Week 5 Paper Review

Niharika Shrivastava
School of Computing
National University of Singapore
Singapore, 119077
niharika@comp.nus.edu.sg

Abstract

This is a brief review of [1], [2], [3].

1 A Divergence Minimization Perspective on Imitation Learning Methods

Prior literature empirically proved that adversarial IRL approaches outperform BC by a huge margin in the low-data regime, even though at optimality both methods can recover the expert policy exactly. To this effect, the authors propose f-MAX, a theoretical framework that enables comparison between several IRL methods in order to understand their underlying algorithmic properties and performance-affecting differences from a divergence-minimization perspective.

1.1 Unified Framework

Using f-MAX, various IL algorithms are unified by writing them as the minimization of some statistical divergence between the expert policy and the learnt policy, e.g., AIRL minimizes the reverse KL divergence. This shows that while BC tries to match only the conditional action distributions, other IRL methods in addition try to match the marginal state distributions. Moreover, BC minimizes the forward KL divergence, which is different from divergences that exhibit more mode-seeking behaviour. Combining these give explicit explanations for the difference in performances of BC and IRL methods and conclusively emphasize the importance of expert marginal-state distribution matching for improved learning rather than action-distribution matching.

1.2 FAIRL

FAIRL is the forward KL divergence counterpart of AIRL. Due to its mode-covering properties rather than mode-seeking, it starts to place mass in low-probability regions and gradually moves towards modes of the expert's state-action distribution. Since FAIRL outperforms BC in spite of being a forward KL divergence formulation, it conclusively supports the performance gain of IRL methods as the result of their objectives explicitly encouraging the policy to match the marginal state distribution of the expert in addition to the matching of conditional action distribution, instead of the direction of KL divergence used.

Lastly, they demonstrate how several tasks can be formulated as a state-only marginal matching problem using simple state distributions, even though their performance in the experiments didn't outperform SOTA.

Preprint. Under review.

2 SQIL: Imitation Learning via Reinforcement Learning with Sparse Rewards

The authors propose SQIL, a regularised version of Behavioural Cloning (BC), and empirically show how it exhibits comparable performance to complex IRL methods, thereby eliminating the need to use learned reward functions. The agent is given a constant reward of +1 for matching demonstrated state-action pairs and a reward of 0 otherwise, along with implicit information about state transition dynamics. This serves as an incentive to imitate the expert in demonstrated states over a long horizon and to take actions that lead it back to demonstrated states when it strays from the demonstrations.

SQIL is equivalent to augmenting BC with a regularization term (sparsity prior) that incorporates information about the state transition dynamics into the imitation policy, and thus enables long-horizon imitation. It is built on soft Q-learning, which assumes that expert behaviour follows the maximum entropy model. This formulation enables high Q values for actions that lead to states from which the demonstrated states are reachable, thereby solving the state-distribution mismatch problem in BC.

Simulation experiments on tasks with low-dimensional, continuous observations and unknown dynamics show that SQIL outperforms BC and achieves competitive results compared to GAIL, while being simple to implement on top of existing off-policy RL algorithms.

3 Concept2Robot: Learning Manipulation Concepts from Instructions and Human Demonstrations

The authors propose a framework for multi-task policy learning to acquire manipulation concepts using natural language instructions and video demonstrations. It's a two-step framework where individual policies are first trained exclusively on singular tasks via RL using natural language instructions and the starting scene of the environment as the input, followed by learning a unified multi-task policy by aggregating the singular policies using imitation learning. The output of this framework is a complete trajectory implemented by the robot in order to complete the task successfully. A supervised video classifier is used to evaluate the single-task policies which also serves as a reward signal for the RL.

The framework is trained and validated on 78 complex manipulation tasks and several design choices are discussed via ablation studies in support of the end-to-end framework. Specifically, it is shown that it is essential to capture the velocities of the motion along with its goal poses and a video classifier is superior to a start-end image frames classifier. Even though the framework shows some generalization in providing natural language instructions, however, it doesn't accommodate more complex out-of-distribution instructions and/or the combination of multiple instructions as one. This paper proves to be a good direction for training multi-task policies but has massive scope for improvement.

References

[1] Ghasemipour, S. K., Zemel, R., & Gu, S. (2019). A Divergence Minimization Perspective on Imitation Learning Methods. ArXiv. https://doi.org/10.48550/arXiv.1911.02256

[2] Reddy, S., Dragan, A. D., & Levine, S. (2019). SQIL: Imitation Learning via Reinforcement Learning with Sparse Rewards. ArXiv. https://doi.org/10.48550/arXiv.1905.11108

[3] Shao L, Migimatsu T, Zhang Q, Yang K, Bohg J. Concept2Robot: Learning manipulation concepts from instructions and human demonstrations. The International Journal of Robotics Research. 2021;40(12-14):1419-1434. doi:10.1177/02783649211046285