Multi-Modal Imitation Learning in Partially Observable Environments

Extended Abstract

Zipeng Fu[1], Minghuan Liu[2], Ming Zhou[2], Weinan Zhang[2]
[1]UCLA, [2]Shanghai Jiao Tong University
fu-zipeng@engineering.ucla.edu,{minghuanliu,mingak,wnzhang}@sjtu.edu.cn

ABSTRACT

We consider imitation learning for agents to learn good policies from expert demonstration without any reward signal. Typical methods focus on single-expert single-task imitation in a fully observable environment. In practice, however, agents mostly make decisions based on their local observations, and the ability to generalize across various experts' behaviors and multiple tasks is crucial for practical imitation learning. In this paper, we propose to take advantage of InfoGAIL with RNN-based belief state representations for multi-modal imitation learning in partially observable environments. We confirm the effectiveness of multi-expert learning of our method in a 2-dimensional environment, in which expert trajectories consist of two human-distinguishable behaviors. Further experimental results in continuous-control locomotion tasks reveal that our method can also disentangle interpretable latent factors in unlabeled multi-task demonstrations.

KEYWORDS

Imitation Learning; Reinforcement Learning; Multi-Task Learning; Interpretable Adversarial Learning; POMDPs

1 INTRODUCTION

By learning from pre-collected expert demonstrations, Imitation learning (IL) alleviates the difficulty in deployment of reinforcement learning (RL), a method starving for interactions with environments paired with appropriate reward functions that are hard to design in many scenarios [2, 18]. Existing IL methods, e.g. Behavior Cloning (BC) [17], Inverse Reinforcement Learning (IRL) [1, 16] and Generative Adversarial Imitation Learning (GAIL) [13], are restricted to learning from clear single-expert demonstrations for a single task under full observability assumption. However, in reality, expert demonstrations are usually collected by several experts, who may have different levels of expertise, expertise in different tasks, or different preferences even at the same situations. Therefore, researchers have proposed methods to tackle the diversity in complex multi-modal expert demonstrations, including multi-expert [15] and multi-task [11, 22] ones. These methods are nonetheless still inadequate to be appropriately applied in real-world tasks in partially observable environments.

To that end, in this paper, we adopt recent advances in deep representation learning that enables the agent to encode its partial observation history into a belief state to help decision making, including recurrent-neural-network based methods [8, 10] and attention based ones [5, 23]. Specifically, we extend InfoGAIL to Partially Observable Markov Decision Process (POMDP) settings with RNN-based belief state representations to help solve real-world imitation learning problems. Unlike other approaches, where decoupled learning is used [9], we jointly learn the belief state representations with the policy and critic modules. Experiments are conducted in two partially observable environments, including a 2D environment and a continuous-control locomotion environment. We show the effectiveness of our method to disentangle human-interpretable latent factors and learn from unlabeled multi-expert behaviors along with multi-task demonstrations in a self-supervised fashion.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 2020, Auckland, New Zealand

https://doi.org/doi

2 PRELIMINARIES

2.1 Partially Observable Environments

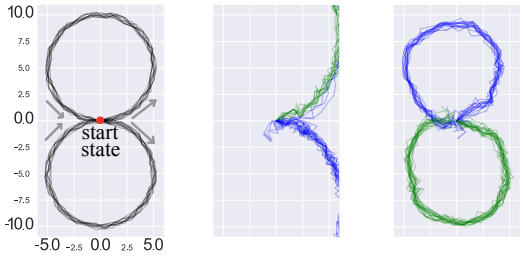
We formulate a partially observable enviroment as a POMDP, a 6-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{Z}, O \rangle$, where $\mathcal{S}$, $\mathcal{A}$, $\mathcal{Z}$ denote the state, action and observation space respectively. $\mathcal{P}$ is the dynamic function. At each time step t, the agent only perceives the partial observation $z_t \in \mathcal{Z}$ w.r.t the underlying state $s_t \in \mathcal{S}$, characterized by the observation function $O : \mathcal{S} \to \mathcal{Z}$. The agent chooses an action relying on its policy $\pi(a_t|z_t) : \mathcal{Z} \times \mathcal{A} \to [0, 1]$. The reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ provides the reward feedback. The objective of the agent is to maximize its discounted expected return $\mathbb{E}_{a_t \sim \pi(\cdot|s_t), s_t \sim p(\cdot|s_{t-1}, a_{t-1})} \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \right]$. Specifically, when the observation function O becomes an identity mapping such that $z_t = s_t$, s_t is fully observable to the agent, which leads to a classic Markov Decision Process (MDP).

2.2 Multi-Modal Imitation Learning

Many recently proposed methods in imitation learning are based on Generative Adversarial Imitation Learning (GAIL), showing high effectiveness in learning the agent's policy. To tackle unlabeled multi-modal demonstrations and inspired by InfoGAN [6], Li et al. [15] and Hausman et al. [11] both propose to correlate the learned policy with latent variables c through mutual information regularization. In this paper, we mainly follow InfoGAIL [15] with a formal objective as:

$$\min_{\theta, \psi} \max_{\omega} \mathbb{E}_{\pi_\theta}[\log(D_\omega(s, a))] + \mathbb{E}_{M_E}[\log(1 - D_\omega(s, a))] - \lambda_1 L_{\mathrm{MI}}(\pi, P_\psi) - \lambda_2 H(\pi_\theta), \tag{1}$$

where M_E is the expert policy, L_{MI} is the variational lower bound of the mutual information between latent code c and the state-action pairs generated from π_θ. Define P_ψ as the approximated posterior



(a) Expert (b) InfoGAIL (c) Ours

Figure 1: Multi-expert 2D trajectories. Unlabeled Experts' demonstrations have two modes: starting from (0, 0), go anti-clockwise or clockwise. Blue and Green represent two distinguishable sets of optimal behaviors. Our method precisely recovers the multi-modal policy while InfoGAIL fails.

to be optimized, MI as the mutual information, and it follows:

$$L_{\mathrm{MI}}(\pi, P_\psi) = \mathbb{E}_{c\sim p(c), a\sim\pi_\theta(\cdot|s,c)}[\log P_\psi(c|s,a)] + H(c), \quad (2)$$
$$\leq \mathrm{MI}(c;(s,a)). \quad (3)$$

3 METHOD

In real-world environments, it is difficult to reveal full states to construct MDP settings for agent learning, thus POMDP is better for modeling.

A general technique for learning in a POMDP relies on Recurrent Neural Networks (RNN) [10], which embeds the history trajectory $\tau_{t-1} = \{(z_i, a_i) \mid i = 0, \ldots, t-1\}$ as a hidden state b_{t-1}. Combining the current observation z_t, action a_t and previous state embedding b_{t-1}, the RNN network produces a new state embedding b_t. Formally, we can represent the state embedding b_t with RNN parametererized by ϕ as below:

$$b_t(\tau_t;\phi) = \mathrm{RNN}_\phi(b_{t-1}(\tau_{t-1};\phi), z_t, a_{t-1}). \quad (4)$$

Specifically, we extend InfoGAIL to learn from multi-modal demonstrations in partially observable environments with learning the underlying state representation using separate RNN models in all modules (π, D, P). Asides, the traditional minimax objective of GAN suffers from mode collapse and vanishing gradient [4, 19]; thus we apply Wasserstein GAN [3] to stabilize training. The objective is as below:

$$\min_{\theta,\psi}\max_{\omega} \mathbb{E}_{a\sim\pi_\theta(\cdot|b(\tau;\theta),c)}[D_\omega(b(\tau;\omega),a)] - \lambda_1 L_{\mathrm{MI}}(\pi_\theta, P_\psi)$$
$$- \mathbb{E}_{a\sim M_E}[D_\omega(b(\tau;\omega),a)] - \lambda_2 H(\pi_\theta), \quad (5)$$

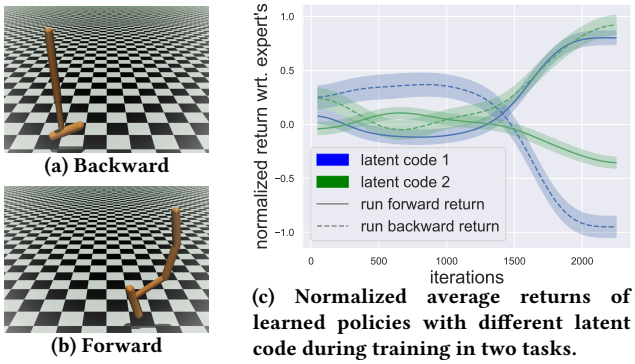
where L_{MI} is the variational lower bound:

$$L_{\mathrm{MI}}(\pi_\theta, P_\psi) = \mathbb{E}_{c\sim p(c),\pi_\theta}[\log P_\psi(c|b(\tau;\psi))] + H(c), \quad (6)$$
$$\leq \mathrm{MI}(c; b(\tau;\psi)). \quad (7)$$

Unlike InfoGAIL, we aims to maximize the mutual information between latent code and belief representations of generated trajectories. Unlike belief-module imitation learning in [8], our method requires no belief regularization to avoid mode collapses, since the belief state representation for π, D and P are modeled via three independent RNN models: $b(\cdot;\theta)$, $b(\cdot;\omega)$ and $b(\cdot;\psi)$.

4 EXPERIMENTS

We choose Proximal Policy Optimization (PPO) [21] with Generalized Advantage Estimation (GAE) [20] to train agents' policies for all experiments. During the training procedure, Adam optimizer [14] is used to for learning parameter θ and ψ, and RMSProp [12] for parameter ω, where we set γ as 0.99, λ in GAE as 0.95, and the learning rate is 2×10^{-4}. We use GRUs [7] with 128 hidden cells and bootstrapped random update [10], such that the architecture has approximately the same number of network parameters and similar training computational complexity as InfoGAIL. Latent code is uniformly sampled.



(a) Backward

(b) Forward

(c) Normalized average returns of learned policies with different latent code during training in two tasks.

Figure 2: Multi-task MuJoCo control in Hopper-v3.

4.1 Multi-Expert 2D Trajectories

We first confirm multi-expert learning of our method in a 2D environment, in which expert demonstrations consist of two distinguishable sets of optimal behaviors. Starting from the origin, both anti-clockwise rotation around $(0, 5)$ and clockwise rotation around $(0, -5)$ are optimal for returning back. The observation at each time step is a 2D coordinate of the current position of the agent, and the action is the moving direction with a unit-length movement. The log standard deviation of noise in moving direction is 1. The setting is harder than the similar experiment in [15] for less observations. As shown in Fig. 1, we plot the trajectories sampled from learned policies with our methods against InfoGAIL. In our implementation, BC and GAIL do not converge, which randomly end up with one mode and neglect the other, thus are not presented.

4.2 Multi-Task MuJoCo Control

In this section, we show our method can effectively learn policies from multi-task expert demonstrations conducted on Hopper-v3, a continuous control task simulated with MuJoCo [24]. As shown in Fig. 2a and Fig. 2b, the agent is required to learn to go Forward and Backward and receives rewards for both tasks respectively. We set the partial observation as xyz-positions and angles, but no xyz-velocities or angular velocities. We form the unlabeled expert demonstrations with equivalent number of trajectories sampled with experts in two tasks separately.

We use the true rewards obtained from the simulated environment for the evaluation. Fig. 2c presents the normalized average returns w.r.t. the expert returns for both tasks during training. The policy learned with latent code 1 converges to gain near-expert returns for the Forward task, while the code 2 specializing for the Backward. Our method succeeds in disentangling different tasks during learning from unlabeled expert demonstrations with partial observations, and recovers expert behaviors with over 76% returns in both tasks.

REFERENCES

[1] Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*. ACM, 1.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, International Convention Centre, Sydney, Australia, 214–223. http://proceedings.mlr.press/v70/arjovsky17a.html

[4] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. 2017. Generalization and equilibrium in generative adversarial nets (gans). In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 224–232.

[5] Elaheh Barati, Xuewen Chen, and Zichun Zhong. 2019. Attention-based Deep Reinforcement Learning for Multi-view Environments. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1805–1807.

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*. 2172–2180.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[8] Tanmay Gangwani, Joel Lehman, Qiang Liu, and Jian Peng. 2019. Learning Belief Representations for Imitation Learning in POMDPs. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*. 383. http://auai.org/uai2019/proceedings/papers/383.pdf

[9] David Ha and Jürgen Schmidhuber. 2018. Recurrent World Models Facilitate Policy Evolution. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 2450–2462. http://papers.nips.cc/paper/7512-recurrent-world-models-facilitate-policy-evolution.pdf

[10] Matthew Hausknecht and Peter Stone. 2015. Deep recurrent q-learning for partially observable mdps. In *2015 AAAI Fall Symposium Series*.

[11] Karol Hausman, Yevgen Chebotar, Stefan Schaal, Gaurav Sukhatme, and Joseph J Lim. 2017. Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets. In *Advances in Neural Information Processing Systems*. 1235–1245.

[12] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. [n. d.]. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. ([n. d.]).

[13] Jonathan Ho and Stefano Ermon. 2016. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 4565–4573. http://papers.nips.cc/paper/6391-generative-adversarial-imitation-learning.pdf

[14] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1412.6980

[15] Yunzhu Li, Jiaming Song, and Stefano Ermon. 2017. Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Information Processing Systems*. 3812–3822.

[16] Andrew Y Ng and Stuart J Russell. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 663–670.

[17] Dean A Pomerleau. 1991. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation* 3, 1 (1991), 88–97.

[18] S.J. Russell, S.J. Russell, P. Norvig, and E. Davis. 2010. *Artificial Intelligence: A Modern Approach*. Prentice Hall. https://books.google.com/books?id=8jZBksh-bUMC

[19] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*. 2234–2242.

[20] John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. 2016. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. http://arxiv.org/abs/1506.02438

[21] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[22] Mohit Sharma, Arjun Sharma, Nicholas Rhinehart, and Kris M. Kitani. 2019. Directed-Info GAIL: Learning Hierarchical Policies from Unsegmented Demonstrations using Directed Information. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. https://openreview.net/forum?id=BJeWUs05KQ

[23] Ivan Sorokin, Alexey Seleznev, Mikhail Pavlov, Aleksandr Fedorov, and Anastasiia Ignateva. 2015. Deep attention recurrent Q-network. *arXiv preprint arXiv:1512.01693* (2015).

[24] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5026–5033.