面向深度强化学习的对抗攻防综述

刘艾杉[1)] 郭骏[1)] 李思民[1)] 肖宜松[1)] 刘祥龙[1)2)3)] 陶大程[4)]

[1)](北京航空航天大学复杂关键软件环境全国重点实验室 北京 100191)

[2)](中关村实验室 北京 100094)

[3)](合肥综合性国家科学中心数据空间研究院 安徽 230000)

[4)](京东探索研究院 北京 100176)

摘 要 深度强化学习技术以一种端到端学习的通用形式融合了深度学习的感知能力与强化学习的决策能力，在多个领域得到了广泛应用，形成了人工智能领域的研究热点. 然而，由于对抗样本等攻击技术的出现，深度强化学习暴露出巨大的安全隐患. 例如，通过在真实世界中打印出对抗贴纸便可以轻松地使基于深度强化学习的智能系统做出错误的决策，造成严重的损失. 基于此，本文对深度强化学习领域对抗攻防技术的前沿研究进行了一次全面的综述，旨在把握整个领域的研究进展与方向，进一步推动深度强化学习对抗攻防技术的长足发展，助力其应用安全可靠. 结合马尔科夫决策过程中可被扰动的空间，本文首先从基于状态、 基于奖励以及基于动作角度的详细阐述了深度强化学习对抗攻击的进展；其次，通过与经典对抗防御算法体系进行对齐，本文从对抗训练、 对抗检测、 可证明鲁棒性和鲁棒学习的角度归纳总结了深度强化学习领域的对抗防御技术；最后，本文从基于对抗攻击的深度强化学习机理理解与模型增强的角度分析了对抗样本在强化学习领域的应用并讨论了领域内的挑战和开放研究方向.

关键词 对抗样本；对抗攻击；对抗防御；深度强化学习；模型鲁棒性

中图法分类号 TP391

A Survey on Adversarial Attacks and Defenses for Deep Reinforcement Learning

LIU Ai-Shan[1)] GUO Jun[1)] LI Si-Min[1)] XIAO Yi-Song[1)] LIU Xiang-Long[1)2)3)] TAO Da-Cheng[4)]

[1)](State Key Laboratory of Software Development Environment, Beihang University, Beijing, 100191)

[2)](Zhongguancun Laboratory, Beijing, 100194)

[3)](Institute of Dataspace, Hefei, Anhui, 230000)

[4)](JD Explore Academy, Beijing, 100176)

Abstract With the spreading of deep learning, deep reinforcement learning technique has been widely used and drawn extensive research attention in multiple research fields such as robots, games, and auto driving, etc. It is a new learning paradigm associated with the development of deep neural networks, that integrates the perception of deep learning and decision making of reinforcement learning. However, adversarial examples, visually imperceptible perturbations that could mislead deep learning into wrong predictions, have emerged and highly challenged the safety of deep reinforcement learning algorithms and applications especially in the safety-critical scenarios. For example, by simply printing and sticking an adversarial patch on the traffic sign in the real-world scenario, the adversary could easily deceive deep reinforcement learning based auto driving systems into wrong directions and decisions, which would cause severe damage to human lives once successfully attacked. Therefore,

本课题得到科技创新2030新一代人工智能重大项目(No. 2020AAA0103502)，国家自然科学基金 (No. 62022009, 62206009)的资助. **刘艾杉**，博士，助理教授，计算机学会(CCF)会员，主要研究领域为对抗样本、模型鲁棒性、可信赖人工智能.E-mail: liuaishan@buaa.edu.cn. **郭骏**，硕士研究生，主要研究领域为可信赖人工智能. **李思民**,博士研究生，主要研究领域为强化学习、可信赖人工智能. **肖宜松**，博士研究生，主要研究领域为可信赖人工智能. **刘祥龙**（**通信作者**），博士，教授，计算机学会(CCF)高级会员，主要研究领域为计算机视觉，可信赖人工智能.E-mail: xlliu@buaa.edu.cn. **陶大程**,博士，教授，主要研究领域为计算机视觉，人工智能.

extensively studying adversarial examples is highly beneficial for evaluating and better understanding the robustness of deep reinforcement learning, and further increase the safety and reliable applications of reinforcement learning in the safety-critical scenarios. Based on the current circumstances, to better understand and further promote the development of deep reinforcement learning, this paper therefore provides a comprehensive and systematic survey on the research development of adversarial attacks and defenses in deep reinforcement learning area. The primary goal of this paper is to better understand the development and future directions of deep reinforcement learning field, and further promote the studies of adversarial attacks and defenses of deep reinforcement learning, which we hope to lead the safer applications. This paper first presents the preliminary backgrounds including deep reinforcement learning, adversarial examples, and related datasets and benchmarks in deep reinforcement learning field. Based on the perturbing spaces of Markov decision process in deep reinforcement learning, we analyze and summarize adversarial attacks in deep reinforcement learning from the perspectives of state-based, reward-based, and action-based attacks. We then illustrate the framework of adversarial attacks in deep reinforcement learning. By aligning deep reinforcement learning defenses with traditional adversarial defenses framework (e.g., adversarial training, adversarial detection, etc.), we then summarize the adversarial defenses for deep reinforcement learning from adversarial training, adversarial detection, certified robustness, and robust learning. Though similar to the traditional adversarial defense strategies, these methods show quite different implications and application paradigms. Moreover, this paper investigates interesting and meaningful topics for the applications of adversarial examples in the deep reinforcement learning fields, including model robustness understanding and exploiting adversarial attacks for better model performance in deep reinforcement learning. Finally, this paper highlights the open issues and future challenges in the deep reinforcement learning field from four main perspectives, including deep reinforcement learning robustness (theories), multi-agent deep reinforcement learning attacks and defenses (techniques), benchmarks and environment for deep reinforcement learning attacks and defenses (platforms), and physical world adversarial attacks and defenses on deep reinforcement learning (applications). We hope this paper could help the researchers to better understand the framework of adversarial machine learning in the deep reinforcement learning field, and further promote the development and applications of deep reinforcement learning in the safety-critical scenarios in the future.

Key words Adversarial example; Adversarial attacks; Adversarial defenses; Deep reinforcement learning; Model robustness

1 引言

人工智能技术是引领新一轮科技革命和产业变革的战略性技术，已经成为世界各国抢占战略制高点、开展科技竞争的核心领域. 这其中，深度强化学习(Deep Reinforcement Learning, DRL)融合了强化学习的自我激励决策能力和深度学习的抽象表征感知能力，通过赋予智能体自监督学习机制，在不断地与环境交互过程中修正策略并使用深度神经网络的强大表征能力拟合复杂高维的环境特征，形成了人工智能领域新的研究热点. DRL 这种通用性较强的端到端感知控制系统展示出了人类专家级别的能力，并在公共安全、金融经济、国防安全等领域得到了应用，发挥了极其关键的作用[1-4]. 例如，2017 年基于 DRL 的 AlphaGo 系统在复杂的围棋比赛中击败了人类世界围棋冠军[5]; AlphaStar 在星际争霸游戏比赛中战胜了多位人类职业电竞选手，证明了 DRL 在复杂空间中的有效性; DRL 在商业领域的推荐系统中也大放异彩. 这些都充分地展示了深度强化学习技术的重要性、实用性以及非凡的应用价值. 然而，由于现实应用场景的开放性，以大数据训练和经验性规则为基础的深度强化学习方法面临环境的动态变化、输入的不确定性、甚至是恶意攻击等问题，暴露出稳定性、安全性等方面的安全隐患. Christian Szegedy 等人[6]在 2013 年首次发现并提出了出现在计算机视觉领域的对抗样本(Adversarial examples). 这种样本隐藏了微小的恶意噪声，人眼无法区分但会导致人工智能算法模型产生错误的预测结果，对其安全性和可靠性构成了严重的威胁. 除了计算机视觉领域，研究学者还发现对抗样本对于自然语言处

理、 深度强化学习等不同领域和类型的人工智能算法和系统都能够产生较强的迷惑性和攻击性. 更为重要的是, 对抗样本可以在没有目标模型具体信息的条件下轻易地攻破智能系统并迫使其产生攻击者期望的任何输出.

在军事领域和民用公共安全领域存在着大量以深度强化学习为基础的智能应用场景, 如: 智能无人机控制[7]、 智能视觉导航[8]、 车联网计算控制[9]、 异构工业任务控制[10]等, 这些安全攸关的场景对于人工智能的安全、 可靠、 可控有极高的需求. 然而, 基于深度强化学习的智能算法都极易受到对抗噪声的干扰产生不可预期的错误, 甚至可能被误导产生严重的安全问题. 例如, 对抗噪声的攻击可以造成真实世界的自动驾驶系统错误地识别路牌、 做出错误的决策行为, 引发危险事故; 自动导航机器人在遇到对抗噪声攻击后就会执行错误的决策, 执行错误的路径预测, 无法达到预设终点; 在多智能体博弈场景中, 攻击者还能利用某个智能体的对抗行为来诱导其他智能体产生错误的动作、配合, 使其最终输掉博弈比赛[11-17]. 可以看到, 对抗攻击的出现对于深度强化学习的安全、 可靠、稳定应用提出了极大的挑战. 因此, 系统性地分析归纳深度强化学习对抗攻防研究发展脉络和未来方向, 对于深刻认识深度强化学习鲁棒性的研究进展与方向、 进一步解决研究不足之处并推动安全可靠深度强化学习技术的发展都显得尤为重要. 然而, 学术界对于深度强化学习对抗安全的综述性研究却仍十分滞后: 研究人员[12,18-20]于 2018 年和 2020 年对深度强化学习的对抗攻防进行了初步的总结探讨, 然而这些研究距发表至今已数年有余, 缺乏对大量较新研究成果的涵盖,对于领域未来发展脉络的把握也已不足. 在此背景下, 为了系统全面地梳理 DRL 对抗攻防的发展思路、 进一步支撑和推动高安全和可信赖深度强化学习技术的发展, 本文针对深度强化学习算法模型的对抗攻防开展了系统的综述性研究, 从面向深度强化学习的对抗攻防技术的发展现状、 研究历程、 未来趋势进行了详细的讨论.

本文围绕面向深度强化学习的对抗攻防技术展开研究和讨论, 其组织结构如下: 第 1 章介绍本文的研究背景、研究内容等; 第 2 章主要从强化学习和对抗样本两个角度对相关预备知识和概念进行介绍及定义; 第 3 章从基于状态、基于奖励以及基于动作这三个角度对 DRL 的对抗攻击技术进行讨论和分析; 进一步, 第 4 章主要从对抗训练、 对抗检测以及可证明鲁棒性这三个角度对 DRL 的对抗防御算法进行讨论和总结; 在第 5 章中, 本文又进一步归纳并讨论了基于对抗的强化学习机理理解和模型增强, 如: 对抗增智等; 第 6 章结合深度强化学习对抗攻防领域的挑战进行了讨论和分析; 最后, 第 7 章给出本文的结论和未来研究方向.

2 预备知识

2.1 强化学习

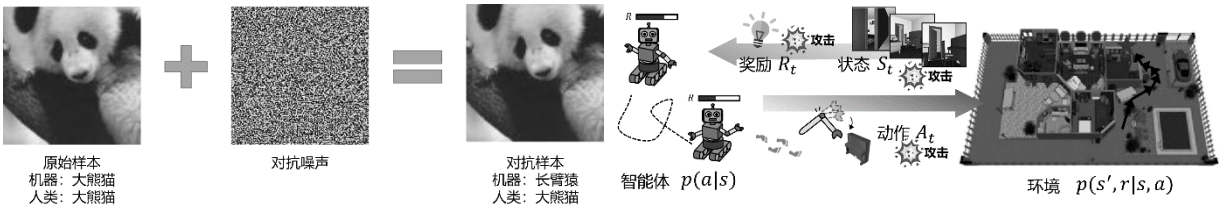
强化学习作为一种机器学习范式, 指的是智能体在环境中探索, 以达成最大化特定目标的学习范式. 强化学习中最重要的两个要素分别为环境(environment)与智能体(agent): 智能体基于环境中的状态(state)进行探索, 给出动作(action); 进一步, 智能体的动作将对环境的状态产生改变, 并产生探索的奖励(reward)[21].

作为理想化的理论框架, 强化学习可以被形式化为马尔可夫决策过程(Markov Decision Process, MDP), 以一个四元组<$\mathcal{S}$, $\mathcal{A}$, $\mathcal{T}$, R>表示. 其中, $S_t \in \mathcal{S}$表示t时刻环境的状态, $A_t \in \mathcal{A}$表示t时刻智能体的动作, $\mathcal{T} = p(s', r|s, a) = p(S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a)$表示环境的状态转移函数, $R_t(S_t = s, A_t = a)$表示t时刻智能体获得的奖励. 强化学习的目标为最大化具有衰减的目标G为 $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$, 其中, $\gamma \in [0,1]$为衰减因子. 基于马尔可夫决策过程, 智能体希望学习策略$\pi(s) = p(A_t = a|S_t = s)$, 以最大化目标$G$. 由于强化学习需要智能体在环境中不断探索以获得最优策略, 具有试错(trial-and-error)的特点; 由于强化学习的最终目标与后续决策有关, 具有延迟奖励(delayed reward)的特点.

在强化学习训练过程中, 基于智能体的策略$\pi(s, a)$可以定义状态-价值函数$V_\pi(s)$与动作-价值函数$Q_\pi(s, a)$. 其表达式分别为:

$$V_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right], \forall s \in \mathcal{S}. \#(1)$$

$$Q_\pi(s, a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a] \# = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s, A_t = a\right], \# \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \#(2)$$



(a) 计算机视觉领域对抗攻击　　(b) 深度强化学习领域对抗攻击

图 1　对抗样本攻击示例. 计算机视觉领域的对抗样本. 在计算机视觉领域, 对抗样本是包含人眼无法识别但却对于深度学习当前决策具有攻击性的样本. 在强化学习领域, 对抗样本攻击使得强化学习智能体产生最差长期奖励.

则$V_\pi(s)$表示当前时刻处于的状态可以获得的最终奖励, $Q_\pi(s,a)$代表与当前时刻、 状态下选取特定动作可以获得的最终奖励. 为训练强化学习智能体, 可以使用基于值函数(value-based)的蒙特卡洛(Monte Carlo)方法[22]和时间差分(Temporal Difference, TD)方法[23]或使用基于策略的(policy-based)策略梯度(Policy Gradient)[24]方法等.

与传统的强化学习相比, 深度强化学习使用深度神经网络估计值函数, 并在 actor-critic 算法中使用深度神经网络构建强化学习策略. 由于深度神经网络的强表征能力与强非线性函数拟合能力, 深度强化学习具有更强的学习能力, 可以收敛到更好的解; 深度强化学习具有更强的泛化能力, 可以一定程度上泛化到未见过的状态-动作对; 深度强化学习具有更强的表征能力, 可以求解一些传统强化学习无法求解的问题, 如围棋. 目前常用的深度强化学习算法包括基于值的深度 Q 网络(Deep-Q Network, DQN)[3],基于策略的近端策略优化算法(Proximal Policy Optimization, PPO)[25]及融合了基于值与基于策略方法的 actor-critic 类方法, 包括 A3C 及 SAC[26,27].

可以看到, 强化学习是一种不同于经典的监督学习的机器学习范式. 监督学习试图基于训练数据预测其标签, 并正确泛化至未经过训练的数据; 但在强化学习中, 由于延迟奖励, 当前状态下的最优动作往往难以定义, 且在智能体与环境交互的过程中, 足以代表训练环境的数据往往难以获取, 使得监督学习很难被用于解决强化学习问题. 此外, 强化学习的试错特点更适合在一个全新的环境中, 在不依赖数据标签的情况下进行探索.

2.2　对抗样本

2013 年, 美国谷歌公司的 Szegedy 等研究人员[6]第一次发现并定义了出现在计算机视觉领域的对抗样本. 这种微小的噪声对于人眼无法产生影响, 但是却会直接造成深度神经网络模型产生错误预测. 在计算机视觉分类任务中, 对抗样本的定义如下:

$$f_\theta(\mathbf{x}_{adv}) \neq \mathbf{y} \quad s.t. \quad \|\mathbf{x}-\mathbf{x}_{adv}\| \leq \epsilon. \#(3)\#$$

其中, $\mathbf{x}$是原始的数据样本, $\mathbf{x}_{adv}$是含有对抗噪声的对抗样本, $\mathbf{y}$是原始样本$\mathbf{x}$的类别标签, $\|\cdot\|$用来衡量$\mathbf{x}$和$\mathbf{x}_{adv}$的差别距离足够小, 但是神经网络f_θ对对抗样本$\mathbf{x}_{adv}$进行了错误分类.

如图 1(a)所示, 在计算机视觉领域, 由于这种人类视觉感知的特殊性, 对抗样本所带来的安全性问题是极其严峻的. 除了计算机视觉领域, 研究者还发现对抗样本对于自然语言处理、 语音识别等不同领域, 统计机器学习、 深度学习、 强化学习等不同类型的人工智能算法和系统, 都能够产生很强的迷惑性和攻击性, 可以在没有目标模型具体信息的条件下进行黑盒攻击. 更为重要的是, 对抗样本不仅仅存在于实验室研究的数字环境中, 其在真实世界中也同样具有很强的攻击能力[28-32]. 目前, 研究人员对于对抗样本攻击进行了广泛的研究, 成功地针对不同真实智能应用进行了攻击, 例如: 自动驾驶[29,33]、 人脸识别[34]、 机器人导航[16]、 自动零售[35,36]、 目标检测[37,38]等.

由于强化学习的延迟奖励特性与目标导向特性, 诱使智能体做出与当前策略给出的最优动作不同的动作不一定会导致最差的长期影响, 在不同时刻进行的攻击同样对最终目标具有不同的影响力. 因此, 与传统计算机视觉中对抗攻击思路不同, 仅仅令模型产生更多次的决策错误并不能最大化地影响强化学习的最终目标函数. 相反, 强化学习中的攻击者假设可以使用的干扰集$B(\cdot)$, 其最终目标为生成干扰噪声$\nu(\cdot)\sim B(\cdot)$以最小化智能体目标函数, 即$\min\limits_{\nu(\cdot)} G_t = \sum_{k=0}^{\infty}\gamma^k R_{t+k+1}$. 如图 1(b)所示, 根据马尔可夫决策过程四元组中可以被扰动的空

间分类，对抗攻击者可以直接在状态$\mathcal{S}$、奖励R以及动作$\mathcal{A}$中进行扰动从而影响模型最终的决策. 回顾公式 3 中经典的对抗样本定义，$\mathbf{x}_{adv}$分别对应于在强化学习的马尔科夫决策过程中的状态空间、奖励空间和动作空间中加入噪声而形成的对抗输入.

2.3 强化学习实验环境与平台

进行面向深度强化学习的对抗攻防研究必须要有良好的环境和数据集的支撑. 不同于传统提供数据集进行训练和测试的机器学习任务，强化学习所有的训练和测试数据均来源于智能体与环境的交互. 针对这种交互式环境的需求，许多知名企业和研究单位针对强化学习任务开发了多款研究环境与平台，开放给强化学习领域的研究者使用.

目前强化学习领域最知名的平台当属 OpenAI Gym[39]. Gym 是非营利组织 OpenAI 于 2016 年发布的一款强化学习研究平台，其提供了丰富而多样的强化学习环境，以部分观测的马尔科夫决策过程形式建模并给出交互接口. 更重要地是，Gym 定义了一种 Python 环境下与强化学习环境交互的通用范式，其 API 格式被后来的强化学习环境广泛采用. Gym 种包含的环境可以分为五类: 经典控制、算法型任务、雅达利游戏、棋牌类游戏、二维和三维机器人控制. 自提出以来，Gym 一直在更新迭代，提供新的强化学习环境和更丰富的功能，目前已有数十种强化学习环境可供使用. 比较经典的环境如 CartPole、Pong、MsPacman 等广泛出现在相关研究论文中.

MuJoCo[40]是一款用于机器人控制研究的高级物理引擎，属于 DeepMind 公司，其也作为一款强化学习实验平台为众人所知. MuJoCo 作为底层引擎，其本身能够提供复杂的控制细节，但需要用户对场景内的可控制单位进行细致的设定，因此也有研究人员制作了完备的场景供大家使用. 如 OpenAI Gym 中，机器人控制部分的环境均基于 MuJoCo 作为底层引擎进行搭建，可以配合 Gym 本身的强化学习接口进行使用.

开放赛车模拟器(The Open Racing Car Simulator, TORCS)[41]是一款多平台的赛车模拟器，可以用于普通的赛车游戏，也可以作为强化学习环境用于研究. TORCS 提供了完整的模拟功能，从赛车的性能定制、车辆控制到 3D 环境渲染，均有接口与界面作为支撑. 该环境也可与 OpenAI Gym 提供的强化学习接口良好适配，为强化学习研究提供便利.

除了以上提到的平台，也有多智能体强化学习任务专用的游戏平台. 例如: PettingZoo[42]，一款类似于 Gym 的多智能体任务平台，其包含了许多类型的环境，除了雅达利游戏与经典控制任务外，还有具有大量可控制智能体的环境；PySC2 平台[43]，DeepMind 公司与暴雪游戏公司基于星际争霸 2 游戏合作开发的一款多智能体游戏平台，整合了星际争霸 2 游戏的机器学习 API 并作为 Python 环境发布；进一步，Whiteson Research Lab 发布了一款简化版的星际争霸 2 多智能体挑战(StarCraftII Multi-Agent Challenge, SMAC)[44]，其已成为合作型多智能体强化学习任务中最常用的平台之一.

以上介绍的强化学习实验环境的覆盖范围十分广泛，从离散动作到连续动作、基础电子游戏到真实场景模拟、简单决策任务到复杂规划任务均有涉及. OpenAI Gym 提供的平台接口通用，包含环境多样，可扩展性强，且一直在更新迭代，是各大强化学习论文中最常使用的实验平台；但其缺点是: 默认提供的环境大多为电子游戏或简单任务，所得到的强化学习模型不具备现实意义，无法真正用于某个生产场景中，因而 Gym 一般作为衡量强化学习算法能力的基准环境进行使用. PettingZoo 的定位与 Gym 类似，但其主要针对于多智能体任务进行设置和优化，其知名度和普及度略逊于 Gym. MuJoCo 和 TORCS 偏向于真实场景，提供现实世界的建模功能，其提供的仿真功能均有真实物理定律作为支撑. 若在这些环境上进行恰当的设置，得到的强化学习模型可以用于物理世界中执行真实任务；但它们的缺点是: 运算开销大，需要用户自行设置大量参数，且对深度学习框架和代码的适配性不如 OpenAI Gym. PySC2 环境完全基于星际争霸 2 游戏进行开发具备即时战略游戏独特的决策系统，强化学习算法需要处理的任务复杂度远高于 Gym 等环境内置的任务，因而该环境时常用于评估一些大型算法的决策能力；SMAC 环境是基于 PySC2 二次开发的合作对战环境，其任务相对简单，也经常在各种论文中作为合作型多智能体强化学习任务的基准环境出现.

3 面向深度强化学习的对抗攻击技术

基于上文所述，本文将对强化学习的攻击分为基于状态$\mathcal{S}$、基于奖励R以及基于动作$\mathcal{A}$三种攻击

方式, 并按照这三种方式进行归纳总结(如表 1 所示). 其中, 基于状态$\mathcal{S}$的攻击通过扰动智能体观测或者改变智能体观测结果, 从而诱使智能体做出最小化目标函数的决策; 基于奖励R的攻击通过微小地扰动智能体训练过程中的奖励函数, 从而影响智能体的全局策略; 基于动作$\mathcal{A}$的攻击直接对智能体的动作进行微小扰动, 从而大幅影响智能体的目标函数, 或通过训练具有对抗策略的智能体从而影响其他智能体决策. 对应至公式 3 中经典对抗样本$\mathbf{x}_{adv}$的定义, 强化学习中的对抗攻击分别从$\mathcal{S}$、 R和$\mathcal{A}$三个空间中加入噪声进行对抗攻击. 从攻击者的角度来看,基于状态和奖励的攻击需要能够获取到模型的控制权, 相比基于动作的攻击更加困难一些.

3.1 基于状态的对抗攻击

在这一节中, 本文将梳理和归纳基于状态的深度强化学习对抗攻击算法. 我们将基于状态的攻击(如图 2)分为两类: 基于观测的对抗攻击与基于环境的对抗攻击. 其中, 基于观测的对抗攻击主要通过扰动智能体的观测值s, 从而改变智能体策略$\pi(s)=p(s|a)$来实现攻击; 基于环境的对抗攻击在环境中添加对智能体观测值s的扰动的同时, 还要求此扰动符合状态转移方程$\mathcal{T}=p(s^{'},r|s,a)$; 对于算法开销而言, 如果攻击方式仅对强化学习的单步决策进行攻击, 则攻击者通过规则直接确定强化学习需要扰动的变量, 并使用模型梯度直接生成可以攻击强化学习策略网络的噪声, 攻击开销较小. 如果攻击方式需要对强化学习的整体策略进行规划, 则攻击者所做出的决策则需要通过求解马尔可夫决策过程, 即训练一个攻击者具有的强化学习智能体获取. 随后, 攻击者在攻击阶段基于其训练的强化学习智能体生成目标噪声. 这类方法由于需要训练强化学习智能体, 攻击开销中等.

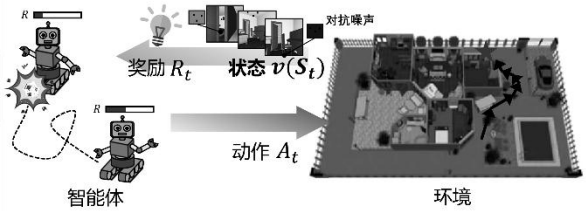


图 2 基于状态的对抗攻击算法示意图. 对抗攻击者通过扰直接动智能体的观测值s或在环境中添加对抗噪声来影响智能体的最终决策.

3.1.1 基于观测的对抗攻击

扰动智能体观测的对抗样本的目标为: 对于智能体状态s, 给定一系列允许的扰动$B(s)$, 令$v(s)\in B(s)$扰动后的智能体观测状态, 其使得强化学习达成的目标G最小:

$$\min_{\theta} G_t=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1},\#$$
$$where\ \ v_\theta(s)\in B(s), s^{'}\sim\mathcal{T}(s,a), a\sim\pi\big(v(s)\big),\#(4)$$

其中, θ为$v(s)$具有的参数. 通过生成添加在状态s上, 且在允许扰动$B(s)$范围内的噪音, 攻击者的目标是最小化被攻击者的总奖励函数. 值得注意的是, 在此过程中, 真实环境的状态转移方程$p(s^{'},r|s,a)$不变. 具体而言, $v(s)$通过扰动智能体的观测从而扰动智能体行为$a_t\sim\pi(a|v(s_t))$, 最终对于强化学习策略产生影响. 由$v(s)$与$\pi(a|s)$定义的马尔可夫决策过程中的$\tilde{V}_{\pi\circ v}(s)$与$\tilde{Q}_{\pi\circ v}(s,a)$仍遵循贝尔曼方程. 若假设强化学习可以估计最优$V^*(s)$与$Q^*(s,a)$, 则其最优扰动$v^*(s)$可被简化为:

$$\tilde{V}^*_{\pi\circ v^*}(s)=\min_{v^*}\tilde{V}^*_{\pi\circ v^*}(s),\#\#$$
$$\tilde{Q}^*_{\pi\circ v^*}(s,a)=\min_{v^*}\tilde{Q}_{\pi\circ v^*}(s,a).\#(5)$$

传统的对抗攻击通过为模型的输入添加特定构造的对抗噪声从而使得模型输出错误结果. 相似地, 在深度强化学习中, 通过向智能体的观测添加对抗噪声从而实现对 DRL 的对抗攻击是一种非常直观的思路. 2017 年, Huang 等人[45]第一次将对抗攻击方法引入深度强化学习中. 如图 3 所示, 作者使用简单的 FGSM 对抗攻击, 针对于深度强化学习智能体的输入空间加入微小的对抗噪声, 可以轻松地使得包括 DQN、 TPPO、 A3C 在内的多种不同的经典强化学习算法做出错误的决策. 通过在 Atari 2600 游戏集[39]中的 Chopper Command、 Pong、 Seaquest、 Space Invaders 四种游戏中进行包含白盒、 黑盒场景在内的广泛测试, 作者发现强化学习和传统的监督学习一样, 对于对抗样本这种微小的噪声也显示出非常脆弱的表现. 该文章也指出, 在深度强化学习领域, 对于不同策略和不同的训练算法, 对抗攻击都具有一定的攻击迁移性.

在上述实验的基础上, 研究者们改进并提出了新的攻击算法. Lin 等人[46]指出, 对强化学习过程中的每一帧进行攻击虽然具有较强的攻击效果, 但这种攻击方式过于“视觉明显”, 且强化学习模型攻击是否成功应由最终奖励值而非通过改变动作的百分比决定. 因此, 该论文提出策略化时间攻击与诱导攻击两种攻击方式, 用于高效地攻击强化学习

表 1　对抗攻击技术总览

文献	针对的强化学习算法	实验环境	测试阶段攻击		训练阶段攻击		攻击技术	攻击开销
			白盒攻击	黑盒攻击	白盒攻击	黑盒攻击		
Behzadan 等人[13]	DQN	pong	✓	✓		✓	基于状态的攻击	低
Liu 等人[16]	PACMAN-RL+Q[52]	EQA-v1 dataset[52]		✓	✓	✓	基于状态的攻击	低
Huang 等人[45]	DQN, TRPO, A3C	chopper command, pong, seaquest, space invaders	✓	✓		✓	基于状态的攻击	低
Lin 等人[46]	DQN, A3C	pong, seaquest, mspacman, chopper command, qbert	✓				基于状态的攻击	低
Inkawhich 等人[47]	DQN, PPO	pong, breakout, space invaders, seaquest		✓			基于状态的攻击	低
Russo 等人[48]	DQN, DRQN, DDPG	CartPole, discrete MountainCar, continuous MountainCar, continuous Lunar-Lander		✓			基于状态的攻击	中
Tretschk 等人[49]	DQN	pong	✓				基于状态的攻击	中
Xiang 等人[50]	Q-learning	自动寻路		✓			基于状态的攻击	低
Bai 等人[51]	DQN	自动寻路			✓		基于状态的攻击	中
Xiao 等人[53]	DQN, DDPG	TORCS[54], Atari(pong, enduro), MuJoCo(half-cheetah, hopper)	✓	✓			基于状态的攻击	中
Behzadan 等人[55]	DQN, A2C, PPO	Cartpole		✓			基于奖励的攻击	低
Zhang 等人[56]	Q-learning	Grid World				✓	基于奖励的攻击	低
Han 等人[57]	DDQN, A3C	软件定义网络	✓	✓	✓		基于奖励的攻击	低
Gleave 等人[14]	PPO	MuJoCo: kick and defend, you shall not pass, sumo humans, sumo ants		✓			基于动作的攻击	中
Hussenot 等人[15]	DQN, Rainbow DQN	pong, space invaders, air raid, HERO	✓		✓		基于动作的攻击	中
Lee 等人[58]	PPO, DDQN	Lunar-Lander, BiPedal Walker			✓		基于动作的攻击	低
Liu 等人[59]	UCB-H[60], UCB-B[60], UCBVI-CH[61]	自定义 MDP	✓	✓			基于动作的攻击	低
Wu 等人[62]	PPO	MuJoCo: you shall not pass, pong		✓			基于动作的攻击	中
Guo 等人[63]	PPO	MuJoCo: kick and defend, you shall not pass, sumo humans, sumo ants, StarCraftII		✓			基于动作的攻击	中
Wang 等人[64]	DDPG	SUMO 交通模拟器[65]				✓	基于动作的攻击	中

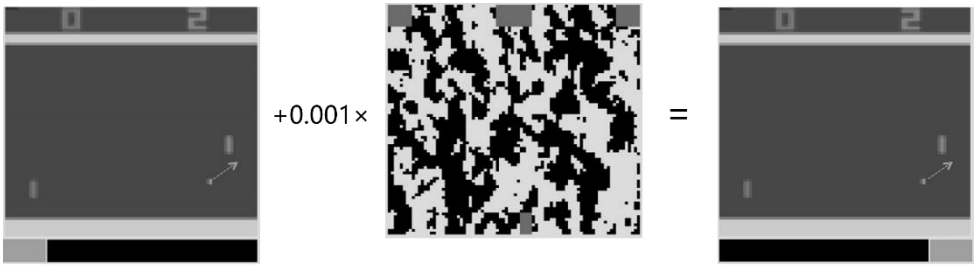


图 3 基于 FGSM 方法生成对抗样本. 其通过在输入图像上加入扰动（如中间部分放大后的对抗噪声扰动所示）对智能体的输入空间观测进行攻击, 从而迷惑智能体决策动作(图片下方框为当前动作决策分布)[45].

智能体. 策略化时间攻击基于强化学习策略给出的动作概率, 在特定的、 危害性最大的时刻攻击. 若某一时刻智能体策略给出某动作的概率极大, 而使用另一动作的概率极小, 则说明此时刻是强化学习的关键决策点.而诱导攻击希望将智能体诱导至某一个特定的位置. 为了达成这一目的, 作者首先预测下一帧的情况, 随后使用规划算法规划下一步诱导智能体到达的位置. 该论文在 Atari 2600 游戏集中的 5 个游戏上实验了他们的攻击方法, 使得智能体奖励大幅下降, 攻击成功率超过 70%.

Inkawhich 等人[47]针对传统对抗攻击方法依赖于白盒模型、 需要智能体训练的环境、 可迁移性弱的问题, 提出模型窥探攻击(model snooping attack). 此攻击假设攻击者无需与模型训练的环境交互和获取模型内部参数, 仅窃听部分模型的动作或奖励函数, 即可对模型进行高成功率的攻击. 攻击者基于窃听到的信息训练一个代理模型(substitute model), 并使此代理模型在一个与目标模型相近的任务上面进行学习. 由于对抗样本的迁移性, 可以成功攻击代理模型的对抗样本同样可以成功攻击目标智能体. 该论文在 Atari 游戏集中的 Pong、 Breakout、 Space Invaders 与 Seaquest 游戏中进行了测试, 表明作者提出的训练方法对于强化学习具有较强的攻击性, 且与攻击使用目标模型所有信息训练的替代模型相比, 其攻击性能下降不大. 相似地, Behzadan 等人[13]也通过引入一个代理模型来使用基于迁移的对抗攻击. 该文章引入了一个攻击者, 在第t时刻, 攻击者会根据前面m个时刻的智能体的观测和行为序列来预测出$t+1$时刻时让智能体做出不好行为所需要修改的像素信息, 然后基于此修改$t+1$时刻的输入图片并作为智能体的输入.

为了进一步提升在黑盒场景的攻击效果, Russo 等人[48]提出了一种基于黑盒的针对强化学习的对抗攻击算法. 该论文认为, 传统的基于梯度的白盒对抗攻击并没有取得最优的攻击效果, 而且在更加真实的情况下, 对抗者无法获取智能体的策略和参数. 因此, 作者提出了一种基于黑盒的对抗攻击方法: 将生成对抗样本视为一个求解马尔可夫过程的问题, 并基于此提出了一种基于 DDPG 算法的优化算法来求解对抗策略, 使其可以对于当前的状态产生微小的噪声来有效攻击智能体.

Tretschk 等人[49]指出, 前人提出的攻击方法大多只攻击强化学习中的一步, 没有考虑强化学习中由于智能体策略带来的长期影响. 基于此, 作者提出对抗 Transformer 网络(Adversarial Transformer Network, ATN). ATN 在固定智能体策略网络的情况下, 在智能体的观测前加入 ATN 模块,从而在智能体策略中的每一步均加入攻击,使得被攻击的强化学习智能体得到任意的最终奖励. 因为扰动了智能体的每一步观测, 该方法具有较强的攻击效果. 其中, ATN 模块在原智能体策略网络固定的情况下进行训练, 可以被视为一个全新的强化学习智能体, 并将原智能体策略网络当作环境中的一部分进行训练.

3.1.2 基于环境的对抗攻击

基于观测的对抗攻击需要在多个时间点直接修改被攻击的智能体的观测, 这意味着攻击者具有被攻击者模型系统的直接访问权限. 显然, 这难以在现实场景下实现. 相反, 在智能体所在的环境中添加扰动在现实情况下更易实现也更为隐蔽, 具有更强的实际意义. 与基于观测的对抗攻击不同的是: 在环境中添加的扰动需要随着智能体在环境中的探索随环境$\mathcal{T}(s',r|s,a)$变化, 而非任意变化. 给定允许对于环境产生的扰动集$B(s)$, 令$v(s)\in B(s)$表示对于真实环境的扰动, 则环境中的状态转移方程相应变为$p(v(s'),r|v(s),a)$. 扰动环境的对抗样本可以被表达为:

图 4　基于环境的对抗攻击示例[45]. 通过在仿真环境中的 3D 目标物体上加入对抗纹理信息，该方法能够有效的迷惑 EQA 机器人，让其对于问题给出错误的回答或执行错误的路径规划.

$$\min_{\theta}\ G_t=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1},\#\#$$
$$where\ v_\theta(s)\in B(s), v(s')\sim\mathcal{T}(v(s),a), a\sim\pi\big(v(s)\big).\#(6)$$

其中，$v(s)$与$v(s')$指的是对于状态的扰动无法随着时间动态改变，而是随着环境的改变而改变. 攻击者对于环境的攻击已经确定，在训练过程中就无法更改.

针对自动寻路的 Q-learning 强化学习算法，Xiang 等人[50]使用了一种基于主成分分析(Principal Component Analysis, PCA)的攻击方法在地图中加入一些额外的障碍点，从而生成这种特殊的对抗样本攻击，使得模型寻路出错. 这篇文章人工分析并定义了对路径产生影响的因子，计算这些因子的值并构建矩阵，对该矩阵进行 PCA 降维，可以计算出概率加权线性组合的权值从而实现攻击. 然而，该方法不同于传统对抗攻击，更多依赖于人工设定的先验条件.

在相似的任务场景下，Bai 等人[51]进一步研究了 DQN 在机器人寻路场景下通过 Q Table 进行白盒攻击. 该论文将 DQN 应用于机器人自动寻路，并分析 DQN 的寻路规则策略，之后提出了一种有效寻找白盒 Q table 中脆弱点的攻击方法. 该论文从隐式对抗方法出发，针对不断试错的 DQN 寻路方法，发现其弱点. 文章采用 15×15 的 Grid-World map 作为实验场景，证实在潜在攻击点的障碍攻击能够干扰训练，延长路径收敛需要的训练时间.

为了使得针对强化学习的攻击在真实环境中具备更高的实用性，Xiao 等人[53]创新性地提出了一种基于环境动力学(environment dynamics)的攻击. 在真实场景中，攻击者往往不能获得被攻击模型的参数、 结构等信息，也不能轻易扰动模型的奖励、动作等内部数据，这篇文章通过随机采样和基于强化学习的对抗性采样等方法修改了环境中的物理属性，在数字世界和物理世界中进行了实验，验证了其方法的有效性. 在此之后，Prithvijit 等人[66]构造了 RobustNAV 的仿真环境，集成了多种视觉和环境动力学的攻击噪声，用于评测深度强化学习导航任务的鲁棒性.

除了上述传统任务外，在复杂的多模态 Embodied Question Answering (EQA)任务中[52]**Error! Reference source not found.**，Liu 等人[16]提出了一种时空融合的对抗攻击方法. 在该任务下，智能机器人需要理解给定的使用自然语言描述的问题，然后通过在仿真环境中进行第一视角导航，完成对应的任务并回答问题. 这类智能体通常使用强化学习的方法进行训练. 研究人员提出的时空融合对抗攻击方法有效利用了路径注意力机制选取智能体最为关注的时序帧，并在其中出现的 3D 目标物体上加入对抗纹理(如图 4 所示)，有效地攻击了强化学习智能体.

3.2　基于奖励的对抗攻击

基于奖励的对抗攻击(如图 5)主要是对目标策略的回报奖励加入对抗噪声进行干扰，影响智能体的学习过程，尽可能减少所学策略的回报，从而达到攻击目的. 假设$v_\theta(r)$为扰动后的奖励函数，则此问题可被形式化为:

$$\min_{\theta}\ G_t=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1},\#$$
$$where\ r\leftarrow v_\theta(r), s'\sim\mathcal{T}(s,a), a\sim\pi(\cdot\,|s).\#(7)$$

此处，$v(r)$表示攻击者对奖励函数的扰动. 虽然训练过程中使用被扰动的奖励函数，但最终评估过程中的奖励函数仍然使用干净的奖励函数计算. 对于算法开销而言，攻击者多基于启发式算法或自适应算法生成对于奖励函数的扰动，算法开销较小. Behzadan 等人[55]从模型窃取的角度展开了研究探索，发现了深度强化学习模型可以被模型窃取的方式所攻击. 文章首先采用 Deep Q-Learning from Demonstrations (DQfD)的方式学习一个模拟的目标策略，将其作为对抗策略的初始值；接着，结合目标策略的回报变化和扰动次数作为对抗策略的回报函数进行 Q-learning 训练. 实验使用基于 DQN、A2C 和 PPO2 算法的智能体作为目标策略进行攻击，证实了该对抗策略的有效性与可迁移性.

除了模型窃取的对抗攻击方式, Zhang 等人[56]将对抗攻击中的数据投毒方法引入强化学习. 在智能体训练阶段, 对于给定的状态和动作得到的奖励, 对抗攻击者通过在奖励中加入对抗噪声使智能体学到错误的策略. 基于这种思想, 文章提出了一种自适应攻击方法, 能够使用较少的训练轮次毒害攻击智能体. 文章还对所提出的训练阶段数据投毒的攻击方法进行了大量的理论证明, 得到了多个在奖励回报中加入噪声的上下界(如: 确保智能体安全的噪声上界、 攻击成功的噪声下界等), 并在机器人自动寻路应用场景下进行了仿真验证.

除了在经典的深度强化学习任务中进行攻击外, Han 等人[57]还探索了在软件定义网络(Software-Defined Networking, SDN)应用场景中对奖励进行对抗攻击.SDN 是指通过使用程序接口集中控制网络的行为.文章中提出了两种攻击 SDN 的方式,使 SDN 中的关键服务器失效:第一种攻击方式为翻转奖励函数的值,通过黑客劫持或传感器错误,在极少的时间内翻转强化学习训练使用的信号;第二种攻击方式为对于强化学习智能体接收到的状态进行扰动,从而使强化学习智能体无法做出最优的策略.

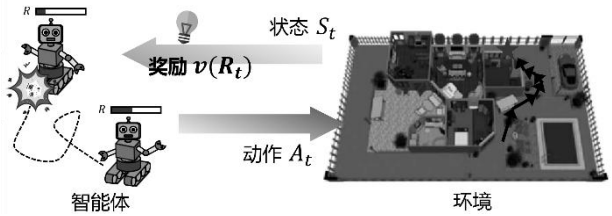


图 5 基于奖励的对抗攻击算法示意图. 对抗攻击者通过在训练过程中对目标策略的回报奖励加入对抗噪声进行干扰, 影响智能体的学习过程, 从而达到攻击目的.

3.3 基于智能体动作的对抗攻击

在基于观测和基于奖励的对抗攻击外, 业内也存在不少工作从强化学习智能体行为动作方面展开对抗攻击的研究(如图 6). 一方面, 可以通过直接扰动智能体策略输出动作的概率来进行攻击; 另一方面, 可以引入另一个智能体, 使其具备对抗性策略并做出攻击性动作, 造成原智能体回报大幅下降.

对于动作进行攻击的对抗样本由允许对于智能体策略$\pi(\cdot|s)$产生扰动的集合$B(\pi)$定义. 令$v(\pi) \in B(\pi)$表示对于智能体策略的扰动, 则此问题可被形式化为:

$$\min_{\theta} G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \#$$
$$where\ s' \sim \mathcal{T}(s,a), a \sim v_\theta(\pi(\cdot|s)). \#(8)$$

其中, 攻击者v直接修改攻击者做出的动作概率$\pi_\theta^v(\cdot|s)$, 其攻击目标为最小化被攻击者的总奖励函数. 对于算法开销而言, 基于动作概率的攻击多将攻击者本身视为一个智能体, 并训练攻击智能体的策略以最大化攻击效果, 算法开销中等.

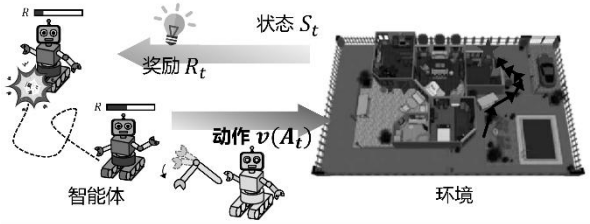


图 6 基于动作的对抗攻击算法示意图. 对抗攻击者通过直接扰动智能体输出动作概率或者引入一个具备对抗性策略的智能体进行攻击.

3.3.1 基于动作概率的攻击

2020 年, Lee 等人[58]首先提出了一套清晰有效的针对智能体动作空间的攻击方法. 文章首先提出了一种“目光短浅”(myopic)的动作空间攻击算法, 即: 攻击只考虑对当前状态下造成的回报奖励影响最大. 之后, 文章将该方法扩展为一种“目光长远”的攻击方式(Look-ahead), 可以连续扰动多个时间点的动作, 考虑这些扰动对回报奖励的总体影响. 具体来说, 为动作加入一个扰动, 该扰动的大小由ℓ_p范数约束, 而其扰动效果由加入扰动后的回报奖励来描述. 该攻击与传统对抗攻击的 FGSM 比较相似, 但只对一个动作进行扰动, 并最小化扰动之后的回报奖励函数. 该方法会用到回报奖励函数的梯度, 但由于回报奖励本身不可导, 因此文章采用了代理回报奖励去代替回报奖励函数, 使得该过程可以求导. 该方法能够在连续动作空间环境取得较好的效果, 但是很难应用在离散动作空间的任务上, 是对抗攻击在强化学习领域的“牛刀小试”.

更进一步, Liu 等人[59]提出通过扰动智能体输出的动作信号来干扰智能体所学到的策略. 在白盒条件下, 文章提出α-portion 攻击方法, 可以使用次线性的损失和代价来攻击. 在黑盒条件下, 文章提出 LCB-H 攻击方法, 是一种可证明有效的攻击方法, 在 UCB-H 上进行了攻击, 以对数级别的代价让 UCB-H 模型按攻击者的设计选择动作.

3.3.2 基于对抗策略的攻击

一系列研究工作尝试训练出一个具有对抗策略(adversarial policy)的强化学习智能体. 具备这种

对抗策略强的智能体将会做出具有对抗攻击性的行为，迫使另一方智能体观测后作出错误的行为.

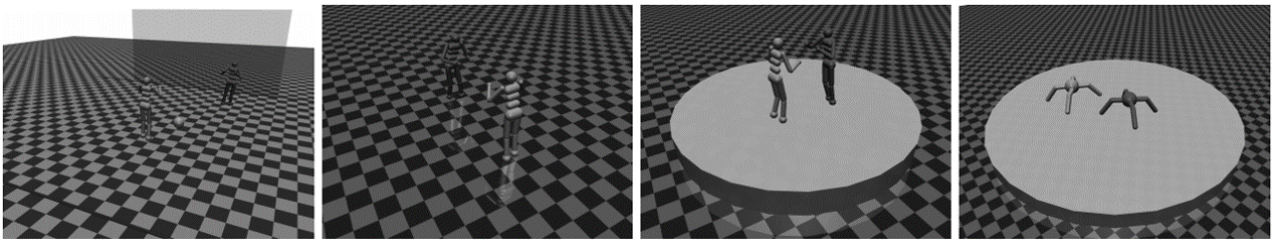


图 7　基于对抗策略的对抗攻击场景[14]. 在多智能体对抗的环境中引入一个带有对抗性策略的智能体，该智能体会做出具有对抗攻击性的行为并诱导另一方智能体产生错误的预测和行为.

Gleave 等人[14]第一次创造性地提出了对抗性策略的概念. 这是一种新的攻击方式，具备这种对抗策略的智能体将会做出具有对抗攻击性的行为，通过对抗性策略在共享环境中采取的行动将诱导另一方智能体产生错误的预测和行为. 该论文在 MuJoCo[40]的四个环境上进行了验证实验，证明了在仿真机器人游戏零和博弈中对抗策略的存在和有效性，如图 7 所示. 与上文类似，Wu 等人[62]着眼于双智能体游戏，通过操纵一方的行为让另一个正常智能体输掉游戏. 这篇文章基于 PPO 算法，指导攻击者训练一个对抗智能体. 文中假设对抗智能体可以获得对手的观测和动作，并通过最大化被攻击智能体采取行动的偏差，根据被攻击智能体的注意力表现指导攻击者改变行为，最终引导对手给出次优动作. Hussenot 等人[15]提出了一种称为 CopyCAT 的目标攻击方法用于攻击强化学习智能体. CopyCAT 方法核心在于：训练一个具有相反效果的对抗策略，通过预先生成的一系列掩码，将正常策略的动作替换为决策出的动作.

为了解决传统对抗策略中游戏双方零和假设的问题，Guo 等人[63]提出了一种新的对抗策略，通过最大化攻击策略的平均期望和最小化受害者的平均奖励，确保对抗性策略的更新不会导致学习目标的随机波动. 该论文重新构建了一个双人游戏，将攻击者和受害者的预期奖励定义为对抗策略的函数，从理论上保证了整个策略学习过程的单调性. 该论文固定了受害者的策略，训练对抗智能体来打败原先的智能体，在传统通过与被攻击智能体交互进行攻击基础上，进一步假设攻击者知道被攻击者的即时奖励，从而实现更强的攻击. 该论文提出的方法可以对复杂的游戏进行有效的、 实际可行的对抗性攻击，并第一次在星际争霸复杂游戏上成功完成了攻击.

更进一步，Wang 等人[64]在自动驾驶控制场景下实现了对抗策略攻击，基于交通物理中已有的原则，提出了一种后门触发器的设计方法. 自动驾驶的场景选在“stop-and-go”场景(即，堵车时汽车的“停下”和“行驶”). 作者提出了两种攻击方式：使得交通车流聚集的拥挤攻击和使得车辆加速导致碰撞的安全攻击. 触发器的设计使得车辆在遇到触发器时进行加速或减速动作. 加速会带来碰撞，而减速会带来拥挤. 其实验结果表明，后门攻击后的模型不会对正常的行驶带来影响，仅会使得累计回报下降 1%，但是却能够在观测到触发器时，引发拥挤或安全问题. 然而，这篇文章中对后门触发器的设计方法与自动驾驶任务高度相关，不具有普适性.

3.4　小结

在本章中，我们系统性地介绍了近年来深度强化学习领域对抗攻击的研究，并从基于状态、基于奖励以及基于动作三个角度对这些工作进行了分类和总结.

(1) 基于状态的攻击算法是针对深度强化学习的对抗攻击中研究最多、范围最广的攻击方式. 已有的工作从黑盒或白盒、 训练或测试阶段等不同角度提出了各种高效的基于状态的攻击算法. 由于和经典的计算机视觉中的对抗攻击方式相似，这类攻击算法适用范围广，攻击效果好，且具有丰富的研究工作作为基础. 然而，这些算法几乎全部派生自传统对抗攻击方法，针对强化学习任务进行了调整，并没有提出具有足够创新性的理论改进；与此同时，由于本身与传统对抗攻击方法的相似性，这些方法也容易被传统对抗防御方法克制.

(2) 基于奖励的攻击算法以奖励函数投毒为基本思想，并针对实际应用场景进行了改进，通过对奖励函数添加噪声或符号翻转来对模型训练过程造成影响，从而实现对抗攻击. 基于奖励的攻击方法往往不限制待攻击模型或算法，适合用于扰动经验回放池中的奖励符号，从而对模型训练带来长期影响. 同时，基于奖励的攻击方法对于在线学习类

的强化学习算法也有可预见的攻击效果.

(3) 基于动作的对抗攻击充分利用了强化学习的特点, 抓住其与传统计算机视觉中任务的不同点进行攻击. 不同于传统的分类任务, 在强化学习任务的马尔可夫过程中, 动作既是上一次策略网络的输出, 也会影响到下一次网络自身的输入. 即对动作的扰动会带来时序层面的影响, 对网络关键输出的扰动价值也远大于传统对抗攻击. 无论是基于动作概率还是基于对抗策略的攻击, 都是强化学习领域中特定的攻击算法, 具有重要价值和挖掘潜力.

可以看到, 由于强化学习训练过程的特殊性, 存在一定量的算法是在智能体训练阶段实施攻击的. 值得注意的是, 这些攻击算法借鉴了投毒攻击[67]的基本思想, 将加入了噪声的样本混入训练过程, 使得智能体模型在最终的测试阶段产生错误预测. 它虽然与在测试阶段直接污染测试数据的对抗攻击不完全一致, 但是其攻击目标是一致的. 在本文中, 笔者将其描述为“训练阶段的对抗攻击”.

总体来看, 针对强化学习的对抗攻击方法研究依然存在不足: 一方面, 现有的大部分工作主要是基于传统对抗攻击算法在强化学习领域的应用, 如何利用强化学习本身特性进行攻击尚有研究空间; 另一方面, 强化学习领域的对抗攻击方法普遍存在迁移性不强、 难以实现的问题, 缺乏在物理世界中的实验.

4 面向强化学习的对抗防御技术

在系统的归纳总结了深度强化学习领域的对抗攻击研究后, 本章进一步分析深度强化学习领域中的对抗防御方法的研究. 与针对攻击的分析不同之处在于, 本文并没有直接基于马尔科夫决策过程四元组进行对抗防御的分类. 相反, 本章结合传统对抗防御方法的分类方式从对抗训练、 对抗检测、可证明鲁棒性、 鲁棒学习等角度出发, 对现有工作进行梳理总结(如表 2 所示). 笔者认为: 首先, 大量的防御方式都是用于防御状态扰动攻击, 直接从状态、 奖励、 动作的维度进行分类可能会造成极度的不平衡, 丧失分类讨论分析的意义; 其次, 这种分类方式可以帮助研究人员将强化学习中的对抗防御手段与经典对抗防御体系进行对齐, 更好地理解在 DRL 领域中的对抗防御算法.

4.1 基于对抗训练的防御

在传统对抗攻防领域, 最常用且最为有效的对抗防御方法便是对抗训练[68-70]. 对抗训练通过为输入图像添加对抗噪声生成对抗样本, 并将其作为训练集的一部分训练模型, 从而有效提升模型的对抗鲁棒性, 其标准定义如下:

$$\min_{\theta} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}\left[\max_{r\in S} L\left(\mathbf{y}, f_{\Theta}(\mathbf{x}_{adv})\right)\right]. \#(9)$$

其中r是干扰噪声量级, x是数据分布D上的样本, y是相对应的标签, $L(\mathbf{x},\mathbf{y})$表示神经网络的损失函数, $\Theta \in \mathbb{R}^{p}$是模型参数集合, 需要优化得到最小化风险$\mathbb{E}$的模型参数. 除了在计算机视觉中有效, 对抗训练的方法同样也适用于强化学习的对抗鲁棒性提升(如图 8). 对抗训练可以被认为是一个最小-最大优化问题, 即攻击者希望生成最具有攻击性的扰动, 从而最小化被攻击者的总目标G_t. 相反, 防御者则希望训练出最鲁棒的模型, 从而使得在攻击者进行攻击的情况下, 最大化自身的总目标G_t. 若攻击者与防御者的策略均为强化学习算法, 则双方的强化学习策略均需要不断更新以适应对方的策略,对抗防御开销极高, 往往无法收敛. 若攻击者基于神经网络梯度[6]生成针对神经网络的噪声, 则攻击者无需不断训练自身的策略, 而是只需针对防御者的策略生成自身的攻击. 这种防御方式虽然仍具有较高开销, 但实际训练中可以接受. 由于对抗训练效果较好, 在实际防御中被广泛采用.

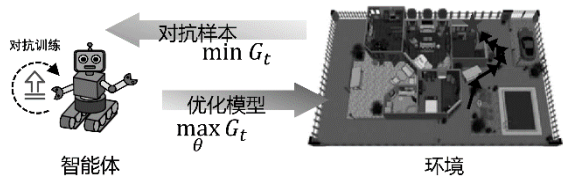


图 8 基于对抗训练防御的算法示意图. 对抗训练是在求解一个最小-最大优化问题. 在训练过程中, 攻击者生成对抗样本给智能体, 智能体在具有攻击性环境下优化自身, 最终智能体模型会变得更加鲁棒.

Kos 等人[71]首先利用 FGSM 对抗攻击在 Atari Pong 任务 A3C 算法上开展基于对抗训练的强化学习鲁棒性研究. 作者通过大量实验发现: 在输入的图像像素上加上对抗噪声和随机噪声都可以使得模型产生错误的行为, 但是对抗噪声效果会更强; 而且使用智能体的奖励函数来指导在对应帧上加入对抗噪声可以有效地提升对抗攻击的有效性和攻击强度. 通过同时使用对抗样本和随机样本重新对抗训练智能体后, 作者发现模型对于 FGSM 攻击的鲁棒性有所提升. 类似地, Mandlekar 等人[72]在训练深度强化学习智能体的过程中, 分别在智能体的

观测特征和环境动力学特征中加入使用 FGSM 攻击方法生成的对抗样本噪声，从而获得鲁棒的学习策略. 相似工作层出不穷，Behzadan 等人[73]通过引入带有随机概率p的FGSM攻击方法在Atari游戏中的 Breakout 与 Pong 两个场景中生成对抗样本，对模型鲁棒性进行增强. Behzadan 等人[74]还提出了一种为智能体观测和模型参数空间添加对抗噪声的方式，对于鲁棒性有明显的提升效果.

表 2 对抗防御技术总览

文献	针对的强化学习算法	实验环境	防御技术	抵御扰动	防御开销
Liu 等人[16]	PACMAN-RL+Q[52]	EQA	对抗训练	状态扰动	高
Kos 等人[71]	A3C	Pong	对抗训练	状态扰动	高
Mandlekar 等人[72]	TRPO	Inverted Pendulum, Half Cheetah, Hopper, Walker	对抗训练	状态扰动	高
Behzadan 等人[73]	DQN	Breakout, Pong	对抗训练	状态扰动	高
Behzadan 等人[74]	DQN	Enduro, Assault, Breakout	对抗训练	状态扰动	高
Pattanaik 等人[75]	DDQN, DDPG	Cartpole, Mountain Car, Hopper, Half Cheetah	对抗训练	状态扰动	高
Chen 等人[76]	A3C	自动寻路	对抗训练	状态扰动	高
Nisioti 等人[77]	DQN	Interconnected Nodes	对抗训练	动作扰动	极高
Lin 等人[78]	DQN	Pong, Seaquest, Freeway ChopperComman, MsPacman	对抗检测	状态扰动	低
Havens 等人[79]	MLAH	InvertedPendulum-v2, MountainCarContinuous-v0, Hopper-v2	对抗检测	状态扰动	低
Fischer 等人[80]	RS-DQN	Freeway, BankHeist, Pong, boxing, road-runner	可证明鲁棒性	状态扰动	中
Tessler 等人[81]	PR-MDP, NR-MDP	Hopper, Walker2d, Humanoid, InvertedPendulum	可证明鲁棒性	动作扰动	高
Wu 等人[82]	Q-learning	CartPole, Pong, FreeWay	可证明鲁棒性	状态扰动	高
Wu 等人[83]	DQN, QR-DQN, C51	Freeway, Breakout	可证明鲁棒性	状态扰动, 奖励扰动	中
Wang 等人[84]	Q-Learning, CEM, SARSA, DQN, PPO, NAF, Dueling DQN, DDPG	CartPole, Pendulum, AirRaid, Alien, Carnival, MsPacman, Pong, Phoenix, Seaquest	鲁棒学习	奖励扰动	高
Gallego 等人[85]	TMDP	Repeated Matrix Games, Friend Or Foe	鲁棒学习	奖励扰动	中
Ying 等人[87]	CPPO	Ant, Halfcheetah, Walker2d, Swimmer, Hopper	鲁棒学习	奖励扰动	中

上述对抗训练方法在生成对抗样本时都使用了简单的 FGSM 攻击算法. Pattanaik 等人[75]从对抗训练的方法和损失函数入手，研究了在强化学习任务下使用新型对抗攻击算法进行对抗训练对模型鲁棒性的提升效果. 作者基于其特殊设计的损失函数应用梯度优化来实现对抗攻击，从而有效提升强化学习算法在训练时对于参数的敏感性，让算法更加鲁棒.

Chen 等人[76]则将对抗训练引入更复杂的智能体自动寻路问题中. 智能体寻路问题与 Atari 等游戏的不同之处在于，智能体没有对于全局的观测，仅能获取局部信息. 针对智能体寻路的特点，作者通过在寻路环境中添加真实存在的障碍进行对抗攻击. 通过对抗训练的模型可以有效降低对抗攻击

成功率, 使智能体做出正常的决策. 其次, Liu 等人[16]也使用对抗训练提升了智能体在 Embodied Question Answering 任务中对于对抗攻击和自然高斯噪声的鲁棒性. 针对更加复杂的多智能体强化学习, Nisioti 等人[77]提出 RoM-Q 算法, 在多智能体强化学习中对于智能体策略噪声进行对抗训练. 在多智能体博弈环境中, 算法选定某智能体, 要求其执行使得 Q 值最小的动作. 鲁棒的策略要求在任意被选择的扰动智能体组合与任意扰动智能体下正常运行. 作者通过实验验证 RoM-Q 对抗训练后的智能体与基线算法相比, 在不同攻击概率下均具有更高的鲁棒性.

4.2 基于对抗检测的防御

对抗检测是对抗防御中一种常见的技术手段, 其通过训练额外的模型对输入样本进行检测. 通过对输入样本的类型进行判别, 将对抗样本直接丢弃, 将正常的干净样本输入给强化学习智能体, 从而在起到防御效果的同时保留了原有的强化学习策略(如图 9). 由于对抗检测本质上是训练一个判别是否存在对抗样本的分类器, 其防御开销较小. 但是, 基于对抗检测的防御仅具有识别对抗样本的功能, 强化学习本身不具有对抗鲁棒性.

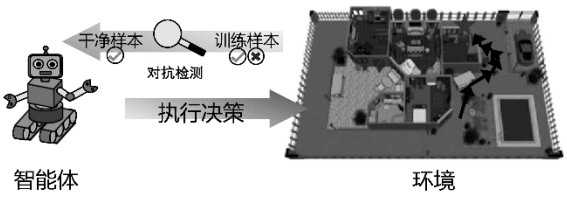


图 9 基于对抗检测防御的算法示意图. 这种防御方法在智能体模型的预处理部分加入一个检测模块, 通过对输入样本的类型进行判别, 将对抗样本直接丢弃, 将正常的干净样本输入给强化学习智能体.

Lin 等人[78]首先提出一个在强化学习场景中利用动作帧进行对抗检测的模型防御方法. 该论文认为传统的计算机视觉领域的对抗样本检测都是基于单帧图像来进行的, 忽略了历史帧的交互关系和信息关联. 因此, 作者根据历史观测与行为数据, 设计了一个模块来预判观测帧. 如果预测值与观测有较大差别, 则认为该帧为对抗样本, 并根据预测进行决策, 否则基于真实观测进行行动. 进一步, Havens 等人[79]提出一种基于元学习的模型无关的层级化攻击检测框架 MLAH, 具有较强适应性的在线防御能力. 与上文的防御方法不同, MLAH 框架基于决策空间进行对抗样本防御, 因此可以直接降低由攻击方法带来的过拟合问题.

4.3 基于可证明鲁棒性的防御

在对抗训练与对抗检测之外, 还有一些学者从可证明鲁棒性等角度出发, 对强化学习对抗防御领域进行研究. 相比其他基于“经验性”的防御手段, 这种防御方式会形式化的推导并给出防御方法带来的鲁棒性下界, 给出“可证明”的模型防御(如图 10). 显然, 这种防御手段在更加严谨的同时也引入了更多的约束条件. 可证明鲁棒性严格证明了模型的鲁棒性下界, 这种证明往往在特殊设计的训练算法下成立. 部分可证明鲁棒性算法训练一个新的强化学习智能体以帮助防御, 具有中等开销. 还有部分可证明鲁棒性算法使用修改后的对抗训练以帮助防御, 具有较高开销. 其具体开销则根据增强鲁棒性的方法而定.

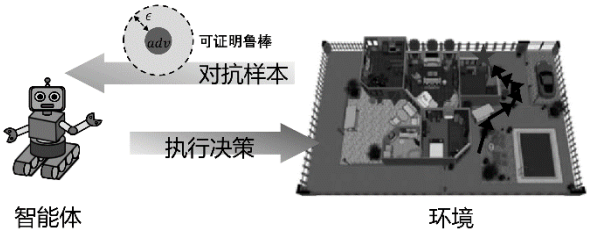


图 10 基于可证明鲁棒性防御的算法示意图. 这种防御方式会形式化的推导并给出防御方法带来的鲁棒性下界, 给出“可证明”的模型防御.

Fischer 等人[80]在模型蒸馏过程中结合现有防御措施(如对抗训练, 可证明的鲁棒学习等), 提出 Robust Student-DQN (RS-DQN), 推导得到了对抗扰动下的算法可证明鲁棒性下界. RS-DQN 是将标准 DQN 拆分为 Student Network (S)网络和 Q 网络. Q 以标准的 DQN 进行训练, S 通过对 Q 蒸馏的方式进行训练, 在蒸馏过程中融合一些其他的防御措施. 当没有攻击存在时, RS-DQN 和 DQN 获得相似的分数, 而在有攻击存在的情况下, 无防御的 DQN 失败, 而 RS-DQN 保持鲁棒. 该论文的实验部分在 Atari 2600 的五个游戏场景上对 Priority Replay、DoubleDQN、 DuelingDQN、 NoisyNet 进行了实验, 对抗训练中的攻击算法选用 PGD 算法. 实验结果证实 RS-DQN 的策略比 DQN 更加鲁棒, 并且可以得到在正负 1 像素扰动下的可证明鲁棒性.

与此同时, Tessler 等人[81]则针对强化学习智能体的鲁棒性和泛化性, 提出了两种鲁棒的马尔可夫决策过程, 分别是概率动作鲁棒的 PR-MDP 与噪声动作鲁棒的 NR-MDP. 这其中, 作者证明了对于 NR-MDP 而言, 最优策略是稳定且确定的; 在

PR-MDP 下，最优策略存在且服从某一随机概率. 通过在 MuJoCo 环境下的大量实验表明，PR-MDP 与 NR-MDP 可以帮助智能体学到更加鲁棒且安全的策略. 此外，即便在攻击者不存在的情况下，经过 PR-MDP 与 NR-MDP 训练的智能体也会在正常环境中取得更好的结果.

2022 年，Wu 等人[82]构建了针对 Q 学习算法的鲁棒性认证框架 CROP，提出了状态级别的鲁棒性证明和累积奖励的下界证明两个标准，以及相应的基于全局平滑和局部平滑思想的证明算法. 作者从理论上证明了在有界对抗状态扰动下，输入状态的认证半径和扰动累积奖励的下界，并在三个 Atari 游戏上对六种经验鲁棒的强化学习算法进行实验评估，证明了提出的鲁棒性认证框架的有效性.

类似地，Wu 等人[83]还将状态行为稳定性和累积奖励界限标准应用于离线强化学习在投毒攻击下的鲁棒性证明. 提出了第一个认证框架 COPA，以证明不同认证标准下离线强化学习算法可以容忍的投毒数量. 作者在 Freeway、 Breakout 游戏场景中使用三种离线强化学习算法对框架进行了全面评估.

4.4 基于鲁棒学习的防御

除了上述防御方法，研究人员还结合强化学习算法和任务本身的特殊性，设计并研究了一系列和强化学习算法本身紧耦合的鲁棒学习方法. 与传统的计算机视觉中的对抗防御方法不同，这一类方法通常结合了强化学习算法本身的特点，设计和研究出与强化学习场景、 算法相适配的防御技术(如图 11). 它们不是通用的，无法直接应用于传统的图像分类任务上，且视具体方法不同而具有不同的开销. 因此笔者将其单独分为一个类别进行讨论和分析.

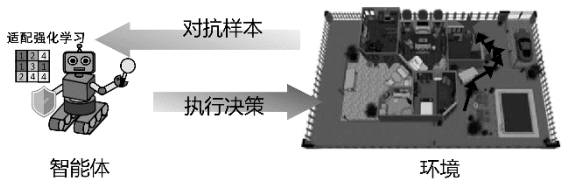


图 11 基于鲁棒学习防御的算法示意图. 与传统的防御算法不同，这一类防御方法通常与强化学习算法本身适配，结合了强化学习算法本身的特点.

强化学习中的回报奖励在动态环境中存在不确定性，甚至可能遭受人为恶意对抗扰动. 针对这一问题，研究人员从奖励观测度量与估计推理角度开发对抗防御方法来提升策略的鲁棒表现. Wang 等人[84]提出了一种在奖励函数遭遇噪声扰动时的鲁棒学习方法. 使用混淆矩阵 (Confusion Matrix)来估计奖励，帮助智能体在噪声扰动环境下更好学习，算法也因此在奖励具有噪声时会有明显的表现提升. 论文中定义了被扰动的奖励，同时使用数学公式证明了对于真实奖励的估计是无偏的，在两个经典控制游戏和多个 Atari 游戏上进行了实验. 为了模拟奖励带有噪声的环境，实验为环境输出的奖励加上了对称或非对称的噪声. 文章使用许多不同的经典强化学习算法如：Q-Learning、 CEM、 SARSA、 DQN、 PPO、 NAF、 Dueling DQN、 DDPG 进行了实验，在不同噪声级别下训练模型，不仅能得到更好的回报，并且还能更快收敛.

同样针对奖励遭受扰动这一问题，Gallego 等人[85]在多智能博弈背景下提出一个鲁棒学习框架，称为受威胁的马尔科夫决策过程 (Threatened Markov Decision Processes, TMDP). 具体来说，作者引入 K 级思维这一概念，将攻击者视为 K-1 级，而决策者在第 K 级去思考问题. 也即决策者会考虑到攻击者的动作后再选择最优动作，决策者的代价函数将攻击者的估计代价纳入考量，从而使得决策者获得更高的鲁棒性. 作者在 Repeated Matrix Games、 Friend Or Foe[86]两个环境中进行实验，证明以二级思维训练的智能体相比传统方法有更好表现.

Ying 等人[87]则认识到现有鲁棒强化学习方法采用奖励方差作为不确定性度量的局限性，提出使用条件风险价值 CVaR 替代风险价值 VaR 作为衡量模型风险的标准，来减轻强化学习面对不确定环境的糟糕表现. 使用 VaR 作为正则项会同时消除表现特别好与特别差的策略，而使用 CVaR 只会消除表现特别差的策略. 进一步，作者提出了一种鲁棒强化学习框架 CPPO (CVaR-Proximal-Policy-Optimization)，在限制模型的 CVaR 值最大化奖励函数. MuJoCo 环境中的实验结果表明，使用 CVaR 的 CPPO 方法可以提高智能体的总奖励函数，且对于智能体内部的不确定性，环境变化均具有较好的适应性，对于对抗样本也具有一定的鲁棒性.

4.5 小结

在本章中，我们针对强化学习领域中的对抗防御方法作了较为全面的回顾，从传统对抗攻防领域中的对抗训练、 对抗检测、 可证明鲁棒性等防御方法出发，对现有工作进行梳理总结.

(1) 一系列工作将传统对抗攻防中的对抗训练

算法迁移应用到深度强化学习领域中. 通过在智能体训练过程中引入干扰噪声(大多选用简单的FGSM对抗攻击算法生成扰动), 对智能体的状态进行扰动进而优化智能体对于噪声的鲁棒性. 然而, 这些方法并未对强化学习本身特性进行更深入的研究, 只是将对抗训练方法的思想迁移至深度学习领域进行应用, 并未取得防御技术的发展突破.

(2) 基于对抗检测的防御方法从分辨干净样本与对抗样本角度入手, 使用专门训练的检测模型分离出干净样本. 其优点在于不改变智能体的原有策略, 但这种方法的通用防御能力相对较弱, 检测器对于训练过程使用的对抗样本会具有较好的检测能力, 而一旦面对未曾在训练中出现的对抗攻击方法, 则难以有效检测出对抗样本. 基于对抗检测的防御方法适合于智能体开箱即用的强化学习场景, 从而在智能体不修改的情况提供防御能力.

(3) 基于可证明鲁棒性的防御方法结合了强化学习决策过程, 通过对智能体鲁棒性下界给出证明(如扰动半径的下界), 在理论层面为智能体鲁棒性进行了保护. 经过鲁棒认证的智能体模型能在认证范围内安全鲁棒, 但这种防御方法也存在一些限制(如: 主要针对ℓ_1, ℓ_2范数约束下的对抗样本), 与基于经验性的鲁棒防御算法的表现(如: 对抗训练)仍有差异. 这种防御方法如果能在更多情况下推广应用(如: ℓ_∞范数下的对抗样本), 将为深度强化学习在理论的鲁棒性提供有力保障.

(4) 与上述防御方法不同的是基于鲁棒学习的防御策略. 这类方法针对强化学习算法的特点, 应用与算法适配的特殊方法(如: 混淆矩阵、 奖励估计等)来进行防御. 这类防御方法与强化学习算法场景紧密耦合, 在其它算法上难以进行通用的适配. 然而, 由于其和强化学习的独特关系, 这个方向具有重要研究价值和挖掘潜力.

目前而言, 针对强化学习领域中对抗防御的研究仍旧存在较大的发展空间: (1) 现有的防御方法大多是传统对抗防御算法在强化学习中的迁移应用, 未来还需要从强化学习本身特性进一步探索; (2) 现有防御方法的泛化能力不足, 需要探索更通用的防御方法, 保障智能体在动态复杂环境面对不确定干扰时的鲁棒表现; (3) 目前主要的防御方法都是针对于状态扰动攻击的加固, 而针对于其他类型对抗攻击的防御较少. 分析其背后原因可看到: 基于状态扰动攻击的定义与计算机视觉领域中的对抗样本较为相似也是相对最早被提出、 有大量研究基础的一种攻击方法, 因此催生了大量的防御算法. 相反, 其他类型的攻击, 如: 基于动作的攻击和基于奖励函数的攻击都有特殊的要求 (如: 要求在零和博弈场景中进行或直接改变奖励函数), 直接防御的难度较大, 相关的防御研究也较少.

5 基于对抗攻击的深度强化学习机理理解与模型增强

在本章中, 我们将介绍和分析在深度强化学习领域除了对抗攻防之外的对抗样本相关的研究工作, 主要分为: 使用对抗样本来分析深度强化学习的脆弱性机理以及提升智能体的任务相关能力两个部分, 如表 3 所示. 可以看到, 这部分研究的第一篇相关论文发表于 2021 年, 是一个仍处于初步探索阶段的新兴方向. 然而, 这个领域的探索向研究人员证明了: 对抗样本对于深度强化学习并非百害而无一益, 通过适当的手段, 对抗攻击也可以变成一种提升对于深度强化学习可解释性和能力的工具. 因此, 这个新兴的领域定会在未来成为深度强化学习对抗攻击领域的一个重要研究方向.

5.1 强化学习中对抗脆弱性的机理分析

在传统的深度学习领域, 存在一系列论文从不同的角度探索深度神经网络的对抗脆弱性机理, 如: 神经元[93,94]、 神经网络路径[95,96]、 图像特征[97,98]、 数据分布[68]等. 在深度强化学习领域, 也有研究人员对攻防作用机理展开了研究, 如: Korkmaz 等人[89]使用两种不同的方法研究了经过对抗训练的深度强化学习策略的表现. 第一种方法使用了傅立叶频谱,研究了对抗训练的策略和普通的策略经过 C&W 攻击方法得到的最小扰动的傅里叶频谱信息, 发现对抗训练策略的扰动更聚焦于低频信息, 这说明对抗训练模型更关注低频区域的信息. 第二种方法定义了用于衡量深度强化学习策略特征敏感性的指标 KMAP 和 HMAP, 并比较了目前最先进的对抗训练方法和普通模型之间特征敏感性的区别. 该研究发现对抗训练方法虽然消除了策略对特定特征的敏感性, 但同时也引入了对新特征的敏感性. 进一步, 作者在 OpenAI 的 Atari 游戏环境下, 使用 DDQN 和 SA-DDQN 作为模型进行实验, 并通过可视化技术开展分析.

5.2 强化学习中的对抗增智

不少研究通过在深度强化学习领域引入对抗

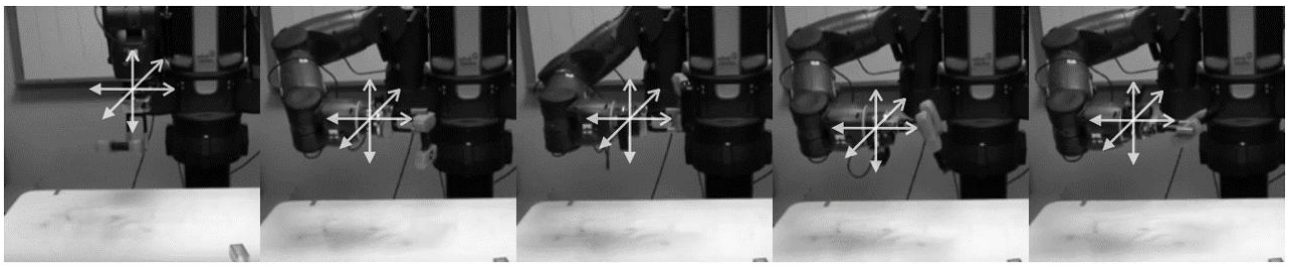


图 12 通过对抗样本提升机械臂准确率[88]. 通过在机械臂拾取任务中引入一个对抗智能体来联合对抗训练提升原本智能体的表现能力. 在训练过程中, 该对抗智能体干扰机械臂, 而机械臂需要在影响下保持抓取的准确度. 这种对抗学习的框架可以有效提升智能机械臂的抓取成功率.

表 3 模型机理理解与对抗增强总览

文献	强化学习算法	实验环境	核心思想	目标
Pinto 等人[88]	DQN	机械臂抓取任务(grasp)	生成式对抗网络	对抗增智
Korkmaz 等人[89]	DDQN 和 SA-DDQN	Brockman, Bellemare	傅立叶频谱分析, 特征敏感性测量	机理理解
Pinto 等人[90]	TRPO	InvertedPendulum, Half Cheetah, Hopper, Swimmer, Walker2D	生成式对抗网络	对抗增智
Ogunmolu 等人[91]	ILQG	Goal Reaching Task	生成式对抗网络	对抗增智
Li 等人[92]	MADDPG	Covert communication, keep-away, physical deception, predator-prey	对抗训练	对抗增智

的思想来有效提升智能体的任务相关能力. 详细地说, 研究人员通过在训练过程中添加对抗智能体作为对手来干扰正常智能体的决策, 使学习到的智能体策略获得更好的任务表现能力(如: 泛化性等).

Pinto 等人[88]利用对抗思想提升了真实智能体机械臂的表现能力. 准确地说, 作者并没有直接将某种对抗攻击方法引入训练过程, 而是在整个任务中引入了一个对抗智能体, 来联合对抗训练提升原本智能体的表现能力. 通过在真实环境中的机械臂抓取任务中进行实验, 实验验证了作者提出的框架可以有效提升智能机械臂的抓取成功率(如图 12 所示). 类似地, Pinto 等人[90]在训练时引入一个对抗智能体, 用于降低正常智能体的回报奖励. 两个智能体依次固定参数和训练, 最终达到纳什均衡, 期望解决强化学习中对于模型训练初始化过于敏感、对于未见过测试环境泛化能力不足的问题. Ogunmolu 等人[91]在训练中引入采用随机策略的对抗智能体, 提出一种极大极小迭代动态博弈框架. 两个智能体执行与对方相反的动作, 试图让策略达到由智能体代价函数定义的凹凸问题的平衡鞍点. 从从某种程度上来说, 这三篇论文都利用了生成式对抗网络 (Generative Adversarial Nets, GAN) [99]来训练一个更加鲁棒的智能体模型.

不同于上述的对抗博弈思想, Li 等人[92]提出了一种多智能体强化学习环境中的新算法, 能够有效提升其表现能力和泛化性. 作者认为深度强化学习策略容易陷入局部最优, 受到环境策略和其他智能体当前策略的影响较大. 基于 MADDPG 算法, 该论文提出了 M3DDPG 方法, 利用了一个最小-最大过程, 最小化其它智能体的 Q 值, 最大化在该条件下自身的 Q 值. 同时, 将对抗训练的思想引入强化学习训练中, 为其他智能体的动作加上一个扰动噪声, 从而实现多智能体条件下的对抗训练. 作者在 particle-world 环境中进行了实验, 选用了其中的 4 个场景进行测试, 证明了 M3DDPG 的表现整体优于 MADDPG.

5.3 小结

本章从对抗增智、 脆弱性机理分析两个方面梳理归纳了对抗样本在强化学习领域中的应用.

(1) 强化学习中对抗脆弱性机理这一主题目前的研究还较少, 相关的研究成果(如: 傅里叶频谱分析、 敏感特征分析等)与传统的计算机视觉中对抗脆弱性机理的分析手段相似. 针对这一主题的研究引发了两个思考: 首先, 是否深度强化学习中的智

能体脆弱性机理和传统的视觉中的脆弱性机理相同; 其次, 结合深度强化学习(尤其是多智能体强化学习)算法本身特点的智能体脆弱性机理理论框架需进一步探索. (2) 相比脆弱性机理研究, 强化学习中应用对抗博弈思想来提升多智能体鲁棒性的工作相对完善, 且对于真实物理世界中的强化学习系统进行了初步的应用验证. 然而, 目前主要的对抗增智方法都借鉴了 GAN 的思想, 和对抗训练的思路仍有一定区别. 如何将对抗训练高效地引入强化学习的训练框架中, 仍是一个需要探索的研究问题.

6 挑战与未来研究方向

6.1 理论: 深度强化学习的普适鲁棒性

由于深度强化学习自身任务独有特性, 针对深度强化学习的对抗攻击与传统计算机视觉领域的对抗攻击存在很大的不同. 例如, 由于强化学习的自举性, 诞生了基于奖励的对抗攻击[55,56], 这间接表明了强化学习的学习过程难以直接泛化至具有噪声的奖励信号; 由于强化学习的时序性, 出现了基于动作概率的对抗攻击方式(隶属于基于动作的攻击) [58], 其核心点在于强化学习智能体的当前决策由一系列之前时刻的行为序列所影响; 由于强化学习具有博弈性, 研究人员也提出了一系列基于对抗策略的攻击方式(同样属于基于动作的攻击) [14,40], 在不改变网络模型本身的前提下, 利用另一个对抗策略找到当前模型的脆弱点. 然而, 传统的计算机视觉中的对抗攻击(通过在图像上加入对抗噪声)只能对应至深度强化学习中的基于状态攻击中的部分内容.

可以看到, 由于强化学习的特性, 相比传统计算机视觉领域, 在该领域中的对抗攻击具有更强的多样性. 基于此, 目前面向深度强化学习的对抗防御研究仍不存在能够防御多种攻击方式的算法和策略. 如表 2 所示, 大部分的防御方法主要针对于基于状态的扰动攻击, 而鲜有研究针对动作、奖励的扰动进行防御, 更不用说针对用多种维度攻击的“普适”鲁棒性. 例如, 被证明最为有效的对抗训练防御手段也只能针对特定种类的单个对抗攻击(状态、 动作、 奖励函数)取得防御效果, 但对于某一类攻击方法鲁棒的防御仍会被另一类攻击方法攻破. 因此, 针对深度强化学习的普适鲁棒性理论研究需要关注的要点包括: (1) 探索基于状态、 动作、 奖励的强化学习对抗攻击噪声在理论上的相似性和一致性 (如: 频域的手段), 建立统一的强化学习对抗攻击框架; (2) 提升强化学习对于多种维度的对抗攻击的防御, 研究强化学习的泛化性与鲁棒性理论之间的关系, 以获得更为普适性的鲁棒强化学习模型.

6.2 技术: 多智能体强化学习的对抗攻防

在强化学习的对抗攻防领域, 已有大量工作对单智能体应用场景进行了研究[15,58,59,71]. 然而, 在适应于复杂协作应用的多智能体强化学习领域(如: 无人集群、 分布式控制、 协作决策系统[100-102]), 对抗攻防的研究投入相对较少.

Lin 等人[103]研究了合作情况下多智能体学习任务的对抗攻击. 该文章选择了典型的中心训练分布式执行算法 QMIX[104], 在星际争霸多智能体环境(SMAC[44])上进行了实验. 对抗攻击者能够获得多智能体其中一个智能体的控制权, 通过该智能体的动作诱导其他智能体做出较差的动作从而降低整体的任务奖励. 同时, 作者也利用动态 JSMA[105]算法对正常智能体的观测进行了扰动, 进一步降低了正常智能体的奖励. Guo 等人[17]指出多智能体强化学习对于多种攻击具有脆弱性, 并对于不同多智能体强化学习算法进行了基于状态、 动作和奖励的鲁棒性评测.

可以看到, 在多智能体强化学习领域, 已有部分研究者意识到多智能体场景中安全鲁棒决策的重要性, 并进行了初步的对抗攻击探索. 然而, 如何对多智能体强化学习进行有效的对抗攻防仍鲜有研究. 因此, 针对多智能体对抗攻防的研究需要关注的要点包括: (1) 设计考虑多智能体群体行为和决策模式的对抗攻击算法, 而不仅仅是在该场景下简单应用传统的对抗攻击算法; (2) 探索和研究在群体中各节点交互决策条件下的多智能体强化学习算法脆弱性原因, 并研究设计出在群体中某些节点被攻击后整体依然正常决策的鲁棒防御算法; (3) 在多智能体场景中, 利用对抗攻击进行智能体间博弈对抗学习, 提升多智能体的任务表现.

6.3 平台: 深度强化学习对抗攻防评测基准与环境

在计算机视觉和自然语言处理领域, 已经涌现出了很多优秀的深度学习对抗攻防评测基准与环境[106-108]. 研究人员可以很方便的利用这些平台的开源代码和模型库进行相关攻防算法的结果复现,

内嵌的标准设置下的排行榜机制(leaderboard)也可以很好的刻画领域内对抗攻防算法的效果. 一个标准和完善的算法鲁棒性评估基准环境, 可以帮助研究人员分析比较不同对抗攻防算法对于算法模型的影响机制, 从而更好地推动鲁棒算法的研究发展. 在强化学习鲁棒性测试基准方面, 也有了一些初步的研究, 例如: Behzadan 等人[109]在碰撞规避机制场景上提出了一项测试基准. 作者在测试中引入一个对抗智能体, 通过迫使正常智能体进入不安全状态发生碰撞, 来完成对智能体碰撞规避策略的测试. 作者提出了完整的工程化测试框架, 并通过对两个碰撞规避策略的案例研究验证了该测试基准的有效性. Behzadan 等人[110]还对强化学习智能体的对抗抵御能力与对抗鲁棒性进行了基准测试. 在测试中, 作者使用两个指标衡量智能体的性能, 分别是对抗预算(Adversarial Budget)与对抗反悔(Adversarial Regret). 对抗预算指攻击方需要付出多少代价才能攻破此智能体, 对抗反悔指对于智能体而言未受扰动时的奖励与受到扰动时奖励之差. 作者对 DQN、 A2C 与 PPO2 三种策略在 CartPole 环境中进行测试鲁棒性测试.

可以看到, 在深度强化学习对抗攻防领域, 已经有学者意识到了构建攻防评测基准的重要性并开始做了初步的尝试. 然而, 目前的攻防评测基准与环境中所包含的强化学习算法以及对抗攻防算法都较少, 评测指标也非常简单. 因此, 针对深度强化学习的对抗攻防基准研究需要关注的要点包括: (1) 在深度强化学习领域构建一个通用的对抗攻防评测基准, 将该领域的攻防算法集成进平台环境中, 从而更好地推动整个 DRL 对抗攻防研究的健康持续发展; (2) 在公正、 统一的实验环境和条件下评测所有 DRL 对抗攻防算法的效果, 给出公平、全面的结果并进行深入的分析, 从而更好地启发新型功法算法的设计研究.

6.4 应用: 面向物理世界的深度强化学习对抗攻防

目前已有许多工作展示了对抗攻击在物理世界中的可行性与效果, 但这些工作主要集中在计算机视觉领域[28-32]. 不同于对抗样本概念定义中的“微小噪声”, 研究人员通过生成不影响人类语义认知的扰动(如:贴片、喷漆), 并将其打印出来放置在真实世界中的指定位置, 使得对抗样本可以对于部署在物理世界中的真实模型系统进行攻击.

然而, 作为当前炙手可热的研究场景, 虽然深度强化学习不断在理论和数字世界中取得建设性突破, 但其在日常生活和工业生产中的应用依旧不足. 有研究人员[111]分析总结了强化学习落地的九大挑战: 真实数据有限、硬件设备延迟、观测动作空间维度过高、系统约束归因困难、交互场景的非稳态、奖励函数设计困难、模型决策时间限制、离线训练困难、系统可解释性差. 这些挑战共同限制了强化学习落地应用的发展, 也是研究者们需要研究解决的问题. 虽然距离大规模应用尚有一定距离, 也有少部分工作尝试在物理世界中实现对深度强化学习模型的攻击. 如 Xiao 等人[53]初步探索了将对抗样本引入物理世界中攻击强化学习智能体的可能性. 该文章提出了基于环境动力学的扰动攻击, 即修改环境中的物理属性等. 文章通过随机采样或利用强化学习的对抗性采样来寻找针对环境动力学的最优扰动, 并同时在数字世界与物理世界进行了他们所设计的实验, 验证了他们攻击方法的有效性.

可以看到, 目前针对物理世界场景中深度强化学习对抗攻防的相关工作依然十分缺乏, 是一个具有潜力且尚未被充分研究的重要方向. 它的难点在于: (1)在真实场景中, 攻击者往往不能轻易扰动模型的奖励、动作等内部数据, 因而攻击者能够采用的扰动手段十分有限, 如: 通过对智能体观测到的环境做出符合环境约束条件的扰动; (2)深度强化学习在真实环境中的应用本身就存在精准度不足的问题, 加入对抗攻击对其的影响难以直接衡量和判断. 因此, 针对物理世界深度强化学习的对抗攻防研究需要关注的要点包括: (1) 设计在数字世界和物理世界攻击能力一致的强化学习对抗攻击算法; (2) 综合、 深入、 全面地评估物理世界中强化学习对抗攻击算法的有效性; (3) 选取高精度的典型强化学习应用场景进行物理世界对抗攻防的验证测试.

7 结论

深度强化学习的广泛应用引起了大量研究对于其对抗鲁棒性的关注. 本文对于深度强化学习领域对抗攻防技术的前沿研究进展进行了一次全面的综述. 本文首先阐述了基于状态、 基于奖励以及基于动作的深度强化学习对抗攻击进展; 本文接着从对抗训练、 对抗检测、 可证明鲁棒性和鲁棒学习的角度归纳总结了深度强化学习领域的对抗防

御技术; 最后, 本文分析了基于对抗样本的深度强化学习机理理解与模型增强并讨论了领域内的未来研究方向.

虽然研究人员在深度强化学习领域开展了大量对抗攻防的研究, 然而领域内还存在多个亟待解决的问题和挑战制约着深度强化学习对抗攻防研究的发展, 如: 面向物理世界的深度强化学习对抗攻防仍鲜有探索、 缺乏统一标准的对抗攻防评测基准环境等. 希望本文能够帮助更多研究人员投身于研究和构建更加安全可靠的深度强化学习技术之中.

参 考 文 献

[1] Kober J, Bagnell J A, Peters J. Reinforcement learning in robotics: A survey. The International Journal of Robotics Research, 2013, 32(11): 1238-1274.

[2] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.

[3] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529-533.

[4] Mao H, Alizadeh M, Menache I, et al. Resource management with deep reinforcement learning//Proceedings of the 15th ACM workshop on hot topics in networks. Atlanta, USA, 2016: 50-56.

[5] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. nature, 2016, 529(7587): 484-489.

[6] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[7] Wang Liang, Wang Wen, Wang Yu-You, et al. Feasibility of reinforcement learning for UAV-based target searching in a simulated communication denied environment. Science in China(Information Sciences), 2020, 50(3):21.(in Chinese)
(汪亮, 王文, 王禹又, 等. 强化学习方法在通信拒止战场仿真环境中多无人机目标搜寻问题上的适用性研究. 中国科学：信息科学, 2020, 50(3):21.)

[8] Ruan Xiao-Gang, Li Peng, Zhu Xiao-Qin, et al. A Visual Navigation Method Based on Goal-Driven Behavior and Space Topological Memory. Chinese Journal of Computers, 2021, 44(3):15.(in Chinese)
(阮晓钢, 李鹏, 朱晓庆, 等. 基于目标导向行为和空间拓扑记忆的视觉导航方法. 计算机学报, 2021, 44(3):15.)

[9] Xu Xiao-Long, Fang Zi-Jie, Qi Lian-Yong, et al. A Deep Reinforcement Learning-Based Distributed Service Offloading Method for Edge Computing Empowered Internet of Vehicles. Chinese Journal of Computers, 2021, 44(12):2382-2405.(in Chinese)
(许小龙, 方子介, 齐连永, 等. 车联网边缘计算环境下基于深度强化学习的分布式服务卸载方法. 计算机学报, 2021, 44(12):2382-2405.)

[10] Liu Xiao-Yu, Xu Chi, Zeng Peng, et al. Deep Reinforcement Learning-Based High Concurrent Computing Off loading for Heterogeneous Industrial Tasks. Chinese Journal of Computers, 2021, 44(12):2367-2381.(in Chinese)
(刘晓宇, 许驰, 曾鹏, 等. 面向异构工业任务高并发计算卸载的深度强化学习算法. 计算机学报, 2021, 44(12):2367-2381.)

[11] Zhang Si-Si, Zuo Xin, Liu Jian-Wei The Problem of the Adversarial Examples in Deep Learning. Chinese Journal of Computers, 2019, 42(8):1886-1904.(in Chinese)
(张思思, 左信, 刘建伟. 深度学习中的对抗样本问题. 计算机学报, 2019, 42(8):1886-1904.)

[12] Behzadan V, Munir A. The faults in our pi stars: Security issues and open challenges in deep reinforcement learning. arXiv preprint arXiv:1810.10369, 2018.

[13] Behzadan V, Munir A. Vulnerability of deep reinforcement learning to policy induction attacks//International Conference on Machine Learning and Data Mining in Pattern Recognition. New York, USA, 2017: 262-275.

[14] Gleave A, Dennis M, Wild C, et al. Adversarial policies: Attacking deep reinforcement learning. arXiv preprint arXiv:1905.10615, 2019.

[15] Hussenot L, Geist M, Pietquin O. CopyCAT: Taking control of neural policies with constant attacks. arXiv preprint arXiv:1905.12282, 2019.

[16] Liu A, Huang T, Liu X, et al. Spatiotemporal attacks for embodied agents//European Conference on Computer Vision. Glasgow, UK, 2020: 122-138.

[17] Guo J, Chen Y, Hao Y, et al. Towards comprehensive testing on the robustness of cooperative multi-agent reinforcement learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans, USA, 2022: 114-121.

[18] Ilahi I, Usama M, Qadir J, et al. Challenges and countermeasures for adversarial attacks on deep reinforcement learning. IEEE Transactions on Artificial Intelligence, 2021, 3(2): 90-109.

[19] Chen Jin-Yin, Zhang Yan, Wang Xue-Ke, et al. A Survey of Attack, Defense and Related Security Analysis for Deep Reinforcement Learning. Acta Automatica Sinica, 2020, 48(1):21-39.(in Chinese)
(陈晋音, 章艳, 王雪柯, 等. 深度强化学习的攻防与安全性分析综述. 自动化学报, 2020, 48(1):21-39.)

[20] Chen T, Liu J, Xiang Y, et al. Adversarial attack and defense in reinforcement learning-from AI security view. Cybersecurity, 2019, 2(1): 1-22.

[21] Sutton R S, Barto A G. Reinforcement learning: An introduction. USA：MIT press, 2018.

[22] Singh S P, Sutton R S. Reinforcement learning with replacing eligibility traces. Machine learning, 1996, 22(1): 123-158.

[23] Tesauro G. Temporal difference learning and TD-Gammon. Communications of the ACM, 1995, 38(3): 58-68.

[24] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms//Proceedings of the 31th International Conference on Machine Learning. Beijing, China, 2014: 387-395.

[25] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

[26] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning//Proceedings of the 33rd International Conference on Machine Learning. New York, USA, 2016: 1928-1937.

[27] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor// Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 1861-1870.

[28] Brown T B, Man é D, Roy A, et al. Adversarial patch. arXiv preprint arXiv:1712.09665, 2017.

[29] Liu A, Liu X, Fan J, et al. Perceptual-sensitive gan for generating adversarial patches//Proceedings of the AAAI conference on artificial intelligence. Honolulu, USA, 2019, 33(01): 1028-1035.

[30] Wang J, Liu A, Yin Z, et al. Dual attention suppression attack: Generate adversarial camouflage in physical world//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8565-8574.

[31] Duan R, Ma X, Wang Y, et al. Adversarial camouflage: Hiding physical-world attacks with natural styles//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 997-1005.

[32] Zhu X, Li X, Li J, et al. Fooling thermal infrared pedestrian detectors in real world using small bulbs//Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 3616-3624.

[33] Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1625-1634.

[34] Nguyen D L, Arora S S, Wu Y, et al. Adversarial light projection attacks on face recognition systems: A feasibility study//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA, 2020: 3548-3556.

[35] Liu A, Wang J, Liu X, et al. Bias-based universal adversarial patch attack for automatic check-out//European Conference on Computer Vision. Glasgow, UK, 2020: 395-410.

[36] Wang J, Liu A, Bai X, et al. Universal adversarial patch attack for automatic checkout using perceptual and attentional bias. IEEE Transactions on Image Processing, 2021, 31: 598-611.

[37] Zhang Y, Foroosh H, David P, et al. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild//International Conference on Learning Representations. New Orleans, USA, 2018.

[38] Huang L, Gao C, Zhou Y, et al. Universal physical camouflage attacks on object detectors//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 720-729.

[39] Brockman G, Cheung V, Pettersson L, et al. Openai gym. arXiv preprint arXiv:1606.01540, 2016.

[40] Todorov E, Erez T, Tassa Y. Mujoco: A physics engine for model-based control//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura, Portugal, 2012: 5026-5033.

[41] Wymann B, Espi é E, Guionneau C, et al. Torcs, the open racing car simulator, http://torcs. sourceforge. net, 2000.

[42] Terry J, Black B, Grammel N, et al. Pettingzoo: Gym for multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 2021, 34: 15032-15043.

[43] Vinyals O, Ewalds T, Bartunov S, et al. Starcraft ii: A new challenge for reinforcement learning. arXiv preprint arXiv:1708.04782, 2017.

[44] Samvelyan M, Rashid T, De Witt C S, et al. The starcraft multi-agent challenge. arXiv preprint arXiv:1902.04043, 2019.

[45] Huang S, Papernot N, Goodfellow I, et al. Adversarial attacks on neural network policies. arXiv preprint arXiv:1702.02284, 2017.

[46] Lin Y C, Hong Z W, Liao Y H, et al. Tactics of adversarial attack on deep reinforcement learning agents. arXiv preprint arXiv:1703.06748, 2017.

[47] Inkawhich M, Chen Y, Li H. Snooping attacks on deep reinforcement learning. arXiv preprint arXiv:1905.11832, 2019.

[48] Russo A, Proutiere A. Optimal attacks on reinforcement learning policies. arXiv preprint arXiv:1907.13548, 2019.

[49] Tretschk E, Oh S J, Fritz M. Sequential attacks on agents for long-term adversarial goals. arXiv preprint arXiv:1805.12487, 2018.

[50] Xiang Y, Niu W, Liu J, et al. A pca-based model to predict adversarial examples on q-learning of path finding//Proceedings of the IEEE International Conference on Data Science in Cyberspace (DSC). Guangzhou, China, 2018: 773-780.

[51] Bai X, Niu W, Liu J, et al. Adversarial examples construction towards white-box Q table variation in DQN pathfinding training//Proceedings of the IEEE International Conference on Data Science in Cyberspace (DSC). Guangzhou, China, 2018: 781-787.

[52] Das A, Datta S, Gkioxari G, et al. Embodied question answering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1-10.

[53] Xiao C, Pan X, He W, et al. Characterizing attacks on deep reinforcement learning. arXiv preprint arXiv:1907.09470, 2019.

[54] Pan X, You Y, Wang Z, et al. Virtual to real reinforcement learning for autonomous driving. arXiv preprint arXiv:1704.03952, 2017.

[55] Behzadan V, Hsu W. Adversarial exploitation of policy imitation. arXiv preprint arXiv:1906.01121, 2019.

[56] Zhang X, Ma Y, Singla A, et al. Adaptive reward-poisoning attacks against reinforcement learning//International Conference on Machine Learning. 2020: 11225-11234.

[57] Han Y, Rubinstein B I P, Abraham T, et al. Reinforcement learning for autonomous defence in software-defined networking//International Conference on Decision and Game Theory for Security. Seattle, USA, 2018: 145-165.

[58] Lee X Y, Ghadai S, Tan K L, et al. Spatiotemporally constrained action space attacks on deep reinforcement learning agents//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 4577-4584.

[59] Liu G, Lai L. Provably efficient black-box action poisoning attacks against reinforcement learning//Advances in Neural Information Processing Systems. 2021, 34: 12400-12410.

[60] Jin C, Allen-Zhu Z, Bubeck S, et al. Is Q-learning provably efficient?//Advances in Neural Information Processing Systems. Montreal, Canada, 2018: 4868-4878.

[61] Azar M G, Osband I, Munos R. Minimax regret bounds for reinforcement learning//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017: 263-272.

[62] Wu X, Guo W, Wei H, et al. Adversarial policy training against deep reinforcement learning//30th USENIX Security Symposium (USENIX Security 21). 2021: 1883-1900.

[63] Guo W, Wu X, Huang S, et al. Adversarial policy learning in two-player competitive games//Proceedings of he 38th International Conference on Machine Learning. 2021: 3910-3919.

[64] Wang Y, Sarkar E, Li W, et al. Stop-and-go: Exploring backdoor attacks on deep reinforcement learning-based traffic congestion control systems. IEEE Transactions on Information Forensics and Security, 2021, 16: 4772-4787.

[65] Krajzewicz D, Erdmann J, Behrisch M, et al. Recent development and applications of SUMO-Simulation of Urban MObility. International journal on advances in systems and measurements, 2012, 5(3&4).

[66] Chattopadhyay P, Hoffman J, Mottaghi R, et al. Robustnav: Towards benchmarking robustness in embodied navigation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 15691-15700.

[67] Gu T, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733, 2017.

[68] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[69] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[70] Liu A, Liu X, Yu H, et al. Training robust deep neural networks via adversarial noise propagation. IEEE Transactions on Image Processing, 2021, 30: 5769-5781.

[71] Kos J, Song D. Delving into adversarial attacks on deep policies. arXiv preprint arXiv:1705.06452, 2017.

[72] Mandlekar A, Zhu Y, Garg A, et al. Adversarially robust policy learning: Active construction of physically-plausible perturbations//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver, Canada, 2017: 3932-3939.

[73] Behzadan V, Munir A. Whatever does not kill deep reinforcement learning, makes it stronger. arXiv preprint arXiv:1712.09344, 2017.

[74] Behzadan V, Munir A. Mitigation of policy manipulation attacks on deep q-networks with parameter-space noise//International Conference on Computer Safety, Reliability, and Security. Vasteraas, Sweden, 2018: 406-417.

[75] Pattanaik A, Tang Z, Liu S, et al. Robust deep reinforcement learning with adversarial attacks. arXiv preprint arXiv:1712.03632, 2017.

[76] Chen T, Niu W, Xiang Y, et al. Gradient band-based adversarial training for generalized attack immunity of A3C path finding. arXiv preprint arXiv:1807.06752, 2018.

[77] Nisioti E, Bloembergen D, Kaisers M. Robust Multi-agent Q-learning in Cooperative Games with Adversaries, http://aaai-rlg.mlanctot.info/2021/papers/AAAI21-RLG_paper_21.pdf, 2021.

[78] Lin Y C, Liu M Y, Sun M, et al. Detecting adversarial attacks on neural network policies with visual foresight. arXiv preprint arXiv:1710.00814, 2017.

[79] Havens A, Jiang Z, Sarkar S. Online robust policy learning in the presence of unknown adversaries//Advances in neural information processing systems. Montreal, Canada, 2018: 9938-9948.

[80] Fischer M, Mirman M, Stalder S, et al. Online robustness training for deep reinforcement learning. arXiv preprint arXiv:1911.00887, 2019.

[81] Tessler C, Efroni Y, Mannor S. Action robust reinforcement learning and applications in continuous control//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 6215-6224

[82] Wu F, Li L, Huang Z, et al. Crop: Certifying robust policies for reinforcement learning through functional smoothing. arXiv preprint arXiv:2106.09292, 2021.

[83] Wu F, Li L, Xu C, et al. COPA: Certifying Robust Policies for Offline Reinforcement Learning against Poisoning Attacks. arXiv preprint arXiv:2203.08398, 2022.

[84] Wang J, Liu Y, Li B. Reinforcement learning with perturbed rewards//Proceedings of the AAAI conference on artificial intelligence. New York, USA, 2020: 6202-6209.

[85] Gallego V, Naveiro R, Insua D R. Reinforcement Learning under Threats//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 9939-9940.

[86] Ying C, Zhou X, Su H, et al. Towards Safe Reinforcement Learning via Constraining Conditional Value-at-Risk. arXiv preprint arXiv:2206.04436, 2022.

[87] Leike J, Martic M, Krakovna V, et al. AI safety gridworlds. arXiv preprint arXiv:1711.09883, 2017.

[88] Pinto L, Davidson J, Gupta A. Supervision via competition: Robot adversaries for learning tasks//IEEE International Conference on Robotics and Automation (ICRA). Singapore, Singapore, 2017: 1601-1608.

[89] Korkmaz E. Investigating vulnerabilities of deep neural policies//Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence. 2021: 1661-1670.

[90] Pinto L, Davidson J, Sukthankar R, et al. Robust adversarial reinforcement learning//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017: 2817-2826.

[91] Ogunmolu O, Gans N, Summers T. Minimax iterative dynamic game: Application to nonlinear robot control tasks//IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain, 2018: 6919-6925.

[92] Li S, Wu Y, Cui X, et al. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 4213-4220.

[93] Dong Y, Su H, Zhu J, et al. Towards interpretable deep neural networks by leveraging adversarial examples. arXiv preprint arXiv:1708.05493, 2017.

[94] Zhang C, Liu A, Liu X, et al. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. IEEE Transactions on Image Processing, 2020, 30: 1291-1304.

[95] Li T, Liu A, Liu X, et al. Understanding adversarial robustness via critical attacking route. Information Sciences, 2021, 547: 568-578.

[96] Wang Y, Su H, Zhang B, et al. Interpret neural networks by identifying critical data routing paths//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8906-8914.

[97] Ilyas A, Santurkar S, Tsipras D, et al. Adversarial examples are not bugs, they are features//Advances in neural information processing systems. Vancouver, Canada, 2019: 125-136

[98] Liu A, Liu X, Yu H, et al. Training robust deep neural networks via adversarial noise propagation. IEEE Transactions on Image Processing, 2021, 30: 5769-5781.

[99] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680.

[100] Stone P, Veloso M. Multiagent systems: A survey from a machine learning perspective. Autonomous Robots, 2000, 8(3): 345-383.

[101] Tesauro G, Kephart J O. Pricing in agent economies using multi-agent Q-learning. Autonomous agents and multi-agent systems, 2002, 5(3): 289-304.

[102] Multiagent systems: a modern approach to distributed artificial intelligence. MIT press, 1999.

[103] Lin J, Dzeparoska K, Zhang S Q, et al. On the robustness of cooperative multi-agent reinforcement learning//IEEE Security and Privacy Workshops. San Francisco, USA, 2020: 62-68.

[104] Rashid T, Samvelyan M, Schroeder C, et al. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 4292-4301.

[105] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings//IEEE European Symposium on Security and Privacy. Saarbrucken, Germany, 2016: 372-387.

[106] Croce F, Andriushchenko M, Sehwag V, et al. Robustbench: a standardized adversarial robustness benchmark. arXiv preprint arXiv:2010.09670, 2020.

[107] Tang S, Gong R, Wang Y, et al. Robustart: Benchmarking robustness on architecture design and training techniques. arXiv preprint arXiv:2109.05211, 2021.

[108] Wang B, Xu C, Wang S, et al. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. arXiv preprint arXiv:2111.02840, 2021.

[109] Behzadan V, Munir A. Adversarial reinforcement learning framework for benchmarking collision avoidance mechanisms in autonomous vehicles. IEEE Intelligent Transportation Systems Magazine, 2019, 13(2): 236-241.

[110] Behzadan V, Hsu W. Rl-based method for benchmarking the adversarial resilience and robustness of deep reinforcement learning policies//International Conference on Computer Safety, Reliability, and Security. Turku, Finland, 2019: 314-325.

[111] Dulac-Arnold G , Levine N , Mankowitz D J , et al. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. Machine Learning, 2021, 110(9):2419-2468.

附录A.

LIU Ai-Shan, Ph.D., assistant professor. His main research interests include adversarial example, model robustness, and the trustworthiness in AI.

GUO Jun, master candidate. His main research interests focus on the trustworthiness in AI.

LI Si-Min, Ph.D. candidate. His main research interests focus on reinforcement learning and the trustworthiness in AI.

XIAO Yi-Song, Ph.D. candidate. His main research interests focus on the trustworthiness in AI.

LIU Xiang-Long, Ph.D., professor. His research interests include computer vision and trustworthiness in AI.

TAO Da-Cheng, Ph.D., professor. His research interests include computer vision and artificial intelligence.

Background

With the spreading of deep learning, deep reinforcement learning technique has been widely used and drawn extensive research attention in multiple research fields such as robots, games, and auto driving, etc. It is a new learning paradigm associated with the development of deep neural networks, that integrates the perception of deep learning and decision making of reinforcement learning. However, adversarial examples, visually imperceptible perturbations that could mislead deep learning into wrong predictions, have emerged and highly challenged the safety of deep reinforcement learning algorithms and applications especially in the safety-critical scenarios. To better understand and further promote the development of this area, this paper therefore provides a comprehensive survey on the adversarial attacks and defenses for deep reinforcement learning.

The research topic primarily belongs to the intersection of adversarial machine learning, reinforcement learning, and the safety of artificial intelligence. There exists a plethora of works that focus on proposing new adversarial attack and defense algorithms on deep reinforcement learning, however, little attempts have been devoted on the comprehensive review of this field. For example, the latest review of DRL attacks and defenses would be dated back to 2020 and 2018, which could not fully include the state-of-the-art literature much less the exposition of challenges or future directions in this field.

In this paper, we comprehensively review the literature of adversarial learning in deep reinforcement learning. The primary goal of this paper is to better understand the development and future directions of deep reinforcement learning field, and further promote the studies of adversarial attacks and defenses of deep reinforcement learning, which we hope to lead the safer applications. This paper first presents the preliminary backgrounds including deep reinforcement learning, adversarial examples, and related datasets and benchmarks in deep reinforcement learning field. Based on the perturbing spaces of Markov decision process in deep reinforcement learning, we analyze and summarize adversarial attacks in deep reinforcement learning from the perspectives of state-based, reward-based, and action-based attacks. We then illustrate the framework of adversarial attacks in deep reinforcement learning. By aligning deep reinforcement learning defenses with traditional adversarial defenses framework (e.g., adversarial training, adversarial detection, etc.), we then summarize the adversarial defenses for deep reinforcement learning from adversarial training, adversarial detection, certified robustness, and robust learning. Moreover, this paper investigates interesting and meaningful topics for the applications of adversarial examples in the deep reinforcement learning fields, including model robustness understanding and exploiting adversarial attacks for better model performance in deep reinforcement learning. Finally, this paper highlights the open issues and future challenges in the deep reinforcement learning field from four main perspectives, including deep reinforcement learning robustness (theories), multi-agent deep reinforcement learning attacks and defenses (techniques), benchmarks and environment for deep reinforcement learning attacks and defenses (platforms), and physical world adversarial attacks and defenses on deep reinforcement learning (applications).

Systematic analysis and survey of attacks/defenses on deep reinforcement learning is highly beneficial on effectively improving the interpretability of deep reinforcement learning, enhancing the security of the model, and further promoting the building of safe and reliable deep reinforcement learning

applications. We hope this paper could help the researchers to better understand the framework of adversarial machine learning in the deep reinforcement learning field, and further promote the development and applications of deep reinforcement learning in the safety-critical scenarios in the future.

This work was supported by The National Key Research and Development Plan of China (2020AAA0103502), National Natural Science Foundation of China (No. 62022009 and 62206009), and the State Key Laboratory of Software Development Environment under (SKLSDE-2022ZX-23).