Structured Interpretable Manipulation of Policies

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Quinlan Dawkins

Spring, 2020.

Technical Project Team Members

Jihyeong Lee

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Hongning Wang, Department of Computer Science

STRUCTURED INTERPRETABLE MANIPULATION OF POLICIES

Quinlan Dawkins & Jihyeong Lee
Department of Computer Science
University of Virginia
Charlottesville, VA, 22903
{qed4wg,jl4wq}@virginia.edu

ABSTRACT

This paper explores the interpretability of adversarial attacks in reinforcement learning (RL) environments. A key feature of adversarial attacks pertains to the limited magnitude of an attack being able to have a significant impact on the output of the target. This is closely tied to model interpretability which involves understanding model outputs and why such a small change in the input is able to completely change the output of a DNN. This naturally leads to a question of interpretability of adversarial attacks. In this research, we explore a technique used in a classification setting for improving group sparsity of adversarial attacks in an RL setting. Additionally we adapt methods used to measure attack interpretability in a classification setting to the RL setting where the objective is to maximize a cumulative reward.

1 INTRODUCTION

Deep neural networks (DNNs) have been widely adopted for solving various machine learning tasks due to their ability to approximate complex functions. In particular, DNNs have been gaining traction in reinforcement learning, thanks to their ability to approximate the equations controlling the reward and actions of an agent interacting within a dynamic environment. Deep Reinforcement Learning (DRL) methods such as the DQN (Mnih et al., 2013) have primarily focused on modeling the approximation of rewards for each action, and have seen success in settings ranging from playing classic arcade video games to optimal control.

Despite their success, previous work has shown that DNNs are vulnerable to adversarial attacks. Attacks, such as those introduced by Carlini & Wagner (2017), have focused on crafting small, human-imperceptible perturbations on inputs that lower the classification accuracy of traditional supervised learning tasks. Since DRL methods rely on the same pattern recognition capabilities used in traditional supervised deep neural networks, they have also been shown to be vulnerable to these types of attacks (Goodfellow et al., 2014). One difference in the RL setting arises from the dynamic interaction between the agent and the environment, since the agent's action could dynamically alter the environment and have compounding consequences across time. This gives attackers another dimension to potentially leverage in creating these attack. One such attack was the Strategically Timed Attack (Lin et al., 2017), where the authors have crafted traditional perturbations that would only be applied on an RL agent at times that would have the greatest effect.

However, the seemingly random and small perturbations of adversarial attacks makes them difficult to interpret. Traditional attacks on supervised learning, such as the Carlini and Wagner attack (Carlini & Wagner, 2017), have relied on constraining the l_p norm of the perturbation on an entire image, most traditionally the l_∞ norm. Recent work by Xu et. al. introduced the Structured Attack (StrAttack) (Xu et al., 2019), which focused on improving the interpretability of these attacks by extracting structural information from the inputs, limiting perturbations to smaller regions rather than an entire image.

Our work aims to extend the concept of interpretablity to the reinforcement learning setting. We propose the Structured Interpretable Manipulation on Policies, which leverages the "greatest timing

effect" of the Strategically Timed Attack (Lin et al., 2017) and the "structured interpretability effect" of the StrAttack (Xu et al., 2019).

2 Related Works

2.1 Strategically Timed Adversarial Attack

Initial work on attacking DRL networks have focused on applying adversarial perturbations on the input at every time step. The Strategically Timed Attack (Lin et al., 2017) aims to limit the amount of perturbation that can be applied across time by posing the scenario as an optimization problem with the goal of minimizing the target agent's total reward with a limited number of perturbations allowed. Due to the computational complexity of solving this optimization problem, the authors instead use a heuristic that measures the difference between the least preferred and most preferred action of the agent. The adversary attacks when the heuristic falls above a given threshold, which forces the adversary to attack when an action is most critical, while reducing the number of total attacks performed by the agent. This method was found to indeed perform much fewer attacks without any significant drop in the adversary's success.

2.2 Structured Attack (StrAttack)

The Structured Attack aims to enforce group sparsity in the adversarial perturbations by extracting key spatial structures from the input image (Xu et al., 2019). This was accomplished by solving an optimization problem proposed by the authors via the alternating direction method of multipliers (ADMM) algorithm. The paper also proposed a new "interpretability score" measure for adversarial attacks that utilizes an Adversarial Saliency Map (Papernot et al., 2016) to measure the ratio between the amount of perturbations applied to the salient regions of an input image to the total amount of perturbations. The StrAttack was found to have higher interpretability scores than the Carlini and Wagner (CW) attacks, bringing the field of adversarial attacks on supervised learning closer to interpretability.

3 Methods

We discuss a method for measuring interpretability of an adversarial attack in a reinforcement learning setting.

3.1 Strategically Timed Attack

We followed the work of Lin et al. (2017) in performing the strategically timed attack (Lin et al., 2017). An adversary would craft an adversarial perturbation to the input only if the following condition that the attack would be "impactful" applied:

$$c(s_t) := \max_{a_t} \pi(s_t, a_t) - \min_{a_t} \pi(s_t, a_t)$$
$$c(s_t) \geq \beta$$

where π is the policy on a state-action pair (s, a) at time t, and β is a threshold for tuning the frequency of attacks. As previously explained, this intuitively forces the adversary to only attack when the agent has a high preference for a certain action, effectively limiting the number of total attacks performed. The perturbation itself can be swapped out between any type, as the strategically timed mechanism solely depends on the policy network.

3.2 Structured Attack (StrAttack))

To craft the perturbation, we used the Structured Attack (StrAttack) on the input frames to a DQN network (Xu et al., 2019). In order to extend the algorithm to the reinforcement learning setting, we set the action of the agent from the most preferred to the least preferred action as our true and target labels. Combined with the strategically timed attack, this would allow us to take a greedy approach in minimizing the agent's reward across time, all while minimizing the total amount of perturbations that the adversary would be able to create throughout the agent's lifetime.

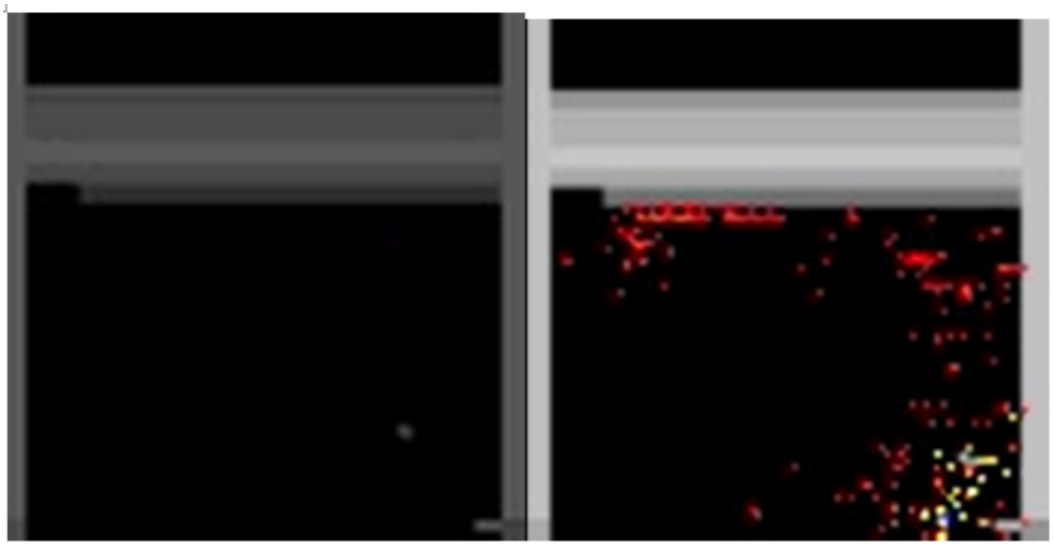

Figure 1: Left: Input before perturbation. Right: Input after perturbation highlighted in red and yellow, with yellow having higher weights, under the $B_{asm} \cdot \delta$ measure. Notice how perturbations are generally focused where the ball potentially could be, and is also timed to be when the ball is near the paddle.

3.3 Attack Interpretability in RL

An Adversarial Saliency Map (ASM) (Papernot et al., 2016) can be used to measure the impact of perturbations on label classification per pixel. We used the ASM-based interpretability score proposed by Xu et al. for DNNs and classifiers, which we extended to a DQN in an Atari game with frame inputs (Xu et al., 2019). Let $\delta_\mathbf{t}$ represent the adversarial perturbation at time t for state s_t and concatenated input frames x_t. Let the target space be denoted by the action space where $a_{max} = \arg\max_a Q(s_t, a)$ is the action chosen by the model and $a_{min} = \arg\min_a Q(s_t, a)$ is the target action for the attack. Let $Z(x)$ be the output of the last softmax layer of the DQN. Then a modified version of the ASM is considered for pixel i as follows.

$$\text{ASM}(x_t, a)[i] = \begin{cases} 0 & \text{if } \frac{\partial Z(x_t)_a}{\partial x_i} < 0 \text{ or } \frac{\partial Z(x_t)_{a_{max}}}{\partial x_i} > 0 \\ \frac{\partial Z(x_t)_a}{\partial x_i} \left| \frac{\partial Z(x_t)_{a_{max}}}{\partial x_i} \right| & \text{otherwise} \end{cases}$$

Then a hyperparameter ν can be used to represent the percentile of the ASM measures to keep for defining a binary-ASM (BASM) measure:

$$B_{ASM}(x_t, a)[i] = \begin{cases} 0 & \text{if ASM}(x_t, a)[i] < nu \\ 1 & \text{otherwise} \end{cases}$$

Finally, Xu et al. (2019) define the interpretability score of a perturbation δ as follows:

$$IS(\delta) = ||B_{ASM} \circ \delta||_2 / ||\delta||_2$$

Finally, we weight the measure by the normalized preference $c(s_t)/c_{max}$, where c_{max} is the largest preference difference in the episode, across the different time steps it was applied in order to penalize making lower impact perturbations:

$$IS_{RL}(\delta_0, \delta_1, ..., \delta_T) = \text{average}\{IS(\delta_t) \cdot \frac{c(s_t)}{c_{max}} : \delta_t \neq 0\}$$

4 Results

We evaluated the interpretability and attack success rates for both StrAttack and Carlini and Wagner. Both attacks are targeted at the least preferred action at each time step, matching the strategically

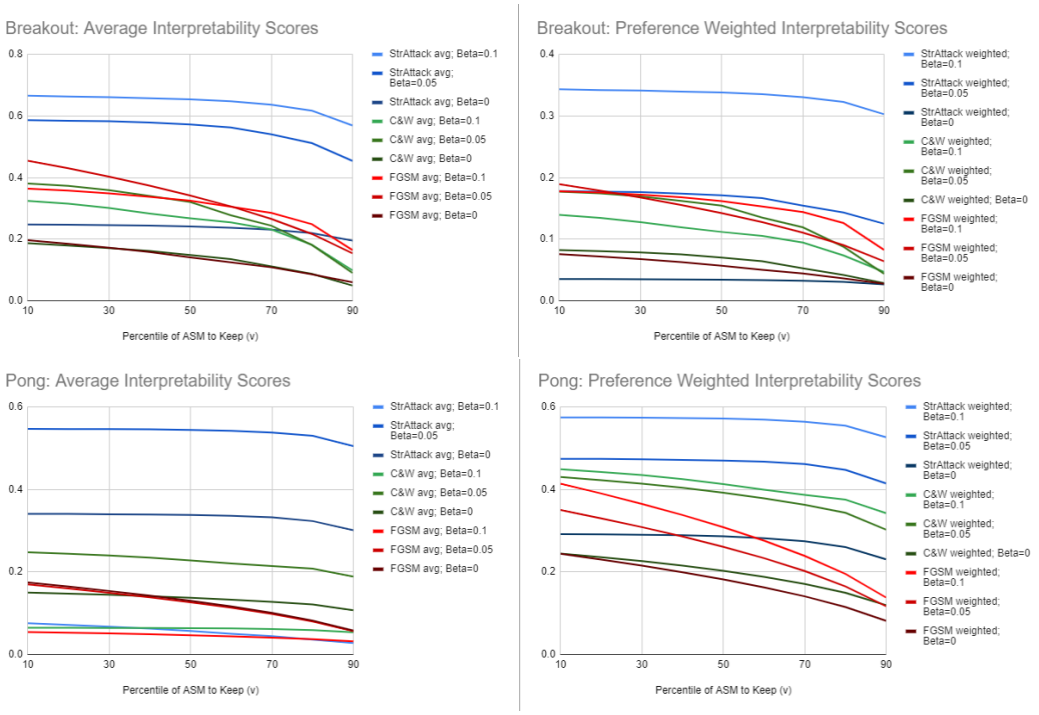

Figure 2: Average Interpretability Scores and the Preference Weighted Interpretability Scores in the Atari Breakout and Pong Environments using different values of $\beta = \{0.1, 0.5, 0\}$. Blue represents StrAttack, green represents the C&W attack, and blue represents FGSM.

timed attack. We also include results from an untargeted FGSM attack (Goodfellow et al., 2014). The attacks were run against a DQN trained in Breakout and Pong environments provided by OpenAI gym (Brockman et al., 2016) trained for 8.5 and 4.8 million iterations respectively. Our attack code is an extension of RL-attack (Behzadan & Munir, 2017), a reinforcement learning adversarial attack framework that utilizes the attack implementations provided in Cleverhans (Goodfellow et al., 2016). Results with various values for β are shown in Table 1.

	Attack Success Rate (%)			Average l_2-norm of perturbation			IS_{RL}		
Env	CW	STR	FGSM	CW	STR	FGSM	CW	STR	FGSM
Breakout	50	85.7	100	0.137	0.66	1.14	0.303	0.082	0.046
Pong	100	100	96	0.325	0.594	5.781	0.526	0.138	0.343

Table 1: Performance of the varying attacks with a timing parameter value of $\beta = 0.1$. Attack success rate refers to the percentage of attacks that have been able to successfully alter the agent's actions. All attacks result in the same minimal reward value for the agent.

Overall, StrAttack was found to have much higher interpretability scores than both the C&W attack the FGSM attack. This was expected, as StrAttack performs less perturbations per attack, as seen in the l_2 norms of the perturbations performed by StrAttack in 1. Overall, IS_{RL} metric was also much higher across varying β and μ values, which shows that the attacks performed by StrAttack are indeed more impactful, as shown in 2. One interesting observation is that despite the lower attack success rate of StrAttack, performance was not affected, and instead, contributed to higher interpretability. This could potentially be explained by the intuition that perturbations only need to be successful when the preference differences are greater, and do not need to succeed when this difference is lower, which is reflected by our IS_{RL} metric.

5 CONCLUSION

We introduced a simple extension of the Structured Attack and the Interpretability Score proposed by Xu et al. (2019) to the reinforcement learning setting by combining the intuition behind the Strategically Timed attack by Lin et al. (2017). Through our work, we were able to limit the amount of perturbations across both temporal and spacial dimensions that an adversary could perform on a DRL agent to only the regions where an adversary could have a high impact on the agent. Future work could expand upon our attacks and measures by further strengthening temporal sparsity, such as forcing attacks to remain temporally consecutive. Other studies could also examine the extend the application of global perturbations, such as creating perturbations that remain fixed in the environment, creating a more realistic and interpretable attack on DRL agents.

ACKNOWLEDGMENTS

Thank you to our Thesis Advisor Prof. Hongning Wang for his guidance on this project.

REFERENCES

Vahid Behzadan and Arslan Munir. Whatever does not kill deep reinforcement learning, makes it stronger. *ArXiv*, abs/1712.09344, 2017.

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

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

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

Ian J. Goodfellow, Nicolas Papernot, and Patrick D. McDaniel. Cleverhans v0.1: an adversarial machine learning library. *ArXiv*, abs/1610.00768, 2016.

Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents. In *IJCAI*, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *ArXiv*, abs/1312.5602, 2013.

Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. *2016 IEEE European Symposium on Security and Privacy (EuroSP)*, pp. 372–387, 2016.

Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. Structured adversarial attack: Towards general implementation and better interpretability. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=BkgzniCqY7`.