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

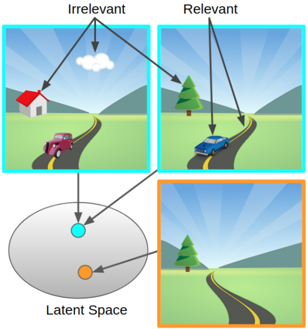
- Not sample efficient: large amount of information loss
- Difficulty with hard exploration tasks and challenges posed by partially observable environment (POMDP)

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 $\rightarrow$ poor knowledge of the environment
- Information available during the decision-making process is neither perfect nor complete

Prior Work and State of the art

- 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] $\rightarrow$ Decoupled from policy and reward signal often leads to the encoding of information that is irrelevant[2]



Source Zhang et al. 2020

[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 for POMDPs (noisy information)
- Notion of making latent representations reflect the state (dis)similarity by adding loss term[3] → Bisumulation 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

tcs research

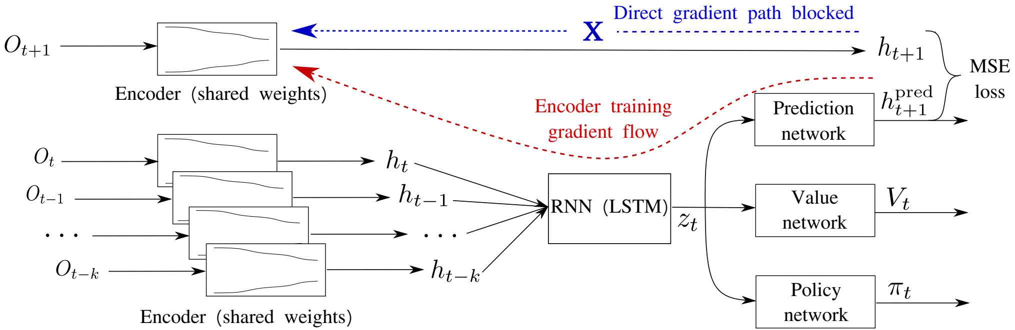
- Proposing a training paradigm and loss function to learn robust latent representations in POMDP settings contextualized on latent representations of belief state
- Ability to modify any off-the-shelf RL algorithm to improve the sample efficiency and exploration characteristics
- Extensive evaluation on two partially observable environments with varying scales

Preliminaries

Markov decision process (MDP)
$(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$, where
$\mathcal{S}$ = state space,
$\mathcal{A}$ = action space,
$\mathcal{T}$ = transition probabilities,
$\mathcal{R}$ = rewards, and
γ = discount factor for future rewards

Partially Markov decision process (POMDP)
$(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \Omega, \mathcal{O}, \gamma)$, where
$\mathcal{S}$ = state space,
$\mathcal{A}$ = action space,
$\mathcal{T}$ = transition probabilities,
$\mathcal{R}$ = rewards,
Ω = observation,
$\mathcal{O}$ = observation function, and
γ = discount factor for future rewards

System Architecture



Augmented Loss Function

tcs research

Policy Gradient Update

$$\Delta\theta \propto \nabla_\theta log(\pi_\theta(a_t/s_t))A(s_t, a_t) + \beta\nabla_\theta H(\pi_\theta(s_t))$$
π_θ = policy parameterized by θ
$A(s_t, a_t)$ = 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, $L_t^{NL} = \eta \times mse(h_{t+1}^{pred}, h_{t+1})$

Augmented Loss Function

Proximal Policy Optimization (PPO) Loss

$$L_t^{PPO} = \hat{\mathbb{E}}[\underbrace{min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)}_{\text{Policy Loss}} + \underbrace{(V_\theta(s_t) - V_t^{targ})^2}_{\text{Value loss}} + \underbrace{\beta H(\pi_\theta(s_t))}_{\text{Entropy}}]$$

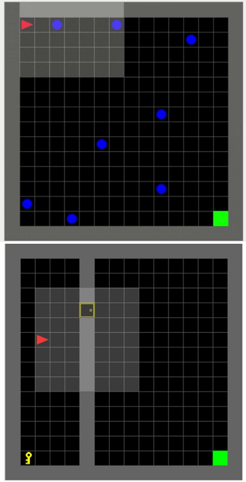
FoLaR Loss

$$L_t^{FoLaR} = L_t^{PPO} + L_t^{NL}$$

where, $L_t^{NL} = min(\eta \times mse(h_{t+1}^{pred}, h_{t+1}), \epsilon)$

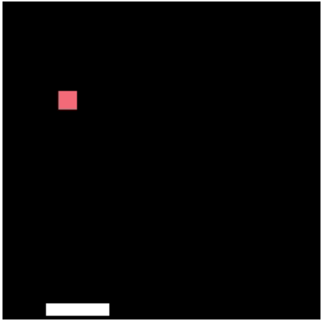
Mini Gridworld

- Structured input
- POMDP
- Visible grid size 7×7
- Dynamic Obstacles: 6×6, 16×16
- Doorkey Gridsize: 6×6, 8×8

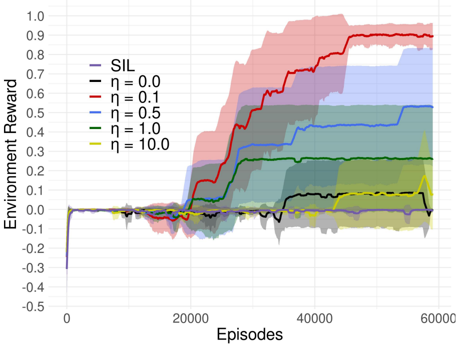


Catcher

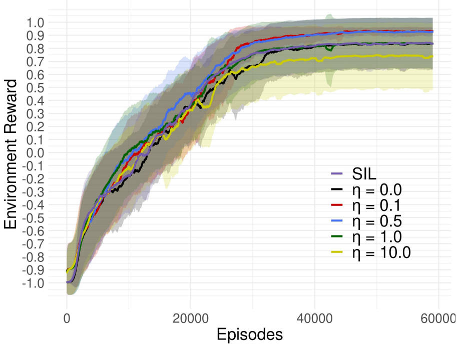
- RGB Pixel inputs
- MDP
- Grid size 32×32



Dynamic Obstacles 6×6

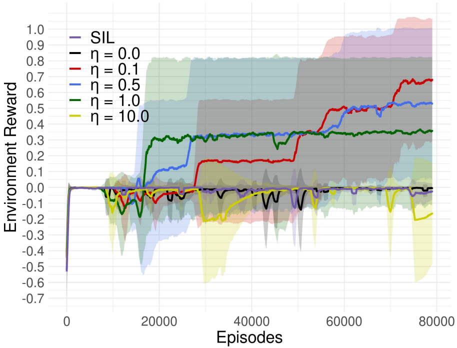


$\beta = 0.01$

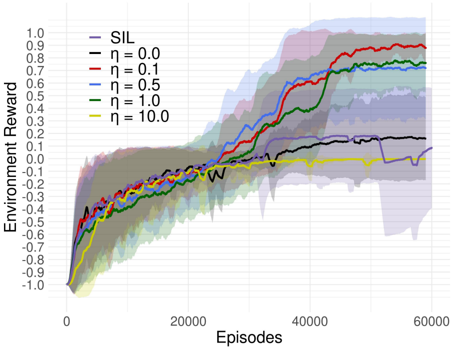


$\beta = 0.2 \rightarrow 0.01$

Dynamic Obstacles 16×16



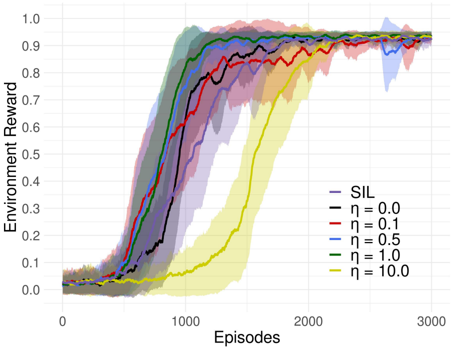
$\beta = 0.01$



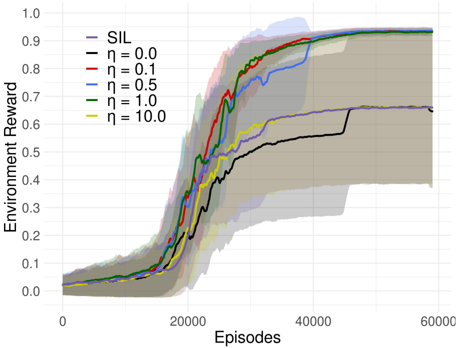
$\beta = 0.2 \rightarrow 0.01$

Training Results

DoorKey 6×6



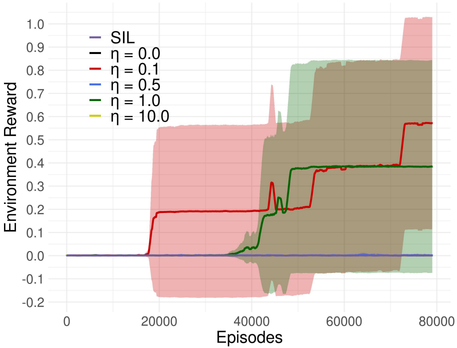
$\beta = 0.01$



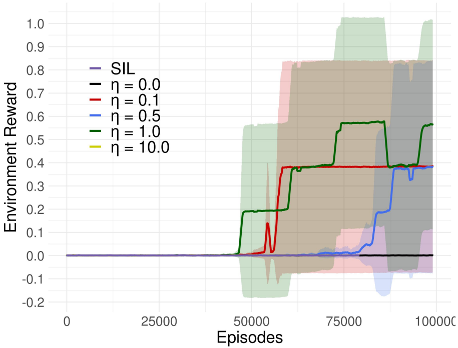
$\beta = 0.2 \rightarrow 0.01$

tcs research

DoorKey 8×8



$\beta = 0.01$

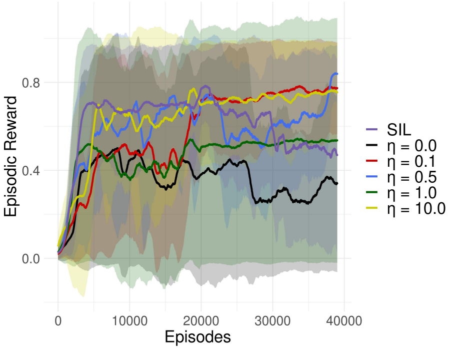


$\beta = 0.2 \rightarrow 0.01$

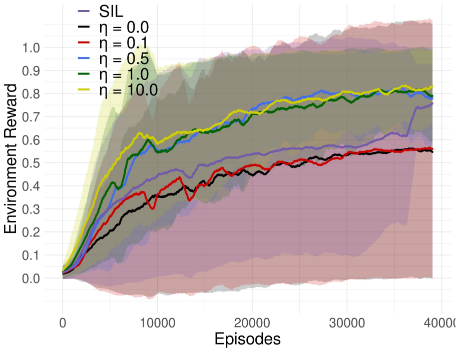
TATA CONSULTANCY SERVICES

tcs research

Catcher



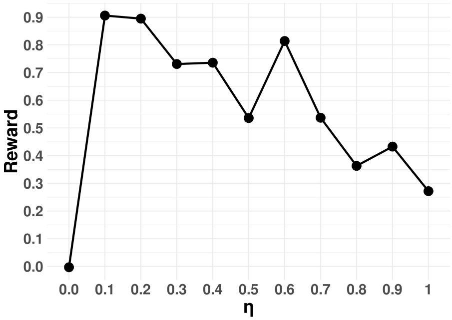
$\beta = 0.01$



$\beta = 0.2 \rightarrow 0.01$

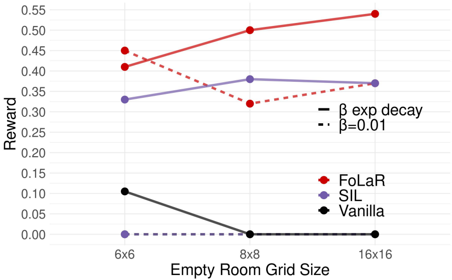
tcs research

Effect of η value, 10 random seeds dynamic obstacles 6×6



TATA CONSULTANCY SERVICES

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 η 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 focussed exploration
- In future work, it would be interesting to look at adaptive η, 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

