Enhancing Centralized Value Functions for Cooperative Multi-agent Reinforcement Learning

Xinghu Yao, Chao Wen, Yuhui Wang, Xiaoyang Tan

AAAI 2020 accepted

Nanjing University of Aeronautics and Astronautics, China
MIIT Key Laboratory of Pattern Analysis and Machine Intelligence

Introduction

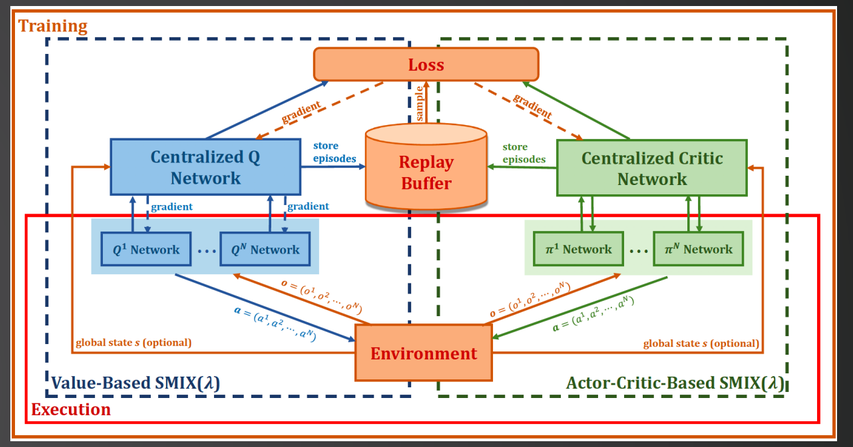
Learning a stable and generalizable centralized value function (CVF) is a crucial but challenging task in multiagent reinforcement learning (MARL), as it has to deal with the issue that the joint action space increases exponentially with the number of agents in such scenarios. This paper proposes an approach, named SMIX(λ), that uses an off-policy training to achieve this by avoiding the greedy assumption commonly made in CVF learning. As importance sampling for such off-policy training is both computationally costly and numerically unstable, we proposed to use the λ-return as a proxy to compute the TD error. With this new loss function objective, we adopt a modified QMIX network structure as the base to train our model. Experiments on the StarCraft Multi-Agent Challenge (SMAC) benchmark demonstrate that our approach not only outperforms several state-of-the-art MARL methods by a large margin, but also can be used as a general tool to improve the overall performance of other CTDE-type algorithms by enhancing their CVFs

Proposed method

The total loss of CVF is as follows:

$$L_t(\theta) = \sum_{i=1}^{N_b} \left[y_i^{tot} - Q_{tot}^{\boldsymbol{\pi}}(\boldsymbol{\tau}, \boldsymbol{a}; \theta) \right]$$

where $y_i^{tot} = \sum_{i=1}^{N_b} \lambda^{n-1} G_t^{(n)}$ is the TD target. $G_t^{(n)} = r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^n E_{\boldsymbol{\pi}} Q(\boldsymbol{\tau}_{t+n}, \boldsymbol{a}_{t+n}; \theta^-)$ is the n-step return.



Theoretical Analysis

The following theorem shows that the proposed SMIX(λ) is consistent with $Q(\lambda)$ algorithm.

Theorem. Suppose we update the value function from $Q_n^{smix}(\boldsymbol{\tau}_t, \boldsymbol{a}_t) = Q_n^{Q(\lambda)}(\boldsymbol{\tau}_t, \boldsymbol{a}_t)$, where n represents the n-th update. Let $\epsilon = \max_{\tau} \|\boldsymbol{\pi}(\cdot\,|\boldsymbol{\tau}) - \boldsymbol{\mu}(\cdot\,|\boldsymbol{\tau})\|_1$, $M = \max_{\boldsymbol{\tau},\boldsymbol{a}} \left| Q_n^{Q(\lambda)}(\boldsymbol{\tau}, \boldsymbol{a}) \right|$. Then, the error between $Q_{n+1}^{smix}(\boldsymbol{\tau}_t, \boldsymbol{a}_t)$ and $Q_{n+1}^{Q(\lambda)}(\boldsymbol{\tau}_t, \boldsymbol{a}_t)$ can be bounded by the expression:

$$\left| Q_{n+1}^{smix}(\boldsymbol{\tau}_t, \boldsymbol{a}_t) - Q_{n+1}^{Q(\lambda)}(\boldsymbol{\tau}_t, \boldsymbol{a}_t) \right| \leq \frac{\epsilon\gamma}{1-\lambda\gamma} M.$$

Results

StarCraft II (SMAC)