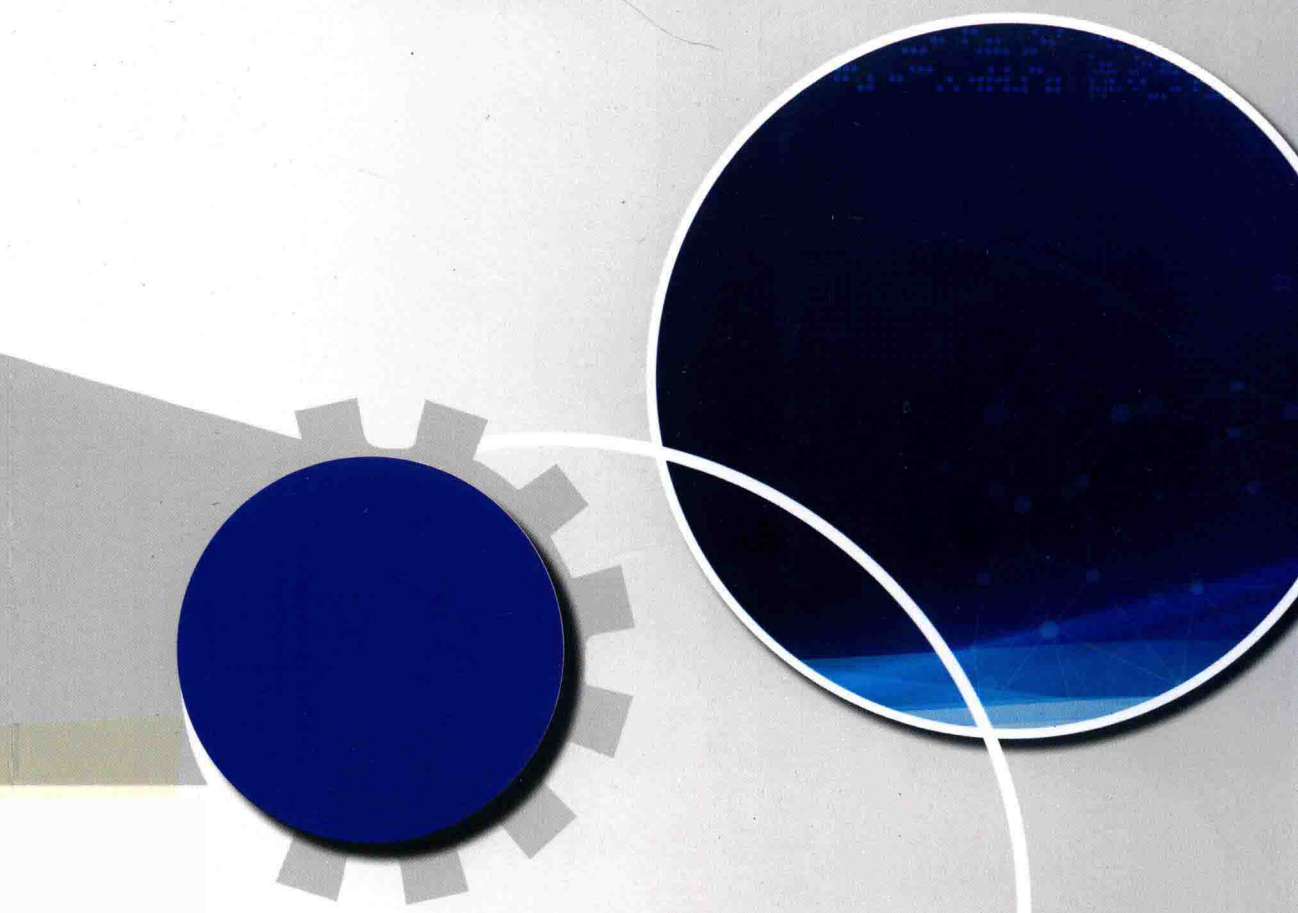
当代机器深度学习方法与应用研究

黄孝平 著

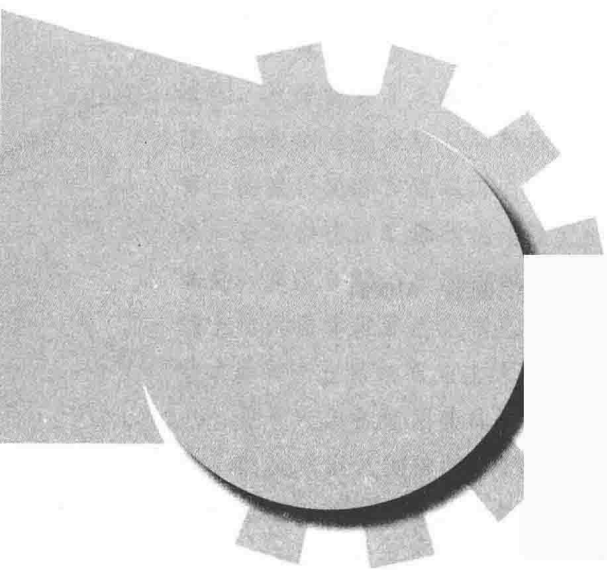
DANGDAI JIQI SHENDU XUEXI
FANGFA YU YINGYONG YANJIU



电子科技大学出版社
University of Electronic Science and Technology of China Press

当代机器深度学习方法与应用研究

黄孝平 著



电子科技大学出版社
University of Electronic Science and Technology of China Press

图书在版编目(CIP)数据

当代机器深度学习方法与应用研究/黄孝平著. --
成都：电子科技大学出版社，2017.11
ISBN 978-7-5647-5261-3

Ⅰ.①当… Ⅱ.①黄… Ⅲ.①机器学习-研究 Ⅳ.
①TP181

中国版本图书馆CIP数据核字（2017）第274607号

当代机器深度学习方法与应用研究

黄孝平 著

策划编辑 杜 倩 熊晶晶

责任编辑 熊晶晶

出版发行 电子科技大学出版社

成都市一环路东一段159号电子信息产业大厦九楼 邮编 610051

主 页 www.uestcp.com.cn

服务电话 028-83203399

邮购电话 028-83201495

印 刷 北京一鑫印务有限责任公司

成品尺寸 185mm × 260mm

印 张 17

字 数 399千字

版 次 2017年11月第一版

印 次 2017年11月第一次印刷

书 号 ISBN 978-7-5647-5261-3

定 价 61.00元

前 言

近年来，几乎整个智能学科的研究者们都注意到一个技术名词——深度学习（deep learning）这个略带神秘色彩的名字。而在此之前，包括Google、Microsoft、Facebook等公司在内的诸多信息科技巨头都已争相在此技术上投入了前所未有的重视力度和战略资源，继而高调宣布布局智能应用领域。学术界和工业界不遗余力地抢占相关研究和技术的制高点，人们并没有感到奇怪，因为所有人都明白：这也许是人类在探索人工智能的伟大旅程和漫漫征途上的重要一刻。

以深度学习为代表的机器学习是当前最接近人类大脑的智能学习方法和认知过程，充分借鉴了人脑的多分层结构、神经元的连接交互、分布式稀疏存储和表征、信息的逐层分析处理机制，自适应、自学习的强大并行信息处理能力，在语音、图像识别等方面取得了突破性进展，在诸多应用领域取得巨大商业成功。随着产业界数据量的爆炸式增长，大数据概念受到越来越多的关注。由于大数据的海量、复杂多样、变化快的特性，对于大数据环境下的应用问题，传统的在小数据上的机器学习算法很多已不再适用。因此，研究大数据环境下的深度学习成为学术界和产业界共同关注的话题。

关于人工神经网络的研究可以追溯到20世纪40年代。在其漫长的历史上经历了数次戏剧性的波折。然而近年来，随着大量数据的获得、先进理论的发现，以及高性能并行计算技术的发展，以深度神经网络为载体的特征学习技术相继在语音、视觉、语言等诸多研究领域中取得了突破性的成果，并且正以不可阻挡之势“入侵”传统技术占领的各个领域。

随着深度学习技术在学术界和工业界得到广泛认可，越来越多的人开始参与到深度学习的相关研究和实践中来。然而，由于存在一定的技术门槛，快速入手深度学习的研究并不是一件容易的事情。其中的一个重要原因是，深度学习中的许多问题非常依赖于实践。然而长期以来，学术界和工业界缺少一款专门为深度学习而设计的，兼具性能、灵活性和扩展性等诸多优势于一身的开源框架。这使得无论是快速实现算法，还是复现他人的结论，都存在着实践上的困难。研究人

员和工程师们迫切需要一套通用而高效的深度学习开源框架。

2006年从单隐层神经网络到深度神经网络模型，迎来了神经网络发展的又一高潮，深度学习及其应用受到了前所未有的重视与关注，世界迎来又一轮人工智能变革的高潮，从谷歌脑到中国脑科学计划，再到互联网+和中国人工智能2.0。深度学习是人工智能及机器学习的一个重要方向，在未来，它将会不断出现激动人心的理论进展和方法实践，深刻影响我们生活的方方面面。

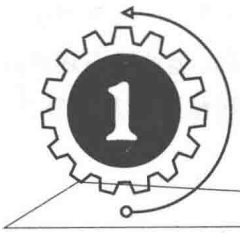
随着研究的不断深入，深度学习已经超越了目前机器学习模型的神经科学观点，学习多层次组合的这一设计原则更加吸引人。

众所周知，感知、认知和决策是衡量智能化的标准，充分发挥深度学习的感知能力和强化学习的决策能力，形成的深度强化学习已在众多应用问题上取得突破，如无人驾驶、计算机围棋程序和智能机器人等。在后深度学习时代，其核心在于生成数据、环境交互和领域迁移，对应着深度生成网络、深度强化学习和深度迁移学习将继续成为人工智能领域的研究热点。另外，根据数据的属性和操作的有效性，衍生的网络包括深度复数域神经网络（如深度复卷积神经网络）、深度二值神经网络和深度脉冲神经网络等。

由于时间的仓促，编者水平有限，本书难免存在不足之处，在此出版之际，我们真诚地希望读者对本书提出宝贵的意见和建议。

编　者

目 录



1 引　言

1.1 机器学习发展简史

机器学习最早可以追溯到对人工神经网络的研究。1943 年，Warren Mc Culloch 和 Walter Pitts 提出了神经网络层次结构模型，确立为神经网络的计算模型理论，从而为机器学习的发展奠定了基础。

1950 年，“人工智能之父”图灵提出了著名的“图灵测试”，使人工智能成为计算机科学领域的一个重要研究课题。

1957 年，康内尔大学教授 Frank Rosenblatt 提出 Perceptron 概念，并且首次用算法精确定义了自组织自学习的神经网络数学模型，设计出了第一个计算机神经网络，这个机器学习算法成为神经网络模型的开山鼻祖。

1959 年，美国 IBM 公司的 A.M.Samuel 设计了一个具有学习能力的跳棋程序，曾经战胜了美国一个保持 8 年不败的冠军。这个程序向人们初步展示了机器学习的能力。

1962 年，Hubel 和 Wiesel 发现猫脑皮层中独特的神经网络结构可以有效降低学习的复杂性，从而提出著名的 Hubel-Wiesel 生物视觉模型，以后提出的神经网络模型均受此启迪。

1969 年，人工智能研究的先驱者 Marvin Minsky 和 Seymour Papert 出版了对机器学习研究具有深远影响的著作 *Perceptron*，虽然提出的 XOR 问题把感知机研究送上不归路，此后的十几年基于神经网络的人工智能研究进入低潮，但是对于机器学习基本思想的论断——解决问题的算法能力和计算复杂性，影响深远，延续至今。

1980 年夏，在美国卡内基·梅隆大学举行了第一届机器学习国际研讨会，标志着机器学习研究在世界范围内兴起。

1986 年，*Machine Learning* 创刊，标志着机器学习逐渐为世人瞩目并开始加速发展。

1982 年，Hopfield 发表了一篇关于神经网络模型的论文，构造出能量函数并把这一概念引入 Hopfield 网络，同时通过对动力系统性质的认识，实现了 Hopfield 网络的最优化求解，推动

了神经网络的深入研究和发展应用。1986年，Rumelhart、Hinton和Williams联合在《自然》杂志发表了著名的反向传播（Backpropagatim，简称BP）算法，首次阐述了BP算法在浅层前向型神经网络模型的应用，如图1-1所示，不但明显降低了最优化问题求解的运算量，还通过增加一个隐层解决了感知器无法解决的XOR Gate难题，该算法成为神经网络的最基本算法。从此，神经网络的研究与应用开始复苏。

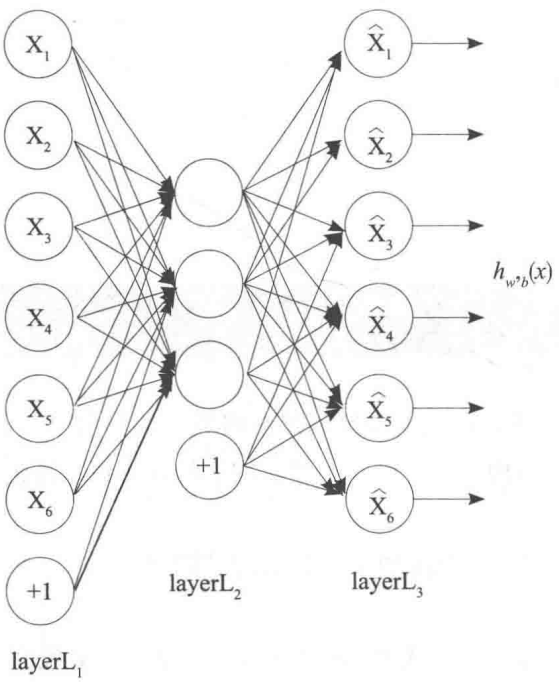


图1-1　浅层稀疏自编码器示意图

1989年，美国贝尔实验室学者Yann Le Cun教授提出了目前最为流行的卷积神经网络（Convolutional Neural Network，CNN）计算模型，推导出基于BP算法的高效训练方法，并成功地应用于英文手写体识别。CNN是第一个被成功训练的人工神经网络，也是后来深度学习最成功、应用最广泛的模型之一。

进入20世纪90年代后，多种浅层机器学习模型相继问世，诸如逻辑回归、支持向量机等，这些机器学习算法的共性是数学模型为凸代价函数的最优化问题，理论分析相对简单，训练方法也容易掌握，易于从训练样本中学习到内在模式，来完成对象识别、任务分类等初级智能工作。基于统计规律的浅层学习方法比起传统的基于规则的方法具备很多优越性，取得了不少成功的商业应用的同时，浅层学习的问题逐渐暴露出来，由于有限的样本和计算单元导致对数据间复杂函数的表示能力有限，学习能力不强，只能提取初级特征。

2006年，在学界及业界巨大需求刺激下，特别是计算机硬件技术的迅速发展提供了强大的计算能力。机器学习领域的泰斗Geoffrey Hinton和Ruslan-Salakhutdinov提出了深度学习模型，主要论点包括：多个隐层的人工神经网络具有良好的特征学习能力；通过逐层初始化来克服训练

的难度，实现网络整体调优。这个模型的提出，开启了深度神经网络机器学习的新时代。

2012 年，Hinton 研究团队采用深度学习模型赢得计算机视觉领域最具影响力的 Image Net 比赛冠军，从而标志着深度学习进入第二个阶段。随着 Hinton、Le Cun 和 Andrew Ng 对深度学习的研究，以及云计算、大数据、计算机硬件技术发展的支撑下，深度学习近年来在多个领域取得了令人赞叹的进展，推出一批成功的商业应用，诸如谷歌翻译、苹果语音工具 Siri、微软的 Cortana 个人语音助手、蚂蚁金服的 Smile to Pay 扫脸技术，特别是谷歌 Alpha Go 人机大战获胜的奇迹等，使机器学习成为计算机科学的一个新的领域。图 1-2 为深度学习随着数据规模的增加可提高预测准确性。

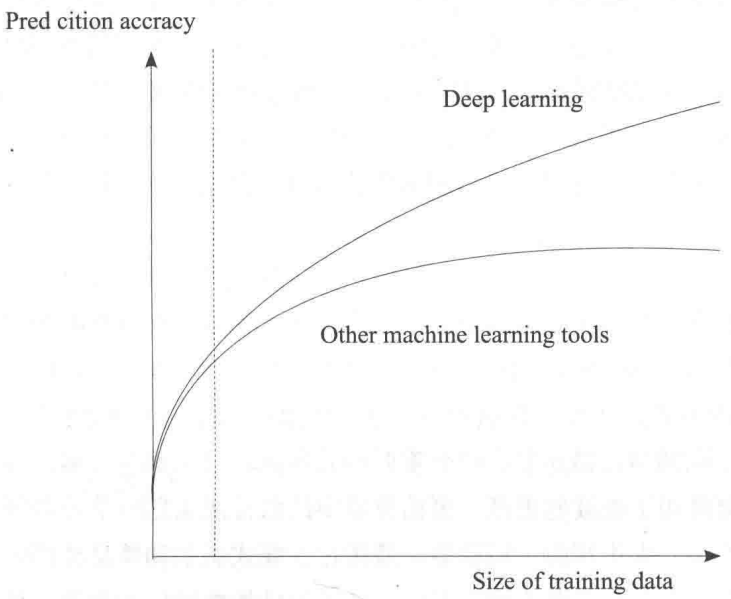


图 1-2　深度学习随着数据规模的增加可提高预测准确性

深度学习是目前最接近人类大脑的分层智能学习方法，通过建立类似于人脑的分层模型结构，突破浅层学习的限制，能够表征复杂函数关系，对输入数据逐层提取从底层到高层的特征，并且逐层抽象，从而建立从底层简单特征到高层抽象语义的非线性映射关系，实现机器学习智能化的进一步提升，成为机器学习的一个里程碑。

1.2 深度学习的定义

我们有必要先了解一些基本概念，下面是一些与深度学习密切相关的概念和描述。

定义 1：机器学习是一类利用多个非线性信息处理层来完成监督或者无监督的特征提取和转化，以及模式分析和分类等任务的技术。

定义 2：深度学习是机器学习的子领域，它是一种通过多层表示来对数据之间的复杂关系

进行建模的算法。高层的特征和概念取决于低层的特征和概念，这样的分层特征叫作深层，其中大多数模型都基于无监督的学习表示。

定义 3：深度学习是机器学习的子领域，它是基于多层表示的学习，每层对应一个特定的特征、因素或概念。高层概念取决于低层概念，而且同一低层的概念有助于确定多个高层概念。深度学习是基于表示学习的众多机器学习算法中的一员。一个观测对象（比如一张图片）可以用很多种方式表示（如像素的一个向量），但是有的表示则可以使基于训练样本的学习任务变得更容易（如判定某张图像是否为人脸图像），选一研究领域试图解决一个问题：哪些因素可以产生更好的表示，以及对于这些表示应该如何学习。

定义 4：深度学习是机器学习的一系列算法，它试图在多个层次中进行学习，每层对应于不同级别的抽象。它一般使用人工神经网络，学习到的统计模型中的不同层对应于不同级别的概念。高层概念取决于低层概念，而且同一低层的概念有助于确定多个高层概念。

定义 5：深度学习是机器学习研究的一个新领域，它的出现将机器学习向人工智能这一目标进一步拉近。深度学习是对多层表示和抽象的学习，它使一些包括如图像、声音和文本的数据变得有意义。

应该注意的是，深度学习是使用深度结构来对信号和信息进行处理，而不是对信号或信息的深度理解，尽管在有的情况下这两个方面可能会比较相似。在教育心理学中，是这样定义深度学习的："深度学习是描述学习的一种方法，其特点是：主动参与、内在激励和个人对意义的探索。"我们应该注意将深度学习与教育心理学中的这些被滥用的术语区别开来。

在上述多个不同的高层描述中有两个重要的共同点：（1）都包含多层或多阶非线性信息处理的模型；（2）都使用了连续的更高、更抽象层中的监督或无监督学习特征表示的方法。深度学习是包括神经网络、人工智能、图模型、最优化、模式识别和信息处理的交叉领域，它今天之所以如此受欢迎，有三个重要原因：其一，芯片处理性能的巨大提升（比如，通用图形处理器）；其二，用于训练的数据爆炸性增长；其三，近来，机器学习和信号 / 信息处理研究有了很大进展，这些都使深度学习方法可以有效利用复杂的非线性函数和非线性的复合函数来学习分布和分层的特征表示，并且可以充分有效地利用标注和非标注的数据。

近年来，活跃在机器学习领域的研究机构包括众多高校，比如多伦多大学、纽约大学、加拿大蒙特利尔大学、斯坦福大学、加州大学伯克利分校、加州大学、伦敦大学学院、密歇根大学、麻省理工学院、华盛顿大学；还有一些企业，如微软研究院（从 2009 年开始）、谷歌（大概从 2011 年开始）、IBM 研究院（大概从 2011 年开始）、百度（从 2012 开始）、Facebook（从 2013 年开始）、IDIAP 研究所、瑞士人工智能研究所等。

这些研究机构将深度学习方法成功地用于计算机领域的众多应用中，其中包括：计算机视觉、语音识别、语音搜索、连续语音识别、语言与图像的特征编码、语义话语理解、手写识别、音频处理、信息检索、机器人学等。

1.3 深度学习的应用领域

目前，深度学习在越来越多的领域表现出优越的性能，尤其体现在图像识别、语音识别和自然语言处理等领域。

1.3.1 图像识别领域

在物体识别问题上，深度学习的优势主要体现在 ImageNet ILSVRC 竞赛上，该竞赛是计算机视觉领域高度权威的竞赛，主要对 1 000 类的物体图像进行识别。2012 年，Geoffery Hinton 和他的学生针对分类问题将分类 Top-5 错误率从原来的 26.2% 降低至 15.3%，取得了当时领先的结果。2013 年，在 ImageNet ILSVRC2013 竞赛中，Clarifai 模型将分类 Top-5 错误率降低至 11.197%；2014 年，在 ImageNet ILSVRC2014 竞赛中，Google Net 通过使用更深的卷积神经网络将分类 Top-5 错误率降低至 6.67%；2015 年，在 ImageNet ILSVRC2015 竞赛中，微软亚洲研究院（MSRA）的深度网络 Deep Residual Network 将分类 Top-5 错误率降低至 3.567%。

在人脸识别领域，深度学习的优势主要体现在 LFW（Labeled Faces in the Wild）竞赛上的识别准确率。LFW 是目前最著名的人脸识别数据库，用来测试非可控条件下的人脸识别准确率，该数据库中的图片是从互联网中获得的，大部分图片在表情、光照、姿态等方面表现出不同的特性，香港中文大学汤晓鸥教授领导的团队设计的 DeepID 算法取得高达 99.53% 的识别准确率。

1.3.2 语音识别领域

语音识别要解决的问题首先就是将语音中的音节识别出来，其次将合适的音节组成文字。上述过程构成了语音识别的两大组成部分：声学模型、语言模型。在很长一段时间内，声学模型使用的是自动机的方法进行划分，最经典的建模方法是隐马尔可夫模型。而在语言模型方面一般分为规则模型和统计模型两种，统计语言模型是用概率统计的方法来揭示语言单位内在的统计规律，其中 N-Gram 简单有效，被广泛使用。最近，基于深度神经网络技术，百度和科大讯飞在语音识别领域都取得了重要突破，百度的 Deep Speech 采用深度学习技术对语音进行识别，它可以在饭店等嘈杂环境下实现将近 81% 的辨识准确率。而同类商业版语音识别系统如 Microsoft Bing、Google 等公司的最高识别率只有 65%。

1.3.3 自然语言理解领域

应用深度学习模型进行自然语言处理，目前主流的做法是应用 Recursive Neural Network（递归神经网络）和 Recurrent Neural Network（循环神经网络）。其中，Recurrent Neural Network 是非常有名的应用于情绪分析的树状神经网络模型，它是包含循环的网络，允许信息的持久化，更加适用于自然语言处理。

1.4 深度学习的成果

1.4.1 Google 的深度学习成果

2015 年 10 月，Google（谷歌）旗下 DeepMind 公司研发了人工智能围棋程序，该程序主要使用深度学习的技术，整体上包含离线学习和在线对弈两个过程。其中离线学习主要利用大量已有棋谱进行训练“价值网络”去计算局面优劣，训练“策略网络”去选择下子位置；在线对弈主要利用“价值网络”计算当前棋面优劣，利用“策略网络”计算当前应该选择的下子位置。2015 年，阿尔法围棋（AlphaGo）以 5 ： 0 的总比分击败欧洲围棋冠军樊麾；紧接着，2016 年 3 月，以 4 ： 1 的总比分击败世界围棋冠军、职业九段选手李世石。

而在此之前，2011 年谷歌就成立了由人工智能和机器学习顶级学者吴恩达（Andrew Ng）领衔的“Google Brain”项目，这个项目利用谷歌的分布式计算框架训练深度人工神经网络。该项目的主要成果是使用包含 16 000 个 CPU 核的并行计算平台，使用基于深度学习算法训练超过 10 亿个神经元的深度神经网络，该系统能够在没有任何先验知识的前提下，自动学习 YouTube 网站上海量的视频数据，训练深度神经网络。吴恩达目前是斯坦福大学计算机科学系和电子工程系副教授、人工智能实验室主任，并担任百度公司首席科学家，负责百度研究院的百度大脑计划。

1.4.2 Microsoft 的深度学习成果

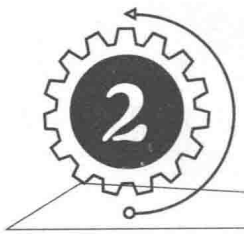
2012 年，微软首席研究官 Rick Rashid 在“21 世纪计算大会”上的英文演讲被实时翻译成与他音色很接近的中文演讲，该功能主要借助于基于深度学习技术实现的自动同声传译系统，自动同声传译过程主要是语音识别、机器翻译和语音合成。

1.4.3 国内公司的深度学习成果

2013 年，百度成立了由知名学者余凯领导的百度深度学习研究院（Instituteof Deep Learning，IDL），主要目标是将深度学习应用于语音识别和图像识别、智能检索等领域。现在，基于深度学习，百度的图像搜索更加准确，百度翻译更加专业，语音识别效果令人十分满意。目前，许多基于深度学习的产品已经面市，例如百度识别 APP，该 APP 主要功能是图像识别和智能检索，其中拍照购物和通过照片匹配度来交友都是该 APP 中比较有特色的功能。百度在“小度机器人”和无人驾驶汽车领域等都取得了重要进展。小度机器人能够通过对话等自然的交互方式，准确理解用户意图，并与用户进行信息和服务等的交流。

阿里巴巴的“拍立淘”是基于“大数据 + 深度学习 + 图像处理”的构思开发的，网购用户通过手机拍照，利用“拍立淘”就能在淘宝中找到非常类似的产品，其搜索准确度和用户满意度非常高。

LinkFace（脸云科技）在2014年开创了基于深度学习的人脸检测算法，支持人脸检测、人脸识别、人脸关键点检测等全套技术，在FDDB数据集上的人脸识别准确率高达99.5%。图森通过深度学习引擎，研发了图像识别和语义分析技术，为企业搭建了自己的图片识别服务，根据企业的实际业务设计了分类标签系统，精准描述企业图片分类需求。该公司还研发了基于摄像头的智能驾驶解决方案。



2 大数据机器学习系统

2.1 大数据机器学习系统研究背景

随着互联网技术渗透到商业和生活的各行各业，由网络所带给人们生活的便捷性大大提高。如今，在网络上浏览新闻，用软件即时聊天和视频，在电子商网站上完成商品购买与金融交易，这一切都可以在瞬间完成。然而，当信息爆炸式地增长时，人类有限的处理能力就会出现瓶颈，这时人们就迫切需要新闻网站能够自动给出感兴趣的新闻、聊天软件能实现智能交互，以及网站上能自动给出想要的商品。与此同时，云计算和物联网技术在这一年里在我国得到了前所未有的重视，如图 2-1 为我国云计算和物联网市场规模现状及预测。

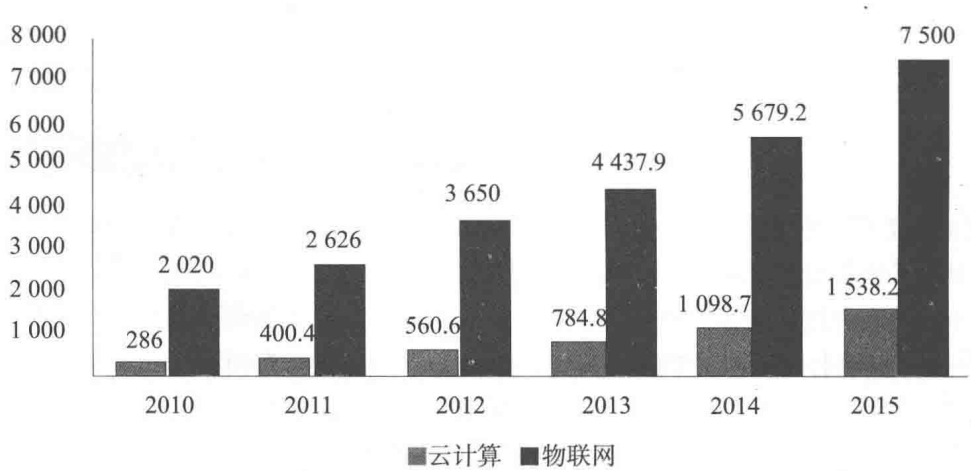


图 2-1　2010—2015 年云计算和物联网市场规模现状及预测

形形色色的智能设备已经进入了人们日常的生活，如 Google Glass、智能手表、微型机器人等，这都表示着人类对于机器智能的追求已经越来越高。更为重要的，网络的普及、硬件设

备的蓬勃发展，都将使得这样的趋势在将来还会持续增高。

当然，市场需求驱动技术提升，越来越多的开发者感受到大数据的魅力，迫切为自己的应用增添更多色彩。他们希望在保持原有产品特色的同时，能够方便地加入智能元素，促进与用户交互的趣味性。这样，种种能够应用大数据的场景均被挖掘出来：分析用户行为、网络安全、智能交互以及客户服务等，而这些模块都有可能进一步分化成为独立的产品或服务，应用公司可以单一的购买或采用。此类的早期实践者有 Google、Facebook 等，他们将自己的产品建设成为服务供外部开发者使用，因而成为服务提供商。因此，开发者也仍然可以在构建自己的核心竞争产品的同时，将购买的机器学习服务置入自己的应用中，分析用户的购买行为、猜测遗失客户的心理等，来提升自身的产品体验，优化产品的使用流程。

互联网已经成为现代生活所必不可缺少的一部分，而互联网的丰富多彩也主要是从网络化时代开始的。如表 2-1 为互联网上的一分钟发生的事情。

表 2-1 互联网一分钟发生的事情

内容	项目	数量
YouTube	用户上传视频总时长	48 小时
移动 Web 网络	新用户	217 个
WordPress 博客	发布新博文	347 篇
新网站	新增	571 个
FourSquare 网站	签到	2 083 次
Flickr 网站	用户新增照片	3 125 张
Instagram	用户分享新照片	3 600 张
Tumblr	发布新博文	27 778 篇
Facebook 网站	品牌或组织收到“喜欢”数目	34 722 个
Apple	接收 APP 下载请求	47 00 次
Twitter	用户发布 tweets	100 000 条
在线购物	顾客花费	272 070 美元
Facebook	用户分享内容	684 478 条
Google	收到查询请求	2 000 000 次
Email	发送邮件	204 166 667 封

资料来源：网络

数据量的日益暴增，如何高效存储、组织和分析这些海量数据成为当前学术界和工业界共同研究的热点问题。分布式计算，可以将计算任务分布在大量通过网络连接在一起的许多计算机上，共同协作完成计算任务，有望成为解决这一问题的有效手段。

机器学习和数据分析是将大数据转换成有用知识的关键技术，有研究表明，在很多情况下，处理的数据规模越大，机器学习模型的效果会越好。目前，国内外业界和学术界专家普遍认同的观点是，越来越多的海量数据资源加上越来越强大的计算能力，已经成为推动大数据时代人工智能技术和应用发展的动力，将基于大数据的机器学习和人工智能推上了新一轮发展浪潮，让大数据机器学习（big data machine learning）成为全球业界和学术界高度关注的热点研究领域。随着大数据时代的来临，Google、Facebook、微软、百度、腾讯等国内外著名企业均纷纷成立专门的基于大数据的机器学习与人工智能研发机构，深入系统地研究基于大数据的机器学习和智能化计算技术。

由于大数据机器学习和数据挖掘等智能计算技术在大数据智能化分析处理应用中具有极其重要的作用，在2014年12月中国计算机学会（China Computer Federation，CCF）大数据专家委员会上百位大数据相关领域学者和技术专家投票推选出的“2015年大数据十大热点技术与发展趋势”中，结合机器学习等智能计算技术的大数据分析技术被推选为大数据领域第一大研究热点和发展趋势。

大数据分析挖掘处理主要分为简单分析和智能化复杂分析两大类。简单分析主要采用类似于传统数据库OLAP的处理技术和方法，用SQL完成各种常规的查询统计分析；而大数据的深度价值仅通过简单分析是难以发现的，通常需要使用基于机器学习和数据挖掘的智能化复杂分析才能实现。由于大数据机器学习在具体实现时通常需要使用分布式和并行化大数据处理技术方法，也有人将大数据机器学习称为“分布式机器学习”（distributed machine learning）或“大规模机器学习”（large-scale machine learning）。

大数据机器学习，不仅是机器学习和算法设计问题，还是一个大规模系统问题。它既不是单纯的机器学习，也不是单纯的大数据处理技术所能解决的问题，而是一个同时涉及机器学习和大数据处理两个主要方面的交叉性研究课题。一方面，它仍然需要继续关注机器学习的方法和算法本身，即需要继续研究新的或改进的学习模型和学习方法，以不断提升分析预测结果的准确性；与此同时，由于数据规模巨大，大数据机器学习会使几乎所有的传统串行化机器学习算法难以在可接受的时间内完成计算，从而使得算法在实际应用场景中失效。因此，大数据机器学习在关注机器学习方法和算法研究的同时，还要关注如何结合分布式和并行化的大数据处理技术，以便在可接受的时间内完成计算。为了能有效完成大数据机器学习过程，需要研究并构建兼具机器学习和大规模分布并行计算处理能力的一体化系统。因此，领域内出现了“大数据机器学习系统”或者“分布式学习系统”的概念，并进行了诸多大数据机器学习系统的研究与开发工作。