MULTI-AGENT REINFORCEMENT LEARNING WITH CENTRALIZED CRITIC IN COLLABORATIVE ENVIRONMENTS

CONTACT

Author: Andrei-Ioan Mija (A.I.Mija@student.tudelft.nl)
Supervisor: Robert Loftin
Responsible Professor: Frans Oliehoek

AFFILIATIONS

TU Delft

1.INTRODUCTION

- Reinforcement learning to train agents in multi-agent collaborative environments through self-play
- In a multi-agent environment, training each agent individually is problematic: all agents learn at once → policies change → non-stationary environment
- Multi-agent with centralized critics → agents' policies become part of the environment → stationary environment.
- Self-play: good results when evaluated with itself, poor results with new partners

2.OBJECTIVE

"Does a multi-agent reinforcement learning algorithm with centralized critics generalize better to new partners compared to a single-agent approach in a collaborative environment?"

We will look at:

- performance during training
- level of generalization

3.METHODOLOGY

Environment:

- Simplified version of the Overcooked game in [1] (Figure 1)

Algorithms:

- Multi-agent with Centralized critics: MAPPO
- Single-agent: PPO [2]
- Behavior cloning (BC) (using human data)

Experimental flow:

- Train PPO & MAPPO agents through self-play
- Compare the performance during training
- Evaluate the performance of algorithms in self-play
- Train BC agent using human data
- Pair PPO & MAPPO with a human model to obtain the level of generalization

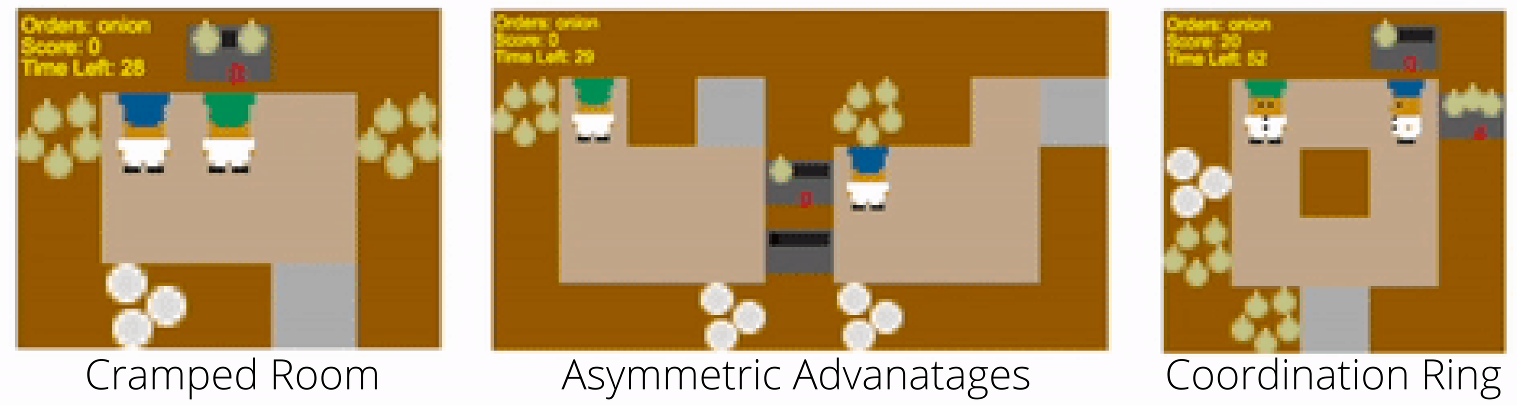

Figure 1: Overcooked layouts used in the study. Source [1]

4.RESULTS

MAPPO (orange) converges twice as fast to an optimal reward value when compared to PPO (blue) (values displayed in millions of timesteps) as per Figure 2:

- Cramped room: ~0.5m (MAPPO) vs. ~1m (PPO)
- Asymmetric Advantages: ~0.6m (MAPPO) vs. ~1.2m (PPO)
- Coordination Ring: ~1m (MAPPO) vs. ~2m (PPO)

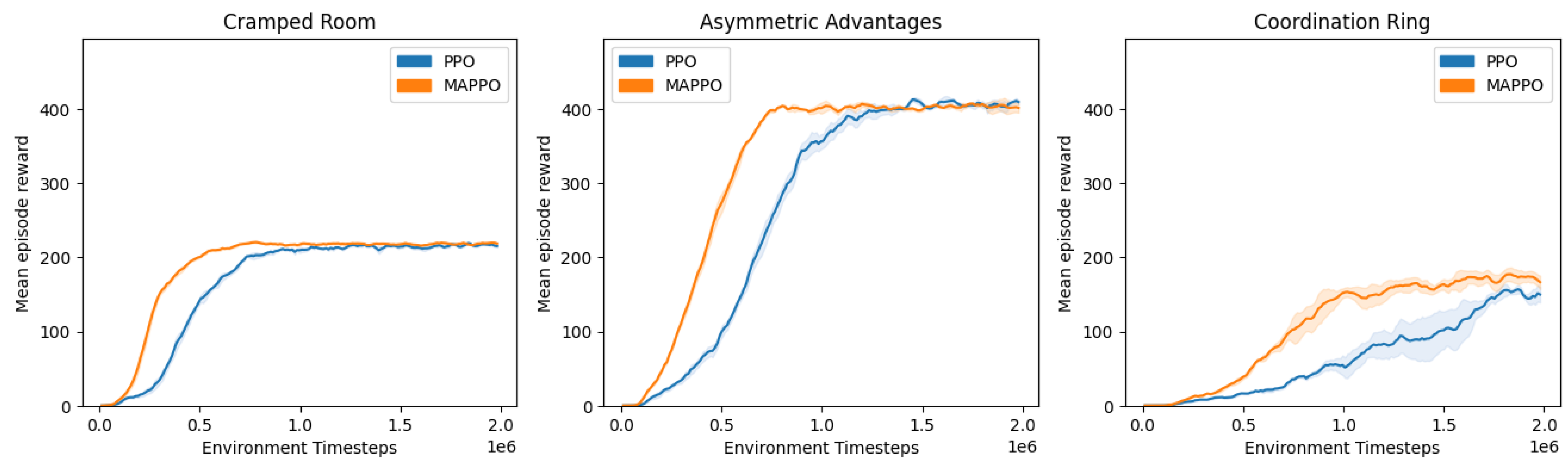

Figure 2: Mean Episode Reward during the training for PPO (blue) and MAPPO (orange) agent pairs

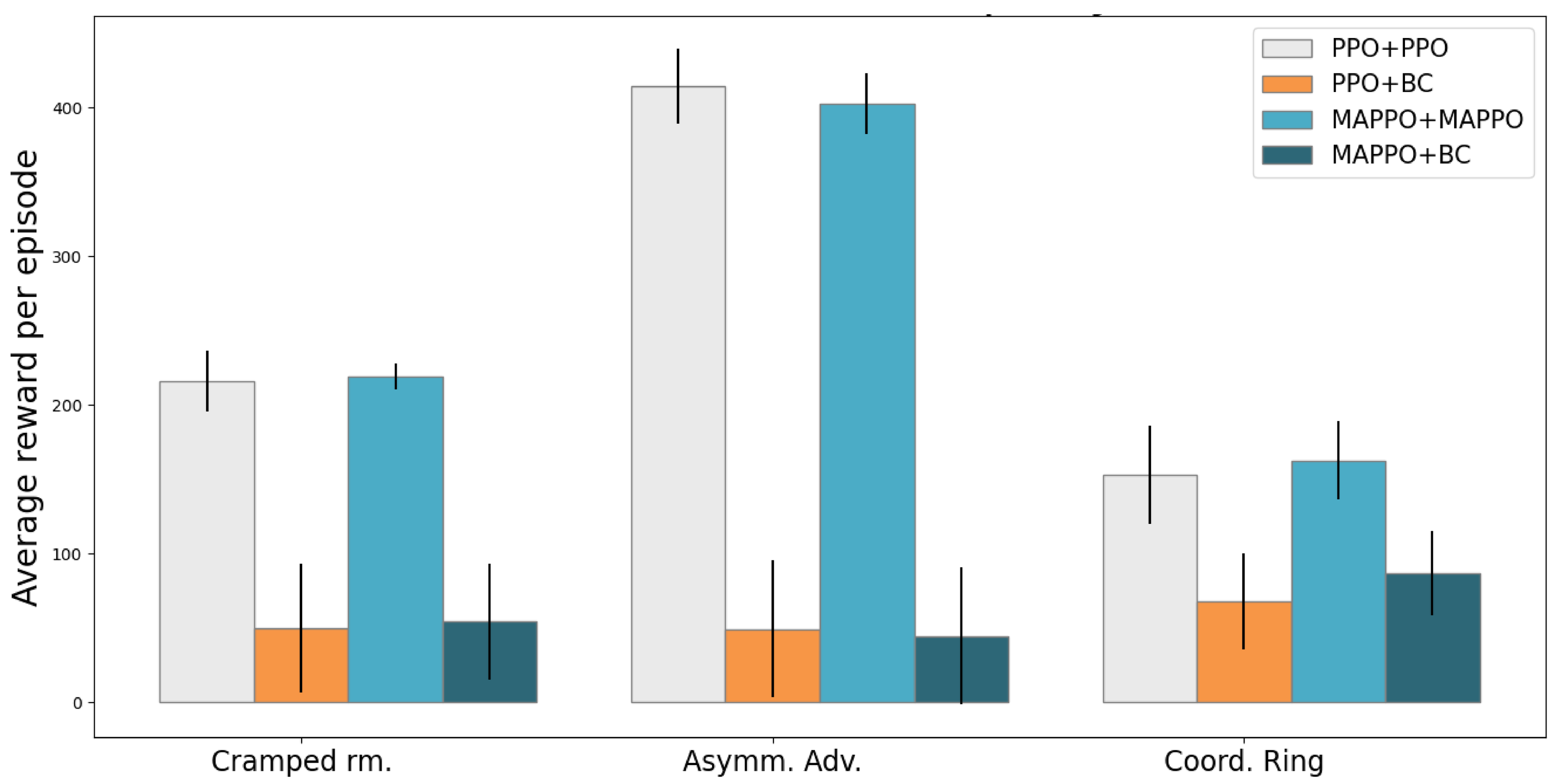

Figure 3: Mean Episode Reward during evaluation for different agent pairs: PPO-PPO (light grey), PPO-BC (orange), MAPPO-MAPPO (light blue) and MAPPO-BC (dark blue)

The multi-agent algorithm with centralized critics does not generalize better than the single-agent one as per Figure 3:

- Depending on the layout, one algorithm performs better than the other
- Whenever an algorithm performs better than the other, the difference in results is negligible
- Difference in performance due to seed initialization
- MAPPO consistently provides a smaller variance than PPO

5.CONCLUSIONS & FUTURE WORK

Summary:

- The centralized critics algorithm does not result in a better level of generalization when compared to its single-agent counterpart
- The multi-agent algorithm trains models twice as fast as the single-agent approach and shows a more consistent performance over all layouts

Future work:

- Use a more complex and potentially more generalizable observation space
- Implement a visual representation for the evaluation to better observe the agent's behaviour
- Use more agent types while training

RELATED LITERATURE

[1] M. Carroll, R. Shah, M. K. Ho, et al., "On the utility of learning about humans or human-ai coordination," CoRR, vol. abs/1910.05789, 2019. arXiv: 1910.05789.
[2] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," CoRR, vol. abs/1707.06347, 2017. arXiv: 1707.06347.