ADVERSARIAL TRAINING BLOCKS GENERALIZATION IN NEURAL POLICIES

Ezgi Korkmaz
ezgikorkmazmail@gmail.com

ABSTRACT

Deep neural networks have made it possible for reinforcement learning algorithms to learn from raw high dimensional inputs. This jump in the progress has caused deep reinforcement learning algorithms to be deployed in many different fields from financial markets to biomedical applications. While the vulnerability of deep neural networks to imperceptible specifically crafted perturbations has also been inherited by deep reinforcement learning agents, several adversarial training methods have been proposed to overcome this vulnerability. In this paper we focus on state-of-the-art adversarial training algorithms and investigate their robustness to semantically meaningful natural perturbations ranging from changes in brightness to rotation. We conduct several experiments in the OpenAI Atari environments, and find that state-of-the-art adversarially trained neural policies are more sensitive to natural perturbations than vanilla trained agents. We believe our investigation lays out intriguing properties of adversarial training and our observations can help build robust and generalizable neural policies.

1 INTRODUCTION

The advancements in deep neural networks lead to wide spread of applications from image classification Krizhevsky et al. (2012) to natural language processing Sutskever et al. (2014), and from speech recognition Hannun et al. (2014) to self learning systems Mnih et al. (2015). Yet the ability to learn from raw high dimensional data brought some of the drawbacks inherited from deep neural networks to deep neural policies.

Szegedy et al. (2014) showed that specifically crafted imperceptible perturbations can lead to misclassification in image classification. After this initial work a new research area emerged to investigate the vulnerabilities of deep neural networks against specifically crafted adversarial examples. While various works studied many different ways to compute these examples Carlini & Wagner (2017); Madry et al. (2018); Goodfellow et al. (2015); Kurakin et al. (2016) several works focused on studying ways to increase the robustness against such specifically crafted perturbations, based on training with the existence of such perturbations Madry et al. (2018); Tramèr et al. (2018); Goodfellow et al. (2015); Xie & Yuille (2020).

As image classification suffered from this vulnerability towards worst-case distributional shift in the input a series of work conducted in deep reinforcement learning showed that deep neural policies are also susceptible towards the specifically crafted imperceptible perturbations Huang et al. (2017); Kos & Song (2017); Pattanaik et al. (2018); Lin et al. (2017); Korkmaz (2020). While one line of work put effort on exploring these vulnerabilities in deep neural policies Korkmaz (2021a;c;b; 2022a), another line in parallel focused making them robust and reliable via adversarial training Pinto et al. (2017); Mandlekar et al. (2017); Zhang et al. (2020).

To be able to build generalizable and robust deep neural policies, in this paper we approach the problem of performance degradation with respect to observed input from a wider perspective of distributional shift and make the following contributions:

- We conduct an investigation on the robustness of state-of-the-art adversarially trained deep neural policies against various types of distributional shift in the input.
- We perform several experiments in the OpenAI Atari baselines.

- We compare the performance drop of the vanilla trained agents with the state-of-the-art adversarially trained agents against natural semantically meaningful perturbations.
- We find that vanilla trained agents are more robust against natural semantically meaningful perturbations than the state-of-the-art adversarially trained agents.

The full version of this paper is published in AAAI Conference on Artificial Intelligence 2023 Korkmaz (2023). In the full version of the paper the natural and adversarial directions are contrasted in terms of their perceptual similarities and their impact on the policy performance. Furthermore, the results reported in the full version of the paper call into question the definition of *robustness* in deep reinforcement learning, and recalls further thinking on the term robustness.

2 Background

2.1 Adversarial Examples

Adversarial examples were introduced in computer vision by Szegedy et al. (2014) based on producing perturbations via a box constrained optimization method. Goodfellow et al. (2015) proposed a fast method called the fast gradient sign method (FGSM) to produce adversarial examples based on the linearization of the cost function used to train the network at the input sample point.

$$x_{\text{adv}} = x + \epsilon \cdot \frac{\nabla_x J(x,y)}{||\nabla_x J(x,y)||_p}, \tag{1}$$

where x represents the input, y represents the labels and $J(x,y)$ represents the cost function used to train the deep neural network. Kurakin et al. (2016) propose the iterative version of the fast gradient sign method inside an ϵ-ball.

$$x^0_{\text{adv}} = x, \tag{2}$$

$$x^{N+1}_{\text{adv}} = \text{clip}_\epsilon(x^N_{\text{adv}} + \alpha \text{sign}(\nabla_x J(x^N_{\text{adv}}, y))) \tag{3}$$

This method is also known as projected gradient descent (PGD) as proposed by Madry et al. (2018).

2.2 Deep Reinforcement Learning and Adversaries

Initially adversarial examples were introduced to the deep reinforcement learning domain by Huang et al. (2017) and Kos & Song (2017) concurrently by utilizing FGSM as proposed by Goodfellow et al. (2015). Korkmaz (2020) proposed to optimize the adversarial example search via Nesterov momentum. Quite recently, **?** showed that adversarial features learned by deep reinforcement learning policies are shared across states, across MDPs and across algorithms. Recently, there are also some studies on the robustness of deep inverse reinforcement learning algorithms Korkmaz (2022b). These new studies investigates the vulnerabilities and adversarial robustness of the state-of-the-art deep inverse reinforcement learning algorithms Korkmaz (2022c). Several studies have been conducted to make deep reinforcement learning policies more robust to such specifically crafted malicious examples. In connection to exposition of these vulnerabilities Mandlekar et al. (2017) use adversarial examples produced via modifying the environment by taking the gradient of the cost function with respect to input at training time to regularize the policy in an attempt to increase robustness. Pinto et al. (2017) model the interaction between the perturbation maker and the agent as a zero-sum Markov game, and propose a joint training algorithm to improve robustness against an adversary that aims to minimize the expected cumulative reward of the agent. Gleave et al. (2020) model the relationship between the adversary and the agent as a zero-sum game where the agent is limited to taking natural actions in the environment rather than minimal ℓ_p-norm bounded perturbations, and proposed a self-playing approach to gain robustness against such an adversary. Finally, Zhang et al. (2020) propose a modified MDP called a state-adversarial MDP with the aim of obtaining theoretically principled robust polices. Recently, Korkmaz (2021c;a) demonstrated the vulnerabilities of state-of-the-art adversarial training techniques in deep reinforcement learning. In our work, we further will show that adversarial training causes problems on the generalization capabilities of deep reinforcement learning agents.

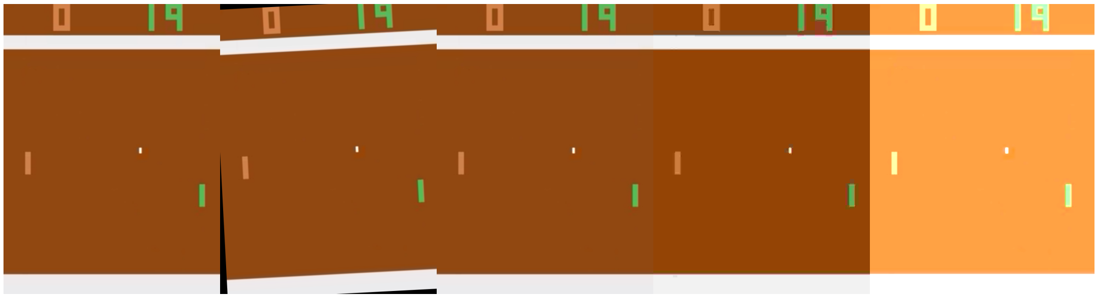

Figure 1: Original frame and environmental modifications. Columns: original frame, rotation, compression artifacts, brightness and contrast.

3 Experimental Setup

In our experiments we use OpenAI Brockman et al. (2016) Atari baselines Bellemare et al. (2013). Our models are trained with DDQN Wang et al. (2016) and Zhang et al. (2020). We test trained policies averaged over 10 episodes. We measure the performance drop of the agent as,

$$\mathcal{I} = \frac{\text{Score}_{\text{clean}} - \text{Score}_{\text{adv}}}{\text{Score}_{\text{clean}} - \text{Score}^{\text{fixed}}_{\text{min}}}. \tag{4}$$

where $\text{Score}_{\text{clean}}$ represents the score from a clean run of the game where no perturbations are introduced to the agent's observations, $\text{Score}^{\text{fixed}}_{\text{min}}$ represents the minimum score available for a given game, and $\text{Score}_{\text{adv}}$ represents a run of the game where perturbations are introduced to agent's observation system.

4 Adversarial Trained Models under Natural Perturbations

In this paper we focus on investigating adversarial training and its effects on generalization of neural policies. In particular, we focus on the state-of-the-art adversarial training method proposed by Zhang et al. (2020). In this paper the authors propose to model the observation perturbations as a modified MDP, which they refer to as a state-adversarial MDP, with the aim of making the agent more robust towards natural measurement errors or adversarial noises. The authors test their adversarially trained models under the PGD attack proposed by Madry et al. (2018). In our paper we test adversarially trained DDQN models proposed by Zhang et al. (2020) under several natural perturbations proposed by Korkmaz (2023). In particular, we introduce several minimal natural perturbations (e.g. rotation, brightness, contrast and compression artifacts) to deep reinforcement learning policy observation systems and investigate the performance drop of the trained policies. Note that these natural modifications are quite small and some of the perturbations are not even recognizable by human perception (see Figure 1). Most importantly, we compare the proposed state-of-the-art adversarially trained agents with vanilla trained agents under the natural semantically meaningful perturbations and find that vanilla trained models are more robust than adversarially trained models against many natural perturbations including brightness, rotation, compression artifacts, and contrast.

While we observe this particular sensitivity increase to natural perturbations in adversarially trained models in baselines like Atari environments where the generalization capabilities of deep reinforcement learning agents is not the primary concern, this sensitivity increase caused by adversarial training might cause severe problems for the environments where generalization is essential for learning reasonable policies. In particular, Cobbe et al. (2019) environment is more challenging and focused on testing generalization capabilities compared to Atari Baselines. The observation on the limitations of adversarial trained neural policies towards such natural perturbations implies that adversarially trained neural policies are going to experience challenges in obtaining reasonable policies in these type of environments proposed by Cobbe et al. (2019).

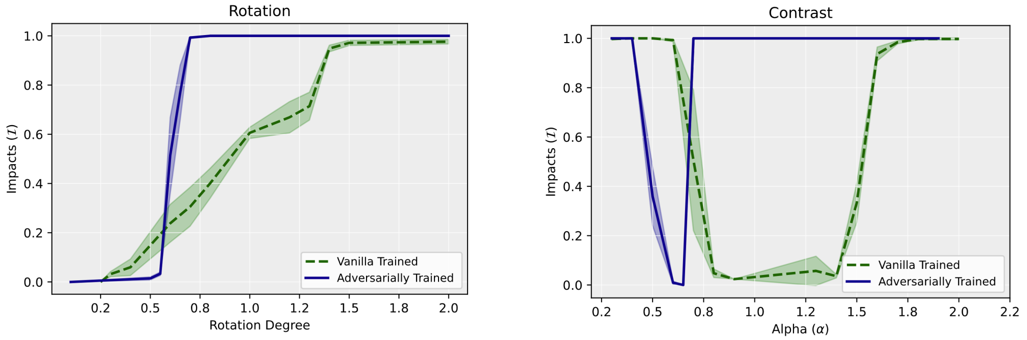

Figure 2: Sensitivity of adversarially trained model and vanilla trained model to the changes in rotation and contrast.

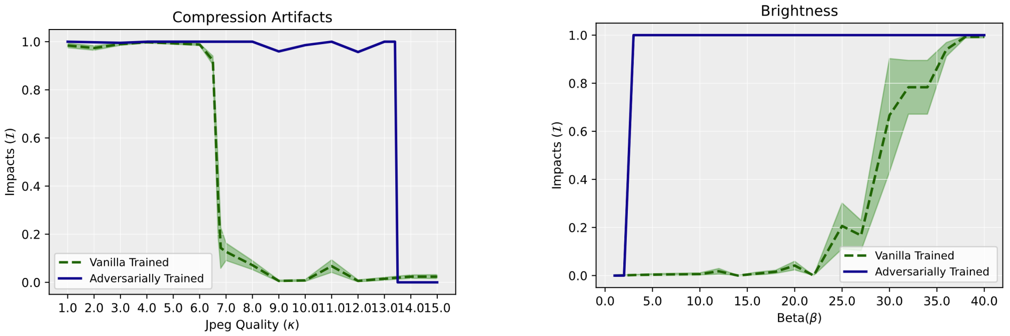

Figure 3: Sensitivity of adversarially trained model and vanilla trained model to the changes in compression artifacts and brightness.

5 Conclusion

In this paper we focused on the generalization capabilities of vanilla trained and the state-of-the-art adversarially trained neural policies. In particular, we tested adversarially trained neural policies under semantically meaningful natural perturbations and we found that vanilla trained deep neural policies are more robust against natural perturbations than adversarially trained deep neural policies. We further argue that this kind of sensitivity increase towards natural perturbations in adversarially trained models can hurt generalization. We belive our study provides a holistic view on the robustness of adversarial training for deep neural policies and can contribute to designing resilient and robust self learning systems for future work.

References

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

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv:1606.01540*, 2016.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *In 2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. *CoRR*, abs/1912.01588, 2019. URL `http://arxiv.org/abs/1912.01588`.

Adam Gleave, Michael Dennis, Cody Wild, Kant Neel, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. *International Conference on Learning Representations ICLR*, 2020.

Ian Goodfellow, Jonathan Shelens, and Christian Szegedy. Explaning and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Diamos Greg, Erich Else, Ryan Prenger, Sanjeev Satheesh, Sengupta Shubho, Ada Coates, and Andrew Ng. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Sandy Huang, Nicholas Papernot, Yan Goodfellow, Ian an Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *Workshop Track of the 5th International Conference on Learning Representations*, 2017.

Ezgi Korkmaz. 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.

Ezgi Korkmaz. Inaccuracy of state-action value function for non-optimal actions in adversarially trained deep neural policies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2323–2327, June 2021a.

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

Ezgi Korkmaz. 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, 2021c.

Ezgi Korkmaz. Deep reinforcement learning policies learn shared adversarial features across mdps. *AAAI Conference on Artificial Intelligence*, 2022a.

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

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

Ezgi Korkmaz. Adversarial robust deep reinforcement learning requires redefining robustness. *AAAI Conference on Artificial Intelligence*, 2023.

Jernej Kos and Dawn Song. Delving into adversarial attacks on deep policies. *International Conference on Learning Representations*, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Yen-Chen Lin, Hong Zhang-Wei, Yuan-Hong Liao, Meng-Li Shih, ing-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 3756–3762, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Ajay Mandlekar, Yuke Zhu, Animesh Garg, Li Fei-Fei, and Silvio Savarese. Adversarially robust policy learning: Active construction of physically-plausible perturbations. *In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3932–3939, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, arc G Bellemare, Alex Graves, Martin Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518: 529–533, 2015.

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, and Bommannan Gautham. Robust deep reinforcement learning with adversarial attacks. *In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2040–2042, 2018.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. *International Conference on Learning Representations ICLR*, 2017.

Ilya Sutskever, Oriol Vinyals, and Quoc V. . Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dimutru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *In Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando. De Freitas. Dueling network architectures for deep reinforcement learning. *Internation Conference on Machine Learning ICML.*, pp. 1995–2003, 2016.

Cihang Xie and Alan L. Yuille. Intriguing properties of adversarial training at scale. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane S. Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.