Emergence of Scenario-Appropriate Collaborative Behaviors for Teams of Robotic Bodyguards

Extended Abstract

Hassam Ullah Sheikh and Ladislau Bölöni
Department of Computer Science University of Central Florida
Orlando, Florida
hassam.sheikh@knights.ucf.edu,lboloni@cs.ucf.edu

ABSTRACT

We are considering the problem of controlling a team of robotic bodyguards protecting a VIP from physical assault in the presence of neutral and/or adversarial bystanders in a variety of scenarios. This problem is challenging due to the large number of active entities with different agendas and dynamic movement patterns, the need of cooperation between the robots as well as the requirement to take into consideration criteria such as social norms in addition to the main goal of VIP safety. We show how a multi-agent reinforcement learning approach can evolve behavior policies that outperform hand-engineered approaches. Furthermore, we propose a novel multi-agent reinforcement learning algorithm inspired by universal value function approximators that can learn policies that exhibit appropriate, distinct behavior in environments with different requirements.

KEYWORDS

Multi-Agent Reinforcement Learning; Robot Team Formation; Multi-Robot Systems

ACM Reference Format:
Hassam Ullah Sheikh and Ladislau Bölöni. 2019. Emergence of Scenario-Appropriate Collaborative Behaviors for Teams of Robotic Bodyguards. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Recent progress in the field of autonomous robotics makes it feasible for robots to interact with multiple humans in public spaces. In this paper, we are considering a practical problem where a human VIP moving in various crowded scenarios is protected from physical assault by a team of bodyguard robots. This problem has been previously explored in [1] where explicitly programmed behaviors of robots were used to carry out the task.

With the recent advancements in the single agent Deep RL [6, 10], there has been a renewed interest in multi-agent reinforcement learning (MARL) [3, 4, 8]. Despite having outstanding performance in multiplayer games like Dota 2 [8] and Quake III Capture-the-Flag [3], MARL algorithms have failed to learn policies that can work in different scenarios [2].

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

Providing physical protection to a VIP through robot bodyguards is a complex task where the robots must take into account the position and movement of the VIP, the bystanders and other robots. The variety of environments and scenarios where the bodyguards need to act presents another challenge. We aim to solve the VIP protection problem through multi-agent deep reinforcement learning while simultaneously learning to communicate and coordinate between the robots. We propose a novel general purpose technique that allows multi-agent learners to learn distributed policies not only over the state space but also over a set of scenarios. We show that our solution outperforms a custom designed behavior, the quadrant load balancing method [1].

2 THE VIP PROTECTION PROBLEM

We are considering a VIP moving in a crowd of bystanders $\mathcal{B} = \{b_1, b_2, \ldots, b_m\}$ protected from assault by a team of robot bodyguards $R = \{r_1, r_2, \ldots, r_n\}$. To be able to reason about this problem, we need to quantify the *threat* to the VIP at a given moment - the aim of the bodyguards is to reduce this value.

Using the threat model defined in [1], the residual threat RT is defined as the threat to the VIP protected by the bodyguards $\mathcal{R}$ from a bystander b. The *cumulative residual threat* to the VIP over the time period $[0, T]$ is defined as:

$$\mathrm{CRT} = \int_0^T 1 - \prod_{i=1}^{k} \left(1 - RT\left(VIP, b_i, R\right)\right) dt \tag{1}$$

Our goal is to use multi-agent reinforcement learning to find a bodyguard behavior that minimizes CRT. Moreover, eq. (1) also forms the basis of our reward function for the VIP protection problem.

3 MULTI-AGENT UNIVERSAL POLICY GRADIENT

To solve the VIP protection problem under various scenarios, we propose *multi-agent universal policy gradient*: a multi-agent deep reinforcement learning algorithm that learns distributed policies not only over state space but also over a set of scenarios.

Our approach uses Universal Value Function Approximators [9] to train policies and value functions that take a state-scenario pair as input. The outcomes are universal multi-agent policies that are able to perform better on multiple scenarios compared to policies trained and tested separately.

The main idea is to represent the value function approximators for each agent i by a single unified value function approximator that generalizes over both state space and the scenarios. For agent

i we consider $V_i(s, g; \phi) \approx V^*_{i_g}(s)$ or $Q_i(s, a, g; \phi) \approx Q^*_{i_g}(s, a)$ that approximate the optimal unified value functions over multiple scenarios and a large state space. These value functions can be used to extract policies implicitly or as critics for policy gradient methods. We extend the idea of MADDPG [5] with a universal functional approximator, specifically we augment the centralized critic with the scenario. Consider N agents with policies $\boldsymbol{\pi} = \{\boldsymbol{\pi_1}, \ldots, \boldsymbol{\pi_N}\}$ parameterized by $\boldsymbol{\theta} = \{\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_N}\}$ learning polices over $\boldsymbol{G}$ scenarios. The *multi-agent universal policy gradient* for agent i can written as

$$\nabla J_{\theta_i} = \mathbb{E}_{s,a,g\sim\mathcal{D}}\left[\nabla_{\theta_i}\pi_i(a_i|o_i, g)\nabla_{a_i}Q^{\pi}_i(s, a_1, \ldots, a_N, g)\right] \quad (2)$$

where $s = (o_1, \ldots, o_N)$, $Q^{\pi}_i(s, a_1, \ldots, a_N, g)$ is a centralized action-value function that takes the actions of all the agents, the state of the environment and the scenario to estimate the Q-value for agent i, $a_i = \pi_i(o_i, g)$ is action from agent i following policy π_i in scenario g and $\mathcal{D}$ is the experience replay buffer.

4 EXPERIMENTS

To investigate the effectiveness of our proposed solution, we designed four scenarios inspired from real world situations of VIP protection and implemented them as behaviors in the Multi-Agent Particle Environment [7]. In each scenario, the participants are the VIP, four robot bodyguards and one or more classes of bystanders. The scenario description contains a number of *landmarks*, points on a 2D space that serve as starting point and destinations for the goal-directed movement by the agents. For each scenario, the VIP starts from the starting point and moves towards the destination landmark. The VIP exhibits a path following behavior, augmented with a social skill metric: when it is about to enter the personal space of a bystander, it will slow down or come to a halt.

(A) **Random Landmark:** landmarks are placed randomly in the area. The bystanders are performing random waypoint navigation using the landmarks as waypoints.
(B) **Shopping Mall:** landmarks representing shops are placed in fixed positions on the periphery of the area. The bystanders visit randomly selected shops.
(C) **Street:** The bystanders are moving towards waypoints that are outside the current area. However, due to their proximity to each other, the position of the other bystanders influence their movement described by laws of particles motion [11].
(D) **Pie-in-the-Face**: In this "red carpet" scenario one bystander takes an active interest in the VIP. The *Unruly* bystander breaks the limit imposed by the line and try to approach the VIP (presumably, to throw a pie in his/her face).

The observation of agent i is the physical state of the closest five bystanders in the environment and verbal utterances of all the agents $o_i = \left[x_{j,\ldots 5}, c_{k,\ldots N}\right] \in O_i$ where x_j is the observation of the entity j from the perspective of agent i and c_k is the verbal utterance of the agent k while g is represented as a 1-hot vector.

In order to verify the claim that MARL algorithms trained on specific scenario fail to generalize over different scenarios, we evaluate policies trained via MADDPG on a specific scenario and tested on different scenarios. From the results shown in Figure 1 we can see that MADDPG policies trained on specific scenarios performed poorly when tested on different scenarios as compared to when tested on same scenario with different seeds. To tackle the generalization problem, we train the agents using multi-agent universal policy gradient and compare with the results of scenario-dependant MADDPG policies and quadrant load balancing(QLB): a hand engineered technique to solve the VIP protection problem. The results can be seen in Figure 2.

	Random Landmarks	Shopping Mall	Street	Pie-in-the-face
Random Landmarks	1.39	1.80	0.39	0.02
Shopping Mall	3.49	1.57	0.40	0.03
Street	2.77	2.00	0.31	0.16
Pie-in-the-face	3.96	3.24	2.64	0.00

Figure 1: A confusion matrix representing the average residual threat values of MADDPG policies trained on a specific scenario when tested on different scenarios over 100 episodes.

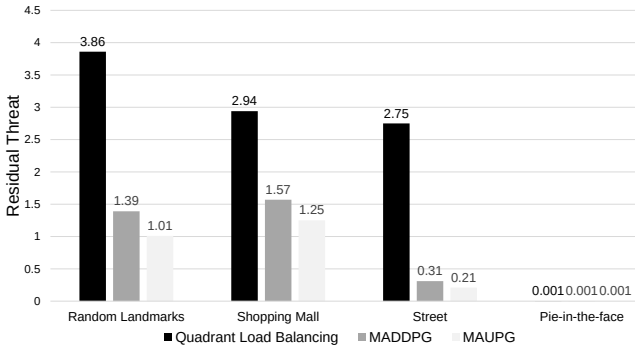

Figure 2: Comparing the average residual threat values for universal policy agents with MADDPG and QLB agents

5 CONCLUSIONS

In this paper, we highlighted the generalization problem faced by multi-agent reinforcement learning across different scenarios. To solve that problem we presented a novel algorithm that generalizes over both state space and scenarios. Using our solution, we solved the problem of providing physical protection to a VIP moving in a crowded space that outperforms state-of-the-art multi-agent reinforcement learning algorithm as well as quadrant load-balancing: a hand engineered technique to solve the VIP protection problem.

Acknowledgement: This research was sponsored in part by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0016 and in part by the National Science Foundation under grant number IIS-1409823. The views and conclusions contained in this document are those of the authors only.

REFERENCES

[1] T.S. Bhatia, G. Solmaz, D. Turgut, and L. Bölöni. 2016. Controlling the movement of robotic bodyguards for maximal physical protection. In *Proc. of the 29th International FLAIRS Conference*. 380–385.

[2] D. L. Cruz and W. Yu. 2014. Multi-agent path planning in unknown environment with reinforcement learning and neural network. In *IEEE Int'l Conf. on Systems, Man, and Cybernetics (SMC)*. 3458–3463.

[3] Max Jaderberg, Wojciech M. Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C. Rabinowitz, Ari S. Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Z. Leibo, David Silver, Demis Hassabis, Koray Kavukcuoglu, and Thore Graepel. 2018. Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. *arXiv preprint arXiv: 1807.01281* (2018).

[4] Siqi Liu, Guy Lever, Nicholas Heess, Josh Merel, Saran Tunyasuvunakool, and Thore Graepel. 2019. Emergent Coordination Through Competition. In *International Conference on Learning Representations*. https://openreview.net/forum?id=BkG8sjR5Km

[5] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems 30*. 6379–6390.

[6] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (26 02 2015), 529–533.

[7] Igor Mordatch and Pieter Abbeel. 2017. Emergence of Grounded Compositional Language in Multi-Agent Populations. *arXiv preprint arXiv:1703.04908* (2017).

[8] OpenAI. 2018. *OpenAI Five*. https://blog.openai.com/openai-five/.

[9] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. 2015. Universal Value Function Approximators. In *Proc. of the 32st Int'l Conf. on Machine Learning(ICML)*. 1312–1320.

[10] David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529 (2016), 484–503.

[11] Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. 1995. Novel type of phase transition in a system of self-driven particles. *Physical review letters* 75, 6 (1995), 1226.