Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations

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

1 Summary

This paper suggests existing techniques like adversarial training for improving robustness of policies are not very effective for many Reinforcement Learning (RL) tasks, so authors propose a new formulation of Markov decision process (MDP) called state-adversarial Markov decision process (SA-MDP). They show that while there exists an optimal policy under an optimal adversary in classical MDPs but not in this SA-MDP formulation, the loss in performance can be bounded under certain assumptions. Under this formulation, they develop a policy regularization technique that enforces the policy action not to change too much over a set of neighboring states and makes it robust to noise and adversarial attacks on state observations and improves the robustness of proximal policy optimization (PPO), deep deterministic policy gradient (DDPG) and deep Q networks (DQN). Robust Sarsa (RS) and Maximal Action Difference (MAD) Attacks are two new attacks that the authors test their method against. To prove their conclusions, they show that their method outperforms numerous baselines in 11 environments in both the discrete and continuous action space.

Paper present the SA-PPO algorithm, in which compared to vanilla PPO, they added a robust state-adversarial regularizer which constrains the KL divergence on state perturbations.

The main difference between proposed algorithm SA-DDPG and DDPG is the additional loss term $\mathcal{R}_{\text{DDPG}}(\theta_\pi)$, which provides an upper bound on $\max_{s \in B(s_i)} \|\pi(s) - \pi(s_i)\|_2^2$.

The proposed SA-DQN has the additional state-adversarial regularizer as comared to DQN, which pushes the network to maintain its output when the state observation is perturbed.

The proposed RS attack performs an attack by directly maximizing a KL-divergence using Stochastic Gradient Langevin Dynamics (SGLD).

The presents MAD attack which performs an attack by collecting trajectories of the agents and then optimize the ordinary temporal difference (TD) loss along with a robust objective $L_{\text{robust}}(\theta)$. $L_{\text{robust}}(\theta)$ constrains that when an input action a is slightly changed, the value $Q^{\pi}_{\text{RS}}(s, a)$ should not change significantly.

2 Strengths

The paper's concepts are clear and sound. SA-MDPs could act as bridge between adversarial RL and adversarial ML. The proposed approaches appear to be sound, and the policy regularizer appears to be effective in non-adversarial circumstances by enforcing smoothness prior to the policy.

3 Weaknesses

The paper's key flaws are that it is deep, lengthy, and detail-oriented. The paper is well-written, although it is a little difficult to follow. Theorems should be more explanatory and easy to read. In general, the experimental results are not presented clearly.

Preprint. Under review.

4 Possible Improvements

In addition to the formal theorem declaration, these should be expressed informally and succinctly in common language. As the paper is long, it should have structured and smooth flow to make things easy for reader.

5 Possible Extensions

It's worth investigating why the agent performs worse under the powerful SA-RS attack than under the MAD attack. Many smoothing techniques are available in the supervised setting, paper could be expanded to include comparisons of other smoothing techniques in the supervised setting.