From Explicit Communication to Tacit Cooperation: A Novel Paradigm for Cooperative MARL

Extended Abstract

Dapeng Li
Institute of Automation, Chinese Academy of Sciences School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
lidapeng2020@ia.ac.cn

Zhiwei Xu
Institute of Automation, Chinese Academy of Sciences School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
xuzhiwei2019@ia.ac.cn

Bin Zhang
Institute of Automation, Chinese Academy of Sciences School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
zhangbin2020@ia.ac.cn

Guangchong Zhou
Institute of Automation, Chinese Academy of Sciences School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
zhouguangchong2021@ia.ac.cn

Zeren Zhang
Institute of Automation, Chinese Academy of Sciences School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
zhangzeren2021@ia.ac.cn

Guoliang Fan
Institute of Automation, Chinese Academy of Sciences School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
guoliang.fan@ia.ac.cn

ABSTRACT

Centralized training with decentralized execution (CTDE) is a widely used learning paradigm that has achieved significant success in complex tasks. Drawing inspiration from human team cooperative learning, we propose a novel paradigm that facilitates a gradual shift from explicit communication to tacit cooperation. In the initial training stage, we promote cooperation by sharing relevant information among agents and concurrently reconstructing this information using each agent's local trajectory in a self-supervised way. We then combine the explicitly communicated information with the reconstructed information to obtain mixed information. Throughout the training process, we progressively decrease the proportion of explicitly communicated information, facilitating a seamless transition to fully decentralized execution without communication.

KEYWORDS

Reinforcement Learning; Multi-agent System; Tacit Cooperation.

ACM Reference Format:
Dapeng Li, Zhiwei Xu, Bin Zhang, Guangchong Zhou, Zeren Zhang, and Guoliang Fan. 2024. From Explicit Communication to Tacit Cooperation: A Novel Paradigm for Cooperative MARL: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand.

1 INTRODUCTION

Cooperative multi-agent reinforcement learning (MARL) has made significant progress in practical applications in recent years, such as traffic light control [1, 18], autonomous driving [11], game playing [2, 15], and multi-robot control [3, 8]. To effectively address multi-agent learning problems, various algorithms have emerged. Among these methods, the paradigm of centralized training with decentralized execution (CTDE) has gradually become the most concerned MARL paradigm due to its scalability and ability to handle non-stationary problems. The CTDE paradigm serves as a hybrid approach that combines the advantages of both centralized [5, 6] and decentralized [13] learning methods. The fundamental concept of the CTDE paradigm is that agents can access global information in a centralized manner during the training process while operating solely on local observations in a decentralized manner during execution. Based on this approach, many MARL algorithms [4, 7, 9, 12, 16, 17, 19] have demonstrated exceptional performance in some complex decision-making tasks [10].

Drawing inspiration from human teamwork, this paper proposes a novel paradigm that can transition from explicit communication to **TA**cit **CO**operation (**TACO**). This paradigm enables agents to share relevant information via an attention mechanism during the initial training stage while simultaneously reconstructing this information using local observations in a self-supervised manner. We then obtain mixed information by weighted sums of the reconstructed information and true information. As training progresses, the accuracy of the reconstructed information steadily improves. Consequently, we can reduce reliance on communication by gradually decreasing its proportion in the mixed information, ultimately achieving the ability to infer teammate's intentions without actual communication.

2 FROM EXPLICIT COMMUNICATION TO TACIT COOPERATION

2.1 The TACO Framework

Communication Abstract Module: The communication abstract module applies a self-attention mechanism [14] to aggregate highly relevant information from other teammates. Given the hidden state of agent i and agent j, the attention weight $w_{i,j}^{att}$ for agent i to agent j can be computed by using a bilinear mapping and then normalizing it with a softmax function, as shown below:

$$w_{i,j}^{att} = \frac{\exp(h_j^T W_k^T W_q h_i)}{\sum_{j \neq i} \exp(h_j^T W_k^T W_q h_i)}, \tag{1}$$

Tacit Reconstruct Module: The tacit reconstruct module is responsible for approximating the relevant attention information based on the agent's own local history trajectory. To achieve this, we use a two-layer fully connected network with the Relu activation function for simplicity. Specifically, the reconstruct network takes the hidden state h_i of agent i as input and outputs an approximation $\hat{v}_i$ of the actual relevant attention information v_i.

From Communication to Tacit Cooperation: To ensure a successful transition from communicate to tacit, we obtain the mixed information $\bar{v}_i$ by taking the weighted average of the real attention information v_i and the reconstructed information $\hat{v}_i$ as follows:

$$\bar{v}_i = (1 - \alpha)\hat{v}_i + \alpha v_i. \tag{2}$$

The mixed weight α starts with an initial value α_{init} and using a simple linear decreasing schedule, given by $\alpha_t = \max(\alpha_{init} - t\Delta\alpha, \alpha_{\min})$, which update during each training step. Therefore, as the training progresses, the proportion of explicit communication information in the mixed information gradually decreases. To make sure the agent can entirely transmit to fully tacit before training is completed, we usually set the $\alpha_{init} = 1$, $\alpha_{\min} = 0$, and $\Delta\alpha \geq \frac{1}{t_{\max}}$. The mixed information $\bar{v}_i$ and the hidden state h_i are concatenated to input MLP to obtain $Q_i(\tau_i, u_i, \bar{v}_i)$. The mixing network decomposes the joint action value function Q_{tot} into the individual action value estimation Q_i.

2.2 Overall Learning Objective

We now introduce the learning objectives of TACO, which include two parts: the reinforcement learning part that tries to minimize the TD error, and the mixed information part that attempts to minimize the reconstruct error.

The reinforcement learning part end-to-end optimizes the same loss function as QMIX [9]:

$$\mathcal{L}_{\mathrm{TD}} = \left(Q_{tot}(s, \boldsymbol{\tau}, \boldsymbol{u}) - y^{tot}\right)^2, \tag{3}$$

where $y^{tot} = r + \gamma \max_{\boldsymbol{u}'} Q_{tot}\left(s', \boldsymbol{\tau}', \boldsymbol{u}'\right)$. To achieve the goal of enforcing the reconstructed information $\hat{v}_i$ to be as close as possible to its corresponding true attention information v_i, TACO also includes a mixed information part that minimizes reconstruct loss. The similarity loss between the true relevant attention information v_i and the reconstructed information $\hat{v}_i$ is measured by using the MSE loss function:

$$\mathcal{L}_{\mathrm{Rec}} = \frac{1}{n}\sum_{i=0}^{n} \mathrm{MSE}(v_i, \hat{v}_i) = \frac{1}{n}\sum_{i=0}^{n} (v_i - \hat{v}_i)^2. \tag{4}$$

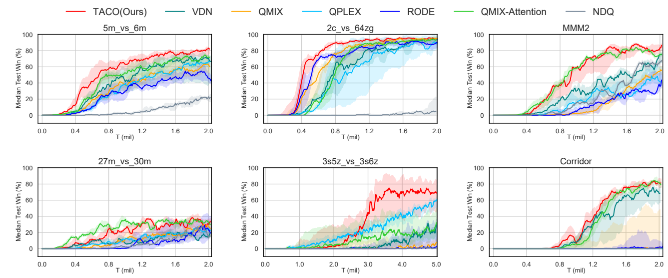

Figure 1: Performance comparison with baselines in different SMAC scenarios.

Both two parts are optimized simultaneously during training. Thus, the total loss function can be written as:

$$\mathcal{L}_{tot} = \mathcal{L}_{\mathrm{TD}} + \beta \mathcal{L}_{\mathrm{Rec}}, \tag{5}$$

where the β is a weighting term. For different complex scenarios, we can set different β to change the proportion of reconstruct loss in the gradient update. It should be noted that the abstract module is updated by both the TD loss and the reconstruct loss gradients in Eq. (5).

3 EXPERIMENTS

We applied our method and baselines to the StarCraft II Multi-Agent Challenge benchmark, which includes a series of scenarios representing different levels of challenge.

As shown in Figure 1, there is not much difference between QMIX and QMIX-Attention in some relatively simple scenarios (*5m_vs_6m* and *2c_vs_64zg*). However, the NDQ method performs poorly and has low learning efficiency, possibly due to its constraints on message passing and message instability. The performance of TACO is similar to that of QMIX-Attention and even exceeds QMIX-Attention in *5m_vs_6m*. In the super hard scenarios, the classic CTDE method performs poorly due to a lack of effective communication, whereas QMIX-Attention performs well. However, QMIX-Attention's success is mainly due to its lack of communication restrictions. The TACO method can achieve or even exceed the performance of QMIX-Attention without utilizing actual communication during the end of the training, which significantly enhances its practicality.

4 CONCLUSION

In this paper, we propose a simple and effective multi-agent collaboration training paradigm called TACO. This approach allows agents gradually replace explicit communication with reconstructed information, ultimately achieving efficient cooperation under fully decentralized execution. Experimental results show that the TACO method can achieve close or even better performance than the same baseline using communication or global information without sharing information.

5 ACKNOWLEDGMENTS AND DISCLOSURE OF FUNDING

This project was supported by Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDA27050102.

REFERENCES

[1] Itamar Arel, Cong Liu, Tom Urbanik, and Airton G Kohls. 2010. Reinforcement learning-based multi-agent system for network traffic signal control. *IET Intelligent Transport Systems* 4, 2 (2010), 128–135.

[2] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Christopher Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. *CoRR* abs/1912.06680 (2019). arXiv:1912.06680

[3] Coline Devin, Abhishek Gupta, Trevor Darrell, Pieter Abbeel, and Sergey Levine. 2017. Learning modular neural network policies for multi-task and multi-robot transfer. In *2017 IEEE International Conference on Robotics and Automation, ICRA 2017, Singapore, Singapore, May 29 - June 3, 2017*. IEEE, 2169–2176. https://doi.org/10.1109/ICRA.2017.7989250

[4] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual Multi-Agent Policy Gradients. In *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*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 2974–2982.

[5] Carlos Guestrin, Michail G. Lagoudakis, and Ronald Parr. 2002. Coordinated Reinforcement Learning. In *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8-12, 2002*, Claude Sammut and Achim G. Hoffmann (Eds.). Morgan Kaufmann, 227–234.

[6] Jelle R. Kok and Nikos Vlassis. 2006. Collaborative Multiagent Reinforcement Learning by Payoff Propagation. *J. Mach. Learn. Res.* 7 (2006), 1789–1828. http://jmlr.org/papers/v7/kok06a.html

[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 *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 6379–6390.

[8] Laëtitia Matignon, Laurent Jeanpierre, and Abdel-Illah Mouaddib. 2012. Coordinated Multi-Robot Exploration Under Communication Constraints Using Decentralized Markov Decision Processes. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada*, Jörg Hoffmann and Bart Selman (Eds.). AAAI Press. http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5038

[9] 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 *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 4292–4301.

[10] Mikayel Samvelyan, Tabish Rashid, Christian Schröder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob N. Foerster, and Shimon Whiteson. 2019. The StarCraft Multi-Agent Challenge. *CoRR* abs/1902.04043 (2019). arXiv:1902.04043 http://arxiv.org/abs/1902.04043

[11] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. 2016. Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving. *CoRR* abs/1610.03295 (2016). arXiv:1610.03295

[12] 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. 2017. Value-Decomposition Networks For Cooperative Multi-Agent Learning. *CoRR* abs/1706.05296 (2017). arXiv:1706.05296

[13] Ming Tan. 1993. Multi-Agent Reinforcement Learning: Independent versus Cooperative Agents. In *Machine Learning, Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, MA, USA, June 27-29, 1993*, Paul E. Utgoff (Ed.). Morgan Kaufmann, 330–337. https://doi.org/10.1016/b978-1-55860-307-3.50049-6

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[15] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Çaglar Gülçehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nat.* 575, 7782 (2019), 350–354. https://doi.org/10.1038/s41586-019-1724-z

[16] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2021. QPLEX: Duplex Dueling Multi-Agent Q-Learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=Rcmk0xxIQV

[17] Rose E Wang, Michael Everett, and Jonathan P How. 2020. R-MADDPG for partially observable environments and limited communication. *arXiv preprint arXiv:2002.06684* (2020).

[18] Cathy Wu, Aboudy Kreidieh, Eugene Vinitsky, and Alexandre M. Bayen. 2017. Emergent Behaviors in Mixed-Autonomy Traffic. In *1st Annual Conference on Robot Learning, CoRL 2017, Mountain View, California, USA, November 13-15, 2017, Proceedings (Proceedings of Machine Learning Research, Vol. 78)*. PMLR, 398–407. http://proceedings.mlr.press/v78/wu17a.html

[19] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre M. Bayen, and Yi Wu. 2021. The Surprising Effectiveness of MAPPO in Cooperative, Multi-Agent Games. *CoRR* abs/2103.01955 (2021). arXiv:2103.01955