PRIORITIZED EXPERIENCE REPLAY

Google DeepMind, 2016 ICLR

Bio-Medical Computing Laboratory

TAEHEUM CHO

25th JUN, 2018

INTRODUCTION

- How prioritizing which transitions are replayed can make experience replay more efficient and effective than if all transitions are replayed uniformly.
- Some transitions may not be immediately useful to the agent, but might become so when the agent competence increases (Schmidhuber, 1991).
- Propose to more frequently replay transitions with high expected learning progress, as measured by the magnitude of their temporal-difference (TD) error.

PRIORITIZED REPLAY

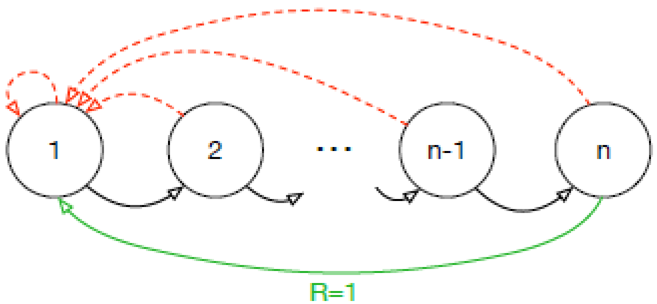
- PRIORITIZING WITH TD-ERROR
  - The central component of prioritized replay is the criterion by which the importance of each transition is measured(TD error δ).
  - TD error , which indicates how 'surprising' or unexpected the transition is.
  - The TD-error can be a poor estimate in some circumstances as well, e.g. when rewards are noisy
  - The transition with the largest absolute TD error is replayed from the memory.

$$V(S_t) \leftarrow V(S_t) + \alpha \underline{\Big[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\Big]}.$$

" They learn a guess from a guess"

PRIORITIZED REPLAY

STOCHASTIC PRIORITIZATION

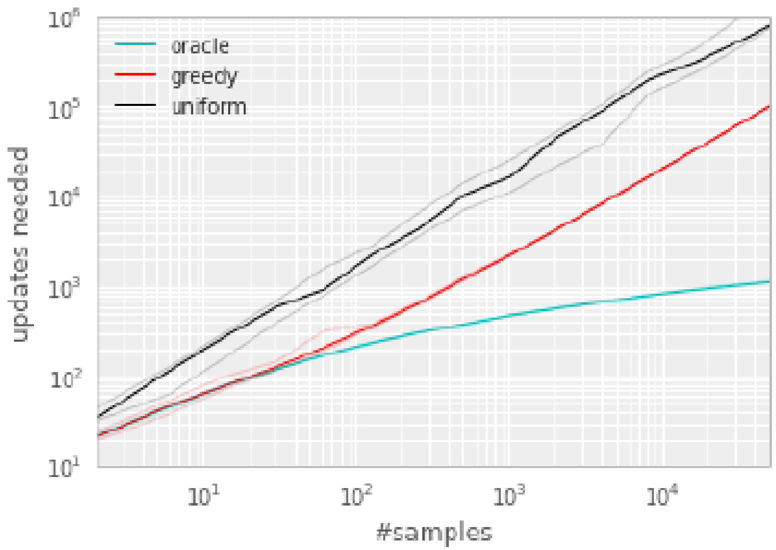


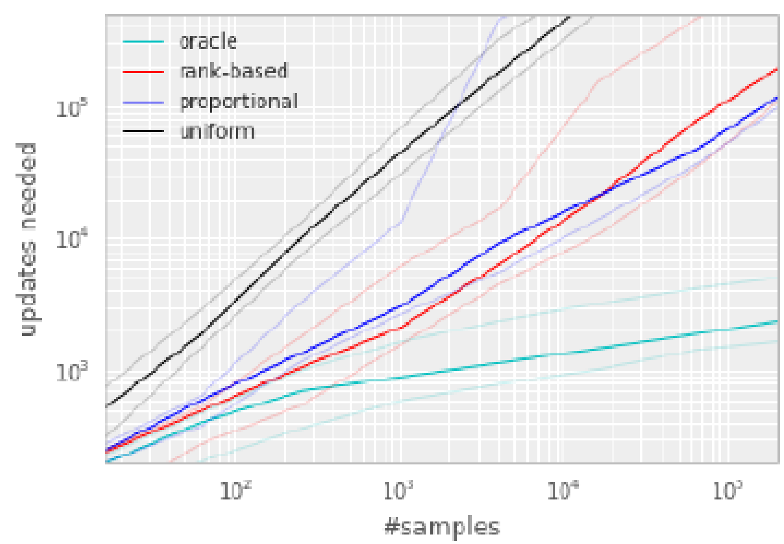
1. there are two actions, a 'right' and a 'wrong' one
2. Taking the 'right' action progresses through a sequence of n states at the end of which lies a final reward of 1, reward is 0 elsewhere.

PRIORITIZED REPLAY

STOCHASTIC PRIORITIZATION

oracle: Priority learned in advance (impossible in real world)





PRIORITIZED REPLAY

- STOCHASTIC PRIORITIZATION
  - Greedy TD-error prioritization has several issues.
  - Transitions that have a low TD error on first visit may not be replayed for a long time (which means effectively never).
  - Further, it is sensitive to noise spikes (e.g. when rewards are stochastic), which can be exacerbated by bootstrapping.
  - lack of diversity that makes the system prone to over-fitting.

- To overcome these issues, they introduce a stochastic sampling method that interpolates between pure greedy prioritization and uniform random sampling.

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

PRIORITIZED REPLAY

Variant 1: proportional prioritization

$$p_i = |\delta_i| + \epsilon.$$

ϵ is a small positive constant that prevents the edge-case of transitions not being revisited once their error is zero. δ is the TD-error

Variant 2: rank-based prioritization

$$p_i = \frac{1}{\text{rank}(i)}$$

rank(i) is the rank of transition i when the replay memory is sorted according to δi

latter is likely to be more robust, as it is insensitive to outliers

They expected Variant 2 is likely to be more robust, as it is insensitive to outliers.

But, both variants perform similarly in practice.

Both variants of stochastic prioritization lead to large speed-ups over the uniform baseline

PRIORITIZED REPLAY

- ANNEALING THE BIAS
  - In typical reinforcement learning scenarios, the unbiased nature of the updates is most important near convergence at the end of training, as the process is highly non-stationary anyway, due to changing policies, state distributions and bootstrap targets.

PRIORITIZED REPLAY

- ANNEALING THE BIAS
  - Prioritized replay introduces bias because it changes this distribution in an uncontrolled fashion. We can correct this bias by using importance-sampling (IS) weights.
  - Note that the choice of this hyperparameter interacts with choice of prioritization exponent ; increasing both simultaneously prioritizes sampling more aggressively at the same time as correcting for it more strongly.
  - In practice, we linearly anneal from its initial value 0 to 1.

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

$$V(S_t) \leftarrow V(S_t) + \alpha\left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right].$$

These weights can be folded into the Q-learning update by using $w_i\delta_i$ instead of δ_i

PRIORITIZED REPLAY

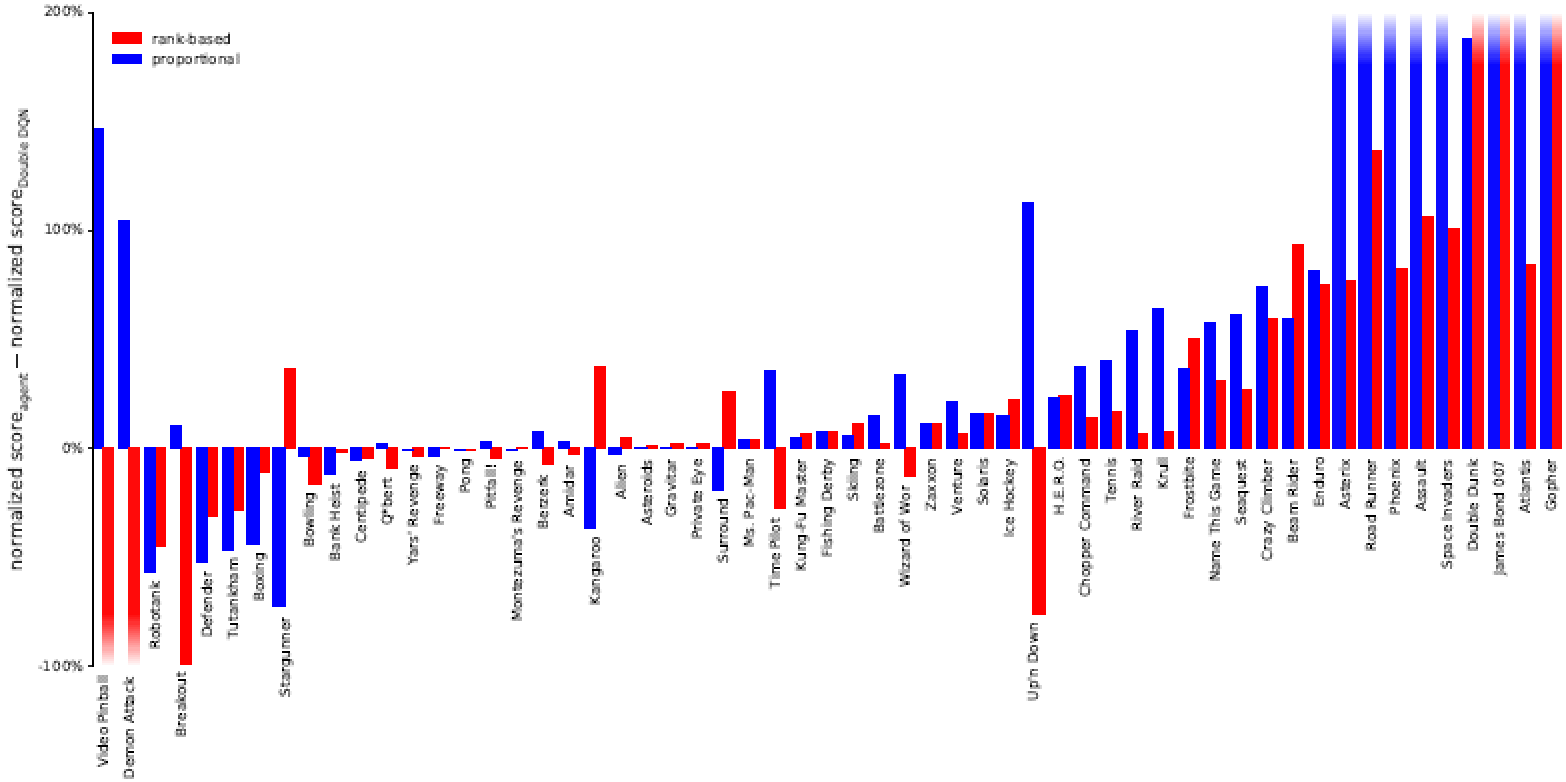


Figure 3: Difference in normalized score (the gap between random and human is 100%) on 57 games with human starts, comparing Double DQN with and without prioritized replay (rank-based variant in red, proportional in blue), showing substantial improvements in most games. Exact scores are in Table 6. See also Figure 9 where regular DQN is the baseline.