R.A. Stringer

Deep Q-Theory

The deep Q network acts as a function approximator. It produces a vector of all possible action values, and selects actions from the maximum value of those options. The algorithm is fed back a reinforcement signal of the change in game score at each time step, so that it can generalize which actions prove most fruitful. By training through hundreds or thousands of episodes, the agent learns which actions in a variety of situations are worthwhile, even in complex state spaces such as the video game environment in this project.

This approach to reinforcement learning was first publicized by Deepmind researchers in 2015. https://storage.googleapis.com/deepmind-media/dqn/DQNNaturePaper.pdf

Vanilla Q-Network

The first implementation in the model is a convolutional neural network of two hidden layers and one fully-connected layer, with RELU activation functions to map state to values learned during training.

$$Q^*(s,a) = \max_{\pi} \mathbb{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots | s_t = s,\, a_t = a,\, \pi\right],$$

which is the maximum sum of rewards r_t discounted by γ at each time step t, using behaviour policy $\pi = P(a|s)$, after experiencing observation (s) and taking action (a). (Deepmind authors, 2015).

Duelling Q-Network

Since the values of most states don't vary across actions, the duelling Q-network estimates state values and captures the difference actions make in each state as an 'advantage'. The advantage for each action, and the state values then branch into their own fully-connected layers in the model. The final Q values are derived by adding the values, advantages, and subtracting the mean of the advantages.

Hyperparameters

The discount factor of 0.995, initial epsilon-greedy value of 1 and final epsilon-greedy value of 0.1 mimic those in the Deepmind paper.

Results

The vanilla approach solved the environment (achieving a score of 13 or more in 100 episodes) after around 500 episodes. It then scored 13-14 consistently during test play.

The duelling network solved the environment in a similar number of episodes, however scored significantly during test play, around 20-23+.

Future Work

I would like to implement the 'rainbow' algorithm – a combination of several deep Q-learning approaches – and chart the results.