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

Zefang Zong¹², Meng Zheng³, Yong Li¹², *, Depeng Jin¹²

¹Beijing National Research Center for Information Science and Technology
²Department of Electronic Engineering, Tsinghua University, Beijing, China
³Hitachi (China) Research & Development Corporation

Abstract

Cooperative Pickup and Delivery Problem (PDP), as a variant of the typical Vehicle Routing Problems (VRP), is an important formulation in many real-world applications, such as on-demand delivery, industrial warehousing, etc. It is of great importance to efficiently provide high-quality solutions of cooperative PDP. However, it is not trivial to provide effective solutions directly due to two major challenges: 1) the structural dependence between pickup and delivery pairs require explicit modeling and representation. 2) the cooperation between different vehicles is highly related to solution exploration and is difficult to model. In this paper, we propose a novel multi-agent reinforcement learning-based framework to solve the cooperative PDP (MAPDP). First, we design a paired context embedding to well measure the dependency of different nodes considering their structural limits. Second, we utilize cooperative multi-agent decoders to leverage the decision dependence among different vehicle agents based on a special communication embedding. Third, we design a novel cooperative A2C algorithm to train the integrated model. We conduct extensive experiments on a randomly generated dataset and a real-world dataset. Experiments result shown that the proposed MAPDP outperforms all other baselines by at least 1.64% in all settings, and shows significant computation speed during solution inference.

Introduction

Vehicle Routing Problem (VRP) has shown its importance in formulating many real-world applications, including express systems, industrial warehousing, and on-demand delivery (Zong et al. 2021). A fleet of vehicles/couriers are managed to fulfill given demands, while the goal is to optimize the routing plan to reduce traveling expenses. In real-world scenarios, a given demand order usually has its own origination and a designated delivery destination. Furthermore, routing tasks are usually assigned to multiple vehicles simultaneously, which forms the Pickup and Delivery Problem (PDP) (Savelsbergh and Sol 1995). For instance, on-demand delivery couriers need to pick up the food first from the restaurants and then deliver it to the corresponding customers, as shown in 1. While in industrial manufacturing, large amounts of materials, productions also need to be transported from specific factories and warehouses at multiple sites (Li et al. 2021b). Efficiently generating high-quality solutions of cooperative PDP can help reduce operating expenses, improve public efficiency and thus bring significant benefits.

Owing to the NP-hard nature, the cooperative PDP along with other VRP variants is still difficult to be optimally solved by exact methods (Toth and Vigo 2002; Madsen, Fisher, and Jornsten 1997). Even though numerous heuristic-based methods are developed to compute near-optimal solutions, the solution generation process remains time-consuming and there is still potential to find more approximate ones. The recent development of deep reinforcement learning (DRL) offers its effectiveness in solving many combinatorial optimization problems including VRPs, and thus brings another perspective to solve PDP (Bello et al. 2016; Kool, van Hoof, and Welling 2019; Nazari et al. 2018; Chen and Tian 2019; Lu, Zhang, and Yang 2019a). Benefiting from learning a parameterized model instead of relying on manually constructed rules to search for solutions, DRL has shown its appealing performance in typical routing problems. Besides, by splitting the phases of training and online inferring, DRL can generate results with much faster computation. Inspired by both high solution quality and inference speed, it is prospective to well solve cooperative PDP by constructing a DRL framework.

However, most existing DRL based methods can only be applied to typical VRPs where the demands share a common final delivery site, i.e., either to be picked up or to be delivered only. They are less effective when facing cooperative PDP with more detailed constraints and challenges from the following two aspects. First, it is not trivial to measure and structural constraint of PDP among different nodes. Compared to typical VRPs, complicated relations among all nodes should be considered. Representation upon the relation between paired pickup and deliveries should differ from that between unpaired ones. How to efficiently measure such relations is essential for framework effectiveness. Second, it is difficult to coordinate the cooperation within the vehicle fleet. As all vehicles working simultaneously to accomplish the pickup and delivery tasks, the potential decision space of one is directly influenced by others. Optimizing the coordinated routing plans as the global objective is different from optimizing the one of a single vehicle, where the dependence

*Corresponding Author
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
To tackle the challenges above, we propose a novel end-to-end cooperative multi-agent RL (MARL) based framework to solve cooperative PDP (MAPDP). MAPDP learns to generate the next node to visit step by step for each vehicle agent and finally outputs a complete routing plan. First, we develop a paired context embedding to represent each pickup or delivery node in each order instance. The paired context embedding well models the inter-relation in-between that will further influence the potential solution space. Second, we utilize cooperated decoders to leverage the decision dependence among different vehicle agents. Given the shared instance encoding results from former module designs, each vehicle agent obtains its own partial action results based on a special communication embedding. Such multi-agent cooperation is trained and could be executed in online inference via a centralized controlling framework. Third, We also design a cooperative A2C algorithm for the integrated model training, where a joint critic value net estimates the state value.

To summarize, the main contributions of our work are listed as follows:

- To the best of our knowledge, we are the first to explore the cooperative PDP with multiple vehicle agents using MARL.
- We design a centralized MARL framework to generate cooperative decisions. We design a paired context embedding to well capture the inter-dependency of heterogeneous nodes, and train different agents based on a communication embedding via a specially designed cooperative A2C algorithm.
- We evaluate the effectiveness of MAPDP on two datasets. It outperforms other baselines by at least 1.64% in all experiment settings and shows significant computation speed during solution inference.

The remainder of this paper is organized as follows. We first introduce the related works in Section 2. We then formulate cooperative PDP in Section 3, and introduce our methodology in Section 4. The experiment results are presented in Section 5. Finally, we conclude our paper in Section 6.

Related Works

In this section, we review the existing literature on conventional heuristic methods for solving PDP and learning based methods for routing problems. Besides, we also introduce the recent research in multi-agent RL.

Heuristics for PDP

As an important variant of the typical vehicle routing problem (VRP), PDP considering pickups and deliveries was first studied in (Savelsbergh and Sol 1995), and exact methods were put forward early as solutions (Ruland and Rodin 1997). However, trying to generate exact optimality suffers from heavy computation given then exponential complexity. As a substitution, more researchers tend to utilize heuristics to generate approximate optimal solutions, which could significantly increase the efficiency with small quality costs. A tabu-embedded simulated annealing algorithm was proposed for solving large-scale PDP with time windows (Ropke and Cordeau 2009). An adaptive large neighborhood search heuristic method was presented in (Ropke and Pisinger 2006) to solve PDP, incorporating regret insertion method and six removal strategies.

Even though the heuristic-based methods could generate near-optimal solutions within a more reasonable time instead, they heavily rely on hand-crafted rules and are greatly limited by human experience. Furthermore, the online inference time of these methods is still not satisfactory when facing high dynamics.

RL based Methods for Routing Problems

Due to the potential that taking feedback reward as training signals to action attempts when interacting with outside environments, RL has shown its effectiveness in solving many decision-making problems. Bello et al. first propose an RL-based algorithm to solve combinatorial optimization problems, including the famous Traveling Salesman Problem (TSP) (Bello et al. 2016). Such an algorithm borrows the idea from the constructive heuristics and utilizes an agent to directly generate solutions from scratch in an end-to-end manner. Following this idea, Nazari et al. utilize RNN structures to further expand its capability on solving capacitated VRP by generating decision sequences (Nazari et al. 2018). Kool et al. (Kool, van Hoof, and Welling 2019) further propose a attention based network to fully capture the in-between relationships between different nodes. The suitable neural network design greatly improves the solution quality, and more routing variants are further evaluated. Xin et al. develop a multi-decoder based framework to generate fine-tuned solutions via a special beam-search.

Despite the idea of directly generate final solution from scratch as the output results for VRP, researchers also attempt to augment RL into local solution improvement. Chen et al. (Chen and Tian 2019) formulates the solution improvement process as keeping rewriting based on the current ones. Following this idea, Lu et al. (Lu, Zhang, and Yang 2019b) further combined RL with Operation Research operators to keep updating the current solutions based on multiple in-hand operators.

However, most existing RL based approaches can only solve typical VRPs, while the cooperative PDP problem with structural dependency and vehicle cooperation requires specific modeling. Even though Li et al. (Li et al. 2021b) pro-
posed a framework to solve dynamic PDP, the agent model is only used to dispatch new orders to one of the vehicles, while the real routing process is processed via enumeration. Such a pure dispatching framework suffers from limits on exploration to global optimization. Li et al. (Li et al. 2021a) proposed a heterogeneous attention based network to solve single-vehicle PDP, but finding solutions for cooperative PDP remains unsolved. In contrast, our proposed MAPDP could well model both pickup-delivery dependency and vehicle cooperations.

Multi-Agent Reinforcement Learning

In many practical decision-making problems, cooperation or competition among different agents requires specific modeling, thus the multi-agent reinforcement learning (Busoniu, Babuška, and De Schutter 2010; Lowe et al. 2017) has attracted much attention. Tampuu et al. (Tampuu et al. 2017) analyzed the cooperation and competition among two agents in reinforcement learning by carefully designing the reward schemes. Lin et al. (Lin et al. 2018) designed a multi-agent model to manage multiple vehicles operating simultaneously in the city. Lee et al. (Lee and Lee 2019) further improved cooperative models by mixing demonstrations from the centralized policy.

As for the variants of VRPs, Zhang et al. (Zhang et al. 2020) proposed a MARL-based framework to solve VRP with soft time windows for a fleet of vehicles. However, their multi-agent modeling relies on manually crafted rules in which vehicles take turns to make decisions. Furthermore, all vehicles share the same policy network, which makes the framework a single agent control indeed. In contrast, we develop independent policy networks without any manual rules in vehicle coordination and thus could enlarge the exploration space of all vehicle agents.

Problem Formulation

In this section, we provide the mathematical formulation of cooperative PDP.

With $N$ delivery nodes and $N$ pickup nodes are included in a given problem instance, the node set including the initial depot $v_0$ where all vehicle starts can be represented as $V = \{v_0, v_1, ..., v_N, v_{N+1}, v_{N+2}, ..., v_{2N}, v_{2N+1}\}$, and $v_{2N+1}$ is the copy of the depot. For simplification, we assume that $v_i(1 \leq i \leq N)$ has parcels to be picked-up and $v_i(N < i \leq 2N)$ is the target to be delivered to. $v_i$ and $v_{2N+1}$ form a corresponding pair. The spatial distances between different nodes can be further represented as $E = \{e_{ij}\}$, where $0 \leq i < 2N, 0 \leq j \leq 2N$. Each pickup order has a demand volume represented as $d_i$, where $d_0 = d_{2N+1} = 0$. A demand is noted as a positive value for pickups and negative for deliveries, where $d_i > 0$ if $1 \leq i \leq N, d_i < 0$ if $N + 1 \leq i \leq 2N$ and $d_i = -d_{i+N}$. All PDP tasks are assigned to $K$ vehicles with $C_k$ denoting the individual capacity of the $k$-th vehicle.

Let $x_{ijk} \in \{0, 1\}$ denote whether the vehicle $k$ travels directly from node $v_i$ to node $v_j$, and $T_i$ as the arrival time at node $v_i$. $S \subseteq V$ denotes a consecutive routing sequence from $v_0$ and ends at $v_{2N+1}$ and does not include $v_0$ in the middle, the cooperative PDP can be further formally formulated as follows:

$$\begin{align*}
\min & \sum_{k=1}^{K} \sum_{i=0}^{2N} \sum_{j=1}^{2N+1} e_{ij} x_{ijk}, \\
\text{s.t.} & \sum_{k=1}^{K} \sum_{j=1}^{2N+1} x_{ijk} = 1, \forall i \in [0, 2N], \\
& \sum_{k=1}^{K} \sum_{i=0}^{2N} x_{ijk} = 1, \forall j \in [1, 2N+1], \\
& \sum_{i \in S^k} d_i \leq C_k, \forall S^k \subseteq S, \forall k \in [1, K], \\
& \sum_{j=1}^{2N+1} x_{i,j,k} = \sum_{j=0}^{2N+1} x_{i+j+N,k}, \forall k \in [1, K], i \in [1, N], \\
& T_i \leq T_{i+N}, \forall i \in [1, N].
\end{align*}$$

The overall objective is to minimize the total traveling distance of all vehicles. Constraint (2) and (3) guarantee that each node is visited and only visited once. (4) guarantees that a vehicle never carries parcels out of its capacity limit. (5) satisfies the structural limit that each parcel should be delivered by the same vehicle as it was picked up, and (6) guarantees that a pickup is always the precondition of its own delivery.

Methodology

To solve the above formulated cooperative PDP, we take advantage of MARL to explicitly learn the effective cooperation among different vehicles. We develop an end-to-end framework, MAPDP, to generate partial solution sequence continuously by combining current routing actions of different agents together, as shown in Figure 2. Different agent networks share a common public context encoder to capture problem instance representations and learn their own policy via independent decoders. We further train the entire network by computing a common critic value as an approximation to the cooperation quality.

Multi-Agent Reinforcement Learning Setting

The construction for PDP solutions can be formulated as a sequence generation process. The sequence is completed gradually and can be modeled as a Markov Decision Process (MDP). We define the essential elements within as follows.

- **State**: The state of agent $k$ at step $t$ includes the remaining available capacity $C_k^t$ the current traveling trajectory $S_k^t$. Specifically, the current location, i.e., the last node visited by agent $k$ is represented as $v_{I_k^t}$, where $I_k^t$ is the node index. Note that $v_{I_k^t} = v_0$ and $C_k^0 = C_k$. In the cooperative PDP setting, we assume that all vehicles can communicate via a centralized control so that all states are fully observable.
• **Action:** The action at step \( t \) for vehicle agent \( k \) is to determine a node as its next target, represented as \( v(k, t) \).

• **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: \( S_k^{t+1} = (S_k^t; \{v(k, t)\}) \), \( C_k^{t+1} = C_k^t - d_{I_k^t} \), where \( ; \) means concatenating the partial solution with the new selected node.

• **Reward:** To optimize the overall routing solution quality, all agents share a common objective, which is to minimize the accumulated traveling distance of all agents in the entire episode. In each decision step, the one-step reward \( r_k^t = -e_{I_k^t} I_k^t+1 \) is the negative of the length of the newly established arc. The final episode reward \( R \) can be computed as \( R = \sum_{k=K}^{K} \sum_{t=0}^{T-1} r_k^t \), where \( T \) is the decision step amount in a complete episode and \( I_k^0 = 0 \) means that all vehicles start from the depot \( v_0 \).

**Paired Context Embedding and Context Encoding**

Due to the structural dependency in-between the nodes, it is critical to measure the inter-node relationship between one and another and provide effective representations for such a structural dependency in-between, we develop a paired context embedding to better characterize the attributes of different nodes.

Given the demand \( d_i \) of node \( v_i \) and its original 2-D location information \( L_i \), which could be used to calculate node-wise Euclidean distances directly, we formulate the original node embeddings by concatenating the two features and map them into one dense vector as \( x_i = W_x [L_i, d_i] + b_x \).

When an agent decides whether to adopt a pickup request and head for its pickup location \( v_i \), it means that the agent also has to visit the corresponding paired delivery \( v_{i+N} \) afterwards. Thus, a complete representation on \( v_i \) should consider the information of its paired delivery \( v_{i+N} \). Meanwhile, the agent is only allowed to visit \( v_{i+N} \) when \( v_i \) is already visited. A policy evaluation upon \( v_{i+N} \) does not rely on \( v_i \) any more. Motivated by this, we update the representation of pickup nodes by concatenating their paired deliveries, i.e. \( x_i = [x_i; x_{i+N}] \). We further compute the linear projection of such augmented representations to generate the final paired context embeddings:

\[
\begin{align*}
    h_i^0 &= \begin{cases} 
        W_0^x x_i + b_0^x, & i = 0, \\
        W_p^x [x_i; x_{i+N}] + b_p^x, & 1 \leq i \leq N, \\
        W_d^x x_i + b_d^x, & N + 1 \leq i \leq 2N,
    \end{cases} \\
    h_i^{\ell} &= BN^{\ell}(h_i^{\ell-1} + MHA^{\ell}(h_1^{\ell-1}, h_2^{\ell-1}, \ldots, h_{2N}^{\ell-1})), \\
    h_i^{\text{hat}} &= BN^{\text{hat}}(h_i^{\ell} + FF^{\text{hat}}(h_i^{\ell})).
\end{align*}
\]

The core of an attention layer above is the multi-head attention block and can be defined as follows:

\[
\begin{align*}
    Q^h_i, K^h_i, V^h_i &= W_Q^h h_i, W_K^h h_i, W_V^h h_i, \\
    A^h_i &= \text{softmax}(Q^h_i K^h_i^T / \sqrt{d_k}) V^h_i, \\
    MHA_i &= \text{Concat}(A_1^h, A_2^h, \ldots, A_H^h) W_O,
\end{align*}
\]

where \( h = 1, 2, \ldots, H \) and \( d_k = d_h / H \). \( H \) is to amount of attention heads, \( Q_i^h, K_i^h, V_i^h \) are the query, key and value vectors respectively. \( W_O \) is the projection matrix used to project the final MHA output. The final embedding of each
node $h_i = h^L_i$ can thus be obtained from the consecutive $L$ attention layers. Furthermore, we also generate a graph embedding $\overline{h} = \frac{1}{2N} \sum_{i=0}^{2N} h_i$ as the average of all nodes in the problem instance, which represents the global aggregated information.

It is worthy to note that even though different agents generate their individual policies via different decoders, the paired context embedding and context encoding upon all nodes are shared. This is because, in a fully cooperative PDP scenario, different vehicle agents can share their observations under centralized agent control. Thus the representations of nodes from the environment are consistent and can be learned jointly. In addition, learning a shared context encoder also accelerates the training stage significantly and helps reduce computational costs.

Cooperative Multi-Agent Decoders

Given both node-wise and global representations, each vehicle agent learns its own policy via its individual decoder network, as depicted in Figure 2. Each agent decoder selects the next node $v_t$ to visit at step $t$ based on the current observations of all agents. The agents work in a cooperative way, and each one of them only visits a part of the overall pickup-delivery pairs within each episode for a complete solution construction.

For better cooperation between different agents, we maintain a communication layer to record the updated states of different agents as follows:

$$Comm^t = [h_1; C_1^t; h_2; C_2^t; \ldots; h_K^t; C_K^t]$$ (13)

Leveraging the up-to-date communication embedding, each agent decoder utilizes an MHA-based structure to evaluate the probability of selecting each node at step $t$. We first generate a special context embedding $h_{k,(c)}^t = [\overline{h}; h_1; C_1^t; Comm^t]$ for agent $k$ to concatenate necessary information used in decision making, including the static global representation, the agent’s current state and the states of others. The context embedding $h_{k,(c)}^t$ is further taken as the single query vector and processed via another MHA layer. Finally, the output $g_k^t$ is used to compute the compatibility of choosing a node, and the final probabilities are computed via softmax. Such a decoding process is as follows:

$$g_k^t = MH A_{k,(c)}(h_1, h_2, \ldots, h_{2N})$$ (14)

$$Q_k^t, K_k^t = W_{Q,k}g_k^t, W_{K,k}h_t$$ (15)

$$u_k^t = D\text{tanh}(Q_k^t K_k^t / \sqrt{d_k})$$ (16)

$$p_{\theta_k,\phi}(v(k, t)) = \text{softmax}(Mask^t(u_k^t))$$ (17)

where $W_{Q,k}$ and $W_{K,k}$ are the weight matrices of the last single-head attention, $D=10$ is the clip rate for better exploration (Bello et al. 2016). $Mask^t$ resets all compatibility value of unfeasible (already visited) nodes to $-\infty$. In addition, $\theta_k$ is the parameter set for the agent decoder $k$, and $\phi$ is used to parameterize the common paired context embedding and context encoder.

During the cooperation of different agents, it is possible that several agents make the same decision to the same node. However, due to the constraint that each node can only be visited once as shown in equation (2) and (3), we design a special fleet handler to resolve such a conflict. It randomly maintains the action of one agent from all candidates to the node and keeps the others stay at their current location $v_{t_k}$. When all delivery and pickup nodes are visited and all agents return to the initial depot $v_0$, the episode ends.

Training via Cooperative A2C

The Advantage Actor-Critic (A2C) (Konda and Tsitsiklis 2000) is a well-known policy gradient approach that has shown its effectiveness. Leveraging the cooperation between different vehicle agents, we design a special cooperative A2C to train the proposed MAPDP.

Besides generating individual policies via individual decoders, we also formulate a centralized critic network $V_{\omega}^c(s)$ to estimate the state-values. In detail, we compute a weighted sum based on the output policy $p_{\theta_k,\phi}(v(k, t))$ at step $t$ and the node-wise embeddings $h_i$. Then the weighted sum vector of all agents is processed with linear projection into a single vector $v_c$ with $d_c = 128$. Finally, we obtain a single critic value from $v_c$ through two dense layers. The entire critic network is parameterized by $\omega$.

We then compute the loss of both networks and update $\theta_k, \gamma, \omega$ as follows:

$$A^\pi(s, a) = r(s, a) + V_{\omega}^c(s') - V_{\omega}^c(s)$$ (18)

$$\nabla \mathcal{L}(\theta_k) = \mathbb{E} \nabla \theta_k \log \pi_{\theta_k}(A^\pi(s, a))$$ (19)

$$\nabla \mathcal{L}(\phi) = 1/K \sum_{k=1}^{K} \mathbb{E} \nabla \phi \log \pi_{\theta_k}(A^\pi(s, a))$$ (20)

$$\nabla \mathcal{L}(\omega) = \mathbb{E} \nabla \omega (A^\pi(s, a))^2$$ (21)

Experiments and Evaluation

In this section, we conduct extensive experiments on two datasets to answer the following research questions:

- **RQ1:** How does our proposed MAPDP perform on cooperative PDP compared to other heuristics and RL based methods?
- **RQ2:** How balance is the cooperation between different agents when incorporating the fleet handler?
- **RQ3:** How effective is the multi-agent formulation and the communication embedding?

**Random Generated Dataset.** We first generate a random dataset with randomly distributed node locations with demands for efficient performance comparison. The location of node $v_i$, $\mathbf{L} = (x_i, y_i)$, is uniformly sampled from a $5 \times 5$ square. Both the $x$ and $y$ coordinates are uniformly distributed in $(0, 5)$. The demand volume of a pickup node $d_i$ is uniformly sampled from $(1, 10)$, and the capacity limit of each vehicle is 10.

**Real-World Dataset.** We collect real-world data from an online logistic platform providing services in Guangdong, China, including more than 100 thousand order
Performance Comparison

Baselines We first compare our DRLPR with the three widely recognized heuristics methods:

- **Ant Colony Optimization (ACO)** (Gambardella, Taillard, and Agazzi 1999) constructs number of ant colonies are established to model and optimize the objective functions.
- **Tabu Search** (Glover 1990) as a classic heuristic involves an enormous exploration space and keeps searching local solutions in the neighborhood based on the current one.
- **OR Tools** (Google 2021) is Google’s vehicle routing problem solver that utilizes a set of metaheuristics.

We also compare the most up-to-date RL-based approaches:

- **RL-VRP** (Nazari et al. 2018) proposed an RL-based framework using an encoder-decoder framework. The model utilizes an RNN model and generates the solution in a sequential manner.
- **AM-VRP** (Kool, van Hoof, and Welling 2019) utilize the transformer structure in both encoder and decoder. The agent network is trained via REINFORCE, and the results are decoded via a greedy method.
- **MDAM** (Xin et al. 2021) updates the encoding continuously via an embedding glimpse layer and further generates solutions based on a set of decoders with an additional training loss considering the KL divergence.

Since all RL baselines are only able to solve the standard CVRP, we manually add an additional mask for pickup-delivery constraints at the decoder of the three methods to guarantee that the output solutions are feasible. As for the cooperation among different vehicles, we set a vehicle assignment order so that the vehicles will take turns to be assigned with the decoded nodes.

Overall Comparison The comparison results of the overall performance are shown in Table ??.

9985
Besides the significant performance on solution quality, MAPDP also shows fast computation speed in solution generation. Compared to the heuristic baselines, we notice that MAPDP is able to infer solutions faster with a maximum of $400+ N = 100$. In real-world application scenarios, the logistic order requests may change continuously and even rapidly, including previous orders canceled and new orders accepted. A fast response-ability is of great significance for an online policy-making system to be adaptive to any new changing dynamics.

**Case Studies on Vehicle Cooperation**

To further investigate the cooperation between different vehicles in solving cooperative PDP, we compute the individual traveling distance of each vehicle with $N = 50$ in both datasets in two case studies, as shown in Figure 3. Besides, since making decisions simultaneously might meet conflicts with other agents, the fleet handler is used to deal with such conflicts and keep the other vehicles halting where they are. We also compute and report the halting ratio $= \text{halting times} / T$ of each agent $k$.

We notice that the balance among different agent’s workload is greatly affected by the dataset. In the random dataset where all nodes distribute randomly and are centered within the unit square, different agent decoder ends up with similar workloads. However, in the real-world dataset where nodes distribute unbalanced and even some locate distantly, agents show significant differences. This is because, in the same decision step, some vehicles might be assigned with one of the distant nodes, which greatly increases their total traveling distance. The agent with the longest traveling distance we investigated traveled 212 spatial units, while the one with the shortest trajectory only traveled 103 units. Meanwhile, in both datasets, we demonstrate that the agent that was halted the most ends up with the lowest workload. This is because such an agent skips the decision step while others continue to travel on the instance graph.

**Ablation Studies on Multi-Agent Design**

We further investigate how effective each part of the MAPDP design is as shown in Table 2. Two additional variants of MAPDP are trained and evaluated following the same evaluation protocols: MAPDP-SP stands for the simplified model where all agent decoders share the same parameters. Thus all agents become homogeneous to generate individual solutions. MAPDP-NC stands for the multi-agent framework without consideration on the communication embedding. It generates context embeddings only based on the agent’s own state and the global graph embedding.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>2N=20</th>
<th>2N=50</th>
<th>2N=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>MAPDP</td>
<td>24.87</td>
<td>52.59</td>
<td>88.97</td>
</tr>
<tr>
<td></td>
<td>MAPDP-SP</td>
<td>24.99</td>
<td>53.61</td>
<td>89.78</td>
</tr>
<tr>
<td></td>
<td>MAPDP-NC</td>
<td>26.89</td>
<td>68.78</td>
<td>108.12</td>
</tr>
<tr>
<td>Real</td>
<td>MAPDP</td>
<td>624</td>
<td>890</td>
<td>1705</td>
</tr>
<tr>
<td></td>
<td>MAPDP-SP</td>
<td>639</td>
<td>943</td>
<td>1721</td>
</tr>
<tr>
<td></td>
<td>MAPDP-NC</td>
<td>731</td>
<td>1033</td>
<td>1896</td>
</tr>
</tbody>
</table>

Table 2: Ablation study on the multi-agent structure design.

Results show that both MAPDP variants are outperformed by the original MAPDP. MAPDP-SP is slightly outperformed and is still superior to many other baselines. This shows that the multi-agent modeling along with a comprehensive communication mechanism is the core of performance improvement, while heterogeneous training can further slightly improve its effectiveness based on pure parameter sharing. However, MAPDP-NC is even outperformed by many other baselines. This is because, in a fully cooperative scenario, up-to-date communication with other agents is critical to effective coordination.

**Conclusion**

In this paper, we propose a multi-agent reinforcement learning framework to solve cooperative pickup and delivery problems (MAPDP). We design a special paired context embedding to explicitly represent the structural dependency among different nodes within the instance graph. We develop special cooperative multi-agent decoders to learn the individual policies of different agents. We also design a cooperative A2C algorithm for the integrated model training, where a joint critic value net estimates the state value. Extensive experiments demonstrate that MAPDP outperforms other baselines in all experiment settings and shows significant computation speed during solution inference.
Acknowledgements
This work was supported in part by The National Key Research and Development Program of China under grant 2020AAA0106000, the National Natural Science Foundation of China under U1936217, 61971267, 61972223, 61941117, 61861136003.

References


