Review-8

Hindsight Credit Assignment

Assigning credit to past decisions has been a challenging problem due to high variance and bootstrapped estimates. The work addresses this open problem by assigning credit to past decisions in hindsight. More specifically, the work aims to answer the question *"given a state x, how does choosing an action a affect the returns?"*. Taking into account the four main hindrances for efficient credit assignment, the work proposes a novel Hindsight Credit Assignment (HCA) scheme based on conditioned future states (HCA|State) and future returns (HCA|Return). The proposed HCA scheme is found to be efficient in assigning credit to past actions in comparison to policy gradient.

The significance of an action in a given trajectory is challenging to estimate. High variance in value estimates yields randomness in trajectories. In the case of partial observability, temporal difference learning is hindered by bootstrapped estimates which lead to inaccurate approximations. Additionally, a single trajectory may not incorporate information about all actions. To that end, the novel HCA scheme consists of future conditionals which weigh the importance of an action with respect to agent's policy. (HCA|State) consists of a state-conditional hindsight distribution which quantifies the relevance of a past action to a future state. Similarly, (HCA|Return) comprises of the reward-conditional hindsight distribution which conditions actions on future rewards. Hindsight distributions in both schemes can be generalized to time-independent distributions which yield the probability of taking an action in a some future state.

The work compares HCA to policy gradient on three distinct tasks highlighting the problem of credit assignment in RL. HCA outpeforms policy gradient in policy adaptation and reducing variance. Additionally, HCA demonstrates suitability in learning without a temporal component. However, the scheme presents two shortcomings. Firstly, (HCA|State) and (HCA|Return) depict equivalent performance in noise reduction, indicating that the significance of their individual components has little effect on agent's actions. Secondly, (HCA|Return) depicts comparable performance to the baseline policy gradient in the delayed effect task as a result of modeling the conditional noise distribution. This presents a hindrance in scalability of the algorithm to large action spaces wherein the agent may have a variety of options to select in its initial state.