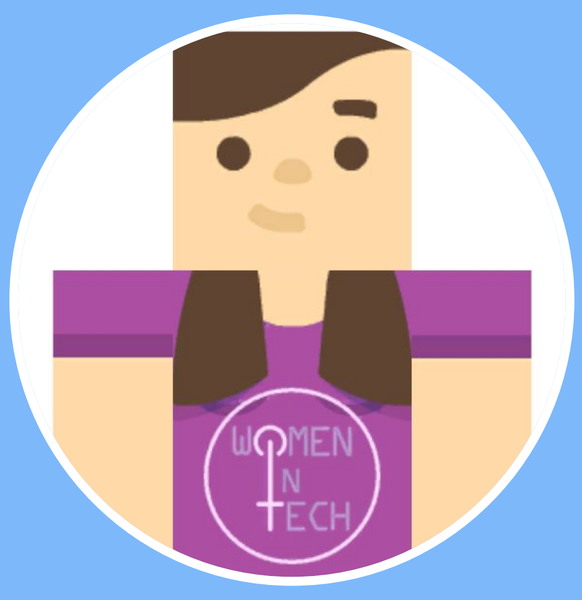
Evaluating Observations vs. Training Duration in Reinforcement Learning

Lauren DeMaio | Muskaan Moinuddin

Dr. Carmine Guida
Seidenberg School of CSIS, Pace University



ABSTRACT

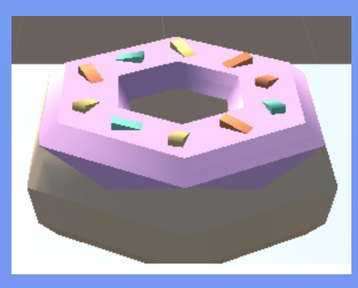
Reinforcement learning refers to a reward based machine learning training method. An AI *Agent* is provided with observations about its environment and in return it makes actions. The agent can be rewarded positively or negatively based on the results of its actions and completing a goal. Large amounts of observations may help to decrease training time, however, will increase the size of the model. We evaluated different levels of observations and compared it with the amount of episodes needed for our agent to successfully complete its goal 100 times in a row.

RESEARCH QUESTIONS

- What is the relationship between increased task complexity (more jumps required) and training time?
- What is the minimal and most effective information needed to train the agent?

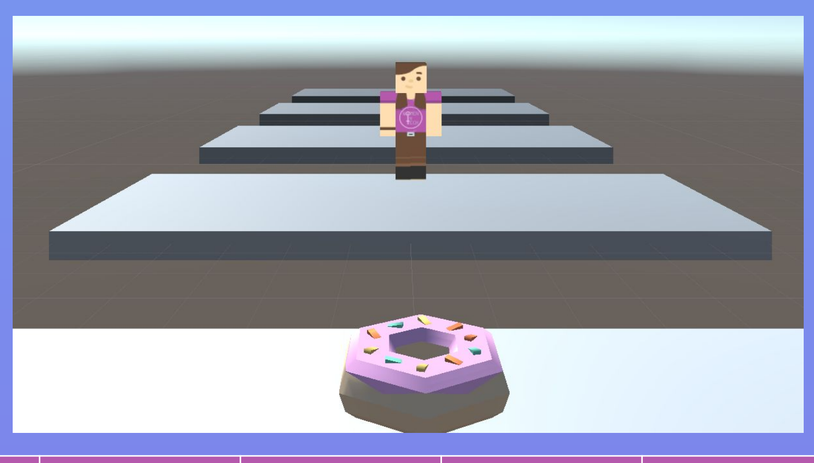
REWARDS

Our reward system incentivised the agent for not only completing the task of reaching the goal (+12.0), but for decreasing the distance between itself and the goal (+0.02) and, likewise, the agent was penalized for increasing that distance (-0.05). Each platform held the same reward amount (+4.0); the agent was only able to retrieve this reward once for each platform. If the agent fell below the platforms or jumped too high it would receive a negative reward for its failure (-20.0).



DESIGN AND IMPLEMENTATION

We performed multiple experiments involving a complex obstacle course designed to test the capabilities of our reinforcement learning agent, named Laurskaan. The course consists of a starting platform and a goal platform, which contains the reward. To investigate training duration related to observations, we gradually added between 0 to 4 intermediate platforms with gaps in between. Our initial test used fixed static platforms. The table below shows the observations provided to the agent and the number of intermediate platforms. The values indicate the average number of episodes required for the agent to complete the task successfully 100 times in a row from 5 trials.



Observations Provided / No. of Intermediate Platforms	0	1	2	3	4
Agent Pos. & Vel.	3303	6806	10108	10510	9449
Agent Pos. & Vel. and Po. of Next Platform	3577	3226	5185	4329	6258
Agent Pos. & Vel., All Platform Pos., Goal Pos.	3875	3191	3701	4031	3974
Agent Pos. & Vel., All Platform Pos., Goal Pos., and IsGrounded	3713	3071	3616	3833	3969
Agent Pos. & Vel., All Platform Pos., Goal Pos., IsGrounded, and Next Platform	3493	3209	3894	4049	3788

Aside from a fixed environment, we experimented with a varying shift of the platforms between each episode. For this, the platforms were randomly placed at heights between -1.0 and 1.0, always leaving the starting and goal platforms fixed. For this experiment the agent was completely observant of the environment.

Varying Platforms / No. of Intermediate Platforms	0	1	2	3	4
Agent Pos. & Vel., All Platform Pos., Goal Pos., IsGrounded, and Next Platform	5059	6594	7736	13252	17040

RESULTS

In the end, our agent was successful in reaching the goal 100 times in a row for every variation of platforms in each observation test. We notice that when the agent is given more information about its environment it learns in fewer episodes. The test that took the fewest number of episodes on average was the fourth test (agent position & velocity, all platform positions, goal position, and an isGrounded observation). Note that fewer episodes does not necessarily mean less training time, since the model takes in more information with additional sensors it can often have a longer training duration. We discovered that increasing the number of fixed platforms does not dramatically increase the number of episodes it takes to reach 100 in a row; however, when we added the extra layer of varying the platforms between -1.0 to 1.0, the addition of platforms substantially increased the number of episodes.

DISCUSSION

Our experiment evaluates on testing complete vs. partial observations regarding the state of the agent's environment. We chose to provide our agent with a discrete set of behaviors in order to limit the variability of the model and speed up training time. The discoveries here suggest that the closer we get to our agent being completely observant, the fewer episodes it takes to train. Aside from test one (agent position & velocity), the difference in averages is slim, the rest of the test averages vary by no more than 1000 episodes. For each level of observation provided, it remained consistent that one intermediate platform took the least number of episodes to train. From here we can approach the question of what observations we are willing to sacrifice in order to maintain a quick training duration.

FUTURE WORK

Our future trajectory involves crafting a dynamically evolving environment to enhance the agent's capabilities. Additionally, our experiments used coordinates of objects in the environment and we aim to explore novel sensor types, expanding the agent's observation capabilities for improved adaptability in a real-time environment.