Karlsruhe Institute of Technology

Swarm Reinforcement Learning with Graph Neural Networks

Bachelor's Thesis
of

Christian Burmeister

KIT Department of Informatics
Institute for Anthropomatics and Robotics (IAR)
Autonomous Learning Robots (ALR)

Referees: Prof. Dr. Techn. Gerhard Neumann
Prof. Dr. Ing. Tamim Asfour

Advisor: Niklas Freymuth

Duration: Juli 17st, 2021 — January 17st, 2022

Erklärung

Ich versichere hiermit, dass ich die Arbeit selbstständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht habe und die Satzung des Karlsruher Instituts für Technologie zur Sicherung guter wissenschaftlicher Praxis beachtet habe.

Karlsruhe, den 17. January 2022

Christian Burmeister

Zusammenfassung

Einseitige deutsche Zusammenfassung (*Abstract*) der Abschlussarbeit. Unabhängig von der Sprache der Abschlussarbeit *muss* eine deutsche Zusammenfassung verfasst werden.

Abstract

The one-page abstract of the thesis.

Table of Contents

Chapter 1.

Introduction

Stuff to talk about in introduction.

- Applications or real-world problems that require a solution.
- MARL: Good for certain applications like
- GNN more and more popular
- Using GNN for MARL
- What has research focused on?
- Some examples from research what can be done.
- What we set out to do what our basic goal was, that should be a natural conclusion of what we talked about above.
- Last item: What the chapters will talk about, what we will talk about the structure of sections
- Structure:
  - Applications or real-world problems that require a solution.
  - Recent applications and research in MARL and GNN
  - What is GNN, What is MARL, what can GNNs it do for MARL? (the main thing we want to talk about, more conceptually)
  - What is my approach I want to talk about here? What was our goal?
  - My work relative to other work. What has other research focused on?
  - talking about the structure of the thesis

Chapter 2.

Preliminaries

This chapter will introduce the necessary concepts that need to be understood. The baseline is a bachelor's degree in computer science without any assumptions made about the elective studies. Topics:

- RL
  - MDP
- MARL
  - PoMDP
- NN
- vanilla message-passing GNN

Chapter 3.

Related Work

20 referenced papers. 2-3 sections

- RL
  - Swarm RL (max, Robin)
  - PPO
  - TRL
- GNN
  - GNNs
  - GATs
  - MeshGraphNets

Deisenroth et al. (2013)

Chapter 4.

Swarm Reinforcement Learning with Graph Neural Networks

This Chapter is moreso a deep dive into the actual solution of the Swarm RL with GNN Algorithm. Our Architecture and stuff.

4.1. Definition of the Problemdomain

Chapter 5.

Experiments

5.1. General Setup

Talk about my code base what it is based on etc. What I use.

- Optuna
- DAVIS
- Code from: Bayesian and Attentive Aggregation for Multi-Agent Deep Reinforcement Learning

5.2. Tasks

5.2.1. Rendezvous

- Goal.
- Basic Environment Structure (Torus,)
- Visual: example task completion, with timesteps and total environment
- Agent-Model (Dynamics, Collision), Evader-Model (Strategy)
- Reward
- Observation => Data, Culling and Graph

5.2.2. Dispersion

- Goal.
- Basic Environment Structure (Torus,)
- Visual: example task completion, with timesteps and total environment
- Agent-Model (Dynamics, Collision), Evader-Model (Strategy)
- Reward
- Observation => Data, Culling and Graph

5.2.3. Single Evader Pursuit

- Goal.
- Basic Environment Structure (Torus,)
- Visual: example task completion, with timesteps and total environment
- Agent-Model (Dynamics, Collision), Evader-Model (Strategy)
- Reward
- Observation => Data, Culling and Graph

5.2.4. Multi Evader Pursuit

- Goal.
- Basic Environment Structure (Torus,)
- Visual: example task completion, with timesteps and total environment
- Agent-Model (Dynamics, Collision), Evader-Model (Strategy)
- Reward
- Observation => Data, Culling and Graph

5.3. Experiments

Let those experiments run over all environments where applicable.

- PPO Hacks vs No Hacks
  - Environments: Rendezvous
  - value-function-clipping (0.0 - 1.0), 1.0 = no clipping
  - normalize rewards
  - ?reward-clipping: graph-normalized constructor: reward-clip = 5, currently no parameter
  - observation-normalization
  - global gradient clipping: max-grad-norm
  - ?tanh (insted of LeakyReLU)
- Number of Hops:

  - Environments: Rendezvous, Pursuit-Multi
  - Environment: Culling Methods: more culling vs less culling, num-agents and dynamics?
  - Network: num-blocks, latent-dimension?, aggregation-function?
- het-neighbor-aggregation: aggr(aggr()) vs conat(aggr())
  - Environments: Pursuit-Single, Pursuit-Multi
  - Environment: Base-Pursuit-Multi with 3+ Hops?
  - Network: latent-dimension, aggregation-function, num-blocks
- random number of agents
  - Environments: Rendezvous
  - Environment: Rendezvous: Culling Methods: more culling vs less culling
  - Network: latent-dimension, num-blocks
- random number of agents + random number of evaders
  - Environments: Pursuit-Multi
  - Environment: Multi-Pursuit: Culling Methods: more culling vs less culling
  - Network: latent-dimension, num-blocks
- dispersion: reward-type and aggregation-function
  - Environments: Dispersion
  - Environment: Culling Methods: more culling vs less culling
  - Network: latent-dimension, aggregation-function
- ?pursuit: reward-type???
  - Environments: Single-Pursuit
  - Environment: nothing?
  - Network: latent-dimension, aggregation-function

Chapter 6.

Evaluation

6.1. PPO Hacks vs No Hacks

- value-function-clipping (0.0 - 1.0), 1.0 = no clipping
- normalize rewards
- reward-clipping: graph-normalized constructor: reward-clip = 5, currently no parameter
- observation-normalization
- global gradient clipping: max-grad-norm
- tanh (insted of LeakyReLU)

6.2. Number of Hops

- Environment: Culling Methods: more culling vs less culling, num-agents and dynamics?
- Network: num-blocks, latent-dimension?, aggregation-function?

6.3. Neighbor Aggregation Type

aggr(aggr()) vs conat(aggr())

- Environment: Base-Pursuit-Multi with 3+ Hops?
- Network: latent-dimension, aggregation-function, num-blocks

6.4. Randomized Number of Agents and Evaders

random number of agents

- Environment: Rendezvous: Culling Methods: more culling vs less culling
- Network: latent-dimension, num-blocks

random number of agents + random number of evaders

- Environment: Multi-Pursuit: Culling Methods: more culling vs less culling
- Network: latent-dimension, num-blocks

6.5. Dispersion

- Environment: Culling Methods: more culling vs less culling, reward-type
- Network: latent-dimension, aggretation-function

Chapter 7.

Conclusion and Future Work

7.1. Conclusion

- What comparison we made on different tasks.
- What have we observed
- What have we shown of results.
- Basically a summary of "evaluation":

Number of Hops

Neighbor Aggregation Type

Randomized Number of Agents and Evaders

PPO Hacks vs No Hacks.

Dispersion.

7.2. Future Work

More can be done to expand on the work already finished in this bachelor thesis.

All of the experiments in this paper only considered using Proximal Policy Optimization (PPO) Schulman et al. (2017), as it is a very common baseline training algorithm. However recent

research shows that other methods might lead to better results for mult-iagent Reinforcement Learning. Specifically Trust Region Layers (PG-TRL) Otto et al. (2021) is an alternative that is able to be atleast on par with PPO, while requiring less code-level optimizations. It is noted that in experiments using sparser rewards the fact that TRL has better exploration over PPO improves the results significantly. Though the base paper only explores single-agent problems. Ruede et al. (2021) explores Trust Region Layers (PG-TRL) Otto et al. (2021) for multi-agent tasks. The author explains that given an multi-agent cooperative task the agents are rewarded as a group, which makes the reward more sparse. Therefore creating correlation of a single agents action and the group reward is harder. The hyper parameters used for TRL were based on searches for PPO and no extensive testing for TRL was done. Even then TRL was able to perform similar to PPO.
Creating further experiments based on the architecture established in this thesis would very likely benefit from TRL.

Furthermore Ruede et al. (2021) also used more complex multi-agent task than we used in our experiments. In Box Clustering there are agents and multiple boxes. Each box is assigned to one cluster. The goal is to move the boxes, so that the distance between the boxes in a given cluster is minimal. Optimal solutions will require that the agents work together to move the boxes and that they split the work between them. In his thesis his approach worked well for two clusters of boxes, but fell appart with 3 clusters. Here the agents were only able to move one cluster correctly. Only after increasing the batch size and environment steps per training steps tenfold the agents were able to consider more than 2 clusters. As explained above, our approach is structuraly similar to Ruede et al. (2021) as both can be described with message-passing of GNNs. It was shown that more complex tasks, especially with tight communication ranges, benefitet hugely from multiple message passing hops. So one can assume that we would be able to solve Box Clustering better.

- Transfer Learning for GNNs?

Bibliography

M. P. Deisenroth, G. Neumann, and J. Peters. *A survey on policy search for robotics*. now publishers, 2013.

F. Otto, P. Becker, V. A. Ngo, H. C. M. Ziesche, and G. Neumann. Differentiable trust region layers for deep reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qYZD-AO1Vn.

R. Ruede, G. Neumann, T. Asfour, and M. Huettenrauch. Bayesian and attentive aggregation for multi-agent deep reinforcement learning. *Autonomous Learning Robotics (ALR)*, 2021. URL https://phiresky.github.io/masters-thesis/manuscript.pdf.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Appendix A.

Example Appendix

This is an example for an appendix.