Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [MABL: Bi-Level Latent-Variable World Model for Sample-Efficient Multi-Agent Reinforcement Learning]
Date: [23-10-2024]

Make sure your critique Address the following points:
1. The problem the paper is trying to address
2. Key contributions of the paper
3. Proposed algorithm/framework
4. How the proposed algorithm addressed the described problem
Note: Be concise with your explanations. Unnecessary verbosity will be penalized. Please don't exceed 2 pages.

---

1 The problem the paper is trying to address

The paper addresses the problem of **high sample complexity** in *Multi-Agent Reinforcement Learning (MARL)*, especially in *partially observable environments*. Existing methods struggle to efficiently encode **global information** into their **latent states** during training, which leads to low learning efficiency. Most existing approaches either assume *centralized execution*, which is impractical, or fail during decentralized execution due to the absence of global context during training. The paper proposes a novel approach, **MABL (Multi-Agent Bi-Level Latent-Variable World Model)**, which enhances **sample efficiency** by introducing a hierarchical structure that encodes both **global** and **agent-specific** information, while ensuring decentralized policy execution.

2 Key contributions of the paper

- **MABL: Multi-Agent Bi-Level Latent-Variable World Model**: The paper introduces a novel model-based MARL algorithm called **MABL**, which learns a bi-level latent-variable world model. It encodes essential global information at the upper level and agent-specific information at the lower level. This structure allows for centralized training with decentralized execution.

- **Improved Sample Efficiency**: The method significantly improves **sample efficiency** by utilizing synthetic trajectories generated by the latent-variable world model, outperforming state-of-the-art methods in empirical benchmarks like SMAC, Flatland, and MA-MuJoCo.

- **Compatibility with Model-Free MARL Algorithms**: MABL can be combined with any model-free MARL algorithm for policy learning, making it a flexible approach for multi-agent settings.

- **Hierarchical Latent Space for Multi-Agent Coordination**: The bi-level model structure enables better representation learning by capturing both global and local dynamics, which improves coordination among agents in multi-agent environments.

3 Proposed algorithm/framework

Algorithm Steps:

1. **Environment Interaction:**
   - Agents interact with the environment using the policy $\pi_\theta(a_{i,t}|z_{a,t}, h_{a,t})$ to collect real data, which is stored in a buffer D.
2. **Model Training:**
   - Sample data from buffer D.
   - Train the bi-level world model using the ELBO loss to learn both global and agent-specific latent states.
3. **Synthetic Trajectory Generation:**
   - Generate synthetic trajectories by propagating the learned latent states using the transition dynamics models.
4. **Policy Learning:**
   - Train the policy π_θ using any model-free MARL algorithm (e.g., MAPPO) on the synthetic trajectories.
5. **Decentralized Execution:**
   - During execution, agents use only their local latent state ($z_{a,t}$) and agent-specific observation to make decisions independently.

4 How the proposed algorithm addressed the problem

1. Incorporation of Global Information:

- MABL introduces a **bi-level structure** with a **global latent state** (z_g) and an **agent latent state** (z_a).
- This global latent state is only used during training to enhance learning efficiency and is not required during execution.

2. Decentralized Execution:

- While the global latent state informs the agent-specific latent state during training, the agent latent state (z_a) is used exclusively for **decentralized execution**.

3. Improved Sample Efficiency:

- By using a **latent-variable world model**, MABL generates **synthetic trajectories** for training. This drastically reduces the number of real environment interactions needed to learn good policies, thus improving sample efficiency.

4. Compatibility with Any Model-Free MARL Algorithm:

- MABL can be integrated with any **model-free MARL algorithm** (e.g., MAPPO) for policy learning. This modularity allows the method to be used with existing state-of-the-art MARL techniques while benefiting from the improved sample efficiency of the world model.