FoLaR: Foggy Latent Representations for Reinforcement Learning with Partial Observability

Hardik Meisheri and Harshad Khadilkar

Planning and Control Research Group, TCS Research, Mumbai, India
July 2021, *IJCNN*
Introduction

Challenges with off-the-shelf RL algorithms

- Not sample efficient: large amount of information loss
- Difficulty with hard exploration tasks and challenges posed by partially observable environment (POMDP)

Motivation in POMDP context

- Model-free algorithms very difficult to apply in partially observable settings, at least partly due to the violation of Markov assumptions
- Possible overfitting to noise → poor knowledge of the environment
- Information available during the decision-making process is neither perfect nor complete
Learning better representations by imposing a constraint of reconstructing the entire state or observation from the latent representations, forcing the model to encode all the information into dense representations\(^1\) → Decoupled from policy and reward signal often leads to the encoding of information that is irrelevant\(^2\)

---

\(^1\)Ha and Schmidhuber 2018; Hafner et al. 2019; Lee et al. 2019.

\(^2\)Zhang et al. 2020; Gelada et al. 2019.
Prior Work and State of the art

- Learning decoupled the representations learning from policy improvement Lange and Riedmiller 2010; Lange, Riedmiller, and Voigtländer 2012; Subramanian et al. 2020 → Not suitable for POMDPs (noisy information)

- Notion of making latent representations reflect the state (dis)similarity by adding loss term\(^3\) → Bisimulation is computationally intensive metric and how to adopt to POMDPs

- Self Imitation Learning (SIL)\(^4\) proposes to learn latent representations better by leveraging past experiences to solve harder exploration problems. → Considered as baseline

---

\(^3\)Zhang et al. 2020; Gelada et al. 2019.

\(^4\)Oh et al. 2018.
Our Intuition

We hypothesize that augmenting latent representations with predictive loss (prediction in latent space not reconstruction) and learning end-to-end generates better policy and sample efficiency as opposed to decoupled learning of representations and policy in POMDPs.
Key Contribution

- Proposing a training paradigm and loss function to learn robust latent representations in POMDP settings contextualized on latent representations of belief state.
- Ability to modify any off-the-shelf RL algorithm to improve the sample efficiency and exploration characteristics.
- Extensive evaluation on two partially observable environments with varying scales.
Markov decision process (MDP) 
\((S, A, T, R, \gamma)\), where 
\(S\) = state space, 
\(A\) = action space, 
\(T\) = transition probabilities, 
\(R\) = rewards, and 
\(\gamma\) = discount factor for future rewards

Partially Markov decision process (POMDP) 
\((S, A, T, R, \Omega, O, \gamma)\), where 
\(S\) = state space, 
\(A\) = action space, 
\(T\) = transition probabilities, 
\(R\) = rewards, 
\(\Omega\) = observation, 
\(O\) = observation function, and 
\(\gamma\) = discount factor for future rewards
System Architecture

Encoder (shared weights) → \(O_{t+1}\) → Direct gradient path blocked

Encoder (shared weights) → \(O_t\) → \(O_{t-1}\) → \(O_{t-k}\) → \(h_t\) → \(h_{t-1}\) → \(h_{t-k}\) → RNN (LSTM) → Prediction network → \(h_{t+1}^{\text{pred}}\)

MSE loss

Value network → \(V_t\)

Policy network → \(\pi_t\)
Augmented Loss Function

Policy Gradient Update

\[ \Delta \theta \propto \nabla_{\theta} \log(\pi_{\theta}(a_t/s_t)) A(s_t, a_t) + \beta \nabla_{\theta} H(\pi_{\theta}(s_t)) \]

\(\pi_{\theta}\) = policy parameterized by \(\theta\)

\(A(s_t, a_t)\) = Advantage Function

Augmented Policy Gradient Update

\[ \Delta \theta \propto \nabla_{\theta} \log(\pi_{\theta}(a_t/s_t)) A(s_t, a_t) + \beta \nabla_{\theta} H(\pi_{\theta}(s_t)) + \Delta L_t^{NL} \]

where, \(L_t^{NL} = \eta \times \text{mse}(h_{t+1}^{\text{pred}}, h_{t+1})\)
Augmented Loss Function

**Proximal Policy Optimization (PPO) Loss**

\[ L_{t}^{PPO} = \mathbb{E}\left[ \min\left( r_{t}(\theta)\hat{A}_{t}, \text{clip}(r_{t}(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_{t}\right) \right] + (V_{\theta}(s_{t}) - V_{t}^{\text{targ}})^{2} + \beta H(\pi_{\theta}(s_{t})) \]

- **Policy Loss**
- **Value Loss**
- **Entropy**

**FoLaR Loss**

\[ L_{t}^{FoLaR} = L_{t}^{PPO} + L_{t}^{NL} \]

where, \[ L_{t}^{NL} = \min(\eta \times \text{mse}(h_{t+1}^{pred}, h_{t+1}), \epsilon) \]
Environments

Mini Gridworld
- Structured input
- POMDP
- Visible grid size $7 \times 7$
- Dynamic Obstacles: $6 \times 6$, $16 \times 16$
- Doorkey Gridsize: $6 \times 6$, $8 \times 8$
Environments

Catcher
- RGB Pixel inputs
- MDP
- Grid size $32 \times 32$
Training Results

Dynamic Obstacles $6 \times 6$

\[
\beta = 0.01
\]

\[
\beta = 0.2 \rightarrow 0.01
\]
Training Results

Dynamic Obstacles $16 \times 16$

$\beta = 0.01$

$\beta = 0.2 \rightarrow 0.01$
Training Results

DoorKey $6 \times 6$

$\beta = 0.01$

$\beta = 0.2 \rightarrow 0.01$
Training Results

DoorKey $8 \times 8$

$\beta = 0.01$

$\beta = 0.2 \rightarrow 0.01$
Training Results

Catcher

\[ \beta = 0.01 \]

\[ \beta = 0.2 \rightarrow 0.01 \]
Ablation Study

Effect of $\eta$ value, 10 random seeds dynamic obstacles $6 \times 6$

![Graph showing the effect of $\eta$ value on reward](image)
Testing on out-of-distribution environments
Conclusion and Summary

- We presented FoLaR, a method to learn robust latent representations in partially observable environments.
- Loss function $L_t^{NL}$ which can be augmented with any on-policy off the shelf algorithm to improve its exploration and convergence characteristics.
- The hyperparameter $\eta$ controls the magnitude of prediction loss, and can be tuned based on the predictability of the environment, but most reasonable values result in superior performance compared to baselines.
Conclusion and Summary

- FoLaR performs especially well in hard exploration tasks and larger grid sizes where entropy coefficient is kept static, indicating improved latent representations that lead to more focused exploration.

- In future work, it would be interesting to look at adaptive $\eta$, which could learn better policies even faster.
Thank You!

Hardik Meisheri: hardik.meisheri@tcs.com, hardik.meisheri@gmail.com
Harshad Khadilkar: harshad.khadilkar@tcs.com, harshadk@iitb.ac.in