第53卷第5期 南京航空航天大学学报 Vol. 53 No. 5
2021年10月 Journal of Nanjing University of Aeronautics & Astronautics Oct. 2021

DOI:10.16356/j.1005-2615.2021.05.001

元强化学习综述

谭晓阳[1,2], 张 哲[1,2]

(1.南京航空航天大学计算机科学与技术学院/人工智能学院,南京 211106;
2.模式分析与机器智能工业和信息化部重点实验室,南京 211106)

摘要: 元强化学习是指自动从一组相关任务中学习强化学习所需归纳偏置的相关理论和方法,对于提高强化学习算法在困难场景下的样本效率和泛化能力具有重要用途。本文提出一种新的元强化学习框架,指出设计和分析一个元强化学习算法需要同时考虑学习经验(相关任务)、归纳偏置及学习目标3个独立因素及这3个因素之间的依赖关系。在此基础上对该领域的研究现状进行了分析和总结,特别对近年来元强化学习若干文献进行了分析和归类,并详细阐述了几种代表性算法的原理及各自特点。本文还对元强化学习常用的实验环境和性能评价方法进行了介绍,对该领域的不足和未来的发展方向进行了讨论和分析。

关键词: 元强化学习;样本效率;泛化性;归纳偏置

中图分类号: TP181 **文献标志码:** A **文章编号:** 1005-2615(2021)05-0653-11

Review on Meta Reinforcement Learning

TAN Xiaoyang[1,2], *ZHANG Zhe*[1,2]

(1. College of Computer Science and Technology/College of Artificial Intelligence, Nanjing University of Aeronautics & Astronautics, Nanjing 211106, China; 2. MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China)

Abstract: Meta reinforcement learning (Meta-RL) aims at automatically learning induction bias for a new reinforcement learning task from a set of different but related tasks. It plays an important role in improving the sample efficiency and generalization of reinforcement learning algorithm in difficult scenarios. This paper first introduces a framework in which three key components of Meta-RL are identified, i.e., learning experience (related tasks), inductive bias and learning objective. Based on this, current research progress in this field is analyzed and reviewed, and the principles and characteristics of several representative algorithms are described. The paper also gives a detailed account of commonly used benchmark environments and performance evaluation methods for meta-RL. The limitation of current research and potential future development directions are also discussed.

Key words: meta reinforcement learning; sample efficiency; generalization; inductive bias

强化学习作为机器学习领域的一个重要分支,是解决序贯决策问题的重要范式。该范式从一个完整的、交互式的和目标导向的智能体出发,通过感知环境变化获得反馈信号,并以此选择合适的动作与环境交互以实现既定目标。深度强化学习(Deep reinforcement learning, DRL)将深度学习技

基金项目: 国家自然科学基金(61976115,61732006)资助项目;全军共用信息系统装备预研基金(315025305)资助项目;南京航空航天大学"人工智能+"研究基金(NZ2020012,56XZA18009)资助项目。

收稿日期: 2020-10-11;**修订日期:** 2021-03-10

作者简介: 谭晓阳,男,教授,博士生导师,研究方向:深度强化学习,多智能体系统。

通信作者: 张哲,E-mail:zhangzhe@nuaa.edu.cn。

引用格式: 谭晓阳,张哲. 元强化学习综述[J]. 南京航空航天大学学报,2021,53(5):653-663. TAN Xiaoyang, ZHANG Zhe. Review on meta reinforcement learning[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2021, 53(5):653-663.

术引入到强化学习领域中,使得构建高度复杂的参数化非线性目标函数(如值函数、策略函数或转换模型)成为可能,从而极大地促进了强化学习领域的进步。从战胜人类顶尖棋手的AlphaGo[1]开始,深度强化学习快速地吸引了大量研究者的关注,并广泛地应用在自动驾驶、推荐系统、机器人控制以及智慧医疗等相关领域。

深度强化学习在某些方面取得了成功,但也存在着缺陷与不足。一个重要的方面是基于深度强化学习的方法需要大量的数据样本提供给模型进行训练。例如,利用深度强化学习模型的智能体在Atari[2]游戏环境下需要1 800万帧的训练数据才能超过人类玩家水平。在医疗影像、机器人控制和推荐系统等现实场景中,收集到如此庞大的数据集几乎是不可能的,在另外一些场合,则存在收集数据的代价过大的问题,以上因素都限制了深度强化学习在这些领域内发挥更大的作用。

为了解决深度强化学习样本复杂度高的问题,研究者们从不同的角度出发,提出了一系列的方法和机制来提高深度强化学习的样本效率。常见的方法有异策略学习[3-7]、基于模型的强化学习[8-9]、迁移强化学习[10-11]、多任务强化学习[12-14]和连续强化学习[15-17]等。异策略和基于模型的方法都是从数据利用的角度出发,前者提高历史数据的利用率,后者则是利用仿真模型产生数据提供给智能体进行训练。迁移强化学习、多任务强化学习以及连续强化学习都是跨任务的强化学习。迁移强化学习是将单个或多个源任务上的学习经验通过知识迁移的方式运用到目标任务的学习中,从而促进在目标任务下的学习。多任务强化学习则同时学习多个相关任务,利用这些任务之间的共性结构,使得联合学习能够进一步提升智能体性能表现并提高样本效率。连续强化学习假设训练任务依次输入,需要智能体在整个生命周期上连续提高其学习技能,通过不断迁移之前学习的知识到当前任务上以促进学习。

与上述方法密切相关的是元强化学习,它是一类自动从相关任务中学习强化学习所需归纳偏置的相关理论和方法的统称,对于提高强化学习算法在困难场景下的样本效率和泛化能力具有重要用途。本文通过对近几年元强化学习的研究进行总结和归纳,回顾并梳理了该领域的发展脉络。

1 元强化学习框架

元学习是一种新的机器学习范式,从模型在相关任务上的训练过程中获取学习经验,并结合这些经验知识提升模型在未知学习任务上的表现,换言之,就是一个“学会学习”的过程。这一概念最早起源于心理学和脑神经科学[18],20世纪90年代,研究者们将其引入机器学习领域[19-22],提出了“自参考”“快慢知识”和“突触学习规则”等元学习相关的早期概念,实质上都是模仿人类利用过往经验在新的场景下快速学习的能力,因此元学习也可看作是一种知识迁移的工具。

把元学习运用到强化学习领域的方法统称元强化学习(Meta reinforcement learning, Meta RL),其目的是自动从一组相关任务中学习有益的归纳偏置,为后续强化学习服务。由于目前深度强化学习算法在现实应用中普遍面临样本效率低、学习速度慢和泛化能力低(每个新任务需重新训练)的性能瓶颈,元强化学习算法通过试图自动为深度强化学习任务提供有益偏置的方法,为上述问题的解决指出了一条有希望的途径,因而日益成为强化学习领域的研究热点,涌现出大量文献。

为了对这些文献进行总结和分类,以加深对该领域研究现状的了解,本文提出了如图1所示的元强化学习框架。在该框架中,一个元强化学习算法主要包含3个成分:学习经验(相关任务)、归纳偏置及学习目标:算法首先在相关任务上进行元学习,自动抽取这些任务背后的共同模式,即归纳偏置,然后将其应用于具体的强化学习算法。简而言之,元强化学习的核心在于如何有效地发现、表示和重用相关任务背后共同的知识,使得新的强化学习任务不必每次从头开始学习。

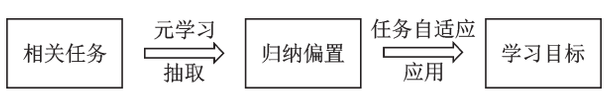

图1 元强化学习流程框架图

Fig.1 Framework of meta reinforcement learning process

值得注意的是以上3个模块之间并不是彼此独立,而是彼此依赖的:一方面,强化学习算法的学习目标决定了其需要何种归纳偏置,而这进一步决定了应该从哪些相关任务中去获取这些归纳偏置;另一方面,相关任务决定了能够从中获取什么样的归纳偏置,而后者在一定程度决定了能够适用何种强化学习算法。因此,对于一个元强化学习问题,需全面进行考虑上述3个方面。

元强化学习算法一般涉及两类学习器:一类是基学习器;另一类是元学习器。其中基学习器负责对单个任务的学习,并向元学习器反馈归纳偏置的效果;而元学习器在基学习的反馈基础上,从相关

任务中搜索最优的有用模式。因此整个学习过程包括两个阶段:一是基学习器在某种归纳偏置的基础上进行个体学习的阶段,称为基学习阶段,又叫内循环阶段,这个阶段为第2个阶段提供反馈;第2个阶段是元学习阶段,又叫外循环阶段,它利用第1阶段的反馈来搜索最优的归纳偏置信息。在具体的元强化学习算法中,通常将相关任务视为问题空间上的一个任务分布上的采样,再利用元学习方法从采样的相关任务中学习或抽取有用的归纳偏置。最后根据不同的学习目标,将其嵌入到具体的基强化学习器。

2　元强化学习框架分析与归纳

基于图1的元强化学习流程,本节分别对元强化学习的相关任务、归纳偏置以及学习目标3个部分作进一步的分析介绍,并从这些角度出发,分别对当前的研究进行归纳和分类。

2.1　元强化学习相关任务

元强化学习中的相关任务即学习经验对元学习器所学习到的归纳偏置的性能好坏,将在每个相关任务的经验上进行估计。对于元学习范式下的监督学习而言,可以用训练损失作为对不同归纳偏置的奖励或评价准则,这些奖励的大小同时也反映了任务经验的好坏。如果以测试集上的损失作为元学习的性能度量,则任务中不仅应包含训练集,用于每个任务在归纳偏置下进行自适应学习,同时也应包含测试数据集,结合每个任务学习的自适应模型来评估特定归纳偏置的性能好坏。

在元强化学习中,任务经验较为复杂,可能包含关于任务环境的全部信息,如状态空间、动作空间、转移模型和奖励函数等。这些信息既可以用于单独学习不同的马尔科夫决策过程(Markov decision process, MDP)中某个方面的模式,如状态空间表示,也可以将全部信息映射到一个全局空间中,作为一个“任务向量”[23-24]。强化学习中的任务还可包括环境以外的辅助信息,特别是关于智能体的信息。例如值函数、策略函数以及该任务训练过程中的信息,如梯度、学习率等,这些均可作为元学习的目标。

如何定义何为“相关任务”?目前文献中尚无统一接受的定义。这不仅取决于各任务的性质,也取决于目标任务本身。尽管很难验证,实践中通常假设用于学习的相关任务来自同一问题空间上的某个任务分布,任务本身可视为这一分布的独立采样。其次,元强化学习的目标(使强化学习算法具有在相关环境也能获得良好性能的能力)决定了测试环境和训练环境应具有一定的共同模式。理论上,测试任务应和训练任务来自同一分布(如果不是,则通常属于迁移强化学习考虑范畴)。例如,这些任务之间尽管奖励函数不同,但状态迁移机制相同,或反之。在常见的迷宫任务中,任务之间往往仅是迷宫布局和目标位置不同,但任务的目标一致。

除了人工选择元强化学习的相关任务,也可以依据所要学习的归纳偏置来自动学习和生成相关任务。由于相关任务的数目对元学习性能具有重要影响,自动生成相关任务可在一定程度上缓解该问题。代表性的工作有以下几个方面:

(1) 无监督方法。最近的研究尝试利用无监督学习的方式自动设计任务分布,增加训练任务的多样性以提高元强化学习的性能表现。具体思路是利用一个隐变量及其互信息最大化条件构造一个伪奖励函数,然后以这个伪奖励函数构造不同的任务供元强化学习算法训练。这种方法使智能体更加充分地探索任务状态空间,同时当给定真实的奖励函数后,通过伪奖励函数训练的元强化学习算法也能快速自适应。基于这种思想,一些研究[25-26]利用包含隐变量的自适应策略构造互信息目标,还有一些研究[27]则利用基于隐变量的状态转移模型来使互信息最大化。

(2) 自动课程学习方法。域随机化(Domain randomization)方法的目标是通过调整模拟场景,缩小在模拟任务下学习的模型与真实场景中模型之间的间隔。实现这一目标的一个典型方法是自动课程学习,根据智能体在当前模拟任务中的学习表现自动地调整模拟环境的参数配置。具体而言,自动课程学习[28-29]的基本思路是先采样一些简单的模拟任务进行元强化学习,当学习的性能收敛后,扩充模拟环境的参数变量,从而采样到更加困难的任务提供给元强化学习进行训练,并不断迭代元强化学习和任务采样过程。这种由易到难的学习方法不仅简化了训练同时也避免了困难的人工调参过程。

2.2　元强化学习的归纳偏置

在机器学习领域,归纳偏置通常是指学习器在对未知样本进行推理与预测时所依赖的一组假设。相关研究指出,没有学习器能在无归纳假设的前提下进行合理的泛化。实际上人类之所以能进行快速的归纳推断,一个主要原因就是人类在持续学习任务的过程中,不仅学习概念和动作技能,而且学习归纳偏置,即如何进行泛化。

归纳偏置的形式可以是专家对问题的洞察和经验,以假设空间等形式引入学习系统。但这种方

式的主要缺点是受制于专家自身的能力,因而更为理想的方式是找到一种方式,能够从以往的相关学习任务中自动学到有用的归纳偏置。换言之,元学习就是偏置学习,因而本文将不加区别地使用元学习、偏置学习和学会学习等术语。从这个角度,“元强化学习”就是自动从一系列相关强化学习任务中学习有用的归纳假设,从而增强学习器在新的未见强化学习任务上的学习效率和泛化性能。

对元学习器而言,学习归纳偏置是其提高学习能力的手段,从相关任务所获取的归纳偏置最终为提高基学习器在新学习任务上的性能服务。由于不同的基学习器所使用的算法不同,其依赖的归纳偏置也不同,这就决定了元学习的学习目标也不同。可以根据元强化学习所抽取的归纳偏置的不同来对元学习器算法进行分类。例如有的元强化学习方法使用非参数方法,通过保留过去的经验记忆作为未来推理的基础[30];有的通过学习共同的表示来为新任务提供先验知识,也有的致力于为新任务学习参数空间中的初始化条件来加快其学习速度。

(1) 学习样本。这些方法主要为新任务提供额外训练样本,用于基于模型或免模型(Model-free)的强化学习。这些样本并非当前策略生成,既可以视为对新任务的“异策略”数据,也可视为对新任务施加的某种方式的约束。如文献[31]提出利用元学习方法直接学习一个独立的探索策略,根据该探索策略采样的异策略数据对目标策略进行训练,从而提高学习效率。该方法根据目标策略训练前后的性能提升作为元奖励信号训练探索策略。上节所述的虚拟相关任务生成方法也可视为通过为新任务提供学习样本的方式来提供归纳偏置。

(2) 假设空间。假设空间是强化学习任务潜在输出的集合,一个好的假设空间应以较高的概率包含从同一任务分布中抽取的新任务的解。理论上,见到足够多的相关任务上充分的样本后,元学习器应能为新任务学习到一个好的假设空间,这既是目标也是巨大的挑战。

假设空间的形式,既可以是参数形式,也可以是非参数形式。一个参数形式的假设,如人工神经网络,通常可以用某个特定的泛函形式进行表示。定义这个泛函形式的先验知识可能包括:(1)输入输出空间的表示。例如,可以将任务自适应建模为一个任务推断过程[23-24,32],即学习一个推断网络,通过推断网络将输入的数据样本映射到一个任务隐变量空间作为自适应信息;(2)函数参数化形式(如网络结构、基函数形式等)。例如一些研究[33-36]利用不同类型的循环神经网络来构造可以存储和利用过往学习记忆的循环网络策略,在每个不同任务上重置模型隐状态并连续输入多条采样轨迹来更新学习策略模型。

另一方面也可利用以往任务的样本来直接学习经验记忆以构造非参数化的假设表示。也有的方法直接存储经验用于比较和查询[30],而并不用于构造假设表示,这种方法对于连续状态环境而言会造成困难,因为其状态空间是无限集。

(3) 学习算法。学习算法的输入为训练数据,输出为假设空间中的某个元素。这种映射一般通过优化某个目标性能实现。具体算法中,可以通过对以往相关任务的学习,从参数的初始化位置、关于目标额外的约束和学习率,甚至迭代方向等角度来学习先验偏置。例如,一些研究学习一个最优初始化策略[37-42]或状态转移模型[43],再根据不同的更新方式自适应到目标任务。还有一些算法直接学习一个损失函数[44]替代传统的强化学习目标或者学习一个额外的辅助损失函数[45],与传统强化学习目标函数结合来约束基强化学习器的学习目标。

2.3 元强化学习目标

如前所述,从元学习器的角度,元强化学习的目标是从相关任务中学习有用的归纳偏置,提供给后续新的强化学习任务使用。但从基强化学习器的角度,则需要考虑使用归纳偏置的目的、使用何种归纳偏置以及怎样使用归纳偏置等一系列问题。

本节从元强化学习器的角度,将元强化学习的目标大致分成两类:提高学习效率和提高泛化能力。强化学习中的学习效率可以通过样本复杂度来衡量,即为达成某个学习性能所需要的经验样本量,也就是指智能体与环境在线交互次数。由于环境的不确定性以及明确监督信号的缺失,强化学习算法所需的样本复杂度通常远高于传统的监督学习算法。利用相关强化学习任务中包含的有益偏置,是解决上述问题的途径之一。例如Florensa等结合随机神经网络和信息论正则化工具,从预训练环境中学习一组既具有多样性,又具有通用性和可迁移性的技能,用于减少一组下游任务的样本复杂度[46];Gupta等利用元学习方法来进行探索策略学习[47],以提高学习效率。这一问题的挑战在于难以定义什么是好的探索策略,特别是对稀疏延迟奖励的任务而言。作者解决前一问题的思路是将探索问题刻画化为学习任务相关的结构化动作噪声分布的问题;同时在训练阶段为每个任务构造稠密的奖励信号(因为此时目标状态已知),用于提供良好探索策略的经验。

元强化学习的第2个目标是提升基强化学习

器的泛化能力。即利用元学习提取的归纳偏置来提高强化学习算法在新任务上的性能。依据统计机器学习的基本理论，一个算法在有限样本上学习后的泛化性能主要由两个方面决定：一个是样本无关的偏置；另一个是训练偏差。例如在Q-learning中，给定任务M下，学习的Q值估计函数$\hat{Q}$的泛化性能上界可记为

$$\| \hat{Q} - Q^* \|_{d^*} \leqslant \varepsilon_{\text{app}}(\mathcal{H}, Q^*) + \varepsilon_{\text{est}}(M) \quad (1)$$

式中：Q^*为最优动作值函数；d^*为π^*在状态空间上的访问频率；$\mathcal{H}$为假设空间；ε_{app}表示由于假设偏置所导致的近似误差；ε_{est}表示有限样本下的估计误差。

为了提高强化学习的泛化性能，有必要同时考虑上述两个因素：一方面度量由于归纳偏置的引入（例如通过改变状态空间表示而间接影响假设空间，或利用特定的深度网络架构来直接决定新任务的函数形式）对近似误差的影响；另一方面，由于知识的迁移，可能降低单任务的学习难度，从而降低估计误差。在实践中，可以通过比较基学习器的渐进性能来判断元强化学习器所输出的归纳偏置是否有利于单个任务泛化性能的提高。

除了从基任务的角度来分析元强化学习的目标，基任务的特点也影响了不同形式的元强化学习算法，当前研究根据基任务的不同主要可以分为以下几类：

（1）免模型元强化学习。这类方法利用对相关任务的学习，在策略网络的参数空间中学习归纳偏置，即元策略[33,37-39,47]。在新的任务下，从这个元策略出发，通过不同的自适应方法得到针对当前任务的最优策略。一些研究[33,38,47]根据最常见的“演员-评论家”框架，利用策略梯度法对学习到的元策略在新任务下进行自适应更新得到当前任务最优策略。为减少基于策略梯度的元强化学习方法存在的计算困难问题，有研究基于进化策略[38]的方法元学习策略网络的归纳偏置，并能以更少的样本代价自适应到最优策略。

（2）基于模型的元强学习。以基于模型的强化学习作为基任务的元强化学习算法目的是在状态转移模型空间学习有用的归纳偏置。在新任务上先根据自适应方法得到任务最优状态转移模型，再根据模型产生的数据进行策略规划或学习。如文献[32]学习一个包含任务隐变量的神经网络来表示状态转移元模型以及任务推断网络，通过推断任务信息并自适应状态转移模型后规划智能体策略。与之类似，文献[48]进一步利用高斯过程对动态模型的不确定性进行建模。还有一些研究[40]用集成模型的方法表示状态转移模型的不确定性，利用集成模型代表不同的相关任务再结合相应的元强化学习方法直接学习最优策略。

（3）元模仿强化学习。模仿学习这类基任务是从专家演示数据中学习到最优策略，最直接的方法是行为克隆，直接学习智能体策略来模仿专家的行为。最简单的方式是利用循环神经网络框架[49]抽取归纳偏置并作为初始化模仿策略，或者使用基于梯度的元学习方法[41]替代循环神经网络学习器。另一种模仿学习方法则通过专家样本推断出奖励函数，再根据这个奖励函数进行学习最优策略，这种方式也称为逆强化学习。结合逆强化学习的元学习框架最终目的是在元奖励函数空间学习一个归纳偏置，根据不同的专家演示能自适应到相应的奖励函数，并以此进行强化学习。基于这个思想，一些工作[50-51]结合对抗最大熵逆强化学习框架进行元强化学习。

（4）异策略元强化学习。前述的免模型元强化学习算法通常使用同策略（on-policy）算法[52-53]，即假定元学习内外两个层次的优化问题的训练数据来自同一分布。相比之下，异策略通常有着更好的样本效率和性能表现。因此一些工作尝试将异策略方法结合到元强化学习框架中，在文献[23]中，通过同策略数据对任务进行推断，同时利用异策略数据训练元策略和元价值函数，提高了训练效率和最终性能。文献[54]中则借鉴倾向性估计的思想，扩充了用于自适应的可用数据量，进一步提高元训练过程的效率。

3　代表性算法

本节以几个典型的元强化学习算法为例，详细介绍上节所述的元强化学习框架如何具体应用到实际算法设计中。如前所述，元学习可以理解为偏置学习，所学的归纳偏置记为参数θ，任务集为$D_{\text{meta-train}} \triangleq \{D_i^{\text{tr}}, D_i^{\text{ts}}\}$，$D_i^{\text{tr}}$，$D_i^{\text{ts}}$分别为第$i$个任务的训练数据和测试数据集，可视为MDP分布$P(M)$上的一个采样，则元学习的目标就是使习得的偏置在这组n个相关MDP任务上的平均性能最大化，即有

$$\begin{aligned} &\theta^* = \arg\max_{\theta} \sum_{i=1}^{n} \log P(D_i^{\text{ts}} | \phi_i) \\ &\text{s.t.} \quad \phi_i = f_\theta(D_i^{\text{tr}}) \text{ for each task } i \end{aligned} \quad (2)$$

式中：ϕ，θ分别为基学习器和元学习器参数；$f_\theta(D_i^{\text{tr}})$表示基学习器利用第$i$个任务的训练样本执行内循环训练过程，$\log P(D_i^{\text{ts}}|\phi_i)$表示该任务的内循环结束后在相应测试集上的外循环反馈得分。以EPG（Evolved policy gradient）[44]算法为例。该算法始终维护一组“精英任务”，元学习的归纳偏置就

是这组精英任务的分布参数。在元学习过程中,先从分布中采样一个任务,然后按正常的强化学习训练该任务,并将得到的最优策略在任务环境中的平均回报作为当前偏置在该任务上的性能度量。根据度量结果,在外循环阶段算法调整归纳偏置,使得在任务分布上的平均性能最大化。具体算法流程如下:

(1) 给定归纳偏置参数θ,从分布中$P(M)$中采样任务M,初始化其策略为$\varphi\sim P_\varphi(\theta)$;

(2) 执行内循环,优化该任务下的最优策略:$\varphi^*=\arg\max_\varphi \mathbb{E}_{\tau\sim M,\pi_\varphi}(R_\tau)$,其中$R_\tau$为策略$\pi_\varphi$在$M$下执行得到的轨迹累计回报;

(3) 当内循环结束后,每个任务M的性能$F_M(\theta)$用其最终学习的策略在该任务上的期望回报来度量:$F_M(\theta)\triangleq E_\tau[R_\tau|M,\pi_{\varphi^*}]$,则元学习的目标可定义为在任务分布上的期望性能,有

$$J(\theta)=\mathbb{E}_{M\sim P(M)}[F_M(\theta)]=\mathbb{E}_{M\sim P(M)}[\mathbb{E}_\tau[R_\tau|M,\pi_{\varphi^*}]] \tag{3}$$

由于偏置θ与其性能度量$J(\theta)$之间的泛函形式并不明确,难以直接计算该目标函数关于θ的梯度$\nabla_\theta J(\theta)$。解决这个问题的一种常见方法是黑盒优化,在EPG算法中采用了如下的进化策略(Evolution strategy,ES)方法来进行估计,有

$$\nabla\hat{J}(\theta)\simeq\nabla_\theta E_{\varepsilon\sim N(0,I)}[J(\theta+\sigma\varepsilon)]=\frac{1}{\sigma}E_{\varepsilon\sim N(0,I)}[J(\theta+\sigma\varepsilon)]\varepsilon\simeq\frac{1}{k\sigma}\sum_{k=1}^{K}J(\theta+\sigma\varepsilon_k)\varepsilon_k \tag{4}$$

式中:第1个等号用高斯扰动(进化变异)的方式来近似估计策略梯度;第2个等式关于扰动参数计算梯度,最后用蒙特卡洛来采样来近似期望。整个偏置梯度可以视为关于随机进化方向ε_k的加权平均,其中提升偏置性能较好的方向获得更大权重。这种方法也可看作为一组相关任务维护一个先验的策略分布,在新任务中,从该分布中采样作为初始策略而不需要从头开始学习,与MAML[37]算法相似。

值得指出,虽然一般而言无法计算$J(\theta)$关于θ的梯度,但在某些场景下,可以设计一个完全可微的训练过程。这种训练过程适用于某一任务分布下的所有任务,使得平均而言能够获得较高的期望回报。因此这一训练过程本身就是一种归纳偏置,并且由于可微,所以极大简化了元学习过程,这一路线的典型算法是Wang等[34]提出的利用LSTM(Long-short temporal memory)循环网络来构造训练过程。假设LSTM的参数为θ,则元学习的目标函数为

$$J(\theta)=\mathbb{E}_{M\sim P(M)}\left[\mathbb{E}_\tau\left[\sum_{t=0}^{T}r(s_t,a_t)|M,\pi_\theta\right]\right] \tag{5}$$

式中:π_θ为基于整个观测历史$H_{t-1}=\{s_0,a_0,\cdots,s_{t-1},a_{t-1}\}$的策略$\pi_\theta(a_t|s_t,H_{t-1})$。该策略本身用一个可微LSTM网络来近似,可以在一组相关任务上通过梯度下降进行训练,即有

$$\nabla J(\theta)=\nabla E_M[F_M(\theta)]\simeq\frac{1}{K}\sum_{i=1}^{K}\nabla_\theta F_M(\hat{\theta}+\sigma\varepsilon_i) \tag{6}$$

式中:$\theta\sim N(\hat{\theta},\sigma^2 I)$;$\varepsilon_i\sim N(0,I)$;$K$为任务数。式(6)可以理解为对复杂任务损失曲面的卷积,令外层优化更容易。注意该方法本质是用多个任务经验来训练一个共享的LSTM策略网络,与用LSTM来表示一个学习算法的元学习方法[55]具有根本不同。本方法可以理解为用LSTM网络本身来学习和保存多个相关任务的经验,而不是为新任务单独学习一个扩展的任务经验字典[30],后者更为直接,但在状态空间较大时的泛化性和伸缩性较差。

基于LSTM策略的方法给出了可微的$J(\theta)$函数形式,使得可以用传统的梯度反传方法来优化归纳偏置,但其缺点在于学习这样的优化器往往需要较大的样本复杂度。为了降低学习代价同时仍然保持目标函数对参数的可微性质,可以对单个任务加以约束和简化。例如MAML算法中,当采样一个任务M后,用$P_\varphi(\theta)$来初始化策略π_φ,但要求每个任务的优化策略π_{φ^*}与初始策略之间的距离尽可能小,则有目标函数

$$\begin{aligned}&J(\theta)=\mathbb{E}_{M\sim P(M)}[F_M(\theta)]=\\&\quad\mathbb{E}_{M\sim P(M)}[\mathbb{E}_\tau[R(\tau)|M,\pi_{\varphi^*}]]\\&\text{s.t.}\quad\varphi^*=\theta+\alpha\nabla_\theta F_M(\pi_\theta)\end{aligned} \tag{7}$$

基于链式法则有

$$\nabla_\theta J(\theta)=\nabla_\theta\sum\nolimits_{M_i\sim P(M)}\frac{\partial F_M(\pi_{\varphi^*})}{\partial\pi_{\varphi^*}}\frac{\partial\pi_{\varphi^*}}{\partial\theta} \tag{8}$$

式中:$\frac{\partial F_M(\pi_{\varphi^*})}{\partial\pi_{\varphi^*}}$为策略梯度;$\frac{\partial\pi_{\varphi^*}}{\partial\theta}$有解析表达式。相比基于LSTM的全可微方法,这种方法仍然保持元学习的目标函数可微的优点,但需学习的参数更少,学习效率更高。

4 元强化学习环境和评估标准

元强化学习的相关任务通常用一个任务分布中的采样来表示,因此其学习环境的构建以及最终算法的性能评估标准与一般的强化学习设置下有所不同。目前有一些开源的元强化学习研究环境,如Sonic the Hedgehog[56]、CoinRun[57]和Meta-

World[58]用于训练和对比各种元强化学习算法的性能表现。

其中 Sonic the Hedgehog 和 CoinRun 是基于视频游戏的元强化学习环境，且都是离散动作环境，两者分别包含 58 和 2^{32} 个强化学习任务。而 Meta-world 则是一个仿真机械臂控制任务的元强化学习环境，属于连续动作环境，总共包含 50 种机械臂控制任务，如“推”、“拉”、“开门”等动作。图 2 所示是 Meta-World 平台上元强化学习的一个任务分布示例，其目标是完成机械臂放置物品到指定位置的任务。每个采样任务上，机械臂状态和动作空间、转移概率模型都是一致的，所不同的是需放置物品的目标位置不同，即奖励函数不同。目前大量元强化学习研究都采用这种方式，即通过随机采样动力学模型或奖励函数的超参数设置来区分不同的采样任务。

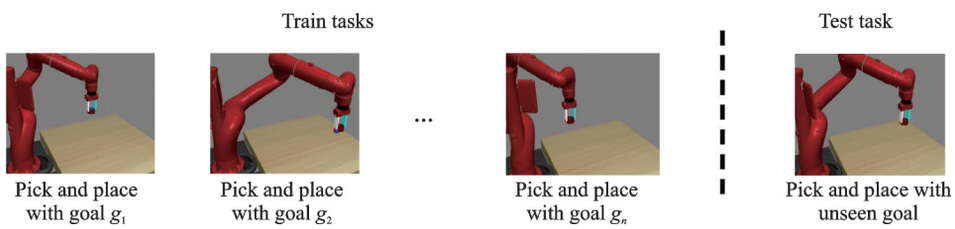

图 2　Meta-World 环境中相关任务示例

Fig.2　Examples of related tasks in Meta-World

元强化学习通过在训练任务上学习到有效的归纳偏置信息，再利用学习到的归纳偏置来提高强化学习智能体在新的测试任务上的学习效率和性能表现。因此，评价不同元强化学习算法性能的优劣就是评价其学习到的归纳偏置信息对智能体学习测试任务时的帮助。Langley[59]提出了几种性能指标（图 3），从以下方面来评估元强化学习算法。

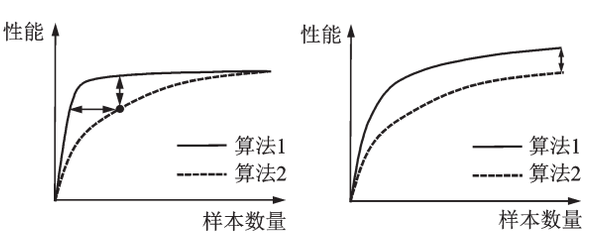

图 3　知识迁移的性能指标

Fig.3　Some performance indicators about transfer of knowledge

4.1　学习速率

元强化学习的一个主要动机是降低强化学习过程的样本复杂度，提高算法泛化性能。因此，度量元强化学习算法性能好坏的一个重要指标就是其在测试任务上的学习速率。如图 3(a)所示，强化学习条件下，学习速率可以用样本复杂度，通过两个方面来反映：(1)达到某个设定的期望性能所需的训练样本数量，即给定一个性能阈值，通过对比不同算法在测试任务上自适应训练达到指定阈值性能时所需训练样本数来比较算法性能，间接反映所学到的归纳偏置的质量：一般所需训练样本数越少，算法样本复杂度越低，说明元学习的归纳偏置越好；(2)相同训练样本数量下算法性能差异，即使用相同数量的训练样本进行自适应训练后，不同算法的性能差异代表了学习速率的快慢，也间接说明了元学习得到的归纳偏置好坏：相同数量的训练样本下，自适应算法性能越高，元学习的归纳偏置性能越好。

需要指出，以上两个准则都是对归纳偏置性能的间接度量，不应作绝对化理解。因为算法在测试环境下的性能本质上是多种因素的非线性函数，如归纳偏置、探索策略、优化方法和基学习器泛化能力等。其中任一因素的改变都会导致算法性能的改变，而实践中往往很难对此施加严格的控制，因此性能改善不能绝对化理解为相应的归纳偏置性能一定有用。

4.2　收敛性能

除了判断元强化学习速率，如图 3(b)所示利用元强化学习得到的归纳偏置在测试任务上最终的收敛性能也是衡量算法表现的评估标准。当元强化学习将学到的归纳偏置应用到不同的强化学习算法中时，其收敛性能也会有明显的差异。通常情况下需要结合学习速率和性能收敛对不同的元强化学习算法进行综合比较。

本文根据相关任务、归纳偏置、元学习算法和性能评估这几个方面对一些代表性算法进行简要归纳，具体总结见表 1，同时本文在表 2 列出了相关平台和算法的项目开源地址。

表1 一些代表性算法归纳对比

Table 1 Summary and comparison of some representative algorithms

算法	相关任务	归纳偏置	元学习算法	性能评估
RL2[33]	一组随机布局和目标的视觉迷宫任务	相关任务经验记忆	循环神经网络	1 000个随机测试任务中导航成功率
LSTM-A2C[34]	一组随机目标的视觉迷宫任务	相关任务经验记忆	结合LSTM的“演员-评论家”算法	视觉MDP任务上的累积奖励
MAML[37]	一组随机方向/速度的Mujoco任务	策略网络初始参数设置	基于梯度的内、外循环优化	智能体按目标方向和速度运动的得分
ProMP[38]	一组随机方向和速度的Mujoco任务	策略网络初始参数设置	结合低方差估计梯度的内、外循环优化	智能体按目标方向和速度运动的得分
ES-MAML[39]	一组随机方向和速度的Mujoco任务	策略网络初始参数设置	基于进化策略梯度的内、外循环优化	智能体按目标方向和速度运动的得分
MAESN[46]	一组随机目标的机械爪推块任务	结构化探索策略	自适应探索噪声隐变量	机器人推块任务的目标距离得分
PEARL[23]	一组随机方向和速度的Mujoco任务	任务隐变量推断网络	异策略算法元训练	智能体按目标方向和速度运动的得分
MQL[54]	一组随机方向和速度的Mujoco任务	“演员-评论家”网络初始设置	异策略数据自适应	智能体按目标方向和速度运动的得分
GMPS[42]	一组随机目标的四足机器人运动任务	策略网络初始参数设置	行为克隆方法元训练	机器人运动任务的目标距离得分
NEC[30]	一组Atari游戏任务	可微神经字典	Q-learning更新记忆模块	Atari游戏任务下的平均性能得分
CaDM[32]	一组随机方向和速度的Mujoco任务	情景感知状态转移模型	双向预测损失训练转移模型	智能体运动得分和状态预测误差
MIL[41]	一组多示例实体机器人操控任务	策略网络初始参数设置	基于模仿学习的内、外循环优化	实体机器人操控任务上的成功率
PEMIRL[50]	一组随机目标的仿真机器人操控任务	情景感知奖励函数	互信息正则化的最大熵逆强化学习	测试任务上得分均值和标准差比较
CARML[26]	一组随机生成的视觉迷宫导航任务	策略网络初始参数设置	无真实奖励信号的元学习	视觉迷宫任务下成功率性能曲线
RCAN[60]	一组现实场景下机械臂抓取任务	规范仿真场景生成器	对抗生成网络	仿真场景和真实场景下的平均抓取成功率

表2 相关算法及实验环境项目地址

Table 2 Experimental environment URL of some algorithms

算法	项目开源地址	算法	项目开源地址	算法	项目开源地址
CoinRun	https://github.com/openai/coinrun	Meta-World	https://github.com/rlwork-group/metaworld	MAML	https://github.com/cbfinn/maml_rl
ProMP	https://github.com/jonasroth-fuss/ProMP	PEARL	https://github.com/katerakel-ly/oyster	GMPS	https://github.com/russell-mendonca/GMPS
CaDM	https://github.com/younggyo-seo/CaDM	PRMIRL	https://github.com/ermon-group/MetaIRL		

5 结 论

本文对最近元强化学习研究的进展进行了回顾和总结。首先介绍了元强化学习的总体框架以及相关概念。在此基础上,从相关任务、归纳偏置、学习目标3个关键方面对当前研究做进一步的分析和分类,同时列举并介绍了相应的代表性工作。最后还介绍了当前一些开源的元强化学习实验环境以及算法性能评估标准。

目前元强化学习研究方兴未艾,还没有形成一套完整且成熟的研究体系,同时也给了研究者们深度挖掘的机会。尽管新的方法和思想不断涌现,但仍然存在以下几个方面值得更加深入的探究:(1)计算复杂度。元强化学习面临一个关键问题就是计算复杂度高。内循环过程除了利用任务样本进行优化自适应外,还需要额外的计算用于评估特定归纳偏置在当前任务上的效果,从而为外循环过程提供奖励。每个任务上的评估往往需要昂贵的计算代价。如何在保证一定计算精度的前提下,简化或改进内循环过程的优化及评估方法是解决元强化学习计算复杂度问题的一个重要方向。(2)鲁棒性问题。机器学习中一个很重要的概念就是算法的鲁棒性,因为现实中的数据通常存在大量噪声。具体到元强化学习问题,单个任务的性能不仅取决于归纳偏置,而且还与其他因素,如模型初始化、训练数据和优化算法等密切相关,因此用归纳

偏置在任务上的学习表现作为其评价标准本质上带有大量噪声。如何设计更加鲁棒的元强化学习算法、减少噪声的影响,也是一个值得研究的方向。(3)优化复杂性。单个强化学习任务的目标函数本身就具有一定复杂性,在多个任务上评估所学归纳偏置的性能并优化进一步加剧了这个复杂性,使得元学习阶段关于归纳偏置的损失曲面本质上高度非线性,给优化求解带来了巨大的困难。因此如何提高元学习器的优化性能也是未来研究的关键。(4)实用性问题。现有元强化学习方法对实际应用通常具有较强的限制,例如要求训练任务的分布不能太复杂,相关任务之间不能相差太大、训练任务和测试任务分布具有一定的一致性等。如何放松这些限制,使得元强化学习适用于更广泛的场合,是值得研究的问题。最近一些研究从"从多个任务中学习如何评价策略好坏"的角度,为全新任务学习有用的归纳偏置[41,61-62],是对这一方向的有益尝试。

参考文献:

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

[2] BELLEMARE M G, NADDAF Y, VENESS J, et al. The arcade learning environment: An evaluation platform for general agents[C]//Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina:[s.n.], 2015: 4148-4152.

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

[4] WEN C, YAO X, WANG Y, et al. SMIX(λ): Enhancing centralized value functions for cooperative multi-agent reinforcement learning[C]//Proceedings of the Thirty-Fourth Conference on Artificial Intelligence. New York, USA:[s.n.], 2020: 7301-7308.

[5] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[C]//Proceedings of International Conference on Machine Learning. Beijing, China:[s.n.], 2014: 387-395.

[6] FUJIMOTO S, HOOF H V, MEGER D, et al. Addressing function approximation error in actor-critic methods[C]//Proceedings of International Conference on Machine Learning. Stockholm, Sweden:[s.n.], 2018: 1587-1596.

[7] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//Proceedings of International Conference on Machine Learning. Stockholm, Sweden:[s.n.], 2018: 1856-1865.

[8] JANNER M, FU J, ZHANG M, et al. When to trust your model: Model-based policy optimization[C]//Proceedings of Neural Information Processing Systems. Vancouver, BC, Canada:[s.n.], 2019: 12498-12509.

[9] KURUTACH T, CLAVERA I, DUAN Y, et al. Model-ensemble trust-region policy optimization[C]//Proceedings of International Conference on Learning Representations. Vancouver, BC, Canada:[s.n.], 2018.

[10] BARRETO A, BORSA D, QUAN J, et al. Transfer in deep reinforcement learning using successor features and generalised policy improvement[C]//Proceedings of International Conference on Machine Learning. Stockholm, Sweden:[s.n.], 2018: 501-510.

[11] LAROCHE R, BARLIER M. Transfer reinforcement learning with shared dynamics[C]//Proceedings of the Thirty-First Conference on Artificial Intelligence. San Francisco, USA:[s.n.], 2017: 2147-2153.

[12] OH J, SINGH S, LEE H, et al. Zero-shot task generalization with multi-task deep reinforcement learning[C]//Proceedings of International Conference on Machine Learning. Sydney, NSW, Australia:[s.n.], 2017: 2661-2670.

[13] YANG Z, MERRICK K E, ABBASS H A, et al. Multi-task deep reinforcement learning for continuous action control[C]//International Joint Conference on Artificial Intelligence. Melbourne, Australia:[s.n.], 2017: 3301-3307.

[14] DERAMO C, TATEO D, BONARINI A, et al. Sharing knowledge in multi-task deep reinforcement learning[C]//Proceedings of International Conference on Learning Representations. Addis Ababa, Ethiopia:[s.n.], 2020.

[15] ROLNICK D, AHUJA A, SCHWARZ J, et al. Experience replay for continual learning[C]//Proceedings of Neural Information Processing Systems. Vancouver, BC, Canada:[s.n.], 2019: 350-360.

[16] KAPLANIS C, CLOPATH C, et al. Continual reinforcement learning with multi-timescale replay[EB/OL]. (2020-04-16)[2020-10-16]. https://arxiv.org/abs/2004.07530.

[17] ZHU F, CHANG X, ZENG R, et al. Continual reinforcement learning with diversity exploration and adversarial self-correction[EB/OL].(2019-06-21)[2020-06-21]. https://arxiv.org/abs/1906.09205.

[18] HARLOW, HARRY F. The formation of learning

sets[J]. Psychological Review, 1949, 56(1): 51-65.

[19] SCHMIDHUBER J. Evolutionary principles in self-referential learning[D]. Munchen, Germany: Technische Universitat Munchen, 1987.

[20] HINTON G E. Using fast weights to deblur old memories[C]//Proceedings of Conference Cognitive Science. Erlbaum:[s.n.], 1987: 177-186.

[21] BENGIO Y, BENGIO S, CLOUTIER J, et al. Learning a synaptic learning rule[C]//Proceedings of International Joint Conference on Neural Network. [S.l.]: [s.n.], 1991: 969.

[22] SCHMIDHUBER J. A neural network that embeds its own meta-levels[C]//Proceedings of IEEE International Conference on Neural Networks. San Francisco, CA, USA:[s.n.], 1993: 407-412.

[23] RAKELLY K, ZHOU A, FINN C, et al. Efficient off-policy meta-reinforcement learning via probabilistic context variables[C]//Proceedings of International Conference on Machine Learning. Long Beach, California, USA:[s.n.], 2019: 5331-5340.

[24] HUMPLIK J, GALASHOV A, HASENCLEVER L, et al. Meta reinforcement learning as task inference[EB/OL]. (2019-05-15)[2019-10-22]. https://arxiv.org/abs/1905.06424.

[25] GUPTA A, EYSENBACH B, FINN C, et al. Unsupervised meta-learning for reinforcement learning[EB/OL]. (2018-06-12)[2020-04-30]. https://arxiv.org/abs/1806.04640.

[26] JABRI A, HSU K, EYSENBACH B, et al. Unsupervised curricula for visual meta-reinforcement learning[C]//Proceedings of Neural Information Processing Systems. Vancouver, BC, Canada:[s.n.], 2019: 10519-10530.

[27] SHARMA A, GU S, LEVINE S, et al. Dynamics-aware unsupervised discovery of skills[C]//Proceedings of International Conference on Learning Representations. Addis Ababa, Ethiopia:[s.n.], 2020.

[28] AKKAYA I, ANDRYCHOWICZ M, et al. Solving Rubik's cube with a robot hand[EB/OL]. (2019-10-16)[2020-10-16]. https://arxiv.org/abs/1910.07113.

[29] MEHTA B, DELEU T, RAPARTHY S C, et al. Curriculum in gradient-based meta-reinforcement learning[EB/OL]. (2020-02-19)[2020-10-19]. https://arxiv.org/abs/2002.07596.

[30] PRITZEL A, URIA B, SRINIVASAN S, et al. Neural episodic control[C]//Proceedings of International Conference on Machine Learning. Sydney, NSW, Australia:[s.n.], 2017: 2827-2836.

[31] XU T, LIU Q, ZHAO L, et al. Learning to explore with meta-policy gradient[C]//Proceedings of International Conference on Machine Learning. Stockholm, Sweden:[s.n.], 2018.

[32] LEE K, SEO Y, LEE S, et al. Context-aware dynamics model for generalization in model-based reinforcement learning[C]//Proceedings of International Conference on Learning Representations. Addis Ababa, Ethiopia:[s.n.], 2020.

[33] DUAN Y, SCHULMAN J, CHEN X, et al. RL^2: Fast reinforcement learning via slow reinforcement learning[EB/OL]. (2016-11-09) [2019-11-10]. https://arxiv.org/abs/1611.02779.

[34] WANG J X, KURTH-NELSON Z, TIRUMALA D, et al. Learning to reinforcement learn[EB/OL]. (2016-11-17) [2017-01-23]. https://arxiv. org/abs/1611.05763.

[35] SANTORO A, FAULKNER R, RAPOSO D, et al. Relational recurrent neural networks[C]//Proceedings of Neural Information Processing Systems. Montreal, Canada:[s.n.], 2018: 7299-7310.

[36] MISHRA N, ROHANINEJAD M, CHEN X, et al. A simple neural attentive meta-learner[C]//Proceedings of International Conference on Learning Representations. Vancouver, BC, Canada:[s.n.], 2018.

[37] FINN C, ABBEEL P, LEVINE S, et al. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of International Conference on Machine Learning. Sydney, NSW, Australia:[s.n.], 2017: 1126-1135.

[38] ROTHFUSS J, LEE D, CLAVERA I, et al. ProMP: Proximal meta-policy search[C]//Proceedings of International Conference on Learning Representations. New Orleans, LA, USA:[s.n.], 2019.

[39] SONG X, GAO W, YANG Y, et al. ES-MAML: Simple Hessian-free meta learning[C]//Proceedings of International Conference on Learning Representations. Addis Ababa, Ethiopia:[s.n.], 2020.

[40] CLAVERA I, ROTHFUSS J, SCHULMAN J, et al. Model-based reinforcement learning via meta-policy optimization[C]//Proceedings of Conference on Robot Learning. Zurich, Switzerland: [s.n.], 2018: 617-629.

[41] YU T, FINN C, XIE A, et al. One-shot imitation from observing Humans via domain-adaptive meta-learning[C]//Proceedings of International Conference on Learning Representations. Vancouver, BC, Canada:[s.n.], 2018.

[42] MENDONCA R, GUPTA A, KRALEV R, et al. Guided meta-policy search[C]//Proceedings of Neural Information Processing Systems. Vancouver, BC, Canada:[s.n.], 2019: 9656-9667.

[43] NAGABANDI, A, CLAVERA I, SIMIN L, et al. Learning to adapt in dynamic, real-world environ-

ments through meta-reinforcement learning[C]//Proceedings of International Conference on Learning Representations. New Orleans, LA, USA:[s.n.], 2019.

[44] HOUTHOOFT R, CHEN Y, ISOLA P, et al. Evolved policy gradients[C]//Proceedings of Neural Information Processing Systems. Montreal, Canada:[s.n.], 2018: 5400-5409.

[45] ZHOU W, LI Y, YANG Y, et al. Online meta-critic learning for off-policy actor-critic methods[EB/OL]. (2020-03-11) [2020-11-02]. https://arxiv. org/abs/2003.05334.

[46] FLORENSA C, DUAN Y, ABBEEL P, et al. Stochastic neural networks for hierarchical reinforcement learning[C]//Proceedings of International Conference on Learning Representations. Toulon, France:[s.n.], 2017.

[47] GUPTA A, MENDONCA R, LIU Y, et al. Meta-reinforcement learning of structured exploration strategies[C]//Proceedings of Neural Information Processing Systems. Montreal, Canada:[s.n.], 2018: 5302-5311.

[48] SAEMUNDSSON S, HOFMANN K, DEISENROTH M P, et al. Meta reinforcement learning with latent variable gaussian processes[C]//Proceedings of Uncertainty in Artificial Intelligence. Monterey, California, USA:[s.n.], 2018: 642-652.

[49] DUAN Y, ANDRYCHOWICZ M, STADIE B C, et al. One-shot imitation learning[C]//Proceedings of Neural Information Processing Systems. Long Beach, CA, USA:[s.n.], 2017: 1087-1098.

[50] YU L, YU T, FINN C, et al. Meta-inverse reinforcement learning with probabilistic context variables[C]//Proceedings of Neural Information Processing Systems. Vancouver, BC, Canada:[s.n.], 2019: 11749-11760.

[51] GHASEMIPOUR S K, GU S, ZEMEL R S, et al. SMILe: Scalable meta inverse reinforcement learning through context-conditional policies[C]//Proceedings of Neural Information Processing Systems. Vancouver, BC, Canada:[s.n.], 2019: 7881-7891.

[52] WANG Y, HE H, TAN X, et al. Trust region-guided proximal policy optimization[C]//Proceedings of Neural Information Processing Systems. Vancouver, BC, Canada:[s.n.], 2019: 626-636.

[53] WANG Y, HE H, TAN X, et al. Truly proximal policy optimization[C]//Proceedings of Uncertainty in Artificial Intelligence. Tel Aviv, Israel:[s.n.], 2019: 113-122.

[54] FAKOOR R, CHAUDHARI P, SOATTO S, et al. Meta-q-learning[C]//Proceedings of International Conference on Learning Representations. Addis Ababa, Ethiopia:[s.n.], 2020.

[55] RAVI S, LAROCHELLE H. Optimization as a model for few-shot learning[C]//Proceedings of International Conference on Learning Representations. Toulon, France:[s.n.], 2017.

[56] NICHOL A, PFAU V, HESSE C, et al. Gotta learn fast: A new benchmark for generalization in RL[EB/OL]. (2018-04-10)[2020-04-23]. https://arxiv. org/abs/1804.03720.

[57] COBBE K, KLIMOV O, HESSE C, et al. Quantifying generalization in reinforcement learning[C]//Proceedings of International Conference on Machine Learning. Long Beach, California, USA: [s. n.], 2019: 1282-1289.

[58] YU T, QUILLEN D, HE Z, et al. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning[EB/OL]. (2019-10-24)[2021-06-14].https://arxiv.org/abs/1910.10897.

[59] LANGLEY P. Transfer of knowledge in cognitive systems[C]//Proceedings of Workshop on Structural Knowledge Transfer for Machine Learning, International Conference on Machine Learning. Pittsburgh, Pennsylvania:[s.n.], 2006.

[60] JAMES S, WOHLHART P, KALAKRISHNAN M, et al. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks[C]//Proceedings of Computer Vision and Pattern Recognition. Long Beach, California, USA:[s.n.], 2019: 12627-12637.

[61] SUNG F, ZHANG L, XIANG T, et al. Learning to learn: Meta-critic networks for sample efficient learning[EB/OL]. (2017-06-29) [2020-06-29]. https://arxiv.org/abs/1706.09529.

[62] KIRSCH L, STEENKISTE S V, SCHMIDHUBER J, et al. Improving generalization in meta reinforcement learning using learned objectives[C]//Proceedings of International Conference on Learning Representations. Addis Ababa, Ethiopia:[s.n.], 2020.

(编辑:刘彦东)