深度强化学习方法应用于组合优化问题的研究综述：以旅行商问题为例

人工神经网络课程论文

陈翰墨 2020214276

1 引言

在计算机科学和多个工程应用领域，组合优化问题都有着大量的应用，传统上解决此类问题的方法大部分为启发式搜索算法。近年来, 随着深度强化学习的兴起和逐渐深入, 由于强化学习和组合优化问题设定的相似性, 有许多研究开始利用深度强化学习去解决组合优化问题。

作为最有代表性的组合优化问题, 旅行商问题（Traveling Salesman Problem，TSP）早已被证明为是 NP-hard 问题。在这个问题上, 相较于传统的启发式搜索算法，强化学习算法可以给出更具一般性的解，同时也更能够满足实时性的要求。本文概括了自 2015 年以来利用深度强化学习求解旅行商问题的研究成果，总结了其共性以及目前存在的不足, 同时对未来的研究方向进行了展望。

2 组合优化概述

2.1 组合优化问题的定义

优化问题通常可分为离散优化和连续优化两类问题，其中离散优化问题又称为组合优化问题。组合优化问题通常是在一个离散的（有限的或者可数的）的解空间内寻找最优解的过程，其中最优解的定义基于给定的损失函数或者收益函数，对应的优化目标分别是最小化损失函数或者最大化收益函数。

常见的组合优化问题包括旅行商问题（Traveling Salesman Problem，TSP），车间作业调度问题（Job-shop Scheduling Problem, JSSP），背包问题（Knapsack），最小顶点覆盖（Minimum Vertex Cover，MVC）与最大独立集（Maximum Independent Set,MIS）等。

一般的组合优化问题的可以形式化定义为下列问题：假设 V 是全体可行解构成的集合，而 $f : V \mapsto \mathbb{R}$ 为所定义的损失函数，组合优化问题旨在求解使得损失函数最小的可行解,

$$v^* = \underset{v \in V}{\operatorname{argmin}} f(v) \tag{1}$$

以旅行商问题（TSP）为例，问题的优化目标是，对于一个图和所有起点，找到经过图中所有顶点后回到起点的最短路径。或者说，在一个全连接的带权图中找到最短的哈密顿回路（Harmilton circuits）。因此对于给定的图 $G = (V, E)$，可行解集合为 $V = H(G)$ 为 G 中所有的哈密顿回路，而对于任意哈密顿回路 $v \in V$，损失函数定义为该哈密顿回路中所有边权重之和, 即 $f(v) = \sum_{e \in v} w(e)$。

2.2 组合优化问题的求解方法

一般地，对组合优化问题的求解可以分为两类，精确方法和近似方法。

其中精确方法是用于求精确的全局最优解的一类方法。由于大部分组合优化问题本质上是一个整数线性规划或者混合线性规划问题，因此在求解时可以利用传统运筹学中的求解整数线性规划的算法，主要算法包括分支定界法和动态规划法。但是由于许多组合优化问题可行解的数量呈指数型增长趋势，因此精确方法大多只在数据量较小的时候使用，在数据量较大的时候因为复杂度过高而变得不实用。

针对精确算法复杂度过高的问题，尤其是对于某些已经被证明为是 NP-hard 的组合优化问题，提出了近似算法，即不要求求得的解一定为全局最优解，转而求解得到局部最优解或近似最优解的算法。根据对求得最优解的限制，近似算法可以进一步分为两类，第一类是要求求得的解为局部最优解的算法如贪心算法、局部搜索算法和松弛算法等。而第二类为启发式算法，是根据一个事先给定的启发式规则对可行解空间进行搜索的方法，包括模拟退火方法和群体智能方法如遗传算法，蚁群算法，粒子群算法等。启发式方法通常能够在一个可接受的时间内收敛到较好的解，但是很难在理论上保证解的质量。

由于组合优化问题在运筹学，计算机科学，交通管理，生产和供应链管理领域等都有着广泛的应用，因此针对组合优化问题的求解算法一直是近年来的热点话题。早期的研究主要针对算法的理论和应用展开，随着近些年来数据量的爆炸增长，也对算法的实时性提出了更高的要求。

3 强化学习在组合优化上的应用

3.1 强化学习概述

强化学习是机器学习领域的一类方法的统称，通过建模一个智能体与环境交互的过程来模仿人类学习的过程如图 1。

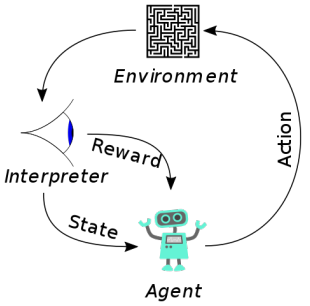

图 1: 强化学习中智能体与环境的交互

它的基本模型是马尔可夫决策模型（Markov Decision Process，MDP），提供了一种对序贯决策问题进行建模的数学框架。马尔可夫决策过程的形式化定义为一个五元组 $M =< S, A, R, P, \gamma >$，其中

- S 为状态空间，$s_t \in S$ 为 t 时刻可能的状态。对于组合优化问题而言，状态空间通常由两种定义方式，第一种可以定义成当前子集内的可能解（比如，对于旅行商问题而言，可以定义为部分顶点集合的可能路径），第二种方式是定义为当前问题的次优解，然后通过迭代的方式改进这个解。
- A 为动作空间，$a_t \in A$ 为 t 时刻可能采取的所有动作（或者行动）的集合，可进一步划分为连续动作空间和离散动作空间。而对于组合优化问题而言，动作空间是离散的。
- $R : S \times A \to \mathbb{R}$ 为奖励函数，即在特定

状态和动作下的收益

- P 为转移概率矩阵，$p(s_{t+1} \mid s_t, a_t)$ 为在给定的状态和动作下到达另一个状态的概率
- γ 为折扣率或贴现率，表示对未来收益的评估时未来收益相较于当前收益的衰减。

对应上述定义，一个智能体的策略可以定义为如下两种形式，

1. 确定性策略 $\pi : S \mapsto A$ 为在特定状态下采取的行动。
2. 随机策略：$\pi = \pi(a_t|s_t)$ 即在特定状态下在动作空间上的一个概率分布。

而强化学习的目标是寻找一个最优策略使得期望收益最大，即

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R\left(s_t, a_t\right)\right] \quad (2)$$

强化学习的算法大体上可以分为基于模型（Model-based）的方法和不基于模型（Model-free）的方法，而不基于模型（Model-free）的方法又可进一步分为基于价值函数（Value-based）的方法和基于策略（Policy-based）的方法。

3.2 利用深度强化学习求解旅行商问题

近年来，随着深度学习在各领域的广泛应用，在诸如自然语言处理和计算机视觉等等领域都取得了与人类相当甚至超越人类的能力，而深度强化学习也在围棋等领域取得了超越人类的成就。由于强化学习的设置与组合优化问题有着相似性：组合优化中离散的可行解空间与强化学习中的动作空间较为相似，同时优化的目标也较为一致；同时与传统的方法不同的是，强化学习并不依赖于具体问题的设置针对某一类问题可以以提供一个更为一般的解法，同时也具有较好的实时性。

最先尝试利用深度强化学习解决旅行商问题是在 2015 年 Vinyals 等人提出的 Pointer Network(Ptr-Net) 中，将旅行商问题视为输入一个特征序列如各个点之间的距离，输出另一序列比如各个点的访问顺序的问题，从而类比为自然语言处理领域的机器翻译问题，借鉴了这一领域中的 Seq2Seq 模型，利用监督学习的方法训练网络，在旅行商问题上取得了较好的效果。传统的组合优化问题的思路都是基于迭代的优化算法，而该模型直接利用神经网络输出最后结果，在这一领域具有开创性的意义。严格来说这一方法并不能算是强化学习，但是由于是首个利用深度学习来解决旅行商问题的，同时思路上对之后的研究也有很大启发，因此将其作为这一领域研究的开始。

由于 Vinyals 等人的研究中采用的是监督式的学习方法，因此需要事先得到训练样本即给定的图模型和对应的解，这就导致了两点问题：第一是导致最终得到的解的质量受限于训练使用的解，第二是通过其他方法训练样本需要大量时间。为了改进这个问题，2017 年 Bello 等人将强化学习方法引入到 Pointer Network 中，其中马尔可夫决策过程的建模为，状态 s_t 为一个 p 维的图嵌入向量，代表 t 时刻已访问的节点，而行动 a_t 则为从当前尚未访问的节点中选取下一个时刻访问的节点，而收益则为负的两个节点之间的距离。针对该强化学习问题采用了 REINFORCE 算法进行训练，同时引入 Critic 网络作为基线来降低训练过程中的方差，最终的结果超过了 Vinyals 等人的研究，甚至在某些设定下接近全局最优解。在这个研究的基础上，Nazari 等人将此模型的应用从静态的旅行商问题拓展至动态的车辆寻路问题（VRP），将训练

时间降低了 60%，不仅在旅行商问题上之前 Bello 等人的模型表现接近，同时也在动态 VRP 问题上取得了比传统算法更好的效果。

与此同时，在 2017 年，Dai 等人将图神经网络（Graph Neural Networks）引入到了组合优化问题中，利用图神经网络表示当前解的图结构，利用 DQN 算法计算尚未访问的节点中各个节点对应的 Q 值，对于下一个节点的选取采取贪心算法从而逐步得到完整的解，在包括旅行商问题在内的多个组合优化问题上取得了与之前方法接近甚至更优的效果。

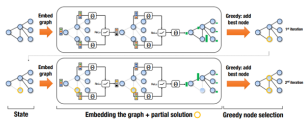

图 2: 利用图神经网络求解组合优化问题

随着 Tranformer 模型在自然语言处理领域取得的巨大成功，在 2018 年，Deudon 等人在 Bello 等人工作的基础上，将编码层的 Seq2Seq 模型换成了 Tranformer 模型从而将注意力机制引入了 Pointer Network，依旧采用 REINFORCE 强化学习算法和 critic 作为基线。在此基础上，还试图通过将传统的 2-Opt 启发式方法与强化学习方法相结合得到更好的结果。

在 2019 年，Kool 等人在引入了 Tranformer 模型的基础上，改进了之前的 Pointer Network 结构，而采用了一种相似的 encoder-decoder 架构，同时将原有的强化学习算法中 critic 的基线换成了 rollout 的基线，使得 Attention 机制不仅能处理旅行商问题，还能处理旅行商的问题的多种变种比如 PCTSP 和 SPCTS 等等。

在这些工作的基础上，Ma 等人将 Pointer Network 和图神经网络相结合，提出了 Graph Pointer Network（GPN），其中编码层包含 Point Encoder 和 Graph Encoder 两部分，同时采用了了分层强化学习 (Hierarchical RL, HRL) 的方法，在大规模的旅行商问题上超越了 Kool 等人的结果。

图 3: 利用 GPN 求解组合优化问题

4 总结与展望

总体上而言，近年来深度强化学习方法在组合优化这一领域的研究取得了较大进展，在包括旅行商问题在内的多个问题上取得了与传统方法接近甚至优于原有方法的效果。大致上，强化学习方法相较于传统方法的优势主要在以下几点：

- 强化学习方法具有更好的泛化能力，传统方法只适用于某一设定下的问题，而强化学习方法可以用于某一类问题甚至是多种问题。
- 强化学习方法在大规模问题时具有更好的求解精度。
- 强化学习方法具有更好的实时性。由于许多强化学习方法具有端到端的特点，因此在训练完成部署之后，输出结果的实时性比传统方法更好。

但是，目前的研究也有以下的不足，

- 与传统的启发式搜索方法一样，缺乏理论上对所给出的解的质量的保证。
- 在模型方面，对样本的利用效率依然是制约最终结果的关键瓶颈。因此，在收集样本和训练网络方面依然需要大量的时间。
- 缺乏统一的测试环境和指标，也难以进一步应用在实际的工业生产环境中。

参考文献

[1] Mittal A, Dhawan A, Manchanda S, et al. Learning heuristics over large graphs via deep reinforcement learning[J]. arXiv preprint arXiv:1903.03332, 2019.

[2] Khalil E, Dai H, Zhang Y, et al. Learning combinatorial optimization algorithms over graphs[C]//Advances in neural information processing systems. 2017: 6348-6358.

[3] 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 [J/OL]. 自动化学报:1-22.https://doi.org/10.16383/j.aas.c200551.

[4] Vinyals O, Fortunato M, Jaitly N. Pointer networks[J]. Advances in neural information processing systems, 2015, 28: 2692-2700.

[5] Bello I, Pham H, Le Q V, et al. Neural combinatorial optimization with reinforcement learning[J]. arXiv preprint arXiv:1611.09940, 2016.

[6] Nazari M, Oroojlooy A, Snyder L, et al. Reinforcement learning for solving the vehicle routing problem[C]//Advances in Neural Information Processing Systems. 2018: 9839-9849.

[7] Deudon M, Cournut P, Lacoste A, et al. Learning heuristics for the tsp by policy gradient[C]//International conference on the integration of constraint programming, artificial intelligence, and operations research. Springer, Cham, 2018: 170-181.

[8] Kool W, van Hoof H, Welling M. Attention, Learn to Solve Routing Problems![C]//International Conference on Learning Representations. 2018.

[9] Ma Q, Ge S, He D, et al. Combinatorial optimization by graph pointer networks and hierarchical reinforcement learning[J]. arXiv preprint arXiv:1911.04936, 2019.

[10] Barrett T, Clements W, Foerster J, et al. Exploratory combinatorial optimization with reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(04): 3243-3250.

[11] Or Rivlin. Reinforcement Learning for Combinatorial Optimization[EB/OL]. https://towardsdatascience.com /reinforcement -learning-for- combinatorial-optimization- d1402e396e91,2019.

[12] Mazyavkina N, Sviridov S, Ivanov S, et al. Reinforcement learning for combinatorial optimization: A survey[J]. arXiv preprint arXiv:2003.03600, 2020.

[13] Li Z, Chen Q, Koltun V. Combinatorial optimization with graph convolutional networks and guided tree search[C]//Advances in Neural Information Processing Systems. 2018: 539-548.

[14] Mittal A, Dhawan A, Manchanda S, et al. Learning heuristics over large graphs via deep reinforcement learning[J]. arXiv preprint arXiv:1903.03332, 2019.

[15] Cappart Q, Goutierre E, Bergman D, et al. Improving optimization bounds using machine learning: Decision dia-

grams meet deep reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 1443-1451.

[16] Miki S, Yamamoto D, Ebara H. Applying deep learning and reinforcement learning to traveling salesman problem[C]//2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE). IEEE, 2018: 65-70.