Inducing Cooperation through Reward Reshaping based on Peer Evaluations in Deep Multi-Agent Reinforcement Learning

AAMAS 2020

Ming,Will,Yang

Intelligent Transportation Group
State Key Laboratory of Internet of Things for Smart City
Faculty of Science and Technology
University of Macau

2020/11/14

Contents

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Contents

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Motivation

semi-cooperative

agents may each have its own separate reward function (the joint reward scenario is a special case), but is willing to cooperate if an incentive for cooperation is appropriately provided.

social welfare

the sum of the rewards of each agent across the entire episode.

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Contents

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Problem Setting

Stochastic game

a multi-agent task with each agent having its own individual reward by modeling it with a stochastic game

Semi-cooperative task

semi-cooperative tasks as the set of tasks where each agent may have a separate reward function but may benefit from cooperative strategies such as the prisoner's dilemma and the stag hunt game.

Goal: Maximizing the social welfare

the sum of the rewards of each agent across the entire episode.

$$\pi^* = \left(\pi_1^*\left(u_1 \mid o_1\right), \pi_2^*\left(u_2 \mid o_2\right), \ldots, \pi_n^*\left(u_n \mid o_n\right)\right)$$

$$\pi^* = \arg\max_{\pi} \sum_{a \in \mathcal{A}} \mathrm{E}_{s \sim \rho^{\pi}, \boldsymbol{u} \sim \pi}\left[r_a(s, \boldsymbol{u})\right]$$

Contents

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Solution

1. Change of Games via Reward Reshaping
2. Reward Update with Peer Evaluation
   1. Peer evaluation signal
   2. Reshaping reward from peer evaluation
3. Peer-evaluation based Dual DQN (PED-DQN)

Change of Games via Reward Reshaping

reward function

in the game G^t:

$$\hat{r}^t = \left(\hat{r}^t_a : a \in \mathcal{A}\right)$$

at time step $t = 0, 1, \cdots$

1. Compute the optimal π^t policy from G^t
2. Evaluate how well-coordinated G^t is by evaluating π^t
3. Update from G^t to G^{t+1} by updating from $\hat{r}^t$ to $\hat{r}^{t+1}$, using the 'evaluation feedback' from (2)
4. Increment t and go to (1)

Framework

Policy Update: $\pi^{t+1} = F\left(\pi^t, \hat{r}^t\right)$
Reward Update: $\hat{r}^{t+1} = H\left(\hat{r}^t, \pi^t\right)$

Reward Update with Peer Evaluation

Peer evaluation signal:

counterfactual evaluation signal (CES) z_k^t

for agent k:

$$z_k^t\left(o_k^t, o_k^{t+1}, u_k^t, r_k^t\right) := r_k^t + \gamma Q_k^{\pi^t}\left(o_k^{t+1}, \pi_k^t\left(o_k^{t+1}\right)\right) - Q_k^{\pi^t}\left(o_k^t, u_k^t\right)$$

Reshaping reward from peer evaluation:

agent a and agent k are peers.

For agent a:

$$Z_a^t\left[o_a^t, u_a^t\right] = \frac{1}{|K_a|}\sum_{k \in K_a} z_k^t$$

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

continue

sample mangy times:

$$\hat{Z}_a^t\left[o_a^t, u_a^t\right] \approx \frac{1}{|K_a|} \sum_{k \in K_a} \times$$
$$\mathbb{E}_{(o,o',u,r)\sim\pi^t : o_a = o_a^t, u_a = u_a^t}\left[z_k\left(o_k, o_k', u_k, r_k\right)\right]$$

moving average:

$$\hat{Z}_a^{t+1}\left[o_a^t, u_a^t\right] \leftarrow (1-\alpha)\hat{Z}_a^t\left[o_a^t, u_a^t\right] + \alpha Z_a^t\left[o_a^t, u_a^t\right]$$

final the reshaped reward $\hat{r}_a^t[o_a^t, u_a^t]$:

$$\hat{r}_a^t\left[0_a^t, u_a^t\right] = r_a^t + \beta\hat{Z}_a^t\left[o_a^t, u_a^t\right]$$

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Example: Evaluation feedback exchange

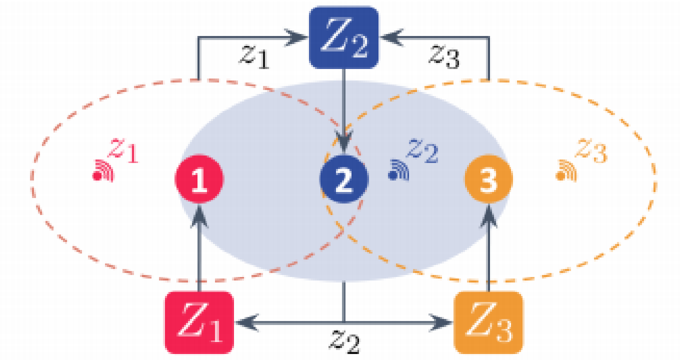


Figure: Example: Evaluation feedback exchange

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Peer-evaluation based Dual DQN (PED-DQN)

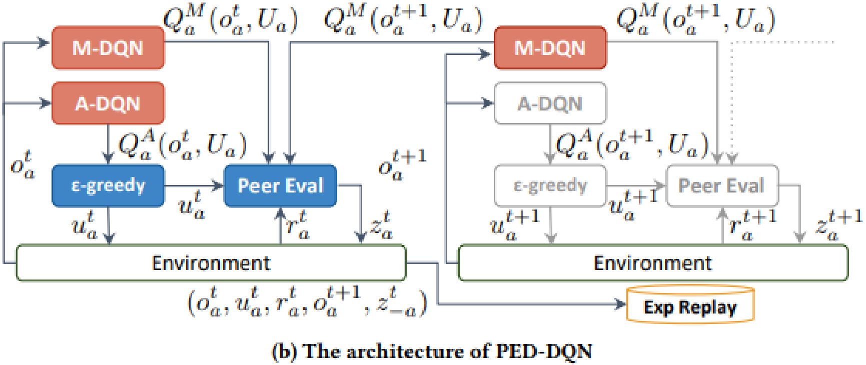


Figure: The architecture of PED-DQN

Contents

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Case study: Prisoner's dilemma

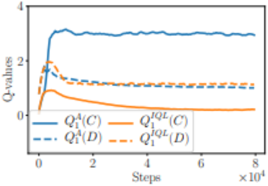
Table 1: Prisoner's dilemma

(a) original payoff

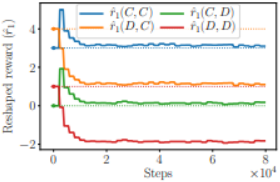
	C	D
C	3, 3	0, 4
D	4, 0	1, 1

(b) reshaped payoff

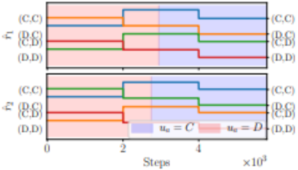
	C	D
C	3.11, 3.13	0.10, 1.13
D	1.11, 0.13	-1.90, -1.87



(a) IQL learned defection while ours learned cooperation



(b) Reshaped rewards $\hat{r}_1(D,C)$ and $\hat{r}_1(D,D)$ penalize defection



(c) Peer evaluations steered the agents'

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Experiments

Algorithms:

- QMIX
- IDQ PED-DQN
- PE:single network
- Pro DQN:directly use reward

$$z_a = r_a$$

$$\hat{r}_a = r_a + \frac{\beta}{|K_a|} \sum_{k \in K_a} r_k$$

Environments:

- Resource share
- Partially cooperative pursuit (PCP)

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Experiments

please read the paper to check the experiments details.

Contents

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Thoughts

communication & reward shaping