TiLD: Third-person Imitation Learning by Estimating Domain Cognitive Differences of Visual Demonstrations

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
gaoy@nju.edu.cn

Yiyu Chen
Nanjing University
Nanjing, China
yiyuiii@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 ***t**hird-person **i**mitation **l**earning by estimating **d**omain 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

ACM Reference Format:
Zixuan Chen, Wenbin Li, Yang Gao, and Yiyu Chen. 2023. TiLD: Third-person Imitation Learning by Estimating Domain Cognitive Differences of Visual Demonstrations: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Imitation Learning (IL) provides a range of learning framework that enables agents to learn a policy by mimicking the expert behavior, independent of reward signals or interactions with the environment, which greatly improves sample efficiency and saves computational overhead [9]. Behavioral cloning (BC) [3, 11], inverse reinforcement learning (IRL) [1, 5, 8] and adversarial imitation learning [4, 6] are recognized as the three most important IL working lines, while they impose the somewhat unrealistic requirement that the demonstrations should be provided from the *first-person* point of view with respect to the agent: the agent is provided with a sequence of state action pairs that it should have taken. The key challenge of applying third-person demonstrations in IL is that the observations from third-person perspective are different from agent's own observations in terms of angle, background, color and other factors. The lack of correspondence between the two kinds of observations makes it impossible for the agent to directly use the third-person demonstrations to imitate. In this paper, we propose the *Third-person Imitation Learning by estimating Domain cognitive differences (TiLD)* of the visual demonstrations, which take advantage of the domain difference to infer significant regions of motion in successive observations and makes agents to better learn a policy by inferring expert's behavior from a third-person perspective. In TiLD, the domain cognitive difference module aligns as much domain information as possible between two different observation perspectives, such as the angle and background differences. Thus, making it easier for agents to imitate the third-person expert visual demonstrations in complex IL tasks,

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_{π_E} and M_{π_θ}. Suppose further there exists a set of trajectories $\rho = \{(\tau_1, \cdots, \tau_n)\}_{i=0}^{n}$ which were generated under a policy π_E acting optimally under some unknown reward R_{π_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 ρ a policy $\pi_\theta = f(\rho)$ which acts optimally with respect to R_{π_θ} [13]. To handle the third-person setting, specifically, we first differentiate the two adjacent observations (o, o') by difference of Gaussians (DoG) [16] to achieve the information of the moving object related to behavior features, *i.e.*, $\mathrm{DoG}(o', o)$, and then extract the features as the input of the discriminator from $\mathrm{DoG}(o', o)$ by a feature extractor D_{FE}. This process is used to infer and estimate the domain cognitive differences. By 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, retain the most relevant features by domain cognitive difference,and greatly reduce the difference between τ_E (third-person) and τ_π (first-person) from the discriminator perspective.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

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:

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

where $\sigma = D_{FE}\big(\mathrm{DoG}(o', o)\big)$, and D_{FE} 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_ω and π_θ, and thus negatively affecting the learning of π_θ. The way we address the above problem is to introduce a latent variable c, which is obtained by mapping the input samples σ to a stochastic encoding $c \sim \varphi_E(c|\sigma)$, into the training of discriminator. φ_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:

$$\begin{aligned}\min_{\pi_\theta}\max_{D_\omega}&\mathbb{E}_{\pi_\theta}\big[\log D_\omega(\sigma)\big] + \mathbb{E}_{\pi_E}\big[\log(1 - D_\omega(\sigma))\big]+\\ &\varepsilon\Big(\mathbb{E}_{\tilde{\pi}}[\mathrm{KL}[\varphi_E(c|\sigma)||h(c)]] - I_u\Big),\end{aligned} \tag{2}$$

where ε is updated adaptively. Notably, in this final version of the objective function, $\tilde{\pi} = \pi_\theta$, 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.

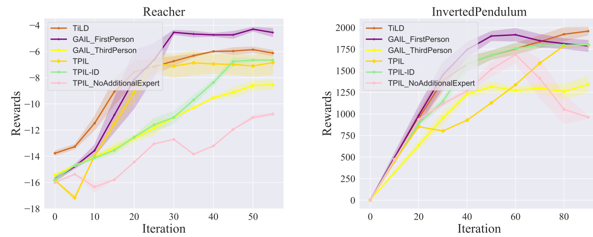

Figure 1: TiLD vs. five baselines: 1) GAIL with first-person demonstrations, 2) GAIL with third-person demonstrations, 3) TPIL with additional expert demonstrations, 4) TPIL-ID, 5) TPIL without additional expert demonstrations.

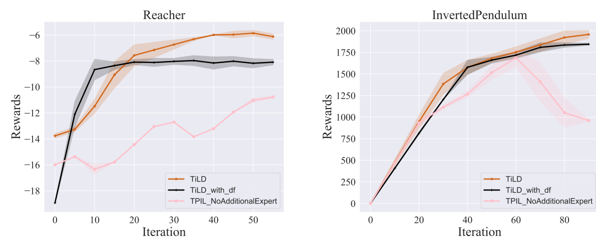

Figure 2: The ablation study for verifying the effectiveness of domain difference module and information-theoretic regularization term.

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 ***t**hird-person **i**mitation **l**earning by estimating **d**omain 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 Developement Plan of Jiangsu Province (No.BE2021028), Shenzhen Fundamental Research Program (No.2021Szvup056), and in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *ICML*. 1.
[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410* (2016).
[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* (2016).
[4] Chelsea Finn, Paul F. Christiano, Pieter Abbeel, and Sergey Levine. 2016. A Connection between Generative Adversarial Networks, Inverse Reinforcement Learning, and Energy-Based Models. *CoRR* (2016).
[5] Chelsea Finn, Sergey Levine, and Pieter Abbeel. 2016. Guided cost learning: Deep inverse optimal control via policy optimization. In *ICML*. 49–58.
[6] Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems* 29 (2016).
[7] Chong Jiang, Zongzhang Zhang, Zixuan Chen, Jiacheng Zhu, and Junpeng Jiang. 2020. Third-person imitation learning via image difference and variational discriminator bottleneck (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13819–13820.
[8] Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *ICML*, Vol. 1. 2.
[9] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. 2018. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics* 7, 1-2 (2018), 1–179.
[10] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. 2018. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *arXiv preprint arXiv:1810.00821* (2018).
[11] Dean A Pomerleau. 1991. Efficient training of artificial neural networks for autonomous navigation. *Neural computation* 3, 1 (1991), 88–97.
[12] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*. PMLR, 1889–1897.
[13] Bradly C Stadie, Pieter Abbeel, and Ilya Sutskever. 2017. Third-person imitation learning. *ICLR* (2017).
[14] Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057* (2000).
[15] Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*. IEEE, 1–5.
[16] Shoujia Wang, Wenhui Li, Ying Wang, Yuanyuan Jiang, Shan Jiang, and Ruilin Zhao. 2012. An Improved Difference of Gaussian Filter in Face Recognition. *J. Multim.* 7, 6 (2012), 429–433.