北京大学

硕士研究生学位论文

题目：基于多智能体强化学习与图神经网络的物流网络调度方法

姓　　名：李锡涵

学　　号：1601214547

院　　系：信息科学技术学院

专　　业：计算机科学与技术（智能科学与技术）

研究方向：网络嵌入与强化学习

导　　师：童云海 教授

二〇一九年五月

版权声明

摘要

复杂物流网络内的资源平衡调度是物流领域中最普遍和重要的问题之一。传统的资源平衡调度广泛采用基于运筹学的多阶段组合优化方法。然而，未来供需的严重不确定性、非凸的复杂业务约束、运输线路的高度复杂性以及运筹学对单一操作的低解释性使得传统运筹学方法在复杂的现代物流网络中遭遇困境。本文提出了一种新型的多智能体强化学习方法来应对这些挑战，通过新颖的状态和奖励设计促进多智能体间的协作，并引入图神经网络和注意力机制，动态生成和表征协作关系，从而提供了一个更为高效、灵活、可解释的运输调度解决方案，并通过广泛的实验证实其优越性。本文的主要工作总结如下：

（一）将复杂运输网络中的资源平衡问题形式化为随机博弈，并由此提出一个合作的多智能体强化学习框架，将网络中的大量运输工具建模为协作的多智能体，同时提出三个层次的合作指标，为改善强化学习中状态和奖励的设计提供指导，以更好地促进复杂物流网络中运输工具智能体的合作。通过这种方式，我们为复杂物流网络中的资源平衡问题提供了一个端到端和高性能的解决方案。该方法不仅对强不确定性的供需预测更加稳健，而且与传统的多阶段运筹学方法相比具备更高的性能和灵活性。

（二）将图神经网络和注意力机制的方法引入多智能体强化学习的框架内，首先从复杂的物流网络中动态抽取出协作图，然后通过注意力机制学习物流网络中包括智能体在内的各对象的实时协作关系，且以权重的方式对外表征，最后通过图卷积的方式对协作信息进行整合，增强了模型的智能性和可解释性。

（三）在现实海洋物流场景中重要的空集装箱调度任务（ECR）上进行了广泛的实验。实验结果表明，我们提出的新方法可以有效学习和表征智能体之间的协作关系，并促进智能体之间的合作。与基于运筹学组合优化方法的传统解决方案相比，我们的方法在性能和可解释性上均带来显着改进。

关键词：强化学习，多智能体系统，图神经网络，注意力机制，物流网络

Resource Balancing in Complex Logistics Networks with Multi-Agent Reinforcement Learning and Graph Attention

Xihan Li (Intelligent Science and Technology)
Directed by Prof. Yunhai Tong

ABSTRACT

Resource balancing within complex transportation networks is one of the most important problems in real logistics domain. Traditional solutions on these problems leverage operational optimization (OR) with demand and supply forecasting. However, the high complexity of transportation routes, severe uncertainty of future demand and supply, non-convex business constraints together with low interpretability of single action make it extremely challenging for traditional OR methods to adapt to the scenarios in the modern complex logistics network. In this work, we propose a novel sophisticated multi-agent reinforcement learning approach to address these challenges. We introduce an innovative cooperative mechanism for state and reward design resulting in more effective and efficient transportation, as well as introducing graph neural network and attention mechanism into the framework, dynamically generating and characterizing collaborative relationships. By this means we provide a more efficient, flexible and interpretable transport scheduling solution, and its superiority is confirmed by extensive experiments. Our major contributions can be summarized as follows:

1. We formulate the resource balancing problem in a complex transportation network as a stochastic game. Then we introduce a cooperative multi-agent reinforcement learning framework as an end-to-end and high-capability solution to the resource balancing problem. Also we propose three levels of cooperative metrics to provide guidance to improve state and reward design, in order to better promote the cooperation in the complex logistics network. This method is not only more robust to the imperfect SnD forecasting but yields higher capability and flexibility compared with the traditional multistage OR-based methods.
2. We introduce graph neural network and attention mechanism into the MARL framework. Firstly, the collaboration graph is dynamically extracted from the complex logistics network, and then the cooperation relation between objects in the logistics

network (including the agents) are learned through the attention mechanism. The real-time collaborative relationship is externally characterized in a weighted manner. Finally, the collaboration information is integrated by means of graph convolution, which enhances the intelligence and interpretability of the model.

3. We conduct extensive experiments on the empty container repositioning task in the scenario of real-world ocean logistics. The result demonstrates that our new approach can stimulate cooperation among agents and lead to much better performance. Compared with traditional solutions based on combinatorial optimization, our approach can give rise to a significant improvement in terms of both performance and interpretability.

KEYWORDS: Reinforcement Learning, Multi-Agent System, Graph Neural Network, Attention Mechanism, Logistics Network

目录

第一章　绪论

1.1　研究背景：传统运筹学方法在现代复杂物流网络中面临的挑战

物流业是现代社会经济发展的基石。随着物流业的快速发展，资源供需（Supply and Demand, SnD）不平衡已成为许多实际物流场景中最普遍而重要的问题之一。例如，在海洋运输领域，由于世界贸易的分工化和差异化，空集装箱的供需非常不平衡[1]；在快递服务业，地区内运输车的供需存在严重的不平衡现象；在快速增长的共享汽车和共享单车领域，由于各种时空因素，共享汽车和共享单车的供需也非常不平衡[2, 3]。在此背景下，有效的资源平衡调度方法成为解决物流业资源供需不平衡问题的关键。不成功的调运方案将导致大量资源需求无法被满足，进一步导致客户满意度降低、资源短缺成本增加和收入下降。

传统的资源平衡调度广泛采用基于运筹学（Operational Research, OR）的方法[1]。这些方法通常是多阶段的：首先，使用时间序列预测的方法来估计每个资源节点的未来供需；然后，采用组合优化方法来找到每个资源节点在未来的最优操作，以最小化预定目标（通常是资源短缺造成的总成本）；最后，通过对基于运筹学模型获得的原始解决方案进行修正，来生成最终可行的执行计划。然而，实际物流场景中未来供需的严重不确定性、非凸形式的复杂业务约束、运输网络的高复杂性和运筹学对单个调度操作的低解释性使得使用传统的运筹学解决方案在实际物流场景中的应用中遭遇了严峻的挑战，具体如下：

未来供需的不确定性：主要是由多个外部的高度动态的因素引起的，包括时间和空间因素，比如特殊节日或事件、新兴市场的变化、不稳定的政策[1]等。此外，由于基于运筹学的模型与未来的供需之间固有的相互依赖性，这种不确定性甚至可能更加恶化。特别是，未来的供需可能会被运筹学模型生成的行动计划所大幅影响，而运筹学模型又很大程度上依赖于未来的供需。于是，未来供需的不确定性大大增加了准确供需预测的难度，使得传统的多阶段运筹学方法的性能受到影响。

非凸形式的复杂业务约束：真实的物流场景中往往存在各种重要但复杂的商业规则，比如动态阶梯定价、贸易管制规则等。一方面，尽管这些商业规则可以被描述和模拟，但却很难仅用线性的或凸的约束来表达，因此使用传统的基于运筹学的方法（如线性规划和凸优化）来精确地建模和解决问题变得非常困难。另一方面，忽略这些必要的约束是不可接受的，因为它会导致模型和现实场景之间出现巨大的差距，导致显着的性能下降，甚至是产生不可行的执行计划。为了适应实际场景中的各种复杂业务约束，物流业不得不想出各种办法为运筹学的模型打上各种“补丁”，使得实际

生产环境中的运筹学模型变得庞大、臃肿而脆弱，在性能、效率、灵活性和可维护性上都难以令人满意。

运输网络的高复杂性：真实物流场景中的运输网络通常非常复杂，包括各种类型的站点和复杂的连接线路，因此在构建有效的基于运筹学的模型时，站点之间的复杂依赖性成为另一个重要挑战。具体而言，为了使用运筹学方法建模这些复杂的依赖关系，我们所需的约束和变量的数量会极大，使得在可接受的时间内求解优化模型中的个体和集体目标变得非常困难。

运筹学对单个调度操作的低解释性：实际的物流场景依然比以上的描述更为复杂，因此，物流领域依然依赖于调度员这一人工职位对各种实际信息进行综合，进行最终的实际调度操作，而将算法生成的调度方案作为辅助性的重要参考。然而，运筹学方法基于全局大尺度时空范围优化的特性使得其所生成的单一调度操作往往难以被人类可理解的方式解释，换言之，调度员仅知道模型所做出的调度决策，却难以理解其决策背后的逻辑动机。这使得其生成的调度方案对调度员的参考价值有所折扣。

随着以深度学习为代表的人工智能领域的迅猛发展，大量交通物流领域的国内外龙头企业，如滴滴出行[3]、摩拜单车[2]和OOCL①，都开始探索基于现代人工智能技术的物流资源调度的新方法。

1.2 本文的主要贡献

为了应对运筹学方法面临的上述挑战，本文结合多智能体强化学习及图神经网络，提出了一种高性能、可解释的物流网络资源平衡调度方法。

首先，我们将复杂物流网络中的资源平衡问题形式化为一个随机博弈（Stochastic Game），然后提出一种新的合作的多智能体强化学习（Multi-Agent Reinforcement Learning, MARL）框架。通过智能体集合、联合动作空间、状态集、奖励函数、转移概率函数和折扣因子的特别设计，我们的多智能体强化学习框架提供了一个端到端的高效能解决方案。其不仅可以补偿不完美的预测结果，以避免多阶段运筹学方法中的误差放大，而且还能够根据实际业务规则优化执行计划，以实现复杂的约束。

然后，相比于在一些较为简单的物流方案下应用MARL而言，盲目使用强化学习方法由于无法加强高度依赖的智能体之间的合作，可能无法在复杂的物流网络中产生令人满意的结果。为了应对这一挑战，我们进一步引入三个层次的合作指标，并相应地改进状态和奖励设计，以更好地促进复杂物流网络中智能体的合作。

最后，为了让模型动态学习智能体间的协作关系，并以可理解的方式对外表征这种协作关系，我们将图神经网络和注意力机制的方法引入多智能体强化学习的框架

① OOCL and MSRA Embrace AI in Digital Transformation https://www.oocl.com/eng/pressandmedia/pressreleases/2018/Pages/23apr18.aspx

内，增强了模型的智能性和可解释性。

为了测试MARL框架的性能并进行比较，我们在复杂的海洋运输网络中的空集装箱调度（ECR）任务下测试了我们的方法。海洋物流运输对于世界经济至关重要，80%的国际贸易是通过海运进行的[4]。到目前为止，海运是在全球范围内运输大宗商品和原材料的最具成本效益的方式。而空集装箱作为海运物流的关键资源，其调度的有效性直接影响海洋物流运输的效率。大量实验表明，我们的方法可以实现近乎最优的资源调度结果，与传统的运筹学对比方法相比，在性能上可以获得显着的改进。

我们的主要贡献可归纳如下：

- 将复杂运输网络中的资源平衡问题形式化为随机博弈；
- 提出一个合作的多智能体强化学习框架，为复杂物流网络中的资源平衡问题提供端到端和高性能的解决方案。其不仅对强不确定性的供需预测更加稳健，而且与传统的多阶段运筹学方法相比具备更高的性能和灵活性；
- 提出三个层次的合作指标，为改善状态和奖励的设计提供指导，以更好地促进复杂物流网络中的合作；
- 将图神经网络和注意力机制的方法引入多智能体强化学习的框架内，增强了模型的智能性和可解释性；
- 在现实海洋物流行业的场景中的空集装箱调度任务上进行了广泛的实验，证实了模型性能的优越性。

相关研究成果已发表在多智能体领域顶级会议AAMAS-2019[5]。

1.3　本文的组织方式

本文共分五章，具体安排如下：

第一章介绍了传统运筹学方法在现代复杂物流网络中面临的挑战，对本文提出的新方法进行了简介，并介绍了论文的结构安排。

第二章介绍了物流网络资源调度问题的当前进展，并对本文所使用的的技术，如多智能体强化学习和图神经网络等进行了介绍。

第三章提出了基于多智能体强化学习的物流网络资源调度算法，在对物流网络中的资源平衡调度问题进行了形式化定义后，将这一问题进一步形式化为随机博弈，从而引入MARL方法，并介绍促进智能体间合作的三个层次的合作指标。在实验部分，首先介绍了空集装箱调度任务（ECR）及其重要性，然后对包括运筹学方法的多种方法进行了性能比较和结果分析。

第四章提出了结合图神经网络和注意力机制的可解释物流网络资源调度算法。首先介绍了模型可解释性的概念及其重要性，然后介绍将物流网络分解成同构的协作网

络的方法，并由此引入图神经网络及注意力机制方法。在实验部分通过对学习到的特征进行分析来说明模型的可解释性，并同样进行了性能对比。

第五章对全文工作进行了总结与展望，概括了研究成果，并说明了后续研究的若干方向。

第二章　国内外研究现状

本章介绍传统运筹学方法及强化学习方法在网络调度与资源平衡问题上已取得的一些成果及其局限性。

2.1　物流网络中的调度问题与运筹学

物流网络中的调度问题（如本工作中重点关注的资源平衡问题）作为调度问题的一个分支，在运筹学（Operational Research, OR）领域已有全面的研究 [6–9]。在后文中我们重点使用了真实海运物流场景中的空集装箱调度问题（Empty Container Repositioning，ECR）作为实验任务（详见3.3.1节）。在这一问题上，[9] 提出了一个物流优化系统，通过基于需求预测和安全库存控制的多类别网络流模型来优化海运物流网络的供需。[1] 提供了运筹学领域在ECR问题上进展的详细综述。然而，正如1.1节所述，未来供需的严重不确定性、非凸的复杂业务约束、运输线路的高度复杂性以及运筹学对单一操作的低解释性使得传统运筹学方法在复杂的现代物流网络中遭遇困境。

2.2　强化学习

2.2.1　深度强化学习

随着深度学习的迅猛发展，深度Q学习（Deep Q-Learning）[10] 等深度强化学习方法在建模和解决许多智力挑战性问题方面取得了巨大成功，例如视频游戏 [10] 和围棋 [11]。然而，这些方法并没有广泛应用于复杂的现实世界应用，特别是那些具有高维动作空间并需要多个智能体之间合作的情况。

近年来，在深度强化学习巨大成功的推动下，已经提出了一些基于强化学习的方法来解决资源平衡问题，尤其是重新平衡同质的、灵活的车辆。[2] 提出了一种深度强化学习算法来解决共享单车的重新平衡问题，该算法学习了一种定价策略来激励用户重新平衡环境内的共享单车。[3] 提出了一个上下文多智能体强化学习框架，以解决在线乘车共享平台的再平衡问题，其中每辆出租车都被视为一个网格世界中的智能体，学习移动到邻近格子的动作。[12] 提出了一种网约车平台下的学习和规划方法，其结合了强化学习技术和组合优化算法，前者用于学习，后者用于规划。以上工作成功地建模并处理了大规模和现实世界的交通场景。然而，与本文中复杂物流网络中的资源平衡问题相比，上述工作的场景中环境更加宽松，智能体的依赖性简单明了。因此，这些方法很难应用于解决资源平衡问题。

2.2.2 多智能体强化学习

多智能体强化学习（Multi-Agent Reinforcement Learning，MARL）是多智能体系统（Multi-Agent System，MAS）与强化学习结合的一个重要的研究方向，研究当多个智能体在同一个环境中交互的情况下，使用强化学习方法学习合作或竞争或两者兼有的模式。

要在资源平衡问题中应用多智能体强化学习，一个主要障碍是处理智能体协作间的复杂依赖性，这种依赖性主要是由复杂的物流网络结构引起的。在经典的多智能体系统领域，多智能体之间的协作已经有一些富有成效的工作。比如FF-Q [13], Nash-Q [14] 和 Correlated-Q [15] 都是实现收敛和最优化的著名方法。然而，所有这些方法都采用联合动作（Joint Action）。由于具有大量智能体的现实多智能体系统往往联合动作空间极大，这些方法很难应用于这样的场景中。类似的限制发生在其他基于联合动作或最佳响应（Best Response）的方法 [16, 17]。其他一些工作 [18–20] 在MARL中应用基于势能的奖励塑造（Potential-based Reward Shaping）来刺激合作。这些方法可以在自己的场景中实现性能提升。但是，在资源平衡场景中，智能体的行为对最终结果具有长期和不可估量的影响，使得我们需要去更深入地理解问题并设计奖励。

2.3 图神经网络

近年来，深度学习方法开始越来越多地应用到图数据领域。以下对本文涉及到的一些图神经网络的方法进行简要介绍，更详细的综述可见[21]和[22]。

2.3.1 图卷积网络

图卷积网络（Graph Convolutional Network，GCN）[23] 。是在图像上已有成功运用的卷积操作在图结构上的扩展，通过将当前节点的特征与其邻居节点的特征进行整合和变换来生成当前节点的新的表示①。具体而言，设$A \in \mathbb{R}^{N\times N}$是归一化的图邻接矩阵表示，$H \in \mathbb{R}^{N\times D}$是图的每个顶点的M维特征（每一行为D维向量），则典型的图卷积操作可定义为：

$$H' = f(H, A) = \sigma(AHW) \tag{2.1}$$

其中$W \in \mathbb{R}^{D\times F}$是待训练的权值矩阵，$\sigma$是非线性的激活函数，$H'$是图卷积操作生成的下一层的特征矩阵（每一行为F维向量）。可见，这样的结构支持多层串联，将当前层输出的H'作为下一层的H输入即可。

① 常用的归一化方式如$A' = D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}}$，其中$D$为顶点的度的对角矩阵，$I$为单位矩阵。

2.3.2　图注意力网络

图注意力网络（Graph Attention Network，GAT）[24] 将注意力机制加入到图卷积的过程中。在图卷积操作中，顶点的特征只能按照邻接矩阵中固定的权值进行加权，而在图注意力网络中，两个具有连边的顶点之间的权值通过对两个顶点的特征应用注意力机制进行计算。具体而言，若h_i和h_j是顶点i和j的特征，则这两个顶点之间的注意力权值通过共享的注意力机制

$$e_{ij} = \begin{cases} a(Wh_i, Wh_j), & j \in \mathcal{N}_i, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

进行计算，其中W是待训练的权值矩阵，$\mathcal{N}_i$是顶点i的邻居集合。通过仅对邻居顶点计算注意力系数，网络的结构信息被加入到模型之中。这里共享的注意力机制a可以是一个单层的前馈神经网络

$$a(x, y) = \sigma(\mathbf{a}^T[x \| y]) \tag{2.3}$$

其中$\|$是连接运算符，$\mathbf{a}$是待训练的权重向量，σ是非线性的激活函数。然后，对注意力机制进行归一化：

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})} \tag{2.4}$$

最后，将$A' = [\alpha_{ij}]_{N \times N}$作为新的图邻接矩阵进行图卷积即可，注意这里的图卷积复用了前面的权值矩阵W。

第三章　基于多智能体强化学习的物流网络资源调度算法

3.1　物流网络资源平衡问题的形式化定义

在本节中，我们形式化地定义复杂物流网络中的资源平衡问题。

典型的物流网络可以定义为$G = (P, R, V)$，其中 P、R 和 V 分别代表站点，线路和运输工具的集合。进一步来说:

- P 中的每个元素 P_i 代表一个可以存储资源并生成相应需求和供给的站点。我们将 P_i 的库存初始资源量表示为 C_i^0 ，使用 C_i^t，D_i^t 和 S_i^t（$t = 1 \cdots T$）来分别代表不同时间的库存量，资源需求量和供应量；
- R 中的每条线路 R_i 是物流网络中的一个闭合回路，由一系列连续站点 $\{P_{i_1}, P_{i_2}, \cdots, P_{i_{|R_i|}}\}$ 组成，其中 $|R_i|$ 是线路 R_i 的总站点数，$P_{i_{|R_i|}}$ 的下一个目的地是 P_{i_1}。每条线路都可以与网络中的其他线路相交；
- 在每条线路 R_i 上，有一组固定的运输工具 $V_{R_i} \subseteq V$，其中每一个 $V_j \in V_{R_i}$ 都具有初始位置、时长函数 $d_j(P_u, P_v) : P \times P \rightarrow N^+$（给定起始站点 P_u 和目的地站点 P_v ，输出运输时间）以及容量 Cap_j^t（该运输工具可以装载的最大资源数）。当运输工具到达站点时，它可以从站点加载资源或将其资源卸到站点。

资源平衡的目标是最小化所有站点的资源短缺量之和。在特定时间 t ，站点只能使用前一天的资源存量，即 C_i^{t-1} 来满足当天的需求 D_i^t。[①]一旦库存不足，就会出现短缺。由此，我们将资源短缺量表示为 $L_i^t = \max\left(D_i^t - C_i^{t-1}, 0\right)$. 由此，所有站点的资源短缺量之和可以表示为：

$$\min L = \sum_{P_i \in P, t \in T} L_i^t \tag{3.1}$$

在当前的需求处理完之后，新的资源供应和从运输工具上卸下的的资源将被添加到当前站点的库存中，因此我们可以将新库存量表示为 $C_i^t = \max\left(C_i^{t-1} - D_i^t, 0\right) + S_i^t - \sum_{j=1}^{|V|} I(i, j, t) x_j^t$ ，其中 $x_j^t \in N$ 表示在 t 时刻加载到运输工具 V_j 的资源数量。x_j^t 可以是负数，以表示从运输工具中卸下的资源量，指示函数 $I(i, j, t)$ 定义为：

① 这是因为在现实环境中，当天新获得的供给资源或者从运输工具上卸下的资源往往因为内部调度和维护流程的原因而无法即时用于满足需求。这种逻辑可以随着特定的应用场景而改变，并且不会影响我们的算法框架。

$$I(i,j,t)=\begin{cases}1, & \text{运输工具 } V_j \text{ 在 } t \text{ 时刻到达站点 } P_i\\ 0, & \text{其他.}\end{cases} \tag{3.2}$$

我们进一步将 $C_{V,j}^t$ 定义为 t 时刻运输工具 V_j 的资源量。显然，$C_{V,j}^t = C_{V,j}^{t-1} + x_j^t$。

3.2　一种合作的多智能体强化学习框架

如前所述，传统的资源平衡解决方案采用供给和需求预测以及组合优化方法。然而，供给与需求的不确定性，复杂的业务约束和运输网络的高度复杂性使得这些方法遭遇困境。为了解决这些问题，在本节中，我们首先将复杂物流网络中的资源平衡建模为随机博弈，然后提出一种新的合作多智能体强化学习框架来解决该问题。

3.2.1　随机博弈视角下的资源平衡问题

资源平衡问题可以形式化地建模为随机博弈 $\mathcal{G}=(N,\mathcal{A},\mathcal{S},\mathcal{R},\mathcal{P},\gamma)$，其中 N 是智能体集合、$\mathcal{A}$是联合动作空间、$\mathcal{S}$ 是状态集、$\mathcal{R}$ 是奖励函数、$\mathcal{P}$ 是转移概率函数、γ 是折扣因子。具体如下：

- **智能体集合** N。我们将每个运输工具定义为一个智能体，这样有两个主要的优势：（1）当每个运输工具的智能体沿着特定线路连续循环行驶时，它可以感知到整个线路内的更大范围的信息，以便对自身进行优化从而最大化其自身的奖励，即最小化资源短缺，这对整条线路都有益处。（2）由于沿同一线路行驶的多个运输工具智能体通常共享相似的环境，因此它们可以自然地共享相同的策略，以显着降低多智能体强化学习中的模型复杂性，从而促进学习过程。
- **联合动作空间**$\mathcal{A}$。我们将运输工具智能体 V_j 到达站点 P_i 时装载或卸载资源的过程定义为动作。与 [25] 类似，我们采用事件驱动强化学习的理念。更具体地说，我们将每次智能体到达站点视为触发事件，并且只有在触发事件发生时，智能体才需要采取行动。在此事件驱动的设置下，我们使用 a_j^t 来表示智能体 $N_j \in N$ 在 t 时刻的到达事件中所采取的操作。对于智能体 N_j，我们将其操作空间定义为 $A_j=[-1,1]$，其中 $a_j^t \in [-1,0)$ 表示从运输工具上以 a_j^t 的比例卸下一部分资源，$a_j^t \in (0,1]$ 表示以 a_j^t 的比例将资源装载到运输工具上，$a_j^t=0$ 表示不进行任何装载或者卸载操作。然后，联合动作空间定义为 $\mathcal{A}=A_1\times A_2\times\cdots\times A_{|N|}$，其中$|N|$是智能体的数量。在 t 时刻可以卸下或装载的资源总量通常是由 C_i^t、Cap_i^t、$C_{V\text{fij}}^t$ 以及一些其他外部因素限制所共同决定的。具体由领域相关的商业逻辑控制。
- **状态集** $\mathcal{S}$。状态 $\mathcal{S}$ 是一个有限集，代表整个物流网络的所有可能情况。注意，

从实际的观点来看，由于极大的状态空间和由无关信息引入的潜在噪声，因此智能体不必基于整个状态信息采取行动。我们将在本节后面详细介绍实际的状态设计。

- **奖励函数** $\mathcal{R}$。资源平衡问题的目标是最小化所有站点的累积资源短缺量。对于每个单独的动作，即在站点处装载或卸下一些资源，其影响往往在其后续时段才表现出来。为了模拟这种延迟奖励，通常可以利用奖励塑造（Reward Shaping）来指导学习过程 [26]，其中的一种典型方式是衡量做或不做此行为的最终累积短缺的差异。但是，这种奖励在实践中很难计算。因此，我们找到了其他更现实的奖励塑造方法，这也将在本节后面的部分讨论。
- **转移概率函数** $\mathcal{P}$。其定义为映射 $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$，可以通过 S, R, V 的定义以及特定物流网络中供给和需求的潜在分布来指定。

3.2.2 加入合作度量的状态与奖励设计

在将资源平衡问题制定为随机游戏之后，我们考虑如何将MARL方法应用于实际问题，这时，我们就需要对状态和行动的奖励进行具体的设计，以促进合作并提高性能。依据智能体的合作感知范围，我们建立了三个层次的合作指标：自我感知、区域感知和区域合作。自我感知的智能体是完全自私和短视的，只考虑即时信息和利益。具有区域感知能力的智能体则具有更广阔的视野，并根据属于其区域的信息做出决定。最后，具有区域合作能力的智能体的视野可以超越他们自己所在的线路，并通过与相邻线路上的智能体进行协作，来达到更加全局意义上的资源平衡，使得资源可以从资源充足的线路流向资源缺乏的线路。

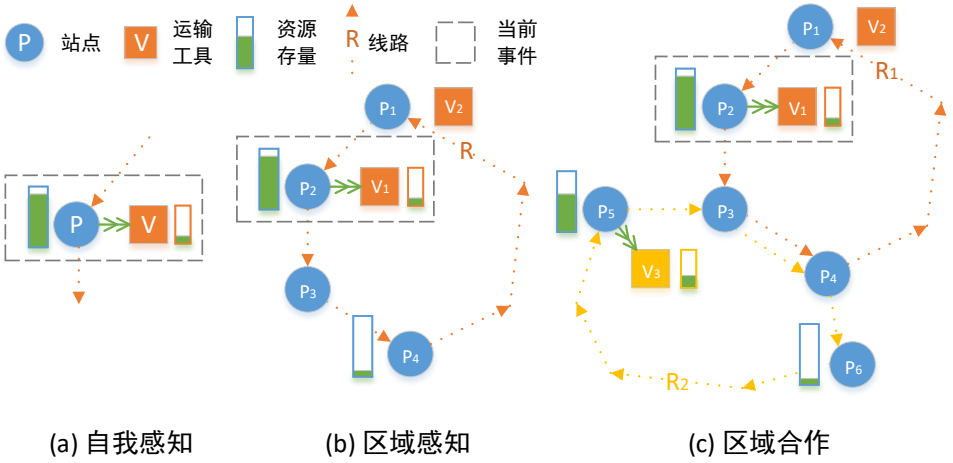
3.2.2.1 自我感知的智能体

当智能体 V_j 到达站点 P_i 时，V_j 的一个很自然的选择是根据自身和站点 P_i 的信息做出决策。关于此行动的奖励，一个直接的指标是考虑在下一个运输工具智能体到达 P_i 之前是否会发生短缺。显然，这是一个非常短视的智能体。

假设运输工具智能体V_j的第 k 次到达事件的时间是 t_k，到达站点是 P_i。站点 P_i 的状态 $s_{P,i}^{t_k}$ 可以表示如下：

- 当前站点的可用资源量 $C_i^{t_k}$.
- 当前站点可用资源量的历史信息 $\phi\left(C_i^1, \cdots, C_i^{t_k-1}\right)$ 以及资源短缺量的历史信息 $\psi\left(L_i^1, \cdots, L_i^{t_k-1}\right)$.
- 其他领域相关的信息，比如站点的编号值、停泊时间长短等。

其中 $\phi(\cdot)$ 和 $\psi(\cdot)$ 表示一些统计函数（Mean，Median等）或更高级的序列数据处理模型（如RNN）。具体实现应取决于应用场景。



(a) 自我感知　　(b) 区域感知　　(c) 区域合作

图 3.1　三种合作度量的图示。(a) 自我感知的智能体 V 仅考虑 (P, V) 的信息来做出决策；(b) 区域感知的智能体 V_1 基于其所在区域内的信息作出决策。当其注意到其所在线路 R 中的港口 P_4 资源存量低时，能够在当前港口 P_2 装入更多的资源；(c) 区域合作的智能体 V_1 可以观察到超出其所在线路的范围。当其注意到相邻线路 R_2 上的站点 P_6 需要支援时，其可以在当前站点 P_2 载入更多的资源并随后在中转站点 P_3 或 P_4 卸下。

对于运输工具 V_j，状态 $s_{V,j}^{t_k}$ 可以包括：

- 当前运输工具装载的可用资源量 $C_j^{t_k}$.
- 当前运输工具的空余空间大小 $Cap_j^{t_k} - C_j^{t_k}$.
- 其他领域相关的信息，比如运输工具的ID、类型等。

整合上述信息，我们得到自我感知智能体的状态 $s_I = \left[s_{P,i}^{t_k}, s_{V,j}^{t_k}\right]$。

自我感知的智能体只关心从 t_k 到 t_k' 的时间间隔中是否发生短缺，其中 $t_k' \geq t_k$ 代表下一辆车到达 P_i 的时间。此外，受传统方法中安全库存（Safety Stock）的启发，如果不发生短缺，我们会增加一个小的正奖励。此奖励根据函数 $f : N \to R$ 计算，具有递减的边际收益②。目的是鼓励智能体在站点上预备一些（具有上限的）安全库存。综上，奖励可以表示如下：

$$r_I = f\left(C_i^{t_k'}\right) - g\left(\sum_{t=t_k}^{t_k'} L_i^t\right), \tag{3.3}$$

其中 $g : N \to R$ 是由总资源短缺量而定义的损失。

3.2.2.2　区域感知的智能体

根据问题的定义，运输工具需要在特定的线路上行驶，并且对平衡自己区域内（也就是线路中的站点）的供给与需求负责。而如果智能体只能进行自我感知，在调度中不考虑其线路中的其他站点和运输工具的话，则不能很好地平衡其线路内的资源

② 例如，$f(x) = \sum_{i=0}^{x} \beta^i$ for $0 < \beta < 1$。

供给与需求。因此，我们引入了区域感知的智能体，以最大限度地减少智能体所在线路中所有站点的总的资源短缺量。具体来说，对于线路 R_q 上的智能体 V_j ，我们希望智能体获取线路上邻近环境的准确信息，这写邻近的信息更有可能影响当前的决策。由此，我们在状态中添加额外的邻近信息如下：

- n 个后继站点的信息 $\{s^{t_k}_{T,i'}|P_{i'} \in \mathrm{Sc}_{i,j}(n)\}$ 。其中 $\mathrm{Sc}_{i,j}(n)$ 是运输工具 V_j 经过站点 P_i 后接下来将要访问的 n 个站点的集合；
- m 个后继运输工具的信息 $\{s^{t_k}_{V,j'}|V_{j'} \in \mathrm{Fu}_{i,j}(m)\}$ 。其中 $\mathrm{Fu}_{i,j}(m)$ 是站点 P_i 在 V_j 到达后接下来将要到达的 m 个运输工具的集合。

我们可以看到，n和m越大，可用于决策的信息就越多。然而，在实践中，我们通常设置较小的 n 和 m 值来控制由不重要信息而引入的模型复杂性和噪声。为了弥补潜在的信息丢失，我们为线路 R_q 引入了整体统计区域信息 $s^{t_k}_{R,q}$，包括：

- 线路 R_q 上所有站点的可用资源量信息 $\Phi\left(\left\{C^{t_k}_i|P_i \in R_q\right\}\right)$
- 线路 R_q 上所有站点的资源短缺量信息 $\Psi\left(\left\{\psi\left(L^1_i,\cdots,L^{t_k-1}_i\right)|P_i \in R_q\right\}\right)$

和 $\phi(\cdot)$ 与 $\psi(\cdot)$ 类似, $\Phi(\cdot)$ 和 $\Psi(\cdot)$ 是基于序列数据的统计函数或模型。

我们将上述所有信息与 s_I 整合起来，以获得区域状态 s_T。区域感知的智能体将根据状态 s_T 做出决定。

3.2.2.3 区域协作的智能体

在真实的物流网络中，不平衡的情况也可能发生在不同的线路上：可能某些线路的供应很多，需求很少，而其他一些线路则相反，需求量大，供不应求。在这种情况下，只尝试在单一线路的范围内平衡供给与需求是不够的。为了解决这个问题，智能体应该学习协作，通过与相邻线路的智能体合作来解决不平衡的问题。

为此，我们需要考虑有关相邻线路的更多信息。假设一个事件为 (P_i, V_j, R_q)，并令 CR_q 表示与线路 R_q 有公共站点的相邻线路的集合。首先，需要加入所有相邻线路的统计信息 $\Phi_r\left(\left\{s^{t_k}_{R,p}|R_p \in \mathrm{CR}_q\right\}\right)$ 来表征相邻线路的总体情况。此外，我们在智能体到达转运站点（即不止一条线路通过的站点）时添加额外信息，即 $\Phi_n\left(\left\{s^{t_k}_{R,p}|R_p \in \mathrm{RT}_i\right\}\right)$ ，其中 RT_i 是通过站点 P_i 的线路集合。我们将上述所有信息与 s_T 整合，得到协作状态 s_D。

为了鼓励合作，我们通过考虑相邻线路的资源短缺量来扩大奖励。对于线路 R_q 上的智能体 V_j，其行为不仅会影响其自身线路上的奖励，还会影响 CR_q 中邻近线路中的智能体的奖励，尤其是在线路相交的转运站点。将相邻线路加入奖励设计的考虑后，我们定义：

$$r_D = \alpha r_I + (1-\alpha) r_C \tag{3.4}$$

其中α是一个软超参数，r_C 定义如下：

$$\begin{aligned} r_C = & f\left(\xi_1\left(\left\{C_i^{t_k'} | P_i \in R_p, R_p \in \mathrm{CR}_q\right\}\right)\right) \\ & - g\left(\xi_2\left(\left\{L_i^t | t_k \leq t \leq t_k', P_i \in R_p, R_p \in \mathrm{CR}_q\right\}\right)\right) \end{aligned} \tag{3.5}$$

其中 $\xi_1(\cdot)$ 和 $\xi_2(\cdot)$ 是统计函数或模型。

三个层次的合作度量如图3.1所示。算法1展示了资源平衡问题的整个协作MARL框架。从第4行到第13行，智能体通过函数调用与环境交互，并收集转移记录。需要强调的是，GetState $\left(S_{j,k}, P_i \text{fi} V_j\right)$ 是指基于当前事件 (P_i, V_j) 和全局环境快照 $S_{j,k}$ 构造状态的过程。此快照包含触发事件时环境的完整信息。GetDelayedReward $\left(S_{j,k-1}, S_{j,k}\right)$ 是指基于这两个快照之间发生的短缺来计算延迟奖励的过程。GetState$(\cdot)$ 和 GetDelayedReward$(\cdot)$ 的详细实现将基于所采用的合作度量级别来确定。

Algorithm 1 合作的多智能体强化学习框架

1: 对每个智能体 V_j 初始化经验回放池（Replay Memory） D_j ，大小为 M
2: 对每个智能体 V_j 初始化动作-值函数 Q_j ，随机初始化函数的权值 θ_j
3: **for** episode ← 1 to MAX **do**
4: ResetEnvironment()
5: **while** 环境运行尚未结束 **do**
6: // k 表示智能体 V_j 的第 k个事件
7: $(P_i, V_j, k) \leftarrow$ WaitingEvent()
8: $S_{j,k} \leftarrow$ GetEnvironmentSnapshot()
9: $s_k \leftarrow$ GetState $\left(S_{j,k}, P_i, V_j\right)$
10: $r_{k-1} \leftarrow$ GetDelayedReward $\left(S_{j,k-1}, S_{j,k}\right)$
11: StoreExperience$\left(D_j, (s_{k-1}, a_{k-1}, r_{k-1}, s_k)\right)$
12: $a_k \leftarrow \epsilon$-Greedy$\left(\arg\max_a Q_j(s_k, a)\right)$
13: Execute(P_i, V_j, a_k)
14: **end while**
15: **for** $l \leftarrow$ 1 to MAX-TRAIN **do**
16: **for** each V_j in V **do**
17: 从 D_j 中采样一个批次的数据 (s, a, r, s')
18: 计算 $y \leftarrow r + \gamma \max_{a'} Q_j(s', a'; \theta_j)$
19: 更新智能体 V_j 的Q网络
$\theta_j \leftarrow \theta_j - \nabla_{\theta_j}(y - Q_j(s, a; \theta_j))^2$
20: **end for**
21: **end for**
22: **end for**

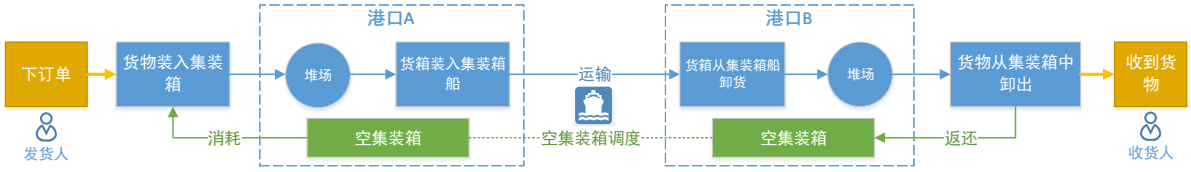


图 3.2　ECR问题中的集装箱运输链。蓝线表示货箱流量，绿线表示空箱流量。所有流程都在实际物流方案中特定业务逻辑的控制之下。

3.3　实验

为了评估我们提出的方法的有效性，我们在海运集装箱运输场景中进行了资源平衡调度实验。在此场景中，资源平衡主要对应于空集装箱调度（Empty Container Repositioning, ECR）问题。本节将首先介绍ECR的背景，然后在真实海洋物流网络的一部分上展示实验结果。

3.3.1　ECR问题

集装箱是海运物流行业的基石。在海洋物流的场景下，资源平衡问题对应于空集装箱调度问题（Empty Container Repositioning，ECR）。ECR的目标是通过在海洋物流网络内的预定线路上航行的集装箱船来对空集装箱进行重新调度，以满足港口的动态运输需求。由于世界贸易的不平衡性，空集装箱的需求和供给非常不平等 [1]，中国等以出口为主的国家对空集装箱有大量需求，而美国等以进口为主的国家则容易囤积大量卸货后的空集装箱。因此，对空集装箱进行调度非常必要。根据 [27] 的统计，2009年海运空集装箱调度的估计成本约为200亿美元，空集装箱运输量为5000万，这表明了海运物流行业中ECR的必要性。

ECR问题可以自然地建模为前文定义的物流网络中的资源平衡问题，其中空集装箱即为需要平衡的资源。更具体地说，港口，集装箱船和预定的船舶航线分别对应于前文资源平衡问题中的站点 P，运输工具 V和线路 R。在 t 时刻的港口 P_i 对空箱的需求量和供应量分别对应于 D_i^t 和 S_i^t。

3.3.1.1　ECR问题的特点

尽管如此，ECR问题仍有一些自身的特点。在ECR问题中，外部需求量 D_i^t 和供应量 S_i^t 由运输订单集合 O 确定，运输订单集合 O 也是外部和动态的。O 中的每个订单 o 为四元组 (P_u, P_v, n, t_o)，分别表示出发港口，目的地港口，所需空箱的数量和下单时间。订单的集装箱运输链（Container Transportation Chain）可以描述如下，如图 3.2 所示：当订单 (P_u, P_v, n, t_o)在 t_o 时刻下单，出发港口 $D_u^{t_o}$ 的外部需求量将加上 n，这意味着 P_u 需要提供 n 个空箱以在 t_o 时刻接单。如果成功接单，货物将装入这些空箱，从而空箱转换成货箱，这些货箱将等待船只将其运至目的港口 P_v。当集装箱船 V_j 到达

出发港口 P_u 时，如果目的地港口 P_v 在 V_j 所属的线路 R_k 上，则该订单的货箱将装入该集装箱船。③当载货集装箱在 t'_o 时刻于目的地港口 P_v 卸下时，货箱中的货物也会被卸下，并在 $t'_o + t_{\text{ret}}$ 时刻作为空箱返还到 P_v，其中 t_{ret} 是一个常量。因此，目的地港口 P_v 在 $t'_o + t_{\text{ret}}$ 时刻的外部供给量 $S_v^{t'_o+t_{\text{ret}}}$ 也将增加 n。

ECR问题的特点总结如下：

- 空箱是可重用的，作为货物的集装箱在港口间循环流动；
- 货箱和空箱共享集装箱船。也就是说，集装箱船 V_j 的空余容量 Cap_j^t 会随着船上货箱的数量而动态变化；
- 如果在下订单的时候出发港口的空箱数量无法满足订单需求，则整个订单都会失败。对于一个单独订单 o 的资源短缺量 L_o 定义为

$$L_o = \begin{cases} n, & n > C_u^{t_o} \\ 0, & otherwise. \end{cases}$$

3.3.1.2　使用运筹学方法解决ECR问题的困难性

基于运筹学解决ECR问题的的第一个难点是未来需求-供给预测的不确定性。如前文所述，这种不确定性主要是由多种外部动态因素引起的，例如市场变化。这种不确定性还会因基于运筹学的模型与未来需求-供给预测之间固有的耦合关系（预测值影响模型的调度决策，模型的调度决策反过来又影响预测值）而加剧。一般基于运筹学的方法都需要基于未来一段长时间的需求-供给预测来生成调度计划，因此，中长期预测的严重不确定性以及模型与预测之间固有的耦合关系将导致基于运筹学的方法性能欠佳。

第二个主要困难是由集装箱运输链中的某些业务逻辑引起的。一个难以通过运筹学方法建模的典型且重要的业务逻辑是集装箱的状态变化，即从空箱到货箱，以及从货箱到空箱的过程。如果要精确地建模这一过程，基于运筹学的方法必须考虑集装箱的状态更改。然而，这在现实情况中是相当困难的，因为这些状态变化完全由航运公司的操作员控制，并且根据不同的客户群体和不同的地区而具有完全不同的规则。因此，作为ECR问题中的“黑盒”部分，这样的业务逻辑难以通过传统的运筹学方法精确建模。在现实世界的集装箱运输中，还存在更多复杂的商业逻辑，例如区域政策，同样难以通过运筹学的方法建模。

为了使用基于运筹学的方法，我们必须放宽相应的约束并采用一些近似方法，包括：

③ 不失一般性，我们只考虑非转运订单，即假定P_u 和 P_v 始终在同一条线路上。转运订单可以被看做多个单独的非转运订单。

- 空箱和货箱的运输过程是分离的。集装箱的状态变化预先确定，而不是由业务逻辑（实际场景中的非线性甚至“黑盒”过程）动态决定，通过分解未来的订单信息来简化未来的需求-供给预测。
- 不保持单个订单的原子性。在我们的场景中，如果剩余的空箱数量不足，即使差值非常小，整个订单也将全部失败。在运筹学模型中，由于订单被分解成了固定的需求-供给量，因此无法保证这一性质。

3.3.2 实验设置

在以下实验中，我们基于某商业公司的现实服务线路，提取亚洲、北美和欧洲之间主要的海运网络。该网络由4条线路，17个港口和31艘集装箱船组成。如图 3.3 所示：

- R1：太平洋-大西洋航线，94天，14艘集装箱船，线路为：欧盟、纽约、萨旺纳、洛杉矶、奥克兰、横滨、上海、神户、东京、奥克兰、洛杉矶、萨旺纳、纽约、欧盟；
- R2：中亚-南亚航线，共60天，9艘集装箱船，线路为：阿拉伯、新加坡、泰国、盐田、洛杉矶、奥克兰、上海、宁波、盐田、新加坡、阿拉伯；
- R3：日本-美国航线，33天，5艘集装箱船，线路为：神户、东京、洛杉矶、奥克兰、东京、神户；
- R4：日本-中国-新加坡航线，19天，3艘集装箱船，线路为：东京、神户、台湾、香港、蛇口、新加坡、蛇口、香港、台湾、东京。

所有船只均在线路上接近均匀分布，发船间隔约一周。在初始状态，根据某海洋物流公司的历史统计数据，在17个港口分布有3000个空箱，所有船只都是空的，没有任何空箱或货箱。在模拟环境中，基于同一公司提供的信息，所有17个港口的供给与需求分布如图 3.4所示。每艘船的容量为200个集装箱，即每艘船的货箱和空箱的总量不得超过200。为了帮助我们进行MARL方法的训练，我们基于来自商业海洋物流公司的真实历史数据建立了模拟的ECR环境。

为了衡量我们的方法的性能，我们使用*满足率*（Fulfillment Ratio）作为度量指标，即整个环境在所有时间步长中成功运输的集装箱数量与总集装箱运输请求数的比率（每次模拟环境运行400个时间步长，其中一个时间步长对应于现实中的一天）。在现实世界中，还有许多其他类型的集装箱调度成本，包括装载/卸载成本，存储成本等。然而，在所有这些成本中，由满足率衡量的（未满足订单而导致的）订单损失成本是占主导地位的，因为它将直接影响订单接受率，从而影响运输公司的收入和声誉。因此，我们在这项研究中专注于尽可能减少订单损失成本。实际上，MARL也可以通过奖励塑造和特别的动作空间设计来自然地建模其他类型的成本，这是我们未来的工作

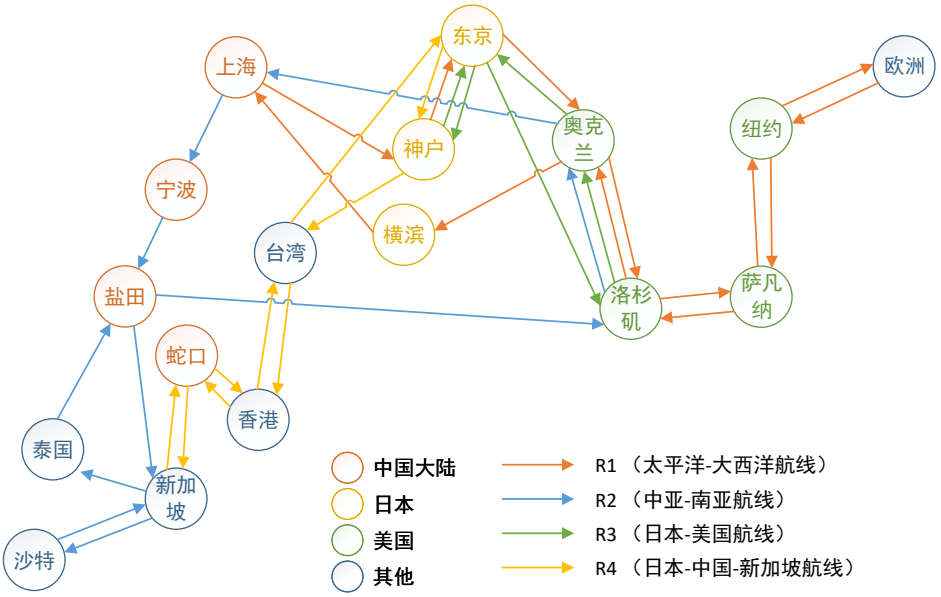


图 3.3　亚洲，北美和欧洲之间的主要海洋运输网络。

图 3.4　所有17个港口的需求和供给分布。

方向之一。

3.3.3　比较方法

在以下实验中，我们比较了ECR问题的以下方法：

- **无调度**：对空集装箱不做任何调度。集装箱的流动只取决于载货集装箱的运输。
- **基于规则的库存控制（Inventory Control，IC）**：根据库存管理理论，此方法在每个港口P_i基于供给和需求的历史信息设置两个库存阈值，安全阈值F_i^s和超额阈值F_i^e（$F_i^s \le F_i^e$）。当船只V_j在时段t到达P_i时，它将尝试通过装卸集装箱来使得空集装箱库存量C_i^t维持在$[F_i^s, F_i^e]$范围内。形式上的表述是：假设$x_{i,j}^t$是从港口P_i装载的集装箱数量（负值意味着向该港口卸货），其满足：

$$x_{i,j}^{t} = \begin{cases} \min(C_i^t - F_i^e, Cap_j^t - C_{V,j}^t, C_i^t), & C_i^t > F_i^e, \\ -\min(F_i^s - C_i^t, C_{V,j}^t), & C_i^t < F_i^s, \\ 0, & v. \end{cases}$$

- **在线线性规划（Linear Programming, LP）**：使用上面提到的一些近似方法，通过采用3.1节中的数学定义，可以使用线性规划建模ECR问题。不过，由于线性规划的模型经过了简化，使得其所生成的解很难直接使用。在这里，我们使用 [28] 中描述的滚动窗口策略（Rolling Horizon Policy）来解决这个问题：划每次在较长的时间窗口范围内生成空箱调度计划，但是只执行开头的部分计划，而抛弃后面的计划，执行完后再重新生成新的长时间窗口范围的计划，重复这一过程直到结束。这就是所谓的在线线性规划方法。请注意，我们提出的端到端的MARL方法直接与模拟的环境交互，而没有明确的预测阶段，因此，为了进行适当的比较，我们使用精确的未来订单信息来替换线性规划模型中预测未来需求的部分，以便消除导致预测不确定的外部因素的影响，这可以被视为一个相对理想的条件。有关在线线性规划的更多详细信息可以在附录中找到。由于3.3.1.2节提到的模型与预测的耦合关系，这里的未来需求仍然不是100%准确的，因此在后面的方法与库存控制方法相结合，以改善此部分的不确定性带来的问题。
- **带有库存控制的在线线性规划**：在此对比算法中，我们采用了 [9] 提出的方法，将线性规划模型与库存控制相结合。此方法根据供给与需求的历史信息为每个港口P_i设置安全阈值F_i^s，然后约束 $L_i^t = \max\left(D_i^t - (C_i^{t-1} - F_i^s), 0\right)$。
- **自我感知的MARL（SA-MARL）**：这是3.2.2.1节中描述的自我感知的MARL模型。对于站点（港口）状态$s_{P,i}^{t_k}$，$\phi(\cdot)$是平均函数，而$\psi(\cdot)$是求和函数。对于运输工具（船舶）状态$s_{V,i}^{t_k}$，我们将船上货箱的数量作为额外的领域相关的信息。至于奖励，我们设置$f(x) = 1 - 0.5^x$和$g(y) = 5y$，其中x和y的计算方法如公式(3.3)。
- **区域感知的MARL（TA-MARL）**：这是3.2.2.2节中描述的区域感知的MARL模型。对于后继的站点信息，m和n都设置为1. $s_{R,q}^{t_k}$中的$\Phi(\cdot)$和$\Psi(\cdot)$设置为平均函数。
- **区域合作的MARL（DA-MARL）**：这是3.2.2.3节中描述的区域合作的MARL模型。其中$\Phi_r(\cdot)$和$\Phi_n(\cdot)$为平均函数，式3.4中的$\alpha = 0.5$，$\xi_1(\cdot)$和$\xi_2(\cdot)$是2层平均函数 $\textsc{Avg}\{\textsc{Avg}\{\sum_{t=t_k}^{t'_k} L_i^t | P_i \in R_p\} | R_p \in \textsc{Cr}_q\}$。
- **离线线性规划（上界）**。在这种情况下，空箱资源的短缺量将直接作为优化目

表 3.1　不同空集装箱调度方法的性能比较

方法	满足率 (%)		
	80% 集装箱	100% 集装箱	150% 集装箱
无调度	26.58 ± 0.90	29.87 ± 0.85	38.25 ± 1.07
库存控制	58.30 ± 0.93	61.07 ± 0.98	68.63 ± 0.98
在线线性规划	76.28 ± 1.54	85.75 ± 1.34	94.48 ± 1.00
加入库存控制的在线线性规划	81.09 ± 1.21	88.99 ± 0.89	96.30 ± 0.80
SA-MARL	65.39 ± 1.20	72.04 ± 1.57	84.21 ± 1.45
TA-MARL	75.25 ± 1.38	83.48 ± 0.94	93.75 ± 0.69
DA-MARL	**82.04 ± 1.69**	**95.97 ± 0.63**	**97.70 ± 0.98**
离线线性规划（上界）	98.32 ± 0.60	98.95 ± 0.31	99.42 ± 0.25

表 3.2　DA-MARL中不同延迟参数 k 下的性能比较

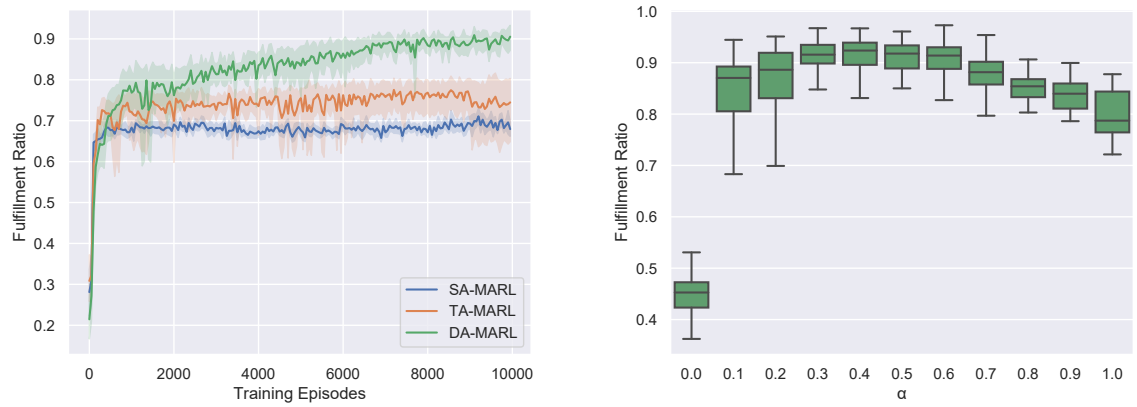
k	满足率	k	满足率
1	95.87 ± 0.65	20	94.52 ± 0.89
5	95.76 ± 0.67	30	93.23 ± 1.76
10	95.49 ± 0.65	40	90.39 ± 2.50
15	94.71 ± 0.93	50	85.87 ± 3.23

标，而使用线性规划模型求解。这一方法预先知道所有未来订单的信息，且不在模拟环境中运行。这可以看作是该问题的一个上界。也就是说，任何方法都不太可能达到比这更好的性能。

我们使用ϵ-greedy探索策略在所有MARL方法上训练10000个回合（episode）。ϵ在前8000个回合中从0.5到0.01线性退火，并在之后的2000个回合中固定为0.01。我们使用*Adam*优化器，设置学习率为10^{-4}。批次（batch）大小固定为32。同一线路中的所有智能体共享相同的Q网络，每个Q网络由2层的多层感知机（MLP）参数化，每层的神经元个数分别为16和16，由ReLU函数激活。由于DQN仅适用于离散动作空间，我们将连续动作空间$A_i = [-1, 1]$均匀地离散化为21个动作，即$A_i' = \{-1, -0.9, \cdots, 0.9, 1\}$。

3.3.4　实验结果分析

为了比较上述所有方法，我们在100个随机初始化的环境中运行我们训练好的模型以及对比方法。对于对比方法，我们使用网格搜索以找到合适的参数。对于MARL方法，我们训练模型10次，并选择表现最佳的模型。为了测试我们的框架所学习到的策略的稳健性，我们把在100%（3000）空集装箱设置下训练好的模型置于集装箱总量为80%（2400个集装箱）和150%（4500个集装箱）的环境下进行测试。实验结果总结在表 3.1中，其中我们报告了满足率的平均值和标准差。我们可以看到，DA-MARL方



(a) (b)

图 3.5 （a）MARL方法训练过程中的收敛性比较。 X轴是回合的数量。（b）区域合作MARL中不同α值的性能比较。 X轴是α。两个图中Y轴均为满足率。

法在所有初始集装箱数值的设定下都表现最佳。甚至TA-MARL方法也可与传统的在线线性规划方法相媲美。SA-MARL在我们的MARL方法中性能最差，但仍然优于基于规则的库存控制。该稳健性测试表明，智能体已经学会了有效的策略来应对剧烈的环境变化。经过训练的DA-MARL模型仍然比在线线性规划及其加入库存控制的版本表现更好，而后两者其实都是能够基于变化的环境进行决策的。

MARL方法的收敛性比较如图 3.5a所示。每种MARL方法均训练10次，我们报告训练期间的性能平均值和标准差。我们可以看到，所有MARL方法在前1000个回合都能非常快速地收敛。之后，DA-MARL比其他方法在训练中获得更大的性能提升。

在DA-MARL中，式3.4中的α是控制自身区域奖励和合作奖励之间比例的重要参数。我们使用不同的α训练模型，结果显示在图 3.5b中。由于时间限制，每个模型均训练了5次，每个训练过的模型都进行了100次测试。结果显示只有 r_I（$\alpha = 1$）和只有 r_C（$\alpha = 0$）单独作为奖励时的性能都不佳，而它们的组合（在我们的例子中为 $\alpha = 0.4$ ）对于实现更好的性能至关重要。

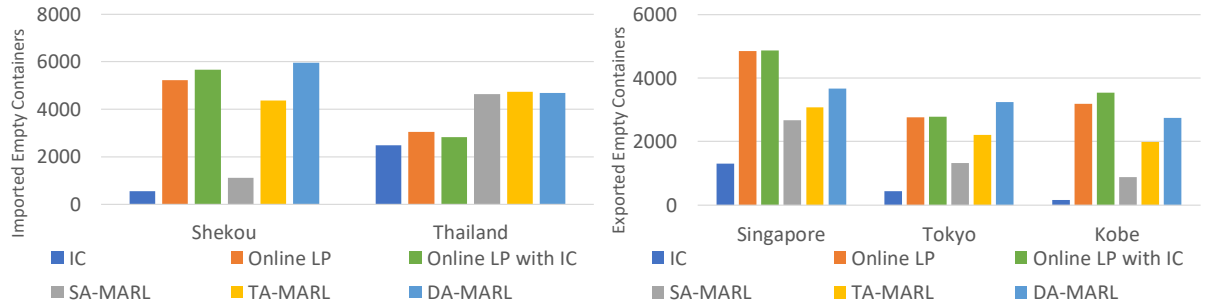
通信是在多智能体系统中达成合作的关键部分，在我们的区域合作MARL设计中，共享邻近线路和转运线路的信息 $\Phi_r(\cdot)$ 和 $\Phi_n(\cdot)$ 对于提高模型性能至关重要。但是，这些信息有可能无法在实际场景中实时传输，即智能体只能访问这些信息的过时版本。表 3.2显示所有智能体只能访问 k 天前的这些信息时的满足率。结果表明，当延迟在合理范围内（$k \leq 20$）时，我们提出的方法可以稳健地运行而没有明显的性能损失。

3.3.4.1 合作能力分析

ECR的主要目标是平衡供给与需求，以便最大限度地减少缺箱港口的空箱短缺量。图 3.6a显示了不同的方法在蛇口（位于深圳）和泰国两个主要的缺箱港口（需

要进口大量空集装箱以用于满足出口订单需求）的进口空箱量。从图 3.3来看，泰国是新加坡在R2航线上的下一个港口而新加坡有较富余的空箱，这意味着即使没有复杂的合作机制也不难获得空箱。但对于蛇口来说，一方面蛇口比泰国需要更多的空集装箱（如图 3.4所示），另一方面其所在的航线R4上，新加坡是唯一自身能够供给集装箱的港口，但也没有足够的集装箱来供应蛇口，所以相比于泰国而言蛇口的情况要困难得多。要使蛇口的空箱需求得以充分实现，只有利用东京和神户作为转运港口而从美国地区运输空箱，这需要地区之间的强大合作能力。图 3.6a显示，所有三种MARL方法在泰国都表现良好，而区域合作的MARL在蛇口明显优于所有其他方法，这表明我们的设计能够满足线路间合作的需求。

对于线路间的合作，转运港口出口空箱的数量非常关键，因为它是蛇口等缺箱港口获得空箱的来源。图 3.6b显示了新加坡，东京和神户的出口空箱数量，这三个港口是我们的环境中不同航线之间的三个主要转运港口。结果表明，随着MARL智能体的合作意识逐渐增强，转运港出口空箱数量也明显增加，这表明我们的合作设计是有效的。加入库存控制的在线线性规划方法作为纯全局优化的方法也可以在转运港口上运行良好。然而，线性规划模型如前分析的一些不利因素（比如和真实环境之间的差距）限制了其整体性能。



(a)　　(b)

图 3.6　（a）不同方法在蛇口和泰国两个主要的缺箱港口的进口空箱数比较。（b）不同方法在新加坡、东京和神户三个主要转运港口的出口空箱数比较。因为“不调度空箱”的方法顾名思义不会有任何空箱的进出口调度，因此在这里省略。

3.4　本章小结

在本章，我们首先将物流网络中的资源平衡问题建模为一个随机博弈过程。在此基础上，我们提出了一个协作的多智能体强化学习框架，其中根据智能体的协作意识范围确定三个级别的协作指标，从而使得运输更为高效。在模拟的海运物流网络环境上的充分实验表明，我们的新方法可以促进智能体之间的合作，并在性能和稳定性方面带来显着改善。

第四章　结合图神经网络和注意力机制的可解释物流网络资源调度算法

4.1　研究背景

在实际的生产环境中，很多时候我们不仅关注模型的性能，同时也关注模型的可解释性（Interpretability），尤其是当人工智能模型作为辅助决策的工具而存在的时候更为如此。我们希望了解模型从数据或者环境中学到了怎样的知识，从而做出了最终的决策。换言之，我们不仅想知道“怎么做”，还想知道“为什么这么做”。例如，假设我们在开发一个判断病人风险的模型，我们很可能不仅希望模型输出一个风险概率，还希望模型给出判断依据，即究竟是基于哪些考虑而得出了这一概率值。如果模型不具备可解释性从而不能给出这些额外的依据信息的话，在很多现实领域中的应用就会受到很大的限制。

在上一章，我们提出了一种基于多智能体强化学习的高性能物流网络资源调度算法。然而，无论是基于强化学习的方法还是基于运筹学的方法，都面临相同的问题，即模型的可解释性不强，具体体现在对于模型生成的一个单独的调度操作，我们很难了解模型究竟是出于怎样的考虑而进行了这样的调度。例如图4.1的情况，假设模型在此时为运输工具 V_1 选择了装载操作，那么仅凭我们朴素的观察，并无从了解模型做出这一操作的原因。

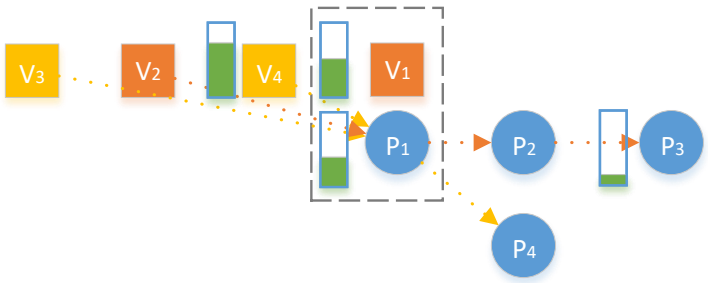


图 4.1　普通的资源调度算法，仅为当前事件生成动作，却没有任何信息帮助人们理解生成的动作。

在实际的物流场景中，物流公司依然依赖于调度员（Operator）这一重要职位来进行各种实际的调度工作 [1]，而各种调度模型给出的调度操作则作为调度员的重要参考。因此，可解释性在物流网络的场景下同样是十分重要的。如果调度员难以理解模型给出的调度操作的意图，可能会让其忽略一些模型正确捕捉到的关键因素，从而做出不够高效的调度操作。由此，我们希望能有一种如图4.2所示的模型，在给出调度操

作的同时，以人类可理解的方式给出当前调度操作的解释性信息（如物流网络中各个对象在当前调度中的权重），从而帮助我们理解模型的决策过程。

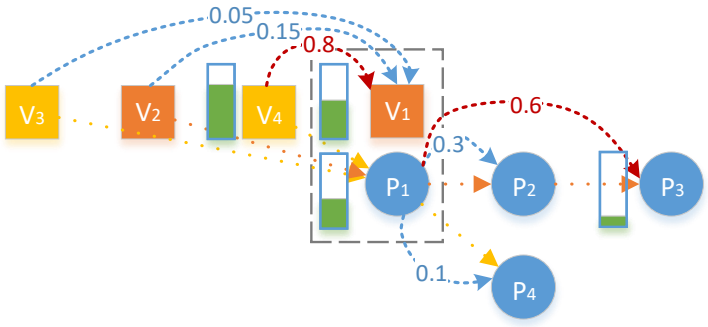


图 4.2　具有可解释性的资源调度算法，在为当前事件生成动作的同时，输出与当前事件相关的各对象在决策中的权重，帮助理解当前的调度决策。比如如果此时模型为运输工具 V_1 选择了装载操作，我们可以理解为模型观察到了后续运输工具 V_4 中的大量资源可用于补充站点 P_1，同时观察到 V_1 所在的线路上有资源缺乏的站点 P_3 亟需补充资源。

4.2　结合图神经网络和注意力机制的强化学习框架

在本节，为了改善现有方法解释性不强的问题，提出了一种结合图神经网络和注意力机制的强化学习框架。在前章提出的方法的基础上，本方法使用图神经网络和注意力机制动态学习物流网络中包括智能体在内的各对象的协作关系，并通过权重的方式对外表征，从而提升了模型的可解释性。

4.2.1　特征的定义

与3.2.2节类似，我们同样需要从物流网络中提取一系列特征用于模型的调度决策。然而，由于在本章我们希望由模型通过图神经网络和注意力机制动态学习物流网络中包括智能体在内的各对象的协作关系，所以我们不再加入上一章人工定义用于促进合作的特征作为状态，而是为物流网络的三个组成部分：运输工具、站点和线路单独定义特征，具体如下：

对t时刻的每个运输工具（智能体）V_i，定义以下特征，记为$h^t_{V,i}$：

- 空余空间比例；
- 空箱比例；
- 货箱比例。

对t时刻的每个站点 P_j，定义以下特征，记为$h^t_{P,j}$：

- 当前空箱数；
- 缺箱数；
- 历史平均空箱数；

- 下一艘船的时长距离；
- 经过的航线数量。

对t时刻的每条线路 R_k，定义以下特征，记为$h^t_{R,k}$：

- 平均空箱数；
- 历史平均空箱数；
- 历史平均缺箱数。

为了避免不同特征的尺度差异过大造成训练困难，对特征均进行了适当的缩放或归一化处理。

4.2.2 协作图的构建

为了将图神经网络方法应用于复杂的物流网络，我们首先将物流网络$G = (P, R, V)$在每个时刻t分解为三个有向加权的同构协作图，分别对应于运输工具、站点和路线。

- **运输工具协作图** $G^t_V = (V, A^t_V)$：G^t_V的顶点集是V。 A^t_V中每个弧$\langle \overrightarrow{V_i, V_j} \rangle$ 的权重表示当前时刻 t 的运输工具 V_j 要到达当前 V_i 所在的位置所需的天数。
- **站点协作图** $G^t_P = (P, A^t_P)$：G^t_P的顶点集是P。每个弧的权重$\langle \overrightarrow{P_i, P_j} \rangle \in A^t_P$表示当前时刻$t$时，位于站点$P_i$的运输工具到达站点$P_j$所需的天数。如果$P_i$中有多辆可以到达站点$P_j$的运输工具，我们累加每个运输工具所需天数的倒数，然后将累加值的倒数作为权值。即$(\sum_{V_k \in P_i} d^{-1}_{V_k})^{-1}$。
- **路线协作图** $G^t_R = (P, A^t_R)$：G^t_R的顶点集是R。 A^t_R中每个弧$\langle \overrightarrow{R_i, R_j} \rangle$ 的权重是指当前时刻t时，位于路线 R_i中的每个运输工具到达路线R_i和R_j的转运站点的天数。如果R_i中有多辆可以到达转运站点的运输工具，我们同样累加每个运输工具所需天数的倒数，然后将累加值的倒数作为权值。即，$(\sum_{V_k \in R_i} d^{-1}_{V_k})^{-1}$。

由于路线R均为环线，运输工具可以在其中无限循环，我们只考虑每个运输工具（智能体）从当前时刻开始走过K个站点的情况来构建这三个协作图。

4.2.3 注意力机制的加入

对于t时刻的每个协作图G^t_*和相应的顶点特征矩阵$H^t_* = [h^t_1, h^t_2, \cdots, h^t_N], h^t_i \in \mathbb{R}^{F_*}$，我们采用图注意机制来表示智能体决策过程中各个顶点的协作关系，然后通过图卷积对特征进行整合，生成每个顶点的上下文特征。

与忽略图的权重信息的标准图注意网络（GAT）[24]不同，我们不仅考虑顶点特征，也考虑顶点之间的权重，因为顶点之间的权重包含关于合作关系强度的重要信息（例如，更近距离的两个运输工具更有可能合作）。我们使用共享参数的注意机制

$a : \mathbb{R}^{F_*} \times \mathbb{R}^{F_*} \times \mathbb{R} \to \mathbb{R}$ 对顶点进行自我注意（self-attention），从而计算注意力系数：

$$e_{ij}^t = \begin{cases} a(h_i^t, h_j^t, \omega_*^t(\langle \overrightarrow{i,j} \rangle)), & j \in \mathcal{N}_i, \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

其中$\omega_*^t(\langle \overrightarrow{i,j} \rangle)$表示协作图$G_*^t$中弧$\langle \overrightarrow{i,j} \rangle$的权重，$\mathcal{N}_i$是顶点$i$的邻居集合。通过仅对邻居顶点计算注意力系数，网络的结构信息被加入到模型之中。注意在这里为了获得更高的模型可解释性，我们不对初始顶点特征应用线性变换，从而让注意机制a中的权重具有具体可解释的含义。

与[24]类似，我们使用由权重向量$\mathbf{a} \in \mathbb{R}^{2F_*+1}$参数化，由LeakyReLU激活（负输入斜率 $\alpha = 0.2$）的单层前馈神经网络来表示注意机制a，即：

$$a(h_i^t, h_j^t, \omega_*^t(\langle \overrightarrow{i,j} \rangle)) = \mathbf{a}^T [h_i^t \| h_j^t \| \omega_*^t(\langle \overrightarrow{i,j} \rangle)] \tag{4.2}$$

其中$\|$是连接运算符。

然后我们使用一个softmax函数来对所有可取的j归一化注意力系数：

$$\alpha_{ij}^t = \frac{\exp(e_{ij}^t)}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik}^t)} \tag{4.3}$$

最后，使用新的协作图$A'^t = [\alpha_{ij}^t]_{N \times N}$（权重由图注意机制生成），我们对顶点特征矩阵$H_*^t$应用图卷积：

$$H_*'^t = H_*^t (A'^t)^T \tag{4.4}$$

其中 $H_*'^t = [h_1'^t, h_2'^t, \cdots, h_N'^t]$ 的上下文特征矩阵（由N个上下文特征向量拼接）。请注意，因为$H_*'^t$将在后面被馈送到由多层感知机参数化的Q网络，其中已经包含线性变换和激活过程，所以我们在这里没有显式地写出卷积操作的线性变换和激活函数。

4.2.4　状态的构建

在t时刻，当运输工具、站点和路线的上下文特征向量都已经生成之后，我们为每个停靠在站点（即t时刻有事件触发）的运输工具智能体构建状态。对于每个事件(V_i, P_j, t)（运输工具V_i在时刻t到达站点P_j），我们将V_i，P_j，R_k在t时刻的特征向量与它们相应的上下文向量作为智能体的状态，即：

$$s_i^t = h_{V,i}^t \| h_{P,j}^t \| h_{R,k}^t \| h_{V,i}'^t \| h_{P,j}'^t \| h_{R,k}'^t \tag{4.5}$$

然后，对于每个运输工具智能体 V_i，将状态传递给由多层感知机参数化的Q网络$Q_i(s, a)$中，以生成所有动作的Q值。

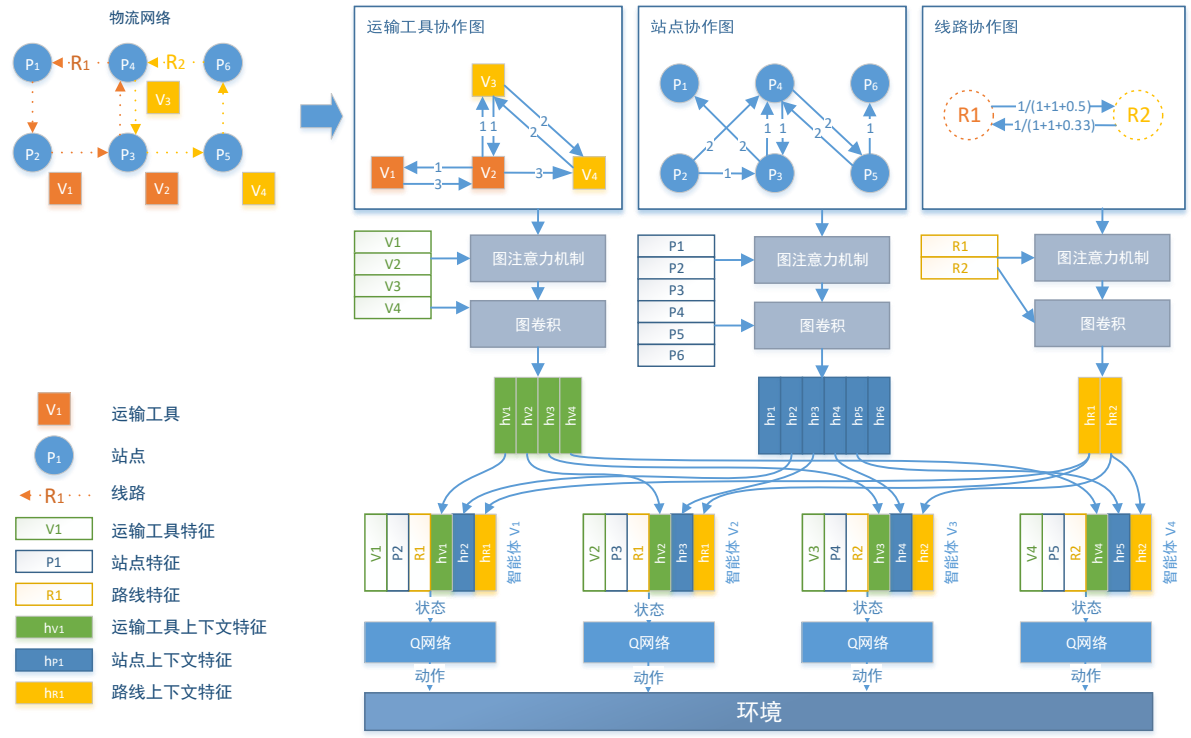


图 4.3　结合图神经网络的物流网络调度模型。

整体模型结构如图4.3所示①。模型的训练依然使用深度Q学习，奖励的设置与3.2.2.3节相同。

4.3　实验

在本节，我们依然使用3.3.1节中介绍的ECR任务进行实验，实验设置与3.3.2节相同。特别的是，所有多智能体共享一个注意力机制模型，以及为了保证注意力机制训练的稳定性，对模型注意力机制涉及的变量均采用10^{-5}的学习率（比其他变量的学习率低一个数量级）。

4.3.1　特征权值分析

为了展示和分析模型所学习到的知识，以验证其可解释性，对于训练好的模型，我们对式4.2中的权值向量进行分析，结果如表4.1所示。由于各特征的尺度有差异，将权值向量中各元素的绝对值进行直接比较意义不大，但我们仍旧可以通过权值的正负来定性分析各个特征项对协作关系的贡献程度。从式4.2的形式可以看出，如果权值为正，说明该权值所对应的特征项对协作关系有正的贡献，即特征项的值越大，协作关系的强度越强，反之则越弱。

① 特别的是，在对站点协作图进行图注意力机制时，还加入了一个特殊特征，即“待使用注意力机制计算权重的站点是否在当前事件的运输工具所在的航线上”。这一特征未在图中体现。

表 4.1　某一性能较优的训练后模型的特征权值

(a) 集装箱船特征权值

特征	权值
空余空间比例	0.63
空箱比例	−0.02
货箱比例	−0.71
与当前船时长距离	−6.41

(b) 港口特征权值

特征	权值
当前空箱数	0.01
缺箱数	0.09
历史平均空箱数	−0.09
下一艘船的时长距离	0.68
经过的航线数量	−0.41
与当前港口时长距离	−5.00
是否在当前船的航线上	1.93

(c) 路线特征权值

特征	权值
平均空箱数	−0.49
历史平均空箱数	0.70
历史平均缺箱数	5.02
与当前航线时长距离	4.27

表 4.2　不同空集装箱调度方法的性能比较

方法	满足率 (%)		
	80% 集装箱	100% 集装箱	150% 集装箱
无调度	26.58 ± 0.90	29.87 ± 0.85	38.25 ± 1.07
库存控制	58.30 ± 0.93	61.07 ± 0.98	68.63 ± 0.98
在线线性规划	76.28 ± 1.54	85.75 ± 1.34	94.48 ± 1.00
加入库存控制的在线线性规划	81.09 ± 1.21	88.99 ± 0.89	96.30 ± 0.80
SA-MARL	65.39 ± 1.20	72.04 ± 1.57	84.21 ± 1.45
TA-MARL	75.25 ± 1.38	83.48 ± 0.94	93.75 ± 0.69
DA-MARL	82.04 ± 1.69	**95.97 ± 0.63**	**97.70 ± 0.98**
GA-MARL	**82.65 ± 1.32**	94.86 ± 0.73	96.70 ± 0.66
离线线性规划（上界）	98.32 ± 0.60	98.95 ± 0.31	99.42 ± 0.25

由表4.1的结果可见，模型学习到的特征权值符合我们的一般常识。例如，我们会常识性地认为，与当前事件接近（即时长距离小）的船或港口更倾向于合作，而表4.1a中的“与当前船时长距离”和表4.1b中的“与当前港口时长距离”也都为显著的负值，即时长距离越短，协作关系的强度越强。我们会常识性地认为，在当前事件的船所在航线轨迹上的后续港口会更倾向于与当前事件有合作关系，而表4.1b中的“是否在当前船的航线上”的特征也为明显正值，即如果在当前航线上（特征的值为1）则协作关系增强。

4.3.2　性能比较

我们将本章所提出的结合图神经网络和注意力机制的多智能体强化学习模型（Graph Attention MARL，GA-MARL）与3.3.3中的方法进行性能比较，结果见表4.2。可见，GA-MARL和我们在前章提出的表现最好的模型DA-MARL性能接近。GA-MARL在80%集装箱的环境下取得最优结果，而在100%集装箱和150%集装箱的环境下略逊于DA-MARL。

4.4 本章小结

在本章，我们将图神经网络和注意力机制的方法引入多智能体强化学习的框架内，首先从复杂的物流网络中动态抽取出协作图，然后通过注意力机制学习物流网络中包括智能体在内的各对象的实时协作关系，且以权重的方式对外表征，最后通过图卷积的方式对协作信息进行整合，增强了模型的智能性和可解释性。实验结果表明，模型学习到的协作知识与我们的预期相符，同时性能与我们第三章提出的最好模型性能接近。

第五章　总结与展望

本文提出了一种新型的多智能体强化学习方法来解决复杂物流网络内的资源平衡调度问题。首先，将复杂运输网络中的资源平衡问题形式化为随机博弈，并由此提出一个合作的多智能体强化学习框架，将网络中的大量运输工具建模为协作的多智能体，同时提出三个层次的合作指标，为改善强化学习中状态和奖励的设计提供指导，以更好地促进复杂物流网络中运输工具智能体的合作。通过这种方式，我们为复杂物流网络中的资源平衡问题提供了一个端到端和高性能的解决方案。该方法不仅对强不确定性的供需预测更加稳健，而且与传统的多阶段运筹学方法相比具备更高的性能和灵活性。然后，将图神经网络和注意力机制的方法引入多智能体强化学习的框架内，首先从复杂的物流网络中动态抽取出协作图，然后通过注意力机制学习物流网络中包括智能体在内的各对象的实时协作关系，且以权重的方式对外表征，最后通过图卷积的方式对协作信息进行整合，增强了模型的智能性和可解释性。最后，在现实海洋物流场景中重要的空集装箱调度任务上进行了广泛的实验。实验结果表明，我们提出的新方法可以有效学习和表征智能体之间的协作关系，并促进智能体之间的合作。与基于运筹学组合优化方法的传统解决方案相比，我们的方法在性能和可解释性上均带来显着改进。

本文的工作可以在以下方面进行扩展：

1. 本文所提出的结合图神经网络和注意力机制的方法本质上是在强化学习的状态层面进行了深入设计，但对于强化学习的另一重要部分——奖励则未有触及。一种可能的思路是将图神经网络和注意力机制所学习到的协作关系通过某种具有延迟的方式反馈到奖励的塑造上，类似于[29]或[30]中所体现的思路。
2. 本文3.1节中所形式化的物流网络资源调度问题仍然是对实际场景的简化。一个重要的简化是，实际场景中往往需要进行有提前量的调度，而不是3.1节中的实时调度。例如，在海运物流网络的ECR问题中，空集装箱的调度不仅需要指定数量，还需要指定每个空集装箱的目标港口（以确定其在集装箱船上的摆放位置）。也就是说，每个空集装箱的目的地是在调度之初就已经提前指定好的。这种“带有提前量的调度”使得问题的动作空间进一步增大。思路是可以通过将文中的每个智能体再次分解成多个子智能体来应对大的动作空间。
3. 本文为了使算法具有更高的稳定性，使用了深度Q学习作为主要的强化学习方法，然而深度Q学习往往适用于离散的动作空间，而本问题的动作空间实质上是连续的。尝试基于策略梯度（Policy Gradient）等适用于连续动作空间的强化

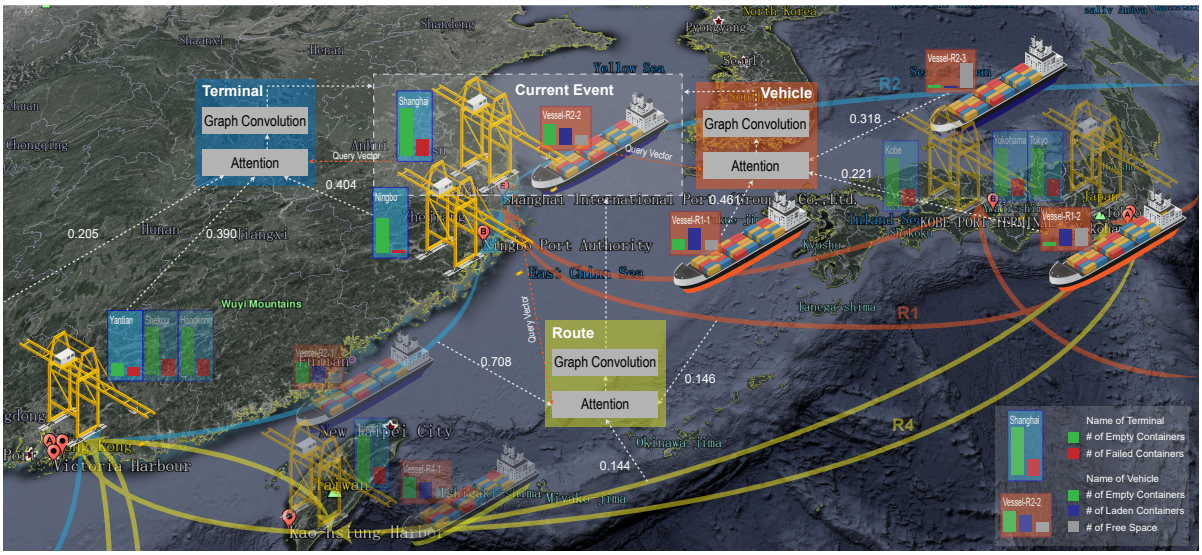


图 5.1　将本文方法部署到实际生产环境中的UI示意图。①

学习方法可能可以获得更优的结果。

4. 本文主要以资源短缺量（式3.1）作为优化目标，但实际物流问题中根据场景不同，还具有别的优化目标。后续工作可以将多个目标进行联合优化。
5. 将算法部署到实际的工业应用场景之中，如图5.1所示。

① 图中的卫星图像来自Google Earth，集装箱船和吊机图标来自Freepik。

参考文献

[1] Dong-Ping Song and Jing-Xin Dong. "*Empty Container Repositioning*". In: *Handbook of Ocean Container Transport Logistics*. Springer, Cham, **2015**: 163–208. https://link.springer.com/chapter/10.1007/978-3-319-11891-8_6, retrieved on 2018-08-18.

[2] Ling Pan, Qingpeng Cai, Zhixuan Fang *et al.* "*Rebalancing Dockless Bike Sharing Systems*". *arXiv:1802.04592 [cs]*, 2018-02. http://arxiv.org/abs/1802.04592, retrieved on 2018-08-10.

[3] Kaixiang Lin, Renyu Zhao, Zhe Xu *et al.* "*Efficient Large-Scale Fleet Management via Multi-Agent Deep Reinforcement Learning*". In: ACM Press, **2018**: 1774–1783. http://dl.acm.org/citation.cfm?doid=3219819.3219993, retrieved on 2018-08-07.

[4] UNCTAD. *Review of maritime transport 2017*, **2017**.

[5] Xihan Li, Jia Zhang, Jiang Bian *et al.* "*A Cooperative Multi-Agent Reinforcement Learning Framework for Resource Balancing in Complex Logistics Network*". In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, **2019**: 980–988. http://dl.acm.org/citation.cfm?id=3306127.3331794, retrieved on 2019-05-13.

[6] Warren B Powell. "*Toward a Unified Modeling Framework for Real-Time Logistics Control*". *Military Operations Research*, **1996**: 69–79.

[7] Teodor Gabriel Crainic and Gilbert Laporte. "*Planning models for freight transportation*". *European journal of operational research*, **1997**, *97*(3): 409–438.

[8] Jing-An Li, Stephen CH Leung, Yue Wu *et al.* "*Allocation of empty containers between multi-ports*". *European Journal of Operational Research*, **2007**, *182*(1): 400–412.

[9] Rafael Epstein, Andres Neely, Andres Weintraub *et al.* "*A Strategic Empty Container Logistics Optimization in a Major Shipping Company*". *Interfaces*, 2012-02: 5–16. http://pubsonline.informs.org/doi/abs/10.1287/inte.1110.0611, retrieved on 2018-08-10.

[10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver *et al.* "*Human-level control through deep reinforcement learning*". *Nature*, 2015-02: 529–533. http://www.nature.com/articles/nature14236, retrieved on 2018-05-08.

[11] David Silver, Aja Huang, Chris J. Maddison *et al.* "*Mastering the game of Go with deep neural networks and tree search*". *Nature*, 2016-01: 484–489. http://www.nature.com/articles/nature16961, retrieved on 2018-05-11.

[12] Zhe Xu, Zhixin Li, Qingwen Guan *et al.* "*Large-Scale Order Dispatch in On-Demand Ride-Hailing Platforms: A Learning and Planning Approach*". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. London, United Kingdom: ACM, **2018**: 905–913. http://doi.acm.org/10.1145/3219819.3219824.

[13] Michael L. Littman. "*Friend-or-Foe Q-learning in General-Sum Games*". In: *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., **2001**: 322–328. http://dl.acm.org/citation.cfm?id=645530.655661.

[14] Junling Hu and Michael P. Wellman. "*Nash Q-learning for General-sum Stochastic Games*". *J. Mach. Learn. Res.* 2003-12: 1039–1069. http://dl.acm.org/citation.cfm?id=945365.964288.

[15] Amy Greenwald and Keith Hall. "*Correlated-Q Learning*". In: *Proceedings of the Twentieth International Conference on Machine Learning*. Washington, DC, USA: AAAI Press, **2003**: 242–249. http://dl.acm.org/citation.cfm?id=3041838.3041869.

[16] Xiaofeng Wang and Tuomas Sandholm. "*Reinforcement Learning to Play an Optimal Nash Equilibrium in Team Markov Games*". In: *Proceedings of the 15th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, **2002**: 1603–1610. http://dl.acm.org/citation.cfm?id=2968618.2968817.

[17] Marc Lanctot, Vinicius Zambaldi, Audrūnas Gruslys *et al.* "*A Unified Game-theoretic Approach to Multiagent Reinforcement Learning*". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California, USA: Curran Associates Inc., **2017**: 4193–4206. http://dl.acm.org/citation.cfm?id=3294996.3295174.

[18] Sam Devlin and Daniel Kudenko. "*Theoretical Considerations of Potential-based Reward Shaping for Multi-agent Systems*". In: *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*. Taipei, Taiwan: International Foundation for Autonomous Agents and Multiagent Systems, **2011**: 225–232. http://dl.acm.org/citation.cfm?id=2030470.2030503.

[19] Sam Devlin, Logan Yliniemi, Daniel Kudenko *et al.* "*Potential-based Difference Rewards for Multiagent Reinforcement Learning*". In: *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*. Paris, France: International Foundation for Autonomous Agents and Multiagent Systems, **2014**: 165–172. http://dl.acm.org/citation.cfm?id=2615731.2615761.

[20] Patrick Mannion, Jim Duggan and Enda Howley. "*Generating multi-agent potential functions using counterfactual estimates*". *Proceedings of Learning, Inference and Control of Multi-Agent Systems (at NIPS 2016)*, **2016**.

[21] Zonghan Wu, Shirui Pan, Fengwen Chen *et al.* "*A Comprehensive Survey on Graph Neural Networks*". *arXiv:1901.00596 [cs, stat]*, 2019-01. http://arxiv.org/abs/1901.00596, retrieved on 2019-01-07.

[22] Jie Zhou, Ganqu Cui, Zhengyan Zhang *et al.* "*Graph Neural Networks: A Review of Methods and Applications*". *arXiv:1812.08434 [cs, stat]*, 2018-12. http://arxiv.org/abs/1812.08434, retrieved on 2018-12-27.

[23] Thomas N. Kipf and Max Welling. "*Semi-Supervised Classification with Graph Convolutional Networks*". *arXiv:1609.02907 [cs, stat]*, 2016-09. http://arxiv.org/abs/1609.02907, retrieved on 2019-01-06.

[24] Petar Veličković, Guillem Cucurull, Arantxa Casanova *et al.* "*Graph Attention Networks*". *arXiv:1710.10903 [cs, stat]*, 2017-10. http://arxiv.org/abs/1710.10903, retrieved on 2019-03-25.

[25] Kunal Menda, Yi-Chun Chen, Justin Grana *et al.* "*Deep Reinforcement Learning for Event-Driven Multi-Agent Decision Processes*". *arXiv:1709.06656 [cs]*, 2017-09. http://arxiv.org/abs/1709.06656, retrieved on 2018-08-19.

[26] Andrew Y. Ng, Daishi Harada and Stuart J. Russell. "*Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping*". In: *Proceedings of the Sixteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., **1999**: 278–287. http://dl.acm.org/citation.cfm?id=645528.657613.

[27] Regina Asariotis, Hassiba Benamara, Hannes Finkenbrink *et al. Review of Maritime Transport, 2011* [techreport], **2011**.

[28] Yin Long, Loo Hay Lee and Ek Peng Chew. "*The sample average approximation method for empty container repositioning with uncertainties*". *European Journal of Operational Research*, 2012-10: 65–75. http://linkinghub.elsevier.com/retrieve/pii/S0377221712003116, retrieved on 2018-09-01.

[29] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza *et al.* "*Asynchronous Methods for Deep Reinforcement Learning*". *arXiv:1602.01783 [cs]*, 2016-02. http://arxiv.org/abs/1602.01783, retrieved on 2018-04-16.

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza *et al.*; ed. by Z. Ghahramani, M. Welling, C. Cortes *et al.* "*Generative Adversarial Nets*". In: *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., **2014**: 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf, retrieved on 2019-05-09.

附录 A　ECR问题的线性规划模型

正文3.3.3节中ECR问题的线性规划模型如下：

$$\min \sum_{P_i \in P, t \in \textsc{Event}(P_i)} L_i^t$$

Subject to

$$C_{P,i}^t = C_{P,i}^{\text{prev}(P,i,t)} - D_i^t + S_i^t - \sum_{j=1}^{|V|} I(i,j,t) x_j^t,$$

$$L_i^t \geq D_i^t - C_{P,i}^{\text{prev}(P,i,k)},$$

$$L_i^t \geq 0,$$

$$\text{for } P_i \in P, t \in \textsc{Event}(P_i);$$

$$C_{V,j}^t = C_{V,j}^{\text{prev}(V,j,t)} + x_j^t,$$

$$0 \leq C_{V,j}^t \leq Cap_j^t,$$

$$\text{for } V_j \in V, t \in \textsc{Event}(V_j);$$

Event(·)表示自变量（可以是船只或港口）预测会发生事件的时刻的集合。这一信息可以通过V，R和线路时长函数$d_j(\cdot,\cdot)$来推断。指标函数$I(i,j,t)$也可以通过类似的方式推断。prev(P,i,t)和prev(V,j,t)分别表示在港口P_i和船只V_j在时刻t事件的上一个事件的时刻。每个港口 $P_i \in P$ 的外部需求D_i^t和外部供给 S_i^t 都由外部预测模型提供。在ECR问题中，Cap_j^t将根据时刻t中V_j的载货量动态变化。对于基于订单的预测模型（也就是模型首先预测未来的订单集合O，然后根据O预测D_i^t和S_i^t），如果我们假设所有外部需求D_i^t都可以被满足，那么Cap_j^t也可以基于O来预测（以便可以估算每个时刻的每个船只的货箱数量）。线性规划模型由GNU线性编程套件（GNU Linear Programming Kit，GLPK）在整数规划模式下求解。

附录 B　ECR环境模拟器

本部分介绍正文3.3.2节中ECR环境的设计细节。

B.1　线路时刻表

每条线路的时间表显示在表 B.1、表 B.2、表 B.3和表 B.4中，均基于实验部分中提到的同一商业公司所提供的信息。所有路线均为环线。为了让船舶能均匀分布，当环境初始化时，船舶不需要在某个港口停泊（即可以为在两个港口之间的行驶状态）。

表 B.1　线路R1的时刻表

港口	地区/城市	转运天数
STN	Europe Union	-
NYC	New York	15
SAV	Sawannah	18
LAS	Los Angeles	31
OAK	Oakland	32
YOK	Yokohama	44
SHA	Shanghai	47
KOY	Kobe	51
TKY	Tokyo	52
OAK	Oakland	67
LAS	Los Angeles	68
SAV	Sawannah	82
NYC	New York	85
STN	Europe Union	94

B.2　港口-集装箱船交互的业务逻辑

当触发事件(P_i, V_j)时，即船只V_j（在路线R_k上）到达港口P_i 时，我们的模拟环境遵循一个基于现实的四阶段业务逻辑来执行动作$a \in [-1, 1]$：

- **货箱卸货**：将所有 V_j 上目的地港口为 P_i 的货箱从船上卸下。注意，由于船上装载的集装箱减少，Cap_j^t将增加至$Cap_j'^t$；
- **空箱卸货（如果 $a < 0$ 则执行）**：将数量为 $[-a * C_{V,j}^t]$ 的空箱从集装箱船 V_j 上卸下；

表 B.2　线路R2的时刻表

港口	地区/城市	转运天数
JEB	Arab	-
SIN	Singapore	3
LCB	Thailand	6
YAT	Yantian	9
LAS	Los Angeles	26
OAK	Oakland	28
SHA	Shanghai	43
NIN	Ningbo	44
YAT	Yantian	46
SIN	Singapore	51
JEB	Arab	60

表 B.3　线路R3的时刻表

港口	地区/城市	转运天数
KOY	Kobe	-
TKY	Tokyo	3
LAS	Los Angeles	17
OAK	Oakland	18
TKY	Tokyo	31
KOY	Kobe	33

- **货箱装载：**将港口 P_i 中目的地港口在航线 R_k 内的货箱尽可能按接收日期的顺序装入船只。同一订单中的多个货箱可以分别运输。类似地，因为船上货箱数量增加，$Cap'_j{}^t$将减少到$Cap'_j{}''^t$；
- **空箱装载（如果 $a > 0$ 则执行）：**将数量为 $[a * \min(Cap'_j{}''^t - C^t_{V,j}, C^t_{P,i})]$ 的空箱装入集装箱船 V_j。

其中[·]表示取整函数。在这种业务逻辑中，货箱的调度优先于空箱，这符合海运集装箱物流行业的实际情况。

表 B.4　线路R4的时刻表

港口	地区/城市	转运天数
TKY	Tokyo	-
KOY	Kobe	2
KHH	Taiwan	5
HKG	Hong Kong	6
SKZ	Shekou	7
SIN	Singapore	11
SKZ	Shekou	14
HKG	Hong Kong	15
KHH	Taiwan	16
TKY	Tokyo	19

附录 C　分区域的模型性能统计

实验部分中八种方法的分区域的性能统计数据分别列于表C.1、 C.2、C.3、 C.4、C.5、C.6、 C.7、C.8。所有方法都经过100次测试，我们在表格中报告平均值。

表 C.1　无调度方法的区域统计

区域/城市	总集装箱数	缺箱数	进口 货箱数	进口 空箱数	出口 货箱数	出口 空箱数	满足率
Shanghai	400.26	0	343.61	0	400.26	0	1
Ningbo	0	0	136.56	0	0	0	/
Yantian	601.08	405.24	134.25	0	195.84	0	0.325 814
Shekou	8010.75	7124.58	785.03	0	886.17	0	0.110 623
Thailand	5008.44	4719.1	109.05	0	289.34	0	0.057 77
Singapore	797.06	0	860.58	0	797.06	0	1
Arab	0	0	0	0	0	0	/
Hong Kong	3991.25	2501.22	1129.67	0	1490.03	0	0.373 324
Taiwan	1403.97	944.97	296.11	0	459	0	0.326 93
Tokyo	2200.88	1038.99	866.4	0	1161.89	0	0.527 921
Kobe	3010.1	1881.84	980.57	0	1128.26	0	0.374 825
Yokohama	0	0	289.72	0	0	0	/
Oakland	199.28	6.12	805.74	0	193.16	0	0.969 289
Los Angeles	1403.94	630.41	762.26	0	773.53	0	0.550 971
Sawannah	1002.21	534.1	457	0	468.11	0	0.467 078
New York	200.86	0	445.1	0	200.86	0	1
EU	1003.48	505.85	279.56	0	497.63	0	0.495 904
合计	29 233.56	20 292.42	8681.21	0	8941.14	0	0.299 626

表 C.2　库存控制方法的区域统计

区域/城市	总集装箱数	缺箱数	进口 货箱数	进口 空箱数	出口 货箱数	出口 空箱数	满足率
Shanghai	397.12	6.81	600.44	52.13	390.31	589.51	0.982 852
Ningbo	0	0	208.77	0	0	352.35	/
Yantian	597.89	9.01	946.32	109.82	588.88	507.6	0.984 93
Shekou	8007.37	6068.76	1321.94	536.18	1938.61	0	0.242 103
Thailand	4994.01	2303.97	108.97	2470.64	2690.04	28.3	0.538 653
Singapore	801.94	1.28	1602.52	108.31	800.66	1314.41	0.998 404
Arab	0	0	0	0	0	269	/
Hong Kong	3998.32	1817.13	2051.84	6.09	2181.19	231.21	0.545 527
Taiwan	1402.76	671.1	650.05	47.26	731.66	111.99	0.521 586
Tokyo	2177.95	4.2	1369.37	1006.98	2173.75	431.89	0.998 072
Kobe	2969.29	6.64	1526.76	1526.89	2962.65	152.97	0.997 764
Yokohama	0	0	463.89	0	0	535.82	/
Oakland	199.29	5.74	2225.95	42.91	193.55	2067.45	0.971 198
Los Angeles	1397.14	41.29	2071.77	127.65	1355.85	804.18	0.970 447
Sawannah	998.67	7.9	883.59	275.88	990.77	131.33	0.992 089
New York	200.87	4.01	862.88	43.47	196.86	745.36	0.980 037
EU	994.69	48.39	459.47	481.44	946.3	177.7	0.951 352
合计	29 137.31	10 996.23	17 354.53	6835.65	18 141.08	8451.07	0.612 14

表 C.3　在线线性规划方法的区域统计

区域/城市	总集装箱数	缺箱数	进口 货箱数	进口 空箱数	出口 货箱数	出口 空箱数	满足率
Shanghai	397.31	14.88	597.56	511.27	382.43	1023.93	0.962 548
Ningbo	0	0	204.94	132.56	0	470.35	/
Yantian	598.71	158.21	1110.7	362.69	440.5	1071.43	0.735 749
Shekou	7962.84	1194.66	1663.6	5213.06	6768.18	142.81	0.849 971
Thailand	4971.08	1760.6	100.3	3051.04	3210.48	66.81	0.645 831
Singapore	803.75	82.85	3172.18	1992.17	720.9	4842.45	0.896 921
Arab	0	0	0	60.54	0	329.54	/
Hong Kong	3982.99	156.27	3664.9	1011.68	3826.72	1164.39	0.960 766
Taiwan	1403.4	101.41	614.38	1108.52	1301.99	536.47	0.927 74
Tokyo	2175.06	112.21	2892.28	1727.29	2062.85	2765.18	0.948 411
Kobe	2975.31	158.49	3196.83	2769.94	2816.82	3194.36	0.946 732
Yokohama	0	0	470.02	122.06	0	624.55	/
Oakland	199.17	5.09	2231.34	979.83	194.08	2844.75	0.974 444
Los Angeles	1392.12	49.25	2129.32	1035.11	1342.87	1760.18	0.964 622
Sawannah	1000.7	11.4	840.71	534.48	989.3	319.59	0.988 608
New York	200.84	1.94	816.29	152.33	198.9	759.67	0.990 341
EU	997.67	143.08	455.15	341.7	854.59	146.09	0.856 586
合计	29 060.95	3950.34	24 160.5	21 106.27	25 110.61	22 062.55	0.859 485

表 C.4　加入库存控制的在线线性规划方法的区域统计

区域/城市	总集装箱数	缺箱数	进口 货箱数	进口 空箱数	出口 货箱数	出口 空箱数	满足率
Shanghai	399.33	21.13	606.51	569.89	378.2	1105.43	0.947 086
Ningbo	0	0	207.5	134.47	0	481.19	/
Yantian	598.65	87.58	1029.14	326.27	511.07	877.82	0.853 704
Shekou	7972.27	677.09	1732.56	5657.99	7295.18	74.89	0.915 069
Thailand	4936.96	1927.75	99.03	2832.29	3009.21	17.63	0.609 527
Singapore	796.67	87.21	3331.11	1853.54	709.46	4859.85	0.890 532
Arab	0	0	0	26.51	0	295.51	/
Hongkong	3981.14	1.53	3896.77	1140.93	3979.61	1248.01	0.999 616
Taiwan	1407.85	25	628.13	1284.43	1382.85	647.28	0.982 242
Tokyo	2194.04	21.85	3093.01	1740.66	2172.19	2786.07	0.990 041
Kobe	2969.91	111.72	3325.12	3043.36	2858.19	3536.31	0.962 383
Yokohama	0	0	465.3	109.34	0	638.74	/
Oakland	199.22	5.01	2221.68	936.29	194.21	2814.07	0.974 852
Los Angeles	1396.98	42.27	2138.31	1026.7	1354.71	1702.91	0.969 742
Sawannah	997.87	8.76	871.68	598.89	989.11	421.21	0.991 221
New York	200.77	5.5	852.01	163.15	195.27	834.53	0.972 605
EU	989.84	64.39	455.31	408.18	925.45	128.4	0.934 949
合计	29 041.5	3086.79	24 953.17	21 852.89	25 954.71	22 469.85	0.889 923

表 C.5　自我感知的MARL方法的区域统计

区域/城市	总集装箱数	缺箱数	进口 货箱数	进口 空箱数	出口 货箱数	出口 空箱数	满足率
Shanghai	396.75	15.46	588.56	327.65	381.29	871.11	0.961 033
Ningbo	0	0	204.44	25.86	0	366.75	/
Yantian	595.06	35.19	1588.92	145.66	559.87	1201.87	0.940 863
Shekou	8010.53	5979.48	1467.91	1101.63	2031.05	646.8	0.253 548
Thailand	4940.34	254.41	96.13	4639.55	4685.93	87.87	0.948 504
Singapore	799.44	89.14	1743.24	1250.79	710.3	2679.4	0.888 497
Arab	0	0	0	224.56	0	491.8	/
Hong Kong	3992.99	1133.04	2252.99	710.83	2859.95	412.74	0.716 243
Taiwan	1405.77	430.5	604.48	453.21	975.27	200.81	0.693 762
Tokyo	2192.7	100.74	1515.5	1663.18	2091.96	1317.72	0.954 057
Kobe	2970.62	180.91	1710.44	1887.26	2789.71	883.26	0.9391
Yokohama	0	0	456.66	10.66	0	531.22	/
Oakland	199.28	7.37	2718.49	188.76	191.91	2693.95	0.963 017
Los Angeles	1396.48	67.75	2731.53	906.05	1328.73	2263.86	0.951 485
Sawannah	1006.08	36.43	839.19	450.62	969.65	289.04	0.963 79
New York	200.75	1.81	831.19	69.71	198.94	731.6	0.990 984
EU	994.14	61.25	449.12	573.6	932.89	272.27	0.938 389
合计	29 100.93	8393.48	19 798.79	14 629.58	20 707.45	15 942.07	0.702 295

表 C.6　区域感知的MARL方法的区域统计

区域/城市	总集装箱数	缺箱数	进口 货箱数	进口 空箱数	出口 货箱数	出口 空箱数	满足率
Shanghai	400.81	20.7	587.02	337.98	380.11	885.46	0.948 355
Ningbo	0	0	206.01	30.21	0	377.1	/
Yantian	597.05	47.7	1633.54	195.44	549.35	1306.06	0.920 107
Shekou	7967.28	2248.07	1447.38	4352.67	5719.21	98.23	0.717 837
Thailand	4939.66	135.07	85.8	4730.73	4804.59	25.76	0.972 656
Singapore	805.05	178.25	2630.17	659.64	626.8	3079.51	0.778 585
Arab	0	0	0	152.09	0	420.43	/
Hong Kong	3997.4	1045.12	3129.24	350.29	2952.28	868.87	0.738 55
Taiwan	1402.58	425.2	605.39	506.39	977.38	270.44	0.696 844
Tokyo	2181.59	169.94	2362.92	1617.55	2011.65	2208.95	0.922 103
Kobe	2969.32	234.94	2671.21	1985.45	2734.38	1986.71	0.920 878
Yokohama	0	0	465.79	16.68	0	542.53	/
Oakland	199.3	5.12	2729.45	226.27	194.18	2744.53	0.974 31
Los Angeles	1393.16	54.74	2726.46	870.76	1338.42	2228.29	0.960 708
Sawannah	998.18	32.36	827.99	558.79	965.82	404.74	0.967 581
New York	200.76	0.04	813.79	124.39	200.72	772.29	0.999 801
EU	993.28	51.05	452.57	505.77	942.23	202.39	0.948 605
合计	29 045.42	4648.3	23 374.73	17 221.1	24 397.12	18 422.29	0.834 133

表 C.7　区域合作的MARL方法的区域统计

区域/城市	总集装箱数	缺箱数	进口 货箱数	进口 空箱数	出口 货箱数	出口 空箱数	满足率
Shanghai	398.2	2.46	591.15	487.31	395.74	1001.32	0.993 822
Ningbo	0	0	205.53	122.52	0	469.3	/
Yantian	596.74	20.2	1614.78	240.68	576.54	1298.11	0.966 149
Shekou	7963.94	383.02	1704.92	5951.09	7580.92	11.96	0.951 906
Thailand	4937.34	205.8	100.41	4670.81	4731.54	57.05	0.958 318
Singapore	800.5	68.58	3395.41	601.86	731.92	3674.31	0.914 329
Arab	0	0	0	216.75	0	484.96	/
Hong Kong	3994.47	117.05	3919.51	543.61	3877.42	881.82	0.970 697
Taiwan	1401.06	47.4	619.78	1025.34	1353.66	405.71	0.966 168
Tokyo	2184.47	64.91	3109.54	2088.9	2119.56	3245.84	0.970 286
Kobe	2975.98	61.25	3380.17	2273.1	2914.73	2754.09	0.979 419
Yokohama	0	0	462.15	21.55	0	541.47	/
Oakland	199.32	7.29	2766.72	449.23	192.03	2955.57	0.963 426
Los Angeles	1399.43	44.17	2763.45	661.21	1355.26	2028.79	0.968 437
Sawannah	1003.12	35.03	887.31	419.99	968.09	196.33	0.965 079
New York	200.8	9.73	853.98	67.35	191.07	711.75	0.951 544
EU	995.32	67.04	447.44	561.59	928.28	202.88	0.932 645
合计	29 050.69	1133.93	26 822.25	20 402.89	27 916.76	20 921.26	0.959 447

表 C.8　图注意力网络MARL方法的区域统计

区域/城市	总集装箱数	缺箱数	进口 货箱数	进口 空箱数	出口 货箱数	出口 空箱数	满足率
Shanghai	399.82	18.69	589.69	239.04	381.13	792.46	0.953 254
Ningbo	0	0	206.21	130.93	0	475.46	/
Yantian	594.99	9.58	1654.44	200.03	585.41	1291.89	0.983 899
Shekou	7959.25	750.13	1686.31	5576.52	7209.12	30.68	0.905 754
Thailand	4926.95	42.62	104.41	4765.27	4884.33	0.25	0.991 35
Singapore	797.9	38.52	3288.54	628.33	759.38	3562.56	0.951 723
Arab	0	0	0	242.08	0	502.6	/
Hongkong	3988.87	178.28	3851.11	735.27	3810.59	1072.42	0.955 306
Taiwan	1404.84	50.87	625.72	984.75	1353.97	365.17	0.963 789
Tokyo	2175.52	57.76	2976.05	2221.01	2117.76	3246.81	0.973 45
Kobe	2966.83	140.28	3279.84	2095.42	2826.55	2578.43	0.952 717
Yokohama	0	0	460.68	84.95	0	603.8	/
Oakland	199.23	6.18	2804.62	298.66	193.05	2853.14	0.968 981
Los_Angeles	1391.74	53.92	2826.82	910.07	1337.82	2334.61	0.961 257
Sawannah	1004.31	40.84	866.93	658.21	963.47	490.2	0.959 335
New_York	200.9	5.41	842.59	152.81	195.49	840.72	0.973 071
EU	996.83	44.27	445.26	614.92	952.56	211.38	0.955 589
合计	29 007.98	1437.35	26 509.22	20 538.27	27 570.63	21 252.58	0.948 581

致谢

光阴如白驹过隙，在燕园的三年时光转瞬即逝。而我也在这三年的时光中收获了宝贵的学识、科研能力和成长经历。在即将踏入未来的新征程之际，我谨向三年来帮助过我的所有人致以衷心的谢意。

感谢我的导师童云海教授。童老师治学严谨、平易近人、包容友善，在科研上给予我悉心的指导、热情的鼓励和充分的自由空间，在实验室设备及学术交流上也给予我充分的支持和信任，使我受益良多。我谨向老师致以最诚挚的敬意。

感谢实验室的李祥泰、庞博琛、白剑刚、陈逸人、丁宇辰、周惠斌、曹琳琳、楚天翔、尤安升、乔康、鄂有君、徐子楠、沈迪曼等同学。和你们的交流不仅愉快，而且经常迸发出新思维的火花，让我受到很多启发，激励着我继续前行。感谢三年来坐在我旁边的邹萌同学，和你的交流与合作让我印象深刻，你在科研上的坚持也始终感染着我。

感谢微软亚洲研究院的张佳和边江博士，罗强、赵之源同学及各位实习生们，以及东方海运的工程师们。你们的热情友善和思想洞见是我经历中的宝贵财富，本论文的研究工作也很大受益于你们的支持和帮助。

感谢谷歌开发者关系团队与TensorFlow工程团队的程路、Soonson Kwon、王铁震、李锐、李双峰、段威、Mike Liang、Paige Bailey和Pryce等Googler，李卓桓、朱金鹏、江骏等GDE①同侪，以及人民邮电出版社的王军花编辑。你们的支持、认可与帮助是我前进的重要动力。

感谢校心理中心的徐凯文副教授，汪春花、聂晶、李晨枫、梁冠琼老师和牟惊雷、黄佳雨、胡海炼同学。和你们一起工作是我的荣幸，希望我的一点微小工作能够惠及北大的同学们。

感谢我长久以来的朋友，UC Davis的吴康隆同学、中科大的李济安同学和豆瓣的王子阳同学。遇见你们是我人生中的幸运。

最后，谨感谢父母对我的养育之恩与对我学业道路的无条件支持，使我能够踏实前行。

① Google Developers Experts `https://developers.google.com/programs/experts/`

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因须要延迟发布学位论文电子版，授权学校在 □ 一年 / □ 两年 / □ 三年以后在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：　　日期：　　年　　月　　日