INF8953DE Project Proposal

Emile Dimas, Pavithra, Vijaya Lakshmi Kuruba

TOTAL POINTS

10 / 10

QUESTION 1

1 Project Proposal **10 / 10**

✓ **- 0 pts** **Good Plan.**

- 1 pts Need plan change.

- 1 pts Be careful with your ablations.

Beyond Prioritized Experience Replay

Pavithra Rajasekar
pavithra.parthasarathy-rajasekar@mila.quebec

Vijaya Lakshmi Kuruba
lakshmiv@mila.quebec

Emile Dimas
emile.dimas@mila.quebec

1 Project title and track:

Ablations/Analysis study of the paper Prioritized Experience Replay (Schaul et al., 2015)

2 Motivation and Problem Definition:

Experience Replay (ER) is nowadays a main component of online reinforcement algorithms. Prior to its adaptation by RL algorithms, RL agents experienced transitions only once and then discarded it. ER is a data structure of size N used by off-policy algorithms that stores experienced transitions. During training, the agent samples uniformly a mini batch of experiences from it and update the agent parameters consequently. ER has two main benefits: First, the uniform sampling causes the agent to process a more diverse mini batch of transitions and thus remove the correlation between the samples. Second, it keeps the agent from forgetting old transitions that would be useful later in the episode. Empirical research has shown robustly that algorithms using experience replay converges faster and to better results.

ER samples transitions uniformly from the buffer. This leads to the use of unnecessary transitions. In the paper Prioritized Experience Replay, the authors present an improvement in ER. The authors used an example to present their intuition: In the Blind Cliff Walk problem, there are n states. The agent must guess the correct action and if it's not the case the agent returns to the starting state. The only possible reward is when the agent reaches the last state. In this setting, most of the transitions are useless. Therefore, in the rare rewards setting, using uniform sampling will not yield great results. It is for that reason that Tom Schaul introduced a prioritized sampling in his paper Prioritized Experience Replay.

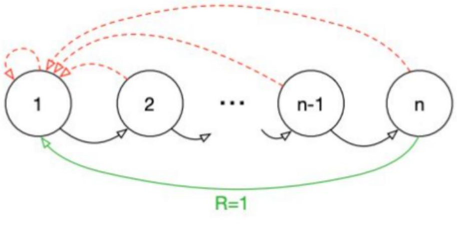

Figure 1: Blind CLlff Walking Setup

3 Brief Literature review for the original research track only:

4 Summary of the paper: Model/Approach details:

As mentioned above, the authors argue that introducing a prioritized sampling in the replay buffer improves the performance of algorithms. The priority used is proportional to a latent value that is not obtainable easily. The authors decided to use the TD error as a proxy metric as it is both easily computed and logically justified. One problem that appears is that first state will almost always have a high error which will cause the related transition to have a high priority. The agent will be biased towards the initial transitions. To counter this problem, a stochastic metric was introduced. The algorithm will sample transitions based on a probability computed as follow:

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}$$

The p_i value are computed using one of the 2 variants:

1. Using a proportional prioritization
2. Using a rank-based prioritization.

Finally, sampling using the probabilities introduced above will introduce a bias. This bias can be corrected using Weighted Importance Sampling. Namely, during training update, we multiply the TD error with some weights computed as follow:

$$w(i) = \left(\frac{1}{N} \cdot \frac{1}{P(i)}\right)^\beta$$

The β term controls the prioritization to be applied. It starts small (0.4-0.6) and is annealed towards one during training. The PER component was tested on the collection of Atari benchmarks (Bellemare et al., 2012).These environments present many problems encountered by the RL algorithms such as delayed credit assignment, partial observability, and difficult function approximation (MnihV et al. (2015); van Hasselt et al. (2015))

The authors used 2 different models: DQN and Double DQN. The replay buffer had a size of 10^6. The mini batch size was 32. Rewards and TD-errors are clipped to fall within [1, 1] for stability reasons. These models were trained first with a standard replay buffer (as a baseline) and then with the PER mechanism. The metric of comparison used was average score per episode, given start states sampled from human traces. The results clearly show that the 2 algorithms using PER performed substantially better than the standard ER. In fact, PER was ahead in 41 of the total 49 games "with the median normalized performance across 49 games increasing from 48% to 106%." The results also showed that mixing Double DQN and PER yielded state of the art performance at the time. Finally, PER increased the rate of convergence in almost all the games.

5 List Research/Analysis Questions that you will pursue.

Our project is an ablation study of the Prioritized Experience Replay paper.The goal of the project is three folds:

In the first place, we will reproduce the prioritized experience replay paper to create base of comparison for our future experiences.

Second, we will attempt to replace the TD error as a proxy for the ideal priority measure of expected learning progress with the following metrics/methods:

- The derivative of the TD error.
- The norm of the weight-change induced by replaying a transition.
- Treating positive TD error transitions and negative ones separately.
- Boosting the replay probabilities of entire episodes, instead of transitions.

Finally, we will try different methods of introducing stochasticity in the sampling process. These procedures are listed below:

- Using a hybrid approach where 2 samples are sampled using different priorities,
- Increasing priorities of transitions that have not been replayed for a while.

6 Experiment setup, Dataset/Environment details:

- OpenAI gym - CartPole : We will use this environment to investigate what extent replay with such prioritized sampling can improve performance on cartpole environment.

7 Plan for contributions by each team member:

In the first place, we will use OpenAI gym - CartPole environment and reproduce the same experiments (proportional prioritization and rank-based prioritization) as in the paper. In the second phase, each team member will take up at least one analysis question as described in section 5 and will carry out the ablation study.

References

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *CoRR*, abs/1207.4708, 2012. URL `http://arxiv.org/abs/1207.4708`.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Mnih V, Kavukcuoglu K, Silver D, and Bellemare MG Graves A Riedmiller M Fidjeland AK Ostrovski G Petersen S Beattie C Sadik A Antonoglou I King H Kumaran D Wierstra D Legg S Hassabis D. Rusu AA, Veness J. Human-level control through deep reinforcement learning. *Nature. 2015 Feb 26;518(7540):529-33. doi: 10.1038/nature14236. PMID: 25719670.*, 2015.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. *CoRR*, abs/1509.06461, 2015. URL `http://arxiv.org/abs/1509.06461`.

1 Project Proposal 10 / 10

- ✓ - **0 pts** **Good Plan.**
- - **1 pts** Need plan change.
- - **1 pts** Be careful with your ablations.