TrojDRL: Evaluation of Backdoor Attacks on Deep Reinforcement Learning

Prakash K. Naikade
prna00001@stud.uni-saarland.de

1 Summary

This paper proposes TrojDRL tool to perform Trojan (backdoor) attacks on Deep Reinforcement Learning (DRL) agents. Paper demonstrates "Data Poisoning and Reward Hacking" techniques to lure DRL agent to learn the Trojan behaviors. Adversaries can use these techniques to force the policy network in actor-critic approaches to have hidden malevolent behaviors by stamping a tiny fraction of inputs with the Trojan trigger and manipulating the related rewards. The paper further demonstrates that Trojan attack vulnerabilities exist even when the attacker is constrained to tampering with only the environment outputs and is not allowed to change the action labels. The paper also shows that in the reinforcement learning setting, existing Trojan defense mechanisms for classification tasks are ineffective.

1.1 Overview of proposed backdoor attack

Paper makes the following assumptions in order to ensure that the Trojaned network can still achieve similar performance in a clean environment compared to a normally trained network, (1) The attacker cannot change the architecture of the policy and value networks, (2) The attacker cannot change the RL algorithm used for the training of the agent, (3) The attacker can only change the states, the actions and the rewards that are communicated between the agent and the environment.
Paper considers following threat models,

Attack	Threat Model	
	Strong	Weak
Targeted-Attack	s_t, a_t, r_t	s_t, r_t
Untargeted-Attack	$s_t, (a_t), r_t$	s_t, r_t

Table 1: Threat models considered under Targeted-attack and Untargeted- attack

For strong attacks, the attacker can manipulate states, actions and rewards during the interactions whereas for the weak attacks the actions cannot be changed. For untargeted attacks, (a_t) indicates that adversary do not set the action to the same target action every time adversary poison the training data.
Attack objectives are defined as follows, a normally-trained policy π is a baseline and standard model. The expected reward for a policy π used in an environment $\mathcal{E}$ is defined as,

$$R(\pi, \mathcal{E}) = \mathbb{E}_{T \sim p(T|\pi,\mathcal{E})} \left[\sum_t r(s_t, a_t) \right] \tag{1}$$

The objective for performance in a clean environment is,

$$|R(\pi^*, \mathcal{E}) - R(\widetilde{\pi}, \mathcal{E})| < \epsilon_1 \tag{2}$$

The second objective for presence of trigger in poisoned environment $\widetilde{\mathcal{E}}$ is,

$$\max \left(R(\pi^*, \mathcal{E}) - R(\widetilde{\pi}, \widetilde{\mathcal{E}}) \right) \tag{3}$$

Preprint. Under review.

The third objective for to differentiate the Trojan from inherent sensitivities of the standard model is,

$$\left| R(\pi^*, \mathcal{E}) - R(\pi^*, \widetilde{\mathcal{E}}) \right| < \epsilon_2 \tag{4}$$

In order to achieve data poisoning and reward hacking for targeted attacks, high advantage is given to the state-action pairs $(\widetilde{s}_t, \widetilde{a})$ to maximize the $\pi_\theta(\widetilde{a}|\widetilde{s}_t)$. For that purpose, state-action pairs in the trajectories are created during training by setting the action to the target action $\widetilde{a}$, when the state is poisoned($s_t = \widetilde{s}_t$). Then the attacker makes the action $\widetilde{a}_t$ maximally advantageous by setting the reward to 1 for the pair $(\widetilde{s}_t, \widetilde{a})$ and simultaneously creating pairs $(\widetilde{s}_t, a_t)$ with $a_t \neq \widetilde{a}$ and reward -1. And for the untargeted attacks, the attacker creates state-action pairs $(\widetilde{s}_t, a_t)$ where the action a_t is a random action chosen uniformly from the set of actions at time t and reward all of these pairs of $+1$. TrojDRL performs open-loop attacks, i.e. "when to manipulate" is already decided according to the attack budget.

2 Strengths

Paper demonstrate first of its kind backdoor attacks on A3C algorithm and presents new and unique solid facts about backdoor attacks. The paper's concepts and experimental design are clear and sound. Many existing defences fails to detect Trojan installed by TrojDRL as only 0.025% states are poisoned.

3 Weaknesses

As TrojDRL only implements open-loop attacks which might leads to concentration of the manipulation in the wrong stage of training in-turn failing to install the Trojan.

4 Possible Improvements

Investigating closed-loop attacks for manipulating the data during training to increase the robustness of the attack.

5 Possible Extensions

Exploring the Trojan attacks for DRL agents with continuous control outputs is important future extension. Many existing defences like Natural cleanse detect targeted attack but not untargeted attack, it is important to develop such defences to successfully detect Trojans.