
Extended Abstract

Zixuan Chen
Nanjing University
Nanjing, China
chenzx@nju.edu.cn

Wenbin Li
Nanjing University
Nanjing, China
liwenbin@nju.edu.cn

Yang Gao
Nanjing University
Nanjing, China
gao@nju.edu.cn

Yiyu Chen
Nanjing University
Nanjing, China
yiyuili@foxmail.com

ABSTRACT

To enable agents to effectively imitate from the third-person visual demonstrations in complex imitation learning (IL) tasks, in this paper, we propose a new IL method, which is named third-person imitation learning by estimating domain cognitive differences (TiLD). The proposed TiLD is able to eliminate the domain cognitive difference between the samples from different perspectives, so as to achieve the purpose of allowing agent to directly learn from the third-person demonstrations. Experimental results indicate that TiLD can achieve significant performance improvements over the existing state-of-the-art IL methods, when dealing with imitation learning tasks with third-person expert demonstrations.

KEYWORDS
imitation Learning; reinforcement Learning; third-person demonstration

1.1 Third-person Imitation Learning by Estimating Domain Cognitive Differences

Under the setting of third-person imitation learning, in our work, we assume that the lack of correspondence between different samples indicates that there is domain differences between them. Formally, the third-person imitation learning setting can be formulated as follows. Suppose that there are two Markov Decision Process $M_{E}$ and $M_{R}$. Suppose further there exists a set of trajectories $\rho = \{\tau_1, \ldots, \tau_n\}$ which were generated under a policy $\pi_E$ acting optimally under some unknown reward $R_{E}$. In the setting of third-person imitation learning, observations are more typically available rather than direct state access, one attempts to recover by proxy through $\rho$ a policy $\pi_R = f(\rho)$ which acts optimally with respect to $R_{R}$ [13]. To handle the third-person setting, specifically, we first differentiate the two adjacent observations $(o, o')$ by difference $\rho = \{(o, o')\}_{i=1}^{n}$, which are generated under a policy $\pi_E$ acting optimally under some unknown reward $R_{E}$. In the setting of third-person imitation learning, observations are more typically available rather than direct state access, one attempts to recover by proxy through $\rho$ a policy $\pi_R = f(\rho)$ which acts optimally with respect to $R_{R}$ [13]. To handle the third-person setting, specifically, we first differentiate the two adjacent observations $(o, o')$ by difference $\rho = \{(o, o')\}_{i=1}^{n}$, which are generated under a policy $\pi_E$ acting optimally under some unknown reward $R_{E}$.
Therefore, the first part of the objective under this assumption, which we called third-person imitation learning by estimating domain cognitive difference (TiLD), then can be written as:

$$\min_{\pi_\theta} \max_{D_\omega} \mathbb{E}_{\pi_\theta} \left[ \log D_\omega (\sigma) \right] + \mathbb{E}_{\pi_\theta} \left[ \log (1 - D_\omega (\sigma)) \right],$$

where $\sigma = D_E (DoG(o', a))$, and $D_E$ is a feature extractor consisting of a convolutional neural network (CNN), which encodes the results of domain difference into a series of low-dimensional, abstract feature representations.

### 1.2 Improved Optimization

For IL tasks involved with complex and less well-specified environments, due to the different observing perspectives, the tilt angles of objects in expert samples and generated samples are different. This will also lead to the discriminator to be too strong and cause the imbalance between $D_\omega$ and $\pi_\theta$, and thus negatively affecting the learning of $\pi_\theta$. The way we address the above problem is to introduce a latent variable $c$, which is obtained by mapping the input samples $\sigma$ to a stochastic encoding $c \sim q_E (c | \sigma)$, into the training of discriminator. $q_E$ represents an Encoder. We utilize an information-theoretic regularization, variational discriminator bottleneck (VDB) [2, 14, 15], to incentivize the model to use $c$ as much as possible, through modulating the accuracy of the discriminator by constraining its information flow [10]. In this way, the objective of our TiLD further becomes:

$$\min_{\pi_\theta} \mathbb{E}_{\pi_\theta} \left[ \log D_\omega (\sigma) \right] + \mathbb{E}_{\pi_\theta} \left[ \log (1 - D_\omega (\sigma)) \right] +$$

$$\epsilon \left( \mathbb{E}_c \left[ KL[q_E (c | \sigma) || h(c)] \right] - I_c \right),$$

where $\epsilon$ is updated adaptively. Notably, in this final version of the objective function, $\pi = \pi_\theta$, that is, we only constrain the information flow from the generated samples to influence the accuracy of discriminator for the generated samples. Since the generated samples are different from the expert samples in the domain, the discriminator can discriminate the generated samples faster and more accurately.

### 2 EXPERIMENTS

To evaluate our algorithm, we conduct the experiments on classical environments in MuJoCo: reacher and inverted pendulum. To construct the third-person imitation learning environment, we first collect expert policies in each environment by running TRPO [12]. Then, we use the expert policies to sample some trajectories, which are composed of some sequences of observations. At the same time, we use a random policy to sample additional expert demonstrations from a third-person perspective. Finally, the state observation angle and environment background are modified to build an environment for agent to make domain differences between the expert demonstrations and generated samples.

We compare TiLD with three baselines to show that our method can effectively imitate from the third-person visual observations without additional expert demonstrations. The results are presented in Fig. 1. From the result we can see that GAIL with the first-person expert demonstrations can get the best performance. The proposed method TiLD can achieve better performance than TPIL and TPIL-ID [7], a simple improved version of TPIL, when using third-person expert demonstrations and is close to GAIL. Moreover, in order to verify the effectiveness of the proposed method, the ablation experimental results using domain difference alone are also given to show that both domain cognitive difference and improved optimization process are useful. The results are shown in Fig. 2, where the performance of TPIL without additional expert demonstrations is mainly used to highlight the effectiveness of the proposed method.

The experimental results in Fig. 1 and Fig. 2 show that our method can effectively solve the third-person imitation learning task without introducing additional expert samples, and can be compared with the reasonable baselines. At the same time, the two innovative points proposed in our method can play a role in improving the performance of the method.

### 3 CONCLUSION

In this work, we propose a novel IL method named third-person imitation learning by estimating domain cognitive differences (TiLD) for third-person imitation learning, which effectively estimates and eliminates most of the domain cognitive difference caused by different observing perspectives without introducing additional expert demonstrations. TiLD allows agent for better learning a policy by observing expert’s behavior from a third-person perspective. For future work, we plan to improve the domain difference module, the next research focus is how to better process the visual observation of the agent, so as to retain more useful information in the visual demonstrations to better guide the agent to perform efficient imitation learning.
ACKNOWLEDGMENTS
This work is supported by Science and Technology Innovation 2030 New Generation Artificial Intelligence Major Project (No.2018AAA0100905), the National Natural Science Foundation of China (No. 62192783), Primary Research and Development Plan of Jiangsu Province (No.BE2021028), Shenzhen Fundamental Research Program (No.2021szsvup056), and in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES