自动化学报
Acta Automatica Sinica
ISSN 0254-4156,CN 11-2109/TP

《自动化学报》网络首发论文

题目：深度强化学习的攻防与安全性分析综述
作者：陈晋音，章燕，王雪柯，蔡鸿斌，王珏，纪守领
DOI：10.16383/j.aas.c200166
收稿日期：2020-04-01
网络首发日期：2020-09-18
引用格式：陈晋音，章燕，王雪柯，蔡鸿斌，王珏，纪守领．深度强化学习的攻防与安全性分析综述．自动化学报. https://doi.org/10.16383/j.aas.c200166

网络首发时间：2020-09-18 15:46:42
网络首发地址：https://kns.cnki.net/kcms/detail/11.2109.TP.20200918.1202.001.html

深度强化学习的攻防与安全性分析综述

陈晋音 [1,2]　章燕 [1]　王雪柯 [1]　蔡鸿斌 [3]　王珏 [1]　纪守领 [4]

摘要　深度强化学习是人工智能领域新兴技术之一，它将深度学习强大的特征提取能力与强化学习的决策能力相结合，实现从感知输入到决策输出的端到端框架，具有较强的学习能力且应用广泛。然而，已有研究表明深度强化学习存在安全漏洞，容易受到对抗样本攻击。为提高深度强化学习的鲁棒性、实现系统的安全应用，本文针对已有的研究工作，较全面地综述了深度强化学习方法、对抗攻击、防御方法与安全性分析，并总结深度强化学习安全领域存在的开放问题以及未来发展的趋势，旨在为从事相关安全研究与工程应用提供基础。

关键词　深度强化学习、对抗攻击、防御、策略攻击、安全性
DOI　10.16383/j.aas.c200166

A Survey of Attack, Defense and Related Security Analysis for Deep Reinforcement Learning

Chen Jin-Yin[1,2]　Zhang Yan[1]　Wang Xue-Ke[1]　Cai Hong-Bin[3]　Wang Jue[1]　Ji Shou-Ling[4]

Abstract　Deep reinforcement learning is one of the emerging technologies in the field of artificial intelligence. It combines the powerful feature extraction capabilities of deep learning with the decision-making capabilities of reinforcement learning to achieve an end-to-end framework from status input to the decision output, which also makes it regarded as an important way to general artificial intelligence. However, existing studies have shown that deep reinforcement learning has security vulnerabilities and is vulnerable to adversarial sample attacks. In order to improve the robustness of deep reinforcement learning and realize the security application of the system, this article comprehensively summarizes deep reinforcement learning methods, adversarial attacks, defense methods and security analysis based on existing research work, and summarizes deep reinforcement learning security The open problems in the field and future development trends are intended to provide a basis for relevant safety research and engineering applications.

Key words　Deep Reinforcement Learning, Adversarial attack, Defense, Policy attack, Security.
Citation　A survey of attack and defense related security analysis for deep reinforcement learning. Acta Automatica Sinica

收稿日期 2020-4-1　　录用日期 2020-9-7
Manuscript received April 1, 2020；accepted September 7, 2020
浙江省自然科学基金（LY19F020025）资助，宁波市“科技创新 2025”重大专项（2018B10063）资助，科技创新 2030—“新一代人工智能”重大项目（2018AAA0100800）资助
Supported by the Zhejiang Provincial Natural Science Foundation of China（LY19F020025）, the Major Special Funding for ``Science and Technology Innovation 2025" in Ningbo（2018B10063）, and the National Key Research and Development Program of China（2018AAA0100800）
本文责任编委 张化光
Recommended by Associate Editor ZHANG Hua-Guang
1. 浙江工业大学信息工程学院 杭州 310023 2. 浙江工业大学网络安全研究院 杭州 310023 3. 华东师范大学软件工程学院 上海 200062 4. 浙江大学计算机科学与技术学院 杭州 310058
1.College of Information Engineering，Zhejiang University of Technology，Hangzhou 310023 2. Institute of Cyberspace Security, Zhejiang University of Technology，Hangzhou 310023 3. School of Software Engineering, East China Normal University Shanghai 200062 4. College of Computer Science and Technology, Zhejiang University, Hangzhou 310058

自 Mnih[1]将深度学习与强化学习结合并提出第一个深度强化学习框架——深度Q网络（Deep Q Network，DQN）以来，深度强化学习（Deep Reinforcement Learning，DRL）方法就被人们视为迈向通用人工智能的必要路径之一。随后，各种强化学习的改进算法不断提出，例如：基于值函数的算法有双重深度Q网络(DDQN)[1-2]、优先经验回放Q网络(Prioritized DQN)[3]、对偶深度Q网络(Dueling DQN)[4]等，基于策略的强化学习算法有异步/同步优势行动者评论者(A3C/A2C)[5]、信任域策略优化(TRPO)[6]、K因子信任域行动者评论者算法(ACKTR)[7]等。基于深度强化学习的应用领域也非常广泛，例如：游戏博弈[8-9]、自动驾驶[10]、医疗健康[11]、金融交易[12]、机器人控制[13]、网络安全[14]、计算机视觉[15-16]等。为加强深度强化学习在安全攸关领域的安全应用，及早发现深度强化学习算法漏洞，防止恶意用户利用这些漏洞进行非法牟利行为。不同于传统机器学习的单步预测任务，深度强化学习系统利用多步决策完成特定任务，且连续决策之间具有高度相关性。总体来说，深度强化学习系统的攻击可针对强化学习算法的五个主要环节展开恶意攻击，包括：环境、观测、奖励、动作以及策略[17]。

Huang[18]最早于2017年对深度强化学习系统存在的漏洞做出了相关研究。他将机器学习安全领域中面临的对抗攻击应用到了深度强化学习模型中，通过在智能体的观测状态添加对抗扰动，令整个深度强化学习系统性能显著下降。随后，针对特定应用，Chen等人[19]在自动寻路任务中通过在环境中添加“挡板状”障碍物，使智能体无法抵达目的地。Tretschk等人[20]通过对抗变换网络修改Pong智能体训练时维护的奖励目标，使智能体的训练朝着游戏失败的方向进行。Ferdowsi[21]在第21届智能交通系统国际会议上提出了此类问题对自动驾驶应用的影响。因此深度强化学习系统真正应用到实际工业界之前，探究深度强化学习系统的脆弱点、提高其防御能力与鲁棒性十分重要。

为了提高深度学习模型的鲁棒性，已有研究提出了较多DRL防御方法，主要包括三个方向：对抗训练、鲁棒学习、对抗检测。例如：Behzadan[22]提出了使用对抗训练实现梯度攻击的防御；Gu等人[23]采用DRL训练对抗智能体，与目标系统的智能体进行零和博弈提升其鲁棒性；Lin[24]借助预测帧模型，通过比较策略对预测帧与当前输出的KL散度概率分布实现攻击检测。

目前，深度强化学习领域的攻防研究还有很大发展空间，针对深度强化学习存在的易受对抗样本攻击等问题，深度强化学习模型的鲁棒性优化以及对抗防御方法也已成为重点关注对象，仍需不断探索。同时由于深度强化学习在安全攸关领域的应用，其存在的策略漏洞也成为了一大安全隐患。为了更好地探究深度强化学习系统的攻防研究现状与未来发展方向，本文针对深度强化学习算法、攻击与防御方法，以及安全性分析展开尽可能全面的综述。

论文接下去章节安排如下：第2节介绍主要的深度强化学习算法，第3节针对强化学习的五个方面介绍攻击方法，第4节介绍相应的防御方法，第5节分析深度强化学习的安全性，第6节相关应用平台及评估指标。最后，总结并列举未来可能的研究方向。

1 深度强化学习方法

强化学习（Reinforcement Learning, RL）是一种智能体通过利用与环境交互得到的经验来优化决策的过程[17]。强化学习问题通常可以被建模为马尔科夫决策过程（Markov Decision Process，MDP），可以由一个四元组表示$MDP=(S,A,R,P)$，其中S表示决策过程中所能得到的状态集合，A表示决策过程中的动作集合，R表示用于对状态转移做出的即刻奖励，P则为状态转移概率。在任意时间步长t的开始，智能体观察环境得到当前状态s_t，并且根据当前的最优策略π^*做出动作a_t。在t的最后，智能体得到其奖励r_t及下一个观测状态s_{t+1}。MDP的目标就是找到最佳的动作序列以最大化长期的平均奖励。深度强化学习则是在强化学习的基础上结合了深度学习强大的特征提取能力，避免了特征人工提取，实现了从原始图像输入到决策结果输出的端到端学习系统。

常用的深度强化学习通常被分为两类：基于值函数的深度强化学习和基于策略梯度的深度强化学习。前者主要通过深度神经网络逼近目标动作价值函数，表示到达某种状态或执行某种动作得到的累积回报，它倾向于选择价值最大的状态或动作，但是它们的训练过程往往不够稳定，而且不能处理动作空间连续的任务；基于策略梯度的深度强化学习则是将策略参数化，利用深度神经网络逼近策略，同时沿着策略梯度的方向来寻求最优策略。策略梯度算法在训练过程中更

加稳定，但是算法实现比较复杂且在通过采样的方式进行学习时会导致方差较大。下面我们对比两类方法中具有代表性的算法，分别对其原理、贡献与不足进行阐述，如表 1 所示。

表 1 经典深度强化学习算法对比

Table 1 Comparison of classic deep reinforcement learning algorithm

分类	算法	原理	贡献	不足
基于值函数	深度 Q 网络 (DQN) [1-2]	使用经验回放机制打破样本相关性； 使用目标网络稳定训练过程	第一个能进行端到端学习的深度强化学习框架	训练过程不稳定； 无法处理连续动作任务；
	双重深度 Q 网络 (DDQN) [3]	用目标网络来评估价值， 用评估网络选择动作	缓解了 DQN 对价值的过估计问题	训练过程不稳定； 无法处理连续动作
	优先经验回放 Q 网络 (Prioritized DQN) [4]	对经验池中的训练样本设立优先级进行采样	提高对稀有样本的使用效率	训练过程不稳定； 无法处理连续动作
	对偶深度 Q 网络 (Dueling DQN) [25]	对偶网络结构，使用状态价值函数，与相对动作价值函数来评估 Q 值	存在多个价值相仿的动作时提高了评估的准确性	无法处理连续动作
	深度循环 Q 网络 (DRQN) [26]	用长短时记忆网络替换全连接层	缓解了部分可观测问题	完全可观测环境下性能表现不足；无法处理连续动作
	注意力机制深度循环 Q 网络 (DARQN)[27]	引入注意力机制	减轻网络训练的运算代价	训练过程不稳定； 无法处理连续动作
	噪声深度 Q 网络 (Noisy DQN)[28]	在网络权重中加入参数噪声	提高了探索效率； 减少了参数设置；	训练过程不稳定； 无法处理连续动作
	循环回放分布式深度 Q 网络（R2D2）[29]	RNN 隐藏状态存在经验池中；采样部分序列产生 RNN 初始状态；	减缓了 RNN 状态滞后性	状态滞后和表征漂移问题仍然存在
	演示循环回放分布式深度 Q 网络（R2D3）[31]	经验回放机制； 专家演示回放缓冲区； 分布式优先采样；	解决了在初始条件高度可变的部分观察环境中的稀疏奖励任务	无法完成记住和越过传感器的任务
基于策略梯度	REINFORCE[33]	使用随机梯度上升法； 累计奖励作为动作价值函数的无偏估计	策略梯度是无偏的	存在高方差; 收敛速度慢
	自然策略梯度（Natural PG）[34]	自然梯度朝贪婪策略方向更新	收敛速度更快； 策略更新变化小	自然梯度未达到有效最大值
	行动者-评论者 (AC)[35]	Actor 用来更新策略； Critic 用来评估策略	解决高方差的问题	AC 算法中策略梯度存在较大偏差
	确定性策略梯度 (DDPG)[36]	确定性策略理论；	解决了连续动作问题	无法处理离散动作问题
	异步/同步优势行动者-评论者 (A3C/A2C) [5]	使用行动者评论者网络结构； 异步更新公共网络参数	用多线程提高学习效率； 降低训练样本的相关性； 降低对硬件的要求；	内存消耗大； 更新策略时方差较大
	信任域策略优化 (TRPO) [6]	用 KL 散度限制策略更新	保证了策略朝着优化的方向更新	实现复杂； 计算开销较大
	近端策略优化 (PPO) [37]	经过裁剪的替代目标函数自适应的 KL 惩罚系数	比 TRPO 更容易实现； 所需要调节的参数较少	用偏差大的大数据批进行学习时无法保证收敛性
	K 因子信任域行动者评论者算法 (ACKTR) [7]	信任域策略优化； Kronecker 因子算法； 行动者评论者结构；	采样效率高； 显著减少计算量	计算依然较复杂

1.1 基于值函数的深度强化学习

基于值函数的 DRL 通过维护更新价值网络参数来得到最优策略，其最初的灵感来源于 RL 中的 Q 学习[35]。Q 学习旨在通过贝尔曼方程，采用时序差分的方式进行迭代更新状态-动作价值函数 Q，使 Q 函数逼近至真实值 Q^*，从而最终得到最优策略：

$$Q_\pi(s,a) = \mathrm{E}_\pi[G_t \mid S_t = s, A_t = a] \quad (1)$$

$$\pi^* = \arg\max_a Q^*(s,a) \quad (2)$$

其中，$Q_\pi(s,a)$ 表示在状态 s 做出动作 a 后，遵循策略 π 的预期回报，G_t 表示从步骤 t 到终止状态的累积回报。尽管已经证明 Q 学习算法在解决一些顺序的决策问题时具有较好的表现，但是它仍然存在许多缺陷：（1）在复杂场景下，状态空间过大会导致 Q 表难以维护；（2）学习过程中，训练样本的高度连续性打破了机器学习的独立同分布要求；（3）由于 Q 学习是一种在线学习方式，一些不常见的样本在使用一次后就被放弃，导致样本使用效率低。

1.1.1 深度 Q 网络

为了克服上述缺点，Mnih 等人[1]首次将 Q 学习与深度神经网络结合，提出深度强化学习 Q 网络（DQN），并且证明经 DQN 训练的智能体在 Atrai 游戏上的技术水平能够达到人类水准。

DQN 采用深度卷积神经网络来逼近 Q 函数，解决了状态空间过大难以维护和特征提取的问题。同时，采用经验回放机制学习使训练数据成为独立同分布，降低了数据间的关联性，而且通过重复利用提高了对样本的利用率。此外，Mnih[2]在 2015 年提出了目标网络机制，目标网络是在原有 Q_θ 之外搭建一个结构完全相同的网络 Q_{θ^-}，减轻了每次 Q 值变化对策略参数的影响，增加了策略训练的稳定性。

1.1.2 深度 Q 网络的改进方法

针对 DQN 存在 Q 值估计偏差过大、训练不稳定等问题，提出了一些改进版的 DQN 方法。Van 等人[3]根据强化学习中的双重 Q 学习构建双重深度 Q 网络（Double Deep Q Network，DDQN），通过评估网络来选择动作、目标网络进行价值评估。针对 DQN 的经验回放机制采用平均随机采样机制，存在稀有样本利用率低的问题，Schaul 等人[4]提出了优先经验回放机制，定义经验优先级，并优先采用级别高的经验。Wang 等人[25]提出了 DQN 的对偶结构（Dueling Network），通过状态价值函数 V 和相对价值函数 A 来评估 Q 值。为了减少隐藏信息的代价，Hausknecht 等人[26]将 DQN 卷积层后的第一个全连接层替换为循环的长短时记忆网络，提出深度循环 Q 网络（Deep Recurrent Q Network，DRQN）。在此基础上，Sorokin 等人[27]加入注意力机制使得智能体在训练过程中关注图像中的某一点进行学习，即：深度注意力机制循环 Q 网络（Deep Attention Recurrent Q Network，DARQN）。Plapper 等人[28]用噪声网络来替代原先的 ε-贪婪探索策略。通过将参数化的自适应噪声加入到的 DQN 网络权重中，驱动智能体探索、简化训练难度。针对使用经验回放机制产生参数滞后而导致的表征漂移等问题，Steven 等人[29]提出了循环回放分布式深度 Q 网络（Recurrent Replay Distributed DQN，R2D2）。R2D2 使用全零状态初始化网络与回放完整轨迹两种方法来比较训练 LSTM[30]的差异，提出状态存储和“Burn-in”方法来训练随机采样的循环神经网络。更进一步，Gaglar 等人[31]提出演示循环回放分布式深度 Q 网络（Recurrent Replay Distributed DQN from Demonstrations，R2D3）。除了经验回放，R2D3 设计了一个专家演示回放缓冲区，学习者通过调整演示和经验之间的比率有效解决了初始条件高度可变的部分观察环境中的奖励稀疏任务。

1.2 基于策略梯度的深度强化学习

由于基于值函数的深度强化学习在处理连续动作空间的场景时需要对动作进行离散化处理，也就需要为众多动作分配 Q 值，给实际应用带来困难，并且 DQN 得到的策略无法处理随机策略问题，基于策略梯度的深度强化学习方法[32]应运而生，包括：异步优势行动者-评论者（Asynchronous Advantage Actor Critic，A3C）[5]、确定性策略梯度（Deterministic Policy Gradient，PGD）[8]和信任域策略优化（Trust Region Policy Optimization，TRPO）[6]以及一些改进方法。

基于策略梯度的深度强化学习通过深度神经网络对策略进行参数化建模：$\pi_\theta(s,a)=p(a\mid s,\theta)$，即对应每个状态采取不同动作的概率。在学习过程中，通过策略梯度直接在策略空间中搜索最优策略。

1.2.1 策略梯度

策略梯度算法的主要思想是将策略 π 参数化为 π_θ，表示对应的状态动作分布概率，然后计算出关于动作的策略梯度，沿着梯度方向来调整动作，最终找到最优策略。策略梯度的定义为：

$$g=\mathrm{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)Q_\pi(s,a)] \quad (3)$$

策略梯度算法中，根据策略的定义不同，又可以分别随机性策略与确定性策略。随机性策略是指在当前状态下，满足策略参数 θ 时的某个概率分布，其对应的动作可能是多

个。而确定性策略则是指对应于每个状态都输出唯一的动作。策略梯度常用于解决深度强化学习的连续控制问题，常见的策略梯度算法包括：REINFORCE 算法[33]、自然策略梯度算法（Natural Policy Gradient，Natural PG）[34]以及行动者-评论者算法（Actor-Critic，AC）[35]等。

1.2.2 异步优势行动者-评论者

基于经验回放的 DRL 算法将智能体与环境的交互数据存储在经验回放池中，训练时进行批量采样，减少了在线强化学习的数据相关性，通常只适用于离线策略强化学习中。针对上述问题，Mnih 等人[5]结合异步强化学习思想提出了异步优势行动者-评论者方法。

A3C 通过创建多个子线程，每个线程中智能体并行地与环境交互，实现异步学习，替代了经验回放机制，解决了在线策略的数据相关性的问题。A3C 在执行过程中采用异步更新网络参数的方式，各线程单独对环境采样并计算梯度，用各自得到的梯度通过累加异步更新到全局模型中，最后将全局模型参数拷贝到各个线程网络中。但是 A3C 的异步更新方式会使得各个线程会以不同的策略去对环境进行采样。对此，Mnih 等人[5]提出了同步的优势行动者-评论者（Advantage Actor Critic，A2C）方法。

相比于 A3C 异步更新全局模型的方式，A2C 中的各线程会将各自的采样计算得到的梯度先进行汇总，再用汇总结果更新全局模型参数。不仅解决了在线策略数据更新的相关性问题，同时使智能体在同一策略下进行交互学习。

1.2.3 确定性策略梯度

由于在连续动作空间中选取确定动作十分困难，为此 Silver[8]提出了确定性策略理论，并证明了确定性策略梯度的存在。Lillicrap 在此基础上结合了 AC 框架以及 DQN 中的机制，提出了深度确定性策略梯度算法（Deep Deterministic Policy Gradient，DDPG）[36]。

DDPG 使用参数为θ_{π}的策略网络和参数为θ_{Q}的动作价值网络分别作为 AC 框架中的行动者和执行者，同时使用经验回放机制进行批处理学习，使用目标网络机制来提高学习过程的稳定性。

1.2.4 信赖域策略优化

为了找到合适的步长使得策略一直向回报增加的方向更新，Schulman 等人[6]提出了信任域策略优化方法，通过 KL 散度来限制策略更新前后的分布差异，令更新步长处于信任域中，使策略的更新会朝着增加回报的方向前进。

理论上，TRPO 能保证更新后的策略比先前策略性能更好，在有限的策略空间中，最终能达到局部或全局最优解。在现实场景中，TRPO 也被证明拥有较好的鲁棒性与实用性。但是由于 TRPO 算法实现十分复杂，且计算代价过大，Schulman 等人[37]随后又提出了改进版本，即近端策略优化（Proximal Policy Optimization，PPO）算法。PPO 提升了采样的复杂度而简化了计算，同时使用了无约束优化，在保持性能同时降低了算法复杂度。Y.Wu 等人[7]结合行动者评论者算法提出了 Kronecker 因子信任域行动者评论者算法（Actor Critic using Kronecker-factored Trust Region，ACKTR），利用 Kronecker 因子减少算法所需的计算量。

2 深度强化学习的攻击方法

随着 DRL 的推广应用，通过攻击方法研究发现 DRL 的安全漏洞也引起广泛关注。为了系统分析各种不同的攻击方法，本文根据强化学习 MDP 中的关键环节对攻击方法进行归类，即：观测攻击、奖励攻击、动作攻击、环境攻击以及策略攻击，其攻击方法主要在 Atari 游戏场景以及自动导航的地图等场景上实现，各个环节攻击的展示如图 1 所示。

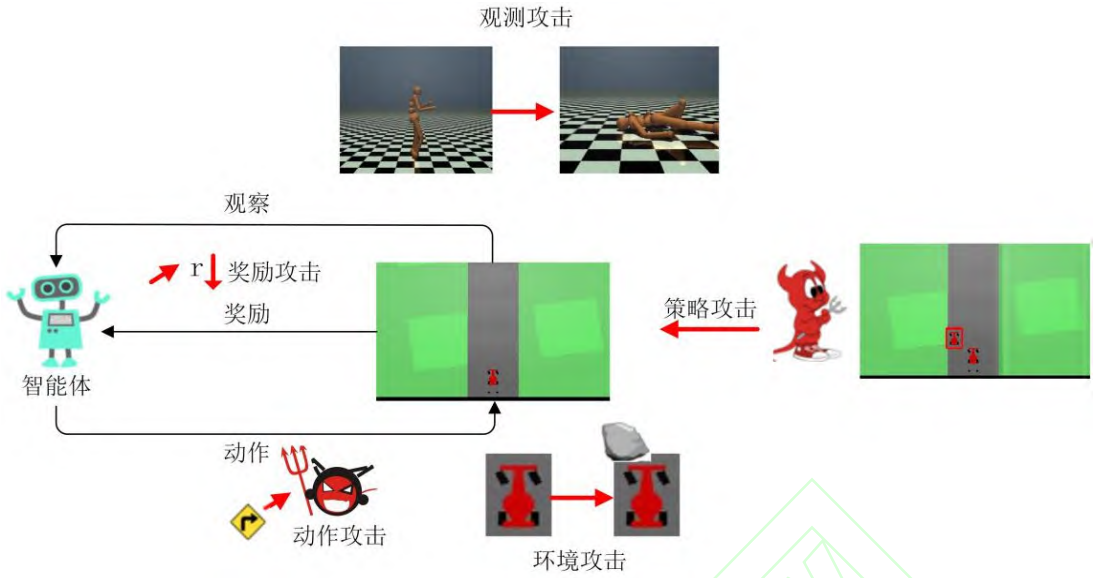


图 1 对 DRL 系统的不同类型攻击

Figure 1 Different types of attacks on DRL system

如图 1 所示，首先，观测攻击指攻击者在智能体所接收到的观测图像上添加扰动，使智能体做出攻击者预期的动作，通常在智能体的图像传感器上添加噪声来实现。不同于观测攻击，环境攻击是直接修改智能体的训练环境，主要通过对环境动态模型的修改以及在环境中加入阻碍物（并非在智能体的传感器上添加噪声）的方式来实现攻击。其次，奖励攻击指修改环境反馈的奖励信号，既可以是通过修改奖励值的符号，也可以使用对抗奖励函数取代原有的奖励函数来实现攻击。再次，策略攻击是指使用对抗智能体来生成目标智能体理解能力之外的状态和行为，继而导致目标智能体进入一种混乱状态。而动作攻击则是指修改动作输出，这种攻击方式可以通过修改训练数据中的动作空间来实现。

本文对 DRL 攻击方法和代表性技术进行了综述与对比，相关方法及其原理简述整理在表 2 中。同时也对攻击成功率进行统计，目前攻击效果统计中，奖励值分析占主流，极少论文提到攻击成功率，其中部分论文中的攻击成功率通过曲线图来展示动态结果，只有两篇论文中的攻击方法给出了具体的成功率数值，相关攻击模型及方法和对应成功率统计在表 3 中。

表 2 深度强化学习的攻击方法

Table 2 Attack methods toward deep reinforcement learning

分类	攻击方法	攻击模型	攻击策略	攻击阶段	对手知识
观测攻击（见 2.1）	FGSM[18]	DQN[1]、TRPO[6]、A3C[5]	在观测上加上 FGSM 攻击	测试阶段	白盒/黑盒
	策略诱导攻击[39]	DQN[1]	训练敌手策略；对抗样本的转移性	训练阶段	黑盒
	战略时间攻击[40]	DQN[1]、A3C[5]	在一些关键时间步进行攻击	测试阶段	白盒
	迷惑攻击[40]	DQN[1]、A3C[5]	通过预测模型诱导智能体做出动作	测试阶段	白盒
	基于值函数的对抗攻击[41]	A3C[5]	在值函数的指导下选择部分观测进行攻击	测试阶段	白盒
	嗅探攻击[42]	DQN[1]、PPO[37]	用观测以及奖励、动作信号来获取代理模型并进行攻击	测试阶段	黑盒
	基于模仿学习的攻击[43]	DQN[1]、A2C[5]、PPO[37]	使用模仿学习提取的专家模型信息进行攻击	测试阶段	黑盒

奖励攻击（见 2.2）	CopyCAT 算法[44]	DQN[1]	使用预先计算的掩码对智能体的观测做出实时的攻击	测试阶段	白盒/黑盒
	基于对抗变换网络的对抗攻击[20]	DQN[1]	加入一个前馈的对抗变换网络使策略追求对抗奖励	测试阶段	白盒
	木马攻击[45]	A2C[5]	在训练阶段用特洛伊木马进行中毒攻击	训练阶段	白盒/黑盒
	翻转奖励符号攻击[46]	DDQN[3]	翻转部分样本的奖励值符号	训练阶段	白盒
	路径脆弱点攻击[47]	DQN[1]	根据路径点 Q 值的差异与直线的夹角找出脆弱点	训练阶段	白盒
环境攻击（见 2.3）	通用优势对抗样本生成方法[19]	A3C[5]	在梯度上升最快的横断面上添加障碍物	训练阶段	白盒
	对环境模型的攻击[48]	DQN[1]、DDPG[36]	在环境的动态模型上增加扰动	测试阶段	黑盒
动作攻击（见 2.4）	动作空间扰动攻击[49]	PPO[37]、DDQN[3]	通过奖励函数计算动作空间扰动	训练阶段	白盒
策略攻击（见 2.5）	通过策略进行攻击[50]	PPO[37]	采用对抗智能体防止目标智能体完成任务	测试阶段	黑盒

表 3 深度强化学习的攻击和攻击成功率

Table 3 Attack success rate toward deep reinforcement learning

攻击模型	攻击方法	攻击阶段	攻击策略	平台	成功率
DQN[1]	CopyCAT 算法[44]	测试阶段	使用预先计算的掩码对智能体的观测做出实时的攻击	OpenAI Gym[74]	60%~100%
	FGSM 攻击[38]	训练阶段	在观测上加上 FGSM 攻击	OpenAI Gym[74]	90%~100%
	策略诱导攻击[38]	训练阶段	训练敌手策略；对抗样本的转移性	Grid-World map[38]	70%~95%
	战略时间攻击[40]	测试阶段	在一些关键时间步进行攻击	OpenAI Gym[74]	40 步以内达到 70%
PPO[37]	通过策略进行攻击[50]	测试阶段	采用对抗智能体防止目标智能体完成任务	OpenAI Gym[74]	玩家智能体成功率下降至 62%和 45%

2.1 基于观测的攻击

2.1.1 FGSM 攻击

Huang 等人[18]最先对通过深度强化学习得到的策略进行攻击，使用机器学习领域常用的快速梯度符号（Fast Gradient Sign Method, FGSM）[38]算法制造对抗扰动并将扰动直接添加到智能体的观测值上，以此对深度学习智能体进行攻击。FGSM 的主要思想是在深度学习模型梯度变化最大的方向添加扰动，导致模型输出错误结果，其数学表达式如下：

$$\eta = \varepsilon sign\ (\nabla_x J(\theta, x, y)) \qquad (4)$$

其中，J 表示损失函数，θ 表示模型参数，x 表示模型输入，y 样本类标（此处指最优动作项），$\nabla J(.,.,.)$ 表示计算损失函数对当前模型参数的梯度，$sign$ 表示符号函数，ε 表示扰动阈值。

实验证明，这种方法在白盒与黑盒设置下均有效。Huang 等人[18]首次尝试并验证了由 DQN、TRPO 以及 A3C 这些算法得到的智能体容易受到对抗性扰动的攻击，且对抗样本在不同强化学习算法得到的模型之间、在相同算法下得到的不同模型之间具有较好的迁移性。但是他的攻击方式依然遵循着机器学习模型在时间上的独立性，而没有考虑到强化学习问题在连续时间上高度的相关性。

2.1.2 策略诱导攻击

Behzadan 等人[39]认为由于深度强化学习系统在学习的过程中依赖于智能体与环境的交互，使得学习过程容易受到可观察环境变化的影响。因此他们使用基于深度学习

分类器的攻击，对 DQN 模型的观测进行了对抗扰动。

在攻击设置中，敌手知道目标模型的输入类型及奖励函数，可以根据目标模型的输入类型建立一个 DQN 副本，通过副本及奖励函数制造对抗样本，使目标 DQN 的训练朝向选择除最优动作 a_t 之外的动作 a_t' 进行学习。这种攻击方式可以视为对深度学习模型中的分类器黑盒攻击的扩展。但是这种攻击依然局限于传统机器学习在时间步上独立计算对抗样本的形式。

2.1.3 战略时间攻击

Lin 等人[40]认为，考虑部分强化学习问题中的奖励信号是稀疏的，对手没有必要在每个时间步都对智能体发起攻击。因此他们提出了一种新颖攻击方式：通过战略性地选择一些时间步进行攻击，以减少目标智能体的预期累积回报。提出了动作偏好函数来衡量当前状态下策略对动作的偏好程度，当偏好程度超过设定的阈值时就制造扰动进行攻击。

实验验证了攻击效果，战略时间攻击可以使用较少的攻击次数达到与 Huang[18]相同的效果。战略时间攻击相比于在所有观测值上都添加扰动的方式更不易被察觉，更具有实用性。

2.1.4 迷惑攻击

Lin 等人[40]提出了迷惑攻击，其目的是从某一时刻下的状态 s_t 开始施加扰动来迷惑智能体，从未观察智能体在 H 步后得到的状态 s_g。迷惑攻击需要知道目标智能体在每一步会选择的动作，以及生成式预测模型获得目标智能体此后可能选择的路径，在这两个前提下，攻击者制造对抗样本来迷惑智能体，使得智能体去往攻击者设定的预期状态 s_g。实验使用由 Carlini 和 Wagner[84]提出的对抗样本生成算法。结果证明，在没有随机动态变化的游戏场景下，40 步以内的迷惑攻击成功率能达到 70%。

这种使智能体做出攻击者所需动作的攻击方式，为面向强化学习系统的多样性攻击提供了新的思路。

2.1.5 基于值函数的对抗攻击

Kos 等人[41]提出了一种值函数指导的攻击方法，其主要思想是借助值函数模块评估当前状态价值的高低，以此来选择是否进行攻击。当值函数对当前状态价值做出的估计高于设定阈值，则对当前状态添加 FGSM 扰动，反之则不进行扰动，以此达到减少攻击成功所需要注入的对抗样本次数。实验证明，在这种攻击方式下，攻击者只需要在一小部分帧内注入扰动就可以达成目的，并且效果比在没有值函数引导下以相似频率注入扰动要更加好。

该方法与 Lin 等人[40]的战略时间攻击想法类似，都追求以更少的攻击次数来实现较好的攻击效果。这类攻击方法考虑到了强化学习场景下一些关键决策时间步对整体的影响，具有一定的指导意义。但是这种方法不能应用在一些单纯依靠策略梯度的场景。

2.1.6 嗅探攻击

Inkawhich 等人[42]提出了嗅探攻击方法，攻击者无法访问目标智能体的学习参数及其与之交互的环境，只能监测到目标智能体接收到的观测值，以及它反馈给环境的动作、奖励信号。基于该假设，给定四种威胁场景 S、SA、SR、SRA，分别对应于只监测状态信号、监测状态及动作信号、监测状态与奖励信号、同时监测三者。在这些场景中，攻击者训练并得到代理模型，以代理模型为基础制造对抗样本。

在一些策略部署在服务器端的场景下，相比于目前大部分需要访问目标智能体学习参数的攻击方法，嗅探攻击的可行性更高。

2.1.7 基于模仿学习的攻击

Behzadan 等人[43]提出使用模仿学习来提取目标模型进而使用对抗样本的迁移性对目标模型进行攻击。模仿学习是一种从专家决策样本中快速学习专家策略的技术。实验证明了对经模仿学习得到的策略有效的对抗样本，对于原目标模型依然适用。

这种攻击方式在思想上与策略诱导攻击方式十分类似，都是在等效模型的基础上使用对抗样本的迁移性进行攻击。不同的是该攻击使用模仿学习加快了等效模型建立的速度，为黑盒设置下对深度强化学习模型的攻击提供了新方案。

2.1.8 CopyCAT 算法

Hussenot 等人[44]提出了 CopyCAT 算法，这一算法可以引导目标智能体遵循攻击者设定的策略。不同于其他针对状态进行的攻击，CopyCAT 算法尝试攻击的是智能体从观测环境到生成状态这一感知过程。该算法的实施分为三个阶段：（1）收集目标智能体与环境交互的数据；（2）根据收集的数

据，采用优化算法为所有的观测感知过程生成掩码；（3）在目标智能体测试阶段，根据攻击者预先设定的策略为智能体添加掩码，更改目标智能体动作所遵循的策略。

该攻击方式并不是简单地为了降低目标智能体地性能表现，而是为了使智能体的行为能遵循攻击者所设定的策略，这种预先设计的策略既可以是使智能体性能恶化的策略，又可以是使智能体性能提升的策略。而且由于掩码是在攻击前预先计算得到的，因此这种攻击方式可以被视为一种实时攻击。相比与 FGSM 等需要在攻击过程中耗费计算资源的攻击方式，CopyCAT 更适合应用于对深度强化学习系统的攻击。

2.2 基于奖励的攻击

2.2.1 基于对抗变换网络的对抗攻击

Tretschk 等人[20]将新型的对抗攻击技术，即对抗变换网络整合到了策略网络结构中，通过一系列的攻击使得目标策略网络在训练时优化对抗性奖励而不再是优化原始奖励。对优化的奖励前后变化如下图2所示，其中绿色区域表示奖励为 1 的区域，暗红色区域表示奖励为 0 的区域。原始奖励 r^0 在球没有击中对手的垫子时给予奖励，对抗奖励在球击中对手垫子中心点时给予奖励。

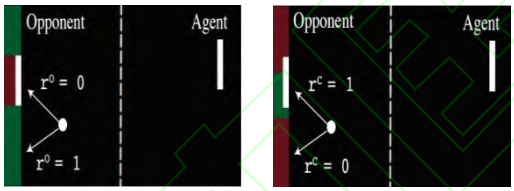


图 2 奖励可视化

Figure 2 Reward visualization

通过实验证明，对状态观测添加一系列的扰动，可以对目标策略网络施加任意的对抗奖励，使目标策略发生变化。这种攻击者存在的可能性令人们对持续学习型深度强化学习系统在工业领域中的应用而感到忧虑。

2.2.2 木马攻击

Kiourti 等人[45]首次提出了在深度强化学习系统的训练阶段使用木马攻击。他们只在 0.025%的训练数据中加入木马触发器，并在合理范围内对这些训练数据中对应的奖励值做出修改。如果目标智能体对这些中毒样本的状态做出了攻击者想要的动作，则给予该数据最大的奖励值；如果没做出攻击者想要的动作，则给予该数据最小的奖励值。

在这种木马攻击下，目标智能体在正常情况下的性能并没有受到任何影响，但是一旦木马触发器被触发，智能体就会执行攻击者预设的行为。

2.2.3 翻转奖励符号攻击

在深度强化学习系统训练过程中，训练样本以 (s,a,s',r) 的形式存放在经验回放池中，其中 s 为当前状态，a 为智能体在此状态下选择的动作，s' 为下一状态，r 为奖励值。在 Chen 等人[46]预设的攻击场景下，攻击者可以翻转经验回放池中 5%样本的奖励值符号，以此来最大化目标智能体的损失函数。

实验结果证明，尽管这种攻击方式可以在短时间内最大化智能体的损失函数，对其性能造成一定的影响，但是在长期训练后，智能体依然可以从中恢复过来。

这种攻击场景可以看做是奖励值信道错误的一种极端情况，例如传感器失灵或被人劫持，因此这种攻击具有一定的实际意义。

2.3 基于环境的攻击

2.3.1 路径脆弱点攻击

针对基于 DQN 的自动寻路系统，Bai 等人[47]提出一种在路径脆弱点上添加障碍物的攻击方法。他们首先利用 DQN 寻找一副地图的最优路径，在 DQN 的训练过程中，通过在路径上相邻点之间 Q 值的变化寻找路径脆弱点，之后借助相邻脆弱点之间连线的角度来辅助计算对抗样本点。最后通过在环境中加入对抗点减缓智能体找到最优路径的时间。

这种攻击方法需要对智能体规划路径上的点进行角度分析，所能应用到的场景受到较大的限制。而且实验最后证明，随着训练次数的增加，智能体依然可以收敛到最优路径。

2.3.2 通用优势对抗样本生成方法

在 A3C 路径查找任务中，智能体在寻路过程中只能获得周围的部分环境信息，因此无法通过在全局地图添加微小的扰动来达成攻击效果。因此，Chen 等人[19]针对基于 A3C 的路径查找任务提出了一种通用的优势对抗样本生成方法，使用这种方法可以为给定的任意地图生成优势对抗样本。这种方法的核心思想是，在智能体训练过程中找到值函数上升最快的梯度带，通过在梯度带上添加“挡板状”的障碍物来使目标智能体无法到达目的地或者在最大程度上延长到达

目的地所需要的时间。

这种攻击在不同规模的地图上进行测试，攻击成功率均在 91.91%以上，证明了这种攻击在不同地图上具有通用性。但是只针对基于 A3C 算法训练的智能体进行试验，尚不足以证明在深度强化学习算法之间的通用性。

2.3.3 对环境模型的攻击

环境动态模型的输入是当前状态及智能体动作，输出为下一状态。Xiao 等人[48]提出了两种对环境动态模型的攻击，希望通过在动态模型上添加扰动使得智能体达到攻击者指定的状态。他们提出了两种攻击方法：（1）随机动态模型搜索，通过随机使用一种动态模型，观察智能体是否会达到指定状态；（2）在现有的动态模型上添加扰动，通过确定性策略梯度的方式不断训练对抗动态模型，直到智能体能达到攻击者指定的状态。

2.4 动作空间扰动攻击

Yeow 等人[49]提出了两种对 DRL 算法动作空间的攻击：第一种方法是一个最小化具有解耦约束的深度强化学习智能体的累积奖励的优化问题，称为近视动作空间攻击；第二种方法和第一种攻击方法的目标相同，但具有时间耦合约束，称为具有前瞻性的动作空间攻击。结果表明，具有时间耦合性约束的攻击方法对深度强化学习智能体的性能具有更强的杀伤力，因为这个方法考虑到了智能体的动态因素。

由于动作空间独立于智能体策略之外，因此这种通过扰乱动作空间以减少智能体所获得的累积回报的方法几乎无法被防御。此类攻击适合应用于连续动作空间任务，但是在面对经过独热编码的离散动作空间任务时难度较大。

2.5 通过策略进行攻击

Gleave 等人[50]提出一种新的威胁算法，攻击者控制着对抗性智能体在同一环境与合法智能体进行对抗。在这种零和博弈场景下，敌人无法操纵合法智能体的观察，但可以在合法智能体遵循自身策略的情形下创建自然观察以作为对抗性输入。这种自然观察并没有包含在合法智能体的训练样本中，因此合法智能体在面对这些自然观察时会显得“手足无措”。

实验中，对抗性对手智能体基于 PPO 训练，受害者智能体基于 LSTM 和 MLP 训练。结果表明，敌人可以通过混淆受害者来赢得比赛，攻击效果如图 3 所示。图中第一行表示正常的对手与受害者的博弈过程，对手采用直接击打受害者的方式进行攻击，而第二行中的对抗性对手在与受害者博弈过程中，采取倒在地上的方式作为攻击手段。如果受害者躲过对手攻击，则受害者获胜，否则对手获胜。对抗性对手在无法保持站立的情况下依然能使受害者陷入一种混乱状态。实验证明，对抗性对手的胜率在 86%左右，而正常对手胜率仅为 47%。

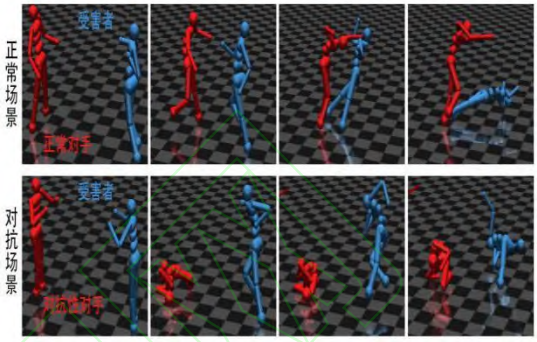


图 3 对抗智能体攻击效果

Figure 3 Adversarial agent attack

2.6 攻击的适用性分析

在本节中，针对基于不同深度强化算法的学习模型及攻击场景，对上述攻击方法的适用性进行分析。

（1）观测攻击：针对环境观测展开攻击的方法中，基于 FGSM[8]的强化学习攻击方法具有较强的攻击迁移能力，实验验证了其生成的对抗样本可以攻击不同的强化学习模型，均有较好的攻击效果。策略诱导攻击[39]、迷惑攻击[40]和基于模仿学习的攻击[43]均通过构建等价模型生成对抗样本，可用于攻击基于不同算法的强化学习黑盒模型。而战略时间攻击通过战略性地选择特定时间点进行攻击，适用于处理离散动作空间算法的学习模型，如 DQN[1,2]和 A3C[5]。基于值函数的对抗攻击[41]通过借助值函数模块评估当前状态价值的高低，从而决定是否进行攻击。因此这种方法不能应用在一些单纯依靠策略梯度的算法构建的学习模型中。嗅探攻击[42]和 CopyCAT 算法[44]分别通过训练不同智能体模型来生成对抗样本与使用掩码让智能体按照预先设定的策略行动来达到攻击的效果，可攻击不同强化学习算法得到的模型，具有一定的攻击迁移性。

（2）奖励攻击：基于对抗变换网络的

攻击[20]通过加入一个前馈的对抗变换网络获得对抗奖励，可实现对强化学习的白盒攻击。木马攻击[45]则在状态训练数据中加入木马触发器，并在合理范围内修改其对应的奖励值，该方法同时适用于不同的算法得到的不同模型。翻转奖励符号攻击[46]可以翻转经验回放池中部分样本的奖励值符号，所以适用于存在经验回放机制的强化学习模型。

（3）环境攻击：路径脆弱点攻击[47]和通用优势对抗样本生成方法[19]都是在自动导航系统上进行攻击，而前者需要对智能体规划路径上的点进行角度分析，所能应用到的场景受到较大的限制；后者则只针对基于A3C 算法训练的智能体进行实验，尚不足以证明在深度强化学习算法之间的通用性。对环境模型的攻击[48]方法是在环境的动态模型上增加扰动，可攻击基于环境动态建模的强化学习模型。

（4）动作攻击：动作空间扰动攻击[49]适合应用于连续动作空间任务，但是在面对经过独热编码的离散动作空间任务时难度较大。

（5）策略攻击：通过训练进行攻击[50]是指通过训练对抗性智能体与目标智能体进行对抗使目标智能体失败，目标智能体可以通过不同强化学习算法训练得到。

3 深度强化学习的防御方法

本节将详细介绍深度强化学习系统为应对各种不同的攻击方法而提出的防御方法，可分为三大类：对抗训练、鲁棒学习、对抗检测。表 4 对现有的主要防御方法做了归纳与比较。同时也对防御成功率进行统计，目前防御效果统计中，奖励值分析占主流，极少论文提到防御成功率，在调研过程中就发现一篇水印授权[65]的对抗检测防御方法给出了对抗样本检测成功率指标，但并没有给出具体数值，文中作者仅给出了检测成功率曲线图。

3.1 对抗训练

对抗训练是指将对抗样本加入到训练样本中对模型进行训练，其主要目的是提高策略对正常样本以外的泛化能力。但是对抗训练往往只能提高策略对参与训练的样本的拟合能力。面对训练样本之外的对抗样本，策略的性能表现依然不尽人意。

表 4 深度强化学习的防御方法

Table 4 Defense methods of deep reinforcement learning

分类	防御方法	防御机制	防御目标	攻击方法
对抗训练（见 3.1）	使用FGSM与随机噪声重训练[41,51]	对正常训练后的策略使用对抗样本与随机噪声进行重训练	状态扰动	FGSM、经值函数指导的对抗攻击（见 2.1）
	基于梯度带的对抗训练[19]	用单一的优势对抗样本进行对抗训练	环境扰动	通用优势对抗样本生成方法（见 2.3）
	非连续扰动下的对抗训练[52]	以一定的攻击概率在训练样本中加入对抗扰动	状态扰动	战略时间攻击、经值函数指导的对抗攻击（见 2.1）
	基于敌对指导探索的对抗训练[53]	根据对抗状态动作对的显著性调整对	状态扰动	战略时间攻击、嗅探攻击（见 2.1）
	基于代理奖励的鲁棒训练[54]	通过混淆矩阵得到代理奖励值以更新动作价值函数	奖励扰动	结合对抗变换网络的对抗攻击（见 2.2）
鲁棒学习（见 3.2）	鲁棒对抗强化学习[55]	在有对抗智能体的情境下利用博弈原理进行鲁棒训练	不同场景下的不稳定因素	在多智能体环境下的对抗策略（见 2.5）
	二人均衡博弈[56]	博弈、均衡原理	奖励扰动	结合对抗变换网络的对抗攻击（见 2.2）
	迭代动态博弈框架[57]	用迭代的极大极小动态博弈框架提供全局控制	状态扰动	FGSM、战略时间攻击、经值函数指导的对抗攻击、迷惑攻击（见 2.1）

	对抗 A3C[23]	在有对抗智能体的情境下进行博弈鲁棒训练	不同场景下的不稳定因素	在多智能体环境下的对抗策略（见 2.5）
	噪声网络[58]	使用参数空间噪声减弱对抗样本的迁移能力	状态扰动	FGSM、策略诱导攻击、利用模仿学习的攻击（见 2.1）
	方差层[59]	用权重遵循零均值分布，并且仅由其方差参数化的随机层进行训练	状态扰动	FGSM、战略时间攻击、经值函数指导的对抗攻击、迷惑攻击（见 2.1）
	基于元学习的对抗检测[60]	学习子策略以检测对抗扰动的存在	状态扰动	FGSM、战略时间攻击、经值函数指导的对抗攻击、迷惑攻击（见 2.1）
	基于预测模型的对抗检测[61]	通过比较预测帧与当前帧之间的动作分布来检测对抗扰动	状态扰动	FGSM、战略时间攻击、经值函数指导的对抗攻击、迷惑攻击（见 2.1）
对抗检测（见 3.3）	水印授权[65]	在策略中加入特有的水印以保证策略不被非法修改	策略篡改	CopyCAT 攻击、策略诱导攻击（见 2.1）
	受威胁的马尔科夫决策过程[67]	在马尔科夫决策过程中加入攻击者动作集并使用 K 级思维模式进行学习	奖励扰动	翻转奖励符号攻击（见 2.2）
	在线认证防御[68]	在输入扰动范围内选择最优动作	状态扰动	FGSM、战略时间攻击、经值函数指导的对抗攻击、迷惑攻击（见 2.1）

3.1.1 使用 FGSM 与随机噪声进行重训练

Kos 等人[41]使用对抗训练来提高深度强化学习系统的鲁棒性。他们先使用普通样本将智能体训练至专家水平，之后将 FGSM 扰动与随机噪声添加至智能体的观测状态值上进行重训练。Pattanaik 等人[51]也采用了这种方法来提高智能体的鲁棒性。

实验证明，经过 FGSM 对抗训练后，智能体在面对 FGSM 扰动时能保持与正常情况下相当的性能。但是这种方法只能防御 FGSM 与随机扰动，在面对其他对抗扰动时依然无能为力。

3.1.2 基于梯度带的对抗训练

Chen 等人[47]针对自己的优势对抗样本攻击方法提出了一种在自动寻路地图场景中基于梯度带的对抗训练方法。该对抗训练方法不同于传统的对抗训练，它只需要在一个优势对抗样本上训练即可免疫几乎所有对此地图的优势对抗攻击。

该实验在基于 A3C 的自动寻路任务下进行。实验结果证明，在一个优势对抗样本地图上进行基于梯度带的对抗训练后，智能体在面对其他优势对抗样本时防御精度能达到 93.89%以上，而且该方法训练所需要的时间远少于传统的对抗训练方法。

3.1.3 非连续扰动下的对抗训练

Behzadan 等人[52]提出了非连续扰动下的对抗训练机制。与传统对抗训练为所有训练样本添加扰动不同，该方法以一定的概率 P 在训练样本中添加 FGSM 扰动。

他们对 DQN 与噪声 DQN 模型进行了此非连续扰动的对抗训练。实验结果表明在 P 为 0.2 和 0.4 的情形下，DQN 与噪声 DQN 均能从扰动中恢复原有的性能。经过此方法重训练得到的智能体在面对测试阶段连续的 FGSM 扰动时，性能表现与正常情况相当。

3.1.4 基于敌对指导探索的对抗训练

Behzadan 等人[53]将 ε 贪婪探索与玻尔兹曼探索结合，提出了敌对指导探索机制。这种探索机制能根据敌对状态动作对的显著性来调整对每个状态抽样的概率。提高非连续对抗扰动对抗训练的样本利用率，同时也能使训练过程更加稳定。

这种方法是非连续扰动下对抗训练的改进，但是这种方法并没有拓展所能防御的攻击类型。

3.2 鲁棒学习

鲁棒学习是训练模型在面对来自训练阶段或者测试阶段时的攻击方法时提高其自身鲁棒性的学习机制。

3.2.1 基于代理奖励的鲁棒学习

由于在现实场景中，通常会因为传感器故障而导致奖励中带有噪声，因此 Wang 等人[54]提出使用奖励混淆矩阵来定义一系列的无偏代理奖励进行学习。使用该代理奖励进行训练能将模型从误导奖励中解救出来，并且训练的收敛速度比基准强化学习算法更快。

实验证明，使用代理奖励值训练得到的

智能体在奖励噪声场景下具有更好的表现。这种代理奖励具有很好的泛化性，可以轻易将其整合到各种强化学习算法中。

3.2.2 鲁棒对抗强化学习

Pinto 等人[55]将建模误差以及训练及测试场景下的差异都看作是系统中的额外干扰，基于这种思想，他们提出了鲁棒对抗强化学习，核心是令一个智能体以扮演系统中的干扰因素，在目标智能体的训练过程中施加压力。他们将策略的学习公式化为零和极大极小值目标函数，目标智能体在学习过程中一边以完成原任务为目标，一边使自己在面对对抗智能体的干扰时变得更加鲁棒。

在 MuJoCo 物理仿真环境中，Pinto 等人[55]证明经过该方法训练得到的智能体在面对额外干扰时具有更好的鲁棒性，考虑到了现实中可能存在的干扰，为深度强化学习系统从模拟环境走向现实环境提供了一份参考方案。

3.2.3 其余基于博弈理论的鲁棒训练

Bravo 等人[56]将受到攻击或损坏的奖励值信道问题建模了强化学习智能体与对手之间的零和博弈问题，并且提出了均衡原则，证明了在具有内部平衡的二人零和博弈情况下，无论观察结果受到的噪声水平如何，训练的时间平均值都将收敛至纳什均衡。

Ogunmolu 等人[57]将深度强化学习智能体与攻击者在训练阶段的对抗交互建模为迭代的最大最小动态博弈框架，通过控制训练过程来使两者达到鞍点均衡。这种方法提高了模型训练的策略在对抗干扰下的鲁棒性。

由于传统 A3C 在正常环境中训练的智能体无法处理一些具有挑战性的场景，因此 Gu 等人[23]提出了一种对抗 A3C 学习框架。与 Pinto 等人[55]类似，对抗 A3C 在学习过程中引入一个敌对智能体，以此模拟环境中可能存在的不稳定因素。目标智能体通过与该敌对智能体博弈训练，最终达到纳什均衡。

3.2.4 噪声网络

Behzadan 等人[58]对噪声网络的防御能力进行了测试。在实验中，他们使用等价模型方法建立了目标网络的副本，以副本为基础制造 FGSM 对抗扰动。

实验证明，在测试阶段，经过噪声 DQN 训练的智能体在面对此类黑盒攻击时，其性能表现要比原始 DQN 训练的智能体更加好；在训练阶段，噪声 DQN 智能体的性能也会随着攻击时间的增长而恶化，但是其恶化速度也比原始 DQN 慢。可以证明，使用噪声网络训练的智能体在面对对抗扰动时具有更好的弹性与鲁棒性。Neklyudov 等人[59]也使用了类似的高斯方差层来提高智能体的探索能力与鲁棒性。

3.3 对抗检测

对抗检测指模型对正常样本与对抗样本加以甄别，并在不修改原始模型参数的情况下处理对抗样本。

3.3.1 基于元学习的对抗检测

Havens 等人[60]介绍了一种元学习优势层次框架，它在只使用优势观察的情况下，能够有效地检测并减轻基于状态信息的对抗攻击。核心思想是使用主智能体监视子策略，通过衡量一定时间内子策略的回报来决定是否继续执行当前子策略。由于主智能体已经对子策略制定了准确的预期，因此一旦攻击者使策略行为发生变化，主智能体就能察觉并转换子策略。

这种学习框架能在时域范围内检测攻击者带来的预期之外的影响。相较于传统深度强化学习系统，提高了受攻击场景下的回报下界。

3.3.2 基于预测模型的对抗检测

Lin 等人[61]提出了一种动作条件帧预测模型，通过比较目标策略对预测帧与当前帧的动作分布差异来判断当前帧是否为对抗样本，如果当前帧被判断为对抗样本，则智能体使用预测帧作为输入并执行动作。实验效果如图 4 所示，该图描述了攻击者对智能体 π_θ 进行连续攻击的场景。在时间步 *t*-1 和 *t*,智能体接受恶意扰动输入 x_{t-1}^{adv} 与 x_{t-1}^{adv} ，并输出会导致性能下降动作分布。给定先前得观测和动作，并结合视觉预测模型得到预测帧 $\hat{x}_t$ ，并通过 $\pi_\theta(\hat{x}_t)$ 得到预测动作分布。比较 $\pi_\theta(x_t)$ 与 $\pi_\theta(\hat{x}_t)$ 两个动作分布，如果两个分布的距离 $D(\pi_\theta(\hat{x}_t), \pi_\theta(x_t))$ 大于阈值 H，则将当前帧视作对抗样本。

Lin 等人[61]将此方法与 Feature Squeezer[62]、AutoEncoder[63]以及 Dropout[64]三类对抗检测方法进行比较。实验结果证明，他们提出的方法能够以 60%到 100%的精度来检测对抗攻击，性能表现优于其他三类方法。

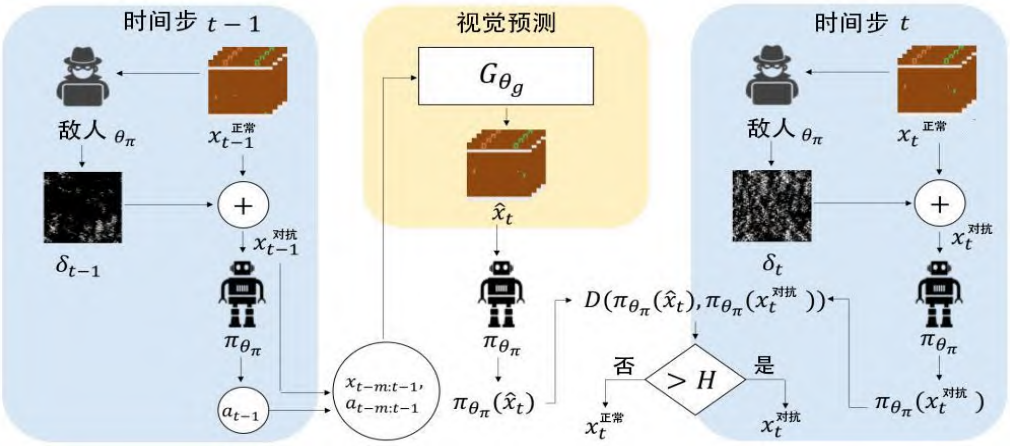


图 4 基于预测模型的对抗检测

Figure 4 Adversarial detection based on prediction model

3.3.3 水印授权

Behzadan 等人[65]将 Uchida 等人[66]提出的水印技术加以修改并应用到了深度强化学习系统中。其核心思想是为策略中对一些特定的状态转移序列加上唯一标识符。同时保证在正常情况下，标识符对策略的性能影响最小。一旦攻击者对策略进行篡改并除法水印，智能体就会中止活动。

3.3.4 受威胁的马尔科夫决策过程

Gallego 等人[67]提出了一种受威胁的马尔科夫决策过程，将攻击者对奖励值产生过程的干扰行为考虑在内。同时提出了一种 K 级思维方式来对这种新型马尔科夫决策过程求解。实验中，攻击者以 1 级思维利用正常的 Q 学习算法降低目标智能体对奖励的获取，目标智能体则以 2 级思维去估计攻击者的行为并尝试获得正向奖励。

实验结果证明，以 2 级思维模型训练的智能体在奖励值干扰下累积回报不断增加，最终实现正向的累积回报；而以传统方式训练的智能体性能不断恶化，最终收敛于最差的累积回报。

3.3.5 在线认证防御

Lutjens 等人[68]提出了一种在线认证的防御机制，智能体能在执行过程中保证状态动作值的下界，以保证在输入空间可能存在对抗扰动的情况下选择最优动作。防御过程中，智能体通过状态观测得到受扰动的状态 s_{adv}，DQN 网络输出状态动作价值 $Q(s_{adv}, a)$。在线认证节点在状态空间中鲁棒阈值 $\pm\varepsilon$，并为每个离散动作计算状态动作价值下限 Q_L，智能体根据最大的动作价值选择相对应的动作 a^*。

实验结果证明，将这种机制添加到 DQN 后，智能体在面对传感器噪声、带目标的 FGSM 扰动时能具有更好的鲁棒性。这种在线认证的防御方式易于集成，而且目前计算机视觉领域的鲁棒性验证工具可以更好地计算状态动作价值的置信下界。

4 深度强化学习的安全性分析

虽然目前已经有了许多对深度强化学习系统的攻防方法，但是攻击与防御方法的效果却很难进行评估。早期往往使用简单的标准对攻击效果进行评估，例如 Atari 游戏中得分的下降，但是这通常不足以表征攻击方法的效果。其次防御方法缺乏泛化性，对当前攻击有效的防御方法在面对其他类型的攻击时可能就失效了。此外，攻击和防御方法都在快速的更新迭代，许多传统的防御方法在面对新出现的攻击方法时都被证明是无效的。例如，在深度学习中，混淆梯度策略的提出，证明了许多防御措施是无效的[69]。由于防御方法泛化能力的不足，众多研究者转而着力研究策略的鲁棒性及策略的安全边界问题，以解决上述的不足。下面介绍模型安全性分析验证方面的一些研究。

4.1 基于等价模型的方法

由于 DNN 网络的复杂性，对学习到的策略网络的鲁棒性等属性进行直接验证是比较困难的。因此，比较直观的想法就是使用等价模型来等效替代策略网络。这种方法对等价模型的要求较高，至少需要满足以下两个条件：（1）等价模型的性能表现能与原来的策略在同一水平线上（或是稍弱一

些）；（2）要求等价模型能够很好地验证安全性、稳定性和鲁棒性等属性。除此之外，还需要考虑到扩展性以及算法复杂度等因素。下面对现有的等价模型方法进行介绍。

4.1.1 决策树等价模型

Bastani 等人[70]提出使用决策树策略来等价 DNN 策略。他们训练的决策树策略能够表示复杂的策略。由于决策树的非参数和高度结构化性质，使用现有的技术可以对其进行有效的验证。但是其中首要的难题就是决策树策略难以训练。对此，他们提出了 VIPER 方法，该方法在模仿学习算法的基础上利用了 Q 函数，将原来的 DNN 策略作为专家策略，最终学习到一颗较小的决策树（小于 1000 个结点），整个流程如图 5 所示。图 5 表明，该方法将强化学习模型建模为 MDP 过程，通过神经网络训练得到相应的策略并将其作为专家策略来训练生成决策树模型，最后将决策树学习生成的策略在该实验场景中验证其有效性。

实验表明，根据使用 DQN 与使用 VIPER 提取的决策树策略进行强化学习任务得到相同回报值的结果，表明学习得到的决策树在 Atari 的 Pong 和 cart-pole 场景下具有较好的表现。并且 Bastani 等人[70]描述了如何手动检查反例来验证决策树策略的正确性、稳定性和鲁棒性，他们表示与传统 DNN 策略相兼容的验证方法相比，决策树等价模型具有更大的扩展性。但是实验所证明的策略属性还不够全面，这是该方法需要在未来进行拓展的方向。

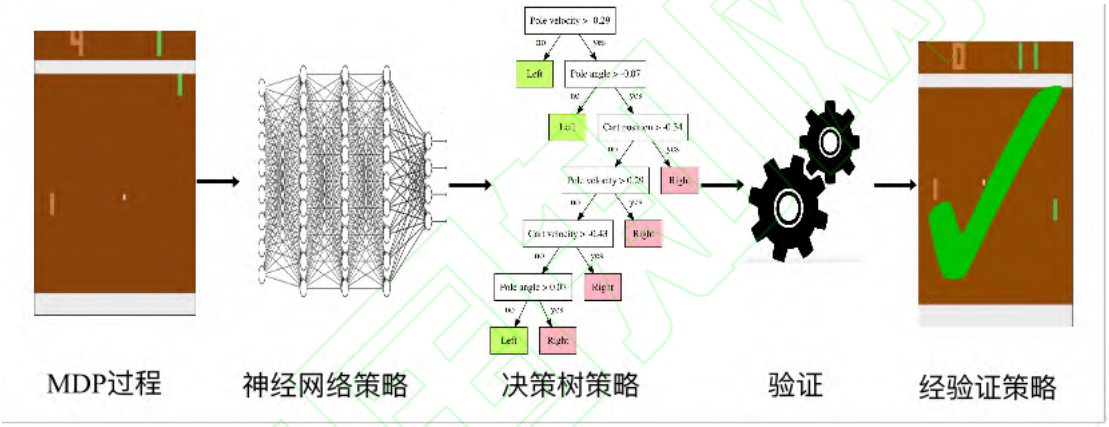


图 5 决策树等价模型验证方法流程

Figure 5 Process of decision tree equivalent model verification

4.1.2 形式化验证技术

Zhu 等人[71]考虑了如何将传统软件系统开发的形式化验证技术用于强化学习的验证问题。该技术不是通过检查和更改神经网络的结构来加强安全性，而是使用黑盒的方法拟合策略，继而得到一个更简单、解释性更强的合成程序。通过反例和句法引导的归纳综合过程来解决神经网络验证问题，并使用一个验证过程来保证程序提出的状态总是与原始规范和部署环境上下文的归纳不变量一致。这个不变量定义了一个归纳属性，该属性将转换系统中可表达的所有可达（安全）和不可达（不安全）状态分开。在此基础之上开发了一个运行监控框架，该框架将合成的程序视为安全盾牌，每当建议的操作可能会导致系统进入不安全区域时，该框架会覆盖此类操作。不安全区域需要根据相应的环境给出，这里根据时间的消耗、能够屏蔽的不安全状态的数量以及达到稳定状态所需要的步数来对合成的确定性程序进行评价。

以上两种方法都是模型本身出发，寻找策略网络的替代模型进行可验证的安全性分析，方法具有可行性。但是我们也需要考虑到在生成等价模型过程中造成的损失。此外可以根据替代模型的优势，在验证某一属性时，进行模型的选择。

4.2 其他方法

除了等价模型的方法外，众多研究者还提出了其他的一些方法。碰撞避免是安全性研究的一个重要方面，如何有效的减少碰撞的发生，是强化学习技术应用在自动驾驶汽车、机器人导航等领域时需要解决的问题。Lütjens 等人[67]在智能体运行过程中对输入状态给定一个范围计算 Q 值的安全下界，以在输入空间由于可能的对手或噪音而导致的最坏情况下，识别并选取最佳操作，并据

此提出了一种防御机制，所得到的策略（添加到训练好的 DQN 网络上）提高了对对手和传感器噪声的鲁棒性，通过调整鲁棒性范围计算碰撞次数的变化以及回报值的变化来衡量模型的性能以及鲁棒性范围的选取。这种方法是事先设定一个安全边界并进行实验验证，与从模型本身得出安全边界有所不同。

同样是在碰撞避免方面的研究，Behzadan 等人[72]提出了一种基于深度强化学习的新框架，用于在最坏情况下对碰撞避免机制的行为进行基准测试，即处理一个经过训练以使系统进入不安全状态的最优对手智能体。他们通过比较两种碰撞避免机制在应对故意碰撞尝试时的可靠性，验证了该框架的有效性。基于碰撞次数以及回报值进行评价，此外还对从开始到产生碰撞的时间进行了测量，时间越长表明这种机制有更强的防碰撞能力。

此外，为了以独立于攻击类型之外的方式评估智能体在测试阶段面对对抗扰动的鲁棒性与弹性，Behzadan 等人[73]提出了衡量深度强化学习策略的弹性与鲁棒性指标。首先定义对抗性后悔的概念，对抗性后悔是指未受干扰的主体在时间 T 获得的回报与受干扰的主体在时间 T 获得的回报的差值，那么弹性指的是造成最大对抗性后悔需要的最小的扰动状态数量，鲁棒性指的是给定最大扰动数量，可以达到的最大对抗性后悔。通过在 Cart-Pole 环境中训练的 DQN、A2C 和 PPO2 智能体上的实验评估，DQN 在较少数量的扰动状态数量下，引起了等量的对抗性后悔，表明其弹性较差，其次是 PPO2 策略，而 A2C 策略的弹性是三者中最强的。对于最大为 10 个扰动状态的情况下，三者的鲁棒性很接近，这是因为在弹性的计算中取得最大的对抗性后悔比较合适的扰动状态数为 7.5，超越这个数量，三者的效果都不是很好，对于固定的最大为 5 个扰动状态的情况下，DQN 的对抗后悔值最大，表明其鲁棒性最差，而 A2C 的对抗后悔值较小，表明鲁棒性最强。

尽管深度强化学习在实验室环境下取得了一个卓越的表现，在没有良好的安全性保证的情况下，深度强化学习在工业领域的落地应用还是有待考虑。

5 应用平台与安全性评估指标

在监督学习中，有如 ImageNet 数据集、LeNet 网络模型作为基准，方便比较学者们的研究成果。在深度强化学习领域与之对应的就是各式各样的环境、算法的实现。本节我们列举部分常用的环境、算法库和攻击方法库，给出了已有论文中在不同模型以及实验平台下的攻击防御安全性评估指标，攻防指标整理在表 6 和表 7 中。本节提供的实验平台算法是已有强化学习研究基础平台，也可作为之后研究的基准。

5.1 深度强化学习的环境基准

OpenAI Gym[74]提供了多种环境，比如 Atari、棋盘游戏等，并且它还提供了统一的环境接口，方便研究人员定制自己想要的环境。Malmo[75]是一个基于流行游戏 Minercraft 的人工智能实验平台，它提供了一系列具有连贯、复杂动态因素的 3D 环境以及丰富的目标任务。OpenSpiel[76]提供了从单智能体到多智能体的零和、合作等博弈场景以及一些分析学习动态和其他常见评估指标的工具。RLBench[77]旨在为机器人学习提供一系列具有挑战的学习环境，它具有 100 项完全独特的手工设计任务。MuJoCo[78]是一个物理模拟引擎，提供了一系列连续动作的模拟任务场景。目前常用的是 OpenAI Gym 游戏平台，已有的大部分实验成果都是在该平台的游戏场景中通过训练、攻击与防御等技术获得的。

5.2 深度强化学习的算法实现基准

OpenAI Baseline[79]提供了几种当下最流行的深度强化学习算法的实现，包括 DQN、TRPG、PPO 等。Rllab[80]提供了各种各样的连续控制任务以及针对连续控制任务的深度强化学习算法基准。Dopamine[81]是用于快速实现强化学习算法原型制作的研究框架，它旨在满足用户对小型、易处理代码库的需求。

5.3 深度强化学习的攻击基准

CleverHans[82]、Foolbox[83]都提供了制造对抗样本和对抗训练的标准化实现，可以用来量化和比较机器学习模型之间的鲁棒性。但是这两者只能用于对深度强化学习中的状态进行攻击，并不能涵盖奖励、动作等强化学习特有的环节。

5.4 深度强化学习的安全性评估基准

安全性评估指标通常用来评价攻击或者防御方法的强弱，以评估模型的鲁棒安全性。我们在表 5 中分别给出现有大部分论文中的攻击和防御的安全性评估指标，分析其评价机制和评价目的。

表 5 深度强化学习的安全性评估指标

Table 5 Security evaluation indicators of deep reinforcement learning

分类	指标	评价机制	评价目的
攻击指标	奖励	根据模型策略运行多个回合，计算累积回合奖励或者平均回合奖励	用于评估攻击方法对模型整体性能的影响
	损失	通过定义含有物理意义的概念来计算其是否到达不安全或者失败场景	用于评估攻击方法对模型策略的影响
	成功率	攻击方法在一定限制条件内可以达到成功攻击的次数比例	用于评估攻击方法的有效性
	精度	模型输出的对抗点中可以成功干扰路径规划的比例	用于评估攻击方法对模型策略的影响
防御指标	平均回报	根据模型策略运行多个回合，计算平均回合奖励	用于评估防御方法对提高模型性能的有效性
	成功率	检测攻击者篡改的策略动作	用于评估防御方法的有效性
	每回合步数	根据模型策略运行多个回合，记录每个回合的存活步数或者平均回合步数	用于评估防御方法对提高模型性能的有效性

表 6 深度强化学习的攻击指标

Table 6 Attack indicators of deep reinforcement learning

分类	攻击方法	攻击模型	平台	奖励	损失	成功率	精度
观测攻击	FGSM[18]	DQN[1]、TRPO[6]、A3C[5]	OpenAI Gym[74]	√			
	策略诱导攻击[39]	DQN[1]	Grid-world[38]	√		√	
	战略时间攻击[40]	DQN[1]、A3C[5]	OpenAI Gym[74]	√		√	
	迷惑攻击[40]	DQN[1]、A3C[5]	OpenAI Gym[74]	√		√	
	基于值函数的对抗攻击[41]	A3C[5]	OpenAI Gym[74]	√			
	嗅探攻击[42]	DQN[1]、PPO[37]	OpenAI Gym[74]	√			
	基于模仿学习的攻击[43]	DQN[1]、A2C[5]、PPO[37]	OpenAI Gym[74]	√			
奖励攻击	CopyCAT 算法[44]	DQN[1]	OpenAI Gym[74]	√		√	
	基于对抗变换网络的对抗攻击[20]	DQN[1]	OpenAI Gym[74]	√			
	木马攻击[45]	A2C[5]	OpenAI Gym[74]	√			
	翻转奖励符号攻击[46]	DDQN[3]	SDN environment[46]		√		
环境攻击	路径脆弱点攻击[47]	DQN[1]	OpenAI Gym[74]	√			√
	通用优势对抗样本生成方法[19]	A3C[5]	Grid-world[38]	√			√
	对环境模型的攻击[48]	DQN[1]、DDPG[36]	OpenAI Gym[74]	√			
动作攻击	动作空间扰动攻击[49]	PPO[37]、DDQN[3]	OpenAI Gym[74]	√			
策略攻击	通过策略进行攻击[50]	PPO[37]	OpenAI Gym[74]			√	

表 7 深度强化学习的防御指标

Table 7 Defense indicators of deep reinforcement learning

分类	防御方法	实验平台	平均回报	成功率	每回合步数
对抗训练	使用 FGSM 与随机噪声重训练[41,51]	OpenAI Gym[74]	√		
	基于梯度带的对抗训练[19]	Grid-world[38]	√		
	非连续扰动下的对抗训练[52]	OpenAI Gym[74]	√		

鲁棒学习	基于敌对指导探索的对抗训练[53]	OpenAI Gym[74]	√		
	基于代理奖励的鲁棒训练[54]	OpenAI Gym[74]	√		√
	鲁棒对抗强化学习[55]	OpenAI Gym[74]	√		
	二人均衡博弈[56]	Grid-world[74]	√		
	迭代动态博弈框架[57]	KUKA youbot[57]	√		
	对抗 A3C[23]	OpenAI Gym[74]	√		
	噪声网络[58]	OpenAI Gym[74]	√		
	方差层[59]	OpenAI Gym[74]	√		
对抗检测	基于元学习的对抗检测[60]	OpenAI Gym[74]	√		
	基于预测模型的对抗检测[61]	OpenAI Gym[74]	√		
	水印授权[65]	OpenAI Gym[74]	√	√	
	受威胁的马尔科夫决策过程[67]	Grid-world[38]	√		
	在线认证防御[68]	OpenAI Gym[74]	√		

6 未来研究方向

本文针对深度强化学习已提出的攻击方法以及为抵御这些攻击而提出的防御措施进行了全面调查。我们还提供了可用于实验的环境、算法以及攻击基准，同时对攻防指标进行整理总结。本节我们针对深度强化学习的攻防方法及安全性分析，探讨其在未来的研究发展方向，从不同角度分析之后可发展的研究内容。

6.1 攻击方法

已有的面向深度学习的攻击方法中，迭代攻击方法的性能相对较优，但是迭代方法计算代价太高，不能满足 DRL 系统实时预测的需求。针对 DRL 的攻击，未来可能从攻击的实时性要求出发，研究基于生成式对抗网络的对抗样本生成方法，经过训练后可生成大量高效的攻击；从攻击的实操角度出发，研究基于模仿学习构建替代模型的方式来缩短攻击准备的时间，以解决 DRL 系统的黑盒替代模型训练代价太大的问题；对于训练阶段进行的攻击，研究 DRL 训练过程的中毒攻击技术，通过在 DRL 系统中的状态、奖励值或是环境模型中嵌入后门触发器实现后门攻击；针对攻击的迁移性，研究攻击方法在不同算法或者不同模型结构上的迁移性，比较其攻击成功率；针对 DRL 的多智能体任务，研究多智能体的协同合作过程中存在的策略漏洞，从而进行策略攻击；从攻击的可解释性出发，研究不同的攻击方法对策略网络中神经元的激活状况的影响，寻找敏感神经元和神经通路来提高攻击的效果。

此外，与传统 DNN 模型类似，一些大型的如金融交易领域的 DRL 系统通常会被部署到云平台上。这些领域的环境模型与训练数据常常具有非常高的价值，攻击者未来可以尝试以访问云平台公用 API 的方式进行模型与训练数据的窃取。

6.2 防御方法

深度学习主要通过修改模型输入、目标函数以及网络结构这三类方法来实现防御效果。但是，深度学习的大多数防御方法不能满足 DRL 的实际应用场景中，尤其是在多智能体的任务场景中。针对 DRL 的防御，之后的研究可能从数据安全的角度出发，研究使用自编码器对受扰动的奖励、观测信号进行数据预处理，提高 DRL 系统面对信号噪声的鲁棒性；从模型鲁棒的角度出发，构建基于模型集成的强化学习环境动态建模方法，通过模型集合来提高模型鲁棒性，生成稳定有效的模型策略；从策略优化的角度出发，研究单个智能体甚至于多个智能体协同合作之间的策略漏洞，体现在模型策略网络的训练过程，以优化模型的策略。

6.3 安全性分析

DL 在攻防的分析上已经提出了许多指标，如对抗类别平均置信度、平均结构相似度、分类精确方差等。而对 DRL 的攻击与防御的实验结果主要还是以简单的平均回合奖励、奖励值的收敛曲线来进行评估。这样单一、表面的指标不能够充分说明 DRL 模型的鲁棒性，未来还需要提出更深层的评估标准，用以展现决策边界、环境模型在防御前后的不同。

目前在 DL 领域，已经有研究人员推出了一些模型测试评估平台，这些平台集成了目前对 DL 模型的攻击方法与防御方法，并以现有的模型安全指标对模型进行安全性分析。DRL 领域也可以结合本身的特点，搭

建相应的攻防安全分析平台，并添加 DRL 特有的测试需求，如对系统的环境建模误差进行分析、针对不同的系统生成标准的连续测试场景等。

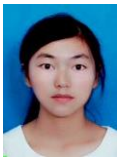
参考文献

1 Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.
2 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529-533.
3 Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning. In: Proceedings of the Thirtieth AAAI conference on artificial intelligence. 2016.
4 Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay. arXiv preprint arXiv:1511.05952, 2015.
5 Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the International Conference on Machine Learning. 2016: 1928-1937.
6 Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization. In: Proceedings of the International Conference on Machine Learning. 2015: 1889-1897.
7 Wu Y, Mansimov E, Grosse R B, et al. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. In: Proceedings of the Advances in Neural Information Processing Systems. 2017: 5279-5288.
8 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.
9 Berner C, Brockman G, Chan B, et al. Dota 2 with Large Scale Deep Reinforcement Learning. arXiv preprint arXiv:1912.06680, 2019.
10 Fayjie A R, Hossain S, Oualid D, et al. Driverless car: Autonomous driving using deep reinforcement learning in urban environment. In: Proceedings of the 2018 15th International Conference on Ubiquitous Robots (UR). IEEE, 2018: 896-901.
11 Prasad N, Cheng L F, Chivers C, et al. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. arXiv preprint arXiv:1704.06300, 2017.
12 Deng Y, Bao F, Kong Y, et al. Deep direct reinforcement learning for financial signal representation and trading. IEEE transactions on neural networks and learning systems, 2016, 28(3): 653-664.
13 Amarjyoti S. Deep reinforcement learning for robotic manipulation-the state of the art. arXiv preprint arXiv:1701.08878, 2017.
14 Nguyen T T, Reddi V J. Deep Reinforcement Learning for Cyber Security. *arXiv preprint arXiv:1906.05799*, 2019.
15 Oh J, Guo X, Lee H, et al. Action-conditional video prediction using deep networks in atari games. In: Proceedings of the Advances in Neural Information Processing Systems. 2015: 2863-2871.
16 Caicedo J C, Lazebnik S. Active object localization with deep reinforcement learning. In: Proceedings of the IEEE international conference on computer vision. 2015: 2488-2496.
17 Sutton R S, Barto A G. Reinforcement learning: An introduction. MIT press, 2018.47-48.
18 Goodfellow I, Huang S, Papernot N, , et al. Adversarial attacks on neural network policies. arXiv preprint arXiv:1702.02284, 2017.
19 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.
20 Tretschk E, Oh S J, Fritz M. Sequential attacks on agents for long-term adversarial goals. arXiv preprint arXiv:1805.12487, 2018.
21 Ferdowsi A, Challita U, Saad W, et al. Robust deep reinforcement learning for security and safety in autonomous vehicle systems. In: Proceedings of International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018: 307-312.
22 Behzadan V, Munir A. Whatever does not kill deep reinforcement learning, makes it stronger. arXiv preprint arXiv:1712.09344, 2017.
23 Gu Z, Jia Z, Choset H. Adversary A3C for Robust Reinforcement Learning. arXiv preprint arXiv:1912.00330, 2019.
24 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.
25 Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning. arXiv preprint arXiv:1511.06581, 2015.
26 Hausknecht M, Stone P. Deep recurrent q-learning for partially observable mdps. In: Proceedings of 2015 AAAI Fall Symposium Series. 2015.
27 Sorokin I, Seleznev A, Pavlov M, et al. Deep attention recurrent Q-network. arXiv preprint arXiv:1512.01693, 2015.
28 Plappert M, Houthooft R, Dhariwal P, et al. Parameter space noise for exploration. *arXiv preprint arXiv:1706.01905*, 2017.
29 Kapturowski S, Ostrovski G, Quan J, et al. Recurrent experience replay in distributed reinforcement learning[J]. 2018.
30 Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. ¨ Neural computation, 9(8):1735 – 1780, 1997.
31 Paine T L, Gulcehre C, Shahriari B, et al. Making Efficient Use of Demonstrations to Solve Hard Exploration Problems[J]. arXiv preprint arXiv:1909.01387, 2019.
32 Sutton R S, McAllester D A, Singh S P, et al. Policy gradient methods for reinforcement learning with function approximation. **In: Proceedings of** Advances in neural information processing systems. 2000: 1057-1063.
33 Graf T, Platzner M. Adaptive playouts in monte-carlo tree search with policy-gradient reinforcement learning[C]//Advances in Computer Games. Springer, Cham, 2015: 1-11.
34 Kakade, Sham M. "A natural policy gradient." Advances in neural information processing systems. 2002.
35 Konda V R, Tsitsiklis J N. Actor-critic algorithms[C]//Advances in neural information processing systems. 2000: 1008-1014.
36 Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.
37 Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
38 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
39 Behzadan V, Munir A. Vulnerability of deep reinforcement learning to policy induction attacks. **In: Proceedings of** International Conference on Machine Learning and Data Mining in Pattern Recognition. Springer, Cham, 2017: 262-275.
40 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.
41 Kos J, Song D. Delving into adversarial attacks on deep policies.*arXiv preprint arXiv:1705.06452*, 2017.
42 Inkawhich M, Chen Y, Li H. Snooping Attacks on Deep Reinforcement Learning. *arXiv preprint arXiv:1905.11832*, 2019.
43 Behzadan V, Hsu W. Adversarial exploitation of policy imitation. *arXiv preprint arXiv:1906.01121*, 2019.
44 Hussenot L, Geist M, Pietquin O. CopyCAT: Taking Control of Neural Policies with Constant Attacks. *arXiv preprint arXiv:1905.12282*, 2020.
45 Kiourti P, Wardega K, Jha S, et al. TrojDRL: Trojan Attacks on Deep Reinforcement Learning Agents. a*rXiv preprint arXiv:1903.06638*, 2019.
46 Han Y, Rubinstein B I P, Abraham T, et al. Reinforcement learning for autonomous defence in software-defined networking. **In: Proceedings of** International Conference on Decision and Game Theory for Security. Springer, Cham, 2018: 145-165.
47 Bai X, Niu W, Liu J, et al. Adversarial examples construction towards white-box Q table variation in DQN pathfinding training. **In: Proceedings of** 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC). IEEE,

2018: 781-787.
48 Xiao C, Pan X, He W, et al. Characterizing attacks on deep reinforcement learning. *arXiv preprint arXiv:1907.09470*, 2019.
49 Lee X Y, Ghadai S, Tan K L, et al. Spatiotemporally Constrained Action Space Attacks on Deep Reinforcement Learning Agents. *arXiv preprint arXiv:1909.02583*, 2019.
50 Gleave A, Dennis M, Kant N, et al. Adversarial policies: Attacking deep reinforcement learning. *arXiv preprint arXiv:1905.10615*, 2019.
51 Pattanaik A, Tang Z, Liu S, et al. Robust deep reinforcement learning with adversarial attacks. **In: Proceedings of** the 17th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2018: 2040-2042.
52 Behzadan V, Munir A. Whatever does not kill deep reinforcement learning, makes it stronger. *arXiv preprint arXiv:1712.09344*, 2017.
53 Behzadan V, Hsu W. Analysis and Improvement of Adversarial Training in DQN Agents With Adversarially-Guided Exploration (AGE). a*rXiv preprint arXiv:1906.01119, 2019*.
54 Wang J, Liu Y, Li B. Reinforcement learning with perturbed rewards. *arXiv preprint arXiv:1810.01032*, 2018.
55 Pinto L, Davidson J, Sukthankar R, et al. Robust adversarial reinforcement learning. **In: Proceedings of** the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 2817-2826.
56 Bravo M, Mertikopoulos P. On the robustness of learning in games with stochastically perturbed payoff observations. *Games and Economic Behavior*, 2017, **103**: 41-66.
57 Ogunmolu O, Gans N, Summers T. Minimax iterative dynamic game: Application to nonlinear robot control tasks. **In: Proceedings of** 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018: 6919-6925.
58 Behzadan V, Munir A. Mitigation of policy manipulation attacks on deep q-networks with parameter-space noise. **In: Proceedings of** the International Conference on Computer Safety, Reliability, and Security. Springer, Cham, 2018: 406-417.
59 Neklyudov K, Molchanov D, Ashukha A, et al. Variance networks: When expectation does not meet your expectations. *arXiv preprint arXiv:1803.03764*, 2018.
60 Havens A, Jiang Z, Sarkar S. Online robust policy learning in the presence of unknown adversaries. **In: Proceedings of** the Advances in Neural Information Processing Systems. 2018: 9916-9926.
61 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.
62 Xu, Weilin, David Evans, and Yanjun Qi. "Feature squeezing mitigates and detects carlini/wagner adversarial examples." *arXiv preprint arXiv:1705.10686*,2017.
63 Meng, Dongyu, and Hao Chen. Magnet: a two-pronged defense against adversarial examples. **In: Proceedings of** the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017.
64 Feinman R, Curtin R R, Shintre S, et al. Detecting adversarial samples from artifacts. arXiv preprint arXiv:1703.00410, 2017.
65 Behzadan V, Hsu W. Sequential Triggers for Watermarking of Deep Reinforcement Learning Policies. arXiv preprint arXiv:1906.01126, 2019.
66 Uchida Y, Nagai Y, Sakazawa S, et al. Embedding watermark s into deep neural networks[C]//Proceedings of the 2017 AC M on International Conference on Multimedia Retrieval. 201 7: 269-277.
67 Gallego V, Naveiro R, Insua D R. Reinforcement Learning under Threats. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 9939-9940.
68 Lütjens B, Everett M, How J P. Certified Adversarial Robustness for Deep Reinforcement Learning. *arXiv preprint arXiv:1910.12908*, 2019.
69 Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.
70 Bastani O, Pu Y, Solar-Lezama A. Verifiable reinforcement learning via policy extraction. **In: Proceedings of** the Advances in neural information processing systems. 2018: 2494-2504.
71 Zhu H, Xiong Z, Magill S, et al. An inductive synthesis framework for verifiable reinforcement learning. **In: Proceedings of** the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation. 2019: 686-701.
72 V. Behzadan and A. Munir. Adversarial reinforcement learning frame work for benchmarking collision avoidance mechanisms in autonomous vehicles. *arXiv preprint arXiv:1806.01368*, 2018.
73 V. Behzadan and W. Hsu. Rl-based method for benchmarking the adversarial resilience and robustness of deep reinforcement learning policies. *arXiv preprint arXiv:1906.01110*, 2019.
74 Brockman G, Cheung V, Pettersson L, et al. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.
75 Johnson M, Hofmann K, Hutton T, et al. The Malmo Platform for Artificial Intelligence Experimentation. **In: Proceedings of** IJCAI. 2016: 4246-4247.
76 Lanctot M, Lockhart E, Lespiau J B, et al. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.
77 James S, Ma Z, Arrojo D R, et al. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020, **5**(2): 3019-3026.
78 Todorov E, Erez T, Tassa Y. Mujoco: A physics engine for model-based control. **In: Proceedings of** 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012: 5026-5033.
79 Dhariwal P, Hesse C, Klimov O, et al. Openai baselines. 2017.
80 Duan Y, Chen X, Houthooft R, et al. Benchmarking deep reinforcement learning for continuous control. **In: Proceedings of** the International Conference on Machine Learning. 2016: 1329-1338.
81 Castro P S, Moitra S, Gelada C, et al. Dopamine: A research framework for deep reinforcement learning. *arXiv preprint arXiv:1812.06110*, 2018.
82 Papernot N, Faghri F, Carlini N, et al. Technical report on the cleverhans v2. 1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2016.
83 Rauber J, Brendel W, Bethge M. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.
84 Carlini N, Wagner D. Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples[J]. arXiv preprint arXiv:1711.08478, 2017.

陈晋音 浙江工业大学网络空间安全研究院副教授，博士生导师，2009 年获得浙江工业大学博士学位。主要从事人工智能安全、网络数据挖掘、智能计算、计算机视觉等方面的教学与科研工作。
E-mail：chenjinyin@zjut.edu.cn
(**Chen Jin-Yin** Associate professor and doctoral supervisor at the Institute of Cyberspace Security, Zhejiang University of Technology. She received her Ph.D. from Zhejiang University of Technology in 2009. She is mainly engaged in teaching and scientific research in artificial intelligence security, network data mining, intelligent computing, and computer vision.)

章 燕 浙江工业大学信息工程学院硕士研究生，主要研究方向为人工智能安全、计算机视觉。
E-mail：2111903240@zjut.edu.cn
(**Zhang Yan** Graduate student of the School of Information Engineering, Zhejiang University of Technology, and her main research methods are artificial intelligence security, computer vision.)



王雪柯 浙江工业大学信息工程学院硕士研究生，主要研究方向为人工智能安全、计算机视觉。
E-mail：17660478061@163.com
(**Wang Xue-Ke** Graduate student of the School of Information Engineering, Zhejiang University of Technology, and her main research methods are artificial intelligence security, computer

vision.)

蔡鸿斌 华东师范大学软件工程学院硕士研究生，主要研究方向为深度学习。
E-mail：hongbincai5330@163.com
(**Cai Hong-Bin** Grgraduate graduate of the School of Software Engineering, East China Normal University, and his main research direction is deep learning.)

王 珏 浙江工业大学信息工程学院硕士研究生，主要研究方向为人工智能安全、计算机视觉。
E-mail：211190321@zjut.edu.cn
(**Wang Jue** Graduate student of the School of Information Engineering, Zhejiang University of Technology, and his main research methods are artificial intelligence security, computer vision.)

纪守领 获美国佐治亚理工学院电子与计算机工程博士学位、佐治亚州立大学计算机科学博士学位，现任浙江大学“百人计划”研究员、博士生导师。目前的研究兴趣包括数据驱动的安全性和隐私性，人工智能安全性和大数据分析。
E-mail：sji@zju.edu.cn
(**Ji Shou-Ling** Received a doctorate in electrical and computer engineering from Georgia Institute of Technology, and a doctorate in computer science from Georgia State University. He is currently a researcher and doctoral supervisor of the "Hundred Talents Program" of Zhejiang University. Current research interests include data-driven security and privacy, artificial intelligence security and big data analysis.