The research on deep reinforcement learning is very trending. Given its strong connection with self-learning, adaptation, and optimization, applying it to multi-agent systems is very attractive for many researchers. In this summary, I would like to list several topics of Multi-Agent Reinforcement Learning (MARL) that are being intensively studied in recent years.

1. Algorithms for the cooperative and competitive environment and framework

Unlike a system that mainly features a single agent, a system consisted of multiple agents is more complex, involving interaction between different agents. The inter-agent relationship can either be cooperative, competitive or even both. So, agents' behaviors, which is influenced not only by the static environment but also by sensing and feedback of peer agents, will evolve adaptively. Agents that adopt cooperative behaviors aims to increase positive team reward while agents that adopt competitive behaviors act for the benefit of oneself. More specifically, cooperation returns a positive value to an agent when its teammate is putting effort into achieving a goal whereas competition returns a negative value.

[4] first points out the limitation of Q-learning in a non-stationary environment that violates Markov assumptions. Then it proposes an extension of the actor-critic policy gradient: the critic takes in the actions of all agents and each Q-value is learned separately. Moreover, the limitation is more severe in terms of competitive settings as it will lead to overfitting. To solve the problem, the paper offers a solution to prevent one policy from growing too strong that only matches certain behaviors of competitors. It trains a collection of sub-policies, randomly selects one for each agent to execute and maximizes the expectation of reward as a whole.

On the other hand, [6] focuses on scaling a cooperative reinforcement learning algorithm to high dimensional raw observations, which is private to every agent and correlated with the state of environment, and to continuous action domains with large numbers of agents, which is a progress from [2].

[9] also identifies the drawbacks mentioned above and aims at a fully decentralized actor-critic algorithm, which relies only on neighboring communication through network. The paper points out some flaws in the SPG and DPG based AC algorithms, that is, these algorithms may vary hugely on gradients or require off-policy that needs to know all agents' policies during learning. Thus, the paper unifies SPG and DPG to come up with expected policy gradient(EPG), which is crucial for improving AC algorithms to be fully decentralized as it offers a joint policy improvement. In the end, the paper also offers a complete convergence analysis.

Another interesting study is [8]. A desirable planning framework is to make reasonable decisions based on inference of external objects and clear self-recognition. In order to obtain such a framework for agents in a given goal space, the key idea of the paper is to interpret intentions of other agents and making decisions based on those beliefs. The subtle difference in interpretation, or "natural" principle, between agents, can lead to perceivable cooperation and reach equilibrium eventually, which matches human intuition. Intent-aware modeling has a good chance of improving the performance for the paper uses probability distributions to approximate the intent of others based on past observation summaries and current state observations. The reasonable relation between future intent and past behavior in the form of probability distribution can be further supported by automata theory and theory of mind. And in the next step, compute all possible goals based on agents' intrinsic and permanent attributes and finally choose the biggest one.

I then read about [11] on curiosity-driven learning in DRL. Since designing proper rewards for different environments is hard to quantify, the paper tries to fill in the blank of the systematic research on learning without extrinsic motivation but only with intrinsic rewards. Using prediction error to quantify rewards, the paper finds out that this kind of learning can really work out with faster learning speed in most human-designed environment because this kind of environment also gives more reward for exploring unknow. Nevertheless, the paper mentions limitation in measuring curiosity with prediction error when handling stochastic dynamics.

The newly published [10] continues to work on obtaining profit-maximized collaboration among agents. The paper adds an observation process adopted by a centralized critic and an attention mechanism to DDPG. The embedded

attention mechanism contains a set of parameters, called attention weights, associated with different sources, measuring the importance of each source. Once other agents change their policies, their attention weights change accordingly, and the agent will be informed and alter its own policy immediately. Thus, these extensions allow agents to generate an ever-changing joint policy adaptively. And since it is an extension from DDPG, it naturally has the same ability to deal with continuous actions. Thus, it solves the problem of the concurrent changing of agents' modeling in a cooperative distributed MARL setting.

2. Multi-robot navigation

Multi-robot navigation is a fundamental issue for real-world application. To let robots safely direct themselves through obstacles and other moving robots, developing a strategy to avoid collision is necessary and also challenging. Specifically, [7] studies the environment with nonholonomic differential drive robot moving on the Euclidean plane. The paper stresses out several problems present in prior centralized server-level methods, decentralized agent-level methods and decentralized sensor-level policies using supervised learning: perfect-sensing, parameter-tuning, scalability, predefined environment and delicate, and so on. Then it puts forward a more decentralized framework. The framework only takes raw data from sensors and uses a policy gradient-based reinforcement learning algorithm to train a complex multi-robot system.

3. Games and self-play

[3] studies a decision-making strategy in poker games in order to scale it to real-world games which are offered with partially flawed information. Nash equilibrium is certainly a preferable strategy; However, it may fail to converge and be easily misled due to too much dependence on external information. Therefore, the paper proposes to learn approximate Nash equilibrium by combining Fictitious Self-Play with neural network function approximation, which is an end-to-end learning strategy. In this method, the agent decides on how to act based on two strategies that are learned through two neural networks respectively. The first network is about action and response, using reinforcement learning to output the approximate best historical response to other agents' behaviors. And the second network produces a model considering the agent's passed strategies of itself using supervised learning.

4. Decisions related to cooperation and competition

When faced with the same task with many peers, what is a preferable policy to maximize your capability and what can possibly influence your decision? [5] takes an interesting and descriptive view towards the result and social effects of policies learned by agents using deep reinforcement learning. And by manipulating different environment settings, such as the abundance of resources, and properties of agents, such as the character or personality parameter, the paper observes the agents' learned tendency to either cooperate or defect and is even able to infer the complexity of certain behaviors from the tendency. The work contributes to understand and model real-world social dilemmas.

5. Solutions to communication learning

In multi-agent systems, learning to communicate by agents themselves is very important, as it will provide state knowledge to paint a whole picture of the world necessary for a reward-maximized decision and it is also unrealistic for human to devise a proper mechanism for communication is. [2] focuses on typical tasks in communication that are fully cooperative and partially observable. It assumes communication settings as centralized learning and decentralized execution and presents two possible learning methods: 1) exploit Deep Recurrent Q-Networks to solve partial observability problem and then either perform independent Q-learning assuming other objects to be static or share parameters among all agents; 2) directly share real-valued gradients using Q-networks during centralized learning.

6. Applications in dynamic traffic schedule

Nowadays, the traffic signal control system cannot adapt to real-time traffic congestion level, which is seen as one essential factor to add to the already huge traffic pressure. In order to address this problem without asking for big financial budgets to improve the infrastructure, [1] turns to work on dynamic traffic signal control policy using MARL in order to optimize the green signal duration and maximize the traffic flow. The paper models each intersection as an

agent and let it only handle surrounding traffic states to prevent dimensionality from growing exponentially. And as learning goes on, the schedule at each intersection will gradually abandon its initial random Round-Robin phase-shift manner and asynchronously adapt to a self-organizing behavior for optimization, effectively minimizing the average delay of the vehicles in the traffic network.

To conclude, recent MARL researchers extend existing reinforcement learning algorithms trying to apply it to non-stationary and high-dimensional multi-agent/robot systems and to ensure robust, simultaneous, independent learning results; Or, they creatively model real-life problems to multi-agent systems and derive (near) optimal strategies from learning to solve the complex problems. And a very common underlying theme for MARL research is to maximize the return from cooperation and competition between agents.

Paper List

[1] Prabuchandran, K. J., Hemanth Kumar AN, and Shalabh Bhatnagar. "Multi-agent reinforcement learning for traffic signal control." 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE, 2014.

[2] Foerster, Jakob, et al. "Learning to communicate with deep multi-agent reinforcement learning." Advances in Neural Information Processing Systems. 2016.

[3] Heinrich, Johannes, and David Silver. "Deep reinforcement learning from self-play in imperfect-information games." arXiv preprint arXiv:1603.01121 (2016).

[4] Lowe, Ryan, et al. "Multi-agent actor-critic for mixed cooperative-competitive environments." Advances in Neural Information Processing Systems. 2017.

[5] Leibo, Joel Z., et al. "Multi-agent reinforcement learning in sequential social dilemmas." Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

[6] Gupta, Jayesh K., Maxim Egorov, and Mykel Kochenderfer. "Cooperative multi-agent control using deep reinforcement learning." International Conference on Autonomous Agents and Multiagent Systems. Springer, Cham, 2017.

[7] Long, Pinxin, et al. "Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning." 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018.

[8] Qi, Siyuan, and Song-Chun Zhu. "Intent-aware multi-agent reinforcement learning." 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018.

[9] Zhang, Kaiqing, Zhuoran Yang, and Tamer Basar. "Networked multi-agent reinforcement learning in continuous spaces." 2018 IEEE Conference on Decision and Control (CDC). IEEE, 2018.

[10] Mao, Hangyu, et al. "Modelling the dynamic joint policy of teammates with attention multi-agent ddpg." Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[11] Burda, Yuri, et al. "Large-scale study of curiosity-driven learning." arXiv preprint arXiv:1808.04355 (2018).