基于深度强化学习的组合优化研究进展

李凯文[1, 2] 张 涛[1, 2] 王 锐[1, 2] 覃伟健[1, 2] 贺惠晖[1, 2] 黄 鸿[1, 2]

摘 要 组合优化问题广泛存在于国防、交通、工业、生活等各个领域, 几十年来, 传统运筹优化方法是解决组合优化问题的主要手段, 但随着实际应用中问题规模的不断扩大、求解实时性的要求越来越高, 传统运筹优化算法面临着很大的计算压力, 很难实现组合优化问题的在线求解. 近年来随着深度学习技术的迅猛发展, 深度强化学习在围棋、机器人等领域的瞩目成果显示了其强大的学习能力与序贯决策能力. 鉴于此, 近年来涌现出了多个利用深度强化学习方法解决组合优化问题的新方法, 具有求解速度快、模型泛化能力强的优势, 为组合优化问题的求解提供了一种全新的思路. 因此本文总结回顾近些年利用深度强化学习方法解决组合优化问题的相关理论方法与应用研究, 对其基本原理、相关方法、应用研究进行总结和综述, 并指出未来该方向亟待解决的若干问题.

关键词 深度强化学习, 组合优化问题, 深度神经网络, 图神经网络, 指针网络

引用格式 李凯文, 张涛, 王锐, 覃伟健, 贺惠晖, 黄鸿. 基于深度强化学习的组合优化研究进展. 自动化学报, 2021, **47**(11): 2521−2537

DOI 10.16383/j.aas.c200551

Research Reviews of Combinatorial Optimization Methods Based on Deep Reinforcement Learning

LI Kai-Wen[1, 2] ZHANG Tao[1, 2] WANG Rui[1, 2] QIN Wei-Jian[1, 2] HE Hui-Hui[1, 2] HUANG Hong[1, 2]

Abstract Combinatorial optimization problems widely exist in various fields such as national defense, transportation, industry and life. For decades, traditional operational research methods are the main means to solve combinatorial optimization problems. However, with the increase of problem size in practical applications and the increasing demands for real-time optimization, traditional methods suffer from great computational burdens, and it is difficult to realize the online solution of combinatorial optimization problems. In recent years, with the rapid development of deep learning technology, the achievements of deep reinforcement learning in AlphaGo, robot and other fields show its strong learning ability and sequential decision-making ability. In view of this, in recent years, a number of new methods using deep reinforcement learning to solve combinatorial optimization problems have emerged, which have the advantages of fast solving speed and strong model generalization ability. It provides a new idea for solving combinatorial optimization problems. Therefore, this paper summarizes and reviews the theoretical methods and application researches of this kind of methods in recent years.

Key words Deep reinforcement learning, combinatorial optimization problems, deep neural network, graph neural networks, pointer networks

Citation Li Kai-Wen, Zhang Tao, Wang Rui, Qin Wei-Jian, He Hui-Hui, Huang Hong. Research reviews of combinatorial optimization methods based on deep reinforcement learning. *Acta Automatica Sinica*, 2021, **47**(11): 2521−2537

收稿日期 2020-07-14 录用日期 2020-11-04
Manuscript received July 14, 2020; accepted November 4, 2020
国家自然科学基金面上项目 (61773390), 湖湘青年英才计划 (2018RS3081), 科技委国防创新特区项目 (193-A11-101-03-01), 国防科技大学自主科研计划 (ZZKY-ZX-11-04) 资助
Supported by National Natural Science Foundation of China (61773390), the Hunan Youth Elite Program (2018RS3081), the National Defense Innovation Zone Project of Science and Technology Committee (193-A11-101-03-01), and the Scientific Project of National University of Defense Technology (ZZKY-ZX-11-04)
本文责任编委 魏庆来
Recommended by Associate Editor WEI Qing-Lai
1. 国防科技大学系统工程学院 长沙 410073 2. 多能源系统智慧互联技术湖南省重点实验室 长沙 410073
1. College of System Engineering, National University of Defense Technology, Changsha 410073 2. Hunan Key Laboratory of Multi-Energy System Intelligent Interconnection Technology, Changsha 410073

组合优化问题 (Combinatorial optimization problem, COP) 是一类在离散状态下求极值的最优化问题, 其数学模型如下所示:

$$\begin{aligned} &\min F(x)\\ &\text{s.t. } G(x) \geq 0\\ &x \in D \end{aligned} \tag{1}$$

其中 x 为决策变量、$F(x)$ 为目标函数、$G(x)$ 为约束条件, D 表示离散的决策空间, 为有限个点组成的集合. 组合优化问题在国防、交通、产品制造、管理决策、电力、通信等领域都有广泛的应用[1], 常见的组合优化问题包括旅行商问题 (Traveling sales-

man problem, TSP)、车辆路径问题 (Vehicle routing problem, VRP)、车间作业调度问题 (Job-shop scheduling)、背包问题 (Knapsack)、最小顶点覆盖问题 (Minimum vertex cover, MVC)、最小支配集问题 (Minimum dominating problem, MDP) 等.

组合优化问题的特点是其决策空间为有限点集, 直观上可以通过穷举法得到问题的最优解, 但是由于可行解数量随问题规模呈指数型增长, 无法在多项式时间内穷举得到问题的最优解[2], 为此数十年来学者对组合优化问题的求解算法进行了大量的研究, 目前求解组合优化问题的方法主要包括精确方法 (Exact approaches) 和近似方法 (Approximate approaches) 两大类:

1) 精确方法是可以求解得到问题全局最优解的一类算法, 主要包括分支定界法 (Branch and bound)[1, 3] 和动态规划法 (Dynamic programming)[4–5], 其均采用分而治之的思想通过将原问题分解为子问题的方式进行求解[2], 通过不断迭代求解得到问题的全局最优解.

2) 近似方法是可以求解得到问题局部最优解的方法, 主要包括近似算法 (Approximate algorithms) 和启发式算法 (Heuristic algorithms) 两类[2]. 近似算法是可以得到有质量保证的解的方法, 包括贪心算法、局部搜索算法、线性规划和松弛算法、序列算法等[6–8]; 启发式算法是利用设定的启发式规则对解空间进行搜索的一类方法, 能够在可行时间内找到一个较好的解, 但是对解的质量没有保证, 文献中用来求解组合优化问题的启发式算法主要包括模拟退火算法[9–10]、禁忌搜索[11–12]、进化算法[13] (如遗传算法[14–15], 差分进化算法[16–17] 等)、蚁群优化算法[18–19]、粒子群算法[20–21]、迭代局部搜索[22–23], 变邻域搜索[24–25] 等.

精确方法可以求解得到组合优化问题的全局最优解, 但是当问题规模扩大时, 该类算法将消耗巨大的计算量, 很难拓展到大规模问题; 相对于精确方法, 近似方法可以在可接受的计算时间内搜索得到一个较好的解. 基于群体智能的进化方法以及局部搜索等方法都是近年来的研究热点, 但是该类方法都是迭代型优化算法, 当问题规模很大时, 大量的迭代搜索仍然会导致较大的计算耗时, 近似方法仍然很难拓展到在线、实时优化问题. 此外, 一旦问题发生变化, 上述方法一般需要重新进行搜索求解, 或者通过不断试错对启发式规则进行调整以获得更好的效果, 计算成本高.

近年来随着人工智能技术的发展, 深度学习技术已经在很多领域打破了传统方法的壁垒, 取得了令人瞩目的突破性进展. 在计算机视觉领域, 十多年前学者们主要利用人工设计的算法进行特征提取以及图像处理, 但如今深度学习已经成为了当前的核心方法, 深度神经网络 (Deep neural networks, DNN) 可以自动地对图像的特征进行学习, 代替了人类的手工算法设计. 作为深度学习另外一个重要的分支, 深度强化学习 (Deep reinforcement learning, DRL) 主要用来做序贯决策, 即根据当前的环境状态做出动作选择, 并根据动作的反馈不断调整自身的策略, 从而达到设定的目标. 近年来深度强化学习在 AlphaGo Zero[26]、Atari[27] 等问题上的表现显示了其强大的学习能力和优化决策能力.

组合优化即在离散决策空间内进行决策变量的最优选择, 与强化学习的 “动作选择” 具有天然相似的特征, 且深度强化学习 “离线训练、在线决策” 的特性使得组合优化问题的在线实时求解成为了可能, 因此利用深度强化学习方法解决传统的组合优化问题是一个很好的选择. 鉴于此, 近些年涌现出了一系列利用深度强化学习方法解决组合优化问题的新方法, 在 TSP、VRP、Knapsack 等组合优化问题上取得了很好的效果. 相对于传统组合优化算法, 基于 DRL 的组合优化算法具有求解速度快、泛化能力强等一系列优势, 该类方法是近年来的研究热点.

由于基于 DRL 的组合优化方法是近年来新兴的研究领域, 尚未有文献对该类方法进行系统的研究和综述, 因此本文对近年来利用 DRL 方法求解组合优化问题的重要模型进行总结回顾, 对该类方法的基本原理进行介绍, 对各个算法的优缺点和优化性能进行总结和比较, 并指出未来该方向亟待解决的若干问题, 旨在为学者在该新兴方向的研究提供指导.

文章的结构组织如下: 第 1 节首先对基于深度强化学习的组合优化方法进行了概述, 对其产生、历史发展、方法分类以及优缺点进行了介绍; 第 2 节对基于深度强化学习解决组合优化问题的基本原理进行介绍; 第 3 节对当前主流的基于深度强化学习的组合优化方法进行了综述, 根据方法的不同类别, 对各个算法的原理、优缺点和优化性能进行了对比介绍; 第 4 节对该类方法在近年来的应用研究进行介绍; 最后对本文进行总结.

1 基于深度强化学习的组合优化: 概述

利用神经网络解决组合优化问题的方法最早可追溯至 Hopfield 等在 1985 年提出的 Hopfield 网络[28], 该神经网络用于求解 TSP 问题以及其他组合优化问题[29], 但是该神经网络网络每次只能学习并解决单个小规模 TSP 问题实例, 对于新给定的一个 TSP

问题需要从头开始再次训练, 相对于传统算法并没有优势.

神经网络真正能够有效解决组合优化问题是在 2015 年, Vinyals 等[30] 将组合优化问题类比为机器翻译过程 (即序列到序列的映射), 神经网络的输入是问题的特征序列 (如城市的坐标序列), 神经网络的输出是解序列 (如城市的访问顺序), Vinyals 等根据该思想, 对机器翻译领域的经典序列映射模型 (Sequence-to-sequence, Seq2Seq) 进行了改进, 提出了可以求解组合优化问题的指针网络模型 (Pointer network, Ptr-Net)[30], 其原理详见第 2 节, 作者采用监督式学习的方式训练该网络并在 TSP 问题上取得了较好的优化效果. 多年来传统的组合优化算法都是以 "迭代搜索"的方式进行求解, 但是 Vinyals 等的模型可以利用神经网络直接输出问题解, 开启了组合优化一个新的研究领域.

自 Ptr-Net 方法被提出后, 近三年来多个基于 Ptr-Net 架构的新方法在 NeurIPS, ICLR 等人工智能顶会上被相继提出[30–36], 在 TSP、VRP、Knapsack、多目标 TSP 等组合优化问题上显示了强大的优化效果, 由于监督式学习需要构造大量带标签的样本, 很难实际应用, 目前大多数研究均利用深度强化学习方法对模型进行训练.

除指针网络模型之外, 近年来随着图神经网络技术的兴起, 部分学者采用图神经网络对组合优化问题进行求解, 与指针网络模型不同的是, 该类方法采用图神经网络对每个节点的特征进行学习, 从而根据学习到的节点特征进行后续的链路预测、节点预测等任务, 其原理详见第 2 节. Dai 等[37] 首次结合图神经网络和深度强化学习方法对 MVC、TSP 等组合优化问题进行了研究, 作者利用图神经网络对各个 "待选节点"的 Q 值进行估计, 每次根据 Q 值利用贪婪策略向当前解插入一个新节点, 直到构造一个完整的解. 文献 [38–41] 使用了不同的图神经网络以及不同的解构造方法对组合优化问题进行求解, 在二次分配问题、最大覆盖问题、MVC 等组合优化问题上取得了较好的效果. 由于图神经网络主要进行节点特征的提取, 部分研究[34–35] 结合图神经网络和指针网络进行组合优化算法的设计, 即首先使用图神经网络进行节点特征计算, 再使用指针网络的 Attention 机制进行解的构造, 在 TSP 等问题上取得了较好的优化性能.

以上方法均为 "端到端 (End-to-end) 方法", 即给定问题实例作为输入, 利用训练好的深度神经网络直接输出问题的解, 其中神经网络的参数一般利用深度强化学习方法训练得到. 相对于传统的迭代型优化算法, 该类端到端方法无需搜索直接输出问题解, 具有求解速度快的优势; 且模型一旦训练完成, 可以对具有相同分布特性的所有问题实例进行求解, 而不需要重新进行训练, 模型具有很强的泛化能力, 而传统算法一旦遇到一个新的问题实例, 则需要从头开始重新进行搜索求解. 因此该类方法为求解组合优化问题提供了一种全新的思路, 具有求解速度快、泛化能力强的优势.

但是由于端到端方法通过神经网络直接输出最终解, 解的最优性很难保证, 上述方法在小规模问题上可以接近最优解, 但是在中大规模问题上与 LKH3[42]、Google OR tools[43]、Gurobi[44]、Concorde[45] 等专业组合优化求解器的优化能力还存在一定差距.

鉴于此, 基于 DRL 求解组合优化问题的另外一类研究是利用 DRL 方法改进传统的精确/近似方法, 如利用机器学习模型对分支定界法 (Branch and bound) 的节点选择和变量选择策略进行优化, Bengio 等[46] 已经对该类方法进行了详细的综述研究, 本文不再赘述. 除了对精确算法进行改进, 近年来兴起的另一类方法是基于深度强化学习对迭代搜索类算法进行改进, 局部搜索/邻域搜索是求解组合优化问题的常用近似方法, 在局部搜索过程中, 学者们通常手工设计各种启发式规则对解进行构造和搜索, 但是随着人工智能技术的发展, 通过神经网络模型代替手工规则设计是未来的发展趋势. 鉴于此, 近几年部分学者研究采用深度强化学习对解搜索的启发式规则进行学习和选择[47–50], 通过学习到的规则进行解的迭代搜索, 根据该思路, 文献 [47, 50] 所提方法在优化效果上达到甚至超过了 LKH3、Google OR tools 等专业组合优化求解器, 文献 [50] 在求解速度上也超越了 LKH3、Google OR tools 等方法.

基于深度强化学习改进的局部搜索方法具有较好的优化效果, 但其本质上仍然是迭代型搜索算法, 求解速度仍然远不及端到端方法; 端到端方法具有求解速度快、泛化能力强的优势, 但是该类方法的缺陷是解的优化效果无法保证, 与传统组合优化方法的优化效果仍然存在一定差距. 目前该两类方法各具优劣, 鉴于深度强化学习近年来在解决组合优化问题上的突出成果, 本文主要总结回顾近些年基于深度强化学习去解决组合优化问题的相关理论方法和应用研究.

2 基于深度强化学习的组合优化: 基本原理介绍

Pointer network 模型和图神经网络模型是基于深度强化学习的组合优化方法中常用的两种模

型, 本节首先对如何利用上述模型求解组合优化问题的基本原理进行介绍, 并对广泛用于训练 Pointer network 模型的 REINFORCE 强化学习算法进行介绍.

2.1 Pointer Network 求解组合优化问题

Pointer network 方法可概括为利用神经网络模型实现序列到序列的映射, 其核心思想是利用编码器 (Encoder) 对组合优化问题的输入序列进行编码得到特征向量, 再利用解码器 (Decoder) 结合 Attention 计算方法以自回归 (Autoregressive) 的方式逐步构造解, 自回归即每次选择一个节点, 并在已选择节点的基础上选择下一个节点, 直到构造得到完整解.

本节以经典指针网络模型[31] 求解 TSP 问题为例, 对该方法的原理进行介绍. 经典指针网络模型的编码器和解码器均为 LSTM (Long short-term memory) 循环神经网络. 利用指针网络模型构造 TSP 解的过程如下:

首先将每个城市的二维坐标转换成高维的节点表征向量 $\boldsymbol{s}_i$, 编码器的 LSTM 依次读入各个城市的 $\boldsymbol{s}_i$, 最终编码得到一个存储输入序列信息的向量 Vector, 同时 LSTM 在计算的过程中可以得到每个城市的隐层状态 $\boldsymbol{e}_i$, 其过程如图 1 所示.

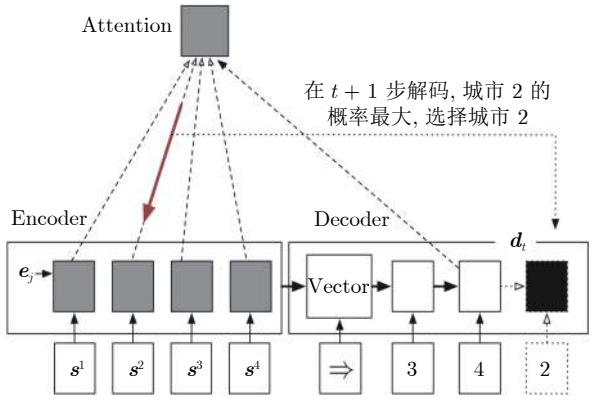


图 1 Pointer network 模型示意图

Fig. 1 Schematic diagram of pointer network model

然后解码器对 Vector 进行解码, 过程如图 1 所示. 在第一步解码过程中, 即 $t=0$ 时, LSTM 读入 Vector, 输出当前的隐层状态 $\boldsymbol{d}_0$, 利用 Attention 机制根据 $\boldsymbol{d}_0$ 和编码器得到的各城市的隐层状态 $\boldsymbol{e}$ 计算选择各个城市的概率, 计算公式见式 (2), 此时可选择概率最大的节点作为第一步选择的城市; 在接下来的解码过程中, 即 $t=1,2,\cdots$ 时, LSTM读入上一步 LSTM 的隐层输出和上一步选择城市的特征向量, 输出当前的隐层状态 $\boldsymbol{d}_t$, 根据 $\boldsymbol{d}_t$ 和各城市的 $\boldsymbol{e}$ 计算选择各个城市的概率, 其计算公式如下, 即 Attention 机制:

$$u_j^t = \boldsymbol{v}^{\mathrm{T}} \tanh\left(\boldsymbol{W}_1 \boldsymbol{e}_j + \boldsymbol{W}_2 \boldsymbol{d}_t\right), j \in (1, \cdots, n)$$

$$P\left(\rho_{t+1} \mid \rho_1, \cdots, \rho_t, X_t\right) = \mathrm{softmax}\left(\boldsymbol{u}^t\right) \tag{2}$$

即利用当前的 $\boldsymbol{d}_t$ 值和每个城市 $\boldsymbol{e}$ 值计算得到第 t 步选择各城市的概率, 其中 $\boldsymbol{W}$ 和 $\boldsymbol{v}$ 均为神经网络的参数. 在每一步解码过程中, 对于每个城市 j, 均可以根据式 (2) 计算得到其 u_j^t 值, u_j^t 代表在第 t 步解码过程中选择城市 j 的概率, 此时可以选择具有最大概率值的节点添加到解当中, 按照该方式不断选择城市, 直至构造得到一个完整解.

因此该深度神经网络模型的输入是城市的坐标序列, 输出是城市的顺序, 通过对该模型参数的训练可以实现问题序列到解序列的准确映射.

2.2 Pointer Network 的强化学习训练方法

对于 Pointer network 深度神经网络模型, 可以通过监督式训练算法或者强化学习算法进行训练, 由于监督式学习方法需要提供大量最优路径的标签数据集, 实际应用较为困难, 因此目前研究中通常以强化学习算法对模型的 $\boldsymbol{W}$ 和 $\boldsymbol{v}$ 等参数进行训练.

强化学习通过试错机制不断训练得到最优策略, 首先需要将组合优化问题建模为马尔科夫过程, 其核心要素为状态、动作以及反馈, 以 TSP 问题为例, 其问题的状态为城市的坐标 s 以及已经访问过的城市, 动作为第 t 步选择的城市 π_t, 所有动作组成的城市访问顺序 π 即为组合优化问题的解, 反馈 r 是路径总距离的负数, 即最小化路径长度, 策略即为状态 s 到动作 π 的映射, 策略通常为随机策略, 即得到的是选择城市的概率 $p_\theta(\pi|s)$, 该随机策略建模为:

$$p_\theta(\pi \mid s) = \prod_{n}^{t=1} p_\theta\left(\pi_t \mid s, \pi_{1:t-1}\right) \tag{3}$$

策略由神经网络参数 θ 进行参数化, 在马尔科夫过程中, 每一步动作的概率为 $p_\theta\left(\pi_t|s, \pi_{1:t-1}\right)$, 即根据已访问过的城市 $\pi_{1:t-1}$ 和城市坐标 s 计算选择下一步访问各个城市的概率, 根据链式法则累乘即可以得到城市坐标 s 到城市最终访问顺序 π 的映射 $p_\theta(\pi|s)$, 该随机策略可以建模为上节所述的指针网络模型, 其参数为 θ.

由于 TSP 问题的优化目标是最小化总的路径长度 $L(\pi)$, 而 REINFORCE 也是以总反馈作为参数更新的标准, 因此现有的研究中通常采用 REINFOR-

CE 强化学习算法对策略的参数 θ 进行训练优化.

REINFORCE 强化学习算法又称基于蒙特卡洛的策略梯度方法, 即不断执行动作直到结束, 在一个回合结束之后计算总反馈, 然后根据总反馈对策略的参数进行更新. 以 TSP 问题为例, 总反馈即为路径总长度的负数 $-L(\pi)$. 可见 REINFORCE 算法天然适用于训练该类问题. REINFORCE 基于以下公式对策略 θ 进行更新:

$$\begin{aligned}&\nabla\mathcal{L}(\theta\mid s)=\\&\mathrm{E}_{\mathrm{p}_\theta(\pi|s)}\left[(L(\pi)-b(s))\nabla\ln p_\theta(\pi\mid s)\right],\\&\theta\leftarrow\theta+\nabla\mathcal{L}(\theta\mid s)\end{aligned}\tag{4}$$

根据链式法则, $p_\theta(\pi|s)$ 为每步动作选择概率的累乘, 则 $\ln p_\theta(\pi\mid s)$ 计算为每步动作选择概率对数的求和, 以该值对参数 θ 计算偏导可得梯度值 $\nabla\ln p_\theta(\pi\mid s)$, $(L(\pi)-b(s))$ 决定了梯度下降的方向, $b(s)$ 代表策略的平均表现 (Baseline), 如果当前策略的表现比 “平均”好, 则对该策略进行正向激励, 反之亦然. 有多种方式对 $b(s)$ 进行估计, 运用较多的方法是新增一个 Critic 神经网络计算 $b(s)$, 即给定一个 TSP 问题 s, 利用 Critic 神经网络估计该问题解的路径长度. Critic 网络与策略网络同步进行训练, 以策略网络训练过程中产生的 $(s, L(\pi))$ 作为训练集对 Critic 进行训练.

REINFORCE 算法通过式 (4) 对 θ 的梯度进行计算并更新, 不断训练从而得到准确的 $p_\theta(\pi|s)$, 即实现组合优化问题序列到解序列的准确映射.

2.3　图神经网络方法求解组合优化问题

图神经网络 (Graph neural network, GNN) 是近年来提出的能够有效处理图结构数据的新方法, 因此部分学者研究如何利用图神经网络对组合优化问题进行建模, 其核心思想是根据每个节点的原始信息 (如城市坐标) 和各个节点之间的关系 (如城市之间的距离), 利用图神经网络方法计算得到各个节点的特征向量, 根据各个节点的特征向量进行节点预测、边预测等任务.

一般将图定义为 $G=(V,E)$, V 代表节点的集合, E 为边的集合. 图神经网络通过不断学习节点的特征、邻居节点的特征、边的特征, 并将其以各种方法进行聚合, 从而最终得到各个节点的特征向量, 根据各个节点的特征向量完成预测、分类等任务. 以经典 GNN[51] 为例, 各个节点的表征以如下公式更新:

$$\boldsymbol{h}_v^{(t)}=\sum_{u\in N(v)}f\left(\boldsymbol{x}_v,\boldsymbol{x}_{(v,u)}^{\boldsymbol{e}},\boldsymbol{x}_u,\boldsymbol{h}_u^{(t-1)}\right)\tag{5}$$

其中 $\boldsymbol{h}_v^{(t)}$ 代表节点 v 的表征向量, $N(v)$ 代表 v 的邻居节点的集合, $\boldsymbol{x}_v$ 是节点 v 的特征, $\boldsymbol{x}_{(v,u)}^{\boldsymbol{e}}$ 是与 v 相连的边的特征, $\boldsymbol{x}_u$ 是邻居节点 u 的特征, $\boldsymbol{h}_u^{(t-1)}$ 是邻居节点 u 在上一步更新的特征向量. 因此该公式根据节点 v 本身的特征、边的特征以及邻居节点的特征对节点 v 的表征向量进行更新, 从 $t=0$ 开始对不断对 $\boldsymbol{h}_v^{(t)}$ 进行更新直到收敛, 从而得到节点 v 的准确特征向量.

根据各个节点的特征向量, 可以进行组合优化问题的求解, 如针对节点选择问题 (如最小顶点覆盖问题), 可以将图神经网络得到的节点特征向量 $\boldsymbol{h}_v^{(t)}$ 以一个全连接层神经网络映射到节点选择概率, 从而根据概率进行节点的选择; 针对边选择问题 (如 TSP 问题), 可以以两个节点的特征向量作为输入, 以一个全连接层神经网络映射得到一个选择概率, 即该两点之间存在边的概率, 从而进行边选择, 值得注意的是, 按照概率进边的选择并不一定可以构成一个完整的哈密顿回路, 因此需要辅以搜索方法进行解的构造.

以文献 [37−38] 所用的方法为例, 模型首先利用 GNN 计算得到各个节点的表征 $\boldsymbol{h}_v^{(t)}$, 将各个节点的 $\boldsymbol{h}_v^{(t)}$ 向量进一步运算得到各个节点的 Q 值. 根据 Q 值以迭代的方式构造解, 即每次选择 Q 值最大的节点添加到当前解当中, 直到构造得到完整解, 通常以 DQN 强化学习方法对该图神经网络进行训练, 从而得到准确的 Q 值估计. 部分文献 [39, 41] 根据 GNN 得到的节点特征向量计算节点/边选择概率, 随后以波束搜索、树搜索的方式根据选择概率进行可行解的构造.

由于图神经网络主要用于计算节点的特征向量, 因此部分学者[34−35] 将图神经网络和指针网络进行结合, 即用图神经网络计算得到的节点特征向量 $\boldsymbol{h}_v^{(t)}$ 代替指针网络 LSTM 编码器计算得到的各节点的隐层输出向量 $\boldsymbol{e}$, 仍然采用式 (2) 的 Attention 机制计算每一步的节点选择概率, 以自回归的方式逐步构造得到完整解.

3　基于深度强化学习的组合优化: 理论与方法

目前基于 DRL 的组合优化方法主要分为基于 DRL 的端到端算法和基于 DRL 的局部搜索改进算法两大类, 其中端到端算法主要包括基于 Pointer network 的端到端方法和基于图神经网络的端到端方法两类, 本文主要对以上方法在近年来的研究进展进行综述介绍, 对各类方法代表性算法的原理、优化性能、优缺点进行对比和介绍, 对各类方法未

来的研究方向进行了分析, 各个算法的总结如表 1.

3.1 基于 Pointer Network 端到端方法

3.1.1 方法综述

Vinyals 等[30] 最早在 2015 年提出了 Pointer network 模型进行组合优化问题求解, 该文章也开启了利用深度神经网络进行组合优化问题求解的一系列研究, 该模型借鉴机器翻译领域中的 Seq2Seq 模型求解组合优化问题, 即利用基于深度神经网络的编码器将组合优化问题的输入序列 (如城市坐标) 进行编码, 然后通过解码器和注意力计算机制 (Attention) 计算得到各节点的选择概率, 并以自回归的方式逐步选择节点, 直到得到完整解 (如城市访问的顺序), 该方法的详细原理见第 2 节. 作者采用监督式学习的方法对该模型进行训练, 即利用大量 "TSP 城市坐标−最优路径"的样本对该 Pointer network 的参数进行离线训练, 利用该训练好的模型, 可以求解与训练集具有相同分布特性的任意 TSP 问题. 相对于传统的局部搜索或者启发式搜索方法, 该端到端模型不需要进行迭代搜索, 具有泛化能力强、求解速度快的优势, 论文实现了对至多 50 个城市的小规模 TSP 问题的快速求解.

由于 Vinyals 等[30] 提出的方法采用监督式方式进行训练, 导致其得到的解的质量永远不会超过样本的解的质量, 并且事先构造训练样本需要耗费大量时间, 因此限制了其应用于更大规模的组合优化问题. 鉴于此, Bello 等[31] 采用强化学习方法训练 Pointer network 模型, 他们将每个问题实例视为一个训练样本, 以问题的目标函数作为反馈信号, 采用 REINFORCE 强化学习算法进行训练, 并引入 Critic 网络作为 Baseline 以降低训练方差. 论文对至多 100 个城市的 TSP 问题以及 200 物品的 0−1 背包问题对该模型进行了测试, 结果发现该模型在

表 1 现有算法模型、训练方法、求解问题、以及优化效果比较

Table 1 Comparison of model, training method, solving problems and performance with existing algorithms

方法类别	研究	模型以及训练方法	求解问题及优化效果
基于 Pointer network 的端到端方法	2015 年 Vinyals 等[30]	Ptr-Net + 监督式训练	30 TSP问题: 接近最优解, 优于启发式算法. 40, 50-TSP: 与最优解存在一定差距. 凸包问题、三角剖分问题.
	2017 年 Bello 等[31]	Ptr-Net + REINFORCE & Critic baseline	50-TSP: 优于文献 [30]. 100-TSP: 接近 Concorde 最优解. 200-Knapsack: 达到最优解.
	2018 年 Nazari 等[32]	Ptr-Net + REINFORCE & Critic baseline	100-TSP: 与文献 [31]优化效果相近, 训练时间降低约60 %. 100-CVRP/随机CVRP: 优于多个启发式算法.
	2018 年 Deudon 等[33]	Transformer attention + REINFORCE & Critic baseline	20, 50-TSP: 优于文献 [31]. 100-TSP: 与文献 [31]优化效果相近.
	2019 年 Kool 等[34]	Transformer attention + REINFORCE & Rollout baseline	100-TSP: 优于文献 [30−33, 37, 40]. 100-CVRP、100-SDVRP、100-OP、100-PCTSP、SPCTSP: 接近 Gurobi 最优解, 优于多种启发式方法.
	2020 年 Ma 等[35]	Graph pointer network + HRL	20, 50-TSP: 优于文献 [31, 37], 劣于文献 [34]. 250, 500, 1 000-TSP: 优于文献 [31, 34]. 20-TSPTW: 优于OR-Tools、蚁群算法.
	2021 年 Li 等[36]	Ptr-Net + REINFORCE & Critic baseline & 分解策略/参数迁移	40, 100, 150, 200, 500-两目标/三目标TSP : 优于 MOEA/D、NSGA-II、MOGLS.
基于图神经网络的端到端方法	2017 年 Dai 等[37]	structure2vec + DQN	1 200-TSP: 接近文献 [31]. 1 200-MVC (最小顶点覆盖): 接近最优解. 1 200-MAXCUT (最大割集): 接近最优解.
	2020 年 Manchanda 等[38]	GCN + DQN	2k至20k-MCP (最大覆盖问题): 优于文献 [37]. 10k, 20k, 50k-MVC: 优于文献 [37].
	2018 年 Li 等[39]	GCN + 监督式训练 + 引导树搜索	实际数据集 MVC、MIS (最大独立点集)、MC (极大团)、Satisfiability (适定性问题): 优于文献 [37].
	2017 年 Nowak 等[40]	GNN + 监督式训练 + 波束搜索	20-TSP: 劣于文献 [30].
	2019 年 Joshi 等[41]	GCN + 监督式训练 + 波束搜索	20, 50, 100-TSP: 略微优于文献 [30−31, 33−34], 优于文献 [37].
深度强化学习改进的局部搜索方法	2019 年 Chen 等[47]	Ptr-Net + Actor-critic	20-CVRP: 达到最优解. 50, 100-CVRP: 优于文献 [32, 34]、OR-Tools. 作业车间调度: 优于OR-Tools、DeepRM.
	2019 年 Yolcu 等[48]	GNN + REINFORCE	实际数据集 Satisfiability、MIS、MVC、MC、图着色问题: 更少搜索步数得到最优解、但单步运行时间长于传统算法.
	2020 年 Gao 等[49]	Graph attention + PPO	100-CVPR: 优于文献 [34]. 100-CVPRTW: 优于多个启发式方法. 400-CVRPTW: 劣于单个启发式方法, 优于其他.
	2020 年 Lu 等[50]	Transformer attention + REINFORCE	20, 50, 100-CVRP: 优于文献 [32, 34, 47], 以及优于 OR Tools、LKH3. 且运行时间远低于 LKH3.

50 城市 TSP 问题上超越了 Vinyals 等监督式训练得到的模型, 并可以在 100 城市的 TSP 问题上接近最优解, 在背包问题上达到了最优解.

进一步的, Nazari 等[32] 将 Pointer network 拓展至具有动态特性的 VRP 问题, 作者将输入分为两部分, 包括静态输入 (顾客位置/坐标) 和动态输入 (顾客需求), 由于考虑到在输入端改变顾客的顺序不会影响问题的求解, 因此作者将 Encoder 输入层的 LSTM 替换成简单的一维卷积层, 从而可以有效降低计算成本. 在仍然采用 REINFORCE 强化学习算法进行训练的情况下, 他们的模型将训练时间降低了 60 %, 在 TSP 问题上与 Bello 等的模型[31] 取得了几乎相同的优化效果, 并且在 VRP、随机 VRP 问题上取得了比 Clarke-Wright savings 和 Sweep 经典启发式搜索算法更好的优化效果.

相对于传统的 Seq2Seq 模型, 近年来 Transformer 模型[52] 在自然语言处理领域取得了巨大的成功, Transformer 的 Multi-head attention 机制可以使模型更好地提取问题的深层特征, 鉴于此, 多个最新的研究借鉴了 Transformer 模型进行了组合优化问题求解的研究.

Deudon 等[33] 借鉴 Transformer 模型改进了传统的指针网络模型, 其编码层采用了与 Transformer 模型编码层相同的结构, 即利用 Multi-head attention 方法计算得到节点的特征向量; 其解码层没有采用 LSTM, 而是将最近三步的决策进行线性映射得到参考向量, 从而降低模型复杂度, 其 Attention 计算方式与传统 Pointer network 模型相同, 仍然采用经典的 REINFORCE 方法对该模型进行训练, 文章仅对 TSP 问题进行了求解, 作者首先利用该训练好的神经网络模型输出初始解, 随后在该初始解的基础上进行一个简单的 2OPT 局部搜索, 结果发现这种方式可以有效提高解的质量.

Kool 等[34] 在 2019 年指出, 虽然 Deudon 等[33] 的模型结合局部搜索可以提高性能, 但是其神经网络模型本身与传统的 Pointer network 模型相比并没有显著的优势, 鉴于此, Kool 等借鉴 Transformer 模型, 提出了一个可以利用 Attention 机制求解多种组合优化问题的新方法[34], 在 TSP、Capacitated VRP (CVRP)、OP (Orienteering problem)、PCTSP (The prize collecting TSP) 等问题上性能超越了前述介绍的所有 Pointer network 模型[30−33], 并且高度接近 Concorde、LKH3、Gurobi 等专业求解器得到的最优解. 该方法的改进主要包括两方面: 1) 与文献 [33] 相同, 该模型的编码层采用了和 Transformer 模型相同的 Multi-head attention 机制, 但解码层和 Attention 机制存在很大不同, 首先该模型每一步的解码过程中考虑的是第一步所做的决策和最近两步的决策, Deudon 等[33] 模型 Attention 的计算方式仍然和经典指针网络相同, 而该模型采用了 Transformer 模型的 Self-attention 计算方法, 增加了更多计算层以提高模型的表现; 2) 另外, 文章对强化学习训练算法进行了改进, 前述所有文章均采用 REINFORCE 算法结合 Critic-baseline 的方式进行训练, 即增加一个 Critic 神经网络来估计 $b(s)$. 作者指出同时训练两个神经网络是低效的, 而且 Critic 很难得到 $b(s)$ 的准确估计, 因此文章设计了一种 Rollout baseline 来代替 Critic 神经网络: 即在之前训练过程中得到的所有策略模型里, 选择在测试集中表现最好的模型作为基线策略, 并采用贪婪方式进行动作选择, 将利用该基线策略对状态 s 求解得到的目标函数值作为 $b(s)$, 如果当前策略比历史最优策略的表现好, 则进行正向激励, 从而对当前策略进行评价和参数更新. 实验证明该训练方法的收敛能力明显优于传统方法. 经过以上改进, 该方法的优化性能超越了之前所有的端到端模型.

进一步的, Ma 等[35] 结合指针网络和图神经网络设计了一种图指针网络 (Graph pointer network, GPN) 用来求解大规模 TSP 问题以及带时间窗约束的 TSP 问题. 该模型的编码器包含两部分: Point encoder 以及 Graph encoder, Point encoder 对城市坐标进行线性映射, 并输入到 LSTM中得到每个城市的点嵌入, Graph encoder 通过 GNN 图神经网络对所有城市进行编码, 得到每个城市的图嵌入. 模型根据图嵌入和点嵌入, 基于 Attention 机制计算每一步城市选择的概率, 并引入 Vector context 提高模型的泛化能力. 文章采用分层强化学习方法 (Hierarchical RL, HRL) 对模型进行训练. 在 50-TSP 问题上训练得到的模型可以有效求解 250, 500, 750, 1 000 等规模的 TSP 问题, 在大规模 TSP 问题上超越了 Kool 等[34] 的方法, 但是在 100 以内规模的 TSP 问题上仍然劣于文献 [34]. 文章并对带时间窗约束的 TSP 问题进行了实验, 性能超越了 OR-tool 以及蚁群算法, 证明了分层强化学习训练方法在处理约束问题上的有效性.

3.1.2 总结

以上为按时间线对各个代表性方法的介绍, Vinyals 等[30] 最早提出了求解组合优化问题的 Pointer network 模型, Bello 等[31] 最先提出采用强化学习方法对该模型进行训练. 目前 Kool 等[34]的方法在 100 规模以下的 TSP 问题上取得了当前业界最

优 (State-of-the-art, SOA) 的优化效果, 超越了其他基于 Ptr-Net 模型的方法[30−33]. Ma 等[35]的方法在小规模 TSP 问题上劣于[34], 但是在大规模 TSP 问题 (250, 500, 1 000) 上超越了文献 [34], 各方法详细的性能对比详见表 1. 值得注意的是, 上述方法在 50 规模以上的 TSP 问题上均未达到 Concorde、LKH3 等求解器得到的最优解.

3.2 基于图神经网络的端到端方法

3.2.1 方法综述

Dai 等[37] 在 2017 年首先研究了如何采用图神经网络对组合优化问题进行求解, 作者采用 structure2vec 图神经网络对当前解的图结构进行建模, 并根据图神经网络计算剩余可选节点中各个节点的 Q 值, 随后基于贪婪策略根据 Q 值选择一个新的节点添加到当前解中, 直至得到完整解. 作者采用了深度 Q 学习 (Deep Q-learning, DQN) 算法对该图神经网络的参数进行训练, 以使模型输出准确的 Q 值估计. 文章首先在 50~100 节点的 MVC、MAXCUT、TSP 问题上对该模型进行了训练, 将训练好的模型在多达 1 200 个节点的上述问题上进行了测试, 以 CPLEX 求得的解作为最优解对模型的优化能力和求解速度进行了研究, 实验结果表明该方法在 TSP 问题上取得了接近 Bello 等[31] 方法的效果, 在 MVC、MAXCUT 问题上得到了接近最优解的优化效果, 且超越了多个基准算法.

Mittal 等[38] 采用了与 Dai 等[37] 相同的模型架构对大型组合优化问题进行求解, 即结合图神经网络、DQN 以及贪婪策略进行解的构造, 作者采用了图卷积神经网络 (Graph convolutional networks, GCN) 对图结构进行建模, 在 20k 规模的最大覆盖问题 (MCP)、50k 规模的 MVC 问题上进行了模型测试, 实验发现该模型在大规模问题上的表现比 Dai 等[37] 的模型获得了 41 % 的优化能力的提升.

Li 等[39] 采用图神经网络对最小顶点覆盖问题、最大独立点集 (Maximal independent set, MIS)、极大团 (Maximal clique, MC)、适定性问题 (Satisfiability) 进行了研究, 由于所研究问题均为点选择问题, 与 TSP 问题不同, 对节点选择的顺序无要求, 因此文章没有采用逐步添加节点的方式构造解, 而是使用 GCN 图神经网络直接输出所有点选择概率的估计值, 并基于该估计值以引导树搜索的方式构造可行解. 为了解决问题可能存在多个最优解的情况, 文章采用 Hindsight loss 方式输出多个概率分布, 在此基础上进行树搜索, 并采用局部搜索的方式对解进行再处理. 文章与 Dai 等[37] 的模型以及测试问题的多个基准方法进行了对比, 在优化效果上均超越了对比算法.

以上方法均为对选择各个节点的概率进行估计, 文献 [40−41] 利用图神经网络对选择各个 “边” 的概率进行估计, 以 TSP 问题为例, 利用图神经网络模型输出一个邻接矩阵, d_{ij} 代表两点之间存在边的概率, d_{ij} 值大则节点 i 和 j 大概率相连. 随后根据各个边出现概率的估计值, 使用波束搜索 (Beam search) 的方式构造最终的可行解. 文献 [40−41] 均采用监督式方法进行训练, 即利用 LKH3 或 Concorde 求解器构造大量 “坐标−最优路径” 的训练数据, 根据最优解的真实邻接矩阵和图神经网络输出的邻接矩阵计算交叉熵, 以交叉熵为损失函数训练模型. Nowak 等[40] 使用的是经典 GNN 图神经网络模型[51], 该模型的优化效果没有超越传统的启发式方法以及指针网络模型, Joshi 等[41] 采用的是 GCN 图神经网络, 该模型在 20, 50, 100 规模 TSP问题上的优化效果略微超越了 Kool 等[34] 的方法, 接近 LKH3、Concorde 等求解器得到的最优解, 但是该方法的求解时间长于 LKH3、Concorde 等方法, 在泛化能力上该方法也不及 Kool 等[34] 的方法.

3.2.2 总结

指针网络模型主要用于求解 TSP、VRP 等具有序列特性的组合优化问题 (即该类问题的解与节点的顺序有关), 由于指针网络利用 Attention 机制以自回归的方式对解进行构造, 因此适用于求解序列组合优化问题. 而基于图神经网络的方法由于得到的是节点的特征向量, 自然地可以计算得到节点选择的概率, 因此在 MVC、MIS 等顺序无关的点选择问题上多有应用, 针对 TSP 等序列优化问题, 一类方法是仍然以自回归的方式逐步选择节点[37], 另一类方式是根据节点的特征向量计算边选择的概率, 然后利用波束搜索等方法构造解[41].

由于 TSP 问题是文献中研究组合优化问题的经典算例, 表 2 对上述端到端模型在不同规模 TSP 问题上的优化性能进行了实验对比, 结果取自各个文献中的实验数据, 各个模型采用 TensorFlow 或者 Pytorch 深度学习工具平台实现, 由于文献 [34, 41] 是各类方法的 SOA 模型且均在相同型号的 1080Ti-GPU 上进行的实验, 因此对文献 [34, 41] 的求解时间也进行了对比.

其中 Greedy 是采用贪婪策略构建 TSP 问题的解, 即每次选取具有最大选择概率的城市; 2OPT 是对模型得到的解进行进一步的 2OPT 局部搜索以提高解的质量; BS 是采用 Beam search 波束搜索的方式根据边选择的概率构造解. 通过实验结果

表 2　端到端模型在 TSP 问题上优化性能比较
Table 2　Comparison of end-to-end model on TSP

方法类别	模型	TSP-20	TSP-50	TSP-100
最优	Concorde	3.84	5.70	7.76
基于指针网络 (Attention 机制)	Vinyals 等[30]	3.88	7.66	—
	Bello 等[31]	3.89	5.95	8.30
	Nazari 等[32]	3.97	6.08	8.44
	Deudon 等[33]	3.86	5.81	8.85
	Deudon 等[33] + 2OPT	3.85	5.85	8.17
	Kool 等[34] (Greedy)	**3.85 (0 s)**	**5.80 (2 s)**	**8.12 (6 s)**
	Kool 等[34] (Sampling)	**3.84 (5 m)**	**5.73 (24 m)**	**7.94 (1 h)**
	Dai 等[37]	3.89	5.99	8.31
基于图神经网络	Nowak 等[40]	3.93	—	—
	Joshi 等[41] (Greedy)	3.86 (6 s)	5.87 (55 s)	8.41 (6 m)
	Joshi 等[41] (BS)	**3.84 (12 m)**	**5.70 (18 m)**	**7.87 (40 m)**

可见 Kool 等[34] 的方法是当前基于 Attention 机制的端到端方法的 SOA 模型, 采用简单的贪婪策略即可在短时间内实现对 TSP 问题的高效求解; Joshi 等[41] 利用图神经网络结合波束搜索对 TSP 问题进行求解, 其优化效果超越了 Kool 等[34] 的模型, 但是该方法耗时过长.

由于图神经网络能够有效处理很多组合优化问题的图结构, 近年来利用图神经网络求解组合优化问题的研究呈上升趋势, 但该类方法仍然有很多问题待解决, 例如波束搜索通常需要大量搜索时间, 并且大多研究仍然采用监督式方式进行训练, 需要构造大量标签样本, 实际应用困难. Ma 等[35] 将图卷积神经网络和指针网络相结合, 但是该方法在 100 规模的 TSP 问题上仍然劣于 Kool 等[34] 的方法, 但是在大规模 TSP 上存在优势, 未来如何将指针网络的 Attention 机制和图神经网络相结合是一个重要的研究点.

3.3　深度强化学习改进的局部搜索方法

虽然端到端方法可以通过深度神经网络模型直接输出问题的解, 实现组合优化的快速求解, 但是其优化效果与 LKH3、Google OR tools 等专业求解器相比仍有一定差距. 局部搜索 (Local search) 是求解组合优化问题的经典方法, 当前的局部搜索算法主要是通过人工对搜索的启发式规则进行设计, 以获得更好的优化效果, 鉴于近年来深度强化学习在在各领域取得的瞩目的学习能力, 学者们开始研究利用深度强化学习方法来自动学习局部搜索算法的启发式规则, 从而比人工设计的搜索规则具有更好的搜索能力.

3.3.1　方法综述

Chen 等[47] 于 2019 年提出了一个基于深度强化学习的组合优化问题搜索模型 NeuRewriter, 它和局部搜索具有相似的算法流程, 即首先随机构造一个可行解, 在该初始解的基础上通过局部搜索不断提高解的质量. 相比于传统算法所采用的人工设计的启发式规则, 作者利用深度强化学习方法对局部搜索的策略进行训练, 利用学习到的策略对搜索过程进行引导. 其策略由两部分构成: Region-picker 和 Rule-picker, 以作业车间调度问题为例, 首先利用 Region-picker 选定一个工序, 其次利用 Rule-picker 对该工序的操作策略进行决策, 如与另一个工序进行调换. 文章利用 Actor-critic 方法对 Region-picker 和 Rule-picker 策略进行了训练, 其优化效果在作业车间调度问题上超越了 DeepRM 和 Google OR-tools 求解器, 在 VRP 问题上超越了 Google OR-tools 求解器.

Yolcu 等[48] 采用深度强化学习改进的局部搜索方法对适定性问题 (Satisfiability) 进行了研究, 仍然采用局部搜索的求解框架, 利用深度强化学习对局部搜索中变量选择算子进行学习, 作者采用图神经网络对变量选择的策略进行参数化, 利用 REINFORCE 算法更新图神经网络的参数, 实验显示相对于传统的启发式算法, 该方法可以在更少的步数内找到最优解, 但是运行时间却远长于传统算法.

Gao 等[49] 基于大规模邻域搜索框架对组合优化问题进行求解, 作者利用深度强化学习方法对大规模邻域搜索的 Destroy 和 Repair 算子进行学习, 采用图注意力神经网络 (Graph attention network) 对问题特征进行编码, 并采用基于循环神经网络的解码器输出 Destroy 和 Repair 算子. 具体的, Destroy 算子是从当前解中选择多个节点, 并将其从当前解中移除, Repair 算子是将移除的节点以一定的顺序重新插入到当前解中, 因此该模型对 Destroy 算子的点集选择策略和 Repair 算子的排序策略进行学习. 文章采用 Proximal policy optimization (PPO) 算法对模型进行训练, 并用来解决 CVRP、带时间窗的 CVRP 等问题, 实验表明该方法在 100 节点的 CVRP 问题上优化效果超越了 Kool 等[34] 的方法, 并在 400 节点的大规模 CVRP 问题上具有比传统启发式算法更快的收敛性能, 在优化能力上接近但未达到最优解, 但本文并没有提供求解时间对比.

Lu 等[50] 于 2020 年提出了一种 Learn to improve (LSI) 组合优化问题求解方法, 该方法不只在优化效果上超越了 LKH3、Google OR-tools 等组合

优化求解器, 其求解速度也超越了上述专业求解器. 作者首先对 LSI 框架进行设计, 算法总体流程仍然采用局部搜索的方式, 在每一步搜索过程中, 算法决定是继续提升当前解还是对当前解进行扰动, 因此算法包括两个算子: 提升算子和扰动算子, 作者采用了 9 种不同的提升算子作为算子库, 采用深度强化学习训练提升算子的选择策略, 每次迭代, 算法根据问题特征和当前的解, 利用学习到的策略从算子库中选择提升算子, 从而不断提升当前解的质量, 如果达到局部最优, 算法对当前解进行扰动. 论文通过实验证明该方法在 20-, 50-, 100-CVRP 问题上超越了当前 State-of-the-art 的 LKH3 求解器, 并且其求解速度也远超 LKH3 算法.

3.3.2 总结

深度强化学习改进的局部搜索方法是自 2019 年以来最新提出的一类组合优化方法, 主要用于求解 VRP 等路径优化问题, 表 3 对该类方法以及端到端模型在求解不同规模 VRP 问题上的优化能力进行了比较, 结果取自各个文献的实验数据. 各算法均在 GPU 上运行 (各算法使用不同型号 GPU, 但运算时间不存在较大差距), 均采用 Pytorch 深度学习工具实现.

表 3 多个模型在 VRP 问题上优化性能比较

Table 3 Comparison of models on VRP

模型	VRP-20	VRP-50	VRP-100
LKH3	**6.14 (2 h)**	**10.38 (7 h)**	**15.65 (13 h)**
Nazari 等[32]	6.40	11.15	16.96
Kool 等[34] (Greedy)	6.40 (1 s)	10.98 (3 s)	16.80 (8 s)
Kool 等[34] (Sampling)	6.25 (6 m)	10.62 (28 m)	16.23 (2 h)
Chen 等[47]	6.12	10.51	16.10
Lu 等[50]	**6.12 (12 m)**	**10.35 (17 m)**	**15.57 (24 m)**

从实验对比可以看出, 深度强化学习改进的局部搜索方法在优化能力上优于当前性能最好的端到端模型, Lu 等[50] 模型的优化性能甚至超越了 LKH3 专业组合优化求解器, 且算法运行时间数倍少于 LKH3; 但是该方法的运算时间仍然远长于端到端模型, Kool 等[34] 的模型采用简单的贪婪策略可见在数秒内运算得到接近最优解的方案, 具有快速在线求解的优势. 可见两类不同的方法具有不同的优势, 需要根据不同应用场景和问题规模进行选择.

3.4 基于深度学习的多目标组合优化方法

目前绝大多数基于深度强化学习解决传统优化问题的研究都是针对单目标优化问题, 而使用深度强化学习方法解决传统多目标优化问题的研究很少, 值得注意的是, “多目标 (深度) 强化学习”的概念[53–54] 很早就被提出并且存在很多文献对其进行研究, 但是其研究的是如何设计具有多个奖励信号的强化学习算法, 例如如何利用强化学习算法控制潜艇寻找目标[53], 其中需要最大化寻找到目标的数量和最小化寻找的时间, 其研究的主体是多目标强化学习算法, 而不是如何利用强化学习方法解决传统的多目标优化问题.

针对该问题, Li 等[36] 最近提出了一个简单却有效的框架 DRL-MOA, 尝试使用深度强化学习方法对传统的多目标优化问题进行求解, 该方法在 2 个、3 个和 5 个目标的多目标 TSP (MOTSP) 问题上取得了显著优于传统 MOEA/D 和 NSGA-II 多目标优化算法的优化效果, 且优于多目标局部搜索算法 MOGLS, 并且随着问题规模的扩大 (如 200-、500-MOTSP 问题), 该方法的优化效果显著超越了传统优化算法, 且具有数倍于传统算法的求解速度. 该方法借鉴 Pointer network 模型采用端到端的求解框架, 采用基于分解的思想将多目标问题分解为多个子问题, 并将每个子问题建模为 Pointer network 模型, 多个子模型利用基于邻居的参数迁移的方法进行协同训练, 文章利用随机生成的 40 城市 TSP 问题进行模型训练, 一旦模型训练好, 可以求解任意生成的 100、200、500 城市的 TSP 问题, 而不用重新训练模型, 具有较强的泛化能力. 得益于端到端的求解框架, 求解速度快以及泛化能力强是该方法的优势, 且该方法的思想很容易迁移到其他多目标优化问题的求解中, 但是文章仅对多目标 TSP 问题进行了实验研究, 对其他多目标组合优化问题以及更为普遍的多目标连续优化问题没有进行研究, 并且由于该类方法神经网络模型个数与权重个数成正比, 如何提高该类方法的训练效率也是未来的研究方向.

3.5 基于深度学习的组合优化方法总结

从上述方法可见, 端到端模型具有求解速度远超传统优化算法的优势, 也是近年来研究较多的一类方法, 模型一旦训练完成, 可以对任意同类型的问题进行求解, 具有很强的泛化能力, 但是很难保证解的优化效果, 尤其随着问题规模的扩大, 其优化能力与传统优化算法之间的差距会不断扩大. 深度强化学习改进的局部搜索方法是近年来兴起的另外一类方法, 其本质上仍然是启发式搜索算法, 但是没有采用人工设计的搜索规则, 而是利用深度强化学习算法对搜索规则进行学习, 因此该方法具有较强的优化能力, 其优化效果可以超越传统的优化

算法, 但是其求解时间仍然远慢于端到端模型, 因此决策者需要根据优化效果和求解速度之间的权衡来选择不同的方法.

由于端到端模型可以采用监督式和强化学习方法进行训练, 文献 [55−56] 对监督式和强化学习训练方法进行了详细的实验对比和分析, 论文发现强化学习训练方法收敛比监督式训练方法慢, 但强化学习得到的模型具有更强的泛化能力.

由于存在多种组合优化问题, 不同文献的研究重点不同, 导致存在多种不同的模型方法. 例如文献 [32, 34, 50] 等偏重于解决 TSP、VRP 等路径优化问题, 其中节点选择的顺序对结果有很大影响, 因此基于 Attention 机制的方法在此类问题上有较好的效果. 并且对于复杂的路径选择问题, 如 CVRP 问题, 目前的研究均采用 Attention 机制, 而没有单纯采用图神经网络的方法, 可见 Attention 机制在处理具有序列特性的组合优化问题上具有较好的性能; 而文献 [37, 39, 48] 等偏重于解决 MVC、MAXCUT 等问题, 即点选择问题, 该类问题对节点的顺序没有要求, 此种情况下图神经网络在该类问题上应用较多; 同时, 结合图神经网络和 Attention 机制的方法在 TSP 等路径优化问题上也取得了较好的效果[35, 49].

鉴于此, 为了更清晰地对求解不同类型组合优化问题的不同模型方法进行比较, 本节对解决不同组合优化问题的不同文献进行统计, 并对各个研究所采用的模型进行归纳, 结果如表 4 所示, 其中部分文献是 2020 年刚提出的新方法, 在模型和实验结果上都有突出的特点和表现并且被多次引用, 但是仅发表了预印版 (在审) 而暂未通过同行评审, 因此上文并没有对这些文献进行详细介绍, 仅列出供读者查阅并以星号 (*) 标注.

由表 4 可见, 近年来图神经网络结合各种搜索方法 (波束搜索、树搜索) 在各种组合优化问题上得到了广泛的应用, 其主要应用于没有序列特性的组合优化问题, 如 MVC、MAXCUT 等. 而基于 Attention 机制的指针网络方法是解决具有序列决策特性组合优化问题的主要方法, 如 TSP、VRP 等问题.

针对基于指针网络的端到端模型, 由于其核心是借鉴机器翻译领域的 Attention 机制, 因此追踪当前自然语言处理领域的前沿成果是提升指针网络模型性能的重要思路, 如 Kool 等[34] 借鉴了 Transformer 模型中的 Multi-head attention 机制, 使得其模型在组合优化问题上取得了当前业界最优的效果. 同时, 如何改进编码器和解码器的神经网络模型结构也是提高模型性能的一个重要研究方向.

针对基于图神经网络的端到端模型, 由于图神经网络是当前人工智能领域的研究热点, 如何从众多模型中选择改进适合求解不同组合优化问题的图神经网络模型是一个重要的研究方向, 同时波束搜索、树搜索耗时长也是制约该类方法的一个问题, 如何高效地将图神经网络和 Attention 机制相结合是未来可行的研究思路.

针对深度强化学习改进的局部搜索方法, 目前的研究仍然处于起步阶段, 但已经取得了超越传统组合优化求解器的成果, 如何提高解搜索的效率以及扩大启发式算子的搜索空间是未来提升算法性能的重要研究方向.

表 4　不同组合优化问题求解算法统计与比较

Table 4　Summary and comparison of algorithms on different combinatorial optimization problems

组合优化问题	文献	模型细节
TSP 问题	[30–36]	基于 Ptr-Net 架构 (Encoder-decoder-attention)
	[37]	GNN + DQN
	[40–41]	GNN + 监督式训练 + 波束搜索
VRP 问题	[32, 34]	基于Ptr-Net 架构 (Encoder-decoder-attention)
	[47, 49–50]	DRL 训练局部搜索算子. [47]: Ptr-Net 模型, [49]: Graph attention 模型, [50]: Transformer attention 模型.
最小顶点覆盖问题 (MVC)	[37–38, 48]	GNN + RL
	[39]	GNN + 监督式训练 + 树搜索
最大割集问题 (MaxCut)	[37]	GNN + DQN
	[57]	Message passing neural network (MPNN) + DQN
	[58] *	CNN&RNN + PPO
适定性问题 (Satisfiability)	[39, 48]	GNN + 监督式训练/RL
最小支配集问题 (MDS)	[48]	GNN + RL
	[59] *	Decision diagram + RL
极大团问题 (MC)	[39, 48]	GNN + 监督式训练/RL
最大独立集问题 (MIS)	[39]	GNN + 监督式训练 + 树搜索
	[60] *	GNN + RL + 蒙特卡洛树搜索
背包问题 (Knapsack)	[31]	Ptr-Net + RL
车间作业调度问题	[47]	LSTM + RL 训练局部搜索算子
装箱问题 (BPP)	[61] *	LSTM + RL
	[62] *	NN + RL + 蒙特卡洛树搜索
图着色问题	[48]	GNN + RL
	[63] *	LSTM + RL + 蒙特卡洛树搜索

4　深度强化学习解决组合优化问题: 应用综述

组合优化问题广泛存在于工业、制造、通信、交通等各个领域, 随着在各个领域中实际问题规模的

不断扩大以及对算法求解时间的严格要求, 传统运筹优化方法很难在可接受时间内实现问题的在线求解, 基于深度强化学习的组合优化方法作为近年来提出的一类前沿方法, 具有求解速度快、泛化能力强的优势, 本节对近年来该类方法的应用研究进行综述. 首先对应用较多的网络与通信领域的研究进行综述, 其次对交通、电网等其他领域的应用研究进行介绍.

4.1 网络与通信领域

由于网络与通信领域存在多种典型的组合优化问题, 如资源分配、路由拓扑优化等, 因此基于深度强化学习的组合优化在网络与通信领域存在较多的应用.

1) 资源分配

在网络与通信领域, 资源分配问题是指将有限的 CPU、内存、带宽等资源分配给不同的用户或任务需求, 如虚拟网络功能部署问题、网络资源切片问题等.

网络功能虚拟化技术 (Network function virtualization, NFV) 通过标准的 IT 虚拟化技术将网络功能虚拟化, 是当前网络通信的前沿技术, 虚拟网络功能 (Virtual network function, VNF) 是 NFV 架构中的虚拟网络功能单元, VNF 的放置与部署问题是当前网络通信领域研究较多的一类问题, 鉴于传统的 VNF 部署方法通常需要数十分钟才可以完成优化过程, 近年来涌现出利用深度强化学习实现 VNF 智能在线部署的多个研究. 文献 [64] 将 VNF 部署问题建模为混合整数规划问题, 并将其转化为马尔科夫过程, 在满足不同的端到端时延需求的前提下, 以最小化总任务时间为目标, 文章以 DQN 强化学习方法对模型进行训练, 从而实现 VNF 在线部署, 该方法在收敛性能以及优化能力上优于多个基准方法. 文献 [65–66] 基于 GNN 图神经网络对 VNF 网元资源需求进行预测, 以提高 VNF 部署的准确性. 文献 [67] 考虑 VNF 网元资源分配的特性, 指出传统强化学习方法很难处理 VNF 部署问题中的大规模离散决策空间探索问题, 因此文章对 DDPG (Deep deterministic policy gradient algorithm) 强化学习算法进行了改进, 提出了增强 DDPG 算法, 该方法的优化能力超过了传统 DDPG 方法以及整数规划方法. 文献 [68] 采用 Ptr-Net 的 Encoder-decoder 架构对 VNF 部署问题进行了求解, 其 Encoder 和 Decoder 均采用 LSTM模型, 文章利用拉格朗日松弛将该约束问题转化为无约束问题, 并采用基于蒙特卡洛的策略梯度方法对模型进行训练. 通过与约束问题求解器 Gecode和传统启发式方法对比, 实验显示该方法的优化性能在大多数小规模和大规模问题上均优于对比算法.

文献 [69] 基于深度强化学习对无线边缘计算网络的切片策略进行了研究, 文章设计了一种 D-DRL 分布式强化学习框架, 采用一个中心协调器和多个分布式智能体对切片策略进行协同优化, 并采用 DDPG 强化学习算法对模型进行训练. 传统网络切片方法通常会受到不确定的资源需求、不确定的服务时间等不确定性因素影响, 文献 [70] 将网络切片过程建模为半马尔科夫过程, 采用 Deep double Q-learning 方法对切片策略进行优化, 以克服 DQN 收敛慢的缺点, 模型在训练时可以充分地对大规模决策空间进行探索, 从而能够在在线优化时有效处理不确定的状态信息, 进行实时在线响应, 将不同的网络资源分配给不同种类的用户, 实验证明该方法的长期优化能力相比于当前最优的方法提高 40 %, 且在线优化耗时可以忽略不计.

文献 [71] 对雾计算中的复杂资源分配问题进行了研究, 将雾计算建模为马尔科夫过程, 以在既定时延内满足用户最大需求为目标, 利用 DQN 强化学习方法对雾计算中的资源分配进行在线求解, 可以得到接近最优解的优化效果, 并优于传统的启发式资源分配方法.

2) 拓扑与路由优化

在通信网络或者无线传感网络中, 通常需要对路由策略、传感器的连接拓扑进行优化, 以降低通信时延和成本.

文献 [72] 基于深度强化学习方法对无线通信网络的路由策略进行研究, 文章采用图神经网络对通信网络的图结构进行建模, 对当前网络信息进行编码, 并输出选择不同节点的 Q 值, 采用 DQN 算法对图神经网络进行训练. 实验显示该方法具有很强的泛化能力, 一旦模型训练完成, 能够对任意结构的网络进行路由策略的在线优化. 文献 [73] 基于 DRL 和蒙特卡洛树搜索提出了一种 DRL-TC 方法, 利用该方法对无线自组织传感网络的通信拓扑连接进行优化, 文章采用深度神经网络对问题进行建模, 利用强化学习方法对其进行训练, 利用该神经网络的输出指导蒙特卡洛树搜索的过程, 从而得到最优的通信拓扑连接. 文献 [74] 对无线传感网络中移动代理的路由策略进行了研究, 采用 Ptr-Net 模型输出移动代理的最优路径, 文章利用 Actor-critic 算法对其进行训练, 实验显示该方法能够有效地对无线传感网络的流量进行控制, 降低能量消耗. 文献 [75] 基于深度强化学习方法对 D2D 无线通信网络的链路选择策略进行优化, 文章采用 Ptr-Net 模型对链路进行选择, 其 Encoder 和 De-

coder 均使用 LSTM模型, 并利用策略梯度方法对模型进行训练, 实验证明该方法能够在更短计算时间内达到和传统方法相似的优化性能.

3) 计算迁移

计算迁移 (Computation offloading) 是通过将部分计算任务从本地迁移到远程设备以解决移动终端资源受限问题的一个有效途径, 由于无线信道状况变化较快, 需要快速进行策略相应, 而传统的数值优化方法很难实现在线快速优化, 鉴于此, 多个文献 [76–78] 采用深度强化学习实现计算迁移策略的在线优化. 文献 [76] 提出了一种基于深度强化学习的 DROO 计算迁移框架, 基于 DQN 算法对移动边缘计算网络对在线计算迁移策略进行优化, 在优化时间和优化效果上优于传统整数规划算法. 文献 [77] 基于深度强化学习方法对多址边缘计算的计算迁移策略进行了研究, 文章采用 Seq2Seq 模型对策略进行建模, 利用 PPO 算法对模型进行训练, 在不同任务数量的情况下均接近最优解, 且超越了多个基准方法. 文献 [78] 将移动边缘计算网络中的计算迁移问题转换成了多维多背包问题, 利用多指针网络 (MPtr-Net) 对策略进行建模, 并采用 REINFORCE 强化学习方法对该指针网络进行训练, 最后采用波束搜索得到最终的方案. 实验表明该方法的优化性能较基准启发式算法提高了 25 %, 且运行时间优于 OR-tool.

4.2 其他领域

1) 交通领域

在货物配送领域, 随着电商规模的不断扩大, 当前的配送路径优化方法很难做到城际交通规划系统的在线实时响应, 鉴于此, 文献 [79] 利用深度强化学习方实现了配送路径的在线快速生成, 文章设计了一种基于 Struct2Vec 图神经网络和 Ptr-Net 注意力网络的模型, 采用图注意力机制对路径生成的过程进行建模, 并采用策略梯度方法对该模型进行训练, 文章基于德国科隆市的城市交通路网对该方法进行了测试, 实验显示该方法可以在可接受时间内实现配送路径的快速生成, 且在相同求解时间内优于传统的整数规划方法和启发式方法. 在网约车领域, 订单的分配和司机载客区域的分配是一个复杂的组合优化问题, 传统运筹优化方法很难处理大规模订单的在线调度和响应, 文献 [80] 利用深度强化学习方法对该问题进行了研究, 文章参考 Attention 机制对深度神经网络的结构进行了设计, 并分别研究了 DQN 和 PPO 训练算法在不同场景下的表现, 实验表明 DQN 和 PPO 训练方法在不同的客流需求场景各具优势, 且能够实现在线实时优化. 文献 [81] 采用深度强化学习方法对交通信号灯控制策略进行优化, 以信号灯持续时间作为优化变量, 结合决策网络、目标网络、Double-Q-Learning 等深度强化学习方法对模型进行训练, 取得了比传统交通信号灯控制方法更好的效果.

2) 生产制造领域

在生产制造领域存在大量组合优化问题, 近年来基于深度强化学习的组合优化模型在生产制造领域也多有应用. 文献 [47, 82] 对经典的车间工作流调度问题进行了研究, 采用深度强化学习方法对局部搜索的解搜索规则进行学习, 利用 Actor-critic 深度强化学习算法对搜索规则进行优化学习, 实验表明该两个模型在优化性能上均超越了传统启发式局部搜索方法; 置换车间工作流调度问题是对是对流水车间调度问题的进一步约束, 文献 [83−84] 均采用指针网络模型对该问题进行了研究, 文献 [83] 采用了经典的指针网络模型, 并利用 CPLEX 求解器构造大量样本对该模型进行监督式训练, 实验结果表明 Attention 机制能够有效提高解的质量, 文献 [84] 采用了 Multi-head attention 机制进行建模, 并利用 REINFORCE 强化学习算法对模型进行训练, 实验表明该模型超越了多个启发式搜索算法和传统的指针网络模型.

3) 高性能计算领域

人工智能模型的训练是一个耗时极长的任务, 合理地对计算资源进行规划和调度能够有效提高计算效率, 随着神经网络规模的不断扩大, 多 CPU 和多 GPU 混合训练是当前通用的一个方法, 将神经网络模型的不同计算功能部署在不同的计算设备上对训练速度有很大的影响, 该问题是一个典型的组合优化问题, 文献 [85−86] 利用深度强化学习方法对该部署问题进行了研究, 文献 [85] 采用了经典的 Ptr-Net 模型架构对问题进行建模, 并利用策略梯度方法进行分布式训练, 实验证明该方法可以将 Tensorflow 计算图的训练速度提高 20 % 以上. 文献 [86] 在此基础上提出了一个分层模型, 首先 Grouper 层将计算图中的不同计算部分进行分组, 然后 Placer 层根据 Grouper 层得到的分组结果输出部署方案, Placer 层仍采用 Encoder-decoder 模型, 实验证明该方法可以将 Tensorflow 计算图的训练速度提高 60 %.

4) 微电网能量管理领域

在微电网能量管理问题中, 用电、储能等设备的启停控制是典型的离散优化问题, 部分学者采用不同的深度强化学习方法对该问题进行了研究. Francois-Lavet等[87] 使用深度强化学习方法对微电网能量管理问题进行了研究, 文章考虑了一个包含

短期储能、长期储能以及光伏发电的微电网系统, 将微电网能量管理问题建模为马尔科夫过程, 以最小化用电费用为目标, 以充电、放电、无操作作为动作空间, 以卷积神经网络对该问题进行建模, 采用 DQN 算法进行训练, 实验发现该训练好的模型能够有效地提高能源利用率, 降低用电费用, 但是文章没有和基准方法进行对比. 文献 [88] 构建了一个包含光伏发电、储氢装置、蓄电池的孤岛型复合能源系统, 并将复合储能系统的协调控制转化为序列决策问题, 文章仍然采用卷积神经网络对问题进行建模, 采用 Double DQN 深度强化学习方法对该系统的协调控制策略进行优化, 实验表明该方法与 DQN 算法相比具有更快的收敛速度. 文献 [89] 利用 Double Q-learning 深度强化学习方法对空调和通风系统的温度调节、启停等策略进行优化, 从而在保证良好的温度舒适度和空气质量的前提下达到最低的能量消耗, 以实现最优能量管理优化, 文章在实际环境中对该模型进行测试, 结果发现该方法与传统的温度调节方法相比更具有优越性. 文献 [90] 利用深度强化学习算法对对楼宇的智能能量管理问题进行了研究, 对楼宇中空调、电视、电动汽车等用电设备的启停进行控制, 文章利用深度神经网络对该问题进行建模, 分别研究了 DQN 和 DPG (Deep policy gradient) 训练算法在该问题上的表现, 实验表明 DPG 策略梯度方法能够更加有效地实现削峰填谷和降低能源消耗.

由于基于深度强化学习的组合优化方法是近年来刚提出的一类方法, 其应用研究较少. 近年来的应用研究对 Ptr-Net、图神经网络模型以及其他深度神经网络模型均有应用, 对 DQN、Double DQN、Actor-critic、PPO、DDPG 等先进的深度强化学习方法也均有涉及. 但由于各个应用研究都是针对各自不同的实际问题进行建模, 其模型的结构、状态空间、动作空间都有较大区别, 很难在文献之间进行横向比较, 且大多数文献的实验对比较为匮乏, 虽能体现出算法的优化效果, 但对算法的优化性能分析较少, 只有少量文献对算法结果与最优解之间的差距进行了分析. 目前的应用研究大多应用在传统算法很难进行在线优化的背景下, 基于深度强化学习的优化算法具有求解速度快、泛化能力强的优势, 由于工业、制造、交通等领域存在大量组合优化问题, 因此该类方法具有广泛的应用背景.

5 结论与展望

实际生产生活中存在很多组合优化问题, 已有大量研究对各种组合优化方法进行了研究, 但是面对大规模复杂组合优化难题时, 现有方法很难在可接受时间内找到满意解, 难以满足很多问题在线优化的需求. 而近年来基于深度强化学习的组合优化方法在多种组合优化问题上展示出了良好性能, 具有较强的泛化性能和快速的求解速度, 为组合优化问题的在线求解提供了新的思路. 因此本文从理论方法和应用研究两个层面综述了近些年关于基于深度强化学习组合优化方法的相关研究, 以期为未来该领域的研究提供较好的支撑.

现有研究已经显示出深度强化学习在解决组合优化问题方面的优势, 但是相关研究还比较浅显, 当前的研究尚属于起步阶段, 仍然存在一系列问题. 要构建基于 DRL 解决组合优化问题的理论方法体系, 还需从如下几个方面开展研究:

1) 在模型方面. 在当前的研究中, 直接采用深度神经网络模型输出的解通常较差, 大部分文献都需要进一步通过波束搜索、局部搜索、采样策略等方式进一步提升解的质量, 这说明当前的模型仍然有很大的提升空间, 未来需要进一步对求解组合优化问题的深度神经网络模型进行研究, 如何有效结合图神经网络和 Attention 机制是一个较好的研究点.

2) 在研究对象方面. 当前文献研究的问题都相对简单, 而实际中的组合优化问题通常具有多目标、多约束、非静态等复杂特性, 当前方法很难对该类问题进行求解, 目前考虑多目标优化、约束优化的文章较少. 未来基于深度强化学习方法对多目标、约束优化、动态优化问题进行研究是一个重要的研究方向.

3) 在深度强化学习训练算法方面. 目前对端到端模型的训练大多采用 REINFORCE、DQN 等传统训练算法, 具有采样效率低、收敛慢等缺陷, 如何根据组合优化问题的特性设计更加高效的强化学习训练算法也是一个未来需要着重研究的内容.

4) 最后, 如何利用基于深度强化学习的组合优化方法来解决工程实际中的在线调度优化问题将会成为未来重要的研究方向.

References

1 Papadimitriou C H, Steiglitz K. *Combinatorial Optimization: Algorithms and Complexity*. Mineola, New York: Dover Publications, 1998.

2 Festa P. A brief introduction to exact, approximation, and heuristic algorithms for solving hard combinatorial optimization problems. In: Proceedings of the 16th International Conference on Transparent Optical Networks (ICTON). Graz, Austria: IEEE, 2014. 1−20

3 Lawler E L, Wood D E. Branch-and-bound methods: A survey. *Operations Research*, 1966, **14**(4): 699−719

4 Bertsekas D P. *Dynamic Programming and Optimal Control*. Belmont, MA: Athena Scientific, 1995.

5 Sniedovich M. *Dynamic Programming: Foundations and Principles* (Second edition). Boca Raton, FL: CRC Press, 2010.

6 Williamson D P, Shmoys D B. *The Design of Approximation Algorithms*. Cambridge: Cambridge University Press, 2011.

7 Vazirani V V. *Approximation Algorithms*. Berlin, Heidelberg: Springer, 2003.

8 Hochba D S. Approximation algorithms for NP-hard problems. *ACM Sigact News*, 1997, **28**(2): 40−52

9 Teoh E J, Tang H J, Tan K C. A columnar competitive model with simulated annealing for solving combinatorial optimization problems. In: Proceedings of the 2006 IEEE International Joint Conference on Neural Network. Vancouver, BC, Canada: IEEE, 2006. 3254−3259

10 Van Laarhoven P J M, Aarts E H L, Lenstra J K. Job shop scheduling by simulated annealing. *Operations Research*, 1992, **40**(1): 113−125

11 Wesley Barnes J, Laguna M. Solving the multiple-machine weighted flow time problem using tabu search. *IIE Transactions*, 1993, **25**(2): 121−128

12 Basu S. Tabu search implementation on traveling salesman problem and its variations: A literature survey. *American Journal of Operations Research*, 2012, **2**(2): 163−173

13 Halim A H, Ismail I. Combinatorial optimization: Comparison of heuristic algorithms in travelling salesman problem. *Archives of Computational Methods in Engineering*, 2019, **26**(2): 367−380

14 Rezoug A, Bader-El-Den M, Boughaci D. Guided genetic algorithm for the multidimensional knapsack problem. *Memetic Computing*, 2018, **10**(1): 29−42

15 Lin B, Sun X Y, Salous S. Solving travelling salesman problem with an improved hybrid genetic algorithm. *Journal of Computer and Communications*, 2016, **4**(15): 98−106

16 Prado R S, Silva R C P, Guimaraes F G, Neto O M. Using differential evolution for combinatorial optimization: A general approach. In: Proceedings of the 2010 IEEE International Conference on Systems, Man and Cybernetics. Istanbul, Turkey: IEEE, 2010. 11−18

17 Onwubolu G C, *Davendra D. Differential Evolution: A Handbook for Global Permutation-Based Combinatorial Optimization. Vol. 175*. Berlin, Heidelberg: Springer, 2009.

18 Deng W, Xu J J, Zhao H M. An improved ant colony optimization algorithm based on hybrid strategies for scheduling problem. *IEEE Access*, 2019, **7**: 20281−20292

19 Ramadhani T, Hertono G F, Handari B D. An ant colony optimization algorithm for solving the fixed destination multi-depot multiple traveling salesman problem with non-random parameters. *AIP Conference Proceedings*, 2017, **1862**: 030123

20 Zhong Y W, Lin J, Wang L J, Zhang H. Discrete comprehensive learning particle swarm optimization algorithm with Metropolis acceptance criterion for traveling salesman problem. *Swarm and Evolutionary Computation*, 2018, **42**: 77−88

21 Nouiri M, Bekrar A, Jemai A, Niar S, Ammari A C. An effective and distributed particle swarm optimization algorithm for flexible job-shop scheduling problem. *Journal of Intelligent Manufacturing*, 2018, **29**(3): 603−615

22 Lourenco H R, Martin O C, Stutzle T. Iterated local search: Framework and applications. *Handbook of Metaheuristics*. Springer, 2019. 129−168

23 Grasas A, Juan A A, Lourenco H R. SimILS: A simulation-based extension of the iterated local search metaheuristic for stochastic combinatorial optimization. *Journal of Simulation*, 2016, **10**(1): 69−77

24 Zhang G H, Zhang L J, Song X H, Wang Y C, Zhou C. A variable neighborhood search based genetic algorithm for flexible job shop scheduling problem. *Cluster Computing*, 2019, **22**(5): 11561−11572

25 Hore S, Chatterjee A, Dewanji A. Improving variable neighborhood search to solve the traveling salesman problem. *Applied Soft Computing*, 2018, **68**: 83−91

26 Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of go without human knowledge. *Nature*, 2017, **550**(7676): 354−359

27 Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529−533

28 Hopfield J J, Tank D W. "Neural" computation of decisions in optimization problems. *Biological Cybernetics*, 1985, **52**(3): 141−152

29 Smith K A. Neural networks for combinatorial optimization: A review of more than a decade of research. *INFORMS Journal on Computing*, 1999, **11**(1): 15−34

30 Vinyals O, Fortunato M, Jaitly N. Pointer networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2015. 2692−2700

31 Bello I, Pham H, Le Q V, Norouzi M, Bengio S. Neural combinatorial optimization with reinforcement learning. In: Proceedings of the 5th International Conference on Learning Representations (ICLR 2017). Toulon, France, 2017.

32 Nazari M, Oroojlooy A, TakacM, Snyder L V. Reinforcement learning for solving the vehicle routing problem. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc., 2018. 9861−9871

33 Deudon M, Cournut P, Lacoste A, Adulyasak Y, Rousseau L M. Learning heuristics for the TSP by policy gradient. In: Proceedings of the 15th International Conference on the Integration of Constraint Programming, Artificial Intelligence, and Operations Research. Delft, The Netherlands: Springer, 2018. 170−181

34 Kool W, Van Hoof H, Welling M. Attention, learn to solve routing problems! In: Proceedings of the 7th International Conference on Learning Representations (ICLR2019). New Orleans, LA, USA, 2019.

35 Ma Q, Ge S W, He D Y, Thaker D, Drori I. Combinatorial optimization by graph pointer networks and hierarchical reinforcement learning. In: Proceedings of the 1st International Workshop on Deep Learning on Graphs: Methodologies and Applications. New York, NY, USA: AAAI, 2020.

36 Li K W, Zhang T, Wang R. Deep reinforcement learning for multiobjective optimization. *IEEE Transactions on Cybernetics*, 2021, **51**(6): 3103−3114

37 Dai H J, Khalil E B, Zhang Y Y, Dilkina B, Song L. Learning combinatorial optimization algorithms over graphs. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017. 6351−6361

38 Manchanda S, Mittal A, Dhawan A, Medya S, Ranu S, Singh A. Learning heuristics over large graphs via deep reinforcement learning. arXiv preprint arXiv: 1903.03332, 2020

39 Li Z W, Chen Q F, Koltun V. Combinatorial optimization with graph convolutional networks and guided tree search. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc., 2018. 537−546

40 Nowak A, Villar S, Bandeira A S, Bruna J. A note on learning algorithms for quadratic assignment with graph neural networks. In: Proceeding of the 34th International Conference on Machine Learning (ICML). Sydney, Australia, 2017.

41 Joshi C K, Laurent T, Bresson X. An efficient graph convolutional network technique for the travelling salesman problem. arXiv preprint arXiv: 1906.01227, 2019

42 Helsgaun K. An Extension of the Lin-Kernighan-Helsgaun TSP Solver for Constrained Traveling Salesman and Vehicle Routing Problems, Technical Report, Roskilde University, Denmark, 2017.

43 Perron L, Furnon V. Google′s OR-Tools [Online], available: https://developersgooglecom/optimization, 2019

44 OPTIMIZATION G. INC. Gurobi optimizer reference manual, 2015 [Online], available: http://wwwgurobicom, 2014

45 Applegate D, Bixby R, Chvatal V, Cook W. Concorde TSP solver. 2006

46 Bengio Y, Lodi A, Prouvost A. Machine learning for combinat-

orial optimization: A methodological tour d′Horizon. arXiv preprint arXiv: 1811.06128, 2020

47 Chen X Y, Tian Y D. Learning to perform local rewriting for combinatorial optimization. In: Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc., 2019. 6278−6289

48 Yolcu E, Poczos B. Learning local search heuristics for boolean satisfiability. In: Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc., 2019. 7992−8003

49 Gao L, Chen M X, Chen Q C, Luo G Z, Zhu N Y, Liu Z X. Learn to design the heuristics for vehicle routing problem. arXiv preprint arXiv: 2002.08539, 2020

50 Lu H, Zhang X W, Yang S. A learning-based iterative method for solving vehicle routing problems. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020.

51 Scarselli F, Gori M, Tsoi A C, Hagenbuchner M, Monfardini G. The graph neural network model. *IEEE Transactions on Neural Networks*, 2009, **20**(1): 61−80

52 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017. 6000−6010

53 Hsu C H, Chang S H, Liang J H, Chou H P, Liu C H, Chang S C, et al. MONAS: Multi-objective neural architecture search using reinforcement learning. arXiv preprint arXiv: 1806.10332, 2018

54 Mossalam H, Assael Y M, Roijers D M, Whiteson S. Multi-objective deep reinforcement learning. arXiv preprint arXiv: 1610.02707, 2016

55 Joshi C K, Laurent T, Bresson X. On Learning paradigms for the travelling salesman problem. In: Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc., 2019.

56 Joshi C K, Cappart Q, Rousseau L M, Laurent T, Bresson X. Learning TSP requires rethinking generalization. arXiv preprint arXiv: 2006.07054, 2020

57 Barrett T, Clements W, Foerster J, Lvovsky A. Exploratory combinatorial optimization with reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, **34**(4): 3243−3250

58 Beloborodov D, Ulanov A E, Foerster J N, Whiteson S, Lvovsky A I. Reinforcement learning enhanced quantum-inspired algorithm for combinatorial optimization. arXiv preprint arXiv: 2002.04676, 2020

59 Cappart Q, Goutierre E, Bergman D, Rousseau L M. Improving optimization bounds using machine learning: Decision diagrams meet deep reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, **331**(1): 1443−1451

60 Abe K, Xu Z J, Sato I, Sugiyama M. Solving NP-hard problems on graphs by reinforcement learning without domain knowledge. arXiv preprint arXiv: 1905.11623, 2020

61 Hu H Y, Zhang X D, Yan X W, Wang L F, Xu Y H. Solving a new 3D bin packing problem with deep reinforcement learning method. arXiv preprint arXiv: 1708.05930, 2017

62 Laterre A, Fu Y G, Jabri M K, Cohen A S, Kas D, Hajjar K, et al. Ranked reward: Enabling self-play reinforcement learning for combinatorial optimization. arXiv preprint arXiv: 1807.01672, 2018

63 Huang J Y, Patwary M, Diamos G. Coloring big graphs with alphagozero. arXiv preprint arXiv: 1902.10162, 2019

64 Li J L, Shi W S, Zhang N, Shen X M. Delay-aware VNF scheduling: A reinforcement learning approach with variable action set. *IEEE Transactions on Cognitive Communications and Networking*, 2021, **7**(1): 304−318

65 Mijumbi R, Hasija S, Davy S, Davy A, Jennings B, Boutaba R. Topology-aware prediction of virtual network function resource requirements. *IEEE Transactions on Network and Service Management*, 2017, **14**(1): 106−120

66 Mijumbi R, Hasija S, Davy S, Davy A, Jennings B, Boutaba R. A connectionist approach to dynamic resource management for virtualised network functions. In: Proceedings of the 12th Conference on Network and Service Management (CNSM). Montreal, Quebec, Canada: IEEE, 2016. 1−9

67 Quang P T A, Hadjadj-Aoul Y, Outtagarts A. A deep reinforcement learning approach for VNF forwarding graph embedding. *IEEE Transactions on Network and Service Management*, 2019, **16**(4): 1318−1331

68 Solozabal R, Ceberio J, Sanchoyerto A, Zabala L, Blanco B, Liberal F. Virtual network function placement optimization with deep reinforcement learning. *IEEE Journal on Selected Areas in Communications*, 2020, **38**(2): 292−303

69 Liu Q, Han T, Moges E. EdgeSlice: Slicing wireless edge computing network with decentralized deep reinforcement learning. arXiv preprint arXiv: 2003.12911, 2020

70 Van Huynh N, Hoang D T, Nguyen D N, Dutkiewicz E. Optimal and fast real-time resource slicing with deep dueling neural networks. *IEEE Journal on Selected Areas in Communications*, 2019, **37**(6): 1455−1470

71 Mseddi A, Jaafar W, Elbiaze H, Ajib W. Intelligent resource allocation in dynamic fog computing environments. In: Proceedings of the 8th International Conference on Cloud Networking (CloudNet). Coimbra, Portugal: IEEE, 2019. 1−7

72 Almasan P, Suarez-Varela J, Badia-Sampera A, Rusek K, Barlet-Ros P, Cabellos-Aparicio A. Deep reinforcement learning meets graph neural networks: Exploring a routing optimization use case. arXiv preprint arXiv: 1910.07421, 2020

73 Meng X Y, Inaltekin H, Krongold B. Deep reinforcement learning-based topology optimization for self-organized wireless sensor networks. In: Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM). Waikoloa, HI, USA: IEEE, 2019. 1−6

74 Lu J Y, Feng L Y, Yang J, Hassan M M, Alelaiwi A, Humar I. Artificial agent: The fusion of artificial intelligence and a mobile agent for energy-efficient traffic control in wireless sensor networks. *Future Generation Computer Systems*, 2019, **95**: 45−51

75 Zhang S, Shen W L, Zhang M, Cao X H, Cheng Y. Experience-driven wireless D2D network link scheduling: A deep learning approach. In: Proceedings of the 2019 IEEE International Conference on Communications (ICC). Shanghai, China: IEEE, 2019. 1−6

76 Huang L, Bi S Z, Zhang Y J A. Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks. *IEEE Transactions on Mobile Computing*, 2020, **19**(11): 2581−2593

77 Wang J, Hu J, Min G Y, Zhan W H, Ni Q, Georgalas N. Computation offloading in multi-access edge computing using a deep sequential model based on reinforcement learning. *IEEE Communications Magazine*, 2019, **57**(5): 64−69

78 Jiang Q M, Zhang Y, Yan J Y. Neural combinatorial optimization for energy-efficient offloading in mobile edge computing. *IEEE Access*, 2020, **8**: 35077−35089

79 Yu J J Q, Yu W, Gu J T. Online vehicle routing with neural combinatorial optimization and deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 2019, **20**(10): 3806−3817

80 Holler J, Vuorio R, Qin Z W, Tang X C, Jiao Y, Jin T C, et al. Deep reinforcement learning for multi-driver vehicle dispatching and repositioning problem. In: Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM). Beijing, China: IEEE, 2019. 1090−1095

81 Liang X Y, Du X S, Wang G L, Han Z. A deep reinforcement learning network for traffic light cycle control. *IEEE Transactions on Vehicular Technology*, 2019, **68**(2): 1243−1253

82 Chen X Y, Tian Y D. Learning to progressively plan. arXiv preprint arXiv: 1810.00337, 2018

83 Zheng P, Zuo L L, Wang J L, Zhang J. Pointer networks for solving the permutation flow shop scheduling problem. In: Pro-

ceedings of the 48th International Conference on Computers & Industrial Engineering (CIE48). Auckland, New Zealand, 2018. 2−5

84 Pan R Y, Dong X Y, Han S. Solving permutation flowshop problem with deep reinforcement learning. In: Proceedings of the 2020 Prognostics and Health Management Conference (PHM-Besancon). Besancon, France: IEEE, 2020. 349−353

85 Mirhoseini A, Pham H, Le Q V, Steiner B, Larsen R, Zhou Y F, et al. Device placement optimization with reinforcement learning. arXiv preprint arXiv: 1706.04972, 2017

86 Mirhoseini A, Goldie A, Pham H, Steiner B, Le Q V, Dean J. A hierarchical model for device placement. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, BC, Canada, 2018.

87 Francois-Lavet V, Taralla D, Ernst D, Fonteneau R. Deep reinforcement learning solutions for energy microgrids management. In: Proceedings of the 13th European Workshop on Reinforcement Learning (EWRL 2016). Barcelona, Spain, 2016.

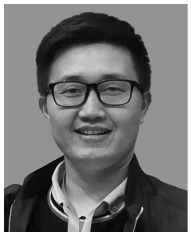
88 Zhang Zi-Dong, Qiu Cai-Ming, Zhang Dong-Xia, Xu Shu-Wei, He Xing. A coordinated control method for hybrid energy storage system in microgrid based on deep reinforcement learning. *Power System Technology*, 2019, **43**(6): 1914−1921
(张自东, 邱才明, 张东霞, 徐舒玮, 贺兴. 基于深度强化学习的微电网复合储能协调控制方法. 电网技术, 2019, **43**(6): 1914−1921)

89 Valladares W, Galindo M, Gutierrez J, Wu W C, Liao K K, Liao J C, et al. Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm. *Building and Environment*, 2019, **155**: 105−117

90 Mocanu E, Mocanu D C, Nguyen P H, Liotta A, Webber M E, Gibescu M, et al. On-line building energy optimization using deep reinforcement learning. *IEEE Transactions on Smart Grid*, 2019, **10**(4): 3698−3708

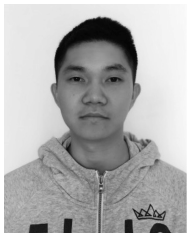
李凯文 国防科技大学系统工程学院博士研究生. 主要研究方向为能源互联网技术, 深度强化学习与优化技术.
E-mail: likaiwen@nudt.edu.cn
(**LI Kai-Wen** Ph. D. candidate at College of System Engineering, National University of Defense Technology. His research interest covers energy internet technology, deep reinforcement learning and optimization.)

张 涛 国防科技大学系统工程学院教授. 主要研究方向为能源互联网技术, 基于计算智能的优化与决策技术.
E-mail: zhangtao@nudt.edu.cn
(**ZHANG Tao** Professor at College of System Engineering, National University of Defense Technology. His research interest covers energy internet technology, optimization and decision making based on computational intelligence.)



王 锐 国防科技大学系统工程学院副研究员. 主要研究方向为能源互联网技术, 计算智能理论与方法, 多目标进化算法. 本文通信作者.
E-mail: ruiwangnudt@gmail.com
(**WANG Rui** Associate professor at College of System Engineering, National University of Defense Technology. His research interest covers energy internet technology, computational intelligence methodologies, multi-objective evolutionary optimizations. Corresponding author of this paper.)

覃伟健 国防科技大学系统工程学院硕士研究生. 主要研究方向为能源互联网技术, 深度强化学习与优化技术.
E-mail: qinweijian@nudt.edu.cn
(**QIN Wei-Jian** Master student at College of System Engineering, National University of Defense Technology. His research interest covers energy internet technology, deep reinforcement learning and optimization.)



贺惠晖 国防科技大学系统工程学院硕士研究生. 主要研究方向为能源互联网技术, 基于计算智能的优化与决策技术.
E-mail: hehuihui@nudt.edu.cn
(**HE Hui-Hui** Master student at College of System Engineering, National University of Defense Technology. His research interest covers energy internet technology, optimization and decision making based on computational intelligence.)

黄 鸿 国防科技大学系统工程学院硕士研究生. 主要研究方向为能源互联网技术, 基于计算智能的优化与决策技术.
E-mail: huanghong@nudt.edu.cn
(**HUANG Hong** Master student at College of System Engineering, National University of Defense Technology. His research interest covers energy internet technology, optimization and decision making based on computational intelligence.)