Robustness of Inverse Reinforcement Learning

Ezgi Korkmaz [1]

Abstract

Reinforcement learning research experienced substantial jumps in its progress after the first achievement on utilizing deep neural networks to approximate the state-action value function in high-dimensional states. While deep reinforcement learning algorithms are currently being employed in many different tasks from industrial control to biomedical applications, the fact that an MDP has to provide a clear reward function limits the tasks that can be achieved via reinforcement learning. In this line of research, some studies proposed to directly learn a policy from observing expert trajectories (i.e. imitation learning), and others proposed to learn a reward function from the expert demonstrations (i.e. inverse reinforcement learning). In this paper we will focus on robustness and vulnerabilities of deep imitation learning and deep inverse reinforcement learning policies. Furthermore, we will layout non-robust features learnt by the deep inverse reinforcement learning policies. We conduct experiments in the Arcade Learning Environment (ALE), and compare the non-robust features learnt by the deep inverse reinforcement learning algorithms to vanilla trained deep reinforcement learning policies. We hope that our study can provide a basis for the future discussions on the robustness of both deep inverse reinforcement learning and deep reinforcement learning.

1. Introduction

Learning from complex state representations was initially achieved by utilizing deep neural networks as function approximators (Mnih et al., 2016). With this initial success reinforcement learning algorithms currently can solve highly complex games (e.g. Go), can learn several robotics tasks (Dosovitsky et al., 2017), and is currently being used in diverse fields from finance to biomedical applications (Yauney & Pratik, 2018).

While various tasks can be learned by reinforcement learning algorithms, the fact that an MDP has to provide a reward function can be quite restrictive for certain types of tasks. To address this several studies focused on proposing algorithms to learn functioning policies without the presence of a reward function. One way to achieve this is to learn an optimal policy from expert trajectories (i.e. imitation learning), and another way is to construct a reward function from observing expert policies (i.e. inverse reinforcement learning).

Initially adversarial perturbations in deep neural networks were discussed by (Goodfellow et al., 2015). The authors of this work demonstrate the effects of introducing invisible adversarial perturbations to the images of the neural network classifiers. Following this the adversarial robustness of deep reinforcement learning policies towards optimized perturbations has been discussed by many studies (Huang et al., 2017; Kos & Song, 2017; Korkmaz, 2020; 2022b). Furthermore, quite recent studies showed that the adversarial perturbations do not need to be specifically optimized, neither for the state observations nor for the MDP, to cause damage to the deep reinforcement learning policy performance (Korkmaz, 2022a). Yet, to the best of our knowledge our paper is the first one to investigate the robustness of deep inverse reinforcement learning and deep imitation learning policies.

In our paper we want to answer several questions:

- *Do state-of-the-art deep imitation learning policies and deep inverse reinforcement learning policies have vulnerabilities towards the state representations they have learnt?*
- *What are the differences in the non-robust features learnt between state-of-the-art deep inverse reinforcement learning policies and vanilla trained deep reinforcement learning policies?*

Hence, to answer these questions in this paper we focus on investigating robustness of deep imitation learning and deep inverse reinforcement learning, and make the following

[1]DeepMind. Correspondence to: Ezgi Korkmaz <ezgikorkmazmail@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning Artificial Intelligence for Agent Based Modelling Workshop, Baltimore, Maryland, USA, PMLR 162, 2022.

contributions:

- We utilize the KMAP algorithm to provide an accurate portrayal of the non-robust features learnt by the state-of-the-art imitation learning and deep inverse reinforcement learning algorithms.
- We conduct experiments in the Arcade Learning Environment (ALE) and we compare the non-robust features learnt by the state-of-the-art imitation learning policy to the vanilla deep reinforcement learning algorithm for high dimensional state representation environments.
- We demonstrate that the policies learnt via vanilla deep reinforcement learning algorithms are more robust compared to the state-of-the-art imitation learning algorithms.

2. Relative Work and Background

2.1. Deep Reinforcement Learning

In this paper we focus on deep reinforcement learning for Markov decision processes (MDPs) given by a set of continuous states S, a set of discrete actions A, a transition probability distribution P on $S \times A \times S$, and a reward function $r : S \times A \to \mathbb{R}$. A policy $\pi : S \to \mathcal{P}(A)$ for an MDP assigns a probability distribution on actions to each $s \in S$. The goal for the reinforcement learning agent is to learn a policy π that maximizes the expected cumulative discounted 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

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

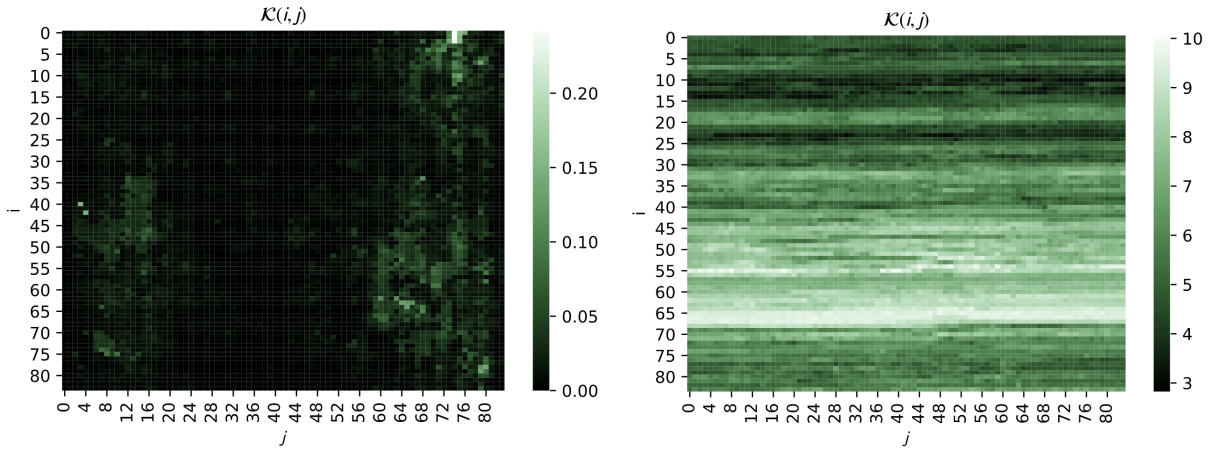
Let $a^*(s) = \arg\max_a Q(s, a)$ denote the highest Q-value for an action in state s. The ϵ-greedy policy of the agent for Q-learning is given by taking action $a^*(s)$ with probability $1 - \epsilon$, and a uniformly random action with probability ϵ.

2.2. Inverse Reinforcement Learning and Imitation Learning

Inverse reinforcement learning focuses on constructing a reward function from a set of observations of expert demonstrations. Thus, once the reward function is learnt from expert trajectories reinforcement learning is used to learn an optimal policy. In particular, in this line of research (Ng & Russell, 2000) shows that multiple different reward function can be constructed for an observed optimal policy. Another way of learning without rewards is imitation learning that focuses on the setting of learning a functioning policy from observing a given set of expert trajectories (Kostrikov et al., 2020). Quite recently, Garg et al. (2021) proposed to construct a single Q-function from the observed expert trajectories to represent both the reward function and the policy (IQ-Learn). This study is the first to achieve the learning of functioning policies via inverse reinforcement learning from highly complex state representations. Furthermore, the authors of this study argue that the fact that a single Q-function is learnt from expert demonstrations is enough to reconstruct the reward function. Thus, this study shows that the predicted rewards from the IQ-Learn algorithm are highly correlated with the true rewards received from the environment.

2.3. Robustness in Reinforcement Learning

The robustness of reinforcement learning policies has been under discussion starting from the initial work of (Huang et al., 2017) mostly focusing on demonstrating the susceptibilities of deep reinforcement learning policies to imperceptible malicious (i.e. adversarial) perturbations produced via the fast gradient sign method proposed by Goodfellow et al. (2015). Several different studies have been conducted so far on optimizing for specifically crafted adversarial perturbations in deep reinforcement learning; however, some studies took a different direction and focused on more natural unoptimized effects on the environment (Korkmaz, 2021d). These studies consider natural semantically meaningful modifications to the environment. While some studies went more along the optimization side of these perturbations (Korkmaz, 2020), several focused on ways to make deep reinforcement learning policies robust to adversarial perturbations. On this line of research several studies modeled the relationship between the deep reinforcement learning policy and intervention made by the adversary as a zero-sum Markov game (Gleave et al., 2020; Pinto et al., 2017). In some of these zero-sum Markov game models the adversary intervention is limited to changing the environment dynamics (Pinto et al., 2017), in others this intervention is limited to a set of natural actions taken by the adversary in the given environment (Gleave et al., 2020). While several theoretically justified adversarial training algorithms were proposed more recently, several concerns and criticisms were raised on their promised robustness capacities and their robustness notions (Korkmaz, 2022a; 2021c;b). In particular, quite recently Korkmaz (2021e) proposed several techniques to investigate and estimate the vulnerabilities of deep reinforcement learning policies by revealing the current non-robust



Vanilla Deep Reinforcement Learning

Deep Inverse Reinforcement Learning

Figure 1. KMAP results of vanilla trained deep reinforcement learning policy and the state-of-the-art deep inverse reinforcement learning policy for Pong.

features learnt by the state-of-the-art adversarial training techniques.

3. Investigating Robustness of Imitation Learning and Inverse Reinforcement Learning

In this section we will use the tools introduced in Korkmaz (2021e;a) to investigate robustness of imitation learning policies. In particular, to highlight the non-robust features learnt by the imitation learning policy trained in high dimensional state representations we will use the KMAP algorithm. In more detail, let $Z_{i,j} : S \rightarrow S$ be the function which sets the i, j coordinate of s to zero and leaves the other coordinates unchanged. $\mathcal{K}(i, j)$ defined as,

$$\mathcal{K}(i,j) = Q(s, \arg\max_a Q(s,a)) - Q(s, \arg\max_a Q(Z_{i,j}(s), a)).$$

Thus, $\mathcal{K}(i, j)$ builds a portrait of the contribution and weight of each observed feature on the learnt representation and the decision that has been made by the policy over visited state observations for an entire episode.

All of the experiments are conducted in the Arcade Learning Environment (ALE) (Bellemare et al., 2013) with the OpenAI version (Brockman et al., 2016). The vanilla trained deep reinforcement learning policy is trained via Deep Double Q-Network proposed by (Hasselt et al., 2016) (the initial idea is proposed in (van Hasselt, 2010)). The deep inverse and deep imitation learning policy is trained via the IQ-Learn algorithm proposed by (Garg et al., 2021). All of the hyperparameters are exactly the same with the original paper.

Figure 1 shows the KMAP results for the vanilla trained deep reinforcement learning policy and the policy trained with imitation learning. We observe that the non-robust feature patterns learnt by imitation learning are completely different and disjoint from the game semantics compared to vanilla trained deep reinforcement learning. While the non-robust feature patterns learnt by the deep inverse reinforcement learning policy are independent from the state representations and MDP semantics, Figure 1 also demonstrates the sparsity of the non-robust features learned by vanilla trained deep reinforcement learning policy when compared to deep inverse reinforcement learning policy. One intriguing takeaway from the results presented in Section 3 is that the exploration plays a foundational role in the representations learnt by the policy. Thus, the fact that non-robust features learnt by the deep inverse reinforcement learning policy are decoupled from the MDP semantics demonstrates the role of the exploration on the non-robust feature patterns learnt by the agent.

4. Conclusion

In this paper we focused on the robustness of deep inverse reinforcement learning policies and deep imitation learning policies. In particular we wanted to answer the following questions: (i) Do the-state-of-the-art deep inverse reinforcement learning and deep imitation learning policies learn non-robust features from the MDP they are trained in?, and (ii) What are the differences in the non-robust features learnt by the deep inverse reinforcement learning policies and deep reinforcement learning policies? We show in the Arcade Learning Environment (ALE) that deep inverse reinforcement learning policies and deep imitation learning policies do learn non-robust features from complex state representa-

tion MDPs. More importantly, when compared to the vanilla trained deep reinforcement learning policies the deep inverse reinforcement learning policies learn non-robust features that are disjoint from the game semantics. Furthermore, the non-robust features learnt by the vanilla trained deep reinforcement learning policies are sparser than the non-robust features learnt by the deep inverse reinforcement learning and deep imitation learning policies. We believe our study can provide an initial basis on understanding the robustness of deep imitation and deep inverse reinforcement learning policies.

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.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv:1606.01540*, 2016.

Dosovitsky, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. Carla: An open urban driving simulator. *In Proceedings of the Conference on Robot Learning (CoRL)*, 78:1–16, 2017.

Garg, D., Chakraborty, S., Cundy, C., Song, J., and Ermon, S. Iq-learn: Inverse soft-q learning for imitation. *Neural Information Processing Systems (NeurIPS) [Spotlight Presentation]*, 2021.

Gleave, A., Dennis, M., Wild, C., Neel, K., Levine, S., and Russell, S. Adversarial policies: Attacking deep reinforcement learning. *International Conference on Learning Representations ICLR*, 2020.

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

Hasselt, H. v., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. *In Thirtieth AAAI conference on artificial intelligence*, 2016.

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. 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 2021c.

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*, 2021d.

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, 2021e.

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. Adversarial attacks against deep inverse reinforcement learning. *International Conference on Machine Learning, ICML Complex Feedback in Online Learning Workshop*, 2022b.

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

Kostrikov, I., Nachum, O., and Tompson, J. Imitation learning via off-policy distribution matching. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Mnih, V., Puigdomenech, A. B., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. *In International Conference on Machine Learning*, pp. 1928–1937, 2016.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In Langley, P. (ed.), *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pp. 663–670, 2000.

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. Robust adversarial reinforcement learning. *International Conference on Learning Representations ICLR*, 2017.

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.

Yauney, G. and Pratik, S. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. *In Machine Learning for Healthcare Conference*, pp. 161–226, 2018.