Learning Structured Communication for Multi-Agent Reinforcement Learning

JAAMAS Track

Junjie Sheng
East China Normal University
Shanghai, China
jarvis@stu.ecnu.edu.cn

Xiangfeng Wang[†]
East China Normal University
Shanghai, China
xfwang@cs.ecnu.edu.cn

Bo Jin
Tongji University
Shanghai, China
bjin@tongji.edu.cn

Wenhao Li
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
liwenhao@cuhk.edu.cn

Jun Wang
East China Normal University
Shanghai, China
jwang@cs.ecnu.edu.cn

Junchi Yan
Shanghai Jiao Tong University
Shanghai, China
yanjunchi@sjtu.edu.cn

Tsung-Hui Chang
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
tsunghui.chang@ieee.org

Hongyuan Zha
The Chinese University of Hong Kong, Shenzhen& Shenzhen Institute of AI and Robotics for Society
Shenzhen, China
zhahy@cuhk.edu.cn

ABSTRACT

This paper investigates multi-agent reinforcement learning (MARL) communication mechanisms in large-scale scenarios. We propose a novel framework, Learning Structured Communication (LSC), that leverages a flexible and efficient communication topology. LSC enables adaptive agent grouping to create diverse hierarchical formations over episodes generated through an auxiliary task and a hierarchical routing protocol. We learn a hierarchical graph neural network with the formed topology that facilitates effective message generation and propagation between inter- and intra-group communications. Unlike state-of-the-art communication mechanisms, LSC possesses a detailed and learnable design for hierarchical communication. Numerical experiments on challenging tasks demonstrate that the proposed LSC exhibits high communication efficiency and global cooperation capability.

KEYWORDS

Learning to Communicate; Multi-Agent Reinforcement Learning; Hierarchical Structure; Graph Neural Networks

ACM Reference Format:
Junjie Sheng, Xiangfeng Wang, Bo Jin, Wenhao Li, Jun Wang, Junchi Yan, Tsung-Hui Chang, Hongyuan Zha. 2023. Learning Structured Communication for Multi-Agent Reinforcement Learning: JAAMAS Track. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

The remarkable benefits of cooperation for multi-agent reinforcement learning achieved through learning to communicate are widely recognized. Despite the proliferation of numerous approaches [1–3, 5], the effectiveness of the learned protocol is hindered as the number of agents increases, leading to inefficiencies in cooperation.

The adeptness of human society in managing communication among a vast number of participants is well-known. The fundamental principle that governs this process is the establishment of a hierarchical communication topology that enables intra- and inter-group communication [4]. Despite the prominence of this approach, the optimal design of a hierarchical communication structure that maximizes communication efficiency while also fostering large-scale cooperation remains largely unexplored.

The Learning Structured Communication (LSC) framework is introduced, which aims to facilitate large-scale cooperation through the learning of a hierarchical communication structure. LSC consists of two primary stages: structure building and communication-based policy learning. The former leverages a distributed cluster-based routing protocol (CBRP [4]) and a learnable weight generator to establish a hierarchical structure, dividing agents into groups (high-level and associated low-level agents) based on their weights. Following the establishment of the hierarchical structure, communication-based policy learning facilitates learning communication and action policies. The communication policy incorporates both inter and intra-group communication strategies. Inter-group communication aids in capturing global information, while intra-group communication facilitates fine-grained message exchanges. The action policy then uses the improved state perception resulting from communication to learn a more effective cooperation strategy.

[†] Corresponding author (xfwang@cs.ecnu.edu.cn).

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

To assess the performance of the proposed LSC, we conducted experiments on large-scale *Battle* scenarios. The empirical findings indicate that LSC consistently outperforms the baselines in terms of both cooperation performance and communication efficiency.

2 METHOD

This section outlines the LSC approach, which is composed of two fundamental stages: structure building and communication-based policy learning.

2.1 Structure Building

The structure building consists of two integral components, the weight generator and the cluster-based routing protocol, CBRP [4]. Each agent generates a communication weight through its weight generator. The weight generator determines the communication importance for each agent and is modeled by a neural network $f_{wg} : o_i \rightarrow w_i$ with parameters θ^w. The CBRP leverages the weights of all agents $\mathbf{w}$ and considers the local geometry to construct the hierarchical communication structure in a distributed fashion.

Concretely, the weight w_i measures an agent's confidence in becoming a high-level agent (HLA). During CBRP execution, each agent checks whether HLAs exist in its receptive field (RF) for multiple rounds and elects itself as a HLA if no HLAs are found. If a HLA detects the presence of other HLAs in its RF, it may opt to downgrade to a low-level agent (LLA). Following a sufficient number of rounds, the CBRP generates a sparse structure where HLAs are not included in the RFs of other HLAs, thereby improving communication efficiency. The hierarchical communication network is established by connecting HLAs across groups and linking each LLA to its respective HLA. This process forms a group consisting of each HLA and the LLAs within its respective RF.

2.2 Communication-Based Policy Learning

Following the establishment of the hierarchical communication structure, communication-based policy learning enables the acquisition of messages and cooperation policies through two sub-modules, namely the hierarchical communicator and the Q-Net-based policy. The hierarchical communicator is implemented as a graph neural network (GNN) ($f_{\theta^{gnn}}$) with parameters θ^{gnn}, while the Q-Net of each agent ($Q^i_{\theta^Q}$) is parameterized by the shared parameter θ^Q. The former is responsible for learning the messages and enhancing overall state perception through message passing. After efficient communication, the Q-Net-based policy then learns an updated policy based on the enhanced state perception.

The hierarchical communicator employs a three-phase communication strategy: intra-group aggregation, inter-group sharing, and intra-group sharing. During the intra-group aggregation, each LLA embeds its local perception (v^l_i) into a message, which is subsequently transmitted to the connected HLAs. The HLAs aggregate the information from all connected LLAs and obtain the group perception (v^h_i). In the inter-group sharing, the HLAs communicate with each other, leveraging the group perception and aggregating the received information to obtain the global perception (v^g_i). During the intra-group sharing, each HLA then embeds its local, group, and global perception as the message and sends it to the connected LLAs. The LLAs update their local perceptions (v^l_i) based on the received information. Subsequent to the three-phase communication, each agent utilizes the Q-Net-based policy to obtain Q values based on its local perception and select an optimal cooperative action.

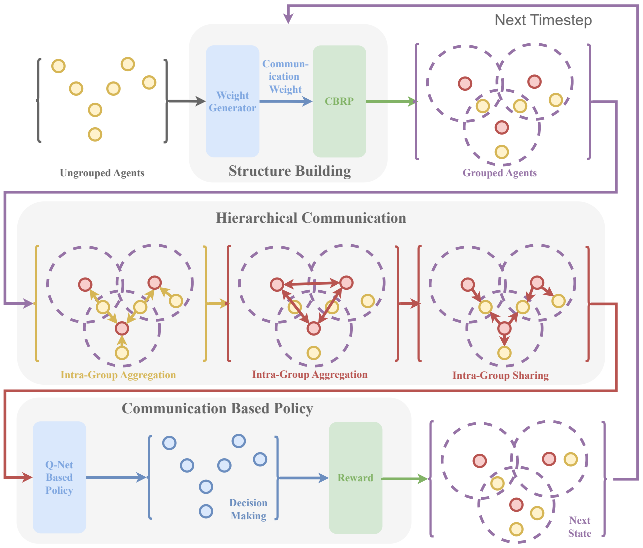

Figure 1: Illustration of the forward pass of LSC. The yellow and red particles denote the high level agents and the low level agents respectively. For each decision step, agents make structure building, hierarchical communication and communication based decisions.

2.3 Training Scheme

This section explains how we train the proposed LSC. The communication-based policy can be learned directly by minimizing the loss function:

$$\ell(\theta^{gnn}) = \mathbb{E}_{\mathbf{o},\mathbf{a},\mathbf{r},\tilde{\mathbf{o}}}\left[\sum_{i=1}^{n}\left(Q^i_{\theta^Q}(f_{\theta^{gnn}}(\mathbf{o}), a_i) - y_i\right)^2\right], \quad (1)$$

where $y_i = r_i + \gamma \max_{\tilde{a}_i} Q^i_{\theta^Q}(f_{\theta^{gnn}}(\tilde{\mathbf{o}}), \tilde{a}_i)$, and r_i is the reward for agent i. We use soft updating schemes with target networks:

$$\theta^{\tilde{Q}} = \tau\theta^{Q} + (1-\tau)\theta^{\tilde{Q}}, \; \theta^{\tilde{gnn}} = \tau\theta^{gnn} + (1-\tau)\theta^{\tilde{gnn}}. \quad (2)$$

The non-differentiability of the CBRP impedes the backpropagation of gradients from the communication-based policy to the weight generator, presenting a significant challenge. We propose an auxiliary reinforcement learning task for weight generation to address this. Each agent's action corresponds to a weight choice in this task, with original observations and rewards and the weight w_i is defined in the discrete action space $\{0, 1, 2\}$. The proposed approach enables a task-driven, closed-loop communication weight generation. For simplicity, we adopt independent deep Q-networks to implement the weight generator. The loss function for the weight generator is defined as follows:

$$\ell(\theta^{w}) = \mathbb{E}_{\mathbf{o},\mathbf{w},\mathbf{r},\tilde{\mathbf{o}}}\left[\sum_{i=1}^{n}(Q_{\theta^w}(o_i, w_i) - y_i)^2\right]. \quad (3)$$

where $y_i = r_i + \gamma \max_{\tilde{w}_i} Q_{\theta^w}(\tilde{o}_i, \tilde{w}_i)$.

3 ACKONWLEDGEMENT

This work was supported in part by Postdoctoral Science Foundation of China (No. 2022M723039), NSFC (No. 12071145), STCSM (No. 22QB1402100), Shenzhen Science and Technology Program (No. JCYJ20210324120011032), a grant from China Academy of LVT, Shanghai Trusted Industry Internet Software Collaborative Innovation Center and a grant from Shenzhen Institute of Artificial Intelligence and Robotics for Society.

REFERENCES

[1] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Neural Information Processing Systems*. 2137–2145.

[2] Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. 2020. Graph Convolutional Reinforcement Learning. In *International Conference on Learning Representations*.

[3] Jiechuan Jiang and Zongqing Lu. 2018. Learning attentional communication for multi-agent cooperation. In *Neural Information Processing Systems*. 7254–7264.

[4] M Rezaee and M Yaghmaee. 2009. Cluster based routing protocol for mobile ad hoc networks. *IEEE International Conference on Computer Communications*, 30–36.

[5] Sainbayar Sukhbaatar, Rob Fergus, et al. 2016. Learning multiagent communication with backpropagation. In *Neural Information Processing Systems*. 2244–2252.