Machine Learning Nanodegree
Capstone Report

David A. Robles

January 31, 2017

1 Definition

1.1 Project Overview

Content.

1.2 Problem Statement

In this project, we will use reinforcement learning with deep learning to make an agent learn to play the game of Connect 4[1] by playing games against itself. In other words, using the formalism used by Mitchell (1997) to define a machine learning problem:

- **Task:** Playing Connect 4.
- **Performance:** Percent of games won against other agents, and accuracy of the predictions on a Connect 4 dataset.
- **Experience:** Games played against itself.
- **Target function:** $Q^{\pi} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, where $\mathcal{S}$ is the set of *states* (board positions) and $\mathcal{A}$ is the set of *actions* (moves), 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:

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A} \tag{1}$$

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

1.3 Metrics

- **Winning percentage.** This metric consists in playing a high number of games (e.g. 100,000) against another agent (e.g. a random agent), and calculating the average of games won by the agent that uses the learned value function.
- **Prediction accuracy.** The learned value function will be used to predict the game-theoretic outcomes (win, loss or draw) of the board positions in the Connect 4 Data Set.

[1] https://en.wikipedia.org/wiki/Connect_Four

2 Analysis

2.1 Data Exploration

Content.

2.2 Exploratory Visualization

Content.

2.3 Algorithms and Techniques

2.3.1 Alpha-beta pruning

Alpha-beta pruning (Knuth and Moore, 1975) is the most common game tree search algorithm for two-player games of perfect information. It extends the minimax algorithm to reduce the number of nodes that are evaluated in the game tree. Instead of calculating the exact minimax values for all the nodes in the game tree, alpha-beta prunes away branches that will not have any effect in the selection of the best move.

Implementation

2.3.2 Q-learning

Implementation

One of the most basic and popular methods to estimate action-value functions is the *Q-learning* algorithm (Watkins, 1989). 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. 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,

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{2}$$

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.

Q-learning is exploration-intensive, which means that it will converge to the optimal policy regardless of the exploration policy being followed, under the assumption that each state-action pair is visited an infinite number of times, and the learning parameter α is decreased appropriately (Watkins and Peter, 1992).

2.3.3 Self-play

Self-play is by far the most popular training method. It is a single policy $\pi(s, a)$ that is used by both players in a two-player game, $\pi_1(s, a) = \pi_2(s, a) = \pi(s, a)$. The first reason for its popularity is that training is quickest if the learner's opponent is roughly equally strong, and that definitely holds for self-play. As a second reason for popularity, there is no need to implement or access a different agent with roughly equal playing strength. However, self-play has several drawbacks, with the main one being that a single opponent does not provide sufficient exploration (Szita, 2011).

2.4 Benchmark

Content.

3 Methodology

3.1 Data Preprocessing

Content.

3.2 Implementation

- Two games were implemented: Tic Tac Toe and Connect 4. - Converting a to an MDP by using a fixed agent. - Show that q-learning learns to play against a fixed opponent using Tic-Tac-Toe. - Show that q-learning learns to play against a fixed opponent using Connect 4.

3.3 Refinement

Content.

3.4 Games

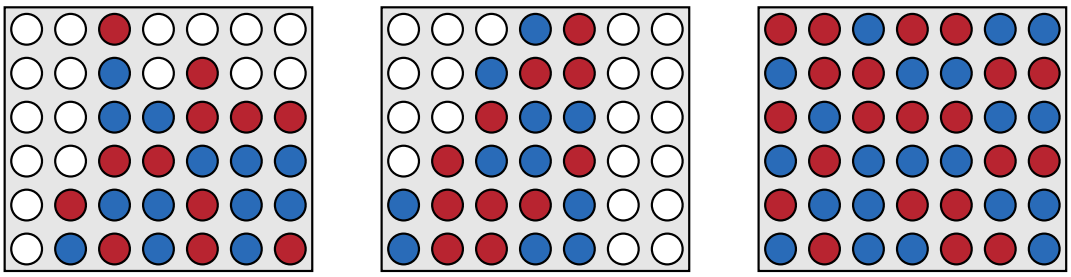
3.4.1 Tic Tac Toe

Implementation

3.4.2 Connect 4

Implementation

Connect 4 is a two-player board game of perfect information where pieces are dropped into the columns of a vertical 6×7 grid with the goal of forming a straight line of 4 connected pieces. There are at most 7 actions per state, since placing a piece in a column is a legal action only if that column has at least one empty location. Figure 1 shows three Connect 4 game positions.



(a) Win for Player 1 **(b)** Win for Player 2 **(c)** Draw

Figure 1: Examples of wins, losses and draws in Connect Four

4 Results

4.1 Model Evaluation and Validation

Content.

4.2 Justification

Content.

5 Conclusion

5.1 Free-Form Visualization

Content.

5.2 Reflection

Content.

5.3 Improvement

Content.

References

Donald E. Knuth and Ronald W. Moore. An analysis of alpha-beta pruning. *Artificial Intelligence*, 6(4), 1975.

T. M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

I. Szita. Reinforcement learning and markov decision processes. In *Reinforcement Learning: State of the Art*. Springer, 2011.

Christopher J.C.H. Watkins and Dayan Peter. Q-learning. *Machine Learning*, 8:279–292, 1992.

C.J.C.H. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, 1989.