MULTI-AGENT EXPLORATION THROUGH INTRINSIC MOTIVATION CREDIT ASSIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse rewards are common in real-world environment, which brings challenges to reinforcement learning, previous works have made great progress in single-agent domain by introducing intrinsic rewards to encourage agent to access relatively new states in their state space, directly applying these methods to multi-agent setting results in agents exploring independently. In this work we attempt to designing intrinsic rewards for individual agent in order to provide more accurate exploration guidance for them. Concretely, we propose Intrinsic-Motivation-Credit-Assignment, a novel mechanism which derive exploration bonus for each agent according to their contribution to the whole system's exploration tendency, which measured by global intrinsic rewards. We verify the effectiveness of our mechanism on three cooperate exploration tasks, for each task we choose different exploration algorithm according to its feature including count-based algorithm, random network distillation and successor feature control. Results show that these algorithms greatly improved their performance with our mechanism in exploration-challenging tasks.

1 INTRODUCTION

Reinforcement learning algorithm has achieve excellent performance in many scenarios, such as Atari gamesMnih et al. (2015), the board game GoSilver et al. (2016), and simulated robotic continuous controlLillicrap et al. (2016), the goal is to learn a policy to maximizes the accumulative reward during agent interacting with the environment. Most of reinforcement learning algorithm's success depend on frequent reward signals from environment, which known as dense rewards, such as expert-designed rewardsWu & Tian (2017), distance to the goalMirowski et al. (2017) or the games' scoresMnih et al. (2015), however, those algorithm often struggle in real-world sparse rewards scenarios, where agent only receive reward signal when accomplish whole task. A widely used method proposed to address the sparse reward issue is using intrinsic motivation as an exploration bias to guide agent to explore unseen region. Many previous works have achieved satisfactory performance by adopting intrinsic reward as exploration bonus, however, most of them focus on single agent settingTang et al. (2017); Pathak et al. (2017); Zhang et al. (2019); Burda et al. (2019); Raileanu & Rocktäschel (2020); Badia et al. (2020), the literature on multi-agent is relatively scarceWang et al. (2020); Jaques et al. (2019); Chitnis et al. (2020); Iqbal & Sha (2019). Since intrinsic rewards are extra learning signal, it may destabilize the learning objective if it accounts too much of the overall rewards or it can not help improving exploration ability of agent if accounts too little, this make multi-agent exploration using intrinsic rewards especially difficult as each agent's exploration bonus need to be carefully designed. This paper aims to take a step towards this goal, we study how to manipulate individual intrinsic rewards as a more accurate exploration guidance for each agent, training with these extra rewards can help a group of agents explore their state space cooperatively.

We propose Intrinsic-Motivation-Credit-Assignment, a novel algorithm which derive exploration bonus for each agent according to their contribution to the whole system's exploration tendency which measured by global intrinsic rewards, IMCA

Credit assignment is another crucial challenge in multi-agent reinforcement learning:in most multi-agent cooperative tasks, agents' joint actions typically generate only global rewards, making it difficult for each agent to deduce its own contribution to the team's success. Previous work using counterfactual baseline like COMAFoerster et al. (2018) to estimate advantage function for updat-

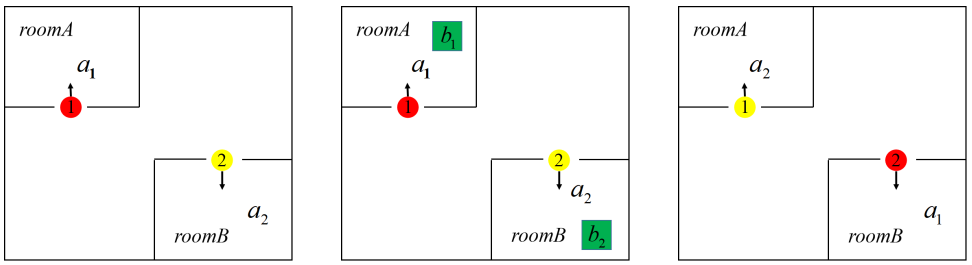


Figure 1: search grid

ing each agent's policy, or decomposed the central state-action value function to get individual value function of each agent and achieve satisfactory performanceSunehag et al. (2018); Rashid et al. (2018). All value networks used in those work are trained with extrinsic reward from environment, none of them consider credit assignment problem in intrinsic rewards. Our motivation is that intrinsic rewards as reward signals have the same shortcoming when applying to training process, i.e., credit assignment, agents trained with same intrinsic rewards are motivated equally, however, their contributions to the overall intrinsic rewards received are different, we take a example to illustrate this issue, as shown in 1(a), two agents, i.e., agent1 and agent 2 are required to search the big room with two small rooms, i.e., room$A$ and room$B$, suppose that agent1 is familiar with room$A$ and agent2 is not familiar with room$B$, that means agent1 step into room$A$ multiple times, and agent2 rarely enter room$B$, more formally, mark the state-action pairs of agent1 and agent2 at current time shown in figure[] as (o_1, a_1) and (o_2, a_2) respectively, and joint state-action pair as (S, U), where $S = \langle o_1, o_2 \rangle$ and $U = \langle a_1, a_2 \rangle$, the joint state-action novelty of (S, U) is very high mainly due to agent2 rarely take action a_2 at state o_2, although we can design high novelty based intrinsic rewards equally for agent1 and agent2, it is more reasonable to distinguish them by their contribution, that is intrinsically motivated agent2 more to visit room$B$ while less for agent1. One may have doubt why not let agent1 and agent2 explore their state space independently, we take another example to illustrate the irrationality of independent exploration, as shown in 1(b), as the same situation as 1(a), the two agents are required to complete the task cooperatively, i.e., press the buttons(green squares) concurrently, in this task, independent exploration is not suitable because agent1 may rarely enter the room$A$ in the later stage duo to fewer exploration bonus, which leads to the rare occurrence of joint state-action pair (S, U), thus they need a exploration strategy which motivate the system to explore all novel joint state-action space.

In addition, note the coordinated nature of multi-agent cooperative tasks, a team of homogeneous agents explore their joint state space cooperatively should take into account the permutation invariance of their position, otherwise may leads to redundant explore actions. For example, the relative locations of n homogeneous agents can be represented as $n!$ different states, when focusing on the relative configuration among agents, n! different states represent the same state, we use a example to illustrate it, as shown in 1(c), compared with figure[1], agent1 and agent2 exchange their positions, i.e., agent1 and agent2 are about to entering the room$B$ and room$B$ respectively, since the two agents are homogeneous, they search the same space from the perspective of whole task, thus the two situations should be treated equally, in another word, agents should be motivated equally using intrinsic rewards.

we claim the irrationality when directly extend single-agent algorithm to multi-agent system, and proposed an Intrinsic-Motivation-Credit-Assignment mechanism, with this mechanism, a group of agents show stronger exploration ability and perform better in cooperative exploration tasks. Also, our mechanism can be easily combined with existing algorithm, we verify the effectiveness of our mechanism using PPO as basic algorithm.

2 RELATED WORK

2.1 INTRINSIC MOTIVATION FOR SINGLE-AGENT EXPLORATION

To overcome sparse reward challenge in real-world scenarios, researchers have long worked on improving exploration in reinforcement learning. One commonly used way is adding intrinsic reward as exploration bonus to encourage agent to visit novel state, a variety of methods were proposed for designing such intrinsic reward, [] use inverse state-action counts in tabular setting[], [] scale those count-based approach to large state space using pseudo state counts, [ICM] trained a model for predicting next state and use predict error as exploration bonus to guide the agent to explore what makes it curious, [RND] use random state embedding network distillation error to define novelty of a state, [SF] calculate distance between successor feature of two consecutive states as exploration bonus while [] use the distance between the learned state representations.

2.2 MULTI-AGENT EXPLORATION

Compare to single agent, exploration is more difficult in multi-agent domain considering the unique cooperative property. [load balancing] study exploration of a large joint action space toward handle a load balancing problem, [coordinate exploration] design several intrinsic reward type and dynamically select one of them per episode to train, [social influence] define an intrinsic reward function based on influence on other agents' actions, which encourages agents to take actions have the biggest effect on other agents' behavior, [influence based] measures the influence of one agent on other agent's transition function (EITI) and rewarding structure(EDTI) and use the mutual information as a regularizer to the learning objective. Some of the above approach use novelty based intrinsic reward as a baseline but none of them consider the

2.3 MULTI-AGENT CREDIT ASSIGNMENT UNDER CENTRALIZE LEARNING WITH DECENTRALISE EXECUTION PARADIGM

Centralised learning with fully decentralised execution is well adopted in multi-agent learning [MADDPG, COMA], agents have access to global information during the training phase while only act upon their own observations during execution phase, in cooperative tasks they always awarded together as a team, however, the global reward received from the environment can't distinguish the contribution of each agent to the team's success, making it difficult to train their local policy. Many previous works have proposed approaches to address this issue and have obtained empirically solid results. [COMA] estimate a counterfactual advantage function for each agent to identify the real advantage of current action compared with other actions, [VDN] decomposed a central state-action value function into a sum of individual agent terms while [Qix] let the global value function be monotonic with each agent's value function.

3 BACKGROUND

3.1 DEC-POMDP

In this work, we restrict ourselves to multi-agent fully cooperative tasks, which is modeled as a decentralized POMDP(Dec-POMDP). A Dec-POMDP is defined as a tuple $\langle \mathcal{S}, U, P, r, O, \mathcal{Z}, n, \gamma \rangle$, n is the number of agents, $\boldsymbol{s} \in \mathcal{S}$ represent the global state and $\boldsymbol{u} \in U$ is the joint actions of each agent. At each time step t, each agent selects an action u^i and form a joint action $\boldsymbol{u}$, resulting in a shared extrinsic reward $r_t(\boldsymbol{s}, \boldsymbol{u})$ for each agent and transfer to the next state $\boldsymbol{s\prime}$ according to the transition function $P\left((\boldsymbol{s}_{t+1}|(\boldsymbol{s}_t, \boldsymbol{u}_t)\right.$. In centralized training and decentralized execution paradigm, agents choose their actions conditioned on their current partial observation o^i, which is drawn from the observation kernel $z^i\left(o_i|s_i\right)$, γ is the discount factor of immediate reward over long-term gain.

3.2 COUNTERFACTUAL BASELINE

Difference rewards has been proved powerful to perform multi-agent credit assignment, in which each agent learns from a shaped reward that compares the global reward to the reward received when the agent's action is replaced with a default action, a simulator is required to estimate the

reward function and the default action is hard to choose, COMA address this issue by estimating an agent-specific advantage function, those functions identify the real contribution of each agent's action to the centralized value function by fixing other agents' actions fixed1, in our work, we adopt the similar idea that is we calculate the advantage functions for each agent which reflect the contribution of their actions to global intrinsic rewards.

$$A^a(s, \mathbf{u}) = Q(s, \mathbf{u}) - \sum_{u'^a} \pi^a\left(u'^a|\tau^a\right) Q\left(s, \left(\mathbf{u}^{-a}, u'^a\right)\right) \tag{1}$$

3.3 Proximal Policy Optimization

Proximal Policy Optimization is a policy-based model-free RL methods, following the *Policy Gradients Theorem* Sutton & Barto (1998), the parameter ϕ of policy π can be updated with the following policy gradients:

$$\nabla_\phi J(\phi) = \mathbb{E}_{\pi_\phi}\left[\nabla_\phi \log \pi_\phi(a_t|s_t) q^{\pi_\phi}(s_t, a_t)\right]. \tag{2}$$

One typical policy gradients algorithm is the REINFORCE Williams (1992) that uses the complete return $G_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i}$ as estimates $\hat{Q}(s_t, a_t)$ of $q^{\pi_\phi}(s, a)$, i.e., Monte Carlo (MC) value estimation. An action-independent baseline $b(s)$, usually the value function $V(s)$, is commonly used to reduce the variance of policy gradients without introducing a bias. To further ensure an effective policy updates, Proximal Policy Optimization (PPO) Schulman et al. (2017) proposes a modified surrogate objective:

$$\mathcal{L}^{PPO}(\phi) = \mathbb{E}_{\pi_{\phi^-}}\left[\min\left(\rho_t \hat{A}(s_t, a_t), \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)\hat{A}(s_t, a_t)\right)\right], \tag{3}$$

where $\hat{A}(s_t, a_t)$ is the estimation of the advantage of old policy π_{ϕ^-}, $A^{\pi_{\phi^-}}(s_t, a_t) = q^{\pi_{\phi^-}}(s_t, a_t) - v^{\pi_{\phi^-}}(s_t)$, and $\rho_t = \frac{\pi_\phi(a_t, s_t)}{\pi_{\phi^-}(a_t, s_t)}$ is the importance sampling ratio. PPO can be viewed as a first-order approximation approach of Trust Region Policy Optimization (TRPO) Schulman et al. (2015). In addition to its simplicity, PPO is intended to be faster and more sample-efficient than TRPO.

4 Method

In order to overcome sparse reward challenge in reinforcement learning, intrinsic reward is commonly induced to improve exploration, we study how to manipulate individual intrinsic rewards for each agent in a multi-agent system, we achieve this goal by combining our mechanism and three widely used intrinsic reward based algorithm, i.e., count-based algorithm, random network distillation and successor feature control. Since those methods are all proposed for single-agent exploration, we extend them to multi-agent setting and show that they perform well combined with our mechanism in cooperative exploration tasks. Naturally, there are two directly way to apply those intrinsic motivation algorithm to multi-agent domain which used as comparison method in ablation experiment: One way is to execute independent exploration, i.e., agents explore the environment by their own according to their local traces, exploration bonus is generated according to their local state information such as local state novelty. Another way is to treat the team of agents as a central agent during centralize training phrase, interacting with environment using global state information and take joint actions, in this way exploration bonus can be designed according to global information such as joint state/state action pairs' novelty, agents are rewarded equally. Our intuition is that on one hand, using global intrinsic rewards generated according to global information to cooperate explore joint state action space is more comprehensive than using local information to do independent exploration. On another hand, when a group of agents executing cooperative exploration, a very rewarded novel joint state-action pair may due to several agents adopt relatively novel actions, thus it's unreasonable to give them the same reward bonus. So there should be a trade-off considering both global exploration bonus and the contribution of each agent to the overall novelty. We address this kind of trade-off by combining counterfactual baseline used in [COMA] and [the decomposition mechanism] used in VDN, which we called [intrinsic-motivation-credit-assignment], we compared to the state-of-art cooperative exploration algorithm[EDTI and EITI] and carry out ablation study both on [our intrinsic-motiation-credit-assignment mechanism] and [permutation invariance update]. [In the follow section, We describe our methods by three steps, first we introduce the overall process in section 3.2, then we describe three global intrinsic reward function used in our experiments, finally we discuss how each method work in our framework.]

4.1 GLOBAL INTRINSIC REWARD FUNCTION

As illustrated [above, section4?], independent exploration with local intrinsic rewards have defects in cooperative tasks, in this paper, we design intrinsic rewards from global perspective(i.e., global intrinsic reward) for representing the overall incentive accurately, we introduce three widely used intrinsic reward design schemes, i.e., count-based algorithm, random network distillation and successor feature control from two perspective, i.e., immediate novelty and long-time novelty as global intrinsic reward function used in our experiments.

Immediate Novelty Novelty based intrinsic reward is widely adopted to encourage agent to visit unseen region so as to gain potential rewards, the more novel the state, the greater the reward, many previous works use local information to evaluate intrinsic motivation which we called immediate novelty, two widely used methods to quantify immediate novelty of a state is count-based algorithm and Random Network Distillation, count-based algorithm design intrinsic reward using inverse state-action count, which has been shown to be effective in speeding up learning in tabular setting. The reward design has following form:

$$R_i = \frac{\beta}{\sqrt{C(s,a)}}, \tag{4}$$

where $N(s,a)$ represents the count of state-action pair (s,a), β is the weight coefficient, due to the simplicity and the poor generalization to large state-action space, we adopt this method in a simple two agent grid-world game.

Random Network Distillation use state predict error of a trained network and a deterministic randomly initialized network to quantify state novelty, due to the strong presentation ability and generalization of the neural network, it performs well in tasks with large state space. The traditional form of RND intrinsic rewards is as follows:

$$R_i = \beta \left\| \hat{f}(s) - f(s) \right\|^2, \tag{5}$$

where $f(x)$ is a fixed network, $\hat{f}(x;\theta)$ is training while agents interacting with the environment. In this work we make a little modification to the form by taking agents' actions into consideration, which help us calculate counterfactual baseline later:

$$R_i = \beta \left\| \hat{f}(s,a) - f(s,a) \right\|^2, \tag{6}$$

Long-time Novelty Instead of evaluating intrinsic motivation using local information, [SF] motivate agent to have diverse trajectories, those intrinsic rewards reflect the

Successor feature is proposed recent years, the successor feature of a state represents the characteristics of states expected to experienced from the beginning of the state using current policy, thus states have similar successor features are expected to have similar future trajectory, a recent work of [SF] proposed to generate exploration bonus according to the gap between two consecutive states' successor features, the larger the gap, the greater the difference in their subsequent trajectories, these exploration bonus always reflected in some narrow channels, through which agent can have diverse trajectories, agent training with those intrinsic reward is encourage to go through gates or channels, thus achieve the purpose of exploration. Their definition of intrinsic reward is:

$$R_i = \beta \left\| \varphi_{\pi,\phi}\left(S_{t+1}\right) - \varphi_{\pi,\phi}\left(S_t\right) \right\|_2^2 \tag{7}$$

where ϕ is a fixed feature network to embed state to vector, φ represents the successor feature network and π is the current policy.

In the following subsections we discuss how to calculate agents' individual intrinsic rewards for each global intrinsic rewards functions respectively.

4.2 GLOBAL INTRINSIC MOTIVATION CREDIT ASSIGNMENT

To combine our mechanism with above three global intrinsic rewards functions, the core work is to calculate the distribution proportion of the global intrinsic rewards according to the contribution of each agent, so intrinsic rewards assigned to each agent a is:

$$r_i^a = p^a \times R_i. \tag{8}$$

Immediate Novelty For the immediate novelty based intrinsic rewards function, let $N(S, U)$ be the novelty of a state-action pair, i.e., the part of formula 4 and 6 that removes the coefficient β. We obtain the difference reward of agent a using follow formula:

$$r^a = N(S, U) - \sum_{u^{a'}} \pi\left(u^{a'}|o^a\right) N\left(S, U^{-a}, u^{a'}\right), \tag{9}$$

where S and U represent the joint state and action, and u^a represents the action of agent a, U^{-a} is agents' joint action except a and π is agent a's current policy. Through this form, the difference reward r^a represent how much the current action of the agent contributes to the global novelty compared with other actions, the bigger the value, the bigger the contribution, negative value means current action is less novel against other actions and is well explored. Since those difference rewards can reflect their relative contribution to the global novelty, they serve as a distribution proportion of the global intrinsic reward after normalization:

$$p^a = \frac{e^{r^a}}{e^{r^1} + \ldots + e^{r^n}}. \tag{10}$$

Long Time Novelty Reward function in 7 is defined on distance of two consecutive global states' successor feature which quantifies the difference of the subsequent trajectories after taking the joint action. To identify each agent's contribution to the overall changes of subsequent trajectories, a feasible solution is fixing other agents still:

$$d^a = \left\|\varphi_{\pi,\phi}\left(S_t^{-a}, s_{t+1}^a\right) - \varphi_{\pi,\phi}\left(S_t\right)\right\|_2^2 \tag{11}$$

d^a represent the influence of the state transition of agent a on the subsequent trajectory when other agents stay still which reflects its contribution to the change of subsequent trajectories, thanks to the l_2 distance is non-negative, it can serve as a distribution proportion directly:

$$p^a = \frac{e^{d^a}}{e^{d^1} + \ldots + e^{d^n}}. \tag{12}$$

In addition, due to the successor feature control will not vanish with training process like novelty based intrinsic rewards, we find in our experiments that agents keep passing through the channel repeatedly, we avoid this situation by taking agents' episodic novelty into consideration, let $N\left(o^a\right)$ represent the visit count of agent a 's local observation o^a, the final internal reward has the following form:

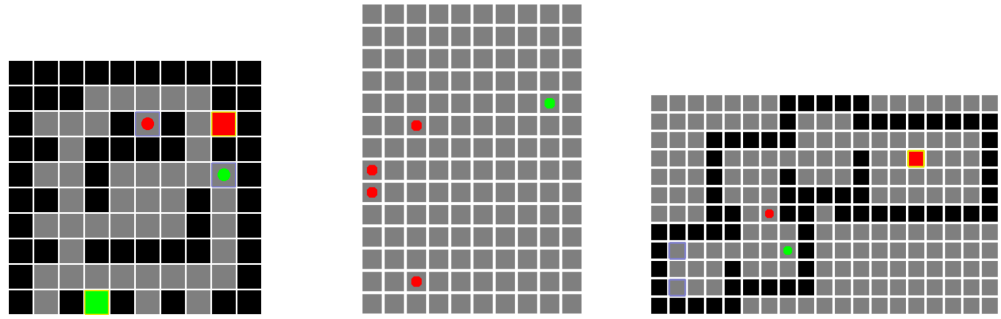
$$r_i^a = \frac{R_i \times p^a}{\sqrt{N\left(s^a\right)}}. \tag{13}$$

4.3 Permutation Invariant update

Consider the coordinated nature of multi-agent cooperative tasks, the position exchange between a team of homogeneous agents has no effect on the completion of the task, for example, a team of sweeping robots have to complete a cleaning task together, as long as one robot sweeps an area, other robots do not need to sweep this area twice in order not to repeat the cleaning, from the perspective of exploration, robot should not be motivated to explore the area where others have visited. For this consideration, we introduce permutation invariant update rule for agents' intrinsic rewards functions, take count-based algorithm for example, $C(\boldsymbol{s}, \boldsymbol{u})$ in 4 represents the visit count of global state-action pair $(\boldsymbol{s}, \boldsymbol{u})$, where $\boldsymbol{s} = \langle o_1, \ldots, o_n\rangle$, $\boldsymbol{u} = \langle u_1, \ldots, u_n\rangle$, let $\boldsymbol{s}' = \langle o_1', \ldots, o_n'\rangle$ is a random permutation sequence of $\langle o_1, \ldots, o_n\rangle$, and $\boldsymbol{u}' = \langle u_1', \ldots, u_n'\rangle$ is the corresponding sequence of $\langle u_1, \ldots, u_n\rangle$, instead of updating $C(\boldsymbol{s}, \boldsymbol{u})$, we updating $C(\boldsymbol{s}', \boldsymbol{u}')$ alternatively, which make global intrinsic rewards R_i and R_i' basically identical when $(\boldsymbol{s}, \boldsymbol{u})$ is a random permutation sequence of $(\boldsymbol{s}', \boldsymbol{u}')$.

5 Experiments

We focus on multi-agent cooperative tasks in gridworld domain and designe three multi-agent scenarios: Cooperative Navigation(2(a)), Predator and Prey(2(b)) and Multi-Room Search(2(c)). Each



(a) Cooperative Navigation (b) Predator and Prey (c) Multi-Room Search

Figure 2: Environment scenarios

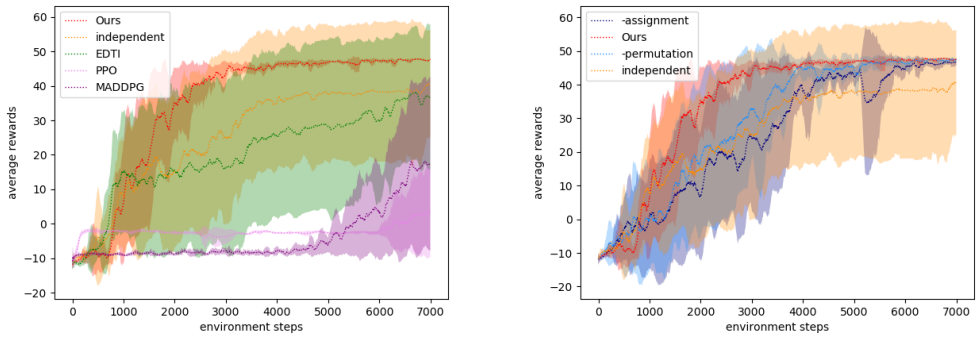


Figure 3: Two agent are required to find the two treasures in the grid world. We plot the mean rewards of

scenario used above has its own characteristics, Cooperative Navigation task is relatively simple due to the small state and action space which count-based exploration is suitable to handle, however, joint state/action space grows exponentially as the number of agents increases in Predator and Prey scenario which highlight the advantages of using neural network to generate exploration bonus as used in Random Network Distillation. Successor feature control encourage agent to go through channels(doors) to reach new states like another room, thus it perform well in environment which have lots of rooms[replace]. We use PPO as basic algorithm to combine with our mechanism, we compared [our algorithm] with PPO and MADDPG as well as state-of-art exploration algorithm EDTI[influence based].

5.1 Cooperative Navigation

In this scenario, two agent(red and green circles) must cooperatively collect all treasures(red and green squares) on the map in order to complete the task, at each time step, agents choose one of five actions: up, down, left, right or stay still, and move to adjacent grids according to the action if not blocked by walk(black grids) or other agents, a reward of 50 is award to each agent when all treasure are collected and agents receive a step penalty of 0.1 whenever take a step.

5.2 Predator and Prey

As shown in 2(b), in this scenario four predator(red circles) aim to catch a prey(greed circle) cooperatively, predators spawn randomly on the bottom of the world, prey on the top. Prey moves randomly and is restricted its position to the top 5 rows, it's caught by the predators only if preda-

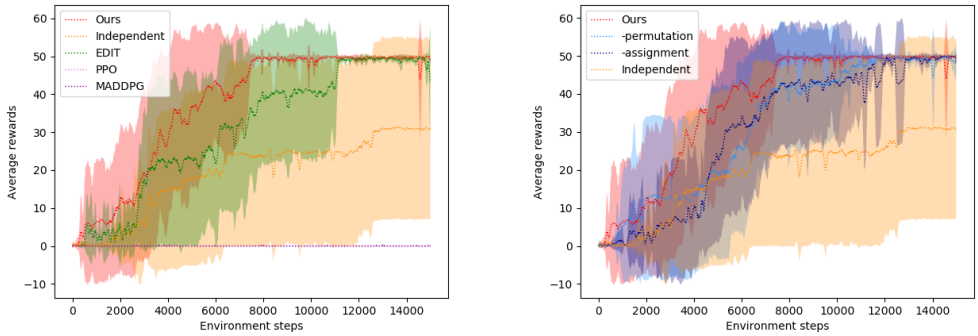


Figure 4: Two agent are required to find the two treasures in the grid world. We plot the mean rewards of

tors/boundaries surround it on all sides. An episode ends either when the prey is caught or after 100 steps, all predators received a reward of 20 when catching the prey.

5.3 Multi-Room Search

In this scenario, agents(red and greed circles) must go through several doors to enter the final room to get the treasure(red square) together. The door connected two adjacent rooms always allows only one agent to pass through which need two agents to pass cooperatively.

6 Conclusion

References

Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andrew Bolt, and Charles Blundell. Never give up: Learning directed exploration strategies. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Rohan Chitnis, Shubham Tulsiani, Saurabh Gupta, and Abhinav Gupta. Intrinsic motivation for encouraging synergistic behavior. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 2974–2982. AAAI Press, 2018.

Shariq Iqbal and Fei Sha. Coordinated exploration via intrinsic rewards for multi-agent reinforcement learning. *CoRR*, abs/1905.12127, 2019.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Çaglar Gülçehre, Pedro A. Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June*

2019, Long Beach, California, USA, volume 97 of *Proceedings of Machine Learning Research*, pp. 3040–3049. PMLR, 2019.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andy Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, Dharshan Kumaran, and Raia Hadsell. Learning to navigate in complex environments. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nat.*, 518(7540):529–533, 2015.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787. PMLR, 2017.

Roberta Raileanu and Tim Rocktäschel. RIDE: rewarding impact-driven exploration for procedurally-generated environments. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=rkg-TJBFPB`.

Tabish Rashid, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4292–4301. PMLR, 2018.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1889–1897. JMLR.org, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nat.*, 529(7587):484–489, 2016.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinícius Flores Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In Elisabeth André, Sven Koenig, Mehdi Dastani, and Gita Sukthankar (eds.), *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, pp. 2085–2087. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2018.

R. S. Sutton and A. G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 2753–2762, 2017.

Tonghan Wang, Jianhao Wang, Yi Wu, and Chongjie Zhang. Influence-based multi-agent exploration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256, 1992.

Yuxin Wu and Yuandong Tian. Training agent for first-person shooter game with actor-critic curriculum learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Jingwei Zhang, Niklas Wetzel, Nicolai Dorka, Joschka Boedecker, and Wolfram Burgard. Scheduled intrinsic drive: A hierarchical take on intrinsically motivated exploration. *CoRR*, abs/1903.07400, 2019.

A Appendix

You may include other additional sections here.