Task Generalisation in Multi-Agent Reinforcement Learning

Doctoral Consortium

Lukas Schäfer
University of Edinburgh
Edinburgh, United Kingdom
l.schaefer@ed.ac.uk

ABSTRACT

Multi-agent reinforcement learning agents are typically trained in a single environment. As a consequence, they overfit to the training environment which results in sensitivity to perturbations and inability to generalise to similar environments. For multi-agent reinforcement learning approaches to be applicable in real-world scenarios, generalisation and robustness need to be addressed. However, unlike in supervised learning, generalisation lacks a clear definition in multi-agent reinforcement learning. We discuss the problem of task generalisation and demonstrate the difficulty of zero-shot generalisation and finetuning at the example of multi-robot warehouse coordination with preliminary results. Lastly, we discuss promising directions of research working towards generalisation of multi-agent reinforcement learning.

KEYWORDS

Reinforcement Learning; Multi-Agent Systems; Generalisation

ACM Reference Format:
Lukas Schäfer. 2022. Task Generalisation in Multi-Agent Reinforcement Learning: Doctoral Consortium. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

Reinforcement learning (RL) [26] is a paradigm of machine learning which enables agents to learn behaviour from interaction with an environment. *Multi-agent reinforcement learning* (MARL) [4, 23] extends this framework to multi-agent systems, i.e. it enables multiple agents to concurrently learn from interaction with the environment as well as interactions with each other. RL methods become increasingly capable in learning complex behaviour [20, 25, 28]. However, their learned strategies are usually highly task-specific. This makes the application of RL in real-world tasks difficult which usually require the learned behaviour to be robust to small perturbations and changes in the environment [1]. In our recent work [24], we proposed a novel decoupling scheme which allows to leverage intrinsically-motivated exploration [2, 21, 22] and train a separate policy while improving the robustness of these typically brittle exploration methods. Ideally, RL should be able to develop a fundamental understanding of the dynamics and interactions within their environment and be able to re-apply such learned knowledge in a related task. However, training a policy in one task and applying it to a new similar task requires to train the policy from scratch as learned behaviour and representations currently do not generalise.

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.

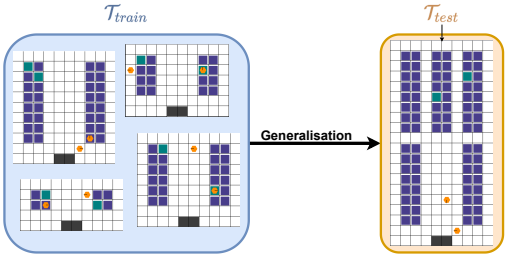


Figure 1: Multi-robot warehouse challenge

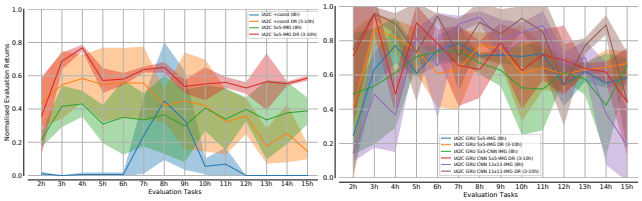
Established techniques from transfer learning [5, 27] try to address this problem by extracting representations, action selection or other components from already learned models. However, these methods are limited in that they require a dedicated transferring procedure for each new task to evaluate in. Meta RL [6–8, 16, 29] instead aims to learn policies which are adaptable using few-shot learning to be effective in testing tasks.

2 WHAT IS GENERALISATION?

The problem of generalisation is present in all branches of machine learning. While it is comparably well understood in supervised learning [13, 19], RL still lacks a unified view on generalisation. Recently, some advances are being made to define and understand generalisation in RL [9, 14, 17]. In the following, we discuss the challenge of generalisation in the setting of MARL.

Individual MARL tasks can be modelled as *partially-observable stochastic games* (POSG) [11]. To evaluate MARL generalisation, the joint policy π over all agents is trained in a set of training tasks $\mathcal{T}_{train}$ and evaluated based on average returns in testing tasks $T \sim \mathcal{T}_{test}$. Note, $\mathcal{T}_{train}$ and $\mathcal{T}_{test}$ might be disjoint but tasks from the training set might also be considered for testing. In this work, we consider zero-shot generalisation of MARL algorithms, i.e. π is directly applied in testing tasks without any further training allowed after the initial training in $\mathcal{T}_{train}$, as well as finetuning, i.e. π is allowed further training in $\mathcal{T}_{test}$. Both these cases of generalisation require agents to learn re-usable skills and representations which transfer to testing tasks.

The most important question with respect to the type of generalisation of a particular problem are the similarities and differences between $\mathcal{T}_{train}$ and $\mathcal{T}_{test}$. Without any further assumptions, training and testing tasks could be arbitrarily different and hence no generalisation could be feasibly expected. Therefore, further assumptions need to be made on the relationship between training and testing tasks to restrict the problem space and allow for focused research. In the following, we focus on *task generalisation* where



(a) Observation encodings (b) Policy architectures

Figure 2: Normalised returns in zero-shot generalisation to warehouses of block height $2-15$ with varying observation encodings, architectures and domain randomisation (DR).

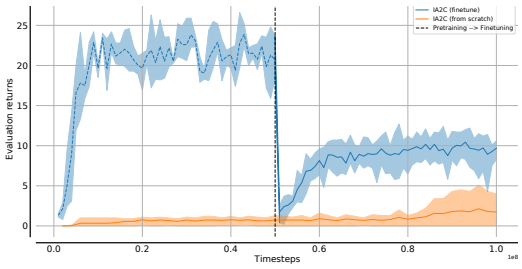


Figure 3: Evaluation returns for IA2C with GRUs and CNNs on 5×5 image observations with finetuning and training from scratch.

training and testing tasks represent the same type of problem, but contain different states. An example would be multi-robot warehouse navigation (Figure 1) where the high-level goal is consistent across all warehouses but encountered states depend on the layout of a particular instance. A diverse set of states has strong implications as it also indirectly affects transition function, rewards and observations which are defined over the set of states.

3 PRELIMINARY EXPERIMENTS

In the multi-robot warehouse (RWARE)[1] [23], visualised in Figure 1, multiple agents (orange) need to navigate a gridworld warehouse, collect randomly requested shelves (green) and deliver them to the dropoff locations (black). While the challenge and required behaviour of agents intuitively remains very similar across varying shapes of warehouses, existing MARL algorithms are unable to achieve such generalisation. We investigated the impact of (1) varying observation encodings, (2) domain randomisation (DR), and (3) varying neural network architectures for policies on zero-shot generalisation capabilities of agents to identify the limitations of existing approaches. Lastly, we also evaluate the ability of agents to finetune representations for task generalisation.

We trained agents using independent synchronous Advantage Actor-Critic (IA2C), i.e. agents are independently trained with A2C [18], in warehouses of similar layout but varying height of blocks of shelves (blue box in Figure 1) for 50 million timesteps.

Observations in RWARE encode a limited grid centered around agents to keep observation dimensions consistent across varying warehouses and focus on most relevant information in the immediate proximity. Preliminary experiments showed that agents which observe absolute locations in the warehouse are unable to generalise at all if trained in a single warehouse instance (blue in Figure 2a). Out-of-(training-)distribution values lead to unstable policies. Slightly changing the observation encoding using an image-encoding of local information without providing coordinates significantly improves generalisation capabilities even without relying on DR (green in Figure 2a). DR is an approach to generalisation in which agents are concurrently trained in a diverse set of tasks. We observed that such training improves generalisation for either observation encoding across warehouses of similar variability as observed in the training set (orange and red in Figure 2a).

We also considered applying convolutional neural networks (CNN) [15] to encode the 2D spatially correlated observations and apply Gated Recurrent Networks (GRU) [3] to preserve memory of the partially-observable information provided to agents [12]. Encoding of image observations using CNNs in itself does not appear to significantly benefit generalisation, but the application of CNNs allowed to learn from observations with increased visibility radius (purple and brown in Figure 2b). The application of GRUs improved performance of all agents, but its benefits appeared to be not specific to generalisation (Figure 2b).

Lastly, we evaluated agents in a larger warehouse after training in smaller warehouses (Figure 1). None of the agents are able to deliver more than a few shelves successfully in this testing task. Observations are sufficiently different from observations encountered at training time which leads to agents becoming stuck.

Overall, our zero-shot experiments demonstrate that carefully selected observation encodings with corresponding architectural choices and DR can improve generalisation across a set of training tasks. However, neither approach is sufficient in achieving generalisation to warehouses of different layouts as visualised in Figure 1. Motivated by these results, we investigated the suitability of finetuning learned representations in this testing task. After training in smaller warehouses for 50 million timesteps, we train agents for further 50 million timesteps in the larger warehouse. Figure 3 compares such finetuned agents with agents only trained in the testing task for 100 million timesteps. We observe that pretrained agents achieve significantly higher returns in the larger warehouse after $\sim$ 20 million timesteps of finetuning compared to agents only trained in the larger warehouse. This indicates that representations learned in the smaller warehouse include useful information for the testing task with agents benefitting from the pretraining procedure.

4 CONCLUSIONS

We demonstrated the challenge of task generalisation in MARL in preliminary experiments. Existing MARL algorithms are shown to be too sensitive to task-specific changes in observations but learned representations appear to include valuable information with promising results after limited finetuning in new testing tasks. These results indicate the possible suitability of meta RL adaptation techniques [6, 8, 16, 29] which we aim to extend for MARL generalisation. Furthermore, we aim to theoretically formalise the relationship of warehouse tasks as shown in Figure 1. Existing approaches, such as contextual formalisms [10] are able to represent the general idea, but do not represent the intuitive similarity of high-level concepts such as dynamics, rewards and observations.

[1]Environment available at https://github.com/uoe-agents/robotic-warehouse

REFERENCES

[1] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. 2019. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113* (2019).

[2] Andrew G Barto. 2013. Intrinsic motivation and reinforcement learning. In *Intrinsically motivated learning in natural and artificial systems*. Springer, 17–47.

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[4] 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*.

[5] Felipe Leno Da Silva and Anna Helena Reali Costa. 2019. A survey on transfer learning for multiagent reinforcement learning systems. *Journal of Artificial Intelligence Research* 64 (2019), 645–703.

[6] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. 2016. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779* (2016).

[7] Rasool Fakoor, Pratik Chaudhari, Stefano Soatto, and Alexander J Smola. 2020. Meta-q-learning. In *International Conference on Learning Representations*.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*. PMLR, 1126–1135.

[9] Dibya Ghosh, Jad Rahme, Aviral Kumar, Amy Zhang, Ryan P Adams, and Sergey Levine. 2021. Why Generalization in RL is Difficult: Epistemic POMDPs and Implicit Partial Observability. In *Advances in Neural Information Processing Systems*.

[10] Assaf Hallak, Dotan Di Castro, and Shie Mannor. 2015. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259* (2015).

[11] Eric A Hansen, Daniel S Bernstein, and Shlomo Zilberstein. 2004. Dynamic programming for partially observable stochastic games. In *AAAI*, Vol. 4. 709–715.

[12] Matthew Hausknecht and Peter Stone. 2015. Deep recurrent q-learning for partially observable mdps. In *2015 AAAI Fall Symposium Series*.

[13] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. 2017. Generalization in deep learning. *arXiv preprint arXiv:1710.05468* (2017).

[14] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. 2021. A Survey of Generalisation in Deep Reinforcement Learning. *arXiv preprint arXiv:2111.09794* (2021).

[15] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1, 4 (1989), 541–551.

[16] Evan Z Liu, Aditi Raghunathan, Percy Liang, and Chelsea Finn. 2021. Decoupling exploration and exploitation for meta-reinforcement learning without sacrifices. In *International Conference on Machine Learning*. PMLR, 6925–6935.

[17] Dhruv Malik, Yuanzhi Li, and Pradeep Ravikumar. 2021. When Is Generalizable Reinforcement Learning Tractable?. In *Advances in Neural Information Processing Systems*.

[18] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. 1928–1937.

[19] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. 2017. Exploring generalization in deep learning. *arXiv preprint arXiv:1706.08947* (2017).

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

[21] Pierre-Yves Oudeyer and Frederic Kaplan. 2009. What is intrinsic motivation? A typology of computational approaches. *Frontiers in neurorobotics* 1 (2009), 6.

[22] Pierre-Yves Oudeyer, Frederic Kaplan, et al. 2008. How can we define intrinsic motivation. In *Proc. of the 8th Conf. on Epigenetic Robotics*, Vol. 5. 29–31.

[23] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. 2021. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

[24] Lukas Schäfer, Filippos Christianos, Josiah Hanna, and Stefano V. Albrecht. 2022. Decoupled Reinforcement Learning to Stabilise Intrinsically-Motivated Exploration. In *International Conference on Autonomous Agents and Multi-Agent Systems*.

[25] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the game of Go without human knowledge. *Nature* 550 (Oct. 2017), 354–. http://dx.doi.org/10.1038/nature24270

[26] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[27] Matthew E Taylor and Peter Stone. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research* 10, 7 (2009).

[28] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[29] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. 2020. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. In *International Conference on Learning Representations*.