Paper: *Rainbow: Combining Improvements in Deep Reinforcement Learning*

Nirajan Koirala

CSC 9010_001

Summary

Hessel et al. perform a empirical study of the combination of six DQN algorithm extensions as well as ablation study on each individual extension to the overall performance on 57 Atari games in this paper. Individually each of these algorithms are found to provide different kinds of performance boost. They show how an agent can combine all these improvements and achieve a new state-of-the-art results on the benchmark suite of various Atari 2700 games. They first provide a background on reinforcement learning where an agent acts in an environment to maximize a scaler reward signal without any direct supervision and the goal is to maximize the expected discounted return by finding a good policy. Agent needs to find the forms of action which balances both exploitation and exploration, so that it won't act in an extremely greedy fashion. Next, a set of extentions of the DQN algorithms which addresses distinct concerns is elaborated.

The first extension, Double Q-learning tackles the overestimation bias of conventional Q-learning by decoupling. Prioritized reply samples transitions with a certain probability relative to the last encountered absolute TD error which increases learning potential of regular DQN algorithm. Another extension named dueling networks is s neural network architecture designed for value based RL. It shares a convolutional encoder with value and advantage streams of computation and merges them together with a special aggregator. Mutistep-learning is another form of extension in which Q-learning is accumulates a single reward and then uses the greedy action at the next step to boot-strap. It minimizes the alternative loss and leads to a faster learning. Distributed RL, learns the probability masses and approximate the distribution of returns instead of the expected return. This return distribution satisfy a variant of Bellman's equation, and it can be derived by first constructing a new support for the target distribution and then minimizing the Kullbeck-Leiber divergence between the distribution d_t and the target distribution d^l_t. Noisy nets deals with the limitation of exploring using greedy policies by using a noisy linear layer that combines a deterministic and noisy stream. Overtime this helps the network ignore the noisy stream allowing state-conditional exploration with a form of self-annealing. All the six extensions are combined into a single integrated agent called Rainbow. They elaborate the process of combining these extensions and then describe the methods and setup used for configuring and evaluating the learning agents.

They follow the training the evaluation procedures of Mnih et al. and van Hasselt et al. and agent's scores are normalized per game. For hyperparameter tuning, they start with the values used in the paper which introduced then this component then tune the most sensitive ones among the hyper-parameters by manual coordinate descent. In figure 2, they compare the performance of their agents with human performance and their agent Rainbow is evidently superior to all other agents in the comparison graph. In the second row of the figure they also provide graphs of rainbow agent without particular extensions. This ablation study is specifically important as it helps to identify where the overall improvements in performance come from. During the ablation study, they found that prioritized

replay and multi-step learning were the two most crucial components of the rainbow agent. Removal of either of these extensions hurt early performance but the removal of multi-step learning also hurt the final performance. Specifically, they found that in the early learning no difference is observed between distributional-ablation, noisy net ablation and the rainbow agent but eventually the agent without distribution and noisy nets starts lagging behind. For dueling network, they did not observe any significant difference in removing the network from the rainbow. In case of double Q-learning, the difference in median performance is found limited with the component sometime harming or helping depending on different games.

Finally they conclude by pointing out that there are other algorithmic components as well which which use purely policy-based RL algorithms. They mention that they didn't include those in this study and discuss about the many possible candidates which could be used in the future.

Strengths

- Paper explores various extensions of DQN algorithms and even provide the ablation studies.
- They provide final performance for each of the individual games for Rainbow, its ablations and baselines.
- The elaboration of their experiments and analysis was very thorough.

Weaknesses

- Organization of the paper could have been made better by breaking into different sections.

Points of Confusion

- They hypothesize that clipping the values of constrained range counteracts the over-estimation bias of Q-learning. I am not sure how?
- The section dealing with integrating the agents was confusing.

Discussion Questions

- Could it be inferred from the figure 4 that some Atari 2700 games which have very similar playing style will have performance drops if some specific algorithmic-extension is removed from rainbow which only helped to gain performance in one of those games?
- How does the hyper-parameter Adam optimizer work and how is it different than RMSProp?
- How does the purely policy-based RL algorithms such as trust-region policy optimization and actor critic methods differ from value-based methods in the Q-learning family?