Playing DOOM with Deep Reinforcement Learning

Yueqiu Sun, Ruofan Wang
New York University, Center for Data Science

Project Overview

In the project, we explored using deep reinforcement learning to play the game DOOM. We implemented Policy Gradient and Deep Q Learning for the “search and hit” scenario and “health gathering” scenario in the game.
In addition, we evaluate different exploration strategies in the “search and hit” scenario.

Exploration Policy

Exploration is vital in training agents as it enables the agent to gather information from the environment to make better decision.

- Random-Policy: Always select a action from a uniform distribution
- ε–Greedy: Agent execute a random action with possibility π, and execute action with the maximum Q value with possibility 1-π

$$a_t = \begin{cases} a_t^* & \text{with probability } 1-\epsilon \\ \text{random action} & \text{with probability } \epsilon \end{cases}$$

- Boltzmann-Policy: Boltzmann uses the value of Q function to assign different probabilities to each action using a softmax distribution. Temperature β control the randomness of the exploration. When $\beta \to \infty$ it becomes random-policy When $\beta \to 0$ it becomes greedy policy.

$$v_\beta(x) = \exp\left(\frac{x}{\beta}\right)$$

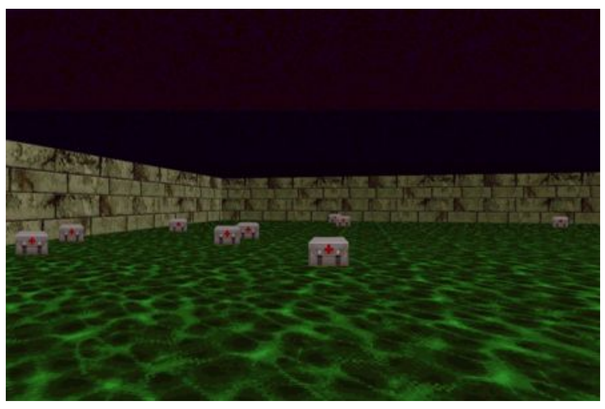
$$\pi_{\text{boltz}}(a|s, \mathcal{Q}, \beta) = \frac{v_\beta(\mathcal{Q}(s,a))}{\sum_{a' \in \mathcal{A}(s)} v_\beta\left(\mathcal{Q}(s,a')\right)}$$

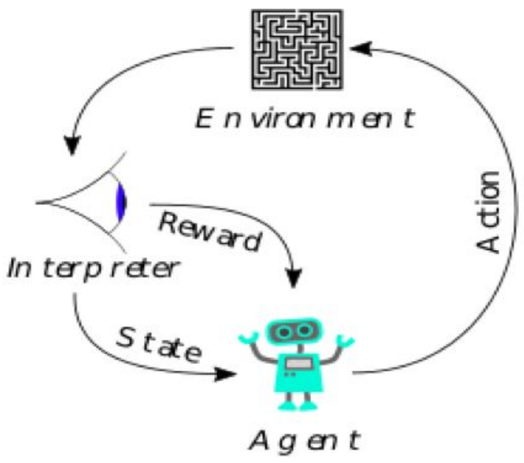
DOOM-Health Gathering

In this scenario, the player can pick up some certain items called 'medkits'. These items can heal some portions of the player's health.

In this setting, the environment is modelled as a Markov Decision Process and the state is not fully observed.

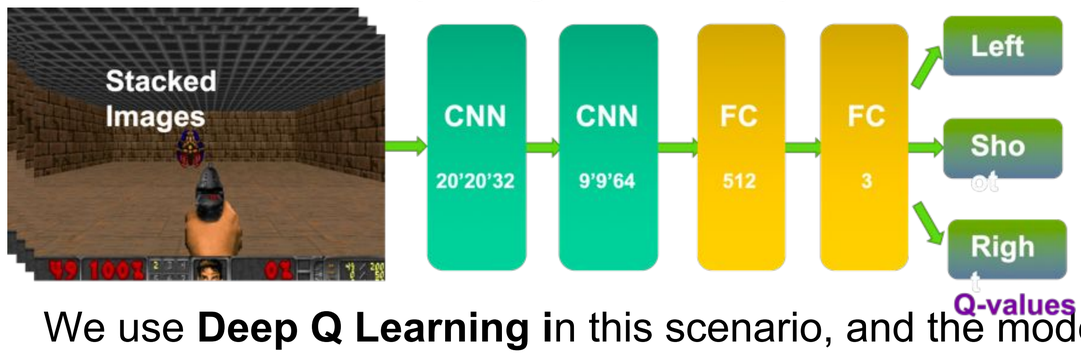
We used policy gradient method to approach this. Policy Gradient is an action-based RL method, where the agent will learn an actor to make actions. The objective function is the expected total rewards and we use gradient ascent the optimize it.





DOOM-Search and Hit

In this scenario, the player aims at hitting the target. The action space contains {left, right and shoot}.

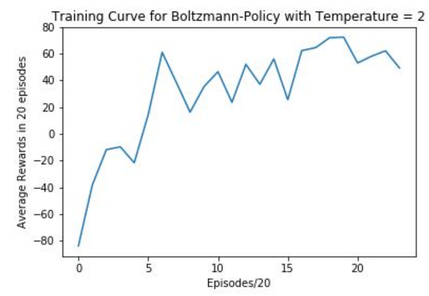


We use **Deep Q Learning i**n this scenario, and the model architecture is shown as above. In contrast to policy gradient, Q learning is a value-based approach. The machine learns a critic and the agents makes a decision using the calculated Q value.

Exploration and exploitation balances between making the best conditional decision and gathering more information via exploring. We evaluate different exploration policies introduced in the exploration policy part using “search and hit” scenario in experiment. The results are shown below

Experiment Result

Training curve for Random-Policy and Boltzmann-Policy with temperature = 2



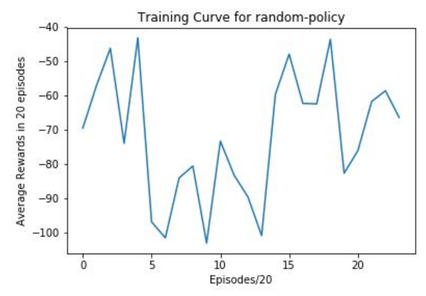


Table 1: Test average rewards for different exploration scheme

Model	Configuration	Test Average Rewards
Baseline	None	-75.18
Human	None	78.21
Random-Policy	None	55.11
ϵ-Greedy-Policy	decay rate = 1e-4	33.76
ϵ-Greedy-Policy	decay rate = 1e-5	70.83
Boltzmann-Policy	temperature $\beta = 0.5$	56.7
Boltzmann-Policy	temperature $\beta = 1$	66.20
Boltzmann-Policy	temperature $\beta = 2$	75.42

The Boltzmann-Policy with temperature = 2 achieves the best result among our models. It suggests that randomness in the exploration policy help the test results.

To our surprise, the random exploration policy outperforms the epsilon-greedy-policy with decay rate = 0.0001. This also shows that exploration without the influence of the Q value is valuable in training agents.