RL Project Proposal Template

Daniel Kruse
github.com/dakru012

1 Motivation

Exploration methods often times add an intrinsic reward to the existing extrinsic environment reward signal. I think this can disrupt or slow down the training process, because the reward signal gets noisy and we learn the exploration signals instead of the actual task at hand. Instead, we could try to decouple this step by actively changing the policy online while selecting the next action, keeping the environment's reward signal unchanged.

2 Related Topics

The project is mostly related to the lecture on exploration, in combination with the online reinforcement learning algorithm of my choice (possible multiple).

Brunner et al. [2018] proposed a similar idea before based on previous work utilizing dynamic model uncertainty to guide the exploration by Pathak et al. [2017]. However, I think there is still room for improvement here by incorporating new ideas, and standards e.g. for the used architecture.

3 Idea

When in state s, for each action a calculate the novelty of s' that will be reached from s by taking action a. Train dynamics model while training that learns to predict s' given s and a. Incorporate ideas from ICM here with inverse dynamics model to get useful embeddings of states. Estimate novelty of the new state s', and shape the given policy accordingly to increase likelihood of ending up in novel states. Reduce this bonus as training goes on, either by variable or because dynamics model naturally gets better at predicting the next state.

In this way, the original reward signal stays intact, potentially enabling the agent to learn it faster. Exploration is not necessarily the goal that the agent should learn but more the path that leads the agent to learn the correct behavior.

4 Experiments

Environments & Metrics Probably start with minigrid environments and possibly extend to Atari if there is enough time. In terms of metrics, mainly interested in training speed so performance versus number of training steps. Maybe visualize the possibly different exploration schemes in some way.

Experimental Scope Do 5-10 seeds, maybe less focus on hyperparameter optimization only if there is still time or performance out of the box is really bad. Compare new exploration method to established baselines: epsilon-greedy, count-based, ICM, Bootstrapped DQN, (...)

Estimated Computational Load Initial PPO run took about 1.5 hours on minigrid env and 5 seeds for 1 million steps on my desktop with CNN policy, (pessimistic?) estimate for proposed method maybe 4h. But, will probably have more resources at hand soon.

5 Timeline

Please tell us how long you think it will take to accomplish all parts of your project. This includes:

- Research
- Implementation
- Experiments
- Analysis
- Reporting

References

Gino Brunner, Manuel Fritsche, Oliver Richter, and Roger Wattenhofer. Using state predictions for value regularization in curiosity driven deep reinforcement learning. *CoRR*, abs/1810.00361, 2018. URL `http://arxiv.org/abs/1810.00361`.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction, 2017. URL `https://arxiv.org/abs/1705.05363`.