计算机工程与应用
*Computer Engineering and Applications*
ISSN 1002-8331,CN 11-2127/TP

# 《计算机工程与应用》网络首发论文

题目：　强化学习在车辆路径问题中的研究综述
作者：　牛鹏飞，王晓峰，芦磊，张九龙
网络首发日期：　2021-10-21
引用格式：　牛鹏飞，王晓峰，芦磊，张九龙．强化学习在车辆路径问题中的研究综述[J/OL]．计算机工程与应用.
https://kns.cnki.net/kcms/detail/11.2127.TP.20211021.1302.014.html

网络首发时间：2021-10-21 16:42:10
网络首发地址：https://kns.cnki.net/kcms/detail/11.2127.TP.20211021.1302.014.html



# 强化学习在车辆路径问题中的研究综述

牛鹏飞[1]，王晓峰[1,2]，芦磊[1]，张九龙[1]

1.北方民族大学 计算机科学与工程学院，银川 750021
2.北方民族大学 图像图形智能处理国家民委重点实验室,银川 750021

**摘 要：**车辆路径问题是物流运输优化中的核心问题，目的是在满足顾客需求下得到一条最低成本的车辆路径规划。但随着物流运输规模的不断增大，车辆路径问题求解难度增加，并且对实时性要求也不断提高，已有的常规算法不再适应实际要求。近年来，基于强化学习算法开始成为求解车辆路径问题的重要方法，本文在简要回顾常规方法求解车辆路径问题的基础上，重点总结基于强化学习求解车辆路径问题的算法，并将算法按照基于动态规划、基于价值、基于策略的方式进行了分类。最后对该问题未来的研究进行了展望。

**关键词**：车辆路径问题；马尔科夫决策过程；强化学习；深度强化学习
**文献标志码**:A 中图分类号:TP301 doi：10.3778/j.issn.1002-8331.2108-0467

## Surveyon Vehicle Reinforcement Learning in Routing Problem

NIU Pengfei[1], WANG Xiaofeng[1,2], LU Lei[1], ZHANG Jiulong[1]

1. Country Department of Computer Science, North Minzu University, Yinchuan 750021, China
2.The Key Laboratory of Images & Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan 750021, China

**Abstract:** Vehicle routing problem is the key technologies in the field of logistics research. Its purpose is to get a lowest cost vehicle routing plan while meeting the customer's needs. However, with the increase of problem sizein logistics transportation, the real-time requirement of solving vehicle routing problem is increasing, and the traddi-tional algorithm can not realize the requirements of the industry gradually. For decades, a number of new methods using reinforcement learning and deep reinforcement learning to solve vehicle routing problem. Base on simple anaysis of conventional methods for solving vehicle routing problem, this review summaries the current algorithms for solying vehicle routing problem based on reinforcement learning. Reinforcement learning algorithms are divided to Dynamic programming, value-based and policy-based. This paper summarizes the theoretical foundation and studying status. Finally, the future development direction of vehicle routing problem based on reinforcement learn-ing and deep reinforcement learning is prospected.

**Key words:** vehicle routing problem; markov decision process; reinforcement learning; deep reinforcement learning

随着经济社会快速发展及交通基础设施的不断完善，城市物流业是当今社会关注的一个重要话题。2020 年全国快递业务量突破 750 亿件，随着构建新发展格局的加快，未来我国快递业务量仍会保持较快的

**基金项目：**国家自然科学基金(62062001,61762019,61862051,61962002);宁夏自然科学基金项目(2020AAC03214, 2020AAC03219,2019AAC03120, 2019AAC03119);北方民族大学重大专项资助(ZDZX201901)。

**作者简介：**牛鹏飞(1997-),男,硕士研究生,CCF 会员，研究方向为组合优化问题;王晓峰(1980-)男,博士,副教授,CCF 会员,研究方向为算法分析与设计、可计算性与计算复杂性;芦磊(1995-)，男，硕士研究生,研究方向为算法分析与设计;张九龙(1997-),男,硕士研究生, CCF 会员,算法设计与分析。

增长。物流业的快速发展使得对超大型物流系统的快速调度提出了更高的要求。车辆路径问题((Vehicle Routing Problem，VRP) 是在车辆数一定的条件下，尽量缩短车辆行驶距离，找到一组总成本最小的路线。同时，根据优化目标不同，可以加入不同约束从而满足不同种类问题的实际需求。

车辆路径问题作为一个众所周知的组合优化问题，最早由 Dantzig 和 Ramser [1]于 1959 年作为卡车调度问题提出的，并被 Lenstra 和 Kan [2]证明是 NP-难问题。经典的车辆路径问题被定义为：有一个仓库（Depot）节点和若干个客户（Customers）节点，已知各个节点在二维空间上的位置和客户的需求，在满足约束条件下，使得车辆从仓库节点出发访问客户节点，满足客户需求，最后返回仓库。在不考虑负载的情况下，VRP 等价于旅行商问题，应用到现实生活中，研究较多的是有容量约束的车辆路径问题（CVRP）[3]。当客户的需求不定时，产生了随机车辆路径问题（SVRP）[4]；当客户对需求提出时间要求时，产生了带时间窗的车辆路径问题（CVRPW）[5]；针对客户当日要求交付的需求，产生了当日交付的车辆路径问题（SDDVRP）[6]。关于 VRP 的详细描述见文献[7]。

多年来大量国内外学者致力于 VRP 的研究，目前求解 VRP 的主要方法分为常规算法和基于强化学习的算法，其中常规算法包括精确算法、启发式算法等。基于强化学习的算法主要包括基于马尔科夫决策过程的强化学习和近年来方兴未艾的深度强化学习。

本文将首先简略概述基于常规算法求解 VRP 的各类算法，再对基于强化学习求解 VRP 的各类模型进行详细的介绍。

## 1 基于常规方法求解 VRP 技术

目前求解 VRP 的常规算法包括精确算法、启发式算法和元启发式算法。

1）精确算法主要包括线性规划法[8]、分支限界法[9]等。这类算法适用于 VRP 规模较小、结构简单的情况，当 VRP 中有较多约束条件时，精确算法无法在有效时间内得到问题的最优解。

2）启发式算法主要包括节约法[10]、线性节约法[11]和插入检测法[12]等。这类算法适用于规模较大的 VRP，面对 CVRP、CVRPTW 等这些有较多约束条件的 VRP 变种问题时，该类算法仍较为快速求解，具有求解效率高、占用内存少的优势，因为该类算法改进目标一直是求解速度，因而问题规模增大时无法得到最优解。

3）元启发式算法主要包括模拟退火算法[13]、禁忌搜索算法[14]、基于群思想的算法[15-18]等。这类算法具有求解速度快、效率高的特点。但这类算法在求解 VRP 时容易陷入局部最优而无法得到全局最优解，以及不容易收敛等问题。

表 1 对上述求解 VRP 的各类常规算法的缺点进行了对比。

**表 1 求解车辆路径问题的常规方法优缺点对比**

Table 1 Comparison of advantages and disadvantages of conventional approaches applied to VRP

| 分类 | 经典方法 | 适用场景 | 优点 | 缺点 |
|---|---|---|---|---|
| 精确算法 | 分支限界法[9]<br>线性规划法[8] | 结构简单的小规模 VRP | 描述简单<br>容易实现 | 问题规模增大时无法得到最优解；问题中约束较多时无法求解 |
| 启发式算法 | 节约法[10]<br>线性节约法[11]<br>插入检测法[12] | CVRP、CVRPW 等约束较多的复杂 VRP | 求解效率高，占用内存少 | 解的质量不高，容易陷入局部最优，问题规模大时无法求解 |
| 元启发式算法 | 禁忌搜索法[13]<br>模拟退火算法[14]<br>蚁群算法[15]<br>粒子群算法[18] | 大规模 VRP | 描述简单<br>易于实现<br>效率高 | 容易陷入局部最优，无法得到全局最优解，收敛速度慢 |

## 2 强化学习概述

### 2.1 强化学习基础

强化学习（ReinforceLearning，RL）是人工智能的一个重要分支，它不需要监督信号来进行学习，而是依赖个体（Agent）在环境（Environment）中的反馈回报信号，依据反馈回报信号对个体的状态和行动进行更正，使得个体逐步实现奖励（Reward）的最大化，从而使得强化学习具有较强的自主学习能力。强化学习的描述见图 1。

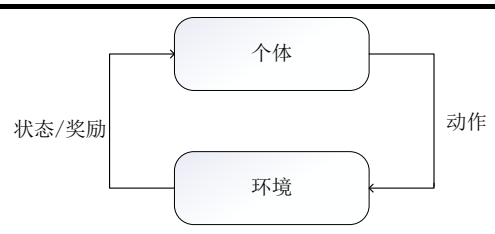


**图 1 强化学习示意图**

Fig.1 Schematic diagram of Reinforce Learning

### 2.2 强化学习算法分类

对强化学习算法的分类可以根据有无模型分为基

于模型（Model-Based）和无模型（Model-Free）的学习算法。在求解 VRP 中常见的基于模型的学习方法有动态规划法；无模型的学习算法主要有基于价值的时序差分算法[18]、Q-learning 算法[19]、DQN 算法[20]等；基于策略的 REINFORC 算法[21]，价值和策略相结合的 Actor-Critic 算法[22]、Advantage Actor-Critic 算法等。图 2 总结了一些已经应用到 VRP 求解中的经典强化学习算法。

## 3 基于模型的算法

在强化学习中“模型”指环境，基于模型的强化学习算法意为通过预先给定或通过学习的方式得到 MDP 模型。最典型的给定环境模型方法是打败围棋冠军柯洁的阿尔法狗算法，通过学习的环境模型方法有 WorldModels 算法[23]。在 VRP 求解中运用最多的基于模型的强化学习算法为动态规划算法，及由动态规划算法衍生出来的近似动态规划算法和深度神经网络动态规划算法。基于模型的算法通过 MDP 模型预测以后可能的状态 $S$ 和动作 $A$，从而对个体行动提供指导，但在现实生活中环境模型可能很复杂以至于难以建模。

### 3.1 动态规划算法

动态规划算法是由美国数学家 Bellman 在研究多阶段决策过程的优化问题时提出的，“动态规划”一词中“动态”意为问题是可以通过一个个子问题去求解，“规划”意为优化策略。在给定一个用马尔科夫决策过程描述的完备环境模型下，其可以计算最优的模型。在强化学习中，动态规划算法的目的是使用价值函数求解最优策略，常规的动态规划算法因“维数灾难”无法有效的求解强化学习问题，但后续其他的方法都是对动态规划算法的改进和近似。运用动态规划算法求解强化学习问题就是求解给定策略 $\pi$ 时对应的价值 $V_*(S)$。价值 $V_*(S)$ 表示为：

$$V_{k+1}(s)=\sum_{a\in A}\pi(a\mid s)(R_s^a+\gamma\sum_{s\in S}P_{ss'}^aV_k(s')) \tag{1}$$

公式表示 $k+1$ 轮的价值 $V_{k+1}(s)$ 可由前 $k$ 轮的 $V_k(S)$ 出来，策略 $\pi(a\mid s)$ 为给定状态 $s$ 时选择动作 $a$ 的概率，$R_s^a$ 为给定状态 $s$ 时选择动作 $a$ 的奖励，$\gamma$ 为折扣率，$\sum_{s\in S}P_{ss'}^aV_k(s')$ 为下一步状态的概率乘以价值函数之和。

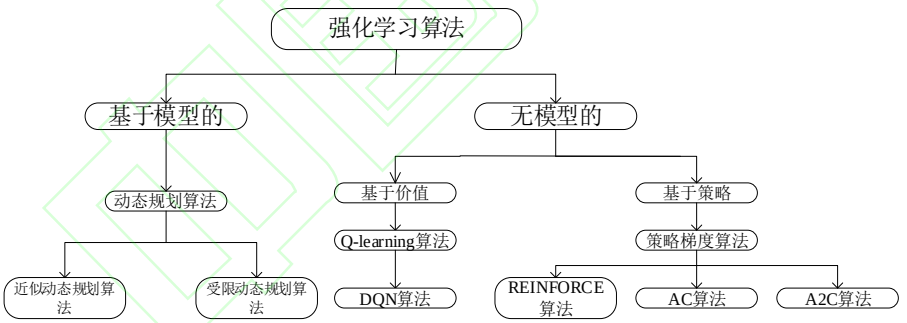


**图 2 强化学习算法分类图**

Fig.2 Classification diagram of reinforcement learning algorithm

#### 3.1.1 基于近似动态规划的方法

Secomandi 等人[24]将首次神经网络近似动态规划（NDP）方法应用到求解带有随机需求的 VRP 中，NDP 是在动态规划中使用神经网络对策略进行近似的新模型，实验结果表明在有随机需求的 VRP 中，基于 rollot 策略的 NDP 算法的性能要优于近似策略迭代的 NDP 算法。Tatarakis 和 Minis[25]对随机需求下有仓库补货的单车辆路径问题（Stochastic Vehicle Routing with Depot Returns Problem，SVRDRP）进行了研究，通过对交付产品的划分，提出了一种近似动态规划算法从而在合理的时间内可得到最优策略。

针对运输和物流中出现的随机优化问题，Powell 等人[26]在 2012 年提出了一个完整的研究框架，其中对近似动态规划（ADP）算法在动态 VRP 的应用做了细致的说明。Çimen 和 Soysal[27]将 VRP 的优化目标从成本最小化更改为排放最小化，从而给出了绿色带时间窗有容量约束的随机车辆路径问题（Green Stochastic Time-Dependent Capacitated Vehicle Routing Problem,GSTDCVRP）的 MDP 模型和基于近似动态规划的启发式算法。

Ulmer 等人[28]利用近似动态规划的方法对价值函数进行近似，从而提出了有求解随机服务请求的车辆路径问题（Vehicle Routing Problem With Stochastic Service Requests，VRPSSR）的 ATB 算法。为降低 VRP 中因客户的随机需求带来的高额计算，Ulmer 等人[29]

针对随机服务请求的单车辆路径问题（Single-Vehicle Routing Problem With Stochastic Service Requests，SVRPSSR），将客户请求服务的时间以及客户自身的空间位置纳入近似动态规划中，从而生成动态的路线策略。

为降低由交通拥堵引起的成本，Kok 等人[30]针对CVRPW，在近似动态规划算法中加入了避免交通拥堵的策略，结果表明该方法能够有效降低通勤中拥堵时长。Secomandi 等人[31]针对只有一辆车的随机需求车辆路径问题（SDVRP），通过有限阶段的 MDP 进行建模，使用部分再优化（Partial Reoptimization）的方法对 SDVRP 进行研究，并通过 PH、SH 两种启发式算法选择 MDP 的状态子集，以此来计算最优策略。Goodson 等人[32]提出了基于 Rollot 策略的近似动态规划框架，并将该框架应用于求解具有随机需求和持续时间限制的多车辆路径问题（Vehicle Routing Problem With Stochastic Demand and Duration Limits，VRPSDL）。

### 3.1.2 基于深度动态规划的方法

Kool 等人[33]提出了结合学习神经启发式和动态规划的深度策略动态规划对 VRP 进行求解，模型根据图神经网络（GraphNeural Network,GNN）得到的每个客户顶点特征向量，通过注意力机制计算每个客户顶点被选中的概率，并将这个概率作为动态规划算法对部分解进行选择的策略，最后根据此策略构造最优解。

### 3.1.3 基于动态规划的方法总结

常规方法通常只能求解静态确定性问题，难以求解带有动态和随机信息的问题。动态规划算法可有效求解静态车辆路径问题和动态车辆路径问题，具有求解范围较广的优势。求解车辆路径问题时，需首先建立 MDP 模型，设计算法求解该模型，并用 Rollout 策略在启发式算法基础上得到最优值函数，但动态规划算法无法解决客户节点规模大的车辆路径问题。因此，学者们设计出近似动态规划算法，利用神经网络的泛化能力，通过价值函数近似或策略函数近似得到奖励函数，从而不用直接求解贝尔曼方程，解决了动态规划算法带来的“维数灾难”问题。学者的改进方向主要是近似动态规划算法中神经网络的结构，其主要区别是针对不同车辆路径问题中的各类相关信息进行编码作为神经网络的输入信息，输入的信息越丰富，奖励函数的近似精度就越高，进而近似动态规划算法的优化效果越好。其次是对 Rollout 策略的改进，以减少模型的计算成本和计算量。

相较于传统运筹学有建模不准确的问题，以近似动态规划算法为代表的基于模型的强化学习算法，可以通过智能体与环境不断交互学习到最优策略。在动态车辆路径问题中动态规划算法可以在于环境交互的过程中不断加入获取的随机信息。基于以上优点使有模型强化学习算法适合求解具有动态结构和随机信息的车辆路径问题。

### 3.1.4 动态规划算法局限性分析

动态规划算法在车辆路径问题等领域中应用较广。但也存在许多问题，比如维数灾难、系统不可知、实时求解效率低、近似动态规划算法虽能有效避免上述问题但也因采用神经网络，其鲁棒性较差。分析原因如下：

（1）维数灾难

现实生活中的车辆路径问题规模较大，以至于通过 MDP 建模以后动作空间和状态空间过大，导致动态规划算法失效。因而动态规划算法在求解大规模车辆路径时性能较差。

（2）系统不可知

动态规划算法可求解静态的车辆路径问题和动态的车辆路径问题但需对问题先建模，但实际场景中的车辆路径问题因系统的状态转移函数未知，从而无法对问题进行建模，或对问题进行过理想化处理，使得动态规划算法应用受限。

（3）实时求解效率低

当下车辆路径问题求解的实时性要求不断提高，即需要在较短的时间内给出问题的解，动态规划算法虽能给出问题的最优解，但耗费的时间较长且求解的效率较低。通过神经网络计算的近似动态规划算法虽能比动态规划算法求解算法快，但因当前计算机的性能，算法实时求解能力仍有待提高。

（4）鲁棒性差

目前改进的动态规划算法均是采用神经网络结构，且使用自举采样的方式获取数据，数据的关联性较高，算法的鲁棒性较差，且算法的抗扰动能力较弱。使得近似动态算法在实际生活中的车辆路径问题应用有限。

### 3.1.5 基于动态规划求解 VRP 分析对比

表2将上述基于动态规划求解VRP的各类模型的训练方法、求解问题、以及优化效果进行了对比。

**表 2 基于动态规划求解车辆路径问题的方法对比**

Table 2 Comparison of approaches of dynamic programming applied to VRP

| 作者及时间 | 强化学习方法 | 求解问题 | 优化效果 |
|---|---|---|---|
| Secomandi 等[24] 2000 | 神经动态规划 | 单车辆的 SDVRP | 基于 rollot 的 NDP 算法的带的策略要优于基于最优策略迭 |
| Tatarakis 和 Minis[26]2009 | 基于决策后的近似动态规划 | 单车辆的 SVRDRP | 该算法在能合理的时间内能获得最优路由策略 |
| Çime 和 Soysal[27], 2017 | 近似动态规划 | 多车辆的 GSTDCVRP | 优化效果好于所选基准算法 |
| Ulmer 等[28] 2018 | 近似动态规划 | 单车辆的 SDVRP | 优于 Anticipatory Insertion、Cost Benefit 等启发式算法 |
| Ulmer 等[29] 2019 | 值函数近似和 rollot 策略结合的近似动态规划 | 单车辆的 SVRPSSR | 基于 rollot 策略的 ADP 算法性能优于所选基准算法 |
| Secomandi 等[31] 2009 | 再优化及受限动态规划 | 单车辆的 SDVRP | 优于基于 rollot 的 NDP 算法 |
| Goodson 等[32] 2013 | 基于各类 rollout 策略的近似动态规划 | 多车辆的 VRPSDL | 与动态分解相结合的 rollout 策略能够显著缩短大规模实例的求解时间 |
| Kool 等[33] 2021 | 深度策略动态规划 | 多车辆的 CVRP | 在客户规模为 100 的 CVRP 上，DPDP 模型的优化效果强于各类基于深度强化学习的算法，并接近 LKH 求解器 |
| Zhang 等[35] 2013 | 基于查找表的 VFA 近似动态规划 | 多车辆的 SVRP | 基于查找表的 VFA 算法的优化效果好于基于 rollot 策略的 ADP 算法 |

## 4 无模型的算法

无模型的强化学习算法是指 MDP 模型中的环境参数未知，即在给定状态条件下个体采取动作以后，未来的状态和奖励值未知。因此，个体不对问题进行建模，而是先和环境进行交互，在不断试错中寻找最优策略。无模型的强化学习算法主要分为基于值函数进行优化的算法、基于策略进行优化的算法、值函数和策略结合进行优化的算法。

### 4.1 基于值函数的算法

基于值函数的强化学习算法通过对值函数进行优化从而得到最优策略。在 VRP 中，值函数是对路径策略 $\pi$ 优劣的评估，值函数分为状态价值函数 $V_*(s)$ 和状态-动作价值函数 $q_*(s,a)$，$V_*(s)$ 表示为：

$$V_*(s)=\mathbb{E}_\pi(R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+...|S_t=s) \tag{2}$$

$q_*(s,a)$ 表示为：

$$q_*(s,a)=\mathbb{E}_\pi(R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+...|S_t=s,A_t=a) \tag{3}$$

在求解 VRP 中，基于值函数的强化学习算法主要有时序差分算法、Q-Learning 算法、DQN 算法，Dueling DQN 算法。

#### 4.1.1 时序差分算法

时序差分算法由 Sutton 等人[14]提出，是强化学习的核心算法之一，它结合了动态规划算法和蒙特卡洛方法的原理，是通过部分状态序列来求解问题的强化学习算法。在时序差分算法中，价值函数的更新是通过两个连续的状态和其的奖励值的差来实现的。最基本的时序差分算法的价值函数更新公式为：

$$V(S_t)\leftarrow V(S_t)+\alpha[R_{t+1}+\gamma V(S_{t+1})-V(S_t)] \tag{4}$$

其中 $\alpha$ 为学习率，$R_{t+1}+\gamma V(S_{t+1})-V(S_t)$ 为时序差分误差，因此使用这种更新方法的时序差分也被称为单步时序差分。

针对带时间窗的动态车辆路径问题（CDVRPTW），Joe 和 Lau[34]提出了 DRLSA 模型，通过基于神经网络的时序差分算法和经验放回策略去近似价值函数，然后运用模拟退火算法生成路径。实验表明，DRLSA 模型在有 48 个客户节点的 CDVRPTW 上优化效果超越了经典的基于值函数近似算法和 MSA 算法。该方法解决了当动态请求普遍存在时，该如何给出有效的路径规划。

时序差分算法作为经典的无模型算法，对模型环境要求低，无需训练结束即可获得各类参数的增量。在规模较小的车辆路径问题中优化效果较好，但收敛速度较慢，作为表格型传统强化学习算法不足以解决复杂的车辆路径问题。

#### 4.1.2 Q-Learning 算法

Q-Learning 算法是 Watkins[19]在 1989 年提出的，该算法求解强化学习问题时，使用两个控制策略，一个策略用于更新动作，另一个用于更新价值函数，核心思想为离轨策略下的时序差分控制。Q-Learning 算法在强化学习的控制问题中应用最为广泛。该算法价值函数的更新公式为：

$$Q(S_t,A_t)\leftarrow Q(S_t,A_t)+\alpha[R_{t+1}+\gamma\max_{\alpha}Q(S_{t+1},a)-Q(S_t,A_t)] \tag{5}$$

其中 $\alpha$ 为学习率，$R_{t+1}$ 为 $t+1$ 步的奖励，$\alpha$ 为状态 $S_{t+1}$ 能

执行的动作。

Zhang 等人[35]提出来一种基于查找表的值函数近似 VFA 模型求解带有随机需求的 VRP。具体来说，作者将当前状态和决策的重要信息存储在一个 Q-表中，并用改进的 Q-learning 算法进行学习。

针对多任务的车辆路径问题，Bouhamed 等人[36]提出了一种基于 Q-learning 算法的多任务代理路由调度框架。该模型首先将与任务相关的时间段定义为一个集合，并据此设计出了相应的 Q-表，再通过 Q-learning 算法对 Q-表进行更新从而对问题进行求解，实验结果表明，该模型在复杂的 VRP 上优化效果接近目前最优方法。

Q-learning 算法因优先执行动作，主动探索能力强。可有效的求解带有随机需求信息的车辆路径问题。但因是把状态和不同的动作存储在 Q 表中并一直更新，易使算法陷入局部最优，降低算法的学习效率，搜索耗时较长。更新 Q 表时 Q 表一直发生动态变化，所以更新的效果不稳定。

### 4.1.3 DQN 算法

DQN 算法是 Mnih 等人[20]提出的，该模型将 Q-learning 算法与深度神经网络相结合起来，通常使用 DNN 或者 CNN 来构建模型，使用 Q-learning 算法训练。DQN 算法对 Q-Learning 的改进主要有以下两个方面：1）DQN 算法利用神经网络逼近值函数。2）DQN 算法利用了经验回放训练强化学习的学习过程。DQN 算法的损失函数如下：

$$L(\theta_i)=\mathbb{E}_{(s,a,r,s')\sim D}[(r+\gamma\max_{a'}Q_{\theta_i}(s',a')-Q_{\theta_i}(s,a))^2] \tag{6}$$

目前，DQN 算法在 VRP 中的应用是一个新兴的研究热点，国内外的主要研究成果有：

针对带有多个仓库（Depots）的车辆路径问题（MDVRP）和动态车辆路径问题，Bdeir 等人[37]提出了基于 DQN 的 RP-DQN 模型，该模型框架如图 3 所示。RP-DQN 模型中为有效降低问题的计算复杂性，编码器不仅对静态的客户节点位置、客户需求进行编码，并对问题中的动态特征信息进行编码，使用 Q-learning 算法对模型进行优化。在客户数量为 20、50、100 的 CVRP 上优化效果均超过了 Kool 等人[46]的方法。作者首次将深度 Q 网络运用到 MDVRP 的求解过程中，在客户数量为 20、50、100 的 MDVRP 上 RP-DQN 模型的优化效果也好于 Kool 等人[46]的方法。总体来说 RP-DQN 模型的泛化能力要高于 Kool 等人[46]提出的 AM 模型。

针对客户当日要求交付（Same-day delivery）的需求，Chen 等人[38]提出了车辆和无人机的当天交付问题（Same-Day Delivery Problem with Vehicles and Drones，SDDPVD），建模过程中作者采用和 Powell[26]相同的模型架构，并使用 Ulmer[39]提出的路由策略来模拟路线的更新和演变，首先将 SDDPVD 建模为 MDP 模型，为了使决策快速有效，作者将动作空间和状态空间进行简化，通过设计 DQN 算法去近似每个特征向量的值。

### 4.1.4DQN 算法总结

DQN 算法作为具有里程碑意义的深度强化学习算法，不同于传统强化学习算法中值函数线性近似方法，使用多层深度神经网络近似代替 Q-learning 算法中的 Q-表，从而可以将高维输入值映射到 Q-空间，DQN 算法通过一个经验池来存储以前的数据，每次更新参数的时候从经验池中抽取一部分的数据来更新，从而打破信息间存在的关联，DQN 算法从而可有效求解有状态、动作高维复杂，数据间彼此有关联的车辆路径问题。

基于 DQN 算法求解车辆路径问题的各类模型主要是通过对 DQN 算法的状态、动作、奖励的表示做出相应的修改，对价值函数进行映射，通过对价值函数做出评价，以此改进初始策略。但 DQN 算法在求解车辆路径问题仍存在很多不足，比如因需对 Q-网络进行最大化操作引起过估计问题，DQN 算法只能求解单车辆的车辆路径问题。

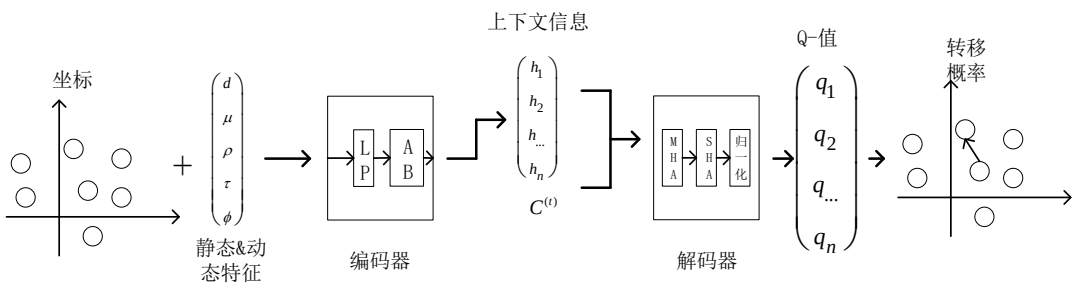


**图 3 RP-DQN 模型示例**

Fig.3 Example of RP-DQN model

#### 4.1.5 DQN 算法局限性分析

DQN 算法因运用经验放回机制和设定一个固定 Q 目标值神经网络，具有较好的收敛性和兼容性。但在车辆路径问题求解中也存在较多问题，例如过拟合问题、样本利用率低、得分不稳定。具体以上问题原因为：

（1）过拟合问题

DQN 算法在训练智能体过程中，会采用 Q 网络最大化的操作，从而出现过度适应当前环境的情况，使算法出现过拟合现象。

（2）样本利用率低

DQN 算法和环境交互的过程中，样本之间有很强的关联性，降低了深度神经网络的更新效率，算法需要很长的时间才能达到合适的得分标准，使得 DQN 算法在车辆路径问题中的数据样本利用率较低。

（3）得分不稳定

使用 DQN 算法求解车辆路径问题时，Q-learning 学习过程中会对 Q 值过高估计，容易产生较高误差，导致算法稳定性较差，得分性能不稳定。

#### 4.1.6 Dueling DQN 算法

针对无模型的强化学习问题，Wang 等人[40]提出了一种新的 DQN 模型：Dueling DQN(DDQN)。不同于 DQN 算法，DDQN 算法把在卷积层得到的特征分为状态值和动作优势函数两部分：

$$Q_\pi(s,a)=V_\pi(s)+A_\pi(s,a) \quad (7)$$

上式中，$Q_\pi(s,a)$ 为状态 $s$ 下选择动作 $a$ 的奖励值，状态值 $V_\pi(s)$ 是对状态 $s$ 的评价，动作优势函数 $A_\pi(s,a)$ 是对状态 $s$ 下各个动作优劣的评价指标。

Zhang 等人[41]为有效解决车辆路径问题中的供需匹配难题，提出了 QRewriter-DDQN 模型。将可用车辆提前调度给需求级别高的客户。QRewriter-DDQN 模型由 Qrewriter 模块和-DDQN 模块构成，DDQN 模块将车辆和客户之间的 KL 分布作为激励函数，从而得到供需之间的动态变化。之后，运用 Qrewriter 模块用来改进 DDQN 生成的调度策略。

#### 4.1.7 Dueling DQN 算法总结

Dueling DQN 算法是通过改进深度神经网络的结构来提高 DQN 算法性能。该算法采用卷积神经网络处理车辆路径问题中的初始信息，并使用两个全连接神经网络把 $Q$ 值划分为状态值和动作优势函数两部分，通过这种改变可以有效的区分出奖励值的来源。因为优势函数存在，Dueling DQN 算法可以让一个状态下的所有动作同时更新，加快了算法收敛速度。尤其是在有大规模动作空间的车辆路径问题中，相较于初始 DQN 算法 Dueling DQN 算法能更加高效的学习价值函数。

### 4.2 基于策略的方法

以上基于值的算法是通过值函数近似方法对价值函数进行参数化比较，从而使个体选择相应动作。另一种常见的是基于策略的算法，直接对策略进行优化，寻找最优策略。常用于 VRP 的基于策略的算法有蒙特卡洛 REINFORCE 算法、Actor-Critic 算法等。

基于策略的算法具有策略参数化简单、收敛速度快的优点，且适用于连续或者高维的动作空间。策略就是策略参数化，即 $\pi_\theta$，参数化后的策略为一个概率密度函数 $\theta$。与策略相对应，策略搜索分为随机策略搜索和确定性策略搜索。策略搜索的目的就是找到最优的参数 $\theta$，定义目标函数：

$$\max_\theta J(\pi_\theta)=\max_\theta \sum_r P(r;\theta)R(r) \quad (8)$$

定义策略梯度公式为：

$$\nabla_\theta J(\pi_\theta)=\mathbb{E}_{\pi_\theta}[\sum_{t=0}\nabla_\theta \log \pi_\theta(a_t|s_t)Q_\pi(s,a)] \quad (9)$$

更新策略的参数：

$$\theta \leftarrow \theta\alpha\nabla J(\theta) \quad (10)$$

式中 $Q^\tau(s,a)$ 为状态动作函数，$\sum_{t=0}\nabla_\theta \log \pi_\theta(a_t|s_t)$ 为状态 $s$，动作 $a$ 随参数 $\theta$ 变化最陡的方向。

#### 4.2.1 蒙特卡洛 REINFORCE 方法

蒙特卡洛 REINFORCE 方法是最简单的策略梯度算法[42]，该算法使用价值函数 $V_*(s)$ 来近似代替策略梯度公式里面的 $Q_\pi(s,a)$，首先输入 $N$ 个蒙特卡洛完整序列，用蒙特卡洛方法计算每个时间位置 $t$ 的状态价值 $V_t(s)$，再按照以下公式对策略 $\theta$ 进行更新：

$$\begin{aligned}&\theta \leftarrow \theta+\alpha\nabla_\theta \log \pi_\theta(a_t|s_t)V_t(s)\\&\theta \leftarrow \theta+\alpha\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{k=t+1}^{T}\gamma^{k-t}R_k\end{aligned} \quad (11)$$

$$\begin{aligned}&\nabla\mathcal{L}(\theta\,|\,s)=\mathbb{E}_{p_0(\pi|s)}[L(\pi)-b(s)\nabla \log p_\theta(\pi|s)]\\&0\leftarrow 0+\nabla\mathcal{L}(0\,|\,s)\end{aligned} \quad (12)$$

不断执行动作直到结束，在一个回合结束之后计算总反馈，然后根据总反馈更新参数。$p_\theta(\pi|s)$ 为每一步动作选择概率的累乘，则 $\log p_\theta(\pi|s)$ 则为每一步动作选择概率对数的求和，$\nabla \log p_\theta(\pi|s)$ 为梯度值，$L(\pi)-b(s)$ 为梯度下降的方向，$b(s)$ 为策略的平均表现。

自 Vinyals 等人[43]在 2015 年提出了指针网络（Ptr-Net）模型求解旅行商问题后，在小规模的旅行

商问题上，相较于传统的启发式搜索算法，该模型具有求解更快的特点，这是深度学习在组合优化问题上的首次应用，旅行商问题是 VRP 的一种特例，因此，研究人员开始将指针网络模型应用于 VRP 的求解。

Nazari 等人[44]针对 CVRP 构建 Ptr-Net—REINFORCE 模型，该模型是一个 end-to-end 的深度强化学习模型，通过 Ptr-Net 进行建模，在构建指针网络阶段，Ptr-Net 中的输入为静态值（客户位置）与动态值(客户需求)，其模型结构如图 4 所示。不同于 Ptr-Net 在训练模型时采用监督式方法，作者使用 REINFORCE 算法对模型进行训练，分别在客户数为 10、20、50、100 的 CVRP 数据集上进行了测试，取得了比经典的启发式搜索算法 CW、SW 更好的效果。Xin 等人[45]为解决深度强化学习算法在构造解的过程中无法修改以前决策，提出基于 REINFORCE 算法的分步 SW-Ptr-Net 模型和近似 SW-AM 模型。该方法有效提升了 Ptr-Net 模型和 AM 模型[46]对 CVRP 的优化效果。

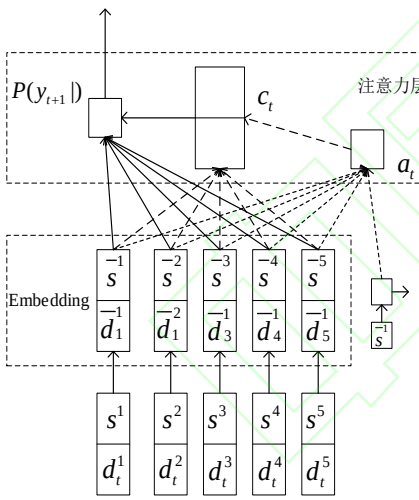


**图 4 Ptr-Net—REINFORCE 模型示例**

Fig.4 Example of Ptr-Net—REINFORCE model

1）基于 Ptr-Net 的深度强化学习模型总结

Ptr-Net 模型使用编码器对车辆路径问题中的初始信息进行编码得到特征向量，再通过解码器对特征向量进行解码，利用注意力机制计算每个客户节点的选择概率，从而构造车辆路径问题的解。所有运用 Ptr-Net 模型构造车辆路径问题的构造过程大致如下：

首先通过 Embedding 层把每个客户节点的地理位置和需求转化为节点表征向量 $s$，传入循环神经网络中得到上下文信息和隐藏层信息。然后通过解码器对上下文信息进行解码，利用注意力机制按照隐藏层中的信息和上下文信息得到每个客户节点的被选概率，每一步都选择被选概率最大的客户节点加入解中，逐步构造车辆路径问题的解。若使用监督式方法训练模型，需要得到已有最优解的车辆路径问题训练集，车辆路径问题是经典的 NP 难问题，因此得到实例的最优解非常困难。且车辆路径问题均可建模成 MDP 模型，使用强化学习算法训练 Ptr-Net 模型是非常合适的。由式（12）可知，REINFORCE 算法是以反向传播作为参数更新的标准，智能体在探索解空间时，可以不断提升初始解的质量。因而，当使用 Ptr-Net 模型求解车辆路径问题时，国内外学者常采用 REINFORCE 算法对模型进行优化。

Ptr-Net 模型这一新型深度神经网络模型，主要解决输入时序与位置相关的离散个体组成输出时序的问题，是求解具有时间序列特性的车辆路径问题的主要深度学习模型。相较于循环神经网络处理具有自回归问题，Ptr-Net 模型对输入序列长度可变时，直接使用归一化方式输出各个客户节点在当前解码位置的概率分布。但 Ptr-Net 模型复杂度较高，且需要大量的采样和局部搜索改善 Ptr-Net 模型得到的初始解，使模型收敛较慢。

Kool 等人[46]首次将 Transformer 模型应用到 VRP 的求解中，和大多数 seq2seq 模型一样，transformer 的结构也是由编码器和解码器组成。在编码器中作者没有使用 Transformer 模型输入时的 positional encoding，从而使得节点嵌入不受输入顺序影响，但仍采用经典 Transformer 模型中的多头-Attention 机制。在解码器中作者为了提高计算效率并未采用 Transformer 模型解码层的多头-Attention 机制，而是只使用一个自-Attention 机制。作者采用加入 Rollout baseline 的 REINFORCE 算法对模型进行训练，并在 CVRP 和 SDVRP 等问题的求解上取得了比基于指针网络模型更好的效果，且与经典的运筹学求解器 LKH3、Gurobi 相比求解效果相差无几。

Falkner 等人[47]针对 CVRPTW，提出了 JAMPR 模型，该模型由多个编码器和一个解码器组成，编码器采用了 self-Attention 计算方法，通过加入两个新的编码器产生上下文信息以及增强联合行动空间，解码器使用 Transform 模型的多头-Attention 机制。使用 REINFORCE 算法对 JAMPR 模型进行训练，模型架构如图 5 所示。在对 CVRP-TW 的三种变体(hard，partly-soft，soft)的实验中，该模型的优化效果要好于现有的元启发式算法和基于 Attention 机制的强化学习模型。

Xin 等人[48]提出了一个多解码器注意模型(Multi-Decoder Attention Model，MDAM)来训练波束搜索（Beam Search）的策略，MDAM 由一个编码器和多个解码器组成，编码器采用和 Transformer 模型相

同的多头-Attention 机制，每个解码器采用节点嵌入的方式来产生访问每个有效节点的概率。使用 REINFORCE 算法对 JAMPR 模型进行训练。在对 CVRP 和 SDVRP 等问题的求解中，相较于所选基准算法该模型的优化效果要更好。为有效地解决具有软时间窗的多车辆路径问题（Multi-vehicle routing problem with soft time windows，MVRPSTW），Zhang 等人[49]提出了多智能体注意力模型(Multi-Agent Attention Model，MAAM)，使用 REINFORCE 算法对 MAAM 模型进行训练。实验结果表明，求解速度优于 Google OR-Tools 求解器和传统的启发式算法。Xu 等人[50]以 AM 模型[46]为基础构建了具有多重关系的 MRAM 模型，更好的获取车辆路径问题的动态特征。

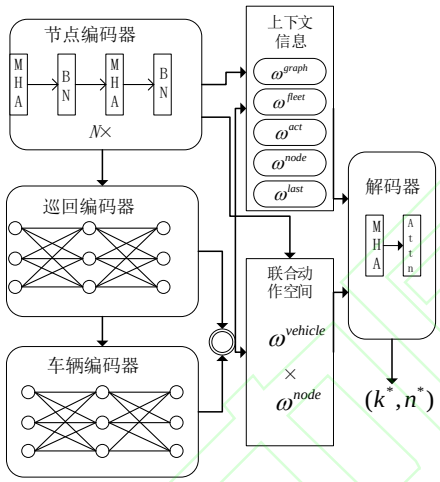


**图 5 JAMPR 模型示例**

Fig.5 Example of JAMPR model

2）基于 Transformer 的深度强化学习模型总结

Ptr-Net 模型中因编码器和解码器使用循环神经网络因而存在不能捕获问题的长程相关性，且模型训练消耗大量时间，因 Transformer 模型中的多头 Attention 可以提取车辆路径问题中更加深层次的特征，所以学者们借鉴 Transformer 模型提出了基于 Attention 机制提出各类新模型，此类深度强化学习模型仅通过 Attention 机制对输入输出的全局依赖关系进行建模，这类模型可以捕获客户节点间的长程相关性且有较高的计算效率。但 Attention 机制需要极大的计算量和内存开销，所以这类模型的主要改进是通过改变编码器和解码器的个数以及编码方式、解码方式、注意力机制来实现的。因 REINFORCE 算法具有自适应性，可自行调节参数，基于 Transformer 的各类模型常使用 REINFORCE 算法作为训练模型的算法，但因为 REINFORCE 算法方差较大，训练结果不稳定，研究人员引入带有基线函数的 REINFORCE 算法，该训练算法加快了模型的收敛速度。

Peng 等人[51]结合动态 Attention 方法与 REINFORCE 方法设计了一种 AM-D 模型来求解 VRP，动态 Attention 方法是基于动态编码器-解码器结构来改进的，改进的关键是动态挖掘实例的结构特征，并在不同的构造步骤中有效地挖掘隐藏的结构信息，模型架构如图 6 所示。。实验结果表明，在客户数量为 20、50、100 的 CVRP 上优化效果均超过了 Kool 等人[46]提出的 AM 方法，并明显减小了最优性差距。Lu 等人[52]提出了基于迭代的 Learn to Improve（L2I）框架，算法首先给出一个可行解，运用基于 REINFORCE 方法的控制器选择改进算子或基于规则的控制器选择扰动算子迭代更新解，经过一定迭代步骤后从所有访问的解决方案中选择最优解。对于 CVRP，该模型不仅在优化效果上超过了 GoogleOR tools、LKH3 等专业运筹学求解器，还是第一个在 CVRP 上求解速度低于 LKH3 求解器的强化学习框架，模型架构如图 6 所示。

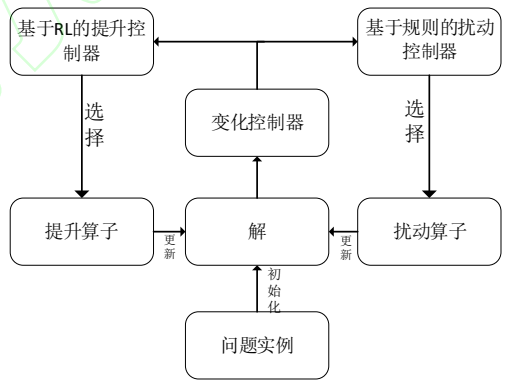


**图 6 L2I 模型框架**

Fig.6 The framework of L2I model

Hottung 和 Tierney[53]提出神经大邻域搜索(NLNS)框架对 CVRP、SDVRP 等问题进行求解，作者运用基于 Attention 机制的神经网络对 LNS 中的毁坏算子和重建算子进行学习，并用 REINFORCE 算法对 NLNS 模型进行训练。实验结果表明，在 CVRP 实例上 NLNS 模型与 LKH3 性能相当；在 SDVRP 实例上，NLNS 能够在拥有 100 个客户的实例上胜过目前最先进的方法。

3）强化学习与局部搜索算法结合的模型总结

基于 Transformer 模型和 Ptr-Net 模型求解车辆路径问题虽然速度较快，但优化效果仍不及专业运筹学求解器。在组合优化问题求解中，局部搜索算法仍然是主流代表算法，局部搜索算法往往涉及到算法参数设置和问题配置，这些都需要算法设计者有高超的算法设计技巧才能保证启发式算子的效果，学者们基于

不同车辆路径问题的特征和算法结构得到合适的参数和策略，运用强化学习方法对局部搜索策略进行训练，扩大了局部搜索算法启发式算子的搜索能力，以此来提高解的质量。目前求解车辆路径问题的最优算法就是基于强化学习的局部搜索算法。

### 4.2.2 REINFORCE 算法局限性分析

随着计算机计算能力不断增加，REINFORCE 算法已成为深度神经网络模型解决车辆路径问题最常用的训练方法之一。REINFORCE 算法相较于基于值函数的算法，可以表示随机策略，当策略不定时，可以输出多个动作的概率分布。但是 REINFORCE 算法也存在很多问题，比如数据利用率低、算法收敛速度慢、方差较大导致算法收敛性变差。分析存在以上的原因如下：

（1）数据利用率低

REINFORCE 算法是回合更新的算法，需要完成整个回合，才能更新状态-动作对。这种更新方式使得整个轨迹的一系列动作被当作整体，若轨迹的收益较低，即使包含一些好的动作，但下次被选的概率仍会下降，使得数据利用率低。

（2）算法收敛速度慢

对于 REINFORCE 算法，车辆路径问题中的每个样本只能被训练一遍，有些能具有高收益的样本没有被重复利用，这不仅浪费计算资源，还增加算法了的收敛时间，使得算法收敛速度慢。

（3）算法收敛性差

在车辆路径问题中，REINFORCE 算法为控制训练个体的时间，不可能遍历所有状态，只能通过采样得到大量轨迹，但这样的做法会造成 REINFORCE 算法与环境交互不充分，那些未被选中的动作有可能对应着较高奖励，因而产生较大方差，导致算法收敛速度变慢。所以经常通过在算法中加入基线函数来避免高方差。

### 4.2.3 Actor-Critic 算法

Actor-Critic 算法是一种基于值方法和基于策略方法相结合而产生的算法。该算法的架构包括两个部分：Actor 部分是基于策略方法的，通过生成动作并和环境交互，用来逼近策略模型 $\pi_\theta(s,a)$；Critic 部分是基于值方法的，判定动作的优劣性，并且选择下一个动作，用来逼近值函数 $Q(s,a)$。Actor 与 Critic 之间互补的训练方式相较于单独的算法更加有效。策略梯度函数为：

$$\nabla_\theta J(\pi_\theta)=\mathbb{E}\left[\sum_{t=0}^{\infty}\psi_t\nabla_\theta\log\pi_\theta(a_t\mid s_t)\right] \tag{13}$$

$$\psi_t=r_t+Q(S_{t+1},A_{+1})-Q(S_t,A_t)$$

Actor 的策略网络的更新公式为：

$$\theta\leftarrow\theta+\alpha\delta\nabla_\theta\log\pi(A_t\mid S_t,\theta) \tag{14}$$

Critic 的值函数网络的更新公式为：

$$\delta\leftarrow G_t-\hat{v}(S_t,w) \tag{15}$$

$$w\leftarrow w+\beta\delta\nabla_w\hat{v}(S_t,w)$$

Zhao 等人[55]提出一种改进的 Actor-Critic 算法对模型进行训练，首先通过路由模拟器生成大量 VRP 实例用于训练和验证，在 Actor 网络的编码过程中将静态特征和动态特征分别编码，在基本 Attention 机制[46]中，模型对输入序列的顺序很敏感。为了克服这个问题，作者采用了图嵌入的方法，Critic 网络由一个 Simple 网络和一个 Actor 网络组成，为加快模型收敛速度，作者还借鉴了图像字幕[56]中的 self critic 思想，据此构成了 adaptive critic 网络，网络架构如图 7 所示。新的 Actor-Critic 模型在客户点数分别为 20、50 和 100 的三个数据集上进行测试，实验结果表明，该模型找到了更好的解决方案，同时实现了 5 到 40 倍的加速。作者还观察到，与其他初始解决方案相比，将深度强化学习模型与各种局部搜索方法相结合，可以以更快的求解速度得到解。

在日常生产生活中，一个客户可能总是有他自己的送货点，比如同城快递服务[57]和拼车服务[58]，而不是像 VRP 那样为所有客户共享一个仓库。所有这类 VRP 可以描述为提货和交货问题（pickup and delivery problem, PDP）。Li 等人[56]提出了一种基于异构 attention 机制的编码器-解码器模型，其编码层在多头 attention 注意力方法中加入了异构 attention 方法以期更早得到优先级较高的关系，采用 Actor-Critic 算法对该模型进行训练。在 PDP 求解中该模型的优化效果要好于传统的启发式算法。Gao[57]提出了基于图注意力网络的模型用去求解 VRP、CVRP。该模型的编码器是由一个 Graph Attention Network 组成，模型首先对每个客户位置进行编码得到每个客户顶点的边缘嵌入和顶点嵌入，在通过 Attention 机制计算出每个客户被选择的概率。解码器采用和 Ptr-Net 一样的架构，为 VLNS 算法的破坏算子提供一个节点子集作为移除候选。依然采用 Actor-Critic 算法对该模型进行训练。当 VRP、CVRP、CVRPTW 等问题的规模较大时（多于 400 顶点）该模型优于传统启发式算法和现有求解 VRP 的深度强化学习模型。

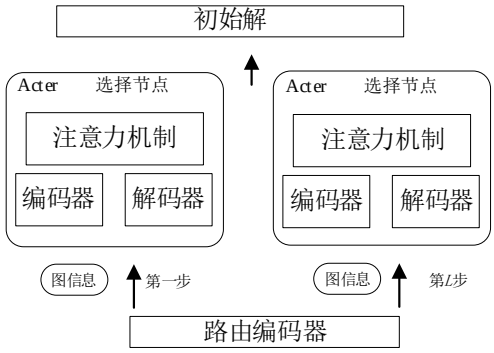


**图 7 adaptive critic 网络示例**

Fig.7 Example of adaptive critic network

1）强化学习与图神经网络结合的模型总结

车辆路径问题是典型的具有图结构的组合优化问题，因图神经网络能够有效求解具有图结构的问题，所以有学者把图神经网络应用到了车辆路径问题的求解中。运用图神经网络求解车辆路径问题的构造过程大致如下：

将车辆路径问题建模为图 $G=(V,E)$，其中 $V$ 代表节点集合，$E$ 代表边集合。图神经网络根据用户节点 $V_i$ 本身的二维空间位置和需求、边的特征、以及节点 $V_i$ 邻居节点的二维空间位置和需求对节点 $V_i$ 特征向量进行更新，从而得到每个节点的特征向量。为加快模型收敛，研究人员通常把图神经网络求得的用户节点的特征向量加入 Transformer 模型和 Ptr-Net 模型的编码器中，在通过 Attention 机制计算出每个客户被选择的概率。因为 Actor-Critic 算法融合了基于值的算法和基于策略的算法的优点，即可解决连续动作空间问题，还能进行单步更新从而加快学习速度，所以在图神经网络中常使用 Actor-Critic 算法训练模型。

**表 3 无模型方法求解车辆路径问题对比**

Table 3 Comparison of approaches of model-free applied to VRP

| 模型 | 机制特点 | 优点 | 缺点 | 优化效果 |
|---|---|---|---|---|
| DRLSA[34] | 在时序查分算法中加入经验放回策略 | 路由策略生成速度快 | 迭代时间较长、执行效率不高 | 优于基于值函数近似的算法和 MSA 算法 |
| VFA[35] | 使用 Q-learning 对基于 ADP 框架的值函数近似算法进行了改进 | 采用有界查找表和有效维护策略有效解决了“维数灾难”、模型具有一定扩展性 | ADP 中所使用的神经网络目前还没有理论上的构造方法，只能利用经验和试错法来设计 | 优于基于 rollot 策略的 ADP 算法 |
| MTA-QL[36] | 通过将任务与时间段联系起来克服时间“维数灾难” | 智能体探索能力较强、可以实现最大化任务覆盖 | 使用贪婪搜索容易得到局部最优解、算法收敛慢 | 接近最优解 |
| RP-DQN[37] | 使用采用 transform 模型，但编码器中删去了 softmax 和掩码 | 对动态特征进行编码，每个步骤中编码器和解码器均执行一次 | Q-表规模较大时，算法迭代时间较久使得算法收敛慢 | 100-CVRP：优于[46] |
| SDPM-DQN[38] | 使用自适应动态规划建模、加入了路由信息 | 使用自适应动态规划有效解决了“维数灾难” | 过估计问题、样本利用率低<br>实时性较低 | 优于 PFA 算法 |
| Ptr-Net[44] | 由 LSTM 作为编码器、解码器 | 不需要迭代搜索，模型求解速度快，编码器使用一层 CNN，降低了计算成本 | 解码器均使用 LSTM，导致神经网络更新较慢，降低模型性能，且模型复杂度较高 | 优于几类经典的启发式算法 |
| AM[46] | 采用 transform 模型 | 解码器解码过程中仅考虑前两步决策，有效缩短计算时间，模型有较强的鲁棒性 | Attention 机制中 head 数量过多、编码器层数影响模型性能 | 优于几类经典的启发式算法并接近几类专业求解器 |
| AM-D[51] | 采用 transform 模型 | 使用动态注意力机制在构造时有效地挖掘隐藏的结构信息 | 模型对局部信息获取能力较弱，编码器只能对静态节点特征编码 | 优于[46]；接近最优解 |
| L2I[52] | 运用基于强化学习的提升算子和基于规则的扰动算子来迭代更新解。 | 运用强化学习指导的提升算子增强了解空间的搜索效率、具有最好的优化效果和求解时长 | 虽有基于规则的扰动算子，但容易得到局部最优解、内存消耗较大 | 优于[44,46,54]；优 Google OR toolsLKH3 求解器；运行时间远低于 LKH3 |
| JAMPR[47] | 模型由多个编码器和一个解码器构成 | 多个编码器可以得到增强动作空间，从而更好的提取问题特征 | 模型内存需求较高，计算复杂度高，学习效率偏低 | CVRPTW：优于[46]；优于 GORT、CVRP：优于[44,46]；优于 GORT,接近于[52] |
| MDAM[48] | 模型由多个解码器和一个编码器构成,采用波束搜索提高解的质量 | 相较于一个解码器，多个解码器有效地增加了找到好的解的可能性 | 使用波束搜索耗时较长，使得模型求解问题时算法收敛慢 | CVRP：优于[44,46,54]；优化效果与 LKH3 求解器一致；SDVRP：优于[44,46] |
| MAAM[49] | 编码器和 Transform 模型一致，解码器部分对每个车辆进行了建模 | 实现了长期记忆，不需要隐藏层传递，加快了计算速度 | 局部信息获取能力较弱，模型收敛慢，易得到局部最优解 | 优于 Google OR tools |
| NLNS[53] | 由毁坏算子和重建算子构成,重 | 运用强化学习算法更高效的搜 | 重建算子提升解的质量效果不明 | CVRP：优化效果与 |

| | | | | |
|---|---|---|---|---|
| | 建算子中使用前馈神经网络作为解码器 | 索解空间，并利用并行计算加快搜索 | 显，方差较大 | LKH3 一致 SDVRP：优于 SplitILS |
| NeuRewriter[54] | 具有和局部搜索算法类似的步骤，先构造解，再通过强化学习提升解的质量 | 采用深度强化学习方法指导搜索算子，扩大了搜索空间 | 构造得到解质量偏低、提升解的质量不明显、学习效率低 | 优于[44,46] |
| Actor-Critic-routing[55] | Actor 网络产生路由策略，Critic 网络调整策略 | 采用自适应性 Critic 网络，加快模型收敛速度，解码器仅使用一层卷积网络，计算成本低 | 使用 LSTM 作为解码器，导致神经网络更新较慢，影响了模型性能 | 优于[44]；优于几类经典的启发式算法 |
| H-AM[56] | 使用 Transform 模型，并在编码器中使用异构 Attention 机制 | 使用异构 Attention 机制可以捕获优先关系，更好的提取特征 | 局部信息获取能力较弱、影响模型性能，训练模型时偏差较大 | 优于[46]；优于 Google OR tools 和模拟退火算法 |
| GAN[57] | GNN 得到节点特征向量，由 Attention 得到选择概率 | 用强化学习算法对节点重新排序，算法收敛速度快 | 使用 LSTM 作为解码器，导致神经网络更新较慢，降低模型性能 | 优于[46]；在大规模 CVRP 问题上优于启发式算法 |
| AM-A2C [58] | 由两个编码器和一个的解码器提取节点的特征 | 两个编码器可以得到更深层次的模型特征 | 使用 RNN 作为解码器，易陷入局部最优解，算法收敛速度慢 | 优于几类经典的启发式算法 |

#### 4.2.4 Actor-Critic 算法局限性分析

Actor-Critic 算法是目前最为流行的强化学习算法之一，结合了基于值算法和基于策略算法的优点，既能应用于连续动作空间，有能进行单步更新。但仍然存在一些问题，比如学习效率低，收敛速度慢。分析存在以上问题的原因如下：

（1）学习效率低

Actor-Critic 算法中涉及到 Actor 网络和 Critic 网络两部分，Actor 网络基于概率选择动作等价于策略函数，Critic 网络等价于价值函数。Actor-Critic 算法每次更新都会涉及到这两个深度神经网络，而且每次都在连续状态中更新参数，参数更新时有很强的相关性。导致算法学习效率低。

（2）收敛速度慢

Actor-Critic 算法中 Critic 网络收敛速度慢，若 Critic 网络不收敛，那么它对 Actor 网络的评估就不准确，导致 Actor 网络难收敛，使得 Actor-Critic 算法收敛速度慢。

#### 4.2.5Advantage Actor-Critic 算法：

Advantage Actor-Critic（A2C）算法又称同步优势 Actor-Critic 算法是 Mnih 等人 2016 年提出的。在 Actor-Critic 算法中 Critic 网络输入 Q 值来对策略梯度进行计算，但这种计算方式存在高方差的问题，从而可以引入 Baseline 的方式减小梯度使得训练更加平稳，通过这种改进就能得到 A2C 算法。A2C 算法的策略梯度函数为：

$$\nabla_\theta J(\pi_\theta)=\mathbb{E}\left[\sum_{t=0}^{\infty}\psi_t\nabla_\theta\log\pi_\theta(a_t\mid s_t)(G_t-b(S_t))\right] \tag{16}$$

$$\psi_t=r_t+Q(S_{t+1},A_{+1})-Q(S_t,A_t)$$

A2C 算法的优势函数为：

$$A(s_t,a_t)=Q(s_t,a_t)-V(S_t) \tag{17}$$

A2C 算法将 Critic 网络对 Q 值的估计改为对优势函数的估计，估算每个决策相较于平均值的优势。A2C 算法的整体架构与 Actor-Critic 算法类似，只是在其基础上加入了并行计算的能力，即整个个体由一个全局网络和多个并行的工作体（Worker）组成，每个工作体都包括一个 Actor-Critic 网络。

**表 4 强化学习总结**

Table 4 Summary in Reinforcement Learning

| 分类 | 算法名称 | 优点 | 缺点 | 适用场景 |
|---|---|---|---|---|
| 基于模型的算法 | 动态规划 | 参数即时更新样本效率高、算法的泛化能力强 | 建模复杂、对环境模型要求高，局限性较大、存在“维数灾难” | 规模较小，建模简单，存在环境动力学的单车辆路径问题 |
| | 近似动态规划 | 有效解决常规的动态规划带来的“维数灾难” | 实时性有待提高、鲁棒性差、收敛速度慢 | 规模较大，存在环境动力学的车辆路径问题 |
| 无模型的算法 | 时序差分 | 对环境模型要求低、无需 episode 结束即可更新参数、只需相关参数即可得到最优解 | 收敛速度较慢，不能保证每个时间点步长的更新 | 规模中等、带有时间窗的车辆路径问题 |
| | Q-learning | 不需要提前输入环境的相关知识、主动探索能力强 | Q-表规模较大时，算法迭代时间较久，易陷入局部最优解，学习效率低 | 规模中等的单车辆路径问题 |
| | DQN | 可求解具有高维数据、复杂环境和任务的问题 | 过估计问题，样本利用率低<br>得分不稳定性、数据探索率较低 | 具有复杂状态、动作空间的单车辆路径问题 |

| | | | |
|---|---|---|---|
| DDQN | 算法优化过程中稳定性较好、解决了Q-learning中的过估计问题、方差小 | 因注重Q函数计算算法性能较差 | 有多个仓库节点的车辆路径问题 |
| REINFORCE | 可以表示随机策略，当策略不定时，可以输出多个动作的概率分布 | 容易得到局部最优解、评估单个策略不充分，方差较大、数据利用率低、算法收敛慢 | 规模较大、具有复杂状态、动作空间的单车辆路径问题 |
| Actor-Critic | 可解决连续动作空间问题，还能进行单步更新从而加快学习速度 | 学习效率较低、收敛速度慢、偏差较大 | 规模较大、具有复杂状态、动作空间、多约束的单车辆路径问题 |
| A2C | 使用优势函数作为评判依据，解决了Actor-Critic学习效率较低的问题 | 在环境训练迭代中稳定性较差 | 具有复杂状态空间的多车辆路径问题 |

Vera 和 Abad[58]针对固定车队规模的容量受限多车辆路径问题（CMVRP），提出了基于 Attention 机制的 Seq2Seq 模型，采用 A2C 算法对模型进行训练，与经典的启发式算法 SW 和 CW 相比该模型生成的解具有更低的标准差。

A2C 算法使用一个优势函数作为评价标准，这个优势函数用来评价当前所选动作与所有动作平均值之间的差值，从而降低了 AC 算法所产生的误差，但运用 A2C 算法求解车辆路径问题时，智能体在环境训练中的稳定性较差。

### 4.3 基于动态规划求解 VRP 分析对比

基于无模型强化学习方法求解 VRP 的对比结果见表 3。

## 5 强化学习算法总结分析

近年来，强化学习在车辆路径问题求解中涌现了很多优秀的算法，在应用过程中这些算法有自己的优势，但也暴露出一些自身局限性，将上述内容进行总结分析，如表 4 所示。

基于模型的强化学习算法求解车辆路径问题时需先构建 MDP 模型，智能体通过与环境的不断交互，智能体对数据的利用率较高。为增强求解问题的实时性，使用 Rollout 框架对初始策略进行迭代更新，逐步逼近最优解。近似动态规划中使用自举采样法训练神经网络，自举采样得到的数据不是独立同分布的，这样的采样方式使模型稳定性降低。

现在的使用深度强化学习模型解决车辆路径问题时主要的创新点是对深度神经网络架构的设计上，即设计更加契合车辆路径问题的数据结构，主要包括像 Ptr-Net 模型和 Transform 模型这种把问题建模为序列形式的输入，然后基于各类 Attention 机制对客户节点的选择优先级进行排序；因车辆路径问题天然是图结构，可基于图神经网络和基于图神经网络的 Attention 机制提取车辆路径问题的特征。通过以上方法得到局部解后以自回归的方式扩展得到最优解。因为监督式学习方法不适用与组合优化问题，学者们采用具有反向学习能力的强化学习算法训练模型。深度强化学习模型求解速度远超传统的启发式算法，模型泛化能力较强。

强化学习算法的优势是智能体通过和环境进行交互可以得到最优策略，从而具备解决大规模复杂组合优化问题的可能性，但其也存在着算法执行时间过长、模型不易收敛等局限性，因此，可用其他启发式算法和强化学习融合，主要是使用强化学习算法对启发式搜索算子进行学习，加快求解速度，提高解的质量，相较于人工设置搜索策略，强化学习算法可以扩大搜索空间和提高搜索策略效率，具有较强的优化能力。

## 6 结论与展望

本文旨在对近年来基于强化学习求解车辆路径优化问题的各类算法进行较为全面的综述，重点分析了各类强化学习算法的机制特点和优劣性。如今在 VRP 的求解中各类深度强化学习模型不断涌现。本文对这些模型的特点和优化效果进行了总结。基于模型的强化学习算法必须对问题进行建模，但往往车辆路径问题建模过程与现实环境总有差距。在基于编码器-解码器的 Transformer 模型和 Ptr-Net 模型以及图神经网络中强化学习算法都是作为训练模型的算法出现，选择强化学习算法时目的性不强。且强化学习自身存在较强的探索和利用困境，个体不仅要学习以前样本中的动作，还要对未来可能带来高奖励的动作进行探索，但车辆路径问题是高度复杂的大型问题，强化学习常用的探索策略常常失效，探索策略的可扩展性较差。强化学习算法常因为经验样本只被采样一次或随着经验的积累，同一个样本被多次抽样的概率越来越小，从而导致样本的采样率较低。计算机本身计算能力也限制了强化学习算法的效果。且目前基于强化学习求解车辆路径问题的模型，主要都是对客户节点小于

100 的问题进行求解，很少有作者对大规模车辆路径问题进行求解。因此，未来基于强化学习模型求解车辆路径问题的工作可从以下方面展开：

（1）建立更小误差的环境模型。对于有模型的强化学习算法，精确的环境模型可以减少个体和环境的交互，可以通过模型仿真生成大量样本数据，使得算法快速学习。通过更加科学的方式定义参数，减少人为设定引起不确定性，可增强模型的可靠性。但当数据量较少的时候，学到的模型误差较大，且现实生活中的车辆路径问题的环境动力学较为复杂，从而使得建立精确模型更为困难。

（2）提高数据采样率。针对无模型的强化学习算法数据采样率低的问题，可以重新设计采样方法，使用离轨策略设计并行架构进行学习；针对有模型的强化学习算法采样率低的问题，可以将模型误差纳入模型设计中，从而提高数据采样率。

（3）设计更加高效的模型。目前求解车辆路径的深度强化学习模型中，运用深度神经网络直接求解得到初始解的质量较差，大多数模型都需要与其他方法结合提升解的质量，使得求解时间变长。因而，未来可以对神经网络的理论基础进行深入研究，设计更加高效的深度网络结构和表示方法，得到更加高效模型。尤其是图神经网络与新型注意力机制结合是未来的重点研究方向之一。

（4）选择更加合适模型训练算法。目前基于编码器-解码器求解车辆路径问题的模型，常直接选用 REINFORCE 算法、Actor-Critic 算法等常规的强化学习算法，不同类型的车辆路径问题优化目标不同，能有效解决复杂问题的多智能体强化学习和分层强化学习尚处在探索阶段，如何选用更新高效强化学习算法作为模型的训练方法，对提升模型性能至关重要。这也是一个重要的改进方向。

（5）提升模型稳定性。现有的深度强化学习模型一旦训练完成，就可以求解同类型的问题，泛化能力较强，但是优化效果无法保证。局部搜索算法可移植性差，但优化效果较好。因此，运用更加高效的强化学习算法提升一些经典局部搜索算法的求解速度，是未来重要的研究内容。

（6）提高解决现实工程问题的能力。车辆路径问题和现实生活息息相关，通常具有多约束、动态变化的特点，当前的强化学习算法和深度强化学习模型可以求解的车辆路径问题约束较少、规模较小。深度学习模型的最大优势就是可以在大规模问题上有良好表现，因此，设计更加契合车辆路径问题的深度神经网络架构是有效求解大规模、复杂车辆路径问题的方法之一，这也是未来可着重研究的方向。

强化学习已成为热门的组合优化问题求解方法之一，但就目前而言，基于强化学习求解车辆路径问题的算法模型并不具备真正的工程应用能力。随着研究不断深入和理论不断创新，强化学习将会和其他最新先进技术结合，让强化学习在车辆路径问题求解中发挥更大作用。

## 参考文献：

[1] DANTZIG G B, RAMSER J H. The truck dispatching problem[J]. Management science, 1959, 6(1): 80-91.

[2] LENSTRA J K, KAN A H G R. Complexity of vehicle routing and scheduling problems[J]. Networks, 1981, 11(2): 221-227.

[3] CLARKE G, WRIGHT J W. Scheduling of vehicles from a central depot to a number of delivery points[J]. Operations research, 1964, 12(4): 568-581.

[4]JUNGER M, REINELT G, RINALDI G. The traveling salesman problem[J]. Handbooks in operations research and management science, 1995, 7: 225-330.

[5] VIDAL T, CRAINIC T G, GENDREAU M, et al. Heuristics for multi-attribute vehicle routing problems: A survey and synthesis[J]. European Journal of Operational Research, 2013, 231(1): 1-21.

[6] ULMER M W, THOMAS B W. Same‐day delivery with heterogeneous fleets of drones and vehicles[J]. Networks, 2018, 72(4): 475-505.

[7] MANDZIUK J. New shades of the vehicle routing problem: emerging problem formulations and computational intelligence solution methods[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2018, 3(3): 230-244.

[8] CAMM J D, MAGAZINE M J, KUPPUSAMY S, et al. The demand weighted vehicle routing problem[J]. European Journal of Operational Research, 2017, 262(1): 151-162.

[9] LU D, GZARA F. The robust vehicle routing problem with time windows: Solution by branch and price and cut[J]. European Journal of Operational Research, 2019, 275(3): 925-938.

[10] LI H, CHANG X, ZHAO W, et al. The vehicle flow formulation and savings-based algorithm for the rollon-rolloff vehicle routing problem[J]. European Journal of Operational Research, 2017, 257(3): 859-869.

[11] MOONS S, RAMAEKERS K, CARIS A, et al. Integrating production scheduling and vehicle routing decisions at the operational decision level: a review and discussion[J]. Computers & Industrial Engineering, 2017, 104: 224-245.

[12] KONG Y, TANG J F, DONG G, et al. An insertion algorithm for vehicle scheduling in picking up and delivering customers to airport[J]. Control Theory & Applications, 2009, 26(1): 92-96.

[13] 穆东, 王超, 王胜春, 等. 基于并行模拟退火算法求解时间依赖型车辆路径问题. 计算机集成制造系统, 2015, 21(6): 1626–1636.
MU D, WANG C, WANG S C, et al. Solving time-dependent vehicle routing problem based on parallel simulated annealing algorithm[J]. Computer Integrated Manufacturing System, 2015, 21(6): 1626–1636.

[14] WANG Y, WU Q, GLOVER F. Effective metaheuristic algorithms for the minimum differential dispersion problem[J]. European Journal of Operational Research, 2017, 258(3): 829-843.

[15]PINTEA C M, CHIRA C, DUMITRESCU D, et al. Senstive ants in solving the generalized vehicle routing problem[J]. International Journal of Computers Communications & Control, 2011, 6(4): 734 –741.

[16]FERNANDEZ-V ARGAS J A, BO NILLA-PETRICIOLET A. Development of a global optimization algorithm in ant colonies with feasible region selection for continuous search spaces[J]. Revista Internacional De Metodos Numericos Para Calculo Y Diseno En Ingenieria, 2014, 30(3): 178 – 187.

[17]JIA Y H, CHEN W N, GU T L, et al. A dynamic logistic dispatching system with set-based particle swarm optimization[J]. IEEE Transactions on Systems Man Cybernetics—Systems, 2018, 48(9): 1607 –1621.

[18] SUTTON R S. Learning to predict by the methods of temporal differences[J]. Machine learning, 1988, 3(1): 9-44.

[19] WATKINS C J C H, DAYAN P. Q-learning[J]. Machine learning, 1992, 8(3-4): 279-292.

[20] MNIH V, KAVUKCUOGLU K, Silver D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602, 2013.

[21] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine learning, 1992, 8(3): 229-256.

[22] PETERS J, SCHAAL S. Natural actor-critic[J]. Neurocomputing, 2008, 71(7-9): 1180-1190.

[23] HA D, SCHMIDHUBER J. World models[J]. arXiv preprint arXiv:1803.10122, 2018.

[24] SECOMANDI N. Comparing neuro-dynamic programming algorithms for the vehicle routing problem with stochastic demands[J]. Computers & Operations Research, 2000, 27(11-12): 1201-1225.

[25] TATARAKIS A, MINIS I. Stochastic single vehicle routing with a predefined customer sequence and multiple depot returns[J]. European Journal of Operational Research, 2009, 197(2): 557-571.

[26] POWELL W B, SIMAO H P, BOUZAIENE-AYARI B. Approximate dynamic programming in transportation and logistics: a unified framework[J]. EURO Journal on Transportation and Logistics, 2012, 1(3): 237-284.

[27] CIMEN M, SOYSAL M. Time-dependent green vehicle routing problem with stochastic vehicle speeds: An approximate dynamic programming algorithm[J]. Transportation Research Part D: Transport and Environment, 2017, 54: 82-98.

[28] ULMER M W, MATTFELD D C, Köster F. Budgeting time for dynamic vehicle routing with stochastic customer requests[J]. Transportation Science, 2018, 52(1): 20-37.

[29] ULMER M W, GOODSON J C, MATTFELD D C, et al. Offline–online approximate dynamic programming for dynamic vehicle routing with stochastic requests[J]. Transportation Science, 2019, 53(1): 185-202.

[30] KOK A L, HANS E W, SCHUTTEN J M J. Vehicle routing under time-dependent travel times: the impact of congestion avoidance[J]. Computers & operations research, 2012, 39(5): 910-918.

[31] SECOMANDI N, MARGOT F. Reoptimization approaches for the vehicle-routing problem with stochastic demands[J]. Operations research, 2009, 57(1): 214-230.

[32] GOODSON J C, OHLMANN J W, THOMAS B W. Rollout policies for dynamic solutions to the multivehicle routing problem with stochastic demand and duration limits[J]. Operations Research, 2013, 61(1): 138-154.

[33] KOOL W, VAN HOOF H, GROMICHO J, et al. Deep Policy Dynamic Programming for Vehicle Routing Problems[J]. arXiv preprint arXiv:2102.11756, 2021.

[34] JOE W, LAU H C. Deep reinforcement learning approach to solve dynamic vehicle routing problem with stochastic customers[C]//Proceedings of the International Conference on Automated Planning and Scheduling. 2020, 30: 394-402.

[35] ZHANG C, DELLAERT N P, ZHAO L, et al. Single vehicle routing with stochastic demands: approximate dynamic programming[J]. Dept. Ind. Eng., Tsinghua Univ., Beijing, China, Tech. Rep, 2013, 425.

[36] BOUHAMED O, GHAZZAI H, BESBES H, et al. Q-learning based routing scheduling for a multi-task autonomous agent[C]//2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS). IEEE, 2019: 634-637.

[37] BDEIR A, BOEDER S, DERNEDDE T, et al. RP-DQN: An application of Q-Learning to Vehicle Routing Problems[J]. arXiv preprint arXiv:2104.12226, 2021.2

[38] CHEN X, ULMER M W, THOMAS B W. Deep Q-learning for same-day delivery with vehicles and drones[J]. European Journal of Operational Research, 2021.

[39] ULMER M W, GOODSON J C, MATTFELD D C, et al. On modeling stochastic dynamic vehicle routing problems[J]. EURO Journal on Transportation and Logistics, 2020: 100008.

[40] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[C]//

International conference on machine learning. PMLR, 2016: 1995-2003.

[41] ZHANG W, WANG Q, LI J, et al. Dynamic fleet management with rewriting deep reinforcement learning[J]. IEEE Access, 2020, 8: 143333-143341.

[42] 刘建伟, 高峰, 罗雄麟. 基于值函数和策略梯度的深度强化学习综述[J]. 计算机学报, 2019, 42(6): 1406-1438.
LIU J W, G F, L X L.Survey of deep reinforcement learning based on value function and policy gradient[J]. Chinese Journal of Computers, 2019, 42(6): 1406-1438.

[43] VINYALS O, FORTUNATO M, JAITLY N. Pointer networks[J]. Advances in Neural Information Processing Systems, 2015, 28: 2692-2700.

[44] NAZARI M, OROOJLOOY A, TAKAC M, et al. Reinforcement learning for solving the vehicle routing problem[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018: 9861-9871.

[45] XIN L, SONG W, CAO Z, et al. Step-wise deep learning models for solving routing problems[J]. IEEE Transactions on Industrial Informatics, 2020, 17(7): 4861-4871.

[46] KOOL W, VAN HOOF H, WELLING M. Attention, learn to solve routing problems[C]//International Conference on Learning Representations, 2018.

[47] FALKNER J K, SCHMIDT-THIEME L. Learning to solve vehicle routing problems with time windows through joint attention[J]. arXiv preprint arXiv:2006.09100, 2020.

[48] XIN L, SONG W, CAO Z, et al. Multi-decoder attention model with embedding glimpse for solving vehicle routing problems[C]//Proceedings of 35th AAAI Conference on Artificial Intelligence. 2021.

[49] ZHANG K, HE F, ZHANG Z, et al. Multi-vehicle routing problems with soft time windows: A multi-agent reinforcement learning approach[J]. Transportation Research Part C: Emerging Technologies, 2020, 121: 102861.

[50] XU Y, FANG M, CHEN L, et al. Reinforcement learning with multiple relational attention for solving vehicle routing problems[J]. IEEE Transactions on Cybernetics, 2021.

[51] PENG B, WANG J, ZHANG Z. A deep reinforcement learning algorithm using dynamic attention model for vehicle routing problems[C]//International Symposium on Intelligence Computation and Applications. Springer, Singapore, 2019: 636-650.

[52] LU H, ZHANG X, YANG S. A learning-based iterative method for solving vehicle routing problems[C]//International Conference on Learning Representations. 2019.

[53] HOTTUNG A, TIERNEY K. Neural large neighborhood search for the capacitated vehicle routing problem[J]. arXiv preprint arXiv:1911.09539, 2019.

[54] CHEN X, TIAN Y. Learning to perform local rewriting for combinatorial optimization[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 6281-6292.

[55] ZHAO J, MAO M, ZHAO X, et al. A hybrid of deep reinforcement learning and local search for the vehicle routing problems[J]. IEEE Transactions on Intelligent Transportation Systems, 2020.

[56] LI J, XIN L, CAO Z, et al. Heterogeneous attentions for solving pickup and delivery problem via deep reinforcement learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2021.

[57] GAO L, CHEN M, CHEN Q, et al. Learn to design the heuristics for vehicle routing problem[J]. arXiv preprint arXiv:2002.08539, 2020.

[58] VERA J M, ABAD A G. Deep reinforcement learning for routing a heterogeneous fleet of vehicles[C]//2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI). IEEE, 2019: 1-6.