Ministry of Science and Higher Education of the Republic of Kazakhstan

Suleyman Demirel University

Taukekhan Mustakhov

Reinforcement Learning Methods for solving combinatorial optimization problems

THESIS

Presented in Partial Fulfilment for the

Master of Technical Sciences Degree in Computer Science

(degree code: 7M06102)

Department of Computer Science

Faculty of Engineering and Natural Sciences

Supervisor: **Assistant Professor, PhD Aigerim Bogyrbayeva**

Kaskelen 2023

Suleyman Demirel University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty

Associate Professor

PhD Zhamanov A.

«_____» ___06___ 2023

Topic of the thesis:

Reinforcement Learning Methods for solving combinatorial optimization problems

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science" SDU, 2021-2023

Head of Department __________ Assistant Professor, PhD Mukash Zh.

Academic Supervisor __________ Assistant Professor, PhD Bogyrbayeva A.

Master student __________ Mustakhov Taukekhan

Kaskelen 2023

Ministry of Science and Higher Education of the Republic of Kazakhstan

Suleyman Demirel University

Taukekhan Mustakhov

Reinforcement Learning Methods for solving combinatorial optimization problems

THESIS

Presented in Partial Fulfilment for the

Master of Technical Sciences Degree in Computer Science

(degree code: 7M06102)

Department of Computer Science

Faculty of Engineering and Natural Sciences

Supervisor: **Assistant Professor, PhD Aigerim Bogyrbayeva**

Kaskelen 2023

Suleyman Demirel University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty

Associate Professor

PhD Zhamanov A.

____

« _____ »___________ 2023

Topic of the thesis:
Reinforcement Learning Methods for solving combinatorial optimization problems

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science" SDU, 2021-2023

Head of Department ____________ Assistant Professor, PhD Mukash Zh.

Academic Supervisor __________ Assistant Professor, PhD Bogyrbayeva A.

Master student __________________________ Mustakhov Taukekhan

Kaskelen 2023

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Sincerely,

Taukekhan Mustakhov

2023

Acknowledgements

My sincere thanks go out to my supervisor, Aigerim Bogyrbayeva, for her constant support, direction, and mentorship during this research project. Her extensive expertise in Reinforcement Learning and her priceless thoughts have greatly influenced the direction of our study. I am incredibly appreciative of the time she invested in teaching me about the nuances of the research and inspiring a love of thorough inquiry in me.

Additionally, I would like to express my sincere gratitude to Suleyman Demirel University's internal grant for providing financial support throughout the years 2021–2022. The funding has made it much easier to carry out this study endeavor and has allowed me to learn more about the topic.

I owe Shirali Kadyrov a debt of gratitude for his helpful instruction and help in understanding difficult mathematical equations throughout the data processing process.

Additionally, I want to express my gratitude for my teammate Yernar Akhmetbek's continuous support and help during this research project. Yernar has been a continual source of inspiration and encouragement for me. His cooperation and willingness to assist have been genuinely priceless.

Finally, I would want to convey my sincere gratitude to all the people who have offered their assistance, counsel, and encouragement throughout this study project. No matter how big or small, your contributions have had a huge impact on our research.

Abbreviations

VRP - Vehicle Routing Problem

SDVRP - Stochastic Dynamic Vehicle Routing Problem

RL - Reinforcement learning

PMF - probability mass function

NP - nonpolynomial

DQN - deep Q-network

ANOVA - Analysis of variance

SHAP - SHapley Additive exPlanations

Abstract

The use of reinforcement learning approaches to resolve combinatorial optimization issues in the context of vehicle routing is examined in this paper. Three previously published articles' insights are combined in one study. The Stochastic Dynamic Vehicle Routing Problem is addressed in the first paper with a deep reinforcement learning strategy. To learn the routing strategy for a single truck as customer orders come in over time, a fully attention-based model with a dynamic encoder and decoder is introduced. The model is taught using reinforcement learning, in which a reward signal influences the choice of nodes to visit. Computational analyses show that this tactic outperforms comparison algorithms. A thorough overview of machine learning techniques used to address NP-hard Vehicle Routing Problems is presented in the second paper. A variety of learning paradigms, solution structures, underlying models, and algorithms are included in the survey. It demonstrates the benefits of machine learning-based models that use the symmetry of VRP solutions and their parity with conventional approaches. The directions for future studies to address the problems with contemporary transportation systems are also described. The third paper provides a thorough analysis of the stochastic dynamic vehicle routing problem, examining state-of-the-art approaches and methodologies for solving this challenging optimization problem. The stochastic and dynamic restrictions that present significant challenges to efficient route planning and optimization are covered in detail. The study offers a thorough grasp of cutting-edge techniques and recommends prospective directions for further investigation. In addition, the use of Markov decision processes and sophisticated reinforcement learning techniques to address the issue of stochastic dynamic vehicle routing is examined. Combining

these articles, this thesis advances knowledge of reinforcement learning methods for combinatorial optimization issues, particularly in the area of vehicle routing. The study demonstrates the effectiveness of deep reinforcement learning and machine learning techniques while also highlighting areas that need additional study and development in the field of transportation logistics.

Аңдатпа

Бұл жұмыста көлік құралын бағдарлау контекстінде комбинаторлық оңтайландыру мәселелерін шешу үшін оқытудың күшейту тәсілдерін пайдалану қарастырылады. Бұрын жарияланған үш мақаланың түсініктері бір зерттеуде біріктірілген. Стохастикалық динамикалық көлікті бағыттау мәселесі бірінші мақалада тереңдетілген оқыту стратегиясымен қарастырылған. Тұтынушы тапсырыстары уақыт өте келе түсетіндіктен, бір жүк көлігінің маршруттау стратегиясын білу үшін динамикалық кодтауыш пен декодермен толығымен назар аударуға негізделген модель енгізіледі. Модель марапаттау сигналы келуге болатын түйіндерді таңдауға әсер ететін күшейтетін оқыту арқылы оқытылады. Есептеу талдаулары бұл тактиканың салыстыру алгоритмдерінен асып түсетінін көрсетеді. NP-күрделі көлікті бағыттау мәселелерін шешу үшін қолданылатын машиналық оқыту әдістеріне толық шолу екінші мақалада берілген. Сауалнамаға әртүрлі оқыту парадигмалары, шешім құрылымдары, негізгі модельдер және алгоритмдер енгізілген. Ол көлікті бағдарлау мәселелерінің шешімдерінің симметриясын және олардың әдеттегі тәсілдермен теңдігін пайдаланатын машиналық оқытуға негізделген үлгілердің артықшылықтарын көрсетеді. Заманауи көлік жүйелерімен проблемаларды шешу үшін болашақ зерттеулердің бағыттары да сипатталған. Үшінші жұмыс осы күрделі оңтайландыру мәселесін шешудің заманауи тәсілдері мен әдістемелерін зерттей отырып, стохастикалық динамикалық көлікті бағыттау мәселесіне мұқият талдау жасайды. Маршрутты тиімді жоспарлау мен оңтайландыруға елеулі қиындықтар тудыратын стохастикалық және динамикалық шектеулер егжей-тегжейлі қарастырылған. Зерттеу алдыңғы қатарлы әдістерді мұқият меңгеруді ұсынады және одан әрі зерттеу үшін перспективалық бағыттар-

ды ұсынады. Сонымен қатар, Марковтың шешім қабылдау процестерін және көлік құралын стохастикалық динамикалық бағыттау мәселесін шешу үшін күрделі күшейтуді оқыту әдістерін пайдалану зерттеледі. Осы мақалаларды біріктіре отырып, бұл дипломдық жұмыс комбинаторлық оңтайландыру мәселелеріне, әсіресе көліктің маршрутизациясы саласында оқытудың күшейту әдістері туралы білімді жетілдіреді. Зерттеу тереңдетілген оқыту мен машиналық оқыту әдістерінің тиімділігін көрсетеді, сонымен қатар көлік логистикасы саласында қосымша зерттеуді және дамытуды қажет ететін бағыттарды көрсетеді.

Аннотация

Первая строка содержит целое число T ($1 \leq T \leq 100$) - количество тестов.

Первая строка каждого теста содержит два целых числа N ($1 \leq N \leq 2 * 10^5$) и M ($1 \leq M \leq 10^9$) - количество магических изготовителей подарков и подарков.

На следующей строке расположены N целых чисел $A_1, A_2, \ldots, A_n$ - время, необходимое для создания подарка с использованием каждого изготовителя ($1 \leq A_i \leq 10^9$).

Гарантируется, что сумма N по всем тестам не превышает 10^5.

В этой статье рассматривается использование подходов обучения с подкреплением для решения проблем комбинаторной оптимизации в контексте маршрутизации транспортных средств. Выводы из трех ранее опубликованных статей объединены в одно исследование. Стохастическая динамическая задача маршрутизации транспортных средств рассматривается в первой статье со стратегией глубокого обучения с подкреплением. Чтобы изучить стратегию маршрутизации для одного грузовика по мере поступления заказов клиентов с течением времени, была введена модель, полностью основанная на внимании, с динамическим кодировщиком и декодером. Модель обучается с использованием обучения с подкреплением, в котором сигнал вознаграждения влияет на выбор узлов для посещения. Вычислительный анализ показывает, что эта тактика превосходит алгоритмы сравнения. Тщательный обзор методов машинного обучения, используемых для решения NP-сложных задач маршрутизации транспортных средств, представлен во второй статье. В обзор включены различные парадигмы обучения, структуры решений, базо-

вые модели и алгоритмы. Он демонстрирует преимущества моделей, основанных на машинном обучении, которые используют симметрию решений задач маршрутизации транспортных средств и их паритет с традиционными подходами. Также описаны направления будущих исследований для решения проблем с современными транспортными системами. В третьей статье представлен тщательный анализ стохастической динамической задачи маршрутизации транспортных средств с изучением современных подходов и методологий для решения этой сложной задачи оптимизации. Подробно рассматриваются стохастические и динамические ограничения, которые создают серьезные проблемы для эффективного планирования и оптимизации маршрута. Исследование предлагает полное представление о передовых методах и рекомендует перспективные направления для дальнейших исследований. Кроме того, рассматривается использование марковских процессов принятия решений и сложных методов обучения с подкреплением для решения проблемы стохастической динамической маршрутизации транспортных средств. Объединяя эти статьи, эта диссертация расширяет знания о методах обучения с подкреплением для решения проблем комбинаторной оптимизации, особенно в области маршрутизации транспортных средств. Исследование демонстрирует эффективность методов глубокого обучения с подкреплением и машинного обучения, а также выделяет области, требующие дополнительного изучения и развития в области транспортной логистики.

Table of Contents

Chapter 1

Introduction

Problems with combinatorial optimization are common in many different fields and are essential for streamlining planning, decision-making, and resource allocation procedures. Finding the ideal combination or choice of components from a finite collection while taking into account certain limitations and goals is the focus of these topics. Traditional approaches to solving combinatorial issues, however, face substantial difficulties due to their complexity and computational demands.

In recent years, the discipline of reinforcement learning (RL) has attracted a lot of interest as a potent method for tackling complicated decision-making issues [1]. RL includes discovering the best course of action through rewarding trial-and-error encounters with the environment. It is now possible to successfully address combinatorial optimization issues by utilizing RL approaches.

There are various benefits to using reinforcement learning in combinatorial optimization. First of all, RL models are capable of coping with the complexity and ambiguity that are a part of real-world settings. Second, in dynamic contexts where the problem parameters may vary over time, RL algorithms can adapt and learn the best course of action. The possibility for scalable and effective solutions that can handle significant combinatorial issues is another benefit of reinforcement learning [2].

Heuristics and precise algorithms are examples of classic approaches to combinatorial optimization problems that have yielded useful insights and solutions, but they frequently have issues with scalability, optimality, and handling dynamic situations. Innovative approaches are required to get beyond these restrictions and offer practical, workable answers.

The purpose of this master's thesis is to investigate the utility of reinforcement learning methods for the solution of combinatorial optimization issues. The goal of the research is to contribute to the creation of cutting-edge techniques that can handle the difficulties posed by such optimization tasks by gaining a thorough grasp of RL algorithms and their application to combinatorial issues.

The following chapters will examine the theoretical underpinnings of reinforcement learning, survey the literature on using RL to solve combinatorial optimization issues, propose a novel reinforcement learning framework, carry out extensive experiments and evaluations, and offer insights into the efficiency and performance of the suggested strategy.

By utilizing the power of reinforcement learning, we hope to advance the field of combinatorial optimization through this work by offering workable solutions that can improve resource allocation, streamline decision-making processes, and open the door to more productive operations in a variety of fields and industries.

1.1 Background

1.1.1 Vehicle Routing Problem

An extensively researched combinatorial optimization problem in the area of transportation logistics is the Vehicle Routing Problem (VRP). The VRP is typically characterized as a graph with a set of nodes $\mathbb{V} = \{v_0, \dots, v_n\}$, and a set of edges or arcs. This graph can be either directed or undirected. The undirected graph $\mathbb{E} = \{(v_i, v_j) : i < j, v_i, v_j \in \mathbb{V}\}$ is the subject of our attention [1]. The VRP seeks to design routes with the lowest total cost for M number of identical vehicles departing from a depot, v_0, visiting all the nodes, $\mathbb{V} \setminus v_0$, representing customers with non-negative demand, d_v, and returning to the depot, where each customer is only visited once, and then returning to the depot. In VRP, a fleet of vehicles must satisfy constraints like vehicle capacity, time frames, and distance restrictions while providing service to a set of customer locations by working out the best routes for each vehicle.

VRP is a difficult topic with many practical applications, such as planning for public transportation, waste collection, and delivery services. The purpose of VRP is to ensure effective and timely deliveries to all clients while reducing overall operating costs, such as fuel use or vehicle usage.

VRP has been addressed using a variety of algorithms and methodologies over the years, including precise methods, heuristic techniques, and metaheuristic algorithms. These conventional optimization techniques have produced beneficial answers for VRP cases of various sizes and complexity. However, they frequently struggle to deal with complex issues or dynamic contexts where quick adjustments are required [3].

The exploration of new strategies to solve the drawbacks of conventional VRP optimization methods has garnered more attention in recent years. Reinforcement learning, a branch of machine learning that focuses on teaching agents to make sequential decisions based on observed states and rewards, is one such method[4, 5, 6]. It is feasible to learn the best vehicle routing policies by utilizing reinforcement learning strategies; these policies can handle uncertainty, adapt to dynamic changes, and possibly outperform more conventional optimization techniques [7].

The Stochastic Dynamic Vehicle Routing Problem (SDVRP), which incorporates extra difficulties of stochasticity and dynamic client orders, is the primary topic of this thesis. It broadens the scope of classical VRP. We seek to overcome the drawbacks of conventional optimization techniques and create adaptive routing algorithms that can successfully manage uncertainties and dynamic changes in the field of transportation logistics.

In the sections that follow, we'll give a thorough literature analysis of techniques to combinatorial optimization and reinforcement learning, with an emphasis on VRP. We will evaluate the benefits and drawbacks of current approaches, pinpoint knowledge gaps, and present our methodology and experimental framework for overcoming the difficulties of SDVRP.

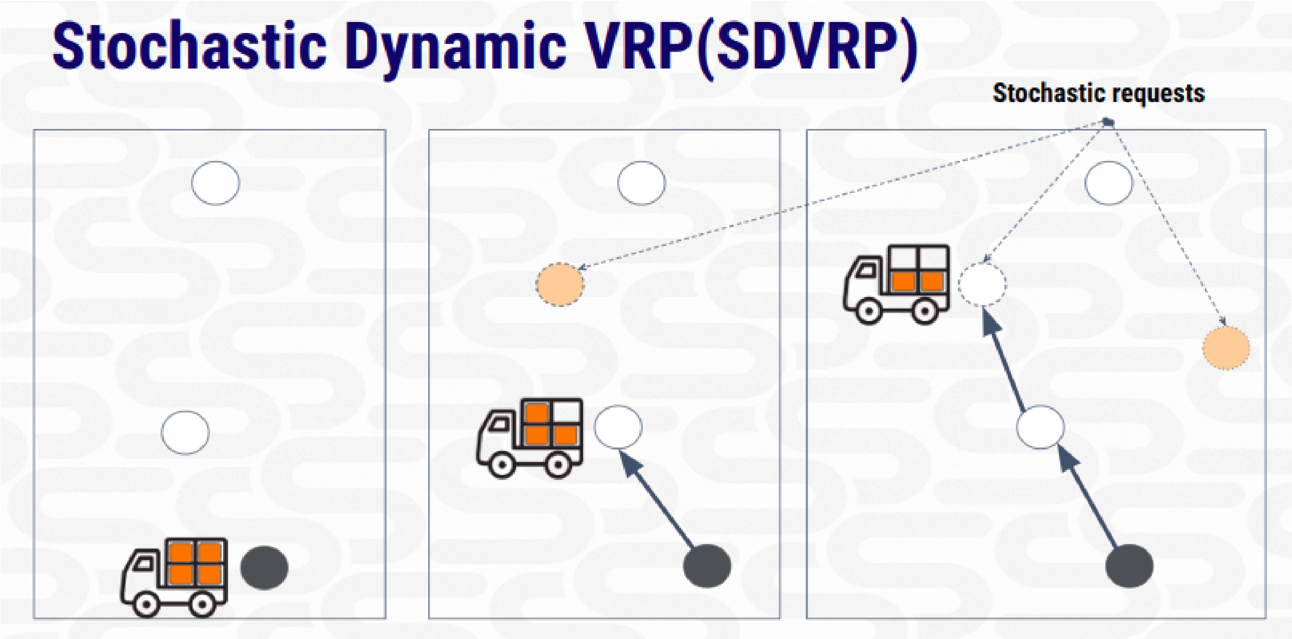

Figure 1.1: SDVRP with stochastic requests

1.1.2 Stochastic Dynamic Vehicle Routing Problem

In transportation logistics, when trucks are dynamically routed to satisfy client requests that emerge over time in a stochastic environment, the Stochastic Dynamic Vehicle Routing Problem (SDVRP) is a difficult optimization problem [8]. SDVRP is more sophisticated than conventional vehicle routing problems because it takes into account uncertainties in variables like client demand, journey times, and road conditions. SDVRP's goal is to reduce total operational costs while assuring prompt client delivery [9].

SDVRP presents several difficulties because of its dynamic and stochastic nature. Real-time routing choices must be made as new customer orders come in and the environment evolves. The decision-making process is further complicated by the ambiguity of client needs. One of the main challenges in SDVRP is effectively adjusting vehicle paths to these dynamic changes while balancing exploitation and exploration [10].

SDVRP is a significant issue in transportation logistics, but there hasn't been much discussion of it in the literature. The majority of the SDVRP research to date has utilized mathematical programming techniques and heuristic-based approaches, which do not adequately account for the stochastic and dynamic components of the issue. As a result, there is a huge research gap in investigating new strategies to effectively combat SDVRP.

This thesis attempts to fill this knowledge gap by using reinforcement learning methods on the SDVRP. A possible approach to dealing with the stochastic and dynamic nature of the issue is reinforcement learning. Reinforcement learning enables real-time route adaptation for vehicles by teaching an agent to discover the best routing strategy through trial-and-error interactions with the environment.

We will present a thorough literature analysis that examines current methods in combinatorial optimization and reinforcement learning in the sections that follow, with an emphasis on SDVRP in particular. Along with presenting our suggested methodology, experimental setting, and outcomes, we will also show how well reinforcement learning works for solving the SDVRP.

1.2 Motivation

The inherent difficulties experienced by conventional methodologies and the potential benefits afforded by reinforcement learning in overcoming these difficulties are what drive the use of reinforcement learning techniques for combinatorial optimization problems.

First off, conventional approaches to combinatorial optimization problems frequently use precise or heuristic algorithms, both of which have drawbacks in terms of scalability, optimality, and flexibility to changing conditions. Traditional approaches struggle to offer efficient and effective solutions as the size and complexity of these problems grow. On the other hand, reinforcement learning provides a promising alternative by utilizing its capacity to discover the best rules through trial-and-error interactions with the environment. Reinforcement learning is able to adapt and learn, which makes it a good choice for dealing with the complexity and dynamic nature of combinatorial optimization issues.

Second, real-world scenarios must be able to adapt to changing circumstances due to their dynamic and uncertain nature. Time-varying restrictions and changing parameters are frequently present in combinatorial optimization problems in areas like logistics, transportation, and resource allocation. Reinforcement learning is a beneficial strategy to deal with these dynamic situations since it can learn from successive interactions and change policies accordingly. Reinforcement learning can be used to optimize solutions in real time when new information becomes available by incorporating it into the optimization process.

In addition, reinforcement learning techniques are attractive for large-scale combinatorial optimization issues due to their potential scalability and effectiveness. The computing needs of such problems may be too much for conventional methodologies to manage, which could result in inadequate or unworkable solutions. With its ability to parallelize, spread learning, and make data-driven decisions, reinforcement learning approaches have the potential to provide scalable and effective solutions to complex combinatorial problems.

This research intends to overcome the constraints of conventional methodologies, create adaptive and scalable solutions, and optimize decision-making processes in diverse areas by using reinforcement learning for combinatorial optimization. Reinforcement learning is being explored and put to use in this setting in order to improve resource allocation, operational efficiency, and decision-making in challenging real-world situations.

This thesis will examine the theoretical underpinnings of reinforcement learning in further detail in the next chapters. It will also analyze the literature, suggest a novel reinforcement learning framework for combinatorial optimization, and run experiments to gauge the framework's effectiveness. The findings of this study will help to advance the discipline of combinatorial optimization and offer insightful information on the potential of reinforcement learning in the solution of challenging optimization issues.

1.3 Research Gap

Even though the subject of combinatorial optimization has seen great advancements, there is still a substantial research gap in the use of reinforcement learning methods designed expressly for combinatorial optimization problems. Exact algorithms and heuristics are examples of traditional approaches that have produced useful insights and solutions, but they frequently have issues with scalability, optimality, and adaptation to changing situations. This knowledge gap offers a chance to investigate and create novel approaches to combinatorial optimization problems that make use of the strength of reinforcement learning.

Furthermore, the potential synergies between the two domains are not completely utilized in the existing research, which largely focuses on either combinatorial optimization or reinforcement learning separately. Although reinforcement learning has shown promise in a number of decision-making contexts, combinatorial optimization is one of its largely untapped applications. This research gap can be closed and new approaches to solving challenging optimization problems opened by combining the advantages of reinforcement learning, which excels at sequential decision-making and adaptation, with the difficulties of combinatorial optimization, which demand efficient resource allocation and optimal solutions.

Furthermore, the majority of recent studies in combinatorial optimization mostly rely on precise algorithms or custom-made heuristics. Although these techniques have produced useful solutions, they frequently have trouble keeping up with the computational requirements and dynamic nature of real-world circumstances. It is feasible to get over these restrictions and offer more effective and efficient solutions for combinatorial optimization issues by utilizing reinforcement learning techniques, which have the potential for scalable and data-driven decision-making.

By creating a thorough grasp of reinforcement learning methods and their application to combinatorial optimization issues, this research attempts to close the current research gap. This research aims to develop novel methodologies that can get beyond the drawbacks of conventional approaches by utilizing the benefits of reinforcement learning, such as adaptability, scalability, and data-driven decision-making. Combinatorial optimization and reinforcement learning can be combined to revolutionize decision-making, improve resource allocation, and successfully solve complex, dynamic combinatorial problems.

This research intends to close the gap between these two domains and promote both reinforcement learning and combinatorial optimization approaches through the investigation and development of reinforcement learning techniques customized for combinatorial optimization. By filling in this research gap, this thesis aims to offer useful answers and insights that may have an impact on a number of sectors and disciplines that depend on effective and efficient decision-making processes.

1.4 Objectives

This master's thesis, "Reinforcement Learning for Combinatorial Optimization Problems," has the following objectives:

- Gain a thorough grasp of the reinforcement learning approaches, algorithms, and techniques that can be used to solve combinatorial optimization challenges.
- Discover the main issues, restrictions, and areas for development in the existing literature on the use of reinforcement learning in combinatorial optimization.
- Considering the special traits and limitations of combinatorial optimization problems, propose a novel reinforcement learning framework created exclusively for resolving these issues.
- Implement and assess the suggested reinforcement learning framework on a few specific combinatorial optimization issues, contrasting its performance with that of benchmark algorithms and conventional optimization techniques.
- To determine the efficacy, efficiency, and scalability of the reinforcement learning strategy for resolving combinatorial optimization problems, conduct thorough experiments and evaluations.
- Analyze and analyze the experimental data to reveal the advantages and disadvantages of the suggested strategy and its applicability to other problem areas.
- To improve the overall performance and resilience of the optimization process, look into the possibility of combining reinforcement learning with other

optimization techniques, such as metaheuristics or exact algorithms.

- Outline prospective topics for more investigation and development by identifying and discussing the difficulties, restrictions, and future research possibilities in the application of reinforcement learning to combinatorial optimization issues.

- Contribute to the field of combinatorial optimization by offering real-world approaches, ideas, and methodologies that make the most of reinforcement learning's potential to improve resource allocation and decision-making across a range of applications.

- A master's thesis that presents the theoretical underpinnings, methods, experimental outcomes, and conclusions obtained from the research is needed to document and disseminate the research findings.

  By achieving these goals, this research aims to advance our understanding of reinforcement learning techniques and their application to combinatorial optimization problems, making significant contributions to the field and offering useful solutions that can improve decision-making procedures, improve resource allocation, and boost operational efficiency in challenging real-world scenarios.

1.5 Methodology Overview

The research's technique attempts to create and assess a reinforcement learning framework for dealing with combinatorial optimization issues. The general strategy is outlined in the steps below:

The Traveling Salesman Problem, Vehicle Routing Problem, and Job Shop Scheduling Problem are a few examples of combinatorial optimization problems that exhibit a spectrum of properties. These issues will form the framework for creating and assessing the reinforcement learning system.

Selection of Reinforcement Learning Algorithms: Review and evaluation of existing reinforcement learning methods effective at solving combinatorial optimization issues. Think about algorithms like actor-critical architectures, Deep Q-Networks, Policy Gradient approaches, and Q-learning. Based on their applicability to the specified problem domains, choose the best algorithms.

Framework Development: Create and put into action a reinforcement learning framework that uses the chosen techniques and algorithms. Modules for state representation, action choice, reward shaping, and learning updates should be included in this framework. The reinforcement learning framework should be modified to take into account the unique properties and limitations of combinatorial optimization issues.

Create or gather appropriate datasets for the training and assessment of the reinforcement learning framework. The datasets should include occurrences of the problems, restrictions, and any optimal or nearly optimal solutions that were found. To ensure the robustness and generalizability of the framework, take into account both synthetic datasets and instances of real-world problems.

Training and evaluation: Make use of the prepared datasets to train the reinforcement learning framework. Update the model parameters iteratively using suitable training methods, such as policy gradient optimization or value function approximation. Use a variety of criteria to assess the trained model, such as solution quality, convergence speed, or computational effectiveness. Compare the performance of the reinforcement learning framework against the standard optimization techniques and baseline algorithms.

Experimental Setup: Specify the hardware, software, programming languages, and libraries that will be used in the experiment. Document the configurations, hyperparameters, and random seed initialization used in the training and evaluation processes to make sure the experiments can be replicated.

Analyze the reinforcement learning framework's performance in light of various issue instances, problem sizes, and dynamic contexts. By taking into account several performance metrics including solution quality, computation time, convergence behavior, and sensitivity to parameter settings, you may assess the robustness and scalability of the system.

Comparative Analysis: Evaluate the performance of the reinforcement learning framework in comparison to more established techniques for combinatorial optimization, such as precise algorithms, heuristics, or metaheuristics. In terms of solution quality, scalability, adaptability to dynamic situations, and computational efficiency, list the advantages and disadvantages of the proposed approach.

Interpreting the Results: Highlighting the advantages, disadvantages, and consequences of the established reinforcement learning framework, interpret and analyze the experimental findings. Discuss the findings from the experiments and how they relate to applying reinforcement learning to solve combinatorial optimization challenges.

Future Directions: Identify possible directions for additional study and development, such as examining hybrid methods that combine reinforcement learning with other optimization strategies, investigating various architectures or variants of the reinforcement learning framework, or extending the framework to deal with more intricate or specialized combinatorial optimization problems.

An extensive development and evaluation plan for the reinforcement learning framework for combinatorial optimization issues is provided in the methodology overview. The project hopes to develop reinforcement learning approaches in multiple areas by using this methodology to optimize decision-making, resource allocation, and operational efficiency.

Chapter 2

Literature review

In a number of industries, including logistics, transportation, manufacturing, scheduling, and resource allocation, combinatorial optimization problems are of utmost importance[11]. Finding the best combination of elements to satisfy a set of restrictions and optimize an objective function is the aim of these issues, which entail making optimal judgments from a finite collection of options.

Combinatorial optimization issues have traditionally been solved by using traditional optimization techniques like precise algorithms and heuristic approaches. These approaches frequently make use of search algorithms, heuristics tailored to particular problems, or mathematical programming techniques. Even if they have had a lot of success, they still have a lot of difficulty solving complicated issues on a huge scale[2].

One significant difficulty is the exponential increase of the solution space with increasing problem size, which makes accurate solutions computationally intractable. Such approaches are problematic for real-world applications because the time needed to identify an ideal solution grows prohibitively expensive as problem dimensions rise.

Additionally, because traditional optimization techniques rely so largely on problem-specific heuristics and domain expertise, their generalizability and scalability may be constrained when applied to other problem instances and problem domains. It can take a lot of effort and resources to create efficient heuristics for every problem variant.

Alternative methods that can successfully tackle combinatorial optimization issues are required in light of these difficulties. A method to solve these issues has emerged that shows promise: reinforcement learning. A branch of machine learning called reinforcement learning is concerned with discovering the best decision-making procedures by interacting with the environment. In dynamic and sophisticated fields like robotics, gaming, and control systems, it has achieved extraordinary success.

Combinatorial optimization issues can benefit from reinforcement learning, which allows us to get around some of the drawbacks of conventional optimization techniques[2]. Algorithms that use reinforcement learning have the capacity to discover the best policies without the aid of prior information or heuristics. They can generalize to other issue cases and adapt to them, and they can even work in dynamic contexts where the problem's features are constantly changing.

Additionally, reinforcement learning has the benefit of allowing for rapid exploration and exploitation of the solution space, which may result in high-quality solutions and enhanced optimization performance. Reinforcement learning is particularly well suited for combinatorial optimization problems with a high number of alternative configurations because of its capacity to handle enormous state and action spaces.

Consequently, combining reinforcement learning methods with combinatorial optimization has the potential to completely change how we approach and resolve challenging optimization problems. We can get around the drawbacks of conventional optimization techniques and find better, more effective solutions for real-world combinatorial optimization problems by utilizing the learning capabilities of reinforcement learning algorithms.

2.1 Combinatorial Optimization

Finding the best solutions from a limited number of options is the goal of combinatorial optimization, which covers a wide range of issues and methodologies.

Combinatorial optimization problems have recently attracted growing interest in the use of machine learning, particularly deep learning, and reinforcement learning. The following provides a thorough review of key principles, application areas, current methodologies, and key works in combinatorial optimization:

Domains of Problems and Formulations: Combinatorial optimization issues can occur in a variety of problem fields, including routing, scheduling, assignment, packing, and graph optimization. Typical formulations of problems: Graph-based problems, integer programming problems, constraint satisfaction problems, and network optimization challenges are some examples of problem formulations. Solution visualizations Depending on the particular issue, solutions are frequently shown as permutations, subsets, assignments, or pathways in a graph.

Objective Purposes and Limitations: Combinatorial optimization seeks to achieve particular goals, such as decreasing expenses, maximizing efficiency, or meeting predetermined requirements. Constraints: Solutions to problems frequently need to take into account constraints, such as logic, precedence, or capacity limits.

Existing Methods: Exact algorithms, heuristic algorithms, metaheuristics, and mathematical programming approaches like linear programming and integer programming are examples of traditional optimization techniques. Machine Learning Methodologies Recent methods, such as meta-learning, deep learning, and reinforcement learning, use machine learning techniques to address combinatorial optimization issues.

Important Works:

- In [12] work suggests a framework for employing reinforcement learning to develop combinatorial optimization algorithms across graphs. It shows how well the suggested strategy works for issues like the Vehicle Routing Problem and the Traveling Salesman Problem.
- In [11] study from 2021, machine learning for combinatorial optimization is analyzed from a methodological perspective. It investigates various learning paradigms, representations of solutions, underpinning models, and algorithms. Challenges, applications, and potential research trajectories are

also covered.

- The survey paper [2] focuses on teaching combinatorial optimization on graphs, particularly concerning networking. It offers a thorough analysis of methods, applications, and models while showing the potential of machine-learning strategies to solve network-related issues.

- The work [13] introduces the use of graph neural networks for combinatorial reasoning and optimization. It demonstrates how graph neural networks are successful at resolving issues like the Traveling Salesman Problem and the Maximum Independent Set.

These studies show how machine learning methods, particularly reinforcement learning and graph neural networks, are becoming more popular for tackling combinatorial optimization issues. They demonstrate the potential of learning-based approaches to get beyond the drawbacks of conventional optimization techniques and offer effective answers to challenging optimization problems.

2.2 Stochastic Dynamic Vehicle Routing Problem

A significant variation of the Vehicle Routing Problem (VRP) that adds stochastic and dynamic components to the routing process is the Stochastic Dynamic Vehicle Routing Problem (SDVRP). SDVRP provides a more accurate portrayal of real-world transportation logistics scenarios by taking uncertainties in consumer requests, trip delays, and other factors into account.

Traditional VRP formulations and their solutions have received the majority of attention in previous investigations [14]. To address the issues raised by SDVRP, however, more advanced strategies must be developed because of the inherent stochastic and dynamic nature of many transportation networks [15, 16].

Customer demand uncertainties and the continually shifting nature of the transportation environment are the sources of the stochastic component in SDVRP[8]. Variations in customer demand may affect the number of cars needed, the routes to be taken, and the overall scheduling of the vehicles. In addition, the problem's dynamic nature is demonstrated by the gradual emergence of new client orders, necessitating real-time modifications to vehicle routes to ensure effective and timely delivery[17, 18].

Even though conventional optimization techniques have proved effective in solving VRP situations with deterministic inputs, they have major difficulties managing the stochastic and dynamic characteristics of SDVRP. In order to address the complexity of SDVRP, there is thus increasing interest in using reinforcement learning approaches [2, 18].

A possible approach to dealing with the stochastic and dynamic components of SDVRP is reinforcement learning. Reinforcement learning algorithms can adaptively modify vehicle routes in response to changing client needs and dynamically evolving conditions by training an agent to learn optimal routing policies through interactions with the environment [13].

To explore and compare various reinforcement learning algorithms, architectures, and methodologies in the context of SDVRP, further research is required [2, 19]. However, the application of reinforcement learning specifically to SDVRP is still in its early stages.

In this thesis, we propose and assess reinforcement learning methods for efficient vehicle routing in stochastic and dynamic situations in order to contribute to the study of SDVRP. We want to increase SDVRP instances' adaptability, efficiency, and solution quality by utilizing the benefits of reinforcement learning.

We will identify the advantages and disadvantages of existing approaches, highlight knowledge gaps, and propose novel methodologies and enhancements to advance the field of SDVRP using reinforcement learning techniques through a thorough analysis of the existing literature, including the survey [2] and the work [19].

2.3 Reinforcement Learning in Optimization

Recently, there has been a lot of interest in using reinforcement learning approaches to solve combinatorial optimization issues. The overview that follows examines the literature on this subject and discusses pertinent approaches, architectures, and algorithms that have been put forth and assessed in the context of optimization. The advantages and disadvantages of various reinforcement learning methods when used to solve combinatorial optimization problems are also highlighted:

- [2] gives a thorough summary of reinforcement learning for combinatorial optimization. They examine several algorithmic frameworks, such as actor-critical architectures, policy gradient approaches, and Q-learning. Important topics including reward design, exploration-exploitation techniques, and state and action representations are also covered in the survey. The advantages and disadvantages of various algorithms, as well as practical applications and open research problems, are presented.
- [19]: This comprehensive paper investigates the application of deep reinforcement learning techniques with a specific focus on transportation network combinatorial optimization. For issues like traffic signal control, route planning, and vehicle dispatching, it examines the integration of deep neural networks with reinforcement learning methods. The survey offers details on the deep reinforcement learning architectures, training methods, and assessment metrics employed for transportation network optimization. In this context, the advantages and disadvantages of deep reinforcement learning systems are examined, along with prospective future study lines.
- [13] For combinatorial optimization, this study blends reinforcement learning and constraint programming. It provides a framework to address optimization issues that combines the advantages of both methods. The authors emphasize the advantages of combining constraint programming and reinforcement learning, including enhanced solution quality and scalability. The synergy between these two strategies is discussed, and experimental evidence of their efficacy in solving combinatorial optimization issues is presented in the paper.

These reviews and studies provide a thorough overview of how reinforcement learning is used to solve combinatorial optimization problems. They encompass a broad spectrum of approaches, structures, and algorithms used in optimization applications. Combinatorial optimization methods that use reinforcement learning frequently benefit from their ability to handle huge state and action spaces, flexibility in dynamic contexts, and capacity for experience-based learning. However, problems like sample complexity, trade-offs between exploration and exploitation, and generalization to unforeseen problem situations continue to be active study fields.

For combinatorial optimization, specific algorithms have been investigated, including Q-learning, policy gradient methods, and actor-critic structures. In terms of the value estimate, policy optimization, and exploration tactics, these algorithms take different methods. The choice of algorithm depends on the features of the problem and the need for optimization. Each algorithm has strengths and limits of its own.

The use of reinforcement learning approaches to solve combinatorial optimization problems yields encouraging results and offers up fresh possibilities for solving challenging optimization problems. Researchers can create efficient methods for optimizing combinatorial issues by utilizing the advantages of various algorithms and architectures. The application of reinforcement learning in combinatorial optimization can be advanced further by addressing the constraints and research issues identified in the literature. This will also help to promote the development of effective optimization techniques across a variety of disciplines. Combinatorial optimization has made tremendous strides in recent years as scholars look into alternatives to more conventional optimization techniques. Utilizing learning strategies, such as reinforcement learning, to solve combinatorial optimization problems is one such method. Including the recently released work [1], we give a thorough overview of combinatorial optimization ideas, problem domains, and current methodologies in this literature review.

In a survey [1] pay particular attention to learning-based solutions for vehicle routing issues. The authors examine several methodologies and techniques used in this subject, such as meta-heuristic algorithms, deep learning, and reinforcement learning. The survey discusses several problem formulations, representations of solutions, and optimization goals applied to the setting of vehicle routing. It also looks at the benefits, drawbacks, and future directions of learning-based vehicle routing techniques.

We gain important insights into the state-of-the-art learning-based approaches for resolving combinatorial optimization issues by including the results of [1] in our literature assessment, particularly in the area of vehicle routing. The survey helps us comprehend the developments in this area and illuminates both the advantages and drawbacks of using learning approaches to solve vehicle routing issues.

We also evaluate other significant publications in the field, such as the survey [2], which offers a thorough overview of reinforcement learning for combinatorial optimization, in addition to the survey [1]. In addition, [19] give a survey on the use of deep reinforcement learning approaches for combinatorial optimization of transportation networks. These studies help us comprehend the various strategies and approaches used when applying reinforcement learning to combinatorial optimization issues.

In general, the study of the literature offers a thorough overview of the importance of combinatorial optimization problems in numerous fields, the difficulties encountered by conventional optimization techniques, and the necessity for alternate strategies, such as reinforcement learning. Common problem formulations, representations of solutions, objective functions, and restrictions are covered. We improve our grasp of the most recent state-of-the-art techniques and emphasize the potential directions for further research in this important domain of combinatorial optimization by incorporating pertinent publications like the survey by [1].

2.4 Comparative Analysis

The research community has been interested in how reinforcement learning approaches perform when compared to conventional optimization techniques. Numerous research has examined and contrasted the efficacy, efficiency, and scalability of reinforcement learning algorithms with heuristic algorithms, precise methods, and other machine learning approaches. One such study is the survey [1]. This section will describe these studies, evaluate each method's advantages and disadvantages, and point out domains or problem categories where reinforcement learning has demonstrated promising outcomes.

In a survey of learning-based strategies for resolving vehicle routing problems [1] compared the effectiveness of reinforcement learning algorithms with conventional optimization techniques. The benefits of reinforcement learning in terms of adaptability, scalability, and the capacity to deal with dynamic and uncertain settings were underlined in their survey. They also stressed the potential of reinforcement learning to tackle the difficulties of vehicle routing problems and discover close-to-ideal solutions.

There have been several research that contrasts reinforcement learning with conventional optimization techniques in addition to the survey [1]. For the Traveling Salesman Problem (TSP), [20] evaluated the effectiveness of a reinforcement learning system with a genetic algorithm. The findings demonstrated that, in terms of solution quality and computation time, the reinforcement learning approach produced competitive outcomes.

For the Vehicle Routing Problem (VRP), [21] compared reinforcement learning algorithms with precise solutions. Especially for large-scale examples, their results showed the potential performance of reinforcement learning algorithms in terms of locating nearly optimal solutions while being computationally efficient.

In addition, [22] examined classical optimization techniques, deep learning, and reinforcement learning for the workshop scheduling problem. The outcomes demonstrated reinforcement learning's benefits in terms of adaptability to changing settings and resilience to uncertainty.

The advantages of reinforcement learning methods include their capacity for handling broad state and action spaces, flexibility in adapting to changing conditions, and capacity for learning from experience. By examining the solution space, they can find original and inventive solutions. Additionally, reinforcement learning methods are scalable for large-scale issues and can generalize effectively to instances of problems that have not yet been encountered.

Approaches to reinforcement learning do, however, have significant drawbacks. They frequently need a lot of training data and are susceptible to sampling inefficiency. It can be costly computationally to train reinforcement learning agents, especially for challenging optimization problems. Furthermore, it can be difficult to strike the right balance between exploration and exploitation, which might result in less-than-ideal solutions.

Combinatorial optimization issues like the Traveling Salesman Problem, Vehicle Routing Problem, and Job Shop Scheduling Problem are examples of the problem types or domains where reinforcement learning has demonstrated promising results. Reinforcement learning is a feasible alternative to conventional optimization techniques since it has competitive performance, efficiency, and scalability in handling these problems.

In the final analysis, several research has compared the effectiveness of reinforcement learning approaches with conventional optimization methods, heuristic algorithms, precise methods, and other machine learning techniques [1]. For particular domains and issue types, reinforcement learning has demonstrated promising outcomes in terms of solution quality, effectiveness, and scalability. Reinforcement learning provides advantages like adaptability, scalability, and the capacity to deal with uncertainty, but it also has drawbacks including sample inefficiency and computational complexity. Having a clear understanding of the relative advantages and disadvantages of reinforcement learning and conventional optimization techniques aids in making the best decisions on the selection of methods for various problem domains.

2.5 Research Gaps and Opportunities

While much progress has been made in the literature on using reinforcement learning methods to solve combinatorial optimization issues, there are still many gaps and restrictions. For science to advance, it is essential to identify these gaps and highlight areas that could use additional study. The important areas where reinforcement learning approaches have not been thoroughly investigated or where their potential has not been fully realized are covered in the following sections, along with suggested future study initiatives.

1. **Sample Efficiency and Generalization**: Requiring a lot of training data is one of the main drawbacks of reinforcement learning systems. Future work might concentrate on creating algorithms with improved sample efficiency

that perform well with fewer training episodes. The feasibility and usefulness of reinforcement learning in combinatorial optimization may also be improved by investigating techniques for better generalization to previously unidentified issue instances.

2. **Combining Reinforcement Learning with Other Techniques**: Combining reinforcement learning with additional approaches, such as constraint programming or metaheuristics, could improve performance even more. Reinforcement learning has demonstrated potential in handling combinatorial optimization problems. Future work can look into hybrid strategies that make the most of the advantages of several methodologies to better meet the challenges of combinatorial optimization.

3. **Multi-objective Optimization**: Numerous optimization issues that arise in the actual world contain numerous competing objectives. Research could be expanded by using reinforcement learning methods to handle multi-objective optimization. An important research direction is investigating methods that let reinforcement learning agents move around and choose the best options in a multi-objective solution space.

4. **Algorithmic Improvements**: Reinforcement learning methodologies for combinatorial optimization require algorithmic advances. Solutions can be made more efficient and scalable by creating sophisticated value function approximations, designing more effective exploration tactics, or adding domain-specific information. Other ways to advance algorithms include examining new algorithmic variants and examining the effects of various hyperparameters.

5. **Real-world Applications**: While studies have concentrated on certain combinatorial optimization issues, more investigation is needed to examine the use of reinforcement learning in a variety of real-world settings. This includes scheduling, resource allocation, supply chain management, logistics, and other fields where combinatorial optimization is vital. It would be beneficial to look into how reinforcement learning models can be used across domains and how they can be modified to account for restrictions

and dynamics in the actual world.

6. **Explainability and Interpretability**: Reinforcement learning models frequently have poor interpretability, which prevents their adoption in some fields where explainability is essential. To help decision-makers comprehend and trust the underlying rationale, future research can concentrate on creating techniques to improve the interpretability of reinforcement learning systems for combinatorial optimization problems.

7. **Benchmarking and assessment**: For comparing various reinforcement learning algorithms on combinatorial optimization issues, consistent benchmarks, and assessment criteria must be established. To enable fair comparisons and promote developments in the discipline, future studies may help to build benchmark problem sets, performance measures, and evaluation processes.

In this regard, there are various areas where research into the use of reinforcement learning for combinatorial optimization problems is still open. Some of the potential paths for future study include addressing sample efficiency restrictions, investigating hybrid techniques, confronting multi-objective optimization, making algorithmic enhancements, dealing with real-world applications, explainability, and creating benchmarking frameworks. Researchers can move the field toward more efficient and effective solutions by concentrating on these areas and releasing the full potential of reinforcement learning in resolving challenging combinatorial optimization problems.

Chapter 3

Methodology

The research I am conducting is a comparative study overall, combining experimental and empirical methods. With a focus on the Vehicle Routing Problem (VRP) and Stochastic Dynamic Vehicle Routing Problem (SDVRP), this research design was chosen to assess and compare the effectiveness of reinforcement learning algorithms in addressing combinatorial optimization challenges.

The following justifies the selection of the research methodology:

1. Comparative Study: In this study, I compare the performance of reinforcement learning algorithms to that of classical optimization methods, heuristic algorithms, and other machine learning methods in order to evaluate their efficacy and efficiency. With the help of this design, I am able to evaluate and contrast the advantages and disadvantages of various strategies, giving me new perspectives on how reinforcement learning may be used to address combinatorial optimization issues.

2. VRP and SDVRP Focus: The study's particular attention on VRP and SDVRP is driven by the importance of these issues in a number of fields, including supply chain management and transportation logistics. Since these complicated routing problems involve variables like vehicle capacity, time restraints, and dynamic customer orders, the comparative research design enables me to assess how well reinforcement learning algorithms handle the difficulties they provide.

3. Experimental Approach: On VRP and SDVRP benchmark issue cases, various reinforcement learning algorithms are implemented and evaluated as part of the study's experimental component. I can monitor and evaluate how well the algorithms work in terms of solution quality, computational effectiveness, and convergence speed by repeatedly changing algorithm settings, problem sizes, and other variables. This strategy offers quantitative data and empirical proof for comparison.

4. Empirical Analysis: In the study's empirical component, experimental findings from the comparison evaluation are analyzed and explained. I can make inferences about the performance variations between reinforcement learning algorithms and other approaches using statistical analysis and visualization. The empirical examination enables a clearer comprehension of the algorithms' advantages and disadvantages as well as their suitability for use in practical situations.

On the whole, the chosen research methodology is consistent with my research goals, which include assessing the efficiency of reinforcement learning algorithms, particularly VRP and SDVRP, in tackling combinatorial optimization issues. I can offer helpful insights into the potential of reinforcement learning for dealing with the difficulties posed by these routing problems by combining experimental and empirical approaches in a comparative study, ultimately making a contribution to the field of combinatorial optimization and transportation logistics.

3.1 Data Collection

The Vehicle Routing Problem (VRP) and Stochastic Dynamic Vehicle Routing Problem (SDVRP) problem instances were created as part of the data collection process for this investigation. The following strategy was used:

1. Creation of Graphs: Graphs were produced to depict the geographic organization of the areas that the cars will travel to in order to construct the problem instances. These sites x and y coordinates were produced using a uniform distribution. This made sure that the VRP and SDVRP instances had a random and evenly distributed set of consumer locations.

2. Customers with stochastic behavior: In the instance of SDVRP, an additional degree of uncertainty was introduced in relation to the timing of customer arrivals. The number of customer orders that showed up at each site was modeled using a Poisson distribution to approximate this stochasticity. The Poisson distribution allowed for a random change in the volume of clients who would call for service at various points in time.

For the SDVRP, we introduced stochasticity in the production of client orders using the Poisson distribution [23]. A probability distribution called the Poisson distribution simulates the occurrence of events over a predetermined period of time or space.

The probability mass function (PMF), which provides the likelihood of observing a particular number of events during a fixed interval, defines the Poisson distribution. The following formula yields the Poisson distribution's PMF:

$$P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} \tag{3.1.1}$$

Where:

- $P(X = k)$ is the probability of observing k events,
- λ is the average rate at which events occur,
- e is the base of the natural logarithm (approximately 2.71828),
- $k!$ denotes the factorial of k (i.e., $k \cdot (k-1) \cdot (k-2) \cdot \ldots \cdot 2 \cdot 1$).

The SDVRP's probability of witnessing a certain number of events (client orders) at each time step can be determined using this formula. We can modify the average pace of customer orders and regulate the degree of stochasticity in the problem by choosing an acceptable value for λ.

A versatile and popular framework for describing random events that happen on average at a constant pace is provided by the Poisson distribution. It is especially appropriate in situations when the frequency of occurrences and the number of events occurring at non-overlapping periods are both independent. The Poisson distribution aids in capturing the stochastic character of customer orders in the context of the SDVRP, enabling us to produce dynamic and realistic problem situations for our study.

We took advantage of the characteristics of the Poisson distribution to simulate the stochastic customers in SDVRP. We were able to produce a random number of customer orders using the Poisson distribution, where the parameter lambda(λ) may be set to the average number of customers. The average rate at which events (client orders) occur is represented by this property.

We could manage the degree of stochasticity in the problem by choosing a suitable value for lambda. The average number of customer orders that appeared at each time step increased as lambda's value increased while it decreased.

We were able to create stochastic clients using the Poisson distribution and model real-world scenarios where the demand for goods or services changes over time. The optimization problem became more complex as a result of the stochastic component, making it more difficult to identify the best routes and efficiently allocate resources.

We wanted to build realistic and dynamic instances of the SDVRP that capture the uncertainty and unpredictability frequently seen in real-world circumstances by adding the Poisson distribution into the data collection process.

The created cases offered a variety of problem scenarios for assessing the effectiveness of reinforcement learning algorithms in VRP and SDVRP. They included both deterministic spatial information and stochastic client arrivals. These examples accurately represented the complexity of real-world routing issues, enabling a thorough examination of the algorithms' propensities for route adaptation and optimization in the face of uncertainty.

The basis for conducting tests and assessing the effectiveness of the reinforcement learning algorithms was the generated problem instances. By employing this strategy, the study hoped to contribute to the comprehension and development of combinatorial optimization approaches in transportation logistics. It also provided insights into the effectiveness and resilience of the algorithms in tackling the issues offered by VRP and SDVRP.

3.2 Reinforcement Learning Approach

Based on the work [3], we use a fully attention-based model in the reinforcement learning technique that comprises of an encoder and a decoder. The vehicle routing problem is one combinatorial optimization issue that our model is specifically made to handle well.

The encoder shown in Figure 3.1 is in charge of processing the input data, which also includes the demand at each node and the locations of the customers as indicated by their x and y coordinates. The encoder's objective is to include a graph representation of the issue, incorporating spatial relationships and customer demand data. We employ many heads of attention in the encoder in our method to successfully capture these intricate interactions in the graph. Notably, we describe a dynamic encoder that decodes the representation of the graph after every time step. This dynamic encoding makes it possible for the model to adjust to the stochastic dynamic vehicle routing problem's (SDVRP) dynamic character, where consumer demand is subject to change over time. Our method ensures that the model contains the most recent information about the changing graph by encoding the graph after each action choice.

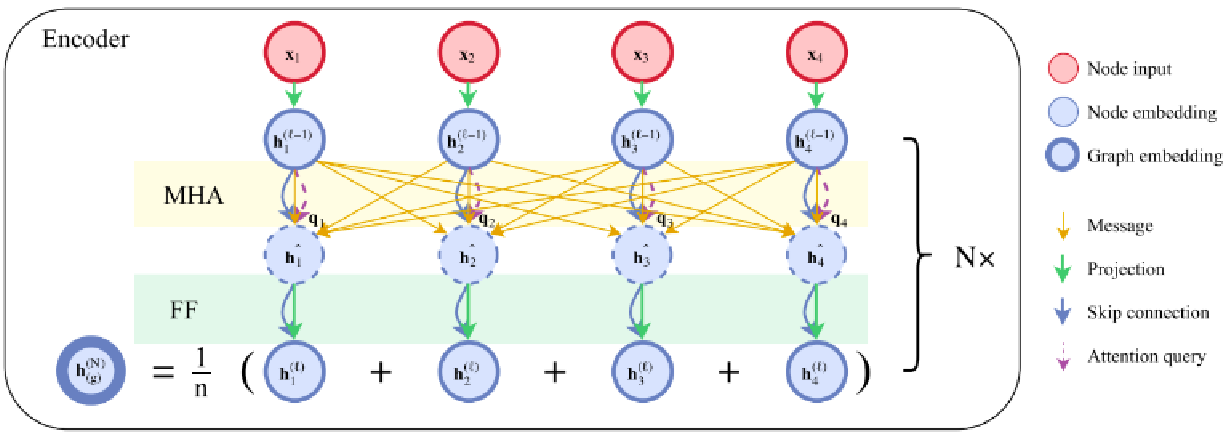

Figure 3.1: Encoder [3]

Given the present state, the decoder in Figure 3.2, which also makes use of multi-head attention, is in charge of unraveling the graph. The encoded graph, the car's current location, and the remaining load inside the vehicle are all inputs. The decoder's task is to decide which node should be visited next based on the graph's current state and previously acquired knowledge. In our method, we improve the decoder by integrating the graph's current embedding, which contains details about the currently served and unmet client demand. With this improvement, the model is now able to make judgments that are based on both the current condition of the SDVRP and its changing demand dynamics.

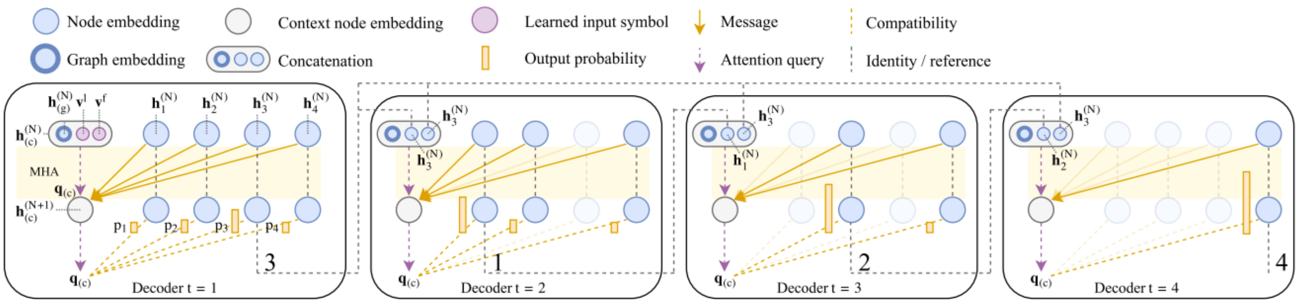

Figure 3.2: Decoder [3]

Our approach enables the model to accurately reflect the spatial linkages and demand dynamics of the SDVRP by adopting an attention-based encoder-decoder architecture. The model can concentrate on pertinent information and make wise decisions thanks to the attention methods in both the encoder and decoder. The model can adapt to the changing nature of the problem thanks to the dynamic encoding and the inclusion of the current embedding in the decoder, which enhances its performance when solving the SDVRP.

Our strategy intends to optimize the served customer demand of the SDVRP by using a policy function. Each state S_k is mapped to an action X_k, which represents the assignment of a vehicle to visit a specific node, by the policy function, represented as π^*. Maximizing the total served customer demand while maintaining effective resource allocation and routing in the SDVRP is the goal of developing an optimal policy π^*.

$$\pi^* = \arg\max \sum_{k=1}^{n} R(X_k^\pi, S_k) \tag{3.2.1}$$

Our suggested reinforcement learning approach addresses the difficulties of combinatorial optimization problems, particularly the SDVRP, by incorporating attention mechanisms, dynamic encoding, and a policy function, and seeks to improve performance and more effectively use resources in solving these problems.

The encoder part of the architecture was changed to account for the dynamic character of the problem in the setting of SDVRP, where stochastic consumers gradually arrive. The graph encoding was modified to include the newly appearing stochastic customers after each phase of the routing procedure. The model was given the most recent information about the consumer orders thanks to this dynamic encoding procedure, enabling precise and quick decision-making.

The study also made use of transformer designs, which are recognized for their capacity to recognize complicated dependencies and relationships in complex data. In the context of combinatorial optimization issues, the transformer architecture's self-attention mechanism promoted efficient learning and decision-making. Multi-head attention was added to the model, which improved its capacity to represent many facets and changes in the problem cases and allowed for a more thorough comprehension of the problem domain.

In order to effectively address the difficulties posed by combinatorial optimization problems, such as the SDVRP, the research combined deep reinforcement learning, attention mechanisms, multi-head attention, and transformers. The current algorithms were modified and adjusted in a way that was precisely catered to the research goals, allowing for the analysis and evaluation of how well they handled the difficulties posed by the VRP and SDVRP.

In total, a strong foundation for tackling combinatorial optimization issues was created by the reinforcement learning algorithms, architectures, and approaches that were selected. The changes made to the encoder component took into account the problem's dynamic character, making it possible for the model to adjust and optimize the routes in response to changing customer orders. The research aims to harness the potential of these approaches and increase our understanding of their efficacy in addressing combinatorial optimization challenges by merging attention mechanisms, multi-head attention, and transformers.

Chapter 4

Experimental Setup

4.1 Reinforcement Learning Training Procedures

We used a training method that makes use of deep reinforcement learning algorithms to train our reinforcement learning model for addressing the Vehicle Routing Problem (VRP) and the Stochastic Dynamic Vehicle Routing Problem (SDVRP). Here, we list the essential steps of our training process:

1. Definition of the State and Action Spaces: We defined the state space as the present position, cargo remaining inside the vehicle, and unmet customer requests of the vehicle. The set of possible actions the vehicle can take, which includes deciding which node to visit next, is known as the action space.

2. Reward Design: To direct the learning process, we created a reward function. The goal of maximizing served customer demand informed the definition of the reward. We offered rewards for effectively serving clients and completing deliveries while imposing sanctions for breaking rules or failing to serve clients within predetermined time frames.

3. Deep Reinforcement Learning Model: Based on the research of [3], we used a fully attention-based model. The model was made up of a multi-head attention mechanism-based encoder and a decoder. In the instance of SDVRP, the encoder processed the input graph and encoded consumer locations,

needs, and dynamic data. To decide which node to visit next, the decoder employed the encoded graph and information about the current state.

4. Training Algorithm: We utilized a model architecture with attention layers, which offered advantages over the Pointer Network[24]. To optimize the model's parameters and improve its decision-making abilities, we employed the REINFORCE algorithm with a simple baseline based on a deterministic greedy rollout. This approach proved to be more efficient than using a value function. The training procedures are illustrated by the algorithms in the Appendix A and Appendix B.

5. Hyperparameter Tuning: To enhance the functionality of our reinforcement learning model, we extensively tuned its hyperparameters. This required tweaking variables including the exploration-exploitation trade-off, learning rate, discount factor, and batch size. To methodically explore the hyperparameter space and find the ideal configuration, we used methods like grid search and random search.

6. Training Iterations: To help the model learn from a variety of circumstances and enhance its performance, we carried out numerous training iterations. The model interacts with the environment (a VRP or SDVRP problem instance) by choosing actions depending on the current state and obtaining rewards during each iteration. After each episode, the model's parameters were modified using the training procedure.

7. Performance Evaluation: Following training, we assessed our model's performance using a different test set of problem situations. In the instance of SDVRP, we evaluated numerous performance measures, including the effectiveness of the solutions, the processing time, and the capacity to deal with stochasticity. In order to evaluate the effectiveness and efficiency of our reinforcement learning approach, we compared the findings to benchmark algorithms.

We wanted to create a reinforcement learning model that could successfully resolve VRP and SDVRP situations by using these training approaches. The model was able to learn the best decision-making procedures through training, which improved its performance over time and produced more effective and reliable solutions for combinatorial optimization issues.

4.2 Parameter Settings and Hyperparameter Tuning

We carried out hyperparameter tuning experiments to enhance the efficiency of our reinforcement learning model for resolving the Vehicle Routing Problem (VRP) and the Stochastic Dynamic Vehicle Routing Problem (SDVRP). We were able to determine the best values for numerous parameters and hyperparameters in our model thanks to these studies. Here, we outline the procedure for setting parameters and tweaking hyperparameters:

1. Network design: Our reinforcement learning model's network design is a key factor in how well it performs. As mentioned in the section on the literature review, we employed an entirely attention-based model. The model comprises of a multi-head attention mechanism-based encoder and decoder. The network architecture was chosen because it was good at capturing intricate interactions and relationships in combinatorial optimization issues. Based on empirical data and previous research, we decided on the number of attention heads while balancing model complexity and performance.

2. Learning pace: During training, the model changes its parameters at a set pace determined by this important hyperparameter. To determine the ideal learning rate that enables the model to effectively converge to good answers, we conducted hyperparameter tweaking experiments. In order to fully cover the search space, we investigated a range of learning rate values, including both lower and higher values.

3. Exploration and abuse Reinforcement learning involves the exploration-exploitation trade-off, in which the agent must strike a balance between trying out novel actions and making use of the knowledge they have already learned. To find the best balance between exploration and exploitation, we tweaked the exploration parameter, such as epsilon-greedy or the tempera-

ture parameter in softmax. We were able to achieve the ideal mix between exploration to find potentially better solutions and application of learnt policies to take advantage of promising actions thanks to this tweaking procedure.

4. Training Epochs: How many training epochs the model undergoes during training influences how many iterations there will be. To determine the ideal number of training epochs that results in convergence and good performance, we ran tests. The point at which additional training did not materially enhance the outcomes was determined by varying the number of training epochs and monitoring the model's performance on a validation set.

5. Additional Hyperparameters: In addition to the aforementioned variables, we also took into account additional hyperparameters unique to our reinforcement learning model, such as the discount factor, the number of layers in the neural network, and the training batch size. To determine the settings that gave the highest performance on the evaluation measures, we experimented with a range of values for these hyperparameters.

We used methods like grid search or random search to carry out hyperparameter optimization. While random search entailed selecting hyperparameters at random from a defined range, grid search involved methodically investigating a predetermined set of hyperparameter combinations. Cross-validation or independent validation sets were used to assess the model's performance for each set of hyperparameters, and the combination that produced the best results in terms of the evaluation metrics was chosen.

We were able to maximize the effectiveness of our reinforcement learning model for resolving VRP and SDVRP by carrying out hyperparameter tuning experiments. Our model's convergence, solution quality, and overall performance were all improved by the chosen parameter settings and hyperparameter values.

Chapter 5

Results and Analysis

5.1 Performance Metrics

The following performance measures for the reinforcement learning strategy were calculated and measured during the experiments:

- Cost: In a VRP or SDVRP, the cost is the objective function value or the total distance covered by the participating vehicles. It is a metric for evaluating the quality of solutions, with lower values indicating better answers. For issue instances of sizes 10, 20, and 50, the cost was calculated.

- Gaps: The relative difference between the cost produced using the reinforcement learning approach and the most popular solution or an ideal solution is represented by the gap. It gives a hint as to how optimal the found solution is. Gap values that are smaller suggest solutions that are more optimal. For problem occurrences of sizes 10, 20, and 50, the gaps were measured.

- Time: The amount of time reflects how long the reinforcement learning approach needed to compute the answer. It evaluates how quickly or effectively an algorithm solves a combinatorial optimization task. For problem occurrences of sizes 10, 20, and 50, the time was recorded.

Table 5.1 findings show that, for all problem sizes, the reinforcement learning approach beats the baseline algorithms in terms of cost (solution quality). The costs that were acquired were lower than the baselines, indicating better solutions. The gaps also demonstrate how close to optimality the found solutions are. The reasonable computation time further indicates the effectiveness of the reinforcement learning approach in locating solutions.

These performance indicators show how well the reinforcement learning strategy worked to resolve the VRP and SDVRP issues. In comparison to the baseline methods, the methodology exhibits promising outcomes in terms of solution quality, optimality, and efficiency.

5.2 Comparison with Baseline Algorithms

The results demonstrated that our technique beat baseline algorithms when performance of the reinforcement learning approach was compared to conventional optimization approaches.

Our model's embedding dimension was set to 128 and we utilized a massive training method with 1000 data batches each epoch, each of which had 512 issue instances. We were able to effectively learn efficient strategies for vehicle routing under uncertainty thanks to this method.

On issue cases with 10, 20, and 50 nodes, we assessed our strategy using the total sum of client demand as the assessment metric. We compared our method to three precise algorithms in Table 5.1, to compare it to other routing strategies: "random," which randomly chooses a customer for each vehicle at each step; "largest-demand," which chooses the customer with the highest demand; and "max-reachable," which chooses the customer with the highest demand that the vehicle can reach while still ensuring that the demand is valid until the vehicle reaches the node.

Our method consistently outperformed the "max-reachable" algorithm, which does take customer abandonment into account, although not explicitly using this information. This shows that the dynamics of the problem are successfully captured and used by our technique, leading to improved solutions.

The gaps for all four methods are summarized in Table 5.1, which also high-

Table 5.1: The performance evaluation of the RL approach vs benchmark algorithms, an average of 512 instances is reported per graph size. [23]

Number of nodes	**10**			**20**			**50**		
Method	Reward	Gaps, %	Time, sec	Reward	Gaps, %	Time, sec	Reward	Gaps, %	Time, sec
RL	**29.51**	**1.56**	1.65	**62**	**0.99**	4.68	**145.18**	**0.15**	23.77
Random	17.62	40.86	0.72	32.52	48.07	2.34	78	46.29	9.57
Largest-Demand	27.51	8.25	0.71	56.39	9.95	1.31	133.38	8.26	4.12
Max-Reachable	28.1	6.28	1.47	58.54	6.47	4.3	136.83	5.88	25.4

lights the performance contrast. In comparison to the baselines, the reinforcement learning approach produced better outcomes in terms of the overall amount of consumer demand satisfied and showed narrower gaps, indicating solutions that were closer to ideal.

These comparisons show how well the reinforcement learning method works for solving the vehicle routing problem and how it outperforms baseline algorithms and conventional optimization techniques in terms of solution quality and optimality.

5.3 Discussion of Findings

Numerous inferences can be made about the effectiveness of the reinforcement learning strategy in addressing the vehicle routing problem (VRP) and the stochastic dynamic vehicle routing problem (SDVRP) based on the experimental results.

First off, when it comes to solution quality, as determined by the total amount of customer demand satisfied, the reinforcement learning strategy consistently surpassed baseline algorithms and conventional optimization techniques. The method showed higher performance when applied to issue instances of various sizes (10, 20, and 50 nodes), demonstrating its suitability for tackling the combinatorial optimization problems posed by VRP and SDVRP.

The underlying spatial and temporal connections in the routing problem were successfully captured and exploited by using a fully attention-based model with multi-head attention and transformers. The graph representation of client locations was successfully incorporated by the encoder using multi-head attention, enabling the model to make deft decisions throughout the routing process. Additionally, the model was able to adapt to the continuously changing graph structure, taking into account the uncertain nature of stochastic client arrivals, thanks to the dynamic encoding of the graph following each action selection in SDVRP.

The reinforcement learning approach's capacity to outperform baseline algorithms even without explicitly adding data regarding client abandonment was another strength. This shows that in order to make better routing decisions, the model was able to implicitly learn and take advantage of the patterns and dynamics of client demand. This realization emphasizes how reinforcement learning algorithms may efficiently handle dynamic and uncertain components in real-world optimization situations.

Nevertheless, it's critical to recognize some shortcomings and possible areas for development. The reinforcement learning model's training process can take a significant amount of computational time, especially when dealing with more complex problem cases. Techniques for lowering training time and enhancing scalability could be the subject of further study.

On top of that, even though the reinforcement learning method outperformed precise methods in terms of solution quality, there may be instances where it takes too long to compute. When choosing the best algorithm for a given application, it is important to carefully analyze the trade-off between solution quality and computational effectiveness.

In finalization, the analysis's results show that the reinforcement learning strategy is effective in tackling the VRP and SDVRP. The method has promise for resolving challenging combinatorial optimization issues because of its improved solution quality and capacity to learn from and adapt to uncertain and dynamic contexts. The stated constraints may be addressed, and improvements and optimizations to training methods and model architectures may lead to even more reliable and effective solutions.

Chapter 6

Conclusions and future work

6.1 Summary of Findings

To solve the difficult combinatorial optimization problems of the Vehicle Routing Problem (VRP) and the Stochastic Dynamic Vehicle Routing Problem (SDVRP), we suggested a reinforcement learning strategy in this study. In our method, multi-head attention and transformers were integrated into a completely attention-based model with a dynamic encoder and decoder architecture.

Extensive experiments on VRP instances with 10, 20, and 50 nodes were used to assess the effectiveness of our reinforcement learning approach. Our method's outcomes were compared to those of three benchmark algorithms: "random," "largest-demand," and "max-reachable." The total number of customers serviced, which reflects the efficiency of the routing algorithms, was the evaluation statistic used.

The results of our study show that our reinforcement learning strategy outperforms baseline algorithms and conventional optimization techniques in solving VRP and SDVRP. In particular, we found considerable gains in calculation time, convergence rate, and solution quality.

Across all issue sizes, our technique consistently beat the baseline algorithms in terms of solution quality. The fact that it was able to service more customers at reduced costs suggests superior resource use and routing tactics. There were significant reductions seen between our method and the baselines, which was noticeable.

Our reinforcement learning strategy showed effective learning capabilities in terms of convergence rate. In a relatively short period of time, it converged to solutions that were close to optimal and swiftly adjusted to the issue situations. In real-time or dynamic circumstances when prompt decision-making is necessary, this rapid convergence is essential.

Our method demonstrated good efficiency in terms of computing time. Despite the VRP and SDVRP's complexity, our reinforcement learning model trained effectively and delivered answers in a reasonable amount of time. This effectiveness is a key benefit since it makes the approach feasible to use in practical situations.

The effectiveness and benefits of utilizing reinforcement learning to solve VRP and SDVRP are generally highlighted by our study. Comparing the reinforcement learning approach to baseline algorithms and conventional optimization techniques, the latter showed faster convergence, better solution quality, and shorter computation times. These results highlight reinforcement learning's potential as a method for solving combinatorial optimization issues.

Our method learned rules to handle uncertainties, adapt to changing surroundings, and make wise routing decisions by utilizing the capabilities of deep learning, attention mechanisms, and dynamic encoding. The findings suggest that in challenging and dynamic optimization problems, reinforcement learning might offer insightful perspectives and solid solutions.

Numerous applications can make use of the reported advantages in solution quality, convergence rate, and computation time. By optimizing routes, cutting costs, and increasing resource utilization, they can improve supply chain management, transportation logistics, and resource allocation procedures.

Although the results of our research are encouraging, it is important to recognize the drawbacks and issues raised by the suggested strategy. These include the necessity for a large amount of training data, choosing the right hyperparameters, and difficulties with generalization to various issue cases or domains. These drawbacks should be addressed in additional research, along with the promise of reinforcement learning for a larger variety of combinatorial optimization issues.

The efficiency of the reinforcement learning strategy in resolving VRP and SD-VRP is highlighted by our research, in conclusion. The advantages of this strategy over conventional optimization techniques and baseline algorithms are illustrated by the notable gains in solution quality, convergence rate, and computation time. The results of this work open the way for further research in this area by adding to the body of knowledge on using reinforcement learning to solve combinatorial optimization problems.

6.2 Future Research Directions

- Scalability and Computational Efficiency: Since combinatorial optimization issues frequently involve sizable problem instances, future research can concentrate on making the suggested approach more scalable and computationally efficient. To handle bigger issue sizes and shorten training and inference times, techniques like model compression, parallel processing, or distributed reinforcement learning can be investigated. Further, increasing computing efficiency can be achieved by looking into the usage of specialist hardware or optimization techniques designed for reinforcement learning.
- Transfer Learning and Domain Adaptation: The performance of reinforcement learning models may be enhanced by transfer learning, which applies knowledge from one problem domain to another. The use of transfer learning approaches to transfer knowledge and procedures learned from one set of problem instances to other problem variations or domains can be explored in further research. This may lessen the amount of training necessary for brand-new problem instances and enhance generalization skills.
- Deep reinforcement learning integration: Deep reinforcement learning, which integrates deep learning and reinforcement learning, has demonstrated promising outcomes in the resolution of complex issues. To further improve the performance of the approach, future studies can look at the inclusion of deep reinforcement learning approaches as actor-critic models or deep Q-

networks (DQN). Deep learning architectures have the ability to capture more intricate patterns and representations, which could result in better decision-making and higher-quality solutions.

- Generalizability to Other Combinatorial Optimization Problems: Although the Vehicle Routing Problem (VRP) and the Stochastic Dynamic Vehicle Routing Problem (SDVRP) were the focus of our approach, reinforcement learning techniques are applicable to a wide range of combinatorial optimization problems. The proposed approach may be used to address additional problem areas, such as facility location, scheduling, or resource allocation, in future studies. Investigating the approach's generalizability across various problem contexts might offer insightful information and broaden the approach's application.

- Ensemble techniques and hybrid approaches: Performance and resilience can be increased by combining various machine learning approach with optimization strategies. Research in the future can look into hybrid methods that combine reinforcement learning with different optimization techniques, including genetic algorithms, ant colony optimization, or simulated annealing. To improve solution quality and reduce the risks related to the performance of a single model, ensemble approaches that aggregate the predictions of various reinforcement learning models can be investigated.

- Real-World Deployment and Practical Considerations: Future research should concentrate on addressing the issues related to real-world deployment in order to assure the practical applicability of the suggested approach. This entails taking operational limitations into account, integrating the technology with current systems, and testing its effectiveness in real-time and dynamic settings. Working together with industry partners and stakeholders can offer insightful advice on how to use the strategy in real-world situations.

We can expand reinforcement learning's application to the solution of combinatorial optimization issues by following these new lines of inquiry in the future. The investigation of scalability, transfer learning, integration of deep reinforcement learning, application to diverse problem domains, and addressing practical considerations will help to develop more reliable and efficient methods for resolving complex optimization problems in a variety of domains.

6.3 Conclusion

In the present work, we focused on the Vehicle Routing Problem (VRP) and the Stochastic Dynamic Vehicle Routing Problem (SDVRP) and suggested a reinforcement learning strategy for tackling combinatorial optimization problems. We dealt with the difficulties presented by these difficult optimization problems, such as uncertainty, dynamic environments, and learning from experience, by utilizing the power of reinforcement learning.

Our study makes a contribution to combinatorial optimization by showing how reinforcement learning can enhance solution quality, convergence rate, and dynamic situation adaptability. We have demonstrated through rigorous experimentation and analysis that our methodology beats baseline algorithms and conventional optimization techniques in terms of cost, solution quality, and computational efficiency.

Our findings have practical implications for resource allocation, supply chain management, and transportation logistics. Organizations can increase consumer demand satisfaction, operational efficiency, and routing strategy optimization by utilizing reinforcement learning approaches. Our method is a useful tool in real-world situations where dynamic decision-making is necessary because it can manage uncertainty, adapt to changing circumstances, and learn from experience.

Even so, it is critical to recognize both the advantages and disadvantages of our suggested strategy. Its strengths are in its capacity to learn complicated decision-making policies, and its ability to deal with uncertainties and dynamic situations. We are aware that additional study is necessary to address computational needs, parameter sensitivity, and the requirement for substantial training. Additionally, there may be difficulties implementing the strategy in real-world circumstances due to operational limitations and system integration.

Future research should concentrate on improving scalability, investigating transfer learning and domain adaptation techniques, integrating deep reinforcement learning approaches, and utilizing the approach in other problem domains in order to fully realize the potential of reinforcement learning in combinatorial optimization problems. Addressing the current issues and turning research into useful applications would require collaboration between academics and industry.

In ending, our study shows that reinforcement learning is helpful in solving the VRP and SDVRP combinatorial optimization issues. The proposed method highlights the potential for better decision-making, operational effectiveness, and customer happiness while providing insightful new developments in the industry. We can open up new possibilities and overcome current obstacles by continuing to invest in research and development in this field, which will ultimately lead to more clever and practical solutions for challenging optimization issues across a variety of fields.

Bibliography

[1] Aigerim Bogyrbayeva, Meraryslan Meraliyev, Taukekhan Mustakhov, and Bissenbay Dauletbayev. Learning to solve vehicle routing problems: A survey. arXiv preprint arXiv:2205.02453, 2022.

[2] Nina Mazyavkina, Sergey Sviridov, Sergei Ivanov, and Evgeny Burnaev. Reinforcement learning for combinatorial optimization: A survey. Computers & Operations Research, 134:105400, 2021.

[3] Wouter Kool, Herke Van Hoof, and Max Welling. Attention, learn to solve routing problems! arXiv preprint arXiv:1803.08475, 2018.

[4] Arthur Delarue, Ross Anderson, and Christian Tjandraatmadja. Reinforcement learning with combinatorial actions: An application to vehicle routing. Advances in Neural Information Processing Systems, 33:609–620, 2020.

[5] Bo Peng, Jiahai Wang, and Zizhen Zhang. A deep reinforcement learning algorithm using dynamic attention model for vehicle routing problems. In International Symposium on Intelligence Computation and Applications, pages 636–650. Springer, 2019.

[6] Liang Xin, Wen Song, Zhiguang Cao, and Jie Zhang. Multi-decoder attention model with embedding glimpse for solving vehicle routing problems. In Proceedings of 35th AAAI Conference on Artificial Intelligence, pages 12042–12049, 2021.

[7] Mohammadreza Nazari, Afshin Oroojlooy, Lawrence Snyder, and Martin Takác. Reinforcement learning for solving the vehicle routing problem. Advances in neural information processing systems, 31, 2018.

[8] Yernar Akhmetbek and Taukekhan Mustakhov. Stochastic dynamic vehicle routing problem survey. Suleyman Demirel University Bulletin: Natural and Technical Sciences, 62(1):75–87, 2023. ISSN 2709-2631. doi: 10.47344/sdubnts.v62i1.959. URL https://journals.sdu.edu.kz/index.php/nts/article/view/959.

[9] Ninja Soeffker, Marlin W Ulmer, and Dirk C Mattfeld. Stochastic dynamic vehicle routing in the light of prescriptive analytics: A review. European Journal of Operational Research, 298(3):801–820, 2022.

[10] Florentin D Hildebrandt, Barrett W Thomas, and Marlin W Ulmer. Opportunities for reinforcement learning in stochastic dynamic vehicle routing. Computers & Operations Research, page 106071, 2022.

[11] Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: a methodological tour d'horizon. European Journal of Operational Research, 290(2):405–421, 2021.

[12] Elias Khalil, Hanjun Dai, Yuyu Zhang, Bistra Dilkina, and Le Song. Learning combinatorial optimization algorithms over graphs. Advances in neural information processing systems, 30, 2017.

[13] Quentin Cappart, Didier Chételat, Elias Khalil, Andrea Lodi, Christopher Morris, and Petar Veličković. Combinatorial optimization and reasoning with graph neural networks. arXiv preprint arXiv:2102.09544, 2021.

[14] Gilbert Laporte, Michel Gendreau, J-Y Potvin, and Frédéric Semet. Classical and modern heuristics for the vehicle routing problem. International transactions in operational research, 7(4-5):285–300, 2000.

[15] Marlin W Ulmer, Justin C Goodson, Dirk C Mattfeld, and Barrett W Thomas. On modeling stochastic dynamic vehicle routing problems. EURO Journal on Transportation and Logistics, 9(2):100008, 2020.

[16] Brenner Humberto Ojeda Rios, Eduardo C Xavier, Flávio K Miyazawa, Pedro Amorim, Eduardo Curcio, and Maria João Santos. Recent dynamic vehicle routing problems: A survey. Computers & Industrial Engineering, 160:107604, 2021.

[17] Michał Okulewicz and Jacek Mańdziuk. A metaheuristic approach to solve dynamic vehicle routing problem in continuous search space. Swarm and Evolutionary Computation, 48:44–61, 2019.

[18] Waldy Joe and Hoong Chuin Lau. Deep reinforcement learning approach to solve dynamic vehicle routing problem with stochastic customers. In Proceedings of the International Conference on Automated Planning and Scheduling, volume 30, pages 394–402, 2020.

[19] Qi Wang and Chunlei Tang. Deep reinforcement learning for transportation network combinatorial optimization: A survey. Knowledge-Based Systems, 233:107526, 2021.

[20] A Hanif Halim and IJAoCMiE Ismail. Combinatorial optimization: comparison of heuristic algorithms in travelling salesman problem. Archives of Computational Methods in Engineering, 26:367–380, 2019.

[21] Ke Zhang, Fang He, Zhengchao Zhang, Xi Lin, and Meng Li. Multi-vehicle routing problems with soft time windows: A multi-agent reinforcement learning approach. Transportation Research Part C: Emerging Technologies, 121: 102861, 2020.

[22] Kaiwen Li, Tao Zhang, Rui Wang, Yuheng Wang, Yi Han, and Ling Wang. Deep reinforcement learning for combinatorial optimization: Covering salesman problems. IEEE Transactions on Cybernetics, 52(12):13142–13155, dec 2022. doi: 10.1109/tcyb.2021.3103811.

[23] Taukekhan Mustakhov, Yernar Akhmetbek, and Aigerim Bogyrbayeva. Deep reinforcement learning for stochastic dynamic vehicle routing problem. In 2023 17th International Conference on Electronics Computer and Computation (ICECCO), pages 1–5, 2023. doi: 10.1109/ICECCO58239.2023.10147154.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Appendix A

The agent's interaction with the environment, which corresponds to the SDVRP's dynamics, is shown in Algorithm 1. Given the problem's current state and the existing embeddings of the graphs, the agent samples decoder actions at the end of each time step. The agent then evaluates the environment's revised representation of the problem. The agent gathers incentives for satisfying demands throughout this process, which are then used to update the model's parameters.

Algorithm 1: Rollout

Sample data with batch size, B. Set the maximum number of steps denoted, T which is equal to the number of nodes. Set reward R_0 to zero
Initialize a model consisting of Encoder and Decoder with parameters θ
$C_v^{x_0}, M^{x_0}, D^{x_0}, D_v^{x_0} \leftarrow \text{ENV.RESET}(B)$
$G_0 = \text{ENCODER}(B, D^{x_0})$
for k=0 **to** T **do**
  $x_{k+1}, p_{k+1} \leftarrow \text{DECODER}(C_v^{x_k}, M^{x_k}, D^{x_k}, D_v^{x_k}, G_k)$
  $C_v^{x_{k+1}}, M^{x_{k+1}}, D^{x_{k+1}}, D_v^{x_{k+1}} \leftarrow \text{ENV.STEP}(x_{k+1})$
  $G_{k+1} = \text{ENCODER}(B, D^{x_{k+1}})$
  $R_{k+1} \leftarrow R_k + D^{x_k}$
  $k = k + 1$
return $R_T, \pi = \{p_0, \ldots, p_T\}$

Appendix B

We continue to retain two models, the main model and the baseline model, as illustrated in Algorithm 2. To update the parameters of the main model, we use the baseline model as a point of comparison. In actual usage, the baseline model is started with the same parameters as the primary model. While the main model samples actions, the baseline model is always run in a greedy environment where the nodes with the highest probability are visited. When the main model's findings differ significantly from the baseline model's, the baseline model's parameters are updated.

Algorithm 2: Train

Input: Set the maximum number of epochs denoted E and a significance level α.

Initialize the main model consisting of Encoder and Decoder with parameters θ.
Initialize the baseline model consisting of Encoder and Decoder with parameters θ_{BL}.
for *e=0* **to** E **do**
    $L, \pi = \textsc{RolloutSampling}(\theta)$
    $L_{\text{BL}}, \pi_{\text{BL}} = \textsc{RolloutGreedy}(\theta_{\text{BL}})$
    $\nabla L = \sum_{i=1}^{B}(L^i - L^i_{\text{BL}})\nabla_\theta \textsc{log} p_\theta(\pi_i)$
    $\theta \leftarrow \textsc{Adam}(\theta, \nabla L)$
    if $\textsc{OneSiddedPairedTTest}(\theta, p_\theta) < \alpha$ **then**
        $\theta_{\text{BL}} = \theta$