深度强化学习研究综述

赵星宇[1] 丁世飞[1,2]

(中国矿业大学计算机科学与技术学院 江苏 徐州 221116)[1]

(中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)[2]

摘 要 作为一种崭新的机器学习方法,深度强化学习将深度学习和强化学习技术结合起来,使智能体能够从高维空间感知信息,并根据得到的信息训练模型、做出决策。由于深度强化学习算法具有通用性和有效性,人们对其进行了广泛的研究,并将其运用到了日常生活的各个领域。首先,对深度强化学习研究进行概述,介绍了深度强化学习的基础理论;然后,分别介绍了基于值函数和基于策略的深度强化学习算法,讨论了其应用前景;最后,对相关研究工作做了总结和展望。

关键词 深度强化学习,深度学习,强化学习,人工智能

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.07.001

Research on Deep Reinforcement Learning

ZHAO Xing-yu[1] DING Shi-fei[1,2]

(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China)[1]

(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)[2]

Abstract As a new machine learning method, deep reinforcement learning combines deep learning and reinforcement learning, which makes that the agent can perceive the information from high dimensional space, train model and make decision according to the received information. Deep reinforcement learning has been widely researched and used in various fields of daily life because of its universality and effectiveness. Firstly, an overview of the deep reinforcement learning research was given and the basic theory of deep reinforcement learning was introduced. Then value-based algorithms and policy-based algorithms were introduced. After that, the application prospects of deep reinfercement learning were discussed. Finally, the related researches were summarized and prospected.

Keywords Deep reinforcement learning, Deep learning, Reinforcement learning, Artificial intelligence

1 引言

深度强化学习(Deep Reinforcement Learning,DRL)是近年来人工智能领域最受关注的方向之一,它将深度学习的感知能力和强化学习的决策能力相结合,直接通过高维感知输入的学习来控制 Agent 的行为,为解决复杂系统的感知决策问题提供了思路。

近年来,学术界和企业界纷纷将深度强化学习算法应用于实际。其中,Google 旗下的 DeepMind 公司将深度强化学习算法应用于游戏中,分别在视频游戏[1]与机器博弈[2]领域取得丰硕的成果。2016 年与 2017 年,DeepMind 公司研制出的围棋博弈系统 AlphaGo 分别与人类顶尖高手李世石和柯洁对战,并都以较大优势取得了胜利,展现出了极强的实战能力,极大地震撼了社会各界。

除视频游戏和机器博弈之外,深度强化学习算法还被人们应用于机器人[3-6]、机器翻译[7]、视频预测[8]、文本生成[9-10]、控制优化、文本游戏[11]、自动驾驶[12]、目标定位[13]等多个领域中,展现出了强大的学习能力和适应能力。因此,深入研究和分析深度强化学习算法,对于推进人工智能方法的发展及拓展其应用具有重要的意义[14]。

文中第 2 节介绍深度学习和强化学习的基础理论及相关的模型和算法;第 3 节对深度强化学习研究的基本情况进行概述,并介绍和分析深度强化学习的主要算法;第 4 节对深度强化学习算法的实际应用进行探讨;最后对深度强化学习的研究进行总结与展望。

2 基础理论

2.1 深度学习

深度学习(Deep Learning,DL)的概念起源于对人工神经网络的研究。2006 年,Hinton 及其学生 Salakhutdinov 提出

到稿日期:2017-06-12 返修日期:2017-08-20 本文受国家自然科学基金(61379101,61672522),国家重点基础研究发展计划(2013CB329502)资助。

赵星宇(1994-),男,硕士生,CCF 会员,主要研究方向为强化学习和深度学习;**丁世飞**(1963-),男,教授,博士生导师,CCF 高级会员,主要研究方向为智能信息处理、人工智能与模式识别、机器学习与数据挖掘、粗糙集与软计算、粒度计算等,E-mail:dingsf@cumt.edu.cn(通信作者)。

了深度网络的概念，开启了深度学习研究的热潮[15]。深度学习的本质是计算观测数据的分层特征或表示，其中高层特征或因子由低层得到[16]。深度学习通过学习一种深层的非线性网络结构，实现复杂函数的逼近，学习数据集的本质特征[17]。深度学习的模型主要有深度信念网络(Deep Belief Network，DBN)、卷积神经网络(Convolutional Neural Network，CNN)、循环神经网络(Recurrent Neural Network，RNN)等。

DBN 的训练由预训练和反向微调两部分组成。预训练过程是一种无监督的贪心逐层训练方法，首先训练最底层的受限玻尔兹曼机(Restricted Boltzmann Machine，RBM)，在其后的训练过程中将下一层的 RBM 作为上一层 RBM 的输入，依次逐层训练所有的 RBM，即可获得深度置信网络的初始权值[18]。完成预训练后，DBN 再利用带标签的数据，通过反向传播算法调整网络的参数，完成整个训练过程。其结构如图 1 所示。

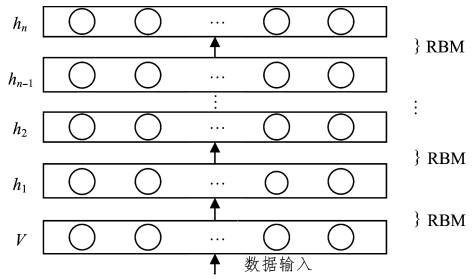

图 1 深度信念网络的结构

Fig. 1 Structure of deep belief network

CNN 属于区分性训练算法。典型的卷积神经网络主要由输入层、卷积层、池化层、全连接层和输出层组成。输入层用于接收数据，输入的数据通常为原始图像；卷积层每个神经元的输入与前一层的一部分局部神经元相连，并提取该局部的特征；下采样层依据一定的下采样规则对特征图进行采样；输出层神经元的类型可以根据实际应用设计。

RNN 的神经元间的连接构成有向图，一次处理一个输入序列元素，每个节点同时包含序列元素在过去时刻的历史信息。RNN 常见的结构有长短期记忆人工神经网络(Long-Short Term Memory，LSTM)[19]、门循环单元人工神经网络(Gated Recurrent Units，GRU)[20]等。

2.2 强化学习

作为一种重要的机器学习方法，强化学习(Reinforcement Learning，RL)采用了人类和动物学习中的“尝试与失败”机制，强调在与环境的交互中学习，利用评价性的反馈信号实现决策的优化[21]。由于强化学习在学习过程中不需要给定各种状态下的教师信号，因此其在求解复杂的优化决策问题方面有着广泛的应用前景。

强化学习的过程是一个试探与评价的过程。在强化学习中，Agent 在环境 s 下选择并执行一个动作 a，环境接受动作后变为 s'，并把一个奖赏信号 r 反馈给 Agent，Agent 再根据奖赏信号选择后续动作。强化学习的基本框架如图 2 所示。

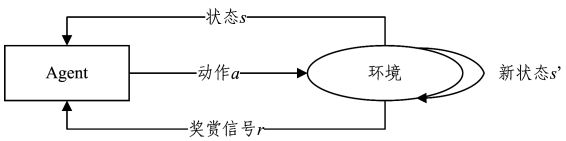

图 2 强化学习的基本框架

Fig. 2 Basic framework of reinforcement learning

强化学习可以分为基于值函数的强化学习和基于策略的强化学习。在基于值函数的强化学习中，最常用的学习算法为 Q 学习算法[22]，其迭代公式如下：

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha[r_{t+1}+\gamma\max_a Q(s_{t+1},a)-Q(s_t,a_t)] \tag{1}$$

其中，$Q(s_t,a_t)$为 t 时刻的状态-动作值。r 为奖赏值，γ 为折扣因子。

在基于策略的强化学习中，最常用的是策略梯度算法。其参数更新的基本形式如下[23]：

$$\theta\leftarrow\theta+\alpha\nabla_\theta\log\pi_\theta(s_t,a_t;\theta)v_t \tag{2}$$

其中，θ 为动作选择策略的参数，α 为学习率，$\pi_\theta(s_t,a_t)$为 t 时刻的动作选择策略，而 v_t 为该策略的评价值。

结合基于值函数的算法和基于策略函数的算法，可以得到一种新的强化学习算法——行动者-评论家(Actor-Critic，AC)算法[24]。在行动者-评论家算法中，行动者(Actor)基于策略选择函数，根据状态选择策略；而评论家(Critic)对 Actor 当前的策略进行评价，并指导 Actor 进行策略的改进。Actor-Critic 算法能够结合多种不同的值函数方法和直接策略选择方法，具有比传统的基于策略函数的强化学习算法更快的收敛速度。

3 深度强化学习的主要算法

3.1 深度强化学习算法的概述

深度强化学习是将深度学习与强化学习相结合的一种全新算法，实现了从感知到动作的端到端的学习。输入图像、文本、音频、视频等，通过 DRL 构建的深度神经网络的处理，可以实现直接输出动作，无须手工干预。

在深度 Q 网络被提出以前，人们就对各种深度学习模型进行了许多研究，其中有不少学者将深度学习与强化学习相结合并应用到实际中[25-26]。2013 年，DeepMind 公司的 Mnih 等提出了开创性的深度 Q 网络(Deep Q-network，DQN)。通过 DQN，Agent 仅通过从图像中获取信息就能学会玩视频游戏[27]。

DQN 被提出以后，深度强化学习得到了广泛的关注，人们开始对其进行更深层次的研究，并将其应用到实际应用中[3-13]。近年来，深度强化学习的成果层出不穷，最具代表性的有 DeepMind 公司于 2015 年和 2016 年连续在 *Nature* 上发表的关于深度强化学习的论文[1-2]，这标志着深度强化学习的研究和应用进入了一个新的阶段。

当今，深度强化学习的研究正处于快速发展的阶段，每年都有很多新算法被提出。总体而言，为人们所广泛认可的深度强化学习算法的研究方向主要包括 DQN 及其相关改进、

基于策略的深度强化学习算法(如 DDPG,DPPO 等)以及一些其他的研究工作(如 A3C,UNREAL 等)。表 1 列出了深度强化学习的主要已有算法。

表 1　深度强化学习的主要算法

Table 1　Main algorithms of deep reinforcement learning

算法名称	提出年份	算法类型
深度 Q 网络(DQN)	2013	基于值函数
改进的深度 Q 网络(NatureDQN)	2015	基于值函数
深度确定性策略梯度算法(DDPG)	2015	基于策略
信赖域策略优化算法(TRPO)	2015	基于策略
异步优势行动者评论家算法(A3C)	2016	基于策略
无监督辅助强化学习(UNREAL)	2016	基于策略
分布式近似策略优化算法(DPPO)	2017	基于策略

3.2　DQN 及其相关改进

近几年的深度强化学习算法的研究主要围绕 DQN 的相关研究和改进展开。DQN 将卷积神经网络与 Q 学习相结合,并引入经验回放机制,使得计算机能够直接根据高维感知输入来学习控制策略。2013 年,Mnih 等利用 DQN 训练计算机,成功使得计算机在 7 款 Atari 游戏中的 3 款上超过了人类专家的水平[27]。

深度 Q 网络采用带有参数 θ 的 Q 值函数 $Q(s,a;\theta)$来逼近值函数。在环境 ε 下,当迭代次数为 i 时,损失函数 $L_i(\theta_i)$ 的定义表示为:

$$L_i(\theta_i)=E_{s,a\sim\rho(\cdot)}[(y_i-Q(s,a;\theta_i))^2] \tag{3}$$

其中,$\rho(\cdot)$表示在给定环境下 s 选择动作 a 的概率分布。y_i 表示第 i 次迭代 Q 值函数的目标,其定义表示为:

$$y_i=E_{s'\sim\varepsilon}[r+\gamma\max_{a'}Q(s',a';\theta_{i-1}\mid s,a)] \tag{4}$$

其中,r 为环境反馈给 Agent 的奖赏值,γ 为折扣因子。由式(4)可知,学习的目标取决于网络权值。而网络权值的更新公式为:

$$\nabla_{\theta_i}L_i(\theta_i)=E_{s,a\sim\rho(\cdot);s'\sim\varepsilon}[(r+\gamma\max_{a'}Q(s',a';\theta_{i-1})-Q(s,a;\theta_i))\nabla Q(s,a;\theta_i)] \tag{5}$$

人们围绕 DQN 做了许多研究和改进工作。Mnih 等将迭代式更新引入 DQN 中,降低了目标计算与当前值的相关性[1];Hasselt 等将双 Q 学习应用于 DQN 中,提出了双 DQN(DoubleDQN)算法,有效地避免了过于乐观的值估计[28];Wang 等提出了决斗模型(Dueling DQN),将状态值和动作优势值区分开,使得网络架构和 RL 算法能够更好地结合在一起[29];Schaul 等使用 DQN 对经验的优先次序进行处理,使用经验优先回放(Prioritized Experience Replay)技术实现了高效的学习[30];此外,Osband[31],Hasselt[32],Lakshminarayanan[33],Munos[34],Vincent[35]等也分别从不同角度对 DQN 进行了研究,并提出了相关的改进方法。

3.3　基于策略的深度强化学习算法

尽管基于 Q 学习算法的 DQN 已在许多领域取得了不错的效果,但是在面对连续动作空间时,输出离散状态-动作值的 DQN 显得十分无力。此时,人们将策略梯度方法引入深度强化学习中。基于策略梯度方法,Lillicrap 等于 2015 年提出了深度确定性策略梯度(Deep Deterministic Policy Gradient,DDPG)算法[36]。DDPG 是深度强化学习应用于连续控制强化学习领域的一种重要算法,将确定性策略梯度算法(DPG)[37]与 Actor-Critic 框架相结合,提出了一个任务无关的模型,并使用相同的参数解决了 20 多个连续控制的仿真问题。DDPG 采取经验回放机制,通过目标网络的参数不断与原网络的参数加权平均进行训练,以避免振荡。

除 DDPG 算法之外,另一个著名的基于策略的算法是 Heess 等于 2017 年提出的分布式近似策略优化算法(Distributed Proximal Policy Optimization,DPPO)[38-39]。DPPO 是信赖域策略优化算法(Trust Region Policy Optimization,TRPO)[40]的改良版本,适用于许多领域,是一种通用的优化思想。DPPO 算法引入了旧策略和更新之后的策略所预测的概率分布之间的 KL 差异,使得更新前后的策略不会相差太大,避免了参数训练的震荡,并据此来控制参数更新的过程。此外,Zhang[41],Duan[42],Balduzzi[43],Heess[44]等也针对策略梯度方法在深度强化学习中的应用进行了研究,并取得了一定的成果。

3.4　其他相关研究工作

除了关于 DQN 和策略梯度方法的研究外,人们对深度强化学习的算法及模型架构还做了许多相关研究。其中,比较著名的包括异步优势行动者评论家(A3C)算法[45]。A3C 算法是由 Mnih 等于 2016 年提出的,该算法是深度强化学习算法的集大成者,融合了之前几乎所有的深度强化学习算法。A3C 算法采取了不同的 actor-learners 并行探索环境的方法,每个 actor-learner 独自探索并在线更新全局策略参数。利用这种方法,可以不再依赖经验池来存储历史经验,极大地缩短了训练的时间。

此外,人们还从其他角度对深度强化学习的算法及模型架构进行了研究。Jaderberg 等提出了无监督辅助强化学习(Unsupervised Reinforcement and Auxiliary Learning,UNREAL)算法,通过训练多个辅助任务来改进算法,极大地提高了算法的性能[46];Finn 等[47]对逆向深度强化学习进行了研究;Oh 等[48]提出了一种基于记忆的深度强化学习模型。此外,Kulkarni[49],Houthooft[50],Fernández[51],Bellemare[52],Schaul[53]等也从不同角度对深度强化学习的算法及模型架构进行了研究,并取得了引人关注的成果。

4　深度强化学习算法的实际应用

4.1　计算机博弈

计算机博弈是人工智能领域最具挑战性的研究方向之一,其研究为人工智能带来了很多重要的方法和理论。2016 年 3 月,DeepMind 公司研制出的围棋博弈系统 AlphaGo 在与世界围棋冠军李世石的对战当中,以 4∶1的大比分取胜[2];2017 年 1 月,AlphaGo 的升级版 Master 在与世界顶尖围棋大师的对战中全部取得了胜利。但是,破解完全信息博弈游戏对于完全破解计算机博弈而言是远远不够的。相比于完全信息博弈游戏,不完全信息博弈游戏具有更多的未知性,给研究者带来的挑战也更加巨大。尽管人们已对不完全信息博弈游戏(如德州扑克游戏)进行研究并取得了许多成果,但目前仍

然不能使计算机在较为复杂的环境下战胜人类。对于人工智能的研究者来说,对不完全信息博弈游戏的研究仍是一个充满挑战性的方向。

4.2 视频游戏

利用算法训练计算机,使计算机自主学会玩游戏并达到较高水平,是深度强化学习在视频游戏中的主要应用。2015年,DeepMind公司利用Atari平台上的49款游戏对深度Q网络进行了测试,发现通过DQN的训练,计算机能够在其中的29款游戏中取得超过人类职业玩家75%的得分[1]。除了Atari平台,人们也基于其他视频游戏对深度强化学习进行了研究。Lample[54],Kempka[55],Oh[48]等研究了深度强化学习在Doom和Minecraft游戏中的应用,而DeepMind则将它们的下一个挑战设定为Starcraft 2游戏[56]。目前,视频游戏是深度强化学习算法最好的试验田之一,人们对视频游戏的研究在近期会有非常快速的发展。

4.3 个人助理机器人

如今,人们越来越希望能够方便地通过移动端搜索到自己关注的内容,并使移动设备通过文本、语音等媒介将相关信息告知用户。随着移动端对话范式的兴起,通过深度强化学习算法实现助理机器人距现实越来越近。助理机器人通过搜索功能来搜索相关信息,通过过滤功能排除掉无用信息,通过社交功能将信息告知给用户。通过深度强化学习和庞大的数据集可以训练助理机器人,使其能够实现这些功能。个人助理机器人可以深入到人们的生活中,这是一个非常有前景的人工智能研究领域。

4.4 家用机器人

随着时代的发展,人们逐渐使用机器人来代替人类进行各种繁琐或危险的工作。其中,家用机器人成为了大众最钟爱的选择。相比于工业机器人,家用机器人面对的环境更为复杂,不但需要接受人类的语音指令、辨别人类的指令,而且需要在不同的环境下就人类的指令完成相应的工作。Zhu等提出了一种家用机器人的实现方法,将知识迁移技术与深度强化学习相结合,首先构建一个非常好的仿真环境,并在该环境中对机器人进行训练,训练完毕后,再将其移植到现实世界中[57]。该方法在高质量虚拟环境中训练算法,再将训练好的模型迁移到现实应用中,为家用机器人提供了一个很好的框架。尽管还面临着不小的挑战,但是家用机器人投入实际应用已指日可待。

4.5 人机对话

人机对话的一个典型模型是Sutskever于2014年提出的SEQ2SEQ。该模型使用了两个LSTM网络,首次将深度神经网络模型运用于翻译与智能问答这一类序列型任务中,利用给定的对话历史生成了一个最大可能性的响应[58]。但该算法倾向于生成高度一般的回应,并且容易陷入到死循环中。针对这些问题,Li等[59]对该模型进行了改进,将深度强化学习的方法引入其中,使用策略梯度方法来优化模型,使得生成的有效对话数量和多样性均得到了一定的提高。此外,许多企业或个人也均致力于人机对话领域的研究和开发,该领域在近几年势必有着快速的发展。

结束语 文中对深度强化学习进行了全面分析,包括深度强化学习的研究概况、主要算法以及实际应用等。在许多领域,深度强化学习都表现出了巨大的潜力,研究成果不断涌现,各种算法层出不穷。如今,基于改进深度Q网络的深度强化学习模型已经较为完善,策略梯度方法得到了广泛应用,而机器学习领域的其他算法也被不断地应用到深度强化学习算法的相关模型中。但作为机器学习的一个新兴领域,深度强化学习现仍处于发展阶段,仍有很多问题值得进一步深入研究。

1)将迁移学习运用到深度强化学习中,是提高深度强化学习算法通用性的一条极佳的途径。Zhu和Parisotto等对深度强化学习中的知识迁移做了相关的研究,并取得了一定的成果[57,60]。但是,对于复杂度较高的场景而言,知识迁移还是一项充满挑战性的工作。如何更好地在复杂场景中将知识迁移技术运用到深度强化学习中,是一个重要的课题。

2)深度强化学习是一种模仿人类行为思路的方法,它使得机器能够与人一样从高维感知输入进行学习。但是,由于缺乏使其对应到人脑机理中的生理学基础,使其更深层次的研究受到了一定的限制。将深度强化学习算法对应到人脑机理的生理学研究,是一项难度极高的工作,也是一项重要的工作。

3)将更多深度学习模型运用到深度强化学习算法中,也是一个重要的课题。目前,人们对基于卷积神经网络、循环神经网络等深度学习模型的深度强化学习算法的研究较多,但将其他一些深度学习模型(如深度信念网络、深度稀疏编码等)运用到算法模型中的研究还很少[61]。随着深度强化学习研究的不断发展,将会有更多的深度学习模型被用于该领域算法的研究中。

4)深度强化学习不但实现了获得高维数据的方法,还实现了根据数据训练模型的具体过程。通用人工智能(Artificial General Intelligence,AGI)是指机器能够在没有编码特定领域知识的情况下解决不同种类的问题,做出类似人类的判断与决策。如何在比较复杂的环境、很少的样本及稀缺的外界激励下根据所掌握的知识做出正确的决策,是实现通用人工智能的重要问题,也是深度强化学习未来的重要研究方向之一。

近年来,DeepMind和OpenAI等公司将深度强化学习技术应用到游戏、医疗、机器人等领域,对这些领域的发展起到了重要的推动作用。随着科学技术水平的逐步提高,深度强化学习的相关研究势必对人们的生活产生越来越大的影响,为人类的进步作出更大的贡献。

参 考 文 献

[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] LEVINE S，PASTOR P，KRIZHEVSKY A，et al. Learning Hand-Eye Coordination for Robotic Grasping with Large-Scale Data Collection[C]// International Symposium on Experimental Robotics. Springer，Cham，2016：173-184.

[4] ZHANG M，MCCARTHY Z，FINN C，et al. Learning deep neural network policies with continuous memory states[C]// Proceedings of the International Conference on Robotics and Automation. Stockholm，Sweden，2016：520-527.

[5] LEVINE S，FINN C，DARRELL T，et al. End-to-end training of deep visuomotor policies[J]. Journal of Machine Learning Research，2016，17(39)：1-40.

[6] LENZ I，KNEPPER R，SAXENA A. Deepmpc：learning deep latent features for model predictive control[C]// Proceedings of the Robotics Scienceand Systems. Rome，Italy，2015：201-209.

[7] SATIJA H，PINEAU J. Simultaneous machine translation using deep reinforcement learning[C]// Proceedings of the Workshops of International Conference on Machine Learning. New York，USA，2016：110-119.

[8] OH J，GUO X，LEE H，et al. Action-conditional video prediction using deep networks in atari games[C]// Advances in Neural Information Processing Systems. 2015：2863-2871.

[9] GUO H. Generating text with deep reinforcement learning [C]// Proceedings of the Workshops of Advances in Neural Information Processing Systems. Montreal，Canada，2015：1-9.

[10] LI J，MONROE W，RITTER A，et al. Deep reinforcement learning for dialogue generation[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Austin，USA，2016：1192-1202.

[11] NARASIMHAN K，KULKARNI T，BARZILAY R. Language Understanding for Text-based Games Using Deep Reinforcement Learning[J]. Computer Science，2015，40(4)：1-5.

[12] SALLAB A，ABDOU M，PEROT E，et al. Deep reinforcement learning framework for autonomous driving[J]. Electronic Imaging，2017，2017(19)：70-76.

[13] CAICEDO J，LAZEBNIK S. Active Object Localization with Deep Reinforcement Learning[C]// IEEE International Conference on Computer Vision. IEEE，2015：2488-2496.

[14] ZHAO D B，SHAO K，ZHU Y H，et al. Review of deep reinforcement learning and discussions on the development of computer Go[J]. Control Theory and Applications，2016，33(6)：701-717. (in Chinese)
赵冬斌，邵坤，朱圆恒，等. 深度强化学习综述：兼论计算机围棋的发展[J]. 控制理论与应用，2016，33(6)：701-717.

[15] HINTON G，SALAKHUTDINOV R. Reducing the Dimensionality of Data with Neural Networks[J]. Science，2006，313(5786)：504-507.

[16] DENG L，YU D. Deep learning：methods and applications[J]. Foundations and Trends in Signal Processing，2014，7(3/4)：197-387.

[17] BENGIO Y，LECUN Y. Scaling learning algorithms towards AI [J]. Large-scale Kernel Machines，2007，34(5)：1-41.

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

[19] HOCHREITER S，SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation，1997，9(8)：1735-1780.

[20] CHO K，VAN MERRIËNBOER B，GULCE-HRE C，et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha：Association for Computational Linguistics，2014：1724-1734.

[21] GAO Y，CHEN S F，LU X. Research on Reinfocerment Learning Technology：A Review[J]. Acta Automatica Sinica，2004，30(1)：86-100. (in Chinese)
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.

[22] WATKINS C. Learning from delayed rewards[D]. Cambridge：King's College，1989.

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

[24] KONDA V，TSITSIKLIS J. Actor-critic algorithms[C]// Advances in Neural Information Processing Systems. 2000：1008-1014.

[25] LANGE S，RIEDMILLER M. Deep auto-encoder neural networks in reinforcement learning[C]// Neural Networks (IJCNN)，The 2010 International Joint Conference on Computational Science and Optimization. IEEE，2010：1-8.

[26] LANGE S，RIEDMILLER M，VOIGTLÄNDER A. Autonomous reinforcement learning on raw visual input data in a real world application[C]// International Joint Conference on Neural Networks. IEEE，2012：1-8.

[27] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Playing atari with deep reinforcement learning[C]// Proceedings of Workshops at the 26th Neural Information Processing Systems 2013. Lake Tahoe，USA，2013：201-220.

[28] HASSELT H，GUEZ A，SILVER D. Deep Reinforcement Learning with Double Q-Learning[C]// AAAI. 2016：2094-2100.

[29] WANG Z，FREITAS N，LANCTOT M. Dueling network architectures for deep reinforcement learning[C]// Proceedings of the International Conference on Machine Learning. New York，USA，2016：1995-2003.

[30] SCHAUL T，QUAN J，ANTONOGLOU I，et al. Prioritized experience replay[C]// Proceedings of the 4th International Conference on Learning Representations. San Juan，Puerto Rico，2016：322-355.

[31] OSBAND I，BLUNDELL C，PRITZEL A，et al. Deep exploration via bootstrapped DQN[C]// Advances in Neural Information Processing Systems. 2016：4026-4034.

[32] HASSELT H，GUEZ A，HESSEL M，et al. Learning functions across many orders of magnitudes[C]// Proceedings of the Advances in Neural Information Processing Systems. Barcelona，

Spain,2016:80-99.

[33] LAKSHMINARAYANAN A,SHARMA S,RAVINDRAN B. Dynamic frame skip deep q network[C]// Proceedings of the Workshops at the International Joint Conference on Artificial Intelligence. New York,USA,2016.

[34] MUNOS R,STEPLETON T,HARUTYUNYAN A,et al. Safe and efficient off-policy reinforcement learning[C]// Advances in Neural Information Processing Systems. 2016:1054-1062.

[35] FRANÇOIS-LAVET V,FONTENEAU R,ERNST D. How to discount deep reinforcement learning: towards new dynamic strategies[C]// Proceedings of the Workshops at the Advances in Neural Information Processing Systems. Montreal, Canada, 2015:107-1160.

[36] LILLICRAP T,HUNT J,PRITZEL A,et al. Continuous control with deep reinforcement learning[J/OL]. https://arxiv.org/abs. 1509.02971.

[37] SILVER D,LEVER G,HEESS N,et al. Deterministic policy gradient algorithms[C]// Proceedings of the 31st International Conference on Machine Learning. 2014:387-395.

[38] SCHULMAN J,WOLSKI F,DHARIWAL P,et al. Proximal Policy Optimization Algorithms[J/OL]. https://arxiv.org/abs/1707.06347.

[39] HEESS N,DHRUVA T,SRIRAM S,et al. Emergence of Locomotion Behaviours in Rich Environments[J/OL]. https://arxiv.org/abs/1707.02286.

[40] SCHULMAN J,LEVINE S,MORITZ P,et al. Trust Region Policy Optimization[C]// International Conference on Machine Learning. Lille: International Machine Learning Society, 2015: 1889-1897.

[41] ZHANG T,KAHN G,LEVINE S,et al. Learning deep control policies for autonomous aerial vehicles with mpc-guided policy search[C]// 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE,2016:528-535.

[42] DUAN Y,CHEN X,HOUTHOOFT R,et al. Benchmarking deep reinforcement learning for continuous control[C]// International Conference on Machine Learning. 2016:1329-1338.

[43] BALDUZZI D,GHIFARY M. Compatible Value Gradients for Reinforcement Learning of Continuous Deep Policies[J/OL]. https://arxiv.org/abs/1509.03005.

[44] HEESS N,WAYNE G,SILVER D,et al. Learning continuous control policies by stochastic value gradients[C]// Advances in Neural Information Processing Systems. 2015:2944-2952.

[45] MNIH V,BADIA A,MIRZA M,et al. Asynchronous methods for deep reinforcement learning[C]// International Conference on Machine Learning. 2016:1928-1937.

[46] JADERBERG M,MNIH V,CZARNECKI W,et al. Reinforcement learning with unsupervised auxiliary tasks [J/OL]. https://arxiv.org/abs/1611.05397.

[47] FINN C,LEVINE S,ABBEEL P. Guided cost learning:Deep inverse optimal control via policy optimization[C]// International Conference on Machine Learning. 2016:49-58.

[48] OH J,CHOCKALINGAM V,SINGH S,et al. Control of memory,active perception,and action in Minecraft[C]// Proceedings of the International Conference on Machine Learning. New York,USA,2016:2790-2799.

[49] KULKARNI T,NARASIMHAN K,SAEEDI A,et al. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation[C]// Advances in Neural Information Processing Systems. 2016:3675-3683.

[50] HOUTHOOFT R,CHEN X,DUAN Y,et al. VIME:Variational information maximizing exploration[C]// Advances in Neural Information Processing Systems. 2016:1109-1117.

[51] FERNÁNDEZ F,VELOSO M. Probabilistic policy reuse in a reinforcement learning agent[C]// Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems. Istanbul,Turkey,2015:720-727.

[52] BELLEMARE M,SRINIVASAN S,OSTROVSKI G,et al. Unifying count-based exploration and intrinsic motivation[C]// Proceedings of the Conference on Neural Information Processing Systems. Barcelona,Spain,2016:1471-1479.

[53] SCHAUL T,HORGAN D,GREGOR K,et al. Universal value function approximators[C]// Proceedings of the 32nd International Conference on Machine Learning. Lugano, Switzerland, 2015:1312-1320.

[54] LAMPLE G,CHAPLOT D. Playing FPS Games with Deep Reinforcement Learning[C]// AAAI. 2017:2140-2146.

[55] KEMPKA M,WYDMUCH M,RUNC G,et al. Vizdoom:A doom-based ai research platform for visual reinforcement learning[C]// 2016 IEEE Conference on Computational Intelligence and Games (CIG). IEEE,2016:1-8.

[56] VINYALS O,EWALDS T,BARTUNOV S,et al. StarCraft II: A New Challenge for Reinforcement Learning[J/OL]. https://arxiv.org/abs/1708.04782.

[57] ZHU Y,MOTTAGHI R,KOLVE E,et al. Target-driven visual navigation in indoor scenes using deep reinforcement learning [C]// 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE,2017:3357-3364.

[58] SUTSKEVER I,VINYALS O,LE Q. Sequence to sequence learning with neural networks[C]// Advances in Neural Information Processing Systems. 2014:3104-3112.

[59] LI J,MONROE W,RITTER A,et al. Deep reinforcement learning for dialogue generation[J/OL]. https://arxiv.org/abs/1707.06347.

[60] PARISOTTO E,BA J,SALAKHUTDINOV R. Actor-mimic: deep multitaskand transfer reinforcement learning[C]// Proceedings of the International Conference on Learning Representations. San Juan,Puerto Rico,2016:156-171.

[61] CHEN X G,YU Y. Reinforcement Learning and Its Application to the Game of Go[J]. Acta Automatica Sinica,2016,42(5): 685-695. (in Chinese)
陈兴国,俞扬. 强化学习及其在电脑围棋中的应用[J]. 自动化学报,2016,42(5):685-695.