Explaining cooperation in multi-agent reinforcement learning

Loïc Cordeiro Fonseca, Yannick Molinghen, and Tom Lenaerts

Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium

Introduction The increasing use of deep Reinforcement Learning (RL) policies to solve large scale decision-making problems has come with a loss in interpretability and resulted in a growing interest for Explainable RL. However, there are few works that tackle explainability in multi-agent RL (XMARL), let alone cooperation mechanisms. This master thesis explores Reward Decomposition [5,3, RD] and Soft Decision Trees [2,1, SDT] policy distillation as two XMARL techniques to investigate cooperation mechanism and answer the question whether cooperation is a "happy by-product" of a selfish policy or if cooperation is effectively encoded in the agents' policies.

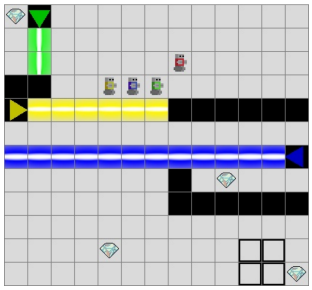

Fig. 1: LLE level representation.

Environment Setup We use the Laser Learning Environment (LLE) [4] in our experiments due to the need for agents to cooperate to complete the collaborative task. Agents have one main challenge in LLE: coloured lasers bar the way in differing spots. An agent can block a laser of its own colour by walking into it, but dies if it is of a different colour, thereby prematurely ending the episode with a punishment (-1). Along the way to the exit (+1), agents can collect gems (+1) and are rewarded the first time they enter a laser beam (+1). When all agents have reached an exit, an extra reward (+1) is collected.

Reward Decomposition We decompose LLE's reward signal to account for these 5 reward components and categorize them as either cooperative (death, end, laser) or selfish (gem, exit), effectively vectorizing the signal and the corresponding value and action-value functions. We train agents with Value Decomposition Network [6, VDN] and analyse the decomposed Q-values in states where cooperation is key to see the importance of each reward component for a given decision. Our results in Figure 2 (right) highlight our first contribution. It shows that agents significantly take in consideration the death of their allies via Q_{death} even when they are not at risk (as shown in Figure 1), i.e. show that cooperation is encoded in the agents' policies.

Our second contribution lies in a novel use of RD to get global insights on the agents' policies. By collecting the estimated decomposed value of each state encountered over the course of the training, we can observe an agent's long-term objective prioritisation over time, as shown Figure 2 (left). With that we can

quantify the impact of long-term rewards and by classifying them as cooperative or selfish, conclude that once cooperation is achieved agents shift their focus on selfish long-term objectives.

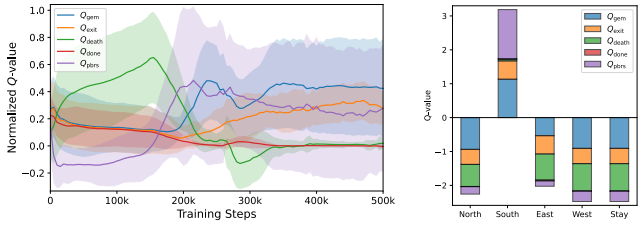

Fig. 2: Agent 0 Q-value prioritization over the course of its training (left); Agent 0 transition Q-values (right).

MARL policy distillation Our third contribution is the application of MARL policy distillation in SDT that take a 2-dimensional representation of the agent observation as input and outputs a probability distribution over the agent's actions. To train the SDT, we collect a curated dataset of greedy state-action pairs from a trained agent by letting the agent interacting with the environment. By plotting a heatmap of the filtered inputs over the grid world, we can visualize the important features in an agent's decision-making. In Figure 3, we show the point of view of the yellow agent (with a red frame) who can block the horizontal laser just below. We can see that the yellow agents pays positive attention to the three laser cells that are below the three adjacent agents but negative attention to the two leftmost tiles of the same beam. This, again, is an indicator that agents takes other agents into consideration and that cooperation is not a "happy by-product" of the optimization process.

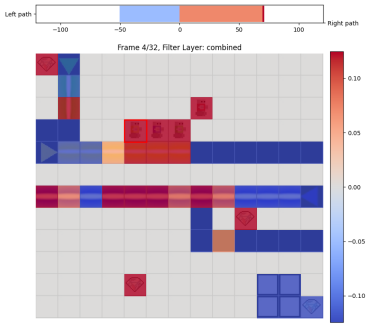

Fig. 3: Visualization of an SDT filter over the game-board from the perspective of agent 0, which can block the laser in front of the agents.

Lastly, we implemented an interactive interface, which allows the replaying of episodes with the overlaid SDT visualization. Users can see the filters over the course of an episode for individual agents and verify the predicted distilled action against the original one. If RD was active for the selected episode, users may also show the decomposed Q-values.

Conclusion These results indicate that local coordinating behaviours in MARL are not merely a by-product of optimization, but explicitly encoded in agent policies at local scale. However, agents are motivated by selfish incentives on a global scale.

References

1. Coppens, Y., Efthymiadis, K., Lenaerts, T., Nowé, A.: Distilling deep reinforcement learning policies in soft decision trees. In: International Joint Conference on Artificial Intelligence (2019), https://api.semanticscholar.org/CorpusID:201700841
2. Frosst, N., Hinton, G.: Distilling a neural network into a soft decision tree (2017), https://arxiv.org/abs/1711.09784
3. Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., Doshi-Velez, F.: Explainable reinforcement learning via reward decomposition. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence Workshop on Explainable Artificial Intelligence (2019)
4. Molinghen, Y., Avalos, R., Van Achter, M., Nowé, A., Lenaerts, T.: Laser learning environment: A new environment for coordination-critical multi-agent tasks. In: BNAIC 2023. BeNeLux Artificial Intelligence Conference (2023)
5. Russell, S., Zimdars, A.: Q-decomposition for reinforcement learning agents. In: Proceedings of the Twentieth International Conference on Machine Learning. vol. 2, pp. 656–663 (01 2003)
6. Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W.M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J.Z., Tuyls, K., Graepel, T.: Value-decomposition networks for cooperative multi-agent learning based on team reward. Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS **3**, 2085–2087 (2018)