Milestone #2 Progress

Kevin, Haiwen, Eric, Phil

Current Tested Model: DRQN

DRQN to solve POMDP:

1. Use Recurrent Neural Network approximates Q-value

2. DRQN takes **observation history** vs. DQN takes **observation points**

3. Better than DQN in solving POMDP

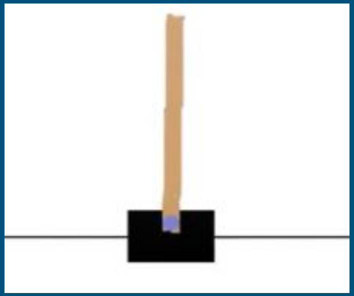
Test Environment #1: Partial Cartpole

Original Cartpole State Space:

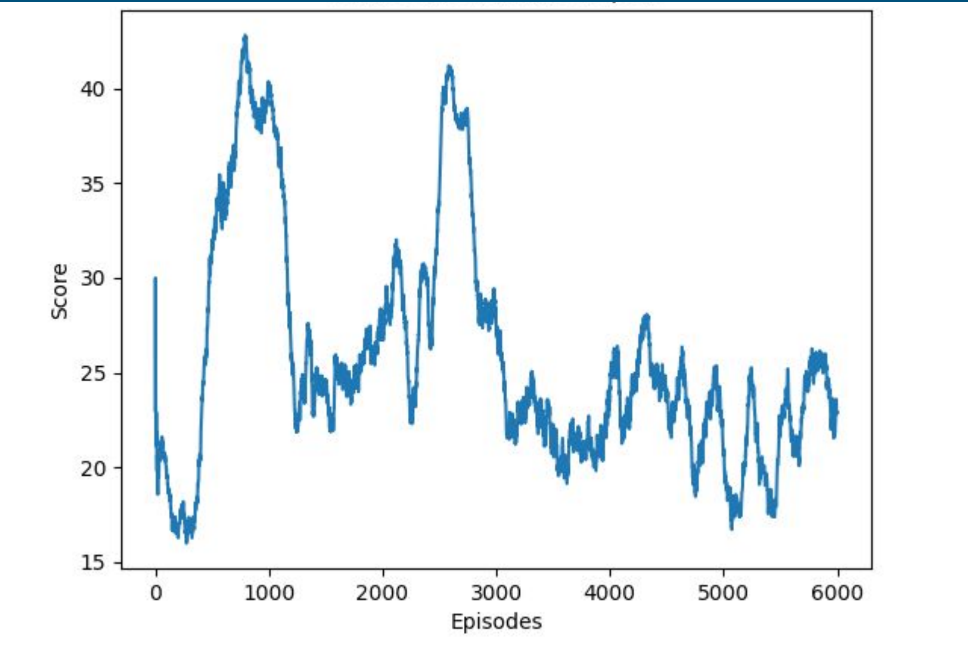
[car position, car velocity, pole angle, pole angular velocity]

Partial Cartpole Observation Space:

[car position, ~~car velocity,~~ pole angle, ~~pole angular velocity~~]



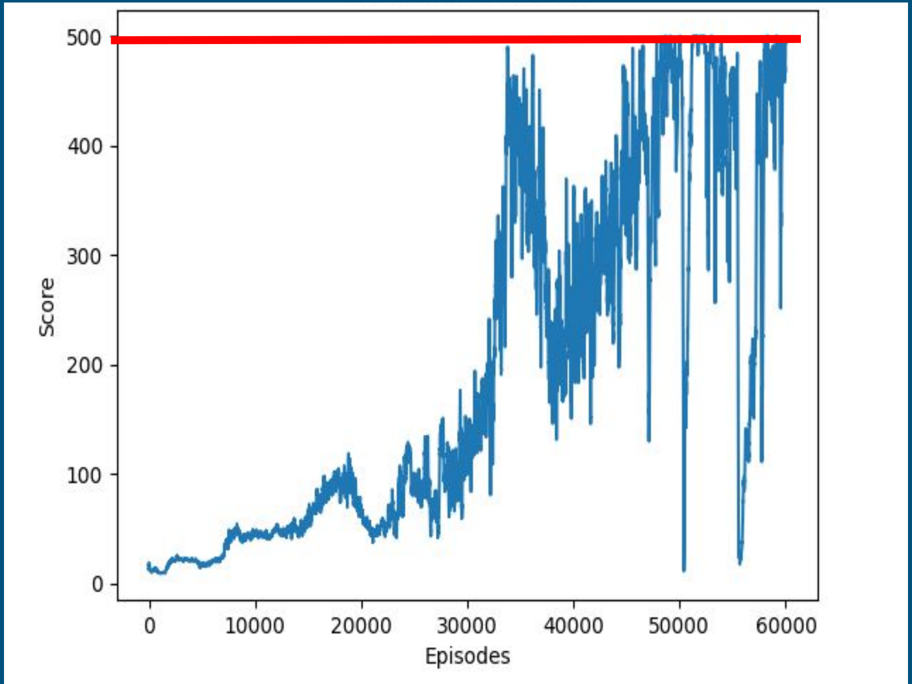
DQN vs. DRQN in Partial Cartpole -- DQN



DQN Performance:

1. Optimal policy should converge to **500**
2. DQN average score around **20**
3. Not showing convergence in 6000 Episode

DQN vs. DRQN in Partial Cartpole -- DRQN



DRQN Performance:

1. Reach **495** in terms of moving average per 100 episodes

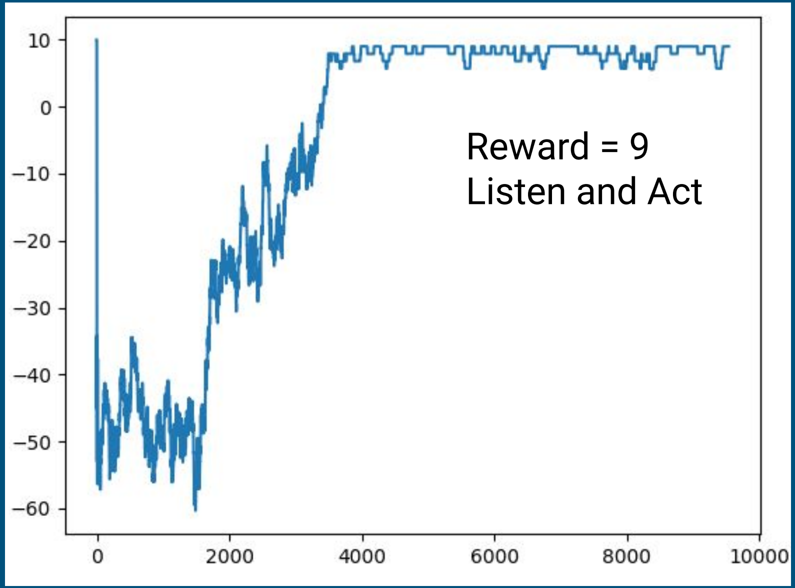
2. Converge to the Optimal Policy

Note: Total Training Step is Controlled in the test

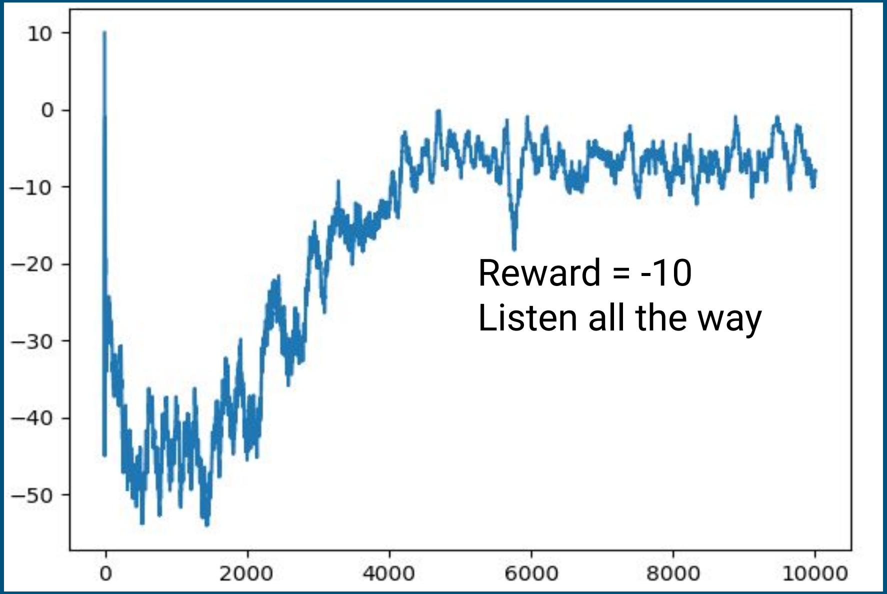
Test Environment #2: Tiger

DQN vs. DRQN in Tiger without Extension

DQN reward: **Optimal Policy**



DRQN reward: **Bad policy**



Why DRQN Performs Worse in Tiger?

Problem:

Short Sequence Experience hurts learning of RNN

Solution:

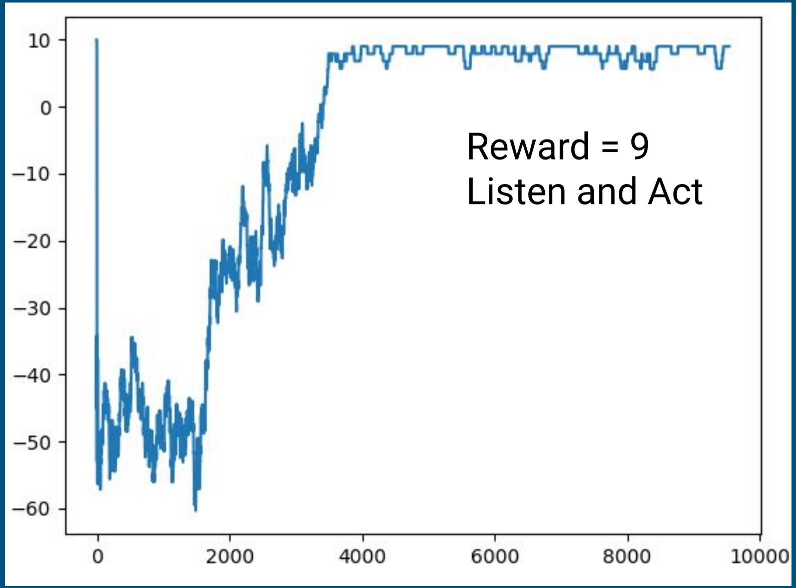
We force the agent to listen 5 times first before opening the door

Result:

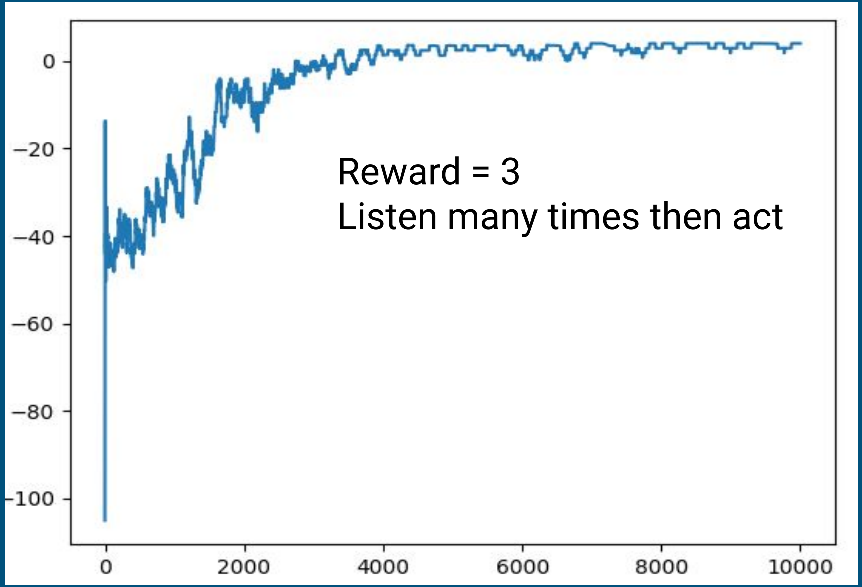
With this solution, DRQN does converge, but because we force to much listening, the agent also learns to listen multiple times, which leads to a smaller average reward

DQN vs. DRQN in Tiger with Extension

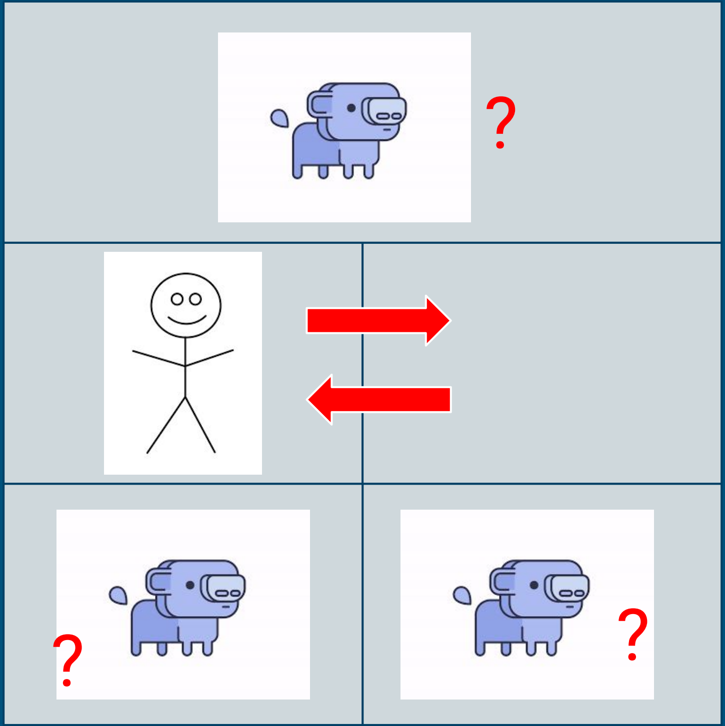
DQN reward: **Optimal Policy**



DRQN reward: **Sub-Optimal policy**



Test Environment #3: EasyWumpus



Action: Stay, Move, Shoot up, Shoot down

State: 2*3

Observation: Stench if 1 block away

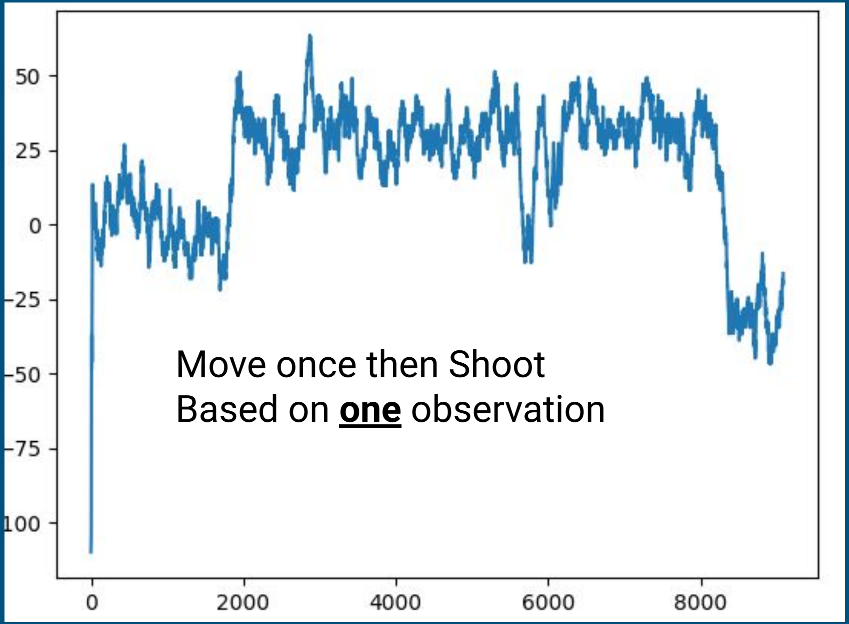
Reward:

1. Stay=Move=-1
2. Shoot right = 100
3. Shoot wrong = -100

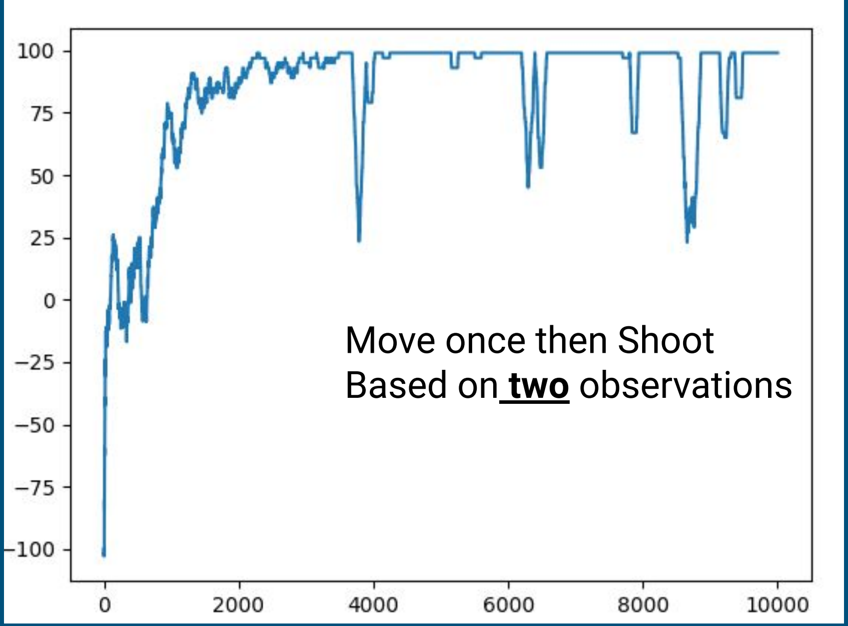
Observation Accuracy = 100%

DQN vs. DRQN in EasyWumpus

DQN Score: Non-Convergence. Around **<u>0</u>**



DRQN Score: Converge to Optimal **<u>99</u>**



Conclusion

1. Optimal Policy requires multiple observations: DRQN has absolute better performance than DQN. (Cartpole & Wumpus)

2. Optimal Policy requires single observation: DRQN has worse performance than DQN

3. Extending experience can improve DRQN when sequence is too short

Next Stage

1. Test DRQN and DQN on Standard Wumpus Environment

2. Proceed to milestone #3, deep variational reinforcement learning model. Compare whether generative models will have better performance than RNN based models.