Australasian Transport Research Forum 2024 Proceedings
27-29 November, Melbourne, Australia
Publication website: https://australasiantransportresearchforum.org.au/

Optimizing the initialization of fleet deployment for mobility-on-demand systems: A deep reinforcement learning approach

Zhengfei Zheng[1], Peibo Duan[2], Zheng Liang[1], Hai Yang[1]

[1]Department of Civil and Environmental Engineering, the Hong Kong University of Science and Technology, Hong Kong, China

[2]Software College, Northeastern University of China, Shenyang, China

Email for correspondence (presenting author): zzhengak@connect.ust.hk

1 Introduction

Mobility-on-Demand (MoD) systems, such as taxi and ridesharing services, have transformed urban transportation by providing convenient and reliable options for commuters, tourists, and residents. While much research has focused on optimizing vehicle dispatching to maximize fleet revenue over time, the challenge of determining the optimal number of vehicles to deploy and their initial locations at the start of a period remains unresolved. This paper addresses this gap by proposing a deep reinforcement learning-based model for initial fleet deployment. Fleet size management and vehicle deployment are crucial for several reasons. An excessive supply of vehicles can lead to financial losses for ridesharing companies, increased idle time for drivers, and reduced revenue. Additionally, an oversized fleet can cause traffic congestion, energy wastage, and environmental pollution. Conversely, an insufficient vehicle supply cannot effectively serve passengers, highlighting the importance of deploying an appropriate number of vehicles. The starting locations of these vehicles are also critical to avoid competition among vehicles in the same area and ensure efficient service across the city.

In this work, we proposed a new methodology to address such a issue.Our model comprises two components: a Deep Reinforcement Learning-based State Valuing (DRLSV) model and a Gradient Ascent-based Deploying (GAD) algorithm. The DRLSV model learns a state-value network that approximates the expected total revenue a vehicle will earn from the current time to the end of the period, considering passenger demand fluctuations, MoD dispatching policies, and vehicle interactions. The GAD algorithm deploys vehicles based on the state-value network to maximize the fleet's total expected revenue over the long term. GAD balances serving more passengers by deploying more vehicles and reducing costs by deploying fewer vehicles, resulting in more efficient fleet deployment.

2 Problem statement

To ensure clarity and consistency, this section provides definitions of the notations and terminologies used throughout this paper. Initially, we partition the city into uniform grids to obtain an $I \times J$ grid map, which we denote as G. Each grid is represented as $g_w \in G, w = 1,2,3 \cdots, IJ$. Based on the grid map, we formally define requests from passengers and the problem investigated in this paper.

Definition 1 (request): In this paper, we define a request as a quadruple represented by $r_w = (r_w.\tau, r_w.g_o, r_w.g_e, r_w.rv)$. This notation indicates that a passenger sent a ride request at timestamp $r_w.\tau$, with the pick-up location in grid $r_w.g_o$ and the destination in grid $r_w.g_e$. After the request has been completed, a reward of $r_w.rv$ was paid to the driver. Here, r_w denotes the passengers' request in grid g_w and the abbreviation "rv" denotes the reward a vehicle will receive if it is designated to serve this specific request. However, it is important to

note that the waiting time for the request cannot exceed a threshold ϑ. If the waiting time exceeds this threshold, the passenger may cancel the request or receive compensation from the MoD company.

Our objective is to learn the optimal initialization of fleet deployment for each episode, which involves historical requests from passengers and the dispatching policy adopted by the MoD system to continually dispatch available vehicles to passengers. Specifically, we aim to determine how many vehicles should be deployed throughout the city and where these vehicles should start at the beginning of the episode. The ultimate goal is to maximize the expected total revenue of the fleet while ensuring reasonable deployment costs.

3 Methodologies

This section details the methodologies employed in our paper. As discussed in Section 1, there is currently no existing metric to measure the quality of fleet deployment since the goal is to maximize long-term revenues instead of immediate ones, which is common in most current research. We assume that each vehicle can only serve one passenger at a time. Even if the fleet initially covers many requests, the requests that the vehicle can reach will change after dropping off the passenger at their destination, and so will the requests served by the fleet. Therefore, our first step is to design a metric that can measure the quality of initial fleet deployment. By optimizing this metric, we can learn the optimal fleet deployment for the MoD system in each episode.

3.1 Evaluation on the initialization of fleet deployment

If we can estimate the anticipated total revenue V_w of a vehicle v_w in a given episode Ep based on its initial location, we can determine the optimal fleet deployment by utilizing $Eq.(1)$.

$$max_D \Phi(D) = \sum_{w=1,\cdots,n} V_w - cn \tag{1}$$

In $Eq.(1)$ the first term represents the anticipated total revenue of the fleet throughout the episode, while the second term pertains to the total operating cost of all vehicles. The objective of the problem is to optimize a balance between revenue and cost. The fleet deployment D that we aim to optimize has an impact on the long-term revenue of each vehicle in the episode, i.e., $V_1, V_2, \cdots, V_n$, and the total cost, where $n = |D|$ represents the total number of vehicles deployed.

3.2 Markov decision process

A reinforcement learning model is typically formulated as a Markov Decision Process (MDP). This framework outlines how an agent interacts with its environment over time. Throughout this interaction, the agent learns an optimal policy aimed at maximizing the long-term cumulative reward. An MDP consists of six components, denoted by a tuple as $\{S, A, TR, R, \pi, \gamma\}$ where the state space S defines the possible states of the agent, the action space A contains the actions the agent can take, $TR: S \times A \times S \to \mathcal{R}^+$ is the transition probability that describes the likelihood of an agent taking action a_t under state S_t and transitioning to the next state S_{t+1}. The immediate reward function $R: S \times A \times S \to \mathcal{R}$ indicates the immediate reward received by the agent after taking an action under a state and transitioning to the next state. The policy $\pi: S \times A \to \mathcal{R}^+$ describes the probability of the agent taking each action under each state, while $\gamma \in (0,1]$ is a discount parameter.

3.3 Deep reinforcement learning based state valuing

In DRLSV, we treat each vehicle as an individual agent, and its state is defined based on **Definition 2**. According to this definition, vehicles that are located in the same grid at the same time share a common state. This approach is reasonable since the MoD system treats all vehicles equally when dispatching them to passengers, and only distinguishes them by their regional locations.

Definition 2 (state): The state of a vehicle is represented as a triplet $S_{w,t} = \{D_t, g_w, t\}$. Here, D_t describes the distribution of available vehicles scattered throughout the city at time period t, and it can be represented as a matrix. g_w represents the grid in which the vehicle is currently located and can be represented as a one-hot vector. t denotes the current time step and can also be represented as a one-hot vector.

By obtaining the state-value function $V(S_{w,t})$, we can infer the optimal fleet deployment using $Eq.(2)$ as following, which is a simple variant of $Eq.(1)$. In $Eq.(2)$, D has the same meaning as in $Eq.(1)$, i.e., it represents the distribution of the fleet at time $t = 0$. D_{gw} represents the entry in D corresponding to grid g_w, and it describes the number of vehicles that have been initially deployed in g_w. $S_{w,0} = \{D, g_w, 0\}$ is the state of grid g_w at time $t = 0$.

$$\max_{D} \Phi(D) = \sum_{w=1}^{I \times J} D_{g_w} \times V(S_{w,0}) - c \times |D| \tag{2}$$

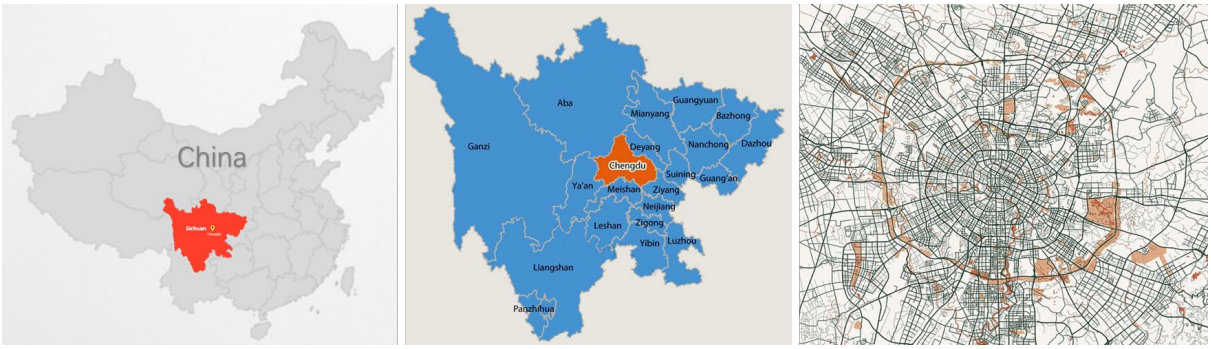
3.4 Deploying fleet

Solving $Eq.(2)$ is not trivial because for a vehicle located in a specific grid g_w, if any entry of D changes, such D_j, its expected total revenue in the episode will also change, regardless of either $j = w$ or $j \neq w$. In this subsection, we propose a Gradient Ascent based Vehicle Deploying algorithm, i.e., GAD, to solve $Eq.(2)$.

4 Evaluation on the effectiveness and efficiency of GAD

This paper's evaluation was conducted on a study site located in Chengdu, China, primarily covering the urban area as shown in **Figure 1**. Chengdu, the capital city of southwestern China's Sichuan province, is sitting on the Chengdu Plain at the western edge of the Sichuan Basin, it spans 192 km east to west and 166 km north to south, covering 12,390 km². With a population of about 21 million (2020 census), Chengdu is the fourth most populous city in China and a traditional hub of Western China.

Figure 4: Location, geographical map and road network of Chengdu, China



We trained the DQN model using the data spanning two weeks (from 7th November to 20th November 2016). We trained the DQN model for previously identified three distinct periods with respect to weekdays and weekends. We have the following results:

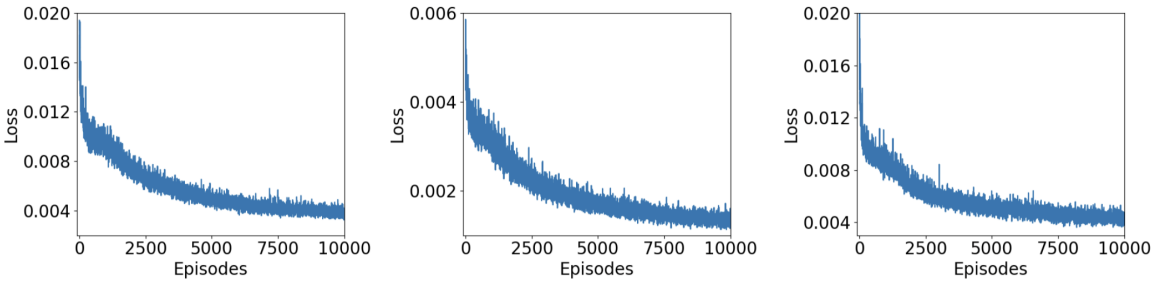
- **Figure 2** shows the loss while training the DQN model. The value of loss function decreases over time as the model learns to make better decisions based on the input states.

5 Conclusions

In this paper, we propose a Deep Reinforcement Learning-based model (DRLD) to learn the optimal fleet deployment for a city in different episodes, with the aim of maximizing the expected total revenue of the fleet. We conduct extensive experiments based on real-world data from Chengdu, China to validate the superiority of our proposed model compared to

baseline approaches. Our model helps reduce operational costs and idle times by preventing oversupply during off-peak hours..

Figure 2: Loss function of model training in the periods of workdays



(a) 8 am-10 am (b) 11 am-1 pm (c) 6 pm-8 pm

Acknowledgments

We would like to acknowledge a grant from RGC Theme-based Research Scheme (TRS) T41-603/20R, and a research grant (project HKUST16207222) from the Hong Kong Research Grants Council.