# 基于课程学习的深度强化学习研究综述

林泽阳,赖 俊,陈希亮
(陆军工程大学 指挥控制工程学院,江苏 南京 210007)

**摘 要**:作为解决序贯决策的机器学习方法,强化学习采用交互试错的方法学习最优策略,能够契合人类的智能决策方式。基于课程学习的深度强化学习是强化学习领域的一个研究热点,它针对强化学习智能体在面临高维状态空间和动作空间时学习效率低、难以收敛的问题,通过抽取一个或多个简单源任务训练优化过程中的共性知识,加速或改善复杂目标任务的学习。论文首先介绍了课程学习的基础知识,从四个角度对深度强化学习中的课程学习最新研究进展进行了综述,包括基于网络优化的课程学习、基于多智能体合作的课程学习、基于能力评估的课程学习、基于功能函数的课程学习。然后对课程强化学习最新发展情况进行了分析,并对深度强化学习中的课程学习的当前存在问题和解决思路进行了总结归纳。最后,基于当前课程学习在深度强化学习中的应用,对课程强化学习的发展和研究方向进行了总结。

**关键词**:强化学习;深度学习;深度强化学习;课程学习;迁移学习
**中图分类号**:TP181 **文献标识码**:A **文章编号**:1673-629X(2022)11-0016-08
doi:10.3969/j.issn.1673-629X.2022.11.003

# An Overview of Deep Reinforcement Learning Based on Curriculum Learning

LIN Ze-yang, LAI Jun, CHEN Xi-liang
(School of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China)

**Abstract**: As a machine learning method to solve sequential decision making, reinforcement learning adopts interactive trial-and-error method to learn the optimal strategy, which can fit human intelligent decision-making mode. Deep reinforcement learning based on curriculum learning is a new research hotspot in the field of reinforcement learning. Aiming at the problems of low learning efficiency and hard convergence in high-dimensional state space and action space faced by reinforcement learning agents, by extracting common knowledge of one or more simple source task training in the process of optimization, the learning of complex target tasks can be accelerated or improved. Firstly, we introduce the basic knowledge of curriculum learning and summarize the latest research progress of curriculum learning in deep reinforcement learning from four perspectives, including the curriculum learning based on network optimization, curriculum learning based on multi-agent cooperation, curriculum learning based on the ability evaluation, curriculum learning based on the functions. Then we analyze the latest development of curriculum reinforcement learning and summarize the existing problems and solutions of curriculum learning in deep reinforcement learning. Finally, based on the application of current curriculum learning in deep reinforcement learning, the development and research direction of curriculum reinforcement learning are summarized.

**Key words**: reinforcement learning; deep learning; deep reinforcement learning; curriculum learning; transfer learning

## 0 引 言

强化学习(Reinforcement Learning, RL)作为机器学习分支之一,在人工智能领域具有重要地位[1]:智能体在环境中通过“交互-试错”获取正/负奖励值,调整自身的动作策略,从而生成总奖励值最大的动作策略模型[2]。

传统强化学习方法在有限状态空间和动作空间的任务中能够取得较好的收敛效果[3],但复杂空间状态任务往往具有很大的状态空间和连续的动作空间,尤其当输入数据为图像和声音时,传统强化学习很难处理,会出现维度爆炸问题[4-5]。解决上述问题的一个方法,就是将强化学习和深度神经网络(Deep Neural Network, DNN)结合,用多层神经网络来显式表示强化学习中的值函数和策略函数[6]。

**收稿日期**:2021-12-22 **修回日期**:2022-04-25
**基金项目**:国家自然科学基金资助项目(61806221)
**作者简介**:林泽阳(1995-),男,硕士研究生,研究方向为深度强化学习、课程学习;通讯作者:赖 俊(1979-),男,硕士,副教授,研究方向为人工智能、计算机仿真。

深度强化学习(Deep Reinforcement Learning,DRL)将深度学习的感知能力和强化学习的决策能力相结合[7],近年来在人工智能领域迅猛发展,例如 Atari 游戏[8-9]、复杂机器人动作控制[10-11],以及围棋 AlphaGo 智能的应用[12]等,2015 年机器学习领域著名专家 Hinton、Bengio、Lecun 在《Nature》上发表的深度学习综述一文将深度强化学习作为深度学习的重要发展方向[13]。

尽管在过去三十年间取得很大进步,但由于标准强化学习智能体的初始设定都是随机策略,在简单环境中通过随机探索和试错,能够达成较好的训练效果[14]。但在复杂环境中由于状态空间的复杂性、奖励信号的稀疏性,强化学习从环境中获取样本的成本不断提高,学习时间过长,从而影响了智能体的有效探索[15]。

解决上述问题的一个有效途径,就是将课程学习(Curriculum Learning,CL)和深度强化学习相结合[16]。2009 年,以机器学习领军人物 Bengio 为首的科研团队在国际顶级机器学习会议 ICML 上首次提出课程学习的概念[17],引起机器学习领域的巨大轰动。课程学习借鉴人类从简单到复杂的学习思想,首先在任务集中筛选出部分简单任务进行学习以产生训练课程,而后在剩余的复杂任务中利用训练课程进行学习,最后在整个训练集中进行训练。将课程学习和深度强化学习相结合,可以有以下两个方面的作用[18]:(1)可以加快训练模型的收敛速度,避免训练初期对于复杂任务投入过多训练时间;(2)提高模型的泛化能力,增强对复杂任务的学习能力。

该文首先对课程学习进行简要描述,从四个角度对深度强化学习中的课程学习进行了分类整理,之后对近三年的基于课程学习的深度强化学习新算法进行了总结分析,最后讨论了基于课程学习的深度强化学习的发展前景和挑战。

## 1　基于课程学习的深度强化学习

课程学习的目标是自动设计和选择完整序列的任务(即课程) $M_1,M_2,\cdots,M_t$ 对智能体进行训练,从而提高对目标任务的学习速度或性能[19],课程学习流程如图 1 所示。

课程马尔可夫决策过程(Curriculum Markov Decision Process,CMDP)[20]是一个 6 元组 $(S,A,p,r,\Delta s_0,S_f)$,其中 $S$ 是状态空间集,$A$ 是动作空间集,$p(s'\mid s,a)$ 代表智能体在状态 $s$ 时采取动作 $a$ 后转移到状态 $s'$ 的概率,$r(s,a,s')$ 代表在状态 $s$ 采取动作 $a$ 到达状态 $s'$ 所获得的即时奖励,$\Delta s_0$ 代表初始状态分布,$S_f$ 代表最终状态集。

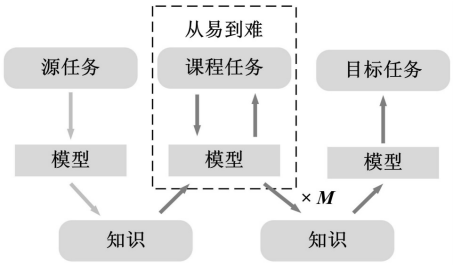


图 1　课程学习流程

常见的课程创建方法有以下两种[21]:(1)在线创建课程,根据智能体对给定顶点样本的学习进度动态添加边;(2)离线创建课程,在训练前生成图,并根据与不同顶点相关联的样本的属性选择边。课程设计流程如图 2 所示。

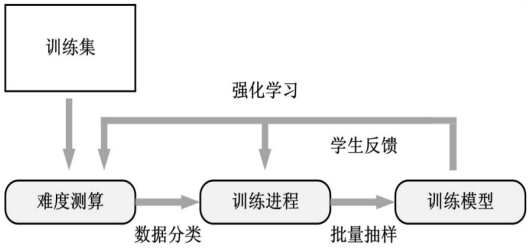


图 2　课程设计流程

课程学习方法可认为包括三部分[22]:任务生成、排序和迁移学习。任务生成是创建一组好的中间任务的过程,从中获取经验样本。排序研究了如何在一组经验样本上创建部分排序 $D$,也就是说,如何生成课程图的边。迁移学习主要研究如何将知识从一个或多个源任务直接转移到目标任务。为了评价源任务迁移到目标任务的性能优劣[23-24],有以下指标可以量化。(1)学习速度提升。即智能体在迁移知识的前提下能够以多快的速度学习到最优策略,从而在目标任务上实现预期的性能值 $G_0\geqslant\delta$,其中 $\delta$ 是总任务期望的性能阈值。(2)初始性能提升。通过从源任务进行迁移,观察智能体在学习过程中对目标任务的初始性能提升来衡量迁移效果。(3)渐近性能提升。通过比较智能体在使用迁移与不使用迁移时目标任务收敛后的最终性能来衡量迁移效果。

## 2　深度强化学习中的课程学习研究进展

对于强化学习智能体来说,自主学习一项复杂任务需要很长的时间。在深度强化学习中应用课程学习,可以通过利用一个或多个源任务的知识来加速或改善复杂目标任务的学习[25]。

Felipe 等人提出了新方法[26]:(1)将目标任务划分为简单任务;(2)在尽量小的专家经验支持下,根据面向对象的任务描述自动生成课程;(3)使用生成的课程来跨任务重用知识。实验表明在人工指定和生成

子任务方面都取得了更好的性能。

为了提高多智能体的学习性能,Jayesh 等人应用前馈神经网络(Feedforward Neural Network,FNN)完成协同控制任务[27],包括离散和连续动作任务,Daphna 等人提出了推断课程(Inference Curriculum,IC)的方法[28],从另一个网络迁移学习的方式,接受不同任务的训练。为了解决从稀疏和延迟奖励中学习的局限性问题,Atsushi 提出了一种基于渐进式神经网络(Progressive Neural Network,PNN)的课程学习方法[29],带参数的模块被附加上预先确定的参数,该策略比单组参数的效果更好。

### 2.1 基于网络优化的课程学习

传统课程学习对于小规模的多智能体强化学习性能提升明显,但在大规模多智能体环境中,由于环境和智能体之间的复杂动态以及状态-动作空间的维度爆炸,这仍然具有挑战性,所以如何更好地学习和产生更有效的任务课程是课程学习的研究重点。

王维埙等人设计了一种新的动态多智能体课程学习(Dynamic Multi-agent Curriculum Learning,DyMA-CL)来解决大规模智能体学习的问题[30],从一个小规模的多智能体场景开始学习,逐步增加智能体的数量。网络设计里有三种迁移机制:缓存复用(Buffer Reuse,BR)、基于 KL 散度的课程蒸馏(Curriculum Distillation,CD)和模型重载(Model Reload,MR)。

DyAN 的网络结构如图 3 所示,由于不同课程间智能体数量以及观测维度变化,缓存复用和基于 KL 散度的课程蒸馏机制不能直接用于 DyMA-CL 框架中,王维埙等人提供了一个语义映射函数 $\varphi(\cdot)$,将语义信息从每个智能体的观察值中抽取出来,从而找出不同状态空间之间的映射关系。

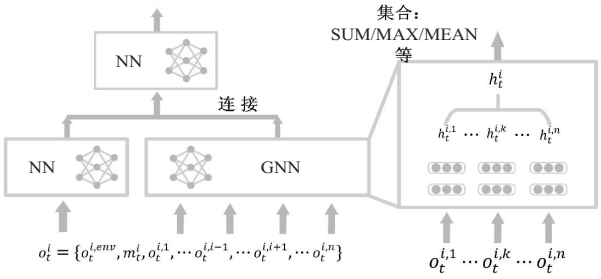


图 3 DyAN 的网络结构

传统的课程学习主要是针对单一类型智能体和固定的动作空间进行设计,Wu 等人引入主从智能体的概念[31],采用异步策略共享感知网络,在不同的动作空间内同时训练多个智能体。

主从智能体以异步方式同时学习相应的控制策略,以不同的频率运行,其中主智能体占用一半的线程,从智能体共享其余的一半线程。

### 2.2 基于多智能体合作的课程学习

不同的多智能体合作控制问题需要智能体在实现各自目标的同时为全局目标的成功做出贡献。这种多目标多智能体的设置给目前针对单一的全局奖励设置的算法带来两个问题[32]:(1)需要高效的学习探索,既要实现智能体的个体目标,又要为其他智能体的成功而进行合作;(2)不同智能体的行动和目标之间相互作用的信度分配。

为解决这两个问题,Yang 等人推导出一种基于多目标多智能体的梯度策略算法[33],并采用信度分配函数进行局部信度分配,使用一个增强函数来连接价值函数和策略函数。多目标多智能体的梯度策略如图 4 所示。

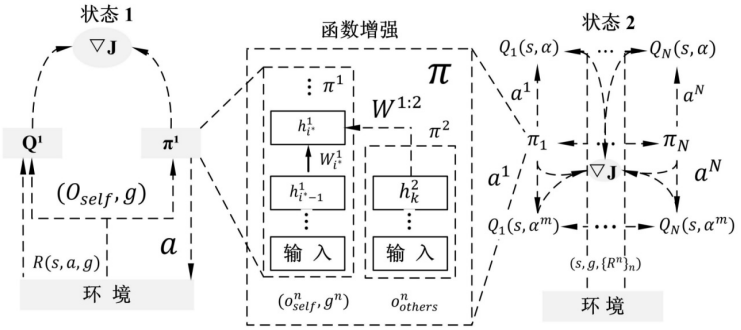


图 4 多目标多智能体的梯度策略

阶段 1:作者在 $N=1$ 和随机目标采样的诱导式 MDP 中训练了一个演员 $\pi^1(a \mid o,g)$ 和一个评论家 $Q^1(s^1,a,g)$,与完整的多智能体环境相比,这种方法使用的样本数量要少得多。

$$L(\theta_{Q_c}) = E_\pi[(R_t^n + \gamma Q_n^\pi(s_{t+1},\alpha_{t+1}^m;\theta_{Q_c}) - Q_n^\pi(s_t,\alpha_t^m;\theta_{Q_c}))^2] \quad (1)$$

$$\nabla_\theta J(\pi) = E_\pi[\sum_{m,n=1}^{N}(\nabla_\theta \log \pi^m(\alpha^m \mid o^m,g^m))A_{n,m}^\pi(s,a)] \quad (2)$$

阶段 2:马尔可夫博弈是用所有 $N$ 个智能体实例化的,将训练好的 $\pi_1$ 参数还原,实例化第二个神经网络 $\pi_2$,用于智能体 $o_{\text{others}}^n$ 处理,并将 $\pi_2$ 的输出连接到 $\pi_1$ 的选定隐藏层。

在多智能体游戏中,随着智能体数量的增加,环境的复杂性会呈指数级增长,所以在大规模智能体的前提下学习好的策略尤其具有挑战性。为解决这一挑战,Long 等人引入了进化种群课程(Evolutionary Population Curriculum,EPC)[34],使用种群进化的方法来解决整个课程中的一个客观错位问题[35]:早期训练的规模较小智能体模型,未必是应用到后期大规模智能体训练的最佳模型。Long 等人在训练的各个阶段维护多个智能体集,对各个智能体集进行混合匹配和微调,筛选出最佳适应性的智能体集进入下个阶段。种群不变 $Q$ 函数如图 5 所示。

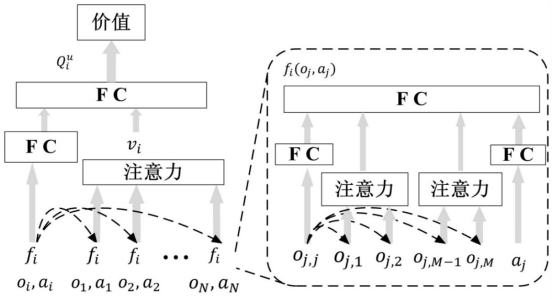


图 5　种群不变 $Q$ 函数

如图 5 所示,左半部分中,作者利用注意力机制组合来自不同观察动作编码器 $f_i$ 的嵌入,右半部分是 $f_i$ 的详细说明,作者还利用注意力模块将 $M$ 个不同的实体组合到一个观察值中。

在强化学习中,以往的任务排序方法都以减少模型训练时间并达到给定性能水平为目标进行探索。Francesco 等人定义了一个通用的任务排序优化框架[36],并评估了常用的元启发式搜索方法在多个任务上的性能。

给定一个评估指标 $P:C_{\leqslant L}\times M\rightarrow\mathbb{R}$ ,它为一个特定的最终任务评估课程,考虑找到一个最优课程 $C$ 的问题,如下:

$$P(c^*,m_f)\geqslant P(c,m_f)\ \forall c\in C_{\leqslant L} \tag{3}$$

### 2.3　基于能力评估的课程学习

与其他自监督的强化学习方法(如内在驱动方法)相比,多智能体的竞争可能会随着环境复杂性的提高而更加激烈,并导致智能体产生类似于人类技能的行为[37]。Bowen 等人提出一种以迁移和微调作为定量评估目标能力的方法[38],并且在一组特定领域的智力测验中将捉迷藏智能体和内在驱动与随机初始化基准值进行了比较。

在复杂的任务中,比如那些组合行动空间大的任务,随机探索的效率太低,当前的学习进展比较缓慢。Gregory 等人使用一个渐进增长的动作空间的课程来加速学习[39],智能体可以通过最初限制其动作空间来设置内部课程。Gregory 的方法使用非策略强化学习来同时估计多个动作空间的最优值函数,并有效地将数据、值函数估计和状态表示从受限的动作空间迁移到完整的任务。

$$V_i^*(s)\leqslant V_j^*(s)\ \forall s\quad \text{if}\quad i<j \tag{4}$$

因为每个动作空间都是较大动作空间的严格子集,因此在最坏的情况下,智能体总是可以退回到使用更受限制的动作空间的策略。

课程学习方法通常依靠启发式方法来估计训练实例的难度和模型的学习能力[40]。John P 等人提出了基于能力评估的课程学习动态数据选择(Dynamic Data Selection for Curriculum Learning via Ability Estimation,DDaCLAE)策略[41],该策略在每个训练阶段根据模型在该阶段的能力评估动态选择最佳训练实例。

算法 1:DDaCLAE
输入:数据( $X,Y$ ),模型 $\varphi$ ,难度 $D$, num_epochs
输出:训练好的模型 $\varphi$
1: for $e$ in num_epochs do
2: $\hat{Y}=\varphi(X)$
3: $\hat{\theta}_e=\text{score}(Y,\hat{Y},D)$
4: $X_e,Y_e=\{(x,y);b_x\leqslant\hat{\theta}_e\}$
5: $\text{train}(\varphi,X_e,Y_e)$
6: end for
7: procedure SCORE( $Y,\hat{Y},D$ )
8: $Z=\forall_{y\in Y}I[y_i=\hat{y}_i]$
9: $\hat{\theta}_e=\arg\max_\theta p(Z\mid\theta,b)$
10: return $\hat{\theta}_e$
11:end procedure

DDaCLAE 的训练过程见算法 1,John P 等人使用评分函数估计模型能力,使用完整的训练集而不是更新模型参数来获取响应数据。John P 等人发现,在 GLUE 分类任务上,使用学习困难参数的模型优于基于启发式的课程学习模型。

### 2.4　基于功能函数的课程学习

通过课程来训练智能体以提高智能体的性能和学习速度,Andrea 等人提出了一种基于任务复杂度的自动课程生成方法[42],引入了不同的进程函数,包括基于智能体性能的自主在线任务进程。与其他基于任务的课程学习方法不同,这种方法的进阶函数决定了智能体在每个中间任务上应该训练多长时间。通过在网格世界[43]和复杂模拟导航领域[44]中与两种最先进的课程学习算法的性能进行对比分析,证明了自动课程生成方法的优点和广泛的适用性。

传统课程学习的数值方法只提供了最初的启发式解决方案,几乎不能保证它们的质量。Francesco 等人

定义了一个新的灰盒函数[45],该函数包含一个合适的调度问题,可以有效地用来重构课程学习问题。

通过引入灰盒函数 $\psi:\mathbb{R}^{n\times n}\rightarrow R$,可以用参数 $(u,p)$ 来计算课程 $c$,并返回遗憾值 $P_r(c)$。利用灰盒函数 $\psi$,问题可以重新表示为:

$$\underset{(u,p)\in\mathbb{R}_+^n\times\mathbb{R}_+^{n\times(n-1)}}{\text{Minimize}}\ \psi(u,p) \tag{5}$$

计算 $(\bar{u},\bar{p})$ 的良好估计对于获得良好的数值性能是至关重要的。Francesco 等人提出了一种通过假设证明的方法,如果假设成立,那么对于任意 $(i,j)$,当 $i\neq j$ 时:

$$U(m_i,m_j)=\bar{u}_i+\bar{u}_j-\sum_{k=1,k\neq i}^{n}\bar{p}_{ik}-\sum_{k=1,j\neq k\neq i}^{n}\bar{p}_{jk}+\bar{U} \tag{6}$$

$$U(m_i)=\bar{u}_i-\sum_{k=1,k\neq i}^{n}\bar{p}_{ik}+\bar{U} \tag{7}$$

$$U(m_j)=\bar{u}_j-\sum_{k=1,k\neq j}^{n}\bar{p}_{jk}+\bar{U} \tag{8}$$

## 3 算法分析与总结

强化学习是处理序列决策任务的流行范式[46],尽管在过去的三十年中取得了许多进步,但在许多领域的学习仍然需要与环境进行大量的交互,导致模型的训练时间过长,收敛速度过慢。为了解决这个问题,课程学习被用于强化学习,这样在一个任务中获得的经验可以在开始学习下一个更难的任务时加以利用。然而,尽管课程学习理论、算法和应用研究在国内外已普遍开展,并且也已经取得了较多的研究成果[47-48],但仍然有许多问题还亟待解决。

### 3.1 强化学习中的课程学习算法理论分析与对比

在算法和理论方面,传统课程学习对于小规模的多智能体强化学习性能提升明显,但在大规模多智能体环境中,由于环境和智能体之间的复杂动态以及状态-行动空间的爆炸,因此在实际问题的解决上进展不大[49]。得益于深度神经网络的数据处理能力,使用深度神经网络表示回报函数,避免了特征提取工作,当前基于课程学习的深度强化学习算法在实验场景中应用于 StarCraft[50]、grid-world[51]、hide-and-seek[52]、Sokoban[53]等经典强化学习问题的解决。随着课程学习技术的发展,算法在智能决策[54]、困难编队下的合作导航[55]、在 SUMO 交通模拟器中协商多车辆变道[56]以及在 Checkers 环境下的战略合作[57]等领域也取得了一定的成功。

该综述分四个角度对目前强化学习中的课程学习方法进行分类并介绍,希望能够为相关研究人员提供一点帮助。为方便了解和对比,该文分析、对比了这几类方法的优缺点,并归纳在表 1 中。

表 1 基于课程学习的深度强化学习算法汇总

| 分类 | 优点 | 缺点 | 算法名称 |
|---|---|---|---|
| 基于网络优化的 CL | 适合大规模多智能体场景 | 需要人工生成多主体课程 | DyMA-CL、M-S |
| 基于多智能体合作的 CL | 全局目标和个体目标协作好 | 冲突频繁、方差高、难以推广 | CM3、EPC、任务排序优化框架 |
| 基于能力评估的 CL | 避免从头学习,学习效率高 | 泛化能力差、没有一致的语义 | F-T、GAS、DDaCLAE |
| 基于功能函数的 CL | 泛化能力强,学习探索能力强 | 只能提供最初的启发式解决方案 | PTC、Gray-Box、PS-MAGDS |

(1)基于网络优化的课程学习。解决大规模问题的方法是从小型多智能体场景开始学习,逐步增加智能体的数量,最终学习目标任务。使用多种传输机制以加速课程学习过程,课程设计是影响课程迁移成绩的关键因素。如何选择合适的课程(包括如何决定每个任务的训练步长,如何选择合适的学习模型重新加载等)是至关重要的。如何自动生成多智能体课程可能是目前尚存在的主要局限性,这将在今后的工作中进一步研究[58]。

(2)基于多智能体合作的课程学习。是根据全局目标和个体目标之间的关系进行学习探索,使用信度分配[33]、种群进化课程[34]、任务排序框架[36],通过函数增强方案来连接价值和策略函数的阶段,在具有高维状态空间的多目标多智能体环境中执行高挑战性任务性能较好,缺点是冲突较为频繁、更高的方差和无法维持合作解决方案[59],目前难以推广到非齐次系统或没有已知目标分配的设置的工作。

(3)基于能力评估的课程学习。通过限制其最初行动空间来设置内部课程,使用非策略强化学习同时估计多个行动空间的最优值函数,建立技能、表述和有意义的经验数据集,从而避免从头开始学习,加快学习效率。缺点是集群对每个状态都会改变[60],这可能会干扰泛化,因为没有一致的语义。

(4)基于功能函数的课程学习。通过设定级数函数和映射函数来为智能体量身定制在线课程,通过高斯过程定义智能体函数,学习策略在单位之间共享,以

鼓励合作行为。使用神经网络作为函数逼近器来估计动作-价值函数,并提出一个奖励函数来帮助单位平衡它们的移动和攻击。缺点是只提供最初的启发式解决方案[61],而且质量不能得到保证。

### 3.2　基于课程学习的深度强化学习研究方向

通过对最新课程学习算法理论的研究分析,本节对当前基于课程学习的深度强化学习存在的开放性问题和可能的研究方向进行讨论。

(1)自动创建任务课程。

任务创建是课程学习方法的重要组成部分,任务质量会影响课程的生成质量,任务数量会影响课程排序算法的搜索空间和效率。现有课程学习中的任务大多由人工创建,减少任务创建过程中的人工输入量是未来工作的重要发展方向[62]。

(2)迁移不同类型知识。

课程任务之间,知识必须从一个任务迁移到另一个任务。目前大部分研究中,知识迁移的类型是固定的。例如,Narvekar 等人在任务之间迁移价值函数[63],而 Svetlik 等人迁移成型奖励[64]。这种知识迁移类型的局限性在于,不同的任务对于知识类型的需求可能是不同的,因此可以从不同任务中分别提取知识进行组合。例如,从一个任务中提取一个选项,从另一个任务中提取模型,从而达成更好的学习效果。

(3)课程重用的成本分摊。

当前课程学习方法的另一个局限性是,生成课程的时间可能比直接学习目标任务的时间更长。原因在于,课程通常是为每个智能体和目标任务独立学习的。因此,分摊成本的一种方法是学习一门课程来训练多个不同的智能体[65],或解决多个不同的目标任务。

## 4　结束语

该文对基于课程学习的深度强化学习进行了回顾,由浅入深地对课程学习进行了分析,介绍了课程学习的概念理论、经典算法、研究进展和发展展望等,从基于网络优化的课程学习、基于多智能体合作的课程学习、基于能力评估的课程学习、基于功能函数的课程学习四个角度对强化学习中的课程学习进行了分类梳理、对比分析,最后对基于课程学习的深度强化学习的未来展望进行简要分析。

根据当前深度强化学习中存在的状态空间复杂、维数灾难、学习时间长等问题,课程学习会是未来的一个发展方向。课程学习算法可以将目标任务分解成多个子任务,结合大多数的强化学习算法,使用多种传输机制以加速强化学习进程,大大提高了学习探索效率和通用性。最后,目前课程算法在大规模多智能体场景的研究进展缓慢,其主要原因在于多智能体场景的复杂性。然而大规模多智能体场景更加贴近现实,优质的课程学习算法能够在很大程度上提高学习探索的效率。因此,相信课程学习算法会成为深度强化学习的热门方向,加快深度强化学习的发展速度。

**参考文献:**


[1] MADDEN M G, HOWLEY T. Transfer of experience between reinforcement learning environments with progressive difficulty[J]. Artificial Intelligence Review, 2004, 21(3-4):375-398.

[2] 刘　全,翟建伟,章宗长,等. 深度强化学习综述[J]. 计算机学报,2018,41(1):1-27.

[3] 赵　纯,董小明. 基于深度 Q-Learning 的信号灯配时优化研究[J]. 计算机技术与发展,2021,31(8):198-203.

[4] 陈希亮,曹　雷,何　明,等. 深度逆向强化学习研究综述[J]. 计算机工程与应用,2018,54(5):24-35.

[5] 赖　俊,魏竞毅,陈希亮. 分层强化学习综述[J]. 计算机工程与应用,2021,57(3):72-79.

[6] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. arXiv: 1312.5602,2013.

[7] 万里鹏,兰旭光,张翰博,等. 深度强化学习理论及其应用综述[J]. 模式识别与人工智能,2019,32(1):67-81.

[8] FOGLINO F, CHRISTAKOU C C, GUTIERREZ R L, et al. Curriculum learning for cumulative return maximization[J]. arXiv:1906.06178,2019.

[9] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]//International conference on machine learning. New York: ICML, 2016: 1928-1937.

[10] FANG M, ZHOU T, DU Y, et al. Curriculum-guided hindsight experience replay[J]. Advances in Neural Information Processing Systems, 2019, 19(4):12602-12613.

[11] GU S, HOLLY E, LILLICRAP T, et al. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates[C]//2017 IEEE international conference on robotics and automation (ICRA). New York: IEEE, 2017: 3389-3396.

[12] 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.

[13] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553):436-440.

[14] 任志鹏. 基于自主优先课程学习的深度强化学习算法研究[D]. 南京:南京大学,2018.

[15] NARVEKAR S, PENG B, LEONETTI M, et al. Curriculum learning for reinforcement learning domains: a framework and survey[J]. Journal of Machine Learning Research, 2020, 21(181):1-50.

[16] WANG X, CHEN Y, ZHU W. A survey on curriculum learn-

ing[J]. arXiv:2010.13166,2020.
[17] BENGIO Y,LOURADOUR J,COLLOBERT R,et al. Curriculum learning[C]//Proceedings of the 26th annual international conference on machine learning. New York:ICML,2009:41-48.
[18] 徐 荣. 基于迁移强化学习的无线接入网能耗优化研究[D]. 成都:电子科技大学,2020.
[19] NARVEKAR S,STONE P. Learning curriculum policies for reinforcement learning[J]. arXiv:1812.00285,2018.
[20] IVANOVIC B,HARRISON J,SHARMA A,et al. Barc:backward reachability curriculum for robotic reinforcement learning[C]//2019 international conference on robotics and automation (ICRA). New York:IEEE,2019:15-21.
[21] WANG Y,GAN W,YANG J,et al. Dynamic curriculum learning for imbalanced data classification[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul:IEEE,2019:5017-5026.
[22] HACOHEN G,WEINSHALL D. On the power of curriculum learning in training deep networks[C]//International conference on machine learning. New York:ICML,2019:2535-2544.
[23] TAYLOR M E,STONE P. Transfer learning for reinforcement learning domains:a survey[J]. Journal of Machine Learning Research,2009,10(7):235-258.
[24] MARKOVA V D,SHOPOV V K. Knowledge transfer in reinforcement learning agent[C]//2019 international conference on information technologies (InfoTech). Shanghai:IEEE,2019:1-4.
[25] NARVEKAR S. Curriculum learning in reinforcement learning[C]//Proceedings of the twenty sixth international joint conference on artificial intelligence. Melbourne:IJCAI,2017:5195-5196.
[26] SILVA F L D,COSTA A H R. Object-oriented curriculum generation for reinforcement learning[C]//Proceedings of the 17th international conference on autonomous agents and multiagent systems. New York:ICML,2018:1026-1034.
[27] GUPTA J K,EGOROV M,KOCHENDERFER M. Cooperative multi-agent control using deep reinforcement learning[C]//International conference on autonomous agents and multiagent systems. São Paulo:Springer,2017:66-83.
[28] WEINSHALL D,COHEN G,AMIR D. Curriculum learning by transfer learning:theory and experiments with deep networks[C]//International conference on machine learning. California:PMLR,2018:5238-5246.
[29] SAITO A. Curriculum learning based on reward sparseness for deep reinforcement learning of task completion dialogue management[C]//Proceedings of the 2018 EMNLP workshop SCAI:the 2nd international workshop on search-oriented conversational AI. Brussels:Association for Computational Linguistics,2018:46-51.
[30] WANG W,YANG T,LIU Y,et al. From few to more:Large-scale dynamic multiagent curriculum learning[C]//Proceedings of the AAAI conference on artificial intelligence. New York:AAAI,2020:7293-7300.
[31] WU Y,ZHANG W,SONG K. Master-slave curriculum design for reinforcement learning[C]//Proceedings of the twenty sixth international joint conference on artificial intelligence. [s.l.]:IJCAI,2018:1523-1529.
[32] STONE P,SINAPOV J,TAYLOR M. Curriculum development for transfer learning in dynamic multiagent settings[R]. Austin:University of Texas at Austin,2016.
[33] YANG J,NAKHAEI A,ISELE D,et al. Cm3:cooperative multi-goal multi-stage multi-agent reinforcement learning[J]. arXiv:1809.05188,2018.
[34] LONG Q,ZHOU Z,GUPTA A,et al. Evolutionary population curriculum for scaling multi-agent reinforcement learning[J]. arXiv:2003.10423,2020.
[35] PORTELAS R,ROMAC C,HOFMANN K,et al. Meta automatic curriculum learning[J]. arXiv:2011.08463,2020.
[36] FOGLINO F,CHRISTAKOU C C,LEONETTI M. An optimization framework for task sequencing in curriculum learning[C]//2019 joint IEEE 9th international conference on development and learning and epigenetic robotics (ICDL-EpiRob). Brighton:IEEE,2019:207-214.
[37] WÖHLKE J,SCHMITT F,VAN HOOF H. A performance-based start state curriculum framework for reinforcement learning[C]//Proceedings of the 19th international conference on autonomous agents and multiagent systems. Stanford:AAMAS,2020:1503-1511.
[38] BAKER B,KANITSCHEIDER I,MARKOV T,et al. Emergent tool use from multi-agent autocurricula[J]. arXiv:1909.07528,2019.
[39] FARQUHAR G,GUSTAFSON L,LIN Z,et al. Growing action spaces[C]//International conference on machine learning. California:PMLR,2020:3040-3051.
[40] PORTELAS R,COLAS C,HOFMANN K,et al. Teacher algorithms for curriculum learning of deep rl in continuously parameterized environments[C]//Conference on robot learning. California:PMLR,2020:835-853.
[41] LALOR J P,YU H. Dynamic data selection for curriculum learning via ability estimation[C]//Proceedings of the conference on empirical methods in natural language processing. New York:EMNLP,2020:545-556.
[42] BASSICH A,FOGLINO F,LEONETTI M,et al. Curriculum learning with a progression function[J]. arXiv:2008.00511,2020.
[43] KULKARNI T D,SAEEDI A,GAUTAM S,et al. Deep successor reinforcement learning[J]. arXiv:1606.02396,2016.
[44] MA X,KARKUS P,HSU D,et al. Discriminative particle filter reinforcement learning for complex partial observations[J]. arXiv:2002.09884,2020.
[45] FOGLINO F,LEONETTI M,SAGRATELLA S,et al. A

gray-box approach for curriculum learning[C]//World congress on global optimization. [s. l.]: Springer, 2019: 720-729.

[46] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Massachusetts: MIT Press, 2018.

[47] MANELA B, BIESS A. Curriculum learning with hindsight experience replay for sequential object manipulation tasks [J]. Neural Networks, 2021, 5(3): 66-77.

[48] SRINIVASAN A, BAHDANAU D, CHEVALIER-BOISVERT M, et al. Automated curriculum generation for policy gradients from demonstrations[J]. arXiv: 1912.00444, 2019.

[49] ROMAC C, PORTELAS R, HOFMANN K, et al. TeachMyAgent: a benchmark for automatic curriculum learning in deep RL[J]. arXiv: 2103.09815, 2021.

[50] VINYALS O, EWALDS T, BARTUNOV S, et al. Starcraft ii: a new challenge for reinforcement learning[J]. arXiv: 1708.04782, 2017.

[51] GABOR T, SEDLMEIER A, KIERMEIER M, et al. Scenario co-evolution for reinforcement learning on a grid world smart factory domain[C]//Proceedings of the genetic and evolutionary computation conference. New York: ASME, 2019: 898-906.

[52] CHEN B, SONG S, LIPSON H, et al. Visual hide and seek [J]. arXiv: 1910.07882, 2019.

[53] KARKUS P, MIRZA M, GUEZ A, et al. Beyond tabula-rasa: a modular reinforcement learning approach for physically embedded 3d sokoban[J]. arXiv: 2010.01298, 2020.

[54] HERNANDEZ D, DENAMGANAÏ K, GAO Y, et al. A generalized framework for self-play training[C]//2019 IEEE conference on games (CoG). New York: IEEE, 2019: 1-8.

[55] RACANIERE S, LAMPINEN A K, SANTORO A, et al. Automated curricula through setter-solver interactions[J]. arXiv: 1909.12892, 2019.

[56] FENG D, GOMES C P, SELMAN B. Solving hard AI planning instances using curriculum-driven deep reinforcement learning[J]. arXiv: 2006.02689, 2020.

[57] NETO H C, JULIA R M S. ACE-RL-checkers: decision-making adaptability through integration of automatic case elicitation, reinforcement learning, and sequential pattern mining[J]. Knowledge and Information Systems, 2018, 57(3): 603-634.

[58] ZHANG Y, ABBEEL P, PINTO L. Automatic curriculum learning through value disagreement[J]. Advances in Neural Information Processing Systems, 2020, 33(5): 46-53.

[59] ECOFFET A, HUIZINGA J, LEHMAN J, et al. Go-explore: a new approach for hard-exploration problems[J]. arXiv: 1901.10995, 2019.

[60] KILINC O, MONTANA G. Follow the object: curriculum learning for manipulation tasks with imagined goals[J]. arXiv: 2008.02066, 2020.

[61] KOSTAS J, CHANDAK Y, JORDAN S M, et al. High confidence generalization for reinforcement learning[C]//International conference on machine learning. California: PMLR, 2021: 5764-5773.

[62] TURCHETTA M, KOLOBOV A, SHAH S, et al. Safe reinforcement learning via curriculum induction [J]. arXiv: 2006.12136, 2020.

[63] NARVEKAR S, SINAPOV J, STONE P. Autonomous task sequencing for customized curriculum design in reinforcement learning[C]//Proceedings of the twenty sixth international joint conference on artificial intelligence. Melbourne: IJCAI, 2017: 2536-2542.

[64] SVETLIK M, LEONETTI M, SINAPOV J, et al. Automatic curriculum graph generation for reinforcement learning agents[C]//Proceedings of the AAAI conference on artificial intelligence. New York: AAAI, 2017: 226-232.

[65] NARVEKAR S, STONE P. Generalizing curricula for reinforcement learning[J]. Lifelong Learning Workshop at ICML, 2020, 16(2): 36-47.