CADP: Towards Better Centralized Learning for Decentralized Execution in MARL

Extended Abstract

Yihe Zhou
zhouyihe@zju.edu.cn
Zhejiang University
Hangzhou, China

Shunyu Liu*
shunyu.liu@ntu.edu.sg
Nanyang Technological University
Singapore

Yunpeng Qing
qingyunpeng@zju.edu.cn
Zhejiang University
Hangzhou, China

Tongya Zheng
doujiangzheng@163.com
Big Graph Center, Hangzhou City University
Hangzhou, China

Kaixuan Chen
chenkx@zju.edu.cn
State Key Laboratory of BC&DS, Zhejiang University
Hangzhou, China

Jie Song
sjie@zju.edu.cn
State Key Laboratory of BC&DS, Zhejiang University
Hangzhou, China

Mingli Song
brooksong@zju.edu.cn
State Key Laboratory of BC&DS, Zhejiang University
Hangzhou, China

ABSTRACT

Centralized Training with Decentralized Execution (CTDE) has recently emerged as a popular framework for cooperative Multi-Agent Reinforcement Learning (MARL), where agents can use additional global state information to guide training in a centralized way and make their own decisions only based on decentralized local policies. Despite the encouraging results achieved, CTDE makes an independence assumption on agent policies, which limits agents from adopting global cooperative information from each other during CT. Therefore, we argue that the existing CTDE framework cannot fully utilize global information for training, leading to an inefficient joint exploration and perception, which can degrade the final performance. In this paper, we introduce a novel Centralized Advising and Decentralized Pruning (CADP) framework for MARL, that not only enables an efficacious message exchange among agents during training but also guarantees DE.

KEYWORDS

deep learning; multi-agent reinforcement learning; CTDE

ACM Reference Format:
Yihe Zhou, Shunyu Liu, Yunpeng Qing, Tongya Zheng, Kaixuan Chen, Jie Song, and Mingli Song. 2025. CADP: Towards Better Centralized Learning for Decentralized Execution in MARL: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

*Corresponding author.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Recently, the CTDE framework has been widely used in MARL, including Value Decomposition (VD) methods [4, 6, 8–12, 14] and Policy Gradient (PG) methods [2, 5, 7, 16], which achieves the state-of-the-art performance in different benchmarks. Despite its promising success, we argue that the centralized training in CTDE is not centralized enough. This is to say, the existing CTDE framework cannot take full advantage of global information for centralized training. Specifically, agent policies are assumed to be independent of each other [15], and the existing CTDE framework only introduces global information in the centralized module, while agents are not granted access to global information even when CT.

To address this limitation, prior works have introduced teacher-student frameworks [1, 3, 17], as shown in Figure 1(b). In these setups, teacher agents utilize global state information for training, while student agents, relying on local observations, learn to mimic the teachers' behavior through knowledge distillation. However, these approaches still let agents make decisions without considering others' policies, leading to suboptimal joint exploration and limited expressiveness of the collective policy. In this paper, we propose a novel Centralized Advising and Decentralized Pruning framework, termed as CADP, to enhance basic CTDE with global cooperative information. As depicted in Figure 1(c), CADP enables agents to exchange advice with each other instead of only using global state information during centralized training. To generate the final decentralized policies, we further propose to smoothly prune the dependence among agents.

2 METHOD

Advice Exchanging. The widely adopted CTDE framework only introduces the global state for agents in the mix/critic module, leading to that an agent policy network only perceives its local observation instead of the global states. In contrast, we design a

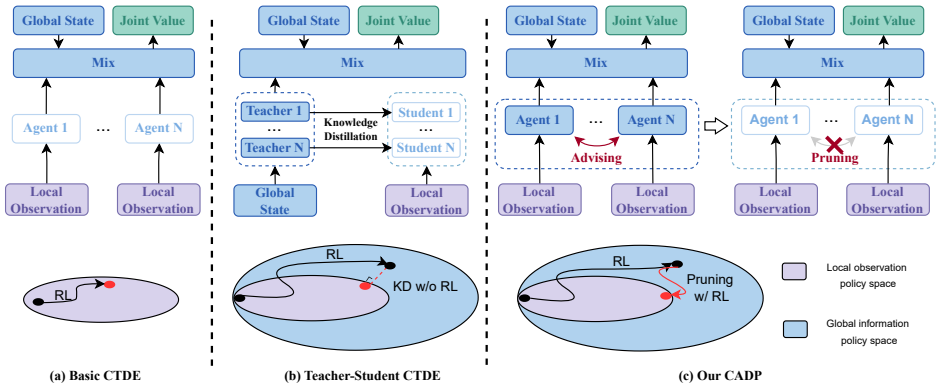

Figure 1: Comparisons between existing frameworks and our CADP.

novel centralized training scheme to augment the agent policy from the local information of an individual agent to the global cooperative information from all agents, inferring better actions.

Formally, we employ an agent's confidence c for all agents to highlight its personalized confidence weights of other agents when receiving interchangeable cooperative advice from them, where the higher confidence corresponds to the more useful information of agents. The whole process can be reduced to a self-attention mechanism [13], where we set messages key k and value v, respectively, while confidence c is considered as the dot product of the key k of other agents and query q of itself. q, k and v are all linear projections of the local observation o. The formula is written as:

$$\alpha_{i,j} := \frac{q_i \cdot k_j}{\sqrt{d_x}}, \quad c_{i,j} := \frac{exp(\alpha_{i,j})}{\sum_{k=1}^{N} exp(\alpha_{i,k})}, \quad z_i := \sum_{j=1}^{N} c_{i,j} \cdot v_j, \quad (1)$$

where k_j means the message key k of agent j and q_i means the query q of agent i, d_x is the scaling coefficient and $c_{i,j}$ is the confidence from agent i to agent j. Finally, the aggregating information z_i of the agent i is obtained by taking the weighted sum of the value according to the confidence $c_{i,1:N} := (c_{i,1}, c_{i,2}, \cdots, c_{i,N})$: where v_j means the value v of agent j. Through this step, each agent refers to the cooperative information of others. Then, we combine the aggregating information z in the previous step with the agent's own local information h, and finally output the action value Q:

$$h_i^t := GRU([z_i, o_i], h_i^{t-1}), \quad Q^i := MLP(h_i^t), \quad (2)$$

where $GRU(\cdot, \cdot)$ stands for Gated Recurrent Unit. Notably, we have incorporated residual connections into the input of the GRU network. This short-circuit mechanism allows us to simultaneously leverage representations with and without advice exchanging, for enhancing training stability and performance. Since our work focuses on the agent policy module, we can adopt different mix modules to generate Q^{tot}. Besides, if the final output of our agent module is a policy distribution π^i instead of Q^i (performing normalization after Equation 2), we can also employ MAPPO [16].

Model Self-Pruning. In pursuit of facilitating decentralized execution, the current model needs to evolve into the decentralized model depending only on itself, rather than global information. In this step, we design a simple yet effective model self-pruning method to achieve this. We claim that if the following conditions are satisfied, the agent model is a decentralized model: $c_{i,1:N} = \mathbf{e}^i, \forall i \in [1, N]$, where $\mathbf{e}^i$ means the i-th standard basis vector (an one-hot vector). In this way, we can just apply the value v to produce the output z without key k and other components. For the convenience of expression, we refer to the model that only requires their own values v without the self-attention mechanism as the decentralized model (CADP (D)). On the other hand, the model using self-attention mechanism is referred to as the centralized model (CADP (C)). A decentralized model actually presents that the agent's confidence $c_{i,1:N}$ of all agents is equal to the one-hot vector $\mathbf{e}^i$ in the DE stage. It is required to smoothly swap from the centralized training with exchanging confidence to the decentralized execution with the one-hot confidence $c_{i,1:N}$. Therefore, we design an auxiliary loss function named pruning loss $\mathcal{L}_p$ to help the decentralized agent gradually alleviate the dependence of other agents, which is given as:

$$\mathcal{L}_p(\theta_a) = \sum_{i=1}^{N} D_{KL}(\mathbf{e}^i \| c_{i,1:N}), \quad (3)$$

where θ_a means the parameters in the agent policy module, and D_{KL} means the Kullback Leibler (KL) Divergence. In the pruning loss, smaller $\mathcal{L}_p$ means that agents rely less on the other agents.

3 CONCLUSION

we argue that the traditional CTDE framework is insufficiently centralized, as it fails to fully utilize global information during training. we propose a novel Centralized Advising and Decentralized Pruning (CADP) framework, enhancing CTDE with global cooperative information. Our focus is not on designing a new communication method but on leveraging agent communication to enable fully centralized training while maintaining decentralized execution. Using just simple and lightweight network architectures to test CADP, we believe it opens avenues for further exploration.

4 ACKNOWLEDGMENTS

This work was supported in part by the Hangzhou Joint Funds of the Zhejiang Provincial Natural Science Foundation of China under Grant No. LHZSD24F020001, in part by the Fundamental Research Funds for the Central Universities under Grant No. 226-2024-00058, and in part by the Zhejiang Province High-Level Talents Special Support Program "Leading Talent of Technological Innovation of Ten-Thousands Talents Program" under Grant No. 2022R52046.

REFERENCES

[1] Yiqun Chen, Hangyu Mao, Tianle Zhang, Shiguang Wu, Bin Zhang, Jianye Hao, Dong Li, Bin Wang, and Hongxing Chang. 2024. PTDE: Personalized Training with Distillated Execution for Multi-Agent Reinforcement Learning. In *International Joint Conference on Artificial Intelligence*.

[2] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual Multi-Agent Policy Gradients. In *AAAI Conference on Artificial Intelligence*.

[3] Yitian Hong, Yaochu Jin, and Yang Tang. 2022. Rethinking Individual Global Max in Cooperative Multi-Agent Reinforcement Learning. In *Annual Conference on Neural Information Processing Systems*.

[4] Zican Hu, Zongzhang Zhang, Huaxiong Li, Chunlin Chen, Hongyu Ding, and Zhi Wang. 2024. Attention-Guided Contrastive Role Representations for Multi-Agent Reinforcement Learning. In *International Conference on Learning Representations*.

[5] Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. 2022. Trust Region Policy Optimisation in Multi-Agent Reinforcement Learning. In *International Conference on Learning Representations*.

[6] Shunyu Liu, Yihe Zhou, Jie Song, Tongya Zheng, Kaixuan Chen, Tongtian Zhu, Zunlei Feng, and Mingli Song. 2023. Contrastive Identity-Aware Learning for Multi-Agent Value Decomposition. In *AAAI Conference on Artificial Intelligence*. 11595–11603.

[7] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Annual Conference on Neural Information Processing Systems*.

[8] Shuang Luo, Yinchuan Li, Jiahui Li, Kun Kuang, Furui Liu, Yunfeng Shao, and Chao Wu. 2022. S2RL: Do We Really Need to Perceive All States in Deep Multi-Agent Reinforcement Learning?. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[9] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. 2020. Weighted QMIX: Expanding Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Annual Conference on Neural Information Processing Systems*.

[10] Tabish Rashid, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning*.

[11] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Hostallero, and Yung Yi. 2019. QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning*.

[12] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, et al. 2018. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *International Joint Conference on Autonomous Agents and Multi-agent Systems*.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Annual Conference on Neural Information Processing Systems*.

[14] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2021. QPLEX: Duplex Dueling Multi-Agent Q-Learning. In *International Conference on Learning Representations*.

[15] Jiangxing Wang, Deheng Ye, and Zongqing Lu. 2023. More Centralized Training, Still Decentralized Execution: Multi-Agent Conditional Policy Factorization. In *International Conference on Learning Representations*.

[16] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative, multi-agent games. In *Annual Conference on Neural Information Processing Systems*.

[17] Jian Zhao, Xunhan Hu, Mingyu Yang, Wengang Zhou, Jiangcheng Zhu, and Houqiang Li. 2022. CTDS: Centralized Teacher with Decentralized Student for Multi-Agent Reinforcement Learning. *IEEE Transactions on Games* 16, 1 (2022), 140–150.