The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Adversarial Robust Deep Reinforcement Learning Requires Redefining Robustness

Ezgi Korkmaz

Abstract

Learning from raw high dimensional data via interaction with a given environment has been effectively achieved through the utilization of deep neural networks. Yet the observed degradation in policy performance caused by imperceptible worst-case policy dependent translations along high sensitivity directions (i.e. adversarial perturbations) raises concerns on the robustness of deep reinforcement learning policies. In our paper, we show that these high sensitivity directions do not lie only along particular worst-case directions, but rather are more abundant in the deep neural policy landscape and can be found via more natural means in a black-box setting. Furthermore, we show that vanilla training techniques intriguingly result in learning more robust policies compared to the policies learnt via the state-of-the-art adversarial training techniques. We believe our work lays out intriguing properties of the deep reinforcement learning policy manifold and our results can help to build robust and generalizable deep reinforcement learning policies.

1 Introduction

Following the initial work of Mnih et al. (2015), the use of deep neural networks as function approximators in reinforcement learning has led to a dramatic increase in the capabilities of reinforcement learning policies (Schulman et al. 2017; Vinyals et al. 2019; Schrittwieser et al. 2020). In particular, these developments allow for the direct learning of strong policies from raw, high-dimensional inputs (i.e. visual observations). With the successes of these new methods come new challenges regarding the robustness and generalization capabilities of deep reinforcement learning agents.

Initially, 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 abilities of deep neural networks against specifically crafted adversarial examples. While various works studied many different ways to compute these examples (Carlini and Wagner 2017; Madry et al. 2018; Goodfellow, Shelens, and Szegedy 2015; Kurakin, Goodfellow, and Bengio 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, Shelens, and Szegedy 2015; Xie and 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 to specifically crafted imperceptible perturbations (Huang et al. 2017; Kos and Song 2017; Pattanaik et al. 2018; Yen-Chen et al. 2017; Korkmaz 2020; Sun et al. 2020; Korkmaz 2021b). While one line of work put effort on exploring these vulnerabilities in deep neural policies, another line in parallel focused on making them robust and reliable via adversarial training (Pinto et al. 2017; Mandlekar et al. 2017; Gleave et al. 2020).

While adversarial perturbations and adversarial training provide a notion of robustness for trained deep neural policies, in this paper we approach the resilience problem of deep reinforcement learning from a wider perspective, and propose to investigate the deep neural policy manifold along high-sensitivity directions. Along this line we essentially seek answers for the following questions:

- *How can we probe the deep neural policy decision boundary with policy-independent high-sensitivity directions innate to the MDP within the perceptual similarity bound?*
- *Is it possible to affect the state-of-the-art deep reinforcement learning policy performance trained in high-dimensional state representation MDPs with policy-independent high-sensitivity directions intrinsic to the MDP?*
- *What are the effects of state-of-the-art certified adversarial training on the robustness of the policy compared to straightforward vanilla training when policy-independent high-sensitivity directions are present?*

Thus, to be able answer these questions, in this work we focus on the notion of robustness for deep reinforcement learning policies and make the following contributions:

- We introduce policy-independent high-sensitivity directions innate to the MDP, and probe the deep reinforcement learning manifold via these policy-independent directions.
- We run multiple experiments in the Arcade Learning Environment (ALE) in various games with high di-

mensional state representation and provide the relationship between the perceptual similarities to base states under policy dependent and policy-independent high-sensitivity directions.

- We compare policy-independent high-sensitivity directions with the state-of-the-art adversarial directions based on ℓ_p-norm changes, and we show that policy-independent high-sensitivity directions intrinsic to the MDP are competitive in degrading the performance of the deep reinforcement learning policies with lower perceptual similarity distance. Thus, the results of this contradistinction of adversarial directions and policy-independent high-sensitivity directions intrinsic to the MDP evidently demonstrates the abundance of high-sensitivity directions in the deep reinforcement learning policy manifold.
- Finally, we inspect state-of-the-art adversarial training under changes intrinsic to the MDP, and demonstrate that the adversarially trained models become more vulnerable to several different types of policy-independent high-sensitivity directions compared to vanilla trained models.

2 Background and Related Work

2.1 Preliminaries

In this paper we consider Markov Decision Processes (MDPs) given by a tuple $(S, A, \mathcal{T}, r, \gamma, s_i)$. The reinforcement learning agent interacts with the MDP by observing states $s \in S$, taking actions $a \in A$ and receiving rewards $r(s, a, s')$. Here s_i represents the initial state of the agent, and $\gamma \in (0, 1]$ represents the discount factor. The probability of transitioning to state s' when the agent takes action a in state s is determined by the Markovian transition kernel $\mathcal{T} : S \times A \times S \to \mathbb{R}$. The reward received by the agent when taking action a in state s is given by the reward function $r : S \times A \times S \to \mathbb{R}$. The goal of the agent is to learn a policy $\pi : S \times A \to \mathbb{R}$ which takes an action a in state s that maximizes the expected cumulative discounted reward $\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t, s_{t+1})$ that the agent receives via interacting with the environment.

$$\tilde{\pi} = \arg\max_{\pi} \sum_{t} \mathbb{E}_{s_t, a_t \sim \mathbb{P}_\pi}[r(s_t, a_t, s_{t+1})] \tag{1}$$

where $\mathbb{P}_\pi$ represents the occupancy distribution of the trajectory followed by the policy $\pi(a_t|s_t)$. Hence, this goal can be achieved via learning the state-action value function via iterative Bellman update

$$Q(s, a) = \mathbb{E}_\pi[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t, s_{t+1})|s_i = s, a_i = a]$$

assigning a value to each state-action pair. In high dimensional state representation MDPs the state-action values are estimated via function approximators.

$$\theta_{t+1} = \theta_t + \alpha(Q_t^{\text{target}} - Q(s_t, a_t; \theta_t))\nabla_{\theta_t} Q(s_t, a_t; \theta_t)$$

where Q_t^{target} is $r(s_t, a_t, s_{t+1}) + \gamma \max_a Q(s_{t+1}, a; \theta_t)$.

2.2 Computing Adversarial Directions

Szegedy et al. (2014) proposed to minimize the distance between the base image and adversarially produced image to create adversarial directions. The authors used box-constrained L-BFGS to solve this optimization problem. Goodfellow, Shelens, and Szegedy (2015) introduced the fast gradient method (FGM),

$$x_{\text{adv}} = x + \epsilon \cdot \frac{\nabla_x J(x, y)}{||\nabla_x J(x, y)||_p}, \tag{2}$$

for crafting adversarial examples in image classification by taking the gradient of the cost function $J(x, y)$ used to train the neural network in the direction of the input, where x is the input, y is the output label, and $J(x, y)$ is the cost function. Carlini and Wagner (2017) introduced targeted attacks in the image classification domain based on distance minimization between the adversarial image and the base image while targeting a particular label. Thus, in deep reinforcement learning the Carlini and Wagner (2017) formulation will find the minimum distance to a nearby state in an ϵ-ball $\mathcal{D}_{\epsilon,p}(s)$ such that,

$$\min_{\hat{s} \in \mathcal{D}_{\epsilon,p}(s)} \|\hat{s} - s\|_p$$
$$\text{subject to } \arg\max_a Q(s, a) \neq \arg\max_a Q(\hat{s}, a)$$

where $s \in S$ represents the base state, $\hat{s} \in \mathcal{D}_{\epsilon,p}(s)$ represents the state when it is moved along the adversarial directions. This formulation attempts to minimize the distance to the base state, constrained to states leading to sub-optimal actions as determined by the Q-network. Note that the Carlini & Wagner formulation has quite recently been used to demonstrate that the state-of-the-art adversarial trained policies share similar, and even in some cases identical, adversarial directions with the vanilla trained deep reinforcement learning policies (Korkmaz 2022). In contrast to adversarial attacks, in our proposed threat model we will not need any information on the cost function used to train the network, the Q-network of the trained agent, or access to the visited states themselves.

2.3 Adversarial Approach in Deep Reinforcement Learning

The first adversarial attacks on deep reinforcement learning introduced by Huang et al. (2017) and Kos and Song (2017) adapted FGSM from image classification to the deep reinforcement learning setting. Subsequently, Pinto et al. (2017) and Gleave et al. (2020) focused on modeling the interaction between the adversary and the agent as a zero-sum Markov game, while Yen-Chen et al. (2017); Sun et al. (2020) focused on strategically timing when (i.e. in which state) to attack an agent using perturbations computed with the Carlini & Wagner adversarial formulation. Orthogonal to this line of research some studies demonstrated that deep reinforcement learning policies learn adversarial directions from underlying MDPs that are shared across states, across MDPs and across algorithms (Korkmaz 2022). While proposing novel

techniques to uncover non-robust features, some recent studies demonstrated the lasting existence of the non-robust features in state-of-the-art adversarial training methods[1] (Korkmaz 2021b).

2.4 Perceptual Similarity Distance

Internal activations of networks trained for high-level tasks correspond to human perceptual judgements across different network architectures (Krizhevsky, Sutskever, and E. Hinton 2012; Simonyan and Zisserman 2015; Iandola et al. 2016) without calibration (Zhang et al. 2018). More importantly, it is possible to measure the perceptual similarity distance between two images with LPIPS matching human perception. Thus, in our experiments we measure the distance moved along the high sensitivity directions from the base states with LPIPS. In particular, $\mathcal{P}_{\text{similarity}}(s, \hat{s})$ returns the distance between s and $\hat{s}$ based on network activations, and results in an effective approximation of human perception. In more detail, the LPIPS metric is given by measuring the ℓ_2-distance between a normalized version of the activations of a neural network at several internal layers. For each layer l let W_l be the width, H_l the height, and C_l the number of channels. Further, let $y^l \in \mathbb{R}^{W_l \times H_l \times C_l}$ denote the vector of activations in convolutional layer l. To compute the perceptual similarity distance between two states s and $\hat{s}$, first calculate the channel-normalized internal activations $\hat{y}^l_s, \hat{y}^l_{\hat{s}} \in \mathbb{R}^{W_l \times H_l \times C_l}$ (corresponding to s and $\hat{s}$ respectively) for L internal layers, and scale each channel in $\hat{y}^l_s$ and $\hat{y}^l_{\hat{s}}$ by the same, fixed weight vector $w_l \in \mathbb{R}^{C_l}$. The last step is then to compute the perceptual similarity distance by first averaging the ℓ_2-distance between the scaled activations over the spatial dimensions, and then summing over the L layers.

3 Moving Through the Deep Neural Policy Manifold via High-Sensitivity Directions

To investigate the deep neural policy manifold we will probe the deep reinforcement learning decision boundary via both adversarial directions and directions innate to the state representations. While the adversarial directions are specifically optimized high-sensitivity directions in the deep neural policy landscape (i.e. worst-case distributional shift) within an *imperceptibility* bound as described in Section 2.3, the natural directions represent intrinsic semantic changes in the state representations within the imperceptibility distance.

Definition 3.1. Let π be a policy in an MDP $\mathcal{M}$ and let S be the set of states in $\mathcal{M}$. Let $\epsilon, \delta > 0$. An (ϵ, δ)-high-sensitivity direction function for π is a function $\xi(s, \pi)$ taking values in S such that $\mathcal{P}_{\text{similarity}}(s, s + \xi(s, \pi)) \leq \epsilon$ for all $s \in S$, and

$$\mathbb{E}_{a_t \sim \pi(s_t + \xi(s_t, \pi), \cdot)} \left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t, s_{t+1}) \right] < \delta \cdot \mathbb{E}_{a_t \sim \pi(s_t, \cdot)} \left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t, s_{t+1}) \right]$$

[1] See (Korkmaz 2021a) for inaccuracy and inconsistency of the state-action value function learnt by adversarially trained policies.

Algorithm 1: Probing Neural Manifold with High-sensitivity Directions within Perceptual Similarity

Input: Policy $\pi(s, a)$, high-sensitivity direction function $\xi(s, \pi)$, internal activations in convolutional layer $y^l \in \mathbb{R}^{W_l \times H_l \times C_l}$, parameters $\epsilon, \delta > 0$.
for $t = 0$ **to** T **do**
  $a_t = \arg\max_{a' \in A(s)} \pi(s_t + \xi(s_t, \pi), a')$
  Sample $s_{t+1} \sim \mathcal{T}(s_t, a_t, \cdot)$
  $\mathcal{P}_{\text{similarity}}(s, s + \xi(s, \pi)) =$
    $\sum_l \frac{1}{H_l W_l} \sum_{h,w} \| w_l \odot (\hat{y}^l_{shw} - \hat{y}^l_{(s+\xi(s,\pi))hw}) \|_2^2$
  $\mathcal{PS} += \mathcal{P}_{\text{similarity}}(s_t, s_t + \xi(s_t, \pi))$
  $\mathcal{R} += r(s_t, a_t)$
end for
Return: Total reward $\mathcal{R}$ and average perceptual similarity $\frac{\mathcal{PS}}{T}$.

Intuitively, $\xi(s, \pi)$ is a high-sensitivity direction function if translating by $\xi(s, \pi)$ in state s causes a significant drop in expected cumulative rewards when executing the policy π. Note that the function $\xi(s, \pi)$ in Definition 3.1 takes the policy π as input, and so is able to use information about the behavior of π in state s in order to compute the direction ξ. We next introduce a restricted version of Definition 3.1 where the function is not allowed to use any information about π.

Definition 3.2. Let S be the set of states for an MDP $\mathcal{M}$, let $\pi \in \Pi$ be a set of policies in $\mathcal{M}$, and let $\xi : S \to S$ be a function on S. Let $\epsilon, \delta > 0$. $\xi(s)$ is a fixed (ϵ, δ)-high-sensitivity direction function if the function $\phi(s, \pi) = \xi(s)$ is an (ϵ, δ)-high-sensitivity direction function for all $\pi \in \Pi$.

To probe the deep reinforcement learning policy landscape we will utilize policy dependent worst-case high-sensitivity directions (i.e. adversarial perturbations) as described in Definition 3.1 and policy-independent directions innate to the MDP as described in Definition 3.2. This probing methodology intrinsically juxtaposes adversarial directions and policy-independent directions with respect to their perceptual similarity distance (see Section 2.4) to the base states and their degree of impact on the policy performance. More importantly, we question the *imperceptibility* of ℓ_p-norm bounded adversarial directions in terms of perceptual similarity distance, and compare this *imperceptibility* notion to the policy-independent high-sensitivity directions intrinsic to the MDP. The fact that policy-independent high sensitivity directions innate to the MDP can achieve ultimately similar or higher drop in the expected cumulative rewards within the perceptual similarity distance brings the line of research focusing on adversarial directions into question. More importantly, the fact that policies trained to resist these adversarial directions and claimed to be "certified robust" are essentially less robust than simple vanilla trained deep reinforcement learning policies, as demonstrated in Section 4, brings the intrinsic trade-off made during training into question.

While it is possible to interpret the outcomes of contrasting worst-case policy dependent high sensitivity directions

ALE MDPs	BankHeist	JamesBond	Pong	Riverraid	TimePilot
C&W Impact	0.982±0.009	0.451±0.231	0.995±0.014	0.928±0.030	0.567 ±0.159
B&C Impact	0.966± 0.030	0.913 ±0.047	1.0±0.009	0.951 ±0.016	0.663±0.239
Blurred Observations Impact	0.979±0.009	0.635±0.200	1.0±0.000	0.946±0.015	0.589±0.150
Rotation Impact	0.997±0.004	0.635±0.189	0.99±0.015	0.942±0.042	0.581±0.158
Shifting Impact	0.985 ±0.005	0.865±0.140	1.0±0.00	0.935 ±0.023	0.623±0.199
DCT Artifacts Impact	0.980 ±0.013	0.884 ±0.128	0.962±0.032	0.803 ±0.051	0.578 ±0.271
PT Impact	0.998±0.003	0.865±0.087	0.996±0.009	0.968±0.006	0.624±0.198
C&W $\mathcal{P}_{\text{similarity}}$	0.0657±0.0073	0.2622±0.0312	0.6134±0.0271	0.2714±0.0285	0.1336± 0.0231
B&C $\mathcal{P}_{\text{similarity}}$	0.0307±0.0039	0.011± 0.0003	0.2190± 0.0046	0.2147±0.0212	0.1045± 0.0031
Blurred Observations $\mathcal{P}_{\text{similarity}}$	0.1672±0.0192	0.0707±0.0074	0.0351±0.0072	0.1442±0.0107	0.2014±0.0645
Rotation $\mathcal{P}_{\text{similarity}}$	0.0520±0.0070	0.0275±0.0016	0.1020±0.0115	0.0422± 0.0033	0.1020±0.0115
Shifting $\mathcal{P}_{\text{similarity}}$	0.0492±0.0046	0.0650±0.0092	0.2455±0.0432	0.0945±0.0032	0.1167±0.0121
DCT Artifacts $\mathcal{P}_{\text{similarity}}$	0.0240±0.0037	0.1325±0.0301	0.2506±0.0559	0.2250±0.0202	0.1592±0.0369
PT $\mathcal{P}_{\text{similarity}}$	0.0398±0.0067	0.012±0.0007	0.0140±0.0018	0.0422±0.0016	0.0440±0.0050
C&W Raw Scores	15.0±2.549	285.0±25.495	-20.8±0.189	1168.0± 140.696	4090.0±347.979
B&C Raw Scores	17.0±1.651	45.0±6.846	-21.0±0.000	744.0±76.957	3180.0±711.027
Blurred Observations Raw Scores	18.0±3.405	190.0±33.015	-21.0±0.000	820.0±72.013	3880.0±329.484
Rotation Raw Scores	2.0±1.264	190.0± 27.203	-20.6±0.209	873.0±201.866	3150.0±482.959
Shifting Raw Scores	13.0±1.449	70.0±20.248	-21.0±0.000	988.0± 89.057	3560.0± 437.538
DCT Artifacts Raw Scores	17.0±3.478	60.0±18.439	-19.4±0.428	2589.0±389.679	3980.0±593.936
PT Raw Scores	1.0±0.948	75.0±12.649	-20.9±0.126	486.0±29.127	3550.0±435.028
B&C $[\alpha, \beta]$	[1.2,40]	[0.9,20]	[1.7,40]	[2.4,-275]	[2.4,-260]
Blurring Kernel Size	5	3	3	5	5
Rotation Degree	1.4	1.6	3	1.8	5
Shifting $[t_i, t_j]$	[1,1]	[0,1]	[2,1]	[1,2]	[2,2]
PT Norm	1	1	3	2	3

Table 1: Impacts on the policy performance, perceptual similarity distances $\mathcal{P}_{\text{similarity}}$ to the base states, and raw scores for the Carlini and Wagner (2017) formulation and policy-independent high-sensitivity directions innate to the environment. We report all of the results with the standard error of the mean.

(i.e. adversarial) and policy-independent high-sensitivity directions as crucially surprising in terms of the security perspective[2], our goal is to provide an exact fundamental trade-off made by employing both adversarial attacks and training techniques. The fact that worst-case directions are heavily investigated in deep reinforcement learning research without clear cost and trade-off of these design choices essentially might create bias on influencing future research directions.

To probe the deep neural policy manifold via policy-independent high sensitivity directions we focus on intrinsic changes that are as simple as possible in the high dimensional state representation MDPs. We categorize these changes with respect to their frequency spectrum and below we explain precisely how these high sensitivity directions are computed.

Low Frequency Policy-Independent High-Sensitivity Directions: For the low frequency investigation we utilized brightness and contrast change in the state representations. We have kept movement along high-sensitivity direction as simple as possible as a linear transformation of the base state where $s(i, j)$ is the ij^{th} pixel of state s, and α and β are the linear brightness parameters

$$\hat{s}(i, j) = s(i, j) \cdot \alpha + \beta. \tag{3}$$

The perspective transform of state representations includes a mapping between four different source and destination pixels given by

$$\hat{s}(i,j) = s\left(\frac{\Gamma_{11}s_i + \Gamma_{12}s_j + \Gamma_{13}}{\Gamma_{31}s_i + \Gamma_{32}s_j + \Gamma_{33}}, \frac{\Gamma_{21}s_i + \Gamma_{22}s_j + \Gamma_{23}}{\Gamma_{31}s_i + \Gamma_{32}s_j + \Gamma_{33}}\right)$$

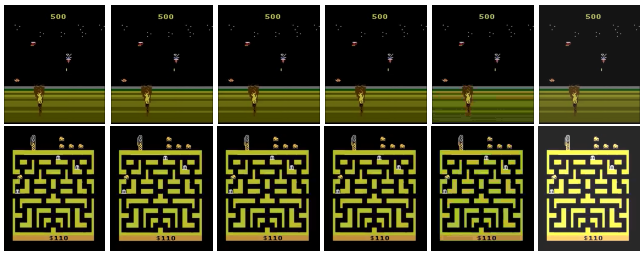
$$\delta_k \begin{bmatrix} s_i^{\text{dst}_k} \\ s_j^{\text{dst}_k} \\ 1 \end{bmatrix} = \Gamma \cdot \begin{bmatrix} s_i^{\text{src}_k} \\ s_j^{\text{src}_k} \\ 1 \end{bmatrix}. \tag{4}$$

The norm of a perspective transformation is defined as the maximum distance that one of the corners of the square moves under this mapping. Note that the perspective transformation has effects on both high and low frequencies as also portrayed in Section 5.

High Frequency Policy-Independent High-Sensitivity Directions: On the high frequency side we included compression artifacts caused by the discrete cosine transform resulting in the loss of high frequency components, also referred to as ringing and blocking artifacts. Another high sensitivity direction considered on the high frequency side of the spectrum is blurring[3]. In particular, median blurring

[2] In terms of the security perspective the research conducted on worst-case high-sensitivity (i.e. adversarial) directions in deep reinforcement learning relies heavily on a strong adversary assumption. In particular, this assumption refers to an adversary that has access to the policy's perception system, training details of the policy (e.g. algorithm, neural network architecture, training dataset), ability to alter observations in real time, simultaneous modifications to the observation system of the policy with computationally demanding adversarial formulations as described in Section 2.2 and in Section 2.3

[3] Note that in the blurring category one might use several different type of blurring techniques as Gaussian blurring, zoom blurring, defocus blur. Yet all these different types of techniques occupy the same frequency band in the Fourier domain.



Base State Shift PT Blur DCT B&C

Figure 1: Base frame and policy-independent high-sensitivity directions. Columns: base frame, shifting, perspective transformation, blurring, discrete cosine transform artifacts, brightness and contrast. Up: JamesBond. Down: BankHeist. The results for the rest of the MDPs in consideration are reported in the full version of the paper.

which is a nonlinear noise removal technique that replaces the base pixel value with the median pixel value of its neighbouring pixels. In this category kernel size k refers to the fact that the median is computed over a $k \times k$ neighborhood of the base pixel. One of the most fundamental geometric transformations leading to high frequency changes rotates the state observation around the centering pixel with corresponding rotation angle reported as degrees. Lastly, on the geometric transformations, shifting is included, which moves the input in the x or y direction with as few pixels moved as possible. This is denoted with $[t_i, t_j]$ as the distance shifted, where t_i is in the direction of x and t_j is in the direction of y.

Figure 1 demonstrates the visual interpretation of moving along these policy-independent high-sensitivity directions innate to the environment described above. While moving along these policy-independent directions is visually imperceptible, we also report exact perceptual similarity distances to the base states computed by Algorithm 1 in Table 1. In more detail, Table 1 shows the raw scores, corresponding performance drops, perceptual similarities to the base states and corresponding hyperparameters for the policy dependent (i.e. adversarial) and policy-independent high-sensitivity directions. Hence, the results in Table 1 demonstrate that the policy-independent high-sensitivity directions cause similar or higher degradation in the policy performance within similar perceptual similarity distance. To compute the results in Table 1, Algorithm 1 described in Section 3 is utilized.

4 Moving Along High-Sensitivity Directions in the Adversarially Trained Neural Manifold

In this section we investigate state-of-the-art adversarially trained deep reinforcement learning policies with policy-independent high-sensitivity directions described in Section 3. In particular, we test State Adversarial Double Deep Q-Network, a state-of-the-art algorithm (Huan et al. 2020). In this paper the authors propose using what they call a state-adversarial MDP to model adversarial attacks in deep reinforcement learning. Based on this model they develop methods to regularize Double Deep Q-Network policies to be certified robust to adversarial attacks. In more detail, letting $B(s)$ be the ℓ_p-norm ball of radius ϵ, this regularization is achieved by adding,

$$\mathcal{R}(\theta) = \max\{ \max_{\hat{s} \in B(s)} \max_{a \neq \arg\max_{a'} Q(s,a')} Q_\theta(\hat{s}, a) - Q_\theta(\hat{s}, \arg\max_{a'} Q(s, a'), -c\}.$$

to the temporal difference loss used in standard DQN. In particular, for a sample of the form (s, a, r, s') the loss is

$$\mathcal{L}(\theta) = L_{\mathcal{H}} \left(r + \gamma \max_{a'} Q^{\text{target}}(s', a') - Q_\theta(s, a) \right) + \mathcal{R}(\theta)$$

where $L_{\mathcal{H}}$ is the Huber loss. Furthermore, we also test the most recent adversarial training technique RADIAL (Oikarinen et al. 2021). In particular, the RADIAL method utilizes interval bound propagation (IBP) to compute upper and lower bounds on the Q-function under perturbations of norm ϵ. In particular, letting $Q^{\text{upper}}(s, a, \epsilon)$ and $Q^{\text{lower}}(s, a, \epsilon)$ be the respective upper and lower bounds on the Q-function when the state s is perturbed by ℓ_p-norm at most ϵ. For a given state s and action a, the RADIAL method utilizes the action-value difference given by

$$Q_{\text{diff}}(s, \hat{a}) = \max(0, Q(s, \hat{a}) - Q(s, a)).$$

The overlap is defined by

$$\mathcal{OV}(s, \hat{a}, \epsilon) = \max(0, Q^{\text{upper}}(s, \hat{a}, \epsilon) - Q^{\text{lower}}(s, a, \epsilon) + \frac{1}{2} Q_{\text{diff}}(s, \hat{a})).$$

The adversarial loss used in RADIAL is then given by the expectation over a minibatch of transitions

$$\mathcal{L}_{\text{adv}}(\theta, \epsilon) = \mathbb{E}_{s,a,s'} \left[\sum_{\hat{a} \in A} \mathcal{OV}(s, \hat{a}, \epsilon) \cdot Q_{\text{diff}}(s, \hat{a}) \right].$$

During training the adversarial loss $\mathcal{L}_{\text{adv}}(\theta, \epsilon)$ is added to the standard temporal difference loss. Note that both of these adversarial training algorithms SA-DDQN and RADIAL appeared in NeurIPS 2020 as a spotlight presentation and NeurIPS 2021 consecutively. Thus, it is of great and critical importance in the lines of AI-safety and in terms of affecting overall research progress and effort to outline both the limitations and the actual robustness capabilities of these algorithms.

Table 2 reports the impact values of the policy-independent high-sensitivity directions introduced to the vanilla trained deep reinforcement learning policies and the state-of-the-art adversarially trained deep reinforcement learning policies for both SA-DDQN and RADIAL. Note that the hyperparameters for Table 2 are identical to the hyperparameters in Table 1 for consistency. Thus, the results in Table 2 are not specifically optimized to affect adversarial training. However, Figure 2 reports the effect of varying the amount of movement along policy-independent non-robust directions, where α stands for contrast, β stands for brightness, and κ for the level of artifacts caused by the discrete cosine transform. Intriguingly, as these parameters for high-sensitivity directions are varied Figure 2 demonstrates that

Environment	BankHeist			Pong		
Training Method	SA-DDQN	RADIAL	Vanilla Trained	SA-DDQN	RADIAL	Vanilla Trained
B&C ($\mathcal{I}$)	0.881±0.010	0.959±0.002	0.971±0.030	1.0±0.000	1.0±0.000	0.996±0.009
Discrete Cosine Transform Artifacts ($\mathcal{I}$)	0.960±0.0014	1.0±0.000	0.984±0.013	1.0±0.000	1.0±0.000	0.962±0.032
Perspective Transform ($\mathcal{I}$)	1.0±0.000	1.0±0.000	1.0±0.003	0.992±0.0034	1.0±0.000	0.996±0.009
Blurred Observations ($\mathcal{I}$)	0.003±0.002	0.985±0.003	0.983±0.009	0.805±0.123	0.901±0.021	1.0±0.000
Rotation ($\mathcal{I}$)	1.0±0.000	0.992±0.000	1.0±0.004	1.0±0.000	1.0±0.000	0.99±0.015
Shifting ($\mathcal{I}$)	1.0±0.000	1.0±0.000	0.989±0.005	1.0±0.000	1.0±0.000	1.0±0.000

Table 2: The effects of moving along policy-independent high-sensitivity directions in state-of-the-art adversarially trained (SA-DDQN and RADIAL) and vanilla trained deep reinforcement learning policy manifolds.

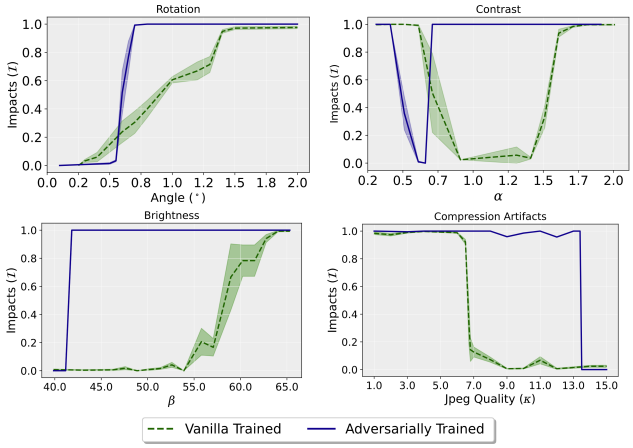


Figure 2: The performance drop results when moved along policy-independent high-sensitivity directions of the state-of-the-art adversarially trained deep reinforcement learning policy manifold and vanilla trained deep reinforcement learning policy manifold with varying degrees of discrete cosine transform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

simple vanilla trained deep reinforcement learning policies are more robust compared to state-of-the-art adversarially trained ones. For instance, modifying brightness with β in the range 3.1 to 20.0 causes impact close to 1.0 (i.e. total collapse of the policy) for the adversarially trained policy, but has negligible impact on the vanilla trained policy.

The results in Figure 2 demonstrate that, across a wide range of parameters, adversarially trained neural policies are less robust to natural directions innate to the MDP than vanilla trained policies. This occurs despite the fact that the central purpose of adversarial training is to increase robustness to imperceptible perturbations, where imperceptibility is measured by ℓ_p-norm. Our results indicate that an increase in robustness to ℓ_p-norm bounded perturbations can come at the cost of a loss in robustness to other natural types of imperceptible high-sensitivity directions. These results call into question the use of adversarial training for the creation of robust deep reinforcement learning policies, and in particular the use of ℓ_p-norm bounds as a metric of imperceptibility.

The fact that adversarial training fails to provide robustness has manifold implications. In particular, from the security point of view the effort put into making robust and reliable policies has been misdirected, resulting in policies that are in fact less robust than simple vanilla training. From the alignment perspective, while adversarial training is built to target and make policies safe against adversarial directions, it actually caused these policies to be misaligned with human perception. In terms of foundational understanding of the policies that are being built, our paper brings the term "robustness" into question. The decrease in resilience to overall distributional shift that "certified robust" adversarial training methods encounter demonstrates the need for further investigation into how robustness should be defined.

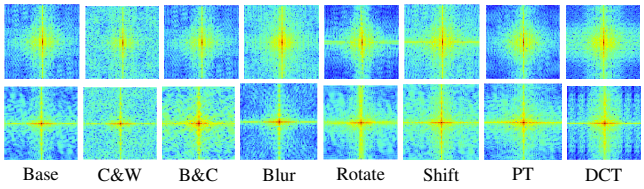


Figure 3: Up: $\mathcal{F}_s(u,v)$ for BankHeist. Down: $\mathcal{F}_s(u,v)$ for Riverraid. Columns: base state observation, the Carlini & Wagner formulation, brightness and contrast, blurred observations, rotation, shifting, perspective transformation, discrete cosine transform artifacts.

5 The Frequency Spectrum of the High-sensitivity Directions

In this section we provide frequency analysis of the policy dependent worst-case high-sensitivity directions and policy-independent high-sensitivity directions intrinsic to the high dimensional state representation MDP. The purpose of this analysis is to provide quantitative evidence that policy-independent high-sensitivity directions cover a broader portion of the spectrum, and thus provide a broader perspective on robustness than policy dependent adversarial directions alone. In particular, the results in Figure 4 and 3 demonstrates how each direction has distinctly different effects in the Fourier spectrum, both policy dependent and policy-independent. In more detail, the frequency spectrum is

$$\mathcal{F}_s(u,v) = \frac{1}{IJ}\sum_{i=0}^{I-1}\sum_{j=0}^{J-1}\hat{s}(i,j)e^{-j2\pi(ui/I+vj/J)} \quad (5)$$

where $\hat{s} = (s + \xi(s,\pi))$. Furthermore, we quantify these effects by measuring, for each type of high-sensitivity direction, the change in total Fourier energy at each spatial frequency level.

$$\mathcal{E}(f) = \sum_{\substack{u,v \\ \max\{u,v\}=f}} |\mathcal{F}_s(u,v)|^2 \quad (6)$$

In Figure 3 we show the Fourier spectrum of the base state s and the states moved towards high sensitivity directions

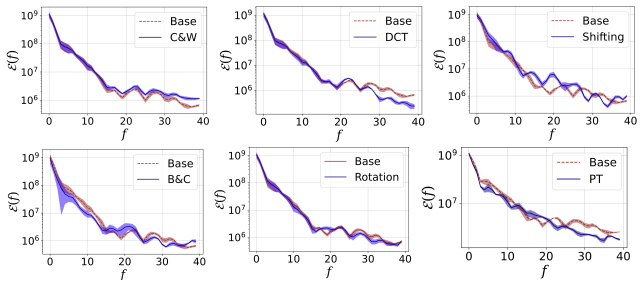


Figure 4: Total energy $\mathcal{E}(f)$ spectrum with various perturbations: worst-case directions (C&W), discrete cosine transform artifacts, perspective transformation, brightness and contrast, shifting, and rotation in RiverRaid.

from the base states $\hat{s}$ with both policy-independent adversarial directions (Carlini and Wagner 2017), and the high-sensitivity directions intrinsic to the MDP. In these spectrums the magnitude of the spatial frequencies increases by moving outward from the center, and the center of the image represents the Fourier basis function where spatial frequencies are zero. To investigate which type of high-sensitivity directions occupy which band in the Fourier domain we compute total energy $\mathcal{E}(f)$ for all basis functions whose maximum spatial frequency is f. Hence, Figure 4 shows the power spectral density of the base state compared to states that diverge from base states along the high-sensitivity direction computed via Algorithm 1 for both policy-independent high-sensitivity directions and policy dependent adversarial directions Carlini and Wagner (2017).

Aside from outlining our methodology, Section 5 serves the purpose of explaining results obtained in Section 4. In particular, training techniques (e.g. adversarial training) solely focusing on building robustness towards high spatial frequency corruptions become more vulnerable towards corruptions in a different band of the spectrum. Figure 4 demonstrates that each policy-independent high-sensitivity direction occupies a different particular band in the frequency domain. In more detail, while the policy dependent adversarial directions increase higher frequencies, the artifacts caused by discrete cosine transform decreases the magnitude of the high frequency band. Along this line both the linear transformation described in 3 and the geometric transformation described in 4 decreases the magnitude of the low frequency band. The fact that Figure 4 demonstrates that high-sensitivity directions indeed capture a broader set of directions in the frequency domain assists in providing a wider notion of robustness compared to solely relying on worst-case distributional shifts.

6 Experimental Details

In our experiments the vanilla trained deep neural policies are trained with Deep Q-Network with Double Q-learning proposed by (Hasselt, Guez, and Silver 2016) with prioritized experience replay (Schaul et al. 2016), and the adversarially trained deep neural policies are trained via the theoretically justified State-Adversarial MDP modelled State-Adversarial Double Deep Q-Network (SA-DDQN), and with RADIAL (see Section 4) with prioritized experience replay (Schaul et al. 2016) with the OpenAI Gym wrapper version (Brockman et al. 2016) of the Arcade Learning Environment (Bellemare et al. 2013). Note that all of the experiments are conducted in policies trained with high dimensional state representations. To be able to compare between different algorithms and different games the performance degradation of the deep reinforcement learning policy is defined as the normalized impact of an adversary on the agent:

$$\mathcal{I} = \frac{\text{Score}_{\text{clean}} - \text{Score}_{\text{adv}}}{\text{Score}_{\text{clean}} - \text{Score}_{\text{min}}^{\text{fixed}}}. \tag{7}$$

$\text{Score}_{\text{min}}^{\text{fixed}}$ is a fixed minimum score for a game, $\text{Score}_{\text{adv}}$ and $\text{Score}_{\text{clean}}$ are the scores of the agent with and without any modification to the agent's observations system respectively. All of the results reported in the paper are from 10 independent runs. In all of our tables and figures we include the means and the standard error of the mean values. More results on the issues discussed in Section 5 are provided in the full version of the paper with additional high-sensitivity analysis of policy gradient techniques, visualizations of the base states and shifts along the high-sensitivity directions intrinsic to the MDP.

7 Conclusion

In this paper we focused on probing the deep neural policy decision boundary via both policy dependent specifically optimized worst-case high-sensitivity directions and policy-independent high-sensitivity directions innate to the high dimensional state representation MDPs. We compared these worst-case adversarial directions computed via the-state-of-the art techniques with policy-independent ingrained directions in the Arcade Learning Environment (ALE). We questioned the *imperceptibility* notion of the ℓ_p-norm bounded adversarial directions, and demonstrated that the states with high-sensitivity directions inherent to the MDP are more perceptually similar to the base states compared to adversarial directions. Furthermore, we demonstrated that the fact that the policy-independent high-sensitivity directions achieve higher impact on policy performance with lower perceptual similarity distance without having access to the policy training details, real time access to the policy's memory and perception system, and computationally demanding adversarial formulations to compute simultaneous perturbations is evidence that high-sensitivity directions are naturally abundant in the deep reinforcement learning policy manifold. Most importantly, we show that state-of-the-art methods proposed to solve robustness problems in deep reinforcement learning are more fragile compared to vanilla trained deep neural policies. We argued for the significance of the interpretations of robustness in terms of the bias it creates in future research directions. Further, while we highlighted the importance of investigating the robustness of trained deep neural policies in a more diverse spectrum, we believe our study can provide a basis for understanding intriguing properties of the deep reinforcement learning decision boundary and can be instrumental in building more robust and generalizable deep neural policies.

References

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

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

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the robustness of neural networks. *In 2017 IEEE Symposium on Security and Privacy (SP)*, 39–57.

Gleave, A.; Dennis, M.; Wild, C.; Neel, K.; Levine, S.; and Russell, S. 2020. Adversarial Policies: Attacking Deep Reinforcement Learning. *International Conference on Learning Representations ICLR*.

Goodfellow, I.; Shelens, J.; and Szegedy, C. 2015. Explaning and Harnessing Adversarial Examples. *International Conference on Learning Representations*.

Hasselt, H. v.; Guez, A.; and Silver, D. 2016. Deep Reinforcement Learning with Double Q-Learning. *Association for the Advancement of Artificial Intelligence (AAAI)*.

Huan, Z.; Chen, H.; Xiao, C.; Li, B.; Liu, M.; Boning, D. S.; and Hseh, C. 2020. 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*.

Huang, S.; Papernot, N.; Goodfellow, Y., Ian an Duan; and Abbeel, P. 2017. Adversarial Attacks on Neural Network Policies. *Workshop Track of the 5th International Conference on Learning Representations*.

Iandola, F. N.; Han, S.; Moskewicz, M. W.; Ashraf, K.; J. Dally, W.; and Keutzer, K. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

Korkmaz, E. 2020. 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.*

Korkmaz, E. 2021a. 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*, 2323–2327.

Korkmaz, E. 2021b. Investigating Vulnerabilities of Deep Neural Policies. *Conference on Uncertainty in Artificial Intelligence (UAI)*.

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

Kos, J.; and Song, D. 2017. Delving Into Adversarial Attacks on Deep Policies. *International Conference on Learning Representations*.

Krizhevsky, A.; Sutskever, I.; and E. Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. 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.

Mandlekar, A.; Zhu, Y.; Garg, A.; Fei-Fei, L.; and Savarese, S. 2017. Adversarially Robust Policy Learning: Active Construction of Physically-Plausible Perturbations. *In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3932–3939.

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. 2015. Human-level control through deep reinforcement learning. *Nature*, 518: 529–533.

Oikarinen, T. P.; Zhang, W.; Megretski, A.; Daniel, L.; and Weng, T. 2021. Robust Deep Reinforcement Learning through Adversarial Loss. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 26156–26167.

Pattanaik, A.; Tang, Z.; Liu, S.; and Gautham, B. 2018. Robust Deep Reinforcement Learning with Adversarial Attacks. *In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2040–2042.

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A. 2017. Robust Adversarial Reinforcement Learning. *International Conference on Learning Representations ICLR*.

Schaul, T.; Quan, J.; Antonogloua, I.; and Silver, D. 2016. Prioritized Experience Replay. *International Conference on Learning Representations (ICLR)*.

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. 2020. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839): 604–609.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv:1707.06347v2 [cs.LG]*.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations, ICLR*.

Sun, J.; Zhang, T.; Xiafei, L.; Ma, X.; Zheng, Y.; Chen, K.; and Liu, Y. 2020. Stealthy and efficient adversarial attacks against deep reinforcement learning. *Association for the Advancement of Artificial Intelligence (AAAI)*.

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

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I. J.; Boneh, D.; and McDaniel, P. D. 2018. 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.

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; Oh, J.; Horgan, D.; Kroiss, M.; Danihelka, I.; Huang, A.; Sifre, L.; Cai, T.; Agapiou, J. P.; Jaderberg, M.; Vezhnevets, A. S.; Leblond, R.; Pohlen, T.; Dalibard, V.; Budden, D.; Sulsky, Y.; Molloy, J.; Paine, T. L.; Gülçehre, Ç.; Wang, Z.; Pfaff, T.; Wu, Y.; Ring, R.; Yogatama, D.; Wünsch, D.; McKinney, K.; Smith, O.; Schaul, T.; Lillicrap, T. P.; Kavukcuoglu, K.; Hassabis, D.; Apps, C.; and Silver, D. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354.

Xie, C.; and Yuille, A. L. 2020. 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.

Yen-Chen, L.; Zhag-Wei, H.; Lao, Y.-H.; Shih, M.-L.; ing Yu Lu; and Sun, M. 2017. Tactics of Adversarial Attack on Deep Reinforcement Learning Agents. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 3756–3762.

Zhang, R.; Isola, P.; Efros, A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *Conference on Computer Vision and Pattern Recognition (CVPR)*.