LEARNING A TRANSFERABLE SCHEDULING POLICY FOR VARIOUS VEHICLE ROUTING PROBLEMS BASED ON GRAPH-CENTRIC REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning has been used to learn to solve various routing problems, however, most of the algorithm is restricted to finding an optimal routing strategy for only a single vehicle. In addition, the trained policy under a specific target routing problem is not able to solve different types of routing problems with different objectives and constraints. This paper proposes an reinforcement learning approach to solve the min-max capacitated multi vehicle routing problem (mCVRP), the problem seeks to minimize the total completion time for multiple vehicles whose one-time traveling distance is constrained by their fuel levels to serve the geographically distributed customer nodes. The method represents the relationships among vehicles, customers, and fuel stations using relationship-specific graphs to consider their topological relationships and employ graph neural network (GNN) to extract the graph’s embedding to be used to make a routing action. We train the proposed model using the random mCVRP instance with different numbers of vehicles, customers, and refueling stations. We then validate that the trained policy solve not only new mCVRP problems with different complexity (weak transferability but also different routing problems (CVRP, mTSP, TSP) with different objectives and constraints (storing transferability).

1 INTRODUCTION

The Vehicle Routing Problem (VRP), a well-known NP-hard problem, has been enormously studied since it appeared by Dantzig & Ramser (1959). There have been numerous attempts to compute the exact (optimal) or approximate solutions for various types of vehicle routing problems by using mixed integer linear programming (MILP), which uses mostly a branch-and-price algorithm appeared in Desrochers et al. (1992) or a column generation method (Chabrier, 2006), or heuristics ((Cordeau et al., 2002; Clarke & Wright, 1964; Gillett & Miller, 1974; Gendreau et al., 1994)). However, these approaches typically require huge computational time to find the near optimum solution. For more information for VRP, see good survey papers (Cordeau et al., 2002; Toth & Vigo, 2002).

There have been attempts to solve such vehicle routing problems using learning based approaches. These approaches can be categorized into supervised-learning based approaches and reinforcement-learning based approaches (Bengio et al., 2020): supervised learning approaches try to map a target VRP with a solution or try to solve sub-problems appears during optimization procedure, while reinforcement learning (RL) approaches seek to learn to solve routing problems without supervision (i.e., solution) but using only repeated trials and the associated reward signal. Furthermore, the RL approaches can be further categorized into improvement heuristics and construction heuristics (Mazyavkina et al., 2020): improvement heuristics learn to modify the current solution for a better solution, while construction heuristics learn to construct a solution in a sequential decision making framework. The current study focuses on the RL-based construction heuristic for solving various routing problems.

Various RL-based solution construction approaches have been employed to solve the traveling salesman problem (TSP) (Bello et al., 2016; Khalil et al., 2017; Nazari et al., 2018; Kool et al., 2018) or the capacitated vehicle routing problem (CVRP) (Nazari et al., 2018; Kool et al., 2018). (Bello et al., 2016; Nazari et al., 2018; Kool et al., 2018) has used the encoder-decoder structure to
sequentially generate routing schedules, and (Khalil et al., 2017) uses graph based embedding to determine the next assignment action. Although these approaches have shown the potential that the RL based approaches can learn to solve some types of routing problems, these approaches have the major two limitations: (1) only focus on routing a single vehicle over cities for minimizing the total traveling distance (i.e., min-sum problem) and (2) the trained policy for a specific routing problem cannot be used for solving other routing problems with different objective and constraints (they show that trained policy can be used to solve the same type of the routing problems with different problem sizes).

In this study, We proposed the Graph-centric RL-based Transferable Scheduler (GRLTS) for various vehicle routing problems. GRLTS is composed of graph-centric representation learning and RL-based scheduling policy learning. GRLTS is mainly designed to solve min-max capacitated multi vehicle routing problems (mCVRP); the problem seeks to minimize the total completion time for multiple vehicles whose one-time traveling distance is constrained by their fuel levels to serve the geographically distributed customer nodes. The method represents the relationships among vehicles, customers, and fuel stations using relationship-specific graphs to consider their topological relationships and employ graph neural network (GNN) to extract the graph’s embedding to be used to make a routing action. To effectively train the policy for minimizing the total completion time while satisfying the fuel constraints, we use the specially designed reward signal in RL framework. The representation learning for graph and the decision making policy are trained in an end-to-end fashion in an MARL framework. In addition, to effectively explore the joint combinatorial action space, we employ curriculum learning while controlling the difficulty (complexity) of a target problem.

The proposed GRLTS resolves the two issues raised in other RL-based routing algorithms:

- GRLTS learns to coordinate multiple vehicles to minimize the total completion time (makespan). It can resolve the first issue of other RL-based routing algorithms and can be used to solve practical routing problems of scheduling multiple vehicles simultaneously. (Kang et al., 2019) also employed the graph based embedding (random graph embedding) to solve identical parallel machine scheduling problem, the problem seeking to minimize the makespan by scheduling multiple machines. However, our approach is more general in that it can consider capacity constraint and more fast and scalable node embedding strategies.

- GRLTS transfers the trained scheduling policy with random mCVRP instances to be used for solving not only new mCVRP problems with different complexity but also different routing problems (CVRP, mTSP, TSP) with different objectives and constraints.

2 FORMULATION

2.1 MIN-MAX SOLUTION FOR MCVRP

We define the set of vehicles $V_V = 1, ..., N_V$, the set of customers $V_C = 1, ..., N_C$, and the set of refueling stations $V_R = 1, ..., N_R$, where $N_V$, $N_C$, and $N_R$ are the numbers of vehicles, customers, and refueling stations, respectively. The objective of min-max mCVRP is minimizing the makespan that is the longest distance among all vehicle’s traveling distance, i.e., $\min \ max_{i \in V_V} L_i$ with $L_i$ being the traveling distance of vehicle $i$, while each vehicle’s one-time traveling distance is constrained by its remaining fuel. The detailed mathematical formulation using mixed integer linear programming (MILP) is provided in Appendix. Figure 1 (left) shows a snapshot of a mCVRP state and Figure 1 (right) represents a feasible solution of the mCVRP.

2.2 DEC-MDP FORMULATION FOR MCVRP

We seek to sequentially construct an optimum solution. Thus, we frame the solution construction procedure as a decentralized Markov decision problem (Dec-MDP) as follows.

2.2.1 STATE

We define the vehicle state $s^v_i, \forall v \in V_V$, the customer state $s^c_i, \forall c \in V_C$, and the refueling station state $s^r_i, \forall r \in V_R$ as follows:
Figure 1: An example of mCVRP. (Left) a snapshot of environment state at time \( t \). The circular range of each vehicle indicates the possible moving range with the current fuel level. (Right) a feasible solution of the environment example.

- State of vehicle \( v \), \( s_v = (x_v, f_v, q_v) \). \( x_v \) is the allocated node that vehicle \( v \) to visit; \( f_v \) is the current fuel level; and \( q_v \) is the number of customers served by the vehicle \( v \) so far.

- State of a customer \( c \), \( s_c = (x_c, v_c) \). \( x_c \) is the location of customer node \( c \) (static). Visit indicator \( v_c \in \{0, 1\} \) becomes 1 if the customer \( c \) is visited and 0, otherwise.

- State of a refueling station \( r \), \( s_r = x_r \). \( x_r \) is the location of the refueling station \( r \) (static).

The global state \( s_t \) then becomes \( s_t = (\{s_v\}_{v=1}^{N_v}, \{s_c\}_{c=1}^{N_C}, \{s_r\}_{r=1}^{N_R}) \).

### 2.2.2 Actions & State Transition

Action \( a_v \) for vehicle \( v \) at time \( t \) is indicating a node to be visited by vehicle \( v \) at time \( t + 1 \), that is, \( a_v = x_{v+1} \in \{V_C \cup VR\} \). Therefore, the next state of vehicle \( v \) becomes \( s_v^{t+1} = (x_v^{t+1}, f_v^{t+1}, q_v^{t+1}) \) where \( f_v^{t+1} \) and \( q_v^{t+1} \) are determined deterministically by an action \( a_v \) as follows:

- Fuel capacity update: \( f_v^{t+1} = \begin{cases} F_v, & \text{if } a_v \in VR \\ f_v - d(x_v, a_v), & \text{otherwise}. \end{cases} \)

- Customer visit number update: \( q_v^{t+1} = \begin{cases} q_v, & \text{if } a_v \in VR \\ q_v + 1, & \text{otherwise}. \end{cases} \)

### 2.2.3 Rewards

The goal of mCVRP is to force all agents to coordinate to finish the distributed tasks quickly while satisfying the fuel constraints. To achieve this global goal in a distributed manner, we use the specially designed independent reward for each agent as:

- **visiting reward**: To encourage vehicles to visit the customer nodes faster, in turn, minimizing makespan, we define customer visit reward \( r_{v, visit} = q_v \). This reward is provided when an agent visits a customer; the more customer nodes a vehicle agent \( n \) visits, the greater reward it can earn.

- **Refueling reward**: To induce a strategic refueling, we introduce refuel reward \( r_{v, refuel} = q_v \times ((F_v - f_v) / (F_v - 1))^\alpha \). We define the refuel reward as an opportunity cost. That is, vehicles with sufficient fuel are not necessary to refuel (small reward). In contrast, refueling vehicles with a lack of fuel is worth as much as visiting customers. In this study, we set \( F_v = 10 \) (which is the equivalent to the total traveling distance that vehicle \( v \) can travel with the fuel tank fully loaded) and \( \alpha = 2 \).

### 2.3 Relationships with Other Class of VRPs

mCVRP, the target problem of this study, has three key properties: 1) the problem seeks to minimize the total completion time of vehicles by forcing all vehicles to coordinate (in a distributed manner), 2) the problem employs fuel capacity constraints requiring the vehicles to visit the refueling stations strategically, 3) the problem considers multiple refueling depots (revisit allowed).
If some of these requirements are relaxed, min-max mCVRP can be degenerated into simpler conventional routing problems:

- **TSP** is the problem where a single vehicle is operated to serve every customer while minimizing the total traveling distance. The agent needs or needs not come back to the depot. This problem does not have capacity constraints.
- **CVRP** (capacity-constrained TSP) is the problem where a single vehicle is operated to serve every customer while minimizing the total traveling distance and satisfying the fuel constraint. The vehicles need to comeback depot to charge.
- **mTSP** (multi-agent TSP) is the problem where multiple vehicles should serve all the customers as quickly as possible. This problem does not have capacity constraints.
- **mCVRP** (multi-agent, capacity-constrained TSP) is our target problem having the properties of both mTSP and CVRP. Additionally, we add more than one refueling depot.

The mathematical formulations for these problems are provided in Appendix. We train the policy using random mCVRP instances with varying numbers of agents and customers and employ the trained policy without parameter changes to solve TSP, CVRP and mTSP to test its domain transferability.

3 Method

This section explains how the proposed model, given a state (a partial solution), assigns an idle vehicle to next node to visit under the sequential decision-making framework (see Figure 2).

3.1 State Representation using Relationship-Specific Multiple Graphs

The proposed model represents the global state $s_t$ using as a weighted graph $G_t = G(V, E, w)$ where $V = \{V_v, V_c, V_R\}$, and $E$ is the set of edges between node $i, j \in V$ and $w$ is weight for edge $(i, j)$ (here, distance $d_{ij}$). Each node corresponding to vehicle, customer, and refueling station will be initialized with its associated states defined earlier.

Although we can assume that all nodes are connected with each other regardless of types and distance, we restrict the edge connection to its neighboring nodes to reduce the computational cost. Specifically, each type of node can define its connectivity range and connect an edge if any node is located within its range as follows (see Figure 2):

\[
e_{v, j} = 1 \quad \forall v \in V_v, \quad d(v, j) \leq R_v = f^v_i
\]
\[
e_{c, j} = 1 \quad \forall c \in V_c, \quad d(c, j) \leq R_c
\]
\[
e_{r, j} = 1 \quad \forall r \in V_R, \quad d(r, j) \leq R_R = \max_{v \in V_v} F^v
\]
That is, the vehicle node $v \in V_v$ connects the edges with nodes that are located within its traveling distance (i.e., the current fuel level $f^v_i$). In addition, the customer nodes $c \in V_c$ connects the edges with nodes that are located within the constant range $R_C$. We set $R_C = 5$ while following a typical hyperparameter selection procedure. Finally, the refueling node $r \in V_R$ connects the edges with nodes within maximum distance that the vehicle with the largest fuel capacity can travel with the full loaded fuel (in this study $F^v = 10$ for all vehicles). Note that the target node $j$ that can be connected to each node can be any types of nodes.

3.2 NODE EMBEDDING USING GNN

The proposed model employs Graph neural network (GNN) [Scarselli et al., 2008] to compute the node embeddings for all nodes. The node embedding procedure starts with constructing the graph $G_i$ out of the global state $s_i = \{(s_i^v)_{v=1}^{N_v}, (s_i^c)_{c=1}^{N_c}, (s_i^r)_{r=1}^{N_r}\}$. The method then compute the initial node embeddings $h_i$ and edge embeddings $e_{ij}$ for all the nodes and edges of $G_i$ by employing encoder network. The sequel will explain how the GNN layer update these node and edge embeddings.

3.2.1 EDGE UPDATE

The edge feature $h_{ij}^\tau$ at $\tau$ iteration is updated using the edge updater function $\phi_E$ as

$$ h_{ij}^\tau = \phi_E(h_{ij}^\tau-1, h_i^\tau-1, h_j^\tau-1), \quad \forall i \in V, \forall j \in N_i $$

where $h_i^\tau-1$ and $h_j^\tau-1$ are node embedding vectors of node $i$ and node $j$ at $\tau - 1$ iteration.

3.2.2 EDGE FEATURE AGGREGATION

The updated edge feature $h_{ij}^\tau$ can be thought of as an message sent from node $j$ to node $i$. Node $i$ aggregates these messages from its neighboring nodes $j \in \{N_V(i), N_C(i), N_R(i)\}$, where $N_V(i)$ is the neighboring vehicle nodes of node $i$, as

$$ \langle \bar{h}_{i,V}^\tau, \bar{h}_{i,C}^\tau, \bar{h}_{i,R}^\tau \rangle = \left( \sum_{j \in N_V(i)} \alpha_{ij} h_{ij}^\tau, \sum_{j \in N_C(i)} \alpha_{ij} h_{ij}^\tau, \sum_{j \in N_R(i)} \alpha_{ij} h_{ij}^\tau \right) $$

where the attention weight $\alpha_{ij} = \text{softmax}(e_{ij})$, where $e_{ij} = f_e(s_i, s_j, w_e)$, scores the significance of node $j$ to node $i$. Note that message aggregation is separately conducted for different types of node and the aggregated messages per type are concatenated.

3.2.3 NODE FEATURE UPDATE

The aggregated edge node embeddings $\langle \bar{h}_{i,V}^\tau, \bar{h}_{i,C}^\tau, \bar{h}_{i,R}^\tau \rangle$ per its neighborhood type are then used to update the node embedding vector $h_i^\tau$ using node update function $\phi_V$ as

$$ h_i^\tau = \phi_V(\bar{h}_{i,V}^\tau, \bar{h}_{i,C}^\tau, \bar{h}_{i,R}^\tau) $$

The node embedding procedure is repeated $H$ (hop) times for all nodes, and the final node embeddings $h_i^H = \{h_i^H\}_{i=1}^N$ is used to determine the next assignment action of an idle vehicle.

3.3 DECISION MAKING USING NODE EMBEDDING

When an vehicle node $i$ reaches the assigned customer node, an event occurs and the vehicle node $i$ computes its node embedding $h_i^H$ and selects one of its feasible actions, choosing one of unvisited customer nodes or refueling nodes around vehicle node $i$. The probability for agent $i$ to choose node $j$, $a_i^t = j$, is computed by the parameterized actor $\pi(a_i^t = j|s_i^t; \phi)$ as:

$$ \pi(a_i^t = j|s_i^t; \phi) = \frac{\exp(F(h_i^H, h_j^H; \phi))}{\sum_{k \in N_V(i) \cup N_R(i)} \exp(F(h_i^H, h_k^H; \phi))} $$

where $F(h_i^H, h_j^H; \phi)$ is the fitness function evaluating the goodness for agent $i$ to choose node $j$ as the next action (i.e., $a_i^t = j$).
3.4 TRAINING GRLTS

We train the proposed model using random mCVRP instances with varying numbers of vehicles, customers. We employ the actor-critic algorithm to train the parameters for the GNN, the critic, and the actor (GRLTS).

We first approximate the action-value function \(Q^\pi(s, a) \approx \hat{Q}^\pi(s, a; \theta)\) using a neural network with the critic parameter \(\theta\). The parameter \(\theta\) for the centralized critic is optimized to minimize the loss \(L\):

\[
L(\theta) = \mathbb{E}_{s, a, r, s' \sim D}[(\hat{Q}^\pi(s, a; \theta) - y)^2]
\]

and the critic \(\theta\) is a state transition memory.

To train the actor network, we use PPO method \cite{Schulman2017} to maximize \(J(\phi)\). PPO aims to maximize the clipped surrogate objective function as follow:

\[
J^{\text{CLIP}}(\phi) = \mathbb{E}_t[\min(R_t(\phi)\hat{A}^t, \text{clip}(R_t(\phi), 1 - \epsilon, 1 + \epsilon)\hat{A}^t)]
\] (9)

where \(R_t(\phi) = \frac{\pi_a(a|s_t)}{\pi_a'(a'|s_t)}\). We compute the advantage estimator \(\hat{A}_t\) by running the policy for \(T\) time steps as

\[
\hat{A}_t = \delta_t + \gamma \delta_{t+1} + \cdots + \gamma^{T-1} \delta_{T-1}
\]

\[
\delta_t = r_t + \gamma V(h_{t+1}^{N_{hop}}; \theta) - V(h_t^{N_{hop}}; \theta)
\] (10)

In addition, equation (9) is added by a value function error and an entropy bonus for sufficient exploration because critic and actor share parameters as follow:

\[
J^{\text{CLIP}}_t(\phi) = \mathbb{E}_t[J^{\text{CLIP}}_t(\phi) - c_1(V_\theta(s_t) - V_{target})^2 + c_2 H(s_t, \pi_\phi(\cdot))]
\] (11)

where \(V_\theta(s_t) \approx V(h_t^{N_{hop}}; \theta)\) from the centralized critic; \(H\) denotes an entropy bonus; and \(c_1\) and \(c_2\) are hyperparameters.

On updating the centralized critic parameter \(\theta\) and the decentralized (but, shared) actor parameter \(\phi\), we use Monte-Carlo simulation. To sequentially update both parameters, we follow an update rule as follow:

\[
\theta^{k+1}, \phi^{k+1} \leftarrow \arg \max_{\theta^*, \phi^*} \mathbb{E}_{s, a, r, s' \sim P_{\pi^*_\theta}} [J^{\text{CLIP}}_t(\phi) - c_1(V_\theta(s_t) - V_{target})^2 + c_2 H(s_t, \pi_\phi(\cdot))]
\] (12)

Algorithm 1 focuses on sequence of the parameter update using above equations.

**Algorithm 1** Parameter update in decentralized actors with PPO and Monte-Carlo simulation

1: for Agent \(i = 1, 2, \ldots, N\) do
2: for Update iteration = 1, 2, \ldots, \(K\) do
3: Evaluate policy \(\pi_{\theta_{old}}\) in environment for an experienced episode
4: Compute \(r_t(\phi)\) in Equation 9
5: Compute advantage estimates \(\hat{A}_1, \ldots, \hat{A}_T\) in Equation 10
6: Optimize surrogate objective in Equation (11) with batch size \(T\)
7: \(\theta_{old} \leftarrow \theta, \phi_{old} \leftarrow \phi\) followed by (12)
8: end for
9: end for

4 EXPERIMENTS

The trained policy is used to solve various types of vehicle routing problems, mCVRP, mTSP, CVRP, and TSP without changing the parameters. For each type except mCVRP, we use two types of data
sets: random instances with varying numbers of vehicles and customers whose locations are sampled randomly, and (2) benchmark problems obtained from the library (mTSP library, CVRP library, and TSP library).

To validate the proposed GRLTS, we develop mixed integer linear programming (MILP) formulations and compute the solutions using CPLEX 12.9 (all the optimization formulations are provided in the Appendix). We also use Google OR-Tools(Perron & Furnon) as the representative heuristic solvers. For the problems where other deep RL based approaches tried to solve (TSP and CVRP), we compare the performance of the proposed approach to those of deep RL baselines. All experiments are conducted on Windows 10, Intel(R) Core(TM) i9-9900K CPU 3.6 GHz, and 32GB RAM. GPU acceleration is not used in testing.

4.1 PERFORMANCE COMPARISON OF mCVRPs

<table>
<thead>
<tr>
<th>Method</th>
<th>mCVRP25</th>
<th>mCVRP50</th>
<th>mCVRP100</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORT</td>
<td>2.51</td>
<td>4.31</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(4.5)</td>
<td>(12.1)</td>
<td>(∞)</td>
</tr>
<tr>
<td>GRLTS</td>
<td>3.15</td>
<td>3.72</td>
<td>2.71</td>
</tr>
<tr>
<td></td>
<td>(0.9)</td>
<td>(4.7)</td>
<td>(11.8)</td>
</tr>
<tr>
<td>Gap (%)</td>
<td>25.5</td>
<td>-13.7</td>
<td>-0.5</td>
</tr>
</tbody>
</table>

Table 1: Performance comparison of mCVRPs on random instances

<table>
<thead>
<tr>
<th>mCVRP25-5</th>
<th>mCVRP100-10</th>
<th>mCVRP400-20</th>
<th>mCVRP2500-50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2 (0.98)</td>
<td>1.88 (12.1)</td>
<td>4.09 (420.2)</td>
<td>11.71 (3,861.2)</td>
</tr>
</tbody>
</table>

Table 2: Scalability test of the trained GRLTS with large-scale mCVRPs

We apply the trained policy (GRLTS) to solve mCVRP with different numbers of vehicles $N_v \in 2, 3, 5$ and the numbers of customers $N_c \in 25, 50, 100$. The number of refueling stations are set to be 4, 5 and 10 in case of $N_c = 25$, $N_v = 50$ and $N_c = 100$. For every combination of $N_v$ and $N_c$, we randomly generate 100 mCVRP instances, each of which has randomly located customer nodes and refueling nodes. The x and y coordinates of each node is randomly sampled from the uniform distributions; $x \sim U[0,1]$ and $y \sim U[0,1]$, respectively. We employ GRLTS and ORT to solve the same 100 random mCVRP instances and compute the average makespan and the computational time required to solve the mCVRP instance with the policies. Table 1 compares the average makespan and computational times.

ORT (Google OR-tool) produces the best results for the small-sized problems with reasonable computational time; however, it requires extensive time or even fails to compute any feasible solution for the large scale problem ($\infty$ means that ORT cannot find a feasible solution). Notably, the GRLTS achieves a better performance than ORT for large-scale problems with a significantly shorter computational time.

To validate the scalability of the proposed model, we further conduct test experiments with the four cases: $N_c = 25$ with $N_v = 5$, $N_c = 100$ with $N_v = 10$, $N_c = 400$ with $N_v = 20$, and $N_c = 2,500$ with $N_v = 50$. Table 2 shows how the makespan and the computational time increases with the size of the problem. By comparing the ratio between the number of customers that each vehicle need to serve, (5: 10: 20: 50), and the makespan, (1.20: 1.88: 4.09: 11.71), we can roughly confirms that the trained model can perform reasonably well even in large sized problem that have never been experience during the training.

For all different sizes of the mCVRP problem, we also compare the performance of the proposed method with the solution computed by CPLEX solver. Because it typically takes a long time to compute the near-optimum solution by CPLEX, we only compare the single mCVRP case. The result is provided in the Appendix.
4.2 Performance Comparison on mTSP

We apply the trained network (without parameter changes) to solve mTSP, which seeks to minimize the total completion time of multiple vehicles (minmax mTSP). This problem is a relaxed version of mCVRP in that it does not require the capacity constraints (we maintain the fuel level to be maximum during execution). We solve the randomly generated 100 mTSP instances for every combination of $N_c \in \{50, 100, 400\}$ and $N_v \in \{2, 4, 6, 8\}$ (or $N_v \in \{10, 20, 30, 40\}$). Table 7 compares the average makespan and computational time. The trained model outperforms ORT in the large-size problems (see $N_c = 100$ and $N_c = 400$ cases in Table 7) in terms of both makespan and the computational time. For large scale problems, GRLTS achieves roughly 28\% shorter makespan than ORT on average with significantly reduced computational time (50\% reduction).

Table 3: Performance comparison on mTSP for random instances

<table>
<thead>
<tr>
<th>$N_c$</th>
<th>mTSP50</th>
<th>mTSP100</th>
<th>mTSP400</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>ORT</td>
<td>3.20</td>
<td>2.21</td>
<td>1.60</td>
</tr>
<tr>
<td></td>
<td>(1.3)</td>
<td>(1.2)</td>
<td>(1.2)</td>
</tr>
<tr>
<td>GRLTS</td>
<td>3.94</td>
<td>2.83</td>
<td>1.47</td>
</tr>
<tr>
<td></td>
<td>(3.1)</td>
<td>(3.2)</td>
<td>(3.0)</td>
</tr>
<tr>
<td>Gap (%)</td>
<td>23.1</td>
<td>28.1</td>
<td>-8.1</td>
</tr>
</tbody>
</table>

Similarly, for all the sizes of mTSP case, we compare the performance of GRLTS to the solution computed by CPLEX in the Appendix. We also employed the trained model to solve mTSP benchmark problems in mTSLbib solving MinMax mTSP. The performance results on 124 instances (four different number of vehicles per 31 different maps) are provided in Table 9 in Appendix.

4.3 Performance Comparison on CVRP (Payload Capacity)

We employ the trained model to solve randomly generated 100 CVRP instances with $N_c = 20$, 50, 100 with a single capacitated vehicle. All nodes are randomly scattered in the unit square of $[0, 1] \times [0, 1]$. To make the trained policy compatible with the CVRP settings, we fix the number of vehicles to be one and use a single refueling node as if it were the depot in CVRP environment. As a result, the single vehicle needs to revisit the depot due to payload capacity (which is equivalent to fuel constraint). The payload capacity is set to be 30, 40 and 50 for CVRP20, CVRP50 and CVRP100, respectively. Demand are uniformly sampled from $\{1, \ldots, 9\}$. These test cases have been widely used by the studies seeking to develop RL-based solvers for CVRP problems.

Table 4 summaries results for the random CVRP environment. L2I and LKH3 are well-known best performing heuristic algorithms developed from the OR community, thus can serve as an oracle for comparing the performance. We also consider other RL-based approaches Kool et al. 2018.
Under review as a conference paper at ICLR 2021

In general, our model is not outperforming other RL-based approaches and OR-tool. However, all RL-based approaches, except our model, are trained under the same environment of CVRP cases (the training and test instances are sampled from distribution). However, our model is trained with the complete different mCVRP environment and tested with the CVRP environment (strong transferability).

We also employed the trained policy to solve the CVRP benchmark problem instances from CVRPLib (Uchoa et al., 2017). Table 10 in Appendix provides all the performance results compared with other RL-baseline models. The results show that our model has better performance than one of the state-of-art RL-based approach (Kool et al., 2018).

4.4 Performance Comparison of TSP

<table>
<thead>
<tr>
<th>Methods</th>
<th>CVRP20 Obj. (%)</th>
<th>CVRP20 Gap (%)</th>
<th>CVRP20 Time</th>
<th>CVRP50 Obj. (%)</th>
<th>CVRP50 Gap (%)</th>
<th>CVRP50 Time</th>
<th>CVRP100 Obj. (%)</th>
<th>CVRP100 Gap (%)</th>
<th>CVRP100 Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concorde</td>
<td>3.84</td>
<td>0.00</td>
<td>1m</td>
<td>5.70</td>
<td>0.00</td>
<td>2m</td>
<td>7.76</td>
<td>0.00</td>
<td>3m</td>
</tr>
<tr>
<td>LKH3</td>
<td>3.84</td>
<td>0.00</td>
<td>18s</td>
<td>5.70</td>
<td>0.00</td>
<td>5m</td>
<td>7.76</td>
<td>0.00</td>
<td>21m</td>
</tr>
<tr>
<td>OR-Tools</td>
<td>3.85</td>
<td>0.26</td>
<td>0s</td>
<td>5.80</td>
<td>1.75</td>
<td>1s</td>
<td>7.99</td>
<td>2.96</td>
<td>3s</td>
</tr>
<tr>
<td>AM (greedy)</td>
<td>3.85</td>
<td>0.26</td>
<td>0s</td>
<td>5.80</td>
<td>1.75</td>
<td>2s</td>
<td>8.12</td>
<td>4.64</td>
<td>6s</td>
</tr>
<tr>
<td>AM (sampling)</td>
<td>3.84</td>
<td>0.00</td>
<td>5m</td>
<td>5.73</td>
<td>0.53</td>
<td>24m</td>
<td>7.94</td>
<td>2.32</td>
<td>1h</td>
</tr>
<tr>
<td>Nearest Insertion</td>
<td>4.33</td>
<td>12.76</td>
<td>1s</td>
<td>6.78</td>
<td>18.95</td>
<td>2s</td>
<td>9.46</td>
<td>21.91</td>
<td>6s</td>
</tr>
<tr>
<td>Random Insertion</td>
<td>4.00</td>
<td>4.17</td>
<td>0s</td>
<td>6.13</td>
<td>7.54</td>
<td>1s</td>
<td>8.52</td>
<td>9.79</td>
<td>3s</td>
</tr>
<tr>
<td>Farthest Insertion</td>
<td>3.93</td>
<td>2.34</td>
<td>1s</td>
<td>6.01</td>
<td>5.44</td>
<td>2s</td>
<td>8.35</td>
<td>7.60</td>
<td>7s</td>
</tr>
<tr>
<td>Bello et al. (2016)</td>
<td>3.89</td>
<td>1.30</td>
<td>3s</td>
<td>5.95</td>
<td>4.39</td>
<td>3s</td>
<td>8.3</td>
<td>6.96</td>
<td></td>
</tr>
<tr>
<td>Khalil et al. (2017)</td>
<td>3.89</td>
<td>1.30</td>
<td>5.99</td>
<td>5.09</td>
<td>8.31</td>
<td>7.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRLTS (Ours)</td>
<td>3.92</td>
<td>2.08</td>
<td>1s</td>
<td>6.32</td>
<td>10.88</td>
<td>3s</td>
<td>8.79</td>
<td>13.27</td>
<td>9s</td>
</tr>
</tbody>
</table>

Lastly, we employ the trained model to solve 100 randomly generated TSP with different number of customers $N_c = 20, 50, 100$ and a single vehicle. Table 5 shows summarized the results. Although GRLTS is not outperforming other RL-based scheduling methods, it shows the reasonably performance that is comparable to some of well known heuristic algorithms. Given that the GRLTS is trained with mCVRP environment and have never seen TSP instances, this result can validate that GRLTS can be transferred to TSP as well.

We also employed the trained policy to solve the TSP benchmark problem instances from TSPLib (Reinelt, 2014). Table 8 in Appendix provides all the performance results compared with other RL-baseline models. The results show that our model has comparable performance with the-state-of-art RL-based approaches (GPN (Ma et al., 2019) and S2V-DQN (Khalil et al., 2017)) and heuristic algorithms, especially for large-scale TSP problems.

5 Conclusion

We proposed the Graph-centric RL-based Transferable Scheduler for various vehicle routing problems using graph-centric state presentation (GRLTS) that can solve any types of vehicle routing problems such as mCVRP, mTSP, CVRP, and TSP. The transferability is achieved by graph-centric representation learning that can generalize well over various relationships among vehicles, customers, and refuel stations (depot). GRLTR is computationally efficient for solving very large-scale vehicle routing problems with complex constraints, which provides potential that such RL-based scheduler can be used for large-scale realistic applications in logistics, transportation, and manufacturing.

References


Laurent Perron and Vincent Furnon. Or-tools. URL [https://developers.google.com/optimization/](https://developers.google.com/optimization/)
G Reinhelt. `{TSPLIB}`: a library of sample instances for the tsp (and related problems) from various sources and of various types. *URL: http://comopt.ifi.uniheidelberg.de/software/TSPLIB95*, 2014.


A APPENDIX

A.1 MILP FORMULATIONS FOR ROUTING PROBLEMS

The MILP formulations for mCVRP, mTSP, and TSP are provided in this section. All problems are formulated to find the optimum routes of vehicles to minimize the makespan. For TSP formulation, the makespan minimization is equal to total distance minimization due to single-vehicle operation. Although mTSP here is same as multi-VRP by its definition, we denote the problem as mTSP because we focus on the makespan minimization problem with multi-vehicle operation.

A.1.1 TSP

A fundamental formulation in the context of routing problems is based on TSP formulation by Miller-Tucker-Zemlin (MTZ) formulation (Miller et al., 1960):

\[
\text{minimize } \sum_{i \in V} \sum_{j \in V, i \neq j} d_{ij} x_{ij} \\
\text{subject to. } \sum_{i \in V} x_{ij} = 1, \quad \forall j \in V: i \neq j, \quad (1)
\]

\[
\sum_{j \in V} x_{ij} = 1, \quad \forall i \in V: i \neq j, \quad (2)
\]

\[
u_i - u_j + |V|x_{ij} \leq |V| - 1, \quad \forall i, j \in V \setminus \text{vStart}, \quad (3)
\]

\[
0 \leq u_i \leq |V| - 1, \quad \forall i \in V \setminus \text{vStart}, \quad (4)
\]

\[
x_{ij} \in \{0, 1\}, \quad \forall i, j \in V, \quad (5)
\]

\[
u_i \in \mathbb{Z}, \quad \forall i \in V \setminus \text{vStart}. \quad (6)
\]

The routing problem is defined in a graph \(G(V, E, w)\) where \(V\), \(E\) and \(w\) are nodes, edges, and weight (distance \(d\) or cost \(c\)). \(x_{ij}\) is 1 if the edge between node \(i\) and node \(j\) exists, otherwise, 0. Here, corresponding distance of the edge is \(d_{ij}\). Constraint (3,4) with dummy variable \(u_i\) is for subtour elimination. As setting \(d_{ij} = 0, \forall j \in \text{vStart}\), the problem can be arbitrary end assumption.

A.1.2 mTSP

The goal of mTSP is to find the sub-routes of multiple vehicles to minimize the makespan. Thus, decision variables are expanded to multi-vehicle settings, and the objective function is modified from A.1.1 to consider the makespan minimization setting (MinMax problem).

\[
\text{minimize } Q \\
\text{subject to. } \sum_{i \in V} \sum_{j \in V} d_{ij} x_{aij} \leq Q, \quad \forall a \in A : i \neq j, \quad (1)
\]

\[
\sum_{j \in V} x_{aij} = 1, \quad \forall a \in A, i \in \text{vStart}_a : i \neq j, \quad (2)
\]

\[
\sum_{a \in A} \sum_{i \in G} x_{aij} = 1, \quad \forall j \in \text{vTask} : i \neq j, \quad (3)
\]

\[
\sum_{i \in G} x_{aij} - \sum_{i \in G} x_{aji} = 0, \quad \forall a \in A, j \in V : i \neq j, \quad (4)
\]

\[
u_{ai} - u_{aj} + |V|x_{aij} \leq |V| - 1, \quad \forall a \in A, j \in V \setminus \text{vStart}_a : i \neq j, \quad (5)
\]

\[
0 \leq u_{ai} \leq |V| - 1, \quad \forall a \in A, i \in V \setminus \text{vStart}_a, \quad (6)
\]

\[
x_{aij} \in \{0, 1\}, \quad \forall a \in A, \forall i, j \in V, \quad (7)
\]

\[
u_{ai} \in \mathbb{Z}, \quad \forall a \in A, i \in V \setminus \text{vStart}_a. \quad (8)
\]

Here, we denoted \(Q\) as the makespan, which is the longest traveling distance among multiple vehicles. By minimizing this maximum traveling distance, the above formulation can minimize the makespan. Therefore, the above formulation is to minimize the makespan. We also allow a vehicle to start a tour at any staring node \(v\text{Start}_a\) (constraint (2)).
A.1.3 mCVRP

We extend the above mTSP formulation in A.1.2 to include the fuel constraint (i.e., allowable traveling distance per sub-tour). In this formulation, vehicles can start their tour at an arbitrary location.

\[
\text{minimize} \quad Q \\
\text{subject to.} \quad \sum_{i \in V} \sum_{j \in V} d_{ij} x_{aij} \leq Q, \quad \forall a \in A : i \neq j, \quad (1) \\
\sum_{j \in V} x_{aij} = 1, \quad \forall a \in A, i \in v_{Start_a} : i \neq j, \quad (2) \\
\sum_{a \in A} \sum_{i \in G} \sum_{j \in G} x_{aij} = 1, \quad \forall j \in v_{Task} : i \neq j, \quad (3) \\
\sum_{i \in G} x_{aij} - \sum_{i \in G} x_{aji} = 0, \quad \forall a \in A, j \in V : i \neq j, \quad (4) \\
u_{ai} - u_{aj} + |V| x_{aij} \leq |V| - 1, \quad \forall a \in A, j \in V \setminus v_{Start_a} : i \neq j, \quad (5) \\
0 \leq u_{ai} \leq |V| - 1, \quad \forall a \in A, i \in v_{Start_a}, \quad (6) \\
x_{aij} \in \{0, 1\}, \quad \forall a \in A, \forall i, j \in V, \quad (7) \\
u_{ai} \in \mathbb{Z}, \quad \forall a \in A, i \in V \setminus v_{Start_a} \quad (8) \\
f_{ai} = F_a, \quad \forall a \in A, i \in v_{Refuel}, \quad (9) \\
f_{ai} - d_{ij} x_{aij} \geq 0, \quad \forall a \in A, \forall i, j \in v_{Refuel}, \quad (10) \\
f_{aj} - f_{ai} + d_{ij} x_{aij} \geq F_a(1 - x_{aij}), \quad \forall a \in A, \forall i, j \in v_{Task}. \quad (11)
\]

Constraint (9) indicates that a vehicle can charge fuel as much as the vehicle’s maximum fuel capacity $F_a$. In addition, constraint (10) requires that a vehicle must have enough remaining fuel to visit a refueling node. Constraint (11) indicates that a vehicle cannot consume (travel) more than its fuel capacity. We allow vehicles to visit refueling nodes as many times as possible by introducing a sufficient number of dummy variables.

A.1.4 CVRP

Similar to the above mCVRP with fuel constraint, we also solve CVRP constrained by payload capacity (not fuel constraint) and with single vehicle ($m = 1$ so that $|A| = 1$). We consider a single depot to start a mission and unload some burdens. Therefore, $v_{Refuel}$ in the MILP of mCVRP becomes same as $v_{Start_a}$. Fuel consumption as much as distance $d_{ij}$ between node $i$ and node $j$ is equivalent to payload capacity consumption as much as demand $d_j$ at node $j$. As a result, constraint (9-11) of the mCVRP model becomes:

\[
\begin{align*}
  c_{ai} &= C_a, & \forall a \in A, i \in v_{Start}, & (9) \\
  c_{aj} - c_{ai} + d_j x_{aij} &\geq C_a(1 - x_{aij}), & \forall a \in A, \forall i, j \in v_{Task}. & (10)
\end{align*}
\]

The constraint (10) in the mCVRP model is dropped because payload capacity is independent in returning to the starting depot.

A.2 Comparison with CPLEX solutions

For a single mCVRP and mTSP instances, we compare the performance of the proposed approach with the exact solution computed by CPLEX. This experiment shows how the solution computed from the proposed method is comparable to the near-optimum solution calculated by the powerful optimization solver. Note that the mCVRP instance for this experiment was generated from the grid environment. The distance between customer nodes are different from the experiments for Table 1.
A.2.1 mCVRP

Table 6: Performance comparison of mCVRPs on random instances

<table>
<thead>
<tr>
<th>N_c</th>
<th>N_v</th>
<th>CPX</th>
<th>ORT</th>
<th>RL</th>
<th>N_v</th>
<th>CPX</th>
<th>ORT</th>
<th>RL</th>
<th>N_v</th>
<th>CPX</th>
<th>ORT</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>2</td>
<td>20 (∞)</td>
<td>15 (3.9)</td>
<td>20 (0.96)</td>
<td>2</td>
<td>-</td>
<td>34 (38.3)</td>
<td>30 (4.8)</td>
<td>5</td>
<td>-</td>
<td>27 (510.3)</td>
<td>29 (11.9)</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>14 (∞)</td>
<td>10 (36.6)</td>
<td>16 (1.01)</td>
<td>3</td>
<td>-</td>
<td>22 (2.2)</td>
<td>22 (4.1)</td>
<td>10</td>
<td>-</td>
<td>-</td>
<td>20 (12.1)</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>12 (∞)</td>
<td>-</td>
<td>8 (0.98)</td>
<td>5</td>
<td>-</td>
<td>14 (1.7)</td>
<td>14 (3.8)</td>
<td>20</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

We apply the already trained model to solve mCVRP with different numbers of vehicles \(N_v \in 2, 3, 5\) and the numbers of customers \(N_c \in 25, 50, 100\). The number of refueling stations are set to be 4, 5 and 10 in case of \(N_c = 25\), \(N_c = 50\) and \(N_c = 100\). For all cases, we compute the total completion time and the computational time (the number in the parenthesis) required to construct a scheduling. We set the limit of the computational time to be 18,000 (sec) for all cases. The symbol ∞ indicate the case where the computational time is reached and the solution at that moment is used in the table. The blanks with the hyphen (-) indicate the case where the algorithm could not find any feasible solution.

First, CPX (CPLEX) can only solve for mCVRPs with \(N_c = 25\); for large scale problems, it cannot compute any feasible solution. ORT (Google OR-tool) produces the best results for the small sized problem with reasonable computational time; however, for the large scale problem it requires extensive amount of time or even fail to compute any feasible solution. Notably, the proposed GRLTS achieves the good performance with the least amount of computational time.

A.2.2 mTSP

For the grid environment, a single mTSP instance was generated and used for comparing the performances of GRLTS to CPLEX and ORT.

Table 7: Performance comparison on mTSP for random instances

<table>
<thead>
<tr>
<th>N_c</th>
<th>CPX</th>
<th>ORT</th>
<th>RL</th>
<th>N_c</th>
<th>CPX</th>
<th>ORT</th>
<th>RL</th>
<th>N_c</th>
<th>CPX</th>
<th>ORT</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>26</td>
<td>32</td>
<td>51 (∞)</td>
<td>57</td>
<td>57</td>
<td>57</td>
<td>10</td>
<td>-</td>
<td>63</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>31</td>
<td>26</td>
<td>54 (∞)</td>
<td>31</td>
<td>26</td>
<td>26</td>
<td>20</td>
<td>-</td>
<td>41</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>24</td>
<td>25</td>
<td>46 (∞)</td>
<td>25</td>
<td>24</td>
<td>24</td>
<td>30</td>
<td>-</td>
<td>44</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>9</td>
<td>8</td>
<td>51 (∞)</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>40</td>
<td>-</td>
<td>39</td>
<td>24</td>
<td></td>
</tr>
</tbody>
</table>

A.3 EXPERIMENT RESULTS ON BENCHMARK PROBLEMS

We apply the trained network to solve benchmark problems for TSP, mTSP and CVRP.

A.3.1 TSP

Table 8 shows results of TSPLib (Reinhelt, 2014) with some RL-based approaches (GPN (Ma et al., 2019) and S2V-DQN (Khalil et al., 2017) and heuristics. The average performance of the proposed model is not outperforming the state of the art RL approaches. However, we observed that the proposed method is scalable in that the performance on the large-scale problem does not degrade compared to other approaches. We compare the performance by problem instance size (eil51 ~ tsp225 Vs. pr226 ~ pch442). Our model showed 1.1% reduction in the optimality gap (15.4% → 14.3%). However, the other RL-based approaches show much higher increases in the optimality gap (Drori et al. (2020): 3.1% → 10.0%, GPN: 14.1% → 33.1%, Drori et al. (2020): 4.7% → 10.6%).
Table 8: Performance comparison on TSP library

<table>
<thead>
<tr>
<th>City</th>
<th>Opt.</th>
<th>RL Approx.</th>
<th>Drori et al. [2020]</th>
<th>GPN</th>
<th>S2V-DQN</th>
<th>Farthest</th>
<th>2-opt</th>
<th>Nearest</th>
</tr>
</thead>
<tbody>
<tr>
<td>eil51</td>
<td>426</td>
<td>471</td>
<td>439</td>
<td>485</td>
<td>439</td>
<td>448</td>
<td>452</td>
<td>514</td>
</tr>
<tr>
<td>berlin52</td>
<td>7542</td>
<td>8269</td>
<td>7681</td>
<td>8795</td>
<td>7734</td>
<td>8121</td>
<td>7778</td>
<td>8981</td>
</tr>
<tr>
<td>st70</td>
<td>675</td>
<td>770</td>
<td>684</td>
<td>701</td>
<td>685</td>
<td>729</td>
<td>701</td>
<td>806</td>
</tr>
<tr>
<td>eil76</td>
<td>538</td>
<td>648</td>
<td>555</td>
<td>591</td>
<td>558</td>
<td>583</td>
<td>597</td>
<td>712</td>
</tr>
<tr>
<td>pr76</td>
<td>108159</td>
<td>135145</td>
<td>112699</td>
<td>118032</td>
<td>111141</td>
<td>119649</td>
<td>125276</td>
<td>153462</td>
</tr>
<tr>
<td>rat99</td>
<td>1211</td>
<td>1368</td>
<td>1268</td>
<td>1472</td>
<td>1250</td>
<td>1319</td>
<td>1351</td>
<td>1565</td>
</tr>
<tr>
<td>kroA100</td>
<td>21282</td>
<td>25282</td>
<td>21452</td>
<td>24806</td>
<td>22335</td>
<td>23374</td>
<td>23306</td>
<td>26856</td>
</tr>
<tr>
<td>kroB100</td>
<td>22141</td>
<td>25939</td>
<td>22488</td>
<td>24369</td>
<td>22548</td>
<td>24035</td>
<td>23129</td>
<td>29155</td>
</tr>
<tr>
<td>kroC100</td>
<td>21294</td>
<td>23397</td>
<td>21555</td>
<td>23494</td>
<td>21886</td>
<td>22361</td>
<td>22754</td>
<td>26950</td>
</tr>
<tr>
<td>kroD100</td>
<td>22068</td>
<td>26326</td>
<td>22267</td>
<td>23467</td>
<td>22820</td>
<td>23604</td>
<td>25325</td>
<td>27587</td>
</tr>
<tr>
<td>kroE100</td>
<td>7910</td>
<td>9237</td>
<td>8243</td>
<td>8844</td>
<td>8305</td>
<td>8652</td>
<td>8832</td>
<td>9941</td>
</tr>
<tr>
<td>rd100</td>
<td>629</td>
<td>753</td>
<td>650</td>
<td>704</td>
<td>667</td>
<td>687</td>
<td>694</td>
<td>825</td>
</tr>
<tr>
<td>lin105</td>
<td>14379</td>
<td>17173</td>
<td>14571</td>
<td>15795</td>
<td>14895</td>
<td>15196</td>
<td>16184</td>
<td>20363</td>
</tr>
<tr>
<td>pr124</td>
<td>59030</td>
<td>66554</td>
<td>59729</td>
<td>67901</td>
<td>61101</td>
<td>61645</td>
<td>61595</td>
<td>69299</td>
</tr>
<tr>
<td>bier127</td>
<td>118282</td>
<td>125233</td>
<td>120672</td>
<td>134089</td>
<td>123371</td>
<td>127795</td>
<td>136058</td>
<td>129346</td>
</tr>
<tr>
<td>ch130</td>
<td>6110</td>
<td>6987</td>
<td>6208</td>
<td>6457</td>
<td>6361</td>
<td>6655</td>
<td>6667</td>
<td>7578</td>
</tr>
<tr>
<td>pr136</td>
<td>96772</td>
<td>111563</td>
<td>98957</td>
<td>110790</td>
<td>100185</td>
<td>104687</td>
<td>103731</td>
<td>120778</td>
</tr>
<tr>
<td>pr144</td>
<td>58537</td>
<td>59197</td>
<td>60492</td>
<td>67211</td>
<td>59836</td>
<td>60259</td>
<td>62385</td>
<td>61651</td>
</tr>
<tr>
<td>ch150</td>
<td>6528</td>
<td>7264</td>
<td>6729</td>
<td>704</td>
<td>693</td>
<td>6866</td>
<td>7439</td>
<td>8195</td>
</tr>
<tr>
<td>kroA150</td>
<td>26528</td>
<td>32492</td>
<td>27419</td>
<td>30260</td>
<td>28076</td>
<td>28789</td>
<td>28313</td>
<td>33610</td>
</tr>
<tr>
<td>kroB150</td>
<td>26130</td>
<td>32438</td>
<td>27165</td>
<td>31914</td>
<td>26963</td>
<td>28156</td>
<td>28603</td>
<td>32825</td>
</tr>
<tr>
<td>u159</td>
<td>42080</td>
<td>51542</td>
<td>43687</td>
<td>52642</td>
<td>45620</td>
<td>46842</td>
<td>42976</td>
<td>53637</td>
</tr>
<tr>
<td>rat195</td>
<td>2323</td>
<td>2774</td>
<td>2384</td>
<td>2686</td>
<td>2567</td>
<td>2620</td>
<td>2569</td>
<td>2762</td>
</tr>
<tr>
<td>d198</td>
<td>15780</td>
<td>16217</td>
<td>17754</td>
<td>19249</td>
<td>16855</td>
<td>16161</td>
<td>16705</td>
<td>18830</td>
</tr>
<tr>
<td>kroA200</td>
<td>29368</td>
<td>34160</td>
<td>30553</td>
<td>34315</td>
<td>30732</td>
<td>31450</td>
<td>32378</td>
<td>35798</td>
</tr>
<tr>
<td>kroB200</td>
<td>29437</td>
<td>35848</td>
<td>30381</td>
<td>33854</td>
<td>31910</td>
<td>31656</td>
<td>32853</td>
<td>36982</td>
</tr>
<tr>
<td>ts225</td>
<td>126643</td>
<td>141754</td>
<td>130493</td>
<td>147092</td>
<td>140088</td>
<td>140625</td>
<td>143197</td>
<td>152494</td>
</tr>
<tr>
<td>tsp225</td>
<td>3916</td>
<td>4536</td>
<td>4091</td>
<td>4988</td>
<td>4219</td>
<td>4233</td>
<td>4046</td>
<td>4746</td>
</tr>
<tr>
<td>pr226</td>
<td>80369</td>
<td>83861</td>
<td>86438</td>
<td>85186</td>
<td>82869</td>
<td>84133</td>
<td>85306</td>
<td>94390</td>
</tr>
<tr>
<td>gil262</td>
<td>2378</td>
<td>2889</td>
<td>2523</td>
<td>5554</td>
<td>2539</td>
<td>2638</td>
<td>2630</td>
<td>3218</td>
</tr>
<tr>
<td>pr264</td>
<td>49135</td>
<td>51836</td>
<td>52838</td>
<td>67588</td>
<td>53790</td>
<td>54934</td>
<td>58115</td>
<td>58634</td>
</tr>
<tr>
<td>a280</td>
<td>2579</td>
<td>3059</td>
<td>2742</td>
<td>3019</td>
<td>3007</td>
<td>3011</td>
<td>2775</td>
<td>3311</td>
</tr>
<tr>
<td>pr299</td>
<td>48191</td>
<td>59257</td>
<td>53371</td>
<td>68011</td>
<td>55413</td>
<td>52110</td>
<td>52058</td>
<td>61252</td>
</tr>
<tr>
<td>lin318</td>
<td>42029</td>
<td>49167</td>
<td>45115</td>
<td>47854</td>
<td>45420</td>
<td>45930</td>
<td>45945</td>
<td>54034</td>
</tr>
<tr>
<td>rd400</td>
<td>15281</td>
<td>16631</td>
<td>16730</td>
<td>17564</td>
<td>16850</td>
<td>16864</td>
<td>16685</td>
<td>19168</td>
</tr>
<tr>
<td>fl147</td>
<td>11861</td>
<td>12768</td>
<td>13300</td>
<td>14684</td>
<td>12535</td>
<td>12589</td>
<td>12879</td>
<td>15288</td>
</tr>
<tr>
<td>pr439</td>
<td>107217</td>
<td>124333</td>
<td>126849</td>
<td>137341</td>
<td>122468</td>
<td>122899</td>
<td>111819</td>
<td>131258</td>
</tr>
<tr>
<td>pch442</td>
<td>50778</td>
<td>61757</td>
<td>55750</td>
<td>58352</td>
<td>59241</td>
<td>57149</td>
<td>57684</td>
<td>60242</td>
</tr>
</tbody>
</table>

Gap 0 15.1 4.9 19.0 6.2 8.4 9.1 24.7

A.3.2 mTSP

Table 9 shows results from mTSPLib solving MinMax TSP. Our model shows 3% longer results than ORT. However, computational time is, on average, about 30 seconds faster, which is about 45%. In addition, we calculate difference of the optimality gap by $N_v = 2, 3, 5, 7$. Our model shows better performance as $N_v$ increases ($N_v = 2$ case: 9.8%, $N_v = 3$ case: 8.9%, $N_v = 5$ case: 2.8%, $N_v = 7$ case: -7.5%)
Table 10 shows results from CVRPLib (Uchoa et al., 2017). Those values are the total traveling distance. Although our model is trained to minimize makespan (MinMax), our model shows better performance than AM in the total distance (MinSum).
Table 10: Performance comparison on CVRP library

<table>
<thead>
<tr>
<th>Instance</th>
<th>Opt</th>
<th>OR-Tools</th>
<th>AM (N=1280)</th>
<th>AM (N=5000)</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-n101-k25</td>
<td>27591</td>
<td>29405</td>
<td>39437</td>
<td>37702</td>
<td>39133</td>
</tr>
<tr>
<td>X-n106-k14</td>
<td>26362</td>
<td>27343</td>
<td>28320</td>
<td>28473</td>
<td>27940</td>
</tr>
<tr>
<td>X-n110-k13</td>
<td>14971</td>
<td>16149</td>
<td>15627</td>
<td>15443</td>
<td>17388</td>
</tr>
<tr>
<td>X-n115-k10</td>
<td>12747</td>
<td>13320</td>
<td>13917</td>
<td>13745</td>
<td>19292</td>
</tr>
<tr>
<td>X-n120-k6</td>
<td>13332</td>
<td>14242</td>
<td>14056</td>
<td>13937</td>
<td>16589</td>
</tr>
<tr>
<td>X-n125-k30</td>
<td>55539</td>
<td>58665</td>
<td>75681</td>
<td>75067</td>
<td>69919</td>
</tr>
<tr>
<td>X-n129-k18</td>
<td>28940</td>
<td>31361</td>
<td>30399</td>
<td>30176</td>
<td>36649</td>
</tr>
<tr>
<td>X-n134-k13</td>
<td>10916</td>
<td>13275</td>
<td>13795</td>
<td>13619</td>
<td>14800</td>
</tr>
<tr>
<td>X-n139-k10</td>
<td>13590</td>
<td>15223</td>
<td>14293</td>
<td>14215</td>
<td>16368</td>
</tr>
<tr>
<td>X-n143-k7</td>
<td>15700</td>
<td>17470</td>
<td>17414</td>
<td>17397</td>
<td>23548</td>
</tr>
<tr>
<td>X-n148-k46</td>
<td>43448</td>
<td>46836</td>
<td>79611</td>
<td>79514</td>
<td>62240</td>
</tr>
<tr>
<td>X-n153-k22</td>
<td>21220</td>
<td>22919</td>
<td>38423</td>
<td>37938</td>
<td>33079</td>
</tr>
<tr>
<td>X-n157-k13</td>
<td>16876</td>
<td>17309</td>
<td>21702</td>
<td>21330</td>
<td>19702</td>
</tr>
<tr>
<td>X-n162-k11</td>
<td>14138</td>
<td>15030</td>
<td>15108</td>
<td>15085</td>
<td>19491</td>
</tr>
<tr>
<td>X-n167-k10</td>
<td>20557</td>
<td>22477</td>
<td>22365</td>
<td>22285</td>
<td>25676</td>
</tr>
<tr>
<td>X-n172-k51</td>
<td>45607</td>
<td>50505</td>
<td>86186</td>
<td>87809</td>
<td>63191</td>
</tr>
<tr>
<td>X-n176-k26</td>
<td>47812</td>
<td>52111</td>
<td>58107</td>
<td>58178</td>
<td>66997</td>
</tr>
<tr>
<td>X-n181-k23</td>
<td>25569</td>
<td>26321</td>
<td>27828</td>
<td>27520</td>
<td>27220</td>
</tr>
<tr>
<td>X-n186-k15</td>
<td>24145</td>
<td>26017</td>
<td>25917</td>
<td>25757</td>
<td>30086</td>
</tr>
<tr>
<td>X-n190-k8</td>
<td>16980</td>
<td>18088</td>
<td>37820</td>
<td>36383</td>
<td>18651</td>
</tr>
<tr>
<td>X-n195-k51</td>
<td>44225</td>
<td>50311</td>
<td>79594</td>
<td>79276</td>
<td>66957</td>
</tr>
<tr>
<td>X-n200-k36</td>
<td>58578</td>
<td>61009</td>
<td>78679</td>
<td>76477</td>
<td>68502</td>
</tr>
</tbody>
</table>

Gap        -       8.06    32.97    31.62    30.22

A.4 Graphical solutions example

This section provides the visual results for several vehicle routing problems solved by the trained policy.

A.4.1 Training results on mCVRP

Figure 3: \(N_c = 100, N_v = 5\) case of mCVRP

Figure 3 depicts the evaluation case \((N_c = 100, N_v = 5\) case of mCVRP) during training process. This test environment is designed to mimic the search and rescue problems, operating multiple drones to search victims distributed over a map.

Green grids are for refueling nodes, blue circles are for vehicles (color becomes black as fuel capacity diminishes), and grey grids are for unvisited customers. If a customer is visited by a vehicle, the color becomes red or white; the red grids indicate the victims, while the white indicates no victim in that cell. These red grids do not have any effect on performance measure, because visiting all customers (grey grids) is top-priority.
At the very beginning of the training process (figure 3 (b)), the trained model cannot have vehicles search over the cells and make them die due to lack of fuel. At the end of the training process (figure 3 (c)), however, all customers are visited by the vehicles.

A.4.2 MVRP (MTSP)

Figure 4 depicts an example solutions of mTSP using CPLEX, OR-Tools, and RL. The RL-based solver is trained under mCVRP environment but we solve mTSP with the trained model.

A.4.3 TSP

Figure 5 depicts an example solutions of TSP using CPLEX, OR-Tools, and RL. The RL-based solver is trained under mCVRP environment but we solve TSP with the trained model.

A.5 Network Architecture and Hyperparameters

Table 11 is summary of hyperparameters used in training process.

<table>
<thead>
<tr>
<th>Parameters (for PPO update)</th>
<th>Value</th>
<th>Parameters (for GRLTS)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimizer</td>
<td>Adam</td>
<td>MLP units</td>
<td>(32, 32)</td>
</tr>
<tr>
<td>betas (\beta_1, \beta_2)</td>
<td>(0.9, 0.999)</td>
<td>Neuron initialization</td>
<td>Kaiming normal</td>
</tr>
<tr>
<td>Learning rate (\eta_{\text{policy}})</td>
<td>(2 \times 10^{-4})</td>
<td>Activation fn.</td>
<td>ReLU</td>
</tr>
<tr>
<td>Learning rate (\eta_{\text{Q-network}})</td>
<td>(2 \times 10^{-4})</td>
<td>Node embedding epoch (N^{\text{tep}})</td>
<td>5</td>
</tr>
<tr>
<td>gamma (\gamma)</td>
<td>0.99</td>
<td>Node feature dim.</td>
<td>5</td>
</tr>
<tr>
<td>clip ratio (\epsilon)</td>
<td>0.2</td>
<td>Edge feature dim.</td>
<td>10</td>
</tr>
<tr>
<td>value function coeffi.</td>
<td>0.5</td>
<td>Max. decision epoch (3 \times N_c)</td>
<td></td>
</tr>
<tr>
<td>entropy coeffi.</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPO update epoch (K)</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training epoch</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A.6 TRAINING PERFORMANCE

Figure (6) shows how the trained policy perform on the thee set of validation problems while the policy is being trained by the random mCVRP inc stances. The three validation problems are 1) 100 customers covered by 5 vehicles, 2) 100 customers by 10 vehicles and 3) 400 customers covered by 20 vehicles. The first row show the performance on the random training instances while the second, third, and fourth row show the performances on the three validation problems.

As training processes progress, the trained model gradually becomes more efficient, that is, a fleet of vehicles visits more customers faster. After about 200 training epochs, the trained model converges, but there are some randomness due to the nature of policy gradient structure. Although the makespan curves in the second column of training case seems relatively constant (the first row row), the level of difficulty is increasing as curriculum becomes harder (The number of refueling is also increasing as curriculum becomes harder).

Comparison of case #1 and case #2 give us two interpretations: 1) case #1 shows the model is trained to visit all customers and 2) in case #2, the model achieves consistent performance as training iterations progress in spite of same visit ratio as 1 around 100 epoch and 350 epoch. Besides, the makespans of both training epochs are very close. Case #2 is relatively easier environment than case #1 in that more vehicles are deployed to serve same number of customers.

![Performance curves](image)

Figure 6: Performance curves of training (1st row) and testing (2nd, 3rd and 4th rows). Note that we plot a reward curve and cover ratio curves in 1st column for training and testing, respectively.

A.7 PSEUDO CODE

Simulation: Problem Instance generation We generate random mCVRP instances and use these training instances to train the proposed model. For training, we randomly select the number of vehicles, customers, and refueling stations. We randomly assign these entities over the grid world with the distance between each grid cell being 1. When transforming the snapshot of the mCVRP to a graph, we use the Manhattan distance between the two cells as the edge weight. In addition, the speed of each vehicle is set to be 1; thus, each vehicle moves over one cell during a single time-tic. Although the world is represented as discrete grid, the mCVRP problem’s state transition is event-based. Whenever the agent reaches the assigned customer, the event occurs, and all the edge distance is updated.

Initially, this grid world is developed to use the trained policy for search and rescue problems, seeking distributed victims over the grid-world. Because each vehicle (drone) can search a specific zone within a certain amount of time, using a cell-based grid world is a reasonable choice.
Once the policy is trained, it can be used for both a discrete world or a continuous world because these environments can be transformed into a graph without any difference. We validated the proposed method using CVRP and TSP instances defined over the continuous state to compare the performance of the proposed model to other deep baseline line models. Other deep baseline line models all use the continuous state.

To boost training, we vary the difficulty level of random mCVRP instances during training process (see Algorithm 2). For the curriculum instance generation in line 9, we compute $N_{curr}$, the number of customers which are assumed to be visited already, as $N_{curr} = N_c \times (1 - 2 \times \text{currLevel}/10)$. Then, we choose $N_{curr}$ tasks randomly and mark as visited and distribute $N_{curr}$ for all agents’ visit number $q_t$.

**Algorithm 2 Training instance generation**

1: Generate random problem instances $N_v, N_c, N_r, x^0_v, x^0_r,$
2: $\text{currLevel} \leftarrow 0$
3: for $i_{\text{training}} = 1, \ldots, N_{\text{training}}$ do
4: for each 100 training iterations do
5: $\text{currLevel} \leftarrow \text{currLevel} + 1$
6: end for
7: Generate a random number $p_{curr} \sim \text{U}[0, 1]$
8: if $p_{curr} \geq 0.5$ then
9: Curriculum instance generation
10: else
11: Non-curriculum instance generation (followed by line 1)
12: end if
13: end for

**Simulation: Episode generation**  As a medium for the interaction of the two components such as environment and the GRLTS, simulation conducts the agents’ actions and stores the state transitions which are used to update the GRLTS. Starting from the problem instance generation (Algorithm 2), simulation assigns the agents’ assignment as $a^t_{agt}$ using the action probability $\pi(\cdot | o^t_{agt})$ computed from the proposed GRLTS. Simulation computes the agent $agt$’s transition time $\Delta_{agt}$ as follow:

$$\Delta_{agt} = \frac{\text{Dist}(x^t_{agt}, a^t_{agt})}{v_{agt}} = \frac{\text{Dist}(x^t_{agt}, a^t_{agt})}{v_{agt}}$$

where velocity of agents are constant as 1 and distance between $i$ and $j$ $\text{Dist}(i,j)$ can be the Manhattan distance or Euclidean distance.
**Algorithm 3** Episode generation

1: Initialize problem instances (based on Algorithm 2)
2: `termination, done ← False, t ← 0, memory ← ∅`
3: `idleAgts ← {1, 2, ..., Nv}, actingAgts ← ∅`
4: **while** `termination` **do**
5:   **for** `agt ∈ idleAgts` **do**
6:     Send an observation $\Pi_{agt} o_t^{agt}$ to a GRLTS
7:   **end for**
8:   Choose an action $a_t^{agt} \sim \pi(\cdot|o_t^{agt})$ where an action probability is from GRLTS
9:   Assign the action $a_t^{agt}$ to the agent
10: **end while**
11: Compute and update transition time $\Delta_{agt}$ for $agt \in idleAgts$ (Equation (13))
12: `actingAgts ← {agt | $\Delta_{agt} = \min_{agt} \Delta_{agt}$}` for all agents
13: `t ← t + \min_{agt} \Delta_{agt}`
14: **for** `agt ∈ actingAgts` **do**
15:   Perform the recently assigned action $a_t^{agt}$ assigned by $\pi(\cdot|o_t^{agt})$
16:   Observe the observation $o_t^{agt}$, reward $r_t^{agt}$ and `done`
17:   Append a state transition tuple $<t', o_t^{agt}, a_t^{agt}, r_t^{agt}, o_t^{agt}>$ to `memory`
18: **end for**
19: **if** `done` **then**
20:   `termination ← True`
21: **else**
22:   `idleAgts ← actingAgts`
23: **end if**

**RL networks: GRLTS** Most of algorithms in the RL model is explained the paper so that we focus on the interaction between simulation and GRLTS networks and internal message exchanges.

**Algorithm 4** GRLTS

1: Receive an observation $\Pi_{agt} o_t^{agt}$ at time $t$ from simulation (line 6 of Algorithm 3)
2: Generate $G_t^{(0)}$ with an input of observations $\Pi_t o_t^{agt}$
3: **for** node embedding iteration $\tau = 1, ..., N_{hop}$ **do**
4:   Compute connectivity $C_{ij}$ for all nodes $i, j$:
5:     For agent nodes $n$, $C_{nj} = f^n$, $\forall j \in V_T \cup V_R$
6:       For refueling nodes $r$, $C_{rj} = F$, $\forall j \in V$
7:       For task nodes $t$, $C_{ij} = 5$, $\forall j \in V$
8:     Store the node embedding vector $h_t^{agt}$ at $memory_i$
9:     Update the connected edges following Equation (3)
10: Aggregate the incoming edge featured at each nodes following Equation (4)
11: Update the node feature with the node updator function $\phi_i$
12: **end for**
13: Compute $Q(h_t^{agt}, a)$ for agent $i$ where $a \in A(\alpha_i) = \{v \in \{V_R \cup V_T\}|C_{iv} = 1\}$
14: Compute $\pi(a|\alpha_i) = \frac{\exp(Q(h_t^{agt}, a))}{\sum_{a \in A(\alpha_i)} \exp(Q(h_t^{agt}, a))}$
15: Send the action probability $\pi(a|\alpha_i)$ to simulation