Reinforcement Learning - A Browser Based Visualisation Tool

Kevin Gleeson

B.Sc.(Hons) in Software Development

March 27, 2019

Final Year Project

Advised by: Mr Martin Hynes
Department of Computer Science and Applied Physics
Galway-Mayo Institute of Technology (GMIT)

Contents

About this project

Abstract This project will help to explain the temporal difference reinforcement learning process by displaying an agents behaviour, performance and Q-Table as it interacts within its environment. The application is a browser based visual tool where a user can interact by tweaking parameters within a form. Once the form is submitted it will then make a request to run a main python script held on a flask server. Once the script has completed the user will be presented with and animation of the agent moving through it's environment. In addition a graph of the agent performance and the q-table will presented to the user for examination. This will aid the user in better understanding the idea of reinforcement learning.

Authors Kevin Gleeson 4th year student studying Software Development at GMIT Galway.

Chapter 1

Introduction

Reinforcement Learning is an unsupervised machine-learning technique that allows an agent to explore and learn from its environment without any prior knowledge of the domain. As the agent moves through the environment it gains knowledge via reward signals gathered by transitioning from state to another based on the action taken from its current state. With each step the agent is only concerned with its current state and what rewards it can gain from transitioning to it's next state.
The agent chooses it's action decision based on what highest reward it can get from the next available states.
The purpose of this application is to demonstrate and explain reinforcement learning through a browser based visualisation tool.
The application will have the following elements on the Browser:

- The agent moving within its environment when the simulation is run. This will be displayed using HTML 5 canvas.
- User input to tweak parameters before each run of the simulation. The parameters that will be available to the user are:
  - The end goal reward
  - The negative trap reward
  - The agent learning rate
  - The learning decay rate
  - The discount factor
  - The Exploration rate
  - The Exploration decay rate
  - The per step reward

- The maximum number of episodes to be run
- The maximum number of agent steps per episode
- Choice of algorithm

What is reinforcement learning?

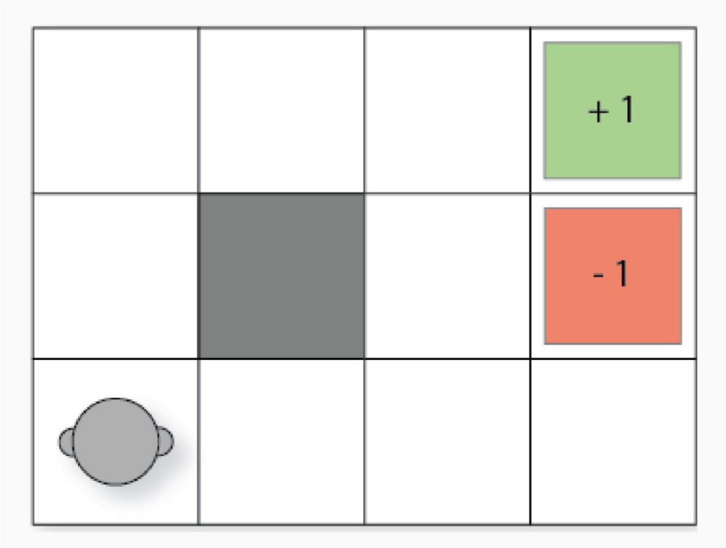

Figure 1.1

Reinforcement learning is the process of rewarding an agent for a decision made within its environment. The reward can be either positive or negative based on the decisions made by the agent as it transitions from one state to another.
For example, if a puppy has no knowledge of the sit command it will not perform the desired action on the first attempt. Each time the puppy sits when commanded its decision is reinforced with a positive treat/reward. If the puppy does not sit the reward is negative (no treat). Eventually after many iterations of training, the dog will associate a treat/reward with that specific command and eventually learn that sitting will get them a treat. The puppy in essence is taking actions to maximise rewards while exploring an unknown environment.
With reinforcement machine learning, this technique is used to train an agent to learn about its environment through trial and error. The environment used for this project will be grid world. The grid world domain is a two dimensional grid with the agent starting at the bottom left of the grid, the

goal state is at the top right of the grid in addition there are traps that the agent needs to avoid while travelling from the start state to it's goal state.

- The agent's actions effect the environment by moving around and exploring.
- The state is what the agent can observe at a given time. In the grid above, the agent can occupy eleven possible squares. We can number theses states from 1 – 11 moving from left to right with the bottom left square being state number 1.
- In the agents initial state (State 1) it knows nothing about its environment and chooses an action of moving left, right, up or down.
- The Epsilon variable sets the probability of choosing a random action. When set to one it will always choose a random action. If set to .8 it will choose the a random action 80% of the time.

This will give the agent a chance to explore the environment depending on what the value is set to.

- Q values are a weighted score attached to an action of a particular state.
- There is a negative reward cost for each move the agent makes in this case -0.04. This will help in getting the best path to the end state.
- The learning rate (alpha) is a value between zero and one determines how much the Q value is updated for each action taken. It will be .5 for this example.
- The discount factor (Gamma) set to .9 is the immediate reward gained for an action taken. The higher the value the more the agent will take the immediate reward.
- The reward cost, gamma and alpha are hyper-parameters chosen by the user.
- There is a formula to follow to update the Q values of each action taken: Q (current state, action) += alpha *[reward + gamma* max value of Q (next state, all possible actions) – Q (current state, action)]
- The Q table is a record of all of the agent's actions taken in a given state. This is the agent's memory and is set to zero when first run. If all Q values are equal, it will Choose one at random.

State	Action left	Action right	Action up	Action down
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0
11	0	0	0	0

If the agent decides to move up one square to state 5 the Q table is updated using the formula above which looks like .5 * -.04 + .9 *0 − 0 = -0.02

State	Action left	Action right	Action up	Action down
1	0	0	-0.02	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0
11	0	0	0	0

If the Agent decides to move down to state one again the value of moving down from state 5 to state 1 is updated to -0.02 also.

State	Action left	Action right	Action up	Action down
1	0	0	-0.02	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	-0.02
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0

Then when back in state one the agent's best choice (highest value) is down, left or right as they are all 0 and higher the -0.02. Eventually all of the actions of a given state will have a value added. The agent will chose the highest value as the optimal path to take to the end goal. Once the agent gets to either end state, the episode is terminated and re-run. When episodes are re-run, the Q-Table will continually update until the optimal path is found and minimal updates will be performed.

references [1, 2, 3, 4, 5, 6, 7, 8]

Chapter 2

Context

- Provide a context for your project.
- Set out the objectives of the project
- Briefly list each chapter / section and provide a 1-2 line description of what each section contains.
- List the resource URL (GitHub address) for the project and provide a brief list of the main elements at the URL.

2.1 Filler

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam mi enim, interdum ut elit lobortis, bibendum tempus diam. Etiam turpis ex, viverra tristique finibus nec, feugiat at metus. Curabitur tempus gravida interdum. Donec ac felis a lorem scelerisque elementum. Vestibulum sit amet gravida tortor, a iaculis orci. Nam a molestie augue. Curabitur malesuada odio at mattis molestie. In hac habitasse platea dictumst. Donec eu lectus eget risus hendrerit euismod nec at orci. Praesent porttitor aliquam diam, eu vestibulum nisl sollicitudin vel. Nullam sed egestas mi.

Quisque vel erat a justo volutpat auctor a nec odio. Sed rhoncus augue sit amet nisl tincidunt, vitae cursus tellus efficitur. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque et auctor dui. Fusce ornare odio ipsum, et laoreet mi molestie sed. Cras at massa sit amet ipsum gravida aliquam. Nulla suscipit porta imperdiet. Fusce eros neque, bibendum sit amet consequat non, pulvinar quis ipsum.

2.1.1 More filler

Donec fermentum sapien ac rhoncus egestas. Nullam condimentum condimentum eros sit amet semper. Nam maximus condimentum ligula. Praesent faucibus in nisi vitae tempus. Sed pellentesque eleifend ante, ac malesuada nibh dapibus nec. Phasellus nisi erat, pulvinar vel sagittis sed, auctor et magna. Quisque finibus augue elit, consequat dignissim purus mollis nec. Duis ultricies euismod tortor, nec sodales libero pellentesque et. Interdum et malesuada fames ac ante ipsum primis in faucibus.

Donec id interdum felis, in semper lacus. Mauris volutpat justo at ex dignissim, sit amet viverra massa pellentesque. Suspendisse potenti. Praesent sit amet ipsum non nibh eleifend pretium. In pretium sapien quam, nec pretium leo consequat nec. Pellentesque non dui lacus. Aenean sed massa lacinia, vehicula ante et, sagittis leo. Sed nec nisl ac tellus scelerisque consequat. Ut arcu metus, eleifend rhoncus sapien sed, consequat tincidunt erat. Cras ut vulputate ipsum.

Curabitur et efficitur augue. Proin condimentum ultrices facilisis. Mauris nisi ante, ultrices sed libero eget, ultrices malesuada augue. Morbi libero magna, faucibus in nunc vitae, ultricies efficitur nisl. Donec eleifend elementum massa, sed eleifend velit aliquet gravida. In ac mattis est, quis sodales neque. Etiam finibus quis tortor eu consequat. Nullam condimentum est eget pulvinar ultricies. Suspendisse ut maximus quam, sed rhoncus urna.

2.2 Filler

Phasellus eu tellus tristique nulla porttitor convallis. Vestibulum ac est eget diam mollis consectetur. Donec egestas facilisis consectetur. Donec magna orci, dignissim vel sem quis, efficitur condimentum felis. Donec mollis leo a nulla imperdiet, in bibendum augue varius. Quisque molestie massa enim, vitae ornare lacus imperdiet non. Donec et ipsum id ante imperdiet mollis. Nullam est est, euismod sit amet cursus a, feugiat a lectus. Integer sed mauris dolor.

Mauris blandit neque tortor, consequat aliquam nisi aliquam vitae. Integer urna dolor, fermentum ut iaculis ut, semper eu lacus. Curabitur mollis at lectus at venenatis. Donec fringilla diam ac risus imperdiet suscipit. Aliquam convallis quam vitae turpis interdum, quis pharetra lacus tincidunt. Nam dictum maximus lectus, vitae faucibus ante. Morbi accumsan velit nec massa tincidunt porttitor. Nullam gravida at justo id viverra. Mauris ante nulla, eleifend vitae sem vitae, porttitor lobortis eros.

Cras tincidunt elit id nisi aliquam, id convallis ex bibendum. Sed vel odio fringilla, congue leo quis, aliquam metus. Nunc tempor vehicula lorem eu ultrices. Curabitur at libero luctus, gravida lectus sed, viverra mi. Cras ultrices aliquet elementum. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Sed metus ante, suscipit sit amet finibus ut, gravida et orci. Nunc est odio, luctus quis diam in, porta molestie magna. Interdum et malesuada fames ac ante ipsum primis in faucibus. Mauris pulvinar lacus odio, luctus tincidunt magna auctor ut. Ut fermentum nisl rhoncus, tempus nulla eget, faucibus tortor. Suspendisse eu ex nec nunc mollis pulvinar. Nunc luctus tempus tellus eleifend porta. Nulla scelerisque porttitor turpis porttitor mollis.

Duis elementum efficitur auctor. Nam nisi nulla, fermentum sed arcu vel, posuere semper dui. Fusce ac imperdiet felis. Aenean quis vestibulum nisl. Integer sit amet tristique neque, at suscipit tortor. Morbi et placerat ante, vel molestie dui. Vivamus in nibh eget massa facilisis accumsan. Nunc et purus ac urna fermentum ultrices eget sit amet justo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Cras elementum dui nunc, ac tempor odio semper et. Ut est ipsum, sollicitudin eleifend nisl eu, scelerisque cursus nunc. Nam at lectus vulputate, volutpat tellus vel, pharetra mauris. Integer at aliquam massa, at iaculis sem. Morbi nec imperdiet odio. In hac habitasse platea dictumst.

Mauris a neque lobortis, venenatis erat ut, eleifend quam. Nullam tincidunt tellus quis ligula bibendum, a malesuada erat gravida. Phasellus eget tellus non risus tincidunt sagittis condimentum quis enim. Donec feugiat sapien sit amet tincidunt fringilla. Vivamus in urna accumsan, vehicula sem in, sodales mauris. Aenean odio eros, tristique non varius id, tincidunt et neque. Maecenas tempor, ipsum et sollicitudin rhoncus, nibh eros tempus dolor, vitae dictum justo massa in eros. Proin nec lorem urna. In ullamcorper vitae felis sit amet tincidunt. Maecenas consectetur iaculis est, eu finibus mi scelerisque et. Nulla id ex varius, ultrices eros nec, luctus est. Aenean ac ex eget dui pretium mattis. Ut vitae nunc lectus. Proin suscipit risus eget ligula sollicitudin vulputate et id lectus.

Chapter 3

Methodology

About one to two pages. Describe the way you went about your project:

- Agile / incremental and iterative approach to development. Planning, meetings.
- What about validation and testing? Junit or some other framework.
- If team based, did you use GitHub during the development process.
- Selection criteria for algorithms, languages, platforms and technologies.

Check out the nice graphs in Figure 3.2, and the nice diagram in Figure **??**.

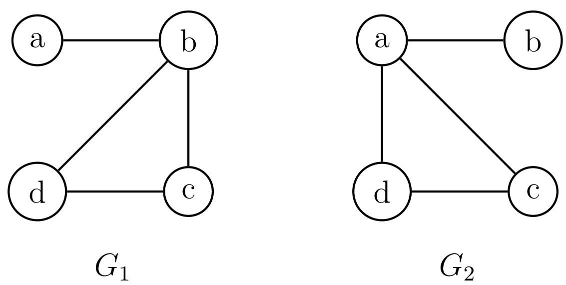

Figure 3.1: Nice pictures

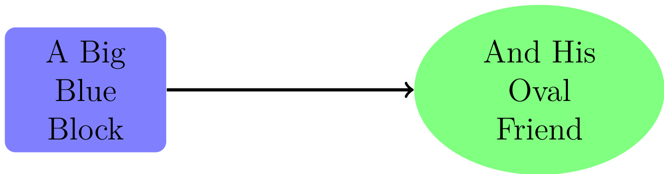

Figure 3.2: Nice pictures

Chapter 4

Technology Review

About seven to ten pages.

- Describe each of the technologies you used at a conceptual level. Standards, Database Model (e.g. MongoDB, CouchDB), XMl, WSDL, JSON, JAXP.
- Use references (IEEE format, e.g. [1]), Books, Papers, URLs (timestamp) – sources should be authoritative.

4.1 XML

Here's some nicely formatted XML:

```xml
<this>
  <looks lookswhat="good">
    Good
  </looks>
</this>
```

Chapter 5

System Design

As many pages as needed.

- Architecture, UML etc. An overview of the different components of the system. Diagrams etc... Screen shots etc.

Column 1	Column 2
Rows 2.1	Row 2.2

Table 5.1: A table.

Chapter 6

System Evaluation

As many pages as needed.

- Prove that your software is robust. How? Testing etc.
- Use performance benchmarks (space and time) if algorithmic.
- Measure the outcomes / outputs of your system / software against the objectives from the Introduction.
- Highlight any limitations or opportuni-ties in your approach or technologies used.

Chapter 7

Conclusion

About three pages.

- Briefly summarise your context and ob-jectives (a few lines).
- Highlight your findings from the evalua-tion section / chapter and any opportuni-ties identified.

Bibliography

[1] A. Einstein, "Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]," *Annalen der Physik*, vol. 322, no. 10, pp. 891–921, 1905.

[2] D. Knuth, "Knuth: Computers and typesetting."

[3] M. Goossens, F. Mittelbach, and A. Samarin, *The LaTeX Companion*. Reading, Massachusetts: Addison-Wesley, 1993.

[4] "Robotics: Ethics of artificial intelligence.," *Nature*, vol. 521, no. 7553, pp. 415 – 418, 2015.

[5] S. Underwood, "Potential and peril: The outlook for artificial intelligence-based autonomous weapons.," *Communications of the ACM*, vol. 60, no. 6, pp. 17 – 19, 2017.

[6] O. DAVIES, "Forecasting the impact of artificial intelligence: An introduction.," *Foresight: The International Journal of Applied Forecasting*, no. 47, pp. 4 – 6, 2017.

[7] P. LEMIEUX, "Rise of the machines?.," *Regulation*, vol. 40, no. 4, pp. 4 – 5, 2017.

[8] M. Ford, "Viewpoint: Could artificial intelligence create an unemployment crisis?.," *Communications of the ACM*, vol. 56, no. 7, pp. 37 – 39, 2013.