Embracing Relational Reasoning in Multi-Agent Actor-Critic

Extended Abstract

Sharlin Utke
University of Warwick
Coventry, United Kingdom
sharlin.utke@warwick.ac.uk

Jeremie Houssineau
Nanyang Technological University
Singapore
jeremie.houssineau@ntu.edu.sg

Giovanni Montana
University of Warwick
Coventry, United Kingdom
g.monatana@warwick.ac.uk

ABSTRACT

Relational reasoning has become an important concept in machine learning and has seen notable progress in its methods like graph neural networks, which highlight the value of capturing intricate relational patterns. While it has shown promise in single-agent reinforcement learning, its potential in the multi-agent landscape remains largely uncharted. Our work aims to bridge this gap, demonstrating the advantages of integrating deep relational learning into multi-agent reinforcement learning. We do so by introducing an actor-critic architecture for centralized learning and decentralized execution that uses relational graph neural networks to imbue a spatial inductive bias. Empirical results highlight improved sample efficiency and asymptotic performance against strong baselines in cooperative tasks with significant spatial complexity.

KEYWORDS

relational reasoning, reinforcement learning, multi-agent systems

ACM Reference Format:
Sharlin Utke, Jeremie Houssineau, and Giovanni Montana. 2024. Embracing Relational Reasoning in Multi-Agent Actor-Critic: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024,* IFAAMAS, 3 pages.

1 INTRODUCTION

In reinforcement learning (RL), solving tasks successfully often requires knowledge of the connection between the agent and the environment objects. However, traditional RL techniques often focus on isolated states and actions, encountering challenges in environments characterized by rich relational structures. Recognizing this gap, some works aim for a more robust understanding of environments by infusing RL with relational reasoning [2–4, 8, 14, 17]. While most of the efforts have been conducted in single-agent settings, multi-agent systems seem to be an even more promising candidate: they require algorithms that can reason about the evolving interplay between agents as well as environment objects. These relationships are multifaceted. On one hand, spatial relations concern the physical positions and orientations of agents relative to each other and objects in the environment. On the other hand, non-spatial relations can arise from communication protocols, task dependencies, or shared goals. Most of the related multi-agent reinforcement learning (MARL) research focuses on inter-agent communication [9, 12, 18], not fully addressing spatial reasoning. However, spatial relations provide a key environmental context by representing relationships fundamental to learning the task. These relations are not limited to interactions between agents but extend to relationships with all entities that populate the environment. Modeling this type of relational information can be highly beneficial for crafting effective strategies.

We make the following key contributions: (a) we propose a novel multi-agent actor-critic architecture incorporating spatial relational inductive biases through relational graph networks; (b) we benchmark our algorithm on 3 different collaborative tasks; (c) we compare against state-of-the-art MARL baselines, demonstrating competitive gains in sample efficiency and asymptotic performance.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 METHODOLOGY

Our objective is to design a multi-agent relational actor-critic architecture (MARC) that integrates a relational reasoning component into the learning process. To realize this, we carefully make specific design choices as described below.

2.1 Relational Observation Encoder

We hypothesize that it is sufficient to only consider the entities relevant to the game dynamics in our graph representation, emphasizing the relation between entities and their attributes. Accordingly, we assume an observation o_i, from which positions and attributes of the agents and environment objects can be extracted.

In line with methods applied in the single-agent literature [8], we propose that enforcing a spatial inductive bias, a structure inherent in many environments, provides superior performance and sample efficiency. Hence, we employ a graph-based representation with multiple, directional relations. Such a graph is formally defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, Z)$. Given N agents and M objects, we define the set of entities as $\mathcal{V} = \{v_1, \ldots, v_N, v_{N+1}, \ldots, v_{N+M}\}$. Moreover, entities in this graph have associated entity features, detailed by $Z \in \mathbb{R}^{d \times |\mathcal{V}|}$, where $d > 0$ denotes the dimensionality of the features.

Each edge $(a, r, b) \in \mathcal{E}$ is defined based on spatial rules, forming the set of relations $\mathcal{R} \ni r$, which are derived from the absolute positions (x_a, y_a) and (x_b, y_b) of two entities a and b. We find that a sufficient spatial inductive bias to complete most tasks is when agents are aware of whether entities are left, right, top, bottom, adjacent or aligned to one another.

To learn a relational representation of the observation that incorporates the connection between the objects, we employ RGCN [15] updates on our constructed graph, as it is adept at dealing with multiple relations. Under this framework, we produce updated entity representations of the entity features Z, influenced by their

respective, relation-specific neighbors. With $z_i \in \mathbb{R}^d$ being the entity features for each entity $i \in \mathcal{V}$, the entity representations are updated as $z_i' = \sigma\big(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_r(i)} |\mathcal{N}_r(i)|^{-1} W_r z_j + W_0 z_i\big)$, with σ an element-wise non-linear activation function. The employed aggregation function uses relation-specific weight matrices W_r and an auxiliary weight matrix W_0 brings in prior entity information. $\mathcal{N}_r(i)$ defines the relation-specific neighbors of entity i and the aggregation process is normalized by the number of relation-specific neighbors $|\mathcal{N}_r(i)|$.

After updating the entity features, we obtain a learned feature matrix Z' that we pass through a feature-wise pooling layer to obtain observation encodings $e(o_i) = \text{max-pool}(Z')$, where Z' implicitly depends on o_i.

2.2 Multi-Agent Relational Actor-Critic

Given the shared encoder for observations, we can feed the resulting relational representation into the MARL procedure. We implement the relational component within the critic to allow agents to share the relational observation encoder during training. Analogous to MAAC [7], each of the N agents possesses their own critic and policy network. Each critic is defined as $Q_{\psi_i}(o_i, a) = f_i(e(o_i), a)$, where f_i is a 2-layered dense neural network, receiving an observation encoding $e(o_i)$ and collective actions $a = (a_1, \ldots, a_N)$. ψ_i represents the shared parameters from e and individual dense layers f_i for each critic. The critics are optimized jointly to minimize the regression loss

$$\mathcal{L}_Q(\psi) = \sum_{i=1}^{N} \mathbb{E}_{(o_i,a,r_i,o_i')\sim D}\big[(Q_{\psi_i}(o_i,a) - y_i)^2\big], \tag{1}$$
$$y_i = r_i + \gamma \mathbb{E}_{a'\sim\pi_{\bar{\theta}}}\big[Q_{\bar{\psi}_i}(o_i', a') - \alpha \log \pi_{\bar{\theta}_i}(a_i'|o_i')\big],$$

where γ is the discount factor and D represents the replay buffer. Following the paradigm of soft actor-critic updates [6], we denote $\bar{\psi}_i$ and $\bar{\theta}_i$ as the individual target critic and policy parameters, respectively, and α defines the temperature parameter balancing entropy and reward maximization. $\pi_{\bar{\theta}} = (\pi_{\bar{\theta}_1}, \ldots, \pi_{\bar{\theta}_N})$ denotes the joint target policy vector, where each target policy $\pi_{\bar{\theta}_i}$, and correspondingly each policy π_{θ_i}, consists of a 3-layered dense neural network.

For individual policy updates, we employ gradient ascent:

$$\nabla_{\theta_i} J(\pi_{\theta_i}) = \mathbb{E}_{o_i\sim D, a\sim\pi_\theta}\Big[\nabla_{\theta_i} \log \pi_{\theta_i}(a_i|o_i) \times \tag{2}$$
$$\Big(-\alpha \log \pi_{\theta_i}(a_i|o_i) + Q_{\psi_i}(o_i, a) - b(o_i, a_{\setminus i})\Big)\Big].$$

To address the challenge of multi-agent credit assignment, we subtract a baseline term $b(o_i, a_{\setminus i})$ to estimate an agent's action net effect [5], here defined as $b(o_i, a_{\setminus i}) = \mathbb{E}_{a_i\sim\pi_{\theta_i}}[Q_{\psi_i}(o_i, (a_i, a_{\setminus i}))]$.

3 EXPERIMENTAL RESULTS

We hypothesize that our algorithm learns effectively, especially in spatially complex coordination tasks under sparse rewards and test this on the following collaborative environments: first, we use the level-based foraging (**LBF**) environment [1], where agents need to collect fruits on a grid. As opposed to the original implementation, we leave the fruits on the grid with a value of −1 after they have been collected. This demands a higher relational reasoning capability from agents, as they must now recognize them as noncollectable obstacles. For testing high cooperation, our experiments run on a 10×10 grid with 4 agents and 4 foods, enforcing cooperation (denoted as 10x10-4p-4f-c). To assess scalability, we extend the environment to a 15×15 grid with 8 agents and 1 fruit, requiring cooperation among a larger number of agents (denoted as 15x15-8p-1f-c). Second, we use **Wolfpack** [11] with 3 agents placed in a 10×10 grid to capture 2 prey. In a departure from the original setup, we introduce sparse rewards by removing additional rewards based on the proximity to prey, significantly weakening the learning signal.

Table 1: Average performance for each task and model over 3 seeds, taken at two different stages: first, after 10^6 steps to indicate sample efficiency, and second showing asymptotic performance. The best values are indicated in bold.

Algorithm	Task		
	LBF-10-4p-4f-c	LBF-15-8p-1f-c	Wolfpack
MAA2C	0.04 | 0.48	0.01 | 0.86	174.1 | 218.5
MAAC	0.02 | 0.14	0.03 | 0.96	176.6 | 297.9
MAPPO	0.10 | 0.53	0.01 | 0.87	202.3 | 220.6
MARC	**0.25** | **0.88**	**0.98** | **0.98**	**270.3** | **346.6**
QMIX	0.00 | 0.34	0.00 | 0.00	4.2 | 292.5

In Table 1, we show average performances attained by our method compared to **MAAC** [7], **MAA2C** [10], **MAPPO** [16] and **QMIX** [13]. MARC indicates a superior sample efficiency, showing a significant difference in performance after only 10^6 steps in all tasks. Asymptotically, our proposed method also outperforms all baselines across tasks. The most significant margin is achieved in the hardest task, LBF-10x10-4p-4f-c, where agents need to coordinate with at least 3 agents, making the coordination effort significantly more challenging and the reward particularly sparse. Also, the fruits remain in the environment, so the agents need to evaluate the level of the fruits properly. The spatial inductive bias introduced in MARC seems to aid the understanding of such complexities.

4 DISCUSSION

In this work, we put forward a relational approach to MARL, showcasing its effectiveness in environments demanding spatial reasoning. The spatial inductive bias inherent in our method not only enhances asymptotic performance but also demonstrates strong sample efficiency compared to our baselines with lesser or no inductive bias. Our method leverages a compact representation of the space and a considered choice of relations, keeping the computational overhead minimal without compromising the informativeness of the representation. Overall, our work lays a strong foundation for further exploration of relational learning in multi-agent settings, while providing a robust framework capable of adapting to diverse environmental conditions and requirements.

ACKNOWLEDGMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC EP/W523793/1), through the Statistics Centre for Doctoral Training at the University of Warwick and from a UKRI Turing AI Acceleration Fellowship (EPSRC EP/V024868/1).

REFERENCES

[1] Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. 2020. Shared Experience Actor-Critic for Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[2] Kurt Driessens and Sašo Džeroski. 2001. Integrating guidance into relational reinforcement learning. In *Machine Learning: ECML 2001*. Springer, Berlin, Heidelberg, 116–127.

[3] Sašo Džeroski, Luc De Raedt, and Hendrik Blockeel. 1998. Relational reinforcement learning. In *Inductive Logic Programming*. Springer Berlin Heidelberg, 11–22.

[4] Sašo Džeroski, Luc De Raedt, and Kurt Driessens. 2001. Relational reinforcement learning. *Machine learning* 43, 1 (2001), 7–52.

[5] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[6] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 1861–1870. https://proceedings.mlr.press/v80/haarnoja18b.html

[7] Shariq Iqbal and Fei Sha. 2019. Actor-Attention-Critic for Multi-Agent Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 2961–2970. http://proceedings.mlr.press/v97/iqbal19a.html

[8] Zhengyao Jiang, Pasquale Minervini, Minqi Jiang, and Tim Rocktäschel. 2021. Grid-to-Graph: Flexible Spatial Relational Inductive Biases for Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '21)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 674–682.

[9] Yaru Niu, Rohan Paleja, and Matthew Gombolay. 2021. Multi-Agent Graph-Attention Communication and Teaming. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (Virtual Event, United Kingdom) *(AAMAS '21)*. International Foundation for Autonomous Agents and Multiagent Systems, 964–973.

[10] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. 2020. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. *arXiv preprint arXiv:2006.07869* (2020).

[11] Arrasy Rahman, Ignacio Carlucho, Niklas Höpner, and Stefano V. Albrecht. 2023. A General Learning Framework for Open Ad Hoc Teamwork Using Graph-based Policy Learning. *Journal of Machine Learning Research* (2023).

[12] Murtaza Rangwala and Ryan Williams. 2020. Learning Multi-Agent Communication through Structured Attentive Reasoning. In *Advances in Neural Information Processing Systems*, Vol. 33. 10088–10098. https://proceedings.neurips.cc/paper/2020/file/72ab54f9b8c11fae5b923d7f854ef06a-Paper.pdf

[13] Tabish Rashid, Mikayel Samvelyan, C. S. D. Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *ArXiv* abs/1803.11485 (2018). https://api.semanticscholar.org/CorpusID:4533648

[14] Scott Sanner and Craig Boutilier. 2009. Practical solution techniques for first-order MDPs. *Artificial Intelligence* 173, 5-6 (2009), 748–788.

[15] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web*, Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam (Eds.). Springer International Publishing, 593–607.

[16] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and YI WU. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 24611–24624. https://proceedings.neurips.cc/paper_files/paper/2022/file/9c1535a02f0ce079433344e14d910597-Paper-Datasets_and_Benchmarks.pdf

[17] Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, Murray Shanahan, Victoria Langston, Razvan Pascanu, Matthew Botvinick, Oriol Vinyals, and Peter Battaglia. 2019. Deep reinforcement learning with relational inductive biases. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HkxaFoC9KQ

[18] Xianjie Zhang, Yu Liu, Xiujuan Xu, Qiong Huang, Hangyu Mao, and Anil Carie. 2021. Structural relational inference actor-critic for multi-agent reinforcement learning. *Neurocomputing* 459 (2021), 383–394. https://www.sciencedirect.com/science/article/pii/S0925231221010481