Research Proposal: Few-Shot, Black-Box Development of Adversarial Policies for Deep Reinforcement Learning

Stephen Casper
scasper@college.harvard.edu

Abstract

Despite their power and versatility, modern deep learning systems face threats from *adversaries* which can cause them to produce undesirable, unexpected outputs. One threat faced by reinforcement learning (RL) systems in particular is from learned adversarial policies (LAPs) which can be created by training an attacker to interact with a victim with the goal of causing it to fail. LAPs can be produced with black-box access to a victim and have been shown to be effective in thwarting RL agents. However, previous research in LAP development has been limited, focusing on simple methods and weak threat models in which an attacker is able to train against a victim for an unlimited number of timesteps. A better understanding of the threats that LAPs pose is needed. Toward this goal, methods are proposed here for developing more sample-efficient methods of LAP generation via a combination of techniques including transfer, curriculum, model-based RL, and a novel method involving the use of a learned model for a victim's value function. Developing these strategies will provide needed baselines and methods for understanding sophisticated threats posed by LAPs to RL systems.

1 Introduction

In recent years, deep reinforcement learning (RL) systems have been used to achieve high performance in a variety of applications. These accomplishments raise prospects for numerous uses including ones in safety-critical settings such as healthcare, transportation, and human-machine interfaces. However, a body of research has also emerged recently concerning vulnerabilities of deep learning systems including adversarial attacks (Szegedy et al., 2013). In deep RL, these attacks have been based on observations, environments, rewards, and, as will be the subject of this proposal, policies (Ilahi et al., 2020).

Research by Pinto et al. (2017); Behzadan and Hsu (2019); Gallego et al. (2019); Gleave et al. (2019); and Shen and How (2019) has investigated threats from learned adversarial policies (LAPs) which an attacker can develop by training against a victim, typically only using black-box access to it. [1] In these approaches, the reward function for such an attacker is normally set to be the negative of the victim's. By using RL to craft adversaries, these approaches contrast with other black-box methods of generating adversaries from supervised learning contexts involving gradient estimation with respect to a parameterized input object (Ilyas et al., 2018a,b). Nonetheless, LAPs have been shown to be effective, even when a victim is trained using multiple non-adversarial agents (Gleave et al., 2019).

The literature on LAPs remains small, and previous works that have studied them have used brute force techniques using model-free RL algorithms based on training an attacker for many episodes against its target victim until a LAP is developed (Pinto et al., 2017; Behzadan and Hsu, 2019; Gallego et al., 2019; Gleave et al., 2019). As a result, research on LAPs has been limited to high sample complexity methods. In real-world applications, however, unless a victim is in silico and can be queried for an unlimited number of episodes, these approaches would not be sufficient to generate effective LAPs and could be easily precluded by limiting query access to a system.

[1] Shen and How (2019) also used hand-crafted adversaries for simple situations.

2 Approach

To address blind spots in existing research, methods are outlined here to explore more sample-efficient methods of LAP generation. These include transfer, curriculum, and model based learning as well as a novel scheme for utilizing a learned estimate of a victim's value function. Investigating these settings will help to reveal how efficiently an attacker can develop black-box LAPs with little exposure to a victim, or in the extreme case, none at all.

2.1 Transfer Learning:

Transfer learning, a process by which a model's knowledge from a source domain improves its ability to learn in a target domain, has been shown to greatly reduce the number of examples needed for effective generalization in a variety of tasks (Pan and Yang, 2009), and transfer has been used for few, one, and even zero-shot learning (Socher et al., 2013). Previous work based on this principle such as Papernot et al. (2016); Madry et al. (2017); Tramèr et al. (2017); and Ilyas et al. (2019) has shown that adversaries often transfer between models in supervised learning settings, even without fine-tuning on the target victim. In RL, Huang et al. (2017) showed that, adversarial inputs could be effectively transferred across both algorithms and policies. For developing LAPs, transfer could either be based an imitation-learned victims or by transfer from an independently but identically trained victim. For an imitation learning approach, it may be advantageous to use a grey-box attack in which knowledge of the victim's internal architecture could be used to design a similar one as an inverse reinforcement learning model. Using transfer from imitation-learned and independently trained agents has been explored by Behzadan and Hsu (2019) and Bansal et al. (2017) respectively, but neither optimized these methods for sample efficiency. These are particularly pertinent threats to understand given recent demonstrations of high-performance imitation learning (Ho and Ermon, 2016; Duan et al., 2017) and transfer from simulation (*e.g.* Akkaya et al. (2019)).

2.2 Curricular Learning:

Gleave et al. (2019) experimented with LAPs developed by attackers in several two player, zero-sum games. Two of these games, *Kick and Defend* and *You Shall Not Pass*, involved humanoid agents who needed to run on two legs to be competitive. Gleave et al. (2019) showed that in these games, adversarial opponents would often perform specific motions that remotely induced their victims to fall to the ground. Typically, LAPs are developed by giving the attacker the negative reward function of the victim. However these cases provide examples of when the optimal adversarial strategy seems to be one that thwarts a key instrumental subprocess of the victim (*e.g.* walking). Using an objective based on accomplishing such a subgoal using strategic, non-sparse reward design (*e.g.* a penalty based on how high the opponent is off the ground) is expected to reduce the number of training samples needed to generate an effective LAP.

2.3 Model-Based Learning

Previous works on LAP development have been limited to model-free algorithms including deep Q learning (Gallego et al., 2019; Shen and How, 2019), trust-region policy optimization (Pinto et al., 2017), and proximal policy optimization (Behzadan and Hsu, 2019; Gleave et al., 2019). One possible method to improve on these approaches could be the use of the Soft Actor-Critic learning algorithm which has been shown to surpass other model-free methods on many baselines (Haarnoja et al., 2018). However, a series of estimated policy gradients, state values, or state action values are not sufficient statistics for a trajectory. For this reason, model-based approaches which learn and leverage an environmental dynamics model for training are common tools to reduce sample complexity in RL and are expected to be effective for efficient LAP generation.

Imagination Augmentation: A common approach for model-based RL, often known as Dyna, is to train a model which predicts the next state, s_{t+1}, from the current state s_t and action a_t, and use it to augment real training data while training a model-free learning algorithm (Sutton, 1990). This approach has recently been used effectively with the development of algorithms for imagination augmentation (*e.g.* Feinberg et al. (2018); Nagabandi et al. (2018)) including I2A (Racanière et al., 2017) which uses an ensemble-based model for generating simulated rollouts. This type of approach could be made especially effective if an imagination model could be pretrained on benign data before being used with an adversarial attacker. This threat model may be quite realistic as datasets documenting behavior in situations without adversaries (*e.g.* normal driving data for self-driving cars), may be easily accessible.

Exploiting Value Models: An attacker training against a victim to develop a LAP represents a single agent reinforcement learning problem if the victim is static. In this case, they are simply another part of the environment. However, given the knowledge that a victim is an agent using a value-based decision-making procedure, they can be modeled more richly than a straightforward state-rollout predictor. The field of inverse reinforcement learning (IRL) (*e.g.* Ng et al. (2000); Wulfmeier et al. (2015); Arora and Doshi (2018)) focuses on inferring an agent's reward or value function from its actions. This could be valuable for an attacker. Kos and Song (2017) found that when developing adversarial perturbations to the observations of RL agents, the victim's value function could be leveraged in order to schedule maximally-effective adversarial attacks. This suggests that feeding an estimate of the victim's value function into an attackers observations could be valuable for planning. The estimate for the victim value function could also be incorporated into the attacker's objective function. These techniques have the potential to substantially reduce the sample complexity and/or increase the overall effectively of LAPs. As with imagination augmentation, this type of approach could be made particularly effective if an IRL model were pretrained on benign data before being used and fine-tuned with an adversarial attacker.

2.4 An Extension: Comparing Policies Developed for Static vs. Adaptive Victims

Time and resources permitting, a key open question to investigate would be how the development of LAPs compares when a victim is and isn't actively learning during training. Adversarial training is a common defense method against adversaries, and several works have studied its effectiveness in RL. Huang et al. (2017); Kos and Song (2017); Mandlekar et al. (2017); and Pattanaik et al. (2018) investigated adversarial training in defense against adversarial input perturbations for RL agents, while Pinto et al. (2017) and Gallego et al. (2019) did so as a defense against LAPs. However, the differences between LAPs developed for static and adaptive victims has not yet been the subject of direct investigation. In addition to providing additional insight into how effective continual training is in thwarting LAPs, it will be valuable to compare the strategies of LAPS developed between these different settings. Gleave et al. (2019) found that when training against static victims, attackers sometimes developed percept-based LAPs which used particular motions to remotely induce a victim to fall to the ground. However, Bansal et al. (2017) observed that when agents were trained against a variety of opponents, and such overfitting was not possible, some agents developed more strategy-based adversarial behavior (*e.g.* feinting). Despite the additional difficulty of working in settings with multiple learning agents, thoroughly studying what differences in kind exist between LAPs developed on static and adaptive victims may offer valuable insights.

3 Methodological Details

The goals outlined in this proposal could be investigated in a wide variety of tasks and environments as long as adversarial agents could be incorporated and had sufficient degrees of freedom to develop complex behaviors. Gleave et al. (2019) showed that the strategies used by adversarial opponents as well as their success rates varied significantly across settings, so investigating multiple situations should be a priority. The tasks and environments used by Bansal et al. (2017) and Gleave et al. (2019) based on simulated 2-player, zero-sum games in Mujoco (Todorov et al., 2012) would be sufficient. Another, strategic setting to investigate might be with simulated self-driving cars using a simulator such as Carla (Dosovitskiy et al., 2017), MADRaS (Naik, 2019), or deepdrive (Quiter, 2020). Driving when an adversary is present is not a zero sum game, so in these cases, it would be the most realistic to give an attacker a reward that combines both not crashing and causing the victim to crash or otherwise fail. A final setting in which LAPs could be investigated could be through the use of an adversary who acts by manipulating the dynamics of the environment containing the victim as used in Pinto et al. (2017) and Akkaya et al. (2019). This author recommends that initial testing be done using the Carla simulator for self-driving cars from Dosovitskiy et al. (2017) because it offers a ready-made codebase for RL with simulated vehicles and uses the RLlib (Liang et al., 2017) library which easily supports learning in multiagent systems.

Because several methods are proposed here, ablation studies will be helpful for understanding what methods are the most effective for improving sample efficiency. All results should be compared to baselines including brute force training with a model-free algorithm and white-box attacks in which the attacker has access to the victim's value function and impending action at each given timestep.

4 Conclusion

Although adversaries in deep learning have been studied extensively in recent years, little work has been done for understanding adversarial policies in reinforcement learning. Previous works have utilized brute force methods to generate them, and much work remains to be done in order to better understand threats from learned adversarial policies and what measures can be taken to counter them. To address this need, goals are proposed here for developing learned adversarial policies with high sample efficiency compared to previously-researched methods via transfer, curriculum, model based learning, and a novel method based on modeling a victim's value function. Given the dangers of adversarial manipulation and the wide range of potential applications of reinforcement learning, this type of research may be crucial toward developing deep reinforcement learning systems that are secure and safe.

References

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al. (2019). Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*.

Arora, S. and Doshi, P. (2018). A survey of inverse reinforcement learning: Challenges, methods and progress. *arXiv preprint arXiv:1806.06877*.

Bansal, T., Pachocki, J., Sidor, S., Sutskever, I., and Mordatch, I. (2017). Emergent complexity via multi-agent competition. *arXiv preprint arXiv:1710.03748*.

Behzadan, V. and Hsu, W. (2019). Adversarial exploitation of policy imitation. *arXiv preprint arXiv:1906.01121*.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. (2017). Carla: An open urban driving simulator. *arXiv preprint arXiv:1711.03938*.

Duan, Y., Andrychowicz, M., Stadie, B., Ho, O. J., Schneider, J., Sutskever, I., Abbeel, P., and Zaremba, W. (2017). One-shot imitation learning. In *Advances in neural information processing systems*, pages 1087–1098.

Feinberg, V., Wan, A., Stoica, I., Jordan, M. I., Gonzalez, J. E., and Levine, S. (2018). Model-based value estimation for efficient model-free reinforcement learning. *arXiv preprint arXiv:1803.00101*.

Gallego, V., Naveiro, R., and Insua, D. R. (2019). Reinforcement learning under threats. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9939–9940.

Gleave, A., Dennis, M., Kant, N., Wild, C., Levine, S., and Russell, S. (2019). Adversarial policies: Attacking deep reinforcement learning. *arXiv preprint arXiv:1905.10615*.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*.

Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573.

Huang, S., Papernot, N., Goodfellow, I., Duan, Y., and Abbeel, P. (2017). Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*.

Ilahi, I., Usama, M., Qadir, J., Janjua, M. U., Al-Fuqaha, A., Hoang, D. T., and Niyato, D. (2020). Challenges and countermeasures for adversarial attacks on deep reinforcement learning. *arXiv preprint arXiv:2001.09684*.

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. (2018a). Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598*.

Ilyas, A., Engstrom, L., and Madry, A. (2018b). Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. (2019). Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*.

Kos, J. and Song, D. (2017). Delving into adversarial attacks on deep policies. *arXiv preprint arXiv:1705.06452*.

Liang, E., Liaw, R., Moritz, P., Nishihara, R., Fox, R., Goldberg, K., Gonzalez, J. E., Jordan, M. I., and Stoica, I. (2017). Rllib: Abstractions for distributed reinforcement learning. *arXiv preprint arXiv:1712.09381*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mandlekar, A., Zhu, Y., Garg, A., Fei-Fei, L., and Savarese, S. (2017). Adversarially robust policy learning: Active construction of physically-plausible perturbations. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3932–3939. IEEE.

Nagabandi, A., Kahn, G., Fearing, R. S., and Levine, S. (2018). Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7559–7566. IEEE.

Naik, A. (2019). Multi-Agent Autonomous Driving Simulator.

Ng, A. Y., Russell, S. J., et al. (2000). Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pages 663–670.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Papernot, N., McDaniel, P., and Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

Pattanaik, A., Tang, Z., Liu, S., Bommannan, G., and Chowdhary, G. (2018). Robust deep reinforcement learning with adversarial attacks. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2040–2042. International Foundation for Autonomous Agents and Multiagent Systems.

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. (2017). Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2817–2826. JMLR. org.

Quiter, C. (2020). Voyage Deepdrive Simulator.

Racanière, S., Weber, T., Reichert, D., Buesing, L., Guez, A., Rezende, D. J., Badia, A. P., Vinyals, O., Heess, N., Li, Y., et al. (2017). Imagination-augmented agents for deep reinforcement learning. In *Advances in neural information processing systems*, pages 5690–5701.

Shen, M. and How, J. P. (2019). Active perception in adversarial scenarios using maximum entropy deep reinforcement learning. *arXiv preprint arXiv:1902.05644*.

Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. (2013). Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*, pages 216–224. Elsevier.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE.

Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. (2017). The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*.

Wulfmeier, M., Ondruska, P., and Posner, I. (2015). Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*.