基于图注意力机制的车辆路径问题研究

王 骊[1] 翁慧颖[1] 孙小江[1]
WANG Li WENG Huiying SUN Xiaojiang

摘 要 车辆路径问题是组合优化中的经典问题，近年来，基于强化学习的深度学习框架已经成为车辆路径问题的主流深度学习框架。提出一种启发式神经网络算法，通过破坏算子产生节点子集，再依据最小成本原则重构序列修复算子。在网络中，编码器由带有注意力机制的图神经网络组成，解码器由带有指针网络的 GRU 组成，所提出的网络由 actor-critic 框架来进行训练。实验结果表明，所提出的模型性能优于经典启发式算法。

关键词 车辆路径问题；组合优化；强化学习；图注意网络

doi：10.3969/j.issn.1672-9528.2024.02.026

0 引言

车辆路径问题（vehicle routing problem，VRP）是一种在应用数学和计算机科学领域研究了数十年的经典的 NP-hard 组合优化问题（combinatorial optimization problem，COP），在无人驾驶、物流调度和交通等领域有着广泛的应用，通过合理地规划路线，提高工作和运输效率。基于物流调度的 VRP 可简易描述为，针对一定数量的客户和相应的货物需求，配送中心组织适当的行车路线发送货物，使得客户的需求得到满足，并能在一定的约束下，达到路程最短、成本最小、耗费时间最少的目标。解决 VRP 问题传统的算法 [1] 主要有精确算法、近似算法、启发式算法等。在考虑到算法精度、模型适用性等方面，这些方法都不能很好地解决路径问题。精确算法可以在小规模 COP 问题上的求解具有较好的结果，但是在求解大规模 COP 问题时，模型的求解速度过慢；近似算法是指对 COP 问题在合理的计算时间内，找到一个合理的近似解，但是大多数 COP 问题没有近似解；启发式算法常见的有蚁群算法 [2]、禁忌算法 [3]，但该类算法针对特定问题进行人工设计，当问题描述发生变化时，算法需要重新设计，且该类算法的解容易陷入局部最优、搜索时间较长。

近年来，深度强化学习（deep reinforcement learning，DRL）[4] 在决策问题中得到广泛应用。深度强化学习方法可以发现数据之下的隐含特征，在求解问题的过程中可以自动学习求解策略。该方法主要是以端到端的形式输出解，模型基于编码器、解码器结构，通过在目标解中添加点和边的策略来构造出完整的解。模型在训练完成后，路径问题的解可以直接输出，无需再次训练。对于求解 VRP 的算法中，DRL 相关理论及其应用已经取得了良好的进展，但是目前还是存在一些问题，例如，DRL 求解 COP 问题的模型泛化能力较差，对复杂路径优化问题的处理有待改进；基于图神经网络（graph neural network，GNN）在拓扑结构以及点与点之间潜在的关系上捕捉不足。

针对上面的问题，本文提出一种新方法，该方法通过迭代逐步收敛到可行解。该方法中，在编码阶段，利用注意力机制对节点嵌入和边嵌入，再由一个改进版本图注意力机制（graph attention network，GAT）[5] 作为其需求的拓扑联合表示；在解码阶段，利用 GRU 来生成最后的结果。本文的方法使得图嵌入的向量空间有着更丰富的计算形式，对图拓扑结构的潜在关系能进行更有效的捕捉。

1 相关工作

深度学习在序贯决策任务中，有着重要的发展。传统的强化学习方法主要依赖于手工设计的特征和函数逼近器来表示状态和动作的价值函数，这在处理复杂的决策问题时面临挑战。深度神经网络可用于近似值函数、策略函数或者动作 - 值函数，从而实现对状态和动作的学习和表示。同时，深度学习可以自动从原始数据中提取有用的特征，避免了手工设计特征的复杂性。Vinyals 等人 [6] 利用神经网络解决 VRP 问题，将其类比为机器翻译过程，提出了指针网络（pointer network，PN）。在该网络中，利用长短期记忆网络（LSTM）作为编码器，注意力机制（AM）作为解码器，从数值坐标中提取特征。PN 采用有监督的方式训练网络，对标签的质量有很高的要求。GNN 是另外一种解决 VRP 问题的算法。

1. 国网浙江省电力有限公司物资分公司 浙江杭州 310003

Scarselli 等人[7]利用 GNN 对节点特征进行学习，通过编、解码器，以自回归的方式逐步构造解，然后根据学习到的特征对后续节点进行预测。Ma 等人[8]将 PN 与 GNN 结合，提出图指针网络（GPN），利用 GNN 得到节点特征，再利用 PN 构造解，使得网络的泛化能力大大加强。受 transformer 架构的启发，Kool 等人[9]以 transformer 为框架，将输入元素分为静态元素和动态向量。在编码阶段，以嵌入的方式对静态元素进行向量表示；在解码阶段，通过循环神经网络将静态元素与动态向量相结合，采用 AM 获得下一个决策的概率分布，这使得路径优化性能大幅增加。由于基于改进 transformer 的模型效果较好，后来的学者在此基础上作了更进一步的研究。Kwon 等人[10]提出一种多目标最优策略框架 POMO，该框架利用 DRL 训练多节点的多头注意力机制，对 TSP 和 VRP 的求解效果良好。Falkner 等人[11]采用修复和破坏算子，通过局部搜索和维护少量候选解来扩展大邻域搜索（LNS）。LNS 主要利用学习模型来重新组合已经被破坏的解，并通过引入随机性来有效地探索大邻域。Ma 等人[12]在此基础上建立了一种双面协同转换模型（DACT）和一种循环编码方法。该方法在一定程度上缓解 PE 算法信息提取不完整的缺陷。上述模型与早期的模型相比，效果有较大的提升，解决直接应用 transformer 求解效率低下的问题。

2 模型

2.1 问题定义

VRP 问题可以定义为一个图$\mathbb{G}=(v,\varepsilon)$，其中 $v=\{0,\cdots,n\}$，节点 $i=0$ 是仓库，$i\in\{0,\cdots,n\}$ 是客户。$\varepsilon=\{a_{i,j}\}$，$i,j\in v$，是 v 中节点的边。客户点 i 的物资需求量为 q_i，在客户点 i 的最早开始时间为 s_i，最晚的结束时间为 e_i。VRP 的目标是找到一个或几个 Hamiltonian 环，使得每个客户节点的物资需求 q_i 能够被满足，由 K 辆汽车完成所有任务，在满足时间窗口和车辆容量限制的情况下，最小化总行驶成本。

节点信息和拓扑结构信息分别包含在节点和边的嵌入中。第 i 个节点信息为一个 8 维向量，这个向量中包含：客户点的开始时间 s_i、客户点的结束时间 e_i、物资需求量 q_i、相关路线的总需求、相关路线到该节点的总需求、沿该路线到该节点的总行驶距离、沿该路线到该节点的行驶时间以及一个偏移量。节点 n_i 到节点 n_j 的边 $a_{i,j}$ 是一个二维向量，它包括两个节点之间的行驶距离以及判断这个边是否存在的二元指标。这些元素经过嵌入编码后，输入到图注意力网络（GAN）[13]中，然后利用 GAN 捕捉节点之间的拓扑关系。

2.2 编码器

在为 VRP 建立的有向图模型中，节点信息的更新不仅和前一节点有关系，还和嵌入的边有关系。GAN 利用注意力机制来更新节点信息，对图中的结构信息有着很强的表征能力。这里用 $e_{i,j}$ 表示边的编码信息，在边和节点的基础上加上一个完全连接层，具体的公式为：

$$\tilde{n}_{q,i}=W_q n_i \tag{1}$$

$$\tilde{n}_{k,j}=W_k n_j \tag{2}$$

$$\tilde{e}_{v,i,j}=W_v e_{i,j} \tag{3}$$

式中：W_q、W_k、W_v 是需要训练的参数。在对 n_i、n_j、$e_{i,j}$ 进行编码后，利用注意力机制来提取相关信息，具体公式如下文所示。

$$Attention(i,j)=softmax(\frac{\tilde{n}_{q,i}\tilde{n}_{k,i}^{T}}{\sqrt{d}})\tilde{e}_{v,i,j} \tag{4}$$

ReLU

式中：W_L 为需要训练的参数，d 为模型的嵌入维数，$softmax(\cdot)$ 和 ReLU 为激活函数，经过注意力机制处理后的新的编码信息 $n_{new,i}$ 的表达式为：

$$n_{new,i}=\tilde{n}_{q,i}+\sum_j \tilde{w}_{i,j}\otimes\tilde{n}_{k,j} \tag{5}$$

最终得到的 $n_{new,i}$ 不仅包含着其他节点的信息还包含着相应边的信息。以上得到的 $n_{new,i}$ 通过一层图注意力机制来得到的，在实际训练过程中，可以串联多层，使得每个节点能有大的感受野，获得的信息更充分，最后在模型的输出前加上一个平均池化层得到最终的输出。模型编码示意图如图 1 所示。

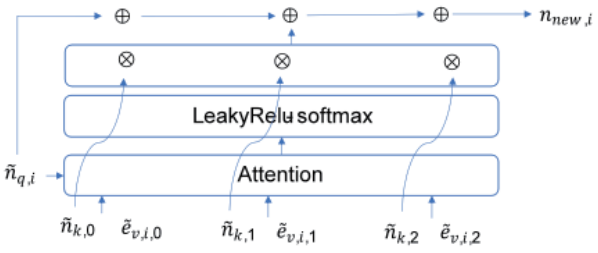

图 1 编码示意图

2.3 解码器

本文的解码器通过 VLNS（very large-scale neighborhood search）来生成目标序列。破坏算子产生节点子集，然后修复算子依据最小成本原则重构这个序列。这个序列子集可以由$P=\pi([\eta_1,\eta_2,...,\eta_M])$表示。

这里的$\eta_m\in\mathcal{V}$ ($m=1,2,...,M$) 是将要移除的候选点，[·] 代表着序列集合，$\pi(\cdot)$ 代表着随机策略。随机策略是一个联合概率分布，可以用链式法则分解如下。

$$P=\pi(\eta_1)\pi(\eta_M|[\eta_1,...,\eta_{M-1}])\prod_{m=2}^{M}\pi(\eta_m|[\eta_1,...,\eta_{m-1}]) \tag{6}$$

在给定$[\eta_1,...,\eta_{m-1}]$后，本文根据循环神经网络来生成η_m和$\pi(\eta_m|[\eta_1,...,\eta_{m-1}]$。编码阶段的生成遵从于指针网络。指针网络的输入包括前面介绍的图神经网络编码 $n_{new,i}$ 以及 GRU 的隐藏层节点。解码过程中，GRU 的输出节点嵌入到注意力机制中，它将计算编码图中的每个节点，然后计算分数，最后应用 softmax 层来得到概率分布。这使得编码层能在任何时候看到整个$\mathbb{G}=(v,\varepsilon)$ 以及最终的有序列表输出。解码器以一定的概率来产生要移除的节点充当销毁算子，并且通过遵循产生的节点的序列顺序来充当修复算子。

2.4 模型训练

本文的优化网络选择强化学习中的 actor-critic 框架。actor 是决策者，根据当前状态选择最优行动，也就是 actor 会观察环境状态来产生一个行动策略，使得当前状态下的选择策略能获得最大奖励。critic 是评价者，对 actor 做出的决策进行评估，并与真实奖励进行评估。定义编码器和解码器中的参数为 θ，在解码步骤 t 中，VRP 的成本函数为策略的奖励，成本函数为总行驶距离和车辆的成本之和，具体公式为：

$$Cost_{VRP}^{(t)} = Dis^{(t)} + C \times K^{(t)} \tag{7}$$

$$r^{(t)} = Cost_{VRP}^{(t)} - Cost_{VRP}^{(t-1)} \tag{8}$$

式中：$K^{(t)}$ 表示在步骤 t 中被使用的车辆数量。C 表示为车辆成本的权重，$r^{(t)}$ 为对应的奖励值。编码器的输出 $Enc^{(t)}$ 作为值网络在步骤 t 的状态，在该步骤估计的状态值为 $v(Enc^{(t)},\phi)$，其中 ϕ 为值网络的参数。在本文的值网络层中，第一层为一个带有 ReLU 激活函数的密集层，第二层为一个线性连接层。值函数中，计算前后两个步骤中的 TD 误差，具体的计算公式为：

$$\delta_{TD}^{(t)} = r^{(t)} + \gamma v\left(Enc^{(t)},\phi\right) - v\left(Enc^{(t-1)},\phi\right) \tag{9}$$

式中：$\delta_{TD}^{(t)}$为 TD 误差，γ为折扣因素，critic 网络的参数更新公式为：

$$\phi = \phi + \alpha_\phi \delta_{TD}^{(t)} \left(\partial v\left(Enc^{(t)},\phi\right)/\partial\phi\right) \tag{10}$$

式中：α_ϕ 为 critic 网络的学习率。actor 网络的训练采用带有裁剪的最近邻优化算法 [14]，在目标函数提升过大时就会进行裁剪，但是不会限制目标函数的下降，这样通过裁剪限制策略进行大幅度的更新导致的分布差异过大。具体公式如下具体公式为：

$$L^{CLIP}(\theta) = E_t[\min(r_t(\theta)\delta_{TD}^{(t)}, clip(r_t(\theta),1-\epsilon,1+\epsilon)\delta_{TD}^{(t)})] \tag{11}$$

式中：$r_t(\theta)$为新策略覆盖旧策略的比率，ϵ为权重限制率为 0.2。此外，解码器状态栏的更新还需要满足以下的关系。

$$Dis^{(t)} < Dis^{(t-1)} - T^{(t)} \cdot \log(Rnd) \tag{12}$$

式中：Rnd 为一个均匀分布在 [0, 1] 之间的随机数，$T^{(t)}$ 为模拟退火中温度，其衰减率为 α_T。

3 实验

本文在带有时间窗口限制和容量限制的 VRP 作为实验，100 个点随机生成，第一个点为仓库，剩下的节点为客户。每个客户的需求量为一个 [1, 9] 的随机值，车的容量为 100。开始时间在 [0, 290] 之间均匀分布，到达时间在 [10, 300] 之间均匀分布，服务时间为 10，仓库没有服务时间且其时间窗口为 [0, 300]。

训练集与测试集之间的比例为 9:1，测试集用来比较基准模型与本文模型之间的性能。8 维节点的嵌入表示维度大小为 N_E，2 维边的嵌入大小为 N_I，编码层的输出大小维度为 N_E。

在解码器中，单元的隐藏层维度为 N_D。Critic 网络为一个全连接层，其隐藏层的维度为 N_C，其输出为一个标量值。在训练过程中，模型的块大小为 64，TD 误差的步数为 10，编码器的层数为 2，节点的嵌入维度大小 N_E，隐藏层维度 N_D、N_C 都为 64，边的嵌入维度 N_I 为 16，学习率为 3e-4。本次的实验是在 Colab 平台上运行的，其运行内存为 12 GB，运行语言为 Python，版本为 3.10，模型由 PyTorch 搭建，版本为 2.0，数据通过 GPU 运行。

本文的对比算法有两类，启发式算法和神经网络优化算法。启发式算法有三种 [15]，分别是随机启发式（random）、自适应大领域搜索（adaptive large neighbourhood search，ALNS）[16]、移除诱导算法（slack induction by string removals，SISR）[17]。随机启发式算法中，算子为一个随机函数；自适应大领域搜索算法在运行时以自适应方式选择最佳启发式；移除诱导算法根据当前路由和节点之间的组合距离来删除节点。神经网络优化算法，主要是基于注意力机制的算法，对于解码，该方法中主要运用贪婪搜索和直接运用 softmax 作为结果。迭代次数统一为 100 次，以平均成本作为评价值。

如表 1 所示，Rondom、ALNS、SISR 为三种启发式算法，AM、AM-Greedy[18] 为两种神经网络算法，其中 AM 解码方式直接运用 softmax 作为结果，AM-Greedy 使用贪婪搜索算法作为解码方式。从表 1 中可以看出，基于神经网络的优化算法，平均成本小于启发式算法，本文提出的模型的平均成本最小，体现了模型的优异。

表 1 模型对比

模型名称	平均成本
Random	1 284.15
ALNS	1 245.26
SISR	1 230.84
AM	1 222.37
AM-Greedy	1 230.51
本文模型	1 196.12

4 结论

本文提出一种解决车辆路径问题的新的模型结构。通过启发方式来生成目标序列，破坏算子产生节点子集，然后修复算子依据最小成本原则重构这个序列。模型由一个编码器和解码器组成：编码器在图注意网络的基础上在节点与边之间加入注意力机制，使得节点和边所带有的信息能够嵌入到一起进行信息传播，堆叠多层编码器，使得拓扑图上的节点的感受野范围更广；在解码器中，通过利用指针网络来连接节点之间的关系，随后利用 GRU 结构来生成最终的序列。在带有时间窗口的 VPR 数据上，本文模型优于三种启发式模型和两种神经网络模型。

参考文献：

[1] 杨笑笑，柯琳，陈智斌 . 深度强化学习求解车辆路径问题的研究综述 [J]. 计算机工程与应用，2023, 59(5):1-13.

[2] YU B, YANG Z Z, YAO B. An improved ant colony optimization for vehicle routing problem[J]. European journal of operational research, 2009,196(1):171-176.

[3] GOEKE D. Granular tabu search for the pickup and delivery problem with time window and electric vehicles[J].European journal of operational research, 2019,278(3):821- 836.

[4] LIU Q, ZHAI J W, ZHANG Z Z, et al. A survey on deep reinforcement learning[J].Chinese journal of computers, 2018, 41(1): 1-27.

[5] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL].(2017-10-30)[2023-10-08]. https://arxiv.org/abs/1710.10903.

[6] VINALYS O, FORTUNATO M, JAITLY N. Pointer networks [EB/OL].(2015-06-09)[2023-10-10].https://arxiv.org/abs/1506.03134.

[7] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J].IEEE transactions on neural networks, 2008, 20(1): 61-80.

[8] MA Q, GE S, HE D, et al. Combinatorial optimization by graph pointer networks and hierarchical reinforcement learning[EB/OL].(2019-11-12)[2023-09-25].https://arxiv.org/abs/1911.04936.

[9] KOOL W, VAN HOOF H, WELLING M. Attention, learn to solve routing problems[EB/OL].(2018-05-22)[2023-10-11]. https://arxiv.org/abs/1803.08475.

[10] KWON Y D, CHOO J, KIM B, et al. Pomo: policy optimization with multiple optima for reinforcement learning [EB/OL]. (2020-10-30)[2023-11-01].https://arxiv.org/abs/2010.16011.

[11] FALKNER J K, THYSSENS D, SCHMIDT-THIEME L. Large neighborhood search based on neural construction heuristics[EB/OL].(2022-05-02)[2023-11-06].https://arxiv.org/abs/2205.00772.

[12] MA Y , LI J, CAO Z, et al. Learning to iteratively solve routing problems with dual-aspect collaborative transformer [EB/OL].(2021-10-06)[2023-09-29].https://arxiv.org/abs/2110.02544.

[13] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[14] 刘娟，万静 . 自然反向最近邻优化的密度峰值聚类算法 [J]. 计算机科学与探索，2021, 15(10): 1888-1899.

[15] 陈红华，崔祜龙，王耀杰 . 基于多种云环境的任务调度算法综述 [J]. 计算机应用研究，2023,40(10): 2889-2895.

[16] STEFAN R, DAVID P. An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows[J].Transportation science, 2006, 40: 455-472.

[17] JAN C, GREET V B. Slack induction by string removals for vehicle routing problems[J]. Transportation science, 2020, 54(2): 417-433.

[18] KOOL W, VAN HOOF H, GROMICHO J, et al. Deep policy dynamic programming for vehicle routing problems[C]// International conference on integration of constraint programming, artificial intelligence and operations research. Cham: Springer International Publishing, 2022: 190-213.

【作者简介】

王骊（1981—），女，浙江杭州人，硕士研究生，副主任，研究方向：仓储配送管理。

翁慧颖（1977—），女，浙江杭州人，硕士研究生，研究方向：物资管理。

孙小江（1982—），男，浙江杭州人，硕士研究生，主任，研究方向：供应履约管理。

（收稿日期：2023-12-04）