VARIATIONAL MODEL-BASED IMITATION LEARNING IN HIGH-DIMENSIONAL OBSERVATION SPACES

Rafael Rafailov[1], **Tianhe Yu**[1], **Aravind Rajeswaran**[2], **Chelsea Finn**[1]
{rafailov, tianheyu, cbfinn}@stanford.edu, aravraj@cs.washington.edu
[1] Stanford University, [2] University of Washington

ABSTRACT

We consider the problem setting of imitation learning where the agent is provided a fixed dataset of demonstrations. While the agent can interact with the environment for exploration, it is oblivious to the reward function used by the demonstrator. This setting is representative of many applications in robotics where task demonstrations may be straightforward while reward shaping or conveying stylistic aspects of human motion may be difficult. For this setting, we develop a variational model-based imitation learning algorithm (VMIL) that is capable of learning policies from visual observations. Through experiments, we find that VMIL is more sample efficient compared to prior algorithms in several challenging vision-based locomotion and manipulation tasks, including a high-dimensional in-hand dexterous manipulation task. All results including videos can be found online at https://sites.google.com/view/vim.

1 INTRODUCTION

The ability of reinforcement learning (RL) agents to autonomously learn by interacting with environment presents a promising approach for learning diverse skills. However, reward specification has remained a major challenge in deployment of RL in practical settings (Amodei et al., 2016; Everitt & Hutter, 2019; Rajeswaran* et al., 2018). The ability to imitate humans or other expert trajectories enables us to avoid the reward specification problem while also circumventing hard exploration challenges in RL. It also presents a more natural way to teach robots various tasks and skills. In this work, we aim to develop a new algorithm that can learn from limited expert demonstration data and scale to high-dimensional observation and action spaces. Behaviour cloning (BC) is a classic algorithm to imitate expert demonstrations (Pomerleau, 1988). It uses supervised learning to greedily match the expert behaviour at demonstrated expert states, without regard for the marginal state distribution. However, due to environment stochasticity or compounding actor errors over time, the agent will drift from the expert state distribution towards out-of-distribution states and ultimately fail to mimic the demonstrations (Ross et al., 2011). This problem is especially prevalent in high-dimensional domains or domains with limited data. While a wide policy coverage by the expert (Spencer et al., 2021) or the ability to repeatedly and interactively query the expert policy (Ross et al., 2011) can circumvent difficulties to an extent, such conditions are seldom met in practical applications. An alternate line of work based on inverse RL (IRL) (Finn et al., 2016b; Fu et al., 2018a) and adversarial imitation learning (Ho & Ermon, 2016) aim to not only match actions at demonstrated states, but also the long term visitation distribution (Ghasemipour et al., 2019). These approaches explicitly train a GAN-based classifier (Goodfellow et al., 2014; Finn et al., 2016a) to estimate the expert state-action distribution, which serves as a reward for training an RL agent. While these methods have achieved substantial improvement over behaviour cloning without additional expert supervision, they are difficult to deploy in realistic scenarios. This is mainly due to three reasons: (1) the need to optimize an RL objective that requires on-policy data collection and contributes towards sample complexity; (2) non-stationarity of the reward function that changes as the RL agent learns; and (3) high-dimensionality of the observation space that requires representation learning.

To overcome the aforementioned difficulties, we develop a new model-based adversarial imitation learning algorithm that trains a variational dynamics model. Our approach has several advantages: (1) model learning serves as a rich auxiliary task for learning visual representations that support efficient learning of control policies in a stable fashion, (2) we optimize the correct (on-policy) imitation learning objective, (3) model-based policy optimization allows us to train an agent on-policy without the sample complexity required by these algorithms. On four simulated environments with image observations and continuous actions, including a vision-based dexterous manipulation task, we find that our algorithm is significantly more performant than a prior state-of-the-art model-free method.

2 Related Work

We begin by reviewing the literature on imitation learning and recent advances in image-based RL.

Imitation Learning. Ho & Ermon (2016) develop GAIL - a practical imitation learning algorithm with a fixed amount of demonstration data, without the ability to further query the expert. The key component is to train a binary classifier $D_\psi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ that discriminates between expert and policy rollouts similar to a GAN. The agent's policy π is then trained using reinforcement learning to maximize $\mathbb{E}_{\tau\sim\pi}[\sum_t \log D_\psi(s, a)] - \lambda\mathcal{H}(\pi)$, where $\mathcal{H}(\pi)$ is an entropy regularizer. Fu et al. (2018a) develop a similar algorithm, but like GAIL it still requires on-policy data. Several off-policy algorithms have been developed (Reddy et al., 2020; Blondé & Kalousis, 2019; Kostrikov et al., 2019) to tackle sample inefficiency, largely by replacing an on-policy RL algorithm (TRPO, PPO) with an off-policy one. However, this introduces a distribution mismatch as it replaces the expectation under the policy state-action marginal distribution with the expectation under the distribution of the replay buffer. While this tends to work in practice, Kostrikov et al. (2020a) show that the distribution mismatch can significantly degrade performance if the buffer distribution is significantly off from the policy distribution. Alternatively Finn et al. (2016b) use a similar approach to reward learning in combination with a learned locally linear dynamics model, which makes policy optimization easier. In this work we extend this approach to high-dimensional POMDPs with arbitrary dynamics.

Learning From Images. Reinforcement learning from images is an inherently difficult task, since the agent needs to learn meaningful visual representations from reward signal alone. A recent line of research (Gelada et al.; Hafner et al., 2019; Lee et al., 2020; Hafner et al., 2020) train a variational model of the image-based environment as an auxiliary task. Lee et al. (2020) use the model for representation learning purposes only and train an off-policy SAC-based algorithm on top of the latent representation. On the other hand, Hafner et al. (2020) use rollouts from the learned model to train an agent in an on-policy fashion within the latent space of the model. We base our algorithm on this approach, as it allows us to train an agent on-policy and target the correct objective of imitation learning, while still maintaining the sample efficiency of off-policy approaches. Separately Kostrikov et al. (2020b) introduce DrQ, a sample efficient approach for off-policy Q-learning from images in a completely model-free way by introducing image-augmentation for both the target and actor Q-networks. The authors report state of the art performance on both asymptotic returns and sample efficiency. We later empirically compare our method to SQIL (Reddy et al. (2020)), an established off-policy imitation learning algorithm combined with DrQ image augmentation.

3 Model-Based Imitation Learning in POMDPs

In this section, we consider learning in image-based domains as a POMDP.

3.1 Modeling via control as inference in POMDPs

Following the control as inference approach (Levine, 2018; Lee et al., 2020), we jointly model the POMDP dynamics and policy learning. We choose choose a joint variational distribution that factors into state inference, latent dynamics, and a policy term. However, unlike prior works (Lee et al., 2020), we base the policy distribution entirely on the inferred latent state. This later allows us to optimize the policy in a model-based fashion under the learned latent model. Our joint variational distribution is as follows:

$$q(\boldsymbol{z}_{1:T}, \boldsymbol{a}_{\tau+1:T}|\boldsymbol{x}_{1:\tau+1}, \boldsymbol{a}_{1:\tau}) = \prod_{t=0}^{\tau} q(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{t+1}, \boldsymbol{z}_t, \boldsymbol{a}_t) \prod_{t=\tau+1}^{T-1} p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t, \boldsymbol{a}_t) \prod_{t=\tau+1}^{T-1} \pi(\boldsymbol{a}_t|\boldsymbol{z}_t)$$

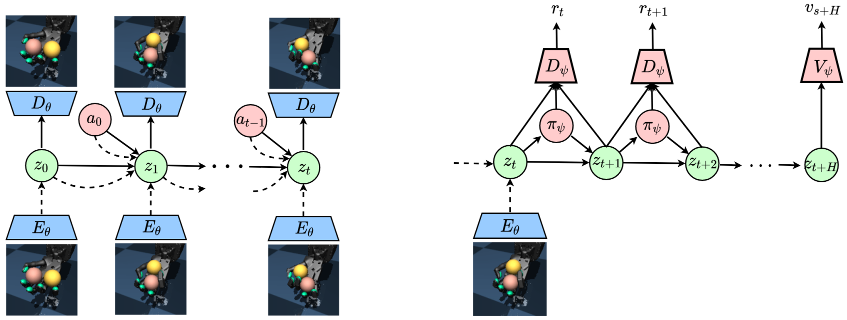

Figure 1: Left: Training procedure for our variational model. Dashed lines represent inference and solid lines represent the generative model. Right: Policy training procedure for our model.

Using this variational distribution, the evidence lower bound becomes:

$$\log p(\boldsymbol{x}_{1:\tau+1}, \mathcal{O}_{\tau+1:T}|\boldsymbol{a}_{1:\tau}) \geq \mathop{\mathbb{E}}_{(\boldsymbol{z}_{1:T}, \boldsymbol{a}_{\tau+1:T})\sim q} \Bigg[\underbrace{\sum_{t=\tau+1}^{T} \big(\log p(\mathcal{O}_t = 1|\boldsymbol{z}_t, \boldsymbol{a}_t) + \log p(\boldsymbol{a}_t) - \log \pi(\boldsymbol{a}_t|\boldsymbol{z}_t)\big)}_{\text{policy objective}}$$

$$+ \underbrace{\sum_{t=0}^{\tau} \big(p(\boldsymbol{x}_{t+1}|\boldsymbol{z}_{t+1}) - \mathbb{D}_{KL}(q(\boldsymbol{z}_{t+1}|\boldsymbol{x}_{t+1}, \boldsymbol{z}_t, \boldsymbol{a}_t)||p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t, \boldsymbol{a}_t))\big)}_{\text{model objective}} \Bigg] \tag{1}$$

All the distributions are represented as diagonal Gaussians with the mean and variance parameterized by the output of neural networks.

3.2 Practical Algorithm and Training

We optimize the two different components of Equation 1 in an alternating fashion as shown in Figure 1. We utilize the latent dynamics model by Hafner et al. (2019), with the exception that we do not include a reward predictor. Following Fu et al. (2018b) we can consider $\log p(\mathcal{O}_t = 1|\boldsymbol{z}_t, \boldsymbol{a}_t)$ as the probability that the state-action pair belongs to an expert "event" distribution. We train a discriminator on top of the latent representation using the following objective:

$$\min_{D_\psi} \mathbb{E}_{q_\theta, \pi_\psi} \Big[\sum_{t=1}^{T-1} \log D_\psi(\boldsymbol{z}_t^\pi, \boldsymbol{a}_t^\pi, \boldsymbol{z}_{t+1}^\pi) + \log(1 - D_\psi(\boldsymbol{z}_t^E, \boldsymbol{a}_t^E, \boldsymbol{z}_{t+1}^E)) \Big] \tag{2}$$

as we find that including the next state helps with alleviating model exploitation. We also add Gaussian input noise (Müller et al., 2019) as it helps improve convergence speed and stability. At this point, we optimize the policy objective as a reinforcement learning problem with a reward function $r(\boldsymbol{z}_t, \boldsymbol{a}_t, \boldsymbol{z}_{t+1}) = D_\psi(\boldsymbol{z}_t, \boldsymbol{a}_t, \boldsymbol{z}_{t+1})$. We can use several algorithms to optimize this component, but we would like to (1) avoid reinforcement learning with a non-stationary reward and (2) use an on-policy training algorithm in order to target the correct objective. Because of these goals we opt for a model-based value expansion approach based on Hafner et al. (2020).

4 Experiments

In our experiments we want to answer several questions: (1) can VMIL successfully scale to environments with image observations, (2) how does VMIL compare to state of the art model-free approaches, (3) how does VMIL perform in different data regimes, (4) can VMIL solve realistic manipulation tasks, (5) can VMIL solve environments with complex interaction and physics?

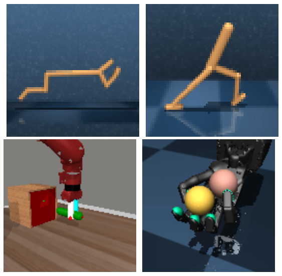

Figure 2: Environments.

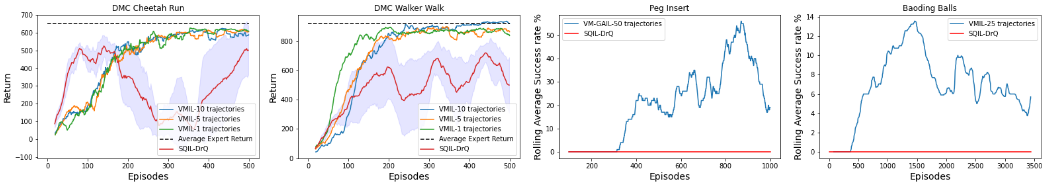

Figure 3: Learning curves for VMIL versus the model-free baseline of SQIL+DrQ.

Tasks. Our test environments are shown in Figure 3.2. We first evaluate our method on the Cheetah and Walker tasks from the DeepMind Control Suite (Tassa et al., 2018). We also consider a Sawyer peg-insertion task, where the arm and peg positions are randomized, which creates a difficult configuration for imitation learning when the configuration is out of the expert distribution. Our final environment is the Baoding balls task from Nagabandi et al. (2019). This is a challenging task for policy learning, even in the state-based case. All observations are from images, while the Baoding balls task additionally includes robot proprioception. The DeepMind Control Suite tasks have between 1 and 10 demonstration rollouts, generated using SAC training from state. The Sawyer task uses 50 demos generated using an analytic controller. Finally the Baoding balls uses 25 demos generated using the method from Nagabandi et al. (2019).

Results. Experiment results are shown in Figure 3. To answer questions (1) and (2) above we compare VMIL to SQIL with DrQ image augmentations, as introduced in Section 2, on the Cheetah and Walker tasks. While the model-free baseline initially outperforms our method, it's performance has high variance between random seeds and exhibits significant instability. We hypothesize that this is due to (1) sparse rewards associated with imitation learning, which prevent the development of reliable image features and (2) the static nature of the reward in the SQIL method, as only expert demonstrations receive positive reward. The addition of a discriminator-based training, similar Kostrikov et al. (2019); Blondé & Kalousis (2019), might help alleviate these issues, however that would introduce the problem of representation learning using non-stationary reward signal. To answer question (3), we vary the number of expert trajectories available to our method, results shown in the first two plots of Figure 4. We find that performance is comparable and we still get stable convergence with only a single expert rollout. To answer question (4), we evaluate our method on a Sawyer peg-insertion tasks, where it reaches over 50% success rate, while the model-free baseline fails to complete the task. Towards the end of training we see meaningful degradation in performance. We hypothesize that this is due to the evolution of the latent space representation in the course of training. A potential solution would be annealing the amount of input noise to the discriminator or exploring alternative stabilization techniques, (Blondé et al. (2020)). To answer question (5) we deploy our method on the Baoding Balls task from Nagabandi et al. (2019). This is a complex task, which is difficult to solve with policy training, even in the state-based domain. We manage to reach up to 14% success rate, while the model-free baseline fails to solve the task, however we encounter similar degradation in performance, as in the peg-insertion task.

5 Conclusion

In this work we presented VMIL, a model-based imitation learning algorithm that works from high-dimensional observations, such as images. Our method achieves better asymptotic returns and is more stable and sample efficient than state-of-the art off-policy model-free approaches in the same domain. The improved sample efficiency and safety of the learning procedure make real-world imitation learning feasible and we plan to evaluate our algorithm on a real robot setting. Although we observe some degradation in performance on environments with out-of-distribution states and more complex dynamics, we believe that by utilizing additional stabilization and reward shaping techniques we can fully solve those cases as well. Finally the use of task-agnostic dynamics models opens many doors to future research, such as offline, multi-task, continual and meta-imitation learning.

REFERENCES

Dario Amodei, Chris Olah, J. Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *ArXiv*, abs/1606.06565, 2016.

Lionel Blondé and Alexandros Kalousis. Sample-efficient imitation learning via generative adversarial nets. *AISTATS*, 2019.

Lionel Blondé, Pablo Strasser, and Alexandros Kalousis. Lipschitzness is all you need to tame off-policy generative adversarial imitation learning, 2020.

Tom Everitt and Marcus Hutter. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *ArXiv*, abs/1908.04734, 2019.

Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *ArXiv Preprint*, 2016a.

Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, pp. 49–58. PMLR, 2016b.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *International Conference on Learning Representations*, 2018a.

Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine. Variational inverse control with events: A general framework for data-driven reward definition. *Conference on Neural Information Processing Systems*, 2018b.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. Deepmdp: Learning continuous latent space models for representation learning.

Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. *Conference on Robot Learning*, 2019.

Ian J. Goodfellow, Jean Pouget-Abadie, M. Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *International Conference on Machine Learning*, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *International Conference on Learning Representations*, 2020.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Conference on Neural Information Processing Systems*, 2016.

Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. *International Conference on Learning Representations*, 2019.

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. *International Conference on Learning Representations*, 2020a.

Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels, 2020b.

Alex X. Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *Conference on Neural Information Processing Systems*, 2020.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *ArXiv Preprint*, 2018.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *Conference on Neural Information Processing Systems*, 2019.

Anusha Nagabandi, Kurt Konolige, Sergey Levine, and Vikash Kumar. Deep dynamics models for learning dexterous manipulation. *Conference on Robot Learning*, 2019.

Dean A Pomerleau. Alvinn: an autonomous land vehicle in a neural network. In *Proceedings of the 1st International Conference on Neural Information Processing Systems*, pp. 305–313, 1988.

Aravind Rajeswaran*, Vikash Kumar*, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018.

Siddharth Reddy, Anca D. Dragan, and Sergey Levine. Sqil: Imitation learning via reinforcement learning with sparse rewards. *International Conference on Learning Representations*, 2020.

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. *AISTATS*, 2011.

Jonathan Spencer, Sanjiban Choudhury, Arun Venkatraman, Brian Ziebart, and J. Andrew Bagnell. Feedback in imitation learning: The three regimes of covariate shift. *ArXiv Preprint*, 2021.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite, 2018.