Behavior Exploration and Team Balancing for Heterogeneous Multiagent Coordination

Extended Abstract

Gaurav Dixit
Oregon State University
Corvallis, USA
dixitg@oregonstate.edu

Kagan Tumer
Oregon State University
Corvallis, USA
kagantumer@oregonstate.edu

ABSTRACT

Diversity in behaviors is instrumental for robust team performance in many multiagent tasks which require agents to coordinate. Unfortunately, exhaustive search through the agents' behavior spaces is often intractable. This paper introduces Behavior Exploration for Heterogeneous Teams (BEHT), a multi-level learning framework that enables agents to progressively explore regions of the behavior space that promote team coordination on diverse goals. By combining diversity search to maximize agent-specific rewards and evolutionary optimization to maximize the team-based fitness, our method effectively filters regions of the behavior space that are conducive to agent coordination. We demonstrate the diverse behaviors and synergies that are method allows agents to learn on a multiagent exploration problem.

KEYWORDS

Adaptive Team Balancing; Quality Diversity; Multiagent Systems

ACM Reference Format:
Gaurav Dixit and Kagan Tumer. 2022. Behavior Exploration and Team Balancing for Heterogeneous Multiagent Coordination: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 2 pages.

1 INTRODUCTION

The ability to reason about the behavior of other agents and adapt is a key component of cooperative behavior in multiagent settings. To work on diverse tasks as a team, agents must specialize and learn complementary behaviors. This presents a need to systematically explore the behavior space and discover specializations that are useful for team coordination.

By evolving a repertoire of behaviors, Quality-Diversity (QD) methods offer a gradient-free evolutionary method to favour diversity over mere optimization [2]. In multiagent settings however, exploring the space of all possible behaviors is intractable due to the potentially large behavior space imposed by multiagent interaction. Moreover, when useful diverse behaviors are discovered by different agents, the relative number of agents with distinct behaviors must also be balanced to maximize the team-wide fitness.

This work introduces Behavior Exploration for Heterogeneous Teams (BEHT), a multi-level training framework for systematic exploration of agents' behavior space to discover diverse behaviors required for coordination. A gradient-based optimizer trains a population of policies to optimize agent-specific rewards. The policies are projected to a behavior space that is inferred by applying a dimensionality reduction method to the policy trajectories. A gradient-free evolutionary algorithm (EA) evolves the policy population to maximize the team-wide objective. The selection of high-fitness teams by the evolutionary process, filters the regions of the behavior space that contribute the most to the team fitness. The trajectories of the policies from the filtered behavior space are then used to relearn the behavior space. This space captures the variance of the high-fitness behaviors and allows to move in the direction of diversity that maximizes the team fitness. *The key insight of this work is that decoupling the search for diversity from fitness optimization (along with iterative refinement of the behavior space) enables the performance-focused team-wide objective to also direct the search for diversity.*

We demonstrate the strength of BEHT on a multiagent rover exploration problem with sparse feedback and a diverse set of goals that requires tight agent coupling. Our approach shows significant performance improvement over current state-of-the-art diversity search methods in multiagent problems.

2 BEHAVIOR EXPLORATION FOR HETEROGENEOUS TEAMS (BEHT)

Behavior Exploration for Heterogeneous Teams (BEHT) is an iterative processes for agent diversity search and team fitness optimization. The iterative process consists of three phases: diversity search, team optimization and behavior space refinement.

BEHT starts with a population of policies trained on dense agent-specific rewards using Proximal Policy Optimization (PPO). The policy trajectories (any data that characterizes the policies can be used) are used as dataset for learning a latent representation using Principal Component Analysis (PCA)[2]. This low dimensional representation is used as the behavior space.

Diversity search: With the agent policies projected in the behavior space, a QD like iteration is performed that involves a policy selection, mutation by perturbing the weights of the policy network, training using PPO and projection in the behavior space. Two key differences set this QD phase apart from a standard QD iteration [1]: 1) A policy mutation is followed by training on the dense agent-specific reward; 2) The behavior descriptor used to project the policies in the behavior space is determined by feeding their trajectories to the trained PCA model. Over the course of several iterations in the QD phase, the mutations progressively fill the behavior space.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online. © 2022 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

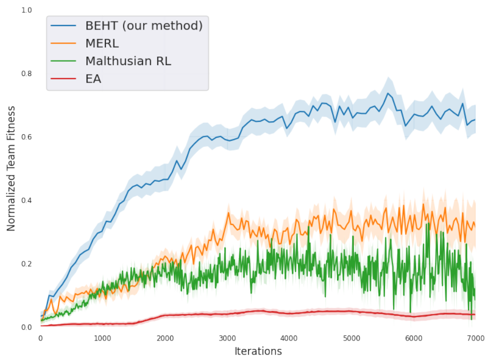

Figure 1: Team fitness on the tightly coupled rover problem.

Team optimization: The next phase of BEHT uses an evolutionary algorithm (EA) to train the policies from the QD phase on the sparse team-wide fitness. A population of teams is created by sampling a set of polices from the policy population. The teams are evaluated on the team-wide objective and are assigned the team reward as fitness at the end of each episode. The EA's selection operator selects a portion of the teams for survival with probability proportional to their fitness. The weights of the policies in the selected teams are probabilistically perturbed through mutation and crossover operators to create the next generation of teams.

After several evolutionary updates, the EA retains the teams with the best performing policies and thus essentially filters out regions of the behavior space that have the highest fitness on the team objective. By implicitly adjusting the population dynamics of teams and policies within the teams via fitness, the EA also balances the number of different behaviors needed for optimal team performance.

Behavior space refinement: This phase updates the latent representation of the policies, the behavior space, to take into account the new high-fitness behaviors. The behavior space is updated (using trajectories of the new policies as the dataset for PCA) to align with the maximum variance (diversity) of the highest performing policies. This update often moves originally distant policies closer together, in which case, the policy with the highest agent-specific reward will be retained. This is a safe replacement since closeness in the behavior space implies that the policies behaved similarly in teams. The next QD phase will now explore the updated behavior space which likely offers regions that were not already covered in the previous phase.

3 EXPERIMENT

We evaluate the performance of BEHT on several scenarios created using a variant of the multiagent heterogeneous rover exploration problem that requires several rovers to simultaneously observe multiple points of interests (POIs) [3]. We introduce three POI variants to add diversity to the goals: 1) A Vanilla POI that rewards agents with a fixed value when observed; 2) a Timed POI which has a constantly declining reward; and 3) a Low-Power POI that is observable stochastically. The trajectory of the policy fed to PCA is a vector of $(o_t, v_t, d_{r,t}, d_{p,t})$ tuples for every time step t, where o_t is the observation radius of the agent, v_t is the speed, $d_{r,t}$ and $d_{p,t}$ are the distances to the closest rover and POI respectively. The tuple captures the rover's POI strategy (which POIs to observe) and the rover's coordination strategy at time t.

We compare our method with two baselines: 1) Multiagent Evolutionary Reinforcement Learning (MERL) - a two-tier architecture the combines gradient-based and EA optimization for learning in multiagent settings [4]; and 2) Malthusian Reinforcement Learning - a method for discovering agent synergies via population dynamics [5]; Our method, combines the strengths of both by optimizing for coordination in tightly coupled settings via diversity search and implicit population management via an EA.

Figure 1 shows the normalized cumulative team fitness. There are 10 agents that start every episode from the center of the environment and 15 POIs (all variants) uniformly distributed throughout the environment. BEHT uses the Euclidean distance to the closest POI as the reward for the QD phase and the cumulative values of the observed POIs as the fitness for the EA. Agents using BEHT acquire sufficient diversity to observe all POI variants and learn to coordinate.

MERL uses the same rewards as BEHT, but learns a sub-optimal team strategy since it fails to account for the needed diversity in agent behavior. Malthusian-RL is configured to train five agent types on three islands. It performs sub-optimally since the team fitness is used for learning population dynamics on the islands but does not explicitly guide the direction of diversity search.

4 DISCUSSION

We introduced BEHT, a multi-level training framework for systematic exploration of the agents' behavior space, required to complete diverse tasks as a coordinated team. The decoupling of the diversity search and fitness optimization allows the transformation of team-wide fitness to direct the search for diversity. This allows agents to progressively explore promising regions of the behavior space that promote team coordination. In future work, we will explore extensions of BEHT to more complex mixed cooperative-competitive settings.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation under grant No. IIS-1815886 and the Air Force Office of Scientific Research under grant No. FA9550-19-1-0195.

REFERENCES

[1] Cédric Colas, Vashisht Madhavan, Joost Huizinga, and Jeff Clune. 2020. Scaling map-elites to deep neuroevolution. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*. 67–75.
[2] Antoine Cully. 2019. Autonomous skill discovery with quality-diversity and unsupervised descriptors. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 81–89.
[3] Gaurav Dixit, Nicholas Zerbel, and Kagan Tumer. 2019. Dirichlet-Multinomial Counterfactual Rewards for Heterogeneous Multiagent Systems. In *2019 International Symposium on Multi-Robot and Multi-Agent Systems (MRS)*. IEEE, 209–215.
[4] Shauharda Khadka, Somdeb Majumdar, Santiago Miret, Stephen McAleer, and Kagan Tumer. 2019. Evolutionary reinforcement learning for sample-efficient multiagent coordination. *arXiv preprint arXiv:1906.07315* (2019).
[5] Joel Z. Leibo, Julien Perolat, Edward Hughes, Steven Wheelwright, Adam H. Marblestone, Edgar Duéñez Guzmán, Peter Sunehag, Iain Dunning, and Thore Graepel. 2019. Malthusian Reinforcement Learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1099–1107.