Analysis of Emergent Behavior in Multi-Agent Environments

Stefanie Anna Baby (stef96)[1], Ling Li (lingli6)[2], Ashwini Pokle (ashwini1)[1]
[1]Dept. of Computer Science, [2]Dept. of Electrical Engineering, Stanford University

CS 234
Winter 2018

Motivation

- Multi-agent environments provide scope for evolution of behaviors like coordination/competition
- This can give rise to emergent phenomena without explicit control of agents
- This project analyses how such behaviors arise and vary

Approaches

- Parameter-Sharing DQN (PS-DQN), PS-DDQN, PS-DRQN
- DQN from expert demonstrations, Prioritized experience replay
- Proximal Policy Optimization (PPO)

Environment

- Games - Pursuit, Battle (MAgent Platform), Gathering
- Partially observable environment; agent has a circle of visibility
- In Pursuit, predators have to cooperate to trap prey; prey try to evade predators
- In Battle, equally competent agents learn to cooperate to kill and defeat the opponents
- In Gathering, two agents compete for resources like food

Parameter Sharing DQN (PS-DQN) and Variants

- A DQN is trained with experiences of all agents of one type
- Each agent receives a different observation and an agent id

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(r + \gamma \max_{a'} Q(s',a';\theta_i^-) - Q(s,a;\theta_i)\right)^2\right]$$

- Hyperparameters: Learning rate 1e-4, experience replay memory 2^{22}, Huber Loss, Adam Optimizer
- DRQN replaces first fully connected layer of DQN with LSTM; this helps it to adapt to non-stationary environment
- DQN with expert demonstrations initializes experience replay buffer with demonstrations from an expert agent
- Prioritized experience replay samples transitions with high expected rate of learning (TD error) more frequently

Proximal Policy Optimization (PPO)

- Maximizes Surrogate Objective

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right]$$

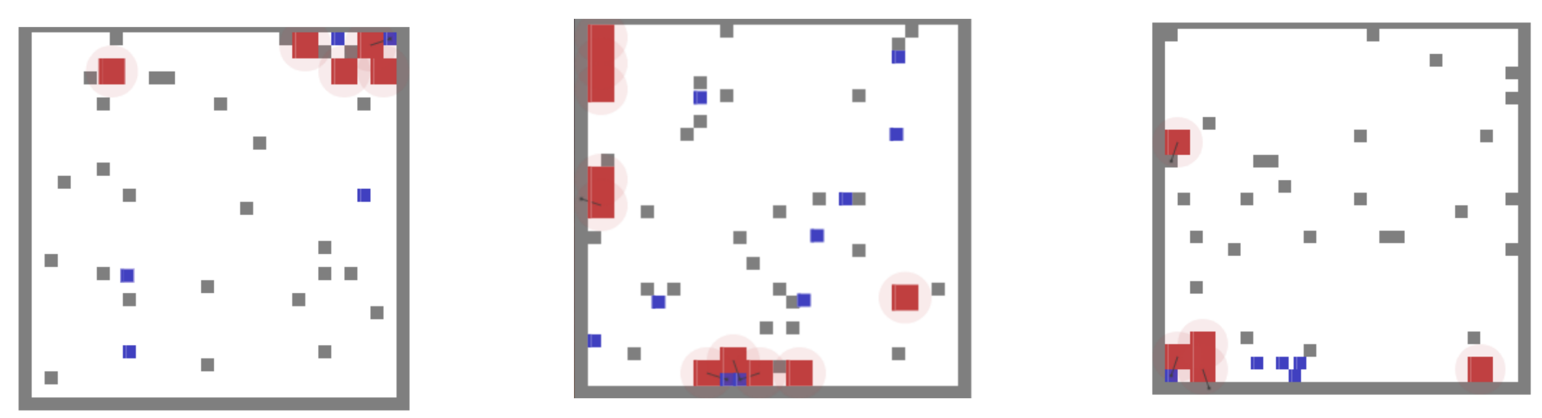
where, $$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$

$\hat{A}_t$ is generalized advantage estimate

- Hyperparameters : epsilon 0.2, Learning rate 1e-4, Adam optimizer
- Clipped probability ratios form a pessimistic estimate of the performance of policy

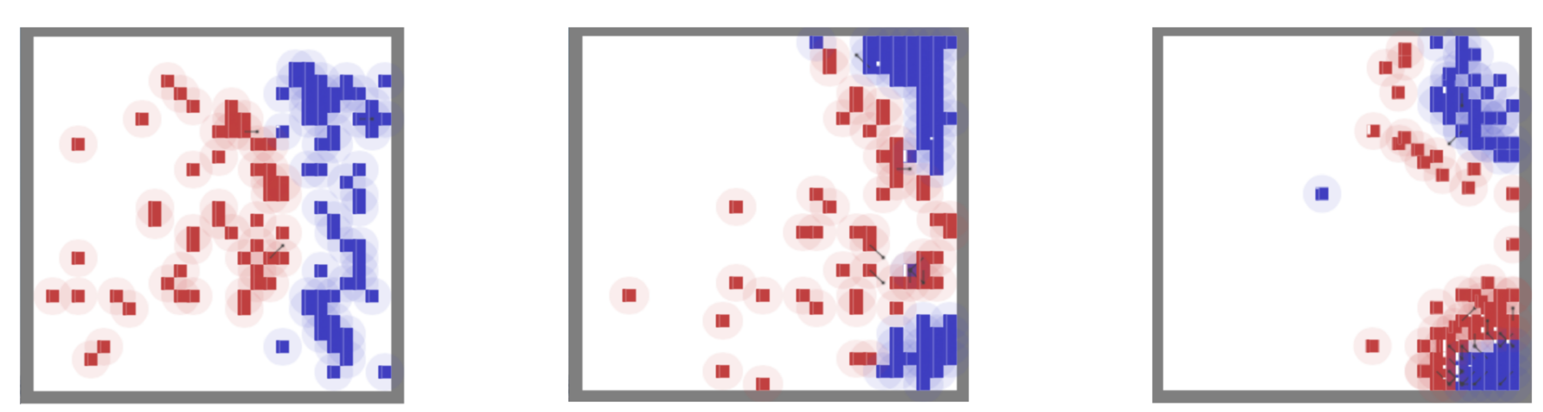
Emergence of Complex Behavior

- **Pursuit**

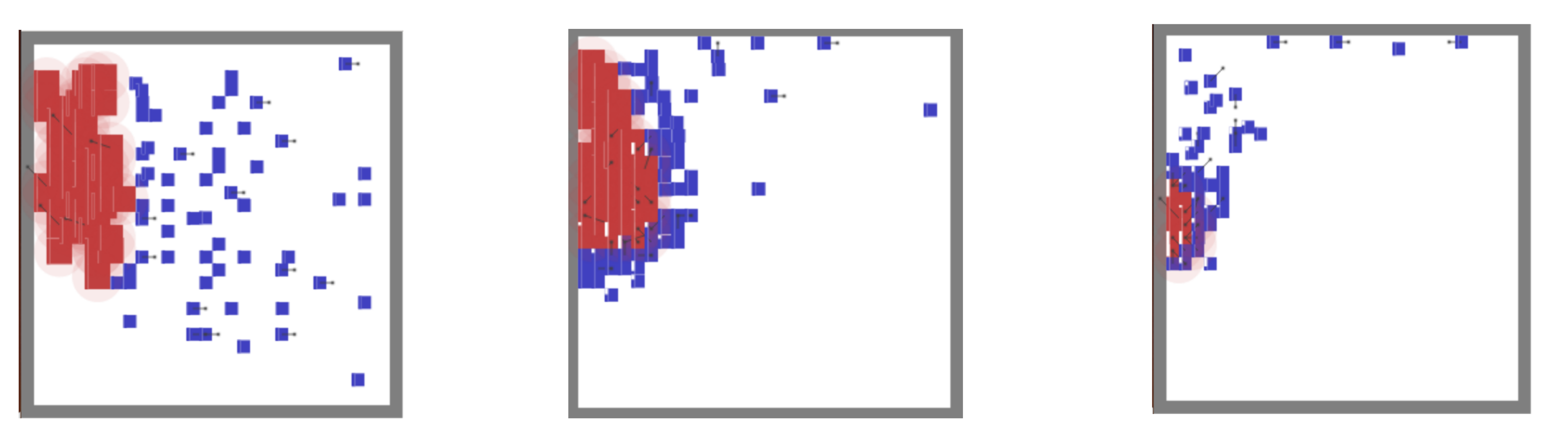


- Strategy : Predators form enclosures to trap prey
- Escape strategies were co-evolved by the prey simultaneously

- **Battle**



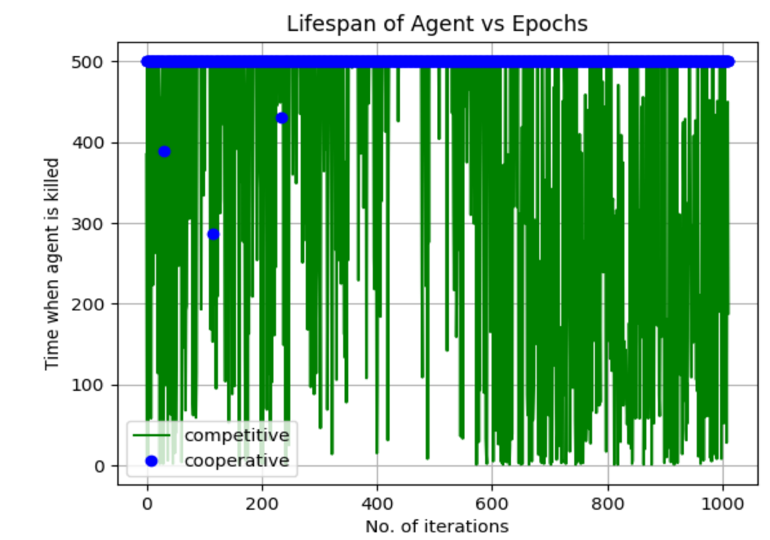
- Strategy : Red agents learned to split and trap blue agents



- Strategy : Blue agents learned to trap red agents
- Defense strategies like escaping an entrapment were learned by the agents as well

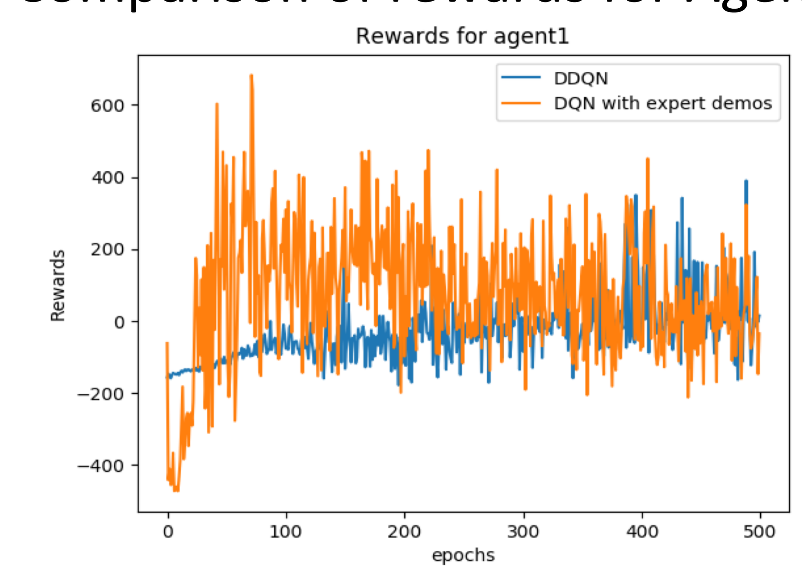
Results

- Emergence of adversarial and non-adversarial behavior in Gathering

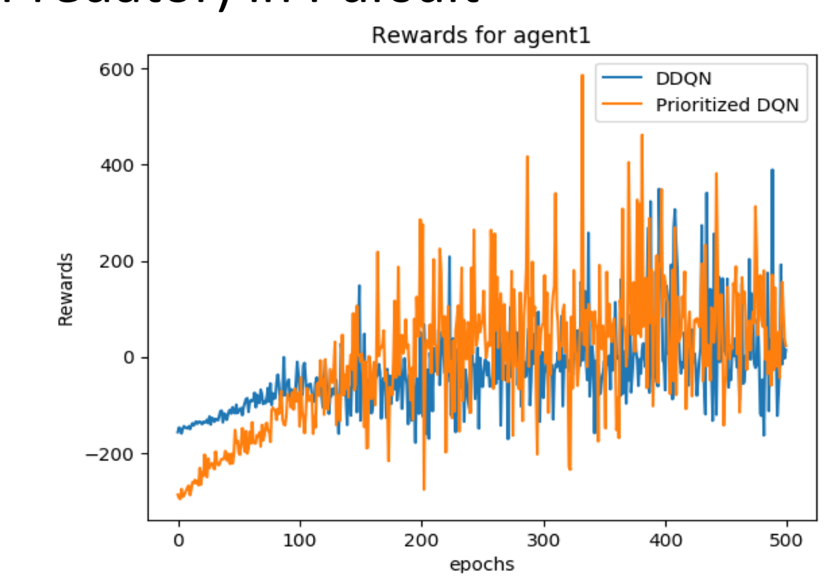


- Scarcity of food caused adversarial behavior, motivating an agent to shoot its opponent
- Abundance of food allowed agents to coexist with minimal shooting

- Comparison of rewards for Agent 1 (Predator) in Pursuit

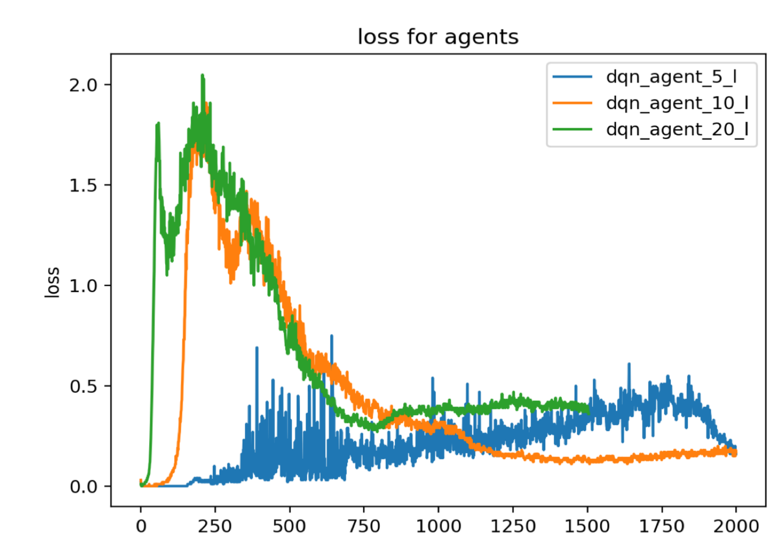


DDQN vs Prioritized DDQN



DDQN vs training with expert demo

- Variation of rewards and loss for blue agents in Battle with team size



Future Work

- Robust analysis of performance of PPO on Gathering environment
- Analyze reasons for negative transfer in case of training DQN with demos

References

[1] Gupta, Jayesh et. al. "Cooperative multi-agent control using deep reinforcement learning." AAMAS 2017.
[2] Zheng, Lianmin, et al. "MAgent: A Many-Agent Reinforcement Learning Platform for Artificial Collective Intelligence." *arXiv:1712.00600* (2017).
[3] Leibo, Joel Z., et al. "Multi-agent reinforcement learning in sequential social dilemmas." AAMAS, 2017.