Reinforcement in Cooperative Games

Deep Learning Approaches

Konstantinos Bardis

May 2022

Table of Contents

1. Preliminaries
2. Experiments and Discussion

Table of Contents

1. Preliminaries
2. Experiments and Discussion

Motivation

- Why bother?
  - Loads of unique applications e.g robotics, vehicle coordination (UAVs), smart appliances, finance, ... made possible
- R&D problems?
  - Many: Establishing objective, non stationarity, curse of dimensionality, API availability...

Decision-making Frameworks

- MDP: Single-agent, full observability
- POMDP: Single-agent, partial Observability
- Dec-POMDP: Decentralized POMDP, for multiple agents

Main Approaches

The primary algorithmic approaches to deal with MADRL problems are

- Centralized: learn a single, centralized policy to produce the joint actions of all agents simultaneously
- Decentralized: every agent optimizes their reward signal independently
- Centralized Training and Decentralized Execution (CTDE): improves upon Decentralized training by learning a centralized critic and using decentralized policies with only local observations during execution

Table of Contents

1 Preliminaries

2 Experiments and Discussion

Overview

- Several (6) environments with focus on cooperation: MPE, Atari & Butterfly suites, obtained from PettingZoo
- Several SOTA Algorithms: PPO, A2C, A3C, IMPALA, SAC
- Discrete Action Spaces
- 2 Learning variants per algorithm: Centralized & Decentralized

Training

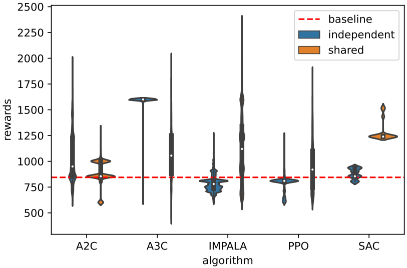
- Hyperparameter tuning with Bayesian Optimization and schedulers to kill bad performing trials early
- Distributed training via RLLib, with several (10) instances of notebooks running in parallel in Kaggle, 4 CPUs each
- Function approximators: Specialized transformer architecture (GTrXL) with GRU cells

Results

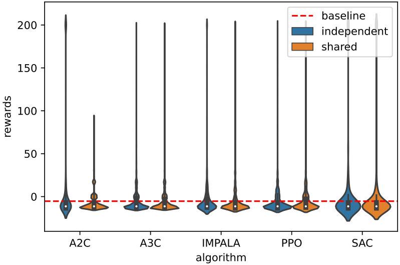
- Overall, MPE very challenging, Atari better, Butterfly worst
- Overall, PPO and SAC extremely slow compared to the others; IMPALA the fastest and with very good comparative performance generally. A2C/A3C also fast but with more mixed performances
- PPO in particular poor performance in all showings, perhaps sensitive to hyperparams
- SAC generally with little variance and often good scores
- IMPALA often (not always) very good performance, several times the best
- A2C/A3C also often good but occasionally very high variance and underperformance
- Overall, no clear winner between the two paradigms

Selected Results

Space Invaders - good performances



Cooperative Pong - awful performances



Future Directions

- Causal inference for better generalization
- Procedural generation for effective evaluation + more APIs generally
- Better theory for dealing with POMDPs
- Curriculum Learning, Hierarchical Approaches