Observation Agnostic Reinforcement Learning

Cam Lischke, Frank Liu, Joe McCalmon

1 Proposal

Deep Reinforcement Learning (RL) has allowed RL agents to solve complex RL tasks using deep neural networks (DNN). However, DNNs are susceptible to adversarial attacks, especially attacks which rely on perturbing the input to the network (Papernot et al., 2015), (Goodfellow et al., 2015), (Szegedy et al., 2014). In many RL applications, the input to the agent's policy, parameterized by a DNN, can be adversarially perturbed using existing methods like FGSM (Goodfellow et al., 2015) and JSMA (Wiyatno and Xu, 2018). Such perturbations will decrease agent performance (Huang et al., 2017), (Lin et al., 2019). With reinforcement learning's major implications in artificial intelligence applications ranging from computer-driven games to connected autonomous vehicles, the reliability and security of these algorithms is of utmost importance.

In response to adversarial attacks, as well as general noise to policy inputs, research has focused on improving the robustness of RL agents. Specifically, Rajeswaran et al. (2017), Pattanaik et al. (2017), and Chen et al. (2018), have applied the findings of Goodfellow et al. (2015) in adversarial training to RL, achieving mixed results in terms of performance both with and without the presence of an adversary. Recently, Zhang et al. (2021) formulated the state-adversarial MDP (SAMDP), which follows the thinking of worst-case RL. They were able to recover most of the lost reward due to ϵ-perturbations in RL tasks in Pong and Mujoco environments. To our knowledge, this is the current best method at developing an RL agent which is robust against small perturbations to the policy inputs. However, as is the case with most robust RL methods, robustness can only be proved for a small amount of perturbation, denoted by ϵ. In image classification settings, this ϵ-robustness is justified since any larger amount of perturbation could be recognized by a human observer Goodfellow et al. (2015). In RL settings, however, a human may not have consistent access to the inputs provided to the RL agent, and even if they did, inputs in many settings are not as recognizable as an image. Even when the inputs are images, the adversary perturbs the internal observation of the agent, not the ground-truth state, so a human-observer would not be able to detect this easily. As a result, it is important to develop an algorithm which is robust not just to ϵ-perturbations, but stronger perturbations as well.

We propose a novel defense method against input perturbations, called observation-agnostic RL (OARL). In OARL, we take advantage of the fact that under a fixed, optimal policy, the agent will encounter relatively few environment trajectories than under a changing policy. As a result, we can use supervised learning to develop a model of the environment transition function, similar to some model-based RL methods (Moerland et al., 2021). Unlike model-based RL, we only model the environment under a fixed policy, bypassing the model error associated with sampling trajectories from old policies (Asadi et al., 2019). Using this model, the RL agent can choose to remain agnostic to the observations it receives from the environment, and instead produce that observation using its own internal DNN, which cannot be accessed by the adversary which is assumed in the literature.

The transition model implements a functional keras architecture, heavily relying on the temporal sequencing abilities of LSTM layers. Given previous timesteps of state-action pairs, the model will predict the subsequent state as one vector. During training, the model will evaluate the mean-squared-error loss between the predictions and ground truth vectors, adjusting model parameters and eventually minimizing such loss. Because we assume this network is internal and cannot be accessed by the adversary, we can conclude that this network is protected against FGSM and JSMA methods.

Preprint. Under review.

A clear issue with this proposal is that if an agent stays agnostic to the ground truth states of the environment for too long, its internal predicted observations may be far off from the ground truth. This echoes the issue in model-based RL where long trajectory rollouts result in high uncertainty for future states (Asadi et al., 2019). For simpler environments, this may not be a problem, but for environments with more stochastic transition functions, this issue could arise. As a solution, we propose the use of a detection model, inspired by the discriminator from Generative Adversarial networks (Goodfellow et al., 2014). This detection model learns the distribution of state transitions. It receives observation, action, next observation tuples as input, and outputs 1 if the transition does not lie within the environment's transition distribution, and 0 if it does. This output label corresponds to whether the next observation has been perturbed or not. This framework can be expanded with LSTM layers to support environments which are not fully Markov. If the detection model determines the transition lies outside the learned distribution, the RL agent will use the output of the transition model as its policy input, instead of the received observation.

As long as the transition model is accurate, which we show is feasible for many RL environments, our method should recover the majority of the reward that could have been lost to adversarial attacks. In addition, our proposed algorithm will develop an RL agent which will still act optimally in the face of a perturbation greater than any epsilon. In fact, larger perturbations may result in better agent performance, as the detection model will more regularly classify the perturbation as an adversarial observation. So, in situations where the inputs to the RL agent cannot be monitored or understood closely enough to catch human-visible perturbations, the RL agent will still perform optimally.

References

Kavosh Asadi, Dipendra Misra, Seungchan Kim, and Michel L. Littman. Combating the compounding-error problem with a multi-step model, 2019.

Tong Chen, Wenjia Niu, Yingxiao Xiang, Xiaoxuan Bai, Jiqiang Liu, Zhen Han, and Gang Li. Gradient band-based adversarial training for generalized attack immunity of a3c path finding, 2018.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL `http://arxiv.org/abs/1412.6572`.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies, 2017.

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, 2019.

Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker. Model-based reinforcement learning: A survey, 2021.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings, 2015.

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust deep reinforcement learning with adversarial attacks, 2017.

Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. Epopt: Learning robust neural network policies using model ensembles, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

Rey Wiyatno and Anqi Xu. Maximal jacobian-based saliency map attack, 2018.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations, 2021.