http://lamda.nju.edu.cn

Learning Belief Representations for Imitation Learning in POMDPs

Presented by Wei-jian Liao

Background

- POMDP is defined as a tuple $< S, A, T, R, O, \Omega >$, where Ω is the observation set and O is the observation model $O(\, o \mid s, a\,)$
- Belief State is originally defined as a probability distribution over the true states of the underlying environment.
  - when the state is discrete and model is known:

$$b'(s') \propto O(\, o \mid s', a\,) \sum_s T(\, s' \mid a, s\,) b(s)$$

- Solving POMDP now means that find a mapping from belief to optimal policy, i.e. $\pi(b)$
- In the context of data-driven methods, we can learn belief from historical observation-action sequence, i.e. , $b_t = \phi(h_t)$ where $h_t = (o_{\le t}, a_{<t})$

Background

- Learning from demonstrations is also regarded as one of transfer learning methods*
- GAIL aims to minimize $D_{JS}[\rho_{\pi}(s,a)||\rho_{E}(s,a)]$, and its min-max objective is :

$$\min_{\theta}\max_{w}\mathbb{E}_{(s,a)\sim\pi_{\theta},\mathcal{T}}\left[\log\left(1-D_{w}(s,a)\right)\right]$$
$$+\mathbb{E}_{(s,a)\sim M_{E}}[\log D_{w}(s,a)]$$

- The main contribution of this paper is to extend GAIL to POMDP setting

*Zhu Z, Lin K, Zhou J. Transfer Learning in Deep Reinforcement Learning: A Survey[J]. arXiv preprint arXiv:2009.07888, 2020

How to do

$$\min_{\theta} \max_{w} \mathbb{E}_{(b,a)\sim\pi_{\theta},T}\left[\log\left(1 - D_{w}(b,a)\right)\right] + \mathbb{E}_{(b,a)\sim M_E}\left[\log D_{w}(b,a)\right]$$

The idea is very simple but there are lots of problem to be solved

- What is the relationship between $D_{JS}[\rho_{\pi}(s)|\rho_{E}(s)]$ and $D_{JS}[\rho_{\pi}(b)|\rho_{E}(b)]$
- How to incorporate belief learning into this framework
- ...and make it work

How to do

Divide the architecture into two modules

- Policy module: $\pi_\theta(a_t|b_t)$ learns a distribution over actions, conditioned on the belief
  - Trained with imitation learning
- Belief module: B_ϕ learns a good representation of the belief $b_t = B_\phi(h_t)$ where $h_t = (o_{\leq t}, a_{<t})$
  - Trained in a task-agnostic manner or in a task-aware manner

Policy Module

We can minimize $D_{JS}[\rho_\pi(s)||\rho_E(s)]$ by minimizing $D_{JS}[\rho_\pi(b,a)||\rho_E(b,a)]$, because

$$D_{JS}[\rho_\pi(s)||\rho_E(s)] \leq D_{JS}[\rho_\pi(b)||\rho_E(b)] \leq D_{JS}[\rho_\pi(b,a)||\rho_E(b,a)]$$

Optimization Objective:

$$\min_\theta \max_w \mathbb{E}_{(b,a)\sim\pi_\theta,\mathcal{T}}\left[\log\left(1 - D_w(b,a)\right)\right] + \mathbb{E}_{(b,a)\sim M_E}[\log D_w(b,a)]$$

Policy Module

Assumption: $p(s|b), p(b'|b,a)$ are both independent of the policy

$$
\begin{aligned}
& D_{JS}[\rho_\pi(b)||\rho_E(b)] \\
&= \mathbb{E}_{b\sim\rho_E(b)}[f(\frac{\rho_\pi(b)}{\rho_E(b)})] \\
&= \mathbb{E}_{b\sim\rho_E(b)}\mathbb{E}_{s\sim p(s|b)}[f(\frac{\rho_\pi(b)p(s|b)}{\rho_E(b)p(s|b)})] \\
&= \mathbb{E}_{s,b\sim\rho_E(s,b)}[f(\frac{\rho_\pi(s,b)}{\rho_E(s,b)})] \\
&= \mathbb{E}_{s\sim\rho_E(s)}[\mathbb{E}_{b\sim\rho_E(b|s)} f(\frac{\rho_\pi(s,b)}{\rho_E(s,b)})] \\
&\geq \mathbb{E}_{s\sim\rho_E(s)}[f(\mathbb{E}_{b\sim\rho_E(b|s)} \frac{\rho_\pi(s,b)}{\rho_E(s,b)})] \\
&= \mathbb{E}_{s\sim\rho_E(s)}[f(\mathbb{E}_{b\sim\rho_\pi(b|s)} \frac{\rho_\pi(s,b)\rho_E(b|s)}{\rho_E(s,b)\rho_\pi(b|s)})] \\
&= \mathbb{E}_{s\sim\rho_E(s)}[f(\mathbb{E}_{b\sim\rho_\pi(b|s)} \frac{\rho_\pi(s)}{\rho_E(s)})] \\
&= \mathbb{E}_{s\sim\rho_E(s)}[f(\frac{\rho_\pi(s)}{\rho_E(s)})] \\
&= D_{JS}[\rho_\pi(s)||\rho_E(s)]
\end{aligned}
$$

$$f(u) = -(u+1)\log\frac{1+u}{2} + u\log u$$

Replace $s \mapsto b', b \mapsto (b,a)$
In the left proof, we can easily get

$$D_{JS}[\rho_\pi(b')||\rho_E(b')] \leq D_{JS}[\rho_\pi(b,a)||\rho_E(b,a)]$$

The independence holds under the trivial case of a deterministic mapping $b' = b$

Belief Module

Model the belief module B_ϕ with an RNN, such that $b_t = B_\phi(b_{t-1}, o_t, a_{t-1})$.

- Task-agnostic learning (separately from policy)
  - to maximize the joint likelihood of the observation sequence conditioned on action, i.e., $\sum_t \log p(o_t | o_{<t}, a_{<t})$
  - autoregressive loss (using unimodal Gaussian generative model):

$$L^{AR}(\phi) = E_R || o_t - g(b_{t-1}^{\phi}, a_{t-1}) ||_2^2$$

Belief Module

- Task-aware learning (jointly with policy)
  - same imitation learning objective naturally can be used

$$L^{IM}(\phi)$$
$$= E_{(h,a)\sim M_E}[\log D^*(B_\phi(h),a)] + E_{(h,a)\sim\pi_\theta(a|B_\phi(h))}[\log(1-D^*(B_\phi(h),a))]$$

  - so the gradient w.r.t ϕ can be approximated as:

$$E_{(h,a)\sim M_E}[\nabla_\phi \log D^*(B_\phi(h),a)] + E_{(h,a)\sim\pi_\theta(a|B_\phi(h)),\mathcal{T}}[\nabla_\phi \log\pi_\phi(a|B_\phi(h))Q^\pi]$$
$$+ E_{(h,a)\sim\pi_\theta(a|B_\phi(h)),\mathcal{T}}[\nabla_\phi \log(1-D^*(B_\phi(h),a))]$$

where $Q^\pi = E_{(h,a)\sim\pi_\theta,\mathcal{T}}\left[\sum_{t'=t}^{\infty}\gamma^{t'-t}\log\left(1-D^*\left(B_\phi(h),a\right)\right)\right]$

Belief Regularization

Overall objective for jointly training policy, belief and discriminator is

$$\min_{\phi,\theta} \max_{\omega} \widetilde{\mathbb{E}}_{(b,a)\sim\mathcal{M}_E}\left[\log D_\omega(b,a)\right] + \widetilde{\mathbb{E}}_{(b,a)\sim\pi,\mathcal{T}}\left[\log(1 - D_\omega(b,a))\right]$$

It may be possible that the belief parameters(ϕ) are driven towards a degenerate solution

Thus, add forward-, inverse- and action-regularization to get non-trivial belief representation

Belief Regularization

Forward regularization

Basic idea is that current belief should be correlated with future true states, conditioned on the intervening future actions

Maximize it

$$
\begin{aligned}
I(b_t;s_{t+k}|a_{t:t+k-1}) &\geq I(b_t;o_{t+k}|a_{t:t+k-1}) \\
&= \mathbb{E}_{a_{t:t+k-1}}\big[H(o_{t+k}|a_{t:t+k-1}) \\
&\qquad - H(o_{t+k}|b_t;a_{t:t+k-1})\big] \\
&\geq \mathbb{E}_{a_{t:t+k-1}}\Big[H(o_{t+k}|a_{t:t+k-1}) \\
&\qquad + \mathbb{E}_{o_{t+k},b_t}\big[\log q(o_{t+k}|b_t;a_{t:t+k-1})\big]\Big]
\end{aligned}
$$

First inequality based on:
If $(X \perp Z \mid Y)$, then $X \to Y \to Z$, and the data processing inequality that $I(X;Z) \leq I(X;Y)$.

And here, we have $o_{t+k} \perp b_t \mid s_{t+k}$, $b_t \to s_{t+k} \to o_{t+k}$

we want that $p(s|b)$ can compeletly characterize the environment

Belief Regularization

Second inequality use a variational approximation q

$$I(o;b|a) = H(o|a) - H(o|b,a)$$
$$= \mathbb{E}_a[H(o|a) + \underline{\mathbb{E}_{o,b}[\log p\,(o|b,a)]}]$$

$$E_{o,b}[\log p\,(o|b,a)]$$
$$= E_{o,b}[\frac{\log p\,(o|b,a)q(o|b,a)}{q(o|b,a)}]$$
$$= E_{o,b}[\log q\,(o|b,a)]$$
$$+ \mathbb{E}_b\mathbb{E}_{o\sim q(o|b)}[\log \frac{p(o|b)}{q(o|b)}] \geq 0$$

Belief Regularization

Thus, now we maximize above mutual information with the surrogate objective:

$$\max_{\phi, q} \mathbb{E}_{\substack{o_{t+k}, b_t, \\ a_{t:t+k-1}}} \left[\log q(o_{t+k} | b_t^{\phi}; a_{t:t+k-1}) \right]$$

Choose q as a unimodal Gaussian (learned function g for the mean and the fixed variance)

$$\mathcal{L}^f(\phi) = \mathbb{E}_{\mathcal{R}} ||o_{t+k} - g(b_t^{\phi}, a_{t:t+k-1})||_2^2$$

Belief Regularization

Inverse regularization

Basic idea is that current belief should be correlated with past true states, conditioned on the intervening past actions

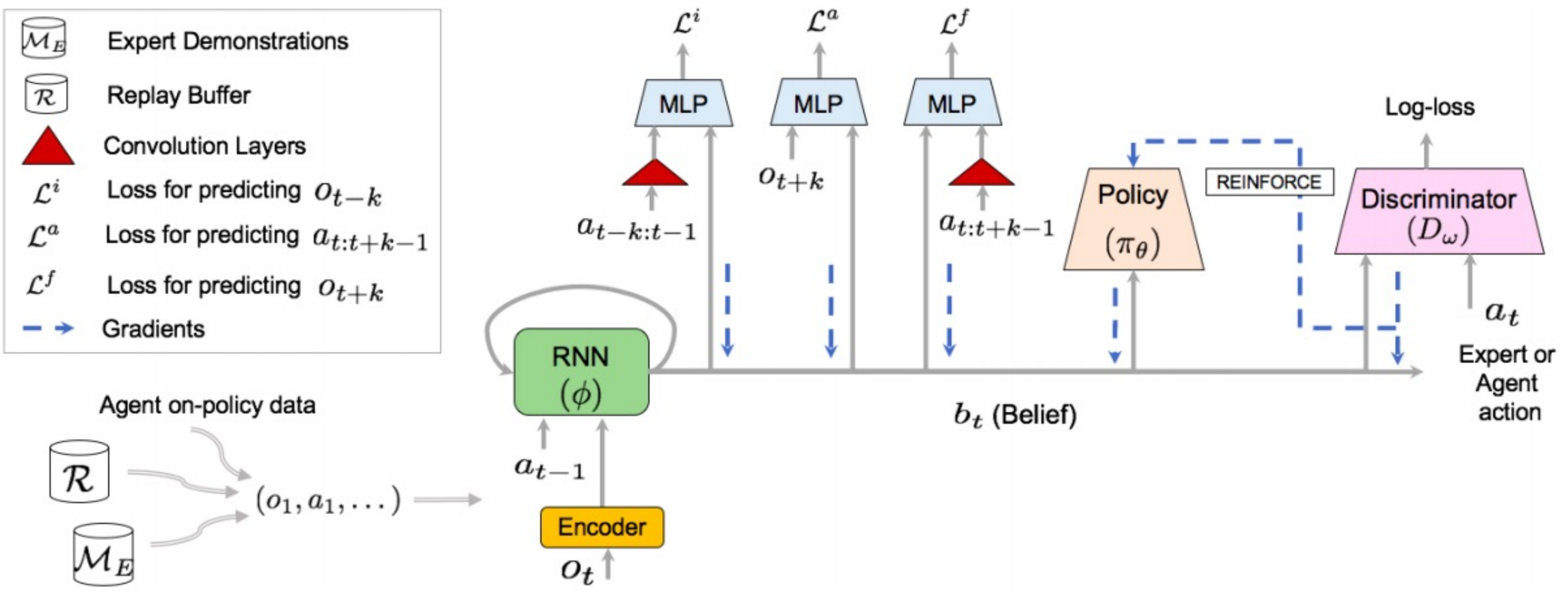
$$\mathcal{L}^i(\phi) = \mathbb{E}_{\mathcal{R}} || o_{t-k} - g(b_t^\phi, a_{t-k:t-1}) ||_2^2$$

Action regularization

Basic idea is that a sequence of *k* subsequent actions should provide information about the resulting true future state, conditioned on current belief, i.e. max $I(a_{t:t+k-1}; s_{t+k} \mid b_t)$

$$\mathcal{L}^a(\phi) = \mathbb{E}_{\mathcal{R}} || (a_{t:t+k-1}) - g(b_t^\phi, o_{t+k}) ||_2^2$$

Overall algorithm



Overall algorithm

Algorithm 1: Belief-module Imitation Learning (BMIL)

```
for each iteration do
    d_π = {}, d_E = {}
    /* Rollout c steps from policy */
    repeat
        Get observation o_t from environment
        a_t ~ π_θ(a_t|b_t), where b_t = B_φ(o_{≤t}, a_{<t})
        r_t = −log(1 − D_ω(b_t, a_t))
        d_π ← d_π ∪ (b_t, a_t, r_t)
        If o_t is terminal, add rollout {o_i, a_i}_{i=0}^{|τ|} to R
    until |d_π| == c;
    /* Update Policy */
    Update θ with policy-gradient (Eq. 4)
    /* Update discriminator ω */
    Fetch (o_t, a_t, ...) of length c from M_E
    Generate belief-action tuples d_E = {(b_i, a_i)}_{i=t}^{t+c−1}
    Update ω with log-loss objective using d_π and d_E
    /* Update Belief Module φ */
    Update φ with ∇_φ L(φ) using d_π and d_E (Eq. 10)
    /* Off-policy Updates */
    for few update steps do
        Fetch (o_t, a_t, ...) of length c from R
        Update φ with ∇_φ(λ_1 L^f + λ_2 L^i + λ_3 L^a)
    end
end
```

$$\nabla_\theta D_{JS}(\theta;\phi) \approx \nabla_\theta \widetilde{\mathbb{E}}_{(b,a)\sim\pi,\mathcal{T}}\big[\log(1-D^*(b,a))\big]$$
$$= \widetilde{\mathbb{E}}_{(b,a)\sim\pi,\mathcal{T}}\big[\nabla_\theta \log \pi_\theta(a|b)\hat{Q}^\pi(b,a)\big], \text{ where}$$
$$\hat{Q}^\pi(b_t,a_t) = \widetilde{\mathbb{E}}_{(b,a)\sim\pi,\mathcal{T}}\Big[\sum_{t'=t}^{\infty}\gamma^{t'-t}\log(1-D^*(b_{t'},a_{t'}))\Big] \tag{4}$$

$$\mathcal{L}(\phi) = \mathcal{L}^{IM} + \lambda_1\mathcal{L}^f + \lambda_2\mathcal{L}^i + \lambda_3\mathcal{L}^a \tag{10}$$

Experiment

Testing environment

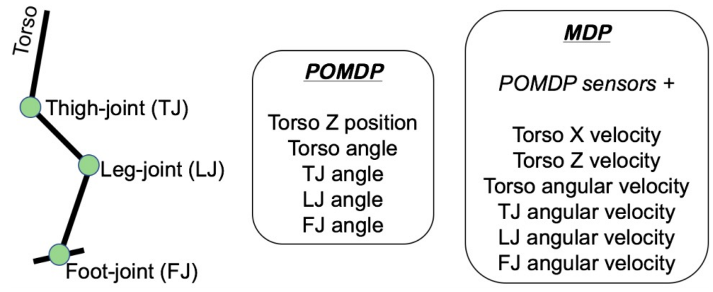


Figure 5: Comparison of sensor information available to the agent in the MDP (original) and the POMDP (modified) settings for Hopper-v2 from the Gym MuJoCo suite.

Environment	MDP sensors ($s \in \mathcal{S}$)	POMDP sensors ($o \in \mathcal{O}$)
Hopper	($\|\mathcal{S}\|$=11) **velocity(6)** + position(5)	($\|\mathcal{O}\|$=5) position(5)
Half-Cheetah	($\|\mathcal{S}\|$=17) **velocity(9)** + position(8)	($\|\mathcal{O}\|$=8) position(8)
Walker2d	($\|\mathcal{S}\|$=17) **velocity(9)** + position(8)	($\|\mathcal{O}\|$=8) position(8)
Inv.DoublePend.	($\|\mathcal{S}\|$=11) **velocity(3)** + position(5) + actuator forces(3)	($\|\mathcal{O}\|$=8) position(5) + actuator forces(3)
Ant	($\|\mathcal{S}\|$=111) **velocity(14)** + position(13) + external forces(84)	($\|\mathcal{O}\|$=97) position(13) + external forces(84)
Humanoid	($\|\mathcal{S}\|$=376) **velocity(23) + center-of-mass based velocity(84)** + position(22) + center-of-mass based inertia(140) + actuator forces(23) + external forces(84)	($\|\mathcal{O}\|$=269) position(22) + center-of-mass based inertia(140) + actuator forces(23) + external forces(84)

Observation is just a subset of the true state(even without noisy), it is indeed kind of frustrating setting

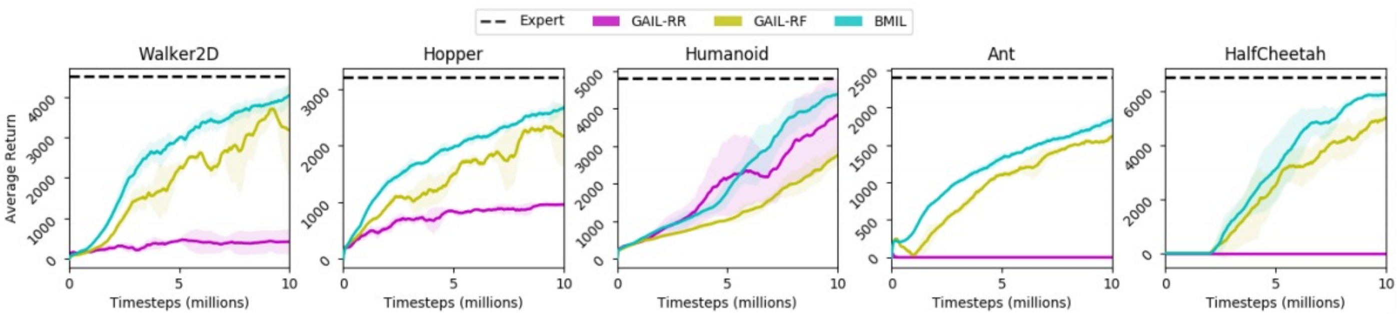
Experiment

	GAIL	GAIL + Obs. stack	BMIL (Ours)	Expert (≈ Avg.)
Inv.DoublePend.	109	1351	**9104**	9300
Hopper	157	517	**2665**	3200
Ant	895	1056	**1832**	2400
Walker	357	562	**4038**	4500
Humanoid	1686	1284	**4382**	4800
Half-cheetah	205	-948	**5860**	6500

Table 1: Mean episode-returns, averaged over 5 runs with random seeds, after 10M timesteps in POMDP MuJoCo.

	GAIL-RR	GAIL-RF	BMIL (Ours)
Inv.DoublePend.	8965	9103	**9104**
Hopper	955	2164	**2665**
Ant	-533	1612	**1832**
Walker	400	3188	**4038**
Humanoid	3829	2761	**4382**
Half-cheetah	-922	5011	**5860**

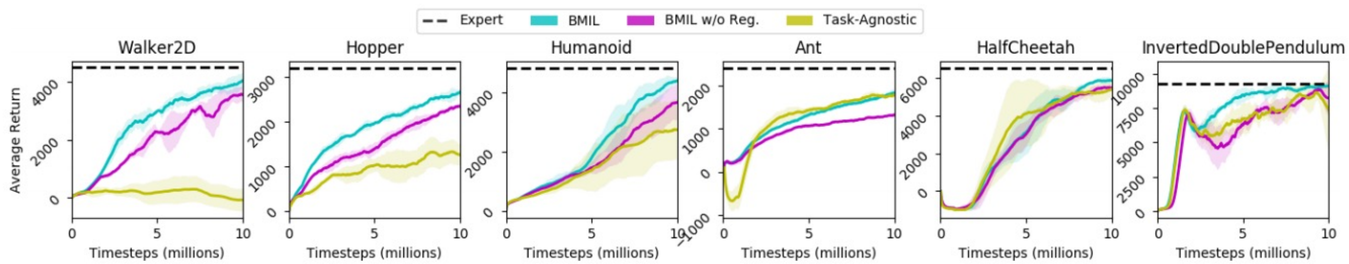
Table 2: Mean episode-returns, averaged over 5 runs with random seeds, after 10M timesteps in POMDP MuJoCo.



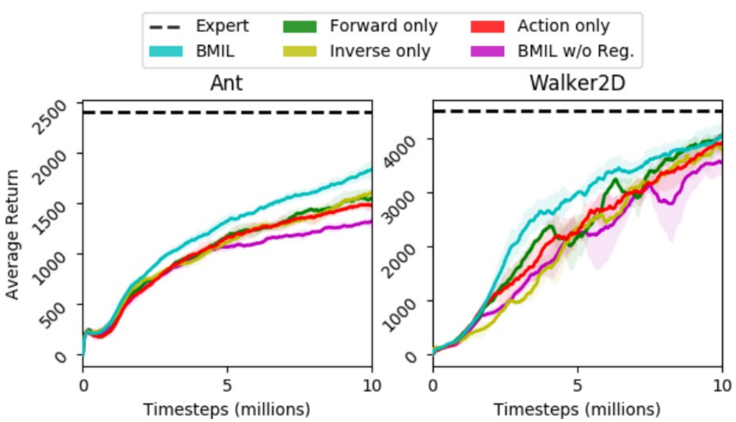
GAIL-RF uses a recurrent policy and a feed-forward discriminator, while in GAIL-RR, both the policy and the discriminator are recurrent.
Unlike BMIL, the belief is not shared between the policy and the discriminator

Ablation Studies

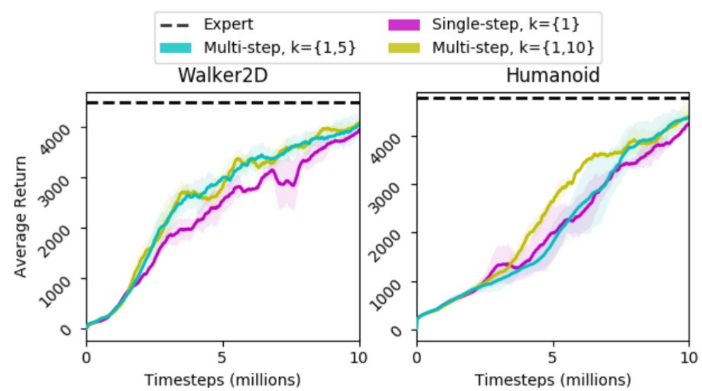
1. How crucial is belief regularization? & Task-aware vs. Task-agnostic belief learning.



2. Are all of L^f, L^i, L^a useful?



3. Are multi-step predictions useful?



Comments

- This paper propose a flexible architecture to do imitation learning in POMDP setting
- Use RNN to represent Belief is naïve
- This work relates to a new research direction that if additional information about environment are given(expert ob-act seq, some true states, etc.), how can we perform better in more challenging POMDP setting?
- For transferring, can we get disentangle belief representation and policy model in POMDP setting, like what they* do in MDP setting.

Zhang A, Satija H, Pineau J. Decoupling dynamics and reward for transfer learning[J]. arXiv preprint arXiv:1804.10689, 2018.

Thanks!