Andrew Garcia
ECE 5984 - Deep Reinforcement Learning
February 21, 2023

Paper Review Assignment #1

Title: Deep Recurrent Q-Learning for Partially Observable MDPs

The authors of the paper explore the capabilities of what they call Deep Recurrent Q-Network (DRQN), an evolution of Deep Q-Networks (DQNs), as a means for solving Partially Observable Markov Decision Processes (POMDPs). They test their models on a specific set of Atari 2600 games such as Pong, Frostbite, and Double Dunk because they all incorporate a moving object as an objective in the game. The main reason for using POMDPs is because the real-world is often filled with incomplete and noisy state information. Traditional Q-Learning techniques are limited on how many past states they can learn from which is why the researchers in this paper implement their video games to be a POMDP for training. This paper does a great job at laying down the foundation of the problem they are trying to solve and how they plan to use their model. It is discovered that when trained with reduced observations, the performance of the DRQN is better than that of a DQN. Therefore, the method outlined in this paper seems to be a valid approach to adaptation in a real-world environment.

The main contribution of this research is adding in "partial observability" to replicate real-world environments. The method for introducing this into the system is by giving each frame the probability of 0.5 to be blank, this is what they refer to here as "flickering." A POMDP is described to have the following features: state, action, transitions, reward, set of observations, and the set of conditional observation probabilities (S, A, T, R, Ω, O). The researchers state that adding recursion to Deep Q-Learning allows the Q-Network to "better estimate the underlying system state." I find this part of the research to be very difficult to understand when they say recursion while talking about the POMDP experiment. I understand it as running the learning model itself thousands of times from start to finish but it can also mean recursion of all actions given each state. Either way, their diagram for the model is very well depicted in Figure 2. In order to convert pixels into something that can be analyzed for learning, they use three layers convolutional filters that feed into a Long Short Term Memory (LSTM) layer to calculate the correct Q-values.

The paper is very convincing as it is remarkable that the DRQN is able to perform well in a POMDP environment and still outperforms a DQN given full information. Overall, it seems like this experiment

produced a great model for robustness in machine learning by demonstrating that it is possible to understand moving objects given static states. It would be very exciting to see the researchers apply the use of DRQNs in larger POMDP like 3-D games where there are millions of possible states with countless different actions.

Title: Prioritized Experience Replay

The researchers in this paper are trying to tackle with the issue that online reinforcement learning loses information each time they update their model or policy, which happens almost daily. Experience Replay is an existing method that addresses the issues with policy updates by playing back what they refer to as "memories" to reduce the amount of experience required to learn. However, instead of playing back all of the stored memories, the researchers in this paper attempt to come up with a method for identifying which ones are best for playing back. As it turns out, using "prioritized replay speeds up learning by a factor of 2," which is amazing in real-time online models where updates are time sensitive.

The main contribution of this paper is providing the metrics for identifying valuable experiences that should be played back, resulting in speeding up learning for Deep Q-Network (DQN) models. This paper utilizes two different metrics (1) Temporal Difference (TD) Error and (2) Stochastic Prioritization, which are tested on the Atari 2600 benchmark suite. TD error is a measure that indicates how unexpected a transition is from one scene to the next is, but using only this has it's downfalls and shortcomings. To make up for TD errors issues, stochastic sampling is introduced to ensure that all samples still have a non-zero change at being used. In the paper's findings, they discovered that adding prioritized replay to DQN "leads to substantial improvement" across the Atari benchmark of various video games.

This papers main strength is that they are very straightforward and to the point about their results at each section. After reading the "Introduction" section of the paper, it was very clear to me what the problem was that they were trying to solve and how they would solve it. However, this paper ironically struggles with a strong conclusion being only a few sentences. I think this may be due to a large Discussion and Extension section where these types of conversations would normally make their way in to.

Overall, I think this research is incredibly convincing with the pages of data that backs up their findings. I also find it amazing that they put in the effort for exploring Prioritization Variants in Appendix A, this level of detail shows the dedication and thoroughness that these researchers have. It would be nice

to see where the authors think this type of research could be applied outside of video games. I think this methodology of prioritizing replays can be beneficial to RL problems where time is very sensitive like social robots or NPCs in a video game. It would be interesting to see how this approach extends to chat bots where policy updates can be so fast that there is no buffer in between message and reply.