CHAPTER 1

第 1 章 绪　论

近些年，深度学习因在许多领域都取得了耀眼的成绩和突破性的进展而受到学术界及工业界的广泛关注，本章将分别从深度学习与机器学习的关系、深度学习与统计学的关系、深度学习框架、深度学习中涉及的优化方法以及深度学习展望五个方面对深度学习进行全面深刻的剖析，旨在为后续学习提供理论铺垫与指导。

1.1　机器学习与深度学习

斯坦福大学终身教授、ImageNet 数据库的缔造者、现任 Google Cloud 首席科学家的华裔科学家李飞飞认为“人工智能将成为新的生产力，成为第四次工业革命的主要推动力之一”。人工智能重在实现机器智能，实现的方式为机器学习。作为人工智能的重要分支，机器学习主要研究的是如何使机器通过识别和利用现有知识来获取新知识和新技能。自 20 世纪 80 年代以来，机器学习已经在算法、理论和应用等方面都取得巨大成功，而被广泛应用于产业界与学术界。简单来说，机器学习就是通过算法使得机器能从大量历史数据中学习规律，从而对新的样本完成智能识别或对未来做预测。

而深度学习是机器学习的一个分支和新的研究领域，如今在大数据的背景下，可用数据量的激增、计算能力的增强以及计算成本的降低为深度学习的进一步发展提供了平台，同时也为深度学习在各大领域中的应用提供了支撑。自 AlphaGo 被提出并成功击败

职业围棋手后，“深度学习”这一概念快速进入人们的视野并在业界引起了轰动，其因强大的特征提取能力以及灵活性而在国内外各大企业中掀起一阵狂潮，在语音识别、图像识别和图像处理领域取得的成果尤为突出。深度学习的本质在于利用海量的训练数据（可为无标签数据），通过构建多隐层的模型，去学习更加有用的特征数据，从而提高数据分类效果，提升预测结果的准确性。

本节将从时期阶段发展和模型结构发展的角度介绍机器学习与深度学习之间的关系，并在此基础上从六个方面对机器学习和深度学习进行对比，从而进一步阐述二者之间的关系。

1.1.1 机器学习与深度学习的关系

机器学习的发展历程大致可以分为五个时期，而伴随着机器学习的发展，深度学习共出现三次浪潮。接下来，以机器学习的发展作为主线来介绍不同时期机器学习与深度学习之间的关系。

第一个时期从 20 世纪 50 年代持续至 20 世纪 70 年代，由于在此期间研究人员致力于用数学证明机器学习的合理性，因此称之为“推理期”。在此期间深度学习的雏形出现在控制论中，随着生物学习理论的发展与第一个模型的实现（感知机，1958 年），其能实现单个神经元的训练，这是深度学习的第一次浪潮。

第二个时期从 20 世纪 70 年代持续至 20 世纪 80 年代，由于在这个阶段费根鲍姆（Edward Albert Feigenbaum）等机器学习专家认为机器学习就是让机器获取知识，因此称之为“知识期”，在此期间深度学习主要表现在机器学习中基于神经网络的连接主义。

第三个时期从 20 世纪 80 年代持续至 20 世纪 90 年代，这个时期的机器学习专家主张让机器“主动”学习，即从样例中学习知识，代表性成果包括决策树和 BP 神经网络，因此称这个时期为“学习期”。在此期间深度学习仍然表现为基于神经网络的连接主义，而其中 BP 神经网络的提出为深度学习带来了第二次浪潮。其实在此期间就存在很好的算法，但由于数据量以及计算能力的限制致使这些算法的良好效果并没有展现出来。

第四个时期从 20 世纪初持续至 21 世纪初，这时的研究者们开始尝试用统计的方法分析并预测数据的分布，因此称这个时期为“统计期”，这个阶段提出了代表性的算法——支持向量机。而此时的深度学习仍然停留在第二次浪潮中。

第五个时期从 20 世纪初持续至今，在这个时期神经网络再一次被机器学习专家重视。2006 年 Hinton 及其学生 Salakhutdinov 发表的论文《 Reducing the Dimensionality of Data with Neural Networks 》标志着深度学习的正式复兴，该时期掀起深度学习的第三次浪潮，同时在机器学习的发展阶段中被称为“深度学习”时期。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的技术以及手工设计功能的 AI 系统。而在此之后，伴随着数据量的爆炸式增长与计算能力的与日俱增，深度学习得到了进一步的发展。

根据机器学习的模型结构，认为机器学习有两次里程碑式的变革。第一次变革为浅层学习，所谓浅层学习是指网络层数较少（多为一层）的人工神经网络。称其为第一次变革主要是因为在此阶段提出了反向传播算法，该算法的提出可以使人工神经网络模型从大量的训练样本中“学习”出统计规律，从而对未知事件做出预测。第二次变革为深度学习，区别于浅层神经网络，深度学习强调了模型结构的深度，同时明确突出了特征学习的重要性，即通过逐层特征变换，将样本在原空间的特征变换到一个新特征空间，从而更加容易地进行分类或预测。

总之，无论从发展历程的角度还是从模型结构的角度出发，深度学习都与机器学习息息相关，并且在机器学习领域中占有重要地位，影响着机器学习的发展趋势。

1.1.2　传统机器学习与深度学习的对比

传统机器学习与深度学习在理论与应用上都存在差异，下面将分别从数据依赖、硬件支持、特征工程、问题解决方案、执行时间以及可解释性这六个方面对传统机器学习与深度学习的差别进行比较。

数据依赖：深度学习和传统机器学习最重要的区别是前者的性能随着数据量的增加而增强。如果数据很少，深度学习算法性能并不好，这是因为深度学习算法需要通过大

量数据才能很好地理解其中蕴含的模式。在这种情况下，使用人工指定规则的传统机器学习占据上风。

硬件支持：深度学习算法严重依赖于高端机，而传统机器学习在低端机上就可以运行。因为深度学习需要进行大量矩阵乘法操作，而 GPU 可以有效优化这些操作，所以 GPU 成为其中必不可少的一部分。

特征工程：特征工程将领域知识输入特征提取器，降低数据复杂度，使数据中的模式对学习算法更加明显，并得到更优秀的结果。从时间和专业性方面讲，这个过程开销很大。在机器学习中，大多数使用的特征都是由专家指定或根据先验知识来确定每个数据域和数据类型。比如，特征可以是像素值、形状、纹理、位置、方向。大多数传统机器学习方法的性能依赖于识别和抽取这些特征的准确度。

问题解决方案：当使用传统机器学习方法解决问题时，经常采取化整为零，分别解决，再合并结果的求解策略。而深度学习主张端到端的模型，即输入训练数据，直接输出最终结果，让网络自己学习如何提取关键特征。如图 1-1 所示为传统机器学习和深度学习对比流程图。

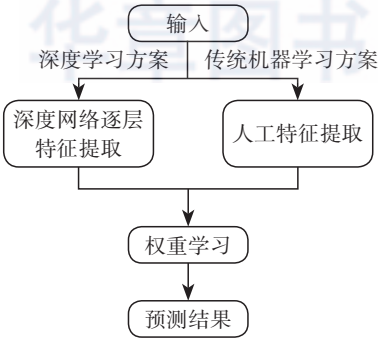

图 1-1　传统机器学习和深度学习对比流程图

执行时间：深度学习需要进行很长时间的训练，因为深度学习中很多参数都需要进行远超正常水平的时间训练，如 ResNet 大概需要两周时间从零开始完成训练，而机器学

习只需要从几秒到几小时不等的训练时间。测试所需要的时间就完全相反，深度学习算法运行只需要很少的时间。

可解释性：假定使用深度学习给文章自动评分，会发现性能很不错，并且接近人类评分水准，但它不能解释为什么给出这样的分数。在运行过程中可以发现深度神经网络的哪些节点被激活，但不知道这些神经元是对什么进行建模以及每层在干什么，所以无法解释结果。另一方面，机器学习算法如决策树按照规则明确解释每一步做出选择的原因，因此像决策树和线性 / 逻辑回归这类算法由于可解释性良好，在工业界应用很广泛。

1.2　统计学与深度学习

统计学是一门古老的学科，其作为机器学习的理论基础这一事实在从 20 世纪 60 年代就开始被学术界所认可。直到 20 世纪 90 年代，伴随着统计学理论的基本成熟，研究者们开始尝试用统计学的方法分析并预测数据的分布，由此产生了著名的支持向量机算法，如今这种算法已被广泛应用于数据分析、模式识别、回归分析等各个领域。

1.2.1　统计学与深度学习的关系

深度学习作为机器学习中重要的分支，因此与统计学同样具有密不可分的关系。通常可以将统计学分为两大类，分别为用于组织、累加和描述数据中信息的描述统计学和使用抽样数据来推断总体的推断统计学。深度学习则是通过大量的样本数据学习总体规则的方法，可见深度学习是统计学对实践技术的延伸。

另外，实际的应用领域中经常需要处理的数据都具有随机性和不确定性，对这些数据最好的描述方式就是通过概率来进行描述。例如，在图像识别中，若要对模糊或残缺的图像进行识别，即在不确定的条件下实现图像的正确识别，基于统计学的深度学习由于可以处理数据的随机性以及不确定性，因此可以在恶劣的条件下实现图像的精准识别。

深度学习的特点在于先设计能够自我学习的神经网络，然后将大量的数据输入网络

中进行训练，通过训练神经网络能够从数据集中学到数据的内在结构和规律，从而对新数据做出预测。

从统计学的角度来看，深度学习用来训练的数据集即为样本，学习的过程即为对总体信息进行估计。对于无监督学习来说，每一个输入样本是一个向量，学习过程相当于要估计出总体的概率分布。而对于监督学习来说，每个输入样本 x 还对应一个期望的输出值 y，称为 label 或 target，那么学习的过程相当于要估计出总体的条件概率分布。这样，当系统遇到新的样本时，就能给出对应的预测值 y。

1.2.2 基于统计的深度学习技术

最典型的基于统计的深度学习技术有受限玻耳兹曼机以及生成对抗式网络。

受限玻耳兹曼机（Restricted Boltzmann Machine，RBM）是一种可用随机神经网络来解释的概率图模型。随机神经网络的核心在于在网络中加入概率因素，而其中的随机是指这种网络中的神经元是随机神经元，其输出只有两种状态（0 或 1），而状态的取值根据概率统计的方法确定。RBM 属于深度学习中常用的模型或方法，其结构如图 1-2 所示。

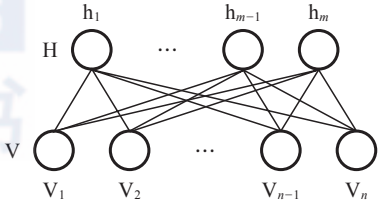

图 1-2 RBM 结构图

其中，下层为输入层，包括 n 个输入单元 v_n，用来表示输入数据；上层为隐藏层，包含 m 个隐藏层单元 h_m，RBM 具有层内无连接、层间全连接的特征，这一特点可以保证 RBM 各层之间的条件独立性。

由于 RBM 为概率模型，而训练 RBM 网络的实质就在于能够使 RBM 所表达出的概率分布尽可能接近真实样本的分布。而实现这个目的 RBM 经典训练算法就是对比散度（Contrastive Divergence，CD）算法，即在每次训练过程中，以数据样本为初始值，通过 Gibbs 采样获取目标分布的近似采样，然后通过近似采样获得目标梯度，取得最终的结果。简单来说，统计学在受限玻耳兹曼机中的应用过程为对图像进行联合分布概率的描述，通过训练可以使 RBM “学”到输入数据的统计规律，从而达到提取特征的目的。

RBM 网络是以统计学为基础进行构建和训练的，是最典型的基于统计的深度学习技术。

生成对抗式网络（Generative Adversarial Networks，GAN）是一种新型网络，是由 Goodfellow 等人在 2014 年提出来的。其基本思想源自博弈论中的二人零和博弈，网络模型由一个生成网络和一个判别网络构成，生成网络用来学习样本的真实分布并用服从某一分布（高斯分布或均匀分布）的噪声生成新的数据分布，判别网络用来判别输入是真实样本还是生成网络生成的样本，通过生成网络与判别网络的对抗学习进行网络的训练。GAN 的优化过程是极小极大博弈（Minimax game）问题，具体是指判别网络的极大化（即判别网络要尽可能区分真实样本和生成网络生成的样本）和生成网络的极小化，即生成网络生成的样本要尽可能“欺骗”判别网络，使其认为是真实的样本，优化目标为达到纳什均衡，使生成网络估测到数据样本的分布。GAN 的计算流程与结构如图 1-3 所示。

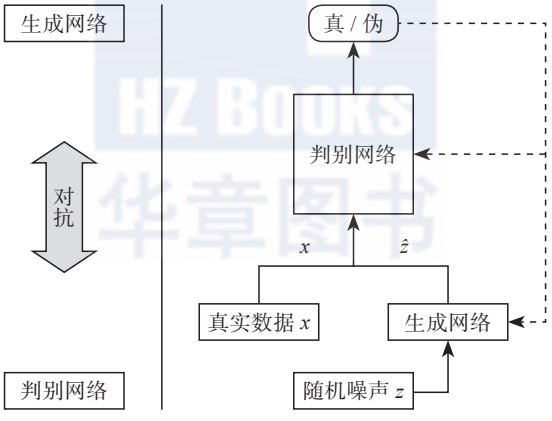

图 1-3　生成对抗式网络流程与结构

生成对抗式网络作为一种基于统计学的新型深度学习技术，通过模型学习来估测其潜在分布并生成同分布的新样本，被广泛应用于图像和视觉、语音与语言、信息安全等领域，如今许多研究者试图将其与强化学习结合进行进一步的研究。

作为深度学习的重要理论基础，未来统计学还有非常大的发展空间。因为深度

学习模型具有较好的非线性函数表示能力，根据神经网络的通用近似理论（universal approximation theory）可知，对于任意的非线性函数一定可以找到一个深度学习网络来对其进行表示，但是“可表示”并不代表“可学习”，因此需要进一步了解深度学习的样本复杂度，即需要多少训练样本才能得到一个足够好的深度学习模型。这些问题都有待于从理论层面进行突破，统计学对深度学习的进一步发展有着十分重要的意义。

1.3 本书涉及的深度学习框架

随着深度学习技术的不断发展，越来越多的深度学习框架得到开发。目前，最受研究人员青睐的深度学习框架有 TensorFlow、Caffe、Torch 和 MXNet。TensorFlow 框架作为一个用于机器智能的开源软件库，以其高度的灵活性、强大的可移植性等特点而成为目前深度学习的主流框架之一；而对于 Caffe，研究者可以按照该框架定义各种各样的卷积神经网络框架，该框架以表达方便、速度快、组件模块化等优势同样成为当今常用的深度学习网络框架；Torch 是一个广泛支持机器学习算法的科学计算框架，其使用简单快速的脚本语言 LuaJIT 以及底层的 C/CUDA 进行实现，因此以易于使用且高效的特点而成为当下流行的深度学习框架；MXNet 是一个以高效和灵活为目的设计的开源深度学习框架，支持命令式编程和声明式编程。这四种框架以各自的优势特点而受到广大研究者的认可，在本书第 2 ～ 5 章将会就这四种框架的理论内容、具体搭建过程（包括涉及的代码描述）以及应用实例进行详细的介绍与分析。

1.4 优化深度学习的方法

目前，深度学习在多种目标分类和识别任务中取得优于传统算法的结果，并产生大量优秀的模型，使用迁移学习方法将优秀的模型应用在其他任务中，可以达到在减少深度学习训练时间的前提下，提升分类任务性能，同时降低对训练集规模的依赖，关于迁移学习及其实例分析将在第 6 章进行详细介绍。

除此之外，随着深度学习模型中网络层数的加深、参数的增多、计算量的加大，计

算速度慢、资源消耗多的问题逐渐成为不可忽视的挑战，以保证深度学习训练精度的同时加快训练速度为目的的并行计算与交叉验证运用而生，这两种方法的详细介绍以及实例分析将在第 7 章进行。

1.5　深度学习展望

随着硬件计算能力的提升以及大规模数据集的出现，深度学习已经成为机器学习中一个重要的领域，下面对深度学习的一些模型进行介绍。

卷积神经网络（Convolutional Neural Network，CNN）是一类适用于处理图像数据的多层神经网络。CNN 从生物学上的视觉皮层得到启发：视觉皮层存在微小区域的细胞对于特定区域的视野十分敏感，这就对应着 CNN 中的局部感知区域。在 CNN 中，图像中的局部感知区域被当作层次结构中的底层输入数据，信息通过前向传播经过网络中的各个层，每一层都由过滤器构成，以便能够获得观测数据的一些显著特征，局部感知区域能够获得一些基础的特征，还能提供一定程度对位移、拉伸和旋转的相对不变性。CNN 通过结合局部感知区域、共享权重、空间或者时间上的降采样来充分利用数据本身包含的局部性等特征，优化网络结构；通过挖掘数据空间上的相关性，来减少网络中可训练参数的数量，以达到改进反向传播算法效率。

长短期记忆（Long Short-Term Memory，LSTM）网络主要适用于处理序列数据。LSTM 网络是一种特殊的 RNN（循环神经网络），但网络本质与 RNN 是一样的。在传统的神经网络模型中，网络的传输是从输入层到隐藏层再到输出层，层与层之间是全连接的，每层之间的节点是无连接的。这其中存在一定的问题，即传统的神经网络对于处理时序问题无能为力。LSTM 网络可以解决长时期依赖的问题，主要是因为 LSTM 网络有一个处理器，其中放置了“三扇门”，分别称为输入门、遗忘门和输出门。一个信息进入 LSTM 网络当中，可以根据规则来判断是否有用，只有符合算法认证的信息才会留下，不符合的信息则通过遗忘门被“遗忘”。所以可以很好地处理序列数据。

受限玻耳兹曼机（RBM）是一种用随机神经网络来解释的概率图模型。RBM 适用于

处理语音、文本类数据。当使用 RBM 建立语音信号模型时，该模型使用对比散度（CD）算法进行有效训练，学习与识别任务关联性更高的特征来更好地得到信号的值。在文档分类问题中，直接将不规范的文档内容作为输入会产生过高的输入数据维数，而无法对其进行处理，因此有必要对文档进行预处理，选择词组出现的频率作为特征项以提取能够表示其本质特征的数据，使用 RBM 可从原始的高维输入特征中提取可高度区分的低维特征，然后将其作为支持向量机的输入进行回归分析，从而实现对文档的分类。

生成对抗式网络（GAN）适用于处理图像数据，估计样本数据的分布，解决图片生成问题。GAN 包含一个生成模型（Generative Model）G 和一个判别模型（Discriminative Model）D，生成模型 G 捕捉样本数据的分布，即生成图片；判别模型 D 是一个二分类器，判别图片是真实数据还是生成的。在训练过程中，首先固定一方，再更新另一个模型的参数，以此交替迭代，直至生成模型与判别模型无法提高自己，即判别模型无法判断一张图片是生成的还是真实的。模型的优化过程是一个二元极小极大博弈问题，在 G 和 D 的任意函数空间中，存在一个唯一的解，G 恢复训练数据分布，D 在任何地方都等于 0.5。该网络可以为模拟型强化学习做好理论准备，在缺乏数据的情况下，可以通过生成模型来补足。

深度学习算法在大规模数据集下的应用取得突破性进展，但仍有以下问题值得进一步研究。

1）无标记数据的特征学习。当前，标记数据的特征学习占据主导地位，但是对于标记数据来说，一个相当困难的地方在于将现实世界的海量无标记数据逐一添加人工标签，是很费时费力且不现实的。所以，随着科学研究的发展，无标记数据的特征学习以及将无标记数据进行自动添加标签的技术会成为研究主流。

2）模型规模与训练速度、训练精度之间的权衡。一般地，在相同数据集下，模型规模越大，则训练精度越高，训练速度越慢。对于模型优化，诸如模型规模调整、超参数设置、训练时调试等，其训练时间会严重影响其效率。所以，如何在保证一定的训练精度的前提下提高训练速度是很有必要的一个研究课题。

3）大规模数据集的依赖性。深度学习最新的研究成果都依赖于大规模数据集和强大的计算能力，如果没有大量真实的数据集，没有相关的工程专业知识，探索新算法将会变得异常困难。

4）超参数的合理取值。深度神经网络以及相关深度学习模型应用需要足够的能力和经验来合理地选择超参数的取值，如学习速率、正则项的强度以及层数和每层的单元个数等，一个超参数的合理值取决于其他超参数的取值，并且深度神经网络中超参数的微调代价很大，所以有必要在超参数这个重要领域内做更进一步的研究。

在许多领域深度学习都表现出巨大的潜力，但深度学习作为机器学习的一个新领域现在仍处于发展阶段，仍然有很多工作需要开展，很多问题需要解决，尽管深度学习的研究还存在许多问题，但是现有的成功和发展表明深度学习是一个值得研究的领域。