DOI: 10.7641/CTA.2016.60173

深度强化学习综述: 兼论计算机围棋的发展

赵冬斌[1†], 邵 坤[1], 朱圆恒[1], 李 栋[1], 陈亚冉[1], 王海涛[1]

(1. 中国科学院自动化研究所 复杂系统管理与控制国家重点实验室, 北京 100190)

刘德荣[2], 周 彤[3], 王成红[4]

(2. 北京科技大学 自动化学院, 北京 100083; 3. 清华大学 自动化系, 北京 100084;

4. 国家自然科学基金委 信息科学部, 北京 100085)

摘要: 深度强化学习将深度学习的感知能力和强化学习的决策能力相结合, 可以直接根据输入的图像进行控制, 是一种更接近人类思维方式的人工智能方法. 自提出以来, 深度强化学习在理论和应用方面均取得了显著的成果. 尤其是谷歌深智(DeepMind)团队基于深度强化学习方法研发的计算机围棋“初弈号–AlphaGo”, 在2016年3月以4:1的大比分战胜了世界围棋顶级选手李世石(Lee Sedol), 成为人工智能历史上一个新里程碑. 为此, 本文综述深度强化学习的发展历程, 兼论计算机围棋的历史, 分析算法特性, 探讨未来的发展趋势和应用前景, 期望能为控制理论与应用新方向的发展提供有价值的参考.

关键词: 深度强化学习; 初弈号; 深度学习; 强化学习; 人工智能

中图分类号: TP273 **文献标识码**: A

Review of deep reinforcement learning and discussions on the development of computer Go

ZHAO Dong-bin[†], SHAO Kun, ZHU Yuan-heng, LI Dong, CHEN Ya-ran, WANG Hai-tao

(1. The State Key Laboratory of Managentment and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijng 100190, China)

LIU De-rong[2], ZHOU Tong[3], WANG Cheng-hong[4]

(2. College of Automation, University of Science and Technology Beijing, Beijing 100083, China;
3. Department of Automation, Tsinghua University, Beijing 100084, China;
4. Department of Information Sciences, National Natural Science Foundation of China, Beijing 100085, China)

Abstract: Deep reinforcement learning which incorporates both the advantages of the perception of deep learning and the decision making of reinforcement learning is able to output control signal directly based on input images. This mechanism makes the artificial intelligence much close to human thinking modes. Deep reinforcement learning has achieved remarkable success in terms of theory and application since it is proposed. ‘Chuyihao–AlphaGo’, a computer Go developed by Google DeepMind, based on deep reinforcement learning, beat the world’s top Go player Lee Sedol 4:1 in March 2016. This becomes a new milestone in artificial intelligence history. This paper surveys the development course of deep reinforcement learning, reviews the history of computer Go concurrently, analyzes the algorithms features, and discusses the research directions and application areas, in order to provide a valuable reference to the development of control theory and applications in a new direction.

Key words: deep reinforcement learning; AlphaGo; deep learning; reinforcement learning; artificial intelligence

1 引言(Introduction)

谷歌公司的人工智能研究团队–深智(DeepMind), 近两年公布了两项令人瞩目的研究成果: 基于Atari视频游戏的深度强化学习算法[1]和计算机围棋初弈号[1][2]. 这些工作打破了传统学术界设计类人智能学习算法的桎梏, 将具有感知能力的深度学习(deep

收稿日期: 2016−03−29; 录用日期: 2016−06−21.

[†]通信作者. E-mail: dongbin.zhao@ia.ac.cn; Tel.: +86 10-82544764.

本文责任编委: 苏剑波.

国家自然科学基金项目(61273136, 61573353, 61533017).

Supported by National Natural Science Foundation of China (61273136, 61573353, 61533017).

[1]初弈号: 谷歌深智团队研发的计算机围棋程序, 国内有很多译名版本, 如“阿尔法围棋”、“阿尔法狗”、或昵称为“狗狗”、“阿发哥”等等. 本文翻译为“初弈号”, 取其“初级、围棋、机器”三大特征, 保留英文原文的朴素感, 也有充满自信、奋发图强之意.

learning, DL)和具有决策能力的强化学习(reinforcement learning, RL) 紧密结合在一起, 构成深度强化学习(deep reinforcement learning, DRL)算法. 其原理框架如图1所示. 这些算法的卓越性能远超出人们的想象, 极大地震撼了学术界和社会各界.

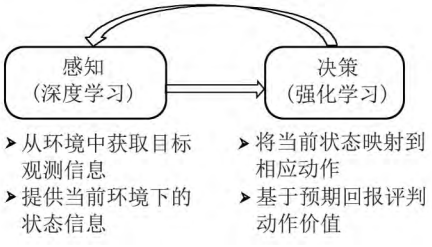

图 1 深度强化学习的框架[3]

Fig. 1 The framework of deep reinforcement learning[3]

深智团队在《Nature》杂志上的两篇文章使深度强化学习成为高级人工智能的热点. 2015年1月的文章[1]提出深度Q网络(deep Q-network, DQN), 在Atari视频游戏上取得了突破性的成果. 深度Q网络模拟人类玩家进行游戏的过程, 直接将游戏画面作为信息输入, 游戏得分作为学习的强化信号. 研究人员对训练收敛后的算法进行测试, 发现其在49个视频游戏中的得分均超过人类的高级玩家. 在此基础上, 深智团队在2016年1月的文章[2] 进一步提出计算机围棋初弈号. 该算法将深度强化学习方法和蒙特卡罗树搜索(Monte Carlo tree search, MCTS)结合, 极大减少了搜索过程的计算量, 提升了对棋局估计的准确度. 初弈号在与欧洲围棋冠军樊麾的对弈中, 取得了5:0完胜的结果. 2016年3月, 与当今世界顶级棋手职业九段李世石(Lee Sedol)进行了举世瞩目的对弈, 最终以4:1获得胜利. 这也标志着深度强化学习作为一种全新的机器学习算法, 已经能够在复杂的棋类博弈游戏中达到匹敌人类的水平.

因此, 深入研究深度强化学习方法, 对于推动人工智能方法的发展, 及其在各个领域中的应用都有非常重要的意义. 本文将从深度强化学习技术和计算机围棋的发展历程两方面展开综述. 主要结构如下: 首先介绍了深度强化学习中的关键技术; 强化学习和深度学习; 然后对深度强化学习发展历程和主要方法进行介绍; 紧接着重点介绍了计算机围棋的历史与现状, 初弈号的原理及其优缺点; 随后分析了深度强化学习的研究趋势和应用前景; 最后作出总结.

2 强化学习(Reinforcement learning)

强化学习是受到生物能够有效适应环境的启发, 以试错的机制与环境进行交互, 通过最大化累积奖赏的方式来学习到最优策略.

强化学习系统由4个基本部分组成: 状态s, 动作a, 状态转移概率$P^a_{s,s'}$和奖赏信号r. 策略$\pi: S \to A$被定义为从状态空间到动作空间的映射. 智能体在当前状态s下根据策略π来选择动作a, 执行该动作并以概率$P^a_{s,s'}$转移到下一状态s', 同时接收到环境反馈回来的奖赏r. 强化学习的目标是通过调整策略来最大化累积奖赏. 通常使用值函数估计某个策略π的优劣程度. 假设初始状态$s_0 = s$, 则关于策略π的状态值函数可定义为

$$V^\pi(s) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_t = \pi(s_t), \quad (1)$$

其中$\gamma \in (0,1)$是衰减因子. 由于最优策略是最大化值函数的策略, 因此可根据下式求得最优策略,

$$\pi^* = \arg\max_\pi V^\pi(s). \quad (2)$$

另一种形式的值函数是状态动作值函数, 定义为

$$Q^\pi(s_t, a_t) = r(s_t, a_t) + \gamma V^\pi(s_{t+1}), \quad (3)$$

此时最优策略可根据下式得到

$$\pi^* = \arg\max_{a \in A} Q^\pi(s, a). \quad (4)$$

强化学习的研究有着悠久的历史. 1992年, Tesauro等成功使用强化学习使西洋双陆棋达到了大师级的水准[4]; Sutton等撰写了第1本系统性介绍强化学习的书籍[5]; Kearns等第1次证明了强化学习问题可以用少量的经验得到近似最优解[6]; 2006年Kocsis等提出的置信上限树算法革命性地推动了强化学习在围棋游戏上的应用, 这可以说是初弈号的鼻祖[7]; 2015年, Littman在《Nature》上对强化学习做了综述[8]. 表1总结了强化学习发展历程中的重要事件. 目前常用的强化学习方法包括蒙特卡罗、Q学习、SARSA学习、TD学习、策略梯度和自适应动态规划等.

表 1 强化学习研究历程

Table 1 Timeline of reinforcement learning research events

1956	Bellman提出了动态规划方法[9]
1977	Werbos提出自适应动态规划方法[10]
1988	Sutton提出了TD算法[5]
1992	Watkins提出了Q学习算法[11]
1994	Rummery等提出了SARSA学习算法[12]
1996	Bertsekas等提出了解决随机过程优化控制的神经动态规划方法[13]
1999	Thrun提出了部分可观测马尔科夫决策过程中的蒙特卡罗方法[14]
2006	Kocsis等提出了置信上限树算法[7]
2009	Lewis等提出了反馈控制自适应动态规划算法[15]
2014	Silver等提出确定性策略梯度算法[16]

2.1 蒙特卡罗方法(Monte Carlo method)

蒙特卡罗方法[14, 17]是一种以概率统计理论为指导的强化学习方法. 它在强化学习中的应用可以追溯到1968年, Michie等用蒙特卡罗方法预测动作值函数[18]. 此后, Barto等讨论了蒙特卡罗方法在策略评估

中的使用, 并用其求解线性方程系统[19]. 通过与环境交互, 从所采集的样本中学习, 获得关于决策过程的状态、动作和奖赏的大量数据(经验), 最后计算出累积奖赏的平均值. 采样越多、累积奖赏的平均值越接近真实的值函数. 因此, 该方法的计算量非常大. 蒙特卡罗作为一种无模型的方法, 它不需要事先知道马尔科夫决策过程(Markov decision process, MDP)的状态转移概率以及奖赏. 蒙特卡罗方法同时还可以与离策略(off-policy)的思想相结合, 得到离策略的蒙特卡罗学习, 能够在执行一个策略的时候评估另一个策略的好坏.

蒙特卡罗方法可以与树搜索结合而形成蒙特卡罗树搜索[20–21], 它是一种用于决策过程的启发式搜索算法. 2006年, Coulom首次提出了蒙特卡罗树搜索这一术语[22]. 2016年, 采用蒙特卡罗树搜索的计算机围棋初弈号击败了人类围棋欧洲冠军[2]. 蒙特卡罗树搜索算法主要包含选择、扩展、模拟和反向传播(back-propagation)4个步骤.

2.2 Q学习, SARSA学习和TD学习(Q learning, SARSA learning and TD learning)

Q学习是最早的在线强化学习算法, 同时也是强化学习最重要的算法之一. 1989年Watkins在其博士论文中提出了Q学习算法[11]. 该算法的主要思路是定义了Q函数(性能函数), 将在线观测到的数据代入到下面的更新公式中对Q函数进行迭代学习, 得到精确解

$$Q_{t+1}(s_t,a_t)=Q_t(s_t,a_t)+\alpha_t\delta_t, \tag{5a}$$

$$\delta_t=r_{t+1}+\gamma\max_{a'}Q_t(s_{t+1},a')-Q_t(s_t,a_t), \tag{5b}$$

其中: t是当前时刻, α_t是学习率, δ_t表示时间差分(temporal difference, TD)误差, a'是状态s_{t+1}能够执行的动作.

Q学习是一种离策略的学习算法. 使用一个合理的策略来产生动作, 根据该动作与环境交互所得到的下一个状态以及奖赏来学习得到另一个最优的Q函数. Q学习在最优控制[23–24]和游戏[1]上有许多的应用. 可以证明, 当满足如下两个条件时, Q学习可以在时间趋于无穷时得到最优控制策略[25–27].

1) $\sum\limits_{t=0}^{\infty}\alpha_t^2<\infty,\sum\limits_{t=0}^{\infty}\alpha_t=\infty$;

2) 所有的状态动作都能够被无限次地遍历.

另一种与Q学习类似的算法是由Rummery和Niranjan提出的SARSA学习[12]. 与Q学习不同的是, SARSA学习是一种在策略(on-policy)学习算法, 它直接使用在线动作来更新Q函数, 其时间差分误差定义为

$$\delta_t=r_{t+1}+\gamma Q_t(s_{t+1},a_{t+1})-Q_t(s_t,a_t). \tag{6}$$

可以看出, SARSA学习更新Q函数时需要用到(s,a,r,s',a')这5部分, 它们构成了该算法的名字SARSA. 在一定条件下, SARSA学习可以在时间趋于无穷时得到最优控制策略[28].

Q学习和SARSA学习都是借助时间差分误差来更新值函数, 它们可以被称为时间差分学习[29]. 这涉及到时间信度分配(temporal credit assignment)问题, 即对于不同时刻的动作, 应该为其分配多少时间差分误差来更新值函数. 为此Sutton提出TD(λ)算法, 解决时间信度分配问题[29].

$$Q_{t+1}(s_t,a_t)=Q_t(s_t,a_t)+\alpha\delta_t e_t(s_t), \tag{7}$$

其中$e_t(s)$代表资格迹(eligibility trace), 定义为

$$e_t(s)=\begin{cases}\gamma\lambda e_t(s)+1, & s=s_t,\\ \gamma\lambda e_t(s), & \text{其他}.\end{cases} \tag{8}$$

TD(λ)建立了蒙特卡罗和时间差分学习的统一框架. 当λ在0到1之间取不同值时, TD(λ)可以转换为不同的方法. $\lambda=1$时TD(λ) 变成了蒙特卡罗, $\lambda=0$ 时TD(λ)变成了时间差分学习.

2.3 策略梯度学习(Policy gradient learning)

上一部分介绍的方法都是基于值函数的方法, 它需要求出值函数, 再根据值函数来选择动作. 另一种是基于策略的方法, 如策略梯度算法[30–31]. 策略梯度是一种直接逼近策略, 优化策略, 最终得到最优策略的方法. 值函数法相比于策略梯度法有两个局限性[32]: 第一, 由值函数法最终得到的是一个确定性的策略, 而最优策略可能是随机的, 此时值函数法不适用; 第二, 值函数的一个小小的变动往往会导致一个原本被选择的动作反而不能被选择, 这种变化会影响算法的收敛性. 策略梯度法又可以分为确定策略梯度算法和随机策略梯度算法. 近些年确定策略梯度算法逐渐受到了人们的关注. 在确定策略梯度算法中, 动作以概率1被执行. 在随机策略梯度算法中, 动作以某一概率被执行. Silver等提出了一种有效的确定策略梯度估计方法[16]. 与随机策略梯度算法相比, 确定策略梯度在高维动作空间上拥有更好的表现. 假设需要逼近的策略是$\pi(s,a;\theta)$, 而且该策略对参数θ可导, 则可定义目标函数和值函数

$$J(\pi_\theta)=\mathrm{E}\sum_{t=1}^{\infty}\gamma^{t-1}r_t|s_0,\pi_\theta, \tag{9}$$

$$Q^{\pi_\theta}(s,a)=\mathrm{E}\sum_{k=1}^{\infty}\gamma^{k-1}r_{t+k}|s_t=s,a_t=a,\pi_\theta. \tag{10}$$

假设从初始状态s_0开始, 依据策略π_θ来选取动作的状态分布为$d^{\pi_\theta}(s)=\sum\limits_{t=1}^{\infty}\gamma^tP(s_t=s|s_0,\pi_\theta)$, 那么可以得到下面的策略梯度定理[32]: 对于任意的马尔科夫决策过程, 均有

$$\nabla_\theta J(\pi_\theta)=\sum_s d^{\pi_\theta}(s)\sum_a\nabla_\theta\pi_\theta(s,a)Q^{\pi_\theta}(s,a). \tag{11}$$

从上式可以看出, 虽然状态分布$d(s)$与策略π_θ有关, 可是$\bigtriangledown_\theta J(\pi_\theta)$却和$\bigtriangledown_\theta d^{\pi_\theta}(s)$无关. 因此策略的变化会使得样本分布发生变化, 但样本分布的变化不会对策略的更新产生影响. 得到策略梯度之后, 便可采用梯度上升等方法来最大化目标函数.

2.4 自适应动态规划(Adaptive dynamic programming)

自适应动态规划(adaptive dynamic programming, ADP) 由 Werbos[10]于 20 世纪 70 年代提出, 在Bertsekas[13]、Lewis[15]、Liu[33]、Zhang[34]等学者的努力下日益发展成熟. 自适应动态规划是一种针对连续状态空间的最优控制方法. 对于比较复杂的问题, 它们的状态空间和动作空间往往是连续的, 规模较大. 由于维度爆炸的缘故, 不能采用传统的查表法来得到性能函数, 此时需要使用函数逼近器, 例如线性函数逼近器和神经网络逼近器等工具来逼近性能函数. 自适应动态规划通过构建动作网络(actor)和评价网络(critic)两个网络来处理复杂的强化学习问题. 在状态$x(t)$时, 动作网络用于选择动作$u(t)$, 评价网络输出值函数对动作网络的动作进行评价, 并对动作网络进行调整, 最终通过被控系统(plant)输出下一时刻状态$x(t+1)$. 其结构如图2所示.

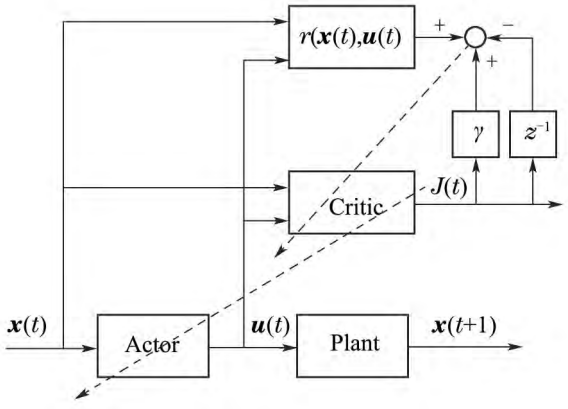

图 2 自适应动态规划结构图

Fig. 2 The flowchart of adaptive dynamic programming

近年来, 自适应动态规划在基础理论和工业领域应用等方面取得了很多研究成果. 离散时间非线性系统[33, 35]和连续状态系统[36–37]的最优控制是自适应动态规划的强项. Jiang等将鲁棒控制和自适应动态规划结合, 将自适应动态规划的应用扩展到了不确定非线性系统[38]. Wu等提出了在线同步策略更新算法, 通过借助神经网络, 成功求解了非线性H_∞控制问题中的HJI方程[39]. Zhao等结合自适应动态规划和经验回放技术解决了模型未知的非零和博弈问题, 并证明了系统稳定性[40]. 在工业领域的应用有机器人[41–42]、游戏[43]和汽车自适应巡航控制[44]等. Wu等对自适应动态规划在机器人领域的应用做了综述[41]. Zhao等成功将自适应动态规划应用到五子棋游戏中, 并达到了和商用五子棋算法相媲美的境地[43].

2.5 强化学习研究的展望(Prospect of reinforcement learning research)

理论研究方面, 强化学习在过去几年取得了丰富的成果. 上文所提算法在面临各种问题时的收敛性和最优性, 研究者们都进行了详细的探索[25, 32, 45–47]. 在使用逼近器解决大规模MDPs或连续状态空间问题时, Sutton和Bhatnagar分别对线性和非线性逼近器的训练结果进行分析, 设计出了使结果准确的训练方法[48–49]. Melo对使用逼近器的收敛性进行了证明[50]. Kearns等利用概率近似正确理论对在线算法的学习时间进行定性的分析, 同样取得了丰富的成果[6, 51]. 随着技术的发展, 神经网络的规模越来越大, 参数越来越多, 对强化学习算法的收敛性提出了更高的要求. 如何保证复杂网络参数获得收敛的训练结果是当前研究者面临的挑战, 同时人们对资源、经济效益的要求越来越高, 对算法结果的最优性也提出了更高的要求. 找到更接近最优结果的算法成为了学术界和工业界共同追求的目标.

强化学习的目标是得到最大的累积奖赏. 为了实现这一目标, 一方面需要“利用(exploitation)”, 即利用已经学习到的经验来选择奖赏最高的动作, 使系统向好的状态转移. 另一方面, 需要“探索(exploration)”, 即通过充分地掌握环境信息, 发现能得到更高奖赏的状态, 避免陷入局部最优. 这就使得强化学习方法陷入了一个“探索–利用”困境: 只有既充分地探索环境, 又利用已学到的知识才能最大化累积奖赏. 因此, 探索–利用的平衡成为了强化学习研究者们一直密切关注的热点[5, 52]. Bernstein等提出的自适应分辨率强化学习方法能够在状态连续的确定性系统中高效探索未知状态[53]. Zhao等提出的连续系统概率近似正确(probably approximately correct)方法能够在有限时间范围内找到最优或近似最优策略[36].

强化学习的学习机制表明它是不断地与环境交互, 以试错的方式学习得到最优策略, 是一种在线学习方法[8]. 然而, 现实中有很多问题需要在离线的评估后再给出决策. 目前的一个研究趋势是用离线估计来处理上下文赌机(contextual bandit) 问题. 例如, 微软研究院的Li等将无偏离线估计的上下文赌机方法成功应用到了推荐系统中[54]. 目前已经有人建议创建强化学习的离线数据库[55], 可是现有的离线估计算法还不够成熟, 有待进一步发展.

强化学习的另一个局限在于合适的奖赏信号定义. 强化学习通过最大化累积奖赏来选择最优策略, 奖赏很大程度上决定了策略的优劣. 现阶段奖赏是由研究人员凭借领域知识定义, 一个不合理的奖赏势必会很大程度上影响最终的最优策略. 有学者已经开始尝试

借助人类的导师信号来改进原有的强化学习算法, 使机器人能够更好地学习到期望动作[56–57]. 因此, 对于奖赏信号的研究将会是强化学习未来发展的一个潜在热点[8].

强化学习和认知科学的交叉研究是另一个研究趋势. 在最近一次的强化学习研究热潮中, 人们发现了多巴胺神经元激活的时间差分机制, 这为计算机科学和认知神经科学建立起了一座桥梁[58]. 此外, 有模型和无模型的强化学习方法在动物决策问题研究上得到了相应的解释[59]. 2013年, 第1届强化学习与决策多学科会议在普林斯顿大学召开, 来自计算机科学、心理学、动物学、神经学等多个领域的研究人员共同探讨了关于动物学习与认知的基本框架. 因此, 以强化学习为切入点的认知模型研究将成为未来的一个研究趋势.

3　深度学习(Deep learning)

深度学习起源于人工神经网络. 20世纪90年代研究人员通过模拟大脑皮层推断分析数据的复杂层状网络结构, 提出了多层感知机的概念, 并且提出优化多层神经网络的反向传播算法, 但是由于受到梯度弥散问题的困扰和硬件资源的限制, 神经网络的研究一直没有取得突破性进展. 2006年, Hinton提出通过自动提取原始数据的层级特征表示来建立输入数据与输出数据之间复杂的函数映射关系. 在文献[60]中, Hinton指出训练深层神经网络的一个基本原则, 即采用非监督方法先对神经网络中间层逐层进行贪婪的预训练, 之后再采用监督方法对整个网络进行精调. 预训练的方法为深度神经网络提供了较好的初始参数, 降低了深度神经网络的优化难度. 深度学习前几年的发展集中在预训练, 提出了多种方法. 到了2010年之后, 随着计算资源和预训练技术的发展, 深度学习在人工智能领域取得了重大突破, 包括语音识别、视觉对象识别及检测等领域[61–63]. 2012年, 微软研究人员建立深度神经网络–隐马尔科夫混合模型并首次成功应用于大词汇量的语音识别系统, 相比于传统的高斯–隐马尔科夫模型, 语音识别错误率相对降低了30%左右[64]. Krizhevsky等首次在大规模数据集ImageNet上应用深度卷积神经网络(convolutional neural network, CNN), 将图像的识别的错误率降低到37.5%, 远远优于之前的方法[65]. 吴恩达(Andrew Ng)负责的“Google Brain”项目采取无监督的学习方式, 从YouTube的视频中学习到了高度抽象的概念, 例如“Google Cat”[66]. 2013年Graves证明, 结合了长短时记忆(long short terms memory, LSTM)的递归神经网络(recurrent neural network, RNN)比传统的递归神经网络在语音处理方面更有效[67]. 2014年至今, 深度学习在很多领域都取得了突破性进展, 发展出了包括注意力(attention)[68], RNN–CNN[69], 以及深度残差网络[70]等多种模型. 2015年LeCun等在《Nature》上面发表了关于深度学习的综述, 总结了深度学习的基本原理和主要优势[71]. 表2总结了深度学习发展历程中的重要事件.

表 2　深度学习研究历程

Table 2　Timeline of deep learning research events

2006	Hinton在DBN中提出了一种逐层预训练方法, 解决了梯度弥散问题[60]
2008	Vincent等提出了降噪自编码器[72]
2011	Rafir等提出了收缩自编码器[73]
2012	微软研究员建立深度神经网络–隐马尔科夫混合模型, 在语音识别领域取得突破[64]
2012	Krizhevsky 等提出应用于ImageNet的AlexNet, 在图像分类领域取得突破[65]
2012	Ng在“GoogleBrain”项目中使用无监督深度学习方法[66]
2015	Xu提出了结合注意力的场景识别[68]
2015	微软研究员He等人提出了拥有152层的深度残差网络[70]

目前典型的深度学习模型包括: 卷积神经网络、深度置信网络(deep belief network, DBN)、堆栈自编码网络(stacked auto-encoder, SAE)和递归神经网络等.

3.1　卷积神经网络(Convolutional neural network)

卷积神经网络由卷积层和下采样层交替层叠而成. 卷积层采用权重共享, 使得网络的参数减少; 下采样层由于采用最大值或均值下采样的方式, 使得图像维度降低, 并且通过卷积和下采样学习到的特征具有平移、旋转不变性[74]. 如图3所示.

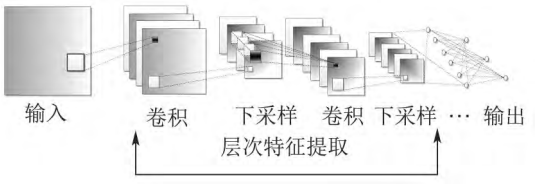

图 3　卷积神经网络结构[74]

Fig. 3　The structure of convolutional neural network[74]

在前向计算中, 图像信息从输入层经过几层卷积和下采样的变换后提取特征, 被传送到全连接层, 得到网络的输出. 向后传播阶段, 卷积神经网络采用误差反向传播算法, 将输出误差反向传递到每一层, 同时利用梯度下降法对每层的参数求导优化. 卷积神经网络架构非常适合处理图像数据, 随着网络层数的增加, 卷积神经网络能够从原始数据中抽取更抽象的特征, 更加有利于图像的识别. 牛津大学研究团队提出的VGG(visual geometry group)网络是目前应用在大规模图像识别任务中的典型深度卷积神经网络, 它的

网络结构多达20层[75]. 2015年, 微软进一步扩张了卷积神经网络的层数, 提出了一种具有152层的深度残差网络, 在ImageNet 上面取得了历史最好的成绩[70].

3.2 深度置信网络(Deep belief network)

深度置信网络由多层受限玻尔兹曼机(restricted Boltzmann machine, RBM)组成, 如图4所示. 受限玻尔兹曼机可看作是一种基于能量的模型(energy-based model, EBM), 通过学习数据的概率密度分布提取抽象特征. 深度置信网络的训练过程包括预训练和微调两部分. 预训练阶段使用非监督贪婪逐层训练的方法获得权值. 首先利用持续对比散度(Persistent contrastive divergence)算法[76], 或快速权重持续对比散度(Fast-weight persistent contrastive divergence)算法[77]充分训练最底层受限玻尔兹曼机; 接下来固定其权值, 把得到的隐层输出作为第2 个受限玻尔兹曼机的输入, 训练第2个受限玻尔兹曼机; 依次逐层训练所有的受限玻尔兹曼机, 即可获得深度置信网络的初始权值. 微调阶段指的是深度置信网络利用带标签数据用反向传播算法对网络的参数进行调整. 深度置信网络可用于识别特征、分类和生成数据. 2012年, Mohamed 采用深度置信网络对深度网络进行预训练, 并应用于语音识别的声学模型中, 将音素识别错误率降低到了20.7%[78].

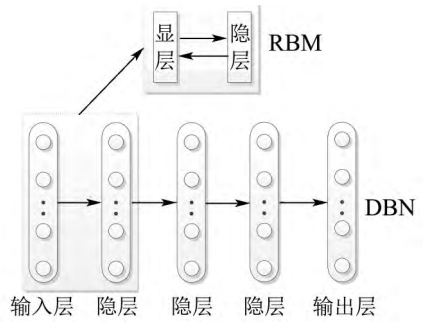

图 4 深度置信网络结构

Fig. 4 The structure of deep belief network

3.3 堆栈自编码网络(Stacked auto-encoder)

堆栈自编码网络的结构由若干个自编码器(Auto-encoder, AE)组成, 如图5所示.

自编码器包括编码层和解码层, 目的是复现输入信号. 堆栈自编码网络通过无监督的训练方法得到每层自编码器的参数权重, 广泛应用于深度神经网络的预训练, 显著地提高了深度神经网络的性能, 在图像和语音领域都取得了巨大的成功[79–80]. 常用的自编码器包括降噪自编码器(denoising auto-encoders, DAEs)[72]和收缩自编码器(contractive auto-encoders, CAEs)[73]等. 降噪自编码器通过向原始输入数据引入随机噪音提高自编码器表示学习的能力. 收缩自编码器在损失函数中加入了惩罚项, 与其他方法相比, 该方法增强了特征的鲁棒性.

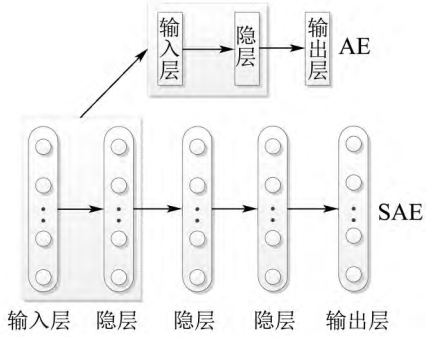

图 5 堆栈自编码网络结构

Fig. 5 The structure of stacked auto-encoder

3.4 递归神经网络(Recurrent neural network)

以上几种模型中, 信息从输入层到隐层, 最后到输出层, 层与层之间是全连接的, 但是每层的节点是无连接的, 这样的连接形式没有考虑数据之间的关联性. 递归神经网络则不同, 它会对前面的信息进行记忆, 并应用于当前层计算输出, 即隐层之间的节点有连接[81], 如图6所示. Schuster考虑前后两个数据的影响, 提出了一种双向递归神经网络, 此结构把前后两个隐层的输出输入到单个网络的隐层[82].

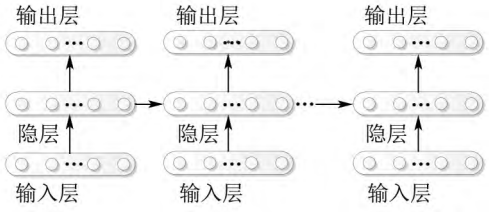

图 6 递归神经网络结构

Fig. 6 The structure of recurrent neural network

递归神经网络一次处理一个输入序列元素, 每个节点同时包含过去时刻序列元素的历史信息. 递归神经网络是非常强大的动态系统, 它一般采用时间反传(backpropagation through time, BPTT)训练算法来解决非长时依赖问题. 但如果递归神经网络的输入序列太长, 则会导致反向传播求导的过程中梯度激增或降为零. 后来人们发现长短时记忆单元可有效解决此问题[71]. 递归神经网络主要用于处理时序数据, 常应用在预测文本和语音识别中, 并取得了不错的效果. 蒙特利尔大学的学者提出了基于循环神经网络的系统RNNsearch[83], 在2014机器翻译大会表现出色, 在某些方面甚至超过基于短语的翻译系统Moses[84].

3.5 深度学习研究的展望(Prospect of deep learning research)

在2015年谷歌推出的图像识别技术的测试中, 拥有多达30层的神经网络错误地将一张满是噪点的图像识别成了香蕉. 早在2014年, 谷歌的3位研究人员就发表了论文来说明神经网络很“好骗”[85]. 论文解释了如何重新构建人类看来毫无异常的图像(对抗图

片)来欺骗神经网络, 同时证明了对抗图片不仅会骗过一个特定的神经网络, 而且可以骗过另一个网络, 甚至是所有的机器学习模型. 这一现象说明虽然深度学习目前在许多领域取得了重大成功, 但仍不够完善. LeCun指出深度学习存在一些局限性, 具体表现在以下几点: 缺乏理论支持、推理能力欠缺、短时记忆能力需加强以及无监督学习能力弱[2].

神经网络的架构规模、参数选择等问题都是由经验来确定. 通过随机梯度下降法优化得到的网络参数可能是一个局部最优值, 还有提升的空间. 深度学习未来的发展还需要更加完善的理论知识支撑.

深度学习由于缺乏逻辑推理能力, 在面对需要复杂推理的任务时受到一定限制. 目前深度学习结合结构化预测的方法是一种解决思路. 未来, 结合复杂推理的深度学习系统是一个具有重要意义的研究方向.

记忆能力对于处理有时间关联的信息具有重要的作用. 包括递归神经网络、长短时记忆在内的记忆模块可以存储大量的包含信息序列之间的内容, 这将有助于深度学习更好地模拟人脑的记忆功能. 如何有效地处理与时间序列有关的信息将是深度学习未来的一个研究方向.

目前大多数的深度学习方法都基于有监督学习, 包括图像分类、语音降噪等. 而在实际生活中, 大部分事物都是未知的、不带标记的, 这就增加了可以发现事物内在结构关系的无监督学习算法的需求. 然而无监督算法在这方面还未取得突破性成果, 未来发展空间巨大. LeCun也提出无监督学习在人类和动物学习中占据主导地位, 它将促进深度学习的进一步发展[71].

4 深度强化学习(Deep reinforcement learning)

在高级人工智能领域, 感知和决策能力都是衡量智能的指标. 然而直接通过学习高维感知输入(如图像、语音等)去控制智能体, 对强化学习来说是一个长期的挑战. 强化学习在策略选择的理论和算法方面已经取得了很大的进步. 其中大部分成功的强化学习应用方案依赖于人工特征的选取, 然而学习结果的好坏严重地取决于特征选取的质量[86].

近期深度学习的发展使得直接从原始的数据中提取高水平特征变成可能. 深度学习具有较强的感知能力, 但是缺乏一定的决策能力; 而强化学习具有决策能力, 对感知问题束手无策. 因此, 将两者结合起来, 优势互补, 为复杂系统的感知决策问题提供了解决思路.

4.1 深度强化学习的早期研究成果(Early research results of deep reinforcement learning)

深度Q网络出现之前, 已经出现了一些类似的研究工作. 它们的主要思路是将神经网络用于复杂高维数据的降维, 转化到低维特征空间便于强化学习处理. Shibata 等将浅层神经网络和强化学习结合处理视觉信号输入, 控制机器人完成推箱子等任务[87–88]. Lange等提出将高效的深度自动编码器应用到视觉的学习控制中, 提出了“视觉动作学习”, 使智能体具有和人相似的感知和决策能力[89]. 之后, Abtahi等将深度置信网络引入强化学习中, 用深度置信网络替代传统的值函数逼近器, 成功地应用在车牌图像的字符分割任务上[90]. 2012年, Lange等将基于视觉输入的强化学习应用到车辆控制中, 这种框架被称为深度拟合Q学习(deep fitted Q learning)[91]. 该算法输入跑道和车辆图像到深度网络, 提取出低维特征用于Q学习, 最后得到合适的控制策略. Koutnik 等将神经演化(neural evolution, NE)方法与强化学习结合, 在视频赛车游戏TORCS[92]中实现了赛车的自动驾驶[93].

4.2 基于卷积神经网络的深度强化学习 (Deep reinforcement learning based on convolutional neural network)

由于卷积神经网络对图像处理拥有天然的优势, 将卷积神经网络与强化学习结合处理图像数据的感知决策任务成了很多学者的研究方向. 深智团队在文献[86]中提出的深度Q网络(deep Q network, DQN), 是将卷积神经网络和Q学习结合, 并集成经验回放技术[94]实现的. 经验回放通过重复采样历史数据增加了数据的使用效率, 同时减少了数据之间的相关性.

深度Q网络是深度强化学习领域的开创性工作. 它采用时间上相邻的4帧游戏画面作为原始图像输入, 经过深度卷积神经网络和全连接神经网络, 输出状态动作Q函数, 实现了端到端的学习控制.

深度Q网络使用带有参数θ的Q函数$Q(s,a;\theta)$去逼近值函数. 迭代次数为i时, 损失函数为

$$L_i(\theta_i)=\mathrm{E}_{(s,a,r,s')}[(y_i^{\mathrm{DQN}}-Q(s,a;\theta_i))^2], \quad (12)$$

其中

$$y_i^{\mathrm{DQN}}=r+\gamma\max_{a'}Q(s',a';\theta^-), \quad (13)$$

θ_i代表学习过程中的网络参数. 经过一段时间的学习后, 新的θ_i更新θ^-. 具体的学习过程根据:

$$\nabla_{\theta_i}L_i(\theta_i)=\mathrm{E}_{(s,a,r,s')}[(r+\gamma\max_{a'}Q(s',a';\theta^-)-Q(s,a;\theta_i))\nabla_{\theta_i}Q(s,a;\theta_i)]. \quad (14)$$

该工作引起了学术界对深度强化学习的关注, 众多研究人员提出了很多改进工作. 深度Q网络使用的的经验回放技术没有考虑历史数据的重要程度, 而是同等频率的回放. Schaul等提出一种带优先级经验回放的深度Q网络, 对经验进行优先次序的处理, 增加重

[2] https://drive.google.com/file/d/0BxKBnD5y2M8NVHRiVXBnOVpiYUk/view

要历史数据的回放频率来提高学习效果, 同时也加快了学习进程[95]. 深度Q网络的另一个不足是它漫长的训练时间, 为此Nair等提出了深度Q网络的大规模分布式架构—Gorila, 极大提高了深度Q网络的学习速率[96]. Guo等提出将蒙特卡罗树搜索与深度Q网络结合, 实现了Atari游戏的实时处理, 游戏得分也普遍高于原始深度Q网络[97].

此外, Q学习由于学习过程中固有的估计误差, 在大规模数据的情况下会对动作的值产生过高估计. Van等提出的双重深度Q网络(double-DQN)将两个Q学习方法运用到深度Q网络中, 有效避免了过高估计, 并且获取到了更加稳定有效的学习策略[98]. Wang 等受优势学习(advantage learning)的启发提出一种适用于无模型强化学习的神经网络架构–竞争架构(dueling architecture), 并以实验证明竞争架构的深度Q 网络能够获取更好的评估策略[99]. 探索和利用问题(文献 [5])一直是强化学习中的主要问题. 复杂环境中的高效探索策略对深度强化学习的学习结果有深远影响. Osband 等提出一种引导(boot-strapped)深度Q网络, 通过使用随机值函数让探索的效率和速率得到了显著的提升[100]. Mnih等提出了异步深度强化学习方法, 在多核CPU上极大提升了训练速度[101].

4.3 基于递归神经网络的深度强化学习(Deep reinforcement learning based on recurrent neural network)

深度强化学习面临的问题往往具有很强的时间依赖性, 而递归神经网络适合处理和时间序列相关的问题. 强化学习与递归神经网络的结合也是深度强化学习的主要形式. Cuccu等提出将神经演化方法应用到基于视觉的强化学习中, 用一个预压缩器对递归神经网络进行训练. 采集的图像数据通过递归神经网络降维后输入给强化学习进行决策, 在基于视觉的小车爬山(Mountain car)任务中获得了良好的控制效果[102]. Narasimhan等提出一种长短时记忆网络与强化学习结合的深度网络架构来处理文本游戏. 这种方法能够将文本信息映射到向量表示空间从而获取游戏状态的语义信息[103].

对于时间序列信息, 深度Q网络的处理方法是加入经验回放机制. 但是经验回放的记忆能力有限, 每个决策点需要获取整个输入画面进行感知记忆. Hausknecht等将长短时记忆网络与深度Q网络结合, 提出深度递归Q网络(deep recurrent Q network, DRQN), 在部分可观测马尔科夫决策过程(partially observable Markov decision process, POMDP)中表现出了更好的鲁棒性, 同时在缺失若干帧画面的情况下也能获得很好的实验结果[104]. 随着视觉注意力机制在目标跟踪和机器翻译等领域的成功, Sorokin等受此启发提出深度注意力递归Q网络(deep attention recurrent Q network, DARQN). 它能够选择性地重点关注相关信息区域, 减少深度神经网络的参数数量和计算开销[105].

深度强化学习通过不断的发展, 在理论与应用方面取得了长足的进步. 表3总结了深度强化学习发展历程中的重要事件. 视频游戏上的成功没有让深度强化学习止步不前, 深智团队把目光转向代表人类智力水平的巅峰—围棋.

表 3 深度强化学习研究历程

Table 3 Timeline of deep reinforcement learning research events

2013	Mnih等提出了深度强化学习的开创性工作DQN, 在视频游戏领域取得突破[86]
2014	Guo等提出DQN与MCTS结合的算法[97]
2015	Nair等提出了适用于DQN的大规模分布式架构—Gorila[96]
2015	Van等提出了双重深度Q网络(double-DQN)[98]
2015	Hausknecht等结合LSTM提出了深度递归Q网络(DRQN)[104]
2015	Sorokin等结合注意力网络提出了深度注意力递归Q网络(DARQN)[105]
2016	深智团队在《Nature》上面发表了基于DRL的计算机围棋程序—初弈号[2]

5 深度强化学习典型应用—初弈号(Typical applications of deep reinforcement learning—AlphaGo)

近期, 使用深度强化学习的初弈号成为人工智能领域的焦点. 与此同时, 计算机围棋也成为人们热议的话题.

5.1 计算机围棋的发展历史与现状(Development history and present situation of computer Go)

计算机围棋起源于20世纪60年代, 长期以来, 它被认为是人工智能领域的一大挑战, 并为智能学习算法的研究提供了一个很好的测试平台. 计算机围棋通过计算一个大约含b^d个落子情况序列的搜索树上的最优值函数来评估棋局和选择落子位置, 其中b是搜索的宽度, d是搜索的深度. 与象棋等具有有限搜索空间的棋类不同, 围棋的计算复杂度约为250^{150}. 如果采用传统的暴力搜索方式, 按照现有的计算能力是远远无法解决围棋问题的[106]. 早期计算机围棋通过专家系统和模糊匹配缩小搜索空间, 减轻计算强度, 但受限于计算资源和硬件能力, 实际效果并不理想.

2006年, 蒙特卡罗树搜索的应用标志着计算机围棋进入了崭新的阶段[7]. 现代计算机围棋的主要算法是基于蒙特卡罗树的优化搜索. Coulom采用这种方法开发的CrazyStone在2006年计算机奥运会上首次夺得九路(9×9的棋盘)围棋的冠军. 2008年, 王一早开发的MoGo在9路围棋中达到段位水平. 2012年, 加藤

英树开发的Zen在19路(19 × 19的全尺寸棋盘) 围棋上以3:1击败二段棋手约翰特朗普. 2014年, 职业棋手依田记基九段让四子不敌CrazyStone, 这在围棋界引起了巨大的轰动. 赛后依田记基表示此时的CrazyStone大概有业余六七段的实力, 但是他依然认为数年内计算机围棋很难达到职业水准. 与此同时, 加藤英树也表示计算机围棋需要数十年的时间才能达到职业水准, 这也是当时大多数围棋领域和人工智能领域的专家持有的观点. 然而, 随着深度学习和蒙特卡罗树搜索方法的结合, 这一观点开始受到挑战. 2015年, Facebook人工智能研究院的Tian结合深度卷积神经网络和蒙特卡罗树搜索开发出的计算机围棋DarkForest表现出了与人类相似的下棋风格和惊人的实力, 这预示着计算机围棋达到职业水准的时间可能会提前[107]. 而2016年3月初弈号的横空出世彻底宣告基于人工智能算法的计算机围棋达到了人类顶尖棋手水准.

5.2 初弈号原理分析(Principle analysis of AlphaGo)

初弈号创新性地结合深度强化学习和蒙特卡罗树搜索, 通过价值网络(value network)评估局面以减小搜索深度, 利用策略网络(policy network)降低搜索宽度, 使搜索效率得到大幅提升, 胜率估算也更加精确[2]. 网络结构如图7所示:

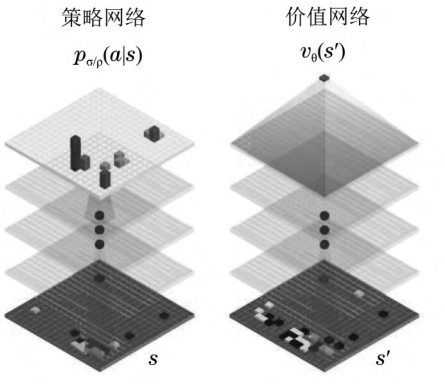

图 7　策略网络和价值网络[2]

Fig. 7　Policy network and value network[2]

策略网络将棋盘状态s作为输入, 经过13层的卷积神经网络输出不同落子位置的概率分布$p_\sigma(a|s)$或$p_\rho(a|s)$, 其中σ和ρ分别表示监督学习和强化学习得到的策略网络, a表示采取的落子选择. 价值网络同样使用深度卷积神经网络, 输出一个标量值$v_\theta(s')$来预测选择落子位置s'时的期望奖赏, θ为价值网络的参数.

初弈号的原理流程主要包含线下学习和在线对弈两部分.

5.2.1 线下学习(Offline learning)

初弈号的线下学习包含3个阶段, 如图8所示.

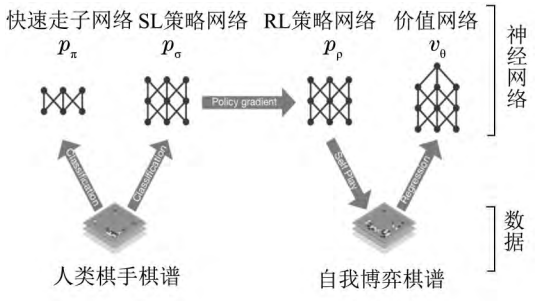

图 8　策略网络和价值网络训练过程[2]

Fig. 8　The training process of policy network and value network[2]

第1阶段, 深智团队使用棋圣堂围棋服务器(Kiseido Go server, KGS)上3000万个专业棋手对弈棋谱的落子数据, 基于监督学习得到一个策略网络, 来预测棋手的落子情况, 称之为监督学习的策略网络p_θ. 训练策略网络时采用随机梯度升序法更新网络权重.

$$\Delta\sigma \propto \frac{\partial \log p_\sigma(a|s)}{\partial \sigma}, \tag{15}$$

在使用全部48个输入特征的情况下, 预测准确率达到了55.7%, 这远远高于其他方法的结果. 同时他们也使用了局部特征匹配和线性回归的方法训练了一个快速走子策略网络p_Π, 在牺牲部分准确度的情况下极大地提高了走棋的速率.

第2阶段在第1阶段结果的基础上, 使用强化学习进一步对策略网络进行学习, 得到强化学习的策略网络p_ρ. 训练过程中先使用监督学习的策略网络对强化学习的策略网络进行初始化, 然后两者通过“自我博弈”来改善策略网络的性能. 训练过程中采用策略梯度算法, 按照预期结果最大值的方向, 更新权重.

$$\Delta\theta \propto \frac{\partial \log p_\rho(a_t|s_t)}{\partial \rho} z_t, \tag{16}$$

其中z_t是在时间步长为t时的奖赏, 胜方为+1、败方为−1. 在与监督学习的策略网络p_θ的对弈中, 强化学习的策略网络p_ρ能够获得80%的胜率.

第3阶段, 使用“自我博弈”产生的棋谱, 根据最终胜负结果来训练价值网络v_θ. 训练价值网络时, 使用随机梯度降序法来最小化预测值$v_\theta(s)$和相应结果z间的差值.

$$\Delta\rho \propto \frac{\partial v_\theta(s)}{\partial \theta}(z - v_\theta(s)), \tag{17}$$

训练好的价值网络可以对棋局进行评估, 预测最终胜负的概率.

5.2.2 在线对弈(Online playing)

初弈号通过蒙特卡罗树搜索将策略网络和价值网络结合起来, 利用前向搜索选择动作, 主要包含5个步骤.

预处理: 利用当前棋盘局面提取特征, 作为深度网络的输入, 最终的初弈号网络输入包含了48个特征层.

选择: 每次模拟时从根节点出发遍历搜索树, 根据最大动作值Q和激励值$u(s,a)$选择下一个节点.

$$u(s,a) \propto \frac{p(s,a)}{1+N(s,a)}, \tag{18}$$

其中$N(s,a)$是访问次数. 遍历进行到步骤L时, 节点记为s_L.

展开: 访问次数达到一定数目时, 叶节点展开, 展开时被监督学习策略网络p_σ处理一次, 此时的输出概率保存为对应动作的前向概率$P(s,a)=p_\sigma(a|s)$, 根据前向概率计算不同落子位置往下发展的权重.

评估: 叶节点有两种评估方式: 价值网络的估值$v_\theta(s_L)$和快速走子产生的结果z_L. 这是因为棋局开始时, 价值网络的估值比较重要, 随着棋局的进行, 局面状态变得复杂, 这时会更加看重快速走子产生的结果[108]. 两者通过加权的方式计算叶节点的估值$V(s_L)$.

备份: 将评估结果作为当前棋局下一步走法的Q值.

$$Q(s,a)=\frac{1}{N(s,a)}\sum_{i=1}^{n}1(s,a,i)V(s_L^i), \tag{19}$$

其中$1(s,a,i)$表示进行第i次模拟时状态动作对(s,a)是否被访问. Q值越大, 之后的模拟选择此走法的次数越多. 模拟结束时, 遍历过的节点的状态动作值和访问次数得到更新. 每个节点累计经过此节点的访问次数和平均估值. 反复进行上述过程达到一定次数后搜索完成, 算法选取从根节点出发访问次数最多的那条路径落子.

初弈号的整个原理流程如图9所示.

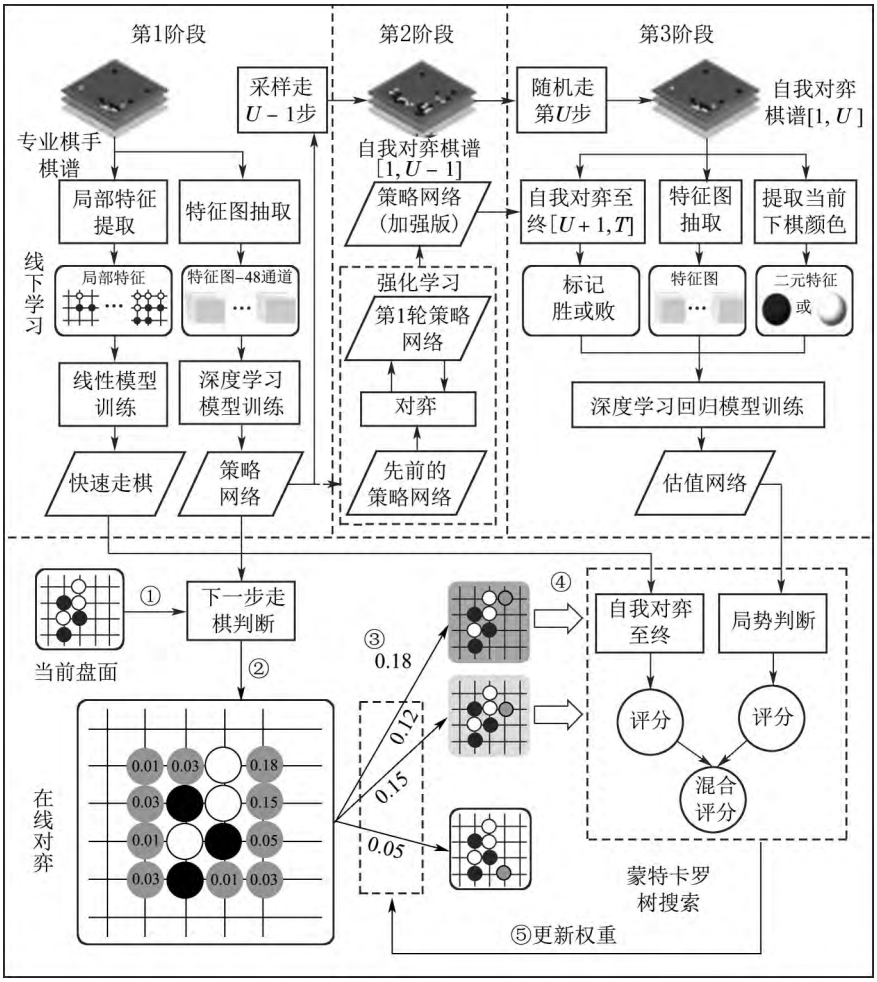

图 9 初弈号原理图[3]

Fig. 9 The principle of AlphaGo

5.3 初弈号性能分析(Performance analysis of AlphaGo)

初弈号成功地整合了上述方法, 并依托强大的硬件支持达到了顶尖棋手的水平. 与此同时, 在与李世石的比赛中, 我们也看到了初弈号不尽完美的一面.

[3]http://www.kddchina.org/#/Content/alphago

5.3.1 成功的原因(Reason of success)

初弈号的成功离不开深度神经网络. 传统的基于规则的计算机围棋方法只能识别固定的棋路, 这类似于背棋谱. 基于深度学习的初弈号自动提取棋谱局面特征并将其有效地组合在一起, 极大增强了对棋谱的学习能力. 其次, 局面评估也是初弈号成功的关键. 价值网络和快速走子网络在局面评估时的互为补充, 能够较好地应对对手下一步棋的不确定性, 对得到更加精确的评估结果至关重要. 此外, 硬件配置的大幅提升也功不可没. 初弈号采用了异步多线程搜索, 用CPU执行模拟过程, 用GPU计算策略网络和价值网络. 最终单机版本初弈号使用了48个CPU和8个GPU, 分布式版本的初弈号则采用了1202个CPU和176个GPU. 正是这些计算机硬件的支持, 才得以让初弈号发挥出强大的实力[2].

5.3.2 打劫问题分析(Analysis of robbery problem)

初弈号在对人类顶尖棋手的对弈中取得了令人瞩目的成绩, 但它也并非完美无缺, 其中打劫能力可能是制约初弈号的一个主要因素. 打劫在围棋对弈中占据着十分重要的地位, 获取最佳的打劫策略一直是计算机围棋的研究难点. 初弈号的研发成员Huang认为价值网络考虑打劫后搜索深度通常会加深, 复杂度也会提高很多, 所以一般的算法选择消劫[109]. Zheng等认为早期棋盘搜索空间大, 此时打劫能够极大地增加初弈号的搜索宽度和深度. 作为应对初弈号的策略, 最好在刚进入中盘时开劫, 并且能在盘面上长时间保持两处以上的劫争, 但随着比赛的进行搜索空间变小, 就应该尽量避免打劫[4]. 关于打劫的问题, 目前尚有争议. 初弈号在和樊麾对弈的第3局有打劫、第5局更是有多次打劫, 和李世石的比赛中也出现打劫, 并没有明显异常的表现. 从算法原理上分析, 打劫有多种, 只有在个别情况(循环劫)才可能产生蒙特卡罗搜索树的节点循环现象. 但也可以采用状态判断和估值来跳出这个循环节点.

5.3.3 第 4 局失利分析(Analysis of losing the 4th game)

初弈号在第4局的失利也让我们认识到它需要改进的地方还很多. 训练初弈号所用的棋谱, 只有小部分是人类职业选手的棋局, 总数上亿的棋局是“自我博弈”产生的, 这远远多于高质量的人类棋谱. 在整个训练数据集中, 低质量的样本占据了绝大多数. 训练样本分布的不均衡可能是导致初弈号失利的原因之一[5]. 蒙特卡罗树搜索本质上是一种随机搜索, 只能在一定的概率下得到正确的搜索结果, 相比于人类基于逻辑推理的方式, 可能会对局势产生非准确的判断. 初弈号在“自我博弈”的过程中使用的是强化学习. 而强化学习的一个突出问题是存在学习的盲区, 即在整个学习过程中, 存在没有被探索到的部分状态空间[110]. 初弈号另一研发成员哈萨比斯赛后也提到其可能存在短暂盲区. 如果找到了初弈号学习的盲区, 就能找到相应的与其对弈的策略.

5.4 初弈号评价(Evaluation of AlphaGo)

围棋因为复杂的落子选择和庞大的搜索空间在人工智能领域具有显著的代表性. 初弈号基于深度卷积神经网络的策略网络和价值网络减小了搜索空间, 并且在训练过程中创新性地结合了监督学习和强化学习, 最后成功地整合蒙特卡罗树搜索算法. 初弈号作为人工智能领域的里程碑, 其智能突出体现在以下4点:

1) 棋谱数据可以完全获取, 知识能够自动表达. 围棋是一种完全信息博弈的游戏; 通过摄像机拍摄即可获得全部的状态信息. 初弈号能够获得完备的数据集, 并且将数据自动地表示成知识.

2) 初弈号能够较好地应对对手下一步棋的不确定性, 按搜索和评价策略进行决策. 通常控制界要先给出系统的很多假设, 比如不确定性在一定的范围之内, 才能证明系统的收敛性或稳定性; 而人工智能是感知与认知交互迭代的方法, 对系统的不确定性不作预先假设, 虽然很难得到理论证明, 但从实践中(搜索和评价)获得成功. 初弈号在应对不确定性中的优秀表现彰显了其智能水平.

3) 以标准赛制产生的人类棋手为智能标准, 设计了较好的智能评价准则. 围棋是一个标准赛制的游戏, 其用段位科学地描述棋手的水平. 因此, 计算机围棋的智能水平很容易通过人类棋手来测试. 通过与职业棋手樊麾和李世石的对弈, 初弈号的智能水平得到了很好的测试.

4) 初弈号通过“自我博弈”产生3000万盘棋, 深度模仿人类顶尖棋手的对弈, 提升系统的智能水平. 初弈号具有强大的自学习能力, 通过深度强化学习的机制不断提高自身水平, 从战胜樊麾到战胜李世石, 经历时间不长, 出乎大多数人意料, 可见自学习在其中发挥了重要作用.

虽然根据公开发表的资料, 初弈号所使用的强化学习、深度学习、蒙特卡罗树搜索等人工智能算法都是已有的、广为人知的方法, 但初弈号与人类

[4] http://www.kddchina.org/#/Content/alphago

[5] http://36kr.com/p/5044469.html

棋手对弈的结果表明, 它已具备了高级智能, 达到了顶级棋手的对弈水准.

6 深度强化学习研究的展望(Prospect of deep reinforcement learning research)

随着深度强化学习的不断发展, 越来越多的实际问题得到了解决. Atari视频游戏上的成功和计算机围棋初弈号的出现都对深度强化学习的发展起到了巨大的推动作用. 然而, 当前的深度强化学习在理论与应用方面依然存在一些不足, 如何解决这些问题并将深度强化学习推向更广阔的应用场景将是未来的研究主题.

6.1 博弈问题(Game problem)

初弈号是目前深度强化学习在解决实际问题时最成功的案例. 但初弈号只是解决了二人零和完全信息博弈的最优决策问题, 还有很多其他博弈问题, 如二人非零和完全信息博弈问题, 多人(多智能体)博弈(包括零和、非零和、完全信息、不完全信息)问题等[111], 都期待着深度强化学习方法能带来大的突破.

其中, 多智能体博弈问题是博弈领域的前沿热点. 多智能体系统[112]是由多个相互联系的智能体组成的系统, 具有自主性、分布性、协调性等特点. 由于深度强化学习对智能体智能水平的巨大提升, 深度强化学习与多智能体系统的结合会进一步加强多智能体博弈的感知和决策水平. Foerster等针对部分可观测马尔科夫决策过程中多智能体非零和博弈问题, 提出了深度分布式递归Q网络, 并成功学习到通信协议[113].

6.2 连续状态动作问题(Continuous state and action problem)

目前, 主流的深度强化学习方法都是针对离散状态和动作的优化问题, 而实际问题中的状态和动作往往是连续的. 对于连续状态和动作的深度强化学习, 虽然已经有初步的研究[114], 但不够深入, 缺少完备的理论支撑, 限制了深度强化学习更广泛的应用. Lillicrap等针对无模型强化学习提出了连续状态和动作的深度强化学习, 成功地应用在机器人控制问题上[115].

自适应动态规划方法是一种有效的决策控制方法, 与传统的面向离散状态动作系统的Q学习等方法相比、更适用于解决连续状态动作系统的控制问题. 但是自适应动态规划在解决高维环境的控制问题时遇到了不可逾越的障碍. 直接使用高维状态输入很难使自适应动态规划的参数收敛, 不能将抽象的图像或音频输入有效地转化成控制策略. 而深度学习可以通过非监督方法进行特征学习, 有效降低输入信息的维数. 因此, 将深度学习和自适应动态规划相结合的深度自适应动态规划(deep ADP, DADP)会为更广泛的复杂系统的智能控制提供新的解决思路. Zhao等将深度卷积神经网络和自适应动态规划方法结合在一起, 在视觉控制任务取得了良好效果[116].

6.3 与其他智能方法的结合(Combination with other intelligent methods)

深度强化学习与其他智能方法的结合是实现高级人工智能的重要途径. 初弈号由于创新性地提出了深度强化学习与蒙特卡罗树搜索结合的方法, 极大地推动了计算机围棋的发展. 但从初弈号第4局的对弈结果来看, 蒙特卡罗树搜索会遇到搜索盲区的问题, 从而导致算法性能下降. 因此, 深度强化学习和更加有效的搜索方法相结合, 会是未来的一个研究方向.

机器学习算法中的一个主要假设是数据在相同的特征空间具有相同的分布. 由于现实环境的复杂多样, 这种假设限制了机器学习算法的应用范围. 迁移学习[117]是一类机器学习方法, 它降低了对这种假设的要求, 能够利用不同领域的数据帮助目标学习. 这类似于人类学习过程中的举一反三. 对于智能体来说, 将不同领域中学到的知识迁移到新的环境是衡量智能体水平的一个关键因素. 深度强化学习可以与迁移学习结合来训练深度策略网络. Parisotto等针对一些游戏之间的相似性, 在深度Q网络中引入迁移学习, 加快了收敛速度, 提高学习性能[118].

6.4 深度强化学习理论分析(Theoretical analysis of deep reinforcement learning)

虽然深度强化学习展现出了广阔的应用前景, 但是在理论层面上对其研究还远远不够. 深智团队在《Nature》上的文章[1]没有给出深度Q网络学习过程收敛性的证明, 其他一些基于深度Q网络的改进工作以及新的深度强化学习框架也没有很好地解决这个问题. 对于控制问题, 控制器的稳定性是实际应用中首要考虑的问题. 而目前缺乏对该方面的研究. 长期来说, 理论上的进步势必会促进深度强化学习的良性发展.

同样的问题反映在对深度强化学习内在机理的理解方面. 虽然深度强化学习为实现高级人工智能提供了解决思路, 但研究人员对其内在机理知之甚少. 此类认知过程的准确理解对进一步完善深度模型提出了不小的挑战[3]. 在不久的将来, 攻克这些难题势必会为高级人工智能带来新的突破.

7 深度强化学习的应用(Application of deep reinforcement learning)

深度强化学习在游戏、棋类问题上展现出强大的能力. 如何迁移到其他领域, 继续发挥深度强化学习的优势, 是研究人员需要考虑和分析的重点.

7.1 游戏(Games)

深智团队在《Nature》上发表的关于深度强化学习在游戏中应用的2篇论文, 从玩Atari游戏的深度Q网络到计算机围棋初弈号, 仅用了短短一年时间. 攻克了围棋并不意味着深度强化学习在游戏行业中应用的终结, 因为围棋只是一种完全信息博弈游戏. 同时Atari中部分策略类游戏的状态和动作是低维的. 而对于状态部分可观测的不完全信息游戏及高维度策略类游戏, 甚至是多玩家博弈的竞技类策略游戏, 预计很快会有新的突破.

7.2 智能驾驶(Intelligent driving)

传感器技术的进步使得智能驾驶得到了快速发展. 目前, 国内外众多研究团队已经将智能驾驶作为重要研究方向并取得了初步的研究成果. 然而类似激光雷达这类传感器的使用大大增加了智能车成本, 极大地阻碍了智能车的普及. 近年来, 基于摄像头的先进驾驶员辅助系统(advanced driver assistance systems, ADAS)逐渐成为智能驾驶的关键技术之一. 该技术基于摄像头获取图像信息, 通过深度学习实时提取路况特征, 设计相应的智能控制算法, 实现稳定可靠的智能驾驶. 深度强化学习作为一类自学习智能控制算法, 可用来解决车辆的复杂非线性系统控制问题. 根据现有深度强化学习在TORCS赛车平台的研究[115]可以预测, 深度强化学习将在智能驾驶领域发挥巨大的作用, 成为降低智能车成本的一个可行方案.

7.3 智能医疗(Intelligent medical services)

健康一直是人们最为关注的话题, 人们在求医问药时都希望得到好医生的指点. 医生需要大量的理论知识学习和临床实践经验才能达到治病救人水平. 深度强化学习能够从基础的知识表示中学习出各个知识之间的关系, 具有强大的自学习能力. 作为深度强化学习的一个具体应用, 深智团队已经与英国的国家医疗服务体系(national health service, NHS)展开合作, 开发出了两款手机应用Streams和HARK[6]. 它们通过监控人体的多种指标来为紧急治疗提供参考数据分析, 并在确诊病情之后, 为医生提供该病症所需的基本操作流程说明和工具资料, 加快医生的看病速度. 然而, 这仅仅是一个开端. 通过人工智能与世界级医疗团队的合作来解决长期存在的医疗问题才是最终目的.

7.4 个人手机助手(Personal phone assistant)

随着智能通讯设备的普及, 手机已经成为了人们生活中不可或缺的一部分, 人们每天与手机发生着成百上千次的交互. 而现阶段人与手机的互动关系为: 人发出指令, 手机执行指令. 手机主动与人进行交互来提供解决方案还没有实现. 在监督式学习的框架下算法很难提升学习性能, 因此深智团队已经开始着手基于庞大的数据集并采用非监督式的学习方法来提升性能, 并籍此希望解决传统手机助手只能针对某些特定的应用场景的问题. 现阶段这方面的应用还处于起步阶段, 需要长时间的探索.

7.5 机器人(Robots)

机器人的应用已经十分广泛, 无论是在工厂、学校, 还是家庭. 近几年, 深度强化学习在机器人领域的应用也有了飞速的发展. 加州大学伯克利分校的Levine等结合卷积神经网络和强化学习, 研制出了能够以纯视觉输入来抓取物体的机器人[119]. 最近, 该团队与谷歌合作, 通过大规模的训练数据, 使机器人能够从一堆物体中识别并抓取某件特定物体[120]. 在美国新成立了一家深度强化学习公司Osaro[7], 该公司致力于使用深度强化学习技术, 融合多传感器信息来提供包括工业机器人、无人机、自动驾驶汽车和物联网等应用领域的解决方案.

7.6 智能制造(Intelligent manufacture)

制造业是一个国家的支柱, 2015年国务院推出“中国制造2025”战略之后, 各个省市都开始思考如何将深度学习、强化学习等人工智能方法应用到制造业. 陕西省召开的国家智能制造战略下的大数据与深度学习研讨报告会, 来自工业界和学术界的人们探讨了在大数据背景下如何将深度学习等技术应用于制造业. 为实现大型工业企业的智能制造, 已开发出上层的企业资源规划(enterprise resource planning, ERP)系统和下层的制造执行系统(manufacturing execution system, MES). 但大多仍止步于实现了企业生产信息的自动化, 尚未能根据这些信息有效感知出企业运转和设备运行的状态, 并在状态感知的基础上进行系统优化决策管理. 深度强化学习在大数据的基础上有机融合了感知和决策, 势必能在智能制造中发挥重要作用.

[6] https://deepmind.com/health.html

[7] http://www.osaro.com/

8 结束语(Conclusion remarks)

深度强化学习是深度学习和强化学习结合的产物, 是一种有效的人工智能算法. 除了玩视频游戏的深度Q网络, 下围棋的初弈号, 深度强化学习还可以在自动驾驶、机器人等其他领域发挥巨大作用. 深度强化学习作为高级人工智能的一种算法, 仍然处于起步阶段. 通过对深度强化学习和初弈号的介绍, 可以看到人工智能只是人类精心设计的算法程序. 它的成功依赖于大量数据样本、计算机硬件和人类的智慧. 人工智能尚没有独立的思维能力和学习能力, 也远没到威胁人类生存发展的地步. 我们更应该清楚地看到, 在深度强化学习等人工智能研究领域, 以谷歌公司的深智团队为代表的国外研究机构处于绝对领先的地位. 其先进的基础理论方法、广泛的日常生活应用、以及潜在的军事领域扩展, 都非常可能进一步加大我们与国外的差距. 因此, 我们亟需更加深入和广泛地开展关于人工智能的前沿理论和算法的基础性研究, 以及面向各种载体或任务的、软硬件相结合的应用性研究. 希望在不久的将来, 我国人工智能的基础理论和应用水平都能产生具有重大国际影响的成果.

致谢 感谢夏中谱对强化学习部分, 吕乐、唐振韬对深度学习和初弈号部分提供的宝贵意见, 感谢卜丽、张启超对全文修改的帮助.

参考文献(References):

[1] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. *Nature*, 2015, 518(7540): 529 – 533.

[2] SILVER D, HUANG A, MADDISON C, et al. Mastering the game of Go with deep neural networks and tree search [J]. *Nature*, 2016, 529(7587): 484 – 489.

[3] AREL I. Deep reinforcement learning as foundation for artificial general intelligence [M] //*Theoretical Foundations of Artificial General Intelligence*. Amsterdam: Atlantis Press, 2012: 89 – 102.

[4] TEAAURO G. TD-Gammon, a self-teaching backgammon program, achieves master-level play [J]. *Neural Computation*, 1994, 6(2): 215 – 219.

[5] SUTTON R S, BARTO A G. *Reinforcement Learning: An Introduction* [M]. Cambridge MA: MIT Press, 1998.

[6] KEARNS M, SINGH S. Near-optimal reinforcement learning in polynomial time [J]. *Machine Learning*, 2002, 49(2/3): 209 – 232.

[7] KOCSIS L, SZEPESVARI C. Bandit based Monte-Carlo planning [C] //*Proceedings of the European Conference on Machine Learning*. Berlin: Springer, 2006: 282 – 293.

[8] LITTMAN M L. Reinforcement learning improves behaviour from evaluative feedback [J]. *Nature*, 2015, 521(7553): 445 – 451.

[9] BELLMAN R. Dynamic programming and Lagrange multipliers [J]. *Proceedings of the National Academy of Sciences*, 1956, 42(10): 767 – 769.

[10] WERBOS P J. Advanced forecasting methods for global crisis warning and models of intelligence [J]. *General Systems Yearbook*, 1977, 22(12): 25 – 38.

[11] WATKINS C J C H. *Learning from delayed rewards* [D]. Cambridge: University of Cambridge, 1989.

[12] RUMMERY G A, NIRANJAN M. *On-Line Q-Learning Using Connectionist Systems* [M]. Cambridge: University of Cambridge, Department of Engineering, 1994.

[13] BERTSEKAS D P, TSITSIKLIS J N. Neuro-dynamic programming: an overview [C] //*Proceedings of the 34th IEEE Conference on Decision and Control*. New Orleans: IEEE, 1995, 1: 560 – 564.

[14] THRUN S. Monte Carlo POMDPs [C] // *Advances in Neural Information Processing Systems*. Denver: MIT Press, 1999, 12: 1064 – 1070.

[15] LEWIS F L, VRABIE D. Reinforcement learning and adaptive dynamic programming for feedback control [J]. *IEEE Circuits and Systems Magazine*, 2009, 9(3): 32 – 50.

[16] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms [C] //*Proceedings of the International Conference on Machine Learning*. Beijing: ACM, 2014: 387 – 395.

[17] CAFLISCH R E. Monte Carlo and quasi-Monte Carlo methods [J]. *Acta Numerica*, 1998, 7: 1 – 49.

[18] MICHIE D, CHAMBERS R A. BOXES: An experiment in adaptive control [J]. *Machine Intelligence*, 1968, 2(2): 137 – 152.

[19] BARTO A G, DUFF M. Monte Carlo matrix inversion and reinforcement learning [C] //*Advances in Neural Information Processing Systems*. Denver: NIPS, 1993: 687 – 694.

[20] BROWNE C B, POWLEY E, WHITEHOUSE D, et al. A survey of Monte Carlo tree search methods [J]. *IEEE Transactions on Computational Intelligence and AI in Games*, 2012, 4(1): 1 – 43.

[21] CHASLOTH G. Monte-Carlo tree search [D]. *Maastricht: Maastricht Universiteit*, 2010.

[22] COULOM R. Efficient selectivity and backup operators in Monte-Carlo tree search [M] //*Computers and Games*. Berlin Heidelberg: Springer, 2006: 72 – 83.

[23] WEI Q L, LIU D R. A new discrete-time iterative adaptive dynamic programming algorithm based on Q-learning [M] //*International Symposium on Neural Networks*. New York: Springer, 2015: 43 – 52.

[24] WEI Q L, LIU D R, SHI G. A novel dual iterative-learning method for optimal battery management in smart residential environments [J]. *IEEE Transactions on Industrial Electronics*, 2015, 62(4): 2509 – 2518.

[25] JAAKKOLA T, JORDAN M I, SINGH S P. On the convergence of stochastic iterative dynamic programming algorithms [J]. *Neural Computation*, 1994, 6(6): 1185 – 1201.

[26] TSITSIKLIS J N. Asynchronous stochastic approximation and Q-learning [J]. *Machine Learning*, 1994, 16(3): 185 – 202.

[27] WATKINS C J C H, DAYAN P. Q-learning [J]. *Machine Learning*, 1992, 8(3/4): 279 – 292.

[28] SINGH S, JAAKKOLA T, LITTMAN M L, et al. Convergence results for single-step on-policy reinforcement-learning algorithms [J]. *Machine Learning*, 2000, 38(3): 287 – 308.

[29] SUTTON R S. Learning to predict by the methods of temporal differences [J]. *Machine Learning*, 1988, 3(1): 9 – 44.

[30] DEGRIS T, PILARSKI P M, SUTTON R S. Model-free reinforcement learning with continuous action in practice [C] //*Proceedings of the American Control Conference*. Montreal: IEEE, 2012: 2177 – 2182.

[31] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning [J]. *Machine Learning*, 1992, 8(3/4): 229 – 256.

[32] SUTTON R S, MCALLESTER D A, SINGH S P, et al. Policy gradient methods for reinforcement learning with function approximation [C] //*Advances in Neural Information Processing Systems*. Denver: MIT Press, 1999, 99: 1057 – 1063.

[33] LIU D R, WEI Q L. Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, 25(3): 621 – 634.

[34] ZHANG H G, LIU D R, LUO Y H, et al. *Adaptive Dynamic Programming for Control: Algorithms and Stability* [M]. New York: Springer, 2012.

[35] ZHAO D B, XIA Z P, WANG D. Model-free optimal control for affine nonlinear systems with convergence analysis [J]. *IEEE Transactions on Automation Science and Engineering*, 2015, 12(4): 1461 – 1468.

[36] ZHAO D B, ZHU Y H. MEC—a near-optimal online reinforcement learning algorithm for continuous deterministic systems [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, 26(2): 346 – 356.

[37] ZHU Y H, ZHAO D B, LI X J. Using reinforcement learning techniques to solve continuous-time non-linear optimal tracking problem without system dynamics [J]. *IET Control Theory & Applications*, 2016, DOI: 10.1049/iet-cta.2015.0769.

[38] JIANG Y, JIANG Z P. Robust adaptive dynamic programming and feedback stabilization of nonlinear systems [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, 25(5): 882 – 893.

[39] WU H N, LUO B. Neural network based online simultaneous policy update algorithm for solving the HJI equation in nonlinear control [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2012, 23(12): 1884 – 1895.

[40] ZHAO D B, ZHANG Q C, WANG D, et al. Experience replay for optimal control of nonzero-sum game systems with unknown dynamics [J]. *IEEE Transactions on Cybernetics*, 2016, 46(3): 854 – 865.

[41] WU Jun, XU Xin, WANG Jian, et al. Recent advances of reinforcement learning in multi-robot systems: a survey [J]. *Control and Decision*, 2011, 26(11): 1601 – 1610.
(吴军, 徐昕, 王健, 等. 面向多机器人系统的增强学习研究进展综述 [J]. 控制与决策, 2011, 26(11): 1601 – 1610.)

[42] MATARIC M J. Reinforcement learning in the multi-robot domain [M] //*Robot Colonies*. New York: Springer, 1997: 73 – 83.

[43] ZHAO D B, ZHANG Z, DAI Y J. Self-teaching adaptive dynamic programming for Gomoku [J]. *Neurocomputing*, 2012, 78(1): 23 – 29.

[44] ZHAO D B, WANG B, LIU D R. A supervised actor – critic approach for adaptive cruise control [J]. *Soft Computing*, 2013, 17(11): 2089 – 2099.

[45] KAKADE S. A natural policy gradient [C] //*Advances in Neural Information Processing Systems*. Vancouver: MIT Press, 2001, 14: 1531 – 1538.

[46] TSITSIKLIS J N, VAN R B. An analysis of temporal-difference learning with function approximation [J]. *IEEE Transactions on Automatic Control*, 1997, 42(5): 674 – 690.

[47] TSITSIKLIS J N, VAN R B. Average cost temporal-difference learning [J]. *Automatica*, 1999, 35(11): 1799 – 1808.

[48] BHATNAGAR S, PRECUP D, SILVER D, et al. Convergent temporal-difference learning with arbitrary smooth function approximation [C] //*Advances in Neural Information Processing Systems*. Vancouver: MIT Press, 2009: 1204 – 1212.

[49] SUTTON R S, MAEI H R, PRECUP D, et al. Fast gradient-descent methods for temporal-difference learning with linear function approximation [C] //*Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal: ACM, 2009: 993 – 1000.

[50] MELO F S, LOPES M. Fitted natural actor-critic: a new algorithm for continuous state-action MDPs [M] //*Machine Learning and Knowledge Discovery in Databases*. Berlin Heidelberg: Springer, 2008: 66 – 81.

[51] BRAFMAN R I, TENNENHOLTZ M. R-MAX – a general polynomial time algorithm for near-optimal reinforcement learning [J]. *The Journal of Machine Learning Research*, 2003, 3(10): 213 – 231.

[52] GAO Yang, CHEN Shifu, LU Xin. Research on reinforcement learning technology: a review [J]. *Acta Automatica Sinica*, 2004, 30(1): 86 – 100.
(高阳, 陈世福, 陆鑫. 强化学习研究综述 [J]. 自动化学报, 2004, 30(1): 86 – 100.)

[53] BERNSTEIN A, SHIMKIN N. Adaptive-resolution reinforcement learning with efficient exploration in deterministic domains [J]. *Machine Learning*, 2010, 81(3): 359 – 397.

[54] LI L H, CHU W, LANGFORD J, et al. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms [C] //*Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. Hong Kong: ACM, 2011: 297 – 306.

[55] NOURI A, LITTMAN M L, LI L H, et al. A novel benchmark methodology and data repository for real-life reinforcement learning [C] //*Proceedings of the 26th International Conference on Machine Learning*. Montreal: ACM, 2009.

[56] LOFTIN R, MACGLASHAN J, PENG B, et al. A strategy-aware technique for learning behaviors from discrete human feedback [C] //*Proceedings of the Association for the Advancement of Artificial Intelligence*. Québec City: AAAI, 2014: 937 – 943.

[57] THOMAZ A L, BREZEAL C. Teachable robots: understanding human teaching behavior to build more effective robot learners [J]. *Artificial Intelligence*, 2008, 172(6): 716 – 737.

[58] NIV Y. Neuroscience: Dopamine ramps up [J]. *Nature*, 2013, 500(7464): 533 – 535.

[59] CUSHMAN F. Action, outcome, and value a dual-system framework for morality [J]. *Personality and Social Psychology Review*, 2013, 17(3): 273 – 292.

[60] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets [J]. *Neural Computation*, 2006, 18(7): 1527 – 1554.

[61] ABDEL-HAMID O, MOHAMED A, JIANG H, et al. Convolutional neural networks for speech recognition [J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, 22(10): 1533 – 1545.

[62] CARLSON B A, CLEMENTS M A. A projection-based likelihood measure for speech recognition in noise [J]. *IEEE Transactions on Speech and Audio Processing*, 1994, 2(1): 97 – 102.

[63] OUYANG W, ZENG X, WANG X. Learning mutual visibility relationship for pedestrian detection with a deep model [J]. *International Journal of Computer Vision*, 2016, DOI: 10.1007/s11263-016-0890-9.

[64] DAHL G E, YU D, DENG L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, 20(1): 30 – 42.

[65] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [C] //*Advances in Neural Information Processing Systems*. Lake Tahoe: MIT Press, 2012: 1097 – 1105.

[66] LE Q V. Building high-level features using large scale unsupervised learning [C] //*Proceedings of the IEEE International Conference on*

Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013: 8595 – 8598.

[67] GRAVERS A, MOHAMED A, HINTON G. Speech recognition with deep recurrent neural networks [C] //*Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver: IEEE, 2013: 6645 – 6649.

[68] XU K, BA J, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention [C] //*Proceedings of the 32nd International Conference on Machine Learning*. Lille: ACM, 2015: 2048 – 2057.

[69] PINHEIRO P, COLLOBERT R. Recurrent convolutional neural networks for scene labeling [C] //*Proceedings of the 31nd International Conference on Machine Learning*. Beijing: ACM, 2014: 82 – 90.

[70] HE K M, ZHANG X, REN S, et al. Deep residual learning for image recognition [C] //*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016.

[71] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. *Nature*, 2015, 521(7553): 436 – 444.

[72] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders [C] //*Proceedings of the 25th International Conference on Machine Learning*. Helsinki: ACM, 2008: 1096 – 1103.

[73] RIFAI S, VINCENT P, MULLER X, et al. Contractive auto-encoders: Explicit invariance during feature extraction [C] //*Proceedings of the 28th International Conference on Machine Learning*. Bellevue: ACM, 2011: 833 – 840.

[74] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition [J]. *Proceedings of the IEEE*, 1998, 86(11): 2278 – 2324.

[75] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL] //*arXiv preprint*. 2015. arXiv:1409.1556[cs.CV].

[76] TIELEMAN T. Training restricted Boltzmann machines using approximations to the likelihood gradient [C] //*Proceedings of the 25th International Conference on Machine Learning*. Helsinki: ACM, 2008: 1064 – 1071.

[77] TIELEMAN T, HINTON G. Using fast weights to improve persistent contrastive divergence [C] //*Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal: ACM, 2009: 1033 – 1040.

[78] MOHAMED A, DAHL G E, HINTON G. Acoustic modeling using deep belief networks [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, 20(1): 14 – 22.

[79] FENG X, ZHANG Y, GLASS J. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition [C] //*Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence: IEEE, 2014: 1759 – 1763.

[80] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion [J]. *The Journal of Machine Learning Research*, 2010, 11(11): 3371 – 3408.

[81] BOULANGER-LEWANDOWSKI N, BENGIO Y, VINCENT P. Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription [C] //*Proceedings of the 29th International Conference on Machine Learning*. Edinburgh: ACM, 2012: 1159 – 1166.

[82] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks [J]. *IEEE Transactions on Signal Processing*, 1997, 45(11): 2673 – 2681.

[83] CHO K, VAN M B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C] //*Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Doha: ACL, 2014: 1724 – 1734.

[84] KOEHN P, HOANG H, BIRCH A, et al. Moses: open source toolkit for statistical machine translation [C] //*Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Stroudsburg: ACM, 2007: 177 – 180.

[85] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [EB/OL] //*arXiv preprint*. 2015. arXiv:1412.6572v3[stat.ML].

[86] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning [C] //*Proceedings of the NIPS Workshop on Deep Learning*. Lake Tahoe: MIT Press, 2013.

[87] SHIBATA K, IIDA M. Acquisition of box pushing by direct-vision-based reinforcement learning [C] //*Proceedings of the SICE Annual Conference*. Nagoya: IEEE, 2003, 3: 2322 – 2327.

[88] SHIBATA K, OKABE Y. Reinforcement learning when visual sensory signals are directly given as inputs [C] //*Proceedings of the International Conference on Neural Networks*. Houston: IEEE, 1997, 3: 1716 – 1720.

[89] LANGE S, RIEDMILLER M. Deep auto-encoder neural networks in reinforcement learning [C] //*Proceedings of the International Joint Conference on Neural Networks*. Barcelona: IEEE, 2010: 1 – 8.

[90] ABTAHI F, ZHU Z, BURRY A M. A deep reinforcement learning approach to character segmentation of license plate images [C] //*Proceedings of the 14th IAPR International Conference on Machine Vision Applications*. Tokyo: IEEE, 2015: 539 – 542.

[91] LANGE S, RIEDMILLER M, VOIGTLANDER A. Autonomous reinforcement learning on raw visual input data in a real world application [C] //*Proceedings of the International Joint Conference on Neural Networks*. Brisbane: IEEE, 2012: 1 – 8.

[92] WYMANN B, ESPI E, GUIONNEAU C, et al. TORCS, The open racing car simulator [EB/OL]. 2014, http://torcs.sourceforge.net.

[93] KOUTNIK J, SCHMIDHUBER J, GOMEZ F. Online evolution of deep convolutional network for vision-based reinforcement learning [M] //*From Animals to Animats 13*. New York: Springer, 2014: 260 – 269.

[94] LIN L J. *Reinforcement learning for robots using neural networks* [D]. Pittsburgh: Carnegie Mellon University, 1993.

[95] SCHAUL T, QUAN J, ANTONOGIOU I, et al. Prioritized experience replay [C] //*Proceedings of the International Conference on Learning Representations*. San Juan: ACM, IEEE, 2016.

[96] NAIR A, SRINIVASAN P, BLACKWELL S, et al. Massively parallel methods for deep reinforcement learning [C] //*Proceedings of the ICML Workshop on Deep Learning*. Lille: ACM, 2015.

[97] GUO X, SINGH S, LEE H, et al. Deep learning for real-time Atari game play using offline Monte-Carlo tree search planning [C] //*Advances in Neural Information Processing Systems*. Montreal: MIT Press, 2014: 3338 – 3346.

[98] VAN H H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning [C] //*Proceedings of the 30th AAAI Conference on Artificial Intelligence*. Phoenix: AAAI, 2016: 1813 – 1819.

[99] WANG Z, FREITAS N, LANCTOT M. Dueling network architectures for deep reinforcement learning [C] //*Proceedings of the 33rd International Conference on Machine Learning*. New York: ACM, 2016.

[100] OSBAND I, BLUNDELL C, PRITZEL A, et al. Deep exploration via bootstrapped DQN [EB/OL] //*arXiv preprint*. 2016. arXiv:1602.04621.

[101] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning [EB/OL] //*arXiv preprint*. 2016. arXiv:1602.01783[cs.LG]

[102] CUCCU G, LUCIW M, SCHMIDHUBER J, et al. Intrinsically motivated neuroevolution for vision-based reinforcement learning [C] //*Proceedings of the IEEE International Conference on Development and Learning*. Trondheim: IEEE, 2011, 2: 1 – 7.

[103] NARASIMHAN K, KULKARNI T, BARZILAY R. Language understanding for text-based games using deep reinforcement learning [C] //*Proceedings of the Conference on Empirical Methods for Natural Language Processing*. Lisbon: ACL, 2015.

[104] HAUSKNECHT M, STONE P. Deep recurrent Q-learning for partially observable mdps [C] //*Proceedings of the AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents*. Arlington: AAAI, 2015.

[105] SOROKIN I, SELEZNEV A, PAVLOV M, et al. Deep attention recurrent Q-network [C] //*Proceedings of the NIPS Workshop on Deep Reinforcement Learning*. Montreal: MIT Press, 2015.

[106] CAI X, WUNSCH II D C. Computer Go: a grand challenge to AI [M] //*Challenges for Computational Intelligence*. Berlin Heidelberg: Springer, 2007: 443 – 465.

[107] TIAN Y D, ZHU Y. Better computer Go player with neural network and long-term pediction [EB/OL] //*arXiv preprint*. 2016. arXiv:1511.06410v3[cs.LG].

[108] TIAN Yuandong. A simple analysis of AlphaGo [J]. *Acta Automatica Sinica*, 2016, 42(5): 671 – 675.
(田渊栋. 阿法狗围棋系统的简要分析 [J]. 自动化学报, 2016, 42(5): 671 – 675.)

[109] HUANG Shijie. *The strategies for Ko fight of computer Go* [D]. Taiwan: National Taiwan Normal University, 2002: 1 – 57.
(黄士杰. 电脑围棋打劫的策略 [D]. 台湾: 台湾师范大学资讯工程研究所, 2002: 1 – 57.)

[110] GUO Xiaoxiao, LI Cheng, MEI Qiaozhu. Deep learning applied to games [J]. *Acta Automatica Sinica*, 2016, 42(5): 676 – 684.
(郭潇逍, 李程, 梅俏竹. 深度学习在游戏中的应用 [J]. 自动化学报, 2016, 42(5): 676 – 684.)

[111] KOLLER D, MILCH B. Multi-agent influence diagrams for representing and solving games [J]. *Games and Economic Behavior*, 2003, 45(1): 181 – 221.

[112] WOOLDRIDGE M. *An Introduction to Multiagent Systems* [M]. New York: John Wiley & Sons, 2009.

[113] FOERSTER J N, ASSAEL Y M, FREITAS N, et al. Learning to communicate to solve riddles with deep distributed recurrent q-networks [EB/OL] //*arXiv preprint*. 2016. arXiv:1602.02672.

[114] GU S, LILLICRAP T, SUTSKEVER I, et al. Continuous deep Q-learning with model-based acceleration [EB/OL] //*arXiv preprint*. 2016. arXiv:1603.00748.

[115] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning [EB/OL] //*arXiv preprint*. 2016. arXiv:1509.02971v5[cs.LG].

[116] ZHAO D B, ZHU Y H, LV L, et al. Convolutional fitted Q iteration for vision-based control problems [C] //*Proceedings of the International Joint Conference on Neural Networks*. VANCOUVER: IEEE, 2016.

[117] PAN S J, YANG Q. A survey on transfer learning [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(10): 1345 – 1359.

[118] PARISOTTO E, BA J L, SALAKHUTDINOV R. Actor-mimic: deep multitask and transfer reinforcement learning [C] //*Proceedings of the International Joint Conference on Neural Networks*. Vancouver: IEEE, 2016.

[119] LEVINE S, WAGENER N, ABBEEL P. Learning contact-rich manipulation skills with guided policy search [C] //*Proceedings of the IEEE International Conference on Robotics and Automation*. Seattle: IEEE, 2015: 156 – 163.

[120] LEVINE S, PASTOR P, KRIZHEVSKY A, et al. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection [EB/OL] //*arXiv preprint*. 2016. arXiv:1603.02199.

作者简介:

赵冬斌　(1972–), 男, 博士, 研究员, 研究方向为强化学习、自适应动态规划、智能交通、机器人、过程控制等, E-mail: dongbin.zhao@ia.ac.cn;

邵　坤　(1991–), 男, 博士研究生, 研究方向为强化学习、深度学习等, E-mail: shaokun2014@ia.ac.cn;

朱圆恒　(1989–), 男, 博士, 助理研究员, 研究方向为强化学习、自适应动态规划等, E-mail: yuanheng.zhu@ia.ac.cn;

李　栋　(1992–), 男, 博士研究生, 研究方向为强化学习、智能控制等, E-mail: lidong2014@ia.ac.cn;

陈亚冉　(1989–), 女, 博士研究生, 研究方向为计算机视觉、机器学习、强化学习等, E-mail: chenyaran2013@ia.ac.cn;

王海涛　(1990–), 男, 硕士, 研究方向为智能控制与计算智能, E-mail: wanghaitao8118@163.com;

刘德荣　(1963–), 男, 博士, 教授, 研究方向为智能控制理论及应用、复杂工业系统建模与控制、计算智能、智能信息处理等, E-mail: derong@ustb.edu.cn;

周　彤　(1964–), 男, 博士, 教授, 研究方向为系统辨识、鲁棒控制系统分析与设计、大规模系统分析与综合、信号处理等, E-mail: tzhou@mail.tsinghua.edu.cn;

王成红　(1955–), 男, 博士, 研究员, 研究方向为控制理论及应用, 系统可靠性理论等, E-mail: wangch@nsfc.gov.cn.