Rapport de stage

Paul Ecoffet

15 avril 2016

Abstract

** Reinforcement learning, navigation, replay, cost of exploration **

1 Introduction

2 Reinforcement learning

Reinforcement Learning is a kind of learning mechanism where an agent tries to maximise a reward by doing some specific action in an environment (R. S. Sutton & Barto, 1998). ** bla bla **

3 Learning by replay

3.1 In reinforcement learning

Reinforcement learning methods like Q-learning or Sarsa converge slowly and need a lot of samples to be efficient. These algorithms, based on the TD(λ), use the sample once to improve their solution, then discard it (Adam, Busoniu, & Babuska, 2012).

To improve the efficiency of reinforcement learning methods, experience replay can be use. Vanseijen and Sutton (2015) offers a good analysis about learning by replay. The goal of replay is to use the maximum of information an experience offers. Each sample is used several times, improving the solution several times. Compared to TD(0), replay techniques give a better convergence to the optimal solution with the same number of experiences. Reinforcement learning methods using replay are more expensive than TD(λ) both in memory and in computations, though the memory and computational power needed can be reduced a lot as Vanseijen and Sutton (2015) shows.

3.1.1 Planning and replay?

RL algorithms with planning use a model of their environment to infer behaviour they haven't done yet. The model is commonly a transition table between the possible state in the environment according to the action done by the agent. For exemple, if the agent has to solve a maze and has access to a transition table, then it can look up the table and know that if he is in the case $(3, 5)$ and it does the action "go to the east", it will be in $(3, 6)$.

The model can be either directly accessible or inferred by the agent. When the model is inferred, the agent builds explicitly the transition table according to its experiences. Dyna is a class of RL algorithms where the model is inferred by the agent. There is two phase during an iteration of Dyna. The first phase is ** bla bla ** and the second is the *planning* phase. During the planning phase, the agent improves its value function estimation with simulations according to its model: The agent generate a sample according to some rules (a probability distribution) and use its model to compute the value of this sample. Then, it updates its solution according to this simulation (R. S. Sutton, Szepesvári, Geramifard, & Bowling, 2012). It is interesting to note that the distribution used has no influence on whether the solution will converge or not, but has an influence on the speed of convergence. Therefore, choosing a good distribution can greatly improve the performance of Dyna. R. S. Sutton et al. (2012) shows that choosing only already explored samples can have a good impact on performances. Therefore, Dyna can be efficient with a selection of already experienced samples, the planning phase is then a *replay* phase.

Vanseijen and Sutton (2015) highlights even more the strong similarity between RL methods using replay and RL methods using planning. First of all, both of them can be considered as model-based algorithm. Indeed, planning methods using an inferred model use an integrated version of samples experienced over time. The sample isn't discarded once the solution is updated. RL algorithms using experience replay ** considered as model-based algo **

In addition to that, Vanseijen and Sutton (2015) shows a complete equivalence in the value function approximation at each iteration between a TD(0) inspired algorithm with replay and Linear Dyna, a model-based RL algorithm. Though Linear Dyna is a model based algorithm and thought as a "looking ahead" algorithm and the replay algorithm as a "looking backward" algorithm, they do the exact same calculation.

- Equivalence between replay and planning as show in Vanseijen and Sutton, 2015.
- Can one use Linear dyna with prioritised sweeping as a learning by replay method?

3.2 In vivo

Replays of sequences have been observed in the rat after navigation tasks (Wilson & McNaughton, 1994; Skaggs & McNaughton, 1996; Davidson, Kloosterman, & Wilson, 2009; Gupta, van der Meer, Touretzky, & Redish, 2010).

- Replay during sleep and during hesitation
- Replay can be either forward or backward
- Replay can also be action that has never be done? –> Planning somewhat?
- Replay aren't necessary about the most recent event?
- Is "surprise" a good criterion?

Gupta et al., 2010

4 Model navigation learning

5 Place cells

Place cells are high level integrative neurons in the hippocampus. They have a specific receptive field

- Highly integrated
- Response depending of the position of the rat
- Involved in replay

6 What are the sequences that are replayed? How are they selected?

Gupta et al. (2010) shows that the sequences that are replayed are not necessarly the most recent experiences. It shows also that the sequences that are replayed can be either forward or backward. In forward replays, the sequence is replayed in the same order as the sequence the rat has done. In backward replays, the sequence is done in the reverse order. What is the explanation for the selection of the sequence which is replayed, and its order? No criteria are proposed in Gupta et al. (2010). Yet, we can make interesting hypothesis knowing that R. S. Sutton et al. (2012) suggests that there are very good samples to replay in a Dyna algorithm, and those sample are the sample where the difference between the expected value of a state and its real value is the greatest. It is consistant with Gupta et al. (2010) which states that the most common replays are the one that leads to the reward or comes from the reward.

7 Limitations

References

Adam, S., Busoniu, L., & Babuska, R. (2012, March). Experience replay for real-time reinforcement learning control. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(2), 201–212. doi:10.1109/TSMCC.2011.2106494

Davidson, T. J., Kloosterman, F., & Wilson, M. A. (2009, August). Hippocampal replay of extended experience. *Neuron*, *63*(4), 497–507. doi:10.1016/j.neuron.2009.07.027

Gupta, A. S., van der Meer, M. A., Touretzky, D. S., & Redish, A. D. (2010, March). Hippocampal replay is not a simple function of experience. *Neuron*, *65*(5), 695–705. doi:10.1016/j.neuron.2010.01.034

Skaggs, W. E. & McNaughton, B. L. (1996, March 29). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science (New York, N.Y.) 271*(5257), 1870–1873.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement learning: an introduction* (Nachdr.). Adaptive computation and machine learning. Cambridge, Mass.: MIT Press.

Sutton, R. S., Szepesvári, C., Geramifard, A., & Bowling, M. P. (2012). Dyna-style planning with linear function approximation and prioritized sweeping. *arXiv preprint arXiv:1206.3285*. Retrieved March 18, 2016, from http://arxiv.org/abs/1206.3285

Vanseijen, H. & Sutton, R. (2015). A deeper look at planning as learning from replay. In *Proceedings of the 32nd international conference on machine learning (ICML-15)* (pp. 2314–2322). Retrieved April 1, 2016, from http://machinelearning.wustl.edu/mlpapers/paper_files/icml2015_vanseijen15.pdf

Wilson, M. A. & McNaughton, B. L. (1994, July 29). Reactivation of hippocampal ensemble memories during sleep. *Science (New York, N.Y.) 265*(5172), 676–679.