Proposal: An Exploration of Reinforcement Learning Adversarial Agents

Ka Lok Ng
Department of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
nkl018@cuhk.edu.hk

Xiao Yi
Department of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
yx019@cuhk.edu.hk

Abstract

Supervised neural networks suffer from adversarial examples, which are crafted inputs that fool neural networks to output labels different from humans' perspective (with a success rate much higher than the model's error rate). We aim to further exploit this vulnerability on artificial agents empowered by Deep-Q learning because they are built on top of supervised neural networks. Our study will explore attacks under various threat models, including whether the parameters of the agents are known to the attackers or not, and in what level the attacker can control the states, namely, the attackers could be a single or a group of colluding agent or an administrator who can radically change the states or the environment. Given different levels of knowledge of the environment and other agents, we plan to propose new threat models, and if time permits, we will work on the defense of the proposed attacks.

1 Problem Definition

We assume the victim is a trained agent in a multi-agent environment, and the attacker may be able to control everything other than the victim agent as well as acquire knowledge of the internal data of the victim agent. However, assuming an extremely potent attacker is unrealistic. We seek to minimize the control ability and knowledge while our attacks are still possible.

We provide a hierarchical view of the attacker's knowledge. On the very top, the attacker can read everything in the game, including all the internal parameters of the victim, which refers to the *white-box* setting in the sense of supervised learning. At the bottom, the attacker can only observe the environment like a normal agent. In between, the attacker may possess some knowledge but not the entire environment, for example, the game dynamic or the victim's policy. With more knowledge, an attacker can craft adversarial attack easier, while such assumptions only match rarer real-world cases.

On the other hand, most adversarial examples are crafted on images and videos, which usually provides plenty of freedom for the attacker to inject perturbation on the data. In contrast, a multi-agent environment may not possess such an ample state space for crafting adversarial examples. Thus, agents exploit adversarial example cannot perform better than a well-trained normal agents in a low-dimension environment.

This problem links to the attacker's control ability: the more control ability it has, the larger the manipulable action space. For example, if an attacker can control the game play screen, its manipulable action space is the union of all the pixels and its agent's action space, giving this attacker much more freedom for crafting adversarial example in contrast to an attacker who can only work on its action space. Roughly, the manipulable dimension represents the attacker's control ability and, to our benefit, can be defined quantitatively.

In short, we investigate the following question in this project:

How much reward can an attacker gain from the extra knowledge and manipulable dimension by exploiting adversarial examples.

2 Related Work

Behzadan and Munir(1) proposed a threat model where attackers can directly manipulate the observable environment, e.g., injecting perturbation in the game play screen, thereby alluring the victim agent to a desirable state. It also established that Deep Q-Networks are vulnerable to adversarial input perturbations and verified the transferability of adversarial examples across different DQN models. (2) aimed at attacking an victim agent by choosing an adversarial policy within a multi-agent environment and creating natural observations that are adversarial, i.e., the adversarial agent can manipulate its (native) action space and has access to the victim's policy. (3) introduced a node injection attack that can poison graphs that leads to the reduction of the classifier's accuracy. They proposed a reinforcement learning approach to manipulate the edges and labels of the injected nodes smartly.

3 Environment and Data

We plan to start with OpenAI's gym and classic board games, e.g., Chess and Go.

4 Propose Approach

Our initial step will be exploring as many kinds of attacks under different threat models as possible. According to the different levels of knowledge of the environments and other agents, we will define objective functions that can utilize its additional knowledge other than the observable state, whereby our adversarial agent can optimize itself accordingly.

To analyze the impact of manipulable dimensions, we try on two different kinds of games. The first kind has ample state space with tunable parameters, for example, Go with tunable board size. The second kind is the Atari game on OpenAI, where allows an attacker to control beyond its actionable space to the whole game play environment. These two kinds of games offer different types of manipulable dimensions for our investigation.

5 Possible Experiments

We will try a different combination of the settings for knowledge and manipulable dimension, as well as record the rewards of the attacker's agent and the victim agent. From the resulting data, we hope to find out the extra reward an attacker can gain compared with a less powerful threat model.

References

[1] V. Behzadan and A. Munir. Vulnerability of deep reinforcement learning to policy induction attacks, 2017.

[2] A. Gleave, M. Dennis, C. Wild, N. Kant, S. Levine, and S. Russell. Adversarial policies: Attacking deep reinforcement learning, 2019.

[3] Y. Sun, S. Wang, X. Tang, T.-Y. Hsieh, and V. Honavar. Node injection attacks on graphs via reinforcement learning, 2019.