深度强化学习图解

[美] 米格尔·莫拉莱斯(Miguel Morales) 著
郭 涛 译

清華大學出版社
北 京

北京市版权局著作权合同登记号 图字：01-2021-4871
Grokking Deep Reinforcement Learning
Miguel Morales
EISBN: 978-1-61729-545-4

图书在版编目 (CIP) 数据

深度强化学习图解 / (美) 米格尔 • 莫拉莱斯(Miguel Morales) 著；郭涛译. —北京：清华大学出版社，2022.5
书名原文：Grokking Deep Reinforcement Learning
ISBN 978-7-302-60546-1

Ⅰ. ①深… Ⅱ. ①米… ②郭… Ⅲ. ①机器学习—图解 Ⅳ. ① TP181-64

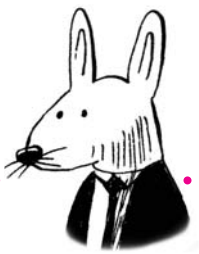
中国版本图书馆 CIP 数据核字 (2022) 第 062675 号

责任编辑：王 军
装帧设计：孔祥峰
责任校对：成凤进
责任印制：刘海龙

出版发行：清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦A座 邮 编：100084
社 总 机：010-83470000 邮 购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经 销：全国新华书店
开 本：170mm×240mm 印 张：27.25 字 数：534千字
版 次：2022年7月第1版 印 次：2022年7月第1次印刷
定 价：139.00元

产品编号：089967-01

专家推荐



AlphaGo 战胜韩国顶尖棋手李世石已成为人工智能领域里程碑式的事件。AlphaGo 核心技术便是深度强化学习。深度强化学习结合了强化学习的决策能力和深度学习的复杂函数表示能力，有望在自动驾驶、新材料研发和机器人控制领域发挥更大作用。然而无论是深度学习、强化学习还是深度强化学习，对于初学者而言学习门槛都很高。本书是一本不可多得的对初学者友好的深度强化学习图书。通过大量的插图和生动有趣的例子，不仅有利于初学者入门，专业读者也能从中受益。

——欧高炎，
北京大数据研究院研究员

本书以图解方式将深度强化学习基础、重要算法和前沿技术进行详尽讲解，降低了学习门槛，拉近了内容与读者的距离，可让读者快速上手。如果你想钻研深度强化学习，请研读本书！

——董豪，
北京大学计算机学院助理教授

2015 年以来，随着深度学习在工业领域的稳步落地，强化学习插上深度学习的翅膀。DeepMind 公司的 AlphaGo 在机器打败人类顶级棋手和电竞选手时风靡全球，其跟踪研究长期霸榜近年来的计算机学术界顶级会议。随后，每年一届的“腾讯开悟多智能体强化学习大赛”在全球顶尖高校院所办得风风火火，世界著名电竞游戏《王者荣耀》团队和腾讯 AI Lab 近年来向学术界开放其深度强化学习研究经验及测试资源，刘渝教授带领优秀大学生获得第二届决赛前十名；整个过程精彩纷呈，影响深远，促进了深度强化学习在广袤无垠的前沿议题的进展，例如通过深度强化学习能近乎 100% 在癌症早期帮助高效检查出患者的恶性肿瘤，其检测准确率和效率已远高于专

业医师；通过深度强化学习，L4 级别自动驾驶高阶方案也正在落地。

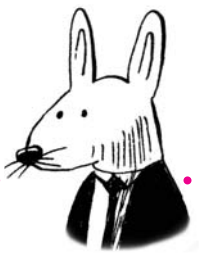
本书的译者和作者长期致力于人工智能和成果转化产研一线，其追索真知的热情和慎思明辨的推敲跃然纸上，给读者带来了公式概化的数学之美和代码凝练的无缝验证，行文严密又不失风趣，连配图都精巧悦目，相信有缘分的读者一定不愿错过这趟探索深度强化学习的奥秘的美好旅程，尽享理论和实践完美结合的沿途风景。

——张东映博士，
华中科技大学土木与水利工程学院，
武汉光电国家研究中心智能存储实验室

本书从深度强化学习的基础理论知识，到代码技术实现做了深入浅出的讲解，经典算法与案例实战教学贯穿始终。此书对深度强化学习爱好者及科研教学人员是非常好的学习材料。

——席武宝，
泰迪科技北京分公司总经理

译 者 序



强化学习是机器学习范式和方法论之一，智能体主要通过不断与环境交互进行自我学习来达到回报最大化，从而解决特定领域的问题。强化学习主要用于智能机器控制、计算机视觉和博弈论等领域。在以往的研究中，强化学习在理论方面遇到一系列挑战，例如探索 - 利用困境、策略优化过程、策略评估等方面的问题。为解决这一系列问题，深度学习的理论和产业界的一些突破性成果为强化学习提供了一种新思路。学者以一种通用的形式将深度学习的感知能力和强化学习的决策能力相结合，并通过端到端的学习方式实现从原始输入到输出的直接控制。由于强化学习仍有不足，新的理论进一步推动深度强化学习的发展，在弥补缺陷的同时扩展了强化学习的研究领域，延伸出模仿学习、分层强化学习和元学习等新的研究方向。

深度强化学习 (Deep Reinforcement Learning，DRL) 是深度学习和强化学习的巧妙结合，是一种新兴的通用人工智能技术，是人工智能迈向智能决策的重要一步，是机器学习的热点，潜力无限，典型的成功案例是 DeepMind AlphaGo 和 OpenAI Five。深度强化学习可看作在深度学习非线性函数超强拟合能力下，构成的一种新增强算法。目前就深度强化学习而言，需要从三个方面进行积累：第一，深度强化学习的理论基础；第二，深度强化学习的仿真平台；第三，产业落地的项目和产品。从深度强化学习库以及框架看，学术界 PyTorch 和工业界 Tensor Flow 深度学习框架都将前沿成果集成进来。目前已有一些经典的深度强化学习文献和著作，但将深度强化学习理论、工具和实战相结合的著作还是很少，本书的出版恰好填补了这方面的空白。

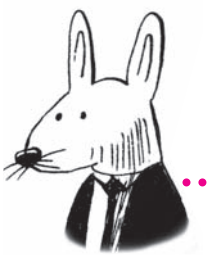
本书图文并茂地对晦涩难懂的深度强化学习理论进行描述，并结合大量的案例和应用程序，引导读者边思考边实践，从而逐步加深对深度强化学习的理解，并将这些新方法、新理论和新思想用于自己的研究。本书可作为从事智能机器人控制、计算机视觉、自然语言处理和自动驾驶系统 / 无人车等领域研究工作的工程师、计算

机科学家和统计学家的参考书。

在翻译本书的过程中，我得到了很多人的帮助。首都师范大学朱琳教授，中国交通通信信息中心孙云华博士，南京大学的王小平博士，吉林大学外国语学院研究生吴禹林，吉林财经大学外国语学院研究生张煜琪、许瀚、王耀珩、张漾予和艾俊等对本书进行了技术审核和校对工作，感谢他们所做的工作。最后，感谢清华大学出版社的编辑，他们做了大量的编辑与校对工作，保证了本书的质量，使得符合出版要求。在此深表谢意。

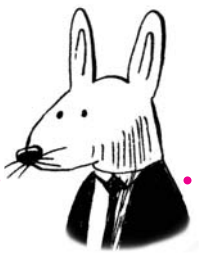
由于本书涉及的广度和深度较大，加上译者翻译水平有限，在翻译过程中难免有不足之处，若各位读者在阅读过程中发现问题，欢迎批评指正。

译者简介



郭涛，主要从事模式识别、人工智能、智能机器人、软件工程、地理人工智能 (GeoAI) 、时空大数据挖掘与分析等前沿交叉研究。担任《复杂性思考: 复杂性科学与计算模型 (第 2 版)》《神经网络设计与实现》和《概率图模型及计算机视觉应用》等畅销书的译者。

推　荐　序



本书是关于强化学习的介绍。出于众多原因，这一主题既不易学习，也不易于教学。首先，这是个相当有技术性的话题。它需要大量的数学知识和理论作为支撑。涉及适量的相关背景而不全然沉浸其中，这本身也是个挑战。

其次，强化学习(Reinforcement Learning，RL)会导致一种概念上的错误。强化学习是一种决策问题的思维方式，也是解决这些问题的工具。我们将 RL 称为“一种思维方式”，是因为 RL 提供了一个决策框架，涉及状态和奖励信号等细节。将其称为“一套工具”，是因为我们在讨论 RL 时，会使用像马尔可夫决策过程(Markov Decision Processes，MDP)和贝尔曼更新(Bellman Update)这样的术语。但当我们用数学工具来验证 RL 思维方式时，我们的思维方式很容易被混淆。

最后，RL可通过多种方式得以实现。由于RL是一种思维方式，我们可通过尝试以一种非常抽象的方式构成框架来讨论它，或将其放入代码中，或将其放入神经元中。你所选择的基板(实现方式)将产生两个难题甚至更多挑战——这便把我们带入深度强化学习中。

专注于深度强化学习会很好地一次性整合所有问题，这便是RL和深度神经网络的背景。两者都值得以各种方式研究和开发，但解答如何在开发工具的背景下解释这两者并非易事。同时不要忘记，理解RL不仅需要理解深度网络中工具实现本身及应用，也要理解RL的思维方式；否则，将不能越过直接研究的例子进行概括。讲授RL也很难，深度强化学习的教学方式中也有很多错误——这便将我们带到了Miguel Morales写的这本书。

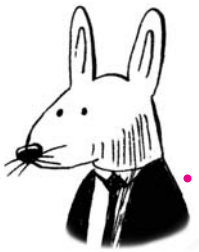
本书的优势在于用地道的技术语言清晰地解释什么是机器学习、深度学习以及强化学习；也让读者了解该领域的大背景、深度强化学习技术的用处，以及 ML、RL 和深度强化学习所体现的思维方式。本书的讲解简洁明了，既是学习指南也是学习参考书，至少对于我们来说，是灵感的来源。

对于这一切，我不感到惊讶，我与 Miguel 相识多年。他曾经是机器学习课程的学习者，后来成为一名教师。我曾在佐治亚理工学院的理科硕士在线课程中，讲授了多个学期的强化学习和决策课程，他一直都是首席教学助理；那段时间，他接触到成千上万的学生。我看着他逐步成长为一名从业者、研究者和教育者。在他的帮助下，佐治亚理工学院的 RL 课程取得了进步。我写推荐序，意在加深学生对强化学习的体会。在此期间，他的帮助也从未停止，他是个天生的老师。

本书展现了他的才华。很高兴能和他一起工作，也很高兴他能写这本书。享受本书吧。我从中学到很多，希望你们也是。

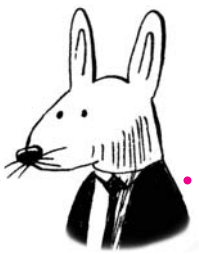
Charles Isbell 教授
佐治亚理工学院计算机系主任

关于作者



Miguel Morales 任职于科罗拉多州丹佛市的 Lockheed Martin 公司导弹和火控自动系统部门，从事强化学习工作。他是佐治亚理工学院强化学习和决策课程的兼职教学助理。Miguel 曾是优达学城机器学习项目的评审人，自动驾驶汽车纳米学位导师，以及深度强化学习纳米学位内容开发人员。他毕业于佐治亚理工学院，主修交互人工智能。

自　序



强化学习是一个令人兴奋的领域，在人类历史上有可能产生深远的影响。一些技术已经影响了世界，改变了人类发展进程，从火到交通、电，再到互联网，每项技术的发明都以复合的方式推动着下一项技术的出现。没有电，个人计算机就不会诞生；没有电，互联网就不会存在；没有电，搜索引擎就不会出现。

对我而言，RL 和人工智能最让人兴奋的方面通常不仅是身边让人激动不已的智能体，而是智能体的出现所产生的深远影响。相信强化学习是一个能自主优化特定任务的强大框架，它有着改变世界的潜力。除了任务自动化，智能机器的创造还将人类智能延伸到未知之地。可以说，如果能确定如何为每个问题找到最优决策，便能理解那些找到最优决策的算法。我能感受到，智能体的出现会使人类的生活变得更加智能。

但现实中，这样的智能体仍遥不可及，为了实现这些疯狂的梦想，我们需要在工作中引入更多思维方式。强化学习当前处于初级阶段，因此一段时间内还有很多工作要做。写这本书的目的是让更多人了解深度 RL 和通用 RL，并为这个领域做出贡献。

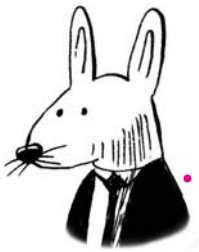
尽管 RL 框架很直观，但对新手来说，大多数资源还不能理解。我的目标不是写一本只提供代码例子的书，更不是创建一个讲授强化学习理论的资源，而是创建一种能够弥补理论与实践之间空白的资源。如你所见，我并不回避公式：如果想了解一个研究领域，它是必不可少的。即使有建立高质量的 RL 解决方案这般实践性的目标，仍然需要理论基础。然而，我也不完全依赖公式。不是所有对 RL 感兴趣的人都喜欢数学，有些人则更喜欢代码和实际例子，所以本书在这个奇妙的领域提供了应用实践。

在这个为期三年的项目中，我大部分的努力都是为了弥补空白。我从不避讳从直观解释理论出发，也并不会简单地给出代码示例。对于这两者，我都非常详细地

去描写。理解教材和课程对于一些人来说很难，但他们可以轻松理解顶级研究人员使用的词汇：为何是这些特定的词，而不是其他的词。还有一些人，他们知道这些词的意思，喜欢阅读公式，却难以理解用代码表示的公式以及公式之间的联系，虽然他们可以轻易理解强化学习应用方面的知识。

最后，我希望你能喜欢本书，更希望它能实现自身的价值，为你助力。我希望你能在了解深度强化学习的过程中学有所得，为这个奇妙的领域做出贡献。如前所述，如果没有大量的现代科技创新，你就没有机会读到本书，但读完本书之后会发生什么，全由你来决定。所以大胆探索吧，去震撼这个世界。

致　　谢



我要感谢佐治亚理工学院的人们，是他们冒着风险为全世界所有人提供了第一个计算机科学硕士在线课程，让大家可以接受高质量的研究生教育。如果不是这些人，我可能就不会撰写本书。

我要感谢院长 Charles Isbell 教授和 Michael Littman 教授，他们整合了一门优秀的强化学习课程。尤其感激 Isbell 教授，他让我成长，给了我很多学习 RL 的机会。同时，师从 Littman 教授，我学会了强化学习的教学方式——把问题分解成三种类型。能得到这两位资深教授的指导，我万分荣幸。

我要感谢佐治亚理工学院 CS 7642 充满活力的老师们，他们与我探讨如何帮助学生学到更多知识，我很享受与他们共度的时光。特别感谢 Tim Bail、Pushkar Kolhe、Chris Serrano、Farrukh Rahman、Vahe Hagopian、Quinn Lee、Taka Hasegawa、Tianhang Zhu 和 Don Jacob 老师，你们都是很棒的队友。我还要感谢之前为这门课程做出巨大贡献的人们，从与你们的交流学习中我也受益匪浅：Alec Feuerstein、Valkyrie Felso、Adrien Ecoffet、Kaushik Subramanian 和 Ashley Edwards。感谢学生们的提问，帮助我发现了那些尝试学习 RL 的人在知识上的差距。在此，我要特别感谢一位匿名的学生，是他把我推荐给 Manning 出版社，建议我写这本书。我撰写本书时常会想起你，感谢你。

我要感谢 Lockheed Martin 公司的员工，感谢他们在我撰写本书时给予的反馈和互动。特别感谢 Chris Aasted、Julia Kwok、Taylor Lopez 和 John Haddon。John 是第一个审阅初稿的人，他的反馈帮助我把写作水平提升到一个新层次。

我要感谢 Manning 提供了框架，让本书顺利付梓。我要感谢 Brian Sawyer 伸出援手，让我打开写作的大门。感谢 Bert Bates 的早期指引，帮助我专注于教学。感谢 Candace West 帮助我从默默无闻到小有成就。感谢 Susanna Kline 帮我紧跟时代的步伐。感谢 Jennifer Stout 在最后阶段为我加油打气。感谢 Rebecca Rinehart 的帮助。

感谢 AI Krinker 提供的反馈，让我在评论中分离出有价值的信息。Matko Hrvatin 一直在关注 MEAP 的发布，并给了我继续写作的额外动力。感谢 Candace Gillhooley 和 Stjepan Jurekovic 帮助我完成本书。感谢 Ivan Martinovic 提供了急需的关于润色文字的反馈。感谢 Lori Weidert 在本书出版前所做的两次调整。感谢 Jennifer Houle 耐心地修改、设计。感谢 Katie Petito 耐心地处理细节。感谢 Katie Tennant 的一丝不苟和最后的润色。还要感谢那些我没提到的幕后人员，是他们让本书的出版成为可能。还有很多需要感谢的人，谢谢大家的辛勤工作。

致所有评论者——Al Rahimi、Alain Couniot、Alberto Ciarlanti、David Finton、Doniyor Ulmasov、Edisson Reinozo、Ezra Joel Schroeder、Hank Meisse、Hao Liu、Ike Okonkwo、Jie Mei、Julien Pohie、Kim Falk Jørgensen、Marc-Philippe Huget、Michael Haller、Michel Klomp、Nacho Ormeño、Rob Pacheco、Sebastian Maier、Sebastian Zaba、Swaminathan Subramanian、Tyler Kowallis、Ursin Stauss 和 Xiaohu Zhu，感谢你们的建议让这本书更加精彩。

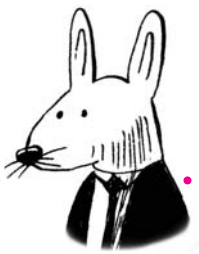
感谢优达学城让我与学生分享我对这个领域的热情，并录制 actor-critic 算法的讲座。特别感谢 Alexis Cook、Mat Leonard 和 Luis Serrano。

感谢 RL 社区的人们帮我完善内容，加深我对这一领域的理解。特别感谢 David Silver、Sergey Levine、Hado van Hasselt、Pascal Poupart、John Schulman、Pieter Abbeel、Chelsea Finn、Vlad Mnih 的讲座；感谢 Rich Sutton 提供这个领域的“黄金副本”(他的教科书)；感谢 James MacGlashan 和 Joshua Achiam 的代码库、线上资源以及当我无处寻求答案时所给予的指导；感谢 David Ha 为我指明前进的道路。

特别感谢 Silvia Mora 帮我塑造了书中所有的人物，并帮我完成几乎每一个项目。

最后，我想感谢我的家人，他们是这个项目的后盾。我认为写书很有挑战性，事实也如此。无论如何，我的妻儿都在陪伴我。谢谢你，Solo，在撰写本书的过程中点亮了我的生活。谢谢你，Rosie，谢谢你给予的爱。谢谢妻子 Danelle，谢谢你所做的一切。有趣的人生中，你们是我完美的队友，我很高兴能遇见你们。

前　言



本书是深度强化学习理论与实践的桥梁。本书内容适用于熟悉机器学习技术，想要学习强化学习的读者。本书开篇将介绍深度强化学习的基础知识。随后深入探索深度强化学习的算法和技术，最后会提供一份具有潜在影响力的先进技术调查。

本书适用对象

熟悉深度强化学习领域、Python 代码、一些数学知识，能够运用大量直观解释和有趣而具体的例子来推动学习的人，会喜欢本书。此外，任何熟悉 Python 的人都能学习到很多知识。即使 DL 知识不扎实，本书也能帮助读者对神经网络和反向传播及相关技术进行简单复习。最重要的是，本书不依赖于其他书，任何想简单了解 AI 智能体和深度强化学习的读者都能通过阅读本书达到目的。

本书结构

本书共 13 章，大致可分为两部分，前一部分包括 1 ～ 5 章，后一部分包括 6 ～ 13 章。

第 1 章是深度强化学习导论，介绍深度强化学习的概念，讲述本书的最佳使用方式。第 2 章涵盖强化学习的概念、数学基础和多智能体强化学习框架等内容。第 3 章通过最佳行为策略算法来解决惯序决策问题，学习妥善平衡短期目标与长期目标的方法。第 4 章以多臂老虎机为例，探索策略，来解决未知转换函数与奖励信号的问题，合理权衡信息收集与信息运用。第 5 章评估智能体的行为。第 6 章介绍在转换函数和奖励函数未知的情况下，构造强化学习环境中的优化策略，训练优化的智能体行为。第 7 章讲述如何基于动态规划思想来优化强化学习，从而获得更高效的实现目标。第 8 章介绍基于价值的深度强化学习。第 9 章探索函数逼近和基于价值的深度强化学习。第 10 章探索高效抽样的价值深度强化学习方法。第 11 章探索策

略梯度和 actor-critic 方法。第 12 章探索高级 actor-critic 方法。第 13 章讨论通用人工智能的未来发展方向。

关于代码

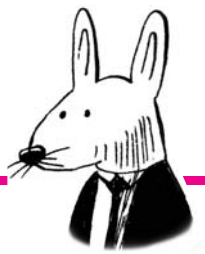
本书在“Python 讲解”栏目中列举了许多源代码示例。源代码使用等宽字体进行格式化，这样就可与普通文本区分开，并添加有序的高亮突出显示，这样能使它更易于阅读。

大多数情况下，原始源代码已被重新格式化，本书添加了换行符、重命名变量并重新调整了缩进，以适应书中可用的页面空间。即使如此，在极少数情况下页面空间也还不够，Python 中包括行连续操作符代码，即反斜杠(\)，指示语句在下一行继续。

此外，源代码中的注释经常会被删除，文本仅描述代码。代码注释指出重要的概念。

可扫描封底二维码下载本书的示例代码。

目　录



第1章　深度强化学习导论　1

1.1　深度强化学习概念　2
1.1.1　深度强化学习：人工智能的机器学习法　2
1.1.2　深度强化学习着重创建计算机程序　5
1.1.3　智能体解决智能问题　6
1.1.4　智能体通过试错提高性能　8
1.1.5　智能体从惯序性反馈中学习　9
1.1.6　智能体从评估性反馈中学习　10
1.1.7　智能体从抽样性反馈中学习　10
1.1.8　智能体使用强大的非线性函数逼近　11
1.2　深度强化学习的过去、现在与未来　12
1.2.1　人工智能和深度强化学习的发展简史　12
1.2.2　人工智能的寒冬　13
1.2.3　人工智能现状　13
1.2.4　深度强化学习进展　14
1.2.5　未来的机遇　17
1.3　深度强化学习的适用性　18
1.3.1　利弊分析　18
1.3.2　深度强化学习之利　19
1.3.3　深度强化学习之弊　20
1.4　设定明确的双向预期　21
1.4.1　本书的预期　21
1.4.2　本书的最佳使用方式　22
1.4.3　深度强化学习的开发环境　23
1.5　小结　24

第2章 强化学习数学基础 27

2.1 强化学习组成 28
2.1.1 问题、智能体和环境的示例 30
2.1.2 智能体：决策者 31
2.1.3 环境：其余一切 32
2.1.4 智能体与环境交互循环 37
2.2 MDP：环境的引擎 38
2.2.1 状态：环境的特定配置 40
2.2.2 动作：影响环境的机制 43
2.2.3 转换函数：智能体行为的后果 44
2.2.4 奖励信号：胡萝卜和棍棒 46
2.2.5 视界：时间改变最佳选择 49
2.2.6 折扣：未来是不确定的，别太看重它 50
2.2.7 MDP扩展 51
2.2.8 总体回顾 53
2.3 小结 54

第3章 平衡短期目标与长期目标 57

3.1 决策智能体的目标 58
3.1.1 策略：各状态动作指示 62
3.1.2 状态–值函数：有何期望 63
3.1.3 动作–值函数：如果这样做，有何期望 64
3.1.4 动作–优势函数：如果这样做，有何进步 65
3.1.5 最优性 66
3.2 规划最优动作顺序 67
3.2.1 策略评估：评级策略 67
3.2.2 策略改进：利用评级得以改善 73
3.2.3 策略迭代：完善改进后的行为 77
3.2.4 价值迭代：早期改进行为 81
3.3 小结 85

第4章 权衡信息收集和运用 87

4.1 解读评估性反馈的挑战 88
4.1.1 老虎机：单状态决策问题 89
4.1.2 后悔值：探索的代价 90

4.1.3　解决MAB环境的方法　91
4.1.4　贪婪策略：总在利用　93
4.1.5　随机策略：总在探索　95
4.1.6　ε贪婪策略：通常贪婪，时而随机　97
4.1.7　衰减ε贪婪策略：先最大化探索，后最大化利用　99
4.1.8　乐观初始化策略：始于相信世界美好　101
4.2　策略型探索　105
4.2.1　柔性最大值策略：根据估计值按比随机选择动作　106
4.2.2　置信上界策略：现实乐观，而非乐观　108
4.2.3　汤普森抽样策略：平衡回报与风险　110
4.3　小结　116

第5章　智能体行为评估　119

5.1　学习估计策略价值　120
5.1.1　首次访问蒙特卡洛：每次迭代后，改善估计　123
5.1.2　蒙特卡洛每次访问：处理状态访问的不同方法　125
5.1.3　时差学习：每步后改进估计　129
5.2　学习从多步进行估算　137
5.2.1　n步TD学习：经过几步后改进估计　138
5.2.2　前瞻TD(λ)：改进对所有访问状态的估计　141
5.2.3　TD(λ)：在每步之后改进对所有访问状态的估计　143
5.3　小结　151

第6章　智能体行为的优化　153

6.1　对智能体强化学习的解析　154
6.1.1　大多数智能体都要收集经验样本　156
6.1.2　大多数智能体都要评估　157
6.1.3　大多数智能体都要优化策略　159
6.1.4　广义策略迭代　160
6.2　学习动作策略的优化　162
6.2.1　蒙特卡洛控制：在每一迭代后优化策略　163
6.2.2　SARSA：在每一步之后优化策略　169
6.3　从学习中分离动作　173
6.3.1　Q学习：学会最优动作，即使我们不选　173
6.3.2　双Q学习：最大值估计值的最大估计值　177
6.4　小结　184

第7章 更有效、更高效地完成目标 187

7.1 学习使用鲁棒性目标优化策略 188
7.1.1 SARSA(λ)：基于多阶段评估，在每一阶段后优化策略 189
7.1.2 Watkin的Q(λ)：再一次，从学习中分离行为 196
7.2 智能体的交互、学习、计划 200
7.2.1 Dyna-Q：学习样本模型 201
7.2.2 轨迹抽样：为不久的将来做计划 206
7.3 小结 219

第8章 基于价值的深度强化学习 221

8.1 深度强化学习智能体使用的反馈种类 222
8.1.1 深度强化学习智能体处理惯序性反馈 223
8.1.2 如果它不是惯序性反馈，那它是什么 224
8.1.3 深度强化学习智能体处理评估性反馈 225
8.1.4 如果它不是评估性反馈，那它是什么 226
8.1.5 深度强化学习智能体处理抽样性反馈 226
8.1.6 如果它不是抽样性反馈，那它是什么 227
8.2 强化学习中的逼近函数 228
8.2.1 强化学习问题能够拥有高维状态和动作空间 229
8.2.2 强化学习问题可以具有连续的状态和动作空间 229
8.2.3 使用函数逼近有很多优点 231
8.3 NFQ：对基于价值的深入强化学习的第一次尝试 233
8.3.1 第1个决策点：选择逼近一个值函数 234
8.3.2 第2个决策点：选择神经网络体系结构 235
8.3.4 第3个决策点：选择要优化的内容 236
8.3.5 第4个决策点：为策略评估选择目标 238
8.3.6 第5个决策点：选择探索策略 241
8.3.7 第6个决策点：选择损失函数 242
8.3.8 第7个决策点：选择一种最优方法 243
8.3.9 可能出错的事情 248
8.4 小结 250

第9章 更稳定的基于价值方法 253

9.1 DQN：使强化学习更像是监督学习 254
9.1.1 基于价值的深度强化学习的普遍问题 254

9.1.2　使用目标网络　256
9.1.3　使用更大网络　259
9.1.4　使用经验回放　259
9.1.5　使用其他探索策略　263
9.2　双重DQN：减少对动作–值函数的高估　269
9.2.1　高估问题　269
9.2.2　将动作选择从动作评估剥离　270
9.2.3　一个解决方案　271
9.2.4　一个更实用的解决方案　272
9.2.5　一个更宽容的损失函数　275
9.2.6　仍可改进之处　280
9.3　小结　281

第10章　高效抽样的基于价值学习方法　285

10.1　Dueling DDQN：具备强化学习意识的神经网络架构　286
10.1.1　强化学习不属于监督学习问题　286
10.1.2　基于价值的强化学习方法的微妙区别　287
10.1.3　利用优点的优势　288
10.1.4　有意识强化学习框架　289
10.1.5　建立一个Dueling网络架构　290
10.1.6　重构动作–值函数　291
10.1.7　连续更新目标网络　293
10.1.8　Dueling网络能为表格带来什么　294
10.2　PER：优先有意义经验的回放　297
10.2.1　更明智的回放经验方法　297
10.2.2　如何较好地衡量“重要”经验　298
10.2.3　利用TD 误差做出贪婪优先级操作　299
10.2.4　随机对优先的经验进行抽样　300
10.2.5　成比例的优先级　301
10.2.6　基于排名的优先级　302
10.2.7　优先偏倚　303
10.3　小结　309

第11章　策略梯度与actor-critic方法　313

11.1　REINFORCE算法：基于结果策略学习　314
11.1.1　策略梯度法简介　314
11.1.2　策略梯度法之优势　315

11.1.3 直接学习策略 319
11.1.4 减少策略梯度方差 320
11.2 VPG：学习值函数 322
11.2.1 进一步减少策略梯度方差 323
11.2.2 学习值函数 323
11.2.3 鼓励探索 324
11.3 A3C：平行策略更新 328
11.3.1 使用actor工作器 328
11.3.2 使用*n*–step估计 331
11.3.3 无障碍模型更新 334
11.4 GAE: 稳健优势估计 335
11.5 A2C: 同步策略更新 338
11.5.1 权重分担模型 338
11.5.2 恢复策略更新秩序 340
11.6 小结 346

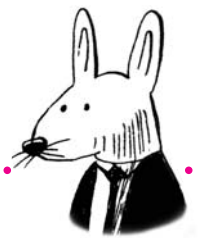
第12章 高级actor–critic方法 349

12.1 DDPG: 逼近确定性策略 351
12.1.1 DDPG使用DQN中的许多技巧 351
12.1.2 学习确定性策略 353
12.1.3 用确定性策略进行探索 356
12.2 TD3：最先进的DDPG改进 358
12.2.1 DDPG中的双重学习 358
12.2.2 平滑策略更新目标 360
12.2.3 延迟更新 363
12.3 SAC: 最大化预期收益和熵 365
12.3.1 在贝尔曼方程中添加熵 365
12.3.2 学习动作–值函数 366
12.3.3 学习策略 366
12.3.4 自动调整熵系数 367
12.4 PPO: 限制优化步骤 372
12.4.1 使用与A2C相同的actor–critic架构 372
12.4.2 分批处理经验 373
12.4.3 剪裁策略更新 377
12.4.4 剪裁值函数更新 377
12.5 小结 382

第13章　迈向通用人工智能　385

13.1　已涵盖的以及未特别提及的内容　386
13.1.1　马尔可夫决策过程　387
13.1.2　规划法　388
13.1.3　Bandit法　389
13.1.4　表格型强化学习　390
13.1.5　基于值函数的深度强化学习　391
13.1.6　基于策略的深度强化学习和actor-critic深度强化学习　392
13.1.7　高级actor-critic技术　392
13.1.8　基于模型的深度强化学习　393
13.1.9　无梯度优化方法　395
13.2　更多AGI高级概念　397
13.2.1　什么是AGI　397
13.2.2　高级探索策略　399
13.2.3　逆强化学习　399
13.2.4　迁移学习　400
13.2.5　多任务学习　401
13.2.6　课程学习　401
13.2.7　元学习　402
13.2.8　分层强化学习　402
13.2.9　多智能体强化学习　402
13.2.10　可解释AI、安全、公平和道德标准　403
13.3　接下来是什么　404
13.3.1　如何用DRL解决特定问题　404
13.3.2　继续前进　405
13.3.3　从现在开始，放下本书　406
13.4　小结　407

第1章 深度强化学习导论



本章内容：

- 了解深度强化学习的概念，以及它与其他机器学习法的不同之处。
- 了解深度强化学习的最新进展，以及它在解决各类问题时的作用。
- 了解本书的主要内容，以及如何最大化利用本书中的知识。

“我能想象到有一天我们之于机器人就如同狗之于人类，我为机器应援。”

——Claude Shannon
信息时代之父与人工智能领域贡献者

追求幸福是人类的天性。从挑选食物到推进事业，我们所做的每一个选择都源于我们想要体验生活中有益的时刻的欲望。无论这些时刻是为了个人的快乐还是更宏大的追求，无论带来的是当下的满足还是长久的成功，它们都是我们对自己重要性和价值的认知。在某种程度上，这些时刻就是我们存在的理由。

实现这些珍贵时刻的能力似乎与智能有关，“智能”定义为获取、应用知识和技能的能力。社会公认的聪明人，不仅能用当下的满足来换取长远目标的实现，而且能用美好的、已确定的未来去换取或许更好的未来，即使后者具有不确定性。需要较长时间才能实现的目标以及长期价值未知的目标通常是最难实现的，而那些能经受住此过程中的挑战的人们是例外，他们是领导者，是社会中的知识分子。

在本书中，你将了解到一种被称为深度强化学习的方法，该方法涉及创建能实现智能目标的计算机程序。本章将介绍深度强化学习，并给出一些建议，让你从本书中得到最大收获。

1.1 深度强化学习概念

深度强化学习(Deep Reinforcement Learning，DRL)是一种人工智能的机器学习方法，着重于创建计算机程序以解决智能问题。DRL 程序的独特属性在于通过试错从反馈中学习，这些反馈通过强大的非线性函数逼近进行抽样，同时具有惯序性和评估性。

下面逐步为你解读这个定义。不要太过于纠结细节，因为我将用整本书来让你理解深度强化学习。以下是对本书内容的介绍，后续章节会反复对其进行详解。

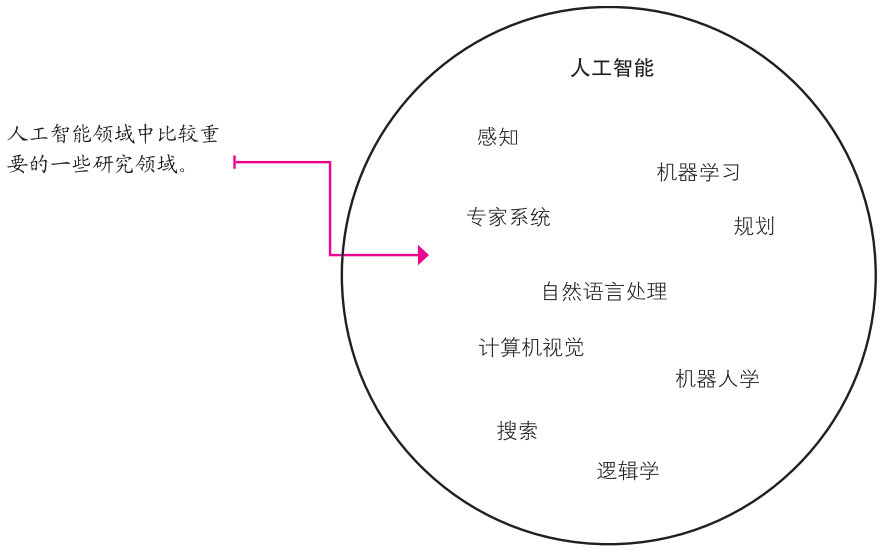
如果我能成功完成本书的目标，那么你在读完后，将能准确地理解这一定义。你将能理解我选用这些词语的原因，以及为何我没有使用更多或更少的文字来定义它。但对于本章而言，你只需要仔细通读一遍。

1.1.1 深度强化学习：人工智能的机器学习法

人工智能(Artificial Intelligence，AI)是计算机科学的一个分支，涉及创建智能的计算机程序。传统上，任何表现认知能力(如感知、搜索、规划和学习)的软件都属于AI的一部分。AI软件的功能示例如下：

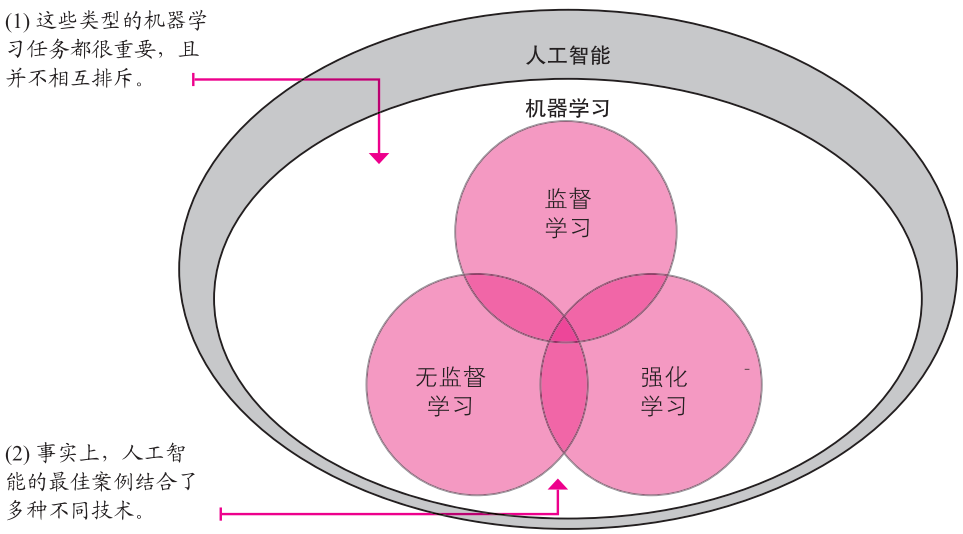
- 搜索引擎的返回页面功能
- GPS 应用程序的路线生成功能
- 智能助手软件的语音识别功能与合成语音功能
- 电子商务网站的产品推荐功能
- 无人机的跟随功能

人工智能子领域



所有表现智能的计算机程序都被认为是人工智能，但并不是所有的人工智能都具有学习能力。机器学习(Machine Learning，ML)是人工智能的一个领域，专注于创建计算机程序，通过从数据中学习的方法解决智能问题。ML 有三个主要分支：监督学习、无监督学习和强化学习。

机器学习的主要分支



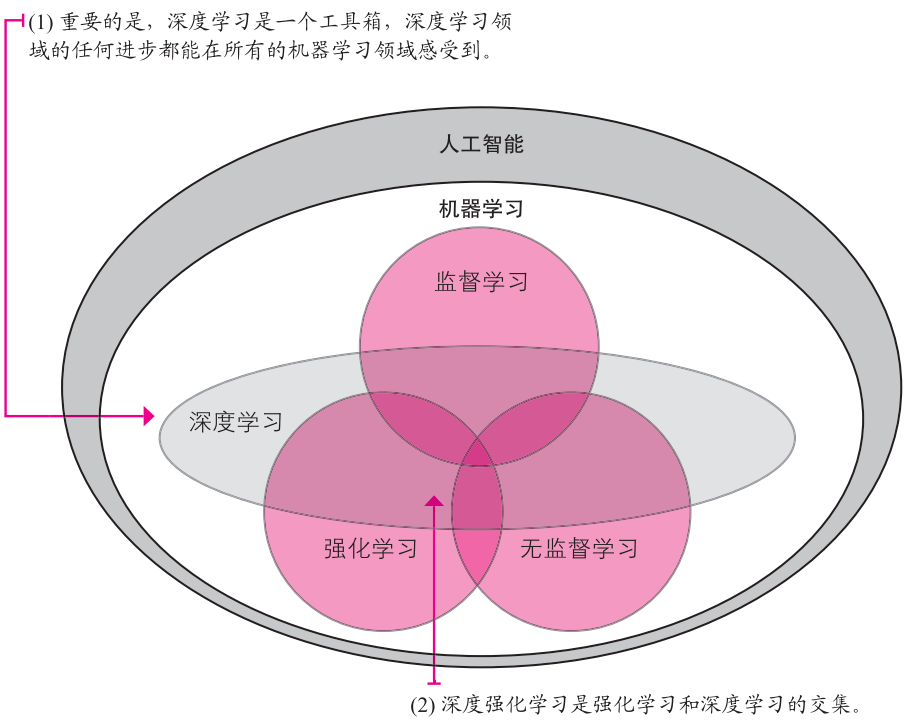
监督学习(Supervised Learning，SL)的工作是从标记数据中学习。在 SL 中，由人决定收集哪些数据以及如何对其进行标记。SL 的目标是归纳数据。手写数据识别应用程序是 SL 的一个经典案例：人工收集并标记带有手写数据的图像，并训练一个模型来正确识别和分类图像中的数字。经过训练的模型有望在新图像中对手写数字进行归纳和正确分类。

无监督学习(Unsupervised Learning，UL)的工作是从未标记数据中学习。即使不再需要标记数据，计算机收集数据的方法仍需要由人类设计。UL 的目标是压缩数据。UL 的一个经典案例是用户细分应用程序：人工收集客户数据，并训练模型将客户聚类。这些聚类对信息进行压缩，揭露潜在客户关系。

强化学习(Reinforcement Learning，RL)的工作是从试错中学习。在该类型任务中，没有人对数据进行标记，也没有人收集数据或明确设计数据集。RL的目标是执行。RL的一个经典案例是 Pong-playing 的智能体；该智能体反复与 Pong 模拟器互动，并通过采取动作和观察动作效果进行学习。经过训练的智能体应该能用成功执行 Pong 的方法执行其他任务。

深度学习(Deep Learning，DL)是一种强大的 ML 最新方法，使用多层次非线性函数逼近，通常是为神经网络。DL并不是 ML 的一个独立分支，所以与前述的那些任务并无不同。DL是技术和方法的集合，使用神经网络完成包括SL、UL和RL在内的ML任务；而DRL只是使用DL完成RL任务。

深度学习是一个强大的工具箱

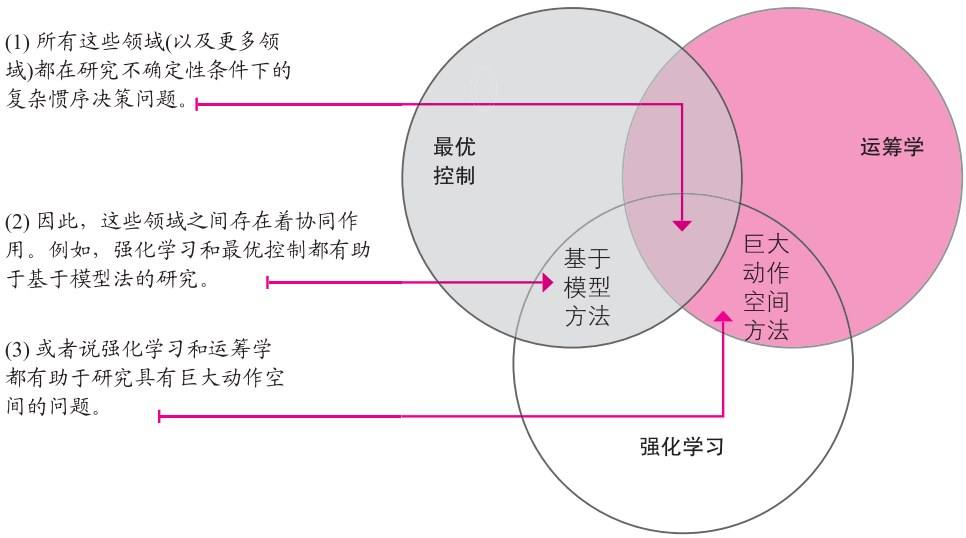


最重要的是 DRL 是解决某种问题的方法。人工智能领域定义这个问题为：创造智能机器。DRL 是解决该问题的方法之一。你会在本书中看到 RL 和其他 ML 方法的对比，但仅会在本章中看到 AI 的定义和历史概述。需要注意的是，RL 领域包含 DRL 领域，因此，尽管我在必要时对二者进行了区分，但当我提到 RL 时，请记住 DRL 也包括在内。

1.1.2 深度强化学习着重创建计算机程序

DRL 的核心在于不确定性条件下复杂的惯序决策问题，但这也是很多领域都感兴趣的话题。例如，控制理论(Control Theory，CT)研究的是控制复杂的已知动态系统的方法。在 CT 中，我们试图控制的系统动态往往是已知的。运筹学(Operations Research，OR)也研究不确定性下的决策问题，但相对于 DRL 中常见的问题，该领域研究的问题往往具有更大的动作空间。心理学研究的是人类行为，在一定程度上，这与“不确定性条件下复杂的惯序决策”问题相同。

相似领域间的协同作用



最重要的是，一个领域会受到其他各个领域的影响。虽然这是好事，但也带来了术语、符号等方面的不一致。我采用计算机科学的方法来解决这一问题，所以这本书的主要内容是构建计算机程序，以解决不确定性条件下的复杂决策问题，同时，本书也会提供代码示例。

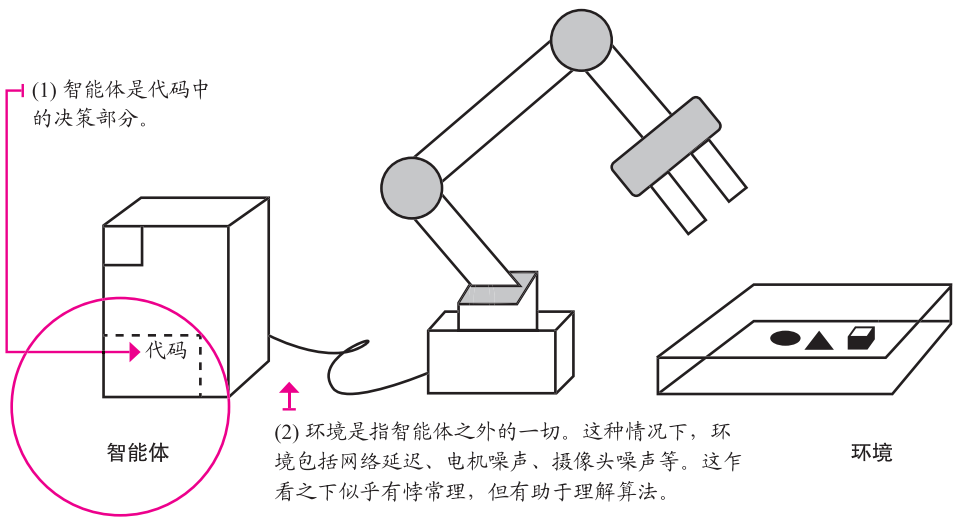
DRL 构建的计算机程序被称为智能体。智能体只做决策，别无其他功能。这意味着，如果你训练机器人捡物，那么机械臂并不是智能体的一部分。只有做出决策的代码才被称为智能体。

1.1.3 智能体解决智能问题

与智能体相对的是环境。环境指的是智能体之外的一切，是智能体无法完全控制的一切。再次，以训练机器人捡起物体为例。要捡起的物体、放置物体的容器、吹过的风以及决策者之外的一切都是环境的一部分。这意味着机械臂也属于环境，因为它并不是智能体。而即使智能体可以控制移动机械臂，实际的机械臂移动却是嘈杂的，因此机械臂也是环境的一部分。

这种智能体和环境之间的严格界线，乍一看有悖于常理；但决策者(即智能体)有且只能有一个任务：做出决定。决策之后的一切都属于“环境”。

智能体与环境的界线

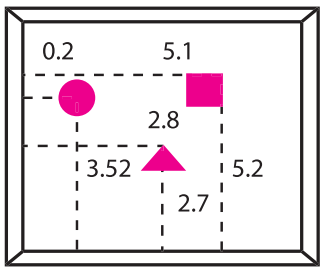


第 2 章对 DRL 的所有组成部分进行了深入调查。以下是第 2 章内容的预览。

环境表示为一组与问题相关的变量。例如，在机械臂例子中，机械臂的位置和速度就是构成环境的变量。这组变量与所有可能的取值被称为状态空间。状态是状态空间的实例，是变量的一组取值。

有趣的是，智能体通常并不能访问环境的实际完整状态。智能体能够观察到的状态被称为观察值。观察值取决于状态，但能为智能体所见。例如，在机械臂案例中，智能体可能只可访问相机图像。虽然存在每个物体的确切位置，但智能体并不能访问这一具体状态。相反，智能体所得的观察值由这些状态衍生。你经常能在文献中看到“观察”和“状态”交替使用，本书亦然。我提前为不一致之处道歉。你只需要知道它们之间的区别，并注意用词，这才是最重要的。

状态与观察

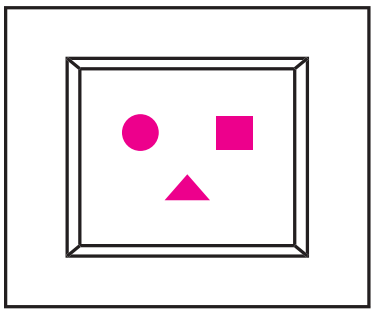


状态：
真实位置

(1) 状态是与当前任务有关的完善、完整的信息。

观察：
只是图像

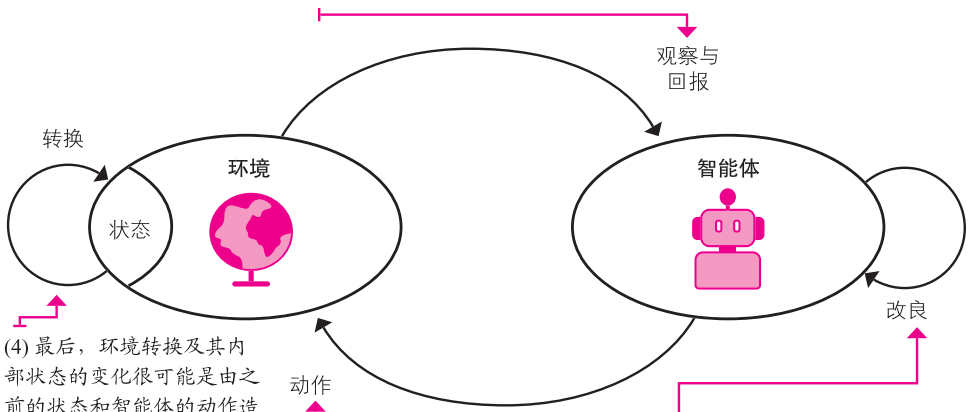
(2) 虽然观察是智能体收到的信息，但这可能是噪声或不完整的信息。



每种状态下，环境都会提供一组动作供智能体选择。智能体就是通过这些动作影响环境的。环境可能会改变状态，回应智能体的动作。进行这种映射的函数称为转换函数。环境也可能发出奖励信号作为对智能体动作的回应。进行这种映射的函数称为回报函数。转换函数和回报函数的集合被称为环境模型。

强化学习循环

(1) 循环从智能体观察环境开始。



(4) 最后，环境转换及其内部状态的变化很可能是由之前的状态和智能体的动作造成的。随后，循环重复。

(3) 随后，它向环境发出一个动作，试图以一种有利的方式控制环境。

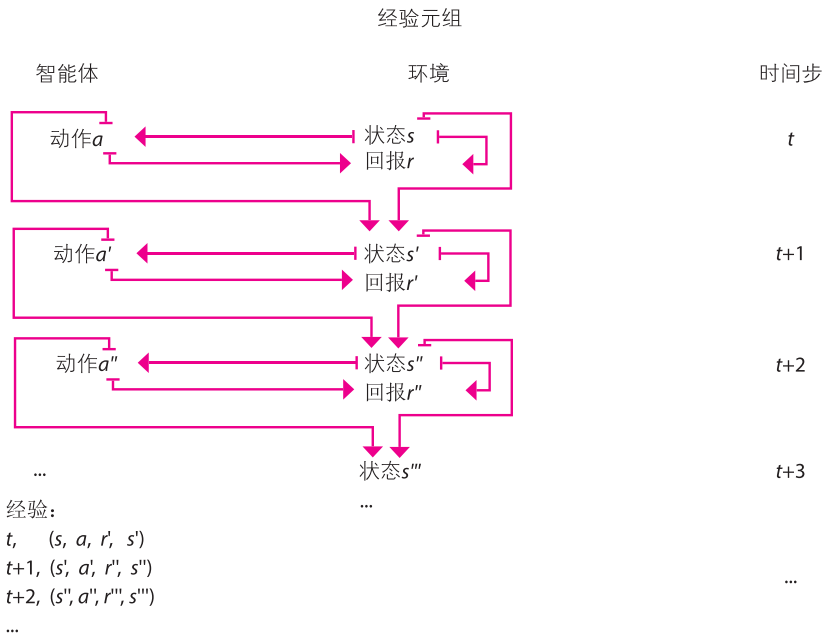
(2) 智能体利用观察和回报来尝试改良任务。

环境的任务通常是确定的，且任务目标由回报函数决定。回报函数的信号可以同时具有惯序性、评估性和抽样性。为了实现任务目标，智能体需要展示出智能，或者至少要展示出常见的与智能有关的认知能力，例如长期思考能力、信息收集能力和概括能力。

智能体的流程有三步：智能体与环境交互、智能体对其动作进行评估、智能体改进其响应速度。智能体可以设计为学习从观察到动作的映射，这种映射称为策略；可以设计为学习映射环境的模型，这种映射称为有模型学习；可以设计为学习评估回报-累计奖励映射，这种映射称为值函数。

1.1.4 智能体通过试错提高性能

智能体和环境之间的交互会持续几个周期。每个周期称为时间步长。每个时间步里，智能体观测环境，采取动作，并得到新的观察和回报。状态、动作、回报和新状态组成的集合称为经验。每一次经验都有机会学习和提高性能。

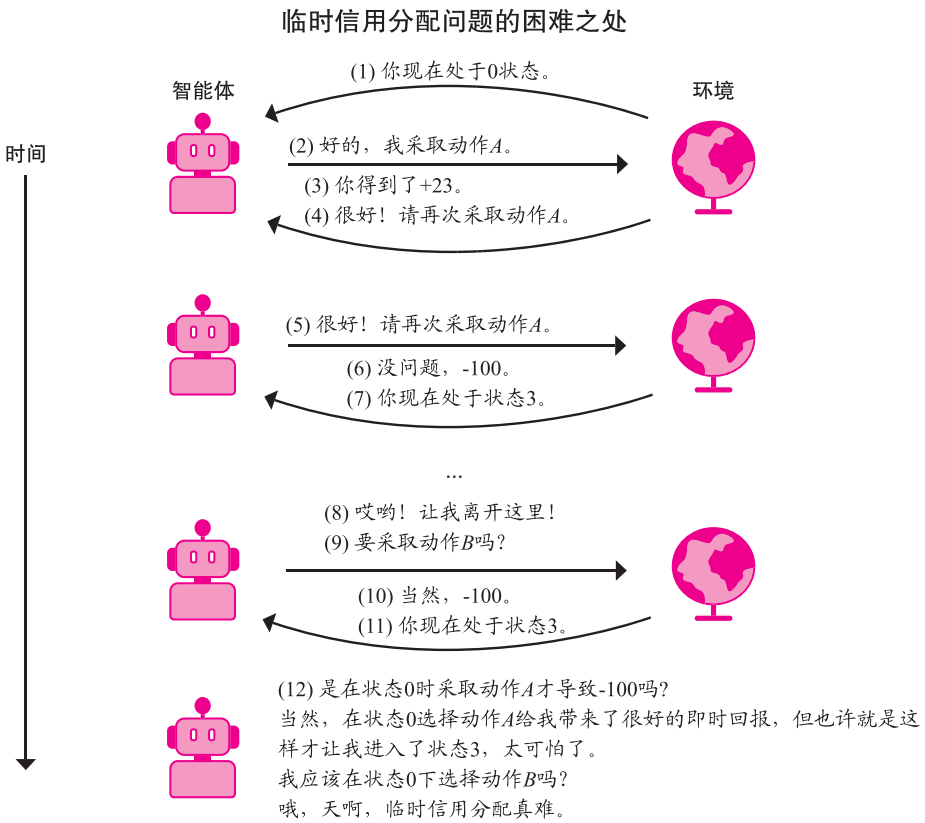


智能体要解决的任务可能自然结束，也可能无法自然结束。自然结束的任务称为偶发任务，例如游戏。相反地，无法自然结束的任务称为持续任务，如学习进步。偶发任务从开始到结束的时间步序列称为迭代。智能体要学习解决一个任务可能需要数个时间步和迭代。智能体通过试错来学习：尝试某事，观察，学习，再尝试其他事情，以此类推。

你将在第 4 章开始学习更多关于周期的内容，其中包括每个场景只有一个时间步的环境。从第 5 章开始，你将学会处理每个场景需要不止一个交互周期的环境。

1.1.5 智能体从惯序性反馈中学习

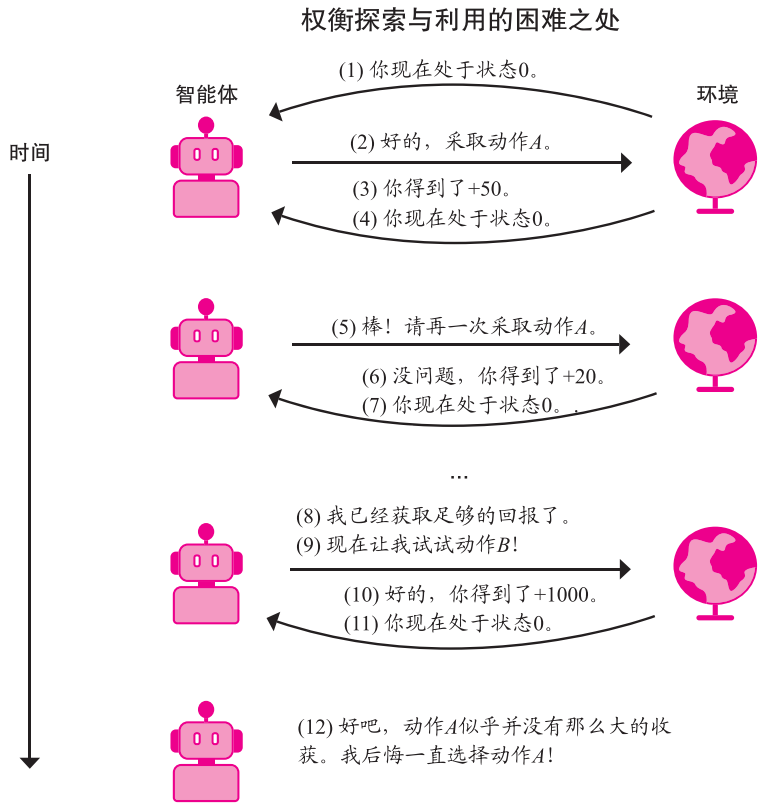
智能体采取的动作可能造成后果延迟。回报可能是稀少的，并且只有在几个时间步之后才表现出来。因此，智能体必须能够从惯序性反馈中学习。惯序性反馈引出了一个问题，被称为临时信用分配问题。临时信用分配问题是一个挑战：它确定哪个状态及/或动作负责回报。当某一问题具有临时成分，并且动作有延迟后果时，将信用分配给回报就变得非常具有挑战性。



在第 3 章中，我们将单独对惯序性反馈进行详细研究。也就是说，你的程序可以从同时具有惯序性、监督性(而非评估性)和详尽性(而非抽样性)的反馈中学习。

1.1.6 智能体从评估性反馈中学习

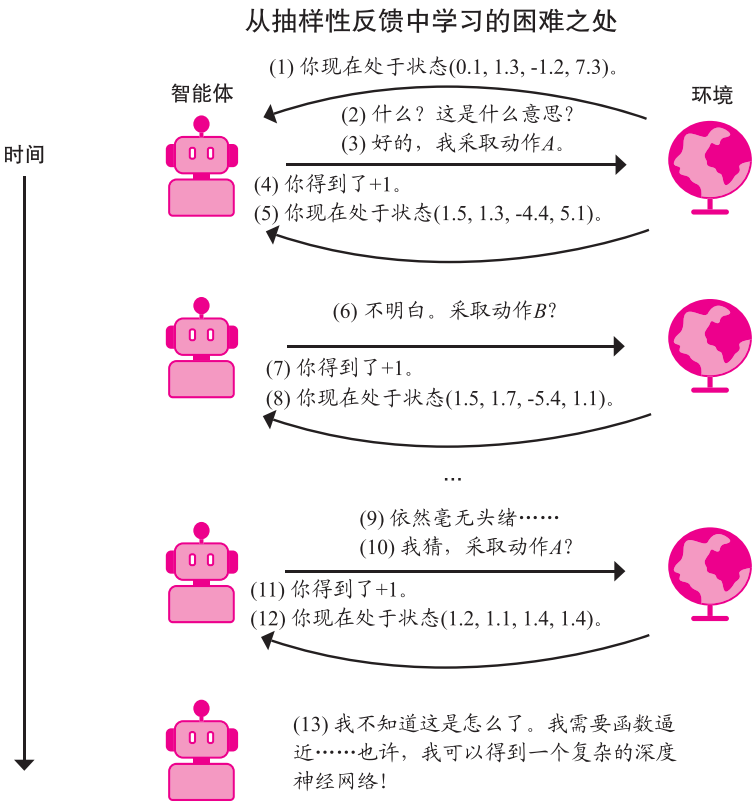
智能体得到的回报可能很微弱，即它可能无法做出监督。回报可能表现出良好性而不是正确性，这意味着它可能不包含其他潜在回报的信息。因此，智能体必须能够从评估性反馈中学习。评估性反馈使得有必要进行探索。智能体必须有能力平衡信息收集和对当前信息的利用，这也被称为权衡探索与利用。



在第 4 章，我们将单独对评估性反馈进行详细研究。也就是说，你的程序将从同时具有一次性(而非惯序性)、评估性和详尽性(而非抽样性)的反馈中学习。

1.1.7 智能体从抽样性反馈中学习

智能体收到的回报只是一个样本，且智能体无法获得回报函数。此外，由于状态和动作空间通常很大，甚至是无限大，尝试从稀疏和微弱的抽样性反馈中学习就变得更困难。因此，智能体必须能从抽样性反馈中学习，并且必须能对其进行泛化。

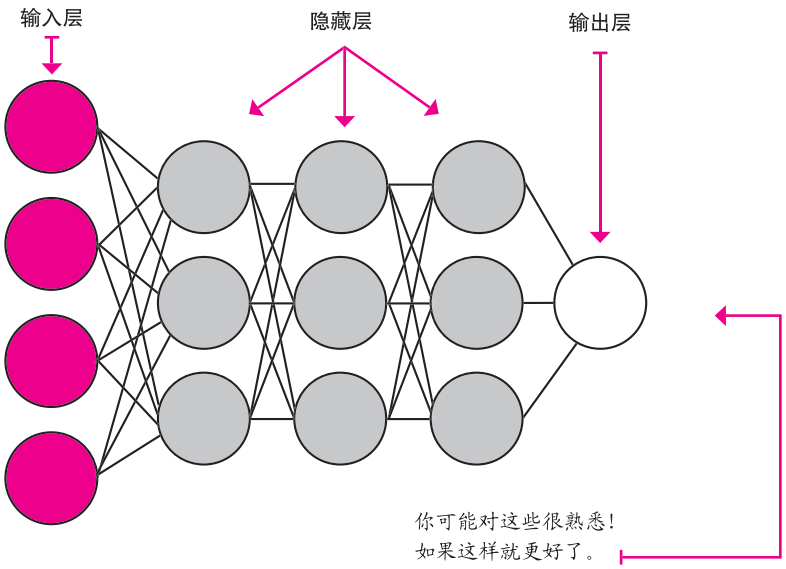


用于逼近策略的智能体称为基于策略学习；用于逼近值函数的智能体称为基于价值学习；用于逼近模型的智能体称为有模型智能体；用于同时逼近策略和值函数的智能体称为 actor-critic。智能体可以接近这些类型中的一个或多个。

1.1.8 智能体使用强大的非线性函数逼近

智能体可以使用从决策树到 SVM，再到神经网络等多种 ML 方法和技术来逼近函数。但本书只使用神经网络，毕竟神经网络就是深度强化学习的“深度”所指。神经网络不一定是解决所有问题的最佳方案，需要谨记神经网络需要大量的数据，且很难解释。然而，神经网络是目前最有效的函数逼近法之一，其性能往往是最好的。

简单前馈神经网络



人工神经网络(Artificial Neural Networks，ANN)是一种多层非线性函数逼近器，受动物大脑中的生物神经网络启发而来。ANN 并不是一种算法，而是由多层数学转换组成的结构，其中数学转换应用于输入值。

第 3 ～ 7 章只研究智能体从详尽性反馈(相对于抽样性反馈)中学习。第 8 章开始，研究完整的 DRL 问题，即使用深度神经网络使得智能体可以从抽样性反馈中学习。需要牢记，DRL 智能体从同时具有惯序性、评估性和抽样性的反馈中学习。

1.2 深度强化学习的过去、现在与未来

历史并不是获取技能的必要条件，但它可以让你了解一个主题的背景，进而帮助你获得动力，从而获取技能。AI和DRL的历史应该能够帮助你明确对这项强大技术未来发展的期望。有时，我觉得关于AI的炒作其实是有成效的，人们变得对 AI 感兴趣。但在此之后，当投入工作时，炒作已然无济于事，这是个问题。虽然我想对 AI 保持兴奋，但还是需要设置切合实际的期望。

1.2.1 人工智能和深度强化学习的发展简史

DRL 的起源可以追溯到很多年前，因为自古以来，人类就对是否存在人类以外的智能生物感到好奇。艾伦 • 图灵在 20 世纪 30—50 年代的工作可能是一个良好开端，这些工作为后来的科学家们提供了关键的理论基础，为现代计算机科学和 AI 铺平了