Learning How to Flock: Deriving Individual Behaviour from Collective Behaviour with Multi-Agent Reinforcement Learning and Natural Evolution Strategies*

Koki Shimada
Computer Science Department
University College London
United Kingdom
k.shimada@cs.ucl.ac.uk

Peter Bentley
Computer Science Department
University College London
United Kingdom
p.bentley@cs.ucl.ac.uk

ABSTRACT

This work proposes a method for predicting the internal mechanisms of individual agents using observed collective behaviours by multi-agent reinforcement learning (MARL). Since the emergence of group behaviour among many agents can undergo phase transitions, and the action space will not in general be smooth, natural evolution strategies were adopted for updating a policy function. We tested the approach using a well-known flocking algorithm as a target model for our system to learn. With the data obtained from this rule-based model, the MARL model was trained, and its acquired behaviour was compared to the original. In the process, we discovered that agents trained by MARL can self-organize flow patterns using only local information. The expressed pattern is robust to changes in the initial positions of agents, whilst being sensitive to the training conditions used.

CCS CONCEPTS

• **Theory of computation → Multi-agent reinforcement learning**

KEYWORDS

Multi-agent systems, Reinforcement learning, Swarm intelligence, Evolution strategies, Neural networks/Deep Learning

ACM Reference format:

K. Shimada and P. Bentley. 2018. SIG Proceedings Paper in word Format. In *Proceedings of ACM GECCO conference, Kyoto, Japan, July 2018 (GECCO'18)*, 8 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

While agent-based models aid in understanding the nature of biological systems, they require the modelling of each individual agent precisely [1]. However, in many cases what we can easily observe is the behaviour of the group, and the behaviour of each individual is hard to model. Therefore, a top-down approach where the behaviour of the individual agents is predicted from the behaviour of the collective group would be useful.

Taking advantage of multi-agent reinforcement learning (MARL), this paper proposes a method to acquire a behavioural policy that will produce an intended collective behaviour as a result of interactions between agents. A reward function is designed that measures the difference between the behaviour of the target system, and the behaviour of the MARL system. Natural evolution strategies (NES) [2] are used to search for optimum parameters of the policy function.

Reinforcement learning has been applied to multi-agent systems previously [3-6]. In these studies, however, reinforcement learning is performed on at most 10 agents. Furthermore, no attempts were made to learn how emergent phenomena arise from interactions between multiple agents.

2 SYSTEM

Reynolds' flocking model was adopted as the target system for this work. Reynolds showed that it is possible to simulate the collective movement of a flock of birds by applying only three forces (*collision avoidance, velocity matching and flock centering*) to each agent [7]. By interacting with each other using only these simple rules, agents perform naturalistic swarming behaviour as a collective.

To allow our target rule-based model to be learned by the MARL model, we developed three metrics to be used in training: 1) entropy of positions and angles e_1, 2) entropy of positions e_2, 3) entropy of angles e_3. These were chosen in order to provide some measure of the overall behaviour of the target flock, without directly providing information relating to the underlying forces of collision avoidance, centring, and velocity matching. Such measurements might be taken by someone unfamiliar with the individual behaviour of the system, who can only observe the collective behaviour.

GECCO '18 Companion, July 15–19, 2018, Kyoto, Japan

ACM ISBN 978-1-4503-5764-7/18/07.
https://doi.org/10.1145/3205651.3205770

The environment is a two-dimensional space; agents are randomly placed in that space and move around with a fixed velocity. Each agent has a field of vision which allows them to recognize other neighbouring agents. For each agent, with positional information $\boldsymbol{s}_i$, in the field of view, the policy function, $\pi(a_i, w_i;\ \boldsymbol{s}_i)$ outputs an angle and a weighting. The $\boldsymbol{s}_i$ consist of three values: the observed agent's heading, the distance to the observed agent and its bearing relative to observer's own heading. After calculating angles and weightings for all objects within the field of view, the weighted average of those angles becomes the direction a_{t+1} in which the agent proceeds in the next time step. This policy function is shared by all agents, but they each make different observations and interpret them as different situations, so the behaviour of each agent can vary.

The MARL model seeks to adapt its own deterministic policy function, π_θ, to approximate the entropy values acquired from the rule-based model, thus imitating its collective behaviour. Since this policy function can be nonlinear and is in general of unknown form, a neural network was chosen. Neural networks are model-free, can approximate nonlinear functions and quantitatively adjust the number of degrees of freedom. The neural network is parameterized by θ, which returns an angle and a weighting for each $\boldsymbol{s}_i$.

The policy function is updated so as to minimise the entropy difference between the rule-based model and the MARL model by using NES. Salimans et al. showed NES can train in a much shorter period than for a recent successful learning method known as A3C [8]. Letting θ_t be a real-valued vector to be optimized at training epoch t and letting the search distribution be Gaussian, the fitness of individual i is:

$$F_i = F(\theta_t + \sigma\epsilon_i) \tag{1}$$

where $\epsilon_i \sim N(0, I)$ is a noise sampled from the distribution, σ is the standard deviation of the distribution and the fitness function F is the sum of the rewards obtained during the simulation period. As we want to update so as to improve the average value of the fitness of all individuals, the parameters are updated as follows:

$$\theta_{t+1} = \theta_t + \alpha\frac{1}{n\sigma}\sum_{i}^{n} F_i\epsilon_i \tag{2}$$

where n is the size of the population and α is the learning rate.

3 RESULTS

We have done two experiments to investigate the trainability of the MARL model. While the MARL model achieves improved metric scores over time, the overall behaviour does not sufficiently resemble the target system. As Fig. 1 (left) shows, the degree of dispersion of the positions of all agents of the trained model differs from that of the rule-based model. It was frequently observed in MARL model simulations that agents self-organised into a cluster instead of flowing like a flock.

Fig. 1 (right) illustrates how a typical pattern is self-organized: at t = 136, a prototype dog-leg pattern is constructed, and after t = 273 a stable flow of agents from right to left can be recognized. Since pattern formations are bottom-up expressions of local interactions between agents, some initial value sensitivity is expected. However, almost the same pattern appears regardless of starting positions provided the same training model is used. Different robust patterns develop for different training models, which is true even if the reward function is unchanged.

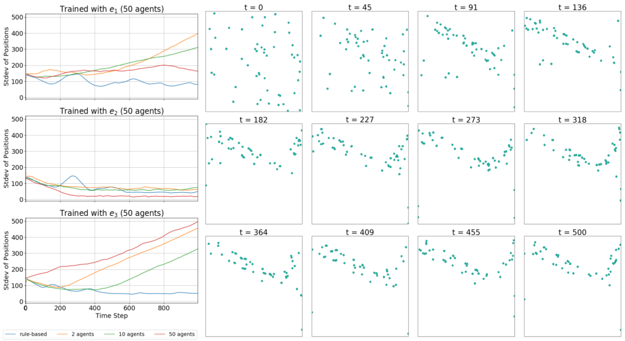

Figure 1: Standard deviation of positions for each reward function (left). The emergence of self-organised flow pattern (right).

4 CONCLUSIONS

In this study, using MARL and NES, we have shown it is possible to train individual agents to maximise a reward function defined by group-level phenomena. The chosen metrics, though appearing to describe intended collective behaviours appropriately, led to different learned behaviours. However, the systems converge to the same patterns regardless of initial values, depending only on the training model used. Thus, a robust behavioural policy with no sensitivity to initial conditions can be designed using MARL.

REFERENCES

[1] Wilensky, U. and Rand, W., 2015. An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo. MIT Press.

[2] Wierstra, D., Schaul, T., Glasmachers, T., Sun, Y., Peters, J. and Schmidhuber, J., 2014. Natural evolution strategies. Journal of Machine Learning Research, 15(1), pp.949-980.

[3] Foerster, J., Assael, Y., de Freitas, N. and Whiteson, S., 2016. Learning to communicate with deep multi-agent reinforcement learning. In Advances in Neural Information Processing Systems (pp. 2137-2145).

[4] Leibo, Joel Z., Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. "Multi-agent reinforcement learning in sequential social dilemmas." In Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, pp. 464-473. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

[5] Lerer, A. and Peysakhovich, A., 2017. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. arXiv preprint arXiv:1707.01068.

[6] Shalev-Shwartz, S., Shammah, S. and Shashua, A., 2016. Safe, multi-agent, reinforcement learning for autonomous driving. arXiv preprint arXiv:1610.03295.

[7] Reynolds, C.W., 1987, August. Flocks, herds and schools: A distributed behavioral model. In ACM SIGGRAPH computer graphics (Vol. 21, No. 4, pp. 25-34). ACM.

[8] Salimans, T., Ho, J., Chen, X. and Sutskever, I., 2017. Evolution strategies as a scalable alternative to reinforcement learning. arXiv preprint arXiv:1703.03864.