Transferable Environment Poisoning: Training-time Attack on Reinforcement Learner with Limited Prior Knowledge

Doctoral Consortium

Hang Xu
Advisor: Zinovi Rabinovich
Nanyang Technological University, Singapore
hang017@e.ntu.edu.sg

ABSTRACT

As reinforcement learning (RL) systems are deployed in various safety-critical applications, it is imperative to understand how vulnerable they are to adversarial attacks. Of these, an environment-poisoning attack is considered particularly insidious, since environment hyper-parameters are significant in determining an RL policy yet prone to be accessed by third parties. In this work, we study an environment-poisoning attack (EPA) against RL at training time. Considering that environment alteration comes at a cost, we seek minimal poisoning in an unknown environment and aim to force a black-box RL agent to learn an attacker-designed policy.

KEYWORDS

Reinforcement Learning; Security; Environment Poisoning

ACM Reference Format:

Hang Xu. 2022. Transferable Environment Poisoning: Training-time Attack on Reinforcement Learner with Limited Prior Knowledge: Doctoral Consortium. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

The security of Reinforcement Learning (RL) has become increasingly significant due to the widespread adoption of RL systems in safety-critical applications, such as autonomous cars [6, 12, 18], smart energy systems [5, 7, 19] and healthcare systems [2, 3, 17]. However, RL policies are typically sensitive to training hyper-parameters [4, 8, 11], where a slight variation of these parameters may cause obvious performance difference. As a result, RL policies are vulnerable to being perturbed by poisoned training hyper-parameters. Among these hyper-parameters, environment hyper-parameters are most susceptible as they can be easily accessed by third parties, which are also termed *causal factors* in physical systems (e.g., gravity and friction) [10, 13, 20]. Therefore, to facilitate the formulation of secure strategies, a study of the threats posed by environment hyper-parameters is necessary.

The success of existing training-time attacks [9, 16, 22] relies on comprehensive prior knowledge of the attacked RL system, including RL agent's learning mechanism (i.e., learning algorithm and policy model) and/or its environment model (i.e., transition dynamics and reward functions). Unfortunately, assuming such an omniscient attacker makes most attack approaches somewhat

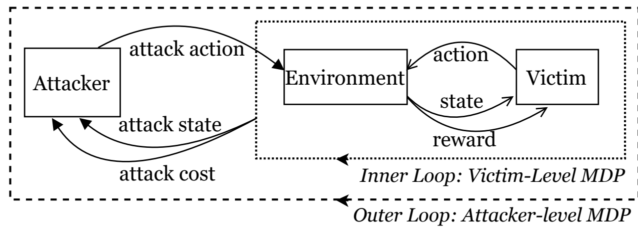

Figure 1: Attack Framework

unrealistic so that their threats to real-world RL-based applications is limited. To alleviate such a limitation, it is imperative to study a novel training-time attack that requires minimal prior knowledge of the RL system.

In this work, we propose a transferable environment-dynamics poisoning attack (TEPA) against RL at training time, assuming only the ability to alter the environment hyper-parameters. We design an attack framework and an optimization objective to seek minimal, adaptive environment poisoning that forces an RL agent to learn an attacker-desired policy. We further demonstrate Transferability property of TEPA and exploit the transferable strategy to poison various RL agents regardless of the types of their learning algorithms. Finally, we empirically show the security threat posed by our TEPA to both tabular-RL and deep-RL algorithms in discrete and continuous environments.

2 PROBLEM STATEMENT

Attack Framework. We adopt a bi-level Markov Decision Process (MDP) architecture [21, 22] illustrated in Figure 1. The task of poisoning a victim RL agent's policy is performed by another RL agent (i.e., the attacker) which operates on a different timescale from the victim. Specifically, with a particular attack frequency, the attacker manipulates the victim's training-environment hyper-parameters in response to the victim's learning progress.

Attack Objective. The attacker's goal is to learn a strategy e that induces the victim to learn an attacker-desired policy with minimized changes to the victim's training environment. Specifically, the attack objective is to minimize the deviation between the victim's policy and the attacker-desired one and, at the same time, minimize the deviation between the poisoned environment and the natural one. Therefore, the attack optimization objective is to minimize the cumulative attack costs, which is denoted as

$$\min_{\sigma} \sum_{i=1}^{\infty} \gamma^i c_i \quad s.t. \quad c_i := \Delta(P_i(s', a'|s, a) || P^*(s', a'|s, a)) \qquad (1)$$

where c_i represents attack cost at the attack epoch i.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

Here, $P_i(s', a'|s, a) = T_{e_i}(s'|s, a)\pi_i(a'|s')$ is a stochastic process [14] over victim's state-action pairs at the i^{th} attack epoch, where the victim follows the policy $\pi_i(a|s)$ in the environment e_i which has been modified by a sequence of attacker's manipulation. Similarly, $P^*(s', a'|s, a)$ represents an ideal stochastic process, where the victim adopts the attacker-desired policy π^* in the natural environment e_0. Thus, $\Delta(P_i||P^*)$ describes the attack cost by capturing the deviation jointly caused by the victim's actual policy $\pi_i(a'|s')$ and its poisoned environment dynamics $T_{e_i}(s'|s, a)$.

3 ATTACK APPROACHES

In this section, we introduce the learning of an attack strategy in both white-box and double-black-box settings, and then we describe the Transferability of TEPA strategy.

White-box Settings. With the prior knowledge of the victim's learning mechanism and its environment dynamics, we measure the attack cost $c_i = D_{KLR}(P_i||P^*)$ using Kullback-Leibler Divergence Rate and compute it following [15]. Thereby the attack strategy can be learned by solving the optimization problem as Equation 1.

Double-Black-Box Settings. To achieve policy compulsion on a *black-box* RL agent in a *black-box* training environment, we first investigate how to infer the internal information of an unknown RL system, and then we learn an adaptive attack strategy based on our proposed approximation of the attack objective.

As shown in Figure 2, given observations of the victim's trajectories τ during its learning process, we jointly train: a) an Encoder-Dual-Decoder network that learns a low-dimensional latent representation z of the victim RL system's internal information; b) an attack strategy σ, conditioned on the latent representation and environment hyper-parameters, that manipulates the victim's policy using minimal environment poisoning.

Based on the inferred representations, we approximate the attack cost c_i as the distance between z and z^* in the latent space, and we measure it using Cosine Similarity [1], denoting it as $\Delta(P_i||P^*) := \Delta(z_i||z^*) = 1 - \frac{z_i \cdot z^*}{\|z_i\|\|z^*\|}$.

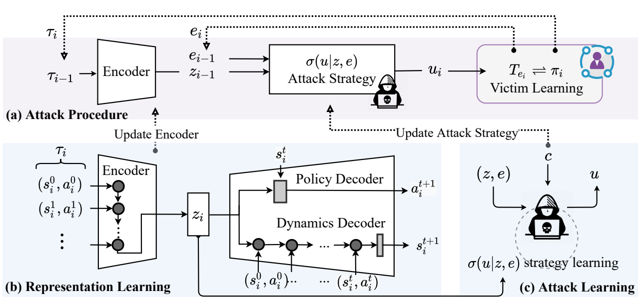

Figure 2: Illustration of double-black-box environment-poisoning attacks: (a) shows the attack procedure; (b) and (c) describe the latent representation learning and attack strategy learning, respectively. The solid line denotes data transfer and the dotted line represents data update.

Since z only captures the environment-dynamics features that have influenced the agent's trajectories, $\Delta(z||z^*)$ cannot measure aggregate changes across the entire environment. Therefore, we additionally measure the aggregate environment changes $\Delta(e, e_0)$ using the normalized euclidean distance between the poisoned hyper-parameter e and the natural one e_0, denoting it as $\frac{\|e_i - e_0\|_2}{\|e_{limit} - e_0\|_2}$ where e_{limit} is the boundary values of environment hyper-parameters.

In summary, the approximation of the attack cost c_i is a combination of $\Delta(z_i||z^*)$ and $\Delta(e_i, e_0)$, denoted as $c_i := (1-\omega) \times \Delta(z_i, z^*) + \omega \times \Delta(e_i, e_0)$ where $\omega \in [0, 1]$ is the weight parameter.

Transferability Property of TEPA. We found *Transferbility* property of TEPA. The attack strategy, which is learned based on a proxy agent, can be transferred to poison other victim agents' policies in the same tasks, even if these victims utilize different algorithms/models. Therefore, *Transferability* allows our attack strategy to be generally effective to various RL agents regardless of their learning algorithms and policy models.

4 EXPERIMENT RESULTS

We evaluate TEPA in 3D grid world where the cell elevation is considered as the environment hyper-parameter which can be manipulated by the attacker. As shown in Figure 3, our TEPA succeeds in poisoning the tabular-RL agent's navigation policy in both white-box and double-black-box settings. We further empirically demonstrate *Transferability* of TEPA strategy as Figure 4. Additionally, we evaluate TEPA against a deep-RL agent in a control task in continuous environments, showing TEPA's feasibility and scalability in terms of the complexity of victim RL systems.

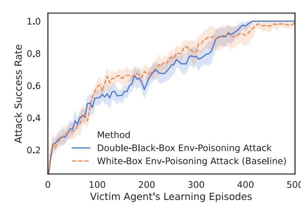

Figure 3: Attack Performance

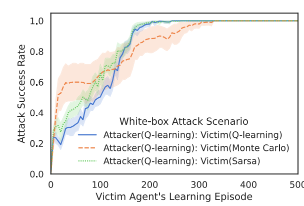

Figure 4: Transferability

5 CONCLUSION & FUTURE WORK

We have proposed a transferable environment-poisoning attack (TEPA) with limited prior knowledge of the vicitm RL system. We have empirically evaluate TEPA against both tabular-RL and deep-RL agents in discrete and continuous environments. Experimental results show that our attack successfully forces an RL agent to learn an attacker-desired policy via minimal changes on its training environment.

Currently, we assume that the attacked RL agent is oblivious to the attack and continues to operate normally throughout the sequence of environment modifications. In the future work, we will discuss the connection between TEPA and existing robust RL methods which consider environment perturbations during its learning process. We will further strengthen TEPA to show its potential impact in real-world RL-based applications. Another significant component for future work is the defence formulation. We aim to develop TEPA as a test-bed core for analyzing RL vulnerabilities to a poisoned environment, and furthermore we will study secure strategies which can prevent an RL agent's policy from being manipulated by poisoned training environments.

REFERENCES

[1] Kenneth Ward Church. 2017. Word2Vec. *Natural Language Engineering* 23, 1 (2017), 155–162.

[2] Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. 2020. Reinforcement Learning for Intelligent Healthcare Applications: A Survey. *Artificial Intelligence in Medicine* 109 (2020), 101964–101964.

[3] Niloufar Eghbali, Tuka Alhanai, and Mohammad M Ghassemi. 2021. Patient-Specific Sedation Management via Deep Reinforcement Learning. *Frontiers in Digital Health* 3 (2021), 17.

[4] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep Reinforcement Learning That Matters. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence*. AAAI, Louisiana, USA, 3207–3214.

[5] Sunyong Kim and Hyuk Lim. 2018. Reinforcement Learning based Energy Management Algorithm for Smart Energy Buildings. *Energies* 11, 8 (2018), 2010.

[6] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. 2021. Deep Reinforcement Learning for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 1, 1 (2021), 1–18.

[7] Sangyoon Lee and Dae-Hyun Choi. 2022. Federated Reinforcement Learning for Energy Management of Multiple Smart Homes with Distributed Energy Resources. *IEEE Transactions on Industrial Informatics* 18, 1 (2022), 488–497.

[8] Roman Liessner, Jakob Schmitt, Ansgar Dietermann, and Bernard Bäker. 2019. Hyperparameter Optimization for Deep Reinforcement Learning in Vehicle Energy Management.. In *ICAART (2)*. SCITEPRESS, Prague, Czech Republic, 134–144.

[9] Yuzhe Ma, Xuezhou Zhang, Wen Sun, and Jerry Zhu. 2019. Policy Poisoning in Batch Reinforcement Learning and Control. In *Proceedings of the 33th Conference on Neural Information Processing Systems*. ACM, Vancouver, Canada, 14570–14580.

[10] Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. 2018. Learning to Adapt in Dynamic, Real-world Environments through Meta-Reinforcement Learning. In *Proceedings of 6th International Conference on Learning Representations*. ICLR, Vancouver, Canada, 1–17.

[11] Xinlei Pan, Weiyao Wang, Xiaoshuai Zhang, Bo Li, Jinfeng Yi, and Dawn Song. 2019. How You Act Tells a Lot: Privacy-leaking Attack on Deep Reinforcement learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. IFAAMS, Auckland, New Zealand, 368–376.

[12] Xinlei Pan, Yurong You, Ziyan Wang, and Cewu Lu. 2017. Virtual to Real Reinforcement Learning for Autonomous Driving. In *Proceedings of 28th British Machine Vision Conference*. BMVA, London, British, 1–13.

[13] Christian Perez, Felipe Petroski Such, and Theofanis Karaletsos. 2020. Generalized Hidden Parameter MDPs: Transferable Model-based RL in a Handful of Trials. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI, New York, USA, 5403–5411.

[14] Zinovi Rabinovich, Lachlan Dufton, Kate Larson, and Nick Jennings. 2010. Cultivating Desired Behaviour: Policy Teaching via Environment-dynamics Tweaks. In *Proceedings of the 10th International Conference on Autonomous Agents and MultiAgent Systems*. IFAAMS, Toronto, Canada, 1097–1104.

[15] Ziad Rached, Fady Alajaji, and L Lorne Campbell. 2004. The Kullback-Leibler Divergence Rate between Markov Sources. *IEEE Transactions on Information Theory* 50, 5 (2004), 917–921.

[16] Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. 2020. Policy Teaching via Environment Poisoning: Training-time Adversarial Attacks against Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Vienna, Austria, 7974–7984.

[17] Elsa Riachi, Muhammad Mamdani, Michael Fralick, and Frank Rudzicz. 2021. Challenges for Reinforcement Learning in Healthcare. arXiv:2103.05612

[18] Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. 2017. Deep Reinforcement Learning Framework for Autonomous Driving. *Electronic Imaging* 2017, 19 (2017), 70–76.

[19] Tomah Sogabe, Dinesh Bahadur Malla, Shota Takayama, Seiichi Shin, Katsuyoshi Sakamoto, Koichi Yamaguchi, Thakur Praveen Singh, Masaru Sogabe, Tomohiro Hirata, and Yoshitaka Okada. 2018. Smart Grid Optimization by Deep Reinforcement Learning over Discrete and Continuous Action Space. In *Proceedings of the 7th World Conference on Photovoltaic Energy Conversion*. IEEE, Hawaii, USA, 3794–3796.

[20] Sumedh A Sontakke, Arash Mehrjou, Laurent Itti, and Bernhard Schölkopf. 2021. Causal Curiosity: RL Agents Discovering Self-Supervised Experiments for Causal Representation Learning. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Online, 9848–9858.

[21] Hang Xu, Rundong Wang, Lev Raizman, and Zinovi Rabinovich. 2021. Transferable Environment Poisoning: Training-time Attack on Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. IFAAMAS, Online, 1398–1406.

[22] Xuezhou Zhang, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. 2020. Adaptive Reward-Poisoning Attacks against Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Vienna, Austria, 11225–11234.