Yixiang REN, Zhenhui YE, Yining CHEN, Xiaohong JIANG, Guanghua SONG, 2023. Soft-HGRNs: soft hierarchical graph recurrent networks for multi-agent partially observable environments. *Frontiers of Information Technology & Electronic Engineering*, 24(1):117-130. https://doi.org/10.1631/FITEE.2200073

Soft-HGRNs: soft hierarchical graph recurrent networks for multi-agent partially observable environments

Key words: Deep reinforcement learning; Graph-based communication; Maximum-entropy learning; Partial observability; Heterogeneous settings

Corresponding author: Guanghua SONG
E-mail: ghsong@zju.edu.cn
ORCID: https://orcid.org/0000-0003-3330-4978

Motivation

1. Human society could be regarded as a large-scale partially observable environment, where everyone has the functions of communicating with neighbors and remembering his/her own experience. Similarly, a large-scale multi-agent system can be modeled as a graph to solve cooperation problems.
2. Deterministic multi-agent deep reinforcement learning (MADRL) strategies in large-scale multi-agent systems suffer from the insufficient exploration and poor robustness. Therefore, we consider designing a stochastic strategy to enhance the exploration capability.
3. Collaboration of heterogeneous agents often has a wider range of application scenarios. We seek to aggregate information from heterogeneous agents effectively.

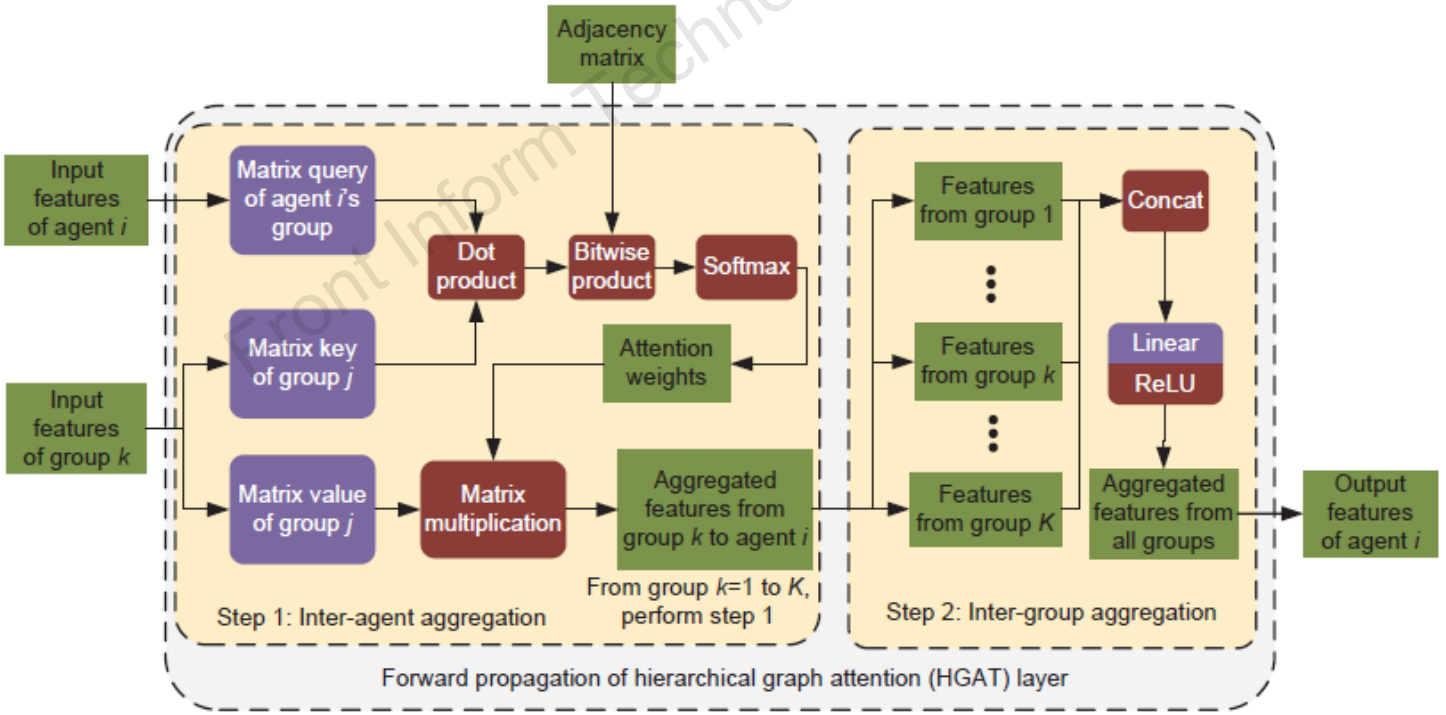
Main ideas

1. Inspired by the observation of human society, we propose a network structure called the hierarchical graph recurrent network (HGRN). HGRN combines the advantages of a graph convolutional network and a recurrent unit in MADRL. It uses spatio-temporal information to handle heterogeneous partially observable environments.
2. We propose two maximum-entropy MADRL algorithms (Soft-HGRN and SAC-HGRN) that introduce a learnable temperature parameter to learn our HGRN-structured policy with a configurable target action entropy.
3. Experiments show that our approach outperforms four MADRL baselines in several homogeneous and heterogeneous environments.

Method

1. To achieve communication between heterogeneous agents, we modify the hierarchical graph attention (HGAT) to better process the graph data with heterogeneous nodes.

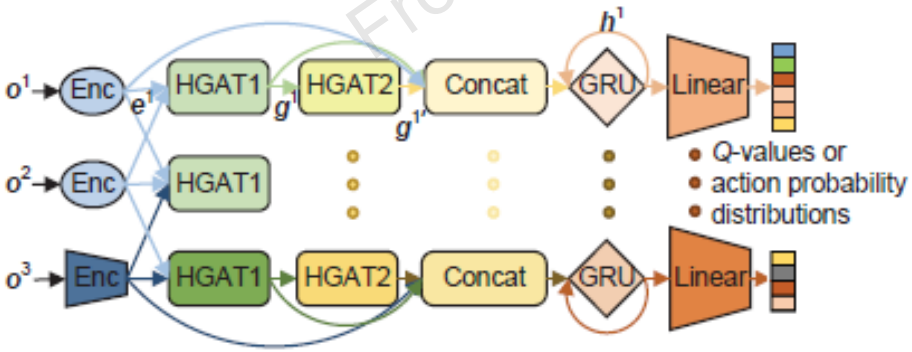
Forward propagation of **HGAT layer** is composed of two steps: the inter-agent and inter-group aggregation.



Method (Cont'd)

2. The overall network structure of our proposed HGRN:

- We use linear **encoders** to map the raw observation of heterogeneous agents to the same dimension.
- HGRN stacks **two HGAT layers** and has a two-hop perceptive field. With the HGAT-based communication channel, each agent can aggregate information from its heterogeneous neighbors to alleviate the information loss in POMDP.
- A gated recurrent unit (**GRU**) is used to record long-term temporal information. After the GRU, a **linear transform** is applied to infer the Q-values for all actions.



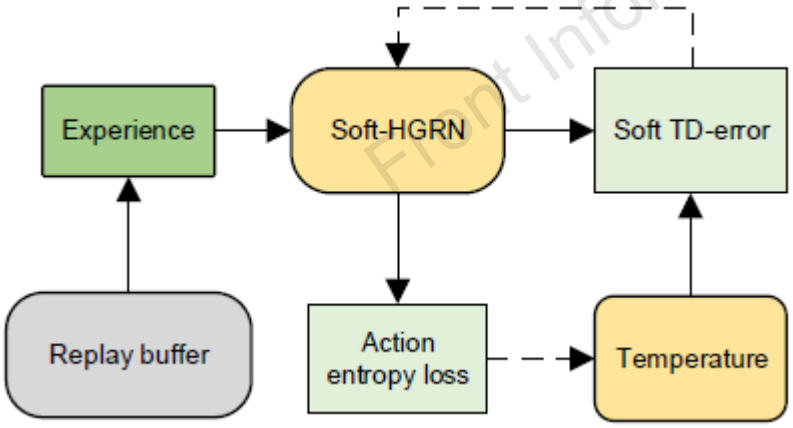
Overall network structure of HGRN

For scalability and sample efficiency, agents in the same group share the same parameters

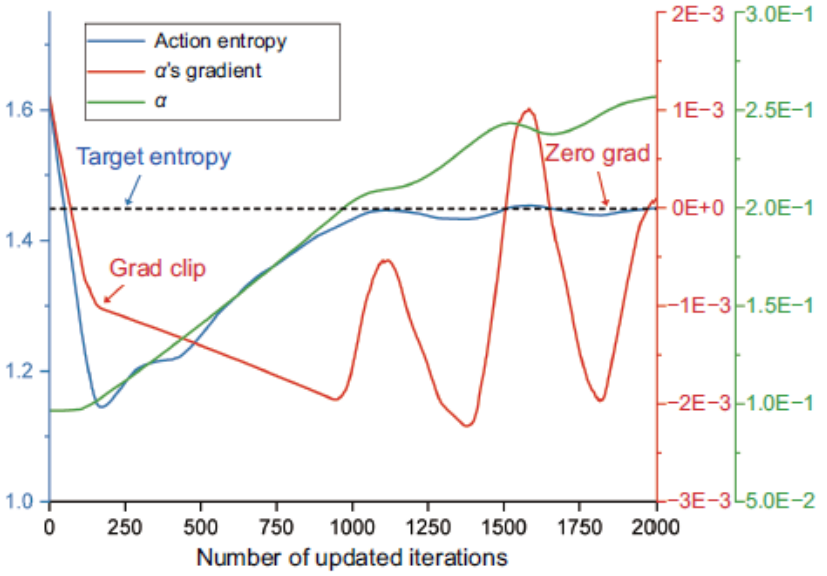
Method (Cont'd)

3. Deterministic policies can easily fall into a local optimum. Therefore, inspired by soft Q-learning, we adopt **maximum-entropy learning** in the multi-agent setting to train an energy-based stochastic model.

We define an action entropy to measure the degree of exploration of the policy. The action entropy is controlled by the temperature hyper-parameter α. We set a target action entropy to adaptively adjust the temperature parameter α.



Flowchart of the Soft-HGRN learning process



Learning curves of action entropy, temperature parameter α, and its gradient

Major results

- Experiments were carried out in four partially observable simulation environments, including three homogeneous scenarios (*UAV-MBS*, *Surviving*, and *Pursuit*) and one heterogeneous scenario (*cooperative treasure collection, CTC*)
- For **homogeneous environments**, our learned model outperforms four MADRL baselines, and ablation studies showed the necessity of each component in our approach.

Table 3 Evaluated episodic reward of different methods in three homogeneous tasks

Algorithm	Reward		
	UAV-MBS	Surviving	Pursuit
DQN	2314 ± 177	-7453 ± 74	5007 ± 150
CommNet	2377 ± 116	-7511 ± 82	5222 ± 60
MAAC	3435 ± 88	-10 ± 46	5927 ± 91
DGN	2808 ± 109	-286 ± 67	4552 ± 50
Soft-HGRN	$\underline{4051 \pm 75}$	$\underline{194 \pm 41}$	6840 ± 20
Soft-HGRN-G	3751 ± 59	-118 ± 90	$\underline{7138 \pm 96}$
Soft-HGRN-R	2935 ± 134	-26 ± 61	5649 ± 166
Soft-HGRN-S	3925 ± 46	-89 ± 35	5570 ± 30
SAC-HGRN	$\mathbf{4072 \pm 77}$	$\mathbf{325 \pm 48}$	7033 ± 55
SAC-HGRN-G	3580 ± 75	-36 ± 63	$\mathbf{7183 \pm 77}$
SAC-HGRN-R	3052 ± 22	141 ± 13	5797 ± 36
SAC-HGRN-S	3723 ± 82	-376 ± 27	-76 ± 1

"-G" means disabling HGAT-based communication among agents; "-R" means removing the GRU-based memory unit from the policy model; "-S" denotes training a deterministic policy instead of a stochastic policy. The best and second best models are indicated by bold and single underline, respectively

Major results (Cont'd)

- For the **heterogeneous environment CTC**, we focused mainly on the impact of HGAT's communication between different types of agents. It can be seen that every HGAT-based model performs better than its GAT-based variant, which demonstrates the effectiveness of the designed hierarchical communication structure.

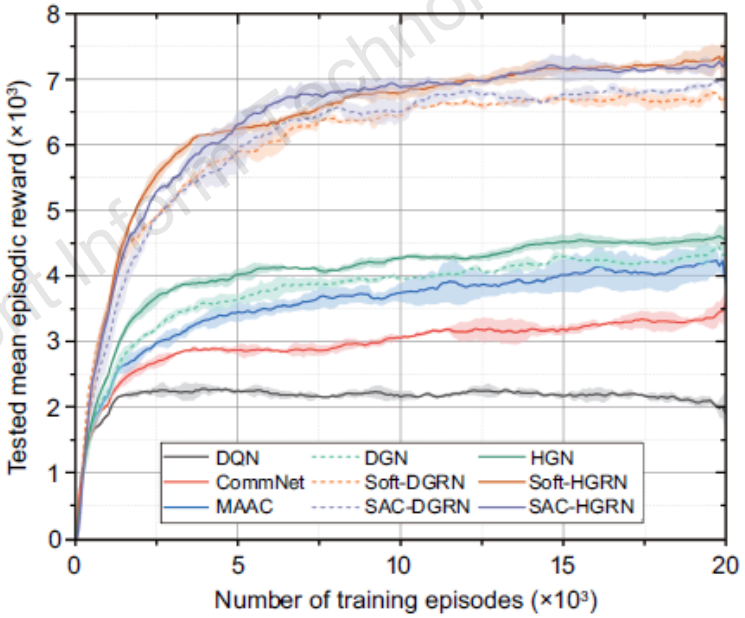


Fig. 5 Learning curves of various algorithms in cooperative treasure collection (CTC). References to color refer to the online version of this figure

Conclusions

1. We have proposed a value-based MADRL model Soft-HGRN and its actor-critic variant SAC-HGRN to address the large-scale multi-agent partially observable problem.
2. The proposed approach consists of two key components: (a) a novel network structure HGRN that could aggregate information from neighbors and history; (b) a maximum-entropy learning technique that could self-adapt the temperature parameter to learn a stochastic policy with configurable action entropy.
3. Experiments on four multi-agent environments demonstrated that the learned model outperforms four MADRL baselines. We have also analyzed the interpretability, scalability, and transferability of the learned model.

Yixiang REN, received the BS degree in flight vehicle design and engineering from Zhejiang University, China, in 2021. He is currently working toward the MSc degree in aerospace information technology with the School of Aeronautics and Astronautics at Zhejiang University. His recent research focuses on multi-agent reinforcement learning, UAV control and person re-identification.

Zhenhui YE, received the BS degree from Zhejiang University, China, in 2020. He is currently pursuing the MSc degree in College of Computer Science and Technology at Zhejiang University. His research interests include practical reinforcement learning, natural language processing and deep learning in real-world applications.

Yining CHEN, received the BSc degree from Sichuan University, China in 2015. He is currently a PhD student at Zhejiang University. His research interests include reinforcement learning and multi-robot system.

Xiaohong JIANG, received her BSc and MSc degrees in computer science from Nanjing University and the PhD degree from Zhejiang University, China, in 2003. She is an associate professor at the College of Computer Science and Technology, Zhejiang University. Her research focuses on distributed systems, cloud computing, and data service.

Guanghua SONG, received the BS degree in computer science from Nanjing University of Science and Technology, China, in 1989, and the MS and Ph D degrees in computer science from Zhejiang University, China, in 1992 and 2003, respectively. He is currently a full professor with School of Aeronautics and Astronautics, Zhejiang University, China. His research interests include swarm intelligence, UAV intelligence, and aerospace information technology.