Jenish Maharjan **Week 12**

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

Summary:

Hessel et al present their experiments to examine six extensions to the DQN algorithm and the empirical studies on the combination of these extensions. They present a detailed ablation study to deduce the contribution of each component to the overall performance. Many extensions have been made to the original Deep Q-Networks Algorithm (DQN) by Mnih et al. that combined Q-learning with convolutional neural networks and experience replay. Since most of these algorithms improve the performance in isolation by focusing on very different issues and are built on the same framework, the authors hypothesize a possibility of combining these algorithms and thus perform empirical studies by combining them.

The authors provide a background of Reinforcement learning, DQN and the extensions made to DQN that they've used in their study. In reinforcement learning, the agent's problem of learning to act in an environment is solved by taking learning a policy that maximized a scalar reward signal. At any discrete time-step, the agent is provided with an observation from the environment based on which the agent selects an action. The agent is rewarded by the environment for the action. The environment also provides a discount value and the next state. The agent selects the action based on a policy that defines a probability distribution over actions for each state. In Deep reinforcement learning, the components agents such as policies and values are represented with deep neural networks that are trained by gradient descent to minimize some suitable loss function. In DQN, the action values of a given state were learned by combining deep networks and reinforcement learning by using a CNN.

Six extensions to DQN that addressed a limitation and improved overall performance were selected for the study. Double DQN (DDQN) that addresses an overestimation bias of Q-learning by decoupling the selection of the action from its evaluation. Prioritized experience reply improves data efficiency by sampling uniformly the replay buffer. The dueling network architecture helps generalize across actions by implementing a value-based reinforcement learning. Multi-step learning shifts the bias variance trade-off. Distributional Q-learning learns categorical distribution of discounted returns. Noisy DQN uses stochastic network layers for exploration – over time, the network is able to learn to ignore the noisy stream at different rates in different parts of the state space allowing state-conditional exploration with a form of self-annealing. The concept of integrated agent is presented in the paper in which the six components are integrated into a single integrated agent called Rainbow. The 1-step distributional loss is replaced with a multi-step variant. The target distribution is constructed by contracting the value distribution in the next state. The multi-step distributional loss is combined with double Q-learning by using the greedy action selected according to the online network as the bootstrap action. The linear layers are replaced with their noisy equivalent layers as in the noisy networks.

For the evaluation of these extensions, the authors follow the training and evaluation procedures of Mnih et al and van Hasselt et al. The average scores of the agent are evaluated during training by suspending learning and evaluating the latest agent for 500K frames. Agents' scores are normalized per game. They use the median human normalized performance for comparison. Two testing regimes – no-ops starts regime and human starts regimes are used to re-evaluate the best agent snapshot at the end of training.

Since the combinational space of hyper-parameters is too large for an exhaustive search, limited tuning of the hyperparameter is performed.

Compared to published baselines, the Rainbow's performance is significantly better. Rainbow is able to achieve a median score of 223% in the no-ops regime and 153% in the human-starts regime after the end of training. Each agent of rainbow was run on a single GPU just like the original DQN setup. Additional experiments were carried out to understand the contribution of the various components of Rainbow in the context of the presented combination. In each of these ablation studies, one component was removed from the combination. The ablation studies showed that prioritized replay and multi-step learning were the two most crucial components of Rainbow. Distributional Q-learning was the third most crucial component followed by Noisy Nets while the effects of removing dueling networks or double Q-learning on the performance were limited. By providing a new way of combining these extensions and performing ablation studies on these, the authors were able to provide a new perspective on these extensions.

Strengths:

- The authors introduced a novel concept of an integrated agent by combining various extensions made to the DQN which maximized the performance of the agent.
- The authors present ablation studies to understand the effects of each extension in the combination which helps better understand the role of the extensions in the improvement in performance.
- The paper is well structured and provides brief background on each of the six extensions which is really helpful.

Weaknesses:

- They could have explained each of the six components in more detail. For someone reading about them for the first time, the descriptions seem inadequate.

Confusions:

- I couldn't understand the relationship between game episodes and frames? Were multiple game episodes combined for training the model?
- I wasn't sure how to read figure 4.

Discussion Questions:

- A short description of each of the six extensions would be really helpful.
- Is the method of ablation exhaustive? Is studying the performance of such components in all possible combinations a feasible method?
- Do the current state-of-the-art DQN models use this sort of combinations?