RL Project Sparse Rewards - Preview

Luan Liebig-Schultz
luan.liebig-schultz@stud.uni-hannover.de

Philipp Link
p.link@stud.uni-hannover.de

Julius Heidmann
julius.heidmann@stud.uni-hannover.de

1 Motivation

One of the key challenges in reinforcement learning, especially in environments with sparse reward structures, is the difficulty for agents to identify and learn from the few meaningful interactions that lead to success. This often results in very slow learning or failure to learn at all. In this project, we want to explore possible ways to make learning in such settings more efficient. Our main interests lie in finding approaches that help the agent better recognize which parts of its behavior were actually responsible for receiving a reward and how to strengthen the learning success based on the few positive trajectories observed.

2 Related Topics

This project connects to several of the topics and techniques discussed in the lecture. Additionally, we are interested in techniques for reward shaping and intrinsic rewards which are especially relevant in sparse environments.

3 Idea

The core idea is to explore different ways to improve agent performance in sparse reward settings, specifically using the Doorkey task in the MiniGrid environment. We are currently looking into architectural changes (e.g. attention-layer) or modifications to the training process. One idea we find interesting is to focus more on positive trajectories early on in the training process. On the other hand we could utilize an attention-layer to focus on the relevant state and actions pairs. The implementation will be based on the Stable Baselines3 framework, but we are still unsure about how flexible it is with respect to modifying policy architectures and training loops.

Algorithm 1 A great RL algorithm.

Require: environment e, algorithm A **return** policy π
  while T **do**RUE
    Train A on e
  end while

$$\pi \in \Pi, \pi : \mathcal{S} \mapsto \mathcal{A} \tag{1}$$

4 Experiments

Environments & Metrics We will focus on the Doorkey-16x16 environment from the MiniGrid suite, as it presents a clear and interpretable sparse reward challenge. The main metrics we plan to

track include steps to threshold and training sample efficiency curve. We also want to get a deeper understanding of the training process itself and the use of positive examples for the agent. Entropy and policy confidence over time could be promising candidates for further analysis.
PPO from stable-baselines3 is a promising option to be the baseline.

Experimental Scope The exact number of experiments will depend on how well we can adapt the policy classes and how stable training is. We are planning to run several short experiments to compare different model types (e.g. attention-layer) or training strategies (e.g. adding intrinsic rewards). For each setup, we plan to run at least 10 seeds.
Moreover we want utilize AutoRL to mitigate the influence of hyperparameters as confounding factors.

Estimated Computational Load All experiments are expected to be lightweight and runnable on a single CPU. Experiments over multiple seeds and HPO could utilize on-premise GPU-Ressources.

5 Timeline

- **Research (2 days):** Gather papers, define key ideas to test, and review Stable Baselines3 structure.
- **Implementation (3 days):** Set up the environment and training scripts, test model variations and integration.
- **Experiments (4 days):** Run selected experiments, log data, and visualize performance.
- **Analysis (2 days):** Compare results, identify trends, and reflect on what worked.
- **Reporting (1 days):** Write final report, summarize findings, and highlight open questions for future work.