Adversarial Robust Deep Reinforcement Learning is Neither Robust Nor Safe

Ezgi Korkmaz

Abstract

The policies trained with deep reinforcement learning are being deployed in many different settings from automated language assistants to biomedical applications. Yet concerns have been raised regarding robustness and safety of deep reinforcement learning policies. To target these problems several works focused on proposing adversarial training methods for deep reinforcement learning and claimed adversarial training achieves safe and robust deep reinforcement learning policies. In this paper, we demonstrate that adversarial deep reinforcement learning is neither safe nor is it robust. While robust deep reinforcement learning policies can be attacked via black-box adversarial perturbations, our results further demonstrate that standard reinforcement learning policies are more robust compared to robust deep reinforcement learning under natural attacks. Furthermore, this paper highlights that robust deep reinforcement learning policies cannot generalize even in the same level with standard reinforcement learning.

1 Introduction

The performance of reinforcement learning algorithms has been boosted with the utilization of deep neural networks as function approximators (Mnih et al., 2015; Wang et al., 2016). Currently, it is possible to learn deep reinforcement learning policies that can operate in large state and/or action space MDPs (Silver et al., 2017; Schrittwieser et al., 2020). From playing one of the most complicated games to interacting with humans via language reinforcement learning is currently being used in many different fields (Schrittwieser et al., 2020; Popova et al., 2018).

Although deep reinforcement learning policies achieved many successes in manifold fields, it has been also observed that there are still many concerns and unanswered questions regarding their safety and reliability (Huang et al., 2017; Korkmaz & Brown-Cohen, 2023; Korkmaz, 2024a). The vulnerabilities of deep neural networks were discussed in many studies starting from the seminal work of Goodfellow et al. (2015); Szegedy et al. (2014).

To target these vulnerabilities a line of work focused on robustifying deep reinforcement learning policies via adversarial training, i.e. robust training, via regularizing the temporal difference loss with worst-case perturbations. Studies that focus on adversarial training claim that via these techniques we can obtain robust and safe deep reinforcement learning policies. In this paper we will challenge the below acknowledged consensus in the field.

> Adversarial robust deep reinforcement learning training ensures robustness and robust deep reinforcement learning policies are safe and robust.

In this paper, we will show that the above acknowledged consensus on the robustness and safety of adversarial robust deep reinforcement learning is in fact false. In particular, in this paper we will highlight and summarize quite recent work on investigating robustness of adversarial training, i.e. robust training, in deep reinforcement learning, and demonstrate that adversarial robust deep

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

reinforcement learning is neither robust nor safe. In particular the contributions of the paper as follows:

- We demonstrate that while deep reinforcement learning policies are vulnerable to black-box adversarial attacks, robust trained deep reinforcement learning policies are in fact not robust and robust deep reinforcement learning policies can be attacked in a black-box setting without having access to the training details of the policy (e.g. algorithm, neural network architecture, training dataset).
- The results reported in our paper demonstrate that the generalization capabilities of standard (i.e. "non-robust") deep reinforcement learning is substantially higher than robust deep reinforcement learning policies.

This paper serves the purpose of concisely explaining and delivering the results stemming from discovering several issues of robust training methods. The results reported in this paper are initially discovered and published in (Korkmaz, 2024a, 2023, 2022a, 2021d). Please see these original papers for more details.

2 Preliminaries and Background

Markov decision processes (MDPs) represented as a tuple of $\mathcal{M}$ $\langle S, A, \mathcal{R}, \mathcal{P}, \gamma, \rho_0 \rangle$ where $s \in S$ represents a state from the state space, $a \in A$ represents an action from the action space, $\mathcal{P}$ represents the transition probability distribution on $S \times A \times S$, $\mathcal{R} : S \times A \to \mathbb{R}$ represents the reward function, $\gamma \in [0, 1)$ represents the discount factor, and the ρ_0 represents the initial state distribution. The objective in reinforcement learning is to learn an optimal policy via interacting with an environment by observing states, taking actions and receiving rewards where a policy $\pi : S \to \mathcal{P}(A)$ for an MDP $\mathcal{M}$ represents a probability distribution on actions in each $s \in S$. The objective is to maximize the rewards

$$R = \mathbb{E}_{a_t \sim \pi(s_t, \cdot)} \sum_t \gamma^t \mathcal{R}(s_t, a_t, s_{t+1}),$$

where $a_t \sim \pi(s_t)$. In Q-learning the learned policy is parametrized by a state-action value function $Q : S \times A \to \mathbb{R}$, which represents the value of taking action a in state s. Learning the optimal state-action value function is achieved via iterative Bellman update

$$Q(s_t, a_t) = \mathcal{R}(s_t, a_t) + \gamma \sum_{s_t} \mathcal{P}(s_{t+1}|s_t, a_t) V(s_{t+1}).$$

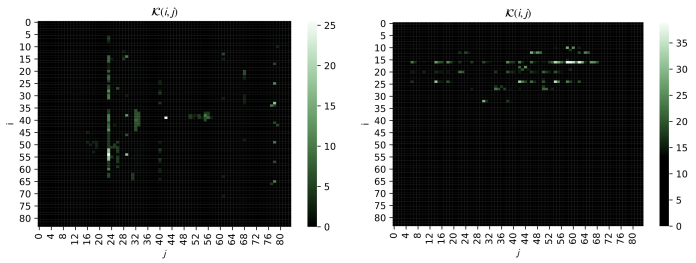
For more details on adversarial optimization techniques in deep reinforcement learning see Korkmaz (2020).

3 Robust Deep Reinforcement Learning Learns Non-Robust Features

In this section we will reiterate the results reported in Korkmaz (2021d). For more details please see the original paper (Korkmaz, 2021d). In particular, the results reported in Figure 1 demonstrates that adversarially robust trained policies still have vulnerabilities compared to vanilla trained policies. Furthermore, robust deep reinforcement learning policies learn a new set of non-robustness compared to vanilla trained deep reinforcement learning policies. This non-robustness is clearly outlined in detail in Figure 1. The bright colors in Figure 1 demonstrates the non-robust regions of these policies[1]. Furthermore, this paper reveals that it is possible to compute adversarial perturbations for robust deep reinforcement learning models Korkmaz (2021d), and these adversarial perturbations computed from robust models are concentrated in lower frequencies (Korkmaz, 2021b). To see the detailed evaluation visit the original paper (Korkmaz, 2021d).

Robust deep reinforcement learning learns a new set of non-robust features compared to standard deep reinforcement learning.

[1] See the main study for non-robustness mapping (Korkmaz, 2021d,a)



Robust Training Vanilla Training

Figure 1: Heatmap results of deep reinforcement learning policy vulnerabilities. Left: Certified robust trained. Right: Vanilla trained deep reinforcement learning policy.

4 Black-Box Vulnerability of Robust Deep Reinforcement Learning

In this section we will provide results for the adversarial vulnerabilities of robust deep reinforcement learning policies. In particular, we will reiterate the results regarding robust deep reinforcement learning reported in (Korkmaz, 2022a). In particular, the results reported in Table 1 demonstrate that the adversarial directions are shared not only across MDPs, but further across algorithms. Note that $\mathbb{A}^{\text{random}}_{\text{alg}+\mathbb{M}}$ represents the adversarial setting where the adversarial direction is transferred both across MDPs and algorithms.

The state-of-the-art adversarially trained deep reinforcement learning polices utilize the State-Adversarial DDQN (SA-DDQN) algorithm proposed by Zhang et al. (2020) with prioritized experience replay. All of the experiments are conducted in the high-dimensional state representation MDPs, i.e. Arcade Learning environment (Bellemare et al., 2013). Vanilla trained deep reinforcement learning policies are trained with deep double Q-learning, i.e. DDQN (van Hasselt, 2010; van Hasselt et al., 2016).

Table 1: Impacts of $\mathbb{A}^{\text{Gaussian}}$, $\mathbb{A}^{\text{random}}_{\text{alg}}$ and $\mathbb{A}^{\text{random}}_{\text{alg}+\mathbb{M}}$ where the perturbation is computed from a policy trained with DDQN and introduced to the observation system of the state-of-the-art adversarially robust trained deep reinforcement learning policy.

MDPs	$\mathbb{A}^{\text{Gaussian}}$	$\mathbb{A}^{\text{random}}_{\text{alg}}$	$\mathbb{A}^{\text{random}}_{\text{alg}+\mathbb{M}}$
RoadRunner	0.023±0.058	0.397±0.024	**0.546±0.014**
Pong	0.019±0.007	**1.0±0.000**	0.659±0.069
BankHeist	0.061±0.012	**0.758±0.042**	0.241±0.009

Robust deep reinforcement learning is not only non-robust, but further can even be attacked via black-box adversarial attacks.

5 Robust Deep Reinforcement Learning is Neither Robust Nor Can Generalize

In this section we will highlight the results regarding how vanilla trained deep reinforcement learning policies can generalize better and further are more robust than adversarial robust deep reinforcement learning policies. In particular, we will reiterate the results discovered in (Korkmaz, 2023) regarding generalization and robust deep reinforcement learning. The results reported in Figure 2 demonstrate clearly the separation between generalization capabilities of adversarial training (i.e. robust deep reinforcement learning) and vanilla training in deep reinforcement learning. In particular, when the environment experiences imperceptible changes the results reported in Figure 2 demonstrate that vanilla trained policies can perform substantially better than robust trained policies. Furthermore, see that the study of Korkmaz (2023) provides the contradistinction between adversarial attacks and natural semantically meaningful changes to the environment within the perceptual similarity metric. The results reported in this paper demonstrate that natural changes made to an environment that are imperceptible as much as the adversarial perturbations can cause more damage on the

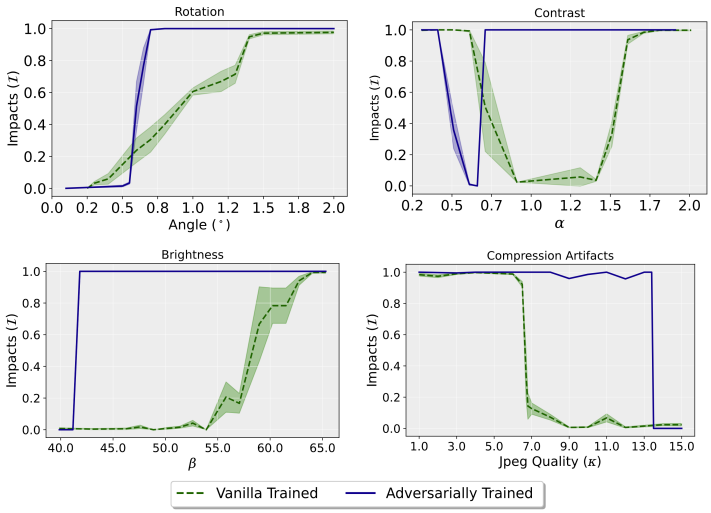


Figure 2: The performance drop results when moved along policy-independent high-sensitivity directions of the state-of-the-art adversarially trained deep reinforcement learning policy manifold and vanilla trained deep reinforcement learning policy manifold with varying degrees of discrete cosine transform artifacts, brightness, rotation, and contrast.

deep reinforcement learning policy performance.[2] Furthermore, the results reported in Figure 1 demonstrate that adversarially robust trained policies still have vulnerabilities compared to vanilla trained policies[3].

Reinforcement learning can generalize better than robust deep reinforcement learning.

In this study we highlight and reiterate the results originally discovered and reported in the following papers. Please visit the original papers for a detailed analysis.

Ezgi Korkmaz. Adversarial Robust Deep Reinforcement Learning Requires Redefining Robustness. AAAI Conference on Artificial Intelligence, **AAAI 2023**.

Ezgi Korkmaz. Deep Reinforcement Learning Policies Learn Shared Adversarial Features Across MDPs. AAAI Conference on Artificial Intelligence, **AAAI 2022**.

Ezgi Korkmaz. Investigating Vulnerabilities of Deep Neural Policies. Conference on Uncertainty in Artificial Intelligence (UAI), Proceedings of Machine Learning Research (PMLR), **PMLR 2021**.

6 Conclusion

In this work we highlight the recent findings on investigations on the safety and robustness of robust deep reinforcement learning. The results reported demonstrate that adversarial robust deep reinforcement learning is in fact neither robust nor safe. While robust deep reinforcement learning policies can be attacked via black-box perturbations (i.e. the adversary does not have access to the training details of the policy of interest), furthermore vanilla trained deep reinforcement learning policies can generalize substantially better than robust deep reinforcement learning policies. These results require further attention on the safety of robust deep reinforcement learning policies and the definitions of robustness in reinforcement learning.

[2]See more details in (Korkmaz, 2021c, 2023). For a more comprehensive analysis on the connection between generalization and adversarial perspective see the recent survey (Korkmaz, 2024b).

[3]See (Korkmaz, 2022c,b) for robustness and safety of inverse reinforcement learning.

References

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research.*, pp. 253–279, 2013.

Goodfellow, I., Shelens, J., and Szegedy, C. Explaning and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Huang, S., Papernot, N., Goodfellow, Ian an Duan, Y., and Abbeel, P. Adversarial attacks on neural network policies. *Workshop Track of the 5th International Conference on Learning Representations*, 2017.

Korkmaz, E. Nesterov momentum adversarial perturbations in the deep reinforcement learning domain. *International Conference on Machine Learning, ICML 2020, Inductive Biases, Invariances and Generalization in Reinforcement Learning Workshop.*, 2020.

Korkmaz, E. Non-robust feature mapping in deep reinforcement learning. *International Conference on Machine Learning, ICML Adversarial Machine Learning Workshop*, 2021a.

Korkmaz, E. Adversarially trained neural policies in fourier domain. *International Conference on Machine Learning, ICML Adversarial Machine Learning Workshop*, 2021b.

Korkmaz, E. Adversarial training blocks generalization in neural policies. *International Conference on Learning Representation (ICLR) Robust and Reliable Machine Learning in the Real World Workshop*, 2021c.

Korkmaz, E. Investigating vulnerabilities of deep neural policies. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021*, volume 161 of *Proceedings of Machine Learning Research (PMLR)*, pp. 1661–1670. AUAI Press, 2021d.

Korkmaz, E. Deep reinforcement learning policies learn shared adversarial features across MDPs. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7229–7238, 2022a.

Korkmaz, E. Spectral robustness analysis of deep imitation learning. *Conference on Neural Information Processing Systems (NeurIPS) Machine Learning Safety Workshop*, 2022b.

Korkmaz, E. The robustness of inverse reinforcement learning. *International Conference on Machine Learning (ICML) Artificial Intelligence for Agent Based Modelling Workshop*, 2022c.

Korkmaz, E. Adversarial robust deep reinforcement learning requires redefining robustness. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 8369–8377. AAAI Press, 2023.

Korkmaz, E. Understanding and diagnosing deep reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning ICML*, Proceedings of Machine Learning Research (PMLR). PMLR, 2024a.

Korkmaz, E. A survey analyzing generalization in deep reinforcement learning. In *arXiv preprint arXiv:2401.02349*, 2024b.

Korkmaz, E. and Brown-Cohen, J. Detecting adversarial directions in deep reinforcement learning to make robust decisions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17534–17543. PMLR, 23–29 Jul 2023.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, a. G., Graves, A., Riedmiller, M., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Popova, M., Isayev, O., and Tropsha, A. Deep reinforcement learning for de-novo drug design. *Science Advances*, 2018.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T. P., and Silver, D. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588, 2020.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, a., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Driessche, v. d. G., Graepel, T., and Hassabis, D. Mastering the game of go without human knowledge. *Nature*, 500:354–359, 2017.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *In Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

van Hasselt, H. Double q-learning. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pp. 2613–2621. Curran Associates, Inc., 2010.

van Hasselt, H., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 2094–2100. AAAI Press, 2016.

Wang, Z., Schaul, T., Hessel, M., Van Hasselt, H., Lanctot, M., and De Freitas, N. Dueling network architectures for deep reinforcement learning. *Internation Conference on Machine Learning ICML.*, pp. 1995–2003, 2016.

Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D. S., and Hsieh, C. Robust deep reinforcement learning against adversarial perturbations on state observations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.