Review-30

Bayesian Action Decoder for Deep Multi-Agent Reinforcement Learning

Reasoning of actions taken by an agent is pivotal for understanding decision-making in multi-agent learning. Interpretable actions allow other agents to understand *why* an agent selected the particular action. To this end, the work presents Baysian Action Decoder (BAD). BAD is a multi-agent algorithm which uses an approximated Baysian update to obtain public beliefs conditioned on actions taken by agents in the environment. The framework results in a public belief MDP with its action spaces as the set of all deterministic partial policies. Agents additionally learn private beliefs about the environment based on the deterministic partial policy being used in the multi-agent system. BAD is validated on a two-step matrix game along with the challenging benchmark of Hanabi wherein it outperforms prior existing methods to yield state-of-the-art performance.

BAD provisions the discovery of efficient communication protocols and resoning mechanims among agents based on public and private beliefs in the modified PUBlic (PuB-) MDP. The PuB-MDP consists of public state features which are accessible to all agents and provision public beliefs. An agent interacting in the environment additionally observes private features which are not accessbile to other agents and construct the private belief of agent. A third-aprty agent, known as the public agent, selects a partial policy as its action. The selected partial policy is used by other agents for executing joint environment actions. Each agent makes use of private beliefs to select an action which construct the joint action. Thus, the partial policy serves as a mapping from private beliefs to executable actions for each agent. Once the joint action is executed in the environment, agents observe the next state. Each state transition depends on the selected action as well as the complete partial policy indicating counterfactual actions for each agent. Practical implementations of public belief updates are carried out using factorized representations of private features which are updated recursively.

BAD demonstrates improved performance on matrix game in comparison to vanilla policy gradient and outperforms pre-existing methods on the Hanabi benchmark. Adoption of BAD on Hanabi further demonstrates its suitability in effectively reasoning future actions based on immediate actions taken by partner agents. While, BAD demonstrates the development of effective communication protocols among agents, it does not provide evidence for scalability of these protocols. For instance, evaluating the variation of returns and reasoned actions by partner agents in the case of increasing number of agents would be an apt addition to the experiment setup. Additionally, reasoning of partner agent's actions could be further studied by making use of semi-private actions which may be accessible to team agents and not to opponents in the environment.

Utilization of effective communication protocols is essential for understanding and executing optimal actions in mnulti-agent learning. BAD presents a suitable communication framework by leveraging public and private features as belief updates. The setup can be further extended to assess its stability and scalability with increasing number of agents. Furthermore, the role of CF gradient updates and alternative beliefs may be explored in detail to extend utility of BAD towards complex environments.

Motivated by the *theory of mind*, the work presents BAD which is a multi-agent algoritm provisioning the discovery of effective communication and reasoning among agents. BAD gives rise to the framework of PuB-MDP with its action spcae consisting of deterministic partial policies used by agents to select actions based on their private beliefs. BAD demonstrates stet-of-the-art performance on Hanabi as a result of apt future actions reasoned by partner agents.