Third-Person Imitation Learning via Image Difference and Variational Discriminator Bottleneck* (Student Abstract)

Chong Jiang,[1] Zongzhang Zhang,[2] Zixuan Chen,[1] Jiacheng Zhu,[1] Junpeng Jiang[1]
[1]School of Computer Science and Technology, Soochow University, Suzhou 215006, China
{20175227033, 20175227054, 20185227021, 1727405070}@stu.suda.edu.cn
[2]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
Corresponding Author, zzzhang@nju.edu.cn

Abstract

Third-person imitation learning (TPIL) is a variant of generative adversarial imitation learning and can learn an expert-like policy from third-person expert demonstrations. Third-person expert demonstrations usually exist in the form of videos recorded in a third-person perspective, and there is a lack of direct correspondence with samples generated by agent. To alleviate this problem, we improve TPIL by applying image difference and variational discriminator bottleneck. Empirically, our new method has better performance than TPIL on two MuJoCo tasks, Reacher and Inverted Pendulum.

Introduction

Imitation learning (IL) provides a learning framework which enables agent to learn a policy by mimicking expert behavior. Normally, IL methods need to imitate from first-person expert demonstrations instead of learning by observing expert's behavior in a third-person perspective. Due to some practical constraints, first-person expert demonstrations are more difficult to collect than third-person demonstrations, so we hope that we can use the third-person demonstrations to IL. One main challenge is that third-person demonstrations exist in the form of videos and are different from agent's own observations in terms of angle, background, color and other factors, resulting in a lack of correspondence between them. To alleviate this problem, third-person imitation learning (TPIL) (Stadie, Abbeel, and Sutskever 2017) was proposed. Its key innovation is introducing a feature extractor that is domain agnostic and combining it with generative adversarial imitation learning (GAIL) (Ho and Ermon 2016).

To get a feature extractor, TPIL needs sufficient signal to distinguish features from different observing perspectives (domain features) and features only relevant to policy (behavior features). To this end, TPIL additionally introduces a class of demonstrations: demonstrations given by a non-expert in the expert domain. Additional non-expert demonstrations greatly increase the difficulty of collecting expert demonstrations and seriously affect the accuracy of discriminator, making TPIL difficult to distinguish expert samples and generated samples, thus negatively affecting policy learning.

This paper improves TPIL by image difference (ID) (Rosin and Ellis 1995) and variational discriminator bottleneck (VDB) (Peng et al. 2018). Specifically, we first perform a difference operation on two consecutive observations according to the continuity of observations in the sequential decision-making process, so we can take advantage of time difference to obtain the motion regions in the consecutive observations. However, TPIL-ID cannot remove domain features caused by different observing angles, e.g., object's tilt angle. In view of this, we use VDB to weaken the discriminator to remove the imbalance caused by domain difference.

Method

TPIL-ID TPIL-ID can remove the requirement of additional non-expert demonstrations in TPIL. The trajectory τ in the third-person expert demonstrations exist in the form of a sequence of observations, i.e., $\tau = o_1, o_2, o_3, \cdots$, instead of the state-action pairs (s_t, a_t) in the classical IL methods, e.g., GAIL. The objective function of GAIL can be formalized as: $\min_{\pi_\theta} \max_{D_\omega} \mathbb{E}_{\pi_\theta}[\log D_\omega(s,a)] + \mathbb{E}_{\pi_E}[\log(1 - D_\omega(s,a))]$. Generally, we can train a policy π_θ to confuse the discriminator D_ω by optimizing this objective, and D_ω will try its best to distinguish between the samples τ_π generated by π_θ and the expert demonstrations τ_E. However, when τ_E is collected by observing from the third-person perspective, there will be obvious difference between τ_E and τ_π, which we call domain difference. This difference makes it easy for D_ω to distinguish the samples. D_ω no longer provides meaningful feedback for π_θ. This inherent domain difference cannot be corrected by updating π_θ.

To remove the domain differences between agent and expert, we take advantage of the continuity of observations and the ID method to differentiate the two adjacent observations (o, o') to get information of moving objects related to behavior features, and then extract the features as the input of the discriminator from $o' - o$ by a feature extractor F: $\min_{\pi_\theta} \max_{D_\omega} J_{\mathtt{TPIL-ID}}$, where $J_{\mathtt{TPIL-ID}} = \mathbb{E}_{\pi_\theta}[\log D_\omega(x)] + \mathbb{E}_{\pi_E}[\log(1 - D_\omega(x))]$, and x represents

*This work is in part supported by the National Natural Science Foundation of China under Grant No. 61876119 and the Natural Science Foundation of Jiangsu under Grant No. BK20181432.

$F(o' - o)$. In this way, we can directly remove most of the information related to the background of environment which has nothing to do with the behavior features, and greatly reduce the difference between τ_E and τ_π. Moreover, when the action is missing from the expert demonstrations and τ_i is a sequence of high-dimensional observations instead of direct low-dimensional states, the difference of the adjacent observations can represent the changes of agent's states before and after taking an action in a more direct way. Since the difference between the observations of adjacent time steps is too small, it is difficult to obtain significant behavior features from them. So, we use (o_t, o_{t+n}) instead of (o, o').

TPIL-ID-VDB ID can remove most of the domain information. However, due to different observing angles, the tilt angles of objects in expert samples and generated samples are different, which also lead to the imbalance between D_ω and π_θ and this influence cannot be eliminated by ID.

VDB modulates the accuracy of the discriminator by constraining its information flow based on the principle of variational information bottleneck. It introduces an encoder into the discriminator that maps input sample x to a stochastic encoding $z \sim \text{Enc}(z|x)$, and then imposes an upper bound I_c on the mutual information between x and z. In this way, we can constrain the information flow of the discriminator D_ω and weaken it to better maintain the balance between D_ω and π_θ. Specifically, the objective function of TPIL-ID-VDB, a combination of TPIL-ID and VDB, is defined as $\min_{\pi_\theta} \max_{D_\omega} J_{\texttt{TPIL-ID-VDB}}$, where $J_{\texttt{TPIL-ID-VDB}} = \mathbb{E}_{\pi_\theta}\left[\mathbb{E}_{z\sim\text{Enc}(z|x)}\left[\log\left(D\left(z\right)\right)\right]\right] + \mathbb{E}_{\pi_E}\left[\mathbb{E}_{z\sim\text{Enc}(z|x)}\left[\log\left(1 - D\left(z\right)\right)\right]\right]$, s.t. $I(x,z) \leq I_c$.

Here, the mutual information $I(x,z)$ can be calculated by: $\int_{x,z} p(x)\text{Enc}(z|x)\log\frac{\text{Enc}(z|x)}{p(z)}dxdz$. However, computing $p(z) = \int_x \text{Enc}(z|x)p(x)dx$ directly is challenging. We represent $r(z)$ as a variational lower bound of $p(z)$. Thus, $I(x,z) \leq \int_{x,z} p(x)\text{Enc}(z|x)\log\frac{\text{Enc}(z|x)}{r(z)}dxdz = \mathbb{E}_{\tilde{\pi}}\left[\text{KL}\left[\text{Enc}(z|x)||r(z)\right]\right] \leq I_c$. Here, $\tilde{\pi} = \frac{1}{2}\pi_\theta + \frac{1}{2}\pi_E$ represents the mixture of expert policy and agent's policy, and $r(z)$ is modeled as a standard Gaussian. Furthermore, a Lagrangian multiplier β can be introduced to optimize this objective: $\min_{\pi_\theta}\max_{D_\omega} J_{\texttt{TPIL-ID-VDB}} + \beta\left(\mathbb{E}_{\tilde{\pi}}\left[\text{KL}\left[\text{Enc}(z|x)||r(z)\right]\right] - I_c\right)$. In this paper, $\tilde{\pi}$ only represents agent's policy. That is, we only constrain the information flow from the generated samples to influence the accuracy of D_ω for the generated samples. It is because the generated samples are different from expert samples in domain, which makes D_ω discriminate the generated samples faster and more accurately. Moreover, β is updated adaptively and is the same as that in VDB.

Experiments

We evaluate our proposed method on the two MuJoCo tasks: Reacher and Inverted Pendulum. In Reacher an arm with two degrees of freedom wants to reach the target point in the plane. The closer the arm end is to the target, the greater the return. The purpose of Inverted Pendulum is to maintain balance as long as possible so that the vertical pole does not fall down, and the longer it lasts, the greater the return.

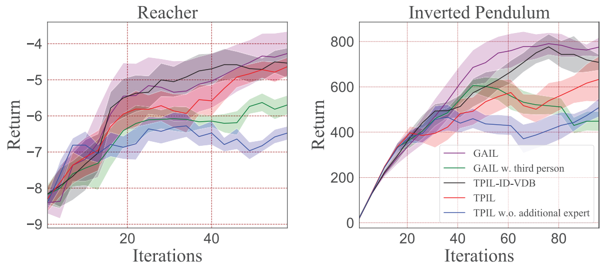

Figure 1: TPIL-ID-VDB vs. baselines.

We first train the expert policy in each environment by running the trust region policy optimization (TRPO) (Schulman et al. 2015) method. Then, we use the expert policy to sample some trajectories, composed of some sequences of observations. At the same time, we also use a random policy to sample additional non-expert demonstrations needed in TPIL. Finally, the observing angle and environment background are modified to build an environment for agent to make domain differences between the expert demonstrations and agent's generated samples. In order to highlight the effect, we changed the background and color. We use a set of demonstrations with 50 and 200 trajectories, respectively, in Inverted Pendulum and Reacher. The lengths of each trajectory of Inverted Pendulum and Reacher are 1024 and 50, respectively. During the training process, these expert trajectories are disrupted into observation pairs (o_t, o_{t+n}), where o_t, o_{t+n} are RGB images and $n = 3$.

Figure 1 compares our method with two baselines: GAIL and TPIL, to show that our method can imitate from the third-person expert demonstrations and does not need additional expert demonstrations. From the figure, we can see that GAIL with the first-person expert demonstrations (purple) can get the best performance. However, when the third-person demonstrations (green) are used, it performs poorly. TPIL (red) can also achieve good performance, but it has to use the additional non-expert demonstrations. The proposed TPIL-ID-VDB (black) achieves better performance than TPIL when using third-person expert demonstrations.

References

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

Peng, X. B.; Kanazawa, A.; Toyer, S.; and Others. 2018. Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow. *arXiv preprint arXiv:1810.00821*.

Rosin, P. L., and Ellis, T. J. 1995. Image difference threshold strategies and shadow detection. In *BWVC*, 347–356.

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

Stadie, B. C.; Abbeel, P.; and Sutskever, I. 2017. Third-person imitation learning. In *ICLR*.