IP Summary - Reinforcement Learning

Aman Mehra

December 23, 2020

This project began with the aim of surveying and understanding the existing literature in Reinforcement Learning. In the first few weeks, I performed a literature survey of existing DQN architectures that are employed in the field. The four major architectures that I surveyed were - DQN [5], Double DQN [3], Dueling Networks [10] and Rainbow [4]. Minh *et al* [5] enabled deep neural networks to be employed as functional approximators for Q learning by handling delayed, sparse and noisy rewards and breaking data correlation through replay memories. The subsequent papers incrementally improved this architecture. Double DQN handled the over estimation bias in Q-learning due the max() function by using a second network to obtain bootstrap values for the Q value target. Dueling networks uses two streams to disentangle the value function and action advantage function with the aim of learning good state estimates without needing to observe all possible actions from this state. Rainbow combines the recent advances in RL architectures, experience replay and rewards to achieve state of the art performance on the Arcade Learning benchmark. Alongside, I also explored the area of intrinsic motivation through the paper on curiosity driven exploration [7].

Subsequently, the focus shifted towards understanding replay memories. Initial works in this domain such as PER [8], CER [12] and PSER [1] were first studied. These focused on TD error based prioritization of transitions in the replay memories. Newer methods follow a different approach by introducing the notion of on-policyness. These works try to maintain more on policy transitions with respect to the current agent policy in the replay memory. The two major works that were analyzed were ERO [11] and ReF-ER [6]. ERO learns a replay policy that chooses samples that maximize the improvement in the cumulative reward of the agent. This approach surpasses TD error based prioritization methods while also contradicting the older paradigm by picking low TD-error transitions instead of high TD-error transitions. This finding puts the notion of on-policy transition selection to the fore. ReF-ER presents an alternate to pick on-policy transitions. It solves the problem by using the importance sampling ratio to classify transitions as near[on] or far[off] policy transitions and back-propagating TD error of only near policy transitions. It also uses KL divergence loss to smoothen the change of the current target policy with respect to an older target policy. Effectively, this paper presents a scheme to smoothen the change in policy of an agent in the later stages of learning. Additionally, an orthogonal

approach of Importance Resampling [9] was analyzed wherein transitions are sampled with a probability proportional to their IS ratio to reduce the overall variance of the update.

In the remaining period we analyzed the findings of a recent paper[2] that studied the interplay between replay memory parameters and different aspects of the RL pipeline to determine what parts enabled improved returns with larger replay memories. Their findings were that n-step returns were the key to seeing improved returns with larger replay memories as the larger memory compensated for the added variance of the update target. Since then we have been trying to reproduce and verify the results of this paper's experiments. The subsequent line of work is to investigate whether this hypothesis holds for low variance multi step returns like tree back up and n-step average DQN bootstraps. Concretely, we want to investigate the hypothesis that - larger replay memories reduce the variance of the target E[Gt] and thus help mitigate the added variance of the n-step return. We aim to study the trends of this variance for different returns to gain a deeper understanding of the phenomenon.

References

[1] Brittain, M., Bertram, J. R., Yang, X., and Wei, P. Prioritized sequence experience replay. *CoRR* (2019).

[2] Fedus, W., Ramachandran, P., Agarwal, R., Bengio, Y., Larochelle, H., Rowland, M., and Dabney, W. Revisiting fundamentals of experience replay. In *International Conference on Machine Learning* (2020).

[3] Hasselt, H. v., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence.

[4] Hessel, M., Modayil, J., van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M. G., and Silver, D. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[5] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature* (2015).

[6] Novati, G., and Koumoutsakos, P. Remember and forget for experience replay. In *Proceedings of the 36th International Conference on Machine Learning* (2019).

[7] Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning* (2017).

[8] Schaul, T., Quan, J., Antonoglou, I., and Silver, D. Prioritized experience replay. In *International Conference on Learning Representations* (2016).

[9] Schlegel, M., Chung, W., Graves, D., Qian, J., and White, M. Importance resampling for off-policy prediction. In *Advances in Neural Information Processing Systems* (2019).

[10] Wang, Z., Schaul, T., Hessel, M., Van Hasselt, H., Lanctot, M., and De Freitas, N. Dueling network architectures for deep reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning* (2016).

[11] Zha, D., Lai, K.-H., Zhou, K., and Hu, X. Experience replay optimization. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19* (2019).

[12] Zhang, S., and Sutton, R. S. A deeper look at experience replay. In *Symposium on Deep Reinforcement Learning at the 31st Conference on Neural Information Processing Systems* (2017).