International Journal of Internet, Broadcasting and Communication Vol.16 No.3 205-211 (2024)
http://dx.doi.org/10.7236/IJIBC.2024.16.3.205

IJIBC 24-3-21

Intelligent Warehousing: Comparing Cooperative MARL Strategies

Yosua Setyawan Soekamto[1,2], Dae-Ki Kang[3,*]

[1] PhD Student, Department of Computer Engineering, Dongseo University, Busan, South Korea
[2] Lecturer, Department of Information Systems, Universitas Ciputra Surabaya
[3] Professor, Department of Computer Engineering, Dongseo University, Busan, South Korea
[*]dkkang@dongseo.ac.kr

Abstract

Effective warehouse management requires advanced resource planning to optimize profits and space. Robots offer a promising solution, but their effectiveness relies on embedded artificial intelligence. Multi-agent reinforcement learning (MARL) enhances robot intelligence in these environments. This study explores various MARL algorithms using the Multi-Robot Warehouse Environment (RWARE) to determine their suitability for warehouse resource planning. Our findings show that cooperative MARL is essential for effective warehouse management. IA2C outperforms MAA2C and VDA2C on smaller maps, while VDA2C excels on larger maps. IA2C's decentralized approach, focusing on cooperation over collaboration, allows for higher reward collection in smaller environments. However, as map size increases, reward collection decreases due to the need for extensive exploration. This study highlights the importance of selecting the appropriate MARL algorithm based on the specific warehouse environment's requirements and scale.

Keywords: *Multi-Agent Reinforcement Learning, Multi-Robot Warehouse Environment, Independent Advantage Actor-Critic, Multi-Agent Advantage Actor-Critic, Value Decomposition Advantage Actor-Critic.*

1. Introduction

In the development of modern business, supply chain management has become increasingly important. With the advancement of technology, businesses have shifted their focus from traditional brick-and-mortar retail stores to online marketplaces. This transition has led to a greater emphasis on managing stock in warehouses rather than in physical stores. Consequently, businesses must now concentrate on the efficient management and delivery of goods based on online orders [1].

Managing goods in a warehouse presents distinct challenges compared to retail stores, requiring effective stock planning and mobilization to maintain space and resource efficiency. Typically, this planning is achieved through business resource planning methods [2]. The actual physical implementation in warehouses is still largely manual, relying on human labor. This manual approach necessitates a large workforce and considerable space for mobilization, which can impact business profits and is considered inefficient.

Manuscript Received: June. 11, 2024 / Revised: June. 18, 2024 / Accepted: June. 23, 2024
Corresponding Author dkkang@dongseo.ac.kr
Tel: +82-51-320-1724, Fax: +82-51-327-8955
Professor, Department of Computer Engineering, Dongseo University, Busan, Korea

Large companies have begun using robots to replace humans in managing and mobilizing warehouse products. Robots can maximize the efficiency of movement and storage space. Although the initial investment in robots is high, the costs can be recouped over time through improved efficiency. Additionally, robots provide enhanced safety and accuracy as they operate according to predefined programs. The effectiveness of robots depends on the intelligence embedded within them [3]. However, robots for warehouse management must utilize artificial intelligence (AI) to make decisions and take actions based on their environment [3], [4]. For this purpose, neural networks, specifically reinforcement learning, are essential. The most suitable AI approach for this case is multi-agent reinforcement learning (MARL) [6].

This research aims to identify the most suitable type of multi-agent reinforcement learning for managing the mobilization of goods in warehouses. MARL can be categorized as cooperative and competitive behaviors. Cooperative MARL involves agents working together as a team, while competitive MARL involves agents competing against each other [7]. Within cooperative MARL, there are two distinct approaches: collaboration and cooperation. Collaboration involves agents working towards a shared goal, whereas cooperation involves supporting each other to achieve individual goals. Although similar, these approaches differ in their emphasis on teamwork. This research seeks to determine which MARL behavior—collaboration or cooperation—is most effective for warehouse management.

2. Literature Review

2.1 Multi-agent Reinforcement Learning (MARL)

Multi-agent reinforcement learning (MARL) has gained significant attention recently due to its ability to model and solve complex problems involving multiple interacting agents. MARL extends traditional single-agent reinforcement learning to environments where multiple agents learn and make decisions simultaneously. This approach is particularly useful when agents must cooperate or compete to achieve their objectives. Notable advancements in MARL have focused on developing algorithms that can handle the challenges of non-stationarity, scalability, and partial observability. For instance, developing the Proximal Policy Optimization (PPO) algorithm and its variants has influenced MARL research. [8] introduced PPO, which balances exploration and exploitation while ensuring stable and efficient policy updates. More recent works have extended PPO to multi-agent settings, demonstrating its efficacy in diverse applications such as robotic control and strategic games [9].

2.2 Cooperative and Competitive Behavior in MARL

Studying cooperative and competitive behaviors in MARL is crucial for understanding how agents interact within shared environments. Cooperative behaviors involve agents working together to maximize a reward, while competitive behaviors involve agents striving to outperform each other. These dynamics can significantly impact the learning process and system performance. Research by [10] provided a comprehensive survey of MARL methods, highlighting the distinctions between cooperative and competitive frameworks. They emphasized the importance of designing reward structures and communication protocols that facilitate effective agent collaboration. Further research by [11] introduced the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm, which is widely adopted for cooperative and competitive tasks. MADDPG allows agents to learn decentralized policies in continuous action spaces, making it suitable for complex, real-world applications.

2.3 Distinguishing Collaboration and Cooperation Goals in Cooperative Behavior

Within the realm of cooperative MARL, it is essential to distinguish between collaboration and cooperation goals. Collaboration involves agents working jointly towards a common objective, often requiring explicit communication and coordination. In contrast, cooperation involves agents pursuing individual goals aligned or beneficial to the group, often relying on implicit coordination. The foundational work distinguishing these concepts, though more recent studies have further explored their implications in MARL contexts [12]. Shared rewards and joint action spaces are typical in collaboration, whereas cooperation may involve shared or individual rewards focusing on complementary actions [13]. Researchers have developed various frameworks to study these dynamics. For example, [14] introduced the concept of counterfactual regret minimization in collaborative settings, which has been instrumental in developing algorithms that can effectively balance collaborative and cooperative strategies [7].

3. Methodology

3.1 System and Environment

The experiments will be conducted using the Multi-Robot Warehouse Environment (RWARE), which offers four predefined layouts: tiny, small, medium, and large. This study will analyze algorithms' performance across the four layouts, running each algorithm for 4000 episodes. In this research the reward discount is 1, means no discount. The experiment will be conducted in high performance computer with Intel Core i9-10900x CPU, two NVIDIA GeForce RTX 4090 GPU, 128 GB DDR5 RAM.

The Multi-Robot Warehouse Environment (RWARE) simulates warehouse settings where agents (robots) are tasked with delivering requested boxes (products) to a designated workstation. RWARE employs a sparse reward system, meaning agents receive a reward only when they successfully deliver and return boxes to the shelves [15]. Each completed task earns the agents 1 point, incentivizing efficient and accurate performance. The appearance of ready products on shelves is randomized, adding an element of unpredictability and challenge to the environment. Agents have five possible actions: turn left, turn right, move forward, load, and unload boxes. Turning left or right changes the agent's direction, while moving forward advances the agent. Loading and unloading actions are performed when the agent is positioned on the shelf tiles. RWARE functions as a partial observation Markov decision process (POMDP) environment, where agents can observe only a limited area—a 3x3 grid centered around themselves. This limited visibility requires agents to make decisions based on incomplete information, mimicking real-world scenarios where full visibility is rarely available. The RWARE environment shown in Figure 1.

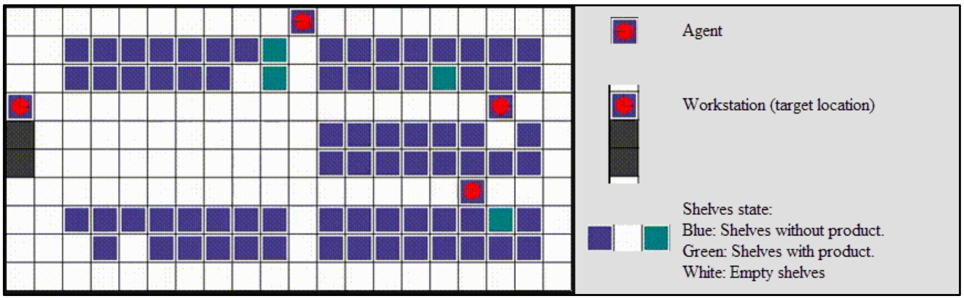
Agent
Workstation (target location)
Shelves state:
Blue: Shelves without product.
Green: Shelves with product.
White: Empty shelves

Figure 1. RWARE Environment

3.2 Independent Synchronous Advantage Actor-Critic

The Independent Synchronous Advantage Actor-Critic (IA2C) algorithm is a decentralized variant of the Advantage Actor-Critic (A2C) algorithm tailored for multi-agent systems. In an independent learning framework, each agent operates with its own actor and critic networks, enabling decentralized training. Each agent independently evaluates its policy and value function without relying on a central controller or shared information [16]. This synchronous action ensures that all agents make decisions and take steps simultaneously, maintaining coordination in their operations. The critic learning (update) function works as follows:

$$\phi_{k+1} = \arg \min_{\phi} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau\in\mathcal{D}_k} \sum_{t=0}^{T} \left(V_{\phi}(o_t) - \hat{R}_t\right)^2 \tag{1}$$

The advantage function of the IA2C algorithm works as follows:

$$A_t = r_{t+1} + \lambda V_{\phi}(o_{t+1}) - V_{\phi}(o_t) \tag{2}$$

The actor (policy) learning works as follows:

$$\nabla_{\theta} J(\theta) \sim \sum_{t=0}^{T-1} \nabla_{\theta} \log \pi_{\theta}(u_t|o_t) A_t \tag{3}$$

The selection of the IA2C algorithm for this research is motivated by its simplicity and effectiveness. The straightforward nature of IA2C makes it a suitable candidate for initial experiments, allowing researchers to establish a baseline performance level.

3.3 Multi-Agent Advantage Actor-Critic

The Multi-Agent Advantage Actor-Critic (MAA2C) algorithm implements the A2C algorithm that follows the centralized training and decentralized execution (CTDE) paradigm. In CTDE, agents share information during the training phase to enhance learning, but their policies during execution are based solely on their local observations. This approach allows agents to benefit from shared experiences and knowledge while maintaining independence during operation [11]. The critic learning (update) function works as follows:

$$\phi_{k+1} = \arg \min_{\phi} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau\in\mathcal{D}_k} \sum_{t=0}^{T} \left(V_{\phi}(o_t, s_t, \mathbf{u}_{\mathbf{t}}^{-}) - \hat{R}_t\right)^2 \tag{4}$$

The advantage function of the MAA2C algorithm works as follows:

$$A_t = r_{t+1} + \lambda V_{\phi}(o_{t+1}, s_{t+1}, \mathbf{u}_{\mathbf{t+1}}^{-}) - V_{\phi}(o_t, s_t, \mathbf{u}_{\mathbf{t}}^{-}) \tag{5}$$

The key difference between MAAC and the Independent Advantage Actor-Critic (IA2C) algorithm lies in handling the critic update function and advantage function. In MAAC, the critic update and advantage function consider the actions of all agents, denoted as $\mathbf{u}_{\mathbf{t}}^{-}$.

3.4 Value Decomposition Advantage Actor-Critic

The Value Decomposition Advantage Actor-Critic (VDA2C) algorithm extends the A2C framework for multi-agent systems. VDA2C leverages the concept of value decomposition, which breaks down the global value function into individual components corresponding to each agent. This approach facilitates more effective learning and coordination among agents by focusing on the contribution of each agent to the overall value function [17]. The critic mixing function works as shown in (7), and therefore, the mixed critic learning function, as shown in (8), as follows:

$$V_{tot}(\mathbf{u}, s; \boldsymbol{\phi}, \psi) = g_{\psi}(s, V_{\phi_1}, V_{\phi_2}, .., V_{\phi_n}) \tag{6}$$

$$\phi_{k+1} = \arg \min_{\phi} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} (V_{tot} - \hat{R}_t)^2 \tag{7}$$

The advantage function of the VDA2C algorithm works as follows:

$$A_t = r_{t+1} + \lambda V_{tot}^{t+1} - V_{tot}^{t} \tag{8}$$

The core idea behind VDA2C is to decompose the global value function into additive value functions for each agent. This decomposition helps understand each agent's contribution to the total reward.

4. Result and Discussion

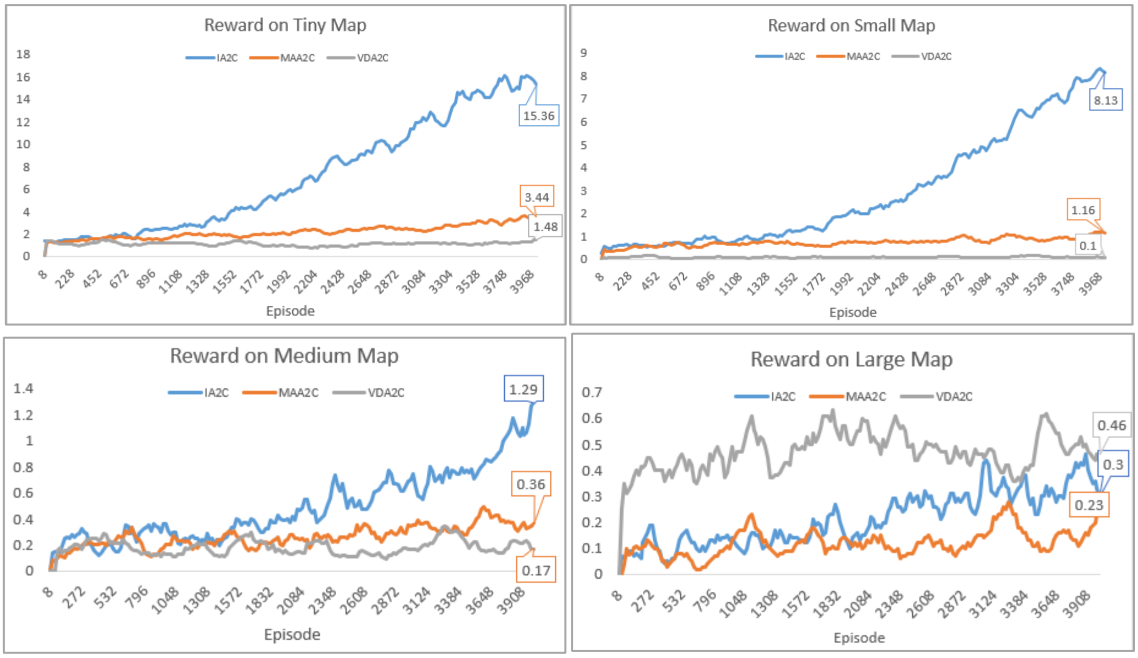

Figure 2. Experiment Results

The experiment shows that IA2C performs better than MAA2C and VDA2C on the tiny, small, and medium maps. This outcome is attributed to IA2C's use of a decentralized (independent) Actor-Critic, in contrast to the centralized critic employed by the other algorithms during the training process. The agents accumulate more rewards on smaller maps due to the reduced space and simpler goal achievement.

5. Conclusion

This research has demonstrated that a cooperative multi-agent reinforcement learning (MARL) approach is essential for managing warehouse environments like RWARE. Our experiments revealed that the IA2C algorithm outperformed MAA2C and VDA2C on tiny, small, and medium maps. However, on the large map, VDA2C showed superior performance. The results indicate that IA2C's independent learning framework, which emphasizes cooperation rather than collaboration, allows agents to collect more rewards. This decentralized approach enables each agent to act independently, optimizing their actions based on individual observations and thus improving overall performance in smaller environments. Moreover, we found that the reward collection decreases as the map size increases. This trend is due to the larger exploration area required on bigger maps, making it more challenging for agents to achieve their goals efficiently.

Acknowledgment

This research was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2023RIS-007) and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2022R1A2C2012243).

References

[1] D. Ivanov, "Digital Supply Chain Management and Technology to Enhance Resilience by Building and Using End-to-End Visibility During the COVID-19 Pandemic," *IEEE Trans Eng Manag*, vol. 71, pp. 10485–10495, 2024, doi: 10.1109/TEM.2021.3095193.

[2] J. Gu, M. Goetschalckx, and L. F. McGinnis, "Research on warehouse operation: A comprehensive review," *Eur J Oper Res*, vol. 177, no. 1, pp. 1–21, Feb. 2007, doi: 10.1016/j.ejor.2006.02.025.

[3] M. C. Gombolay, R. J. Wilcox, and J. A. Shah, "Fast Scheduling of Multi-Robot Teams with Temporospatial Constraints," *IEEE Transactions on Robotics*, vol. 34, no. 1, pp. 220–239, 2018, doi: 10.1109/TRO.2018.2795034.

[4] Y. Liu, X. Tao, X. Li, A. W. Colombo, and S. Hu, "Artificial Intelligence in Smart Logistics Cyber-Physical Systems: State-of-The-Arts and Potential Applications," *IEEE Transactions on Industrial Cyber-Physical Systems*, vol. 1, pp. 1–20, Jun. 2023, doi: 10.1109/ticps.2023.3283230.

[5] M. Akbari and T. N. A. Do, "A systematic review of machine learning in logistics and supply chain management: current trends and future directions," *Benchmarking*, vol. 28, no. 10. Emerald Group Holdings Ltd., pp. 2977–3005, Nov. 05, 2021. doi: 10.1108/BIJ-10-2020-0514.

[6] L. Buşoniu, R. Babuška, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 38, no. 2. pp. 156–172, Mar. 2008. doi: 10.1109/TSMCC.2007.913919.

[7] K. Zhang, Z. Yang, and T. Başar, "Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.10635

[8] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," Jul. 2017, [Online]. Available: http://arxiv.org/abs/1707.06347

[9] M. Zhou *et al.*, "MALib: A Parallel Framework for Population-based Multi-agent Reinforcement Learning," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.07551

[10] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean Field Multi-Agent Reinforcement Learning," Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.05438

[11] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.02275

[12] L. Panait and S. Luke, "Cooperative Multi-Agent Learning: The State of the Art," *Autonomous Agents and Multi-Agent Systems*, vol. 11, no. 1, pp. 378–434, 2005, doi: 10.1007/s10458-005-2631-2.

[13] A. OroojlooyJadid and D. Hajinezhad, "A Review of Cooperative Multi-Agent Deep Reinforcement Learning," Aug. 2019, [Online]. Available: http://arxiv.org/abs/1908.03963

[14] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual Multi-Agent Policy Gradients," May 2017, [Online]. Available: http://arxiv.org/abs/1705.08926

[15] G. Papoudakis, F. Christianos, L. Schäfer, and S. V. Albrecht, "Benchmarking Multi-Agent Deep Reinforcement

Learning Algorithms in Cooperative Tasks," *Conference on Neural Information Processing Systems (NeurIPS)*, Jun. 2021.

[16] V. Mnih *et al.*, "Asynchronous Methods for Deep Reinforcement Learning," *International Conference on Machine Learning (ICML)*, Feb. 2016.

[17] J. Su, S. Adams, and P. A. Beling, "Value-Decomposition Multi-Agent Actor-Critics," *AAAI Conference on Artificial Intelligence (AAAI)*, Jul. 2020.