ASSESSING REINFORCEMENT LEARNING POLICIES VIA NATURAL CORRUPTIONS AT THE EDGE OF IMPERCEPTIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep reinforcement learning algorithms have recently achieved significant success in learning high-performing policies from purely visual observations. The ability to perform end-to-end learning from raw high dimensional input alone has led to deep reinforcement learning algorithms being deployed in a variety of fields. Thus, understanding and improving the ability of deep reinforcement learning policies to generalize to unseen data distributions is of critical importance. Much recent work has focused on assessing the generalization of deep reinforcement learning policies by introducing specifically crafted adversarial perturbations to their inputs. In this paper, we approach this problem from another perspective and propose a framework to assess the generalization skills of trained deep reinforcement learning policies. Rather than focusing on worst-case analysis of distribution shift, our approach is based on black-box perturbations that correspond to minimal semantically meaningful natural changes to the environment or the agent's visual observation system ranging from brightness to compression artifacts. We demonstrate that the perceptual similarity distance of the minimal natural perturbations is orders of magnitude smaller than the perceptual similarity distance of the adversarial perturbations to the unperturbed observations (i.e. minimal natural perturbations are perceptually more similar to the unperturbed states than the adversarial perturbations), while causing larger degradation in the policy performance. Furthermore, we investigate state-of-the-art adversarial training methods and show that adversarially trained deep reinforcement learning policies are more sensitive to almost all of the natural perturbations compared to vanilla trained policies. Lastly, we highlight that our framework captures a diverse set of bands in the Fourier spectrum; thus providing a better overall understanding of the policy's generalization capabilities. We believe our work can be crucial towards building resilient and generalizable deep reinforcement learning policies.

1 INTRODUCTION

Following the initial work of Mnih et al. (2015), the use of DNNs as function approximators in reinforcement learning has led to a dramatic increase in the capabilities of RL agents Schulman et al. (2017); Lillicrap et al. (2015). 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.

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 & 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 to specifically crafted imperceptible perturbations (Huang et al., 2017; Kos & Song, 2017; Pattanaik et al., 2018; Lin et al., 2017; Sun et al., 2020; Korkmaz, 2021). While one line of work put effort on exploring these vulnerabilities in deep neural policies, another line in parallel focused making them robust and reliable via adversarial training (Pinto et al., 2017; Mandlekar et al., 2017; Huan 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 the deep neural policies from a wider perspective, and propose a framework to test a more generic sense of robustness towards minimal perceptually similar perturbations[1]. To be able to achieve this we go beyond ℓ_p-norm bounded pixel perturbations and include semantically meaningful minimal realistic perturbations. By this approach we seek answers to the following questions: (i) How perceptually similar are minimal semantically meaningful perturbed states to the original unperturbed states, and how does this compare to ℓ_p-norm bounded adversarially perturbed states? (ii) What are the differences between adversarial perturbations and minimal natural perturbations introduced to the policy observation in terms of performance degradation of the trained deep reinforcement learning policy? (iii) How does state-of-the-art adversarial training affect the performance degradation caused by perceptually similar minimal natural perturbations compared to vanilla training? To be able answer these questions, in this work we focus on the notion of robustness of trained deep reinforcement learning agents and make the following contributions:

- We propose a framework consisting of a diverse set of minimalistic (i.e. perceptually similar) semantically meaningful natural perturbations.
- We run multiple experiments in the Arcade Learning Environment (ALE) in various games with high dimensional state representation and provide the relationship between the perceptual similarities to unperturbed states under our proposed natural perturbation framework and adversarial perturbations.
- We compare our proposed framework with the state-of-the-art adversarial method based on ℓ_p-norm changes, and we show that our natural perturbation framework is competitive in degrading the performance of the deep reinforcement learning agent with lower perceptual similarity distance.
- We inspect state-of-the-art adversarial training under our proposed framework, and demonstrate that the adversarially trained models become more vulnerable to various natural perturbations compared to vanilla trained models.
- Finally, we investigate the frequency domain of our framework and state-of-the-art targeted attacks. We show that our framework captures different bands of the frequency spectrum, thus yielding a better estimate of the model robustness.

2 Background and Related Work

2.1 Preliminaries

In this paper we consider Markov Decision Processes (MDPs) given by a tuple $(S, A, P, r, \gamma, s_0)$. The reinforcement learning agent interacts with the MDP by observing states $s \in S$, and then taking actions $a \in A$. Here s_0 represents the initial state of the agent, and γ 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 $P : S \times A \times S \rightarrow \mathbb{R}$. The reward received by the agent when taking action a in state s is given by the reward function $r : S \times A \rightarrow \mathbb{R}$. The goal of the agent is to learn a policy $\pi_\theta : S \times A \rightarrow \mathbb{R}$ which takes an action a in state s that maximizes the expected cumulative discounted reward $\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t)$ that the agent receives via interacting with the environment. This goal is achieved via learning the state-action value function $Q(s, a) = \mathbb{E}_\pi[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t) | s = s_0, a_0 = 0]$ assigning a value to each state-action pair. We use $\mathcal{F}(s)$ to denote the 2D discrete Fourier transform of state s in which each frequency is computed via

[1]Perceptual similarity is explained in detail in Section 2.4

$\mathcal{F}(m,n) = \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} s(k,l) e^{-j2\pi(mk/M+nl/N)}$ where k and l are the coordinates of the state, and M and N correspond to ranges of the 2D state representation.

2.2 Crafting Adversarial Perturbations

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

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

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 for image classification. Carlini & Wagner (2017) introduced targeted attacks in the image classification domain based on distance minimization between the adversarial image and the original image while targeting a particular label. In the deep reinforcement learning domain the Carlini & Wagner (2017) formulation is

$$\min_{s_{\text{adv}} \in D_{\epsilon,p}(s)} \|s_{\text{adv}} - s\|_p$$
$$\text{subject to } \arg\max_a Q(s,a) \neq \arg\max_a Q(s_{\text{adv}}, a)$$

where s is the unperturbed input, s_{adv} is the adversarially perturbed input, $a^*(s)$ is the action taken in the unperturbed state, and $a^*(s_{\text{adv}}) = \arg\max_a Q(s_{\text{adv}}, a)$ is the action taken in the adversarial state. This formulation attempts to minimize the distance to the original state, constrained to states leading to sub-optimal actions as determined by the Q-network. 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 & Song (2017) adapted FGSM from image classification to the deep reinforcement learning setting. Subsequently, Mandlekar et al. (2017) used FGSM perturbations for adversarial training of deep reinforcement learning agents. Pinto et al. (2017); Gleave et al. (2020) focused on modeling the interaction between the adversary and the agent, while Lin 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 (2017) adversarial formulation. Quite recently, Huan et al. (2020) proposed to model this dynamic as a State-Adversarial Markov Decision Process (SA-MDP), and the authors claimed the SA-MDP model provides theoretically justified robust deep reinforcement learning agents.

2.4 Perceptual Similarity Distance

Zhang et al. (2018) found that internal activations of networks trained for high-level tasks correspond to human perceptual judgements across different network architectures Iandola et al. (2016), Krizhevsky et al. (2012), Simonyan & Zisserman (2015) without calibration. Furthermore, the authors propose a method to measure the perceptual distance between two images with the Learned Perceptual Image Patch Similarity (LPIPS) metric. We compare the distance between adversarial states s_{adv} and the original states s with the LPIPS metric. We refer to the LPIPS metric as $\mathcal{P}_{\text{similarity}}$ throughout the paper. $\mathcal{P}_{\text{similarity}}(s, s_{\text{adv}})$ returns the distance between s and s_{adv} based on network activations. Zhang et al. (2018) show that $\mathcal{P}_{\text{similarity}}$ results in a reliable approximation of human perception.

In more detail, the LPIPS metric in Zhang et al. (2018) is given by measuring the ℓ_2 distance between a normalized version of the activations of the neural network at several internal convolutional layers. For each convolutional 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 s_0, first calculate the internal activations $y^l, y_0^l \in \mathbb{R}^{W_l \times H_l \times C_l}$ (corresponding to s and s_0 respectively) for L internal layers. Second, unit-normalize the activation vectors in the channel dimension, and denote the resulting normalized activations by $\hat{y}^l$ and $\hat{y}_0^l$. Next scale each channel in $\hat{y}^l$ and $\hat{y}_0^l$ by the same, fixed weight vector $w_l \in \mathbb{R}^{C_l}$. Here w_l can either be a learned vector of weights for layer l or the vector of all ones if no scaling is desired. 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. Formally,

$$\mathcal{P}_{\text{similarity}}(s, s_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0hw}^l) \right\|_2^2 \tag{2}$$

2.5 IMPACT

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:

$$\mathcal{I} = \frac{\text{Score}_{\text{clean}} - \text{Score}_{\text{adv}}}{\text{Score}_{\text{clean}} - \text{Score}_{\text{min}}^{\text{fixed}}}. \tag{3}$$

$\text{Score}_{\text{min}}^{\text{fixed}}$ 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.

3 A GENERALIZATION TESTING FRAMEWORK WITH NATURAL PERTURBATIONS

In our paper we focus on robustness and generalization issues that deep reinforcement learning policies encounter with a contrasting view compared to prior work focusing on worst-case distributional shift within an *imperceptibility* bound (see Section 2.3). We propose a baseline to evaluate deep reinforcement learning policies with realistic and minimal corruptions to the environment with which they interact. We essentially juxtapose adversarial perturbations and natural corruptions with respect to their perceptual similarity distance (see Section 2.4) to the original states and their degree of impact on the policy performance. More importantly, we question the *imperceptibility* of ℓ_p-norm bounded adversarial perturbations in terms of perceptual similarity distance, and compare this *imperceptibility* notion to natural perturbations. While we categorize adversarial perturbations also as a component in the framework majorly concentrated on the high frequencies, we embed several realistic perturbations that aim to cover diverse bands in the frequency spectrum. We highlight that prior work focused on the presence of a strong adversary model that requires prior access to training details of the agent's neural network Huang et al. (2017); Korkmaz (2021), real time access to the agent's perception system Pattanaik et al. (2018); Kos & Song (2017), and highly computationally demanding adversarial formulations for computing simultaneous perturbations Lin et al. (2017); Sun et al. (2020). From the security point of view we emphasize that natural corruptions at the edge of imperceptibility can be more dangerous than a strong adversary assumption [2] without carrying any of these requirements.

In our model we examine several natural environmental changes such as: changes in the brightness of the environment, blurring of the observation, slight rotation of the observation, several geometric transformations and compression artifacts. These changes from our model can be easily linked to naturally occurring changes in the environment[3]. In Table 1 we compare our proposed framework with the state-of-the-art targeted adversarial attack proposed by Carlini & Wagner (2017) in terms of

[2]Strong adversary assumption refers to an adversary that has access to the agent's observation 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.

[3]These natural changes can be linked to the time of day for brightness in a self driving vehicle, or the appearance of reflective objects or shadows. Rotation, perspective transformation, and shifting can be linked to driving on a road with varied terrain. Blurring can be linked to a rainy day, foggy weather or a fogged up camera lens utilized by the agent.

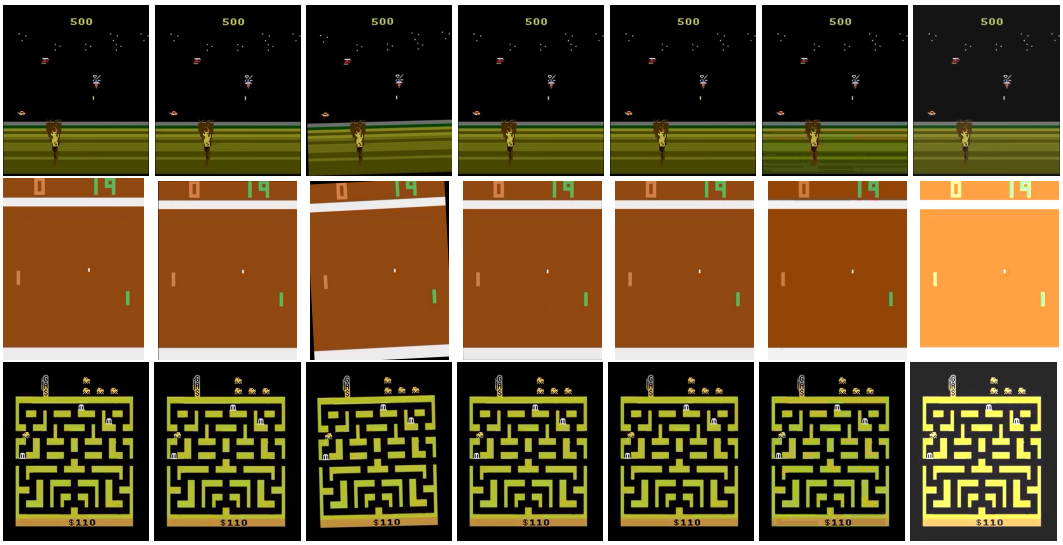

Figure 1: Original frame and environmental modifications. Columns: original frame, shifting, rotation, perspective transformation, blurring, compression artifacts. brightness and contrast. Rows: JamesBond, Pong and BankHeist.

perceptual similarity distances, and the impacts on the policy performance. While in the remainder of this section we explain in detail each component of our proposed framework, we provide all the experimental details in Section 5, as well as results on policy gradients, performance degradation in the time domain, and complementary results for Section 5 in the appendix.

Note that the natural corruptions considered in our framework are as minimalistic as possible. Most of the perturbations from the proposed framework cannot be recognized by human perception (see Figure 1). More formally, the perceptual similarity distances for each corruption, and the resulting policy performance degradation, are given in Table 1.

Brightness and Contrast: To inspect the effects of low frequency corruptions we included brightness and contrast level changes using linear brightness and contrast transformation,

$$s_{\text{adv}}(i,j) = s(i,j) \cdot \alpha + \beta, \tag{4}$$

where $s(i,j)$ is the ij^{th} pixel of state s, and α and β are the linear brightness parameters. In Table 1 we show the impacts and perceptual similarity distances with corresponding α, β values. In all of the games except BankHeist brightness and contrast change results in higher impact than the Carlini & Wagner (2017) formulation, while the perceptual similarity distance of brightness and contrast is lower in every game.

Blurring: To observe the effects of high frequency corruptions we included blurring in our framework. In particular, *median bluring*[4] is a nonlinear noise removal technique that replaces the original pixel value with the median pixel value of its neighbouring pixels. A kernel size k means that the median is computed over a $k \times k$ neighborhood of the original pixel. Only in BankHeist and TimePilot we observe that the perceptual similarity distance required for blurring is higher compared to adversarial perturbations to be able to cause higher impact on policy performance (see Table 1). For the rest of the games impact is higher and perceptual similarity distance is lower for blurring.

Rotation: Rotation is one of the most fundamental geometric changes in an environment which we incorporate in our framework. In Table 1 we show impact values and perceptual similarity distance with corresponding rotation angle in degrees. In all of the games except Pong rotation results in higher impact and orders of magnitude lower perceptual similarity distance than the Carlini

[4]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 blurring techniques occupy the same frequency band in the Fourier domain.

Table 1: Impacts on the policy performance, perceptual similarity distances $\mathcal{P}_{\text{similarity}}$ to the unperturbed states, and raw scores for Carlini & Wagner (2017) formulation and natural perturbation framework components. We report all of the results with the standard error of the mean.

Games	BankHeist	JamesBond	Pong	Riverraid	TimePilot
Carlini & Wagner Impact	0.982±0.009	0.451±0.231	0.995±0.014	0.928±0.030	0.567 ±0.159
Brightness&Contrast Impact	0.966± 0.030	0.913 ±0.047	1.0±0.009	0.951 ±0.016	0.663±0.239
Blurring 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
Compression Artifacts Impact	0.980 ±0.013	0.884 ±0.128	0.962±0.032	0.803 ±0.051	0.578 ±0.271
Perspective Transform Impact	0.998±0.003	0.865±0.087	0.996±0.009	0.968±0.006	0.624±0.198
Carlini & Wagner $\mathcal{P}_{\text{similarity}}$	0.0657±0.0073	0.2622±0.0312	0.6134±0.0271	0.2714±0.0285	0.1336± 0.0231
Brightness&Contrast $\mathcal{P}_{\text{similarity}}$	0.0307±0.0039	0.011± 0.0003	0.2190± 0.0046	0.2147±0.0212	0.1045± 0.0031
Blurring $\mathcal{P}_{\text{similarity}}$	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
Compression 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
Perspective Transform $\mathcal{P}_{\text{similarity}}$	0.0398±0.0067	0.012±0.0007	0.0140±0.0018	0.0422±0.0016	0.0440±0.0050
Carlini& Wagner Raw Scores	15.0±2.549	285.0±25.495	-20.8±0.189	1168.0± 140.696	4090.0±347.979
Brightness&Contrast Raw Scores	17.0±1.651	45.0±6.846	-21.0±0.000	744.0±76.957	3180.0±711.027
Blurring 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
Compression Artifacts Raw Scores	17.0±3.478	60.0±18.439	-19.4±0.428	2589.0±389.679	3980.0±593.936
Perspective Transform Raw Scores	1.0±0.948	75.0±12.649	-20.9±0.126	486.0±29.127	3550.0±435.028
Brightness&Contrast $[\alpha, \beta]$	[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]
Perspective Transform Norm	1	1	3	2	3

& Wagner (2017) formulation. In Pong the impact is comparable and the perceptual similarity distance is lower by a factor of 6.

Shifting: We included several plausible geometric transformations in our natural perturbation framework, the first of which is shifting. In more detail, shifting an image moves the elements of the image matrix along any dimension by any number of elements. For this modification we shift the inputs in the x or y direction with as few pixels shifted as possible. We use $[t_i, t_j]$ to denote the distance shifted, where t_i is in the direction of x and t_j is in the direction of y. In Table 1 we show the impact values and perceptual similarity distances for both Carlini & Wagner (2017) and shifting with corresponding $[t_i, t_j]$ values. For all of the games shifting yields higher impact and lower perceptual similarity distance.

Compression Artifacts: With this natural perturbation component we look at JPEG compression artifacts caused by the discrete cosine transform (DCT) resulting in the loss of high frequency components (ringing and blocking). In Table 1 we show the impact values and perceptual similarities of Carlini & Wagner (2017) and compression artifacts (CA). Only in Pong and Riverraid do we observe a lower impact than Carlini & Wagner (2017) while the perceptual similarity distance is significantly smaller for compression artifacts. While in TimePilot the perceptual similarity distance is higher, in the rest of the games compression artifacts result in higher impact and lower perceptual similarity distance compared to Carlini & Wagner (2017).

Perspective Transformation: The final component of our proposed natural perturbation framework is perspective transformation. Given four points in the plane defining a convex quadrangle, there is a unique perspective transformation mapping the corners of the square to these four points[5]. We define the norm of a perspective transformation as the maximum distance that one of the corners of the square moves under this mapping. Note that for most of the games the perspective norm is small (see Table 1). Hence, the changes caused by the perspective transform are *imperceptible* (e.g. fourth column of Figure 1). Furthermore, for all the games we observe perspective transformation yields higher impact and lower perceptual similarity distance than the Carlini & Wagner (2017) formulation.

[5]See Supplementary Material for details

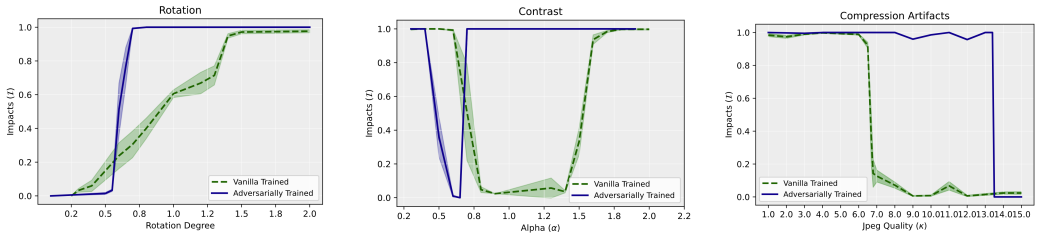

Figure 2: Performance drop of adversarially trained deep reinforcement learning policy and vanilla trained deep reinforcement learning policy under the changes in rotation, compression artifacts, and contrast.

4 Adversarial Training under Natural Corruptions

In this section we investigate state-of-the-art adversarially trained deep reinforcement learning policies within our proposed natural perturbation framework. In particular, we test State Adversarial Double Deep Q-Network, a state-of-the-art algorithm (see Section 2.3). 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 more 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,

$$\mathcal{R}(\theta) = \max\{\max_{\hat{s}\in B(s)} \max_{a\neq a^*(s)} Q_\theta(\hat{s}, a) - Q_\theta(\hat{s}, a^*(s)), -c\}. \tag{5}$$

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

$$\mathcal{L}(\theta) = L_H\left(r + \gamma \max_{a'} Q^{\text{target}}(s', a') - Q_\theta(s, a)\right) + \mathcal{R}(\theta) \tag{6}$$

where L_H is the Huber loss.

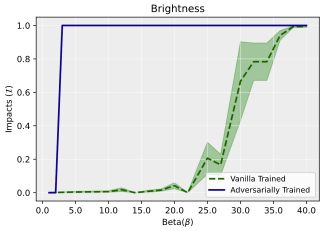

Figure 3: Performance drop of adversarially trained model and vanilla trained model to the changes in brightness.

Table 2 shows the impact values of the components of our proposed framework for the vanilla trained agent and the adversarially trained agent. We find that while the adversarially trained model gains robustness against blurring, no additional robustness is gained against any other component of the framework under adversarial training. Furthermore, in Figure 2 and Figure 3 we show the effect of varying the degrees for rotation, α for contrast, β for brightness, and jpeg quality κ for compression artifacts. We find that, as these parameters are varied, the vanilla trained agent is more robust than the adversarially trained one. For example, modifying brightness with β in the range 3.1 to 20.0 causes impact close to 1.0 (i.e. total failure) for the adversarially trained policy, but has negligible impact on the vanilla trained policy. Thus, not only does our proposed framework capture semantically meaningful perturbations that are not captured by adversarial robustness, but additionally adversarial training actively harms robustness to some of the natural perturbations from our proposed framework.

The results in Figure 2 and Figure 3 demonstrate that, across a wide range of parameters, adversarially trained neural policies are less robust to natural perturbations 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 corruptions. 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 deep reinforcement learning policies are being widely deployed in many different domains: self-driving automobiles Dosovitsky et al. (2017); Wolf et al. (2017), drug design Pereira

Table 2: Impacts of adversarially and vanilla trained policies with natural perturbation framework: brightness &contrast, blurring, rotation, shifting, compression artifacts and perspective transform.

Environment Training Method	BankHeist Adversarially Trained	BankHeist Vanilla Trained	Pong Vanilla Trained	Pong Adversarially Trained
Brightness&Contrast ($\mathcal{I}$)	0.881±0.010	0.971±0.030	0.996±0.009	1.0±0.000
Compression Artifacts ($\mathcal{I}$)	0.960±0.0014	0.984±0.013	0.962±0.032	1.0±0.000
Perspective Transform ($\mathcal{I}$)	1.0±0.000	1.0±0.003	0.996±0.009	0.992±0.0034
Blurring ($\mathcal{I}$)	0.003±0.002	0.983±0.009	1.0±0.000	0.805±0.123
Rotation ($\mathcal{I}$)	1.0±0.000	1.0±0.004	0.99±0.015	1.0±0.000
Shifting ($\mathcal{I}$)	1.0±0.000	0.989±0.005	1.0±0.000	1.0±0.000

et al. (2021); Popova et al. (2018), autonomous aerial vehicles Zhang et al. (2020), medical diagnosis and treatment Thananjeyan et al. (2017); Yauney & Pratik (2018), natural language processing He et al. (2016); Jaques et al. (2017), and industrial control and security Wang et al. (2019); Duan et al. (2020), 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.

5 Perturbations in the Fourier Domain

In this section we provide frequency analysis of our proposed framework and state-of-the-art adversarial formulations. The purpose of this analysis is to provide quantitative evidence that natural perturbations cover a broader concept of robustness than adversarial perturbations alone. In particular, we demonstrate that each natural perturbation has distinctly different effects in the Fourier spectrum, both from other natural corruptions and from adversarial perturbations. Furthermore, we quantify these effects by measuring, for each type of perturbation, the change in total Fourier energy at each spatial frequency level. 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 different band of the spectrum.

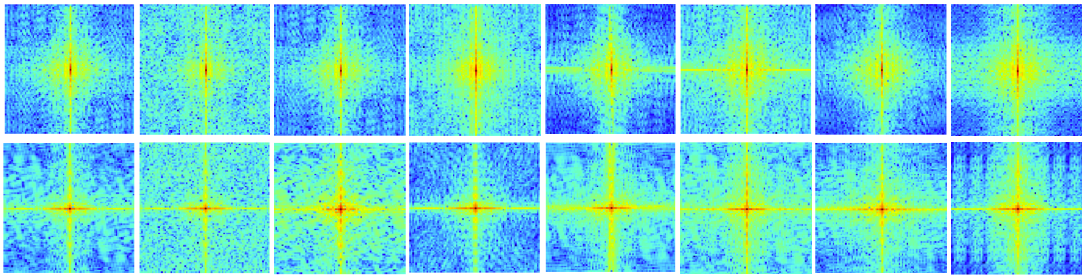

Figure 4: Rows: $\mathcal{F}(s)$ for BankHeist, $\mathcal{F}(s)$ for Riverraid. Columns: unperturbed state, Carlini & Wagner, brightness and contrast, blurring, rotation, shifting, perspective transformation, compression artifacts.

In Figure 4 we show the Fourier spectrum of the original state s and the perturbed states s_{adv} from our proposed framework based on natural perturbations and Carlini & Wagner (2017) formulated perturbations. 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. We provide more detailed description of $\mathcal{F}(s)$ in Section 2.1. To investigate which type of perturbations 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 5 shows the power spectral density of the original state compared to perturbed states computed via components from our proposed natural perturbation framework and Carlini & Wagner (2017). Figure 5 demonstrates that each component from our natural perturbation framework occupies different bands in the Fourier domain. In particular, in Figure 5 we observe that while the Carlini & Wagner (2017) formulation increases the magnitude of the higher frequencies, compression artifacts decrease the magnitude of

the high frequency band. On the other hand, brightness and contrast decreases the magnitude of the low frequency band, and shifting increases the mid-band. Blurring decreases the mid-band and high frequencies together, and perspective transformation decreases the low frequencies and high frequencies while increasing the mid-band.

Figure 5 shows that our proposed framework indeed captures a broader set of directions in the frequency domain. Thus, capturing the susceptibilities towards perturbations in different bands of the frequency domain represents a wider notion of robustness compared to only focusing on worst-case distributional shifts.

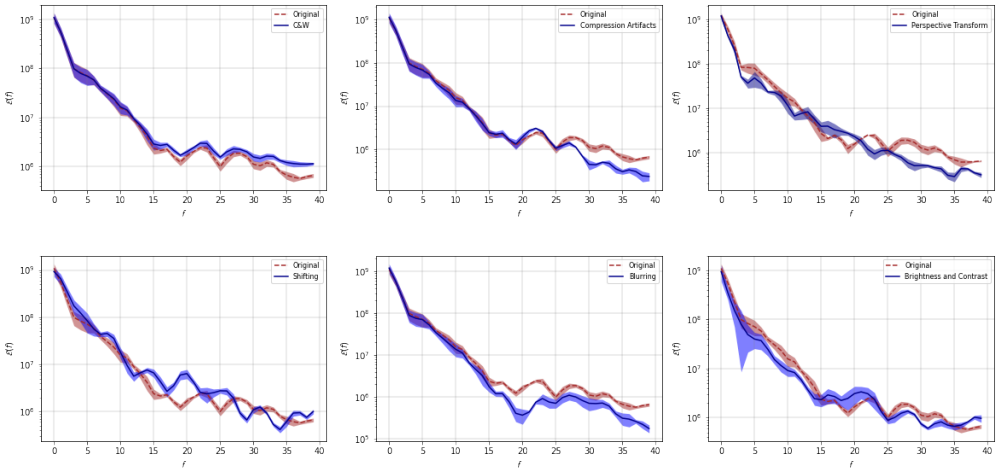

Figure 5: Riverraid total energy $\mathcal{E}(f)$ spectrum with various perturbations: Carlini & Wagner, compression artifacts, brightness and contrast, perspective transformation, shifting, rotation.

Experimental Details: In our experiments the vanilla trained deep neural policies are trained with Double Deep Q- Network Wang et al. (2016) and the adversarially trained deep neural policy is trained via the theoretically justified State-Adversarial MDP modelled State-Adversarial Double Deep Q-Network (SA-DDQN) (see Section 2.3) in the OpenAI Gym Brockman et al. (2016) Arcade Learning Environment Bellemare et al. (2013). We evaluate several trained policies from Arcade Learning Environment with our proposed framework averaged over 10 episodes. In all of our tables and figures we include the means and the standard error of the mean values. See more details in the appendix.

6 Conclusion

In this paper we studied a realistic threat model based on basic environmental changes and proposed a framework to asses the generalization capabilities of deep reinforcement learning policies. We compared our natural perturbation framework with the state-of-the-art adversarial attacks in the Arcade Learning Environment (ALE). We questioned the *imperceptibility* notion of the ℓ_p-norm bounded adversarial perturbations, and demonstrated that the states with minimal natural perturbations are more perceptually similar to the unperturbed states compared to adversarial ones. Moreover, we demonstrated that our framework achieves higher impact on policy performance with lower perceptual similarity distance without having access to the policy training details, real time access to the agent's memory and perception system, and computationally demanding adversarial formulations to compute simultaneous perturbations. Furthermore, we showed that each component of our framework contains distinct bands in the frequency domain, resulting in a better estimate of the generalization capabilities of trained agents. Most importantly, we investigated state-of-the-art adversarial training methods and found that vanilla trained policies are more robust than adversarially trained policies to minimal natural perturbations. We think that the robustness of the trained deep neural policies should be investigated in a more diverse spectrum and we believe our framework can be instrumental towards generalization and robustification of deep reinforcement learning algorithms.

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.

Alexey Dosovitsky, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. *In Proceedings of the Conference on Robot Learning (CoRL)*, 78: 1–16, 2017.

Jiajun Duan, Di Shi, Ruisheng Diao, Haifeng Li, Zhiwei Wang, Bei Zhang, Desong Bian, and Zhehan Yi. Deep-reinforcement-learning-based autonomous voltage control for power grid operations. *IEEE Transactions on Power Systems*, 35(1):814–817, 2020.

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.

Ji He, Jianshu Chen, Xiaodong He, Jianfeng Gao, Lihong Li, Li Deng, and Mari. Ostendorf. Deep reinforcement learning with a natural language action space. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1621–1630, 2016.

Zhang Huan, Hongge Chen, Chaowe Xiao, Bo Li, Mingyan Liu, Duane S. Boning, and ChoJui Hseh. Robust deep renforcement learning against adversaral perturbations on state observations. In Hugo Larochelle, Marc'Aurelo Ranzato, Raia Hadsell, Maria-Florna 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.

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.

Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E. Turner, and Douglas Eck. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. *In Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1645–1654, 2017.

Ezgi Korkmaz. Investigating vulnerabilities of deep neural policies. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021.

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.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan. Wierstra. Continuous control with deep reinforcement learning. *arXivpreprint arXiv:1509.02971*, 2015.

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.

Tiago Pereira, Maryam Abbasi, Bernardete Ribeiro, and Joel P. Arrais. Diversity oriented deep reinforcement learning for targeted molecule generation. *J. Cheminformatics*, 13(1): 21, 2021. doi: 10.1186/s13321-021-00498-z. URL `https://doi.org/10.1186/s13321-021-00498-z`.

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

Mariya Popova, Olexandr Isayev, and Alexander. Tropsha. Deep reinforcement learning for de novo drug design. *Science advances 4*, 78, 2018.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg. Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347v2 [cs.LG].*, 2017.

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

Jianwen Sun, Tianwei Zhang, Lei Xiaofei, Xie Ma, Yan Zheng, Kangjie Chen, and Yang. Liu. Stealthy and efficient adversarial attacks against deep reinforcement learning. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

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.

Brijen Thananjeyan, Animesh Garg, Sanjay Krishnan, Carolyn Chen, Lauren Miller, and Ken. Goldberg. Multilateral surgical pattern cutting in 2d orthotropic gauze with deep reinforcement learning policies for tensioning. *In 2017 IEEE International Conference on Robotics and Automation (ICRA).*, pp. 2371–2378, 2017.

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.

Yuandou Wang, Hang Liu, Wanbo Zheng, Yunni Xia, Yawen Li, Peng Chen, Kunyin Guo, , and Hong Xie. Multi-objective workflow scheduling with deep-q-network-based multi-agent reinforcement learning. *IEEE Access*, pp. 39974–39982, 2019.

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.

Peter Wolf, Christian Hubschneider, Michael Weber, André Bauer, Jonathan Härtl, Fabian Dürr, and Marius Zöllner J. Learning how to drive in a real world simulation with deep q-networks. *IEEE Intelligent Vehicles Symposium (IV)*, pp. 244–250, 2017.

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.

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

Ning Zhang, Juan Liu, Lingfu Xie, and Peng Tong. A deep reinforcement learning approach to energy-harvesting uav-aided data collection. In *2020 International Conference on Wireless Communications and Signal Processing (WCSP), Nanjing, China, October 21-23, 2020*, pp. 93–98. IEEE, 2020.

Richard Zhang, Phillip Isola, Alexei Efros, Eli Shechtman, and Oliver. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.