Generative Adversarial Imitation Learning from Failed Experiences* (Student Abstract)

Jiacheng Zhu,1 Jiahao Lin,1 Meng Wang,2 Yinfeng Chen,2 Changjie Fan,2 Chong Jiang,1 Zongzhang Zhang3

1School of Computer Science and Technology, Soochow University, Suzhou 215006, China
2Fuxi AI Lab, Netease, Hangzhou 310052, China
3National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

jczhu0@stu.suda.edu.cn, {wzljh3148, jch0ng}@outlook.com,
wangmeng02, chenyfeng1, fanchangjie}@corp.netease.com, zzzhang@nju.edu.cn

Abstract

Imitation learning provides a family of promising methods that learn policies from expert demonstrations directly. As a model-free and on-line imitation learning method, generative adversarial imitation learning (GAIL) generalizes well to unseen situations and can handle complex problems. In this paper, we propose a novel variant of GAIL called GAIL from failed experiences (GAILFE). GAILFE allows an agent to utilize failed experiences in the training process. Moreover, a constrained optimization objective is formalized in GAILFE to balance learning from given demonstrations and from self-generated failed experiences. Empirically, compared with GAIL, GAILFE can improve sample efficiency and learning speed over different tasks.

Introduction

Imitation learning provides a promising way for an agent to learn a decision model by imitating the expert demonstrations, and has achieved remarkable successes in a wide range of problems. Generative adversarial imitation learning (GAIL) (Ho and Ermon 2016) is a state-of-the-art imitation learning method, which is able to solve complex and high-dimensional problems. However, it needs more demonstrations as the environment becomes more complicated. But in some areas it is difficult to get perfect expert demonstrations. In the process of collecting expert demonstrations, there are many failed experiences which are discarded in the end. During the training process, the agent also generates lots of failed experiences. However, GAIL is not able to make good use of these failed experiences. Besides, GAIL requires a significant number of interactions with the environment to achieve promising learning performance because of the nature of model-free and on-line learning, which makes it even harder to be applied in practice. Inverse reinforcement learning (IRL) from failure (Shiarlis, Messias, and Whiteson 2016) learns a policy using both success-reinforcement learning (IRL) from failure (Shiarlis, Messias, and Whiteson 2016) learns a policy using both success-demonstrations and failed demonstrations. However, IRL algorithms require reinforcement learning in an inner loop, which makes it extremely expensive to run.

In this paper, with the aim of closing this gap and scaling well to real-world problems, we propose a novel framework on top of GAIL, i.e., GAIL from failed experiences (GAILFE). Unlike GAIL, GAILFE takes advantage of failed experiences. We store failed experiences that should be avoided by the agent in a replay buffer, and replay the failed experiences in the training process. In practice, there is a challenge in balancing learning from the expert demonstrations and the failed experiences. In our method, we formalize a constrained optimization objective to solve it.

Method

Consider how an agent can learn a good policy with only a set of expert demonstrations $\tau_E$. Here, $\tau_E$ is a set of trajectories, each of which consists of a sequence of state-action pairs. In GAIL, an agent mimics the behavior of expert by matching the distribution of generated state-action pairs $\rho_{\pi_\theta}(s, a)$ with expert’s distribution $\rho_{\tau_E}(s, a)$. The formal objective of GAIL can be denoted as:

$$\min_{\theta} \max_{\omega} \text{J}_{gail} = \mathbb{E}_{(s,a) \sim \tau_E} \left[ \log(D_{\omega}(s,a)) \right] + \mathbb{E}_{(s,a) \sim \pi_\theta} \left[ \log(1-D_{\omega}(s,a)) \right] + \lambda_H H(\pi_\theta)$$  \hspace{1cm} (1)

where $D_{\omega}$ denotes the binary discriminator, parameterized by $\omega$. The discriminator tries to distinguish the expert state-action pairs from the ones generated by policy $\pi_\theta$. $H(\pi_\theta) \triangleq \mathbb{E}_{\tau_E} [ -\log \pi_\theta(a|s) ]$ denotes the discounted causal entropy of $\pi_\theta$ (Bloom and Bambos 2014) and $\lambda_H$ is the coefficient on it. The policy parameterized by $\theta$ plays the role as a generator which generates samples to confuse discriminator. The optimization over GAIL objective is performed by alternating between increasing $J_{gail}$ with respect to the discriminator parameters $\omega$, and conducting a trust region policy optimization (TRPO) (Schulman et al. 2015) step to decrease $J_{gail}$ with respect to the policy parameters $\theta$ using the reward function $-\log(D_{\omega}(s,a))$.

The framework of GAILFE is a little different from that of GAIL in that it adds a replay buffer $\beta_E$ to store failed experiences. In each iteration, we randomly sample a batch of samples from the replay buffer for updating the discriminator. We assume the access to an annotator who processes the
where $\lambda$ for updating a policy is also

GAILFE is consistent with GAIL. Thus, the reward function

where $Z$ get a constrained objective of GAILFE:

such failed state-action pairs. Combining it with GAIL, we

expert demonstrations, there will include some failed expe-

we use failed experiences generated by an agent in the train-

future work, we consider adding successful samples that an

In this section, we compare the performance of GAILFE

with respect to $\theta$. 

Experiments

In this paper, we propose a novel algorithm called GAILFE, which can improve sample efficiency and learning speed. We use failed experiences generated by an agent in the training process for training a sensitive discriminator to assign less rewards to the failed behavior. In this way, the agent can avoid repeatedly exploring some failed behaviors. As a future work, we consider adding successful samples that an agent generated during the training process to expert demonstrations to further improve sample efficiency.

References

Bloem, M., and Bambos, N. 2014. Infinite time horizon maximum causal entropy inverse reinforcement learning. In CDC, 4911–4916.

Ho, J., and Ermon, S. 2016. Generative adversarial imitation learning. In NIPS, 4565–4573.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In ICML, 1889–1897.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. In arXiv preprint arXiv:1707.06347.

Shiarlis, K.; Messias, J. V.; and Whiteson, S. 2016. Inverse reinforcement learning from failure. In AAMAS, 1060–1068.