WEEKLY RESEARCH PROGRESS REPORT: 37

Shangshang Wang
School of Information Science and Technology
ShanghaiTech University
wangshsh2@shanghaitech.edu.cn

1 QUOTE OF LAST WEEK'S PLAN

- Systematically review of probablity and oprimization for shcool courses
- Five kinds of Exploration methods in RL
- Survey the way of reusing data (increase sample efficiency) in RL

2 PLANNED ACCOMPLISHMENTS

Finished

- Systematically review of probablity and oprimization for shcool courses
- Five kinds of Exploration methods in RL
- Survey the way of reusing data (increase sample efficiency) in RL
- Self-consistency in path-consistency RL
- Take a deeper look: The role of Replay Buffer

Unfinished

- Self-consistency in path-consistency RL

3 OTHER ACCOMPLISHMENTS

- book: Deep Learning with Pytorch
- paper: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling (Chung et al., 2014)
- paper: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation (Cho et al., 2014)
- paper: Randomized Prior Functions for Deep Reinforcement Learning (Osband et al., 2018)
- paper: A Tutorial on Thompson Sampling (Russo et al., 2018)
- paper: Deep Exploration via Bootstrapped DQN (Osband et al., 2016)

4 ISSUES AND PROBLEM TO SOLVE

- Current version of weekly working plan is not as useful as I thought.
  Several weeks back, I found a problem that the lack of overall view of the classic and cutting edge research of RL tempted me to read papers of many different subfields rather than focus on one at a time. At first, I tried not to work on too much things altogether, but the impecunious confidence of my studying experience got in the way and without proper background of the whole field of the RL (even AI), I almost always got a lot of trouble in learning one specific topic e.g. I need RNN and GRU from NLP world to learn RL models and I need deeper stochastic process knowledge to understand langevin DQN. So, after

some struggling, I decided to learn the "general" knowledge first (things I did in the last 36 weeks, not so satisfying though).

Now, I think have roughly come through the aforementioned dilemma and it is better for me to focus on one small task from now on. Though I haven't avioded zigzags in the last 36 weeks, with a relative richer knowledge I feel at least more confident to work efficently on my own. The plan is very important but I sometimes find myself not following the plan, I think the reason is that the plan is not concrete and practical enough. Since I have decided to switch to smaller but deeper learning on a topic at a time, there is an urge of adjust the plan content e.g. add more comment on exactly what I should next week.

5 NEXT WEEK'S PLAN

- Train on more existing envs: Atari, openai gym, unity RL, street fighters, sonic, paopao kadingche and so on.

  Last week, I read one paper about the PPO and TRPO (IMPLEMENTATION MATTERS IN DEEP POLICY GRADIENTS: A CASE STUDY ON PPO AND TRPO) which emphasizes on the importance of implementation details and variaties, we should not only specifcy the parameters, settings but also we should use more envs to test the correctness and generalization. If I remember corretly, the most used env for is the car pole agent which is far from enough. This job requires me to: first implement some classic algorithms like PPO, Rainbow and DDPG; Second, get access to all aforementioned envs; Third, run the algorithms on them and use different tools like tensorboard and plotting to visualize the results; Fourth, while reading interesting papers I should either find implementation online or do it myself and then deploy it on the envs to check out the result.

  also, I am thinking about use arduino or iphone (there are people doing these!) to further the possibility of RL agents.

REFERENCES

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pp. 4026–4034, 2016.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 8617–8629, 2018.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.