1. Learning Archetecture
    a) Algorithm

Input: batch size k, learning rate η, number of episodes N, initial epsilon for epsilon-greedy policy ϵ_0, epsilon decay rate r_ϵ, discount factor γ, soft update factor τ

Initialize replay memory buffer $\mathrm{H} = \varnothing$, $\epsilon = \epsilon_0$, $t = 0$

for i = 1 **to** N **do**

Observe $S_t = S_0$

while True **do**

Choose action $A_t \sim \pi_{\theta,\epsilon}$

Observe S_{t+1}, R_{t+1}

Store transition $(S_t, A_t, R_{t+1}, S_{t+1})$ in H

If $t \equiv 0 \; mod \; K$ and $len(\mathrm{H}) > k$ **then**

Sample batch H_s from H

Compute TD-error for $(S_j, A_j, R_{j+1}, S_{j+1})$ in H_s

$\delta = R_{j+1} + \gamma \, Q_{target}(S_{j+1}, arg\; max_a Q(S_{j+1}, a)) - Q(S_j, A_j)$

Update weights $\theta \leftarrow \theta + \eta \cdot \delta \cdot \nabla_\theta Q(S_j, A_j)$

Update target network $\theta_{target} = \tau * \theta + (1 - \tau) * \theta_{target}$

Update epsilon $\epsilon = \epsilon \cdot r_\epsilon$

$t = t + 1$

if done **then**

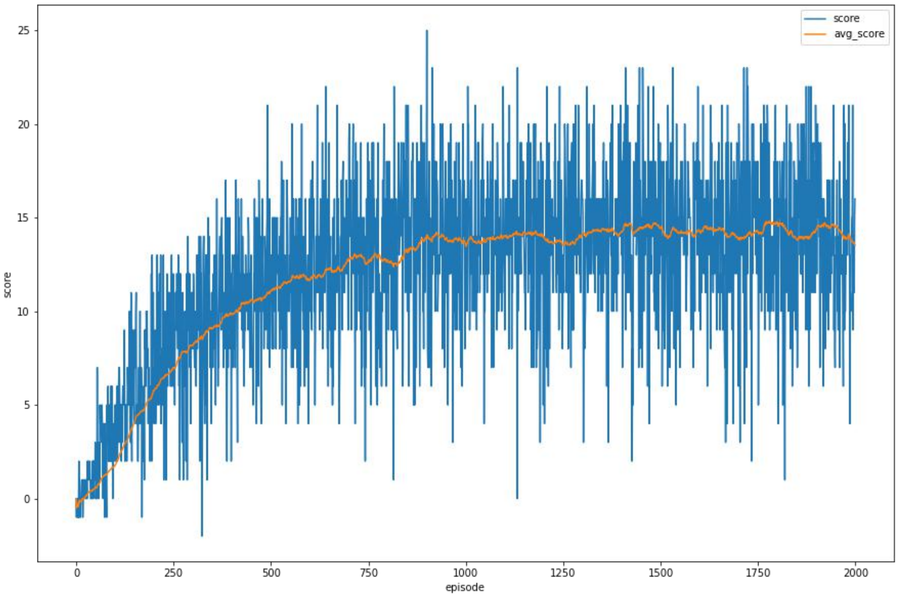
break

end for

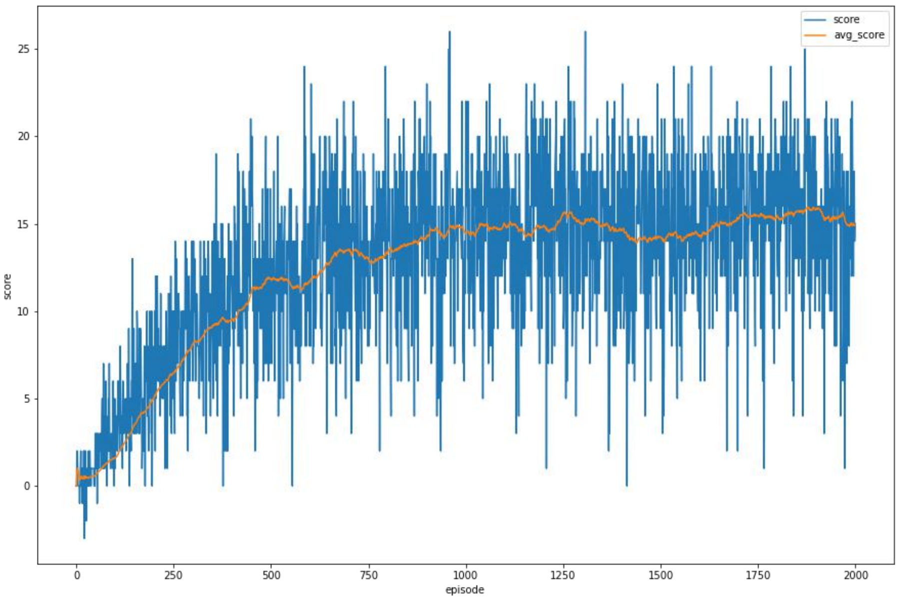
    b) Model Structure
        i. Input size = state size = 37
        ii. Hidden layers(2)
            1. Fully connected with 8*37 rectifiers
            2. Fully connected with 8*37 rectifiers
        iii. Output layer of size 4(=action size)
    c) Hyperparameters

i.	Batch size	64
ii.	Memory buffer size	1e5
iii.	Number of episodes	2000
iv.	Epsilon decay rate	0.995
v.	Target score	13.0
vi.	Discount factor gamma	1e-3
vii.	Learning rate	5e-4
viii.	Update Period	4
ix.	Sampling priority buffer	0.1
x.	SamplingWeightOrderIncreaseSpeed	1.0/1500.0
xi.	SamplingPriorityOrderIncreaseSpeed	1.0/1500.0

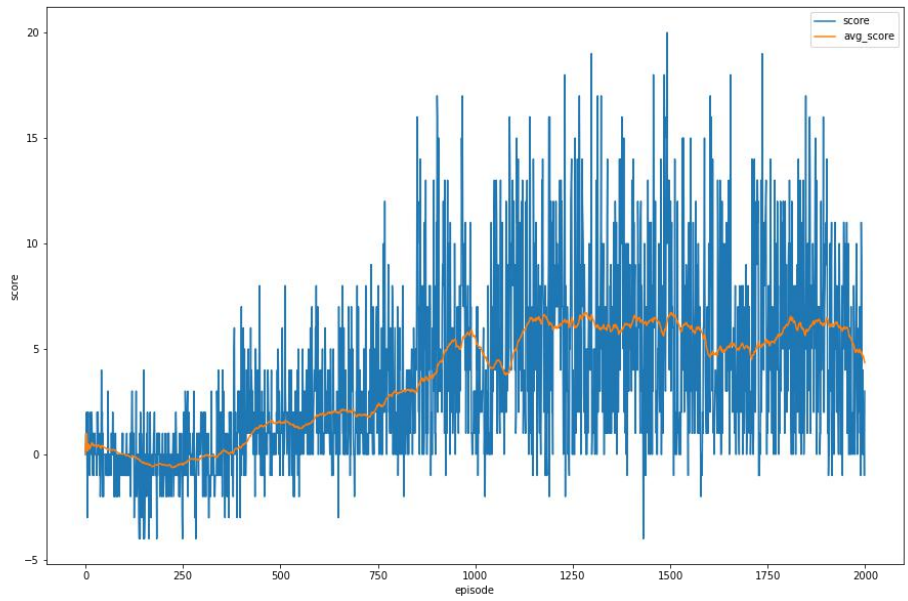
2. Results
    a) Original DQN
        i. Folder: dqn_result_2022_04_02_14_24_08
        ii. Result:



b) Double DQN
   i. Folder: dqn_result_2022_04_02_16_29_25
   ii. Result:



c) Double DQN + Prioritized Experience Replay
   i. Folder: dqn_result_2022_04_11_12_18_33
   ii. Result:



3. Conclusion
    a) Both original DQN and Double DQN can converge fast enough(within 1000 episodes)
    b) Comparing to original DQN, Double DQN reached a highier final score(14.92 vs 13.56)
    c) Prioritized Experience Learning can converge, but the final score is much lower than others(4.35), the reason maybe implementation error or inappropriate hyperparameters
4. Future Improvements
    a) Will try differrent hyperparameters by using google's ML hypertun subsystem Vizier
    b) Will try to correct implementation of Prioritized Experience Replay
    c) Will try to implement Dueling DQN
    d) Will try learning from pixels
    e) Will try Adaptively Parametric ReLU