ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/361549456

Vehicle Routing Problem Using Reinforcement Learning: Recent Advancements

Chapter · June 2022

DOI: 10.1007/978-981-19-0840-8_20

CITATIONS
3

READS
760

3 authors:

Syed Mohib Raza
Newcastle University
2 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Mohammad Sajid
Aligarh Muslim University
43 PUBLICATIONS **542** CITATIONS
SEE PROFILE

Rahul Kumar
178 PUBLICATIONS **3,141** CITATIONS
SEE PROFILE

All content following this page was uploaded by Mohammad Sajid on 02 November 2022.

The user has requested enhancement of the downloaded file.

Vehicle Routing Problem Using Reinforcement Learning: Recent Advancements

Syed Mohib Raza, Mohammad Sajid, and Jagendra Singh

Abstract In the realization of smart cities, the most important component is the smart logistics in which the vehicle routing problem (VRP) plays a significant role. The VRP has been proven to be NP-hard, and this combinatorial optimization problem requires efficiently serving the demands of geographically distributed customers using vehicles with limited capacities in order to optimize travel time or traveled distance. In general, VRP and its variants have been solved using OR-Tools, meta-heuristic as well as local search algorithms. However, these methods need high computational efforts and may offer poor-quality solutions in case of large problem sizes. The deep learning models can also be employed to solve the VRP. This paper explores the recent advancements in solving VRP using reinforcement learning (RL). The paper surveys the different RL approaches used to solve VRP and its variants. The paper also presents the issues and challenges that emerged with the use of RL to solve the VRP variants.

Keywords Reinforcement learning · Vehicle routing problem · Pointer network · Markov decision process

1 Introduction

In 1959, Dantzing and Ramser (1959) introduced the vehicle routing problem (VRP) as the truck dispatching problem [1]. The VRP is one of the most important research problems in the field of combinatorial optimization (CO) due to the requirement for the delivery of services to distributed consumers by millions of emerging and existing organizations. It has a holistic presence in the world, from its importance in logistics, supply chain, disaster management, data traffic management, and other

S. M. Raza · M. Sajid (✉)
Department of Computer Science, Aligarh Muslim University, Aligarh, India
e-mail: sajid.cst@gmail.com

J. Singh
School of Computer Science Engineering and Technology, Bennett University, Greater Noida, India
e-mail: jagendra.singh@bennett.edu.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

D. Gupta et al. (eds.), *Advanced Machine Intelligence and Signal Processing*, Lecture Notes in Electrical Engineering 858, https://doi.org/10.1007/978-981-19-0840-8_20

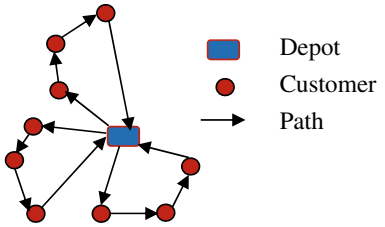

Fig. 1 VRP with nine customers and three routes

fields. VRP has been proved to be an NP-hard problem [2], and it becomes even more complex with other constraints like limited capacity (CVRP), split delivery (VRPSD), stochastic time (SVRP), last-mile delivery with drones (VRP-D), simultaneous pickup and delivery (VRPPD), multiple depots (MDVRP), time windows (VRPTW), and combinations of them [2, 3]. Vehicle routing problem (VRP) requires efficiently serving the demands of geographically distributed customers using vehicles with limited capacities in order to optimize travel time/distance while fulfilling the vehicle's capacity, flow, and other constraints [2–4]. A CVRP instance consisting of nine customers served by the three vehicle routes has been depicted in Fig. 1.

In general, VRP and other combinatorial optimization problems have been solved using exact methods (branch-&-bound, dynamic programming, Google OR-Tools), evolutionary algorithms (GA, ACO, firefly algorithms), heuristics (sweep algorithm, Fisher-and-Jaikumar algorithm, and Clarke-&-Wright saving algorithm,), local search algorithms (swap, exchange, 2-Opt*), and their combinations [2, 5–10]. However, the size of solution space increases exponentially with the increasing size of VRP instances (number of customers). Therefore, these methods need high computational efforts and may offer the poor-quality solutions in case of large problem instances. Recently, the scientific community is addressing combinatorial optimization problems using deep learning models including reinforcement learning (RL) [11–26]. RL is a learning in which the learner (agent) itself learns and discovers "what to do" in order to maximize the reward. The reinforcement learning (RL) works based on two important features, i.e., "*trial-and-error search*" and "delayed reward." To maximize the reward in RL, the learning agent needs to exploit the previously explored actions as well as it also requires exploring better actions in order to exploit them in the future [27].

This paper explores the recent advancements in solving VRP using reinforcement learning (RL) and other deep learning models. The paper surveys the different RL approaches used to solve VRP. Table 1 provides the references of research works

Table 1 Reviewed papers

Year	Number of papers reviewed	The list of papers
2018	1	[11]
2019	5	[12–16]
2020	10	[17–26]

surveyed in this paper [11–26]. This paper also surveys issues and challenges that emerged with the use of RL to solve the VRP variants.

The contents of this paper are organized as follows. The next section presents a gentle introduction to the vehicle routing problem along with its mathematical formulation. Section 3 puts the reinforcement learning and related concepts, whereas Sect. 4 presents the review research for VRP solved using RL-based approaches. Section 5 presents some major issues and challenges which need to be addressed while solving VRP using RL approaches. Lastly, the conclusions of the study are presented in Sect. 6.

2 Vehicle Routing Problem

The general VRP, i.e., capacitated vehicle routing problem (CVRP), can be described in the form of two-dimensional graph $G = (N, A)$. The set $N = \{0, 1, 2, \ldots. n\}$ represents the set of nodes in which node 0 is the central depot, and the remaining nodes (1 to n) are representing the customers. The location of each node i is given as $2-D$ geometric coordinates (x_i, y_i). The set A consists of edges to join nodes with each other and is given as $A = \{(i, j) : i, j \in N, i \neq j\}$, and every edge is characterized by the geometric distance d_{ij} between nodes $i, j \in N$. All customers also have some non-zero demands which are characterized by the set $Q = \{q_i : 1 \leq i \leq N, q_i > 0\}$. The customers are served by a fleet of K homogeneous vehicles which are given as $V = \{v_k : 1 \leq k \leq K\}$. All vehicles are available at depot node initially, and all of them have identical capacity C [4, 8].

The CVRP requires efficiently serving the demands of geographically distributed customers using vehicles with limited capacities in order to optimize travel time/distance while fulfilling the vehicle's capacity, flow, and integrity constraints. For mathematical formulation of the CVRP, assume a decision variable χ_{ijk} whose value may be either 0 or 1. The value 1 of χ_{ijk} indicates that vehicle v_k travels to node j after performing delivery service at node i. The objective of the CVRP is to minimize the sum of distances traveled by all vehicles which depends on the distance between two nodes and the decision variable. The mathematical formulation of CVRP is given as:

$$\text{Min} f_1 = \sum_{v_k \in V} \sum_{i \in N} \sum_{j \in N} \chi_{ijk} * d_{ij} \tag{1}$$

with the following constraints

$$\sum_{v_k \in V} \sum_{j \in N} \chi_{ijk} = 1, \quad \forall i \in N \tag{2}$$

$$\sum_{i \in N} \sum_{j \in N - \{0\}} \chi_{ijk} * q_j \leq H, \quad \forall v_k \in V \tag{3}$$

$$\sum_{j \in N} \chi_{0,j,k} = 1, \quad \forall v_k \in V \tag{4}$$

$$\sum_{i \in N} \chi_{ihk} - \sum_{j \in N} \chi_{hjk} = 0, \quad \forall h \in N - \{0\}, \forall v_k \in V \tag{5}$$

$$\sum_{i \in N} \chi_{i,0,k} = 1, \quad \forall v_k \in V \tag{6}$$

$$\chi_{ijk} \varepsilon \{0, 1\}, \quad \forall i, j \in N, \forall v_k \in V \tag{7}$$

The Eq. (1) minimizes the objective of the problem, i.e., total traveled distance. The constraint (2) ensures the service integrity, i.e., every customer is serviced once using only one vehicle. The constraint (2) ensures that every customer is serviced once using only one vehicle. The constrained (3) ensures the capacity constraints of vehicles, i.e., the sum of the demands assigned to a customer in a route must not exceed the vehicles' capacity. The remaining constraints (4), (5), (6) and (7) are flow and integrity constraints. The flow constraints (4), (5) and (6) ensure that every vehicle leaves the depot, departs from the visited customer, and finally returns to depot. Final, integrity constraint (7) ensures the value of decision must be either zero or one only [4, 8].

2.1 Unit-Square Vehicle Routing Problem

The unit-square VRP is a variant of CVRP which is primarily used for experimental purposes. In unit-square VRP, the locations of depot (central service point) and customers are randomly selected from the two-dimensional unit square, i.e., [0, 1] × [0, 1], and the demands of customers are also randomly chosen from the set {1, 2, 3, ……0.9}. It is also assumed that there is only one vehicle with limited capacity which is responsible for delivering the items. Initially, the vehicle loads the items, visits the customers for delivery, and then goes back to depot for refill if there is any need [11].

3 Reinforcement Learning and Pointer Networks

With the advent of IT, the reliability on data was inevitable, and there is need to utilize the available data in the best possible way. The "learning" ability is considered the most important milestone of human intelligence. It is this human phenomenon that was able to develop and merge with the machine because of the given increase in computing power and data. This learning phenomenon in machines is called machine learning which comes under the umbrella term "artificial intelligence" [25, 27]. In

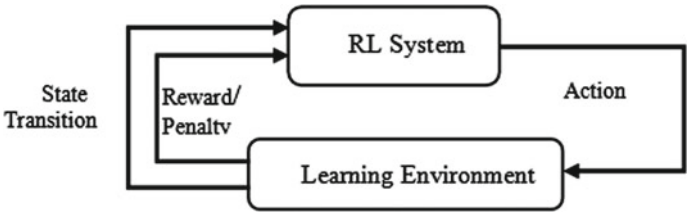

Fig. 2 Framework of reinforcement learning

the review research Sect. 4, many cited models were developed based on RL and pointer networks which are discussed in next sub-sections.

3.1 Reinforcement Learning

In reinforcement learning, the learner (agent) itself learns and discovers "what to do" in order to maximize the reward. To maximize the reward in RL, the learning agent needs to exploit the previously explored actions as well as it also requires to explore new actions in order to exploit them in the future [27]. A generic framework of reinforcement learning is shown in Fig. 2.

The objective of RL is to solve a Markov decision process (MDP) in which the reward is not known. The MDP is a mathematical discrete-time dynamic system model which is used for modeling decision-making in the context of partially random output as well as partial output controlled by the decision maker. The components of MDP are the state of the system at time $t(S_t)$, action taken at time t (A_t), and the reward at time t (R_t).

In RL algorithms, the learner can learn "on-policy" or "off-policy" in discrete/continuous state space using discrete/continuous actions.

3.2 Pointer Networks

The basic idea behind the pointer networks is the attention mechanism, and they can be assumed the simple extension of attention model. Because of being invariant to the length of the encoder sequence, it enables the model to apply to CO problems, where the source sequence is the one that determines the length of the output sequence. The pointer network in every decoding step generates a vector with attention weights corresponding to the length of the input sequence. These weights are used to compute the output sequence [28, 29] (Fig. 3).

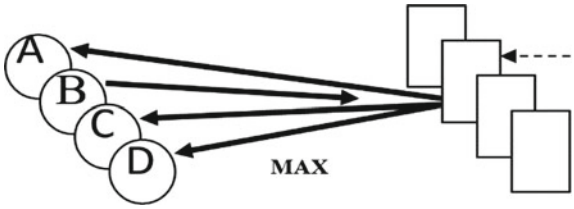

Fig. 3 Framework of pointer networks

4 Review Research

The following Table 2 summarizes our review research clearly and concisely. It presents different VRP types along with the different approaches taken by the respective researchers for solving those problems, indexed along the reference column.

Nalepa [20] puts an overview of the data-driven approaches which were applied in the context of smart delivery systems. All techniques are divided into three groups.

- Use of ML to solve VRP by tuning hyperparameters of existing algorithms.

Table 2 VRP using RL

Year	VRP type	Approach used	References
2018	VRP	Reinforcement learning (Iterative approach)	[11]
2020	Electric VRPTW	Graph embedding with PN architecture	[21]
2019	Online autonomous VRP	Structural graph-embedded PN [28] with Structure2Vec tool encoder	[13]
2020	Gen. VRP	Hybrid DRL	[22]
2020	Gen. VRP	Dynamic attention model and dynamic encoder–decoder architecture plus 2-opt local search	[23]
2020	Global routing problem (GRP)	DRL in a simulated environment	[24]
2019	Gen. VRP	RL with Markov decision process (MDP)	[14]
2020	Dynamic VRP with time windows (DVRPTW) and mixed types of customers	DRL (NN-based temporal difference learning) with AVF and DRLSA (Simulated annealing), MDP	[25]
2019	Warehouse VRP	MCTS-CNN	[15]
2020	Dynamic route planning with minimizing travel time	Deep Q-Network (DQN)	[26]
2019	Capacitated VRP (CVRP)	Iterative learning	[16]
2020	TSP and CVRP	Self-attended deep arch. as policy network	[19]

- Hybrid algorithms exploiting ML for solving the VRP.
- Use of data-driven ML algorithmic techniques for solving VRP.

Furthermore, [20] presents a brief overview of the applications of data-driven algorithms for smart delivery systems and also discusses the application areas for the presented algorithms.

Nazari et al. [11] proposed an end-to-end framework by training a heuristic single policy model for a broad range of problem instances of similar size. He added another layer of encoder/decoder attention mechanism resulting in optimized logic through RL. The proposed RL-based model does not require a distance matrix calculation and also does not need to be retrained for every new instance if the problems are generated from the training distribution making it more practical and well scalable with the increase in the problem size along with superior performance and competitive solution time. Kool et al. [12] proposed an attention model and used DRL to train the model with a simple baseline based on deterministic greedy rollout which outperformed the baseline solutions. Hao et al. [16] proposed learn to improve (L2I) approach which refines solution by learning with the help of an improvement operator, selected by an RL-based controller. The improvement operator is chosen from an RL-specific powerful operator's pool. He used the same settings as [11] for CVRP and also compared with [11] along with OR-Tools, LKH, and with Kool et al. [12] with the obj. to minimize cost, routing plan and without violating vehicle capacity.

Lin et al. [21] solved electric vehicle routing problem with time windows (EVRPTW) using developed attention model. The developed attention model combines pointer network (PN) and a graph embedding technique to parameterize a stochastic policy. The solutions generated for small instances were not optimal but were able to solve, earlier unsolvable large-size instances, providing feasible solutions to the problem. The graph embedding techniques were incorporated with the PN architecture which allows algorithm to synthesize the global as well as local information in order to solve the target problem.

Similar work was done by Yu [13] who transforms VRP to vehicle tour generation problem and proposes a structural graph-embedded PN with Structure2Vec tool incorporation to develop a distributed system for solving an online autonomous VRP called green logistics system. The transportation network and traffic data of Cologne, Germany, were adopted in the test case study; the map data of the city were from OpenStreetMap, and the real-time traffic speed of each road in the network is from [17]. To perform the case study, the capacity and battery charging configurations of Tesla Model S & X and Nissan LEAF are adopted. With the offline parameter training, minimal computation time was incurred, making the researched strategy promising for online VRP, but due to the impractical nature involved in the construction of a supervised training data set for NN, an alternate was devised in the form of using DRL mechanism to fine-tune the NN model parameters thereby showing the outstanding capability of the proposed strategy as compared to conventional mathematical programming-based strategies.

Zhao et al. [22] proposed a DRL model based on an actor which was in turn based on an attention mechanism for generating routing strategies and, adaptive

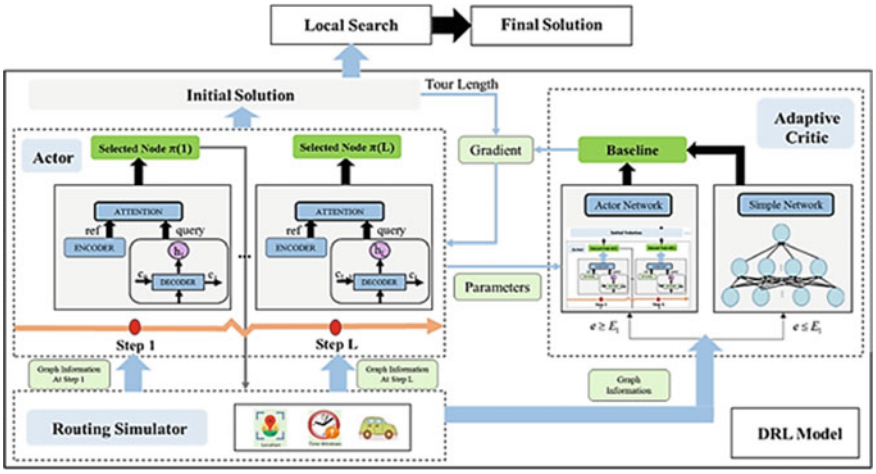

Fig. 4 Hybrid model architecture [22]

critic devised to change network structure along with a routing simulator. The devised model was combined with a local search in order to the quality of offered solutions. Figure 4 shows the proposed for DRL. The proposed model was tested on three datasets with 20, 50, and 100 customer points. The result demonstrated the excellent capability of the DRL model as compared to construction algorithms and previous DRL approaches. Combining the model with local search methods (Google OR-Tools and LNS) also yielded excellent solutions with superior generation speed, comparing to other initial solutions.

Peng et al. [23] presented a dynamic attention model consisting of dynamic encoder–decoder architecture which enables it to explore node features dynamically at different construction steps. The following are the differences between the proposed dynamic architecture and the vanilla architecture:

- The encoder is only used once in the vanilla architecture of AM for VRP; the embedding of each node is fixed, and each node can only represent the input instance's initial state.
- The encoder and decoder are employed alternately to encode the embedding of each node and build a partial solution in AM-dynamic encoder–decoder architecture.

As in [11, 12, 14, 25, 29], solving CO problem was taken as MDP, whereas some authors trained AM-D by policy gradient using reinforcement learning algorithm [30]. The performance of AM-D was compared with AM, and it was notably offered an improved performance of 2.02% for VRP20, 2.01% for VRP50, and 2.55% for VRP100. To further improve the results, the application of local search was used as in [22]. At first, for each instance, AM-D (greedy) creates a solution and then improves the result using the 2-opt local search algorithm. The result showed better performance as compared to AM and other baseline models.

A two-phase solver with geometric clustering was used by [14] to solve the curse of dimensionality. It was derived from the implementation of the "Extended SOLOMON benchmarking dataset" and is compared to Clarke–Wright Savings and sweep heuristic. The overall conclusion was that RL in deterministic, and stochastic demand settings do have better and comparable performance, respectively. Another solution given by [25] was NN-based temporal learning for value function approximation (AVI) and MSA. It was tested on 2-months' worth of historical data and outperformed AVI with 12% improvement and in DRLSA v/s MSA by an avg. of 9.6% with ten-times faster computation time. It also shows better performance with the increase in the degree of dynamism.

The [24] presents a DRL-based GRP solution which was found to be superior when compared to the A* algorithm in most cases for Type II (partial edge depletion). It also contributes toward the development of the GRP sets generator which can generate parameterized GRP sets with different user-defined sizes and constraints.

Another type of DRL-based solution was proposed by Johan Oxenstierna [15]. He applied MCTS-CNN on VRP in warehouses and compared it with VRP two-phase algorithm (a reference algorithm two-phase is provided by SonyTenshiAI) on simulated data to minimize distance. With CNN loss between 0 and 0.1, MCTS-CNN quickly converges on a good or sometimes better solution than two phase. When it was 0.1–0.2, CNN prediction visualizations showed incorrect rational, but still, MCTS search could beat two phase after few seconds, and when it was above 0.2, MCTS-CNN generally showed a weaker solution. The proposed MCTS-CNN algorithm does show high capability on simulated data, but its implementation in a real warehouse management system was premature due to the unavailability of data.

The paper by Nalepa [26] proposed a DRL-based route planning algorithm for pedestrians. The DRL-based algorithm considers and plans the suitable route by considering the travel time and pedestrian flow within the road network by using an agent on a virtual map. It is also called dynamic adjustable route planning (DARP). The initial state is non-congested with uniform distribution of pedestrian flow. Average method performance was recorded after testing ten times. After about ten iterations, average time cost reaches convergence at around −67. In comparison to shortest path selection algorithms, the DRL approach shows a saving of 37.3%-time cost for the best cases. Several nodes are selected at random, and the edges connected to them are set as crowded road segments. The DARP converged at around 15 iterations and saved more travel time for users. Compared with random path selection and shortest path selection, it saved 52.1% and 65.3% time, respectively.

The Euclidean TSP was solved by a learning method based on 2-opt operators and dynamic RL [18]. The learning method also employed neural architecture with graph and sequence embedding. They used the same dataset as in [12], and the results were also compared in terms of solution quality/optimality gap and also in terms of solution quality and sample efficacy. Wu et al. [19] proposed self-attention-based deep architecture as a policy network for guiding the next solution. The results produced by [19] consistently narrowed the optimality gap but [18] achieved 0.01, 0.36, and 1.84% optimality gap for TSP20, TSP50, and TSP100, outshining all previous sampling or search-based RL method in 500 steps although it falls short of

supervised learning method GCN [33]; however, this can be solved by increasing the number of steps.

The solution proposed by Zheng et al. [17] combines three RL techniques (i.e., Q-learning, SARSA, and Monte Carlo) with variable strategy reinforced LKH (VSR-LKH). Learning to choose at each search step is taken care of by the program itself using RL. It is essentially reinforcement on the k-opt process of LKH. It demonstrated excellent performance on 111 TSP benchmarks from the TSPLIB.

5 Major Issues and Challenges

Heuristics for the CO problem are promising cost-saving approaches, but to put this idea in the real world would require better models and better ways of training. Furthermore, hand-crafted rules govern all traditional improvement heuristics, limiting their performance. RL is a great and reliable approach to solve the CO problems, but there are some limitations [11]. The RL algorithm needs to be retrained for new problem instances generated using different distribution. If we talk about AM-D, its training process is a little time-consuming, but it can be solved to some extent by using parallel computing techniques. Recent work in the field of RL is promising, but most of them are intense and require a lot of computing power, and they also face the three-dimensionality curses: For a given state and its action, there are many possibilities for the further research.

- Solving CO-VRP with DRL in itself is a major challenge due to the involvement of volatile research space leading to more volatile time consumption and accuracy.
- The solution space increases exponentially with the increase in problem's input, and thus, the search of optimal solution requires very high computational efforts. More variables are involved in variants of VRP, i.e., VRPTW, CVRP, and VRPPD. To find efficient routes become more challenging.
- The solutions of VRP and its variants require distance matrix calculation, especially in the case of dynamic changing VRPs which is computationally cumbersome.
- One reason for the poor performance of the reinforcement learning algorithm in a stochastic travel time environment is its overfitting during training to the deterministic environment. This highlights the need for generalization, thereby solving the problem, but only in principle, because, in actuality, it is one of the challenges, and the research is in progress.
- Some models are needed to be retrained as soon as the problem instance generation process is moved away from the training distribution leading to a non-feasible/non-dynamic solution approach toward VRP.
- Most of the existing research works don't factor in the realistic simulation settings of the customer environment and stochastic elements.
- Another aspect of research is the scalability aspect of the machine learning models, which currently is not up to the sufficient level.

- In the VRP field, there exists insufficient study of the many real-life characteristics like dynamic traffic environment, demand uncertainty, time period, time windows, service and charging time, and electric vehicles charging requirements.
- The existing reinforcement learning-based platforms can also be employed for scheduling problems [31, 32]

6 Conclusion

This work presents a survey of reinforcement learning-based (RL) approaches proposed to solve the vehicle routing problem (VRP) along with its different variants. The paper also discusses the essential components required to solve VRP using RL-based approaches. Some major issues and challenges that emerged with the use of RL along with the several VRP variants and different solutions are also observed and reported. It has been observed that many state-of-the-art methods have been developed, but they still suffer from a non-optimum quality, slow run time in case of a large-size problems, and lack of model flexibility. It is also required to develop deep learning-based model for the realistic vehicle routing problem with different constraints. Furthermore, scalability is an important issue as well as future research direction that needs to be addressed properly.

References

1. Dantzig, G.B., Ramser, J.H.: The truck dispatching problem. Manage. Sci. **6**, 80–91 (1959)
2. Toth, P., Vigo, D.: Vehicle routing. In: Society for Industrial and Applied Mathematics, Philadelphia, pp. 1–18, PA (2014)
3. Han, M., Wang, Y.: A Survey for vehicle routing problems and its derivatives. IOP Conf. Ser.: Mater. Sci. Eng. **452**, 042024 (2018)
4. Sajid, M., Zafar, A., Sharma, S.: Hybrid genetic and simulated annealing algorithm for capacitated vehicle routing problem. In: 6th IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC), JUIT Solan (2020)
5. Potvin, J.-Y.: State-of-the art review-evolutionary algorithms for vehicle routing. INFORMS J. Comput. **21**, 518–548 (2009)
6. Demir, E., Huckle, K., Syntetos, A., Lahy, A., Wilson, M.: Vehicle routing problem: past and future. In: Wells, P. (ed.) Contemporary Operations and Logistics, pp. 97–117. Springer International Publishing, Cham (2019)
7. Rabbouch, B., Saâdaoui, F., Mraihi, R.: Empirical-type simulated annealing for solving the capacitated vehicle routing problem. J. Exp. Theor. Artif. Intell. **32**, 437–452 (2020)
8. Lin, N., Shi, Y., Zhang, T., Wang, X.: An effective order-aware hybrid genetic algorithm for capacitated vehicle routing problems in internet of things. IEEE Access **7**, 86102–86114 (2019)
9. Altabeeb, A.M., Mohsen, A.M., Ghallab, A.: An improved hybrid firefly algorithm for capacitated vehicle routing problem. Appl. Soft Comput. **84**, 105728 (2019)
10. Cordeau, J.F., Gendreau, M., Laporte, G., Potvin, J.Y., Semet, F.: A guide to vehicle routing heuristics. J. Oper. Res. Soc. **53**, 512–522 (2002)
11. Nazari, M. R., Oroojlooy, A., Snyder, L., Takac, M.: Reinforcement learning for solving the vehicle routing problem. In: Advances in Neural Information Processing Systems, pp. 9860–9870 (2018)

12. Kool, W., Van, H., Welling, M.: Attention, learn to solve routing problems! arXiv:1803.08475 [cs, stat] (2019)
13. Yu, J.J.Q., Yu, W., Gu, J.: Online vehicle routing with neural combinatorial optimization and deep reinforcement learning. IEEE Trans. Intell. Transport. Syst. **20**, 3806–3817 (2019)
14. Kalakanti, A.K., Verma, S., Paul, T., Yoshida, T.: RL SolVeR Pro: reinforcement learning for solving vehicle routing problem. In: 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS), pp. 94–99. IEEE, Ipoh, Perak, Malaysia (2019)
15. Oxenstierna, J.: Warehouse vehicle routing using deep reinforcement learning (2019)
16. Lu, H., Zhang, X., Yang, S.: A learning-based iterative method for solving vehicle routing problems. In: Presented at the International Conference on Learning Representations Sept 25 (2019)
17. Zheng, J., He, K., Zhou, J., Jin, Y., and Li, C.M. : Combining reinforcement learning with Lin-Kernighan-Helsgaun algorithm for the traveling salesman problem (2020)
18. Da Costa, P.R. d O., Rhuggenaath, J., Zhang, Y., Akcay, A.: Learning 2-opt Heuristics for the Traveling Salesman Problem via Deep Reinforcement Learning. In: Pan, S.J., Sugiyama, M. (eds.) Proceedings of The 12th Asian Conference on Machine Learning. pp. 465–480. PMLR, Bangkok, Thailand (2020)
19. Wu, Y., Song, W., Cao, Z., Zhang, J., Lim, A.: Learning Improvement Heuristics for Solving Routing Problems. IEEE Transactions on Neural Networks and Learning Systems. 1–13 (2021)
20. Nalepa, J.: Where machine learning meets smart delivery systems. In: Smart Delivery Systems. pp. 203–226. Elsevier (2020)
21. Lin, B., Ghaddar, B., Nathwani, J.: Deep reinforcement learning for electric vehicle routing problem with time windows. arXiv:2010.02068 [cs, math, stat] (2021)
22. Zhao, J., Mao, M., Zhao, X., Zou, J.: A hybrid of deep reinforcement learning and local search for the vehicle routing problems. IEEE Trans. Intell. Transport. Syst. 1–11 (2020)
23. Peng, B., Wang, J., Zhang, Z.: A deep reinforcement learning algorithm using dynamic attention model for vehicle routing problems. In: Artificial Intelligence Algorithms and Applications: 11th International Symposium, ISICA 2019, Guangzhou, China, November 16–17, 2019, Revised Selected Papers. Springer Singapore, Singapore (2020)
24. Liao, H., Zhang, W., Dong, X., Poczos B., Shimada K., Kara L. B.: A deep reinforcement learning approach for global routing problem (2020) (arXiv:1906.08809)
25. Joe, W., Lau, H.C.: Deep reinforcement learning approach to solve dynamic vehicle routing problem with stochastic customers. In: Proceedings of the International Conference on Automated Planning and Scheduling, vol. 30, pp. 394–402 (2020)
26. Geng, Y., Liu, E., Wang, R., Liu, Y.: Deep reinforcement learning based dynamic route planning for minimizing travel time. arXiv:2011.01771 [cs, eess] (2020)
27. Sutton, R.S., Barto, A.G.: Reinforcement Learning, Second edn: An Introduction. MIT Press (2018)
28. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. arXiv:1506.03134 [cs, stat] (2017)
29. Mazyavkina, N., Sviridov, S., Ivanov, S., Burnaev, E.: Reinforcement learning for combinatorial optimization: a survey. Comput. Oper. Res. **134**, 105400 (2021)
30. Deudon, M., Cournut, P., Lacoste, A., Adulyasak, Y., Rousseau, L.-M.: Learning heuristics for the TSP by policy gradient. In: van Hoeve, W.-J. (ed.) Integration of Constraint Programming, Artificial Intelligence, and Operations Research, pp. 170–181. Springer International Publishing, Cham (2018)
31. Sajid, M., Raza, Z.: Energy-efficient quantum-inspired stochastic Q-HypE algorithm for batch-of-stochastic-tasks on heterogeneous DVFS-enabled processors. Concurrency Comput. Pract Experience **31**, (2019)
32. Sajid, M., Raza, Z.: Energy-aware stochastic scheduler for batch of precedence-constrained jobs on heterogeneous computing system. Energy **125**, 258–274 (2017)

View publication stats