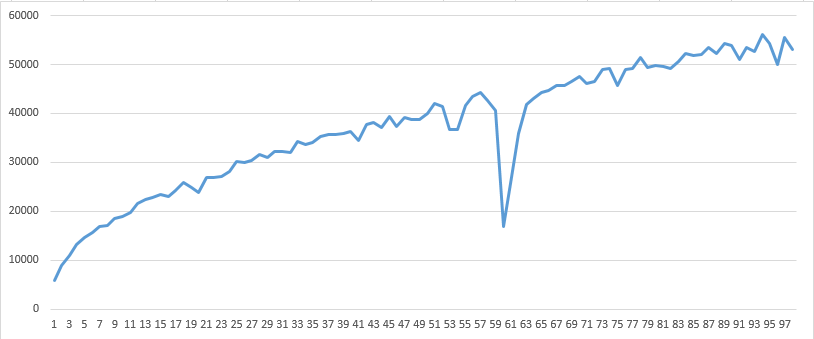
Lab5 Report

0412237

陳逸豐

⚫ Report (70%)

**1.A plot shows episode rewards of at least 100,000 training episodes (10%)**



x-axis is training games(x 1000).

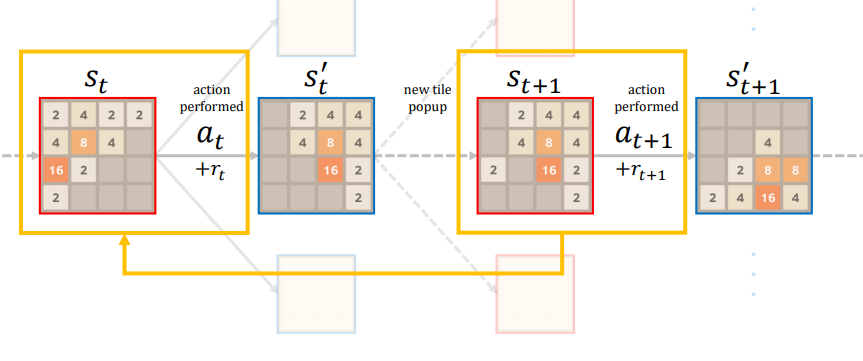
y-axis is reward.

**2.Explain the mechanism of TD(0) (10%)**



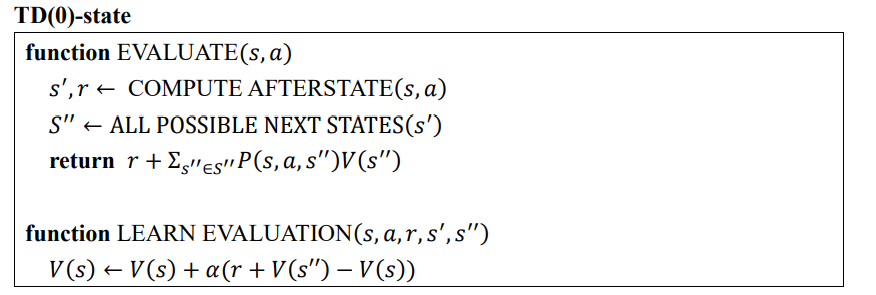
TD(0) estimate the reward by updating the neural network V which will evaluate the state’s reward. In this lab, we replace the neural network with n-tuple network, since it will need lots of resource to store the whole state. In n-tuple network, we extract 4 features which is 6-tuple to represent the whole state. The more game the agent play, the higher accuracy of the TD(0) will be. After it play lots of game, it will learn how to estimate the reward of the state.

**3.Describe how to train and use a V(state) network (20%)**



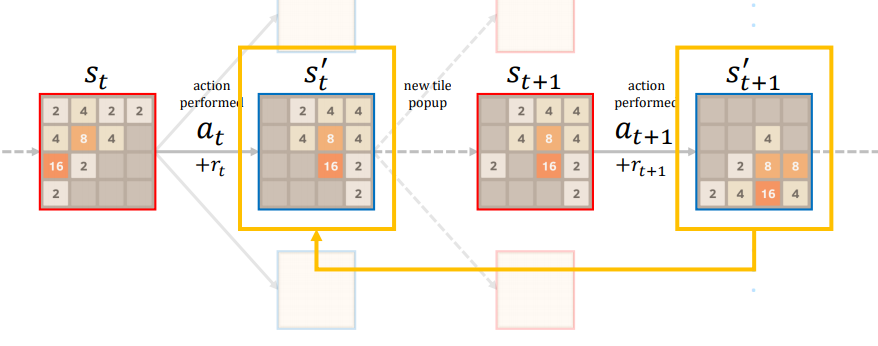
In V(state) network, its input is a state and output is the expected value for applying action on the state. After playing an episode, it records the before-state St, the action at, the reward rt and the next before-state St+1. Then, Update value 𝑄 (𝑆𝑡, 𝐴𝑡) toward TD target 𝑅𝑡+1 + 𝑄 (𝑆𝑡+1, a).

To use a V(state), input a state and decide a way to go which will get max reward.



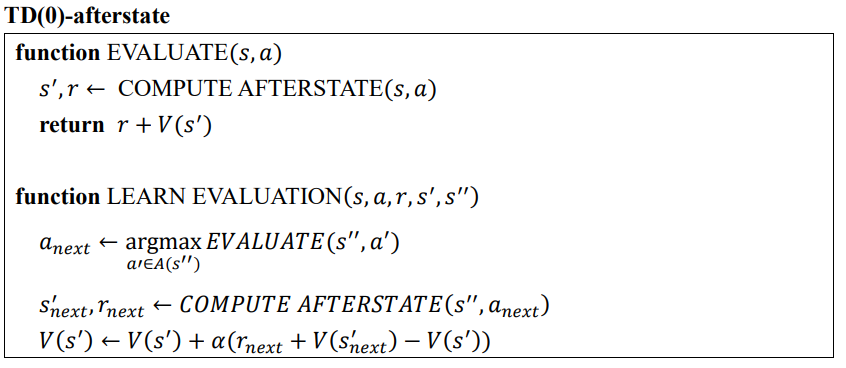
**4.Describe how to train and use a V(after-state)**

**network (20%)**



In V(after-state) network, its input is an after-state and output is the expected value for after-state. After playing an episode, it records the after-state St’, the reward rt’ and the next after-state St+1’. Then, Update value V (St’) toward TD target 𝑅𝑡+1 + V (𝑆𝑡+1’).

To use a V(after-state), use the network four times which includes four ways to move and chose the one get max reward.



**5.Describe how the code work (the whole code) (10%)**

It will new a class ‘learning’ which will update the network. Then, add the feature into learning, feature in this lab is 6-tuple, and we use 4\*6-tuple in the n-tuple network. And it will have 8 possible isomorphisms, since it can rotate in four positions and mirror, so 4(four positions) \* 2(mirror or not). In every episode, it will choose one position from the four position which will get max reward and record the state, action and reward to update the V(state) and the V(after-state). The learning rate(alpha) is set to 0.1

**⚫ Strength (30%)**

**The 2048-tile win rate in 1000 games, ⌈winrate2048⌉** 