The University of Manchester

Department of Computer Science
Project Report 2020

Reinforcement learning for a learnable agent in classic arcade games

Author: Sean J Parker

Supervisor: Dr. Konstantin Korovin

Abstract

Reinforcement learning for a learnable agent
in classic arcade games

Author: Sean J Parker

The aim of the project is to investigate the performance of Gismos and to design and construct a super multi-functional Gismo.
The novel aspects of the new Gismo are described. The abstract should perhaps be about half a page long.
The results of testing, which show the abject failure of the Gismo, are presented.
In the conclusions proposals for rectifying the deficiences are outlined.

Supervisor: Dr. Konstantin Korovin

Acknowledgements

I would like to thank my parents, my school teachers, my friends, my wonderful supervisor, and all my wonderful fellow students for their unswerving support during my project. Without your help none of this would have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

My project aims to replicate some of the reinforcement learning algorithms that can be used to play classic Atari 2600 games. It also compares the results of different tests with these algorithms such as varying the hyperparameters of the network. By observing the effects on the trained agents[1] when the hyperparameters are changed we can deduce a set of optimal values such that the networks can play three different Atari 2600 games. Overall, the main features of the project are the following:

- Agents with raw pixel game data as input, outputting a set of values for the best action.
- Agents attempt to find an optimal model of the environment without any prior knowledge.
- Visualization of the agent "brain" to provide insight into what information the agents is learning.

1.1 Motivation

Over the past 10 years there has been significant improvement in the RL (reinforcement learning) algorithms. One reason is that the computing power has become cheaply available by using discrete graphics cards. For example, for my project I used a Nvidia GeForce GTX 1070 that provides 1920 CUDA cores that can be used to accelerate training of neural networks. Despite this, RL algorithms are massively computationally expensive and hence take a long time to train.

Over recent years one of the pioneers in this area is DeepMind, which was acquired by Google in 2014, and they developed the DQN (deep q-network) algorithm in 2013 which they demonstrated could learn directly from the raw pixel data of games in order to achieve either human-level or super-human level performance.

This research was expanded upon by DeepMind and OpenAI which is based on the original DQN by DeepMind. This research focused on trying to approximate a Q-function and thereby infer the optimal policy. On the otherhand, there has recently been a focus on other methods such as A3C and PPO which instead seek to directly optimise in the policy space of the environment.

[1] Agent. In this case, agent refers to a trained neural network that takes actions in a chosen environment.

1.2 Objectives

Further to what was described in section 1 there was a few main objectives of the project. Firstly, I chose three games on which I decieded to train the agents, Pong, Breakout, and Space Invaders which are shown below in Figure 1.1.

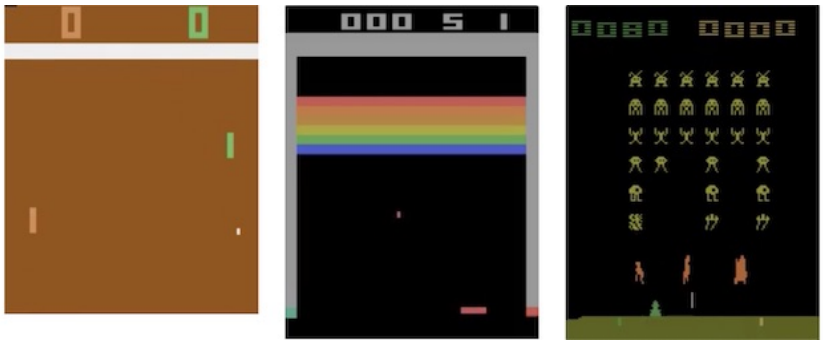

Figure 1.1: Screenshots of Pong, Breakout, Space Invaders (left to right).

Secondly, I wanted to find a way to explore the internals of a trained agent, in order to give further insight into what the agent is trying to learn. The reason for this is a researcher could use this information to determine, for example, where the agent has learnt to focus on the frame. Additionally, it provides a insight into how to optimally tune the hyperparameters which is described in section TODO: include section.

TODO: include gantt chart

1.3 Report structure

My report is divided into three main sections. Firstly, describing the background of the problem, then going onto giving details of my implementation and finally project evaluations/conclusions.

Chapter 2

Background

This chapter will cover some of the background material required for the following sections, it will cover the history of reinforcement learning (RL) and it's evolution to the current state-of-the-art. Additionally, it will cover the related work to this project and also cover some details of the past research papers for which this project has been based upon.

2.1 Reinforcement learning

Reinforcement learning is an area of machine learning has has been under active research since the late 1980s (TODO: ref watkins phd). The first defining algorithm of RL was called *'temporal-difference learning'* often referred to as TD-Learning. This algorithm learns by bootstrapping from the current value function in order iteratively converge towards a optimal policy (i.e. the agent's strategy for taking actions in the environment).

Further work by R. Sutton led to the development of TD-Lambda, an algorithm that was applied to the game of Backgammon, in 1992, by Gerald Tesauro to create TD-Gammon(TODO: ref Tesauri paper). It was a computer program that was shown to compete at expert-human level. The program also found novel strategies that were either unexplored, or dismissed in error as poor strategies. This was the first example of RL aiding in discovery or reconsideration of board game strategies. This trend of RL algorithms helping to improve human play would prove to only continue with DeepMind's AlphaZero program mastering the games of Chess (beating the strongest Chess programs such as Stockfish[1]) and Go.

2.1.1 Deep reinforcement learning

Following from section 2.1 on RL, this section talks about the combination of two areas, deep learning and reinforcement learning methods, called deep reinforcement learning (DRL). Deep learning (DL) is a common class of machine learning methods that has been of much research focus over the past decade and can deal with high-dimensional sensory input; for the case of Atari this is 84x84 greyscale images after pre-processing of the raw Atari frames. On the other hand, reinforcement learning allows us to create an agent which can learn an optimal policy to navigate some environment in order to optimise its reward.

Through the combination of these methods, it has proven to provide solutions to previously intractable problems [1] in areas such as robotics, computer vision and healthcare. For example,

[1] Stockfish. One of the strongest Chess programs based on the CCRL ratings list.

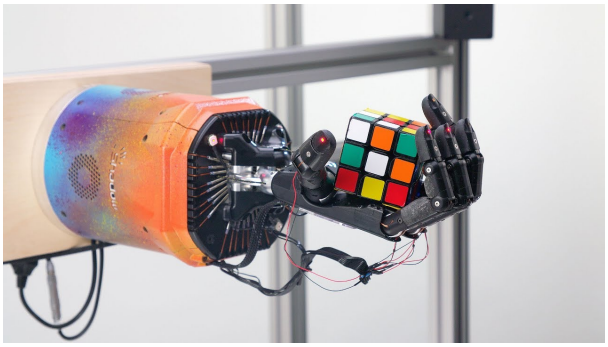

Figure 2.1: OpenAI Robot solving a Rubik's cube

in 2019 OpenAI developed a robotic hand that could solve a Rubik's cube 2.1, trained using deep reinforcement learning. As an extension to DRL, end-to-end reinforcement learning is a method for single layered neural network, trained by reinforcement learning. Figure 2.2 shows, diagramatically, how DL and RL are used together in order to produce a single end-to-end model. In this simple architcture, there are two main components, the agent and the environment. This is a key feature of all DRL methods, an agent observes some state and reward from the environment after taking an action.

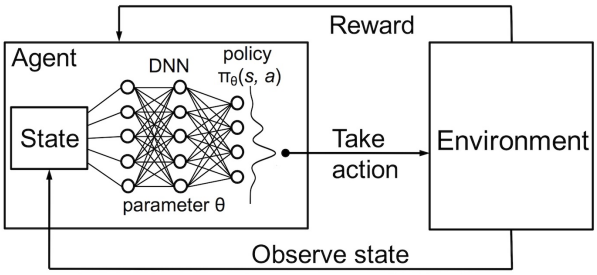

Figure 2.2: Representation of end-to-end RL architctures

As was mentioned in Section 2.1 there has been a renewed focus on the area of DRL. This research trend originated in 2013 when DeepMind showed that using a combination of Q-Learning and deep learning, it can produce agents that can compete with expert-humans in Atari.

2.2 DQN on Atari 2600

As mentioned in previous sections, one of the pioneering companies in the area was DeepMind. In 2013, V.Mnih et.al. while working at DeepMind, released a paper titled "*Playing Atari with Deep Reinforcement Learning*"[2]. This paper combined DRL, convolutional neural networks and a novel strategy called **experience replay**. DeepMind, in 2014, patented Q-Learning and it's application with deep learning on Atari games. Further, they also published further papers expanding on the idea in prestigious journals such as NIPS and Nature. Figure 2.3 shows the structure of a Deep Q-Network as used to play Atari 2600 games.

Experience replay was an important improvement that helped in stabilising the Q-Learning algorithm. It does this by removing the correlations between a observation sequence, i.e. we

don't learn from a chronological series of frames, rather we store all the experience and learn using random samples of this memory.

Over the following few years, this area would see rapid progress with many different advancements by both DeepMind and OpenAI. As was noted with the introduction of Deep Q-Learning (often referred to as '*Deep Q-Networks*' or *DQN*) the algorithm sufferes from overestimating the value of some actions. This can lead to a poor performance on more complex Atari games such as Space Invaders. Further detailed descriptions of the Q-Learning algorithm is provided in Section 3.3.

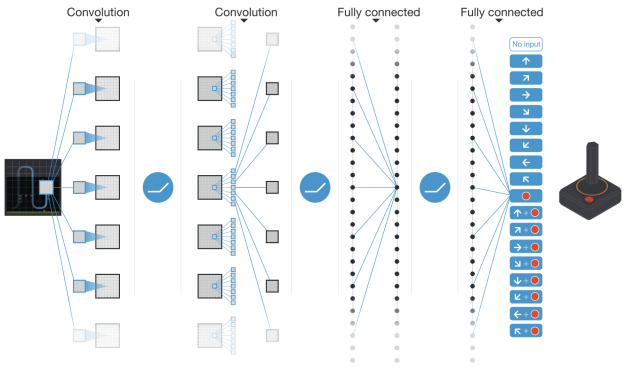

Figure 2.3: Deep Q-Network architcture that from left-to-right is showing how a single frame is processed. First passing through convolutional layers, before being flattened to a single tensor which is connected to a fully connected network. This MLP acts as the Q-function approximator which will produce an output of predicted Q-values for each action.

2.3 CNN Visualisation

Chapter 3

Design

This section will cover the details of the methods and algoithms that were used in the implementation of the project. The previous section covered the history of some of these techniques which were imperative for building the foundation of the methods described in this chapter.

Deep Q-Networks and its enhancements have the majority of focus, as it is the basis of the project. Below is a table of the different experiments and the methods that were used in order to produce the best solution for each tested environment.

Network	Algorithm	Environment
CNN + MLP	Q-Learning	Pong
CNN + MLP	Q-Learning	Breakout
CNN + MLP	Duelling Double Q-Learning	Space Invaders

Table 3.1: Best performing method in given environment

3.1 Markov decision process

This section will describe some of the basics of Markov decision processes (MDP), they are the foundation of the reinforcement learning methods that will be described later in this chapter. First, we need some mathematical definitions of MDPs, these are from David Silver's excellent lecture series on Reinforcement Learning.

Markov decision processes are just markov reward processes with decisions, i.e. At each state S_t, we have a finite set of actions to choose from in order to get to a new state S_{t+1}.

Definition 3.1. A Markov decision process is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$.

- $\mathcal{S}$, finite set of states
- $\mathcal{A}$, finite set of actions
- $\mathcal{P}$, state transition probability matrix,
  $\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$
- $\mathcal{R}$, reward function, $\mathcal{R}^a_s = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$
- γ discount factor, $\gamma \in [0, 1]$

The definition above defines a Markov decision process, which we use as a basis for describing the methods in reinforcement learning. In order to illustrate the idea, let us consider the game of Pong. For simplicity, assume we can encode each frame of the game into the set of states $\mathcal{S}$. In order to play the game, we need to know what the best action to take would be at each frame of the game to move the paddle under the ball, hitting the ball back and (hopefully) scoring a point.

Given a frame of the game S_t and the set of actions we can choose from A_t, we are going to try and maximise our future (expected) reward using the reward function $\mathcal{R}_{S_t}^{A_t}$. We want to choose the best action a, that will result in the maximum future reward. It is important that we don't look at immediate rewards only, since, in Pong we don't get the point until we have hit the ball back to the other side.

In order to look at future rewards, we use the discount factor γ. When γ is close to zero, we are "*myopic*" in our evaluation (we only look for short-term rewards). However, as γ gets closer towards 1, we are "*far-sighted*" in our evaluation.

Overall, we need to know a strategy that provides the best action to take in a given state which maximises the expected total reward. This is called a *policy*, denoted by π.

Definition 3.2. A *policy* π is a distribution over actions given states,

$$\pi(a|s) = \mathbb{P}[A_t = a \mid S_t = s]$$

The policy of an agent fully describes the behaviour of the agent (TODO: add ref) which only depends on the current state, not the history. In order to describe the optimal policy for an agent to follow, we first need some more definitions.

Definition 3.3. G_t is the total discounted reward for time-step t

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Definition 3.4. The *action-value* function $q_\pi(s,a)$ is expected return starting from state s, taking action a, following policy π

$$q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

Definition 3.5. The *optimal action-value* function is denoted by $q_*(s,a)$ and is the maximum action-value function over all possible policies

$$q_*(s,a) = \max_\pi q_\pi(s,a)$$

3.2 Reinforcement learning

3.2.1 Exploration vs Exploitation

3.3 Q-Learning

3.4 Q-Learning improvements

3.4.1 Double Q-Learning

3.4.2 Duelling Q-Learning

References

[1] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath. A brief survey of deep reinforcement learning. *CoRR*, abs/1708.05866, 2017.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning, 2013.