深度强化学习模型攻击与防御方法综述

姓名：Allenpandas，学号：xxxxx，专业：xxxxx

***Abstract*—深度强化学习是人工智能领域新兴技术之一，它将深度神经网络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，使模型具有较强的学习能力。然而，近期研究结果表明深度强化学习存在安全漏洞，极易受到对抗样本攻击，大大降低了模型安全性。**

本文结合童恩栋老师讲授的《强化学习前沿技术》，回顾并归纳了深度强化的对抗攻击方法以及防御策略，通过课程的学习以及课程报告的总结，为日后进一步研究强化学习的防御方法、对强化模型方法进行安全性分析奠定扎实的理论基础。

***Index Terms*—深度强化学习，特征提取，对抗样本攻击，安全性分析.**

引言

近年来，深度强化学习（Deep reinforcement learning, DRL）方法作为人工智能领域新兴技术之一，被广泛应用于游戏博弈 [1] 、自动驾驶 [2] 、医疗健康 [3] 、金融交易 [4] 、机器人控制 [5] 、网络安全 [6] 、计算机视觉 [7] 等领域。各种深度强化学习算法不断改进优化，提出了基于双重深度 Q 网络的强化学习算法 DDQN [8] 、基于优先经验回放 Q 网络的强化学习算法 PrioritizedDQN、信任域策略优化算法 TRPO [9] 以及 K 因子信任域行动者评论者算法 ACKTR [10] 等。

然而，近年来多项研究结果表明深度强化学习存在安全漏洞，极易受到对抗样本攻击，大大降低了模型安全性。Huang 等 [11] 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 [12] 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；Ferdowsi 等 [13] 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。由此可见，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。

本文针对现有的深度强化学习算法、深度强化学习攻击方法进行梳理，同时对深度强化学习防御方法以及安全性分析提出自己的一些简介。本文后续章节安排如下：第 1 节将简要介绍深度强化学习；第 2 节梳理现有的强化学习攻击算法、第 3 节对深度强化学习的防御方法进行总结，最后提出一些自己的若干思考。

I. 深度强化学习

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 [14] 。强化学习问题通常可以被建模为马尔可夫决策过程（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。从任意时间步长 t 时刻开始，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，同时智能体得到奖励 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，实现了从原始输入到决策结果输出的端到端模型。

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算法和基于策略梯度的深度强化学习算法。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。

II. 深度强化学习的攻击方法

为了系统分析深度强化学习各种不同的攻击方法, 本文根据强化学习 MDP 中的关键元素（状态、奖励、动作、环境、策略）对攻击方法进行归类, 即: 基于观测攻击、奖励攻击、动作攻击、环境攻击以及策略攻击。

A. 基于观测的攻击

Huang 等 [11] 最先对通过深度强化学习的观测进行攻击，使用机器学习领域常用的快速梯度符号（Fast gradient sign method, FGSM）[15] 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:

$$\eta = \varepsilon \operatorname{sign}\left(\nabla_x J(\theta, x, y)\right) \tag{1}$$

其中，J 表示损失函数，θ 表示模型参数，x 表示模型输入，y 样本的最优动作项，$\nabla_x J(\cdot, \cdot, \cdot)$ 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。

Lin 等 [16] 考虑到强化学习奖励的稀疏性问题，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，减少目标智能体的预期累积回报。利用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 [11] 相同的效果。由于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，因此更具有实用性。

Kos 等 [17] 提出了一种值函数指导的攻击方法，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 [16] 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。

B. 基于奖励的攻击

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 Han 等 [18] 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, 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。

C. 基于环境的攻击

针对基于深度 Q 网络（Deep Q-network, DQN）的自动寻路系统，Bai 等 [19] 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。

D. 基于动作的攻击

Lee 等 [20] 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。实验结果表明, 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。

E. 基于策略的攻击

Gleave 等 [21] 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。

III. 深度强化学习的防御方法

针对以上提到的 5 种攻击类型，目前深度强化学习的防御方法可以主要分为两类：对抗训练和鲁棒学习。

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。Kos 等 [17] 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，进而提高智能体的鲁棒性。Bai 等 [19] 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。

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。Pinto 等 [22] 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，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，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。

IV. 关于深度强化学习攻击与防御的若干思考

本文针对深度强化学习的攻击方法与抵御这些攻击而提出的防御措施进行了总结，针对深度强化学习的攻防方法及安全性分析有了以下的若干思考：

A. 如何评估量化攻击？

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 游戏中得分下降说明攻击成功，但是这通常不足以表征攻击方法的效果，更无法将多种对抗样本攻击的方法进行量化比较。

B. 如何验证防御效果及泛化性？

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。如何提高模型防御的泛化性，针对多样攻击是未来值得关注的一个方向。

C. 模型验证的评估标准？

深度学习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，而对深度强化学习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深度强化学习模型的鲁棒性，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。

V. 致谢

感谢《网络空间安全前沿》课程的负责人和本学期的全体授课教师。通过这门课，让我对于网络空间安全专业的前沿知识有了更深刻的认识，尤其是当下人工智能领域的新兴技术与网络空间安全专业交叉结合的热点知识，包括但不限于：利用人工智能技术解决传统的安全问题，新型的人工智能技术的攻击、防护与安全性分析等。希望在未来的科研工作中，能够学习更多的相关知识，更好的服务社会。

References

[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, “Mastering the game of go with deep neural networks and tree search,” *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[2] A. R. Fayjie, S. Hossain, D. Oualid, and D.-J. Lee, “Driverless car: Autonomous driving using deep reinforcement learning in urban environment,” in *2018 15th international conference on ubiquitous robots (ur)*. IEEE, 2018, pp. 896–901.

[3] N. Prasad, L.-F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt, "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units," *arXiv preprint arXiv:1704.06300*, 2017.

[4] Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, "Deep direct reinforcement learning for financial signal representation and trading," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 3, pp. 653–664, 2016.

[5] S. Amarjyoti, "Deep reinforcement learning for robotic manipulation-the state of the art," *arXiv preprint arXiv:1701.08878*, 2017.

[6] T. T. Nguyen and V. J. Reddi, "Deep reinforcement learning for cyber security," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[7] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-conditional video prediction using deep networks in atari games," *Advances in neural information processing systems*, vol. 28, 2015.

[8] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[9] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*. PMLR, 2015, pp. 1889–1897.

[10] Y. Wu, E. Mansimov, R. B. Grosse, S. Liao, and J. Ba, "Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation," *Advances in neural information processing systems*, vol. 30, 2017.

[11] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," *arXiv preprint arXiv:1702.02284*, 2017.

[12] T. Chen, W. Niu, Y. Xiang, X. Bai, J. Liu, Z. Han, and G. Li, "Gradient band-based adversarial training for generalized attack immunity of a3c path finding," *arXiv preprint arXiv:1807.06752*, 2018.

[13] A. Ferdowsi, U. Challita, W. Saad, and N. B. Mandayam, "Robust deep reinforcement learning for security and safety in autonomous vehicle systems," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 307–312.

[14] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

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

[16] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, "Tactics of adversarial attack on deep reinforcement learning agents," *arXiv preprint arXiv:1703.06748*, 2017.

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

[18] P. Kiourti, K. Wardega, S. Jha, and W. C. T. Li, "Trojan attacks on deep reinforcement learning agents," *arXiv preprint arXiv:1903.06638*, 2019.

[19] X. Bai, W. Niu, J. Liu, X. Gao, Y. Xiang, and J. Liu, "Adversarial examples construction towards white-box q table variation in dqn pathfinding training," in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*. IEEE, 2018, pp. 781–787.

[20] X. Y. Lee, S. Ghadai, K. L. Tan, C. Hegde, and S. Sarkar, "Spatiotemporally constrained action space attacks on deep reinforcement learning agents," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4577–4584.

[21] A. Gleave, M. Dennis, C. Wild, N. Kant, S. Levine, and S. Russell, "Adversarial policies: Attacking deep reinforcement learning," *arXiv preprint arXiv:1905.10615*, 2019.

[22] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2817–2826.