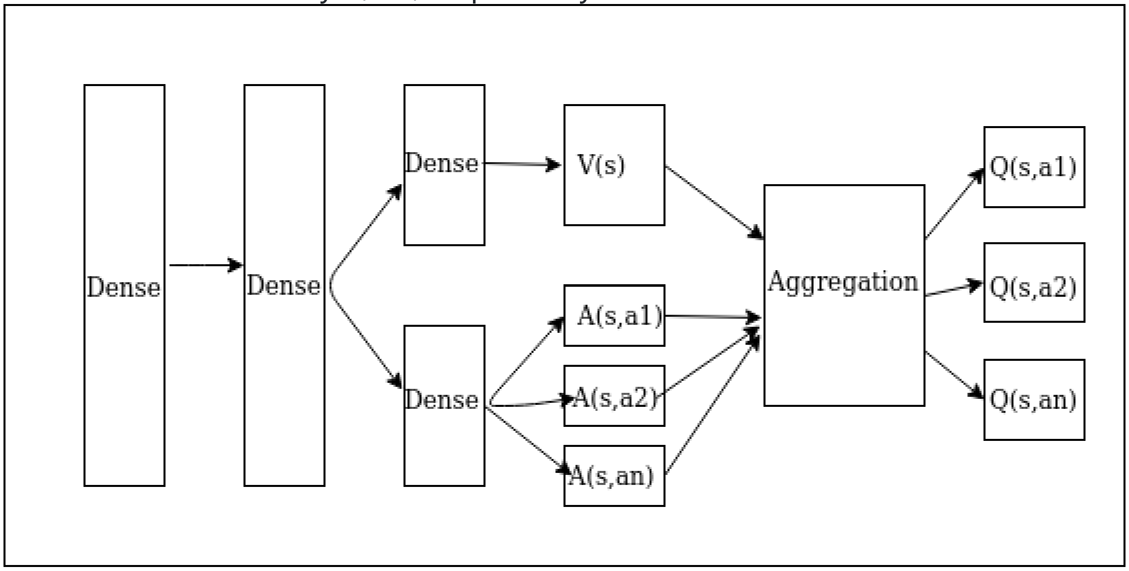
Report of Navigation

1. Define the Deep Q-Network:

- I have use Double DQN and Dueling DQN in this practice
    1. Double DQN seems improve the training performance significantly in 500 episodes
    2. Dueling DDQN does not that significantly, it seems has a slightly advantage in narrowing the variance during the training.

- As we had observed the Vector Observation space size in this environment is 37, so I have defined First hidden layer units as 128 and 2nd hidden layer as 32, small enough to run in CPU, since I have used up my GPU time.

- Then the 2-liner hidden layers for state value and advantage values same as 2nd hidden layer, 32, respectively



- Then concoct the state value with the advantage values as

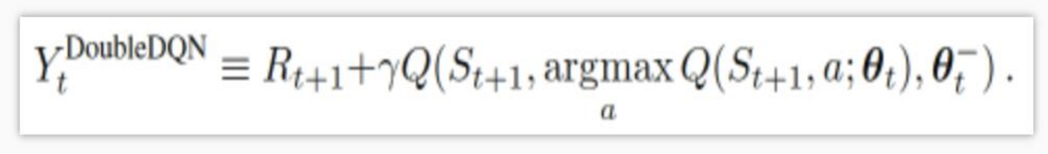
$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + \left(A(s,a;\theta,\alpha) - \frac{1}{|\mathcal{A}|}\sum_{a'} A(s,a;\theta,\alpha)\right)$$

2. Define the Replay Buffer:

- Make a deque for memorizing episode
- Each step of episode will be saved in the buffer
- Sample randomly from the buffer once it was filled more than the batch size
- Have not implement prioritized DDQN here

3. Define the Agent:

- Since DDQN is off-policy learning, I need to define 2 Deep Q-Networks: local and target, the local network will used for generating policy, and the target network will be used as updating the better policy
- As Double DQN, In each learning step, we will find the max value actions in local network, and then use this action to get Q value from the target network.



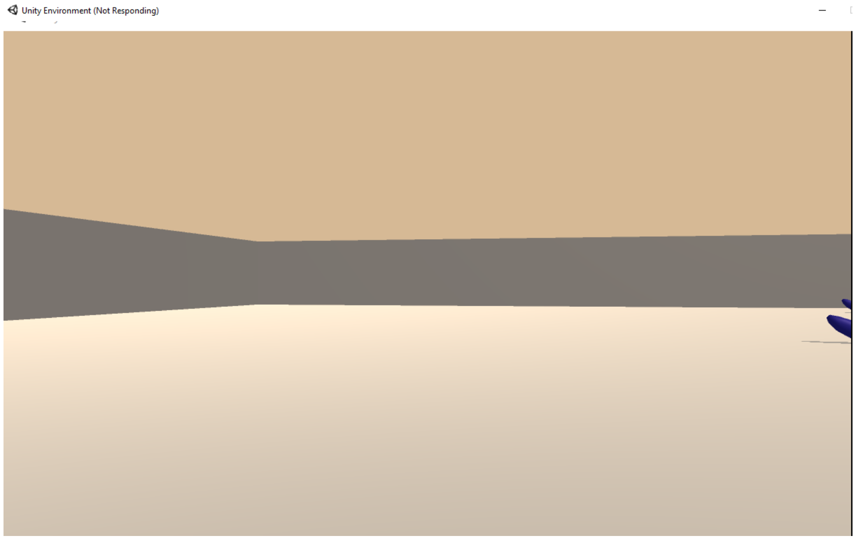
$$Y_t^{\mathrm{DoubleDQN}} \equiv R_{t+1}+\gamma Q(S_{t+1}, \underset{a}{\operatorname{argmax}}\, Q(S_{t+1}, a; \boldsymbol{\theta}_t), \boldsymbol{\theta}_t^-)\,.$$

4. Training the DDQN:

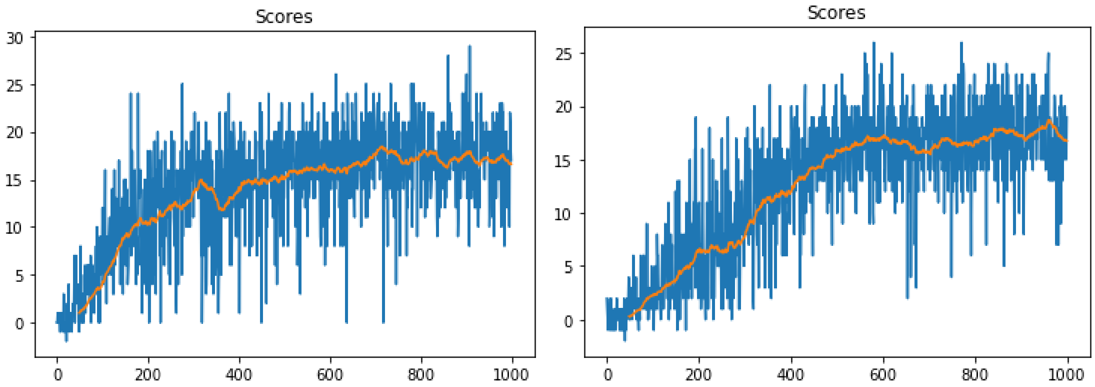
- Epsilon starts from 1 and times 0.97 as decay, this will make it down to 0.002 around 200 episodes
- The navigations go 300 - 1000 steps, so make the minimum epsilon as 0.0005 gets very good result at first 300 episodes
- The model reaches average score as 13 around 300-350 episodes, and 15 in less 500 episodes

5. Interesting observations:

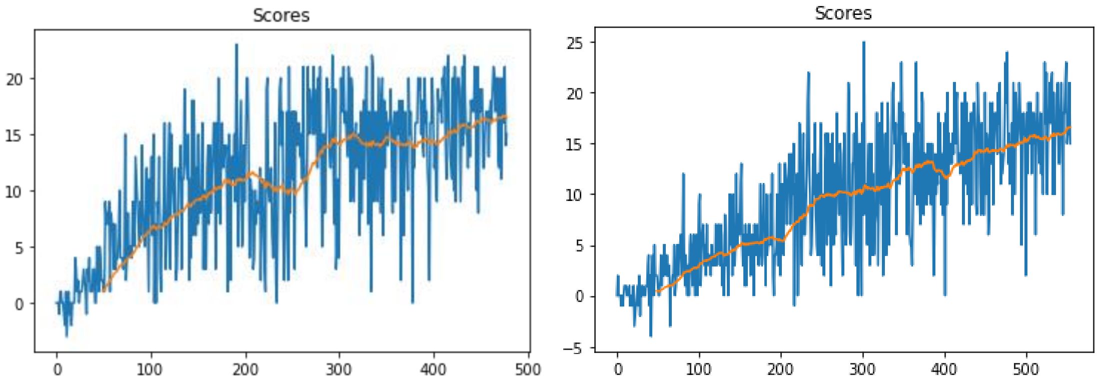
- If the model stops at average score 13, some of them would be stuck in the situation that there is no yellow banana in the view, but the model trained with Dueling DDQN get higher possibility get out of this situation. When the avg score up to 15, most model would know try to turnaround.



- Dueling DQN seems has a slightly advantage in narrowing the variance during the training.



Vanilla DDQN & Dueling DDQN with 1000 episodes with 64x16x16 network



Vanilla DDQN & Dueling DDQN with 500 episodes with 128x32x32 network