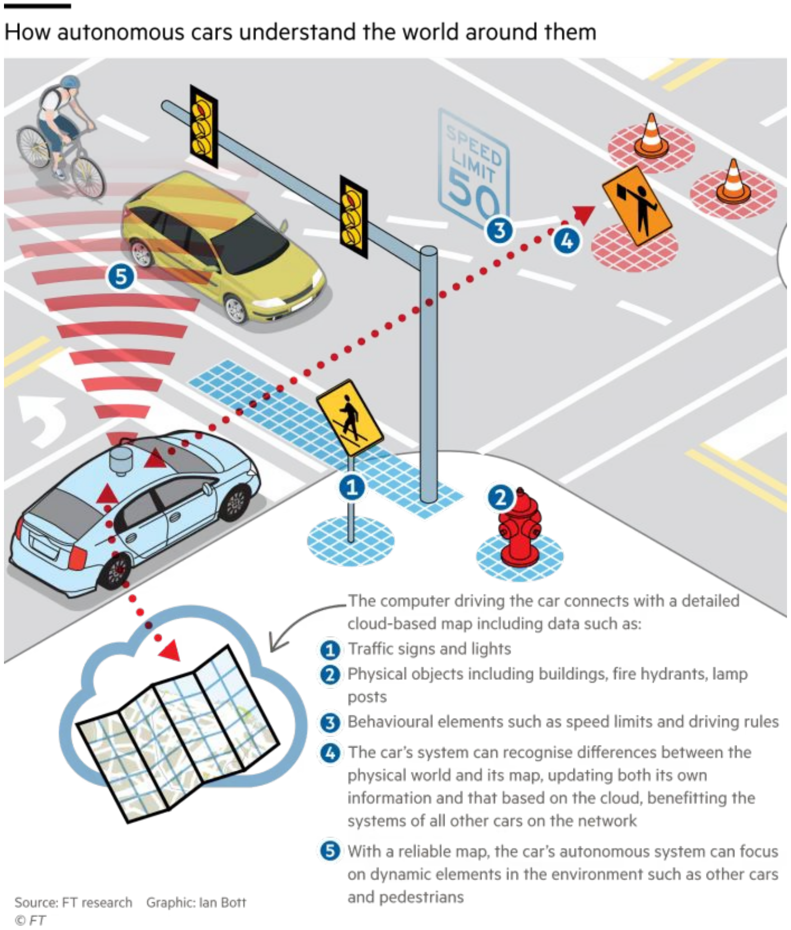
Q-Learning vs Actor-Critic Performance in 2D Highway Environment

Karan Baijal (kb553) and Elida Met-Hoxha (em744)

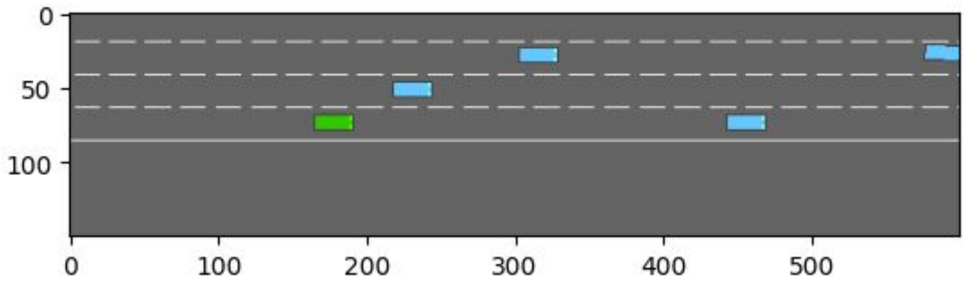
Motivation

- Autonomous Driving is complex - multi-agent interactions and complex kinematics.
- On versus off policy algorithms perform differently.
- More informed decision making, improving safety and reliability of self-driving systems.



Motivation - Highway Environment

- Part of Gymnasium package
- Goal: Drive at high speeds on highway without colliding on obstacles. Driving on right side is rewarded.
- Observation Space: 5*5 matrix
- Action Space: 5 actions
- Rewards: $$R(s,a) = a\frac{v - v_{\min}}{v_{\max} - v_{\min}} - b\,\mathrm{collision}$$



On versus Off Policy

<u>On-Policy</u>

- Learn from current policy's behavior- more stable
- Easier to tune
- Predictable
- Convergence

<u>Off-Policy</u>

- More sample efficient- replay buffer
- Can generalize better
- Can be unpredictable if diverges from target policy

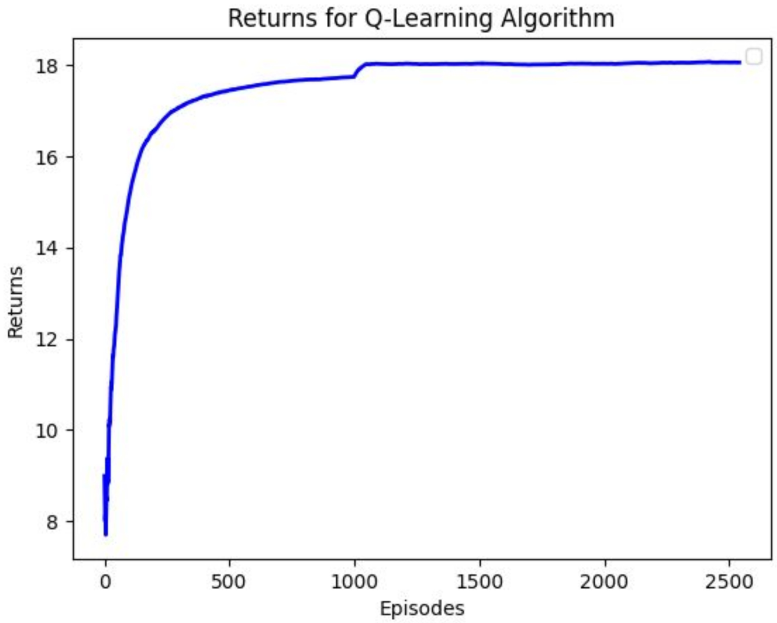
Hypothesis

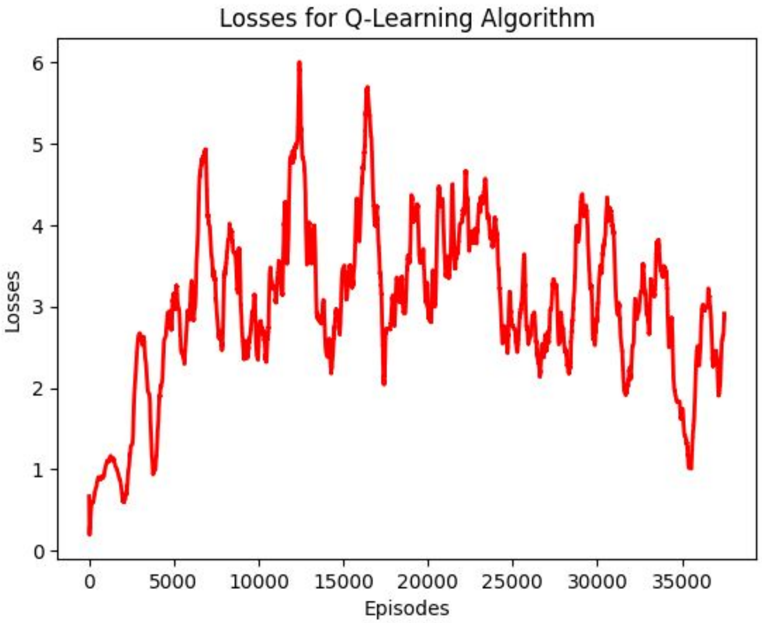
The off-policy algorithm, ie. Q-learning will be more flexible, thus performing better in the ‘highway-fast-v0’ gym environment.

Q-learning

- Q-learning for continuous state spaces
- Based off already existing DQN implementations
- PyTorch framework is used to define a fully connected neural network for the DQN model.
- Action Selection: Epsilon-greedy strategy
- Balance between exploration and exploitation
- Replay Buffer

Q-learning Evaluation

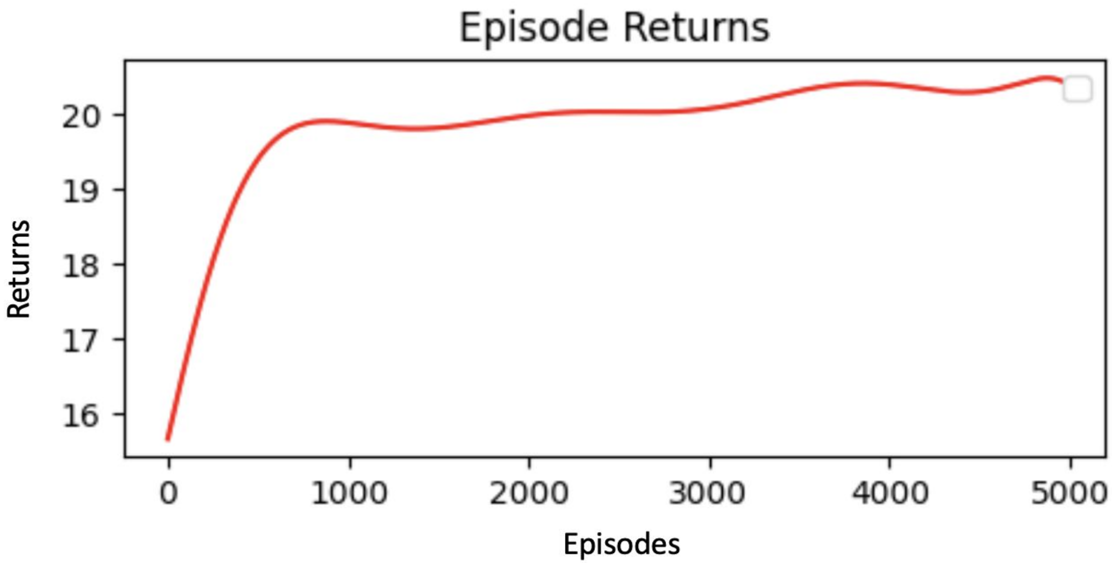




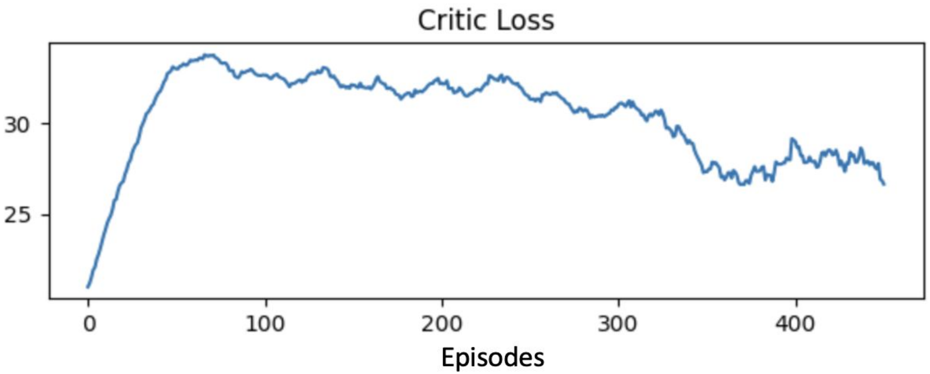
Actor-Critic

- Works on Continuous state spaces better
- Play-off between Actor and Critic Policy
- Code based on CS4756 Assignment
- Added entropy to encourage exploration

Actor-Critic Evaluation

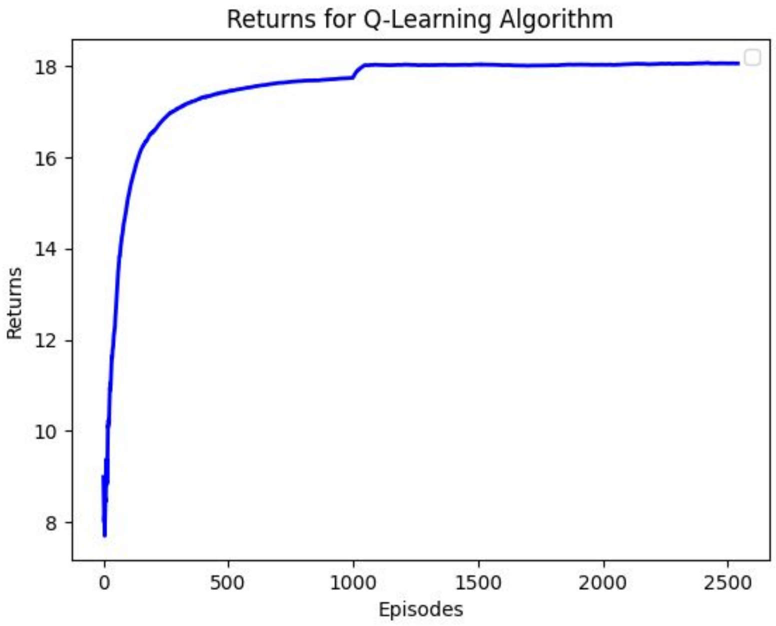


Actor-Critic Evaluation

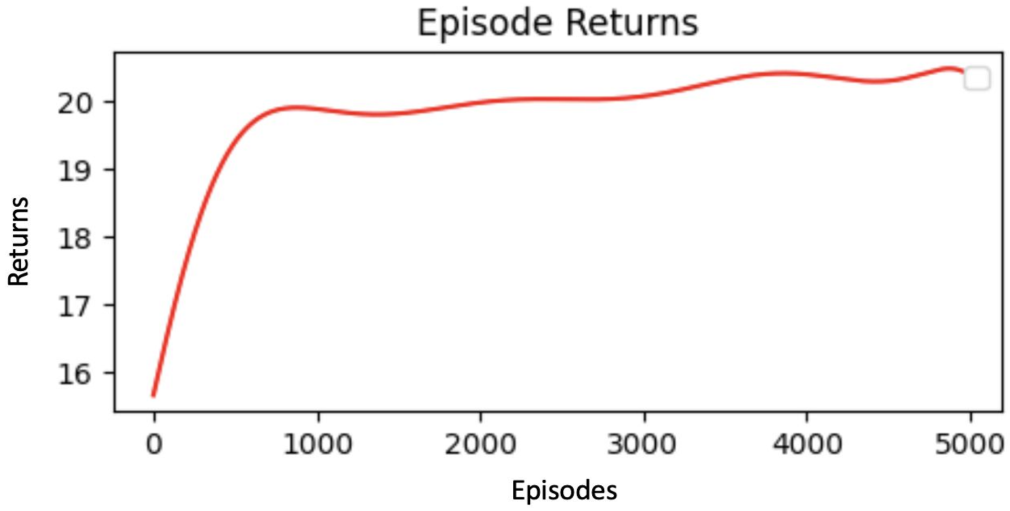


Highway Rendering

Comparison



Returns for Q-Learning Algorithm
18
16
14
12
10
8
Returns
0
500
1000
1500
2000
2500
Episodes



Episode Returns
20
19
18
17
16
Returns
0
1000
2000
3000
4000
5000
Episodes

Takeaways

- Actor-critic had higher rewards.
- A2C was less stable.
- Hypothesis invalidated - On-policy A2C seemed more effective than Off-policy DQN on Highway-v0 environment

Thank You!