Embracing Relational Reasoning in Multi-Agent Actor-Critic

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

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.

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 \(o_t\), 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 \(G = (V, E, R, Z)\). Given \(N\) agents and \(M\) objects, we define the set of entities as \(V = \{v_1, \ldots, v_N, o_{N+1}, \ldots, o_{N+M}\}\). Moreover, entities in this graph have associated entity features, detailed by \(Z \in \mathbb{R}^{d\times |V|}\), where \(d > 0\) denotes the dimensionality of the features.

Each edge \((a, r, b) \in E\) is defined based on spatial rules, forming the set of relations \(R \supset r\), which are derived from the absolute positions \((x_a, y_a)\) and \((x_b, y_b)\) of two entities \(a \text{ and } b\). 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 adept at dealing with multiple relations. Under this framework, we produce updated entity representations of the entity features \(Z\), influenced by their
respectively, 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 \( z_i' = \sigma (\sum_{j \in \mathcal{N}_e(i)} |\mathcal{N}_e(i)|^{-1} W_e z_j + W_0 z_i) \), with \( \sigma \) an element-wise non-linear activation function. The employed aggregation function uses relation-specific weight matrices \( W_e \) and an auxiliary weight matrix \( W_0 \) brings in prior entity information. \( \mathcal{N}_e(i) \) defines the relation-specific neighbors of entity \( i \) and the aggregation process is normalized by the number of relation-specific neighbors \( |\mathcal{N}_e(i)| \).

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 observation encoder during training. Analogous to the spatial component within the critic to allow agents to share relational information during training. Analogous to \( \mathcal{L} \), each of the \( N \) agents possesses their own critic and policy network. Each critic is defined as \( Q_{\psi_i}(o_i, a) = f_t(e(o_i), a) \), where \( f_t \) is a 2-layered dense neural network, receiving an observation encoding \( e(o_i) \) and collective actions \( a = (a_1, \ldots, a_N) \). \( \psi_i \) represents the shared parameters from \( e \) and individual dense layers \( f_t \) for each critic. The critics are optimized jointly to minimize the regression loss

\[
\mathcal{L}_Q(\psi) = \sum_{i=1}^N \mathbb{E}_{(o_i,a_i,r_t,a_{t+1}) \sim D} \left[ (Q_{\psi_i}(o_i, a) - r_t)^2 \right],
\]

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

For individual policy updates, we employ gradient ascent:

\[
\nabla_{\theta_i} J(\pi_{\theta_i}) = \mathbb{E}_{(o_i,a_i) \sim D_{\pi_{\tilde{\theta}}}} \nabla_{\theta_i} \log \pi_{\tilde{\theta}_i}(a_i|o_i) \times \left( -\alpha \log \pi_{\tilde{\theta}_i}(a_i|o_i) + Q_{\psi_i}(o_i, a_i) - b(o_i, a_i) \right),
\]

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

### 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 now recognize them as noncollectable obstacles. For testing high cooperation, our experiments run on a \( 10 \times 10 \) grid with 4 agents and 4 foods, enforcing cooperation (denoted as 10×10-4p-4f-c). To assess scalability, we extend the environment to a \( 15 \times 15 \) 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 \( 10 \times 10 \) 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.

In Table 1, we show average performances attained by our method compared to MAAC [7], MAA2C [10], MAPPO [16] and QMIX [18]. 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


