

Multi-agent System and Application

Xiangfeng Wang

Multi-Agent Artificial Intelligence Laboratory,
School of Computer Science and Technology,
East China Normal University

2021 Spring



华东师范大学计算机科学与技术学院
School of Computer Science and Technology



Table of Contents

Reward in Reinforcement Learning

- Potential Based Reward Shaping

- Intrinsically Motivated Reinforcement Learning

New Challenges in MARL

Credit Assignment

Social Influence

Table of Contents

Reward in Reinforcement Learning

- Potential Based Reward Shaping

- Intrinsically Motivated Reinforcement Learning

New Challenges in MARL

Credit Assignment

Social Influence

Reward in Reinforcement Learning

- ▶ Roles of reward and policy
 - ▶ Reward describes "What the agent should strive to do";
 - ▶ Policy describes "How should the agent behave";
 - ▶ Reward is the supervised information in reinforcement learning
- ▶ Challenges
 - ▶ Sparse reward: Agent can only get reward signal at the end of the game, which brings much difficulty for agent exploration and learning.
 - ▶ Delayed reward: (long-term credit assignment) Agents have to get the reward signal after a few steps and which will cause biases and high variance in training process. (sparse reward is also a delayed reward)

Eventually, these problems are caused by missing supervised signal – reward (Imagine that the supervised learning task on dataset with missing labels.)

Reward Shaping

- **Reward shaping** is an additional reward signal to reinforcement learning agents. It can be described as:

$$\hat{R} = R + F, \quad (1)$$

where F is the reward shaping and \hat{R} is the shaped reward. The additional reward signal can promote exploration, speed up training, reduce variance and etc (based on settings). But reward shaping without well-designed will change the original task and cause uncontrolled influence.

Potential Based Reward Shaping

- ▶ **Potential-based reward shaping**: A shaping reward function: $F : S \times A \times S \longrightarrow \mathbb{R}$ is a **Potential-based reward shaping** if there exists $\phi : S \longrightarrow \mathbb{R}$ s.t.

$$F(s, a, s') = \gamma\phi(s') - \phi(s) \quad (2)$$

for all $s \neq s_0, a, s'$.

- ▶ **Potential-based reward shaping** can maintain the original optimal policy – policy consistency. In another world, it can hold the original task with shaped reward. Which has been proven in ¹
- ▶ Eventually, **Potential-based reward shaping** is equivalent to an initialization of Q function. Which has been proven in ²

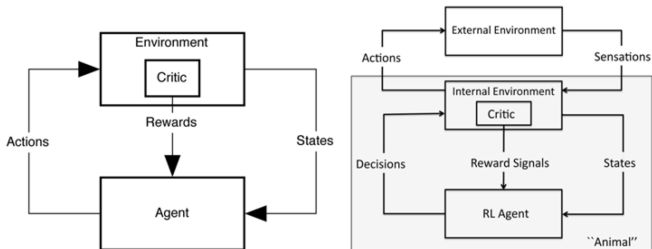
¹Ng A Y, Harada D, Russell S. Policy invariance under reward transformations: Theory and application to reward shaping[C]//ICML. 1999, 99: 278-287.

²Wiewiora E. Potential-based shaping and Q-value initialization are equivalent[J]. Journal of Artificial Intelligence Research, 2003, 19: 205-208.

Intrinsically Motivated Reinforcement Learning

- ▶ Formulation: $\hat{R} = R_{external} + R_{intrinsic}$, where external reward is from environment, intrinsic reward is something the agent enjoys. (Eventually the same as reward shaping)
- ▶ Motivations
 - ▶ “Forces” that energize an organism to act and that direct its activity
 - ▶ Extrinsic Motivation: being moved to do something because of some external reward (such as a prize, etc.)
 - ▶ Intrinsic Motivation: being moved to do something because it is inherently enjoyable (curiosity, exploration, manipulation, play, learning itself. . .)
- ▶ Normal types:
 - ▶ Curiosity Driven
 - ▶ Visitation Counts
- ▶ Usually, intrinsically motivated aims to promote exploration for the agent.

Intrinsically Motivated Reinforcement Learning



VIME: Variational Information Maximizing Exploration³(Curiosity Driven)

- ▶ **Main Idea:** Using the variational information in learning dynamic as the intrinsic reward.
- ▶ Maximizing the sum of reductions in entropy:
$$\sum_t [H(\Theta|\zeta_t, a_t) - H(\Theta|S_{t+1}, \zeta_t, a_t)] = \sum_t I(S_{t+1}; \Theta|\zeta_t, a_t),$$
where $\zeta_t = \{s_1, a_1, \dots, s_t\}$ is the current historical in formations.

- ▶ The agent is encouraged to take actions that lead to states that are maximally informative about the dynamics model:

$$I(S_{t+1}; \Theta|\zeta_t, a_t) = \mathbb{E}_{s_{t+1} \sim P(\cdot|\zeta_t, a_t)} [D_{KL}[p(\theta|\zeta_t, a_t, s_{t+1})||p(\theta|\zeta_t)]] \quad (3)$$

- ▶ Then the shaped reward can be expressed as:

$$r'(s_t, a_t, s_{t+1}) = r(s_t, a_t) + \eta D_{KL}[p(\theta|\zeta_t, a_t, s_{t+1})||p(\theta|\zeta_t)] \quad (4)$$

³Houthoofd R, Chen X, Duan Y, et al. VIME: Variational Information Maximizing Exploration[C]//Neural Information Processing Systems (NIPS). 2016.

Unifying Count-Based Exploration and Intrinsic Motivation⁴ (Visitation Counts)

- ▶ This idea is simple, if some state $s \in S$ hasn't been visited many times, encouraging the agent to visit the state

$$r(\hat{s}_t, a_t) = r(s_t, a_t) + \mathcal{B}(\hat{N}(s_t)), \quad \mathcal{B}(\hat{N}(s_t)) = \sqrt{\frac{1}{\hat{N}(s_t)}}, \quad (5)$$

where $\hat{N}(s_t)$ is the visitation counts of s_t

⁴Bellemare M G, Srinivasan S, Ostrovski G, et al. Unifying Count-Based Exploration and Intrinsic Motivation[C]//NIPS. 2016.

Table of Contents

Reward in Reinforcement Learning

Potential Based Reward Shaping

Intrinsically Motivated Reinforcement Learning

New Challenges in MARL

Credit Assignment

Social Influence

New Challenges in MARL

- ▶ **Exploration** is still a challenge in MARL;
- ▶ Agents need to learn **coordinating** in MARL and consider the **social influence**;
- ▶ From long-term **credit assignment** to multi-agent **credit assignment**;

Eventually, these new challenges are all related to coordination.

But in MARL, reward design/shaping maintains the same formulation as in single agent reinforcement learning.

The different scenarios for these motivations

- ▶ Credit Assignment:
 - ▶ Agents are fully cooperative;
 - ▶ They share a same team reward r_g from the environment.
- ▶ Social Influence:
 - ▶ Agents are semi-cooperative;
 - ▶ Each agent i has its own reward r_i from the environment, but they need to cooperative so as to get a better global welfare.

Table of Contents

Reward in Reinforcement Learning

Potential Based Reward Shaping

Intrinsically Motivated Reinforcement Learning

New Challenges in MARL

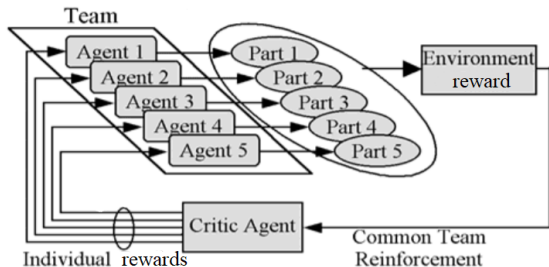
Credit Assignment

Social Influence

Credit Assignment in MARL

- ▶ Credit assignment of MARL is considered in such environment where agents are fully cooperative and share a same team reward which may not be decomposed among agents.
- ▶ So in this setting, credit assignment problem will show:
 - ▶ Agent cannot access its real own contribution to the cumulative team reward;
 - ▶ As the average performance is good enough, some agent will not exploration any more as exploration may reduce the performance. Which is called **lazy agent**.

The basic framework for Credit Assignment by reward shaping




The core is to use the reward shaping/design methods to decomposing the global environment reward into individual rewards so as to show each agent's contribution.

Credit Assignment For Collective Multiagent RL With Global Rewards⁵

- ▶ Main Idea: Using a **difference reward** based **shaped reward** to deal with credit assignment problem in MARL with Collective Decentralized POMDP setting.
- ▶ Collective Decentralized POMDP setting:
 - ▶ agent identities do not matter (agents are homogeneous);
 - ▶ different model components are only affected by agent's local state-action, and a statistic of other agents' states-actions;
 - ▶ a global reward signal r_g that is not decomposable among individual agent.
- ▶ Basic formulation:

$$\text{For each agent } m, r^m = r(s, a) + F^m \quad (6)$$

where $r(s, a)$ is the team reward, F^m is agent m 's difference reward based shaped reward.

⁵Nguyen D T, Kumar A, Lau H C. Credit Assignment For Collective Multiagent RL With Global Rewards[C]//NeurIPS. 2018. 

Credit Assignment For Collective Multiagent RL With Global Rewards

Difference Rewards

- ▶ Difference rewards provide a powerful way to perform credit assignment when there are several agents.
- ▶ Difference rewards (DR) are **shaped rewards** that help individual agents filter out the noise from the global/team reward signal.
- ▶ As such, there is no general technique to compute DRs for different problems.
- ▶ Two difference rewards in this work:
 - ▶ Wonderful Life Utility (WLU);
 - ▶ Aristocratic Utility (AU).

Credit Assignment For Collective Multiagent RL With Global Rewards

Wonderful Life Utility (WLU):

- ▶ For a given joint state-action (s, a) , the WLU based DR for an agent m is defined as:

$$r^m = r(s, a) - r(s, \mathbf{a}^{-m}), \quad (7)$$

where \mathbf{a}^{-m} is the joint-action without the agent m .

- ▶ The WLU DR compares the global reward to the reward received when agent m is not in the system, so that m can get its own contribution. (In real task, such shaped reward need to be approximated as it is hard to access directly.)

Credit Assignment For Collective Multiagent RL With Global Rewards

Aristocratic Utility (AU)

- ▶ For a given joint state-action (s, a) , the AU based DR for an agent m is defined as:

$$r^m = r(s, a) - \sum_{a^m} \pi^m(a^m | o^m(s)) r(s, a^{-m} \cup a^m), \quad (8)$$

where $a^{-m} \cup a^m$ is the joint action where agent m 's action in a is replaced with a^m ; o^m is the observation of the agent; π^m is the probability of action a^m .

- ▶ The AU marginalizes over all the actions of agent m keeping other agents' actions fixed so as to get the contribution of m .

Tips: Difference rewards are task specific! The performance is heavily related to specific task.

Table of Contents

Reward in Reinforcement Learning

Potential Based Reward Shaping

Intrinsically Motivated Reinforcement Learning

New Challenges in MARL

Credit Assignment

Social Influence

Social Influence

- ▶ Social influence intrinsic motivation gives an agent **additional reward** for having a **causal influence** on another agent's actions
- ▶ Social influence intrinsic motivation under the setting that each agent has its **own reward** but they need to coordinating so as to reach a better result.
- ▶ Specifically, it modifies an agent's immediate reward so that it becomes $\hat{r}_t^k = \alpha r_t^k + \beta c_t^k$, where r_t^k is the extrinsic or environmental reward and c_t^k is the causal influence reward.
- ▶ The causal influence reward can promote coordination among agents.

Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning⁶

Main works:

- ▶ Propose a unified method for achieving both coordination and communication in MARL by giving agents an intrinsic reward for having a causal influence on other agents' actions;
- ▶ The causal influence is assessed using counterfactual reasoning

⁶Jaques N, Lazaridou A, Hughes E, et al. Social influence as intrinsic motivation for multi-agent deep reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2019: 3040-3049.

Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning

- For a given joint state-action (s, a) , the causal influence based intrinsic reward for an agent m is defined as:

$$\hat{r}_t^m = \alpha r_t^m + \beta c_t^m \quad (9)$$

where r_t^m is the extrinsic or environmental reward and c_t^m is the causal influence reward.

- The causal influence reward for agent m is

$$\begin{aligned} c_t^m &= \sum_{j=0, j \neq m}^N [D_{KL}[p(a_t^j | a_t^m, s_t^j) || \sum_{\hat{a}_t^m} p(a_t^j | \hat{a}_t^m, s_t^j) p(\hat{a}_t^m | s_t^j)]] \\ &= \sum_{j=0, j \neq m}^N D_{KL}[p(a_t^j | a_t^m, s_t^m) || p(a_t^j | s_t^j)], \end{aligned} \quad (10)$$

where $p(\dots | \dots)$ is the policy (distribution) of specific agent.

Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning

- ▶ Social influence can reduce the variance of policy gradients caused by increasing number of agents by introducing explicit dependencies across the actions of each agent.
- ▶ Intrinsic social influence reward consistently leads to higher collective return.
- ▶ Using counterfactuals can allow each agent understand the effects of their actions on others, this will lead to coordination among agents.

Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning

How to calculate the causal influence reward ?

- ▶ Basic Influence
 - ▶ 1. Using centralized training to compute s_t^m directly from the policy of agent j
 - ▶ 2. Assume that influence is unidirectional: agents trained with the influence reward can only influence agents that are not trained with the influence reward (the sets of influencers and influencees are disjoint, and the number of influencers is in $[1, N-1]$)

But scenario with centralized training is less realistic than a scenario in which each agent is trained independently

Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning

How to calculate the causal influence reward ?

- ▶ Influential Communication
 - ▶ Equip agents with an explicit communication channel.
 - ▶ At each timestep, each agent k chooses a discrete communication symbol m_t^k , these symbols are concatenated into a combined message vector $\mathbf{m}_t = [m_t^0, m_t^1 \dots m_t^N]$
 - ▶ To train the agents to communicate, there need an additional output head of the model of agent to learn a communication policy π_m and value function V_m to determine which symbol to emit (see the figure)

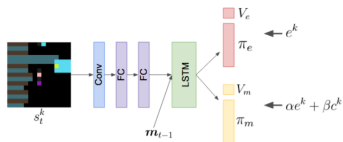


Figure 3: The communication model has two heads, which learn the environment policy, π_e , and a policy for emitting communication symbols, π_m . Other agents' communication messages \mathbf{m}_{t-1} are input to the LSTM.

Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning

How to calculate the causal influence reward ?

- ▶ Modeling Other Agents
 - ▶ Equip each agent with its own internal Model of Other Agents (MOA).
 - ▶ The MOA consists of a second set of fully-connected and LSTM layers connected to the agent's convolutional layer (see the figure in next page), and is trained to predict all other agents' next actions given their previous actions, and the agent's egocentric view of the state: $p(\mathbf{a}_{t+1} | \mathbf{a}_t, s_t^k)$
 - ▶ The MOA is trained using observed action trajectories and cross-entropy loss.

Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning

How to calculate the causal influence reward ?

- ▶ Modeling Other Agents
 - ▶ A trained MOA can be used to compute the social influence reward in the following way: Each agent can “imagine” counterfactual actions that it could have taken at each timestep, and use its internal MOA to predict the effect on other agents.

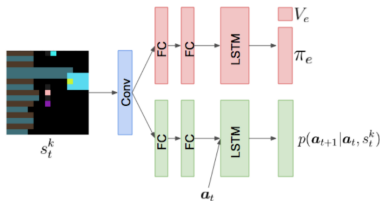


Figure 6: The Model of Other Agents (MOA) architecture learns both an RL policy π_e , and a supervised model that predicts the actions of other agents, a_{t+1} . The supervised model is used for internally computing the influence reward.

Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning

How to calculate the causal influence reward ?

- Performance of different methods to approximate the causal influence reward.

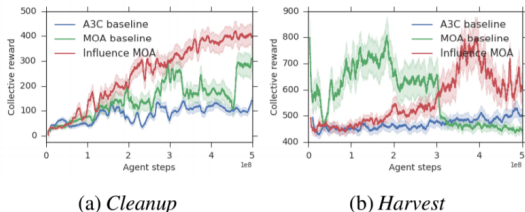


Figure 7: Total collective reward for MOA models. Again, intrinsic influence consistently improves learning, with the powerful A3C agent baselines not being able to learn.

We can find that MOA performs better.

Thank You