

COOPERATIVE CONTROL OF AUTONOMOUS VEHICLE ON HIGHWAYS

by

PENG XU

Submitted in partial fulfillment of the requirements

For the degree of Master of Science

Department of Electrical Engineering and Computer Science

CASE WESTERN RESERVE UNIVERSITY

May, 2018

Cooperative Control of Autonomous Vehicle on Highways

Case Western Reserve University
Case School of Graduate Studies

We hereby approve the thesis¹ of

PENG XU

for the degree of

Master of Science

Dr. Wyatt Newman

Committee Chair, Adviser
Department of Electrical Engineering and Computer Science

Date

Dr. M. Cenk Cavusoglu

Committee Member
Department of Electrical Engineering and Computer Science

Date

¹We certify that written approval has been obtained for any proprietary material contained therein.

*Dedicated to *your
dedication message goes here**

Table of Contents

List of Tables	v
List of Figures	vi
Acknowledgements	vii
Acknowledgements	vii
Abstract	viii
Abstract	viii
Chapter 1. Introduction	1
Lifetime and Degradation Science: Applicability to Renewable energy	1
Chapter 2. Chapter Title	4
Analysis	4
Chapter 3. Experimental 1	11
Section header	11
Methodology	11
Chapter 4. Experimental 2	16
Specimen Selection & Characterization	16
Chapter 5. Discussion	18
Chapter 6. Conclusions	19
Chapter 7. Suggested Future Research	22
Continuation of Exposures and Analysis	22
	iv

Appendix A. Pprofile	24
Appendix B. Specimens	25
Appendix C. Preparation of this document	26
Appendix D. Figures	27

List of Tables

List of Figures

1.1	How an agent interacts with the environment.	2
2.1	A browser screenshot of Coaster Racer Game.	5
2.2	A flowchart of training the Coaster Racer Game Bot.	6
2.3	Deep Neural Network model from DeepMind paper.	7
2.4	LeNet-5.	8
3.1	Cropped input image.	12
3.2	Downsized input image.	12
3.3	Grayscaled input image.	13
4.1	High score by the Game Bot.	16
4.2	A comparison between DQN trained Game Bot and a random Game Bot.	17
7.1	A bigger action space.	23

Acknowledgements

0.1 Acknowledgements

Abstract

Cooperative Control of Autonomous Vehicle on Highways

Abstract

by

PENG XU

0.2 Abstract

1 Introduction

1.1 Lifetime and Degradation Science: Applicability to Renewable energy

1.1.1 Project Overview

Reinforcement Learning is a type of machine learning that allows you to create AI agents that learn from the environment by interacting with it. Just like how we learn to ride a bicycle, this kind of AI learns by trial and error. As seen in Figure 1, the brain represents the AI agent, which acts on the environment. After each action, the agent receives the feedback. The feedback consists of the reward and next state of the environment. The reward is usually defined by a human. If we use the analogy of the bicycle, we can define reward as the distance from the original starting point.

Google's DeepMind published its famous paper Playing Atari with Deep Reinforcement Learning, in which they introduced a new algorithm called Deep Q Network (DQN for short) in 2013. It demonstrated how an AI agent can learn to play games by just observing the screen without any prior information about those games. The result turned out to be pretty impressive. This paper opened the era of what is called ?deep reinforcement learning?, a mix of deep learning and reinforcement learning.

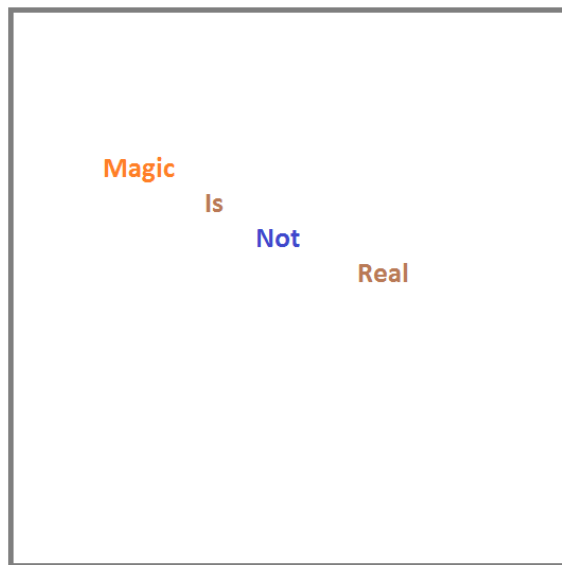


Figure 1.1. How an agent interacts with the environment.

1.1.2 Problem Statement

In this project, a deep reinforcement learning method, Deep Q Network, would be implemented and applied to play a Coast Racer game in OpenGym / Universe using TensorFlow.

Q-learning, a model-free online off-policy algorithm, whose main strength is that it is able to compare the expected utility of the available actions without requiring a model of the environment. In Q-Learning Algorithm, there is a function called Q Function, which is used to approximate the reward based on a state. We call it $Q(s,a)$, where Q is a function which calculates the expected future value from state s and action a. Similarly in Deep Q Network algorithm, we use a neural network to approximate the reward based on the state. We will discuss how this works in detail.

Target function:

$Q^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, where \mathcal{S} is the set of states and \mathcal{A} is the set of actions (turn left, forward or turn right), and \mathbb{R} represents the value of being in a state $s \in \mathcal{S}$, applying a action $a \in \mathcal{A}$, and following policy π thereafter.

Target function representation:

Deep neural network.

Therefore, I seek to build a Q-learning agent trained via a deep convolutional neural network to approximate the optimal action-value function:

$$Q^*(s, a) = \max_{\pi} Q^\pi(s, a), \forall s \in \mathcal{S}, a \in \mathcal{A} \quad (1.1)$$

which is the maximum sum of rewards achievable by a behavior policy π .

1.1.3 Metrics

Since Q learning is recursively defined, we did not consider fitting errors such as RMSE to be good metrics for evaluation. Instead, given our ultimate our goal to make a good game AI, we evaluated gameplay directly in order to gauge performance. We evaluated the performance of our algorithms using two metrics: game length and game score. Game length is the number of moves the player makes before they reach a game over, and game score is the score based on the number of lines they clear. While the game score is ultimately more important, it can only be used to judge the performance of sufficiently advanced networks because there is a large initial barrier to overcome before the algorithm can score with any degree of consistency. Thus, game length, which was typically correlated with game score, is a good metric for showing intermediate performance since an algorithm can incrementally learn to survive longer.

2 Chapter Title

2.1 Analysis

To tackle the problem described in Section ??, we will use Reinforcement learning with Deep Learning to automatically learn evaluation functions by playing games by itself. Unlike other approaches that need a very large dataset, this approach will try to learn to play games without any domain knowledge (no dataset will be used). This is a promising approach for creating game-playing algorithms for playing other two-player games of perfect information.

2.1.1 Data Exploration and Visualization

Coaster Racer Game environment. The browser screenshot as shown in Figure 2 is both for the user and the AI Game Bot. Namely, that is the raw vision of the Game Bot which apparently contains lots of pixels useless for deciding the steering. For more efficient and faster computing, some preprocessing procedures would be done before we throw the image data into a deep learning model, which will be described in later paragraph.

States. For the purposes of winning the game, the state of the game at time t is simply the pixel array of the game screen at that particular time. I experimented with using only a cropped version of the game screen that contained the view of the driver. In order to avoid unnecessary computation and memory usage, we downsample this image. By

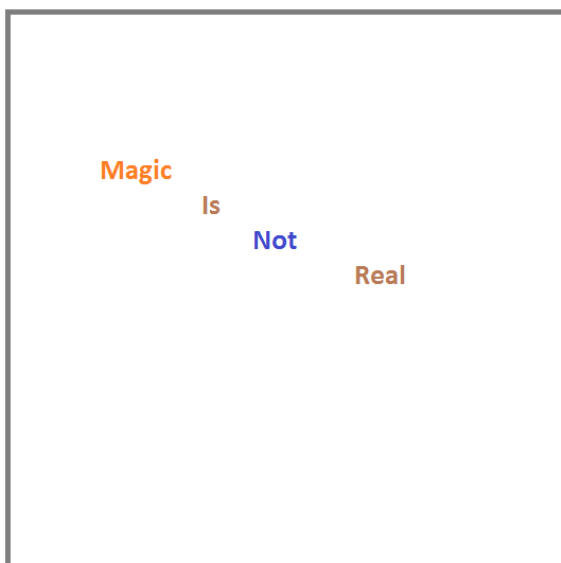


Figure 2.1. A browser screenshot of Coaster Racer Game.

reducing the size of each state, this allows us to fit more training examples in our replay history without running out of memory, which is a crucial component of the experience replay technique we used. In addition, with lower-dimensional input, our network requires fewer matrix multiplications and fewer parameters in the fully connected layers, greatly improving speed.

- State Space: Preprocessed screenshots of the browser loading the game

Actions. Coaster Racer Flash Game is a typical car racing game which is simply controlled by moving forward, left turn and right turn. With this simple setting, we only have an action space of three actions, namely, moving forward, left turn and right turn.

- Action Space: Move Forward action, Turn Left action, Turn Right action

Rewards. Rewards are primarily based off of the game score. From the perspective of human player, we judge a good player or not by how the driving behavior is close to real and safe case. While for a Game Bot, it only cares about the score. For this specific

game, Coaster Racer Game, it unreasonably encourages a dangerous driving behavior. For example, when the car hits a traffic cone it gets 1000 reward, even if the price of hitting a cone is falling down the road. There is no punishment for falling down the road.

An ideal reward function would encourage the car to stay in the middle of the road and to try not to hitting other cars and guard bars. Here I simply cancel the one time 1000 score reward. But there is still room to fine tune the function in the future.

2.1.2 Algorithms and Techniques

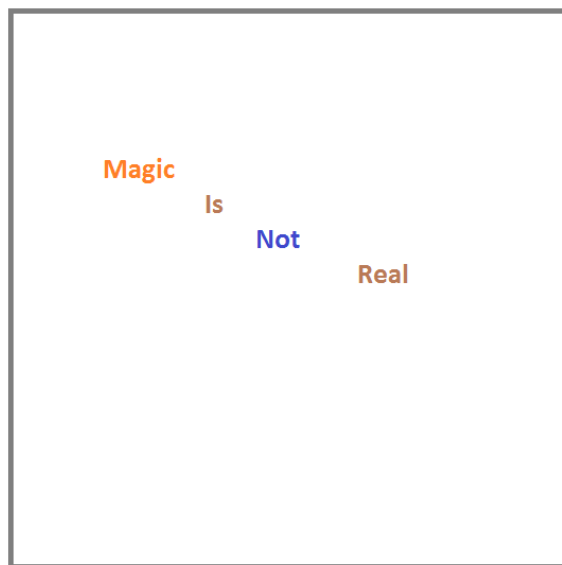


Figure 2.2. A flowchart of training the Coaster Racer Game Bot.

Overall Representation. An overall representation of how the different components relate during a play evaluation, centered around the deep Q-network for playing, the main decision component is shown in Figure 3. Each game screen is resized to a desaturated pixels image, and if you might be wondering why each state is a sequence of four game

screens instead of one, that is because the agent's history is used for better motion perception. Achieving this requires a sequence of preprocessed images to be stacked in channels (like you would stack RGB channels on a colored image) and fed to the network.

Deep Neural Network. Convolutional Neural Networks, or CNNs, are a special type of neural network that has a known grid-like topology. Like most other neural networks they are trained with a variant of the back-propagation algorithm. CNN's strength is pattern recognition directly from pixels of images with minimal processing. We use a convolutional network as a function mapping the preprocessed images to Q values, since the actions are highly based on what would be seen as pixel matrix.

- **DeepMind's Deep Q Network**

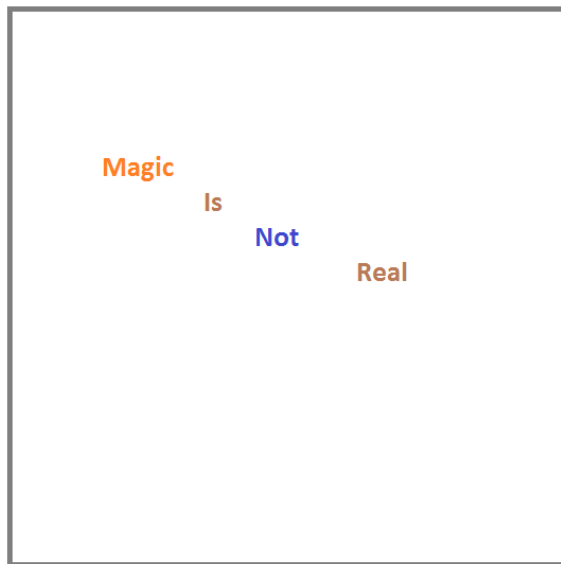


Figure 2.3. Deep Neural Network model from DeepMind paper.

The network's architecture that I firstly used is essentially the same used by DeepMind, except for the first convolutional neural network's input (80x80x4

instead of $84 \times 84 \times 4$, to account for the different input sizes) and the linear layer's output (3 instead of 18, to account for the different number of actions available).

- **LeNet-5.**

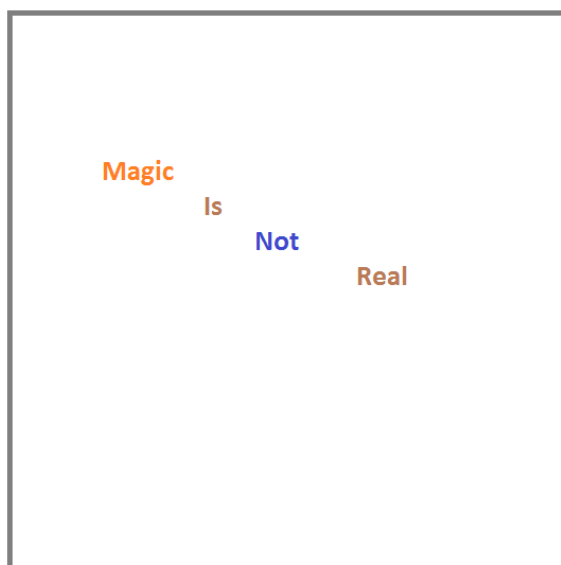


Figure 2.4. LeNet-5.

For an image classification problem, there are several mature models other than the one above such as LeNet, VGG, Inception and Xception. I then tried a LeNet-5 model here and expected it to perform better than the DeepMind one for predicting actions. LeNet is deeper than the DeepMind one to grasp more features and has some dropout layers to avoid overfitting. In practice, it shows an advantage especially on predicting turning actions.

Q-learning. One of the most basic and popular methods to estimate action-value functions is the Q-learning algorithm. It is model-free online off-policy algorithm, whose main strength is that it is able to compare the expected utility of the available actions

without requiring a model of the environment. Q-learning works by learning an action-value function that gives the expected utility of taking a given action in a given state and following a fixed policy thereafter.

A value function estimates what is good for an agent over the long run. It estimates the expected outcome from any given state, by summarizing the total amount of reward that an agent can expect to accumulate into a single number. Value functions are defined for particular policies.

The state value function (or V-function), is the expected return when starting in state s and following policy π thereafter[?],

$$V^\pi(s) = \mathbb{E}_\pi [R_t | s_t = s] \quad (2.1)$$

The action value function (or Q-function), is the expected return after selecting action a in state s and then following policy π ,

$$Q^\pi(s, a) = \mathbb{E}_\pi [R_t | s_t = s, a_t = a] \quad (2.2)$$

The optimal value function is the unique value function that maximises the value of every state, or state-action pair,

$$Q^*(s, a) = \max_{\pi} Q^\pi(s, a), \forall s \in \mathcal{S}, a \in \mathcal{A} \quad (2.3)$$

An optimal policy $\pi^*(s, a)$ is a policy that maximises the action value function from every state in the MDP,

$$\pi^*(s, a) = \operatorname{argmax}_{\pi} Q^\pi(s, a) \quad (2.4)$$

The update rule uses action-values and a built-in max-operator over the action-values of the next state in order to update $Q(s_t, a_t)$ as follows,

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right] \quad (2.5)$$

The agent makes a step in the environment from state s_t to s_{t+1} using action a_t while receiving reward r_t . The update takes place on the action-value a_t in the state s_t from which this action was executed. This version of Q-learning works well for tasks with a small a state-space, since it uses arrays or tables with one entry for each state-action pair.

In this project the policy is using the ϵ -**greedy** policy:

- **ϵ -greedy.** Selects the best action for a proportion $1 - \epsilon$ of the trials, and another action is randomly selected (with uniform probability) for a proportion,

$$\pi_\epsilon(s) = \begin{cases} \pi_{\text{rand}}(s, a) & \text{if } \text{rand}() < \epsilon \\ \pi_{\text{greedy}}(s, a) & \text{otherwise} \end{cases} \quad (2.6)$$

where $\epsilon \in [0, 1]$ and $\text{rand}()$ returns a random number from a uniform distri-

bution $\in [0, 1]$.

2.1.3 Benchmark

This benchmark consists in playing against an agent that takes uniformly random moves. This is the most basic benchmark, but first we have to be sure that our learned evaluation function can play better than a random agent before moving into a harder benchmark. Also, this will help us to detect bugs in the code and algorithms: if a learned value function does not play significantly better than a random agent, is not learning. The idea is to test against this benchmark using Alpha-beta pruning at 1, 2 and 4-ply search.

3 Methodology

3.1 Section header

3.2 Algorithm

Q-learning, a model-free online off-policy algorithm, whose main strength is that it is able to compare the expected utility of the available actions without requiring a model of the environment:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (3.1)$$

3.2.1 Data Preprocessing

- Crop.

The raw input image is captured by Universe in the function "step()" which is simply a screenshot of the browser as shown in Figure 2. Obviously the useful vision information are only from the game screen. If we go further, the top half of the game screen displaying the sky does not change very much and the bottom half of the is heavily effected by the turning actions. Thus a region of interest is carefully chosen by whether considering whether it relates to decide

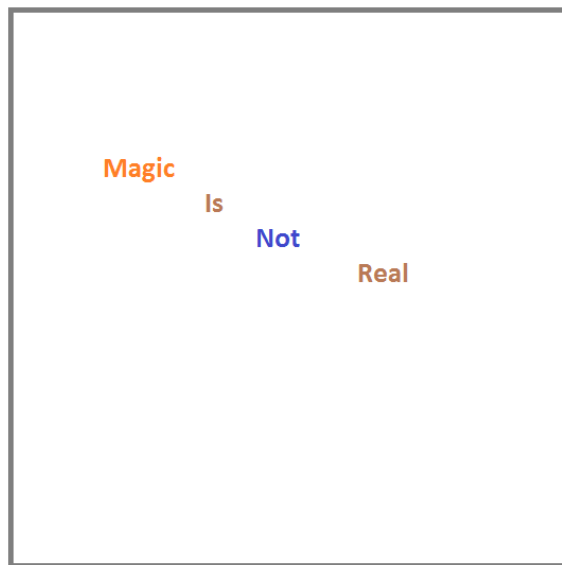


Figure 3.1. Cropped input image.

an expected steering behavior. Then the image is simply cropped by a new window with the "cropFrame()" function. An example cropped image is as shown in Figure 6.

- Downscale the resolution.



Figure 3.2. Downsized input image.

A high resolution is usually redundant for a computer vision task. By resizing with a smaller size, the space information are almost remained and the computing time are greatly saved. The cropped image is then downsized to a smaller size, [80, 80] as in Figure 7. Downscaling the resolution doesn't hurt the

information for turning left or right but highly accelerating the computing since much smaller data are being processed.

- Grayscale.

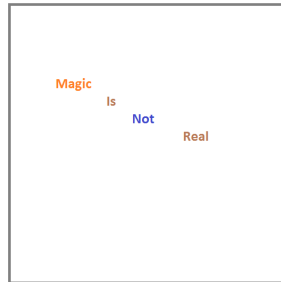


Figure 3.3. Grayscaled input image.

Grayscale processing is another useful technique in computer vision tasks since it is a great help for accelerating the computing and is at least three times faster than that of color image processing. This is because grayscale image has only one color channel as opposed to three in a color image. The color information are usually dumped when unnecessary for the computer vision tasks. As here, the downsized image is gray scaled as in Figure 8, because the color information does not help much for the steering control.

3.2.2 Implementation

Environment Setting. OpenAI Universe is designed with using Docker so it can reset the Game continuously. Docker is a tool that lets you run virtual machines on the local computer. I created an image wrapping up everything necessary, like TensorFlow, OpenCV, Gym and Universe.

To start up a very first Game Environment, only gym and universe modules are needed. Since we are to generate some random behaviors, "random" is also imported.

...

Build the Deep Neural Network. The Deep Neural Network is based on the DeepMind paper except for the changes of input and output sizes.

...

Preprocessing Functions. As described in the previous chapters, the crop, downsample and grayscale operations are done with the following functions:

...

Finetuning Hyperparameters. I played with some hyperparameters:

- The number of time steps
- The capacity of the replay memory,
- The layers of the network and the region of interest.

Currently this process is nothing more than trial and error. I could be better designed by grid search or random search. Well, it could be even easier to design the pipeline of implement a Deep Reinforcement Learning problem but tougher in hyperparameter tuning. Without a deeper understanding of the features of model, it has a high probability to fall into a less efficient experiment process.

3.2.3 Refinement

Relay Memory. Along the first several trials, I experimented with different memory capacities. The memory is a pool where the Game Bot would pick actions by Q values from. A small pool means it would be quick to converge during the training stage while it also limits the capacity it can learn. So there is a tradeoff here.

The Complexity of the Network. The original works well for some simple game such as Atari and CartPole. The Coster Racer Game here is more complex to play because it scores by driving along the way with a high speed. The delay between an action and a

reward is bigger. Through a long time training, it has a bad classification on Left Turn and Right Turn cases.

I replace the DeepMind model with LeNet-5 which then shows a better ability in classifying and quicker convergence.

4 Results

4.1 Specimen Selection & Characterization

4.1.1 Model Evaluation and Validation

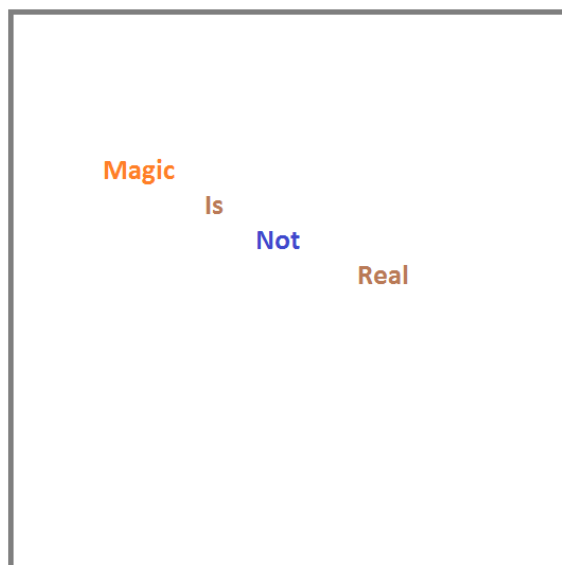


Figure 4.1. High score by the Game Bot.

Figure 9 is a screenshot when the car reach the finishing line and it got a score of 10269. The network was trained over 10,000 time steps and the behavior has been learned gradually. Compared to random bot, it has shown obvious better performance. For better or human level behavior, more training time steps are needed.

4.1.2 Justification

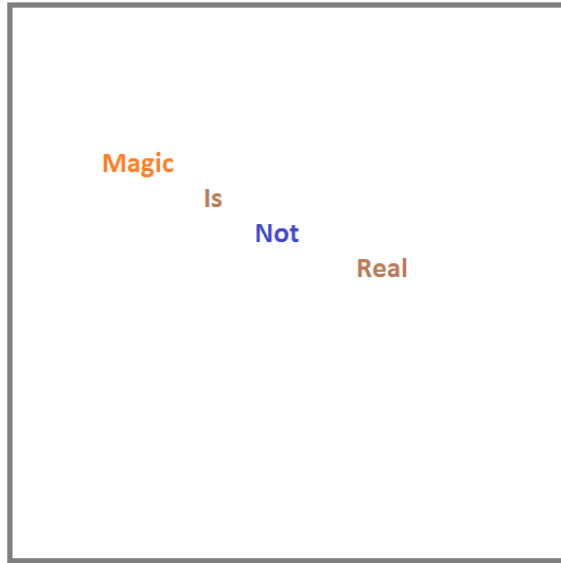


Figure 4.2. A comparison between DQN trained Game Bot and a random Game Bot.

The reward histories shown in Figure 10 indicates how it performs better than a random Game Bot when the DQN one is trained 100000 time steps. It proves a much better ability to earn scores or rewards though there is still a big room to improve. The Game Bot can earn as high as 10269 scores though it still usually ranks the last.

5 Discussion

6 Conclusions

6.0.1 Free-Form Visualization

A Youtube video has also been uploaded with the link <https://youtu.be/VdVA3od4tVs> here and it shows how it performs after 17000 time steps' training. It shows a capacity to stay in the road though it has some difficulty choosing right actions when blocked by the guard bars. The guard covered part of its view and the always acceleration actions made it even hard to get out of the stuck.

6.0.2 Reflection

Difficulties. The most difficult aspect of this project was that is extremely hard to stabilize reinforcement learning with non-linear function approximators. There are plenty of tricks that can be used and hyper-parameters that need to be tuned to get it to work, such as exploration policy, discount factor, learning rate, number of episodes, batch size, experience pool size and initial value.

All these techniques and parameters were selected by trial and error, and no systematic grid search was done due to the high computational cost. More than once it seemed that the implementation of the algorithms and techniques was incorrect, and it turned out that the wrong parameters were being used. A “simple” change such as decreasing ϵ ,

or changing the neural network optimizer made big changes in the performance of the value function.

Also a huge difficulty of a reinforcement learning problem could be the time lag between the action and the reward. When training with grouped actions off of the heuristic reward function, the reward for a given action was immediate, and this network showed the best performance. The next best performance came from a network based off of grouped actions, where actions were only a few steps removed from the next reward. Our worst performance came from the networks trained to estimate actions, where actions were several dozen steps removed from the next rewards.

General Pipeline. In a Deep Q Network setting, there are several elements which we have to be careful to define.

- **Environment:** An environment defines what the agent interacts with. It receives states and actions and generate new states and plays a role as an online data generator.
- **State Space:** A state is the input of the Deep Q Network. In this project, the stats is an image or a pixel array. It will be trained by a Deep Neural Network and predict the next actions.
- **Action Space:** An action space could be discrete and continuous. It defines the all classes that would be generated from the Deep Q Network. A bigger action space indicates a bigger room for an agent to learn and improve but also means a much complexity to train.
- **Reward Functions:** The reward function determines in which way we would like the agent to grow. For example, we would define a bigger reward for the car

to stay in the middle of the road than in the side of the road. It can be a discrete or a continuous function.

- **Deep Neural Network:** The Deep Neural Network is responsible to map the states to the Q values, which are corresponding to different actions by Q functions.
- **Fine tune the Hyperparameters:** By fine tuning the hyperparameters, we try to maximize the ability of the defined Deep Q Network. It can be subtle to modify the hyperparameter values which might change the output in different ways, like effecting the time of the convergence, the prediction accuracy, overfitting or underfitting and robustness.

7 Suggested Future Research

7.1 Continuation of Exposures and Analysis

7.1.1 Improvement

There are at least two aspects we can improve in the next stage,

- (1) **More Detailed Action Space:** Here only three actions are defined and the car is always in acceleration mode which is not the best choice in some cases. When the car is got stuck, acceleration wastes the effort of getting out. And it is easier to pass the sharp turn if it slows down the speed a little bit. Thus we can define more actions as follows.
- (2) **Tuned Reward Functions:** In this project, I almost relied on the game score as the reward resource. In fact, there are some rules to get maximal rewards which I don't think help learning a reasonable driving behavior. For example, there is no punishment on being blocked and falling into the sky. The Game Bot may think it is no big deal to fall into the sky and sometimes it can be rewarded 1000 score if it happen to hit a traffic cone before it falls down. I put my focus on building a working Deep Q Learning structure and had a hard time hacking into the scoring system, so I have to put this task next time.

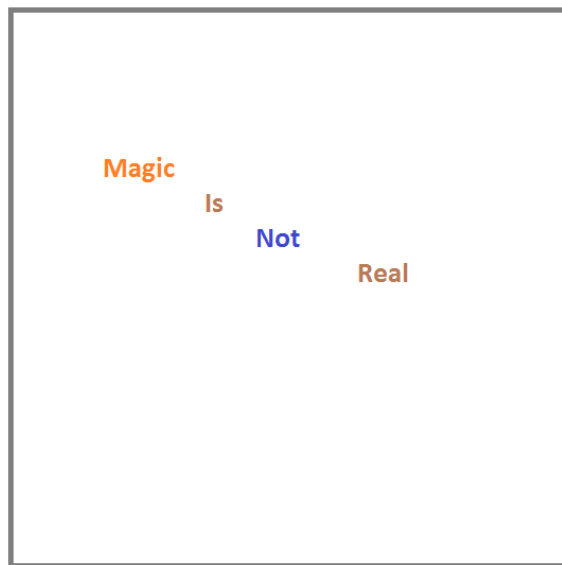


Figure 7.1. A bigger action space.

- (3) **Better benchmarks:** In most of this project we used simple benchmarks, such as playing against random players. While testing against random is probably the first thing to test against (if you can't beat a random player your learning algorithm is not working), it would be better to find a few heuristics and better players that can be used for testing.
- (4) **Incorporate other RL techniques:** The field of RL has been advancing fast in recent years. There are a few new and old techniques that I would like to try, such as asynchronous RL, double Q-learning, prioritized experience replay and Asynchronous Actor-Critic Agents (A3C).

Appendix A

Pprofile

Here is a section for HHP profile stuff, whatever that is...

- Pprofile 1 stuff here
- Pprofile 2 stuff here
- Pprofile 3 stuff here
- Etc...

Appendix B

Specimens

Here is a section to describe specimens used in this research.

- Specimen 1 stuff here
- Specimen 2 stuff here
- Specimen 3 stuff here
- Etc...

Appendix C

Preparation of this document

This document was prepared using pdf \LaTeX and other open source tools. The (free) programs implemented are as follows:

- \LaTeX implementation:

MiK \TeX

<http://www.miktex.org/>

T \TeX Live

<https://www.tug.org/texlive/>

Mac \TeX

<https://tug.org/mactex/>

- T \TeX -oriented editing environments:

Vim Text Editor

<https:http://www.vim.org/>

- Bibliographical:

Bib \TeX

<http://www.bibtex.org/>

Zotero

<https://www.zotero.org/>

Appendix D

Figures

Here is a section for figures in the document.

- Figure 1 stuff here
- Figure 2 stuff here
- Figure 3 stuff here
- Etc...