

DELE CA2

Reinforcement

Learning

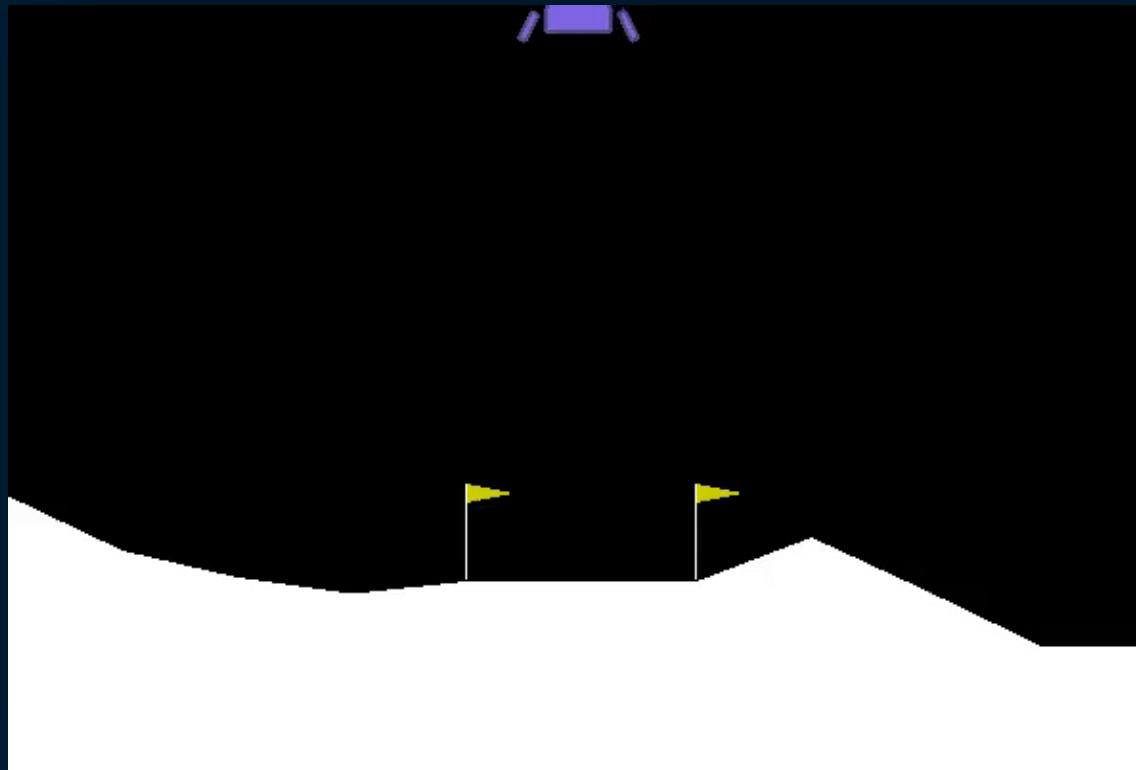
SHI TINGXIAO (P2033444)
Tan Yu Hoe (P2026309)

Exploratory Data Analysis

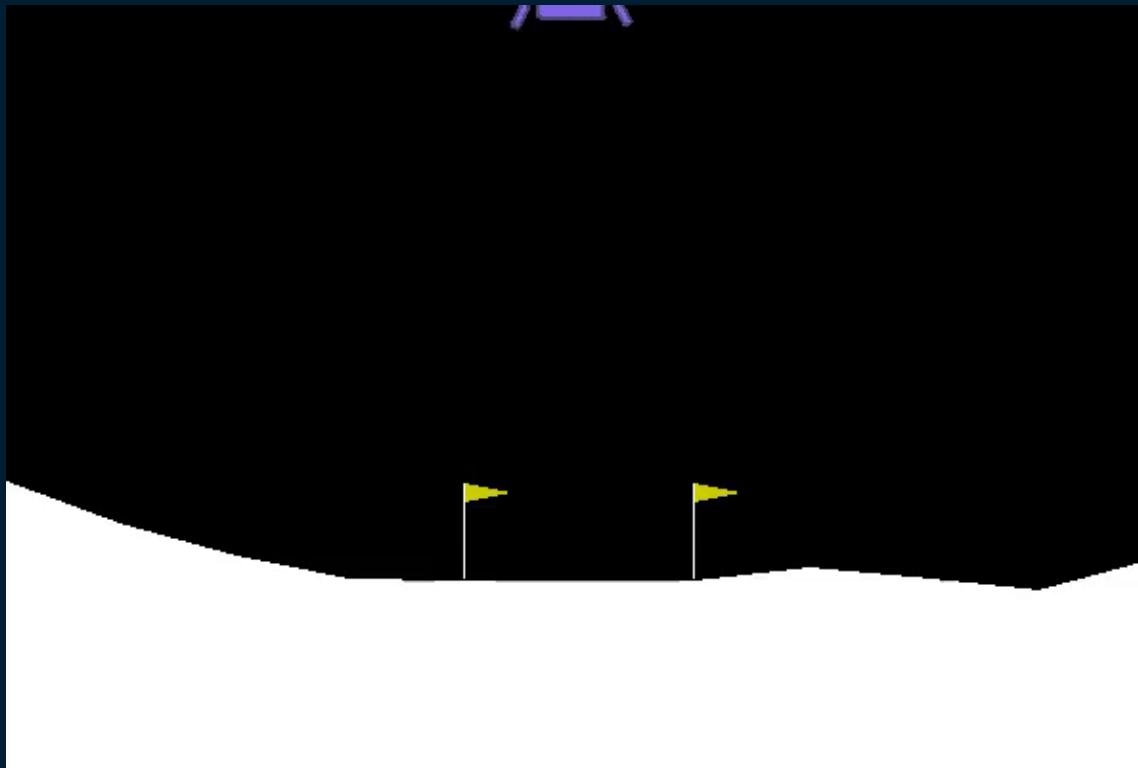
Observation Space

0	Lander's X Coordinate
1	Lander's Y Coordinate
2	X Linear Velocity
3	Y Linear Velocity
4	Angle
5	Angular Velocity
6	Left Leg is touching the ground
7	Right Leg is touching the ground

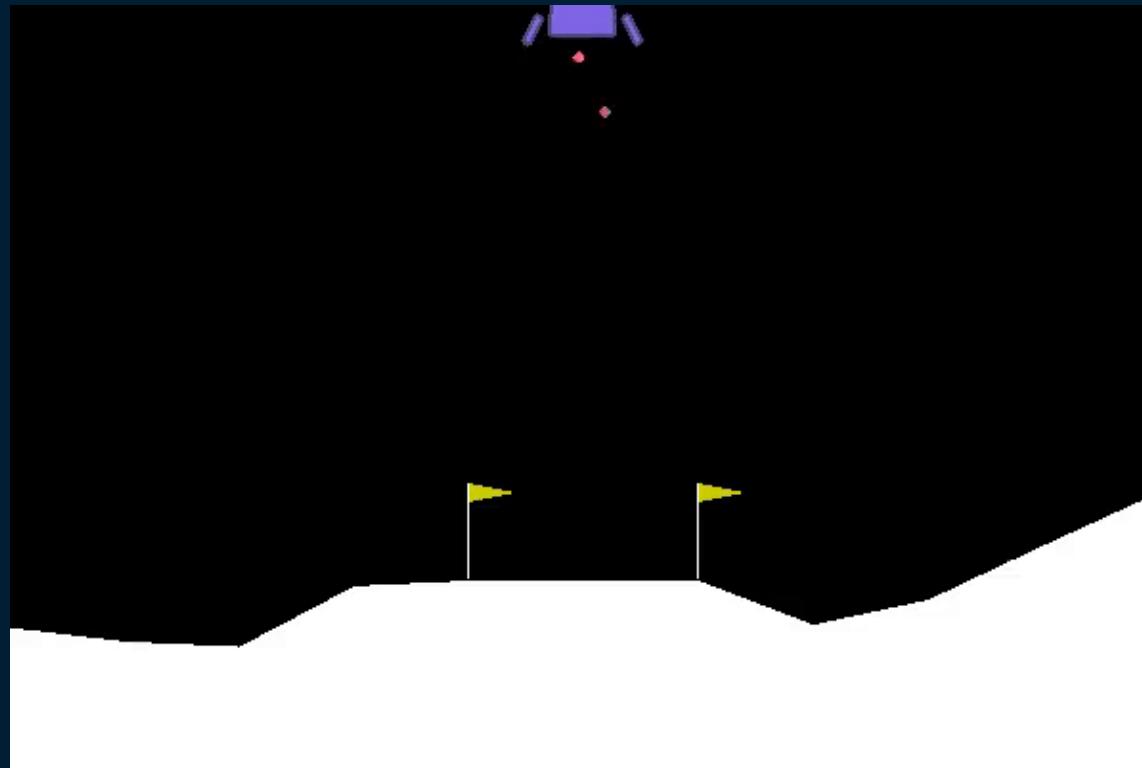
Action 0 Do Nothing



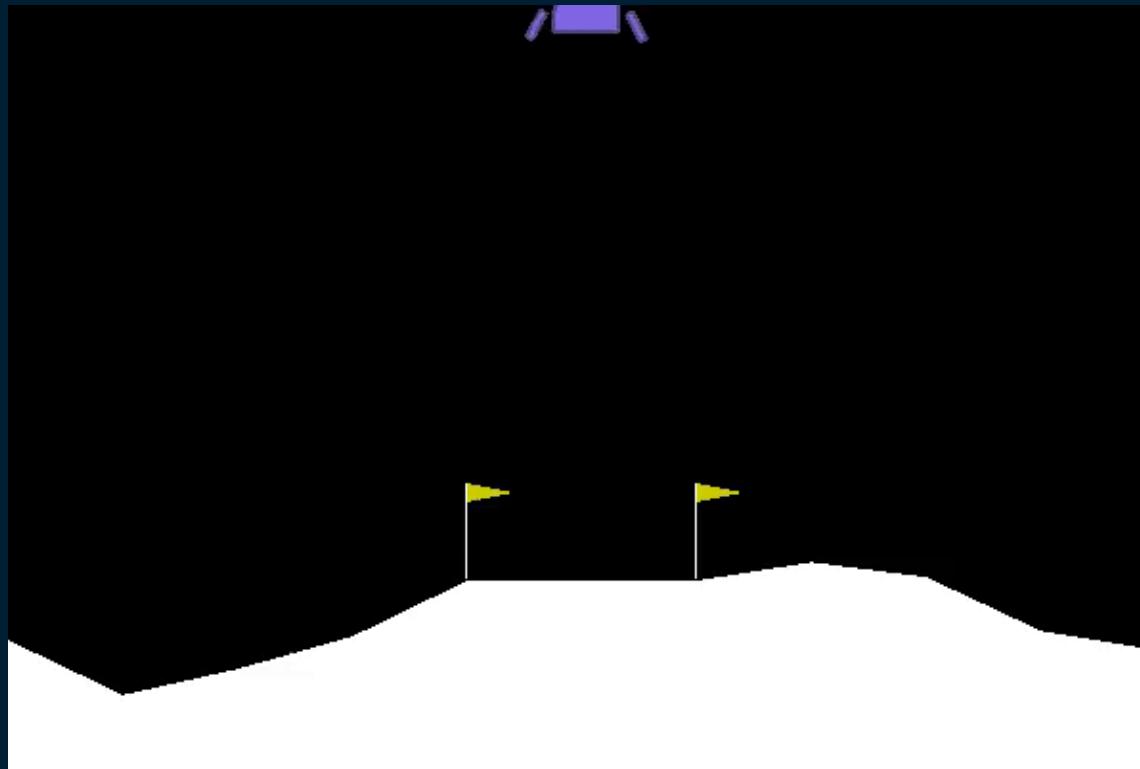
Action 1 Fire Right Engine



Action 2 Fire Main Engine



Action 3 Fire Left Engine



Random Actions



Reward Scenarios

Moving from top of screen to the landing pad and zero speed	+100-140
Moves away from landing pad	Lose Reward
Lander Crashes	-100
Lander comes to rest	+100
Each leg with ground contact	+10
Firing Main Engine	-0.3 per frame
Firing Side Engine	-0.03 per frame
Solving Scenario	≥ 200 points

Episode Termination

Lander Crashes

In contact with
the moon

Lander gets
outside of the
view port

X coordinate > 1

Lander is not
awake

Body doesn't move
and doesn't
collide with any
other body

Models we attempted

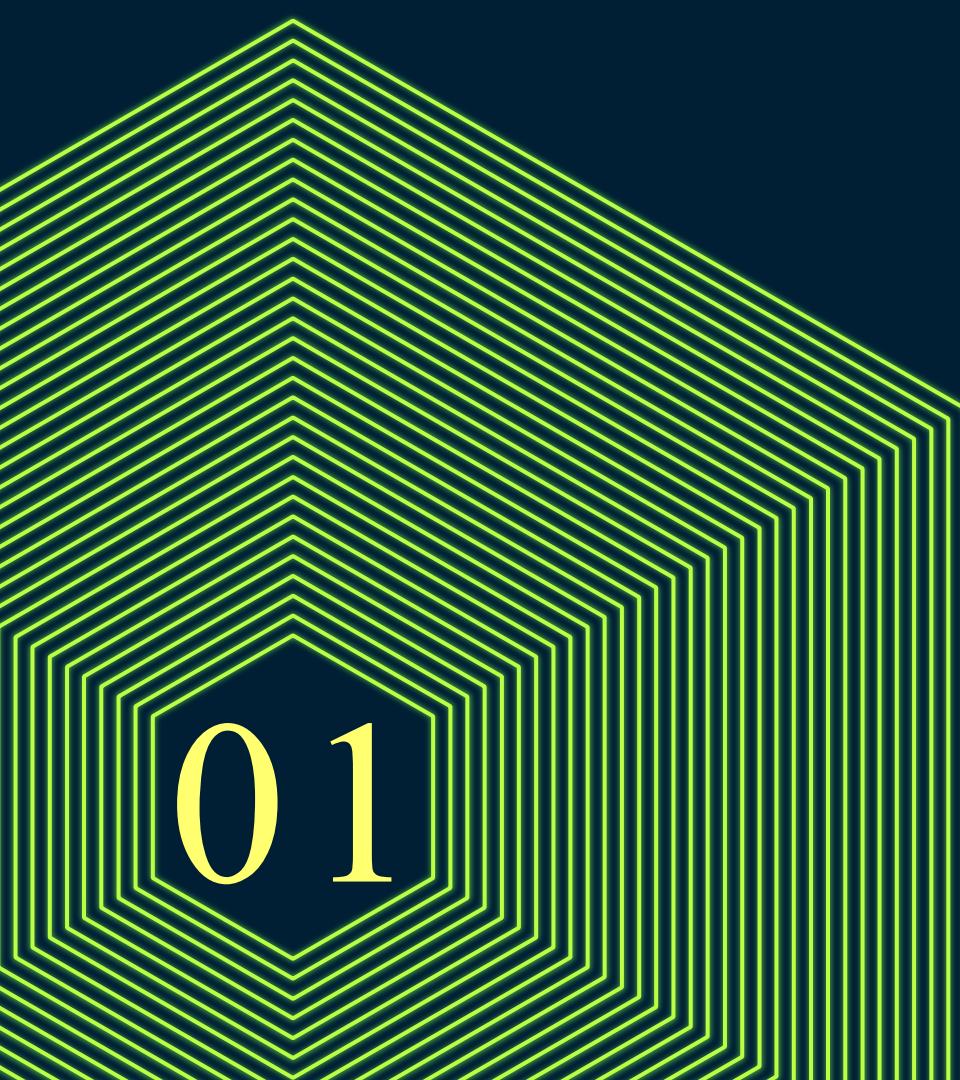
Models that we have attempted

Deep Q-
Network

Dueling Deep
Q-Network

Double
Dueling Deep
Q-Network

Actor
To
Critic



01

Deep Q Learning

Value Based Model



What is Deep Q Learning

- Deep Q Learning is an extended version of Q Learning
- Instead of approximating Q value using a function
- Deep Q Learning uses a neural network to approximate Q



Experience Replay and Replay Buffer

- Store past experiences
- Uses a random subset of these experiences to update the Deep Q-Network
- Contains a collection of experience as a Tuple (State, Action, Reward, New State)
- A buffer is added to the end so that it pushes the oldest experience out



Target Network

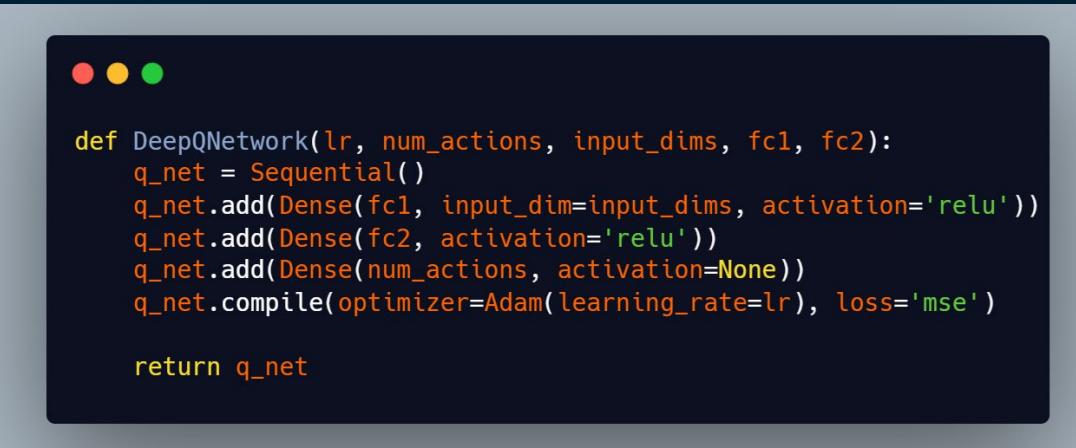
- When neural networks parameters are updated to $Q(s, a)$
- $Q(s', a')$ and other nearby states are also affected.
- Makes training unstable

- Target Network makes a copy of the Deep Q-Net to serve as stable target every N number of steps.
- The Target Network is used to create $Q(s', a')$



Deep Q Network Architecture

- Input layer taking in 8 states
- Two FC Hidden Layers with 256 nodes with ReLU Activation
- FC Output Layer with 4 nodes
- Compiled with Adam Optimizer and MSE Loss



```
● ● ●

def DeepQNetwork(lr, num_actions, input_dims, fc1, fc2):
    q_net = Sequential()
    q_net.add(Dense(fc1, input_dim=input_dims, activation='relu'))
    q_net.add(Dense(fc2, activation='relu'))
    q_net.add(Dense(num_actions, activation=None))
    q_net.compile(optimizer=Adam(learning_rate=lr), loss='mse')

    return q_net
```



DQN Policy

- Takes the highest Q-value of the current state and chooses the action with the high Q-value. (using `argmax()` operation)
- Explores the environment by randomly choosing actions for the first few episodes.

$$a \leftarrow \operatorname{argmax} Q(s, a_s)$$

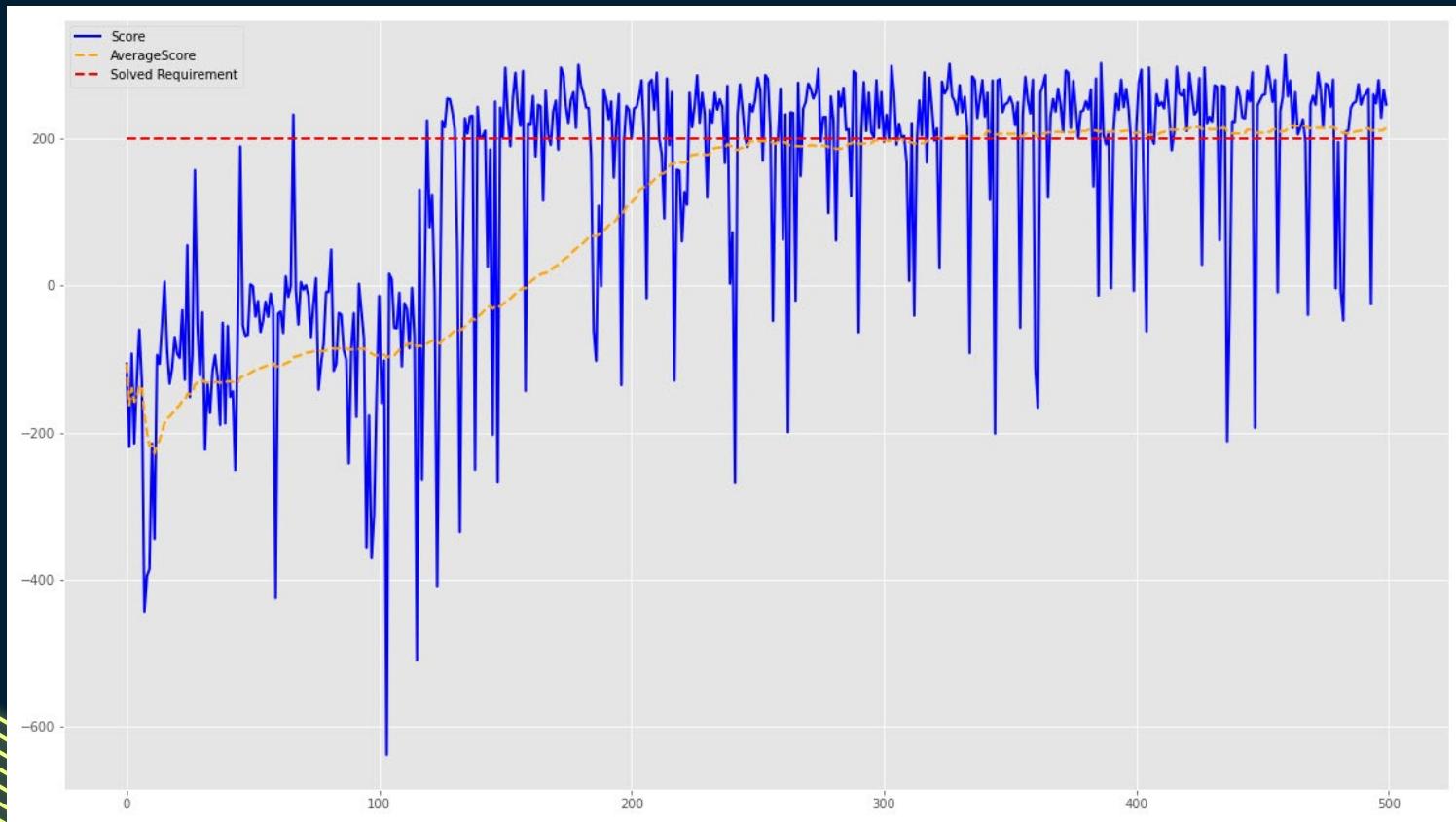
Training Function

For every State

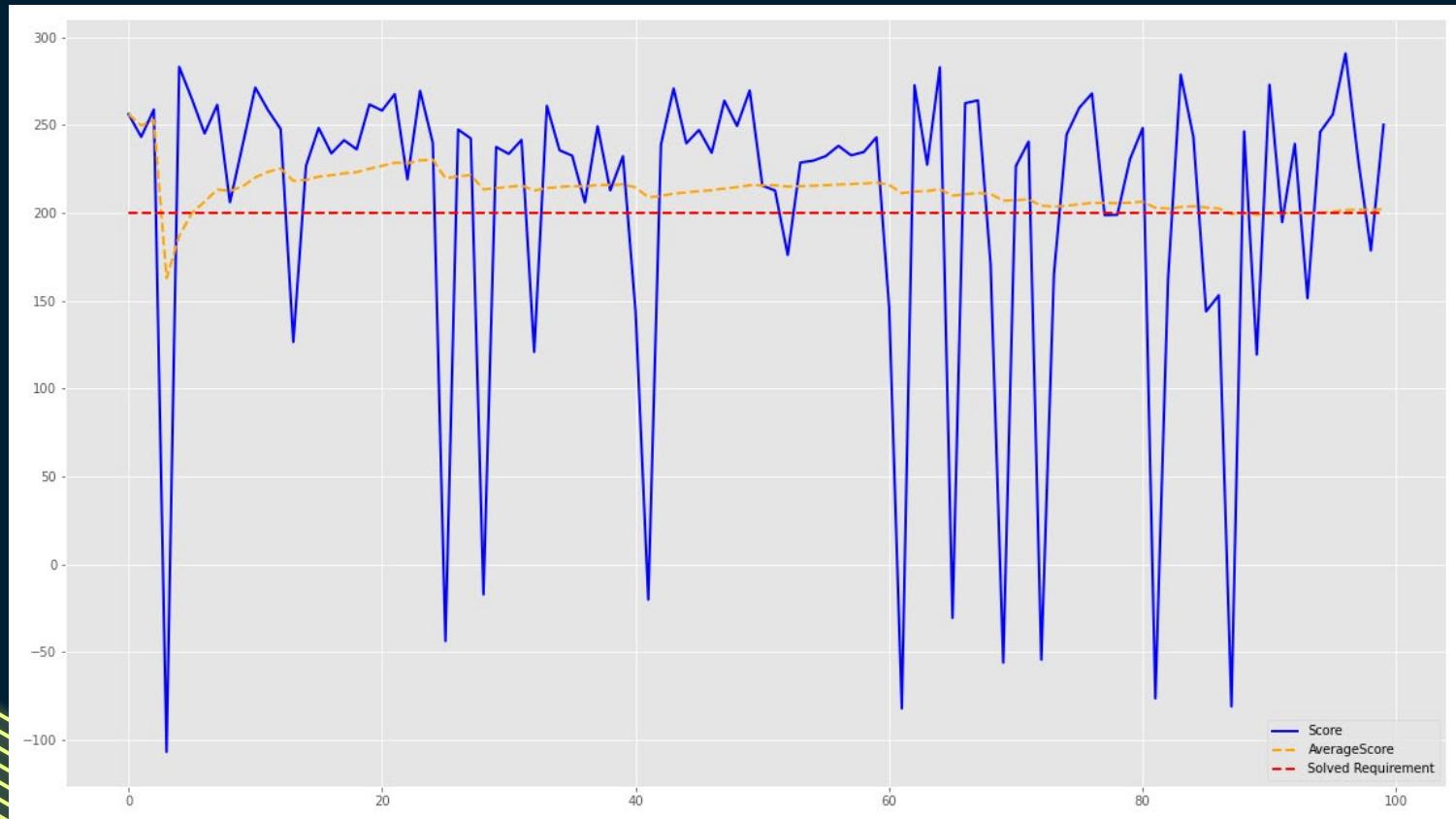
1. Copy QNet over to Target QNet (Every 120 steps)
2. Sample a batch of experience
3. Get Q value for the current state
4. Calculate Target Q Value using reward, discount factor, and the max Q value
5. Network performs backpropagation using state and target Q value.

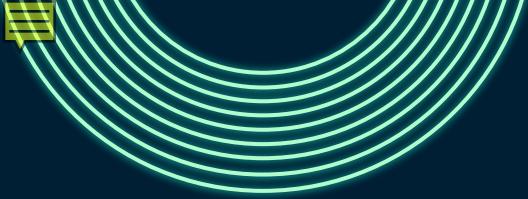
$$Q^{new}(s_t, a_t) \leftarrow \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \underbrace{\left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\substack{\text{estimate of optimal future value}}} - \underbrace{Q(s_t, a_t)}_{\text{old value}} \right)}_{\text{temporal difference}} \\ \text{new value (temporal difference target)}$$

DQN Train 500 Episodes

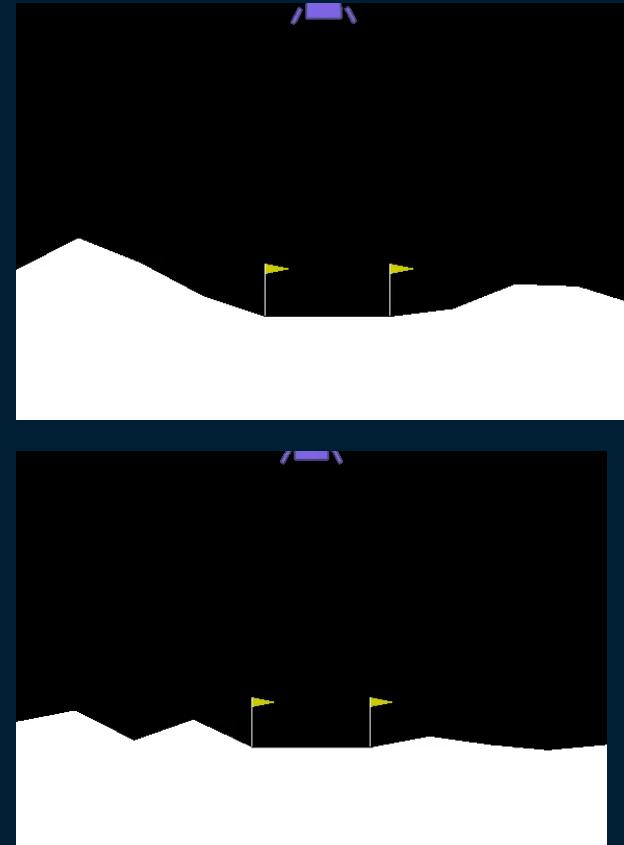
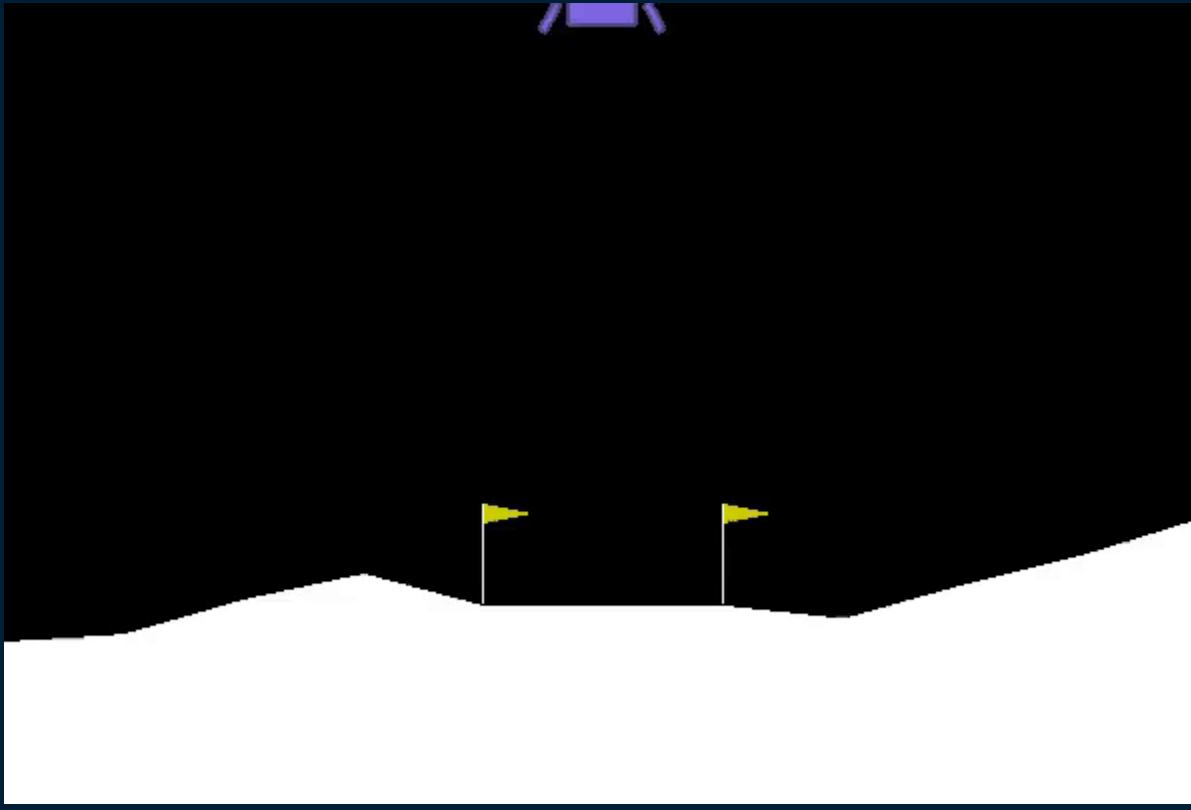


DQN Test 100 episodes



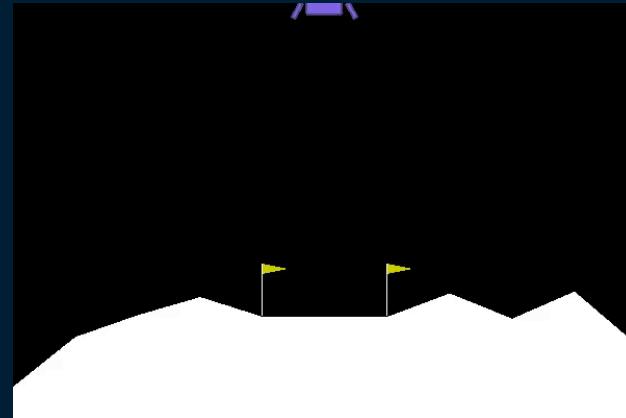
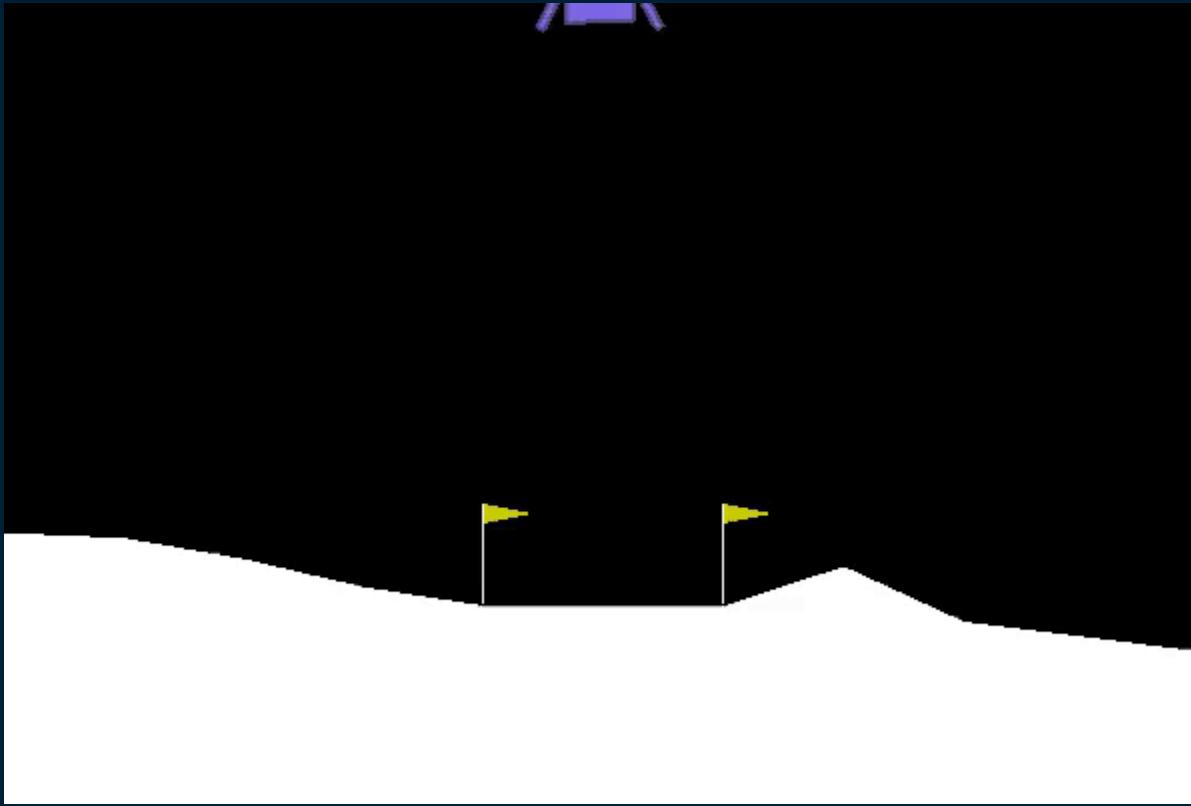


DQN Video Showcase





DQN Video Showcase





The problem of Single DQN

Temporal Difference learning makes DQN overestimate action-values.

- 1) TD target utilised maximization, TD target is bigger than the real action-value

```
for idx in range(done_batch.shape[0]):  
    target_q_val = reward_batch[idx]  
    if not done_batch[idx]:  
        target_q_val += self.discount_factor*q_max_next[idx]
```

- 1) Since TD target is overestimated, bootstrapping propagates the overestimation
- 2) Since each overestimation is bootstrapped, once DQN overestimates, it will be a vicious circle of overestimation.



02

Dueling Deep Q Network

Value Based Model



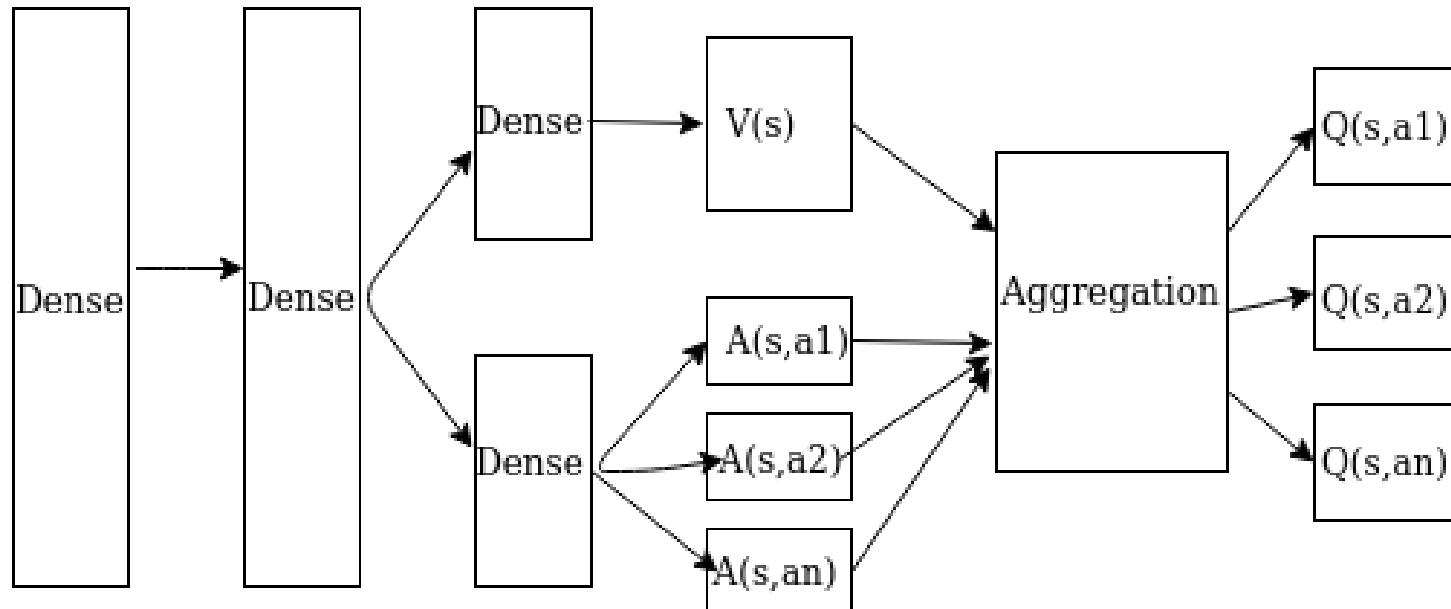
What is Dueling Deep-Qnetwork

- The model is represented by the equation below,

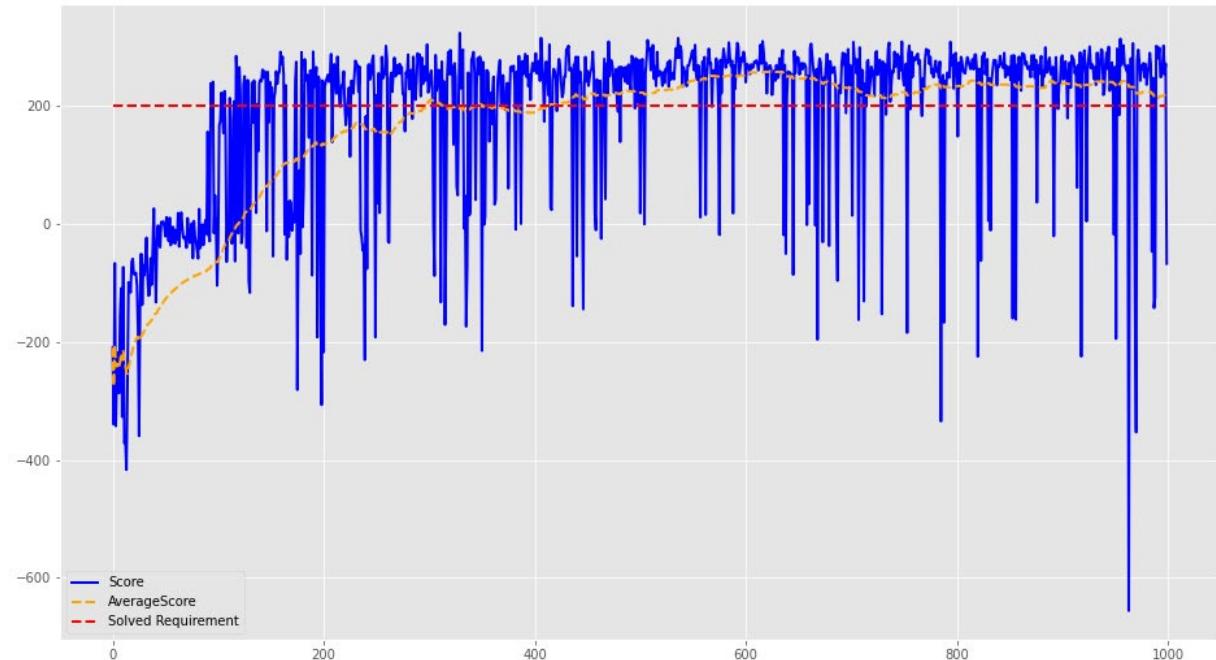
$$Q = V + A - \text{mean of all } A$$

- The model consists of two estimators
 - State value function V
 - State dependent action advantage function A
- These two estimators are combined to generate the action-value function
- The benefit is that the state-value function is quicker and easier to learn → Faster Convergence

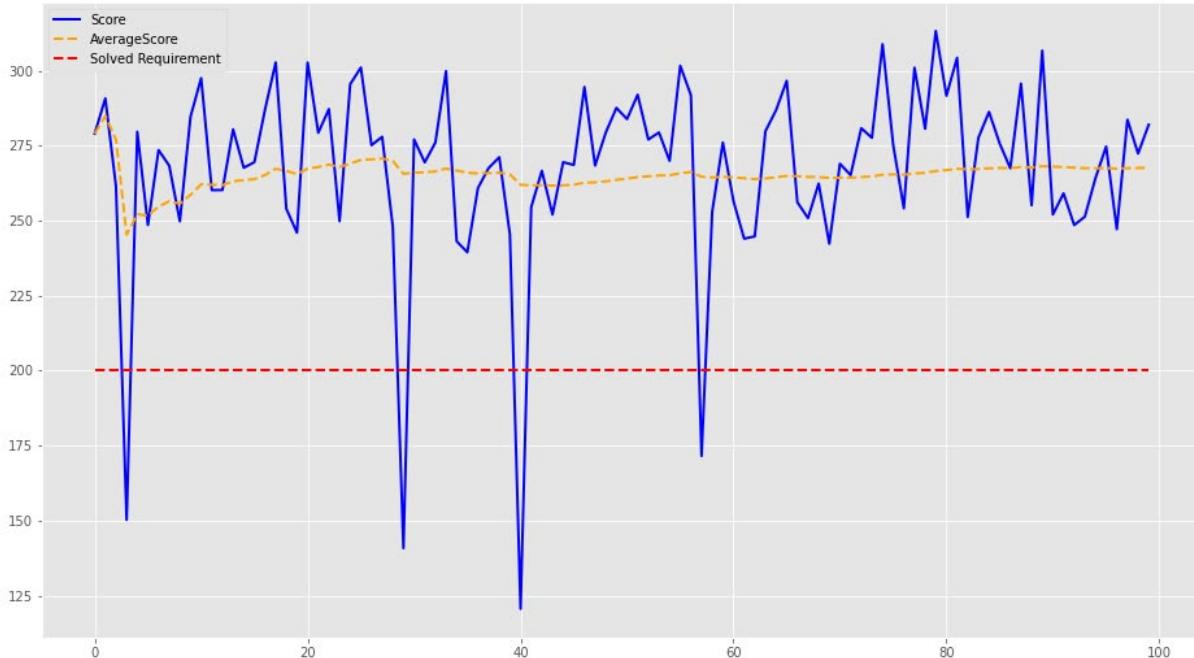
What is Dueling Deep-Qnetwork



DDQN Train 1000 Episodes

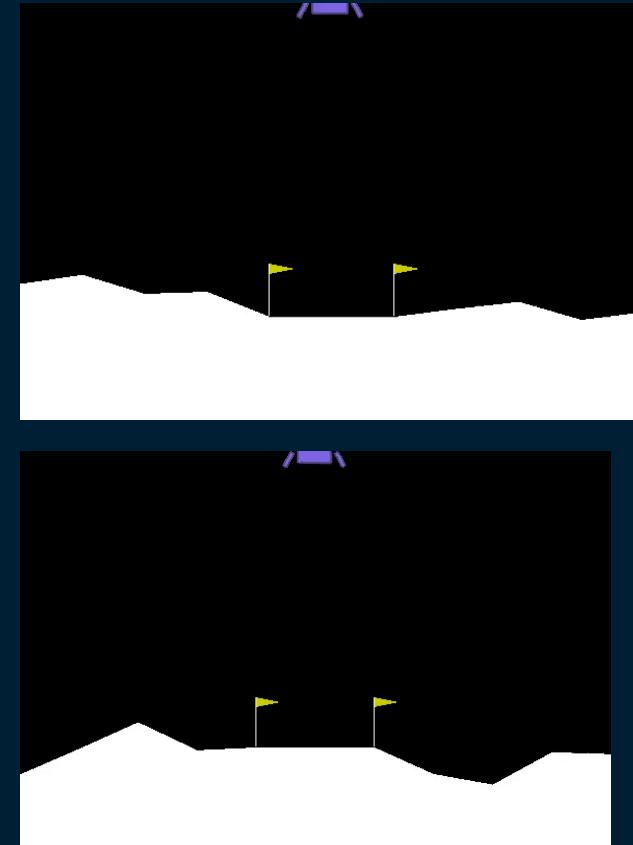
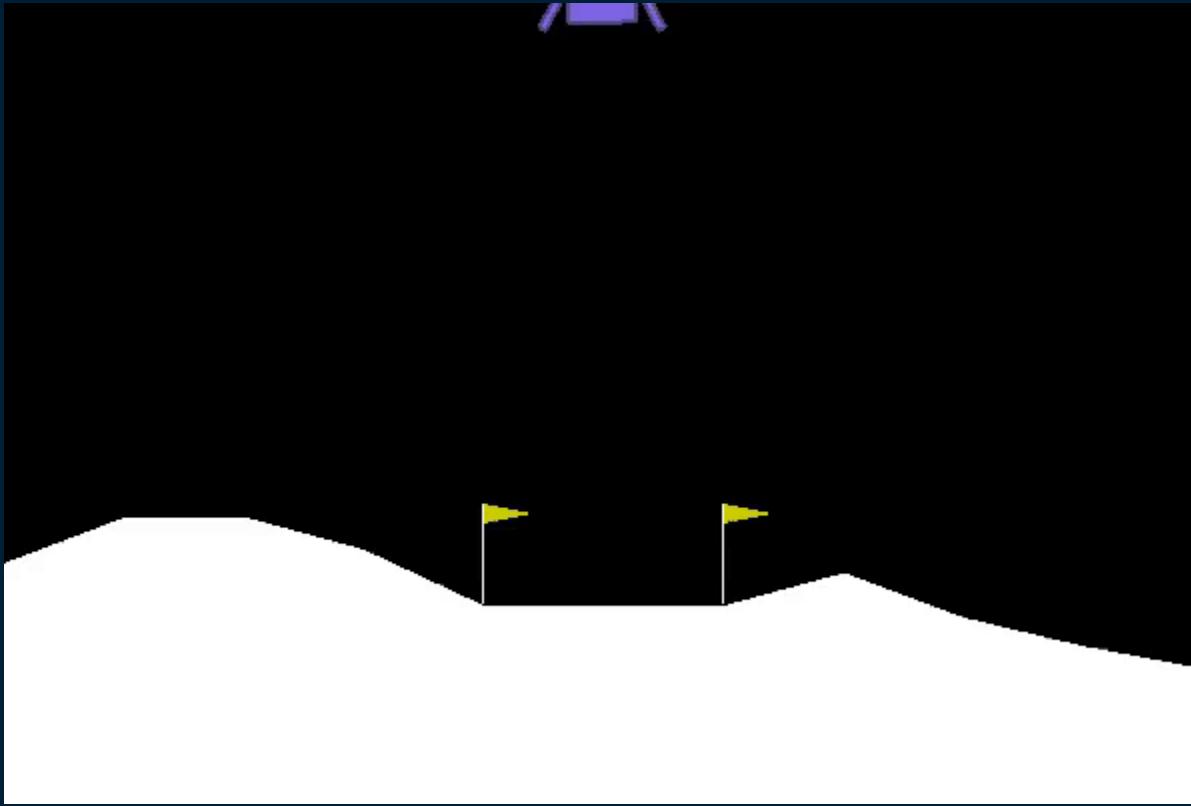


DDQN Test 100 episodes



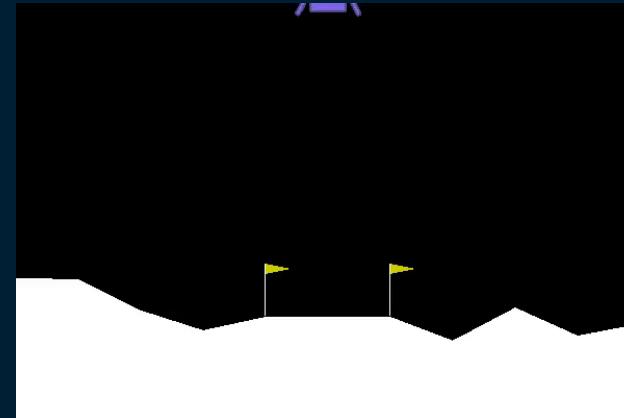
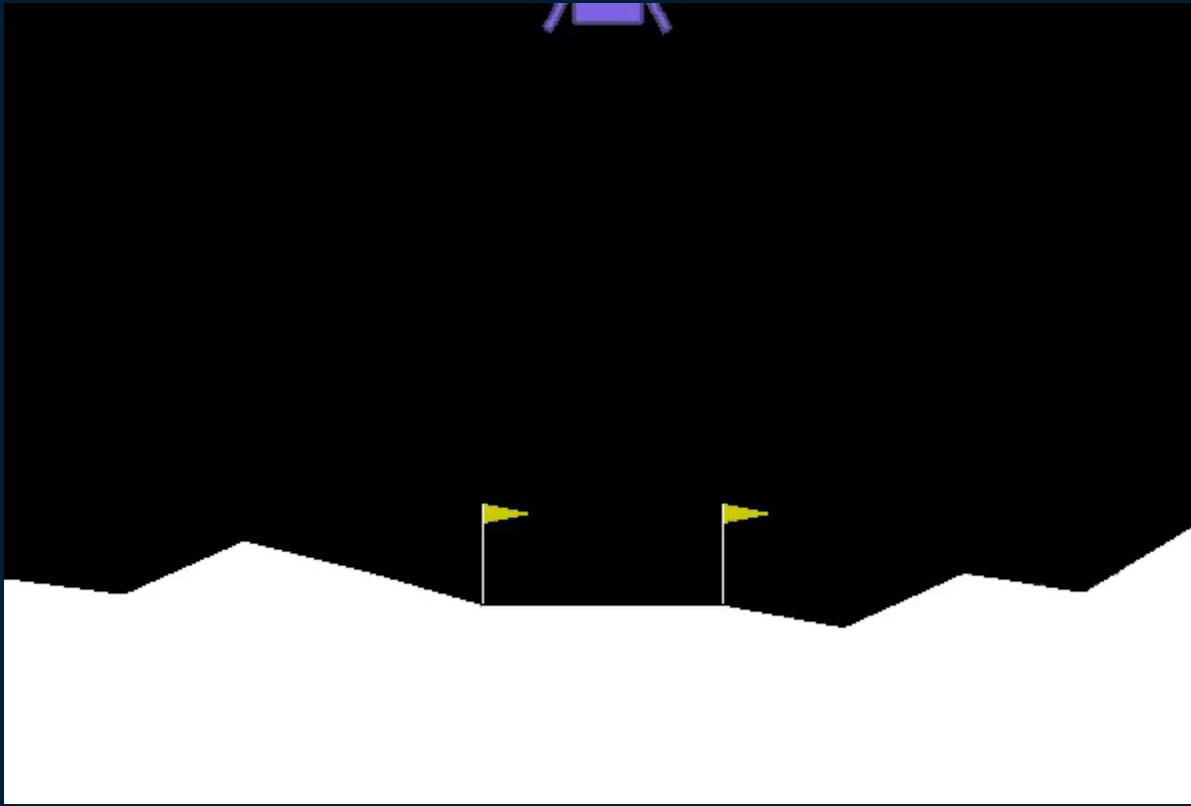


DDQN Video Showcase





DDQN Video Showcase





Cross Comparison DQN vs DDQN



Deep Q-Network

Dueling Deep Q-Network



03

Double Dueling Q-Network

Value Based model



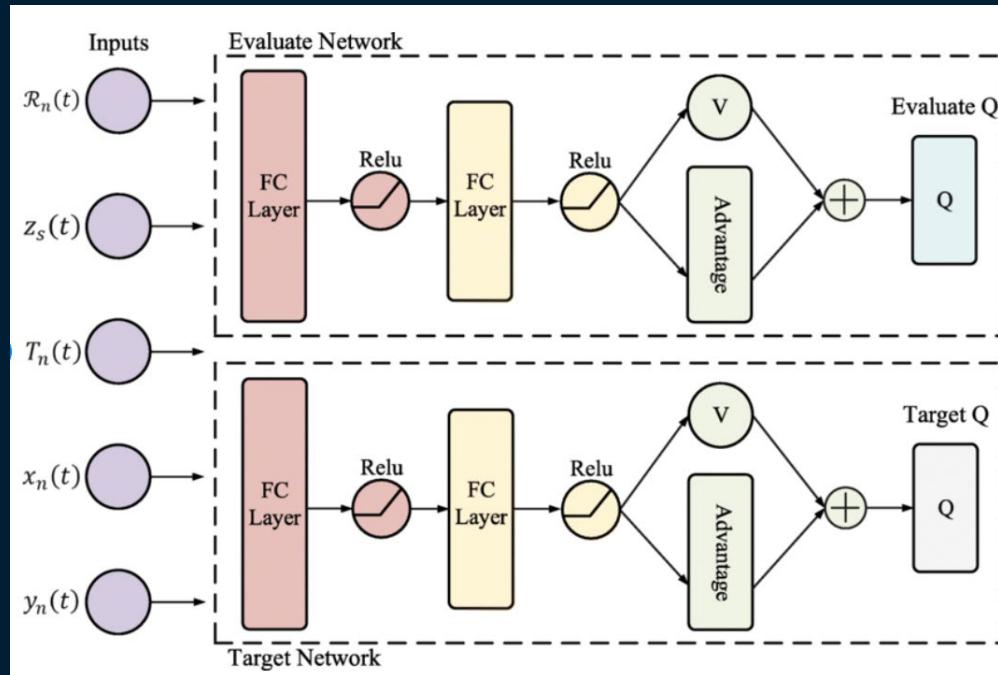
What is Double Dueling Deep Q Learning

- The double dueling DQN is pretty similar to the Dueling DQN, there are two consisting estimator - the state value function, and the state dependent action advantage function.

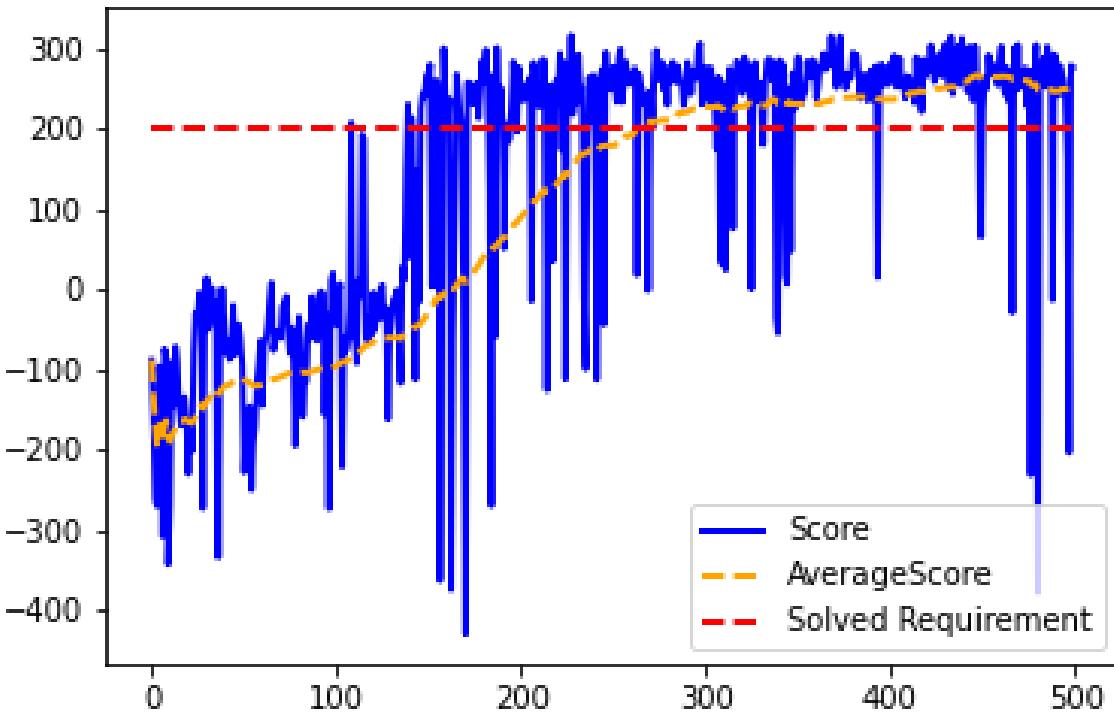
$$Q_1(S, A) \leftarrow Q_q(S, A) + \alpha(R + \gamma Q_2(S', \text{argmax}_a Q_1(S, a)) - Q_1(S, A))$$

- However there is an issue with single DQNs. The first time an agent receives a positive reward, it might continue making the same set action. A solution to solve this problem is to use two action value functions, in other words, two lookup tables. The agent uses one to update the other Q-value, providing an unbiased estimate of $Q(s, a)$.

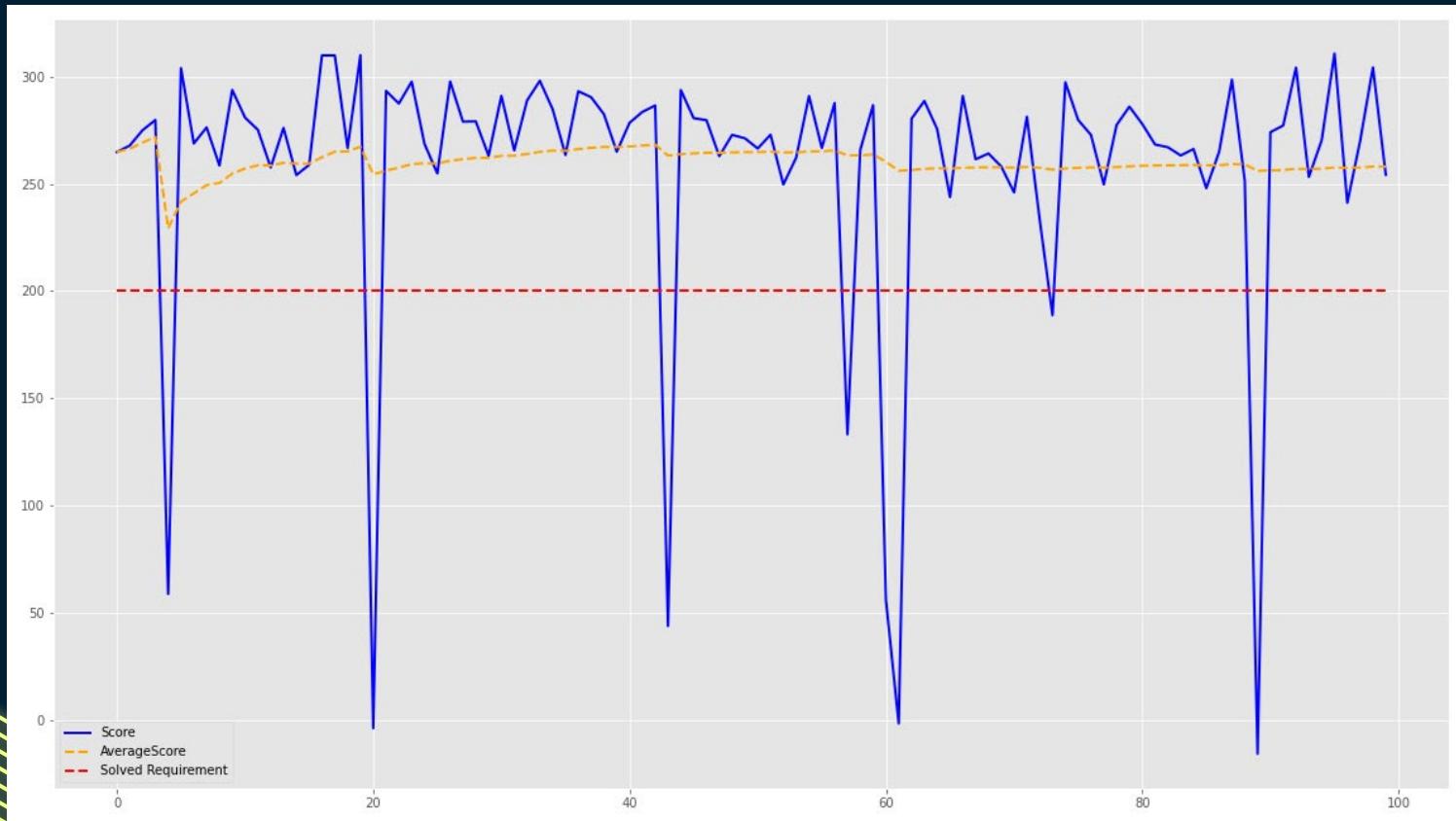
What is Double Dueling Deep Q Learning

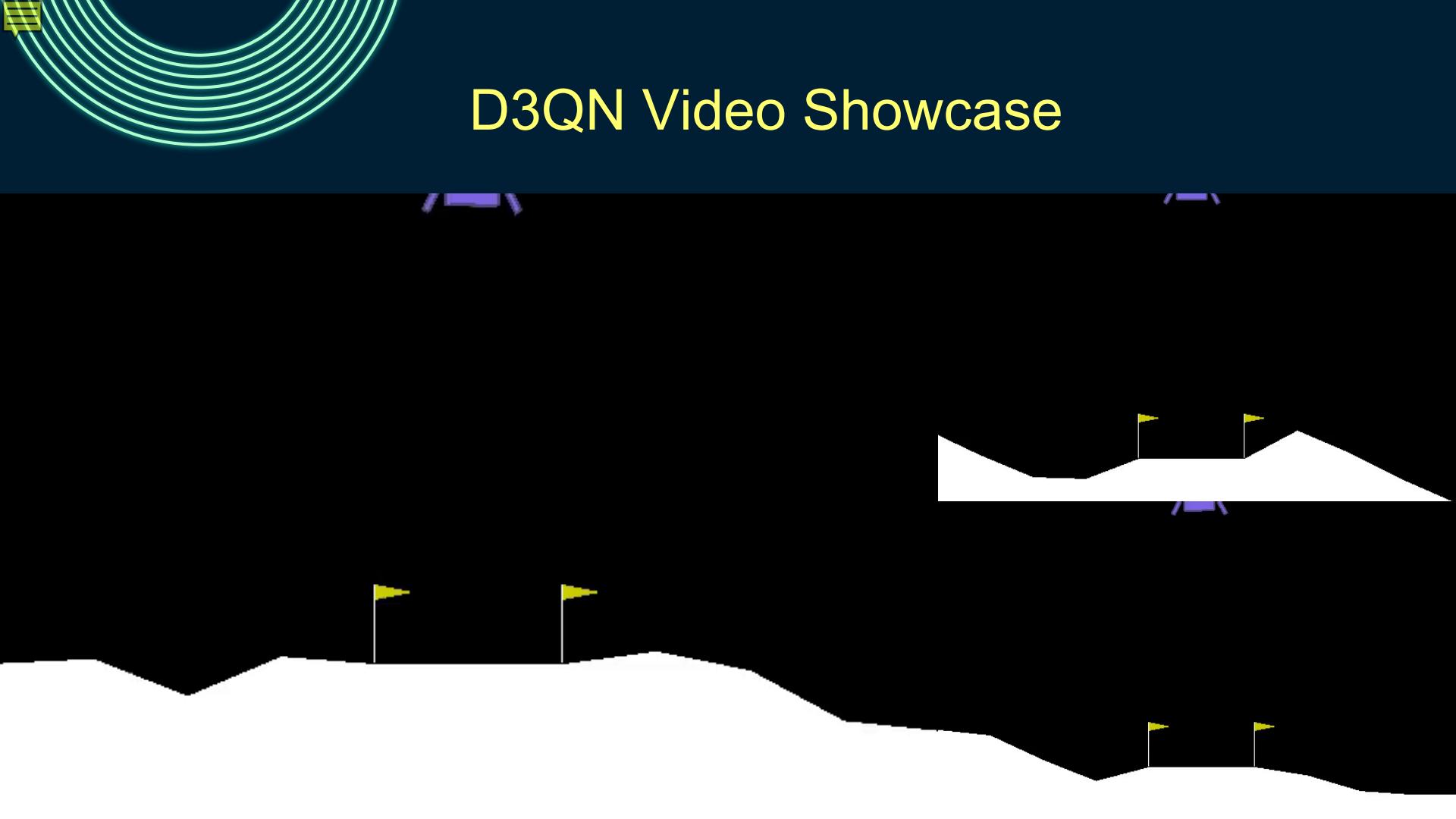


D3QN Train 500 episodes



D3QN Test 100 episodes

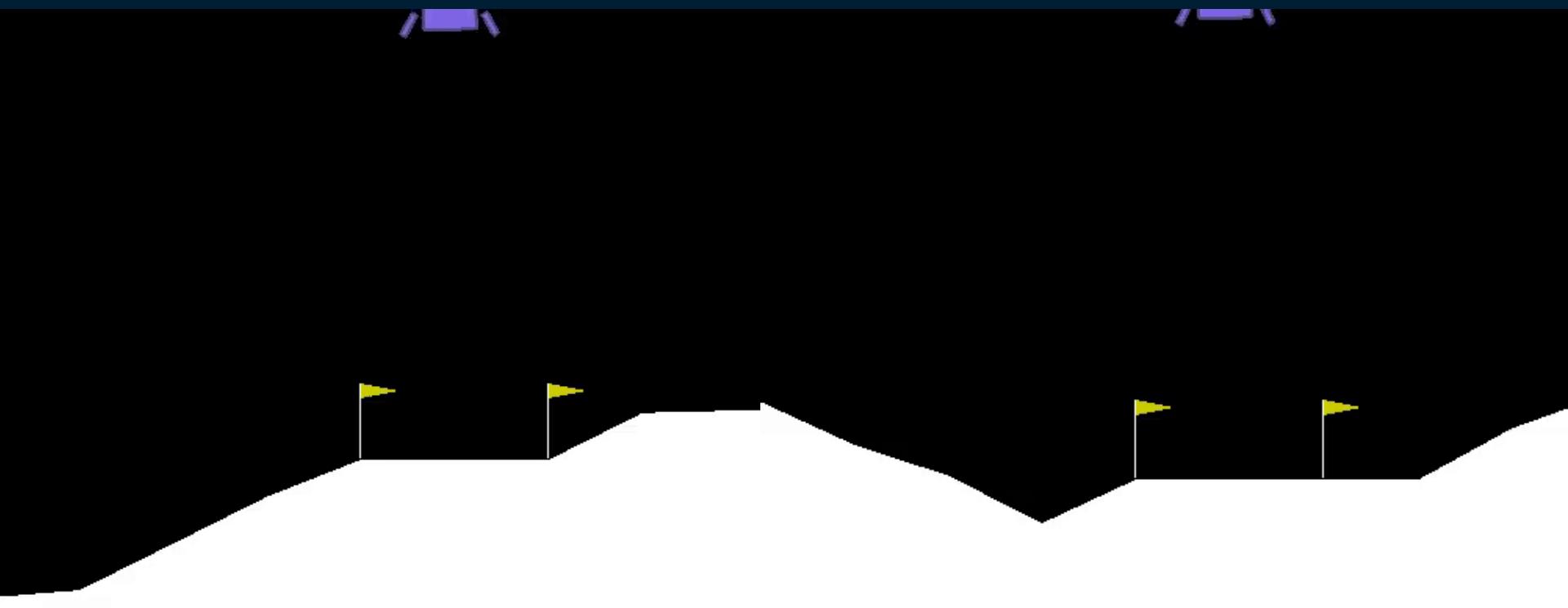




D3QN Video Showcase



D3QN Video Showcase

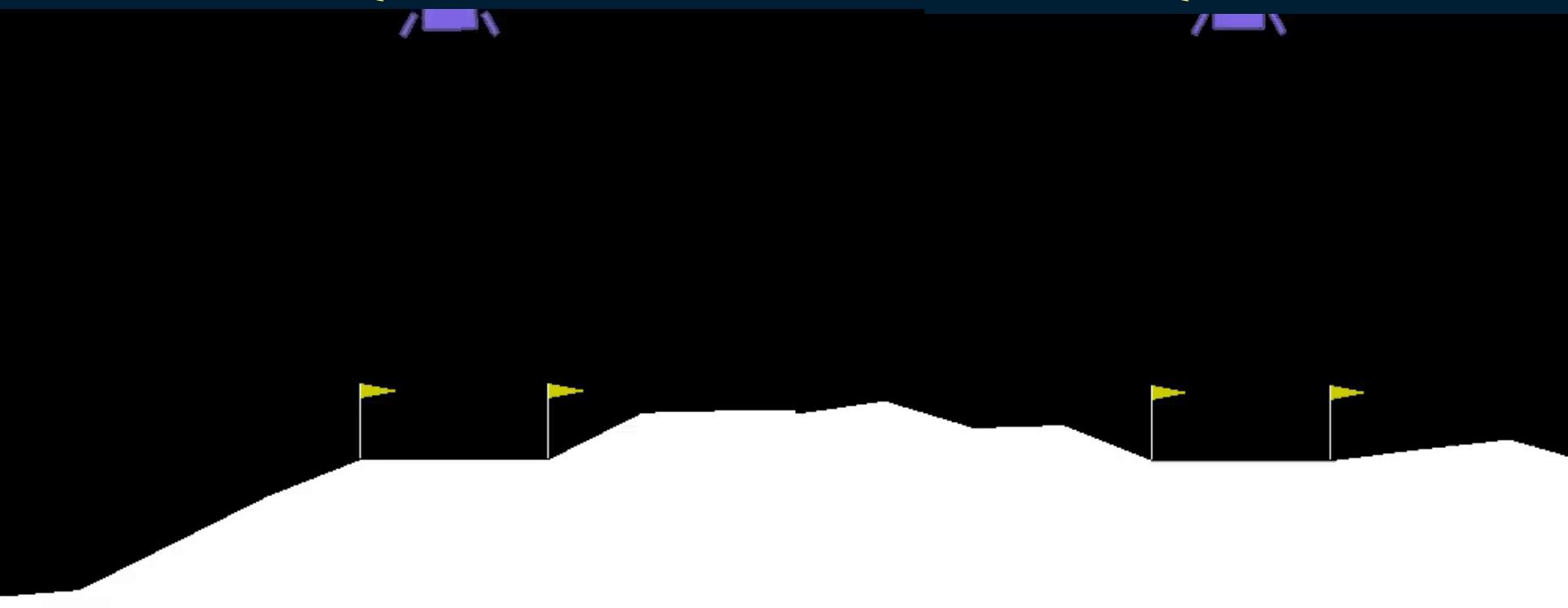




Cross Comparison

D3QN

DDQN





04

Actor to Critic

Value and Policy based model



What is Actor to Critic

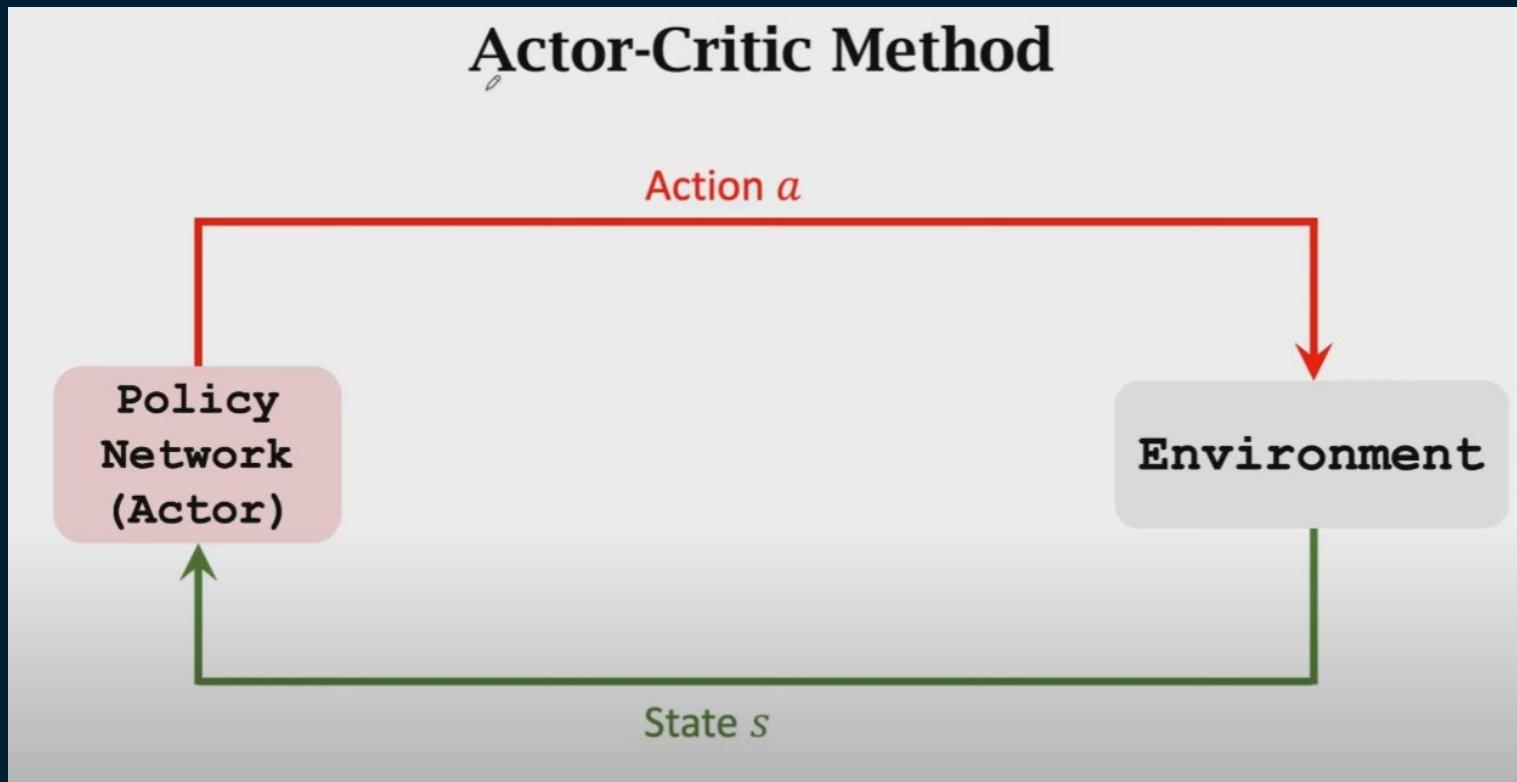
Actor critic network is a combination of policy-based method and value based method

Actor is a policy network, it actively trying to
Search for the best policy in the action table given state s
Policy(s) = argmax a (Q[s,a])

Critic is a value network, it rates the action, and help actor learn

Actor's Role

Actor-Critic Method





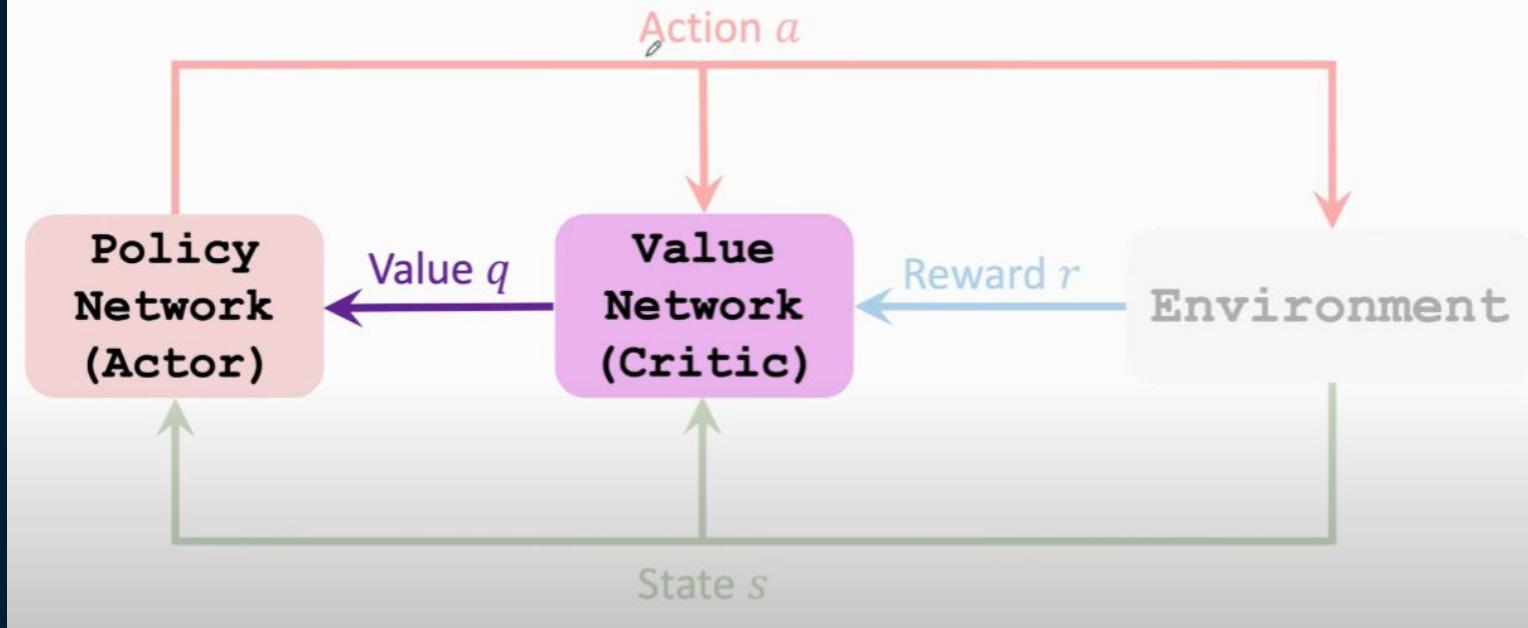
Actor network

With the input of 8 states, the actor network would choose which action to take, the dense layers using relu activation while the last dense uses softmax.

```
def build_model_actor(lr = 0.001, size = [128,128,64]):  
    # input the state and output the action  
    model = Sequential()  
    # dense layers 1  
    model.add(Dense(size[0], input_shape = (8,), activation = 'relu'))  
    # dense layers 2  
    model.add(Dense(size[1], activation = 'relu'))  
    # dense layers 3  
    model.add(Dense(size[2], activation = 'relu'))  
    # choose an action to take  
    model.add(Dense(4, activation = 'softmax'))  
    adam = optimizers.Adam(lr=lr, beta_1=0.9, beta_2=0.999)  
    model.compile(loss = 'categorical_crossentropy', optimizer = adam)  
    return model
```

Critic's Role

Actor-Critic Method





Critic Network

With the input of 8 states, the critic network will produce a

```
def build_model_critic( lr = 0.001, size = [128,128,64]):  
    # input the state and output the rating  
    model = Sequential()  
    # dense layer 1  
    model.add(Dense(size[0], input_shape = (8,), activation = 'relu'))  
    # dense layer 2  
    model.add(Dense(size[1], activation = 'relu'))  
    # dense layer 3  
    model.add(Dense(size[2], activation = 'relu'))  
    # output layer  
    model.add(Dense(1, activation = 'linear'))  
    # compile the model with adam optimizer and 'MSE loss'  
    adam = optimizers.Adam(lr=lr, beta_1=0.9, beta_2=0.999)  
    model.compile(loss = 'mse', optimizer = adam)  
  
    return model
```

Run Episodes

When the state is not done(either success or crash) or action is less than 1000 (to give model time to react to the env) the episode will still be running.

The memory stores the continuous state, reward and action.

```
def run_episode(env, actor, render = False):
    # create the memory list
    memory = []
    # reset the environment
    state = env.reset()
    # get the first action
    episode_reward = 0

    cnt = 0

    done = False

    while not done and cnt < 1000:
        cnt += 1
        if render:
            env.render()
        # predict the action
        action = decide_action(actor, state)
        observation, reward, done, _ = env.step(action)
        episode_reward += reward
        state_new = observation
        memory.append((state, action, reward, state_new, done))
        state = state_new

    return(memory, episode_reward)
```

Training Process

- 1) Actor observe the state $S(t)$
- 2) Actor randomly sample action $A(t)$ from the action space
- 3) Actor perform action $A(t)$ and observe the next state $S(t+1)$ and reward $R(t)$
- 4) update the weight of the critic network using temporal difference
- 5) update the weight of the actor network using the policy gradient

```
1 def decide_action(actor, state):    TINGXIAO, Today + a2c
2     # flatten the state
3     flat_state = np.reshape(state, [1,8])
4     # numpy to choose a action with the probability associated with each action
5     # generated by the actor
6     action = np.random.choice(4, 1, p = actor.predict(flat_state)[0])[0]
7     return(action)
8
9 def run_episode(env, actor, render = False):
10     # create the memory list
11     memory = []
12     # reset the environment
13     state = env.reset()
14     # get the first action
15     episode_reward = 0
16
17     cnt = 0
18     done = False
19
20     while not done and cnt < 1000:
21         cnt += 1
22         if render:
23             env.render()
24         # predict the action
25         action = decide_action(actor, state)
26         observation, reward, done, _ = env.step(action)
27         episode_reward += reward
28         state_new = observation
29         memory.append((state, action, reward, state_new, done))
30         state = state_new
31
32     return(memory, episode_reward)
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
639
640
641
642
643
644
645
646
647
648
649
649
650
651
652
653
654
655
656
657
658
659
659
660
661
662
663
664
665
666
667
668
669
669
670
671
672
673
674
675
676
677
678
679
679
680
681
682
683
684
685
686
687
687
688
689
689
690
691
692
693
694
695
696
697
697
698
699
699
700
701
702
703
704
705
706
707
708
709
709
710
711
712
713
714
715
716
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
737
738
739
739
740
741
742
743
744
745
745
746
747
747
748
749
749
750
751
752
753
754
755
755
756
757
757
758
759
759
760
761
762
763
764
764
765
766
766
767
767
768
768
769
769
770
771
771
772
772
773
773
774
774
775
775
776
776
777
777
778
778
779
779
780
780
781
781
782
782
783
783
784
784
785
785
786
786
787
787
788
788
789
789
790
790
791
791
792
792
793
793
794
794
795
795
796
796
797
797
798
798
799
799
800
800
801
801
802
802
803
803
804
804
805
805
806
806
807
807
808
808
809
809
810
810
811
811
812
812
813
813
814
814
815
815
816
816
817
817
818
818
819
819
820
820
821
821
822
822
823
823
824
824
825
825
826
826
827
827
828
828
829
829
830
830
831
831
832
832
833
833
834
834
835
835
836
836
837
837
838
838
839
839
840
840
841
841
842
842
843
843
844
844
845
845
846
846
847
847
848
848
849
849
850
850
851
851
852
852
853
853
854
854
855
855
856
856
857
857
858
858
859
859
860
860
861
861
862
862
863
863
864
864
865
865
866
866
867
867
868
868
869
869
870
870
871
871
872
872
873
873
874
874
875
875
876
876
877
877
878
878
879
879
880
880
881
881
882
882
883
883
884
884
885
885
886
886
887
887
888
888
889
889
890
890
891
891
892
892
893
893
894
894
895
895
896
896
897
897
898
898
899
899
900
900
901
901
902
902
903
903
904
904
905
905
906
906
907
907
908
908
909
909
910
910
911
911
912
912
913
913
914
914
915
915
916
916
917
917
918
918
919
919
920
920
921
921
922
922
923
923
924
924
925
925
926
926
927
927
928
928
929
929
930
930
931
931
932
932
933
933
934
934
935
935
936
936
937
937
938
938
939
939
940
940
941
941
942
942
943
943
944
944
945
945
946
946
947
947
948
948
949
949
950
950
951
951
952
952
953
953
954
954
955
955
956
956
957
957
958
958
959
959
960
960
961
961
962
962
963
963
964
964
965
965
966
966
967
967
968
968
969
969
970
970
971
971
972
972
973
973
974
974
975
975
976
976
977
977
978
978
979
979
980
980
981
981
982
982
983
983
984
984
985
985
986
986
987
987
988
988
989
989
990
990
991
991
992
992
993
993
994
994
995
995
996
996
997
997
998
998
999
999
1000
1000
1001
1001
1002
1002
1003
1003
1004
1004
1005
1005
1006
1006
1007
1007
1008
1008
1009
1009
1010
1010
1011
1011
1012
1012
1013
1013
1014
1014
1015
1015
1016
1016
1017
1017
1018
1018
1019
1019
1020
1020
1021
1021
1022
1022
1023
1023
1024
1024
1025
1025
1026
1026
1027
1027
1028
1028
1029
1029
1030
1030
1031
1031
1032
1032
1033
1033
1034
1034
1035
1035
1036
1036
1037
1037
1038
1038
1039
1039
1040
1040
1041
1041
1042
1042
1043
1043
1044
1044
1045
1045
1046
1046
1047
1047
1048
1048
1049
1049
1050
1050
1051
1051
1052
1052
1053
1053
1054
1054
1055
1055
1056
1056
1057
1057
1058
1058
1059
1059
1060
1060
1061
1061
1062
1062
1063
1063
1064
1064
1065
1065
1066
1066
1067
1067
1068
1068
1069
1069
1070
1070
1071
1071
1072
1072
1073
1073
1074
1074
1075
1075
1076
1076
1077
1077
1078
1078
1079
1079
1080
1080
1081
1081
1082
1082
1083
1083
1084
1084
1085
1085
1086
1086
1087
1087
1088
1088
1089
1089
1090
1090
1091
1091
1092
1092
1093
1093
1094
1094
1095
1095
1096
1096
1097
1097
1098
1098
1099
1099
1100
1100
1101
1101
1102
1102
1103
1103
1104
1104
1105
1105
1106
1106
1107
1107
1108
1108
1109
1109
1110
1110
1111
1111
1112
1112
1113
1113
1114
1114
1115
1115
1116
1116
1117
1117
1118
1118
1119
1119
1120
1120
1121
1121
1122
1122
1123
1123
1124
1124
1125
1125
1126
1126
1127
1127
1128
1128
1129
1129
1130
1130
1131
1131
1132
1132
1133
1133
1134
1134
1135
1135
1136
1136
1137
1137
1138
1138
1139
1139
1140
1140
1141
1141
1142
1142
1143
1143
1144
1144
1145
1145
1146
1146
1147
1147
1148
1148
1149
1149
1150
1150
1151
1151
1152
1152
1153
1153
1154
1154
1155
1155
1156
1156
1157
1157
1158
1158
1159
1159
1160
1160
1161
1161
1162
1162
1163
1163
1164
1164
1165
1165
1166
1166
1167
1167
1168
1168
1169
1169
1170
1170
1171
1171
1172
1172
1173
1173
1174
1174
1175
1175
1176
1176
1177
1177
1178
1178
1179
1179
1180
1180
1181
1181
1182
1182
1183
1183
1184
1184
1185
1185
1186
1186
1187
1187
1188
1188
1189
1189
1190
1190
1191
1191
1192
1192
1193
1193
1194
1194
1195
1195
1196
1196
1197
1197
1198
1198
1199
1199
1200
1200
1201
1201
1202
1202
1203
1203
1204
1204
1205
1205
1206
1206
1207
1207
1208
1208
1209
1209
1210
1210
1211
1211
1212
1212
1213
1213
1214
1214
1215
1215
1216
1216
1217
1217
1218
1218
1219
1219
1220
1220
1221
1221
1222
1222
1223
1223
1224
1224
1225
1225
1226
1226
1227
1227
1228
1228
1229
1229
1230
1230
1231
1231
1232
1232
1233
1233
1234
1234
1235
1235
1236
1236
1237
1237
1238
1238
1239
1239
1240
1240
1241
1241
1242
1242
1243
1243
1244
1244
1245
1245
1246
1246
1247
1247
1248
1248
1249
1249
1250
1250
1251
1251
1252
1252
1253
1253
1254
1254
1255
1255
1256
1256
1257
1257
1258
1258
1259
1259
1260
1260
1261
1261
1262
1262
1263
1263
1264
1264
1265
1265
1266
1266
1267
1267
1268
1268
1269
1269
1270
1270
1271
1271
1272
1272
1273
1273
1274
1274
1275
1275
1276
1276
1277
1277
1278
1278
1279
1279
1280
1280
1281
1281
1282
1282
1283
1283
1284
1284
1285
1285
1286
1286
1287
1287
1288
1288
1289
1289
1290
1290
1291
1291
1292
1292
1293
1293
1294
1294
1295
1295
1296
1296
1297
1297
1298
1298
1299
1299
1300
1300
1301
1301
1302
1302
1303
1303
1304
1304
1305
1305
1306
1306
1307
1307
1308
1308
1309
1309
1310
1310
1311
1311
1312
1312
1313
1313
1314
1314
1315
1315
1316
1316
1317
1317
1318
1318
1319
1319
1320
1320
1321
1321
1322
1322
1323
1323
1324
1324
1325
1325
1326
1326
1327
1327
1328
1328
1329
1329
1330
1330
1331
1331
1332
1332
1333
1333
1334
1334
1335
1335
1336
1336
1337
1337
1338
1338
1339
1339
1340
1340
1341
1341
1342
1342
1343
1343
1344
1344
1345
1345
1346
1346
1347
1347
1348
1348
1349
1349
1350
1350
1351
1351
1352
1352
1353
1353
1354
1354
1355
1355
1356
1356
1357
1357
1358
1358
1359
1359
1360
1360
1361
1361
1362
1362
1363
1363
1364
1364
1365
1365
1366
1366
1367
1367
1368
1368
1369
1369
1370
1370
1371
1371
1372
1372
1373
1373
1374
1374
1375
1375
1376
1376
1377
1377
1378
1378
1379
1379
1380
1380
1381
1381
1382
1382
1383
1383
1384
1384
1385
1385
1386
1386
1387
1387
1388
1388
1389
1389
1390
1390
1391
1391
1392
1392
1393
1393
1394
1394
1395
1395
1396
1396
1397
1397
1398
1398
1399
1399
1400
1400
1401
1401
1402
1402
1403
1403
1404
1404
1405
1405
1406
1406
1407
1407
1408
1408
1409
1409
1410
1410
1411
1411
1412
1412
1413
1413
1414
1414
1415
1415
1416
1416
1417
1417
1418
1418
1419
1419
1420
1420
1421
1421
1422
1422
1423
1423
1424
1424
1425
1425
1426
1426
1427
1427
1428
1428
1429
1429
1430
1430
1431
1431
1432
1432
1433
1433
1434
1434
1435
1435
1436
1436
1437
1437
1438
1438
1439
1439
1440
1440
1441
1441
1442
1442
1443
1443
1444
1444
1445
1445
1446
1446
1447
1447
1448
1448
1449
1449
1450
1450
1451
1451
1452
1452
1453
1453
1454
1454
1455
1455
1456
1456
1457
1457
1458
1458
1459
1459
1460
1460
1461
1461
1462
1462
1463
1463
1464
1464
1465
1465
1466
1466
1467
1467
1468
1468
1469
1469
1470
1470
1471
1471
1472
1472
1473
1473
1474
1474
1475
1475
1476
1476
1477
1477
1478
1478
1479
1479
1480
1480
1481
1481
1482
1482
1483
1483
1484
1484
1485
1485
1486
1486
1487
1487
1488
1488
1489
1489
1490
1490
1491
1491
1492
1492
1493
1493
1494
1494
1495
1495
1496
1496
1497
1497
1498
1498
1499
1499
1500
1500
1501
1501
1502
1502
1503
1503
1504
1504
1505
1505
1506
1506
1507
1507
1508
1508
1509
1509
1510
1510
1511
1511
1512
1512
1513
1513
1514
1514
1515
1515
1516
1516
1517
1517
1518
1518
1519
1519
1520
1520
1521
1521
1522
1522
1523
1523
1524
1524
1525
1525
1526
1526
1527
1527
1528
1528
1529
1529
1530
1530
1531
1531
1532
1532
1533
1533
1534
1534
1535
1535
1536
1536
1537
1537
1538
1538
1539
1539
1540
1540
1541
1541
1542
1542
1543
1543
1544
1544
1545
1545
1546
1546
1547
1547
1548
1548
1549
1549
1550
1550
1551
1551
1552
1552
1553
1553
1554
1554
1555
1555
1556
1556
1557
1557
1558
1558
1559
1559
1560
1560
1561
1561
1562
1562
1563
1563
1564
1564
1565
1565
1566
1566
1567
1567
1568
1568
1569
1569
1570
1570
1571
1571
1572
1572
1573
1573
1574
1574
1575
1575
1576
1576
1577
1577
1578
1578
1579
1579
1580
1580
1581
1581
1582
1582
1583
1583
1584
1584
1585
1585
1586
1586
1587
1587
1588
1588
1589
1589
1590
1590
1591
1591
1592
1592
1593
1593
1594
1594
1595
1595
1596
1596
1597
1597
1598
1598
1599
1599
1600
1600
1601
1601
1602
1602
1603
1603
1604
1604
1605
1605
1606
1606
1607
1607
1608
1608
1609
1609
1610
1610
1611
1611
1612
1612
1613
1613
1614
1614
1615
1615
1616
1616
1617
1617
1618
1618
1619
1619
1620
1620
1621
1621
1622
1622
1623
1623
1624
1624
1625
1625
1626
1626
1627
1627
1628
1628
1629
1629
1630
1630
1631
1631
1632
1632
1633
1633
1634
1634
1635
1635
1636
1636
1637
1637
1638
1638
1639
1639
1640
1640
1641
1641
1642
1642
1643
1643
1644
1644
1645
1645
1646
1646
1647
1647
1648
1648
1649
1649
1650
1650
1651
1651
1652
1652
1653
1653
1654
1654
1655
1655
1656
1656
1657
1657
1658
1658
1659
1659
1660
1660
1661
1661
1662
1662
1663
1663
1664
1664
1665
1665
1666
1666
1667
1667
1668
1668
1669
1669
1670
1670
1671
1671
1672
1672
1673
1673
1674
1674
1675
1675
1676
1676
1677
1677
1678
1678
1679
1679
1680
1680
1681
1681
1682
1682
1683
1683
1684
1684
1685
1685
1686
1686
1687
1687
1688
1688
1689
1689
1690
1690
1691
1691
1692
1692
1693
1693
1694
1694
1695
1695
1696
1696
1697
1697
1698
1698
1699
1699
1700
1700
1701
1701
1702
1702
1703
1703
1704
17
```

Performance

The performance was far from ideal, after 700 episode, the model was not improving. As all RL models, uncertainty is prevalent. Especially in this case, we need to find the perfect balance between two models in order to make it perform. And we simply ran out of time and computing power.

```
ALR: 2e-06 CLR: 9e-05 episode 658 of 2000 Average Reward (last 100 eps)= -195.52599301400278
ALR: 2e-06 CLR: 9e-05 episode 659 of 2000 Average Reward (last 100 eps)= -191.9333021155691
ALR: 2e-06 CLR: 9e-05 episode 660 of 2000 Average Reward (last 100 eps)= -191.75499694754973
ALR: 2e-06 CLR: 9e-05 episode 661 of 2000 Average Reward (last 100 eps)= -193.12594209924998
ALR: 2e-06 CLR: 9e-05 episode 662 of 2000 Average Reward (last 100 eps)= -191.84445102953157
ALR: 2e-06 CLR: 9e-05 episode 663 of 2000 Average Reward (last 100 eps)= -190.44990034709605
ALR: 2e-06 CLR: 9e-05 episode 664 of 2000 Average Reward (last 100 eps)= -191.47033097885253
ALR: 2e-06 CLR: 9e-05 episode 665 of 2000 Average Reward (last 100 eps)= -191.43658763918566
ALR: 2e-06 CLR: 9e-05 episode 666 of 2000 Average Reward (last 100 eps)= -189.55633828918744
ALR: 2e-06 CLR: 9e-05 episode 667 of 2000 Average Reward (last 100 eps)= -187.75440749070046
ALR: 2e-06 CLR: 9e-05 episode 668 of 2000 Average Reward (last 100 eps)= -188.0288119590154
ALR: 2e-06 CLR: 9e-05 episode 669 of 2000 Average Reward (last 100 eps)= -188.9932788684279
ALR: 2e-06 CLR: 9e-05 episode 670 of 2000 Average Reward (last 100 eps)= -190.62176962248495
ALR: 2e-06 CLR: 9e-05 episode 671 of 2000 Average Reward (last 100 eps)= -190.397969598374
ALR: 2e-06 CLR: 9e-05 episode 672 of 2000 Average Reward (last 100 eps)= -190.69511110511445
ALR: 2e-06 CLR: 9e-05 episode 673 of 2000 Average Reward (last 100 eps)= -190.01140096953898
ALR: 2e-06 CLR: 9e-05 episode 674 of 2000 Average Reward (last 100 eps)= -189.75008438812725
ALR: 2e-06 CLR: 9e-05 episode 675 of 2000 Average Reward (last 100 eps)= -194.16118206153968
ALR: 2e-06 CLR: 9e-05 episode 676 of 2000 Average Reward (last 100 eps)= -193.45029528705695
ALR: 2e-06 CLR: 9e-05 episode 677 of 2000 Average Reward (last 100 eps)= -197.12702651911684
ALR: 2e-06 CLR: 9e-05 episode 678 of 2000 Average Reward (last 100 eps)= -197.2498457418017
ALR: 2e-06 CLR: 9e-05 episode 679 of 2000 Average Reward (last 100 eps)= -196.56136256900194
ALR: 2e-06 CLR: 9e-05 episode 680 of 2000 Average Reward (last 100 eps)= -197.35531800569493
ALR: 2e-06 CLR: 9e-05 episode 681 of 2000 Average Reward (last 100 eps)= -196.0454408507716
ALR: 2e-06 CLR: 9e-05 episode 682 of 2000 Average Reward (last 100 eps)= -195.6432167234526
ALR: 2e-06 CLR: 9e-05 episode 683 of 2000 Average Reward (last 100 eps)= -195.13780059505496
ALR: 2e-06 CLR: 9e-05 episode 684 of 2000 Average Reward (last 100 eps)= -193.67506448293042
ALR: 2e-06 CLR: 9e-05 episode 685 of 2000 Average Reward (last 100 eps)= -193.14369150785996
ALR: 2e-06 CLR: 9e-05 episode 686 of 2000 Average Reward (last 100 eps)= -194.73156345327865
ALR: 2e-06 CLR: 9e-05 episode 687 of 2000 Average Reward (last 100 eps)= -193.52972566680975
ALR: 2e-06 CLR: 9e-05 episode 688 of 2000 Average Reward (last 100 eps)= -194.888572633882718
ALR: 2e-06 CLR: 9e-05 episode 689 of 2000 Average Reward (last 100 eps)= -194.4114778952176
ALR: 2e-06 CLR: 9e-05 episode 690 of 2000 Average Reward (last 100 eps)= -193.01414867592723
ALR: 2e-06 CLR: 9e-05 episode 691 of 2000 Average Reward (last 100 eps)= -193.11469268558065
ALR: 2e-06 CLR: 9e-05 episode 692 of 2000 Average Reward (last 100 eps)= -190.98922799647502
ALR: 2e-06 CLR: 9e-05 episode 693 of 2000 Average Reward (last 100 eps)= -192.82072964139775
```

Conclusion

- Relatively, we can see that the different Q-Networks are more consistent in meeting the solved requirements.
- DQN was also found to be overestimating over a single state, hence our decision to improve to DDQN and D3QN

Further Improvement

- Due to time constraint, we could not fully implement the A2C network.
- We could try to implement more RL algorithm such as **Proximal Policy Optimization (PPO)**, **Variational DQN**, or even Google's **Evolving Reinforcement Learning Algorithm**
- We could also try different hyperparameters such as **Learning Rate**, γ Discount Factor, **Batch Size**
- We could also try to change the Loss Function to Huber Loss to minimise the effect of outliers.



Thanks!