从一般抽象概括到高级抽象表示）。图4说明了这些不同的AI学科之间的关系。图5展示了每个学科如何工作的高层次原理。

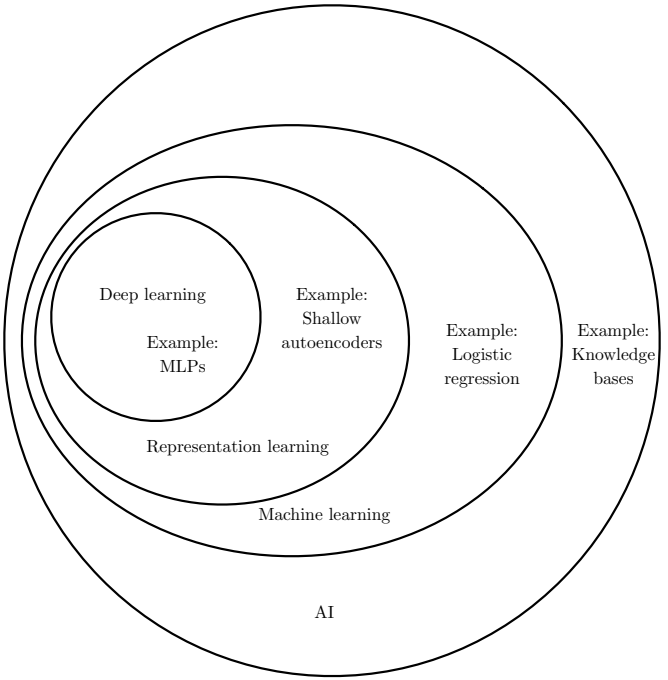

图 4: 维恩图展示了深度学习是一种表示学习，也是一种机器学习，可以用于许多（但不是全部）AI方法。维恩图的每个部分包括一个AI技术的示例。

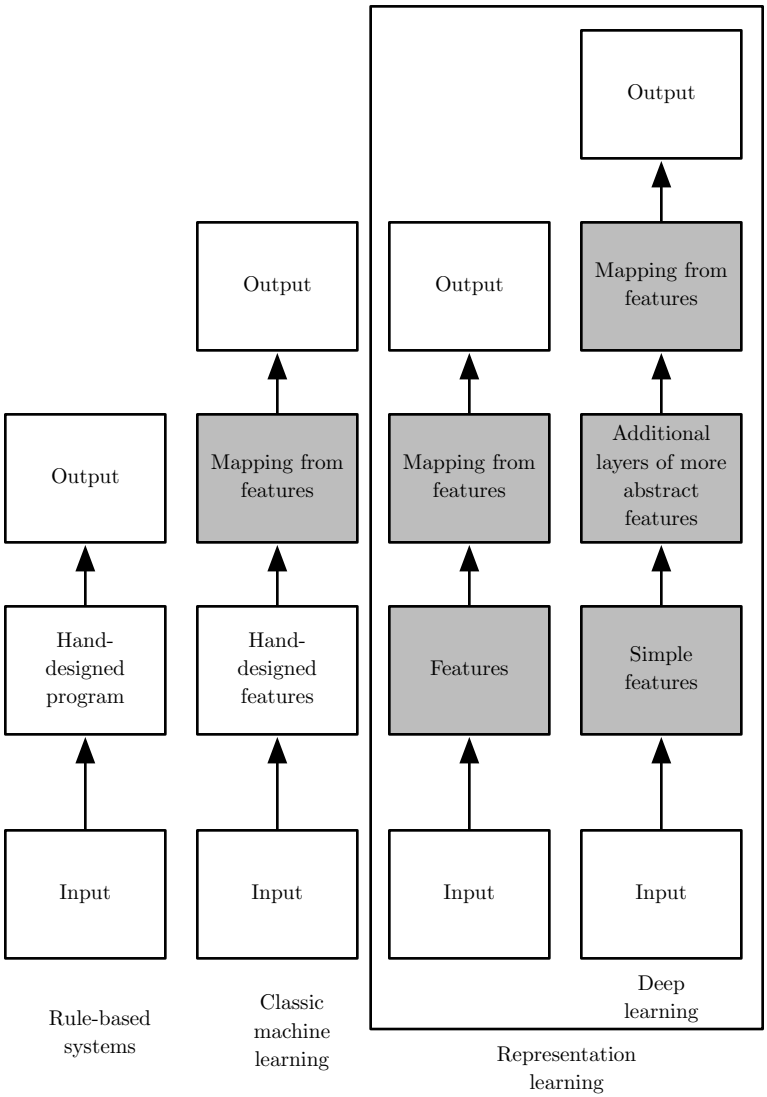

图 5: 流程图展示了AI系统的不同部分如何在不同的AI学科中彼此相关。阴影框表示能从数据中学习的组件。

1 本书面向的读者

这本书对各类读者都有一定用处，但我们主要是为两类受众对象而写的。其中一类受众对象是学习机器学习的大学生（本科或研究生），包括那些已经开始职业生涯的深度学习和人工智能研究者。另一类受众对象是没有机器学习或统计背景但希望能快速地掌握这方面知识并在他们的产品或平台中使用深度学习的软件工程师。深度学习在许多软件领域都已被证明是有用的，包括计算机视觉、语音和音频处理、自然语言处理、机器人技术、生物信息学和化学、电子游戏、搜索引擎、网络广告和金融。

为了最好地服务各类读者，这本书被组织为三个部分。第一部分介绍基本的数学工具和机器学习的概念。第二部分介绍本质上已解决的技术和最成熟的深度学习算法。第三部分讨论某些具有展望性的想法，它们被广泛地认为是深度学习未来的研究重点。

读者可以随意跳过不感兴趣或与自己背景不相关的部分。熟悉线性代数、概率和基本机器学习概念的读者可以跳过第一部分，例如，当读者只是想实现一个能工作的系统则不需要阅读超出第二部分的内容。为了帮助读者选择章节，图6展示了这本书的高层组织结构的流程图。

我们假设所有读者都具备计算机科学背景。也假设读者熟悉编程，并且对计算的性能问题、复杂性理论、入门级微积分和一些图论术语有基本的了解。

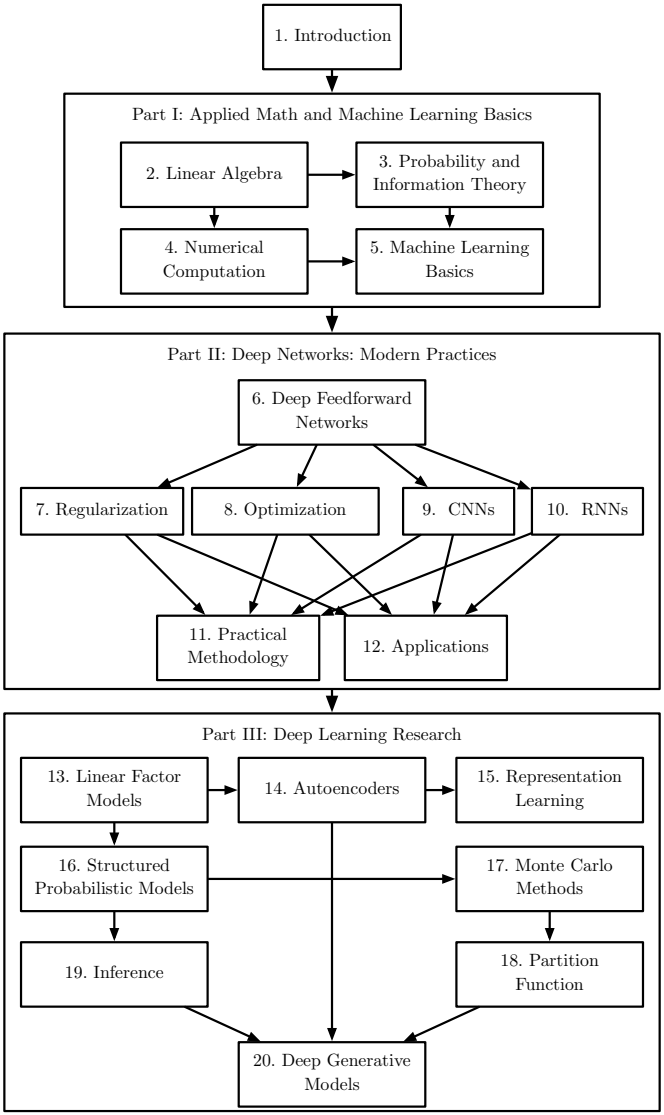

图 6: 本书的高层组织。从一章到另一章的箭头表示前一章是理解后一章的必备内容。

2 深度学习的历史趋势

通过历史背景了解深度学习是最简单的方式。这里我们仅指出深度学习的几个关键趋势，而不是提供其详细的历史：

- 深度学习有着悠久而丰富的历史，但随着许多不同哲学观点的渐渐消逝，与之对应的名称也渐渐尘封。
- 随着可用的训练数据量不断增加，深度学习变得更加有用。
- 随着时间的推移，针对深度学习的计算机软硬件基础设施都有所改善，深度学习模型的规模也随之增长。
- 随着时间的推移，深度学习已经解决日益复杂的应用，并且精度不断提高。

2.1 神经网络的众多名称和命运变迁

我们期待这本书的许多读者都听说过深度学习这一激动人心的新技术，并对一本书提及一个新兴领域的 "历史" 而感到惊讶。事实上，深度学习的历史可以追溯到 20 世纪 40 年代。深度学习*看似*是一个全新的领域，只不过因为在目前流行的前几年它是相对冷门的，同时也因为它被赋予了许多不同的名称（其中大部分已经不再使用），最近才成为众所周知的 "深度学习"。这个领域已经更换了很多名称，它反映了不同的研究人员和不同观点的影响。

全面地讲述深度学习的历史超出了本书的范围。然而，一些基本的背景对理解深度学习是有用的。一般来说，目前为止深度学习已经经历了三次发展浪潮：20 世纪 40 年代到 60 年代深度学习的雏形出现在**控制论** (cybernetics) 中，20 世纪 80 年代到 90 年代深度学习表现为**联结主义** (connectionism)，直到 2006 年，才真正以深度学习之名复兴。图7给出了定量的展示。

我们今天知道的一些最早的学习算法，是旨在模拟生物学习的计算模型，即大脑怎样学习或为什么能学习的模型。其结果是深度学习以**人工神经网络** (artificial neural network, ANN) 之名而淡去。彼时，深度学习模型被认为是受生物大脑（无论人类大脑或其他动物的大脑）所启发而设计出来的系统。尽管有些机器学习的神经网络有时被用来理解大脑功能 (Hinton and Shallice, 1991)，但它们一般都没有被设计成生物功能的真实模型。深度学习的神经观点受两个主要思想启发。一个想法是大脑作为例子证明智能行为是可能的，因此，概念上，建立智能的直接途径是逆向大脑背后的计算原理，并复制其功能。另一种看法是，理解大脑和人类智能背后的原理也非常有趣，因此机器学习模型除了解决工程应用的能力，如果能让人类对这些基本的科学问题有进一步的认识也将会很有用。

现代术语 "深度学习" 超越了目前机器学习模型的神经科学观点。学习*多层次组合*这一更普遍的原则更加吸引人，这可以应用于机器学习框架而不必受神经系统启发。

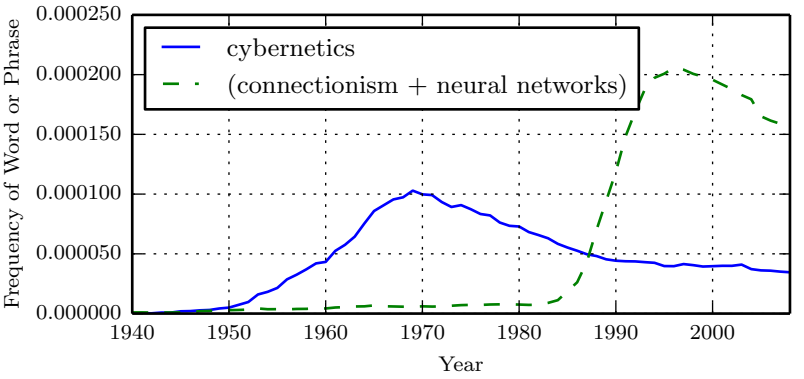

图 7: 根据 Google 图书中短语 “控制论”、“联结主义” 或 “神经网络” 频率衡量的人工神经网络研究的历史浪潮（图中展示了三次浪潮的前两次，第三次最近才出现）。第一次浪潮开始于 20 世纪 40 年代到 20 世纪 60 年代的控制论，随着生物学习理论的发展 (McCulloch and Pitts, 1943; Hebb, 1949) 和第一个模型的实现（如感知机 (Rosenblatt, 1958)），能实现单个神经元的训练。第二次浪潮开始于 1980-1995 年间的联结主义方法，可以使用反向传播 (Rumelhart *et al.*, 1986a) 训练具有一两个隐藏层的神经网络。当前第三次浪潮，也就是深度学习，大约始于 2006 年 (Hinton *et al.*, 2006; Bengio *et al.*, 2007; Ranzato *et al.*, 2007a)，并且现在在 2016 年以书的形式出现。另外两次浪潮类似地出现在书中的时间比相应的科学活动晚得多。

现代深度学习的最早前身是从神经科学的角度出发的简单线性模型。这些模型被设计为使用一组 n 个输入 $x_1, \ldots, x_n$ 并将它们与一个输出 y 相关联。这些模型希望学习一组权重 $w_1, \ldots, w_n$，并计算它们的输出 $f(\boldsymbol{x}, \boldsymbol{w}) = x_1 w_1 + \cdots + x_n w_n$。如图7所示，这第一波神经网络研究浪潮被称为控制论。

McCulloch-Pitts神经元 (McCulloch and Pitts, 1943) 是脑功能的早期模型。该线性模型通过检验函数 $f(\boldsymbol{x}, \boldsymbol{w})$ 的正负来识别两种不同类别的输入。显然，模型的权重需要正确设置后才能使模型的输出对应于期望的类别。这些权重可以由操作人员设定。在 20 世纪 50 年代，感知机 (Rosenblatt, 1956, 1958) 成为第一个能根据每个类别的输入样本来学习权重的模型。约在同一时期，**自适应线性单元** (adaptive linear element, ADALINE) 简单地返回函数 $f(\boldsymbol{x})$ 本身的值来预测一个实数 (Widrow and Hoff, 1960)，并且它还可以学习从数据预测这些数。

这些简单的学习算法大大影响了机器学习的现代景象。用于调节 ADALINE 权重的训练算法是被称为**随机梯度下降** (stochastic gradient descent) 的一种特例。稍加改进后的随机梯度下降算法仍然是当今深度学习的主要训练算法。

基于感知机和 ADALINE 中使用的函数 $f(\boldsymbol{x}, \boldsymbol{w})$ 的模型被称为**线性模型** (linear model)。尽管在许多情况下，这些模型以不同于原始模型的方式进行训练，但仍是目前最广泛使用的机器学习模型。

线性模型有很多局限性。最著名的是，它们无法学习异或（XOR）函数，即 $f([0,1], \boldsymbol{w}) = 1$ 和 $f([1,1], \boldsymbol{w}) = 1$，但 $f([1,1], \boldsymbol{w}) = 0$ 和 $f([0,0], \boldsymbol{w}) = 0$。观察到线性模型这个缺陷的批评者对受生物学启发的学习普遍地产生了抵触 (Minsky and Papert, 1969)。这导致了神经网络热潮的第一次大衰退。

现在，神经科学被视为深度学习研究的一个重要灵感来源，但它已不再是该领域的主要指导。

如今神经科学在深度学习研究中的作用被削弱，主要原因是我们根本没有足够的关于大脑的信息来作为指导去使用它。要获得对被大脑实际使用算法的深刻理解，我们需要有能力同时监测（至少是）数千相连神经元的活动。我们不能够做到这一点，所以我们甚至连大脑最简单、最深入研究的部分都还远远没有理解 (Olshausen and Field, 2005)。

神经科学已经给了我们依靠单一深度学习算法解决许多不同任务的理由。神经学家们发现，如果将雪貂的大脑重新连接，使视觉信号传送到听觉区域，它们可以学会用大脑的听觉处理区域去“看”(Von Melchner *et al.*, 2000)。这暗示着大多数哺乳动物的大脑能够使用单一的算法就可以解决其大脑可以解决的大部分不同任务。在这个假设之前，机器学习研究是比较分散的，研究人员在不同的社群研究自然语言处理、计算机视觉、运动规划和语音识别。如今，这些应用社群仍然是独立的，但是对于深度学习研究团体来说，同时研究许多或甚至所有这些应用领域是很常见的。

我们能够从神经科学得到一些粗略的指南。仅通过计算单元之间的相互作用而变得智能的基本思想是受大脑启发的。新认知机 (Fukushima, 1980) 受哺乳动物视觉系统的结构启发，引入了一个处理图片的强大模型架构，它后来成为了现代卷积网络的基础 (LeCun *et al.*, 1998)（我们将会在之后的章节看到）。目前大多数神经网络是基于一个称为**整流线性单元** (rectified linear unit) 的神经单元模型。原始认知机 (Fukushima, 1975) 受我们关于大脑功能知识的启发，引入了一个更复杂的版本。简化的现代版通过吸收来自不同观点的思想而形成，Nair and Hinton (2010) 和Glorot *et al.* (2011) 援引神经科学作为影响，Jarrett *et al.* (2009a) 援引更多面向工程的影响。虽然神经科学是灵感的重要来源，但它不需要被视为刚性指导。我们知道，真实的神经元计算着与现代整流线性单元非常不同的函数，但更接近真实神经网络的系统并没有导致机器学习性能的提升。此外，虽然神经科学已经成功地启发了一些神经网络架构，但我们对用于神经科学的生物学习还没有足够多的了解，因此也就不能为训练这些架构用的学习算法提供太多的借鉴。

媒体报道经常强调深度学习与大脑的相似性。的确，深度学习研究者比其他机器学

习领域（如核方法或贝叶斯统计）的研究者更可能地引用大脑作为影响，但是大家不应该认为深度学习在尝试模拟大脑。现代深度学习从许多领域获取灵感，特别是应用数学的基本内容如线性代数、概率论、信息论和数值优化。尽管一些深度学习的研究人员引用神经科学作为灵感的重要来源，然而其他学者完全不关心神经科学。

值得注意的是，了解大脑是如何在算法层面上工作的尝试确实存在且发展良好。这项尝试主要被称为“计算神经科学”，并且是独立于深度学习的领域。研究人员在两个领域之间来回研究是很常见的。深度学习领域主要关注如何构建计算机系统，从而成功解决需要智能才能解决的任务，而计算神经科学领域主要关注构建大脑如何真实工作的比较精确的模型。

在 20 世纪 80 年代，神经网络研究的第二次浪潮在很大程度上是伴随一个被称为**联结主义** (connectionism) 或**并行分布处理** (parallel distributed processing) 潮流而出现的 (Rumelhart *et al.*, 1986d; McClelland *et al.*, 1995)。联结主义是在认知科学的背景下出现的。认知科学是理解思维的跨学科途径，即它融合多个不同的分析层次。在 20 世纪 80 年代初期，大多数认知科学家研究符号推理模型。尽管这很流行，但符号模型很难解释大脑如何真正使用神经元实现推理功能。联结主义者开始研究真正基于神经系统实现的认知模型 (Touretzky and Minton, 1985)，其中很多复苏的想法可以追溯到心理学家Donald Hebb在 20 世纪 40 年代的工作 (Hebb, 1949)。

联结主义的中心思想是，当网络将大量简单的计算单元连接在一起时可以实现智能行为。这种见解同样适用于生物神经系统中的神经元，因为它和计算模型中隐藏单元起着类似的作用。

在上世纪 80 年代的联结主义期间形成的几个关键概念在今天的深度学习中仍然是非常重要的。

其中一个概念是**分布式表示** (distributed representation)(Hinton *et al.*, 1986)。其思想是：系统的每一个输入都应该由多个特征表示，并且每一个特征都应该参与到多个可能输入的表示。例如，假设我们有一个能够识别红色、绿色、或蓝色的汽车、卡车和鸟类的视觉系统，表示这些输入的其中一个方法是将九个可能的组合：红卡车，红汽车，红鸟，绿卡车等等使用单独的神经元或隐藏单元激活。这需要九个不同的神经元，并且每个神经必须独立地学习颜色和对象身份的概念。改善这种情况的方法之一是使用分布式表示，即用三个神经元描述颜色，三个神经元描述对象身份。这仅仅需要 6 个神经元而不是 9 个，并且描述红色的神经元能够从汽车、卡车和鸟类的图像中学习红色，而不仅仅是从一个特定类别的图像中学习。分布式表示的概念是本书的核心，我们将在之后的章节中更加详细地描述。

联结主义潮流的另一个重要成就是反向传播在训练具有内部表示的深度神经网络中

的成功使用以及反向传播算法的普及 (Rumelhart *et al.*, 1986c; LeCun, 1987)。这个算法虽然曾黯然失色不再流行，但截至写书之时，它仍是训练深度模型的主导方法。

在 20 世纪 90 年代，研究人员在使用神经网络进行序列建模的方面取得了重要进展。Hochreiter (1991) 和Bengio *et al.* (1994) 指出了对长序列进行建模的一些根本性数学难题，这将在之后的章节中描述。Hochreiter and Schmidhuber (1997) 引入**长短期记忆** (long short-term memory, LSTM) 网络来解决这些难题。如今，LSTM在许多序列建模任务中广泛应用，包括 Google 的许多自然语言处理任务。

神经网络研究的第二次浪潮一直持续到上世纪 90 年代中期。基于神经网络和其他AI技术的创业公司开始寻求投资，其做法野心勃勃但不切实际。当AI研究不能实现这些不合理的期望时，投资者感到失望。同时，机器学习的其他领域取得了进步。比如，核方法 (Boser *et al.*, 1992; Cortes and Vapnik, 1995; Schölkopf *et al.*, 1999) 和图模型 (Jordan, 1998) 都在很多重要任务上实现了很好的效果。这两个因素导致了神经网络热潮的第二次衰退，并一直持续到 2007 年。

在此期间，神经网络继续在某些任务上获得令人印象深刻的表现 (LeCun *et al.*, 1998; Bengio *et al.*, 2001)。加拿大高级研究所（CIFAR）通过其神经计算和自适应感知（NCAP）研究计划帮助维持神经网络研究。该计划联合了分别由Geoffrey Hinton、Yoshua Bengio和Yann LeCun领导的多伦多大学、蒙特利尔大学和纽约大学的机器学习研究小组。这个多学科的 CIFAR NCAP 研究计划还囊括了神经科学家、人类和计算机视觉专家。

在那个时候，人们普遍认为深度网络是难以训练的。现在我们知道，20 世纪 80 年代就存在的算法能工作得非常好，但是直到在 2006 年前后都没有体现出来。这可能仅仅由于其计算代价太高，而以当时可用的硬件难以进行足够的实验。

神经网络研究的第三次浪潮始于 2006 年的突破。Geoffrey Hinton表明名为深度信念网络的神经网络可以使用一种称为贪婪逐层预训练的策略来有效地训练 (Hinton *et al.*, 2006)，我们将在之后的章节中更详细地描述。其他 CIFAR 附属研究小组很快表明，同样的策略可以被用来训练许多其他类型的深度网络 (Bengio and LeCun, 2007a; Ranzato *et al.*, 2007b)，并能系统地帮助提高在测试样例上的泛化能力。神经网络研究的这一次浪潮普及了“深度学习”这一术语的使用，强调研究者现在有能力训练以前不可能训练的比较深的神经网络，并着力于深度的理论重要性上 (Bengio and LeCun, 2007b; Delalleau and Bengio, 2011; Pascanu *et al.*, 2014; Montufar *et al.*, 2014)。此时，深度神经网络已经优于与之竞争的基于其他机器学习技术以及手工设计功能的AI系统。在写这本书的时候，神经网络的第三次发展浪潮仍在继续，尽管深度学习的研究重点在这一段时间内发生了巨大变化。第三次浪潮已开始着眼于新的无监督学习技术和深度模型在小数据集的

泛化能力，但目前更多的兴趣点仍是比较传统的监督学习算法和深度模型充分利用大型标注数据集的能力。

2.2 与日俱增的数据量

人们可能想问，既然人工神经网络的第一个实验在 20 世纪 50 年代就完成了，但为什么深度学习直到最近才被认为是关键技术。自 20 世纪 90 年代以来，深度学习就已经成功用于商业应用，但通常被视为是一种艺术而不是一种技术，且只有专家可以使用的艺术，这种观点持续到最近。确实，要从一个深度学习算法获得良好的性能需要一些技巧。幸运的是，随着训练数据的增加，所需的技巧正在减少。目前在复杂的任务达到人类水平的学习算法，与 20 世纪 80 年代努力解决玩具问题 (toy problem) 的学习算法几乎是一样的，尽管我们使用这些算法训练的模型经历了变革，即简化了极深架构的训练。最重要的新进展是现在我们有了这些算法得以成功训练所需的资源。图8展示了基准数据集的大小如何随着时间的推移而显著增加。这种趋势是由社会日益数字化驱动的。由于我们的活动越来越多发生在计算机上，我们做什么也越来越多地被记录。由于我们的计算机越来越多地联网在一起，这些记录变得更容易集中管理，并更容易将它们整理成适于机器学习应用的数据集。因为统计估计的主要负担（观察少量数据以在新数据上泛化）已经减轻，‘‘大数据’’ 时代使机器学习更加容易。截至 2016 年，一个粗略的经验法则是，监督深度学习算法在每类给定约 5000 个标注样本情况下一般将达到可以接受的性能，当至少有 1000 万个标注样本的数据集用于训练时，它将达到或超过人类表现。此外，在更小的数据集上获得成功是一个重要的研究领域，为此我们应特别侧重于如何通过无监督或半监督学习充分利用大量的未标注样本。

2.3 与日俱增的模型规模

20 世纪 80 年代，神经网络只能取得相对较小的成功，而现在神经网络非常成功的另一个重要原因是我们现在拥有的计算资源可以运行更大的模型。联结主义的主要见解之一是，当动物的许多神经元一起工作时会变得聪明。单独神经元或小集合的神经元不是特别有用。

生物神经元不是特别稠密地连接在一起。如图10所示，几十年来，我们的机器学习模型中每个神经元的连接数量已经与哺乳动物的大脑在同一数量级上。

如图11所示，就神经元的总数目而言，直到最近神经网络都是惊人的小。自从隐藏单元引入以来，人工神经网络的规模大约每 2.4 年扩大一倍。这种增长是由更大内存、更快的计算机和更大的可用数据集驱动的。更大的网络能够在更复杂的任务中实现更高的

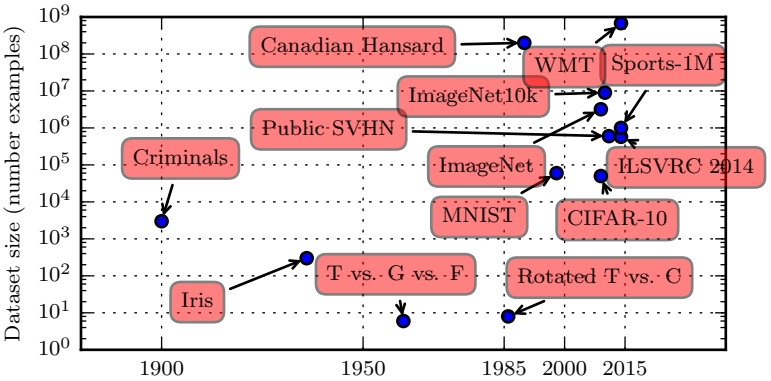

图 8: 与日俱增的数据量。20 世纪初，统计学家使用数百或数千的手动制作的度量来研究数据集 (Garson, 1900; Gosset, 1908; Anderson, 1935; Fisher, 1936)。20 世纪 50 年代到 80 年代，受生物启发的机器学习开拓者通常使用小的合成数据集，如低分辨率的字母位图，设计为在低计算成本下表明神经网络能够学习特定功能 (Widrow and Hoff, 1960; Rumelhart *et al.*, 1986b)。20 世纪 80 年代和 90 年代，机器学习变得更加统计，并开始利用包含成千上万个样本的更大数据集，如手写扫描数字的 MNIST 数据集（如图9）所示 (LeCun *et al.*, 1998)。在 21 世纪初的第一个十年，相同大小更复杂的数据集持续出现，如 CIFAR-10 数据集 (Krizhevsky and Hinton, 2009) 。在这十年结束和下五年，明显更大的数据集（包含数万到数千万的样例）完全改变了深度学习的可能实现的事。这些数据集包括公共 Street View House Numbers 数据集 (Netzer *et al.*, 2011)、各种版本的 ImageNet 数据集 (Deng *et al.*, 2009, 2010a; Russakovsky *et al.*, 2014a) 以及 Sports-1M 数据集 (Karpathy *et al.*, 2014)。在图顶部，我们看到翻译句子的数据集通常远大于其他数据集，如根据 Canadian Hansard 制作的 IBM 数据集 (Brown *et al.*, 1990) 和 WMT 2014 英法数据集 (Schwenk, 2014) 。

精度。这种趋势看起来将持续数十年。除非有能力迅速扩展的新技术，否则至少要到 21 世纪 50 年代，人工神经网络将才能具备与人脑相同数量级的神经元。生物神经元表示的功能可能比目前的人工神经元所表示的更复杂，因此生物神经网络可能比图中描绘的甚至要更大。

现在看来，其神经元比一个水蛭还少的神经网络不能解决复杂的人工智能问题是不足为奇的。即使现在的网络，从计算系统角度来看它可能相当大的，但实际上它比相对原始的脊椎动物如青蛙的神经系统还要小。

由于更快的 CPU、通用 GPU 的出现（在之后的章节中讨论）、更快的网络连接和更好的分布式计算的软件基础设施，模型规模随着时间的推移不断增加是深度学习历史中最重要的趋势之一。普遍预计这种趋势将很好地持续到未来。

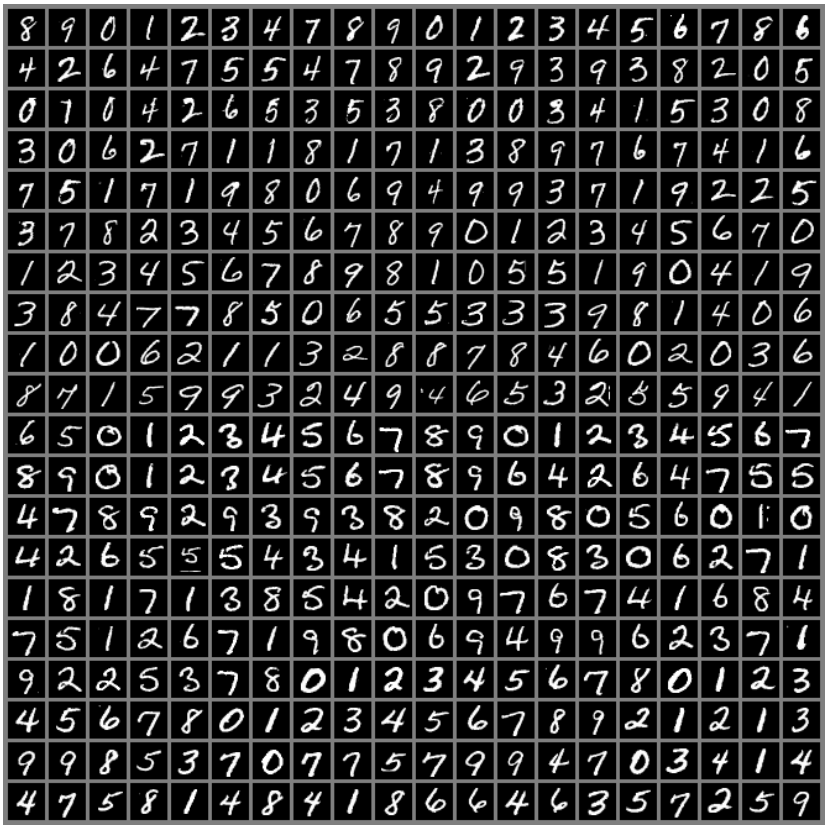

图 9: MNIST 数据集的输入样例。“NIST’’ 代表国家标准和技术研究所 (National Institute of Standards and Technology)，是最初收集这些数据的机构。“M’’ 代表 ‘‘修改的 (Modified)’’，为更容易地与机器学习算法一起使用，数据已经过预处理。MNIST 数据集包括手写数字的扫描和相关标签（描述每个图像中包含 0-9 中哪个数字）。这个简单的分类问题是深度学习研究中最简单和最广泛使用的测试之一。尽管现代技术很容易解决这个问题，它仍然很受欢迎。Geoffrey Hinton 将其描述为 ‘‘机器学习的果蝇’’，这意味着机器学习研究人员可以在受控的实验室条件下研究他们的算法，就像生物学家经常研究果蝇一样。

2.4 与日俱增的精度、复杂度和对现实世界的冲击

20 世纪 80 年代以来，深度学习提供精确识别和预测的能力一直在提高。而且，深度学习持续成功地被应用于越来越广泛的实际问题中。

最早的深度模型被用来识别裁剪紧凑且非常小的图像中的单个对象 (Rumelhart *et al.*, 1986d)。此后，神经网络可以处理的图像尺寸逐渐增加。现代对象识别网络能处

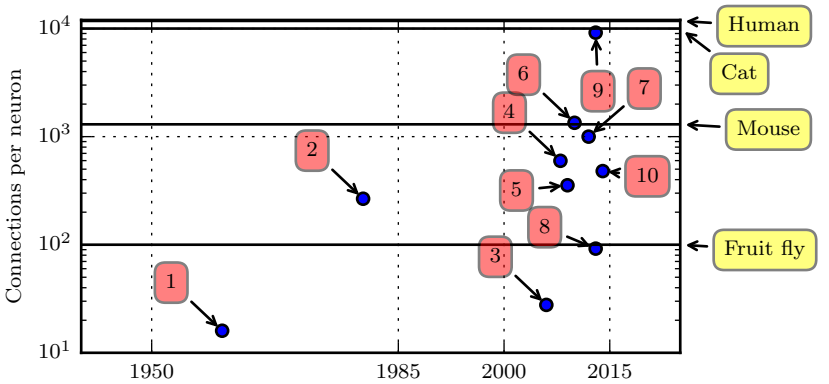

图 10: 与日俱增的每神经元连接数。最初，人工神经网络中神经元之间的连接数受限于硬件能力。而现在，神经元之间的连接数大多是出于设计考虑。一些人工神经网络中每个神经元的连接数与猫一样多，并且对于其他神经网络来说，每个神经元的连接与较小哺乳动物（如小鼠）一样多是非常普遍的。甚至人类大脑每个神经元的连接也没有过高的数量。生物神经网络规模来自Wikipedia (2015)。

1. 自适应线性单元 (Widrow and Hoff, 1960)
2. 神经认知机 (Fukushima, 1980)
3. GPU-加速 卷积网络 (Chellapilla *et al.*, 2006)
4. 深度玻尔兹曼机 (Salakhutdinov and Hinton, 2009)
5. 无监督卷积网络 (Jarrett *et al.*, 2009b)
6. GPU-加速 多层感知机 (Ciresan *et al.*, 2010)
7. 分布式自编码器 (Le *et al.*, 2012)
8. Multi-GPU 卷积网络 (Krizhevsky *et al.*, 2012a)
9. COTS HPC 无监督卷积网络 (Coates *et al.*, 2013)
10. GoogLeNet (Szegedy *et al.*, 2014)

理丰富的高分辨率照片，并且不需要在被识别的对象附近进行裁剪 (Krizhevsky *et al.*, 2012b)。类似地，最早的网络只能识别两种对象（或在某些情况下，单类对象的存在与否），而这些现代网络通常能够识别至少1000个不同类别的对象。对象识别中最大的比赛是每年举行的 ImageNet 大型视觉识别挑战（ILSVRC）。深度学习迅速崛起的激动人心的一幕是卷积网络第一次大幅赢得这一挑战，它将最高水准的前 5 错误率从26.1%降到15.3%(Krizhevsky *et al.*, 2012b)，这意味着该卷积网络针对每个图像的可能类别生成一个顺序列表，除了 15.3% 的测试样本，其他测试样本的正确类标都出现在此列表中的前 5 项里。此后，深度卷积网络连续地赢得这些比赛，截至写本书时，深度学习的最新结果将这个比赛中的前 5 错误率降到了3.6%，如图12所示。

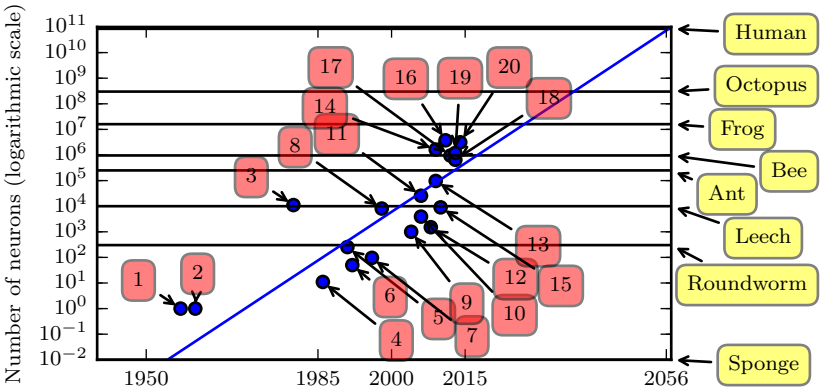

图 11: 与日俱增的神经网络规模。自从引入隐藏单元，人工神经网络的大小大约每 2.4 年翻一倍。生物神经网络规模来自Wikipedia (2015)。

1. 感知机 (Rosenblatt, 1958, 1962)
2. 自适应线性单元 (Widrow and Hoff, 1960)
3. 神经认知机 (Fukushima, 1980)
4. 早期后向传播网络 (Rumelhart *et al.*, 1986b)
5. 用于语音识别的循环神经网络 (Robinson and Fallside, 1991)
6. 用于语音识别的多层感知机 (Bengio *et al.*, 1991)
7. 均匀场sigmoid信念网络 (Saul *et al.*, 1996)
8. LeNet-5 (LeCun *et al.*, 1998)
9. 回声状态网络 (Jaeger and Haas, 2004)
10. 深度信念网络 (Hinton *et al.*, 2006)
11. GPU-加速卷积网络 (Chellapilla *et al.*, 2006)
12. 深度玻尔兹曼机 (Salakhutdinov and Hinton, 2009)
13. GPU-加速深度信念网络 (Raina *et al.*, 2009)
14. 无监督卷积网络 (Jarrett *et al.*, 2009b)
15. GPU-加速多层感知机 (Ciresan *et al.*, 2010)
16. OMP-1 网络 (Coates and Ng, 2011)
17. 分布式自编码器 (Le *et al.*, 2012)
18. Multi-GPU卷积网络 (Krizhevsky *et al.*, 2012a)
19. COTS HPC 无监督卷积网络 (Coates *et al.*, 2013)
20. GoogLeNet (Szegedy *et al.*, 2014)

深度学习也对语音识别产生了巨大影响。语音识别在 20 世纪 90 年代得到提高后，直到约 2000 年都停滞不前。深度学习的引入 (Dahl *et al.*, 2010; Deng *et al.*, 2010b; Seide *et al.*, 2011; Hinton *et al.*, 2012) 使得语音识别错误率陡然下降，有些错误率甚至降低了一半。我们将在之后的章节更详细地探讨这个历史。

深度网络在行人检测和图像分割中也取得了引人注目的成功 (Sermanet *et al.*, 2013;

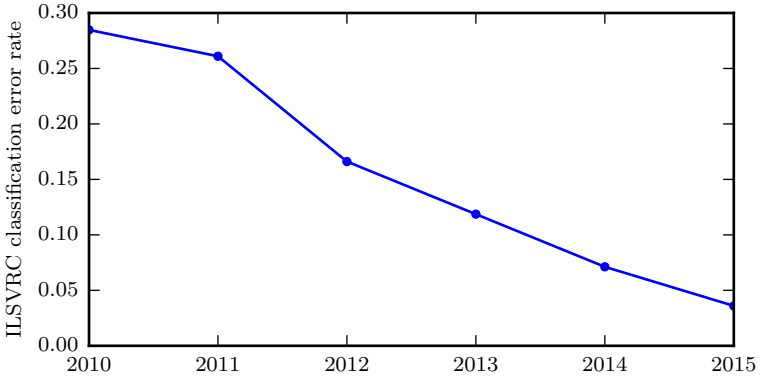

图 12: 日益降低的错误率。由于深度网络达到了在 ImageNet 大规模视觉识别挑战中竞争所必需的规模，它们每年都能赢得胜利，并且产生越来越低的错误率。数据来源于 Russakovsky *et al.* (2014b) 和He *et al.* (2015)。

Farabet *et al.*, 2013; Couprie *et al.*, 2013)，并且在交通标志分类上取得了超越人类的表现 (Ciresan *et al.*, 2012)。

在深度网络的规模和精度有所提高的同时，它们可以解决的任务也日益复杂。Goodfellow *et al.* (2014) 表明，神经网络可以学习输出描述图像的整个字符序列，而不是仅仅识别单个对象。此前，人们普遍认为，这种学习需要对序列中的单个元素进行标注 (Gulcehre and Bengio, 2013)。循环神经网络，如之前提到的LSTM序列模型，现在用于对序列和其他序列之间的关系进行建模，而不是仅仅固定输入之间的关系。这种序列到序列的学习似乎引领着另一个应用的颠覆性发展，即机器翻译 (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2015)。

这种复杂性日益增加的趋势已将其推向逻辑结论，即神经图灵机 (Graves *et al.*, 2014) 的引入，它能学习读取存储单元和向存储单元写入任意内容。这样的神经网络可以从期望行为的样本中学习简单的程序。例如，从杂乱和排好序的样本中学习对一系列数进行排序。这种自我编程技术正处于起步阶段，但原则上未来可以适用于几乎所有的任务。

深度学习的另一个最大的成就是其在**强化学习** (reinforcement learning) 领域的扩展。在强化学习中，一个自主的智能体必须在没有人类操作者指导的情况下，通过试错来学习执行任务。DeepMind 表明，基于深度学习的强化学习系统能够学会玩 Atari 视频游戏，并在多种任务中可与人类匹敌 (Mnih *et al.*, 2015)。深度学习也显著改善了机器人强化学习的性能 (Finn *et al.*, 2015)。

许多深度学习应用都是高利润的。现在深度学习被许多顶级的技术公司使用，包括 Google、Microsoft、Facebook、IBM、Baidu、Apple、Adobe、Netflix、NVIDIA 和 NEC 等。

深度学习的进步也严重依赖于软件基础架构的进展。软件库如 Theano(Bergstra *et al.*, 2010; Bastien *et al.*, 2012)、PyLearn2(Goodfellow *et al.*, 2013)、Torch(Collobert *et al.*, 2011)、DistBelief(Dean *et al.*, 2012)、Caffe(Jia, 2013)、MXNet(Chen *et al.*, 2015) 和 TensorFlow(Abadi *et al.*, 2015) 都能支持重要的研究项目或商业产品。

深度学习也为其他科学做出了贡献。用于对象识别的现代卷积网络为神经科学家们提供了可以研究的视觉处理模型 (DiCarlo, 2013)。深度学习也为处理海量数据以及在科学领域作出有效的预测提供了非常有用的工具。它已成功地用于预测分子如何相互作用从而帮助制药公司设计新的药物 (Dahl *et al.*, 2014)，搜索亚原子粒子 (Baldi *et al.*, 2014)，以及自动解析用于构建人脑三维图的显微镜图像 (Knowles-Barley *et al.*, 2014) 等。我们期待深度学习未来能够出现在越来越多的科学领域中。

总之，深度学习是机器学习的一种方法。在过去几十年的发展中，它大量借鉴了我们关于人脑、统计学和应用数学的知识。近年来，得益于更强大的计算机、更大的数据集和能够训练更深网络的技术，深度学习的普及性和实用性都有了极大的发展。未来几年充满了进一步提高深度学习并将它带到新领域的挑战和机遇。

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Anderson, E. (1935). The Irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, **59**, 2–5.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *ICLR'2015, arXiv:1409.0473*.

Baldi, P., Sadowski, P., and Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, **5**.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D., and Bengio, Y. (2012). Theano: new features and speed improvements. Submited to the Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, http://www.iro.umontreal.ca/ lisa/publications2/index.php/publications/show/551.

Bengio, Y. and LeCun, Y. (2007a). Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*.

Bengio, Y. and LeCun, Y. (2007b). Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Machines*. MIT Press.

Bengio, Y., De Mori, R., Flammia, G., and Kompe, R. (1991). Phonetically motivated acoustic parameters for continuous speech recognition using artificial neural networks. In *Proceedings of EuroSpeech'91*.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, **5**(2), 157–166.

Bengio, Y., Ducharme, R., and Vincent, P. (2001). A neural probabilistic language model. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13 (NIPS'00)*, pages 933–938. MIT Press.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *NIPS'2006*.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010). Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*. Oral Presentation.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, New York, NY, USA. ACM.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, **16**(2), 79–85.

Chellapilla, K., Puri, S., and Simard, P. (2006). High Performance Convolutional Neural Networks for Document Processing. In Guy Lorette, editor, *Tenth International Workshop on Frontiers in Handwriting Recognition*, La Baule (France). Université de Rennes 1, Suvisoft. http://www.suvisoft.com.

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., and Zhang, Z. (2015). MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*.

Ciresan, D., Meier, U., Masci, J., and Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. *Neural Networks*, **32**, 333–338.

Ciresan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2010). Deep big simple neural nets for handwritten digit recognition. *Neural Computation*, **22**, 1–14.

Coates, A. and Ng, A. Y. (2011). The importance of encoding versus training with sparse coding and vector quantization. In *ICML'2011*.

Coates, A., Huval, B., Wang, T., Wu, D., Catanzaro, B., and Andrew, N. (2013). Deep learning with COTS HPC systems. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28 (3), pages 1337–1345. JMLR Workshop and Conference Proceedings.

Collobert, R., Kavukcuoglu, K., and Farabet, C. (2011). Torch7: A Matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, **20**, 273–297.

Couprie, C., Farabet, C., Najman, L., and LeCun, Y. (2013). Indoor semantic segmentation using depth information. In *International Conference on Learning Representations (ICLR2013)*.

Dahl, G. E., Ranzato, M., Mohamed, A., and Hinton, G. E. (2010). Phone recognition with the mean-covariance restricted Boltzmann machine. In *Advances in Neural Information Processing Systems (NIPS)*.

Dahl, G. E., Jaitly, N., and Salakhutdinov, R. (2014). Multi-task neural networks for QSAR predictions. arXiv:1406.1231.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., and Ng, A. Y. (2012). Large scale distributed deep networks. In *NIPS'2012*.

Delalleau, O. and Bengio, Y. (2011). Shallow vs. deep sum-product networks. In *NIPS*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Deng, J., Berg, A. C., Li, K., and Fei-Fei, L. (2010a). What does classifying more than 10,000 image categories tell us? In *Proceedings of the 11th European Conference on Computer Vision: Part V*, ECCV'10, pages 71–84, Berlin, Heidelberg. Springer-Verlag.

Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., and Hinton, G. (2010b). Binary coding of speech spectrograms using a deep auto-encoder. In *Interspeech 2010*, Makuhari, Chiba, Japan.

DiCarlo, J. J. (2013). Mechanisms underlying visual object recognition: Humans vs. neurons vs. machines. NIPS Tutorial.

Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(8), 1915–1929.

Finn, C., Tan, X. Y., Duan, Y., Darrell, T., Levine, S., and Abbeel, P. (2015). Learning visual feature spaces for robotic manipulation with deep spatial autoencoders. *arXiv preprint arXiv:1509.06113*.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, **20**, 121–136.

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

Garson, J. (1900). The metric system of identification of criminals, as used in Great Britain and Ireland. *The Journal of the Anthropological Institute of Great Britain and Ireland*, (2), 177–227.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In *AISTATS'2011*.

Goodfellow, I. J., Warde-Farley, D., Lamblin, P., Dumoulin, V., Mirza, M., Pascanu, R., Bergstra, J., Bastien, F., and Bengio, Y. (2013). Pylearn2: a machine learning research library. *arXiv preprint arXiv:1308.4214*.

Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., and Shet, V. (2014). Multi-digit number recognition from Street View imagery using deep convolutional neural networks. In *International Conference on Learning Representations*.

Gosset, W. S. (1908). The probable error of a mean. *Biometrika*, **6**(1), 1–25. Originally published under the pseudonym "Student".

Graves, A., Wayne, G., and Danihelka, I. (2014). Neural Turing machines. arXiv:1410.5401.

Gulcehre, C. and Bengio, Y. (2013). Knowledge matters: Importance of prior information for optimization. Technical Report arXiv:1301.4083, Universite de Montreal.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *arXiv preprint arXiv:1502.01852*.

Hebb, D. O. (1949). *The Organization of Behavior*. Wiley, New York.

Hinton, G., Deng, L., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, **29**(6), 82–97.

Hinton, G. E. and Shallice, T. (1991). Lesioning an attractor network: investigations of acquired dyslexia. *Psychological review*, **98**(1), 74.

Hinton, G. E., McClelland, J., and Rumelhart, D. (1986). Distributed representations. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 77–109. MIT Press, Cambridge.

Hinton, G. E., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, **18**, 1527–1554.

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**(8), 1735–1780.

Hsu, F.-H. (2002). *Behind Deep Blue: Building the Computer That Defeated the World Chess Champion*. Princeton University Press, Princeton, NJ, USA.

Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, **304**(5667), 78–80.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009a). What is the best multi-stage architecture for object recognition? In *Proc. International Conference on Computer Vision (ICCV'09)*, pages 2146–2153. IEEE.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009b). What is the best multi-stage architecture for object recognition? In *ICCV'09*.

Jia, Y. (2013). Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/.

Jordan, M. I. (1998). *Learning in Graphical Models*. Kluwer, Dordrecht, Netherlands.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*.

Knowles-Barley, S., Jones, T. R., Morgan, J., Lee, D., Kasthuri, N., Lichtman, J. W., and Pfister, H. (2014). Deep learning for the connectome. *GPU Technology Conference*.

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012a). ImageNet classification with deep convolutional neural networks. In *NIPS'2012*.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012b). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS'2012)*.

Le, Q., Ranzato, M., Monga, R., Devin, M., Corrado, G., Chen, K., Dean, J., and Ng, A. (2012). Building high-level features using large scale unsupervised learning. In *ICML'2012*.

LeCun, Y. (1987). *Modèles connexionistes de l'apprentissage*. Ph.D. thesis, Université de Paris VI.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient based learning applied to document recognition. *Proc. IEEE*.

Lenat, D. B. and Guha, R. V. (1989). *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc.

Linde, N. (1992). The machine that changed the world, episode 3. Documentary miniseries.

Lovelace, A. (1842). Notes upon L. F. Menabrea's "Sketch of the Analytical Engine invented by Charles Babbage".

McClelland, J., Rumelhart, D., and Hinton, G. (1995). The appeal of parallel distributed processing. In *Computation & intelligence*, pages 305–341. American Association for Artificial Intelligence.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, 115–133.

Minsky, M. L. and Papert, S. A. (1969). *Perceptrons*. MIT Press, Cambridge.

Mnih, V., Kavukcuoglo, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidgeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, **518**, 529–533.

Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *NIPS'2014*.

Mor-Yosef, S., Samueloff, A., Modan, B., Navot, D., and Schenker, J. G. (1990). Ranking the risk factors for cesarean: logistic regression analysis of a nationwide study. *Obstet Gynecol*, **75**(6), 944–7.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In L. Bottou and M. Littman, editors, *Proceedings of the Twenty-seventh International Conference on Machine Learning (ICML-10)*, pages 807–814. ACM.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. Deep Learning and Unsupervised Feature Learning Workshop, NIPS.

Olshausen, B. and Field, D. J. (2005). How close are we to understanding V1? *Neural Computation*, **17**, 1665–1699.

Ovid and Martin, C. (2004). *Metamorphoses*. W.W. Norton.

Pascanu, R., Gulcehre, C., Cho, K., and Bengio, Y. (2014). How to construct deep recurrent neural networks. In *ICLR*.

Raina, R., Madhavan, A., and Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. In L. Bottou and M. Littman, editors, *Proceedings of the Twenty-sixth International Conference on Machine Learning (ICML'09)*, pages 873–880, New York, NY, USA. ACM.

Ranzato, M., Poultney, C., Chopra, S., and LeCun, Y. (2007a). Efficient learning of sparse representations with an energy-based model. In *NIPS'2006*.

Ranzato, M., Poultney, C., Chopra, S., and LeCun, Y. (2007b). Efficient learning of sparse representations with an energy-based model. In B. Schölkopf, J. Platt, and T. Hoffman,

editors, *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pages 1137–1144. MIT Press.

Robinson, A. J. and Fallside, F. (1991). A recurrent error propagation network speech recognition system. *Computer Speech and Language*, **5**(3), 259–274.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386–408.

Rosenblatt, F. (1962). *Principles of Neurodynamics*. Spartan, New York.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, **27**(3), 832–837.

Rumelhart, D., Hinton, G., and Williams, R. (1986a). Learning representations by back-propagating errors. *Nature*, **323**, 533–536.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986b). Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, chapter 8, pages 318–362. MIT Press, Cambridge.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986c). Learning representations by back-propagating errors. *Nature*, **323**, 533–536.

Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (1986d). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2014a). ImageNet Large Scale Visual Recognition Challenge.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., *et al.* (2014b). Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*.

Salakhutdinov, R. and Hinton, G. (2009). Deep Boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 5, pages 448–455.

Saul, L. K., Jaakkola, T., and Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, **4**, 61–76.

Schölkopf, B., Burges, C. J. C., and Smola, A. J. (1999). *Advances in Kernel Methods — Support Vector Learning*. MIT Press, Cambridge, MA.

Schwenk, H. (2014). Cleaned subset of WMT '14 dataset.

Seide, F., Li, G., and Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. In *Interspeech 2011*, pages 437–440.

Sermanet, P., Kavukcuoglu, K., Chintala, S., and LeCun, Y. (2013). Pedestrian detection with unsupervised multi-stage feature learning. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR'13)*. IEEE.

Sparkes, B. (1996). *The Red and the Black: Studies in Greek Pottery*. Routledge.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS'2014, arXiv:1409.3215*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions. Technical report, arXiv:1409.4842.

Tandy, D. W. (1997). *Works and Days: A Translation and Commentary for the Social Sciences*. University of California Press.

Touretzky, D. S. and Minton, G. E. (1985). Symbols among the neurons: Details of a connectionist inference architecture. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'85, pages 238–243, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Von Melchner, L., Pallas, S. L., and Sur, M. (2000). Visual behaviour mediated by retinal projections directed to the auditory pathway. *Nature*, **404**(6780), 871–876.

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, volume 4, pages 96–104. IRE, New York.

Wikipedia (2015). List of animals by number of neurons — Wikipedia, the free encyclopedia. [Online; accessed 4-March-2015].

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *ECCV'14*.

术语

人工智能 artificial intelligence 2–10, 16, 18

人工神经网络 artificial neural network 12, 13, 20, 21

自编码器 autoencoder 4, 20, 21

信念网络 belief network 21

联结主义 connectionism 12, 13, 15, 17

卷积网络 convolutional network 20, 21

控制论 cybernetics 12, 13

解码器 decoder 4

深度信念网络 deep belief network 16, 21

深度玻尔兹曼机 Deep Boltzmann Machine 20, 21

深度学习 deep learning 2, 5, 7, 10–23

分布式表示 distributed representation 15

回声状态网络 echo state network 21

编码器 encoder 4

样本 example 13, 22

变差因素 factors of variation 4, 5

隐藏层 hidden layer 6, 13

隐藏单元 hidden unit 6, 15, 17, 21

推断 inference 3

知识库 knowledge base 3

线性模型 linear model 14