TURING 图灵程序设计丛书

Packt

Deep Learning with Hadoop

Hadoop 深度学习

[印] 迪帕延·德夫 著　范东来 赵运枫 封强 译

学习用Hadoop在深度神经网络中部署大数据

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

Deep Learning with Hadoop

Hadoop 深度学习

[印] 迪帕延·德夫 著
范东来 赵运枫 封强 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Hadoop深度学习 / （印）迪帕延·德夫
(Dipayan Dev) 著 ; 范东来, 赵运枫, 封强译. -- 北京:
人民邮电出版社, 2018.5
（图灵程序设计丛书）
ISBN 978-7-115-48218-1

Ⅰ. ①H… Ⅱ. ①迪… ②范… ③赵… ④封… Ⅲ. ①
数据处理软件 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第064440号

内 容 提 要

本书主要目标是处理很多深度学习应用的热点问题并向读者披露解决方案的细节。主要内容分为7章：第1章介绍深度学习基础知识，第2章介绍大规模数据的分布式深度学习，第3章介绍卷积神经网络，第4章介绍循环神经网络，第5章介绍受限玻尔兹曼机，第6章介绍自动编码器，第7章介绍如何用Hadoop玩转深度学习。

本书适合人工智能相关专业师生，以及对深度学习在大数据领域的应用感兴趣的软件工程师。

◆ 著　[印] 迪帕延·德夫
译　范东来　赵运枫　封　强
责任编辑　岳新欣
执行编辑　杨　婷
责任印制　周昇亮

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路11号
邮编　100164　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
大厂聚鑫印刷有限责任公司印刷

◆ 开本：800×1000　1/16
印张：8.5
字数：201千字　2018年5月第1版
印数：1-3 500册　2018年5月河北第1次印刷
著作权合同登记号　图字：01-2017-6477号

定价：39.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字20170147号

谨以此书献给我的父亲 Tarun Kumar Deb 和母亲 Dipti Deb。
也献给我的兄长 Tapojit Deb。

前　言

本书将教你如何使用 Hadoop 在深度神经网络中部署大型数据集，以实现最佳性能。

从了解什么是深度学习以及与深度神经网络相关的各种模型开始，本书将向你展示如何配置用于深度学习的 Hadoop 环境。

本书内容

第 1 章，深度学习介绍。深度学习在过去十年间已深入人心，由于功能增强了，其发展速度甚至超过了机器学习。这一章首先介绍了人工智能的现实应用、相关的挑战，以及深度学习为何能够有效地解决这些问题。通过解决一些主要的机器学习问题（如维度诅咒、梯度消失等），深入阐释了深度学习。为了后续各章内容的学习，后半部分讨论了各种类型的深度学习网络。该章主要适用于想了解深度学习的基础知识，但不需要深入了解各个深度神经网络细节的读者。

第 2 章，大规模数据的分布式深度学习。大数据和深度学习无疑是近段时间最热门的两大技术趋势。两者关系密切，过去几年中都呈现出了巨大的发展。这一章首先介绍了如何将深度学习技术用于大量非结构化的数据，并从中提取宝贵的隐藏信息。Google、Facebook、苹果等知名技术公司正在深度学习项目中使用这种大规模数据，以更智能的方式训练一些优秀的深度神经网络。然而，深度神经网络在处理大数据时遇到了一些挑战。这一章将详细说明这些挑战。后半部分介绍了 Hadoop，并探讨了如何使用 Hadoop 的 YARN 及其迭代 Map-Reduce 来实现深度学习模型。接着介绍了深度学习中一个流行的开源分布式框架：Deeplearning4j，并解释了其各种组件。

第 3 章，卷积神经网络。卷积神经网络是一种深度神经网络，广泛应用于顶尖技术产业的各种深度学习项目中。卷积神经网络在图像识别、视频识别、自然语言处理等各个领域都有广泛的应用。卷积是一种特殊的数学运算，是卷积神经网络的重要组成部分。为了学习卷积神经网络，这一章首先用现实生活中的一个示例说明了卷积的概念。接下来，通过对网络的每个组成部分进行说明，深入阐释了卷积神经网络。为了提高网络性能，卷积神经网络具有三个最重要的参数：稀疏连接、参数共享和平移不变性。这一章对这些概念进行了解释，以便更好地理解卷积神经网络。卷积神经网络还有一些关键的超参数，这些超参数有助于确定网络输出图像的维度。这一章还详细讨论了这些超参数之间的数学关系。后半部分重点介绍分布式卷积神经网络，并展示了如

何使用 Hadoop 和 Deeplearning4j 来实现分布式卷积神经网络。

第 4 章，循环神经网络。循环神经网络是一种特殊的神经网络，可作用于长向量序列，以产生不同的向量序列。近年来，它们已成为可变长序列建模中极受欢迎的选择。循环神经网络已经成功应用于语音识别、在线手写识别、语言建模等领域。通过提供一些必要的数学关系和可视化表征，这一章详细阐释了循环神经网络的各种概念。循环神经网络拥有自己的“内存”来存储中间隐藏层的输出。“记忆”是循环神经网络的核心部分，这一章用合适的框图对其进行了讨论。此外，为了克服单向循环神经网络的局限性，这一章引入了双向循环神经网络的概念。随后，为了解决第 1 章中提到的梯度消失问题，讨论了循环神经网络中被称为“长短期记忆”的一个特殊单元。最后，使用 Deeplearning4j 在 Hadoop 中实现分布式深度循环神经网络。

第 5 章，受限玻尔兹曼机。这一章涵盖了第 3 章和第 4 章中讨论过的两种模型，并说明了它们是判别模型，还探讨了名为“受限玻尔兹曼机”的生成模型。在给定隐藏参数时，受限玻尔兹曼机能够随机生成可见的数值。该章首先介绍了“基于能量的模型”这一概念，并阐释了受限玻尔兹曼机和它的关系。此外，还讨论了一种被称为“卷积受限玻尔兹曼机”的特殊受限玻尔兹曼机，它是卷积和受限玻尔兹曼机的组合，有助于提取高维图像的特征。

这一章的后半部分介绍了深度信念网络，这是一种被广泛使用的、由几个受限玻尔兹曼机组成的多层网络。此外，还讨论了如何使用 Hadoop 在分布式环境中实现深度信念网络。最后讨论了如何使用 Deeplearning4j 实现受限玻尔兹曼机以及分布式深度信念网络。

第 6 章，自动编码器。这一章引入了一种称为“自动编码器”的生成模型，这种模型通常用于降维、特征学习或提取。该章首先解释了自动编码器的基本概念及其通用框图。自动编码器的核心结构基本上可分为编码器和解码器两部分。编码器将输入映射到隐藏层，而解码器将隐藏层映射到输出层。基础自动编码器的主要作用是将输入层的某些方面复制到输出层。这一章接着讨论了稀疏自动编码器，它基于隐藏层的分布式稀疏表征。随后深入介绍了包含多个编码器和解码器的深度自动编码器的概念，并提供了适当的示例和框图。该章后半部分对降噪自动编码器和堆叠式降噪自动编码器进行了说明。最后展示了如何使用 Deeplearning4j 在 Hadoop 中实现堆叠式降噪自动编码器和深度自动编码器。

第 7 章，用 Hadoop 玩转深度学习。这一章主要介绍分布式环境中三种最常用的机器学习应用的设计。该章讨论了如何使用 Hadoop 进行大规模的视频处理、图像处理和自然语言处理，阐释了如何在 Hadoop 分布式文件系统中部署大型视频和图像数据集，并使用 Map-Reduce 算法进行处理。对于自然语言处理，该章最后对其设计和实现进行了深入的说明。

阅读背景

我们希望本书的所有读者都具有一定的计算机科学背景。本书主要讨论不同的深度神经网

络，以及其基于 Deeplearning4j 的设计和应用。为了更好地学习本书中的内容，你最好已掌握机器学习、线性代数、概率论、分布式系统和 Hadoop 的基础知识。为了使用 Hadoop 实现深度神经网络，本书广泛应用了 Deeplearning4j。运行 Deeplearning4j 所需的知识可以参考以下链接：https://deeplearning4j.org/quickstart。

读者对象

如果你是想学习如何在 Hadoop 上进行深度学习的数据科学家，那么本书很适合你。对机器学习的基本概念与 Hadoop 有一定的了解，将有助于你充分利用本书。

排版约定

在本书中，你会发现一些不同的文本样式。以下举例说明它们的含义。

嵌入代码、数据库表名、用户输入等用等宽字体表示，例如：“`.build()`函数用于构建层。”

代码块的样式如下所示：

```
public static final String DATA_URL =
  "http://ai.stanford.edu/~amaas/data/sentiment/*";
```

当我们希望你注意代码块中的特定部分时，相关行或项目将以粗体显示：

```
MultiLayerNetwork model = new MultiLayerNetwork(getConfiguration());
Model.init();
```

新术语和重要内容会以黑体字显示。

此图标表示警告或重要事项。

此图标表示提示和技巧。

读者反馈

我们非常欢迎读者的积极反馈。如果你对本书有任何想法或看法，请及时反馈给我们，这将有助于我们出版充分满足读者需求的图书。一般性反馈请发送至电子邮箱 feedback@packtpub.com，并在邮件主题中注明书名。如果你擅长某个领域，并有意编写图书或是贡献一份力量，可以参考我们的作者指南：http://www.packtpub.com/authors。

客户支持

你现在已经是 Packt 的尊贵读者了。为了让你的购买物超所值，我们还为你准备了以下内容。

下载示例代码

你可以使用自己的账户从 http://www.packtpub.com 下载所有已购 Packt 图书的示例代码文件。如果你是从其他途径购买的本书，那么可以访问 http://www.packtpub.com/support 并注册，我们将通过电子邮件向你发送文件。

可以通过以下步骤下载示例代码文件。

(1) 使用电子邮件和密码登录或注册我们的网站。
(2) 将鼠标光标移到网站上方的 SUPPORT 标签。
(3) 单击 Code Downloads & Errata 按钮。
(4) 在搜索框中输入书名。
(5) 选择想要下载代码文件的图书。
(6) 从下拉菜单中选择购书途径。
(7) 单击 Code Download 按钮。

下载文件后，确保使用以下软件的最新版来解压文件：

- WinRAR / 7-Zip（Windows）
- Zipeg / iZip / UnRarX（Mac）
- 7-Zip / PeaZip（Linux)

也可以在 GitHub 上获取本书的代码包，具体网址为 https://github.com/PacktPublishing/Deep-Learning-with-Hadoop。另外，https://github.com/PacktPublishing 上还有其他图书的代码包和视频，你可以自行下载。

下载本书的彩色图片

我们还提供了一份 PDF 文件，其中包含了书中的截屏和图表等彩色图片，以帮助你更好地理解输出的变化。下载网址为 https://www.packtpub.com/sites/default/files/downloads/DeepLearningwithHadoop_ColorImages.pdf。

勘误

虽然我们已尽力确保书中的内容正确无误，但出错仍旧在所难免。如果在书中发现错误，不

管是文本还是代码，希望你能够告知我们，我们将不胜感激。这样一来，你可以让其他读者免受挫败，也可以帮助我们改进本书的后续版本。如果发现任何错误，请访问 http://www.packtpub.com/submit-errata，选择本书，单击 Errata Submission Form 链接，并输入详细说明。[①]经过核实后，你提交的勘误内容将上传到官方网站或添加到现有的勘误表中。

访问 https://www.packtpub.com/books/content/support，在搜索框中输入书名后就可以在**勘误**（Errata）部分查看已经提交的勘误信息。

盗版

任何媒体都会面临版权内容在互联网上的盗版问题，Packt 也不例外。Packt 非常重视版权保护。如果你在互联网上发现了我们作品的非法复制版，不管该复制版以何种形式存在，请立即提供相关网址或网站名称，以便我们寻求补救。

请将可疑盗版材料的链接发至 copyright@packtpub.com。

维护作者的权益就是在维护我们继续为你带来价值的能力，感谢你对此作出的努力。

问题

如果你对本书内容存有任何疑虑，可以通过 questions@packtpub.com 联系我们，我们将尽最大努力解决问题。

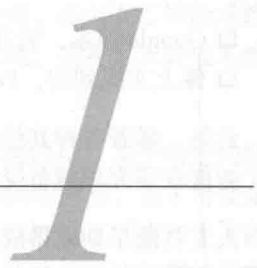
电子书

扫描如下二维码，即可购买本书电子版。

① 中文版勘误可以在 http://www.ituring.com.cn/book/1940 中查看和提交。——编者注

目　录

第 1 章 深度学习介绍



“到目前为止，人工智能的最大危险是人们过早地断定了他们了解人工智能。”

——Eliezer Yudkowsky

你是否想过，为什么即使是最顶尖的选手也很难在国际象棋比赛中战胜计算机？Facebook 如何在数亿张照片中识别出你的脸？你的手机如何识别你的声音并从数百个联系人中呼叫正确的人？

本书的主要目标就是处理这些查询，并提供详细的解决方案。本书可供多种读者阅读使用，但我们的目标读者主要有两类。第一类是学习深度学习和人工智能的本科生或研究生，第二类是对大数据、深度学习和统计建模有一定了解，但想快速了解如何将深度学习应用于大数据领域以及如何将大数据技术应用于深度学习领域的软件工程师。

本章将介绍深度学习的基本概念、术语、特性和主要挑战，主要是为读者打好基础。本章还提出了不同深度网络算法的分类，这些算法在最近十年被研究者广泛采用。本章涉及的主题如下：

- 开始深度学习之旅
- 深度学习的术语
- 深度学习——一场人工智能的革命
- 深度学习网络的分类

自人类文明诞生以来，人们总是梦想着建造人工机器或机器人，它们可以像人类一样运作和工作。从希腊神话人物到古印度史诗，历史中有许多这样的示例，它们清楚地表明了人们的兴趣和意向：创造并拥有人工生命。

在计算机时代的初期，人们就想知道计算机将来能否变得像人类一样聪明。后来，自动化机器开始变得不可或缺，即便在医学领域也是如此。随着这种需求和人们对这一领域的不断研究，**人工智能**（Artificial Intelligence，AI）已经是一种呈蓬勃发展之势的技术，在多个领域，如图像处理、视频处理以及许多其他医学工具，都得到了应用。

虽然人工智能解决了很多日常问题，但没有人知道编码人工智能系统的具体规则。比较直观

的几个问题如下所示。

- Google 搜索，它在理解你输入或说出的内容方面做得很好。
- 像上文提到的，Facebook 在识别人脸方面也做得不错，从而可以理解用户的兴趣所在。

此外，随着各种其他领域的整合，如概率论、线性代数、统计学、机器学习和深度学习等，人工智能在研究领域备受青睐。

人工智能早期取得成功的一个关键原因是，它主要解决的是基本问题，而计算机在解决这些问题时不需要使用大量的知识。例如，1997 年，IBM 的深蓝计算机击败了国际象棋世界冠军 Garry Kasparov[1]。虽然这一成就在当时是很了不起的，但用国际象棋的有限规则来训练计算机绝对不算是一个繁重的任务。用固定且数量有限的规则来训练系统称为计算机的**硬编码知识**。许多人工智能项目都经历了用传统语言来描述世界不同领域的硬编码知识的阶段。随着时间的推移，这种硬编码知识似乎不再适用于处理拥有大量数据的系统。而且，数据遵循的规则数量也在不断地改变。因此，遵循这一系统的大多数项目并没有达到预期的高度。

要想克服这种硬编码知识所面临的困难，这些人工智能系统需要以某种方式从提供的原始数据中概括出模型和规则，而无需外部灌输。系统做这件事的熟练程度被称为**机器学习**。日常生活中有各种各样成功的机器学习实现，以下是几个最常见且最重要的实现。

- **垃圾邮件检测**：对于收件箱中的电子邮件，模型可以检测是将该电子邮件放在垃圾箱还是收件箱中。普通的朴素贝叶斯模型就可以区分这样的电子邮件。
- **信用卡欺诈检测**：模型可以检测在特定时间间隔内执行的多个交易是否是由最初的客户执行的。
- 最受欢迎的机器学习模型之一是 Mor-Yosef 等人在 1990 年提出的，该模型使用了逻辑回归，可以对患者是否需要剖腹产给出建议。

这样的模型有很多，它们是借助机器学习技术实现的，如图 1-1 所示。

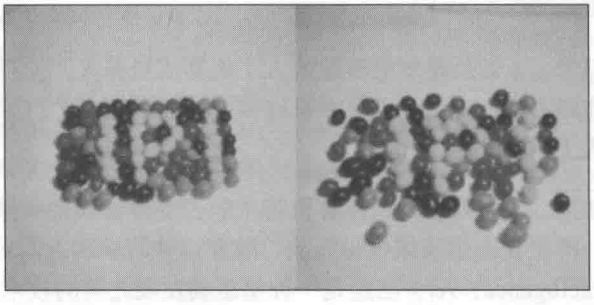


图 1-1　不同类型的表征示例。假设需要训练机器来检测糖豆之间的空隙。图中右侧的糖豆稀疏，人工智能系统可以很容易就确定空隙。图中左侧的糖豆非常紧凑，找到空隙对机器来说是一个非常困难的任务（图像来自 USC-SIPI 图像数据库）

机器学习系统的性能主要取决于向系统提供的数据，这称为数据**表征**。与表征相关的所有信息称为数据的**特征**。例如，如果使用逻辑回归来检测患者是否患有脑肿瘤，那么人工智能系统不会尝试直接诊断患者。相反，相关医生需要将该患者的常见症状输入系统，接着人工智能系统会将输入的信息与用于训练系统的以往信息相匹配。

基于系统的预测分析，人工智能系统会对疾病做出诊断。虽然逻辑回归可以基于给定的特征来学习和决策，但它不能影响或修改特征的定义方式。逻辑回归是一种回归模型，基于自变量的因变量的可能值是有限的，这一点与线性回归不同。因此，如果提供给模型的是剖腹产患者的报告而不是脑肿瘤患者的报告，那么它肯定不能预测出正确的结果，因为给定的特征与训练的数据不匹配。

机器学习系统对数据表征的依赖对我们来说并非是未知的。事实上，大多数计算机理论都是基于数据表征而表现得更好。例如，模式设计会对数据库的质量产生影响。任何数据库查询的执行，哪怕是在百万甚至千万行的数据上查询，只要表格被正确索引，那么速度就会变得极快。因此，人工智能系统对数据表征的依赖也没什么可让人大惊小怪的。

日常生活中也有很多这样的示例，其中数据表征决定了我们的效率。从 20 人中找到一个人明显比从 500 人中找到一个人容易。图 1-1 就是两种类型的数据表征的可视化表示。

因此，如果向人工智能系统传递适当的特征数据，那么即便是最困难的问题也可以得到解决。然而，以正确的方式收集并传递系统期望的数据对于程序员来说是一大难题。

在许多实际场景中，提取特征可能是一件繁琐的事情。因此，数据的表征方式决定了系统智能的主要因素。

如果特征不恰当，那么从人和猫的组合中找到猫将会是非常复杂的任务。我们都知道猫有尾巴，因此可能会想要将有尾部作为突出特征。然而，鉴于尾巴有不同的形状和大小，通常很难用像素值来准确地描述尾部。此外，尾巴有时可能会与人的手部混淆。另外，一些物体的重叠可能会导致看不到猫尾巴的存在，使得图像更加复杂。

从上述讨论可以得出一个确切的结论：人工智能系统的成功主要取决于输入数据的表征方式。除此之外，不同的表征还能捕获与保留能够解释数据差异性的那些隐藏因素（解释性因素）。

表征学习是解决这些特定问题时广泛使用的一种流行学习方法。表征学习可以定义为根据数据当前层的表征来推断下一层数据的表征。在理想情况下，所有的表征学习算法都有一个优势：它捕获隐藏的因素，而这一子集可能适用于每个特定的子任务。图 1-2 中给出了简单的说明，如下所示。

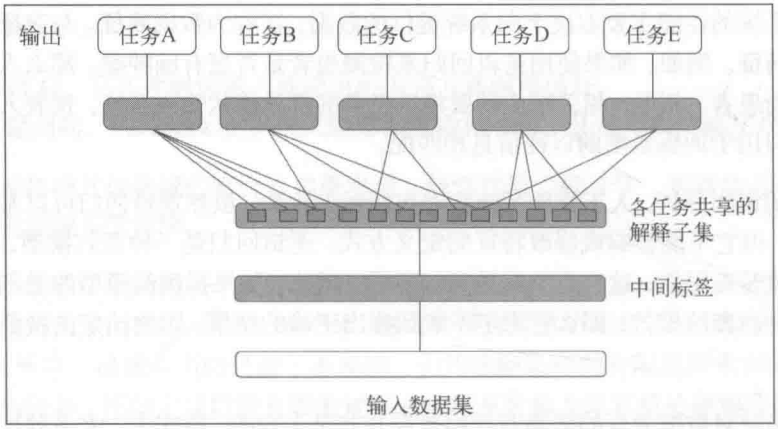


图 1-2　表征学习的简单说明。中间层能够发现解释性因素（蓝色方框中的隐藏层）。有些因素解释了每个任务的目标，有些因素则解释了任务的输入

然而，从大量原始数据中提取高级数据和特征需要人类级别的理解力，这就有了局限。以下就是这样的示例。

- 区分年龄相仿的两个婴儿的哭声。
- 识别猫眼在白天与黑夜的图像。因为猫眼在夜间会发光，所以完成这项工作并不轻松。

在所有这些极端情况中，表征学习并没有异常的表现，而是展现出了威慑行为。

深度学习是机器学习的一个子领域。通过构建多层次的表征或从一系列简单的表征和特征中学习一个具有层次结构的特征集，深度学习能够解决表征学习的主要问题[2,8]。

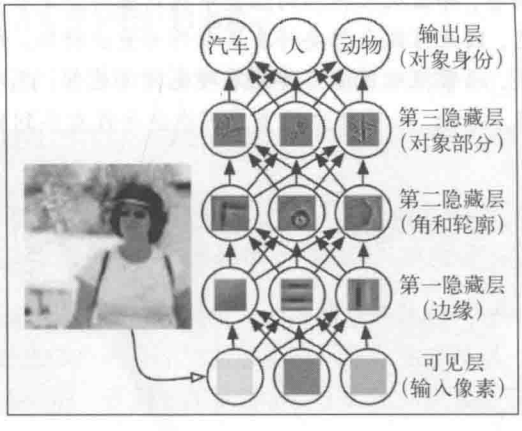


图 1-3　通过识别各种组合（如轮廓、角，可以用边缘来定义），此图展示了深度学习系统如何显示人的图像。图片的转载获得了 Ian Goodfellow、Yoshua Bengio 和 Aaron Courville 的许可，来源于 The MIT Press 出版的 *Deep Learning* 一书

图 1-3 对深度学习模型进行了说明。就像将一系列不同的像素值构成一幅图像一样，用计算机解码这些原始的非结构化的输入数据通常是一件很繁琐的事。在理想情况下，转换像素值来识别图像的映射函数是很难实现的。此外，为这类映射直接训练计算机几乎是不可能的。为了应对这些类型的任务，深度学习会创建连接至期望输出的一系列映射子集，以解决这类难题。映射的每个子集对应于模型的一个层。输入层包含了可以观察的变量，因此处于可见层。我们可以从给定的输入中逐层提取数据的抽象特征。因为这些抽象的值在给定数据中是不可用或不可见的，所以这些层称为隐藏层。

在图 1-3 的第一隐藏层中，通过对相邻像素进行比较学习，可以轻易地识别出边缘。第二隐藏层可以将角和轮廓从第一隐藏层的边缘中区分开来。基于描述角和轮廓的第二隐藏层，第三隐藏层可以识别特定对象的不同部分，最终可以从第三隐藏层中检测出图中存在的不同对象。

深度学习由 Hinton 等人于 2006 年提出[2]；Bengio 等人在 2007 年将其用于解决 MNIST 数字分类问题。最近几年，深度学习经历了从数字识别向自然图像的物体识别的重大转变。除此之外，Krizhevsky 等人在 2012 年使用了 ImageNet 数据集，这实现了深度学习领域的一个重大突破。

本书的范围仅限于深度学习。因此，在进入主题之前，需要先明确与深度学习有关的概念。

在过去的 10 年间，许多学者从不同角度对深度学习进行了定义，但目前还没有一个统一的定论。以下是被广泛接受的两种定义。

- GitHub：深度学习是机器学习研究的一个新领域，其目的是让机器学习更接近其原始目标之一——人工智能。深度学习是学习多层次的表征和抽象的一种方法，有助于理解图像、声音和文本等数据。
- 维基百科：深度学习是机器学习的一个分支，它基于一组算法，通过线性或非线性的转换，尝试使用具有多个处理层的深度图来对数据中的高级抽象进行建模。

上述定义表明，也可以认为深度学习是一类特殊的机器学习。深度学习具有从各种简单特征中学习复杂表征的能力，这使其在数据科学领域获得了广泛的应用。为了进一步了解深度学习，我们列出了本书后面将经常使用的一些术语。接下来将给出各种术语和深度学习使用的重要网络，以帮助你奠定深度学习的基础。

1.1 开始深度学习之旅

开始本书的深度学习之旅前，你应当了解机器学习的所有术语和基本概念。如果已经充分了解了机器学习及其相关术语，那么你可以忽略本节，直接跳转到 1.2 节开始阅读。热衷于数据科学，同时想要彻底学习机器学习的读者，可以阅读 *Machine Learning*（Tom M. Mitchell，1997）[5] 和 *Machine Learning: A Probabilistic Perspective*（2012）[6]这两本书。

注意，神经网络并不会产生奇迹，但合理地使用它们可以产生一些惊人的效果。

1.1.1　深度前馈网络

神经网络可以是循环的或前馈的。前馈网络在其网络图中并没有任何循环结构。具有多层次的网络称为深度网络。简单来说，任何具有两层或更多（隐藏）层的神经网络被定义为**深度前馈网络**或**前馈神经网络**。图 1-4 显示了深度前馈神经网络的一般形式。

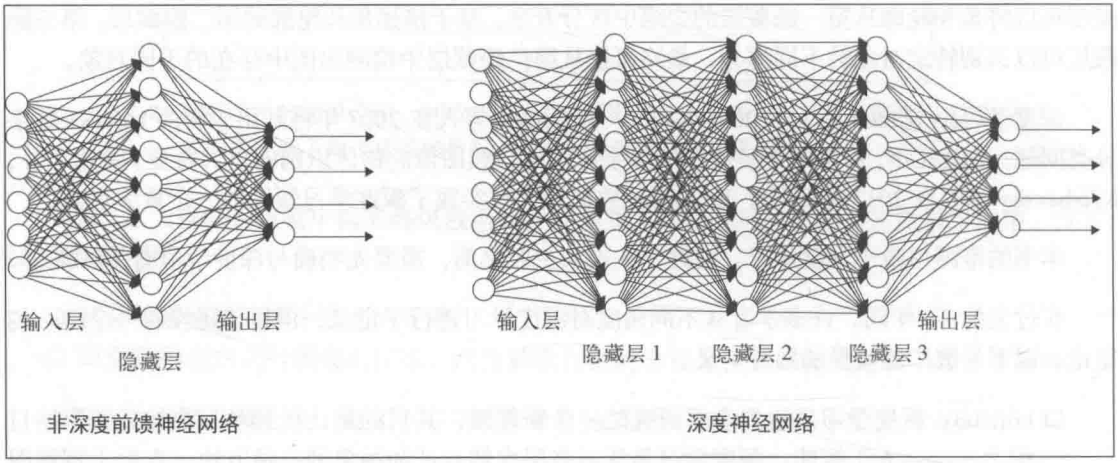


图 1-4　浅层前馈网络和深度前馈网络

深度前馈网络的工作原理是，随着深度的增加，网络可以执行更多的顺序指令。顺序指令可以提供很大的威力，因为它们可以指向较早的指令。

前馈网络的目的是对某个函数 f 进行通用化。例如，分类器 $y=f(x)$ 将输入值 x 映射到类别 y。深度前馈网络将该映射关系修改为 $y = f(x; \alpha)$，并学习参数 α 的值，从而得到最适合的函数值。图 1-4 是深度前馈网络的简单表示，它展示了深度前馈网络与传统神经网络的架构差异。

注意，深度神经网络是具有多个隐藏层的前馈网络。

1.1.2　各种学习算法

数据集被认为是学习过程的基石。数据集可以定义为相互关联的数据的集合，该集合由多个独立的实体组成，但也可以根据使用场景将该集合当作单个实体。数据集的各个数据元素称为**数据点**。

图 1-5 是从社交网络分析中收集到的各种数据点的可视化表示。

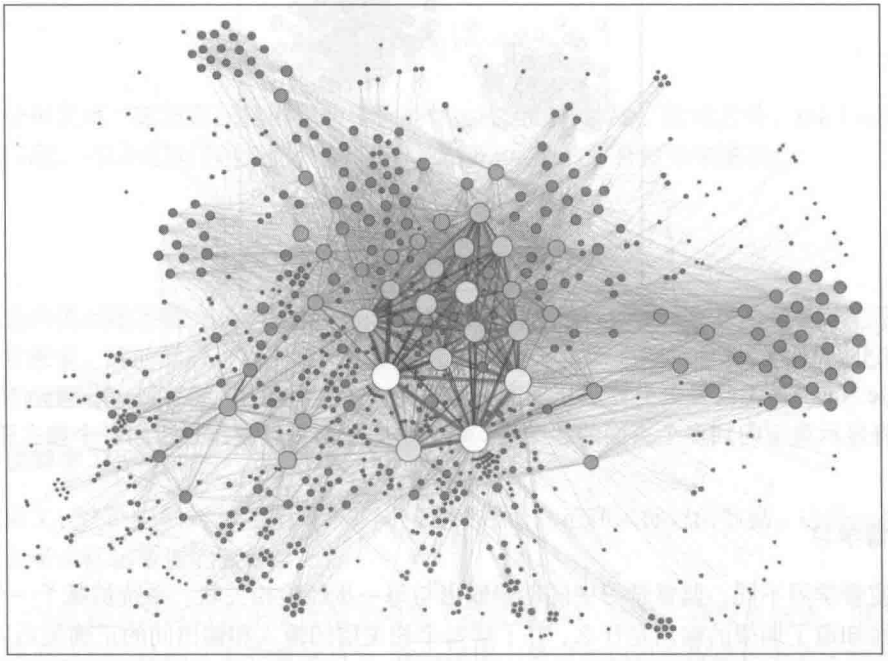


图 1-5 社会网络分析的分散数据点（图片来源于维基百科）

- **未标记的数据**：这部分数据由人为生成的对象组成，这些对象可以从周围环境中轻松获取，如 X 射线、日志文件数据、新闻文章、演讲、视频、推文等。
- **已标记的数据**：这部分数据是通过将一组未标记的数据进行标准化而得到的。这类数据通常是格式化的、已分类的、已标签化的，并且易于人类理解，以便进行进一步的处理。

从最高层面来看，机器学习技术可以根据不同的学习过程分为监督学习和无监督学习。

1. 无监督学习

在无监督学习算法中，给定的输入数据集没有期望的输出。在分析数据集时，系统从其经验中学习有意义的属性和特征。在深度学习中，系统通常会尝试从数据点的整体概率分布中学习。到目前为止，执行聚类的无监督学习算法有多种类型。简单地说，聚类将具有相似数据类型的多个数据点放入同一个簇中。但是，通过这种学习，其最终输出不会提供任何反馈，也就是说，不会有任何老师来纠正你的错误。图 1-6 展示了无监督聚类。