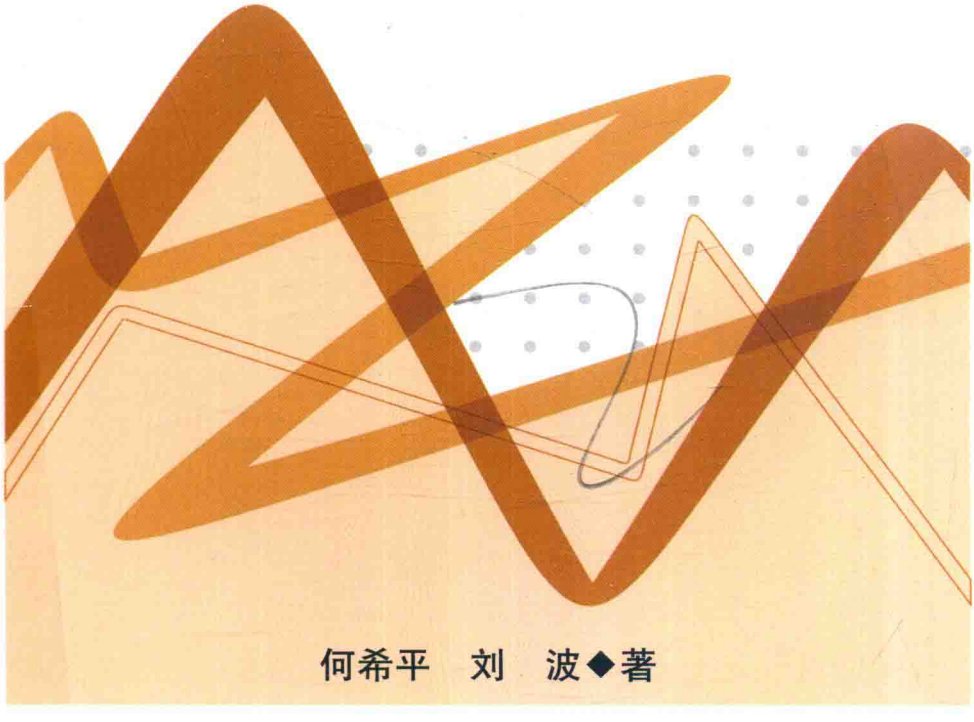
重庆市检测控制集成系统工程实验室资助
电子商务及供应链系统重庆市重点实验室（重庆工商大学）资助

深度学习

理论与实践

SHENDU XUEXI
LILUN YU SHIJIAN



何希平　刘　波◆著

科学出版社

重庆市检测控制集成系
电子商务及供应链系统重庆市重点实验室（重庆工商大学）资助

深度学习理论与实践

何希平　刘　波　著

科学出版社
北京

内 容 简 介

深度学习作为表示学习的重要分支，有着广泛的应用价值。深度学习通常会基于多层的神经网络，它能从大规模数据中提取有效特征来表示数据，从而提高机器学习算法的性能。本书以重庆工商大学等单位的机器学习、图像处理课题为基础，系统地介绍特征选择的基本概念，以及相关的理论和算法，也对深度学习的前沿研究（如区域-卷积神经网络等）和其在计算机视觉中的应用（如目标检测）进行详细介绍，最后对深度学习的发展方向进行展望。

本书理论联系实际，对教学、科研具有重要指导意义，可作为高等院校和科研机构从事机器学习的学者的参考书，也可供从事计算机视觉、语言识别的专业技术人员参考。

图书在版编目（CIP）数据

深度学习理论与实践/何希平，刘波著. —北京：科学出版社，2017.3

ISBN 978-7-03-052104-0

Ⅰ. ①深… Ⅱ. ①何… ②刘… Ⅲ. ①学习系统–研究 Ⅳ. ①TP273

中国版本图书馆 CIP 数据核字（2017）第 050454 号

责任编辑：张 展 孟 锐 / 责任校对：王 翔

责任印制：余少力 / 封面设计：墨创文化

科 学 出 版 社 出版

北京东黄城根北街 16 号

邮政编码：100717

http：//www.sciencep.com

成都锦瑞印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 3 月第 一 版 开本：720×1000 B5

2017 年 3 月第一次印刷 印张：10

字数：205 000

定价：59.00 元

（如有印装质量问题，我社负责调换）

前　言

深度学习作为机器学习的重要分支，近年来得到了非常迅速的发展，已经成为目前研究最热门的技术之一。深度学习的模型是一种基于多层神经网络的层次结构模型，它不仅可从大数据中自动学习表达数据本质与隐含规律的数据特征，而且这类模型还克服了传统学习方法通过手工设计特征算子的局限性。自从Hinton在2006年首次提出深度学习的概念以来，在短短的10年时间里，深度学习已经被用于众多应用领域，并取得了重大突破。例如，2009年微软将深度学习模型引入语音识别中，使得语音识别的准确率提升了25%，远超业界所预测的5%；2011年谷歌在“谷歌大脑”的研究项目中利用自身的大数据和云平台优势建立了一个计算平台，该平台配备1.6万个能并行计算的CPU和10亿个突触神经元，研究人员随机选取1000万个视频作为深度学习模型的训练数据，最终谷歌大脑学习到了一张猫面部图像；深度学习在计算机视觉领域也取得了重大突破。2012年，Hinton的研究小组采用深度学习赢得ImageNet图像分类比赛的冠军。他们的分类精度比排名第2的小组要高10%左右，该小组所使用的特征是用手工方法设计的；而在人脸识别方法中，深度学习也有非凡的表现，人眼在LFW数据集上的识别率是97.53%，而采用深度学习可达到99.47%。深度学习的所有这些突破性进展为人工智能在大数据应用领域的研究开辟了一条坚实的新通道。

为方便理解与推广应用，本书采用通俗易懂的语言和深入浅出的表达方式，对深度学习的基本概念、基本模型结构和相关理论进行介绍。同时也对深度学习在计算机视觉领域的最新研究进展进行一些讨论。对于一些抽象概念，尽量辅以直观图形来进行辅助解释。

本书主要分为4部分。第一部分由第1章和第2章构成，这部分主要介绍深度学习的发展以及在各个领域的应用，并简要介绍深度学习与机器学习、表示学习、神经网络的关系，以及促使深度学习发展的因素。由于神经网络是深度学习的基础，所以在这一部分还介绍神经网络的基本原理和相应的网络结构。神经网络是一种层次结构的机器学习模型，传统的神经网络包含输入层、隐藏层和输出层，其中隐藏层由多个神经元组成，每个神经元都有激活函数。通过这部分的内容不难发现深度学习的本质，是其模型由众多逐层贪婪学习的神经网络所构成；逐层无监督特征学习是深度学习算法的根本特征。第二部分即第3章，这部分介绍实现深度学习的随机梯度下降算法及其改进算法，这类算法是梯度下降算法（如最陡梯度下降法、Newton法、拟Newton法等）的简化版本。这部分也介绍一些

无约束优化和约束优化的基本理论，旨在帮助读者理解与深度学习相关的优化算法。第三部分由第 4～6 章构成，这部分介绍广泛使用的几个深度学习模型：深度信念网和卷积神经网络等。深度信念网由多层受限 Boltzmann 机构成，并采用逐层训练方式进行预训练，即对每个隐藏层网络单独训练，所得到的权重参数作为该层的初始化参数，然后再训练整个网络。这部分还介绍卷积神经网络的基本结构和相关的操作，以及它在图像分类和目标检测上的最新进展。除此以外，这部分还会介绍自编码器，因为深度信念网在进行逐层训练时会采用自编码器的思想。第四部分即第 7 章，这部分主要面向深度学习研究实践，介绍通用的深度学习模型描述、训练与测试的框架平台 Caffe，主要介绍利用 Caffe 如何进行一些基本操作，例如，各种网络参数的意义和读取各个网络层权重的方法，以便研究人员把主要的工作精力放到深度学习的模型设计上，而不是经典算法代码的编写上。

本书第 1～4 章由重庆工商大学计算机科学与信息工程学院的何希平教授编写；第 5～7 章及本书附录，由重庆工商大学计算机科学与信息工程学院的刘波博士编写，刘波博士撰写本书一半的内容，总计 10 万字。特别感谢重庆工商大学电子商务及供应链系统重庆市重点实验室的管理科学与工程专业的两位研究生寇茜茜和丁一楠，她们为本书绘制了图形。

本书编写过程中得到挂靠在重庆工商大学的电子商务及供应链系统重庆市重点实验室、重庆市检测控制集成系统工程实验室的专项基金的支持。此外，本书研究成果还得到如下项目资助：①重庆市教委研究项目“多核正则化机器学习理论研究”，项目号 KJ130709；②重庆工商大学研究项目“基于多核学习的高维数据分析研究”，项目号 2013-56-09；③电子商务及供应链系统重庆市重点实验室研究项目“基于迹比率的特征选择及关键技术研究”；④重庆市教委研究项目“大数据稀疏表示判别字典学习及其应用技术研究”，项目号 KJ1400612；⑤重庆工商大学研究生教改项目“基于二维码的研究生互动教学改革”，项目号 2015YJG0205。在此，对提供资助的单位或机构一并表示感谢！

编写本书的过程也是作者进行研究和不断学习的过程。为了做到概念准确，内容详实可靠，作者查阅了大量的相关文献和资料。但由于时间仓促，作者能力和水平有限，书中内容难免出现差错。若发现问题，欢迎读者通过电子邮件 jsjhxp@ctbu.edu.cn，liubo7971@163.com 与作者取得联系，希望能一起探讨，共同进步。

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

第 1 章　深度学习概述

1.1　什么是深度学习

人工智能（artificial intelligence，AI）是计算机应用的一个重要分支，是通过让计算机来模拟人的智慧，从而实现对人类智能扩展的一套理论和方法。人工智能是对人的思维等信息进行模拟的过程。

人工智能有很多不同的分支，其中重要的分支有机器学习、计算机视觉、机器人、自然语言处理和专家系统等。人工智能的研究目标是使机器具有人的智能，然后完成一些需要人类智能才能完成的复杂工作。人工智能从诞生以来，理论和技术日益成熟，应用领域也不断扩大。

从人工智能发现的阶段来看，20 世纪 50 年代至 70 年代初，人工智能处于“推理期”，那时主要是让计算机具有逻辑推理能力。随着计算机的飞速发展，人们发现让计算机具有逻辑推理能力远远不够，于是在 20 世纪 70 年代中期，人工智能进入了“知识期”，在这个阶段出现了大量的专家系统，但人们发现专家系统面临“知识工程瓶颈”，即由人将知识总结出来再教给计算机相当困难，于是很多研究者想到让机器自己来学习[1]。

机器学习（machine learning，ML）的概念是 20 世纪 50 年代提出的，它是研究如何使用计算机来模拟人类学习活动的一门学科，它的研究目的是让机器像人一样获取新知识和新技能。到了 20 世纪 90 年代中期，统计机器学习迅速发展，并很快占据统治地位。统计机器学习的代表性技术是支持向量机（support vector machine，SVM）[2]。

统计机器学习（后面简称机器学习）是通过数据来建立模型并用该模型对数据进行预测和分析的过程。

机器学习主要分为两个步骤[3]：模型设计与使用模型。所谓模型设计是指从应用环境采集数据，进行预处理，再用一些算法（如支持向量机等）对数据进行训练，并由此得到相应模型。使用模型是指对新采集的数据，采用第一步所得到的模型进行数据分析和处理。图 1-1 为机器学习的流程图。它主要由信息获取、预处理、数据降维（如选择特征）、训练模型、分析和处理数据等 5 部分组成。

下面对各个步骤进行简要说明。

（1）信息获取。信息获取是指在各种应用中，利用相关的设备（如数码照相机、摄像机等）将各种对象信息转换为计算机可接受的信息并保存到存储设备中。另外，信息获取也有可能是收集用户输入的数据（如网站数据、微博数据等）。

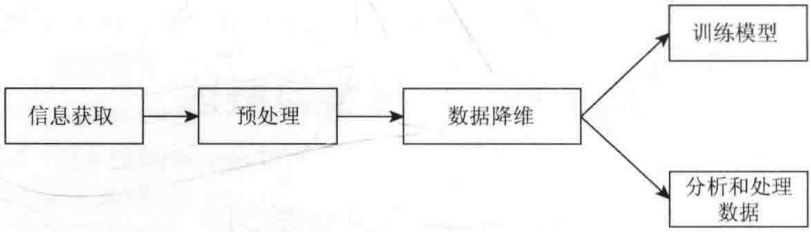


图 1-1　机器学习的流程图

（2）预处理。信息预处理是将噪声从采集到的数据中去除，并对信息数据进行分解、合并，最后将其转化为适合相应算法处理的数据。

（3）数据降维。在很多应用中，所采集的原始数据有很高的维数。为了让算法能高效准确地处理这些高维数据，需通过算法对原始数据的特征进行选择。其目的是用尽量少的特征来描述原始数据，并保持原始数据的特性。通过特征选择，不但可以减少算法的处理时间，还可以提高分类精度。

（4）训练模型。为了对数据进行正确处理，需要根据应用来选择合适的算法，并由此建立相应的机器学习模型。

（5）分析和处理数据。按已确定的模型对新输入数据进行判断并输出相应的分类结果。

在上述 5 个阶段中，特征的好坏对后续步骤中的机器学习算法的精度有很大影响。

另外，在 20 世纪 50 年代后期，曾出现基于神经网络的机器学习方法，但由于其结构复杂、容易出现过拟合等问题，后来慢慢被人们放弃。

但在 21 世纪初，基于神经网络的机器学习方法又卷土重来，这就是近年所掀起的深度学习（deep learning，DL），它逐渐成为机器学习研究中的一个新的领域。简单地讲，深度学习就是很多层的神经网络。深度学习的出现，带来了多种应用性能的巨大提高，包括语音识别、目标检测与识别、药物发现和基因组学等。深度学习的出现有两方面的原因：①计算能力的提高，随着大规模计算集群等高性能设备的出现，人们能快速有效地计算多层神经网络；②大数据的出现，随着计算机的飞速发展，各个领域都会产生大数据，大数据的特点就是数据维度高、数据量巨大。下面简单介绍一些经常产生高维数据的领域[4]。

（1）基因表达数据。美国科学家于 1990 年启动人类基因组计划（human genome project，HGP）。该计划主要是了解生命本质、生命体生长规律、生命之间的联系，存在个体差异的原因，以及认识和理解疾病产生的机制。在人类遗传变异基因中最常见的是单核苷酸的多态性（single nucleotide polymorphism，SNP），它占所有已知多态性的 90%以上。SNP 大量存在于人类基因组中，每 500～1000 个碱基对就有 1 个。据初步统计，它的数量约为 300 万个。SNP 同其他基因数据一起构成

了一个大规模高维度数据。又如，蛋白质和核酸是原生质的重要成分，它们是生命的基础物质之一。蛋白质通过反应来催化生命体，在调节生命体内的新陈代谢、抵御各类细菌入侵以及控制生命体中各种遗传信息等方面都起着非常重要的作用。在生物化学及相关的其他学科（如食品检验、临床检验等）中，蛋白质的分离和定性以及定量分析是很重要的步骤。人体蛋白质数据的维度高达 15154 维。基因芯片又称 DNA 芯片，是研究生物基因的有效工具，它主要研究氨基酸序列（蛋白质序列）和核酸序列（DNA 和 RNA 序列）等。目前，数值型是基因芯片数据的主要表达形式，这些值以矩阵的方式存储，即基因表示矩阵。样本在不同水平下的形式是用该矩阵的一行来表示。相同水平下所有样本的表达形式是用该矩阵的列来表示。基因在特定条件下的表达值就是矩阵中对应的元素值。基因表示矩阵规模庞大，通常涉及数千或数万的基因数，但其样本数非常小，一般只有数十个，这是典型的高维小样本数据。基因选择是微阵列基因数据分析的核心内容，它既是建立有效分类器的关键，又是发现疾病分类的标志。目前，科研工作者正在对该问题进行探索。如何从成千上万个基因中高效地选出有效特征用于分类，一直是基因数据分析的难点[5]。

（2）Web 文本数据。网络的飞速发展，产生了大量 Web 网页。以新浪、搜狐等国内著名的门户网站为例，其主栏目在 100 个以上，在每个主栏目下面，又有很多的子栏目，每个子栏目下 Web 页面的内容也不尽相同。标注这些不同内容的网页，会产生高维数据。另外，电子邮件已成为人们相互交流和通信的一种便捷的工具，但垃圾邮件会影响人们的正常生活。据统计，2012 年二季度中国网民平均每周收到垃圾邮件数量为 15.3 封，中国网民平均每周收到垃圾邮件比例为 34.7%，同比上涨 1.5 个百分点。企业邮箱平均每周收到正常邮件 57.8 封，收到垃圾邮件 29.5 封，垃圾邮件占 33.8%。普通个人邮箱垃圾邮件，垃圾信息过多影响比例为 67.7%，无法发送大附件为 66.0%，企业邮箱垃圾邮件，垃圾信息过多影响比例为 58.1%。如何区别正常邮件（ham E-mail）和垃圾邮件（spam E-mail），是一个重要的研究课题。在用词和行文格式等方面，垃圾邮件与正常邮件不一样。因此可以对邮件内容进行分析，用关键词方法基本可以有效区分垃圾邮件和正常邮件。目前市面上采用的垃圾邮件识别系统（如 Norton AntiSpam、SAproxy Pro 等）都是从每个邮件中抽取特征（也称关键词），然后采用分类算法，对这些邮件进行分类，从而识别垃圾邮件。但这些特征所构成的样本数据维度非常高，而且只有极少数特征对分类有用。由此可见，Web 文本是高维数据。通过对这类文档数据进行特征选择，可以大大提高分类精度和处理效率。

（3）图像数据。图像数据通常都是高维数据。在图像数据中，人脸数据最常见。人脸识别在公共安全、军事安全、国家安全等领域有着十分广阔的应用。同时也在智能监控、智能交通、智能门禁、公安布控中的身份识别与验证、出入境

管理等领域广泛使用。人脸识别是对测试的人脸图像用训练得到的特征表示，然后进行识别。计算机识别人脸的复杂表情是一个极其困难的事情。这是因为：人脸本身存在一定的弹性，会随着人的情绪而不断变化；随着年龄的增长，人脸会变化（变衰老）；由于拍摄人脸的光照、成像角度及成像距离的不同，所得到的人脸图像会差别很大。此外，由于图像设备的精度不断提高，在一般情况下，人脸数据可以达到几百万维，甚至上千万维像素。一般证件上人脸照片的像素也有几万，例如，如果一幅人脸图像的长和宽都是 512 个像素，该图像数据为 262144 维的向量，这是非常高维的数据。人脸图像的高维性使人脸识别变得比较困难。

（4）时间序列数据。在股票分析、证券期货、水文气象、工业过程控制、金融、医疗诊断、科学实验等领域经常按时间顺序记录一系列数据，这些有序数据称为时间序列数据。时间序列数据与静态数据不同，它是按等间距的时间段来获取数据，其值随时间的变化而不同。对时间序列数据分析的应用十分广泛，但难度也相当大。通常如果对时间序列数据的采样频率较高或采样持续的时间较长时，其数据维度相当高。例如，对某个事件，用 $x_1,x_2,\cdots,x_n$ 表示在固定的时间间隔 $t_1,t_2,\cdots,t_n$ 上的取值，可用 $\boldsymbol{X}=x_1,x_2,\cdots,x_n$ 表示该事件。这个 n 维的向量是时间序列数据，它的维数一般很高。将经典的分类或聚类算法用于高维时间序列数据时，会大大增加这些算法的时间复杂度和空间复杂度。因此，在处理这些数据之前，需要对它们进行预处理。

（5）推荐系统中用户评价数据。推荐系统的任务是用网站来联系用户与信息。一方面让信息能够展现在对它感兴趣的用户面前，另一方面帮助用户发现有价值的信息，引导用户获得想要的结果。最典型的推荐系统应用是电子商务领域中的 B2C 模式。商家根据用户的喜好、兴趣，向用户推荐感兴趣商品（如图书、衣服等）。在难以把握顾客的需求时，如果卖家通过向用户推荐商品来满足用户模糊需求，就可以将用户的潜在需求转化为现实需求，从而促进产品销售量。对于从事电子商务的大型网站，如 taobao、china-pub、amazon 等，推荐系统被大量使用。其中 amazon 花了大约 10 年时间来研究推荐系统在电子商务中的应用。一些具有个性化服务的 Web 网站，如 IMDb 和最大的 DVD 租赁商 Netflix，也对推荐系统有很大的依赖。推荐系统能够与用户建立长期稳定的关系，为用户提供个性化服务，对防止用户流失和提高用户忠诚度都有很大的作用。目前的推荐系统可以依赖的数据有客户活跃度信息、用户标签信息、用户对商品的反馈数据、时间上下文信息（如系统时间特性等）、社交网络等。其中用户反馈数据是用户根据对所购商品的感受来对该商品进行评分。这是用户喜好、兴趣最真实的反映。通过该数据可以划分不同的用户群体，从而把对某个用户行为的预测转换为对与该用户有相似行为的群体行为的预测。用户对商品的反馈数据由商品类别、用户爱好等构

此为试读，需要完整PDF请访问：www.ertongbook.com

成。其中，商品类别非常多，一般有几千至几万。因此，推荐系统所涉及的数据通常是高维数据。可用大数据训练具有很多参数的多层神经网络，从而发现大数据中的复杂结构。

近年来，深度学习在各个分支都取得了很大的进步，如深度卷积网络在处理图像、视频、语音和音频方面带来了突破；而递归网络在处理序列数据，如文本和演讲方面表现出快速的发展。后面会详细介绍在这些领域的进展情况。

深度学习的概念源于人工神经网络的研究。含多隐层的多层感知器就是一种深度学习结构。深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。随着抽象等级的增加，表现形式的等级增加；每一个阶段是一种可训练特征的转换；如识别图像，这些阶段分别是像素→边缘→纹理基元→主题→部分→对象，而在文本数据中，这些阶段分别是字符→词→句子→事件。

深度学习的奠基人是加拿大多伦多大学计算机系教授 Hinton，他和他的学生 2006 年在《科学》上发表的论文 *Reducing the Dimensionality of Data with Neural Networks* 中提出了深度学习的相关概念。其主要观点：①多隐层的神经网络具有很好的特征学习能力，所得到的特征更能反映数据的本质，从而有利于可视化或分类；②深度神经网络在训练上的难度，可以通过“逐层初始化”（layer-wise pre-training）来有效克服，而逐层初始化可通过无监督学习实现。从那时起，深度学习受到很多研究机构、大公司的关注，并被投入大量人力和财力来研究。

Google 公司的大脑项目主要就是研究深度学习，它用机器来模拟人脑进行数据处理。这个项目是 2011 年由斯坦福大学的 Andrew Ng 教授主导的，项目利用 Google 的分布式计算框架计算和学习大规模人工神经网络，用 16000 个 CPU Core 的并行计算平台训练学习到含 10 亿参数的深层神经网络（deep neural networks，DNN）的模型（这一网络内部共有 10 亿个节点，自然不能跟人类的神经网络相提并论。要知道，人脑中有 150 多亿个神经元，互相连接的节点也就是突触数更多），能够在没有任何先验知识的情况下，仅通过无标注的 YouTube 的视频来学习并识别高级别的概念，如猫，这就是著名的“Google Cat”。这个项目的技术已经被应用到了安卓操作系统的语音识别系统上。

为了充分利用这些先进的算法，Google 不断扩充自己的深度学习研究领域，如 Google 还在探索如何让机器理解人们的观点和情绪，对大量无标签的数据进行了学习研究。如果未来能够找到一种可行的算法来让机器对无标签的数据进行识别，那将有可能会改变整个计算行业，毕竟现在网络的大部分数据（如 Facebook、Twitter 和谷歌）都是没有标签的。这也正是深度学习技术未来想要实现的目标。利用数万台计算机通过软件模拟人脑中的神经元网络，从而让机器获得与人类相似的学习能

力，如在某些情况下机器能够在无须对数据添加标签的情况下实现自动学习。

2013 年年初，百度成立深度学习研究院（Institute of Deep Learning，IDL），CEO 李彦宏亲自任院长。2014 年 5 月 16 日，Google 大脑项目创始人 Andrew Ng 正式加盟百度，出任百度首席科学家，负责百度深度学习研究院工作，尤其是百度大脑计划。百度大脑融合深度学习算法、数据建模、大规模 GPU 并行化平台等技术，拥有 200 亿个参数，构造起深度神经网络，在政府、NGO、制造、金融、零售、教育等领域开展项目合作。

深度学习给计算机视觉带来了重大突破。在该方法应用于 ImageNet 大赛之前，第一届参赛冠军的准确率（top 5 精度）是 71.8%，而 2011 年是 74.3%。从 2009 年开始，全球多个顶尖小组都参加了 ImageNet 比赛，在 2012 年，传统的计算机视觉算法遇到瓶颈，进展缓慢。但这一年出现了深度学习的方法，对计算机视觉来说是革命性的。2012 年的冠军小组采用了深度学习的方法一举将准确率提升到 84.7%，这给机器学习领域带来了巨大影响，随后世界各大科研团体和公司纷纷投身于深度学习领域。截至目前，这一赛事的精度已经达到 95%以上，这在某种程度上与人眼的分辨能力相当。

深度学习主要用来学习特征，它被认为是表示学习（representation learning）（也称为特征学习）的一个分支[6]。表示学习是通过计算机来学习特征，以更好表示数据，其得到的特征通常比手工设计的特征要好。目标表示学习成为机器学习社区研究的热点。

图 1-2 为人工智能、机器学习、表示学习、深度学习之间的关系。

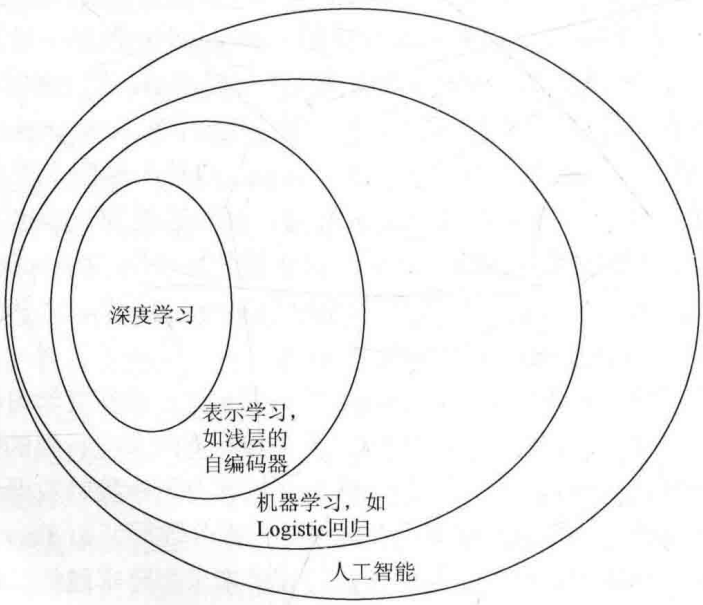


图 1-2　人工智能、机器学习、表示学习、深度学习之间的关系

下面对深度学习的一些特点进行介绍。

1.1.1　深度学习与表示学习

用于机器学习的训练样本都是通过特征（feature）来表示的。特征通常是指对事物或过程进行度量得到的数据。设有 k 个训练样本，每个样本由 n 个特征构成，则用于机器学习的训练集 $\boldsymbol{X}$ 可表示成下面这样的矩阵[4]：

$$\boldsymbol{X}=\begin{array}{c} \\ f_1 \\ f_2 \\ \vdots \\ f_n \end{array}\begin{array}{c} \begin{array}{cccccc} x_1 & \cdots & x_i & \cdots & x_k \end{array} \\ \begin{bmatrix} f_{11} & \cdots & f_{1i} & \cdots & f_{1k} \\ f_{21} & \cdots & f_{2i} & \cdots & f_{2k} \\ \vdots & & \vdots & & \vdots \\ f_{n1} & \cdots & f_{ni} & \cdots & f_{nk} \end{bmatrix}_{n\times k} \end{array}$$

通常对特征处理的方式是特征学习，它主要包含两方面的内容：①特征选择；②特征变换。特征选择是基于某种评价标准从原始特征中选择最优特征子集来表示数据。对具有 n 个原始特征的训练样本集 $\boldsymbol{X}$，给定要选择的特征数 $m(m \ll n)$，按某种评价标准从训练集 $\boldsymbol{X}$ 的原始特征中选择 m 个特征，从而得到新的训练集 $\overline{\boldsymbol{X}}$，使 $\overline{\boldsymbol{X}}$ 最能有效表达训练样本集 $\boldsymbol{X}$。新的训练集 $\overline{\boldsymbol{X}}$ 可用下面的矩阵表示：

$$\overline{\boldsymbol{X}}=\begin{array}{c} \\ f_1 \\ f_2 \\ \vdots \\ f_m \end{array}\begin{array}{c} \begin{array}{cccccc} x_1 & \cdots & x_i & \cdots & x_k \end{array} \\ \begin{bmatrix} f_{11} & \cdots & f_{1i} & \cdots & f_{1k} \\ f_{21} & \cdots & f_{2i} & \cdots & f_{2k} \\ \vdots & & \vdots & & \vdots \\ f_{m1} & \cdots & f_{mi} & \cdots & f_{mk} \end{bmatrix}_{m\times k} \end{array}$$

特征变换是指将样本的原始特征通过某种变换方法（或变换函数）得到新的特征，使得新特征更适合学习任务，这个过程可以看成是学习特征表示的过程。特征变换的方法非常多，常见的有线性变换（如主成分分析法和 Fisher 判别法等）和非线性变换（如基于核方法的主成分分析法等）。在计算机视觉中，图像的特征变换要比通常的特征变换复杂得多，因为图像信息是用矩阵表示，它是一种二维结构，也就是说图像会涉及空间信息，而且图像的特征变换需要考虑光照不变性、尺度不变性、旋转不变性等多种特性。对图像的特征变换通常称为特征提取。

深度学习就是一种特征变换方法，把原始数据通过一些简单的但是非线性的模型转变成为更高层次的，更加抽象的表达，从而完成对样本的重新表示。深度学习可得到各种层次的特征表示。对于分类任务，高层次的表达能够强化输入数据的区分能力方面，同时削弱不相关因素。例如，一幅图像的原始格式是一个像

素数组，那么在第一层上的学习特征表达通常指的是在图像的特定位置和方向上有没有边的存在。第二层通常会根据那些边的某些排放来检测图案，这时候会忽略掉一些边上的一些小的干扰。第三层或许会对那些图案进行组合，从而使其对应于熟悉目标的某部分。随后的一些层会将这些部分再组合，从而构成待检测目标。深度学习的核心方面是，上述各层的特征都不是利用人工工程来设计的，而是使用一种通用的学习过程从数据中学到的。

良好的特征表达，对最终算法的准确性起到非常关键的作用；识别系统的计算和测试工作耗时主要集中在特征提取部分；如果数据被很好表达成了特征，通常线性模型就能达到满意的预测精度。对于具体的特征变换，需要考虑特征表示的粒度、层次结构、数量。

（1）特征表示的粒度。特征表示的粒度大小对机器学习算法影响较大。如一幅图像，基于像素级的特征根本没有价值，因为像素级的特征可能得不到任何有用信息，要用这些特征分类（判断是否为摩托车）会很困难。但如果将特征表示的粒度放大，使表示的特征具有结构性，如将车把手（handle）和车轮（wheel）作为对象，学习算法就很容易区分摩托车和非摩托车。

（2）特征表示的层次结构。很多时候，用特征表示对象时需要有层次性，这就需要特征由浅入深构建。小块的图形可以由基本边构成，这些小块又可以构成复杂的图像。文档由句子构成，句子由词按一定语法规则构成，而词由字构成。

（3）特征表示的数量。在通过层次结构来表示特征时，每一层该有多少个特征呢？任何一种方法，特征越多，给出的参考信息就越多，准确性会得到提升。但特征多意味着计算复杂，探索的空间大，可以用来训练的数据在每个特征上就会稀疏，都会带来各种问题，并不一定特征越多越好。因此需要确定好特征数，这样才能得到好的效果。

1.1.2　深度学习与神经网络

深度学习的结构与普通的神经网络有很多相似之处，主要表现在：它们都有输入层、隐藏层（多层）、输出层，都属于多层网络结构，它们只有相邻层节点之间有连接，同一层以及跨层节点之间相互无连接，都有激活函数（通常会使用Sigmoid 函数）。

1. 神经网络

普通神经网络的层数较少。它采用反向传播（back propagation）算法来计算目标函数的梯度，在此基础上用坐标下降方法来求解目标函数。整个过程：随机设定各层的参数初值，计算目标函数的梯度，确定迭代的步长，然后更新参数直

到收敛。其缺点在于：

（1）比较容易过拟合，参数比较难调整，而且需要不少技巧；

（2）训练速度比较慢，在层次比较少（小于等于 3）的情况下效果并不比其他方法更优。

2. 深度学习

深度学习由深层（很多层）的神经网络构成，这些网络由一个输入层、多个隐层以及一个输出层构成。每层有若干个神经元，神经元之间有连接权重。每个神经元模拟人类的神经细胞，而节点之间的连接模拟神经细胞之间的连接。深度学习采用逐层训练机制来得到各层的参数的初始值。然后反向传播来计算目标函数的梯度。在深度网络（deep networks）（7 层以上）中，残差传播到最前面的层将变得很小，出现所谓的梯度扩散（gradient diffusion）。

（1）深度神经网络模型复杂，需要的训练数据量大，计算量大。一方面，深度神经网络是对人脑的模拟，而人脑包含 100 多亿个神经细胞，这要求深度神经网络中神经元越多效果才越好。从数学的角度看，深度神经网络中每个神经元都包含激活函数（如 Sigmoid、ReLU 或者 Softmax 函数），需要估计的参数量也极大。语音识别和图像识别应用中，神经元达数万个，参数多达上千万个，这样的模型非常复杂，而要求解这样的模型需要很大的计算量。另一方面，深度神经网络需要大量数据才能训练出高准确率的模型。深度神经网络参数量大，模型复杂，为了避免过拟合，需要海量训练数据。基于这两方面因素，训练一个模型耗时惊人。以语音识别为例，目前业界通常使用样本量达数十亿，若用一台普通的计算机，需要数年才能完成一次训练。

（2）深度神经网络需要更高硬件配置支持大模型。目前的实验已经证明：通过增加卷积层的滤波器数量，加大网络的深度等方式可以获得更好的模型，但模型参数也随之增加。以 ImageNet 2012 竞赛冠军[7]的网络为例，其占用 3.99GB 的显存，已接近主流 GPU 的显存容量，试图增大模型则会超过 GPU 显存范围。因此，如何支持更大模型的网络对硬件平台是一个大的挑战。

（3）深度神经网络训练收敛难，需要反复多次实验。深度神经网络是非线性模型，其代价函数是非凸函数，容易收敛到局部最优解。同时，深度神经网络的模型结构、输入数据处理方式、权重初始化方案、参数配置、激活函数选择、权重优化方法等均可能对最终结果有较大影响。另外，深度神经网络的数学基础研究稍显不足。虽然可以通过受限玻尔兹曼机（restricted Boltzmann machines，RBM），或者去噪自编码器（denoising autoencoder，DA）等产生式建模方法初始化网络模型，以达到减少获得局部最优的风险，但仍然不能彻底解决问题。在实际中使用深度神经网络时，要合理地利用海量数据，合理地选择优化方式。上述

原因导致需要技巧、经验，基于大量实验才能训练出一个好的模型。

以上的问题研究人员正在想办法解决。虽然有这么多问题，但深度学习确实能取得很好的效果。下面对深度学习在各个领域取得如此好的效果的原因进行简单介绍。

1）浅层学习的局限

人工神经网络（BP 算法）虽被称为多层感知机，但实际是一种只含有一层隐藏层节点的浅层模型。很多统计机器学习算法（如支持向量机等）都是带有一层隐层节点或没有隐层节点（如 Logistic 回归）的浅层模型。

浅层模型局限性在于：在样本和计算单元有限的情况下对复杂函数的表示能力有限，针对复杂分类问题其泛化能力较差。

深度学习通过构建多隐层的模型和大量训练数据来学习更有用的特征，从而提升分类或预测的准确性。深度模型是手段，对特征进行学习是目的。深度学习强调了模型结构的深度，通常有 5～10 层的隐层节点，而且明确突出了特征学习的重要性，即通过逐层特征变换，将样本的原始特征变换到一个新的特征空间，从而使分类或预测更加容易。与人工构造特征的方法相比，利用大数据来学习特征，更能够刻画数据的丰富内在信息。

在许多情形中，只要两层隐藏层就足够逼近任何给定误差精度的函数。但是这种结构的问题在于所需要的节点数可能变得非常大。在这种情形下，只有通过增加深度来减少节点数量。

2）大脑具有深度结构

目前，人们对视觉皮质进行了很好的研究，并观察到人脑的一系列区域的功能。这些区域按层次连接在一起。在每一区域中，包含一个输入的表示和信号流。每一层表示上一个层的输入，越是上层其抽象程度越高。这个结构包含如下三方面的内容。

（1）人脑视觉机理：1981 年的诺贝尔医学奖获得者 David Hubel 和 Torsten Wiesel 得到了视觉系统的信息处理机制，他们发现了一种被称为“方向选择性细胞”的神经元细胞，当瞳孔发现了眼前的物体的边缘，而且这个边缘指向某个方向时，这种神经元细胞就会活跃。

（2）视觉的层次性：人的视觉系统的信息处理是分层的；高层的特征是低层特征组合得到的，从低层到高层的特征表示越来越抽象，越来越能表现语义或者意图；抽象层面越高，存在的不确定就越少，就越利于分类。

（3）特征表示的粒度：具有结构性（或者语义）的高层特征对于分类更有意义。需要注意的是，大脑中的表示在中间紧密分布并且纯局部，它们是稀疏的：1%的神经元是同时活动的。给定大量的神经元，仍然有一个非常高效的（指数级高效）表示。

此为试读，需要完整PDF请访问：www.ertongbook.com