Paper ID: #3829

Multi-Agent Sparse Interaction Modeling is An Anomaly Detection Problem

Chao Li [1], Shaokang Dong [1], Shangdong Yang [2,3], Hongye Cao [1], Wenbin Li [1], Yang Gao [1]

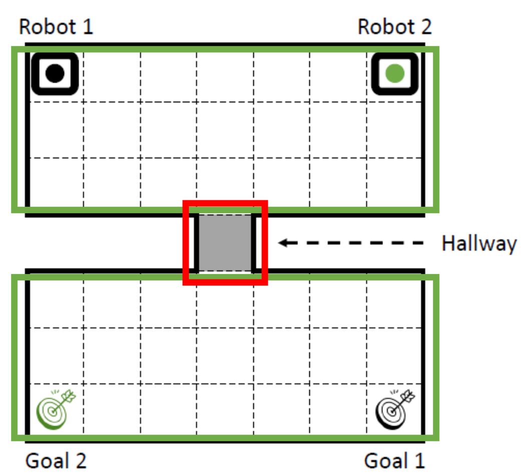
[1] State Key Laboratory for Novel Software Technology, Nanjing University [2] School of Computer Science, Nanjing University of Posts and Telecommunications

[3] Guangxi Key Lab of Multi-Source Information Mining & Security, Guangxi Normal University

chaoli1996@smail.nju.edu.cn

1. Motivation

- Multi-Agent Reinforcement Learning (MARL) heavily suffers from ***sample inefficiency*** problem
  - Algorithm perspective: the trial-and-error paradigm inherent in RL;
  - Task perspective: policy search over enormous state-joint action space.



Sparse Interaction Property

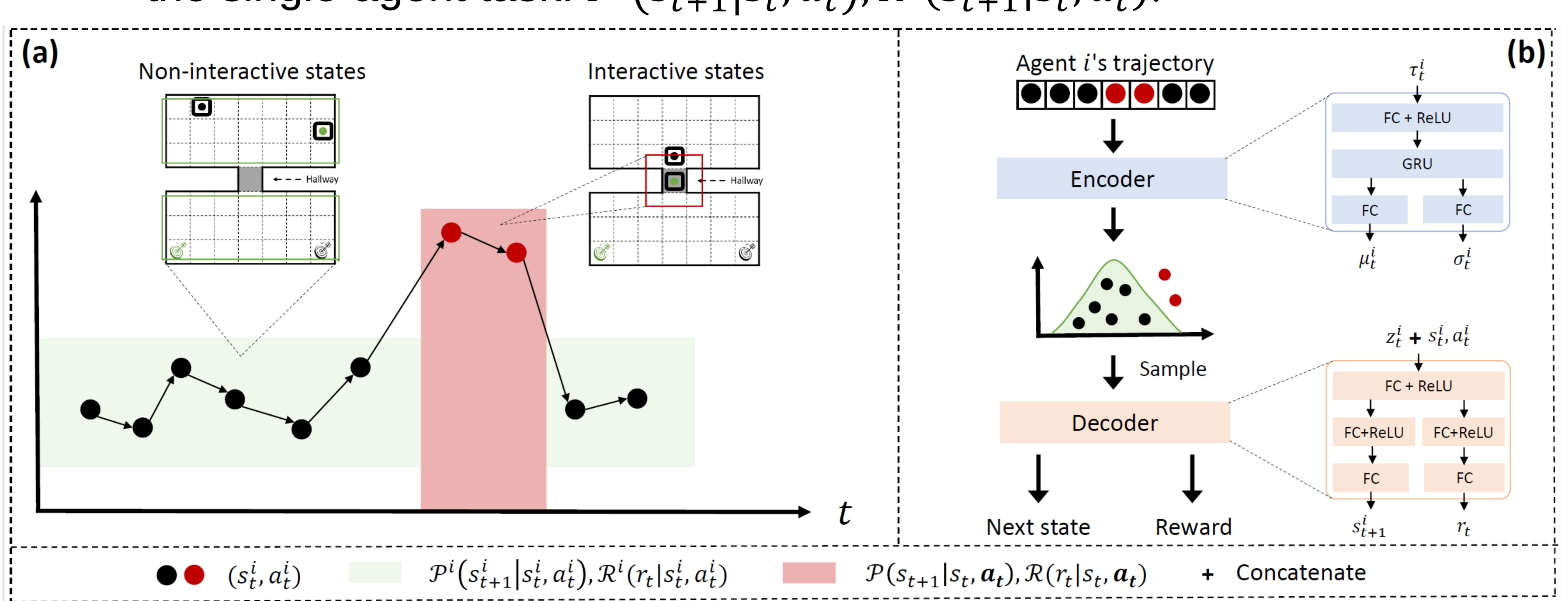
- On rare interactive states where agents need to coordinate, the multi-agent task continues;
- On commonplace non-interactive states where agents can act independently to achieve desired outcomes, task reduces to a single-agent task.

- **Our focus**: improve MARL from the task perspective
  - Model the sparse interaction structure among agents;
  - Utilize this structure to instruct per-agent policy learning.

2. Methodology

2.1 Model the sparse interaction by dynamics

- In per-agent local trajectory, rare interactive states (outliers) adhere to the dynamics of the multi-agent task: $P(s_{t+1}|s_t, \boldsymbol{a}_t), R(r_t|s_t, \boldsymbol{a}_t)$;
- In contrast, common non-interactive states adhere to the dynamics of the single-agent task: $P^i(s_{t+1}^i|s_t^i, a_t^i), R^i(s_{t+1}^i|s_t^i, a_t^i)$.



2.2 Dynamics Distribution Modeling

- We characterize the trajectory dynamics by a latent variable, and the objective is defined as: $\min \mathcal{D}_{\mathrm{KL}}(q(z^i \mid \tau_t^i) \,\|\, p(z^i \mid \tau_t^i))$

$$\mathcal{D}_{\mathrm{KL}}(q(z^i \mid \tau_t^i) \,\|\, p(z^i \mid \tau_t^i)) = \mathcal{D}_{\mathrm{KL}}(q(z^i \mid \tau_t^i) \,\|\, p(z^i)) - \mathbb{E}_{z^i \sim q(z^i|\tau_t^i)} \log p(\tau_t^i \mid z^i) + \log p(\tau_t^i).$$

$$\max \mathbb{E}_{z^i \sim q(z^i|\tau_t^i)} \log p(\tau_t^i \mid z^i) + \mathcal{D}_{\mathrm{KL}}(q(z^i \mid \tau_t^i) \,\|\, p(z^i))$$

$$\max \underbrace{\mathbb{E}_{z^i \sim q(z^i|\tau_t^i)}}_{\text{Encoder}} \sum_{t=0}^{T-1} \underbrace{\log p(s_{t+1}^i, r_t \mid s_t^i, a_t^i, z^i)}_{\text{Decoder (Reconstruction Likelihood)}} - \underbrace{\mathcal{D}_{\mathrm{KL}}(q(z^i \mid \tau_t^i) \,\|\, p(z^i))}_{\text{Kl Divergence between posterior and prior}}.$$

- We achieve this optimization using a VAE-like network.

2.3 Interactive State Discovery

- For interactive states, the reconstruction likelihood is poor
- We deine the prediction discrepancy as follows:

$$Dis(s_t^i, a_t^i) = (s_{t+1}^i - \hat{s}_{t+1}^i)^2 + (r_t - \hat{r}_t)^2.$$

- A large value indicates that current state-action is interactive.

2.4 Interaction-Instructed Exploration

- We use the prediction discrepancy as the intrinsic reward:

$$r^i(s_{t+1}^i, a_t^i) = Dis(s_t^i, a_t^i) = (s_{t+1}^i - \hat{s}_{t+1}^i)^2 + (r_t - \hat{r}_t)^2.$$

Intuition: This intrinsic reward encouarges agents to explore more on interactive states, enhancing their coordinated behaviors and thus accelerating coordinated policy learning.

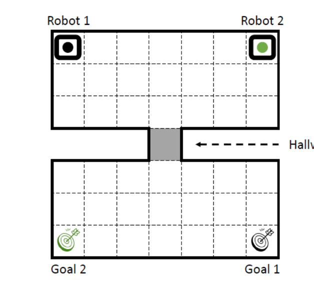
- Consequently, the MARL algorithm aims to maximize the total reward of all agents, which is defined as follows:

$$r(s_t, \boldsymbol{a}_t) = r_t + \alpha \sum_i r^i(s_{t+1}^i, a_t^i),$$

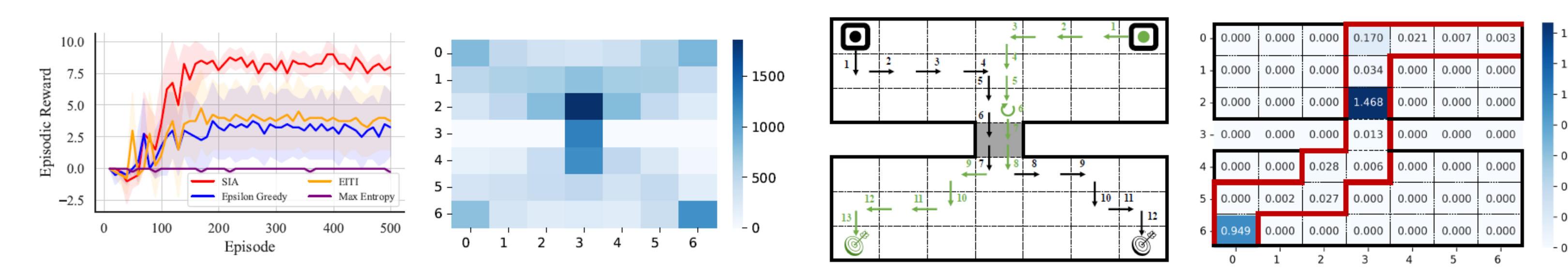
- where r_t denotes the extrinsic task reward.

3. Experiments

- Benchmarks
  - Hall Way (A Didactic Example);
  - StarCraft Multi-Agent Challenge;



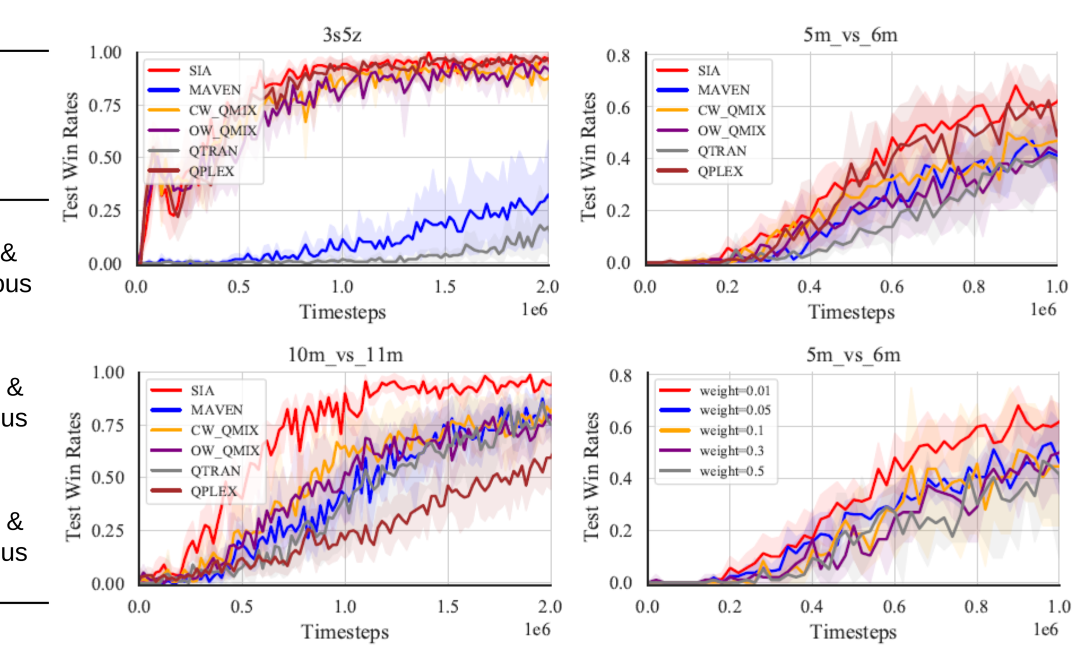
- Results



(a) Comparison results. (b) Visitation heatmap. (c) Final learned policies. (d) Prediction error heatmap.

Summary: SIA successfully identifies interactive states, and the interaction instructed exploration encouarges more exploration on them, leading to superior performance.

Map	Ally vs. Enemy	Type
3s5z	3 Stalkers, 5 Zealots vs. 3 Stalkers, 5 Zealots	Symmetric & Heterogeneous
5m_vs_6m	5 Marines vs. 6 Marines	Asymmetric & Homogeneous
10m_vs_11m	10 Marines vs. 11 Marines	Asymmetric & Homogeneous



Summary: The superior performance on complex tasks further verifies its effectiveness.

4. Conclusion

- Modeling the interaction strcture among agents and utilizing it to improve MARL is promising. More ways are worth further exploration;
- In future, we would focus on the nearly decomposable property of multi-agent tasks to enhance multi-agent coordination on large-scale scenarios !