Review-24

MAVEN: Multi-Agent Variational Exploration

State-of-the-art value-based methods in Multi-Agent Reinforcement Learning (MARL) such as QMIX depict improved performance but suffer as a result of representational constraints imposed by monotonicity in actions. The alternative of continued exploration is a suitable direction which allows the agent to carry out committed exploration. To this end, the work presents Multi-Agent Variational ExploraitoN (MAVEN) which makes use of a latent space for hierarchical control. A hierachical exploration policy yields latent variables for conditioning behaviors in agents. In addition, a Mutual Information (MI) maximization objective enables the learning of diverse behaviors from latent space in trajectories. On a range of challenging exploration scenarios, MAVEN demonstrates improved performance as a result of continued exploration.

Representational constraints induced in QMIX as a result of monotonic mixing prevent it from comprehensively expressing joint actions in various scenarios. The problem is further highlighted in cases where Q-functions are nonmonotonic and depend on ordering of agents' actions. This dependence can be resolved by utilizing continued exploration over temporally-extended horizons. While various methods dwell on curriculum learning, MAVEN directs its attention to variational exploration in a shared latent space. The hierarchical policy yields latent variables which optimize the trajectory from its initial state. While latent variables themselves do not encourage diverse behavior, the MI maximization objective between trajectories and latent variables motivates diverse behaviors. At the same time, the objective imposes the condition of individual agent actions to be close to joint actions. The MI loss is trained using a discriminator which is obtained from lower bounding MI in order to obtain a tractable objective. Lower bounding of dependence is carried out using a variational approximation which serves as the discriminator. Agents execute actions in the environment using the original QMIX framework and maximize joint Q-values using Q-learning which are utilized in further maximizing the MI objective.

MAVEN demonstrates improved exploration on StarCraft II micromanagement scenarios as a result of variational MI objective and hierarchical policy. The 2_corridors task suitably demonstrates the effective exploration carried out by MAVEN as a result of which agents are able to adapt to the closed corridor. Furthermore, ablations carried out on the components of MAVEN highlight the necessity for latent space and variational MI objective. While MAVEN improves exploration performance and yields diverse behaviors, its experimental setup presents a few caveats. Firstly, utilization of MI objective per timestep improves performance as a result of a more spread out exploration budget. While the original MI objective allows joint actions to be narrow and performs equivalently well in practice, the work does not throw light on its limitations in scenarios where sufficiently diverse behaviors are of requirement. Lastly, increasing the size of latent space does not show a definite improvement indicating that the representational constraints of QMIX may still persist and exploration is being carried out only to sidestep them.