DQN

Sheng Zhong

March 27, 2019

The paper introduces the now ubiquitous DQN to do RL with neural nets. Its strength of allowing features of a task to be learned is demonstrably effective. This review will mostly focus on potential improvements. A side note is that it's a bit odd this paper was published in nature and tried really hard to incorporate biological connections to the RL...

One important difference of the DQN to previous attempts to approximate the Q function with a neural net is what the authors call experience replay. What it essentially does is to train the network probabilistically off-policy. Intuitively this prevents local minima where the agent keeps taking similar actions and the reward space also stays in a small region - it forces the network to generalize and explain previous experience resulting from a different policy. This is an important technique with broad applications.

One potential weakness is that the vanilla DQN is a model-free network: it is learning by directly remembering (state sequence,action) mappings to reward rather than some underlying model explaining why the actions lead to rewards. It is learning correlations rather than causations. The cost of this is experience/data inefficiency. An improvement is to adjust the model to learn on a latent space, and recent work on it proved increased data efficiency by orders of magnitude [1].

Another weakness of DQN is its total dependence on a simple extrinsic reward function. For the Atari games this was not a problem, but it would be for real life tasks. A class of methods for addressing this is value learning (aka inverse reinforcement learning): learning the value function from demonstrations, such as [2].

Apart from the existence of the external cost function, having sparse rewards or deceptive rewards would also break the DQN. The 0% performance in Montezuma's Revenge where reward is very sparse (only upon completing a level) proves this. Recent work to address this has been to add instrinsic rewards associated with curiosity and exploring the search space [3]. Additionally it was shown that we can achieve super-human performances by explicitly not exploring similar states and prioritizing exploration around remembered promising states (performance depends heavily on how we define state similarity) [4].

References

[1] S. Sæmundsson, K. Hofmann, and M. P. Deisenroth, "Meta reinforcement learning with latent variable gaussian processes," *arXiv preprint arXiv:1803.07551*, 2018.

[2] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich, "Maximum margin planning," in *Proceedings of the 23rd international conference on Machine learning.* ACM, 2006, pp. 729–736.

[3] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros, "Large-scale study of curiosity-driven learning," *arXiv preprint arXiv:1808.04355*, 2018.

[4] A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune, "Go-explore: a new approach for hard-exploration problems," *arXiv preprint arXiv:1901.10995*, 2019.