

# MAPDP: Cooperative Multi-Agent Reinforcement Learning to Solve Pickup and Delivery Problems

Wangxianyi

LZU

2024 年 5 月 10 日

- ① Introduction
- ② Problem Formulation
- ③ Methodology
- ④ MAPDP
- ⑤ Experiments
- ⑥ Conclusion
- ⑦ References

# 1 Introduction

## 2 Problem Formulation

## 3 Methodology

## 4 MAPDP

## 5 Experiments

## 6 Conclusion

## 7 References

# Background Introduction

- Vehicle Routing Problem (VRP) is crucial in various real-world applications such as express systems, industrial warehousing, and on-demand delivery.
- Cooperative Pickup and Delivery Problem (PDP) is a variant of VRP that plays a significant role in applications like on-demand delivery and industrial logistics.
- Challenges in solving cooperative PDP include structural dependency between pickup and delivery pairs and the need for effective cooperation among different vehicles.
- Existing solutions face difficulties in explicit modeling of dependencies and cooperation, leading to suboptimal performance.

# Research Objectives

- Explore the cooperative Pickup and Delivery Problem (PDP) with multiple vehicle agents using Multi-Agent Reinforcement Learning (MARL).
- Design a centralized MARL framework to generate cooperative decisions by capturing the inter-dependency of heterogeneous nodes.
- Train different agents based on communication embedding using a specially designed cooperative Advantage Actor-Critic (A2C) algorithm.
- Evaluate the effectiveness of the MAPDP framework on different datasets and compare its performance with existing baselines.

# Overview of MAPDP Framework

- MAPDP is a novel cooperative Multi-Agent Reinforcement Learning (MARL) framework designed to solve the Cooperative Pickup and Delivery Problem (PDP).
- The framework utilizes multi-agent cooperation to generate high-quality solutions by sharing a common context encoder and individual decoders for each vehicle agent.
- MAPDP learns to generate the next node to visit for each vehicle agent step by step and outputs a complete routing plan.
- Key components of MAPDP include paired context embedding to represent node dependencies, cooperative decoders for decision dependence, and a cooperative A2C algorithm for model training.

① Introduction

② Problem Formulation

③ Methodology

④ MAPDP

⑤ Experiments

⑥ Conclusion

⑦ References

# Mathematical Modeling of Cooperative PDP

$$\min \sum_{k=1}^K \sum_{i=0}^{2N} \sum_{j=1}^{2N+1} e_{ij} x_{ijk} \quad (1)$$

- $x_{ijk} \in \{0, 1\}$ : whether the vehicle  $k$  travels directly from node  $v_i$  to node  $v_j$ .
- $e_{ij}$ : spatial distances.

$$\sum_{k=1}^K \sum_{j=1}^{2N+1} x_{ijk} = 1, \forall i \in [0, 2N] \quad (2)$$

$$\sum_{k=1}^K \sum_{i=0}^{2N} x_{ijk} = 1, \forall j \in [1, 2N+1] \quad (3)$$



# Mathematical Modeling of Cooperative PDP

$$\sum_{i \in S'} d_i \leq C_k, \forall S' \subseteq S, \forall k \in [1, K] \quad (4)$$

- $d_i$ : Each pickup order has a demand volume.
- $C_k$ : Capacity of the k-th vehicle.
- $S$ : A consecutive routing sequence from  $v_0$  and ends at  $v_{2N+1}$ .

$$\sum_{j=1}^{2N+1} x_{i,jk} = \sum_{j=0}^{2N+1} x_{i+N,jk}, \forall k \in [1, K], i \in [1, N] \quad (5)$$

$$T_i \leq T_{i+N}, \forall i \in [1, N] \quad (6)$$

1 Introduction

2 Problem Formulation

3 Methodology

4 MAPDP

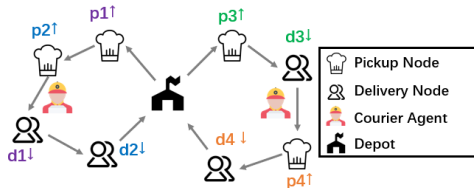
5 Experiments

6 Conclusion

7 References

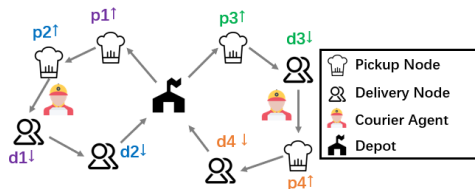
# Explanation of State, Action

- State: At step  $t$ , agent  $k$ 's state includes its remaining capacity  $C_k^t$  and current trajectory  $S_k^t$ . The current location, the last visited node, is denoted by  $v_{l_k^t}$ , where  $l_k^t$  is the node index. All vehicles can communicate centrally, ensuring full observability in the cooperative PDP setting.
- Action: The action at step  $t$  for vehicle agent  $k$  is to determine a



# Explanation of Transition, Reward

- Transition: The transition between adjacent states is to replace every agent to its target node as its current action. Then we update both the trajectory and the remaining capacity of each agent:  $S_k^{t+1} = (S_k^t; \{v_{l_k}^t\})$ ,  $C_k^{t+1} = C_k^t - d_{l_k}^t$ , where ; means concatenating the partial solution with the new selected node.
- Reward: All agents aim to minimize the total travel distance during the episode. At each step, the reward  $r_k^t$  is the negative length of the newly traveled arc, and the final episode reward  $R$  is the sum of all individual rewards  $r_k^t$ , where  $T$  is the total number of decision steps in an episode, and  $I_k^0 = 0$  denotes that all vehicles start from depot  $v_0$ .



- ① Introduction
- ② Problem Formulation
- ③ Methodology
- ④ MAPDP**
- ⑤ Experiments
- ⑥ Conclusion
- ⑦ References

# Overview of MAPDP Framework

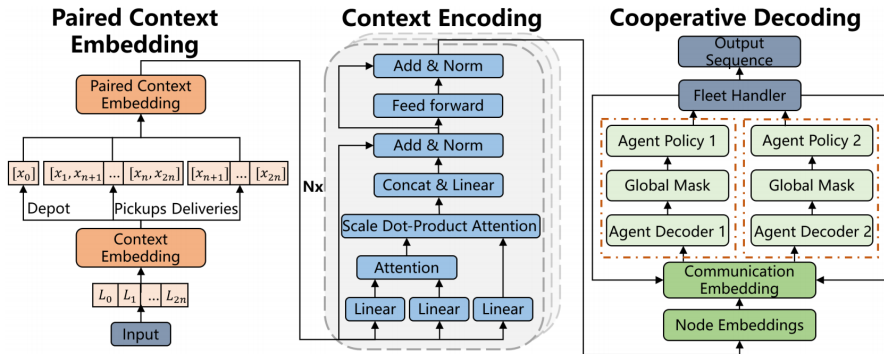


图 1: MAPDP Framework

# Paired Context Embedding

Original 2-D location information :  $\mathcal{L}_i$

Concatenat the two features and map them into one dense vector :

$$x_i = W^x[\mathcal{L}_i, d_i] + b^x.$$

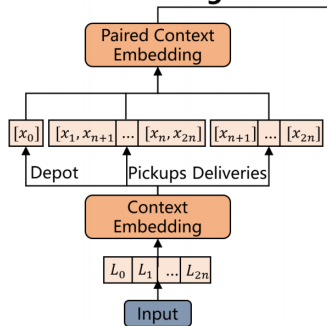
# Paired Context Embedding

$$h_i^0 = \begin{cases} W_0^x x_i + b_0^x, & i = 0, \\ W_p^x [x_i; x_{i+N}] + b_p^x, & 1 \leq i \leq N, \\ W_d^x x_i + b_d^x, & N+1 \leq i \leq 2N, \end{cases} \quad (7)$$

$$\hat{h}_i = BN^\ell(h_i^{\ell-1} + MHA_i^\ell(h_1^{\ell-1}, h_2^{\ell-1}, \dots, h_{2N}^{\ell-1})), \quad (8)$$

$$h_i^\ell = BN^\ell(\hat{h}_i + FF^\ell(\hat{h}_i)). \quad (9)$$

## Paired Context Embedding



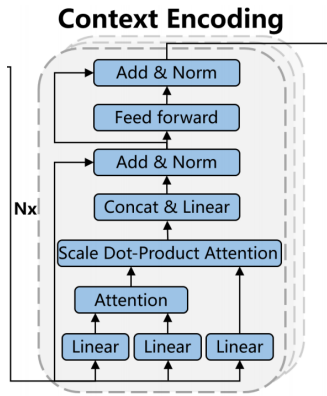


# Context Encoding

$$Q_i^h, K_i^h, V_i^h = W_Q^h h_i, W_K^h h_i, W_V^h h_i, \quad (10)$$

$$A_i^h = \text{softmax}(Q_i^h K^h{}^T / \sqrt{d_k}) V_j^h, \quad (11)$$

$$MHA_i = \text{Concat}(A_i^1, A_i^2, \dots, A_i^H) W_O, \quad (12)$$



# Cooperative Multi-Agent Decoders

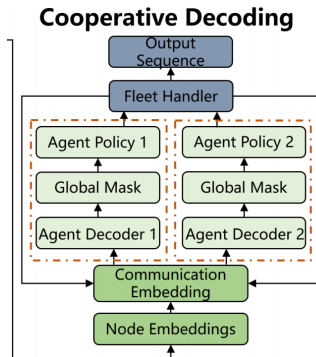
$$Comm^t = [h_{l_1^t}; C_1^t; h_{l_2^t}; C_2^t; \dots; h_{l_K^t}; C_K^t] \quad (13)$$

$$g_k^t = MHA_{k,(c)}(h_1, h_2, \dots, h_{2N}), \quad (14)$$

$$Q_k^t, K_{k,i}^t = W_{Q,k} g_k^t, W_{K,k} h_i, \quad (15)$$

$$u_{k,i}^t = Dtanh(Q_k^{tT} K_{k,i}^t / \sqrt{d_k}), \quad (16)$$

$$p_{\theta_k, \phi}(v(k, t)) = softmax(Mask^t(u_{k,i}^t)), \quad (17)$$



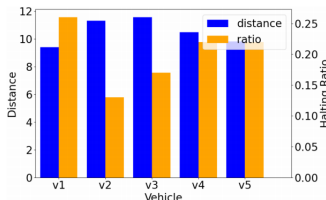
- ① Introduction
- ② Problem Formulation
- ③ Methodology
- ④ MAPDP
- ⑤ Experiments**
- ⑥ Conclusion
- ⑦ References

# Evaluation Results on Different Datasets

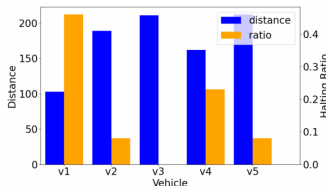
Model	Random Dataset								
	2N = 20, K=2			2N = 50, K=5			2N = 100, K=10		
	Cost	Gap	Time	Cost	Gap	Time	Cost	Gap	Time
ACO (Gambardella, Taillard, and Agazzi 1999)	34.73	39.60%	6min	79.94	52.01%	32min	136.89	53.86%	51min
Tabu Search (Glover 1990)	29.76	19.67%	7min	64.57	22.78%	34min	112.38	26.31%	51min
OR-Tools (Google 2021)	25.91	4.18%	4min	54.64	3.90%	31min	94.25	5.93%	49min
RL-VRP (Nazari et al. 2018)	26.79	7.72%	1s	63.12	20.02%	5s	101.13	13.67%	9s
AM-VRP (Kool, van Hoof, and Welling 2019)	26.64	7.12%	1s	67.41	28.18%	4s	105.91	19.04%	8s
MDAM (Xin et al. 2021)	25.98	4.46%	8s	67.24	27.86%	25s	105.11	18.14%	51s
MAPDP	<b>24.87</b>	<b>0.00%</b>	1s	<b>52.59</b>	<b>0.00%</b>	4s	<b>88.97</b>	<b>0.00%</b>	7s
Model	Real-World Dataset								
	2N = 20, K=2			2N = 50, K=5			2N = 100, K=10		
	Cost	Gap	Time	Cost	Gap	Time	Cost	Gap	Time
ACO (Gambardella, Taillard, and Agazzi 1999)	812	30.13%	6min	1205	35.39%	34min	2054	20.47%	53min
Tabu Search (Glover 1990)	834	33.65%	6min	1197	34.49%	34min	2033	19.24%	51min
OR-Tools (Google 2021)	749	20.03%	4min	1056	18.65%	31min	1811	6.22%	50min
RL-VRP (Nazari et al. 2018)	714	14.42%	1s	1130	26.97%	5s	1842	8.04%	9s
AM-VRP (Kool, van Hoof, and Welling 2019)	661	5.93%	1s	942	5.84%	4s	1759	3.17%	9s
MDAM (Xin et al. 2021)	638	2.24%	8s	941	5.73%	25s	1733	1.64%	52s
MAPDP	<b>624</b>	<b>0.00%</b>	1s	<b>890</b>	<b>0.00%</b>	4s	<b>1705</b>	<b>0.00%</b>	7s

图 2: Comparison of Different Models on Random and Real-World Datasets

# Performance Comparison with Other Methods



(a) Random Dataset.



(b) Real-World Dataset.

图 3: Case studies on vehicle cooperation analysis from two datasets.

- ① Introduction
- ② Problem Formulation
- ③ Methodology
- ④ MAPDP
- ⑤ Experiments
- ⑥ Conclusion**
- ⑦ References

# Conclusion

- The proposed MAPDP framework leverages Multi-Agent Reinforcement Learning (MARL) to effectively solve the Cooperative Pickup and Delivery Problem (PDP) by capturing dependencies and promoting cooperation among multiple vehicles.
- MAPDP outperforms existing baselines by at least 1.64
- The centralized MARL framework, paired context embedding, cooperative decoders, and cooperative A2C algorithm collectively contribute to the success of MAPDP in addressing the challenges of PDP.
- Future research directions may include exploring scalability of MAPDP to larger problem instances, incorporating real-time constraints, and adapting the framework to dynamic environments.

- ① Introduction
- ② Problem Formulation
- ③ Methodology
- ④ MAPDP
- ⑤ Experiments
- ⑥ Conclusion
- ⑦ References**



- [1] B. McMahan, E. Moore, D. Ramage, et al. Communication-efficient learning of deep networks from decentralized data[C]. Artificial intelligence and statistics, 2017, 1273-1282
- [2] Y. LeCun, L. Bottou, Y. Bengio, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324
- [3] Y. Kim, Y. Jernite, D. Sontag, et al. Character-aware neural language models[C]. Proceedings of the AAAI conference on artificial intelligence, 2016, 2741-2749

*Thanks!*