Northumbria University NEWCASTLE

A Scalable Evaluation Framework for Intelligent Agents

Name: Mohanraj Muthumanickam

Supervisors: Lynne Connis, Tom Bartindale

W22056353
MSc in Data Science
KF7029 / MSc Computer Science and Digital Technologies Project
Northumbria University

INTRODUCTION

Reinforcement Learning (RL) It's a powerful form of an artificial intelligence that learns by engaging the environment to achieve certain goals. Although it is mostly used in single-agent problems it is more useful in the multi-agent scenarios where many agents can cooperate or compete. This transition birthed **Multi-Agent Reinforcement Learning (MARL)** since all the agents learn in the same environment. It is also important to note that, unlike in the cooperative multi-agent systems, the MAS and the agents in the competitive MASs work not only for their own benefits pursuing the maximization of their own rewards but also the minimization of their rivals' rewards. This dynamic is particularly apparent in use cases such as the **automated warehouse**, in which multiple robotic agents have to operate, for example, to pick shelves while avoiding obstacles and pauses.

This study compares two key MARL algorithms: **MADDPG with DQN**. However, it is applicable for continuous action spaces and decentralized learning where an agent can learn from others' actions using **MADDPG**. **DQN** that is a value based algorithm is pretty well suited to discrete action spaces. Based on the RWARE environment which models the warehouse activities, both algorithms are compared and tested in terms of learning stability, adaptability, computational complexity and output efficiency. The goal is to find out which algorithm works more effectively in conditions that can be described as ''adversarial' dealing with other agents, in the context of **automating a warehouse** and similar scenarios.

OBJECTIVES

- Efficient and effective evaluation of performance of intelligent agents is essential to establishing their reliability as the use of intelligent agents increases in number.

- In specific, the key objectives of this research could be formulated as: To analyse the two MARL algorithms such as **MADDPG and DQN**, with the focus on the learning efficiency, stability, adaptability, and computational complexity measures and to compare their performance specifically in terms of the mentioned aspects and within the context of the multi-agent, competitive scenarios and the RWARE environment.

METHODOLOGY

MADDPG:

In **MADDPG**, the rewards system is used in the training of the agents to enable them to improve on their strategies in a multi-agent scenario. Since the action space that is addressed by MADDPG is continuous, the rewards are very flexible and the agents are allowed to alter their actions in accordance to feedback from their interactions with the environment, and other agents.

Agents receive positive rewards for successfully completing tasks, such as delivering shelves, while negative rewards (penalties) are assigned for inefficient actions like collisions, unnecessary movements, or delays in task completion. The reward system can also be tailored to either cooperative or competitive settings: Here rewards system is tailored in competitive environments, where individual rewards are assigned based on each agent's ability to outperform others. This flexibility in the reward structure enables agents to optimize their behavior accordingly.

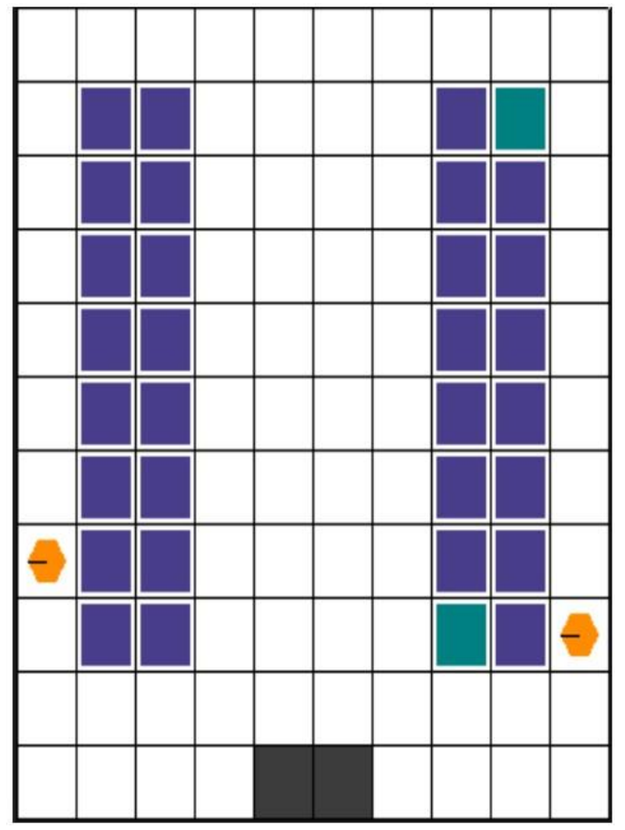

Fig1. Rware environment

DQN

In **DQN** case rewards are straightforward as the method deals only with discrete actions. Q-value used in DQN is the values of the future rewards corresponding to the possible actions from a given state. This value helps the agent choose actions, as it is attempting to optimize the cumulative reward function in the future.

In **DQN**, the reward system consists of positive feedbacks that are issued after the accomplishment of tasks, for instance, delivering a shelf and negative feedbacks or penalties for inefficient movements such as collisions, unnecessary movements, or slow movements. DQN uses an epsilon-greedy policy to mix up exploration (meaning acting entirely at random and choosing an arbitrary action no matter what its Q values) and exploitation (this is choosing the best-known action according to Q-values). After each action, DQN adjusts the Q-values, which are the value functions obtained, with the help of the rewards predicted and the actual rewards which are obtained, this helps the agent to learn about the best action to take with a view of producing the best results in the next times.

RWARE ENVIRONMENT

Rware is a simulation, which is used for research purposes and is a dynamic simulation model. RWARE mimics a workplace environment with activities taking place simultaneously for some of the robots to perform at the same time.

RESULTS AND DISCUSSIONS

In the first graph (MADDPG), both of the agents demonstrate consistent learning over time as smoothed rewards become closer to a peak, or a value of about 160 for either graphs. However, the second graph (DQN) indicates instability with both agents' rewards ranging between 2 and 9 in episodes, which signifies inconsistent learning and performance. MADDPG is found to have better learning behaviour compared to the DQN.

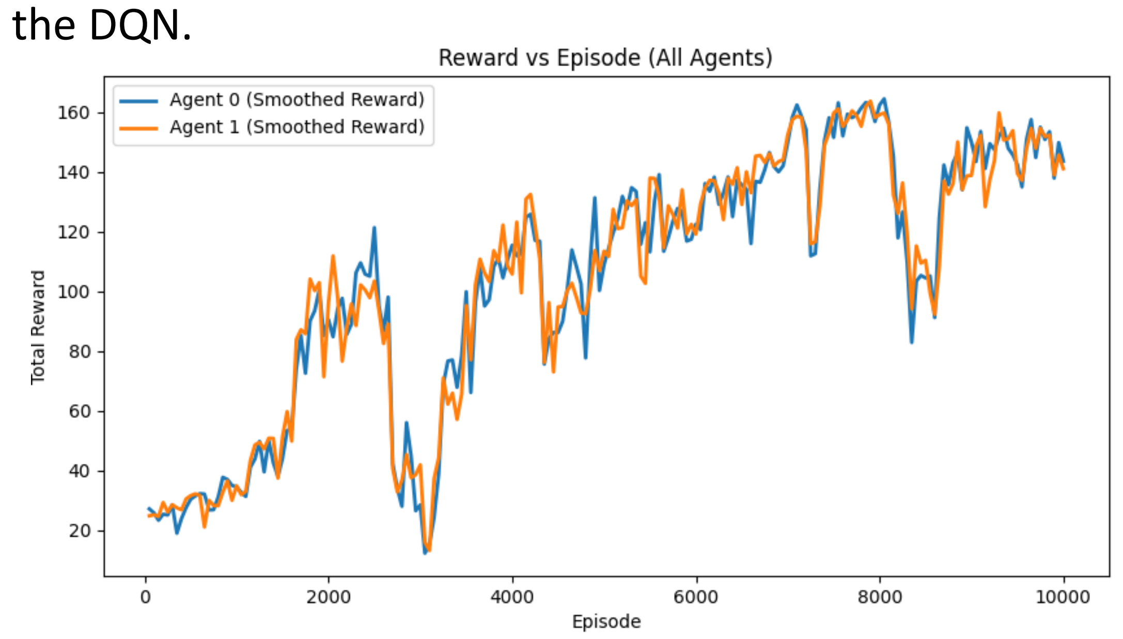

Fig 2: Smoothed Total Reward vs. Episode (All Agents) for MADDPG Algorithm

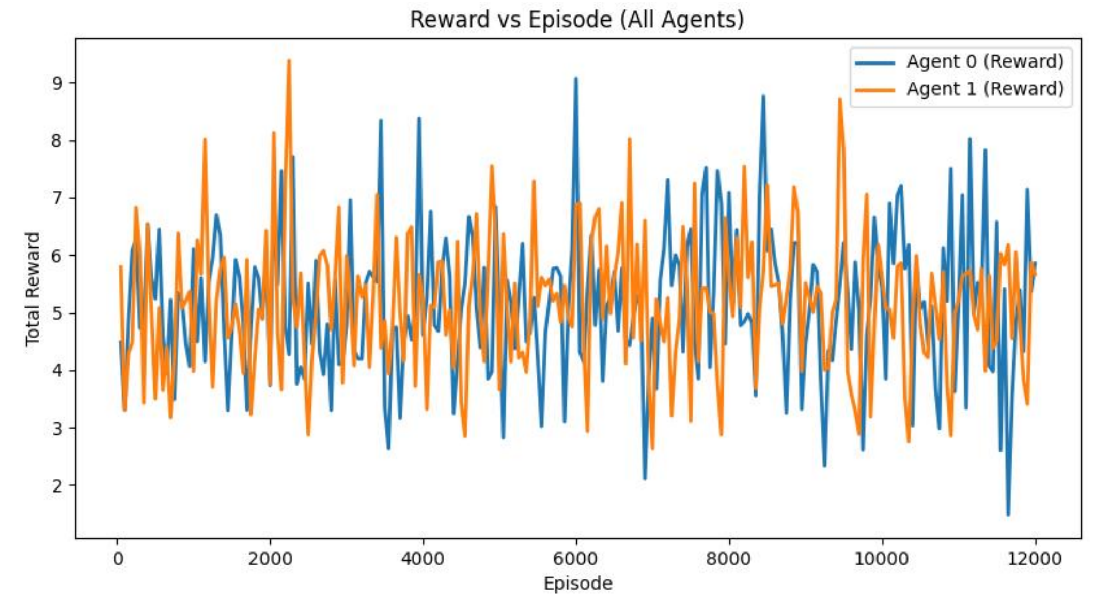

Fig 3: Total Reward vs. Episode (All Agents) for DQN Algorithm

FUTURE WORKS

- MADDPG shows better stability, adaptability, and faster convergence in competitive multi-agent environments. DQN, while computationally lighter, struggles with stability and performance in non-stationary, multi-agent systems.

- Future work could focus on optimizing the computational efficiency of MADDPG and exploring hybrid approaches combining the strengths of both algorithms.

BIBLIOGRAPHY

- Papoudakis, G., Christianos, F., Schäfer, L. and Albrecht, S. V. (2020) 'Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks', arXiv preprint arXiv:2006.07869.
- Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O. and Mordatch, I. (2017) 'Multi-agent actor-critic for mixed cooperative-competitive environments', Advances in neural information processing systems, 30.