*Reinforcement Learning*

Reinforcement learning (RL) is a very popular area of Artificial Intelligence research. It is a machine learning technique in which an agent makes decisions based on the context of an environment. The agent learns through trial and error to maximize its reward based on the feedback generated from its actions and experiences. In Figure 1, a pictorial representation of the reinforcement learning model is represented. Various applications are a good match for the RL approach as the environment is unspecified and the outcome of actions are undetermined. Some real-world applications include autonomous cars and robotics.

Video games have been a popular domain for RL implementation as they provide interesting and complex problems for agents to solve. Once such example [1] is where an agent learned to play the Atari 2600 with deep reinforcement learning. In this project, we sought to implement a RL algorithm capable of learning how to play the Atari breakout game.

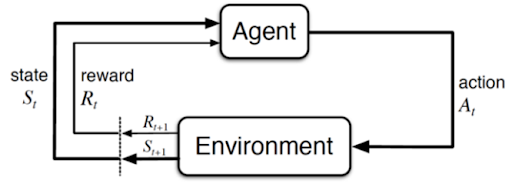


Figure 1. Reinforcement learning cycle

*Markov Decision Process*

Reinforcement learning is about solving Markov Decision Processes (MDPs). An MDP is a formal way to describe a game in terms of states, actions, and rewards. It defines a mechanism of how certain states & agent’s actions lead to other states. Playing Breakout can be formalized as a finite state MDP.

The state is the current situation the agent is in, which can be approximated in the Atari game using the current frame. While a single frame may be useful, it becomes difficult to determine the motion of the ball as going up or going down. Utilizing a single frame may end up violating the Markov property of the MDP, history does not matter. In order to satisfy the Markov property, the previous states must not have any useful information. In the case of Breakout, using a single frame will break the Markov property as the previous frames may help the agent to learn about the speed of objects and infer its acceleration.

The action is a command given to reach a certain state in the game or attain a reward. Using the joystick, one can perform actions like doing nothing, asking for the ball, going left, and going right in the Breakout game.

Last component of the MDPs are the rewards the agent receives upon completion of an action. Rewards are calculated by considering the starting state, action performed, and the end state. The goal of the reinforcement learning program is to learn the optimal behavior in an environment to maximize long term rewards. In the case of Atari 2600, rewards correspond to changes in score as shown in Figure 2.

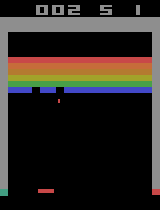
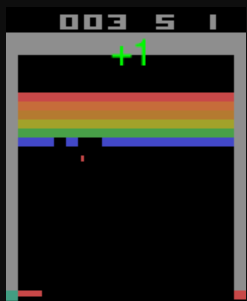
 

Figure 2. Score increases from 2 to 3, a reward of +1 is attained