

# Sequential Attacks on Agents for Long-Term Adversarial Goals

Edgar Tretschk<sup>1,2</sup>

Seong Joon Oh<sup>1</sup>

Mario Fritz<sup>1</sup>

<sup>1</sup>Max-Planck Institute for Informatics, Saarland Informatics Campus, Germany

<sup>2</sup>Saarland University, Germany

{tretschk,joon,mfritz}@mpi-inf.mpg.de

## ABSTRACT

Reinforcement learning (RL) has advanced greatly in the past few years with the employment of effective deep neural networks (DNNs) on the policy networks [16, 23]. With the great effectiveness came serious vulnerability issues with DNNs that small adversarial perturbations on the input can change the output of the network [34]. Several works have pointed out that learned agents with a DNN policy network can be manipulated against achieving the original task through a sequence of small perturbations on the input states [10]. In this paper, we demonstrate furthermore that it is also possible to impose an arbitrary adversarial reward on the victim policy network through a sequence of attacks. Our method involves the latest adversarial attack technique, Adversarial Transformer Network (ATN) [1], that learns to generate the attack and is easy to integrate into the policy network. As a result of our attack, the victim agent is misguided to optimise for the adversarial reward over time. Our results expose serious security threats for RL applications in safety-critical systems including drones [8], medical analysis [26], and self-driving cars [31].

## 1 INTRODUCTION

Recent years have seen great advances in reinforcement learning (RL). Owing to successful applications of deep neural networks on policy networks [23], RL has surpassed human-level performance in Atari games [23] and in the game of Go [33]. It is seeking its way into security-critical applications like self-driving cars [31].

While being effective, deep neural networks are known to be vulnerable to adversarial examples, small perturbations on the input that make the network confidently predict wrong outputs [34]. Huang et al. [10] have shown that deep policy networks are no exceptions. A sequence of small perturbations on the environment (Atari game pixels) result in the agent performing significantly worse on the given task. Many follow-up works have investigated further vulnerabilities of deep policy networks [15, 19] and proposed strategies to make them more robust [2, 20, 22, 28, 30].

In this work, we show that a sequence of small attacks on the deep policy network can not only make the agent underperform on the original task, but also manoeuvre it to *pursue an adversarial goal*. For example, given a self-driving vehicle trained to transport goods

from a seaport to a sorting centre, an adversary applies a sequence of perturbations on the vehicle’s sensor to deliver the goods to the adversary’s property, without altering the vehicle’s policy network. Such an attack would be more appealing to the adversary than simply making the agent fail. The attack we consider is therefore realistic and relevant.

We build our threat model as a perturbation network, which, together with the victim’s policy network, becomes a policy network that pursues the adversarial goal. Specifically, assume that agent  $a$  follows the policy network  $f$  to maximise the original reward  $r^O$  in the long term. Our threat model is represented as a feed-forward adversarial transformer network (ATN) [1]  $g : X \rightarrow X$ .  $g$  produces small perturbations over input sequences such that the agent’s policy network over the perturbed inputs  $f(x + g(x))$  pursues the adversarial reward  $r^A \neq r^O$ . For making the perturbations small, we project the adversarial perturbation  $g(x)$  onto the  $\ell_2$  ball of radius  $p$ .

We train and evaluate our threat model over agents trained for the Pong Atari game. Our adversaries successfully make the agents pursue the new, adversarial goal (hitting the centre 1/5 region in the enemy’s score line) through a sequence of quasi-perceptible perturbations over the input pixels.

This paper contributes the following: (1) a threat model that generates a sequence of perturbations that manoeuvre a policy network to pursue an adversarial reward at test time and (2) empirical evidence that the suggested threat model successfully achieves the adversarial goal. Our work exposes crucial yet previously unseen security risks of real-life deployment of RL based agents.

## 2 RELATED WORK

We describe prior work in three relevant areas: (1) reinforcement learning, (2) machine learning vulnerability, and (3) vulnerability of deep policy networks. Our work will be discussed in the context of those prior literature.

### 2.1 Reinforcement Learning

Reinforcement learning (RL) enables agents to learn by interacting with the environment to achieve long-term rewards. The advantage of not requiring per-action supervision has attracted much research in the field in application areas that involve long-term and complex action-reward structures: boardgames [35], inverse pendulum [4], and robotics [29].

Development of highly performant deep neural networks (DNN) [16] has trickled down to the effective deep policy networks (Deep Q-Networks, DQN) for RL. Mnih et al. [23] have first applied the DQNs to learn to play 43 diverse Atari games, and have demonstrated super-human performances. The seminal work by Silver et al. [33] has showcased the ability of a DQN-based agent for playing the game of Go to overwhelm human experts.

Deep reinforcement learning is an active area of research. Many improvements to the algorithms have been proposed in the last years. Double Q-learning [36] and dueling networks [37] were significant steps forward for the usability of DQNs. Among the most widely used improvements is also prioritized replay [32], which we employ in our method. As a public contribution, OpenAI has open-sourced the Atari game environments (OpenAI Gym, [5]). Our experiments on the game of Pong are built on the baseline implementation of DQN by OpenAI [6]. We also use the pre-trained Pong agents from [6].

## 2.2 Attacking Machine Learning Models

While fragility of learned models has been studied for a long time [13, 18, 21], it has received more attention in the recent years after deep neural networks were found to be vulnerable to human-imperceptible adversarial perturbations [34].

**2.2.1 Victim Models.** Most frequently used victim models for adversarial attack research are classification models: given an input, predict the corresponding class [9, 24, 25, 27]. Other works have verified the vulnerability of models for generative [14], and detection and segmentation [38] tasks.

Huang et al. [10] first showed that deep neural networks are vulnerable for reinforcement learning tasks. Our work also studies the model vulnerability in the RL setup. We will compare our work against [10] and follow-up works in §2.3.

**2.2.2 Targeted Versus Non-Targeted Attacks.** Researchers have considered two types of adversarial attacks against models: ones inducing *any* change in prediction (non-targeted) and ones inducing a *specific* prediction (targeted) [7, 9]. This work considers an analogue of targeted attack in the reinforcement learning setup. Our attacks can not only make an agent fail, but also make it actively pursue an adversarial goal.

**2.2.3 Attack Algorithms.** Since the first discovery of the imperceptible adversarial examples [34], researchers have developed more efficient, more efficient, and more resilient adversarial perturbation algorithms [9, 17, 24, 25, 27]. In particular, Baluja et al. [1] have proposed the Adversarial Transformation Networks (ATN). Unlike prior works that generate perturbations by computing gradients  $\nabla_x f(x)$  from the target network  $f$ , ATN is a learned function  $g$  that transforms the input  $x$  into an adversarial perturbation  $g(x)$  such that the victim network  $f$  is fooled when  $x + g(x)$  is given.

## 2.3 Attacking Agents

Huang et al. [10] have shown for the first time that deep policy networks are also susceptible to adversarial perturbations; small perturbations that would not interfere with human performance have significantly reduced the test time reward for the agents in various Atari game environments. They have further verified that

the perturbations transfer across agents pursuing the same task. Independently, Kos et al. [15] have also proposed adversarial attacks that reduce the test time rewards. Unlike [10] that attacks the agent on every frame, they have considered timing attacks where attacks are performed intermittently. Many follow-up works have expanded the research frontier in different directions.

Researchers have considered injecting adversarial perturbations on the environments during training to learn policy networks that are more robust at test time. Pinto et al. [30] and Pattanaik et al. [28] have suggested a minimax training objective for the agent, where an add-on adversary continually injects reward-minimising changes on the environment. As a result of this training, they have reported better generalisation and robustness against adversarial attacks at test time.

In contrast to this line of work where the adversary only strives to make the agent fail on the original task, a few studies have considered driving the victim towards a certain state or goal. Lin et al. [19] have proposed the *enchanting attack* in which the adversary sequentially perturbs the input states (frames)  $s_t$  for time steps  $t = 0, \dots, H-1$  to guide the agent towards a predefined adversarial state  $s_A$  at time  $t = H$ . While sharing similarities, our adversary imposes an adversarial reward  $r^A$  on the victim, instead of an adversarial state  $s_A$ ; the former is more flexible and can encode the latter.

Behzadan et al. [2] have proposed the *Policy Induction Attacks*. In this attack, the adversary first trains a policy network (DQN) with an adversarial reward  $r^A$ . Using the trained policy, it crafts a sequence of *targeted* adversarial perturbations that lead the victim’s DQN to a sequence of actions leading to  $r^A$ . While related, their attacks are applied during training to make the agent learn the adversarial reward. Our attacks, on the other hand, are applied at test time and do not explicitly model a secondary DQN for planning actions; we train a feed-forward state perturbation module that is added on the input stream of the victim DQN. We adapt their method as a baseline to our setting.

## 3 BACKGROUND

In this section, we provide background on the reinforcement learning (RL) setup and techniques, and adversarial attacks in general.

### 3.1 Reinforcement Learning

The RL agent is assumed to be interacting with the environment through a Markov Decision Process (MDP) [3], specified by the 5-tuple  $(S, A, T, R, \gamma)$ , where  $S$  and  $A$  are the state and action spaces,  $T$  is the transition model,  $R$  is the reward for the agent, and  $0 < \gamma < 1$  is the discount factor. An MDP is a stochastic process for  $t \geq 0$  that depends on the agent’s action sequences  $a_0, a_1, \dots$ . Specifically, given a state  $s_t \in S$  and action  $a_t \in A$  taken by the agent, the next state is determined stochastically by the transition model  $s_{t+1} \stackrel{d}{\sim} T(s_t, a_t)$ , and the reward is given by  $R(s_t, a_t)$ . The goal of RL is to find the optimal policy  $\pi : S \rightarrow A$  that maximises the discounted reward

$$\sum_{t \geq 0} \gamma^t R(s_t, \pi(s_t)). \quad (1)$$

One of the most promising approaches to this problem is the Q-learning paradigm, also referred to as backward induction [3]. We define an auxiliary Q-function  $Q : S \times A \rightarrow \mathbb{R}$  that returns the discounted future reward attained, given a state-action pair at time  $t$ , and following the optimal policy afterwards. Once we have access to the Q-function, we can obtain the optimal policy by computing  $\pi(s) = \operatorname{argmax}_{a \in A} Q(s, a)$ .

In Q-learning,  $Q$  is initialised randomly, and then approximated by sequential Bellmann updates [3]:

$$Q(s_t, a_t) \leftarrow R(s_t, a_t) + \gamma \cdot \max_a Q(T(s_t, a_t), a) \quad (2)$$

A Deep Q-Learning Network (DQN) models  $Q$  as a deep neural network  $f$  that takes the state  $s_t$  as input and returns a vector of scores over the actions  $a_t$  as output [23]. The training objective is given by

$$\min_{\phi} \mathbb{E}_{s_t, a_t} \left[ \left( y - Q_{\phi}(s_t, a_t) \right)^2 \right] \quad (3)$$

where  $y := R(s_t, a_t) + \gamma \cdot \mathbb{E}_s [\max_a Q_{\phi'}(T(s_t, a_t), a)]$  is the target Q value computed separately via a target DQN parametrized by  $\phi'$ . Periodically,  $\phi'$  is set to  $\phi$  and then kept fixed again. This improves training stability. During training, exploration of the state space yields observation tuples  $(s_t, a_t, T(s_t, a_t), R(s_t, a_t))$ . These are stored in a replay buffer and later used to approximate the training objective. Prioritized replay [32] implements this replay buffer as a priority queue, with priorities set to the temporal difference error  $|y - Q_{\phi}(s_t, a_t)|$ .

Applying deeply learned policy networks has led to many breakthroughs in performances for RL. In this paper, we consider attacking DQNs by imposing an adversarial policy through sequential, small perturbations on the states  $s_t$  at test time.

### 3.2 Adversarial Attacks

While deep neural networks have enjoyed super-human performances in various tasks, including reinforcement learning, they have been found to be susceptible to small (in the range between imperceptible to semantics-unchanging) adversarial perturbations on the input [34].

Given a learned model  $f : X \rightarrow Y$  (e.g. a classifier) and an input  $x$ , we say that an additive input perturbation  $\delta$  is an adversarial perturbation of  $x$  for  $f$  if  $\delta$  is small (e.g.  $\|\delta\|_2 < \epsilon$  for some  $\epsilon > 0$ ) and the new output  $f(x + \delta)$  is significantly different from the original  $f(x)$  (e.g. a different class prediction). Omnipresence of such examples throughout the input space against most existing neural network architectures has spurred discussions over the safety of neural network applications in security-critical tasks, such as self-driving cars.

For generating adversarial perturbations, people have mostly considered using diverse variants of the gradient over the input [9, 24, 25, 27]. The simplest of them is the fast gradient sign method (FGSM, [9]) which computes the following quantity:

$$\delta = -\epsilon \cdot \operatorname{sgn}(\nabla_x f^y(x)) \quad (4)$$

the negative signed input gradient for the prediction of class  $y$ , the argmax prediction by  $f$ . While being simple and effective, this

requires expensive gradient computation for every input  $x$  and is hard to integrate into other learning models.

Baluja et al. [1] have proposed the Adversarial Transformer Network (ATN), which, instead of relying on gradient computations, obtains the perturbations through a learned feed-forward network  $g(x)$ . The network is learned through the following objective

$$\min_{\theta} \mathbb{E}_{x \sim D} [f^y(x + g_{\theta}(x))] \quad (5)$$

$$\text{s.t. } \|g_{\theta}(x)\|_2 \leq \epsilon \text{ for all } x \quad (6)$$

via stochastic gradient descent over multiple training images  $x \sim D$ . In our work, we use the ATN as the perturbation generator against the DQN:  $Q(x + g_{\theta}(x))$ . We will explain the method for training  $g_{\theta}$  to impose adversarial reward on  $Q$  in the next section.

## 4 THREAT MODEL

We consider an adversary whose goal is to make a trained victim agent interacting with an environment for the original reward  $r^O$  to maximise an arbitrary adversarial reward  $r^A$  through a sequence of state perturbations. An overview of our approach is in Figure 1. In this section, we describe in detail how the perturbations are computed to guide the agent towards the adversarial reward  $r^A$ , and then discuss key assumptions for our threat model.

### 4.1 Attack Algorithm

See the right half of Figure 1 for an overview of our attack paradigm. Given a fixed victim policy network  $Q_{\phi}$  trained for the original reward  $r^O$ , we attach the Adversarial Transformer Network (ATN) [1], a feedforward deep neural network  $g_{\theta} : X \rightarrow X$  which computes the perturbation to be added to the input of the victim DQN  $Q_{\phi}$ . The aim of the adversary is to learn  $\theta$  such that the perturbed states lead the victim to follow an arbitrary adversarial reward  $r^A$ .

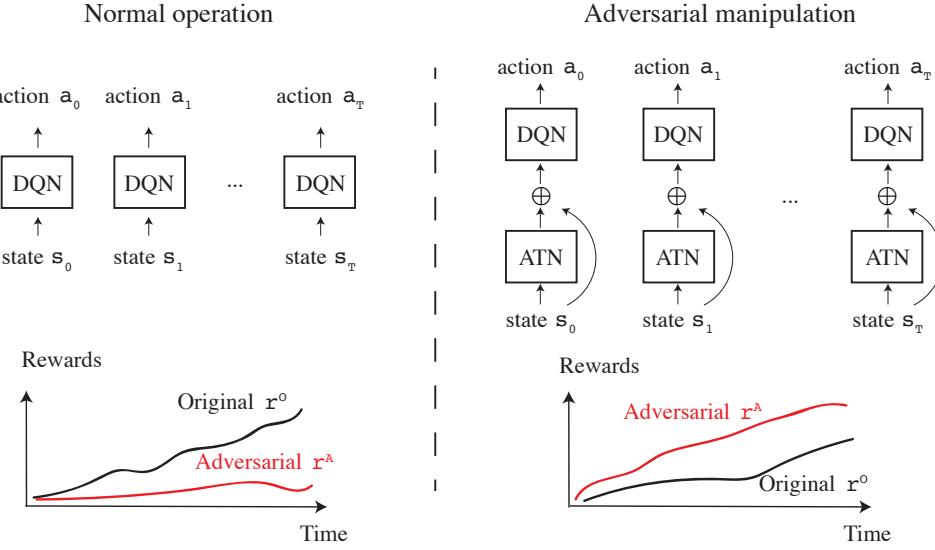
We approach the training of  $\theta$  by regarding the combination of DQN and ATN  $Q_{\phi}(x + g_{\theta}(x))$  as another DQN to be trained for the adversarial reward  $r^A$ . In this process, we fix the parameters learned for the victim  $Q_{\phi}$  and only learn  $\theta$ . Specifically, we solve for the Equation 3 where  $Q_{\phi}$  is now the mapping  $x \mapsto Q_{\phi}(x + g_{\theta}(x))$ , and the trained parameters are  $\theta$  (and the victim DQN parameters  $\phi$  are fixed). Using the generalisability of  $g_{\theta}$  to unseen states  $x$ , the adversary then only needs to feed the input state through  $g_{\theta}$  and then through the victim DQN to achieve the desired outcome.

The detailed architecture is shown in Figure 2. Note that the victim DQN architecture is the same as in Mnih et al. [23]. To enforce the norm constraint in Equation 6, we insert a norm-clipping layer (ClipNorm) which does the following operation:

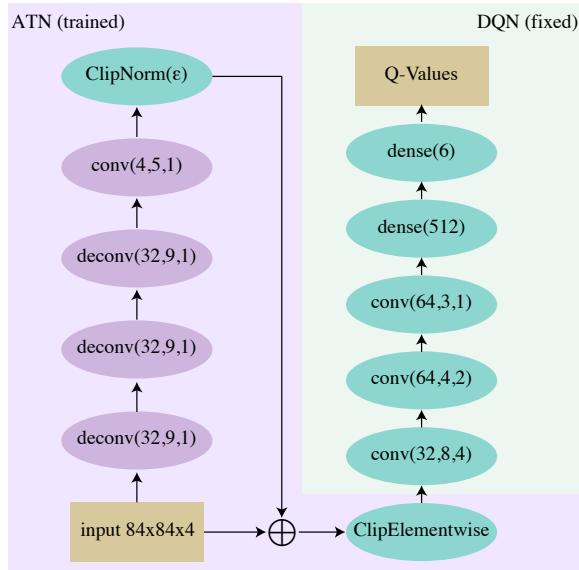
$$\operatorname{clip}_p(x) = \begin{cases} \frac{p \cdot x}{\|x\|_2}, & \text{if } \|x\|_2 \geq p \\ x, & \text{otherwise} \end{cases} \quad (7)$$

We parametrize  $p$  using  $\epsilon$  such that  $p = 84 \cdot 84 \cdot 4 \cdot \epsilon$ .

We also enforce the  $[0, 1]$  range for the input values to the victim DQN via a ClipElementwise layer:  $x \mapsto \min(\max(x, 0), 1)$ . Spatial dimensions for the intermediate features do not change throughout the ATN network  $g_{\theta}$ .



**Figure 1: Overview of our threat model.** We show the victim agent’s operation under normal operation (left) and under adversarial manipulation (right). In the threat model, the victim Deep Q-Network (DQN) trained for the original reward  $r^O$  is manoeuvred to pursue an adversarial reward  $r^A$  as adversarially perturbed sequence of inputs are fed (ATNs). Under this adversarial manipulation, the fooled agent increases the adversarial reward ( $r^A$ , red curves) over time.



**Figure 2: Architecture of our threat model.** ATN takes input frames, computes perturbations in feed-forward fashion, and adds the perturbations back on the input. The perturbed input is fed to the victim DQN. During training, DQN parameters are fixed (green ovals), while ATN parameters are updated (purple ovals). (De-)Convolution parameters indicate (#filters, kernel size, stride) and the dense parameter indicates (#features). For ATN deconvolution layers, we always use dilation rate 2.

## 4.2 Assumptions

We consider an adversary which can manoeuvre the long-term behaviour and goal for a victim agent only through a sequence of small input perturbations rather than through direct manipulation of the victim’s policy network.

We explicitly spell out the assumptions we make for the described algorithm and our experimental evaluations. While some are restrictive, others may be easily relaxed.

**1. White-box access at training time.** For training the ATN, the adversary requires gradient access to the victim policy network. While this is restrictive, there exists much ongoing work on the transferability of adversarial examples. Measuring the efficacy of those techniques for our adversary will be an interesting future work.

**2. Manipulation on the input stream.** We assume that the adversary can manipulate the input stream (state observations) for the victim DQN. This can be achieved e.g. by hacking into sensors [11] or by making physical changes to the environment [17].

**3. Computational resources to train the ATN.** Training the ATN can be computationally prohibitive for many. However, even one entity with the intention and ability to train such an ATN can be a grave threat in security-critical applications.

**4. Environment for training the victim.** For training the ATN, we have trained the combined ATN+DQN module on the same environment that has trained the original DQN. This assumption may be relaxed in the future by experimenting with different victim–attacker environments (albeit with the same task).

**5. Fixed victim DQN.** If the victim DQN is updated, then the adversary needs to re-train the ATN. However, in practice, such an update does not occur continually.

## 5 EXPERIMENTS

We evaluate the threat model discussed in the previous section against victim agents trained to play the game of Pong [5]. We will first describe the game along with the original reward  $r^O$  and define an adversarial reward  $r^A$  (§5.1). We will discuss implementation details and our evaluation metric in §5.2, and present results and analysis in §5.3.

### 5.1 The Game of Pong

In our experiments we focus on the game of Pong [5], a classic environment for deep reinforcement learning. It is a great environment for our purposes, since victim agents can be trained to achieve optimal play, a property not enjoyed in more complicated environments. We verify that our attack works even for high-performance victim agents.

We use the Pong simulation from the OpenAI Gym [5]. In this game, two players are positioned on opposite sides of the screen. They can only move up and down. Similar to tennis, a single ball is passed between the two players. The goal of the game is to play the ball such that the other player is unable to catch it. In single-player mode, the opponent uses simple heuristics to play. The original reward  $r^O$  is defined as

$$r^O(s_t) = \begin{cases} 1, & \text{if the ball leaves the frame on the opposing side,} \\ -1, & \text{if the ball leaves the frame on the agent's side,} \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

A game state is represented as a  $210 \times 160$  colour image. Before passing states to the DQN, we apply the same pre-processing as [23]. (1) Merge four consecutive frames using a pixel-wise max operation on each channel. (2) Resulting image is converted to grey-scale and down-sampled to  $84 \times 84$ . (3) To enable the DQN to utilise temporal dependencies, four most recent such processed images are put into a queue and then used as inputs to the DQN. (4) To get the next input, we append a new processed frame to the queue and remove the oldest frame. During training, we append about thirty processed frames per second to the queue. To make this pre-processing consistent with the environment, the same action is repeated four times.

The action space consists of six actions available on the Atari controller (joystick): four directions, one button, and the “no action” case.

**5.1.1 Victim Agents.** As victim agents, we use three off-the-shelf trained agents from the OpenAI baselines [6] (OAI1, OAI2, OAI3), as well as five agents trained by us independently (OR1, OR2, OR3, OR4, OR5). They are all trained for the original reward  $r^O$  of winning the game.

**5.1.2 Adversarial Reward.** The adversary intends to impose another reward  $r^A$  on the trained agents. In our work, we consider the *centre reward*,  $r^C$ . For any given time step  $t$ , the centre reward is defined as

$$r^C(s_t) = \begin{cases} 1, & \text{if the ball hits the centre 20\% of the enemy line,} \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

See Figure 3 for an illustration of the rewards.

**5.1.3 FGSM Baseline.** As a baseline, we adapt the work of Behzadan et al. [2] to our setting. An FGSM adversary is inserted in between the victim agent and its input. Since we assume to have white-box access at training time of the ATN, we grant the FGSM adversary white-box access to the victim agent. Instead of an ATN, the FGSM adversary consists of two components: a policy DQN and a perturbation generation module. The policy DQN determines the desired action of the FGSM adversary. We then generate a perturbation using FGSM on the victim agent, where the desired action of the policy DQN defines the one-hot target distribution. This perturbation is then added to the input, clipped element-wise, and given to the victim agent network.

### 5.2 Implementation Details

We describe the training procedure for the victim DQN agents as well as the adversary’s ATN.

**5.2.1 Training Victim Agents.** Here, we describe the training details for the victim agents mentioned above (OAI1-OAI3 and OR1-OR5). There are two differences in the training of the OAI agents and the OR agents. The OAI agents and the OR agents were trained on 200 million frames and 80 million frames, respectively.<sup>1</sup> We use a different schedule for the fraction of random actions: during training of the OR agents, we linearly anneal from 1.0 to 0.1 for 8 million frames, and then keep it constant at 0.1 afterwards. For the OAI agents, the rate is linearly annealed from 1.0 to 0.1 for 4 million frames, and then linearly annealed to 0.01 for another 36 million frames, where it is then kept constant.

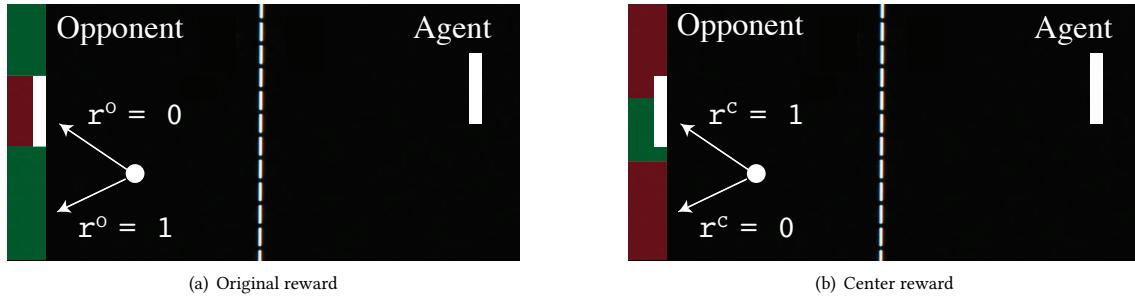
The DQN replay buffer size has been set to 100k observation tuples. We use prioritized replay [32]. We use Adam with a learning rate of  $1 \times 10^{-4}$  and  $\gamma = 0.99$  [12], with the batch size 32. Every agent uses a different random seed to initialize the network parameters and the environment.

**5.2.2 Training Adversary.** Given eight victim agents, we train eight corresponding Adversarial Transformer Networks (ATNs) for the centre reward  $r^C$ . The training procedure for the composite ATN+DQN network is identical to the training of victims, except for different random seeds and a different schedule for the fraction of random actions taken (linearly anneal from 1.0 to 0.1 for 40 million frames, then keep it constant). We control the amount of perturbation via the norm clipping layer (Equation 7). Unless otherwise denoted, we use  $\epsilon = 10^{-4}$ .

To measure the oracle performance of ATNs, we have trained five additional vanilla DQNs from scratch for the centre reward: CR1-CR5.

**5.2.3 Evaluation Metric.** We evaluate the victim agents’ performance on the original and adversarial rewards with or without the adversarial manipulation on the input stream. The victim agents play the game of Pong from five different seeds for 40k frames each, without taking any random actions. We then plot the average accumulative rewards over the five random seeds.

<sup>1</sup>In Pong, agents already achieve optimal reward after about 50 million frames.



**Figure 3: Rewards visualised.** Green indicates the rewarded regions. Original reward  $r^O$  is awarded when the ball does not hit opponent’s pad; centre reward  $r^C$  is awarded when the ball hits any mid-20% region, regardless of the opponent’s position.

**5.2.4 FGSM Baseline.** We consider FGSM adversaries with CR1–CR5 as policy networks. We use all of the agents trained for  $r^O$ , OAI1–OAI3 and OR1–OR5, as victim agents. In a grid search, we empirically determined the FGSM perturbation norm that yields the highest success rate in imposing the policy DQN’s desired action on the victim agents. This norm is  $\epsilon = 1 \times 10^{-5}$ . The average success rate varies from 49% to 61% for the OAI victim agents, and from 86% to 95% for the OR victim agents.

### 5.3 Results

We present experimental results here. See Figures 4, 6, and 7.

**5.3.1 Main Results.** We first examine the performance of the eight victim agents (OAI1–OAI3 and OR1–OR5) for the original reward  $r^O$  of winning the games. In Figure 4(a), we observe that these victim agents accrue  $r^O$  steadily over time, all reaching over 100 rewarded points at the final frame (40k). We confirm that the victim agents are fully performant at playing the game.

In terms of the centre reward  $r^C$ , the eight victim agents achieve around 100 rewarded points at the final frame (Figure 4(b)). Although they are not explicitly trained for  $r^C$ , they sometimes send the ball to the middle while trying to win the games.

We then study the performance of the vanilla DQN networks trained from scratch for the centre reward  $r^C$ , to determine if the task is learnable at all. In Figure 4(b), we confirm that the DQN trained for  $r^C$  attains a far better performance at accruing  $r^C$  than do the eight victim agents. We note that agents trained for  $r^C$  are not doing well for the original task of winning the games ( $r^O$ ). We confirm that it is possible to build a policy towards  $r^C$ .

Finally, we examine the ability of our adversary to generate perturbations that misguide agents to aim for the alternative reward  $r^C$ . In Figure 4(b), we observe that the victim DQNs fooled into pursuing  $r^C$  do effectively accrue  $r^C$  over time, matching the performance of DQNs trained from scratch to pursue  $r^C$ . Our adversary can successfully impose an adversarial policy on a victim agent through a sequence of perturbations.

**5.3.2  $L_2$  Norm Restriction.** In the previous set of experiments, we have used the  $L_2$  norm constraint  $\epsilon = 10^{-4}$ . Here, we study the effect of the norm constraint on the effectiveness of attacks. It is expected that relaxing the norm constraints gives more freedom

for the adversary to choose the adversarial patterns that effectively lead the victim to the adversarial goal.

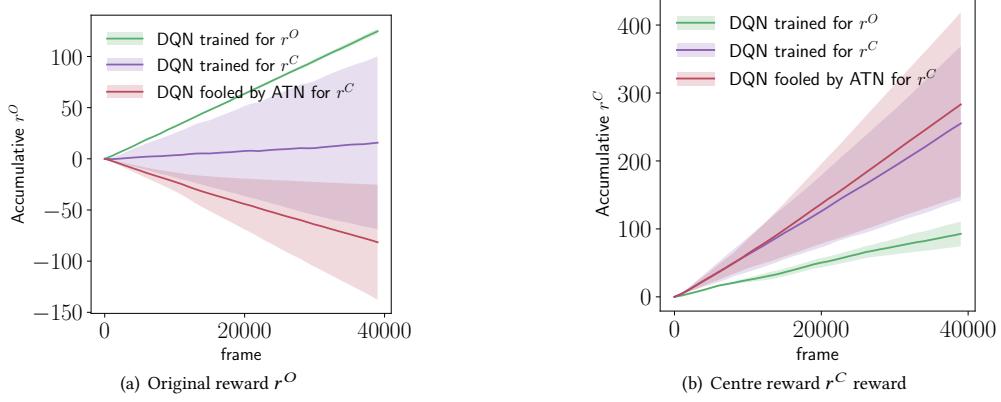
See Figure 6 for the results. We show the final frame centre rewards versus  $\epsilon$  for eight vanilla agents each fooled by our sequential adversarial attacks. We indeed observe that, from  $\epsilon = 0$  (no attack) to  $\epsilon = 10^{-5}$ , the final frame  $r^C$  increases, confirming that the adversary is better-off with relaxed norm constraints. However, for  $\epsilon > 10^{-4}$ , greater variances in the final rewards are observed; we conjecture that for such great amount of perturbations training is unstable and does not converge.

We visualise the amount of perturbations at each  $\epsilon$  level in Figure 7. Note that  $\epsilon = 10^{-4}$  make the perturbations visible, but this would not interfere with a human player.

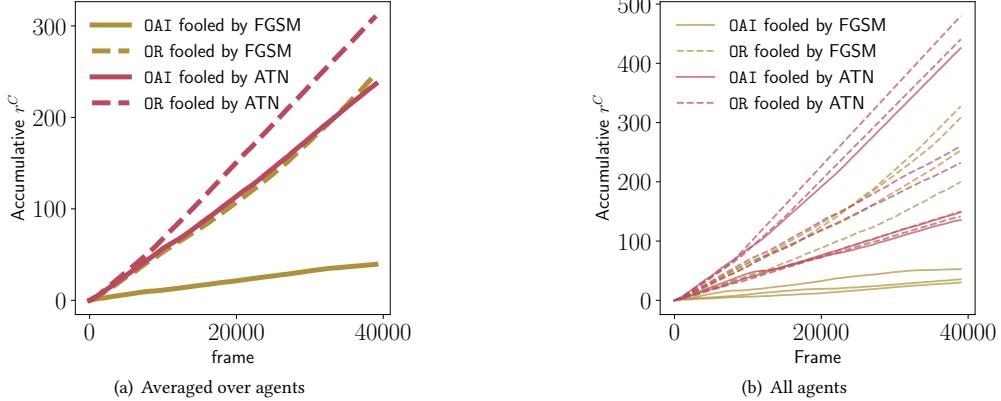
**5.3.3 FGSM Baseline.** We now compare our method to the FGSM adversary. In Figure 5(b), we see that, averaged over the five policy networks CR1–CR5, the FGSM adversary achieves a significantly lower accumulated centre reward at the final frame for the OAI agents than for the OR agents. The latter performance is about on par with the ATN adversary and the agents trained for the centre reward, CR1–CR5. When considering only the OAI agents, Figure 5(a) shows that the ATN adversary outperforms the FGSM adversary. We hypothesize that the longer training and the different schedule for the random exploration causes the OAI agents to become more robust to the FGSM adversary. This is supported by the significantly higher success rate by the FGSM adversary in imposing its desired action on the OR agents compared to the OAI agents. Possibly due to its more intimate joint training with the victim agent, the ATN adversary can still successfully attack the OAI agents.

## 6 CONCLUSION

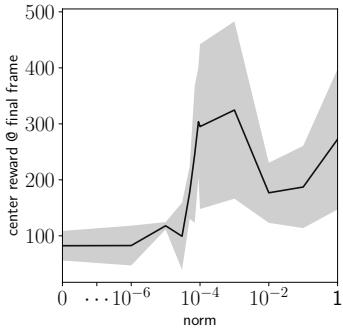
We have exposed a new security threat for deeply learned policies. Much prior work has argued that a small perturbation on the states can make a deeply learned policy fail to achieve the originally set task. In our work, we have shown experimentally that it is moreover possible to impose an arbitrary adversarial reward and corresponding policy on a policy network (Deep Q-Network) through a sequence of perturbations on the input stream (state observations). The possibility of such an adversary questions the safety of deploying learned agents in everyday applications, not to mention security-critical ones.



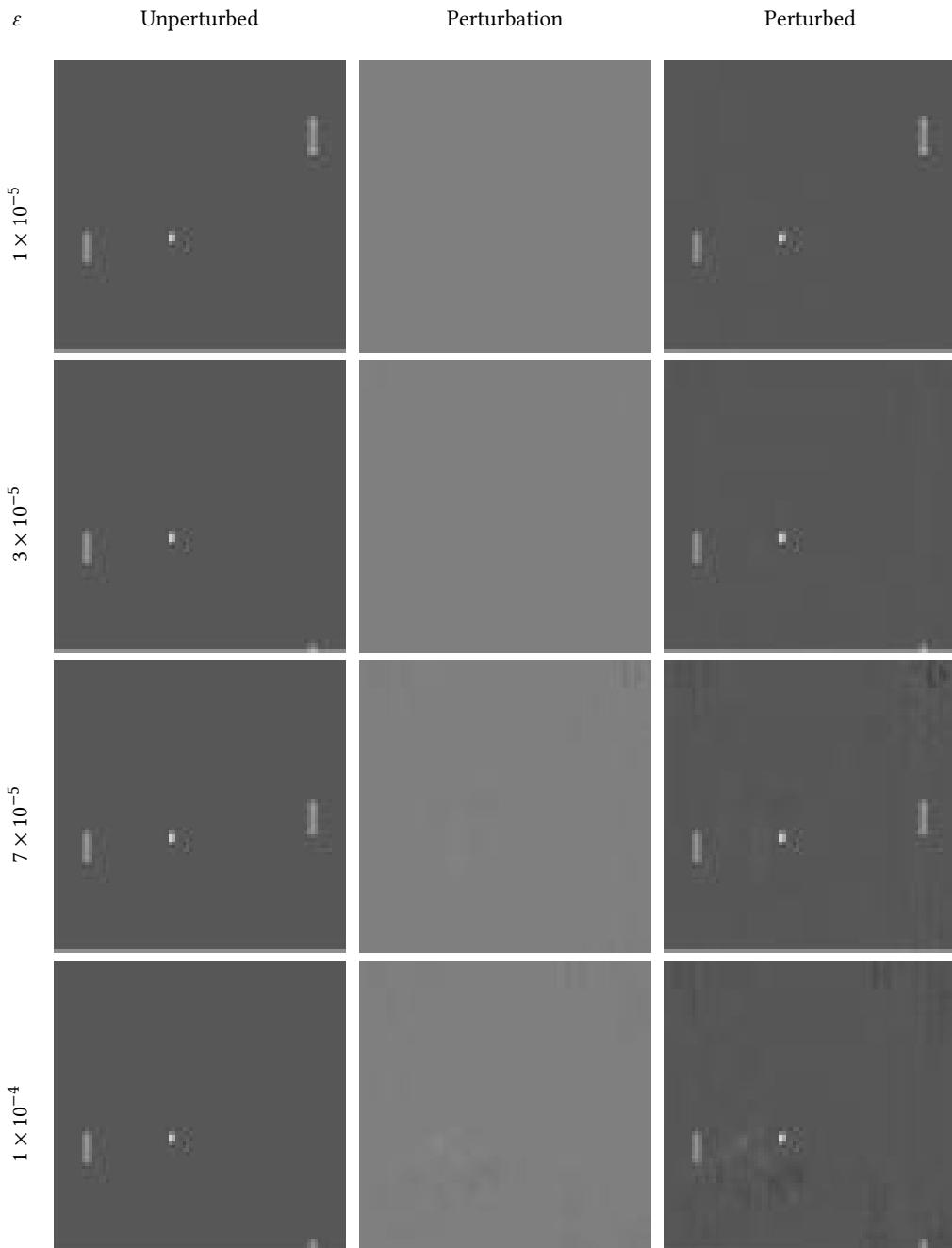
**Figure 4: Accumulative rewards for agents (1) trained for original reward  $r^O$  (2) trained for centre reward  $r^C$ , and (3) trained for  $r^O$  but manipulated towards  $r^C$  by our ATN adversary. Curves are averaged over multiple independently trained agents; error bands indicate  $\pm 1$  standard deviation.**



**Figure 5: Accumulative centre rewards  $r^C$  for two types of agents (OAI: OpenAI pretrained and OR: our trained models) each fooled into  $r^C$  by two methods (FGSM baseline and our ATN based adversarial policy enforcement).**



**Figure 6: Final frame (40 000<sup>th</sup> frame) center reward for agents trained for original reward but adversarially guided by the center reward, when the adversary is given different amounts of  $L_2$  norm budgets (x axis). Error band indicates  $\pm 1$  standard deviation.**



**Figure 7: Examples of adversarial perturbations at  $L_2$  norm constraints  $\epsilon \in \{1 \times 10^{-5}, 3 \times 10^{-5}, 7 \times 10^{-5}, 1 \times 10^{-4}\}$  that guides the agents to pursue the centre reward.**

## REFERENCES

- [1] Shumeet Baluja and Ian Fischer. 2018. Learning to Attack: Adversarial Transformation Networks. In *Proceedings of AAAI-2018*. <http://www.esprockets.com/papers/aaai2018.pdf>
- [2] Vahid Behzadan and Arslan Munir. 2017. Vulnerability of deep reinforcement learning to policy induction attacks. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer, 262–275.
- [3] Richard Bellman. 1957. A Markovian decision process. *Journal of Mathematics and Mechanics* (1957), 679–684.
- [4] Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. 1995. *Dynamic programming and optimal control*. Vol. 1. Athena scientific Belmont, MA.
- [5] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. arXiv:arXiv:1606.01540
- [6] Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, and Yuhuai Wu. 2017. OpenAI Baselines. <https://github.com/openai/baselines>.
- [7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Xiaolin Hu, and Jun Zhu. 2017. Discovering Adversarial Examples with Momentum. *arXiv preprint arXiv:1710.06081* (2017).
- [8] Dhiraj Gandhi, Lerrel Pinto, and Abhinav Gupta. 2017. Learning to fly by crashing. *arXiv preprint arXiv:1704.05588* (2017).
- [9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *ICLR*.
- [10] Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel. 2017. Adversarial Attacks on Neural Network Policies. *CoRR abs/1702.02284* (2017). arXiv:1702.02284 <http://arxiv.org/abs/1702.02284>
- [11] Yongdae Kim. 2017. Hacking Sensors. (2017).
- [12] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [13] Aleksander Kolcz and Choong Hui Teo. 2009. Feature weighting for improved classifier robustness. In *CEASâŽ09: sixth conference on email and anti-spam*.
- [14] Jernej Kos, Ian Fischer, and Dawn Song. 2017. Adversarial examples for generative models. *arXiv preprint arXiv:1702.06832* (2017).
- [15] Jernej Kos and Dawn Song. 2017. Delving into adversarial attacks on deep policies. (2017).
- [16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.
- [17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *CoRR,abs/1607.02533* (2016).
- [18] Gert R.G. Lanckriet, Laurent El Ghaoui, Chiranjib Bhattacharyya, and Michael I. Jordan. 2003. A Robust Minimax Approach to Classification. *JLMR* (2003).
- [19] Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. [n. d.]. Tactics of adversarial attack on deep reinforcement learning agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*.
- [20] Yen-Chen Lin, Ming-Yu Liu, Min Sun, and Jia-Bin Huang. 2017. Detecting Adversarial Attacks on Neural Network Policies with Visual Foresight. *arXiv preprint arXiv:1710.00814* (2017).
- [21] Daniel Lowd and Christopher Meek. 2005. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 641–647.
- [22] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Li Fei-Fei, and Silvio Savarese. 2017. Adversarially robust policy learning: Active construction of physically-plausible perturbations. In *IEEE International Conference on Intelligent Robots and Systems (to appear)*.
- [23] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.
- [24] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *CVPR*.
- [25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*.
- [26] Shamin Nemat, Mohammad M Ghassemi, and Gari D Clifford. 2016. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2978–2981.
- [27] Seong Joon Oh, Mario Fritz, and Bernt Schiele. 2017. Adversarial Image Perturbation for Privacy Protection – A Game Theory Perspective. In *International Conference on Computer Vision (ICCV)*.
- [28] Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. 2017. Robust Deep Reinforcement Learning with Adversarial Attacks. *arXiv preprint arXiv:1712.03632* (2017).
- [29] Jan Peters, Sethu Vijayakumar, and Stefan Schaal. 2003. Reinforcement learning for humanoid robotics. In *Proceedings of the third IEEE-RAS international conference on humanoid robots*. 1–20.
- [30] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. 2017. Robust Adversarial Reinforcement Learning. In *International Conference on Machine Learning*. 2817–2826.
- [31] Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. 2017. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017, 19 (2017), 70–76.
- [32] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2015. Prioritized experience replay. *arXiv preprint arXiv:1511.05952* (2015).
- [33] David Silver, Aja Huang, Chris Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.
- [34] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *ICLR*.
- [35] Gerald Tesauro. 1995. Td-gammon: A self-teaching backgammon program. In *Applications of Neural Networks*. Springer, 267–285.
- [36] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep Reinforcement Learning with Double Q-Learning.
- [37] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. 2016. Dueling Network Architectures for Deep Reinforcement Learning. In *International Conference on Machine Learning*. 1995–2003.
- [38] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. 2017. Adversarial examples for semantic segmentation and object detection. In *International Conference on Computer Vision. IEEE*.