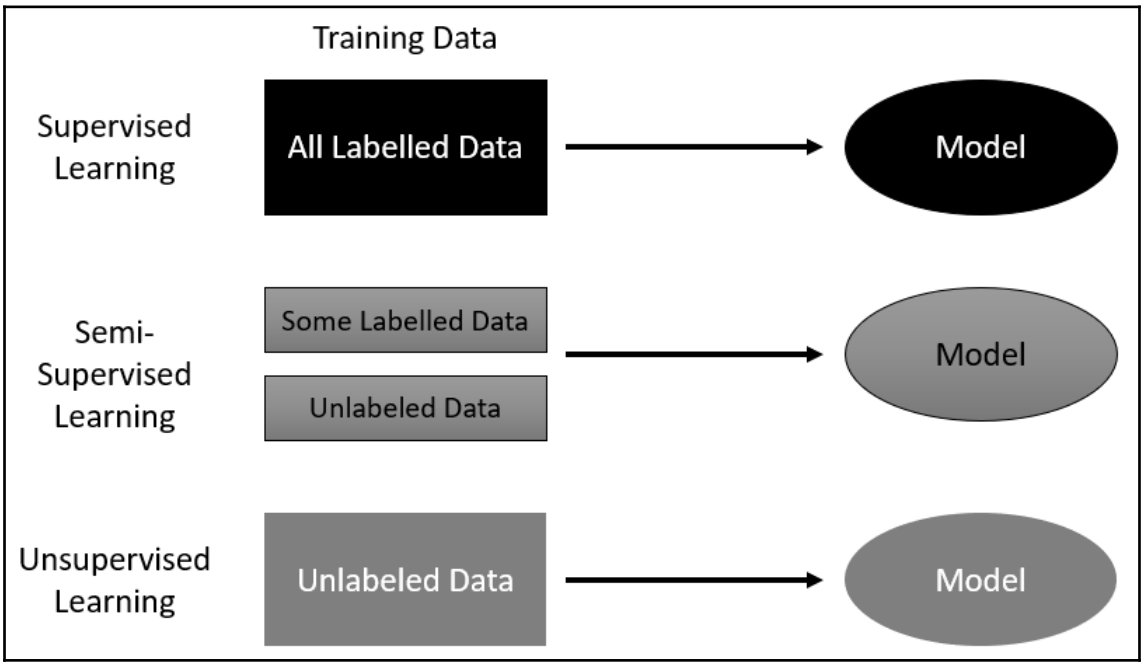
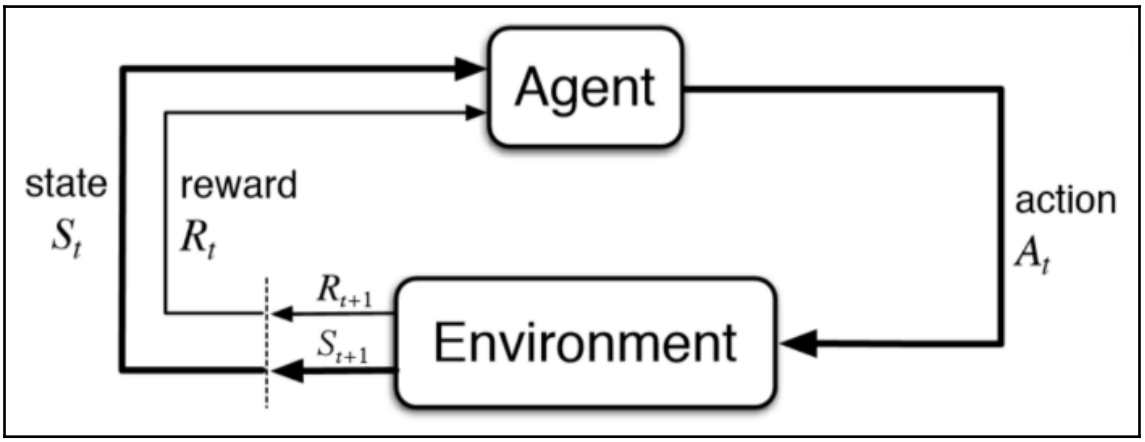
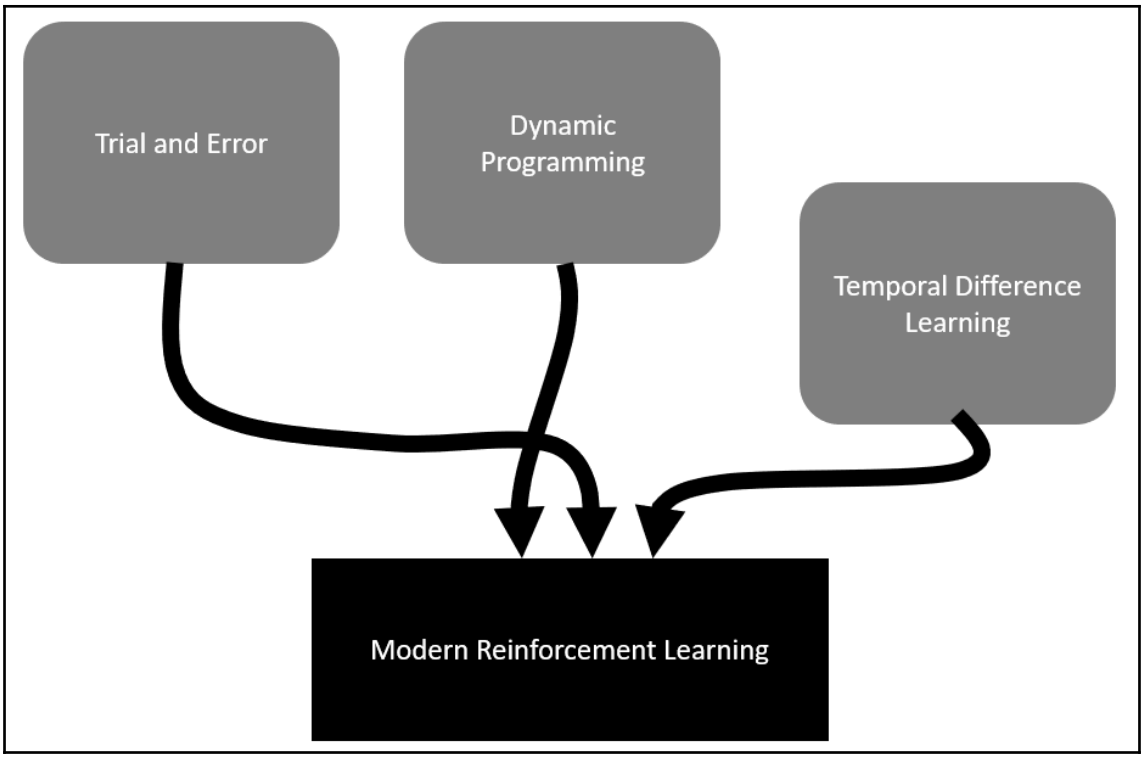
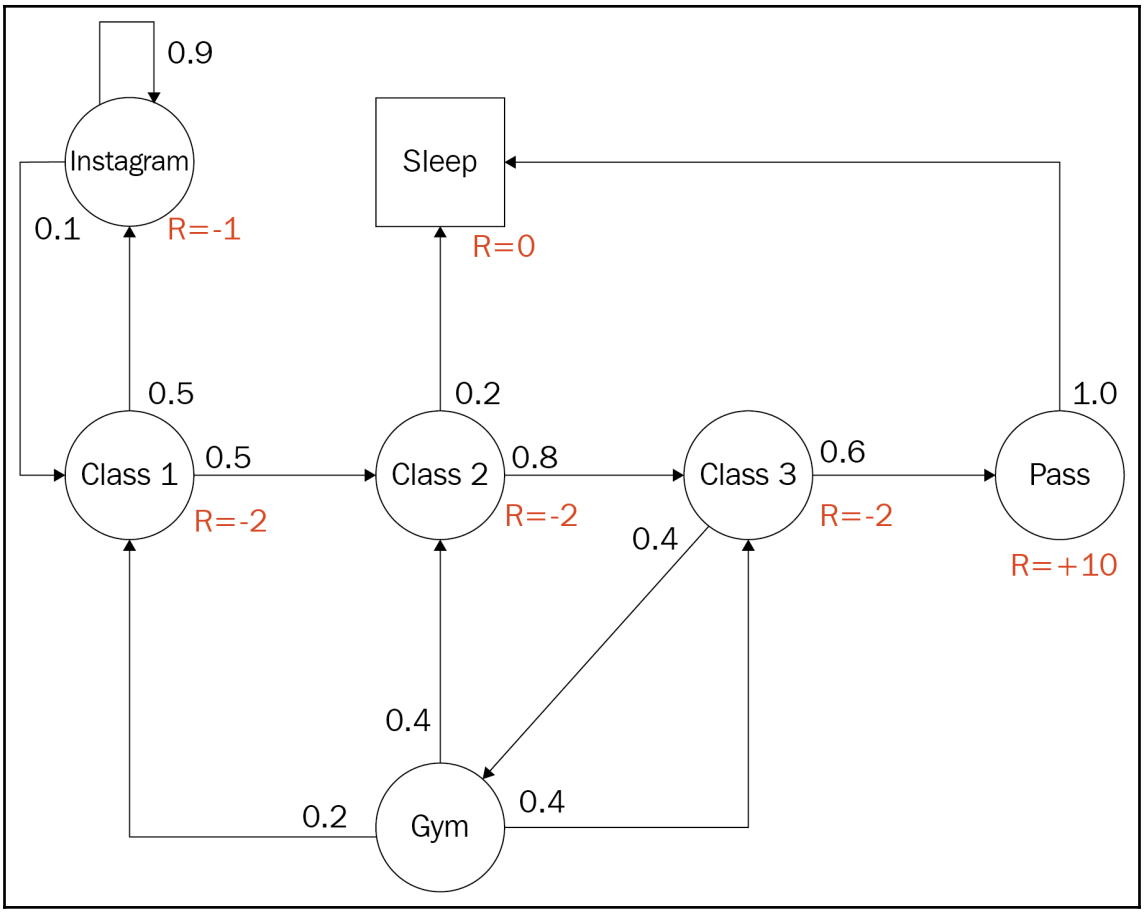
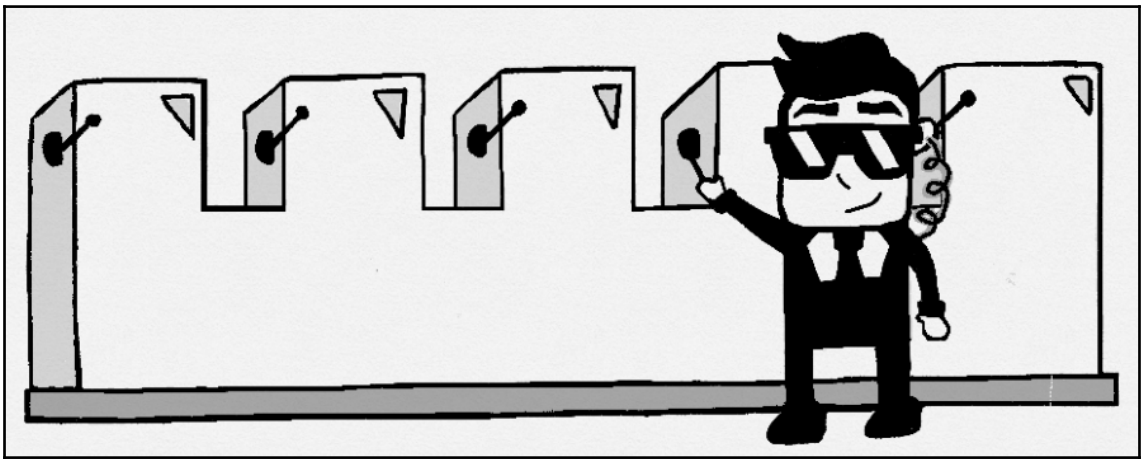
Chapter 1: Understanding Rewards-Based Learning

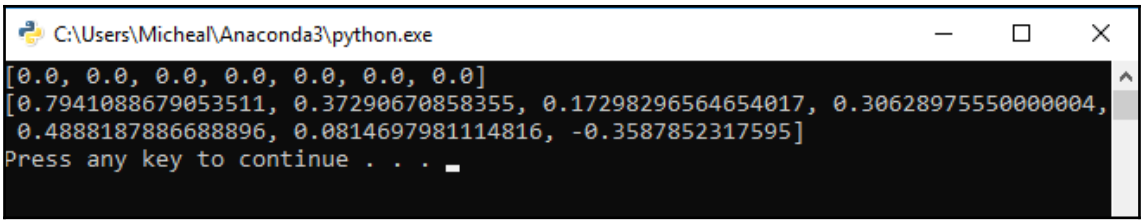


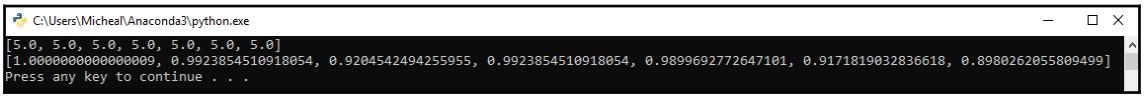



Trial and Error
Dynamic Programming
Temporal Difference Learning
Modern Reinforcement Learning

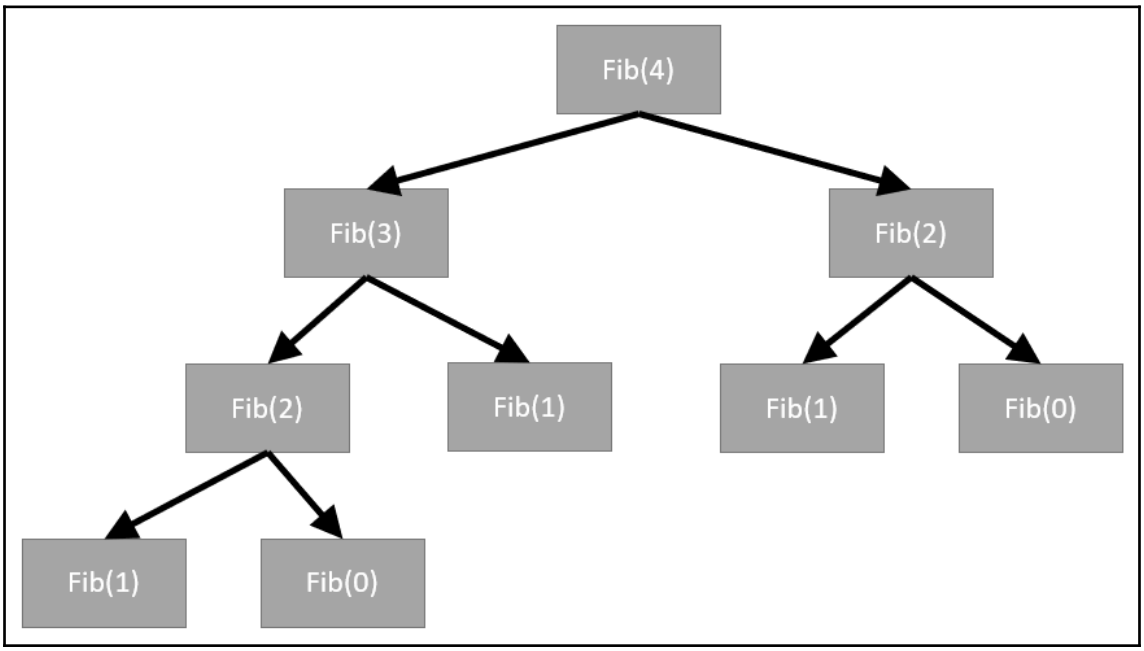

0.9
Instagram
0.1
R=-1
Sleep
R=0
0.5
0.2
1.0
Class 1
0.5
Class 2
0.8
Class 3
0.6
Pass
R=-2
R=-2
R=-2
R=+10
0.4
0.4
0.2
Gym
0.4

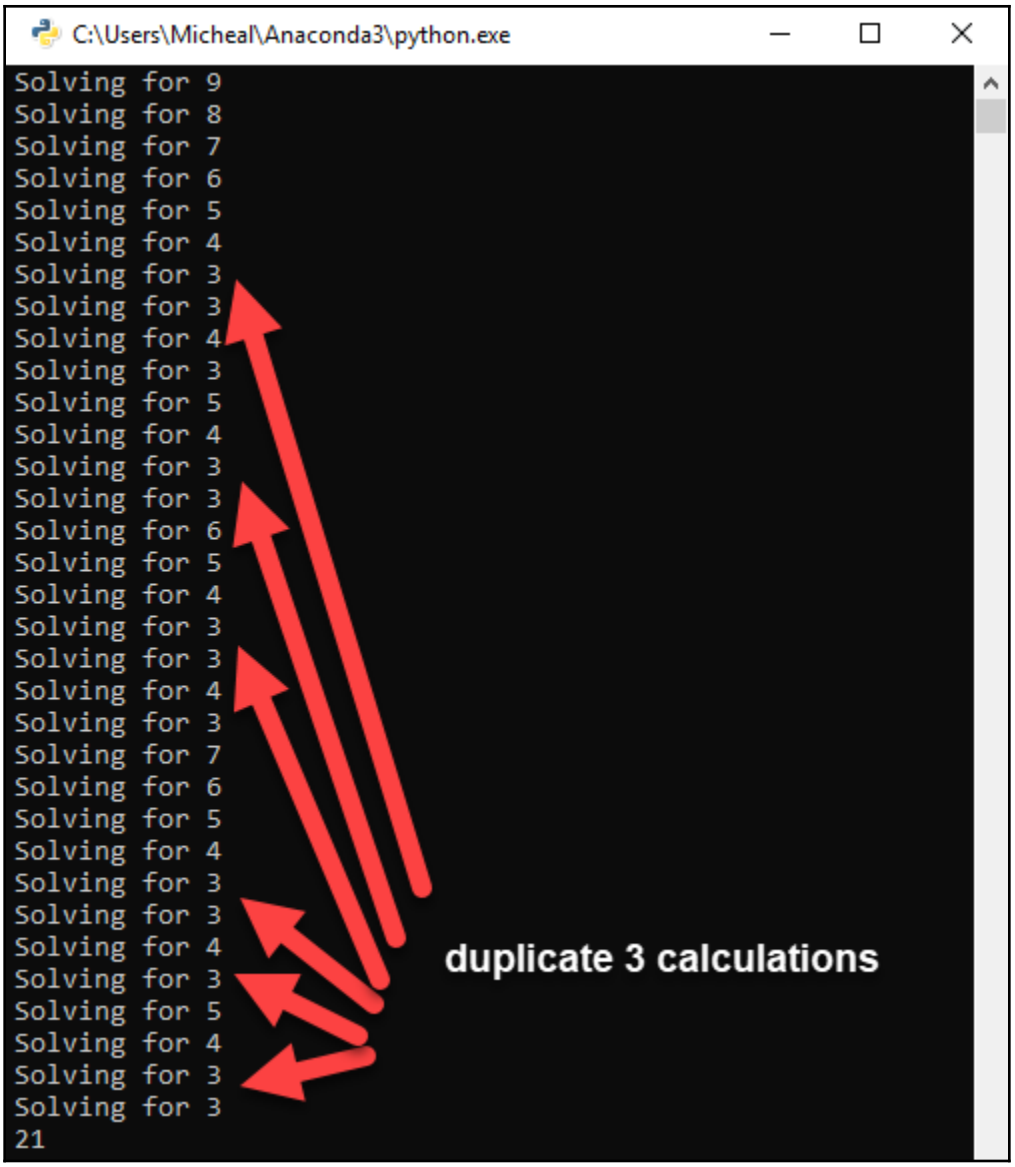


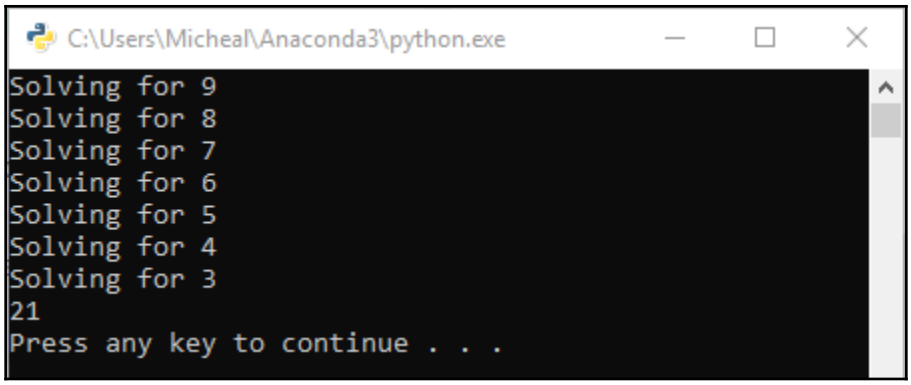

C:\Users\Micheal\Anaconda3\python.exe
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
[0.7941088679053511, 0.37290670858355, 0.17298296564654017, 0.30628975550000004,
0.4888187886688896, 0.0814697981114816, -0.3587852317595]
Press any key to continue . . .

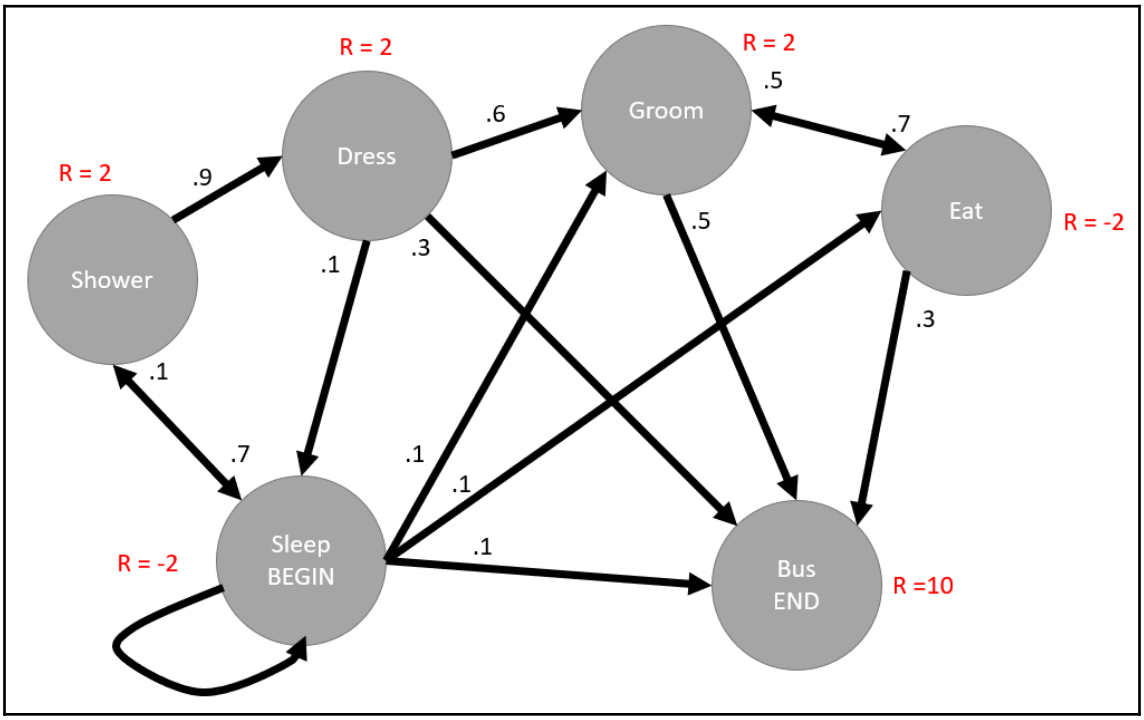

C:\Users\Micheal\Anaconda3\python.exe
[5.0, 5.0, 5.0, 5.0, 5.0, 5.0, 5.0]
[1.0000000000000009, 0.9923854510918054, 0.9204542494255955, 0.9923854510918054, 0.9899692772647101, 0.9171819032836618, 0.8980262055809499]
Press any key to continue . . .

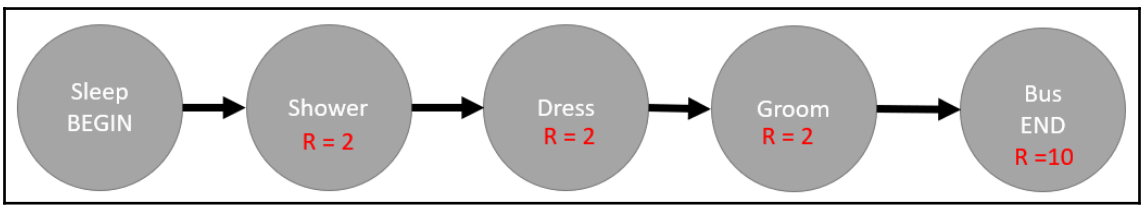
Chapter 2: Dynamic Programming and the Bellman Equation



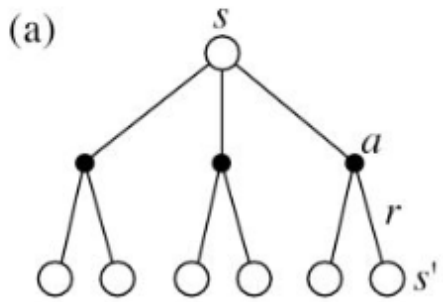

C:\Users\Micheal\Anaconda3\python.exe
Solving for 9
Solving for 8
Solving for 7
Solving for 6
Solving for 5
Solving for 4
Solving for 3
Solving for 3
Solving for 4
Solving for 3
Solving for 5
Solving for 4
Solving for 3
Solving for 3
Solving for 6
Solving for 5
Solving for 4
Solving for 3
Solving for 3
Solving for 4
Solving for 3
Solving for 7
Solving for 6
Solving for 5
Solving for 4
Solving for 3
Solving for 3
Solving for 4
Solving for 3
Solving for 5
Solving for 4
Solving for 3
Solving for 3
21
duplicate 3 calculations


C:\Users\Micheal\Anaconda3\python.exe
Solving for 9
Solving for 8
Solving for 7
Solving for 6
Solving for 5
Solving for 4
Solving for 3
21
Press any key to continue . . .

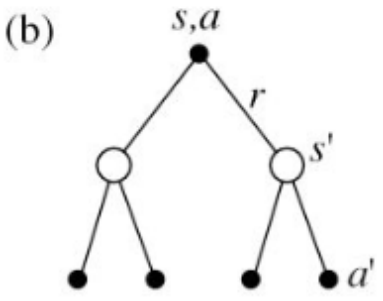

R = 2
R = 2
.5
.6
Groom
.7
Dress
R = 2
.9
Eat
R = -2
Shower
.5
.1
.3
.3
.1
.7
.1
.1
.1
Sleep
BEGIN
R = -2
Bus
END
R =10


Sleep
BEGIN
Shower
R = 2
Dress
R = 2
Groom
R = 2
Bus
END
R =10

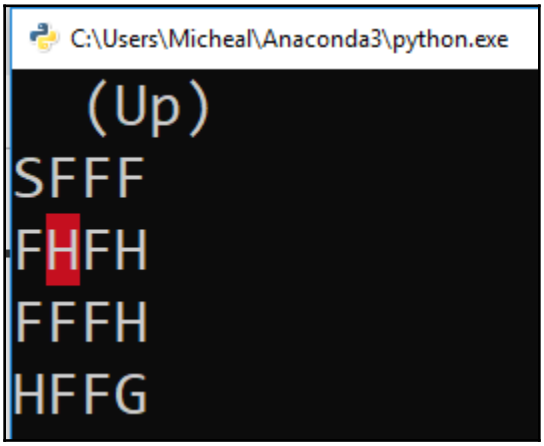
Backup diagrams:

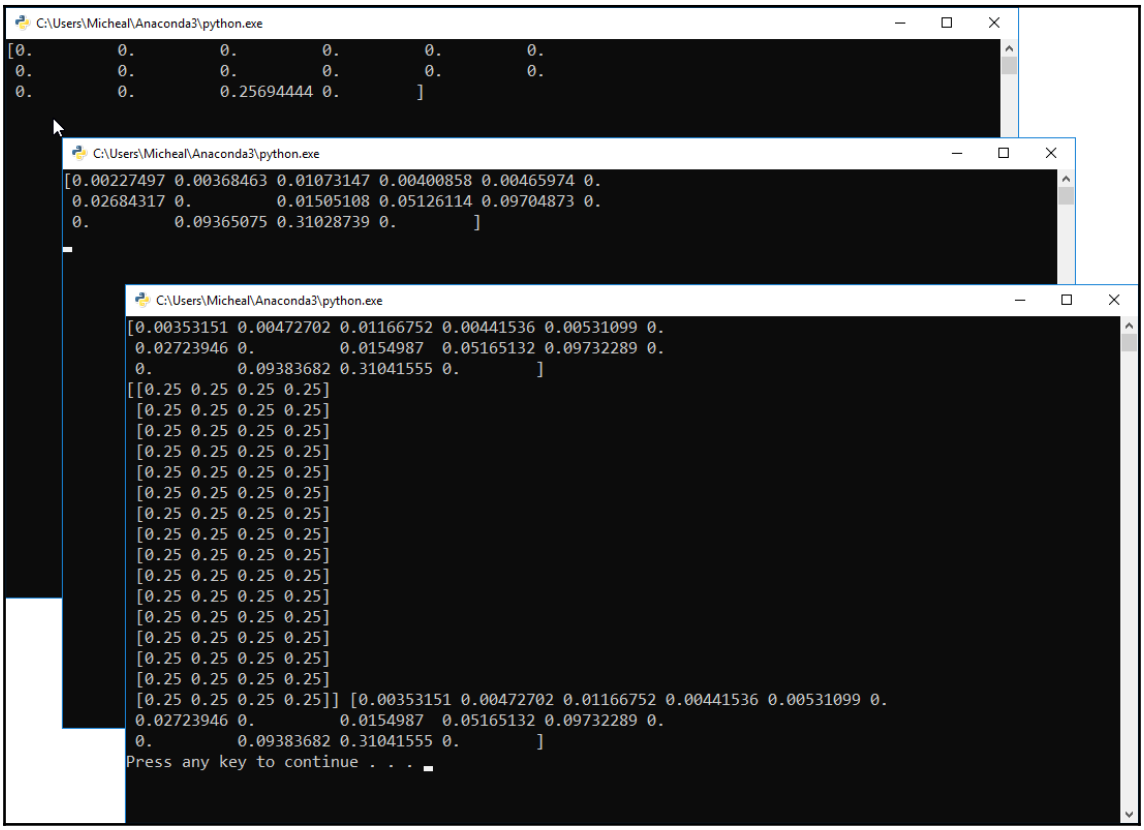


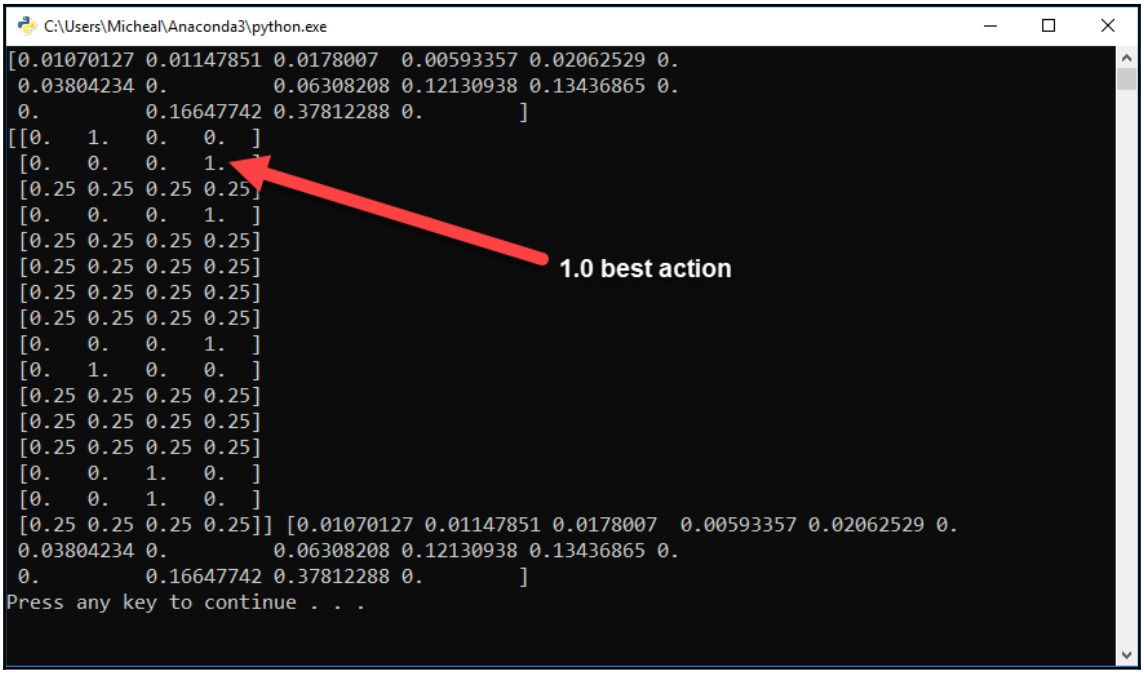
for V^{π}

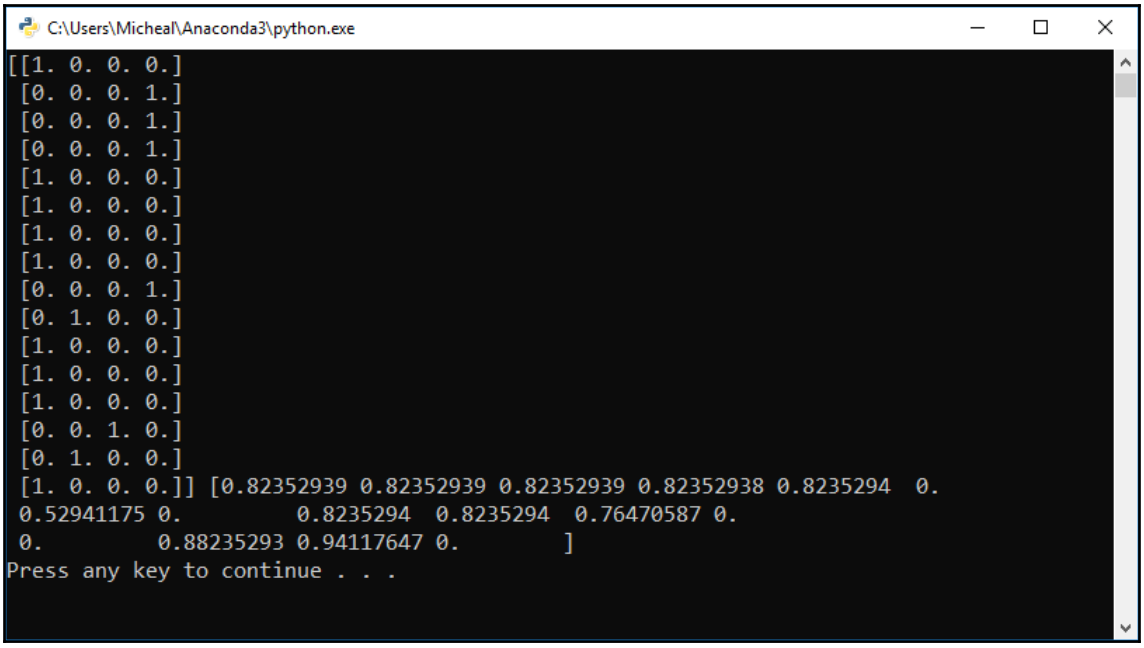


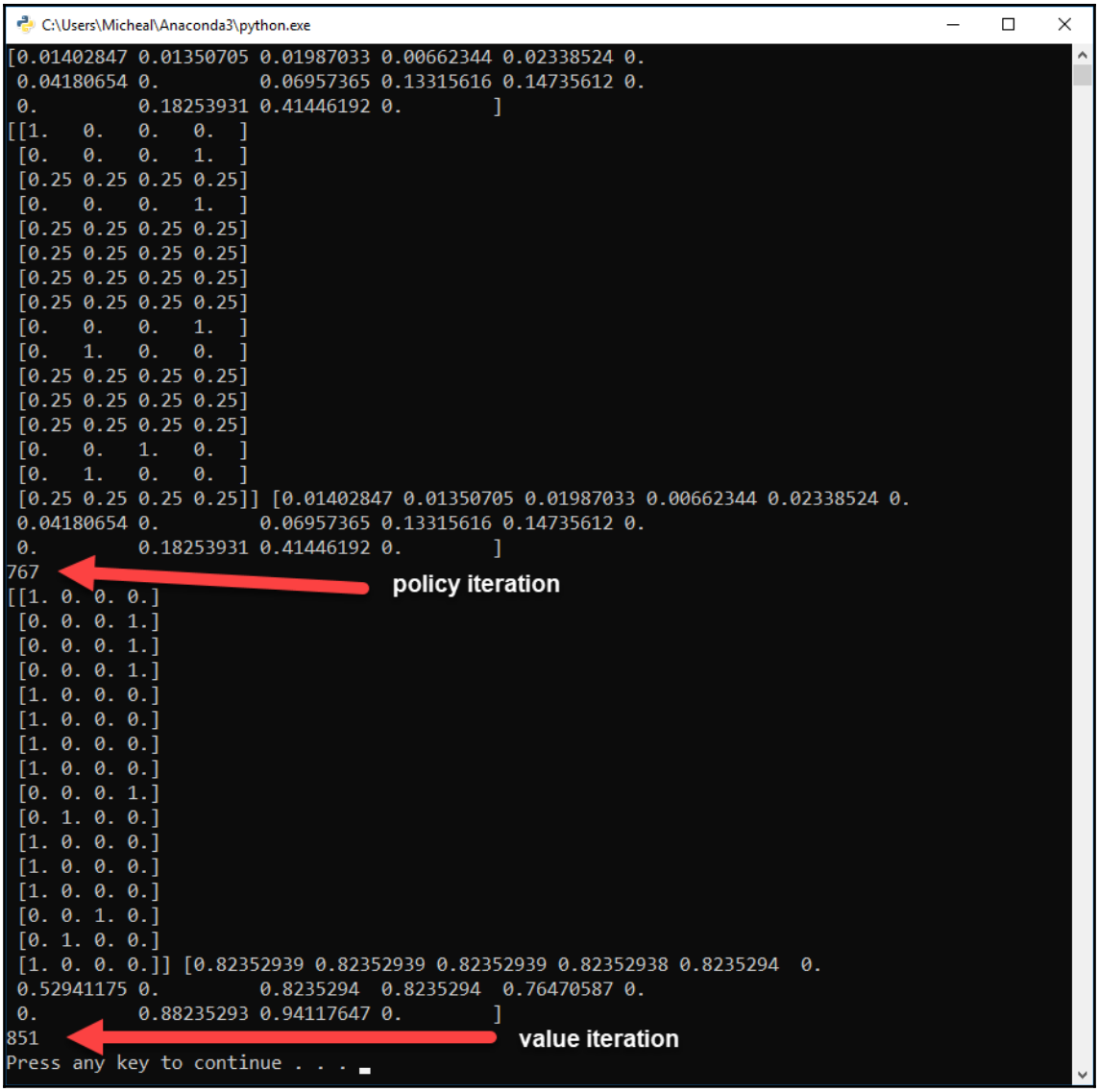
for Q^{π}



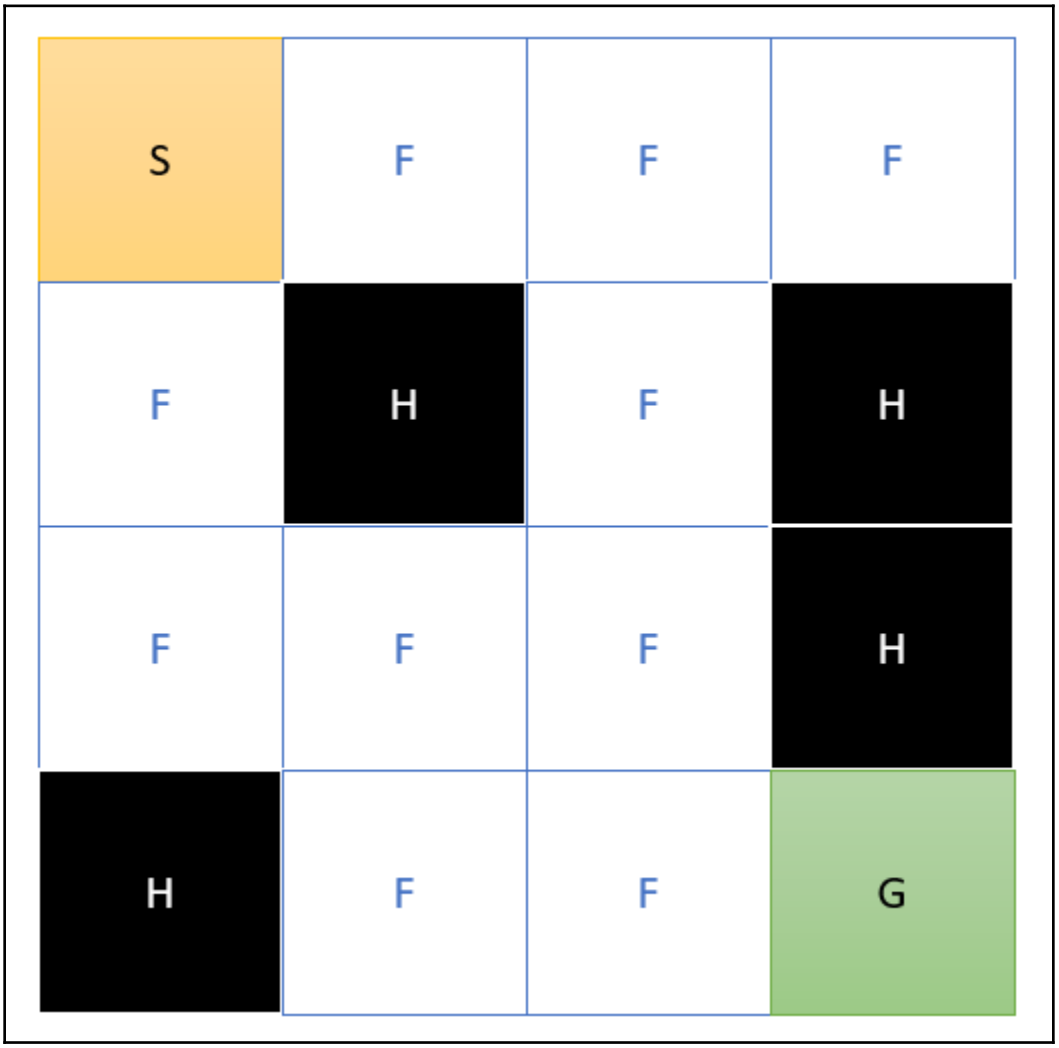

C:\Users\Micheal\Anaconda3\python.exe
[0.   0.   0.   0.   0.   0.
 0.   0.   0.   0.   0.   0.
 0.   0.   0.25694444 0.   ]
C:\Users\Micheal\Anaconda3\python.exe
[0.00227497 0.00368463 0.01073147 0.00400858 0.00465974 0.
 0.02684317 0.   0.01505108 0.05126114 0.09704873 0.
 0.   0.09365075 0.31028739 0.   ]
C:\Users\Micheal\Anaconda3\python.exe
[0.00353151 0.00472702 0.01166752 0.00441536 0.00531099 0.
 0.02723946 0.   0.0154987 0.05165132 0.09732289 0.
 0.   0.09383682 0.31041555 0.   ]
[[0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]] [0.00353151 0.00472702 0.01166752 0.00441536 0.00531099 0.
 0.02723946 0.   0.0154987 0.05165132 0.09732289 0.
 0.   0.09383682 0.31041555 0.   ]
Press any key to continue . . .

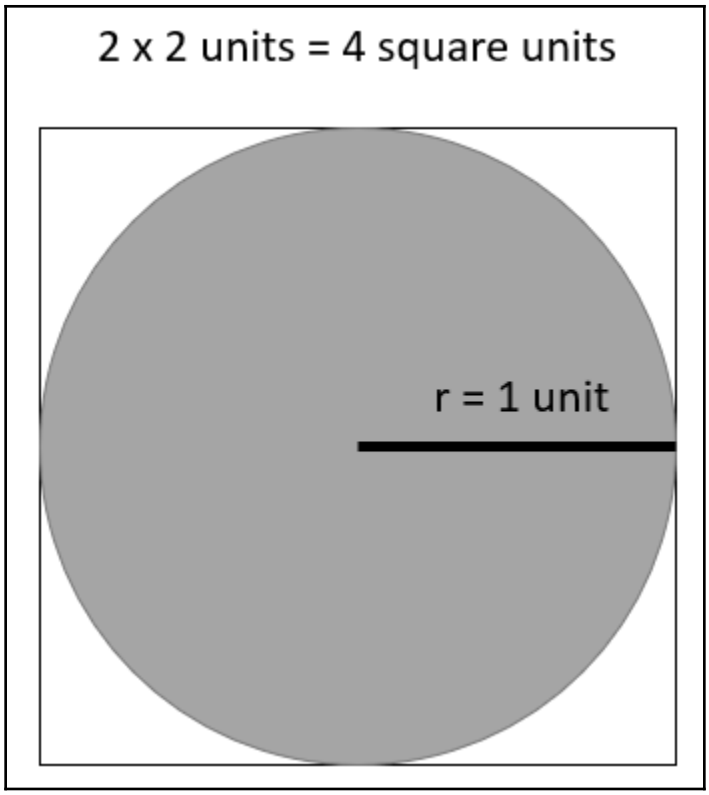

C:\Users\Micheal\Anaconda3\python.exe
[0.01070127 0.01147851 0.0178007 0.00593357 0.02062529 0.
 0.03804234 0.   0.06308208 0.12130938 0.13436865 0.
 0.   0.16647742 0.37812288 0.   ]
[[0.   1.   0.   0.]
 [0.   0.   0.   1.]
 [0.25 0.25 0.25 0.25]
 [0.   0.   0.   1.]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.   0.   0.   1.]
 [0.   1.   0.   0.]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.   0.   1.   0.]
 [0.   0.   1.   0.]
 [0.25 0.25 0.25 0.25]] [0.01070127 0.01147851 0.0178007 0.00593357 0.02062529 0.
 0.03804234 0.   0.06308208 0.12130938 0.13436865 0.
 0.   0.16647742 0.37812288 0.   ]
Press any key to continue . . .
1.0 best action


C:\Users\Micheal\Anaconda3\python.exe
[[1. 0. 0. 0.]
 [0. 0. 0. 1.]
 [0. 0. 0. 1.]
 [0. 0. 0. 1.]
 [1. 0. 0. 0.]
 [1. 0. 0. 0.]
 [1. 0. 0. 0.]
 [1. 0. 0. 0.]
 [0. 0. 0. 1.]
 [0. 1. 0. 0.]
 [1. 0. 0. 0.]
 [1. 0. 0. 0.]
 [1. 0. 0. 0.]
 [0. 0. 1. 0.]
 [0. 1. 0. 0.]
 [1. 0. 0. 0.]] [0.82352939 0.82352939 0.82352939 0.82352938 0.8235294 0.
 0.52941175 0.   0.8235294 0.8235294 0.76470587 0.
 0.   0.88235293 0.94117647 0.   ]
Press any key to continue . . .


C:\Users\Micheal\Anaconda3\python.exe
[0.01402847 0.01350705 0.01987033 0.00662344 0.02338524 0.
0.04180654 0. 0.06957365 0.13315616 0.14735612 0.
0. 0.18253931 0.41446192 0.]
[[1. 0. 0. 0.]
[0. 0. 0. 1.]
[0.25 0.25 0.25 0.25]
[0. 0. 0. 1.]
[0.25 0.25 0.25 0.25]
[0.25 0.25 0.25 0.25]
[0.25 0.25 0.25 0.25]
[0.25 0.25 0.25 0.25]
[0. 0. 0. 1.]
[0. 1. 0. 0.]
[0.25 0.25 0.25 0.25]
[0.25 0.25 0.25 0.25]
[0.25 0.25 0.25 0.25]
[0. 0. 1. 0.]
[0. 1. 0. 0.]
[0.25 0.25 0.25 0.25]] [0.01402847 0.01350705 0.01987033 0.00662344 0.02338524 0.
0.04180654 0. 0.06957365 0.13315616 0.14735612 0.
0. 0.18253931 0.41446192 0.]
767
policy iteration
[[1. 0. 0. 0.]
[0. 0. 0. 1.]
[0. 0. 0. 1.]
[0. 0. 0. 1.]
[1. 0. 0. 0.]
[1. 0. 0. 0.]
[1. 0. 0. 0.]
[1. 0. 0. 0.]
[0. 0. 0. 1.]
[0. 1. 0. 0.]
[1. 0. 0. 0.]
[1. 0. 0. 0.]
[1. 0. 0. 0.]
[0. 0. 1. 0.]
[0. 1. 0. 0.]
[1. 0. 0. 0.]] [0.82352939 0.82352939 0.82352939 0.82352938 0.8235294 0.
0.52941175 0. 0.8235294 0.8235294 0.76470587 0.
0. 0.88235293 0.94117647 0.]
851
value iteration
Press any key to continue . . .

Chapter 3: Monte Carlo Methods




2 x 2 units = 4 square units
r = 1 unit

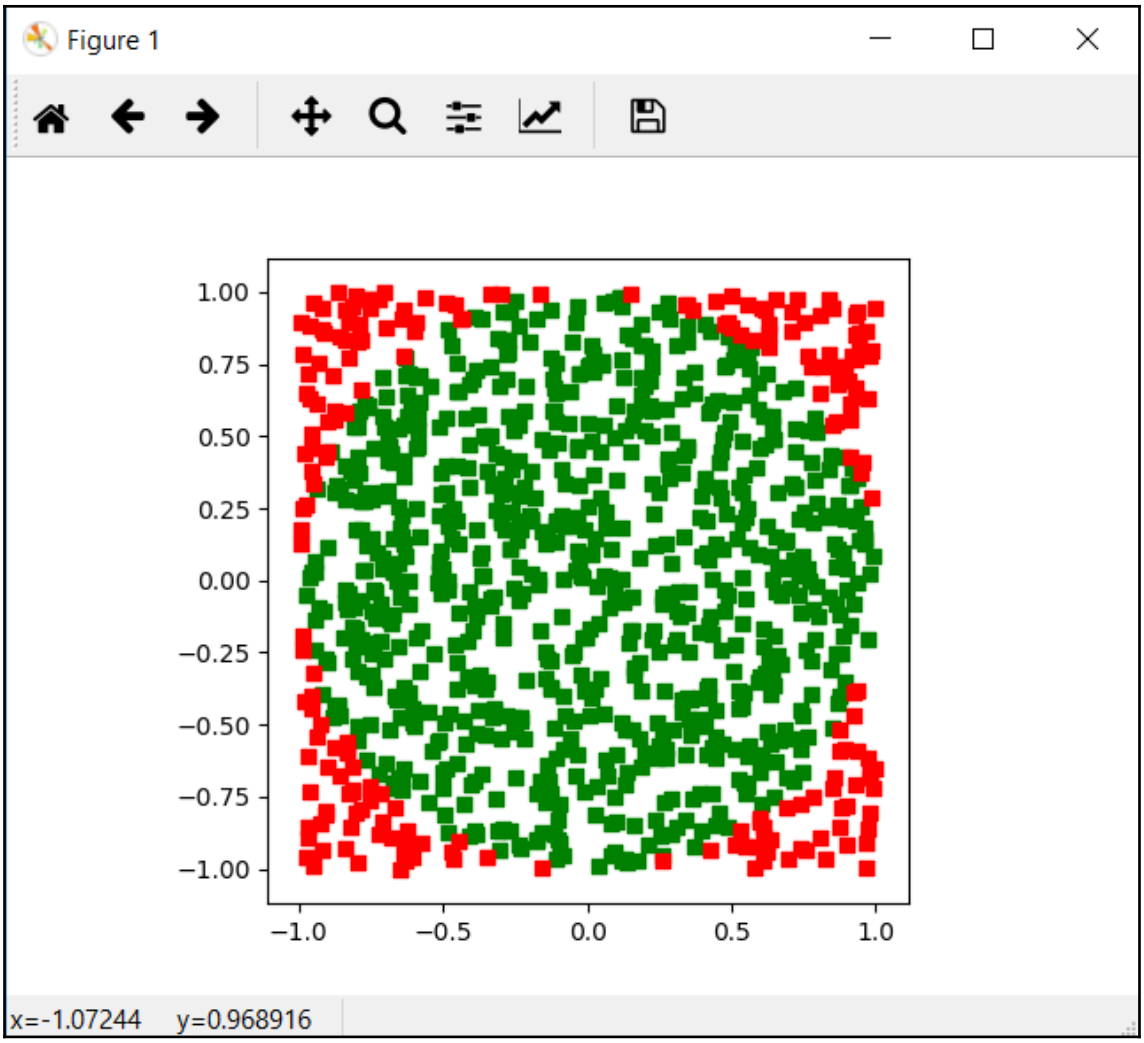
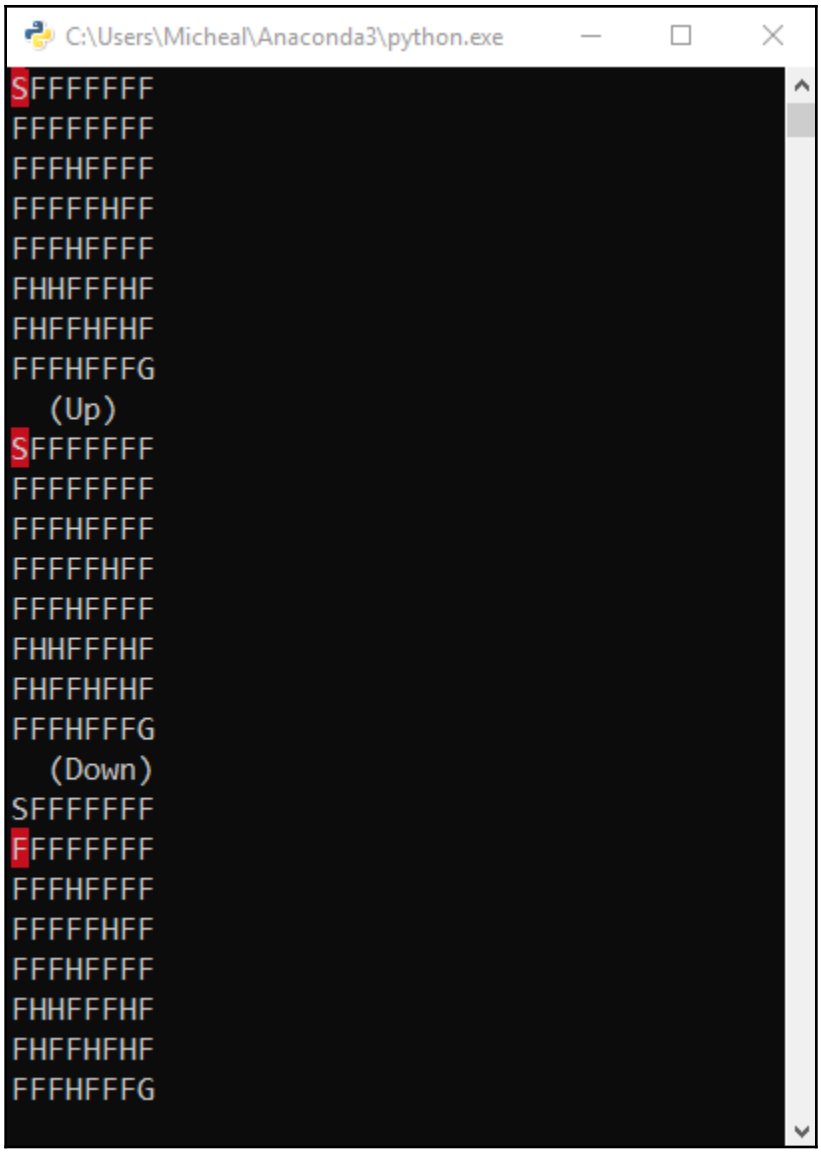
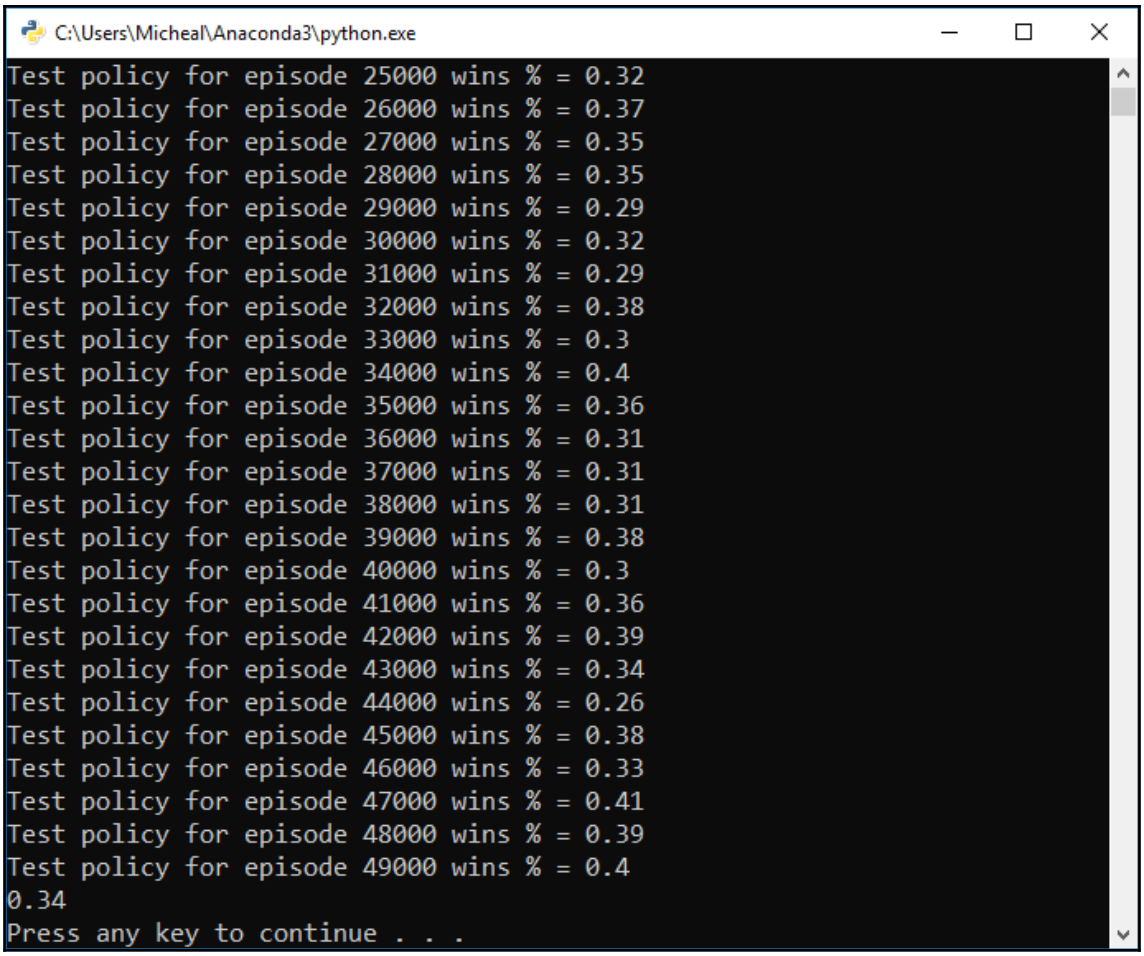


Figure 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.0
−0.5
0.0
0.5
1.0
x=-1.07244 y=0.968916


C:\Users\Micheal\Anaconda3\python.exe
SFFFFFFF
FFFFFFFF
FFFHFFFF
FFFFFHFF
FFFHFFFF
FHHFFFHF
FHFFHFHF
FFFHFFFG
(Up)
SFFFFFFF
FFFFFFFF
FFFHFFFF
FFFFFHFF
FFFHFFFF
FHHFFFHF
FHFFHFHF
FFFHFFFG
(Down)
SFFFFFFF
FFFFFFFF
FFFHFFFF
FFFFFHFF
FFFHFFFF
FHHFFFHF
FHFFHFHF
FFFHFFFG


C:\Users\Micheal\Anaconda3\python.exe
Test policy for episode 25000 wins % = 0.32
Test policy for episode 26000 wins % = 0.37
Test policy for episode 27000 wins % = 0.35
Test policy for episode 28000 wins % = 0.35
Test policy for episode 29000 wins % = 0.29
Test policy for episode 30000 wins % = 0.32
Test policy for episode 31000 wins % = 0.29
Test policy for episode 32000 wins % = 0.38
Test policy for episode 33000 wins % = 0.3
Test policy for episode 34000 wins % = 0.4
Test policy for episode 35000 wins % = 0.36
Test policy for episode 36000 wins % = 0.31
Test policy for episode 37000 wins % = 0.31
Test policy for episode 38000 wins % = 0.31
Test policy for episode 39000 wins % = 0.38
Test policy for episode 40000 wins % = 0.3
Test policy for episode 41000 wins % = 0.36
Test policy for episode 42000 wins % = 0.39
Test policy for episode 43000 wins % = 0.34
Test policy for episode 44000 wins % = 0.26
Test policy for episode 45000 wins % = 0.38
Test policy for episode 46000 wins % = 0.33
Test policy for episode 47000 wins % = 0.41
Test policy for episode 48000 wins % = 0.39
Test policy for episode 49000 wins % = 0.4
0.34
Press any key to continue . . .

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:

$Q(s,a) \leftarrow$ arbitrary

$Returns(s,a) \leftarrow$ empty list

$\pi \leftarrow$ an arbitrary ε-soft policy

Repeat forever:

(a) Generate an episode using π

(b) For each pair s,a appearing in the episode:

$R \leftarrow$ return following the first occurrence of s,a

Append R to $Returns(s,a)$

$Q(s,a) \leftarrow \text{average}(Returns(s,a))$

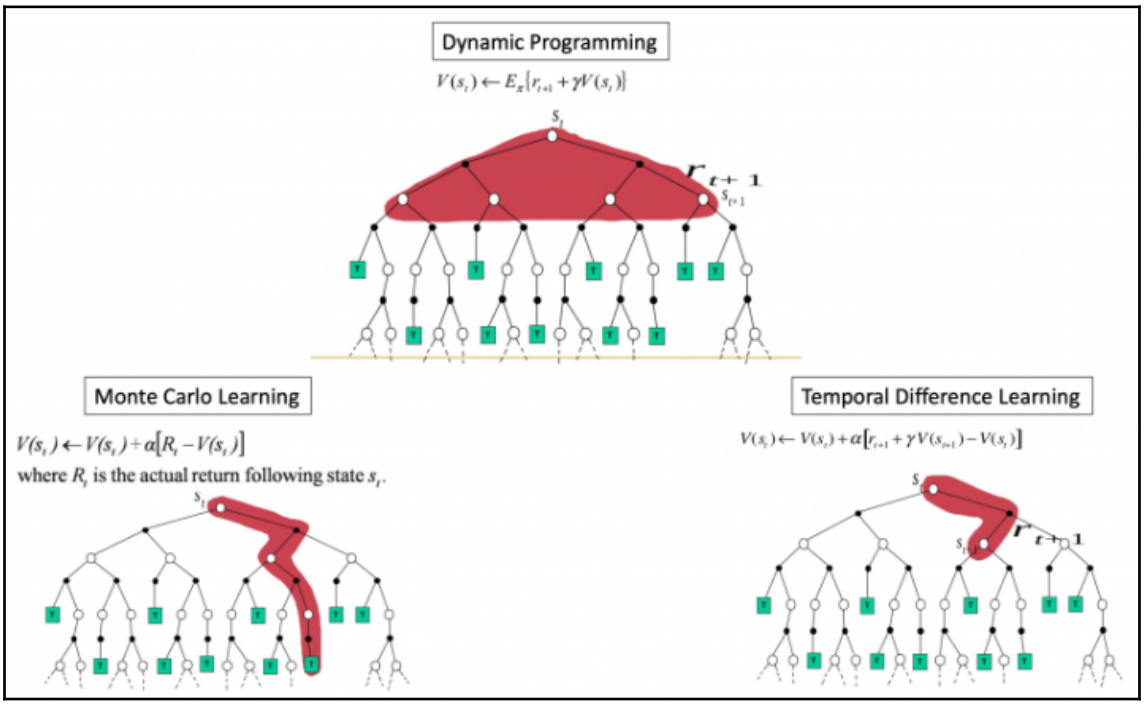
(c) For each s in the episode:

$a^* \leftarrow \arg\max_a Q(s,a)$

For all $a \in \mathcal{A}(s)$:

$$\pi(s,a) \leftarrow \begin{cases} 1-\varepsilon+\varepsilon/|\mathcal{A}(s)| & \text{if } a = a^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq a^* \end{cases}$$

Chapter 4: Temporal Difference Learning




```
C:\Users\Micheal\Anaconda3\python.exe
100%|##################################################################| 10000/10000 [00:21<00:00, 458.54it/s]
[[ 0.         -1.49423831 -1.9958849  -1.99776133]
 [-1.2447176  -1.98811069 -1.98747494 -1.98064978]
 [-1.86806477 -1.98964328 -1.88924666 -1.49072661]
 [-1.98202495 -1.90378281 -1.55590434  0.        ]]
Press any key to continue . . .
```

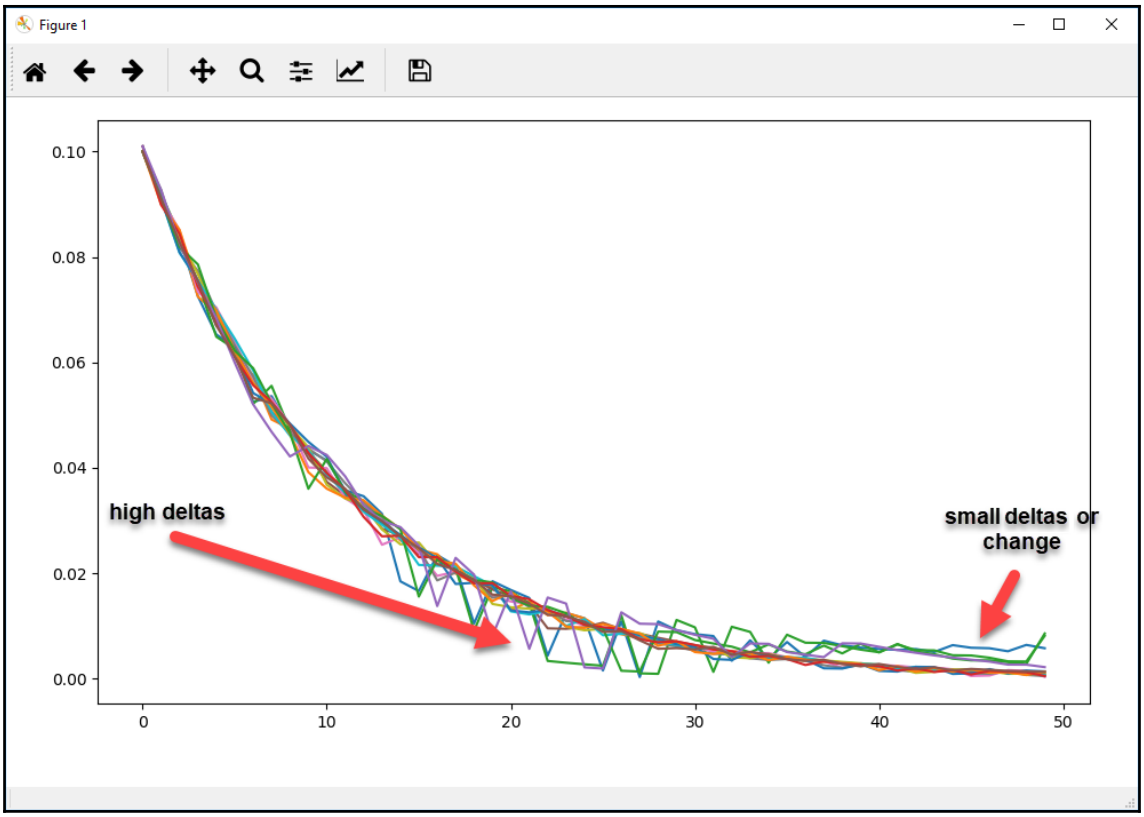
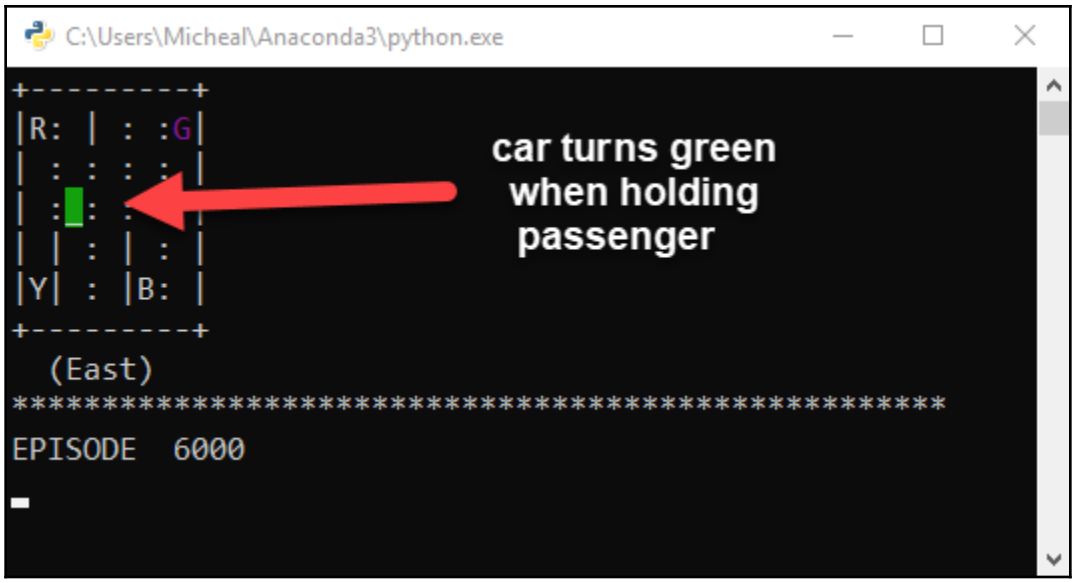


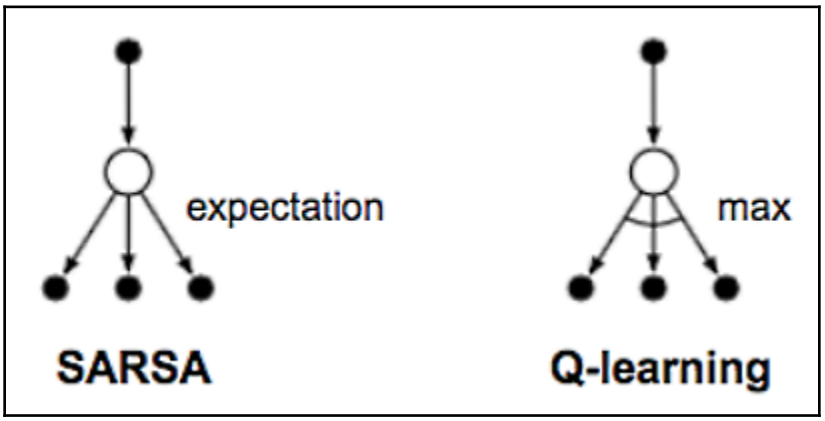
Figure 1
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
high deltas
small deltas or change

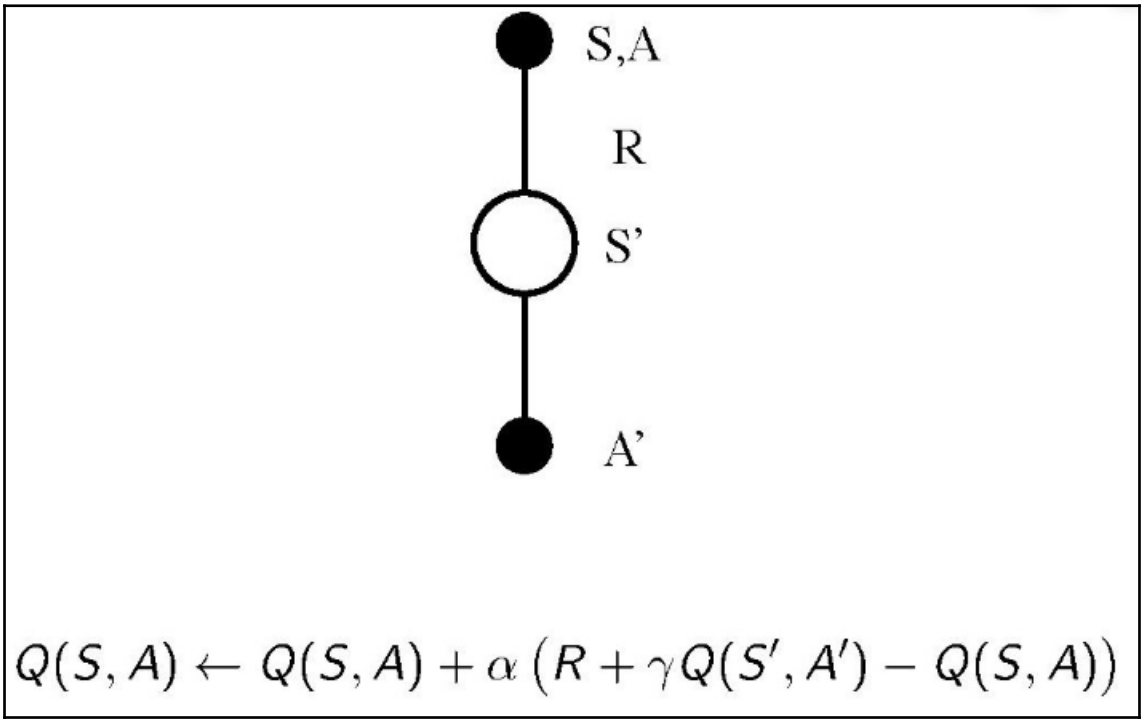

C:\Users\Micheal\Anaconda3\python.exe
(Left)
SFFF
FHFH
FFFH
HFFG
****
EPISODE 10000


C:\Users\Micheal\Anaconda3\python.exe
+---------+
|R: | : :G|
| : : : : |
| : : : : |
| | : | : |
|Y| : |B: |
+---------+
(East)
****
EPISODE 6000
car turns green when holding passenger

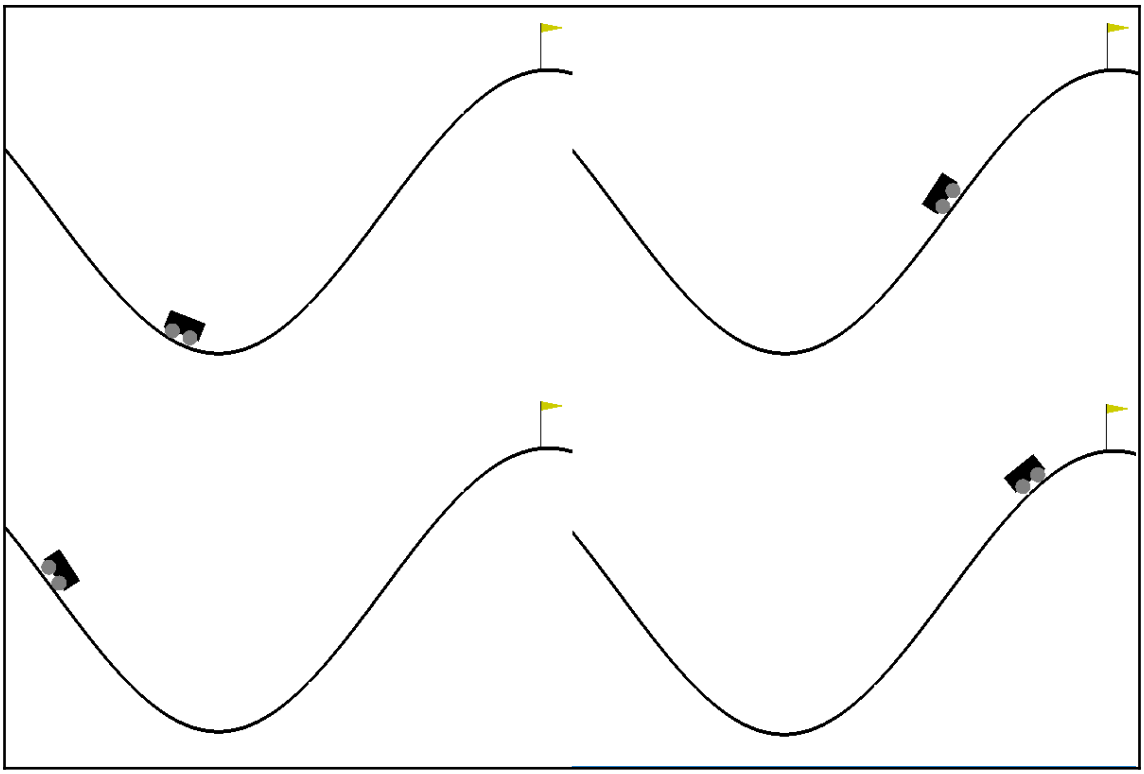

C:\Users\Micheal\Anaconda3\python.exe
Score 4
100%|##########| 50000/50000 [10:15<00:00, 81.21it/s]
[[0.16666667 0.16666667 0.16666667 0.16666667 0.16666667
0.16666667]
[-1.8953181 -1.44880345 -1.8953178 -1.44819268 -0.72512948
-10.44789578]
[-0.72617581 0.4446 -0.84835659 0.44421064 2.33782249
-8.55527336]
...
[-2.14796729 5.15374787 -2.18733027 -2.13442453 -9.94738054
-9.31336954]
[-2.30243816 -1.46591343 -2.30243816 -2.35177552 -10.65814823
-10.18335066]
[12.69124466 -0.89999954 -0.5779 31.35602094 0.16666667
0.16666667]]
Score over time: 12.64
Press any key to continue . . .
Best score averaged over 100 episodes

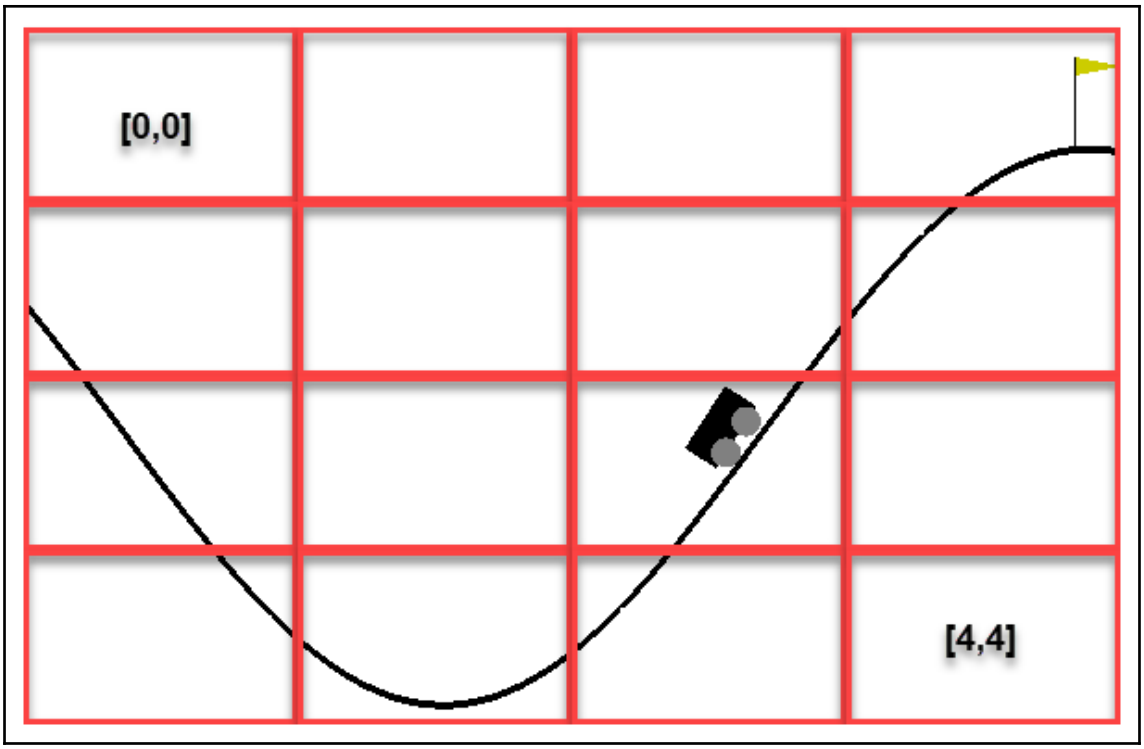
Chapter 5: Exploring SARSA





$$Q(S,A) \leftarrow Q(S,A) + \alpha\left(R + \gamma Q(S',A') - Q(S,A)\right)$$




[0,0]
[4,4]

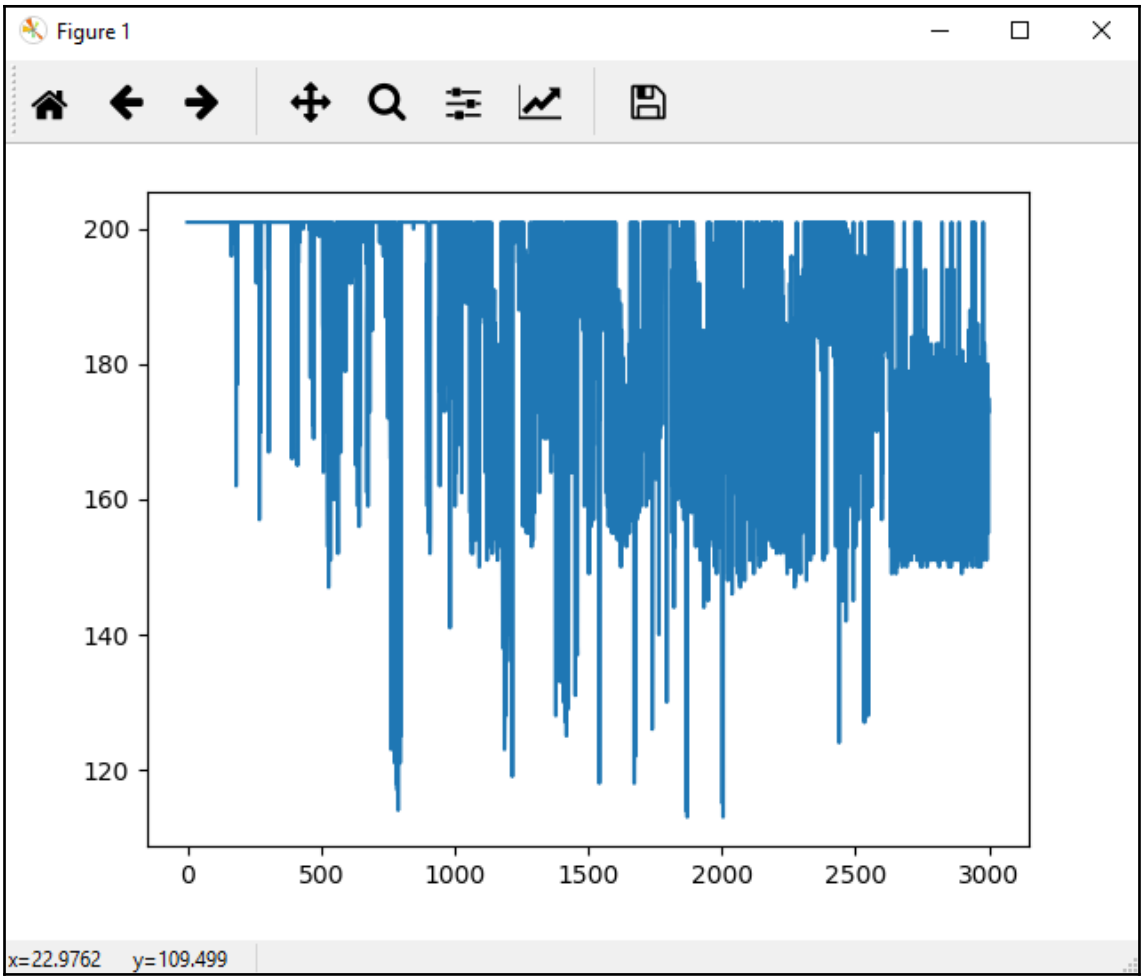


Figure 1
200
180
160
140
120
0
500
1000
1500
2000
2500
3000
x=22.9762 y=109.499

Mountain Car

Observation

Type: Box(2)

Num	Observation	Min	Max
0	position	-1.2	0.6
1	velocity	-0.07	0.07

Actions

Type: Discrete(3)

Num	Action
0	push left
1	no push
2	push right

Cart Pole

Observation

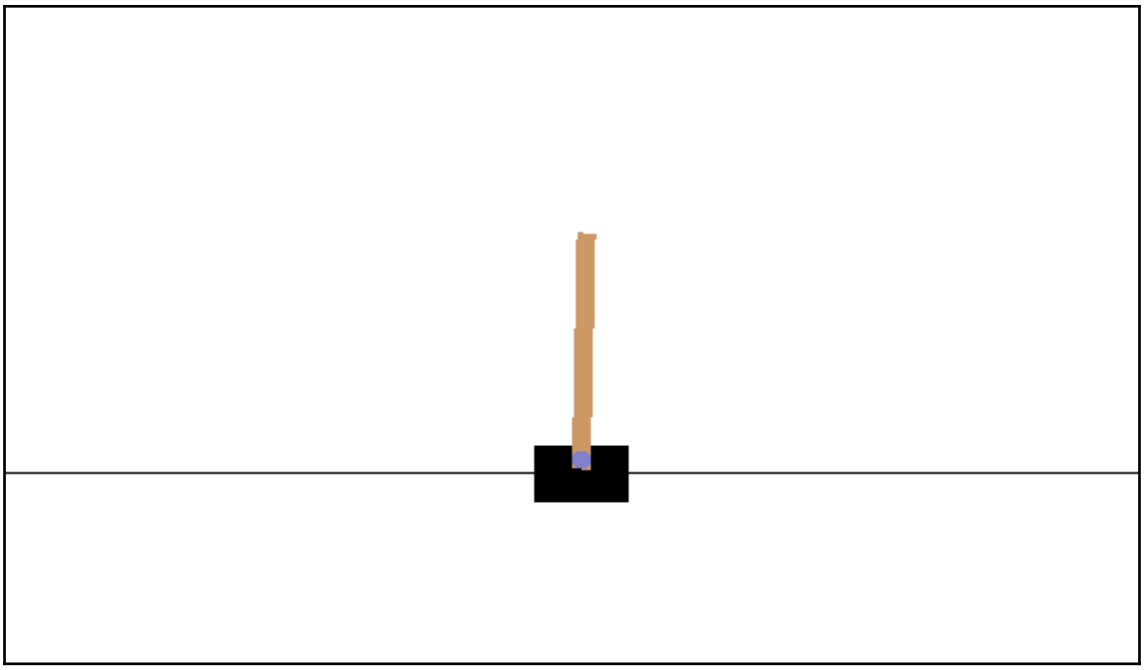
Type: Box(4)

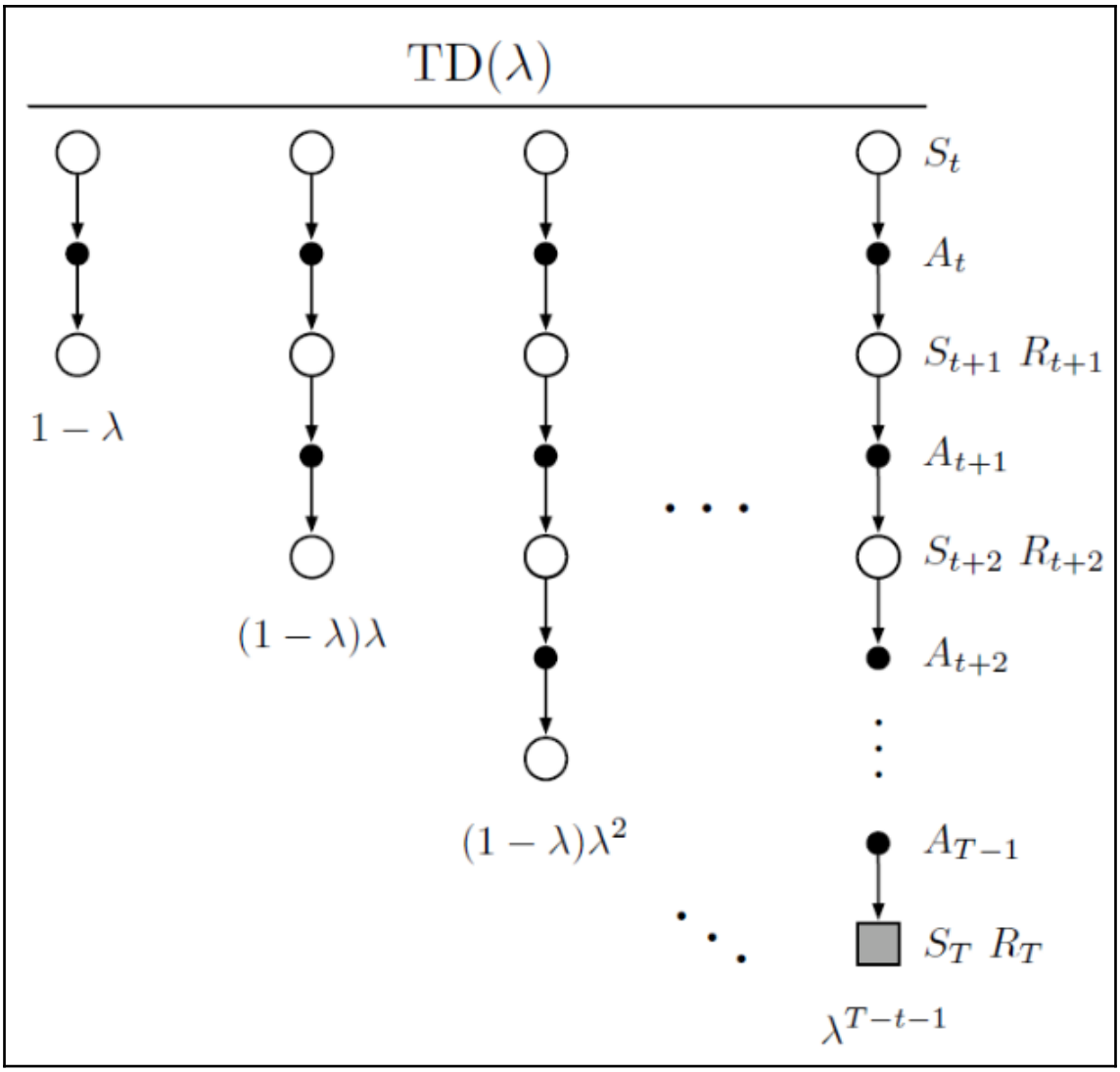
Num	Observation	Min	Max
0	Cart Position	-2.4	2.4
1	Cart Velocity	-Inf	Inf
2	Pole Angle	~ -41.8°	~ 41.8°
3	Pole Velocity At Tip	-Inf	Inf

Actions

Type: Discrete(2)

Num	Action
0	Push cart to the left
1	Push cart to the right




TD(λ)
S_t
A_t
S_{t+1} R_{t+1}
A_{t+1}
S_{t+2} R_{t+2}
A_{t+2}
A_{T-1}
S_T R_T
$1-\lambda$
$(1-\lambda)\lambda$
$(1-\lambda)\lambda^2$
λ^{T-t-1}

Tabular TD(λ)

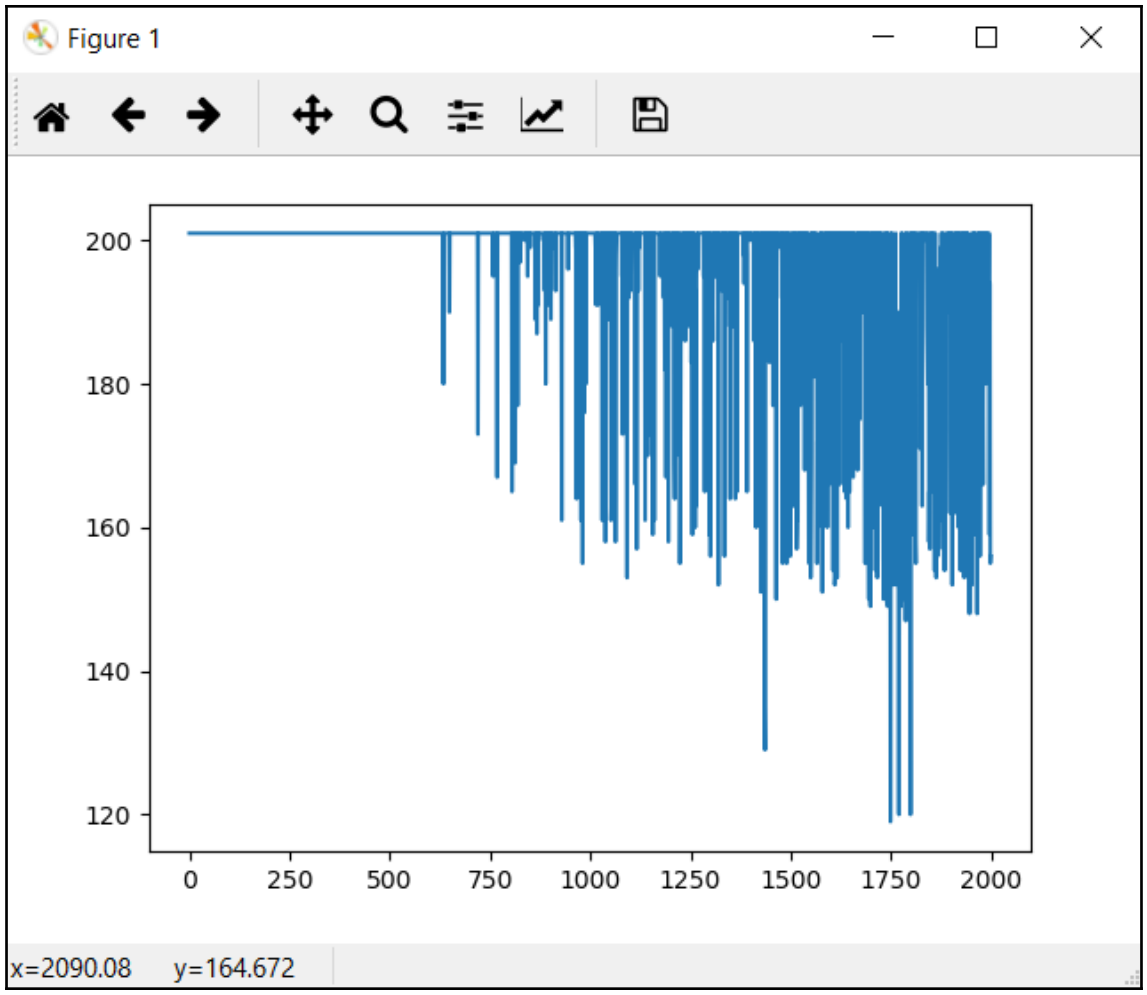
Initialize $V(s)$ arbitrarily and $e(s) = 0$, for all $s \in S$
Repeat (for each episode):
  Initialize s
  Repeat (for each step of episode) :
    $a \leftarrow$ action given by π for s
    Take action a, observe reward, r, and next state s'
    $\delta \leftarrow r + \gamma V(s') - V(s)$
    $e(s) \leftarrow e(s) + 1$
    For all s :
      $V(s) \leftarrow V(s) + \alpha\delta e(s)$
      $e(s) \leftarrow \gamma\lambda e(s)$
    $s \leftarrow s'$
  Until s is terminal

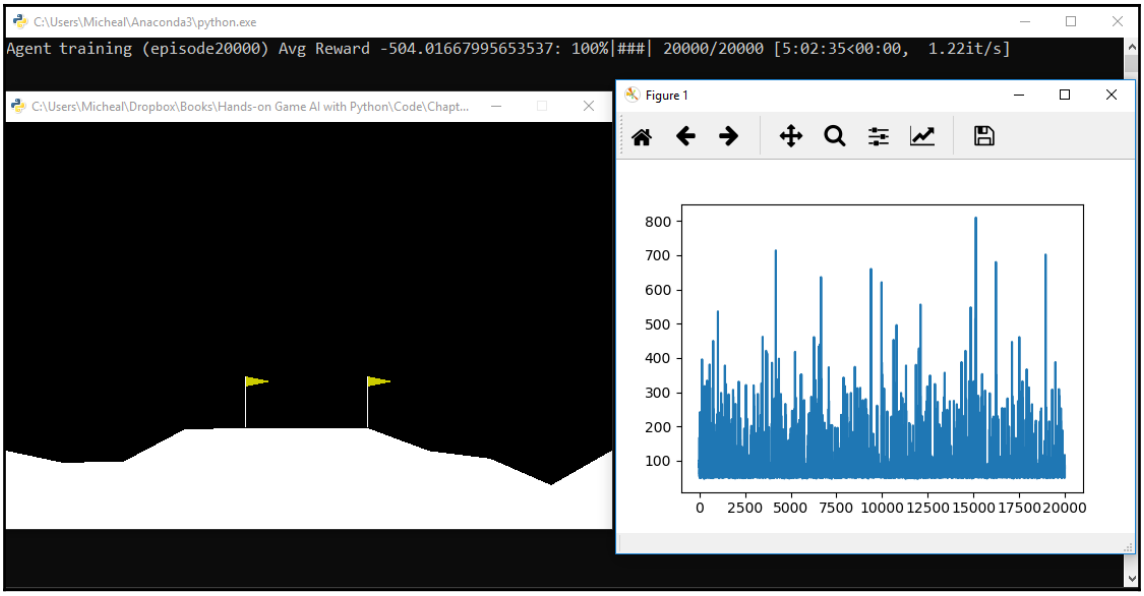
SARSA(λ)

Initialize $Q(s,a)$ arbitrarily and $e(s,a) = 0$, for all s, a
Repeat (for each episode):
  Initialize s, a
  Repeat (for each step of episode):
    Take action a, observe r, s'
    Choose a' from s' using policy derived from Q (e.g. ? - greedy)
    $\delta \leftarrow r + \gamma Q(s', a') - Q(s, a)$
    $e(s,a) \leftarrow e(s,a) + 1$
    For all s,a :
      $Q(s, a) \leftarrow Q(s, a) + \alpha\delta e(s, a)$
      $e(s, a) \leftarrow \gamma\lambda e(s, a)$
    $s \leftarrow s'; a \leftarrow a'$
  Until s is terminal

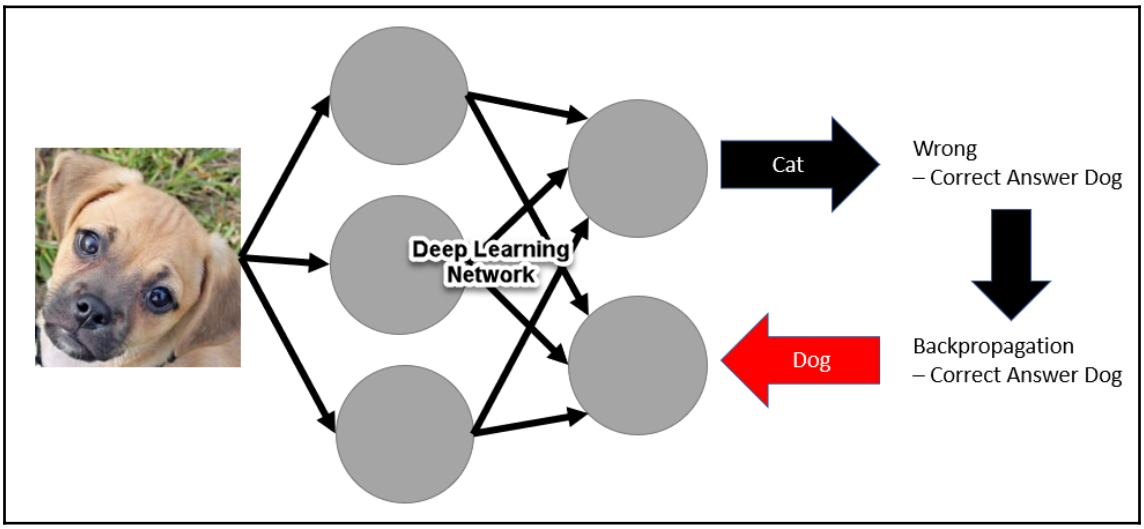
Q(λ)

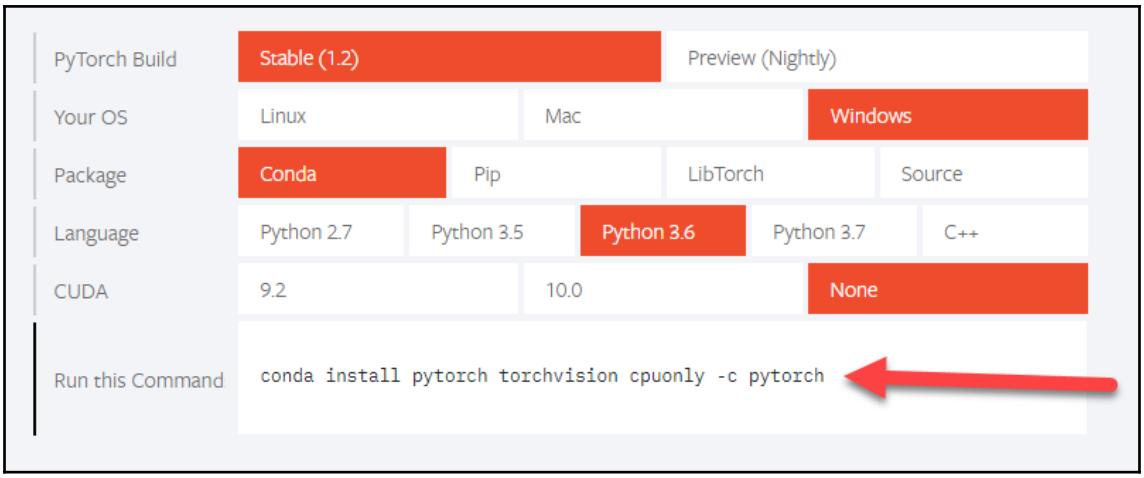
Initialize $Q(s,a)$ arbitrarily and $e(s,a) = 0$, for all s, a
Repeat (for each episode):
  Initialize s, a
  Repeat (for each step of episode):
    Take action a, observe r, s'
    Choose a' from s' using policy derived from Q (e.g. ? - greedy)
    $a^* \leftarrow \arg\max_b Q(s', b)$ (if a ties for the max, then $a^* \leftarrow a'$)
    $\delta \leftarrow r + \gamma Q(s', a') - Q(s, a^*)$
    $e(s,a) \leftarrow e(s,a) + 1$
    For all s,a :
      $Q(s, a) \leftarrow Q(s, a) + \alpha\delta e(s, a)$
      If $a' = a^*$, then $e(s, a) \leftarrow \gamma\lambda e(s,a)$
                  else $e(s, a) \leftarrow 0$
    $s \leftarrow s'; a \leftarrow a'$
  Until s is terminal

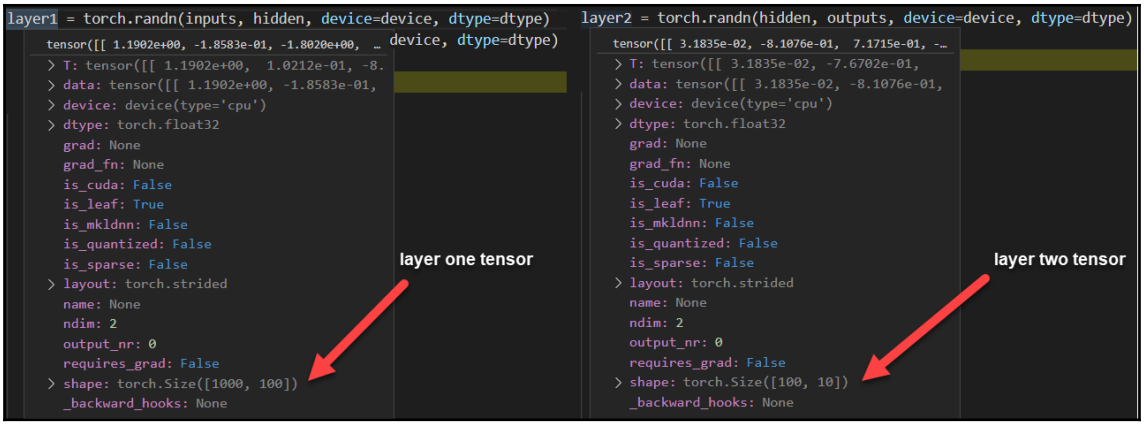



C:\Users\Micheal\Anaconda3\python.exe
Agent training (episode20000) Avg Reward -504.01667995653537: 100%|###| 20000/20000 [5:02:35<00:00, 1.22it/s]
C:\Users\Micheal\Dropbox\Books\Hands-on Game AI with Python\Code\Chapt...
Figure 1
800
700
600
500
400
300
200
100
0
2500
5000
7500
10000
12500
15000
17500
20000

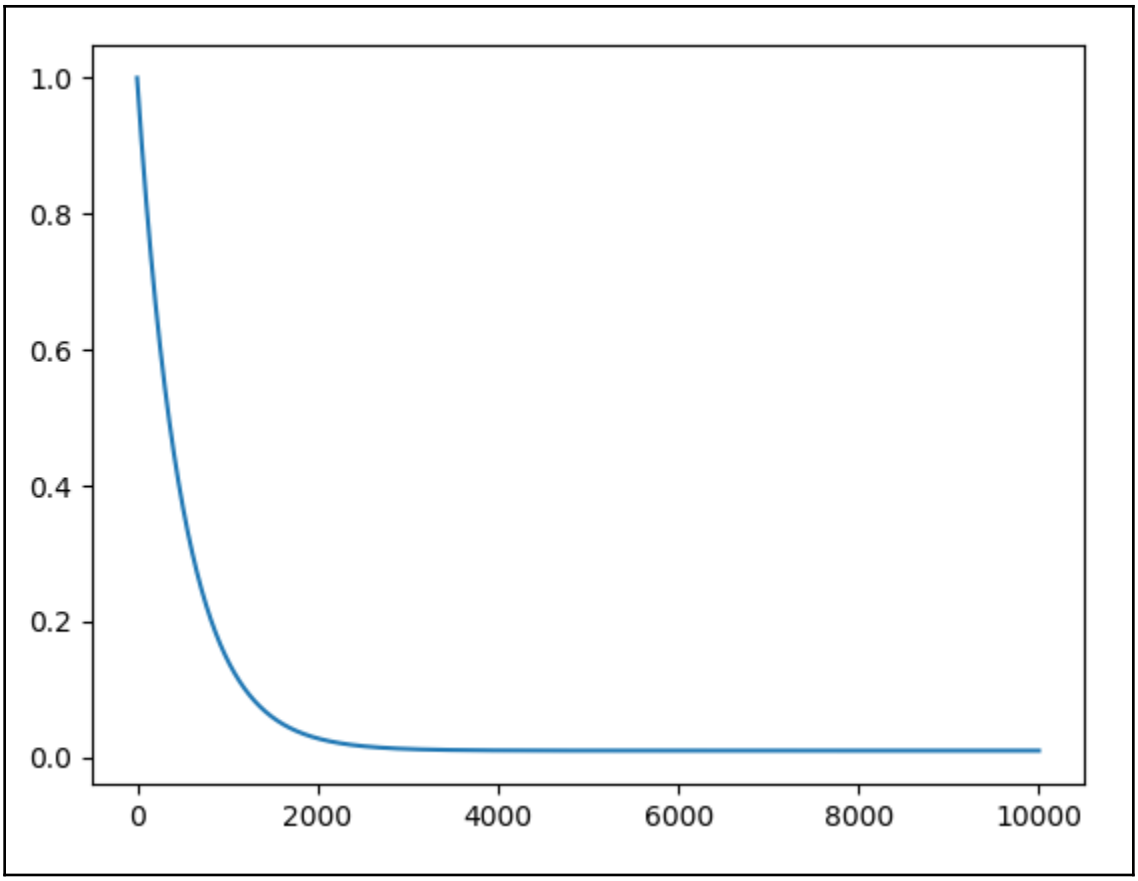
Chapter 6: Going Deep with DQN

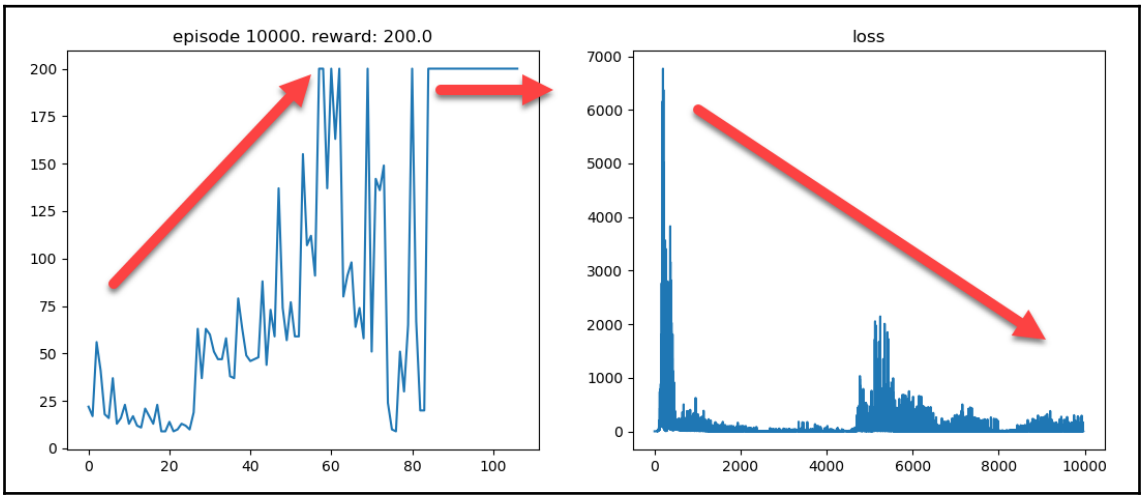


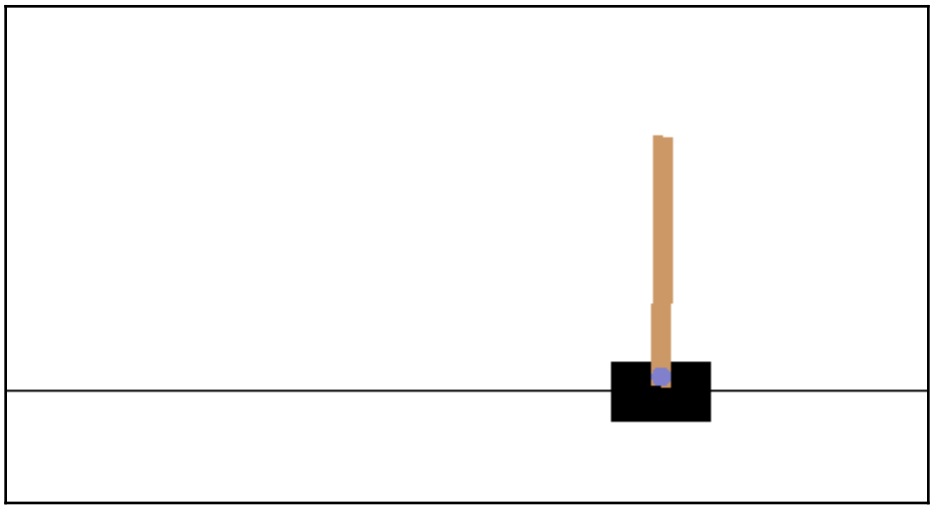


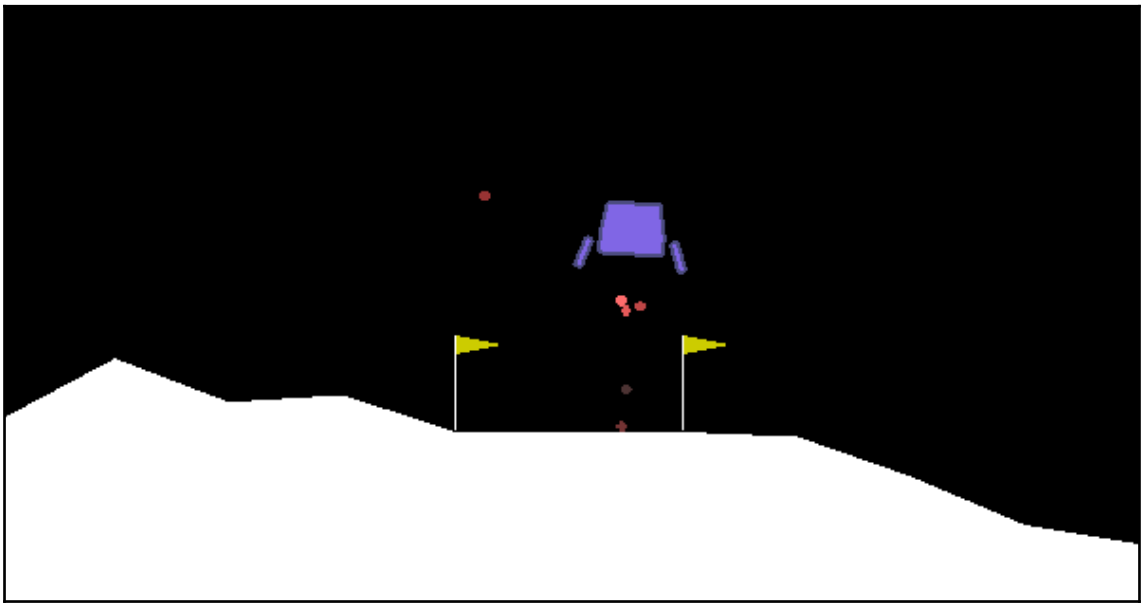



```
99 635.5177001953125
199 3.109295606613159
299 0.026640085503458977
399 0.0005209375522099435
499 6.454912363551557e-05
```

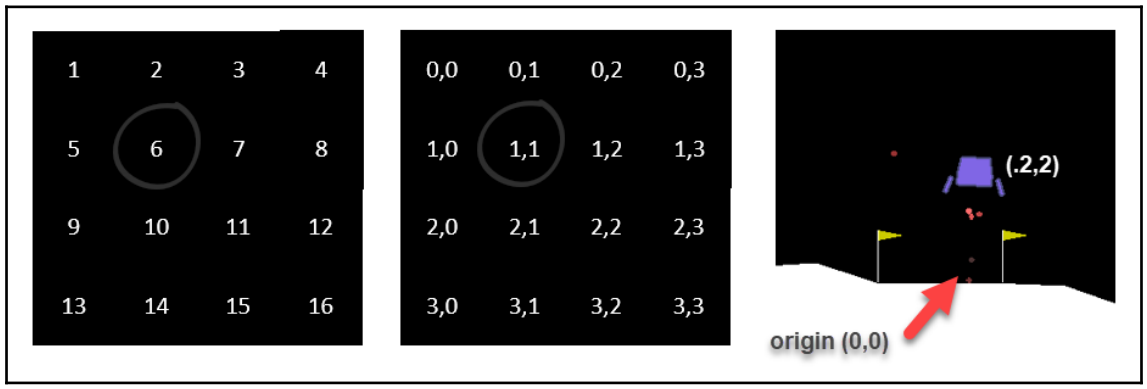

1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000

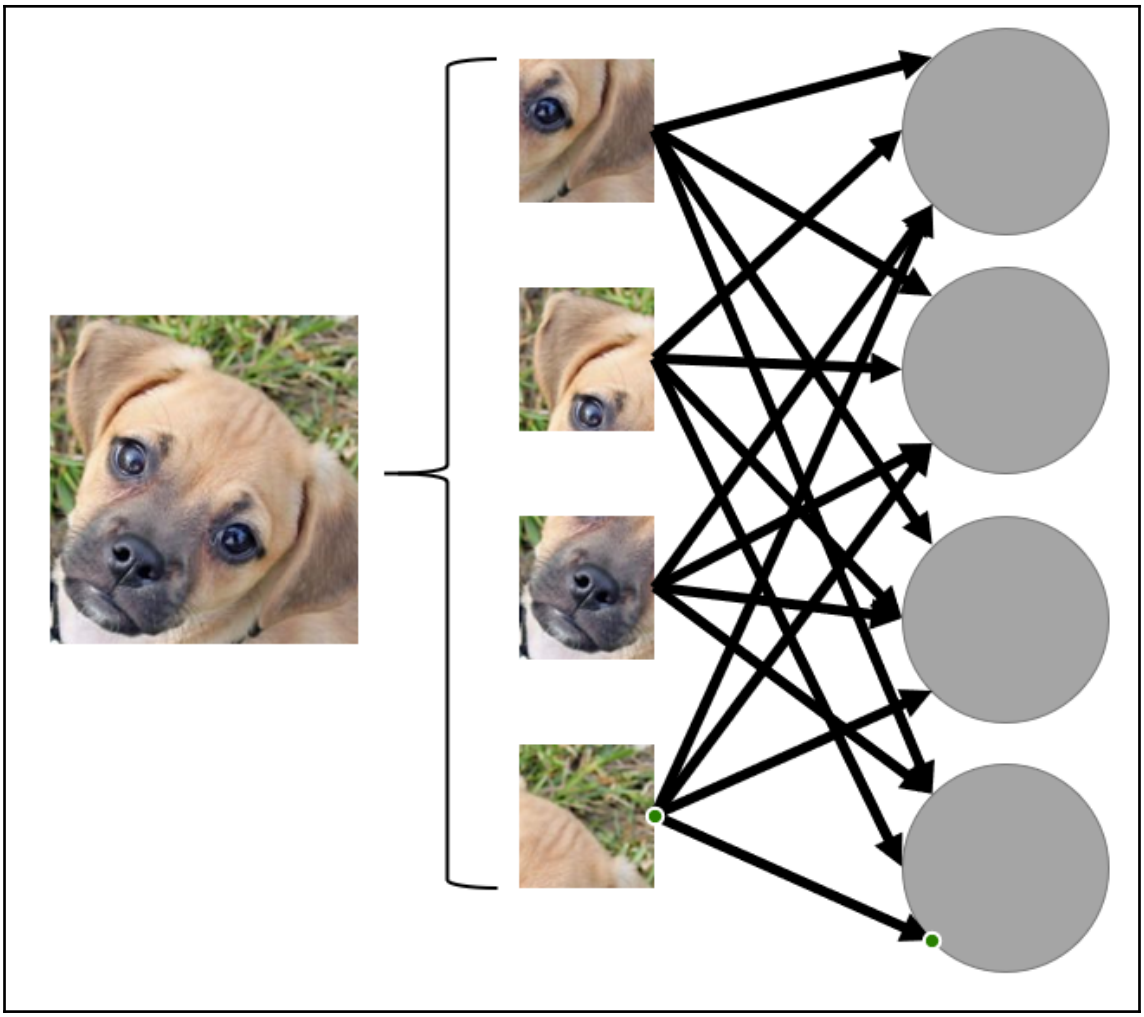

episode 10000. reward: 200.0
200
175
150
125
100
75
50
25
0
0
20
40
60
80
100
loss
7000
6000
5000
4000
3000
2000
1000
0
0
2000
4000
6000
8000
10000

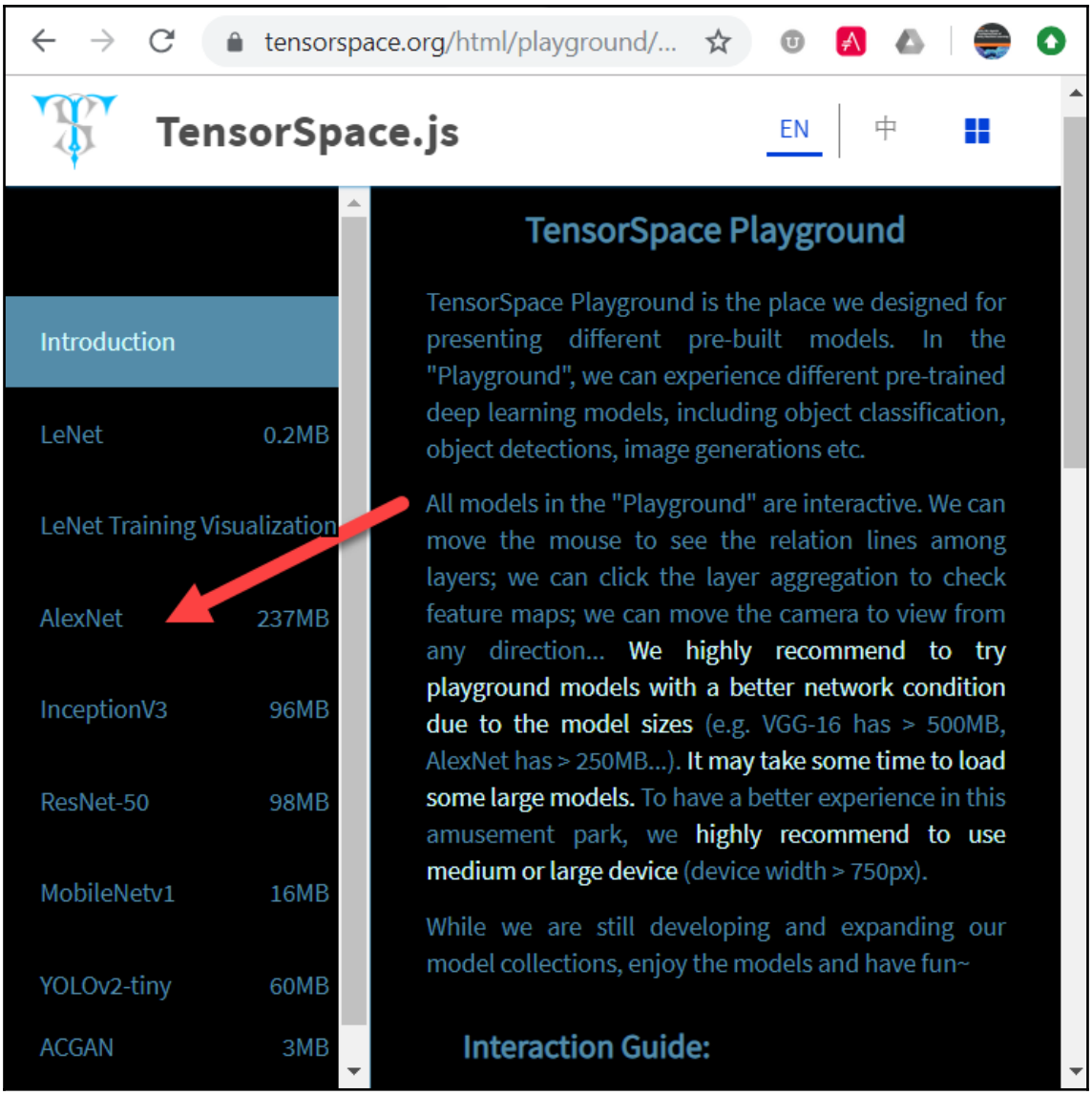


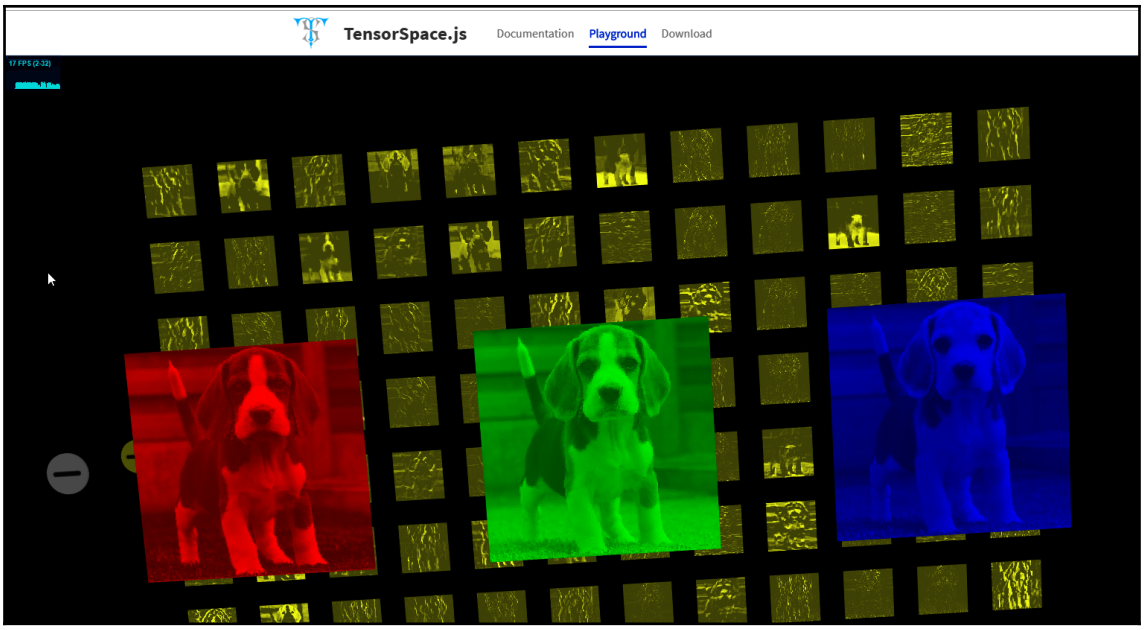


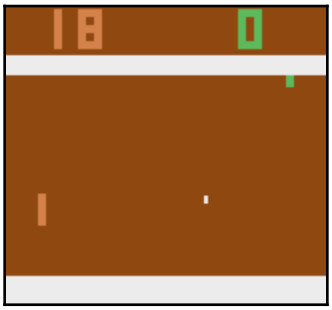
Chapter 7: Going Deeper with DDQN

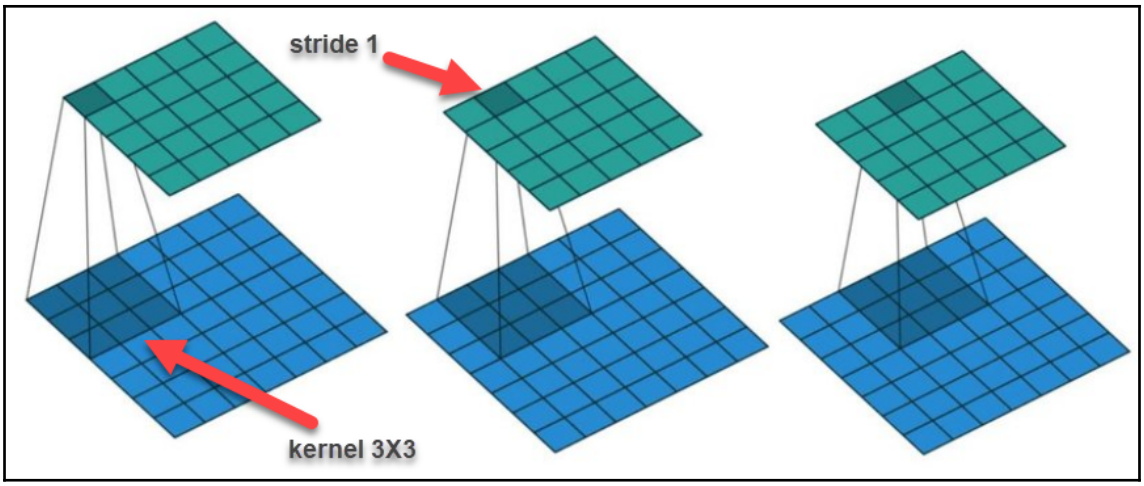


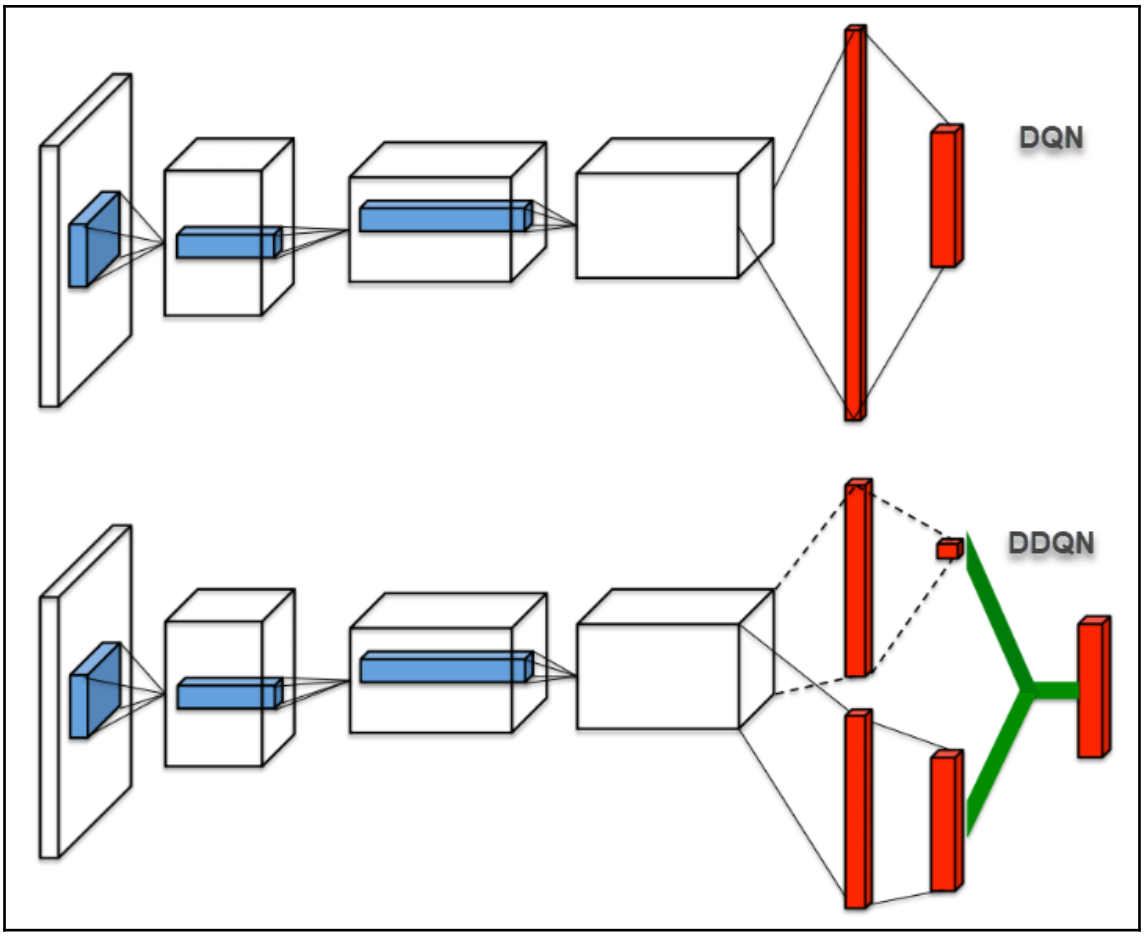


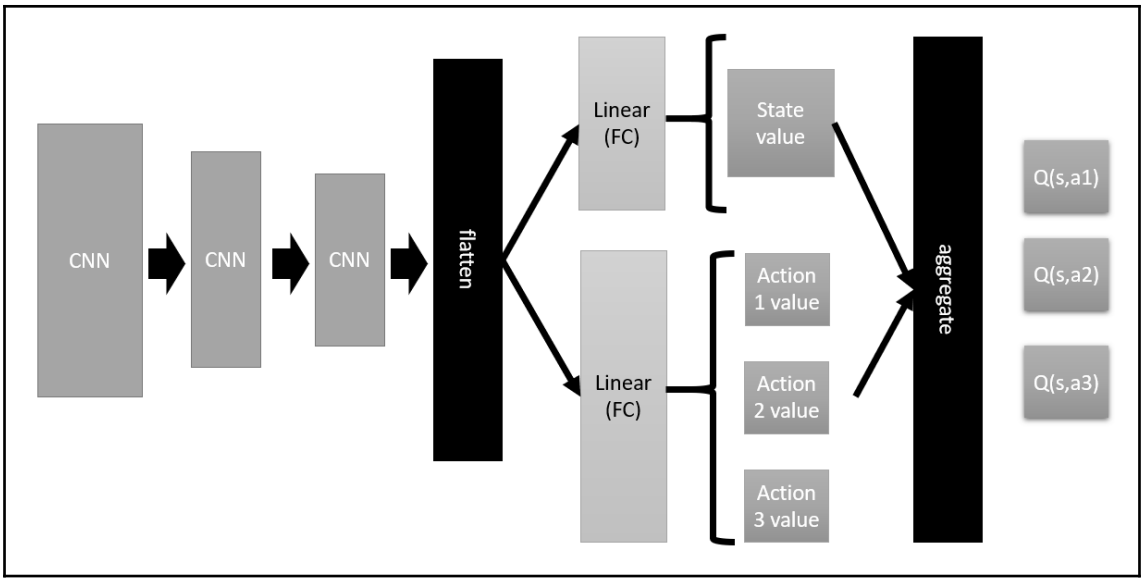

tensorspace.org/html/playground/...
TensorSpace.js
EN
中
Introduction
LeNet 0.2MB
LeNet Training Visualization
AlexNet 237MB
InceptionV3 96MB
ResNet-50 98MB
MobileNetv1 16MB
YOLOv2-tiny 60MB
ACGAN 3MB
TensorSpace Playground
TensorSpace Playground is the place we designed for presenting different pre-built models. In the "Playground", we can experience different pre-trained deep learning models, including object classification, object detections, image generations etc.
All models in the "Playground" are interactive. We can move the mouse to see the relation lines among layers; we can click the layer aggregation to check feature maps; we can move the camera to view from any direction... We highly recommend to try playground models with a better network condition due to the model sizes (e.g. VGG-16 has > 500MB, AlexNet has > 250MB...). It may take some time to load some large models. To have a better experience in this amusement park, we highly recommend to use medium or large device (device width > 750px).
While we are still developing and expanding our model collections, enjoy the models and have fun~
Interaction Guide:

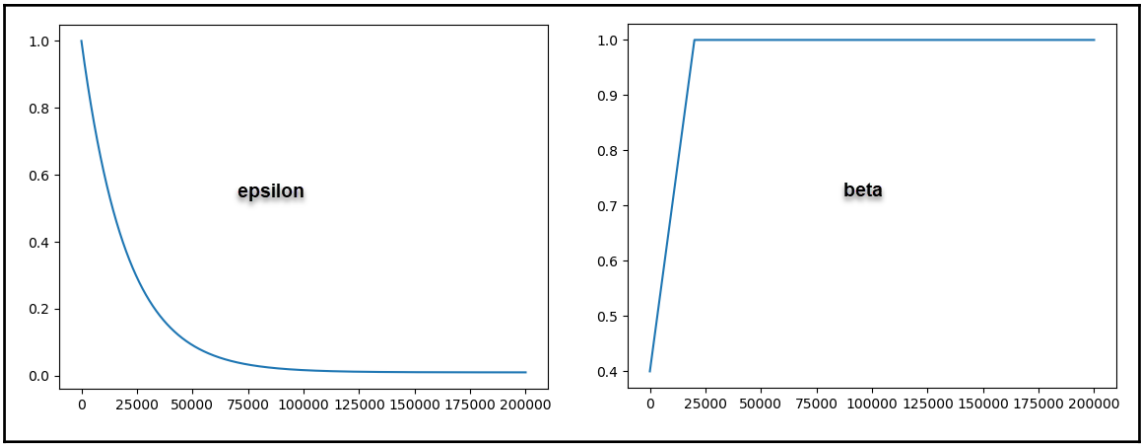

TensorSpace.js
Documentation
Playground
Download
17 FPS (2-32)

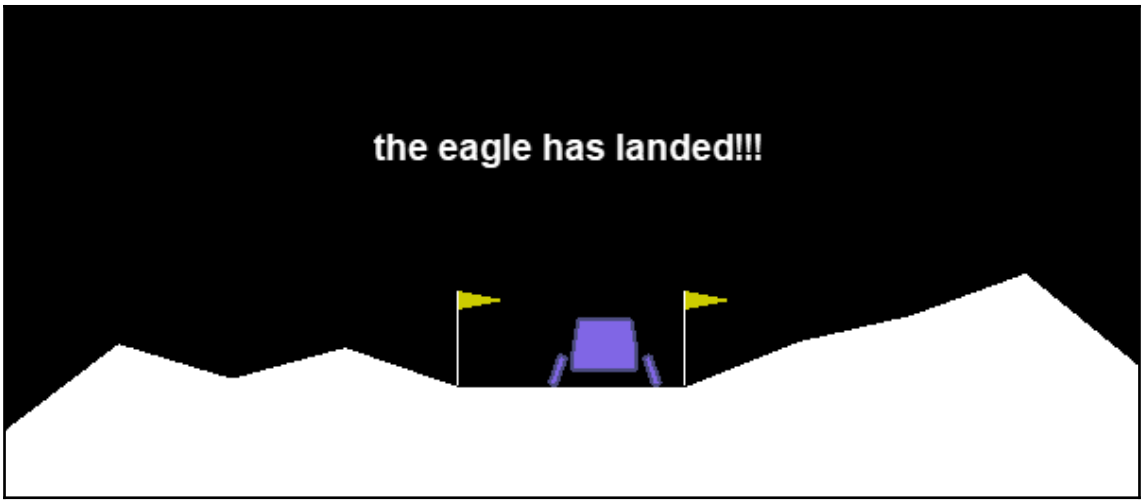



stride 1
kernel 3X3

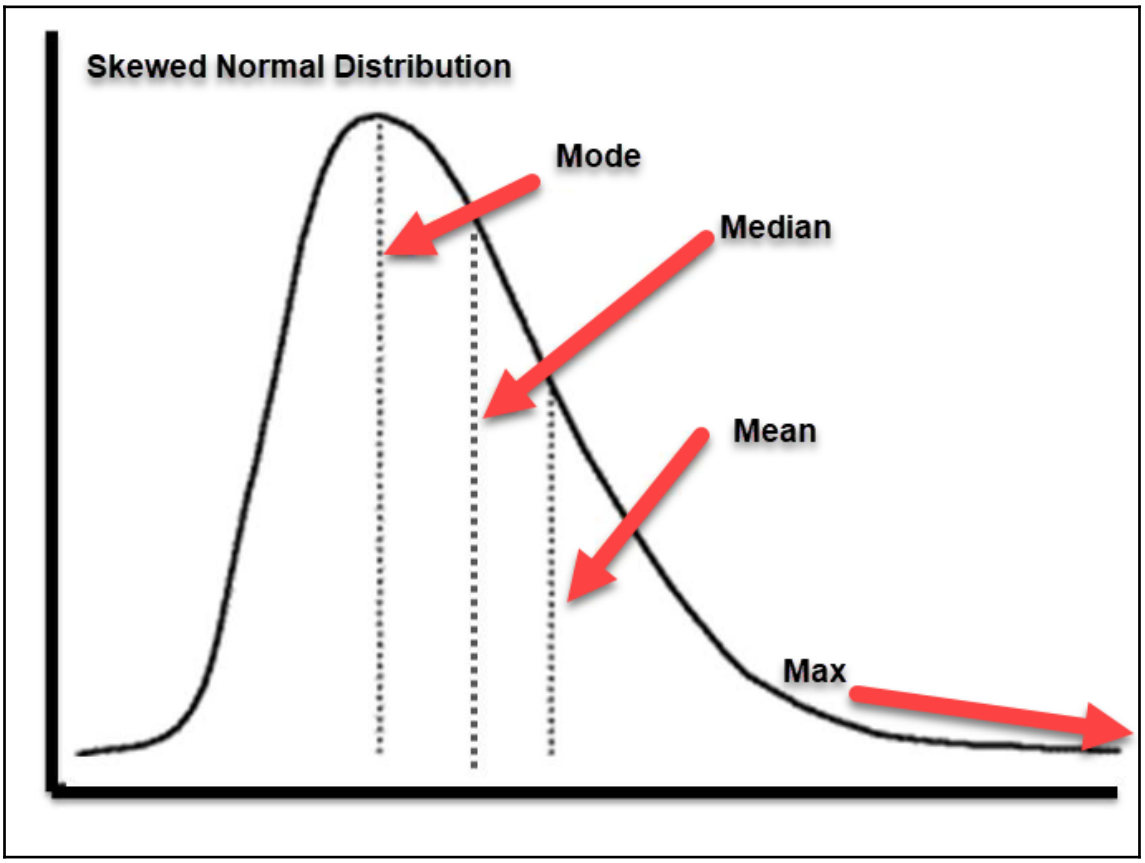

DQN
DDQN

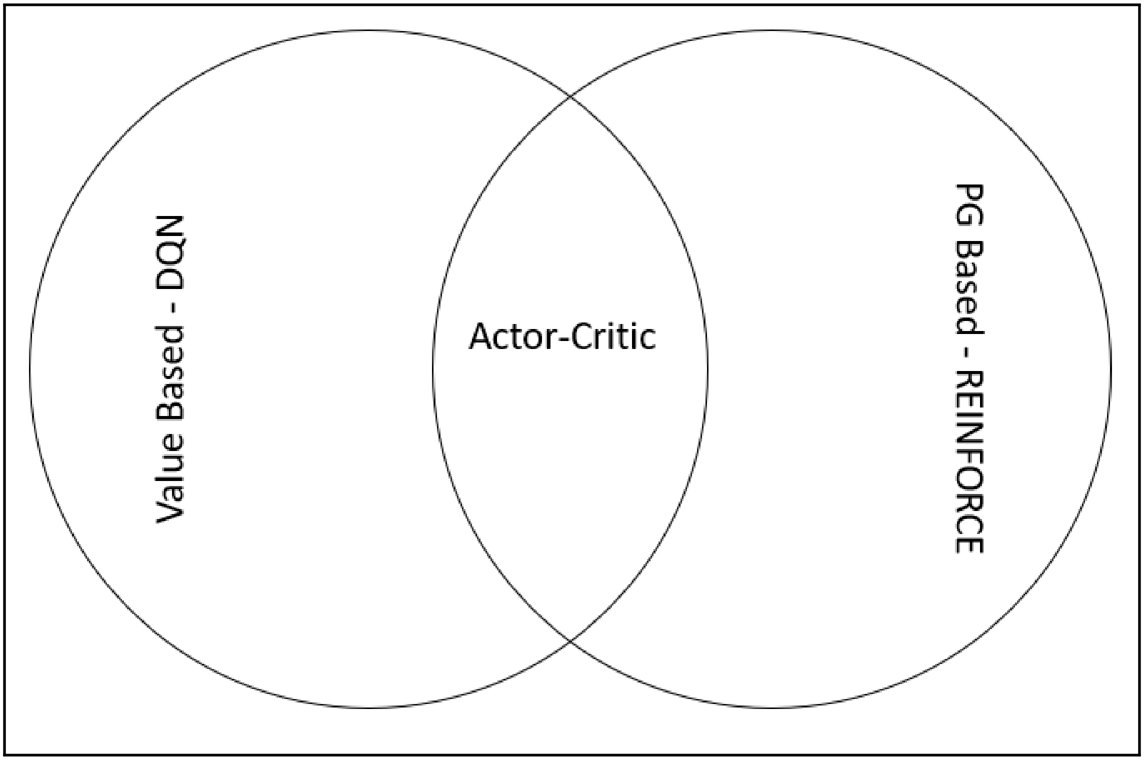

CNN
CNN
CNN
flatten
Linear (FC)
Linear (FC)
State value
Action 1 value
Action 2 value
Action 3 value
aggregate
Q(s,a1)
Q(s,a2)
Q(s,a3)

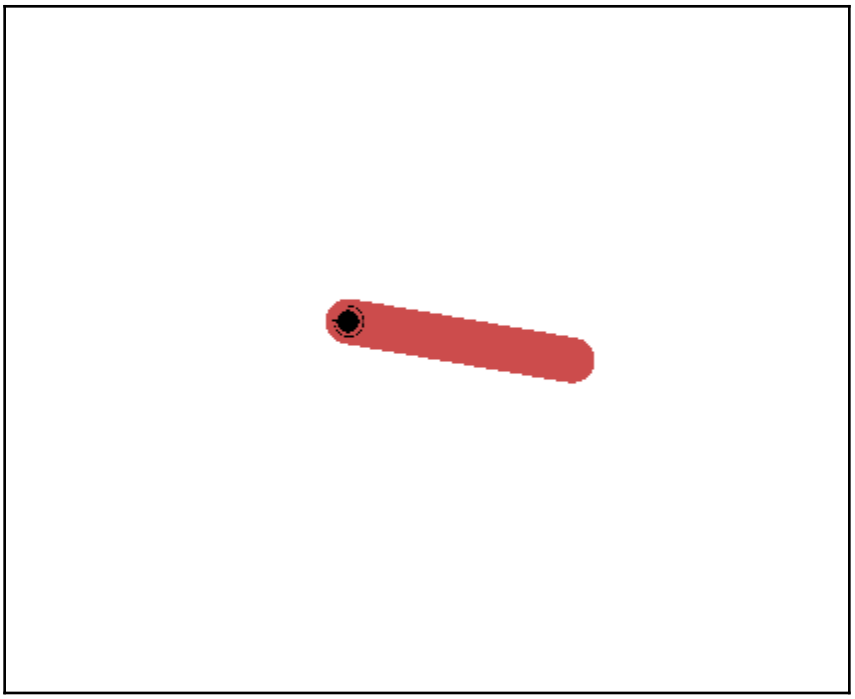

epsilon
1.0
0.8
0.6
0.4
0.2
0.0
0
25000
50000
75000
100000
125000
150000
175000
200000
beta
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
25000
50000
75000
100000
125000
150000
175000
200000

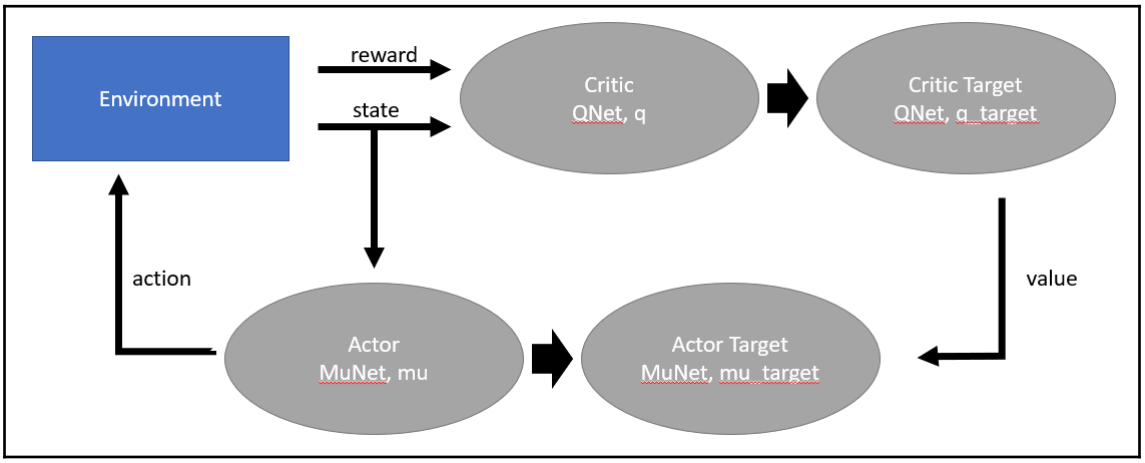

the eagle has landed!!!

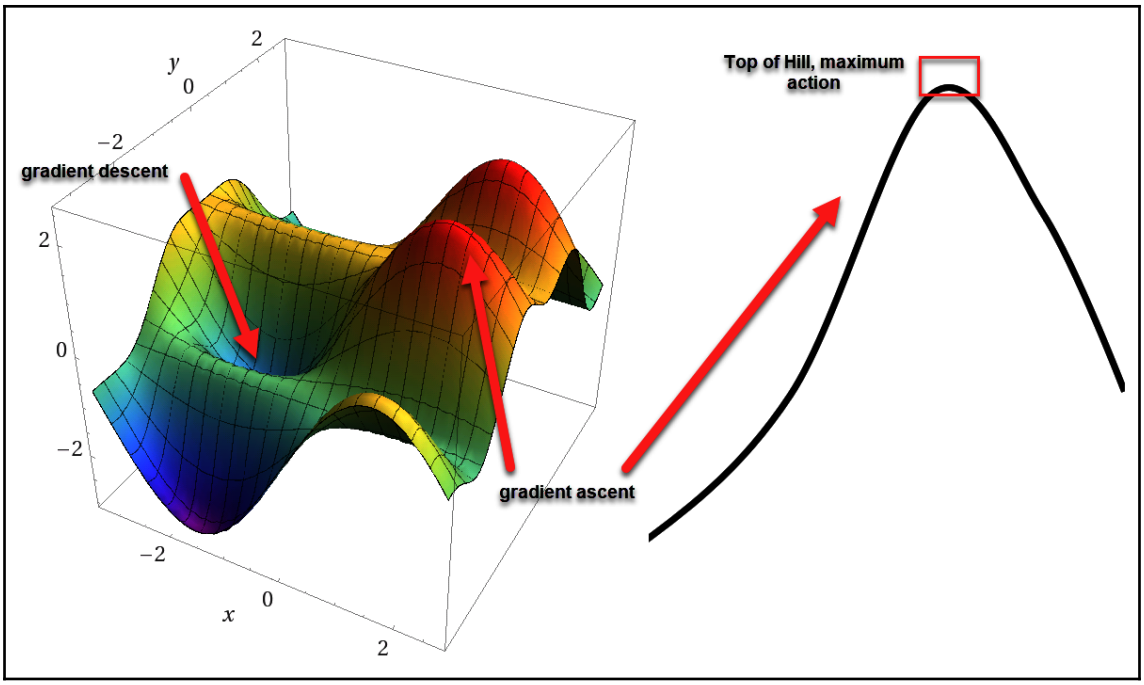
Chapter 8: Policy Gradient Methods

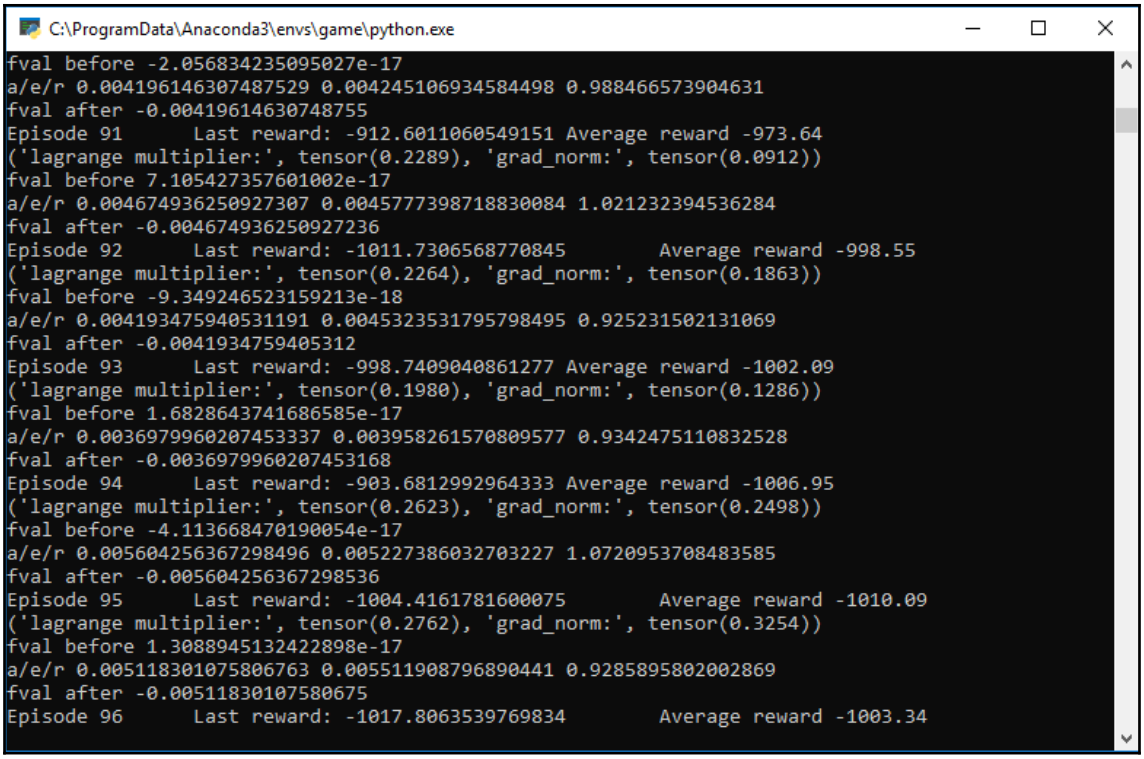



Value Based - DQN
Actor-Critic
PG Based - REINFORCE




Environment
reward
state
Critic
QNet, q
Critic Target
QNet, q_target
action
value
Actor
MuNet, mu
Actor Target
MuNet, mu_target

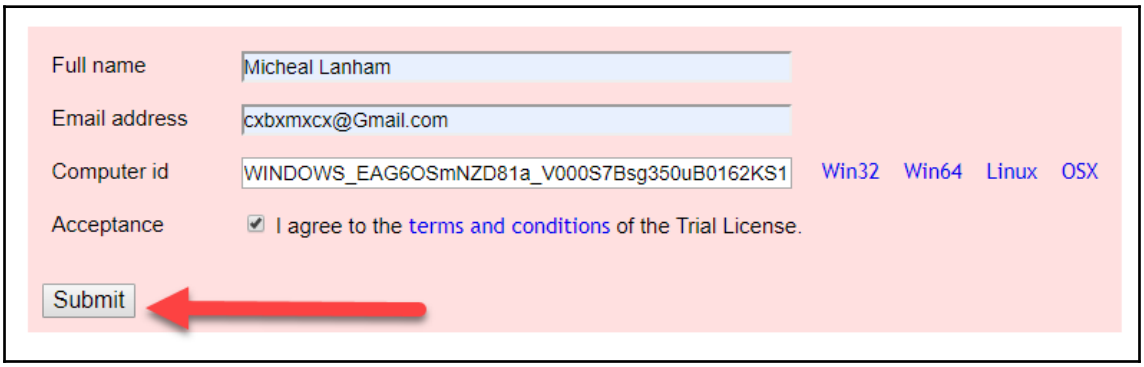

gradient descent
gradient ascent
Top of Hill, maximum action
x
y
2
0
−2

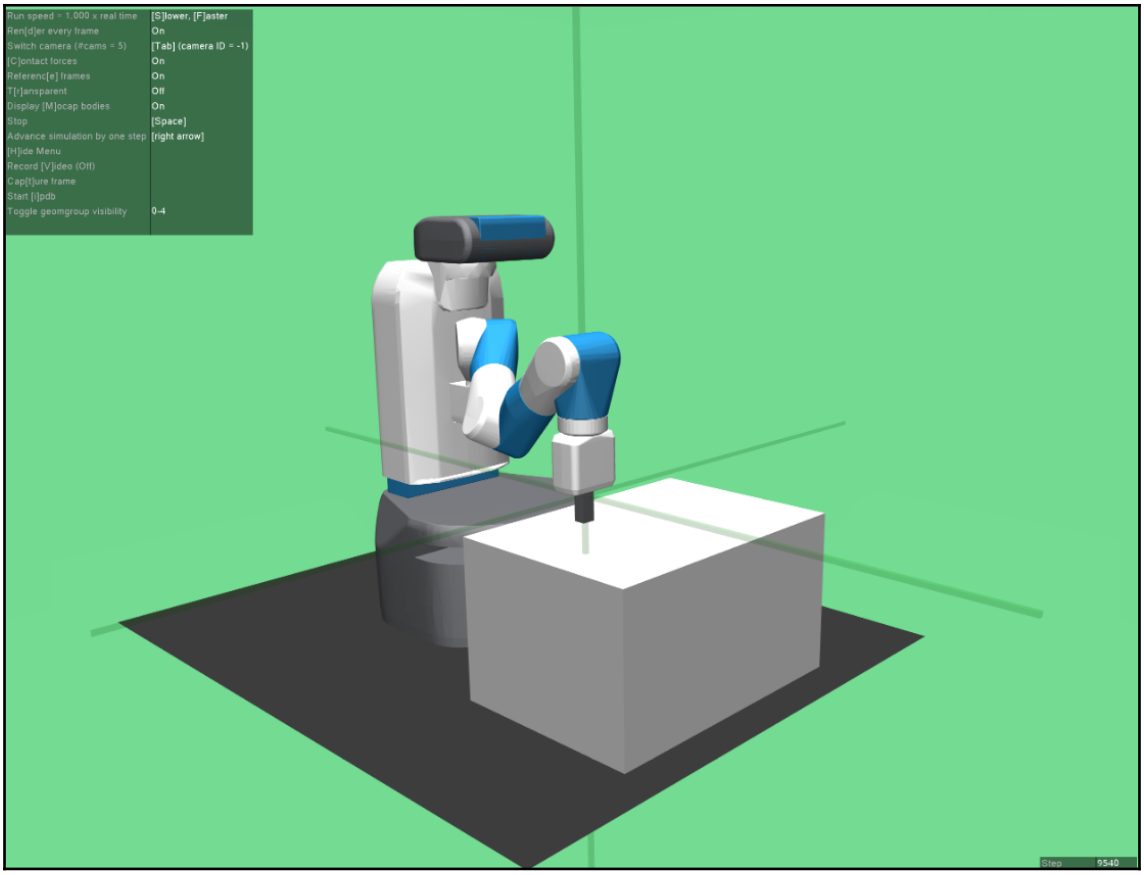

C:\ProgramData\Anaconda3\envs\game\python.exe
fval before -2.056834235095027e-17
a/e/r 0.004196146307487529 0.004245106934584498 0.988466573904631
fval after -0.00419614630748755
Episode 91 Last reward: -912.6011060549151 Average reward -973.64
('lagrange multiplier:', tensor(0.2289), 'grad_norm:', tensor(0.0912))
fval before 7.105427357601002e-17
a/e/r 0.004674936250927307 0.0045777398718830084 1.021232394536284
fval after -0.004674936250927236
Episode 92 Last reward: -1011.7306568770845 Average reward -998.55
('lagrange multiplier:', tensor(0.2264), 'grad_norm:', tensor(0.1863))
fval before -9.349246523159213e-18
a/e/r 0.004193475940531191 0.0045323531795798495 0.925231502131069
fval after -0.0041934759405312
Episode 93 Last reward: -998.7409040861277 Average reward -1002.09
('lagrange multiplier:', tensor(0.1980), 'grad_norm:', tensor(0.1286))
fval before 1.6828643741686585e-17
a/e/r 0.0036979960207453337 0.003958261570809577 0.9342475110832528
fval after -0.003697996020745316
Episode 94 Last reward: -903.6812992964333 Average reward -1006.95
('lagrange multiplier:', tensor(0.2623), 'grad_norm:', tensor(0.2498))
fval before -4.113668470190054e-17
a/e/r 0.005604256367298496 0.005227386032703227 1.0720953708483585
fval after -0.005604256367298536
Episode 95 Last reward: -1004.4161781600075 Average reward -1010.09
('lagrange multiplier:', tensor(0.2762), 'grad_norm:', tensor(0.3254))
fval before 1.3088945132422898e-17
a/e/r 0.005118301075806763 0.005511908796890441 0.9285895802002869
fval after -0.00511830107580675
Episode 96 Last reward: -1017.8063539769834 Average reward -1003.34

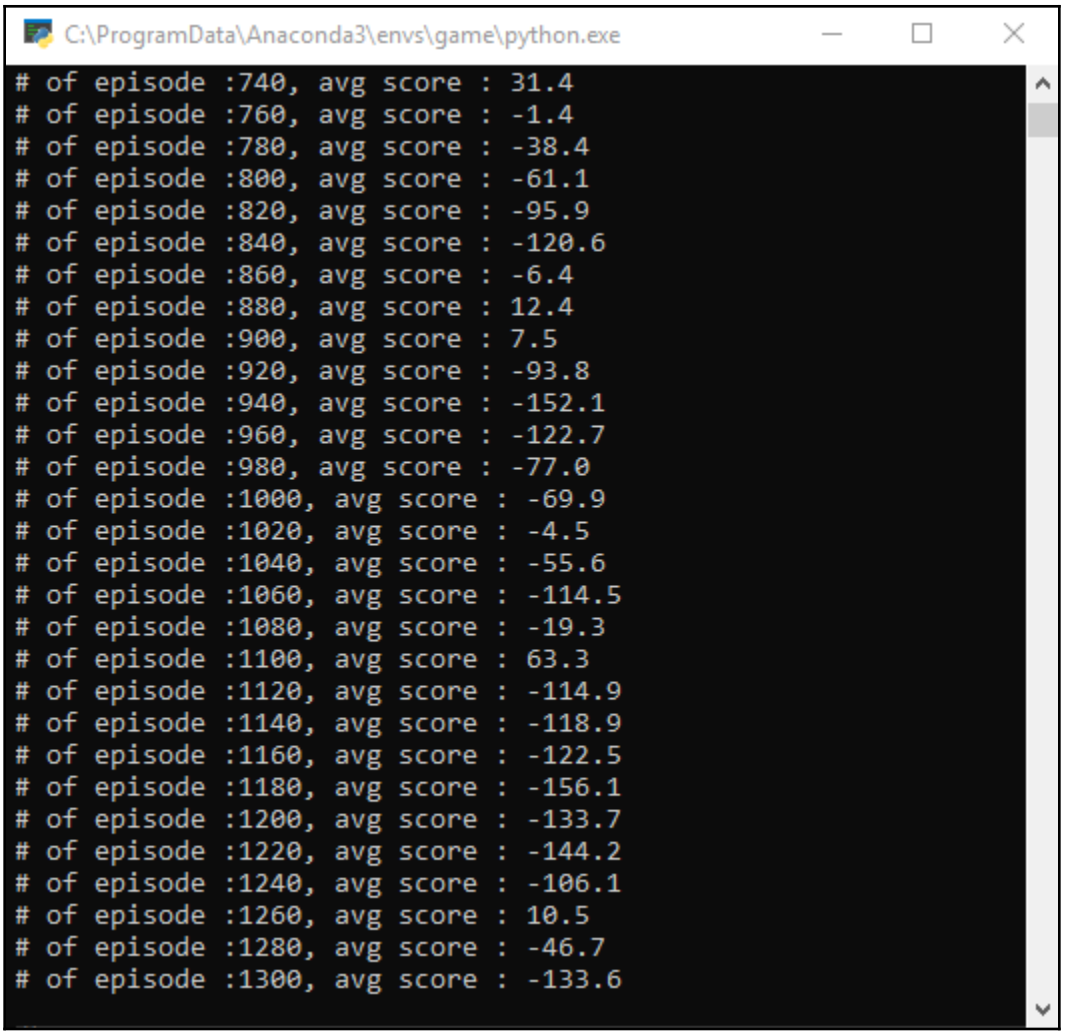
Line search
(like gradient ascent)

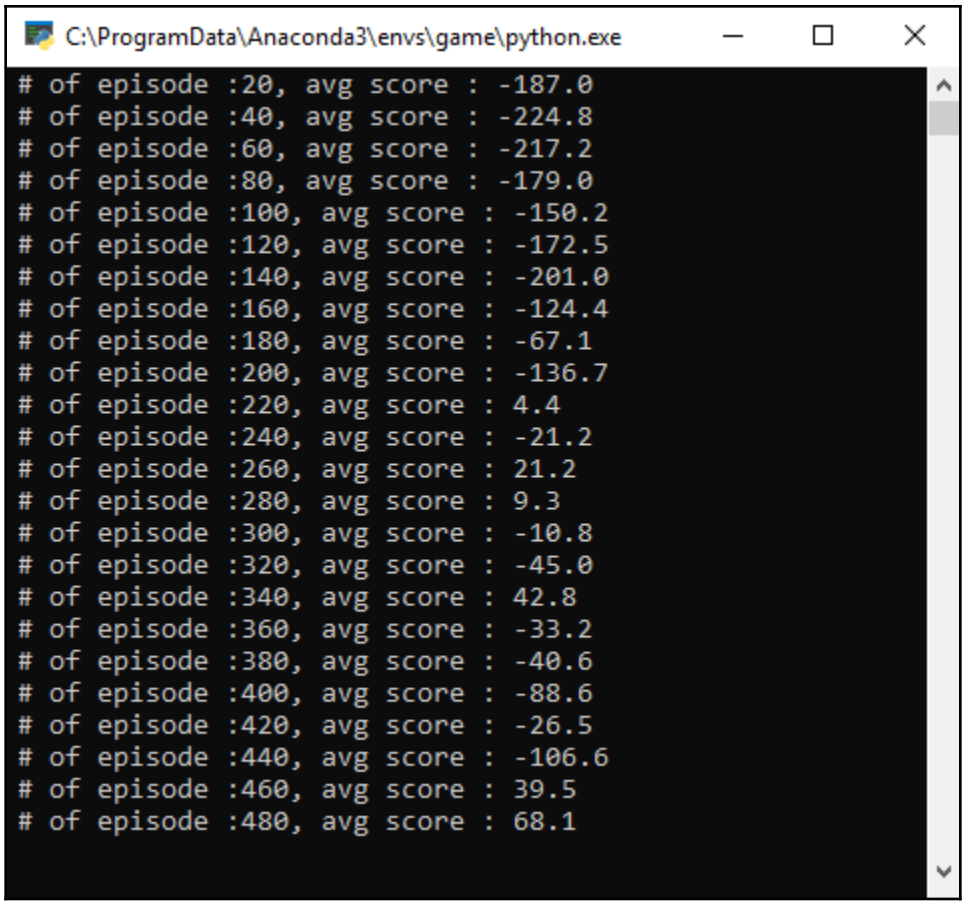
Trust region

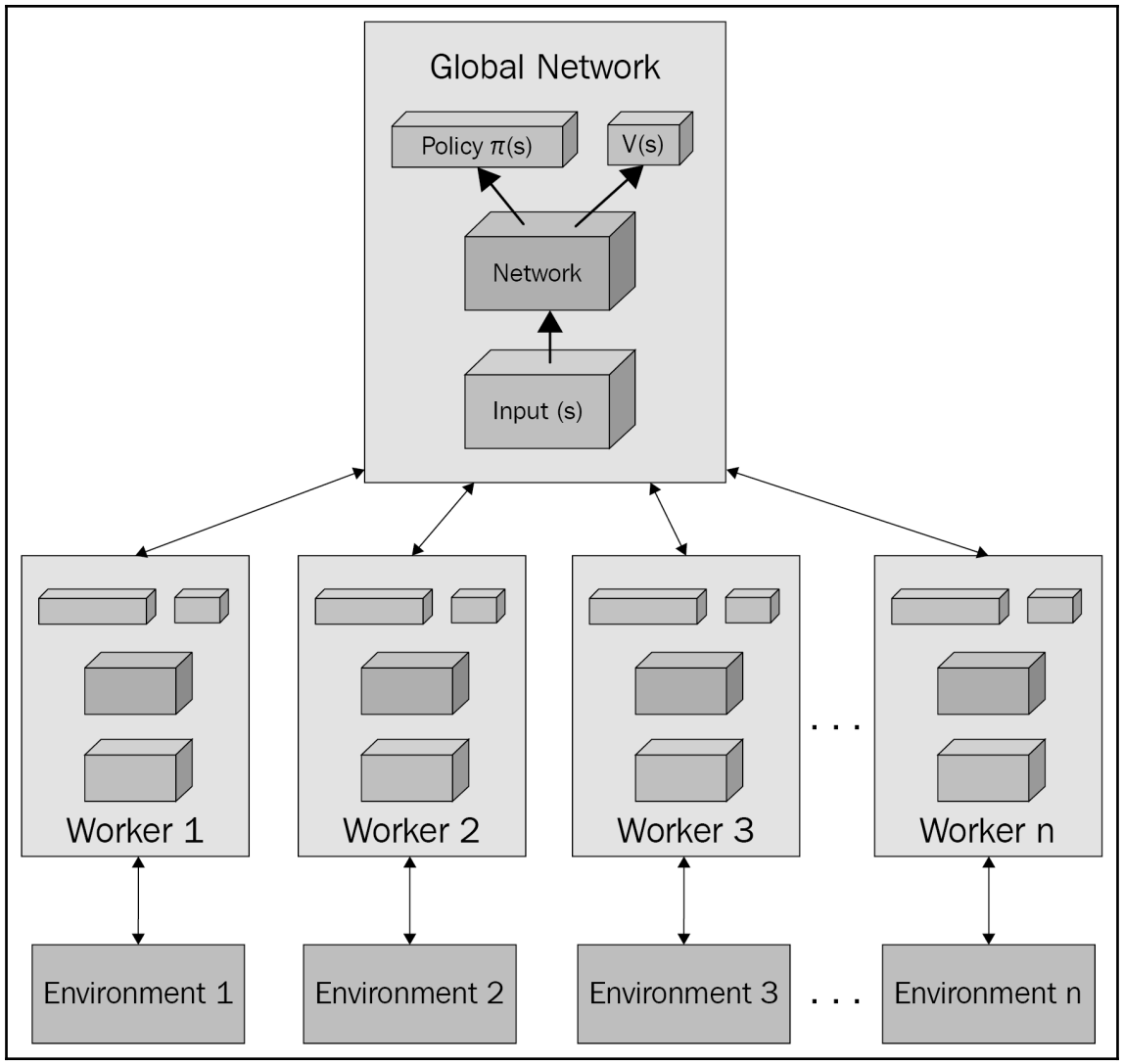
Chapter 9: Optimizing for Continuous Control

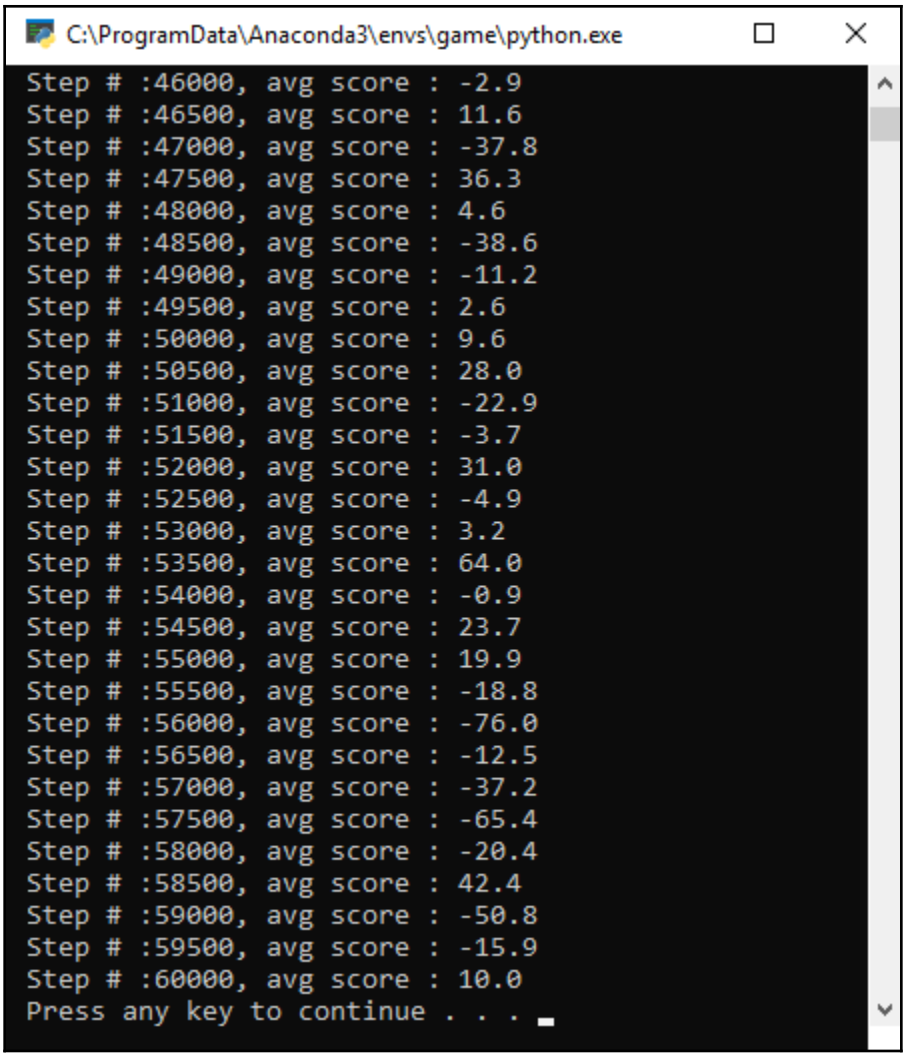


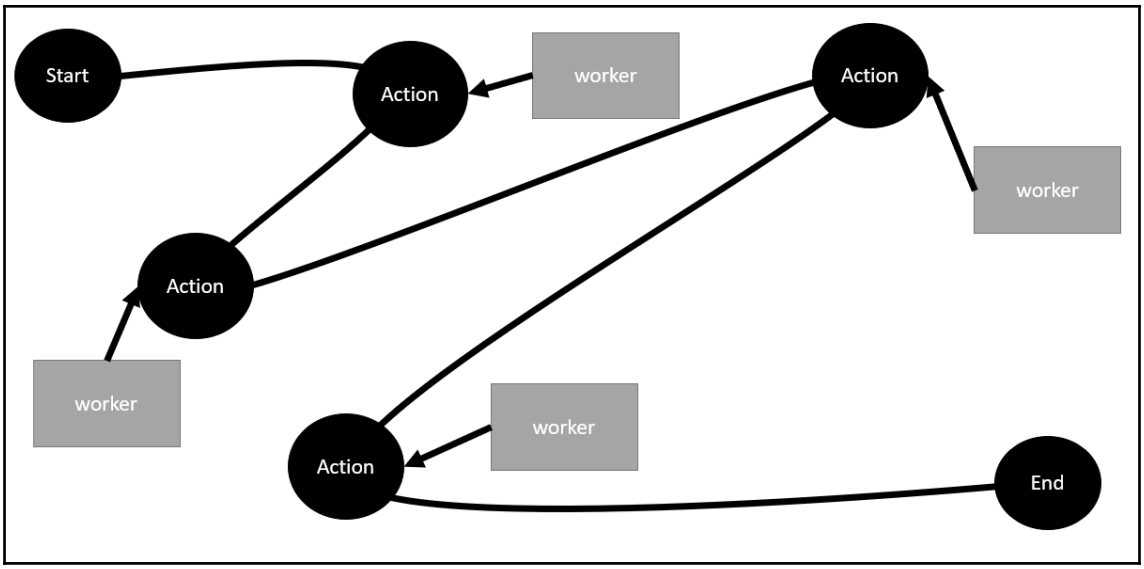

Run speed = 1.000 x real time
[S]lower, [F]aster
Ren[d]er every frame
On
Switch camera (#cams = 5)
[Tab] (camera ID = -1)
[C]ontact forces
On
Referenc[e] frames
On
T[r]ansparent
Off
Display [M]ocap bodies
On
Stop
[Space]
Advance simulation by one step
[right arrow]
[H]ide Menu
Record [V]ideo (Off)
Cap[t]ure frame
Start [i]pdb
Toggle geomgroup visibility
0-4
Step
9540

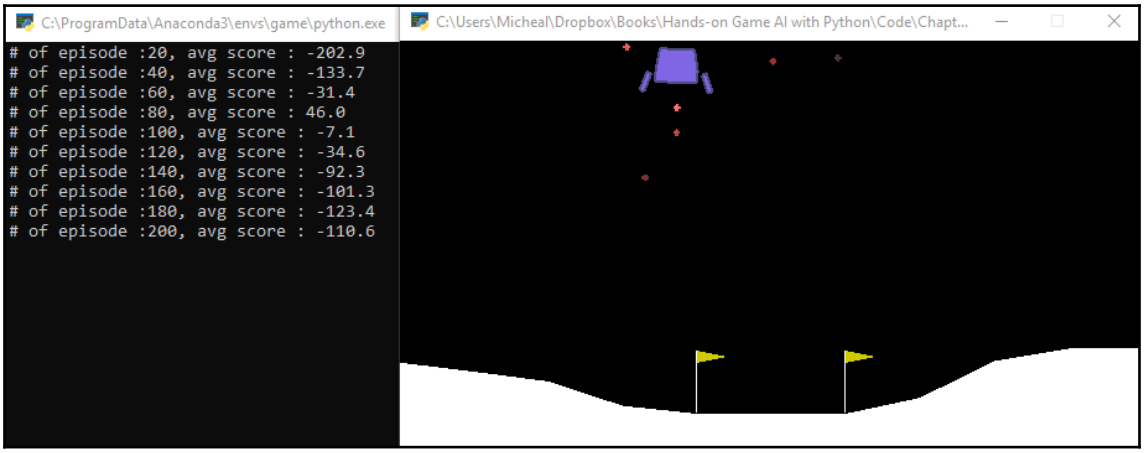

C:\ProgramData\Anaconda3\envs\game\python.exe
# of episode :740, avg score : 31.4
# of episode :760, avg score : -1.4
# of episode :780, avg score : -38.4
# of episode :800, avg score : -61.1
# of episode :820, avg score : -95.9
# of episode :840, avg score : -120.6
# of episode :860, avg score : -6.4
# of episode :880, avg score : 12.4
# of episode :900, avg score : 7.5
# of episode :920, avg score : -93.8
# of episode :940, avg score : -152.1
# of episode :960, avg score : -122.7
# of episode :980, avg score : -77.0
# of episode :1000, avg score : -69.9
# of episode :1020, avg score : -4.5
# of episode :1040, avg score : -55.6
# of episode :1060, avg score : -114.5
# of episode :1080, avg score : -19.3
# of episode :1100, avg score : 63.3
# of episode :1120, avg score : -114.9
# of episode :1140, avg score : -118.9
# of episode :1160, avg score : -122.5
# of episode :1180, avg score : -156.1
# of episode :1200, avg score : -133.7
# of episode :1220, avg score : -144.2
# of episode :1240, avg score : -106.1
# of episode :1260, avg score : 10.5
# of episode :1280, avg score : -46.7
# of episode :1300, avg score : -133.6

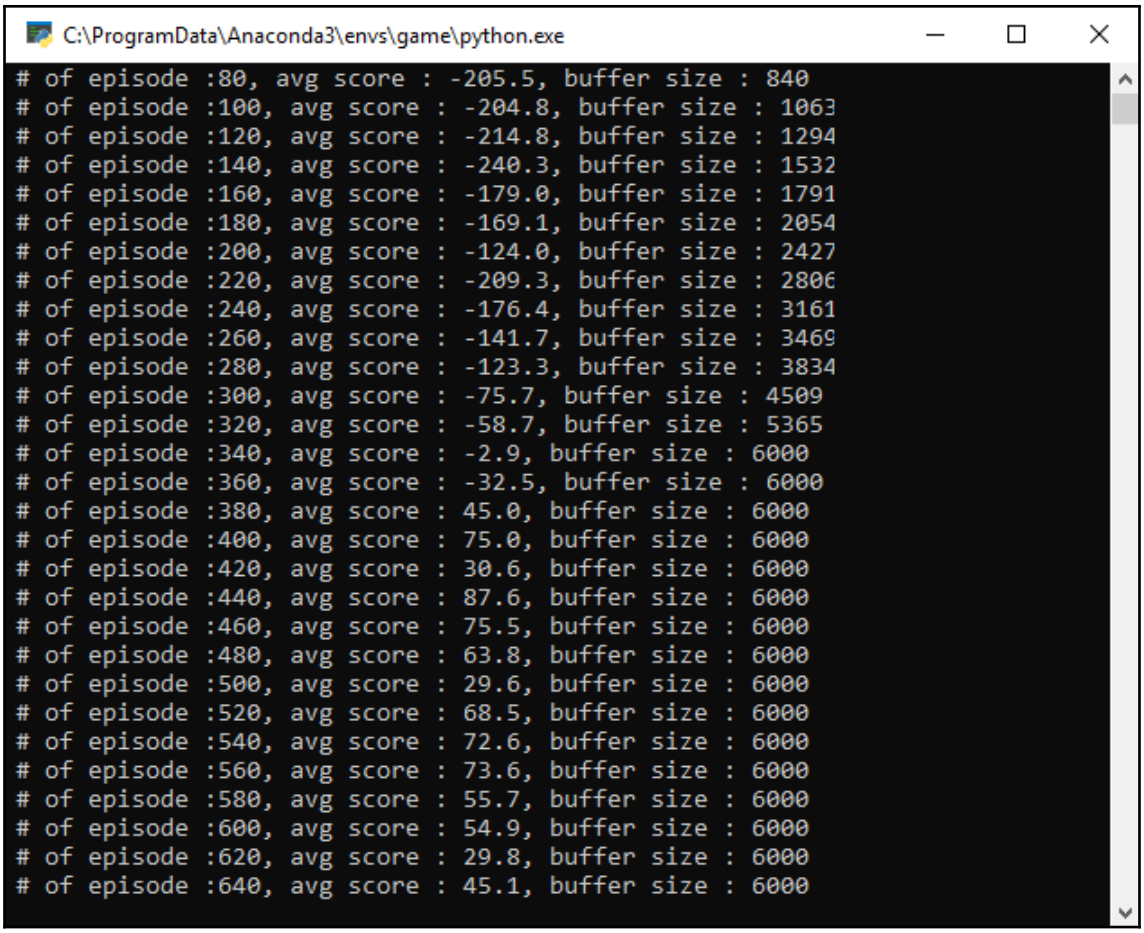

C:\ProgramData\Anaconda3\envs\game\python.exe
# of episode :20, avg score : -187.0
# of episode :40, avg score : -224.8
# of episode :60, avg score : -217.2
# of episode :80, avg score : -179.0
# of episode :100, avg score : -150.2
# of episode :120, avg score : -172.5
# of episode :140, avg score : -201.0
# of episode :160, avg score : -124.4
# of episode :180, avg score : -67.1
# of episode :200, avg score : -136.7
# of episode :220, avg score : 4.4
# of episode :240, avg score : -21.2
# of episode :260, avg score : 21.2
# of episode :280, avg score : 9.3
# of episode :300, avg score : -10.8
# of episode :320, avg score : -45.0
# of episode :340, avg score : 42.8
# of episode :360, avg score : -33.2
# of episode :380, avg score : -40.6
# of episode :400, avg score : -88.6
# of episode :420, avg score : -26.5
# of episode :440, avg score : -106.6
# of episode :460, avg score : 39.5
# of episode :480, avg score : 68.1


Global Network
Policy π(s)
V(s)
Network
Input (s)
Worker 1
Worker 2
Worker 3
Worker n
Environment 1
Environment 2
Environment 3
Environment n

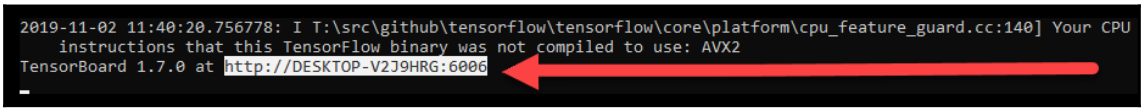

C:\ProgramData\Anaconda3\envs\game\python.exe
Step # :46000, avg score : -2.9
Step # :46500, avg score : 11.6
Step # :47000, avg score : -37.8
Step # :47500, avg score : 36.3
Step # :48000, avg score : 4.6
Step # :48500, avg score : -38.6
Step # :49000, avg score : -11.2
Step # :49500, avg score : 2.6
Step # :50000, avg score : 9.6
Step # :50500, avg score : 28.0
Step # :51000, avg score : -22.9
Step # :51500, avg score : -3.7
Step # :52000, avg score : 31.0
Step # :52500, avg score : -4.9
Step # :53000, avg score : 3.2
Step # :53500, avg score : 64.0
Step # :54000, avg score : -0.9
Step # :54500, avg score : 23.7
Step # :55000, avg score : 19.9
Step # :55500, avg score : -18.8
Step # :56000, avg score : -76.0
Step # :56500, avg score : -12.5
Step # :57000, avg score : -37.2
Step # :57500, avg score : -65.4
Step # :58000, avg score : -20.4
Step # :58500, avg score : 42.4
Step # :59000, avg score : -50.8
Step # :59500, avg score : -15.9
Step # :60000, avg score : 10.0
Press any key to continue . . .

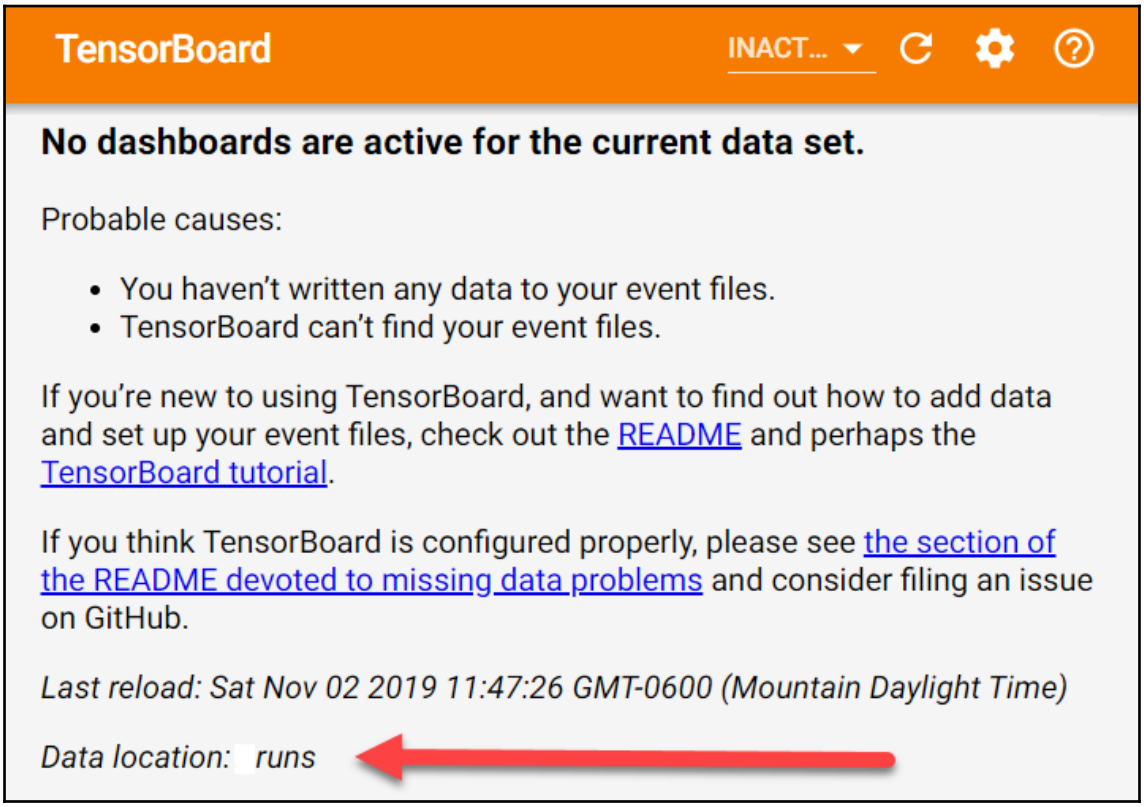

Start
Action
worker
Action
worker
Action
worker
worker
Action
End

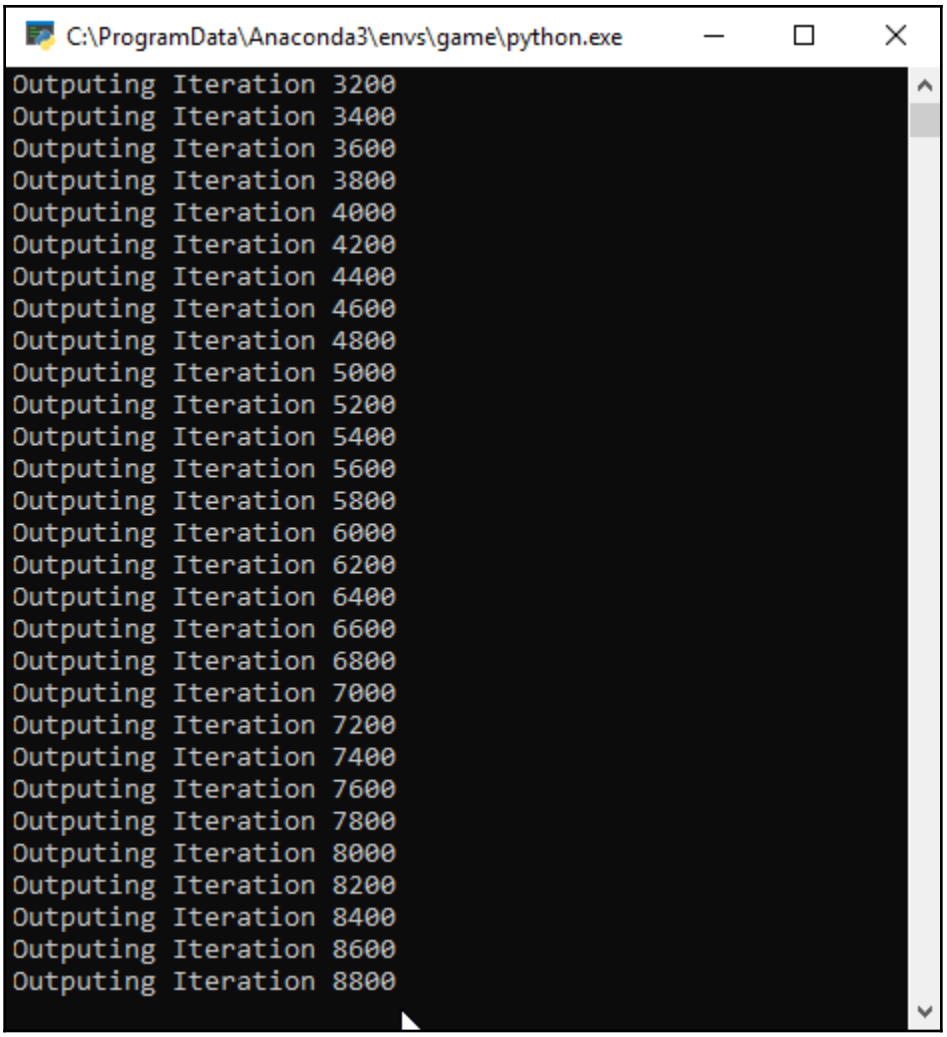

C:\ProgramData\Anaconda3\envs\game\python.exe
# of episode :20, avg score : -202.9
# of episode :40, avg score : -133.7
# of episode :60, avg score : -31.4
# of episode :80, avg score : 46.0
# of episode :100, avg score : -7.1
# of episode :120, avg score : -34.6
# of episode :140, avg score : -92.3
# of episode :160, avg score : -101.3
# of episode :180, avg score : -123.4
# of episode :200, avg score : -110.6
C:\Users\Micheal\Dropbox\Books\Hands-on Game AI with Python\Code\Chapt...

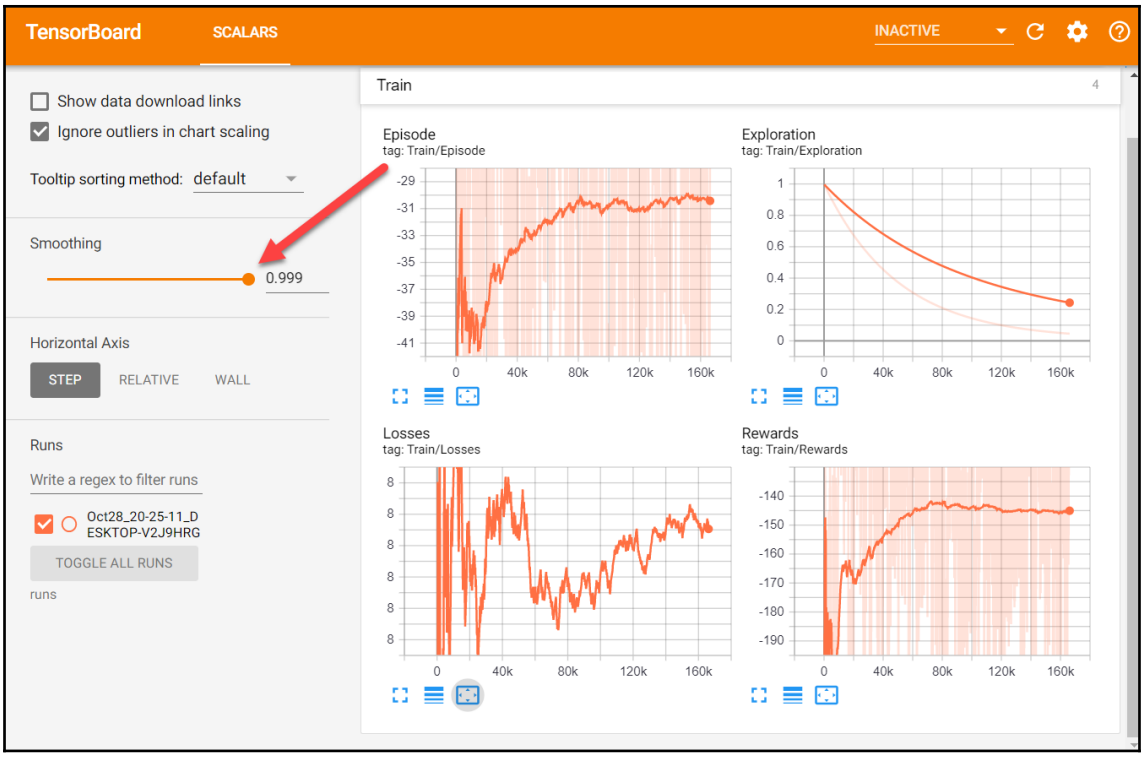

C:\ProgramData\Anaconda3\envs\game\python.exe
# of episode :80, avg score : -205.5, buffer size : 840
# of episode :100, avg score : -204.8, buffer size : 1063
# of episode :120, avg score : -214.8, buffer size : 1294
# of episode :140, avg score : -240.3, buffer size : 1532
# of episode :160, avg score : -179.0, buffer size : 1791
# of episode :180, avg score : -169.1, buffer size : 2054
# of episode :200, avg score : -124.0, buffer size : 2427
# of episode :220, avg score : -209.3, buffer size : 2806
# of episode :240, avg score : -176.4, buffer size : 3161
# of episode :260, avg score : -141.7, buffer size : 3469
# of episode :280, avg score : -123.3, buffer size : 3834
# of episode :300, avg score : -75.7, buffer size : 4509
# of episode :320, avg score : -58.7, buffer size : 5365
# of episode :340, avg score : -2.9, buffer size : 6000
# of episode :360, avg score : -32.5, buffer size : 6000
# of episode :380, avg score : 45.0, buffer size : 6000
# of episode :400, avg score : 75.0, buffer size : 6000
# of episode :420, avg score : 30.6, buffer size : 6000
# of episode :440, avg score : 87.6, buffer size : 6000
# of episode :460, avg score : 75.5, buffer size : 6000
# of episode :480, avg score : 63.8, buffer size : 6000
# of episode :500, avg score : 29.6, buffer size : 6000
# of episode :520, avg score : 68.5, buffer size : 6000
# of episode :540, avg score : 72.6, buffer size : 6000
# of episode :560, avg score : 73.6, buffer size : 6000
# of episode :580, avg score : 55.7, buffer size : 6000
# of episode :600, avg score : 54.9, buffer size : 6000
# of episode :620, avg score : 29.8, buffer size : 6000
# of episode :640, avg score : 45.1, buffer size : 6000

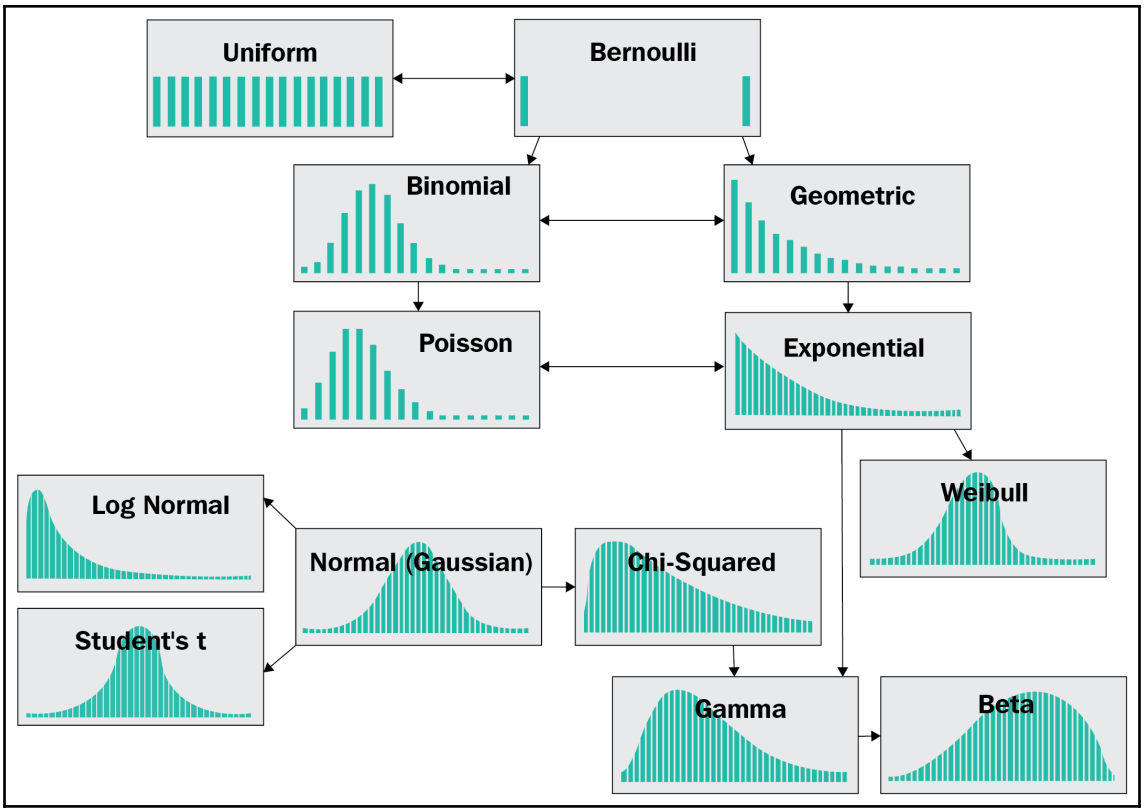
Chapter 10: All about Rainbow DQN

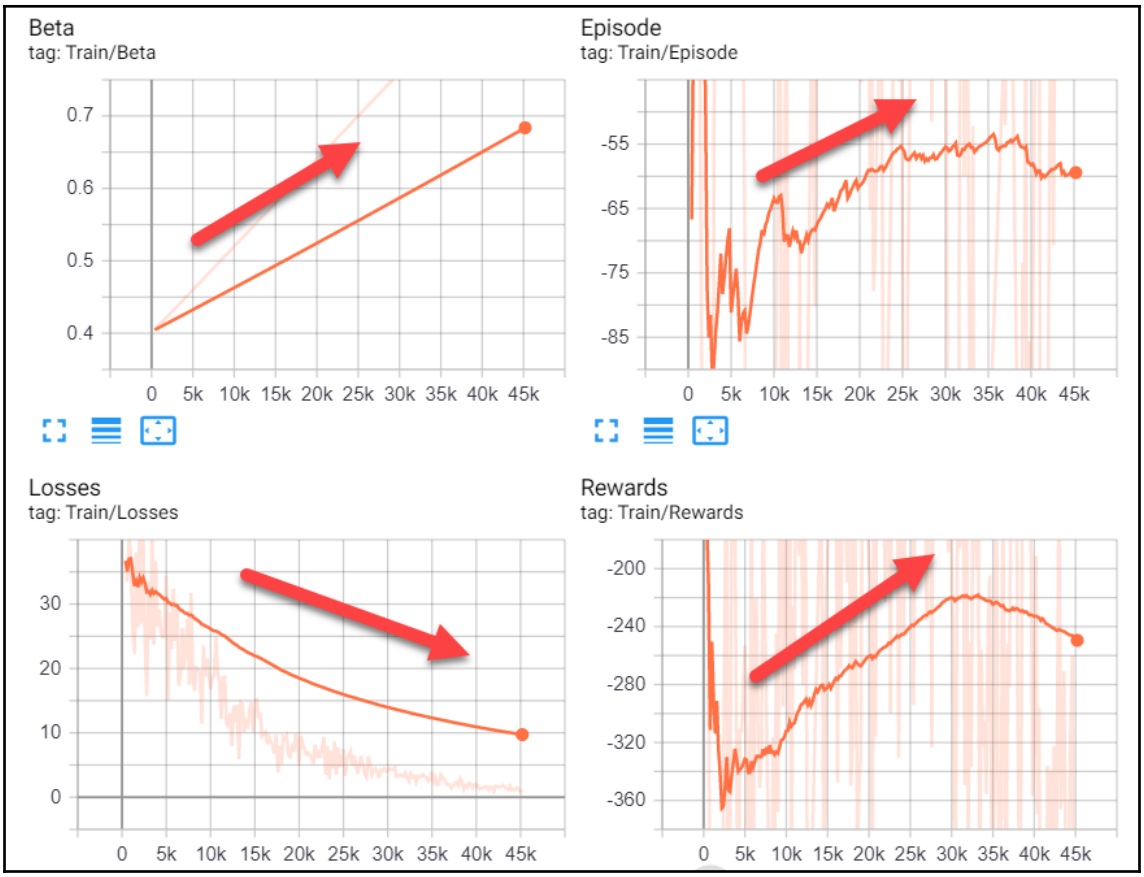


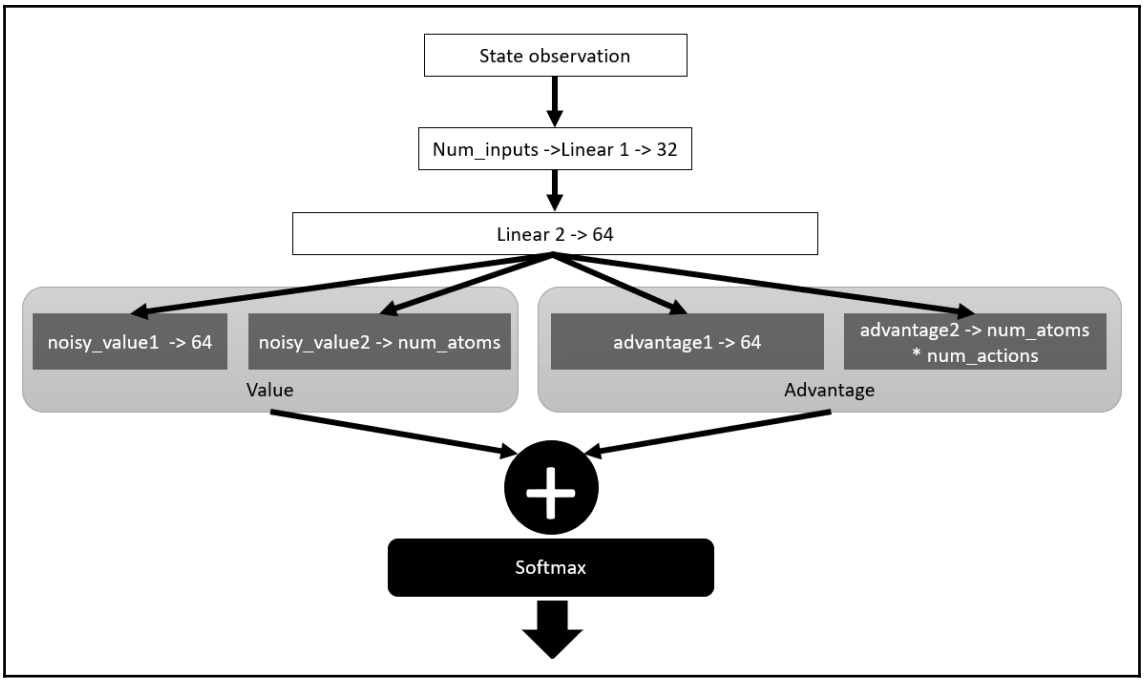


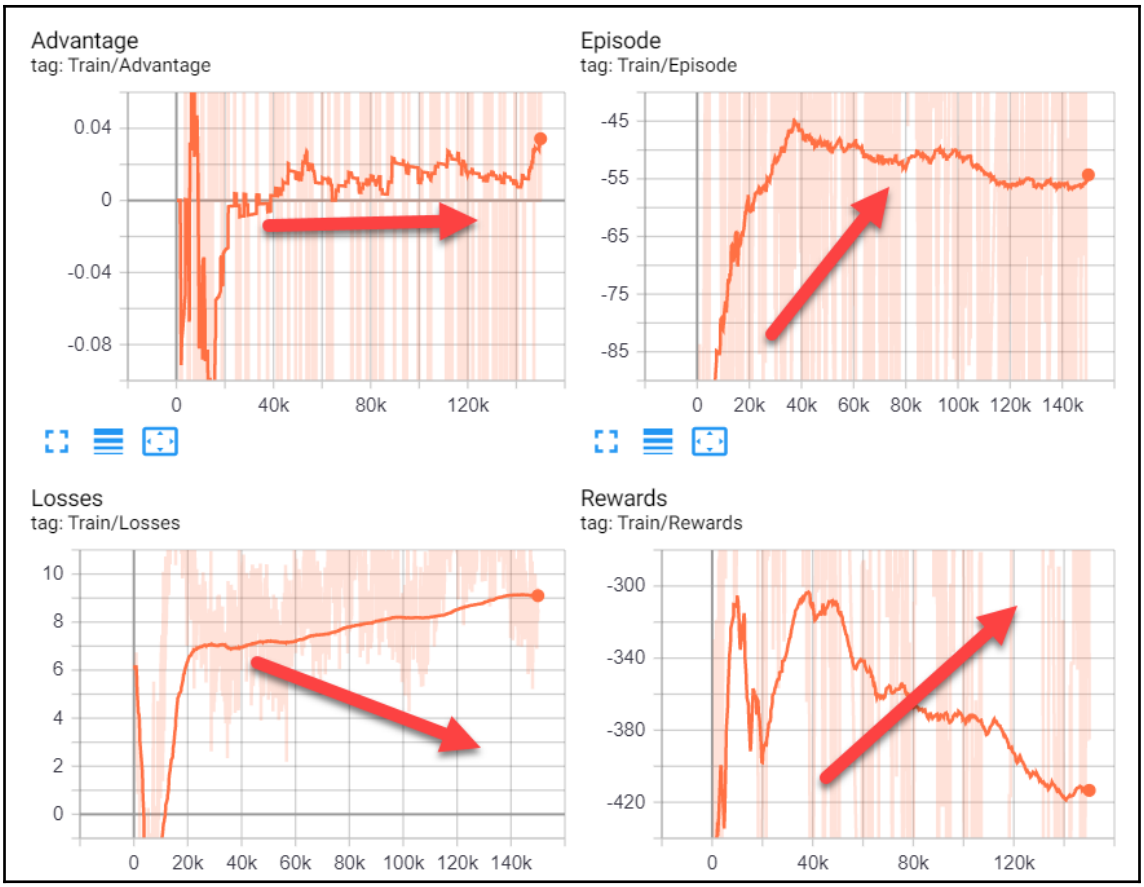

C:\ProgramData\Anaconda3\envs\game\python.exe
Outputing Iteration 3200
Outputing Iteration 3400
Outputing Iteration 3600
Outputing Iteration 3800
Outputing Iteration 4000
Outputing Iteration 4200
Outputing Iteration 4400
Outputing Iteration 4600
Outputing Iteration 4800
Outputing Iteration 5000
Outputing Iteration 5200
Outputing Iteration 5400
Outputing Iteration 5600
Outputing Iteration 5800
Outputing Iteration 6000
Outputing Iteration 6200
Outputing Iteration 6400
Outputing Iteration 6600
Outputing Iteration 6800
Outputing Iteration 7000
Outputing Iteration 7200
Outputing Iteration 7400
Outputing Iteration 7600
Outputing Iteration 7800
Outputing Iteration 8000
Outputing Iteration 8200
Outputing Iteration 8400
Outputing Iteration 8600
Outputing Iteration 8800

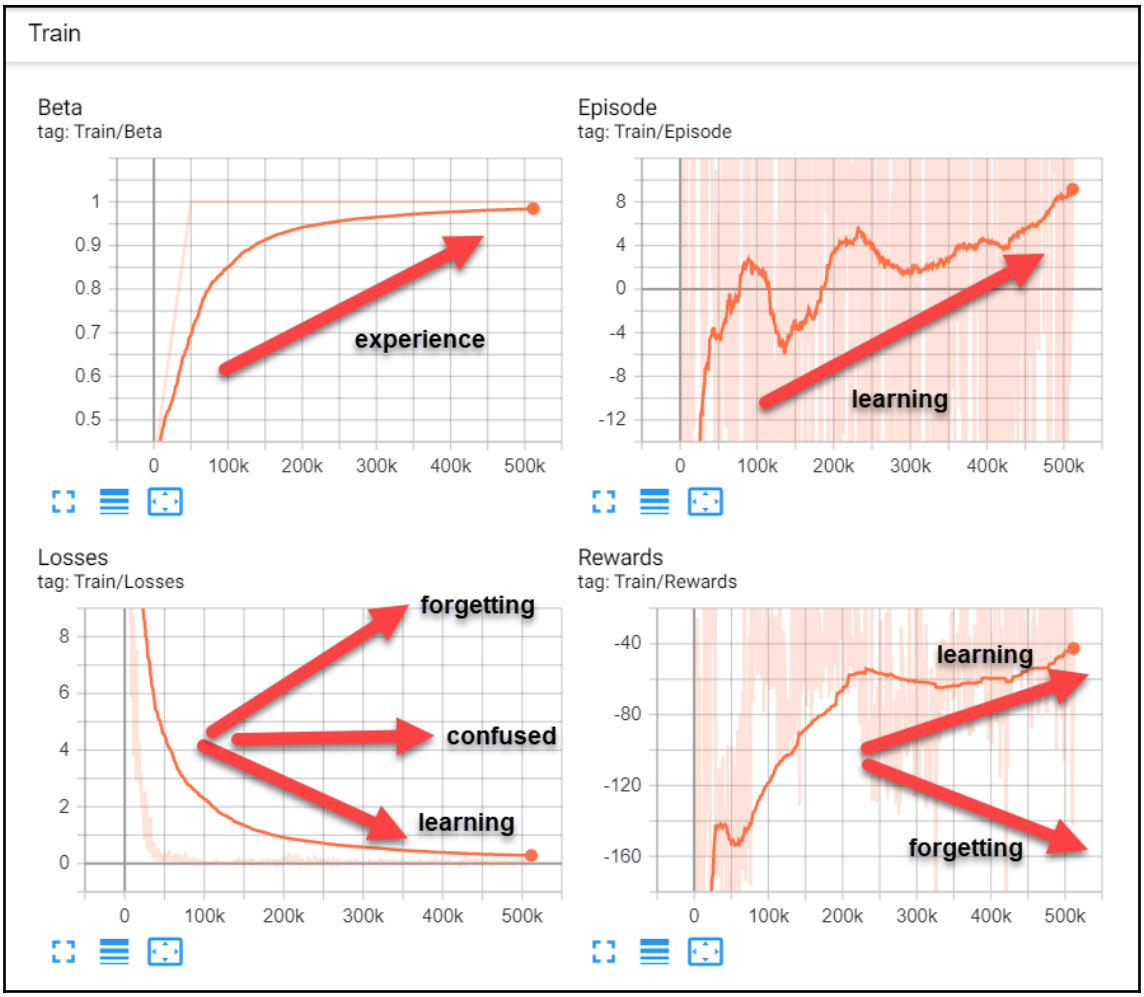

TensorBoard
SCALARS
INACTIVE
Show data download links
Ignore outliers in chart scaling
Tooltip sorting method: default
Smoothing
0.999
Horizontal Axis
STEP
RELATIVE
WALL
Runs
Write a regex to filter runs
Oct28_20-25-11_DESKTOP-V2J9HRG
TOGGLE ALL RUNS
runs
Train
4
Episode
tag: Train/Episode
-29
-31
-33
-35
-37
-39
-41
0
40k
80k
120k
160k
Exploration
tag: Train/Exploration
1
0.8
0.6
0.4
0.2
0
Losses
tag: Train/Losses
Rewards
tag: Train/Rewards
-140
-150
-160
-170
-180
-190


Uniform
Bernoulli
Binomial
Geometric
Poisson
Exponential
Weibull
Log Normal
Normal (Gaussian)
Chi-Squared
Student's t
Gamma
Beta


Beta
tag: Train/Beta
0.7
0.6
0.5
0.4
0 5k 10k 15k 20k 25k 30k 35k 40k 45k
Episode
tag: Train/Episode
-55
-65
-75
-85
0 5k 10k 15k 20k 25k 30k 35k 40k 45k
Losses
tag: Train/Losses
30
20
10
0
0 5k 10k 15k 20k 25k 30k 35k 40k 45k
Rewards
tag: Train/Rewards
-200
-240
-280
-320
-360
0 5k 10k 15k 20k 25k 30k 35k 40k 45k

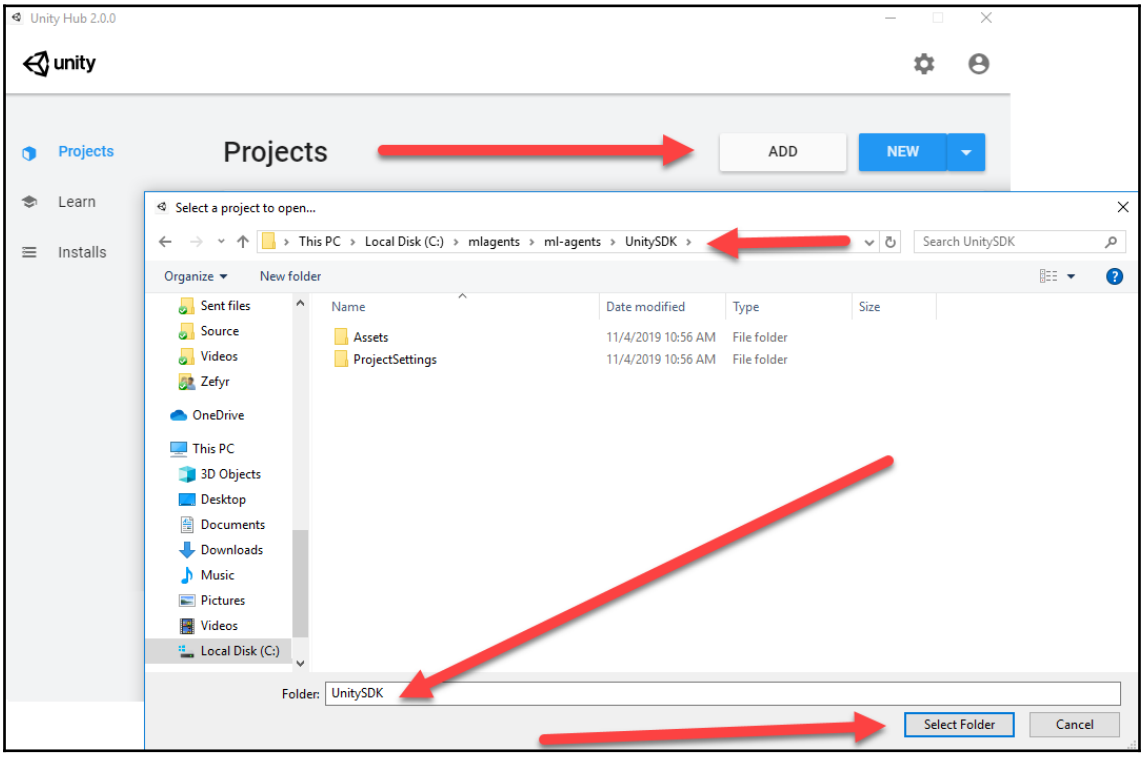

State observation
Num_inputs ->Linear 1 -> 32
Linear 2 -> 64
noisy_value1 -> 64
noisy_value2 -> num_atoms
Value
advantage1 -> 64
advantage2 -> num_atoms * num_actions
Advantage
Softmax

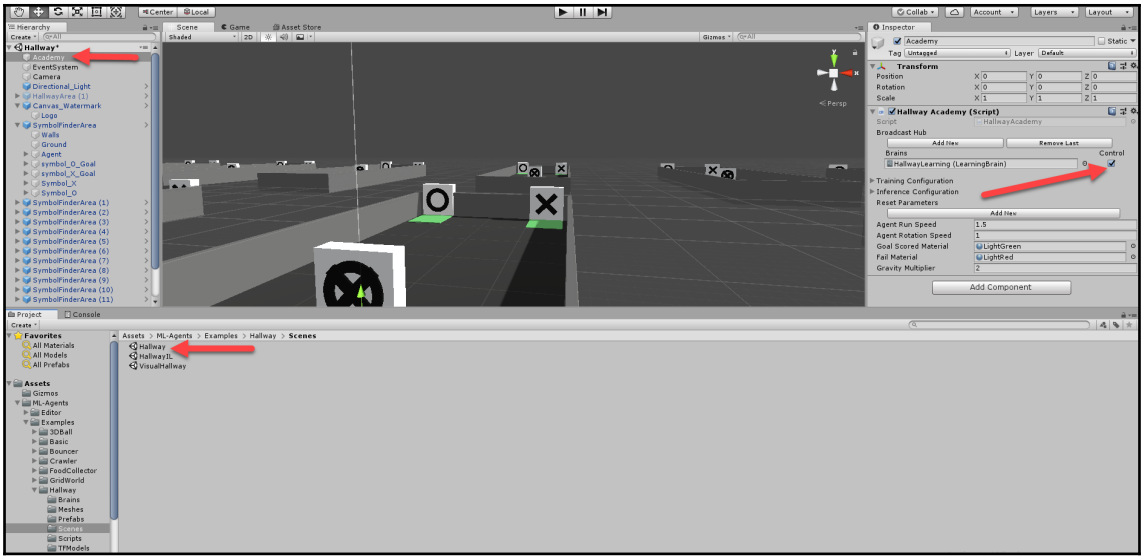

Advantage
tag: Train/Advantage
0.04
0
-0.04
-0.08
0
40k
80k
120k
Episode
tag: Train/Episode
-45
-55
-65
-75
-85
0
20k
40k
60k
80k
100k
120k
140k
Losses
tag: Train/Losses
10
8
6
4
2
0
0
20k
40k
60k
80k
100k
120k
140k
Rewards
tag: Train/Rewards
-300
-340
-380
-420
0
40k
80k
120k

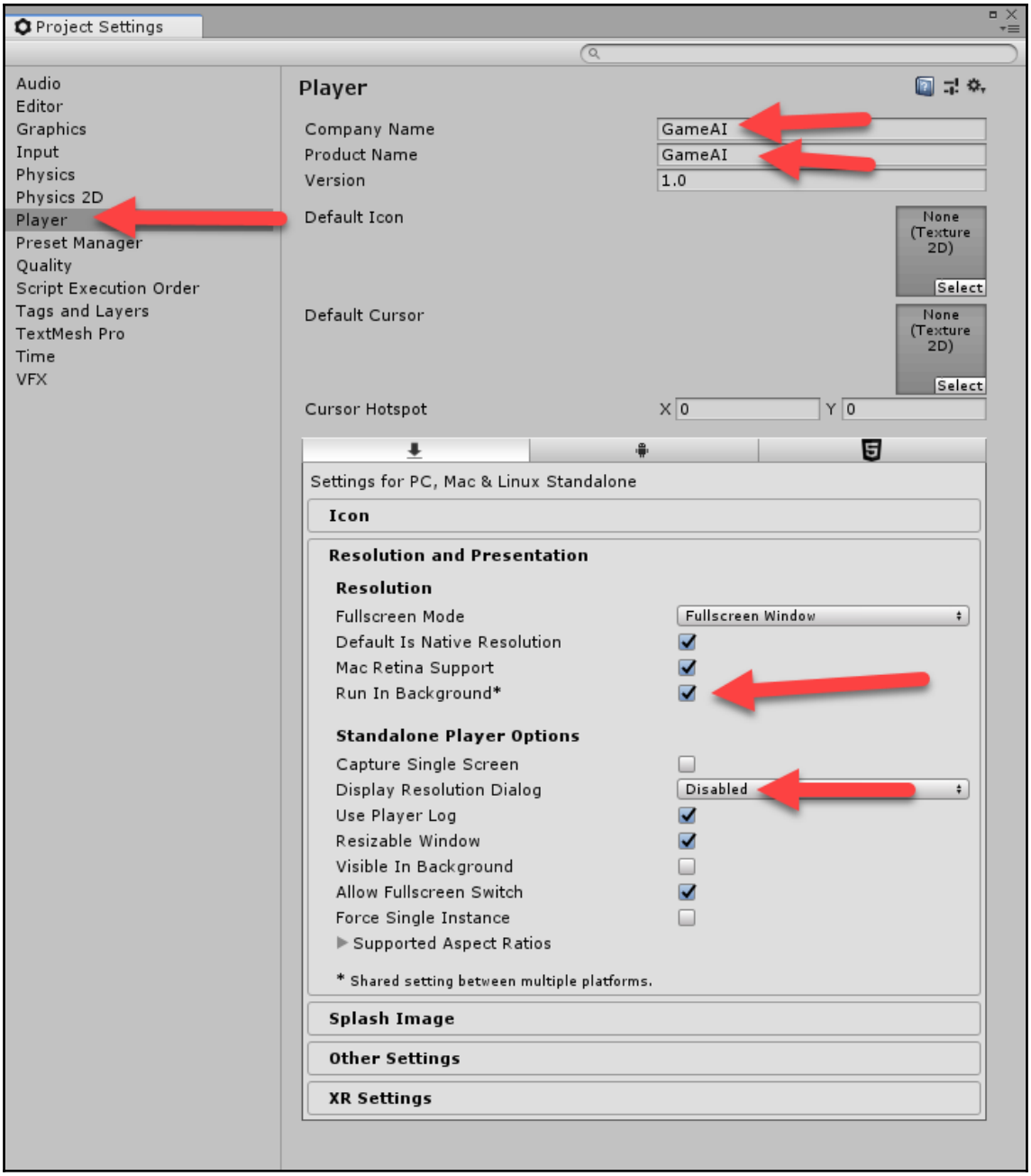

Train
Beta
tag: Train/Beta
1
0.9
0.8
0.7
0.6
0.5
0
100k
200k
300k
400k
500k
experience
Episode
tag: Train/Episode
8
4
0
-4
-8
-12
learning
Losses
tag: Train/Losses
8
6
4
2
0
forgetting
confused
learning
Rewards
tag: Train/Rewards
-40
-80
-120
-160
learning
forgetting

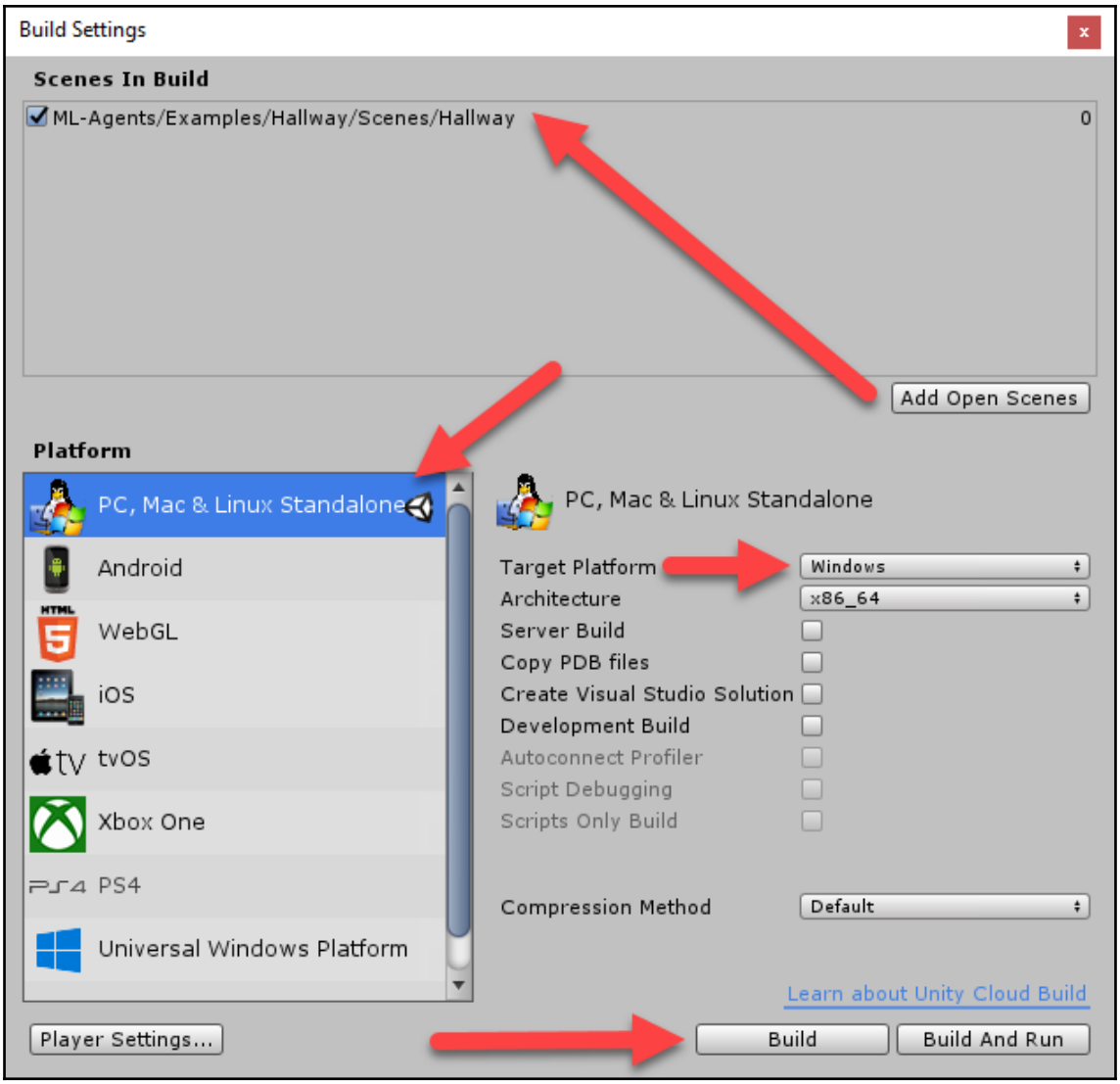

Advantage
tag: Train/Advantage
0.25
0.15
0.05
-0.05
0
40k
80k
120k
Episode
tag: Train/Episode
-40
-50
-60
-70
0
20k
40k
60k
80k
100k
120k
140k
Losses
tag: Train/Losses
12
8
4
0
0
20k
40k
60k
80k
100k
120k
140k
Rewards
tag: Train/Rewards
-200
-240
-280
-320
-360
-400
-440
0
40k
80k
120k

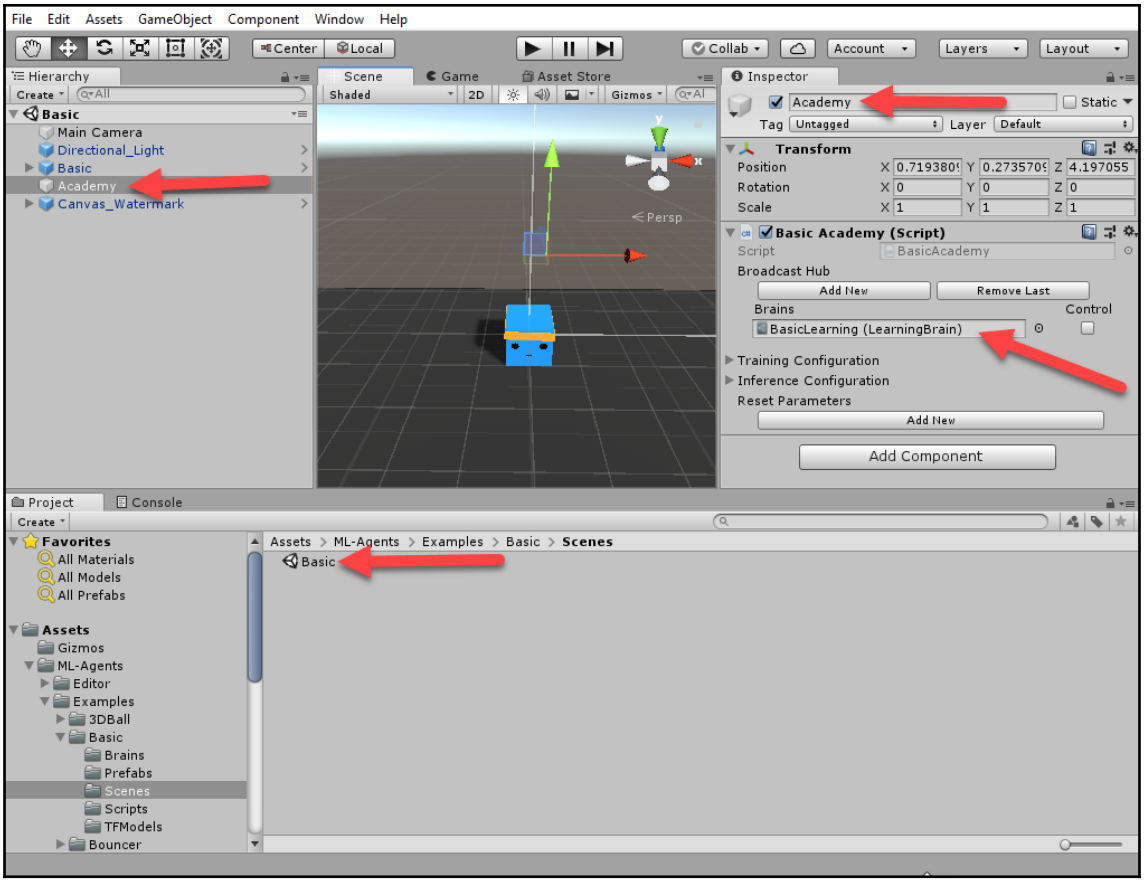
Chapter 11: Exploiting ML-Agents

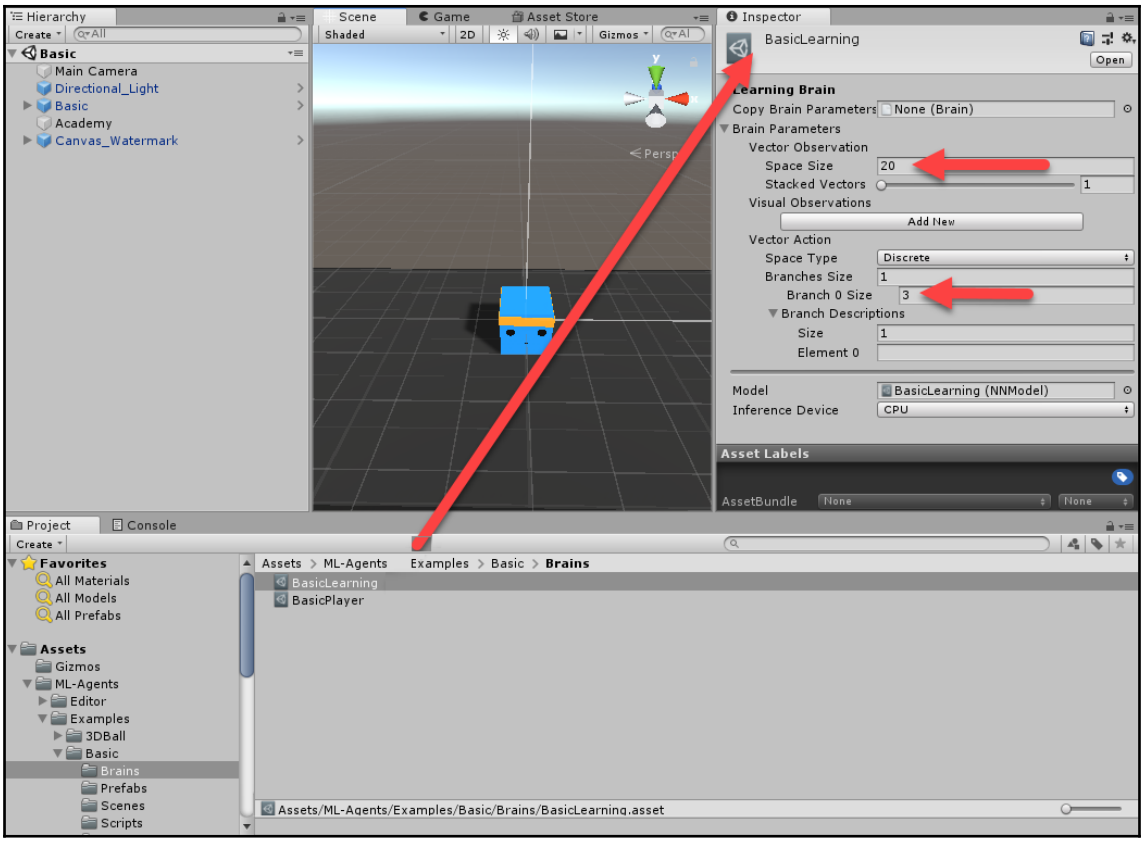


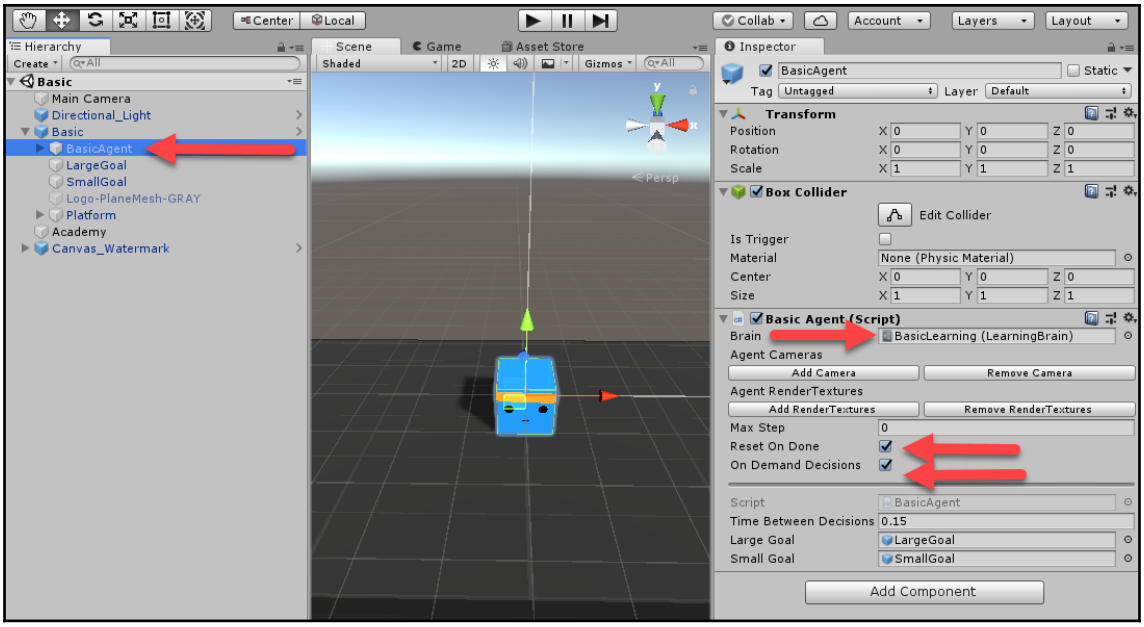

Hierarchy
Scene
Game
Asset Store
Inspector
Hallway*
Academy
EventSystem
Camera
Directional_Light
HallwayArea (1)
Canvas_Watermark
Logo
SymbolFinderArea
Walls
Ground
Agent
symbol_O_Goal
symbol_X_Goal
Symbol_X
Symbol_O
Transform
Position
Rotation
Scale
Hallway Academy (Script)
Broadcast Hub
Add New
Remove Last
Brains
Control
HallwayLearning (LearningBrain)
Training Configuration
Inference Configuration
Reset Parameters
Agent Run Speed
Agent Rotation Speed
Goal Scored Material
Fail Material
Gravity Multiplier
Add Component
Project
Console
Favorites
All Materials
All Models
All Prefabs
Assets
Gizmos
ML-Agents
Editor
Examples
3DBall
Basic
Bouncer
Crawler
FoodCollector
GridWorld
Hallway
Brains
Meshes
Prefabs
Scenes
Scripts
TFModels
Assets > ML-Agents > Examples > Hallway > Scenes
Hallway
HallwayIL
VisualHallway

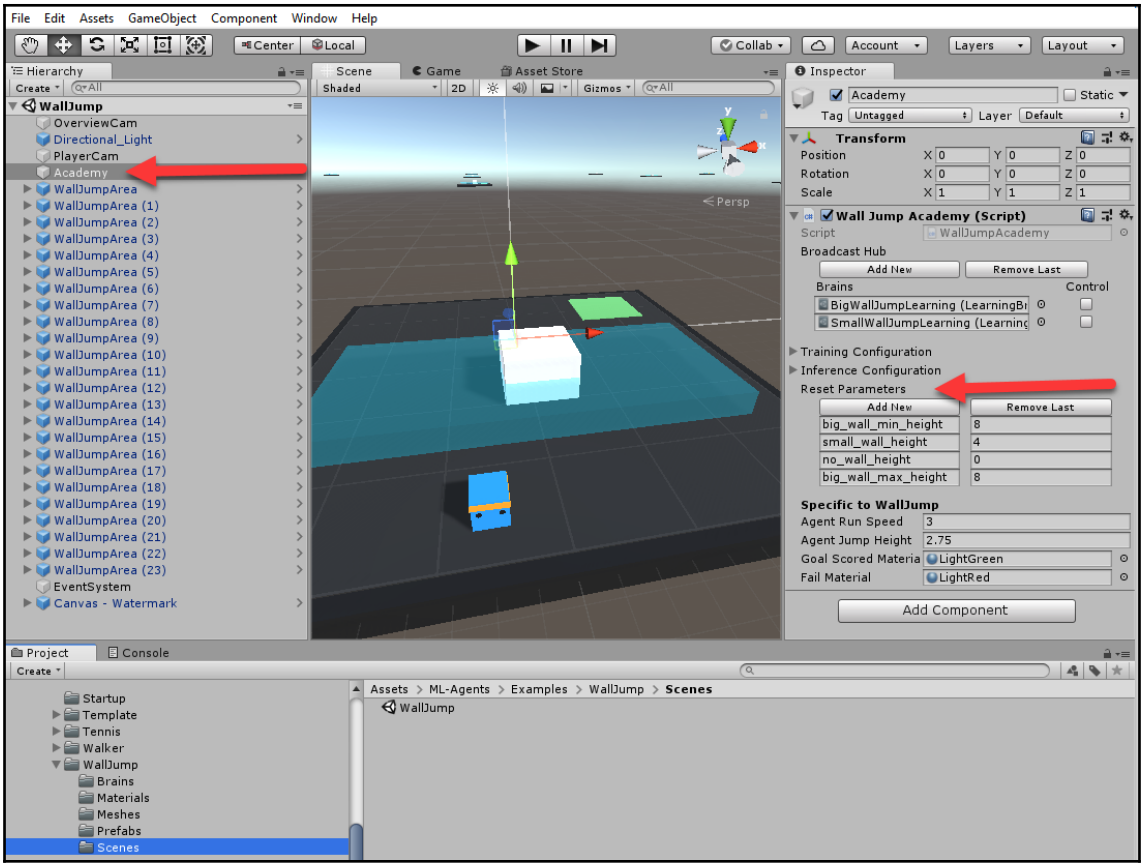

Project Settings
Audio
Editor
Graphics
Input
Physics
Physics 2D
Player
Preset Manager
Quality
Script Execution Order
Tags and Layers
TextMesh Pro
Time
VFX
Player
Company Name
GameAI
Product Name
GameAI
Version
1.0
Default Icon
None (Texture 2D)
Select
Default Cursor
None (Texture 2D)
Select
Cursor Hotspot
X 0
Y 0
Settings for PC, Mac & Linux Standalone
Icon
Resolution and Presentation
Resolution
Fullscreen Mode
Fullscreen Window
Default Is Native Resolution
Mac Retina Support
Run In Background*
Standalone Player Options
Capture Single Screen
Display Resolution Dialog
Disabled
Use Player Log
Resizable Window
Visible In Background
Allow Fullscreen Switch
Force Single Instance
Supported Aspect Ratios
* Shared setting between multiple platforms.
Splash Image
Other Settings
XR Settings

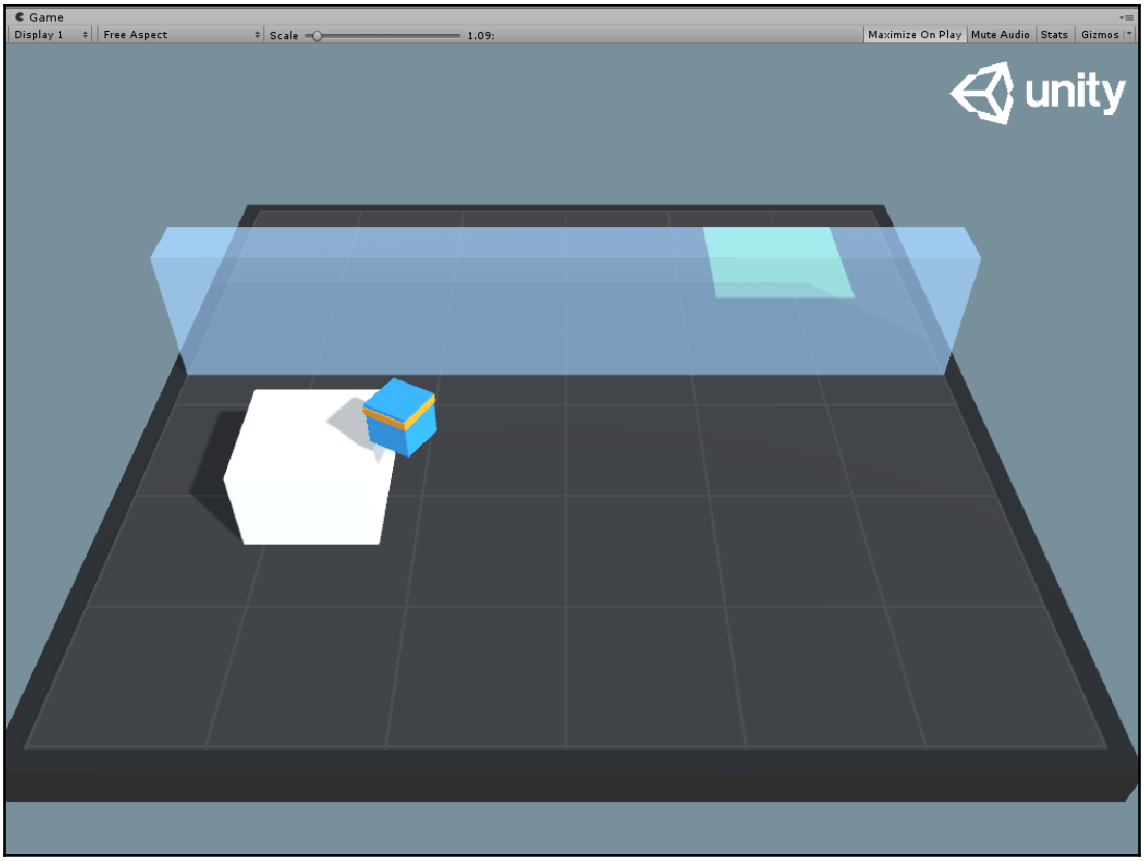

Build Settings
Scenes In Build
ML-Agents/Examples/Hallway/Scenes/Hallway
0
Add Open Scenes
Platform
PC, Mac & Linux Standalone
Android
WebGL
iOS
tvOS
Xbox One
PS4
Universal Windows Platform
PC, Mac & Linux Standalone
Target Platform
Windows
Architecture
x86_64
Server Build
Copy PDB files
Create Visual Studio Solution
Development Build
Autoconnect Profiler
Script Debugging
Scripts Only Build
Compression Method
Default
Learn about Unity Cloud Build
Player Settings...
Build
Build And Run

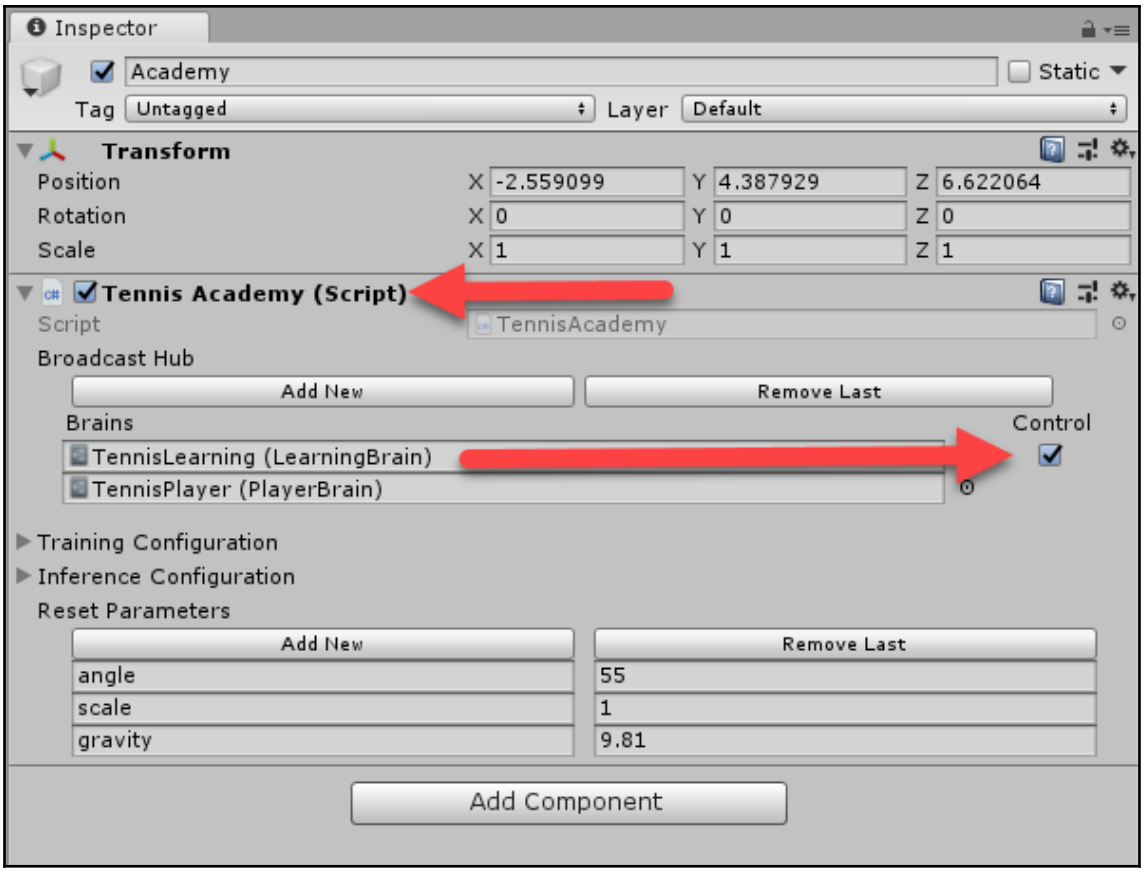

File Edit Assets GameObject Component Window Help
Center Local
Collab Account Layers Layout
Hierarchy
Create All
Basic
Main Camera
Directional_Light
Basic
Academy
Canvas_Watermark
Scene Game Asset Store
Shaded 2D Gizmos
Persp
Inspector
Academy
Static
Tag Untagged Layer Default
Transform
Position
Rotation
Scale
Basic Academy (Script)
Script BasicAcademy
Broadcast Hub
Add New Remove Last
Brains Control
BasicLearning (LearningBrain)
Training Configuration
Inference Configuration
Reset Parameters
Add New
Add Component
Project Console
Create
Favorites
All Materials
All Models
All Prefabs
Assets
Gizmos
ML-Agents
Editor
Examples
3DBall
Basic
Brains
Prefabs
Scenes
Scripts
TFModels
Bouncer
Assets > ML-Agents > Examples > Basic > Scenes
Basic

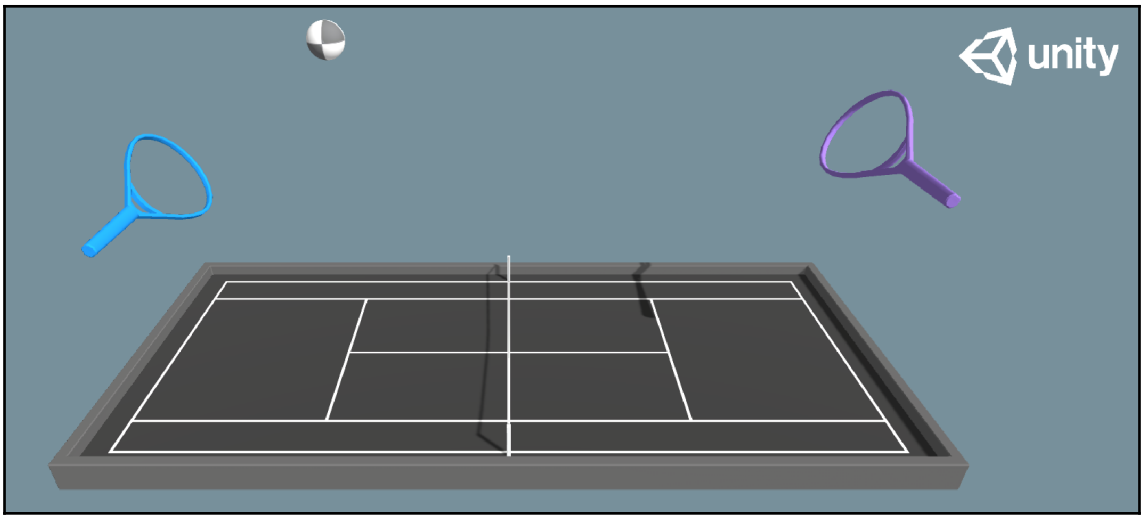

Hierarchy
Create
Basic
Main Camera
Directional_Light
Basic
Academy
Canvas_Watermark
Scene
Game
Asset Store
Shaded
2D
Gizmos
Persp
Inspector
BasicLearning
Open
Learning Brain
Copy Brain Parameters
None (Brain)
Brain Parameters
Vector Observation
Space Size
20
Stacked Vectors
1
Visual Observations
Add New
Vector Action
Space Type
Discrete
Branches Size
1
Branch 0 Size
3
Branch Descriptions
Size
1
Element 0
Model
BasicLearning (NNModel)
Inference Device
CPU
Asset Labels
AssetBundle
None
None
Project
Console
Create
Favorites
All Materials
All Models
All Prefabs
Assets
Gizmos
ML-Agents
Editor
Examples
3DBall
Basic
Brains
Prefabs
Scenes
Scripts
Assets > ML-Agents > Examples > Basic > Brains
BasicLearning
BasicPlayer
Assets/ML-Agents/Examples/Basic/Brains/BasicLearning.asset

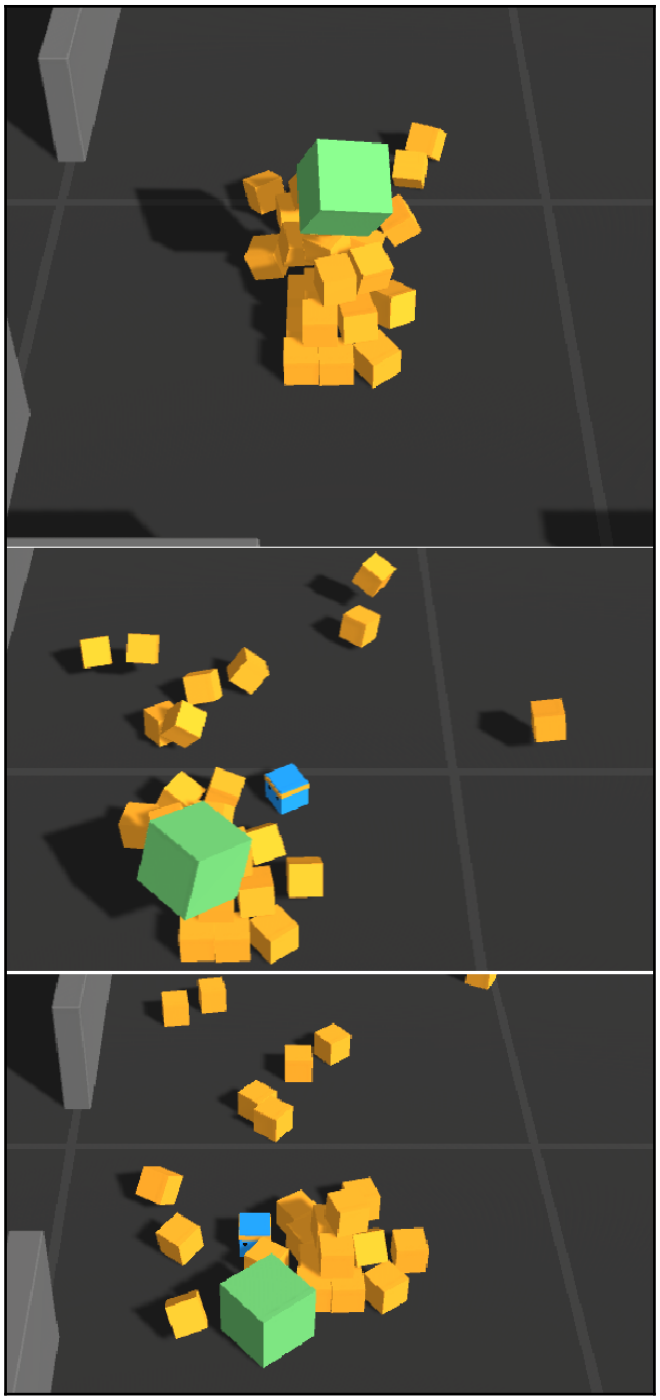

Center
Local
Collab
Account
Layers
Layout
Hierarchy
Create
Basic
Main Camera
Directional_Light
Basic
BasicAgent
LargeGoal
SmallGoal
Logo-PlaneMesh-GRAY
Platform
Academy
Canvas_Watermark
Scene
Game
Asset Store
Shaded
2D
Gizmos
Persp
Inspector
BasicAgent
Static
Tag
Untagged
Layer
Default
Transform
Position
Rotation
Scale
Box Collider
Edit Collider
Is Trigger
Material
None (Physic Material)
Center
Size
Basic Agent (Script)
Brain
BasicLearning (LearningBrain)
Agent Cameras
Add Camera
Remove Camera
Agent RenderTextures
Add RenderTextures
Remove RenderTextures
Max Step
Reset On Done
On Demand Decisions
Script
BasicAgent
Time Between Decisions
0.15
Large Goal
LargeGoal
Small Goal
SmallGoal
Add Component


File Edit Assets GameObject Component Window Help
Hierarchy
WallJump
OverviewCam
Directional_Light
PlayerCam
Academy
WallJumpArea
WallJumpArea (1)
WallJumpArea (2)
WallJumpArea (3)
WallJumpArea (4)
WallJumpArea (5)
WallJumpArea (6)
WallJumpArea (7)
WallJumpArea (8)
WallJumpArea (9)
WallJumpArea (10)
WallJumpArea (11)
WallJumpArea (12)
WallJumpArea (13)
WallJumpArea (14)
WallJumpArea (15)
WallJumpArea (16)
WallJumpArea (17)
WallJumpArea (18)
WallJumpArea (19)
WallJumpArea (20)
WallJumpArea (21)
WallJumpArea (22)
WallJumpArea (23)
EventSystem
Canvas - Watermark
Scene
Game
Asset Store
Persp
Inspector
Academy
Static
Tag Untagged
Layer Default
Transform
Position
Rotation
Scale
Wall Jump Academy (Script)
Script WallJumpAcademy
Broadcast Hub
Add New
Remove Last
Brains
Control
BigWallJumpLearning (LearningBr
SmallWallJumpLearning (Learning
Training Configuration
Inference Configuration
Reset Parameters
big_wall_min_height 8
small_wall_height 4
no_wall_height 0
big_wall_max_height 8
Specific to WallJump
Agent Run Speed 3
Agent Jump Height 2.75
Goal Scored Materia LightGreen
Fail Material LightRed
Add Component
Project
Console
Assets > ML-Agents > Examples > WallJump > Scenes
WallJump
Startup
Template
Tennis
Walker
WallJump
Brains
Materials
Meshes
Prefabs
Scenes

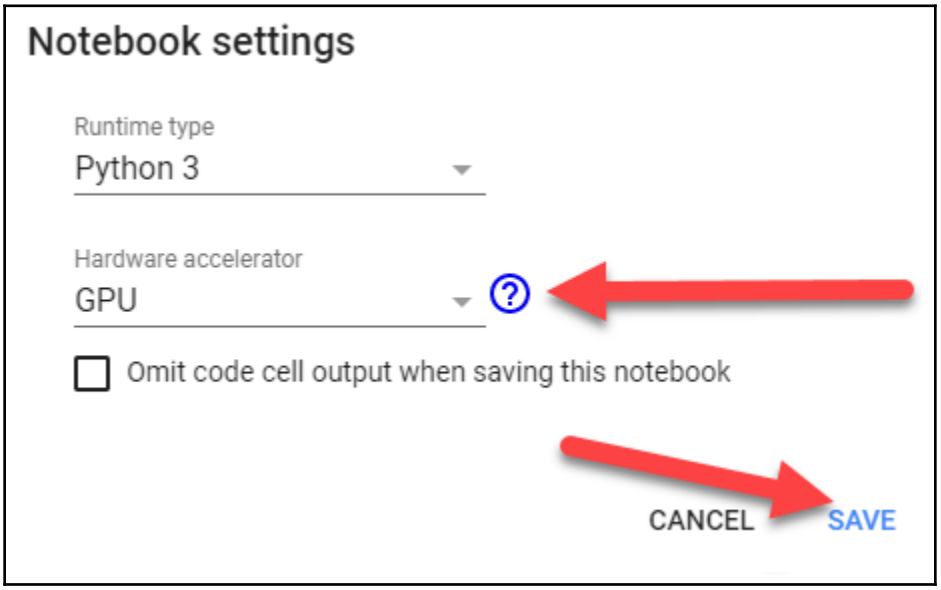

Game
Display 1
Free Aspect
Scale
1.09:
Maximize On Play
Mute Audio
Stats
Gizmos
unity

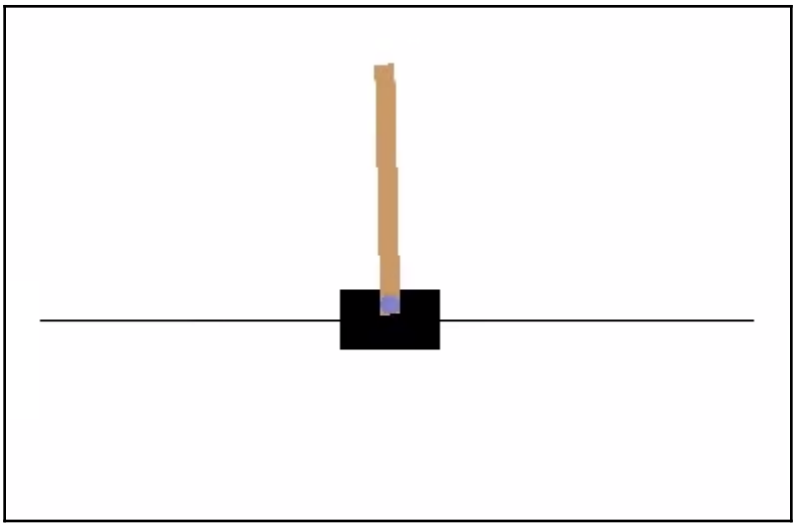

Inspector
Academy
Static
Tag
Untagged
Layer
Default
Transform
Position
X -2.559099
Y 4.387929
Z 6.622064
Rotation
X 0
Y 0
Z 0
Scale
X 1
Y 1
Z 1
Tennis Academy (Script)
Script
TennisAcademy
Broadcast Hub
Add New
Remove Last
Brains
Control
TennisLearning (LearningBrain)
TennisPlayer (PlayerBrain)
Training Configuration
Inference Configuration
Reset Parameters
Add New
Remove Last
angle
55
scale
1
gravity
9.81
Add Component


unity



Chapter 12: DRL Frameworks





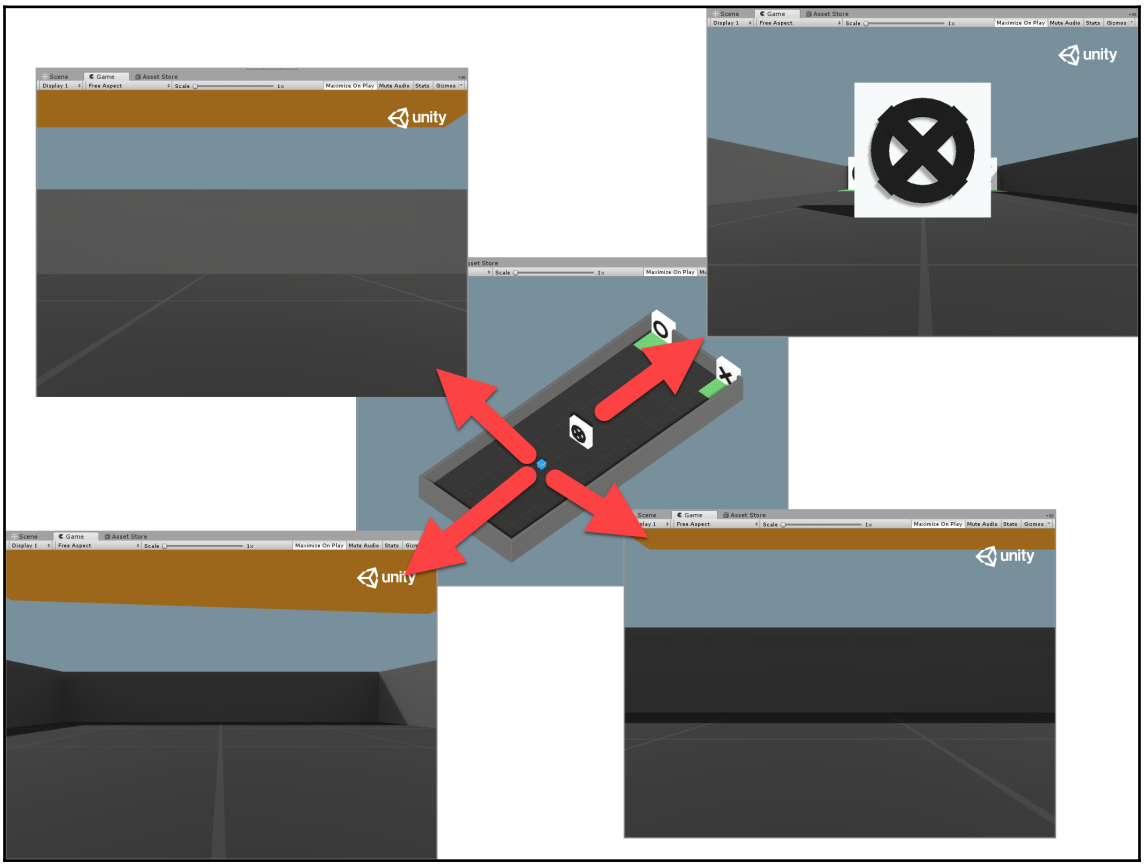
RLLib.ipynb ☆

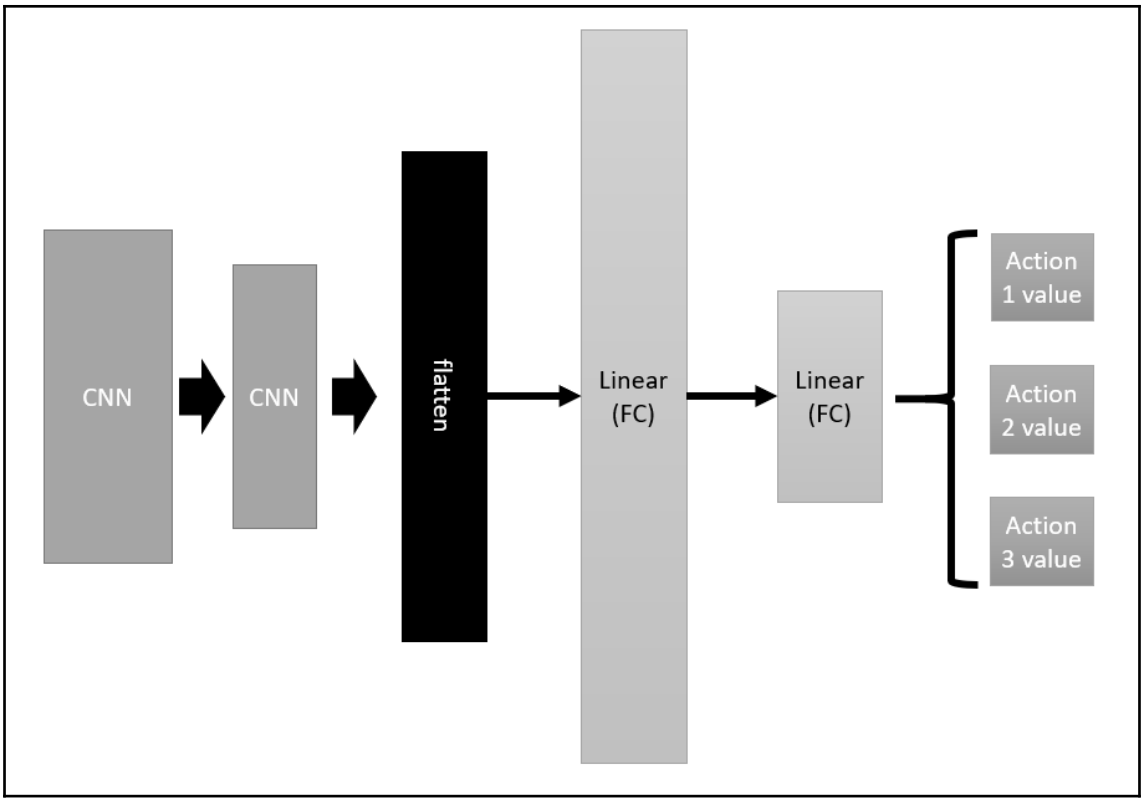
File Edit View Insert Runtime Tools Help All changes saved

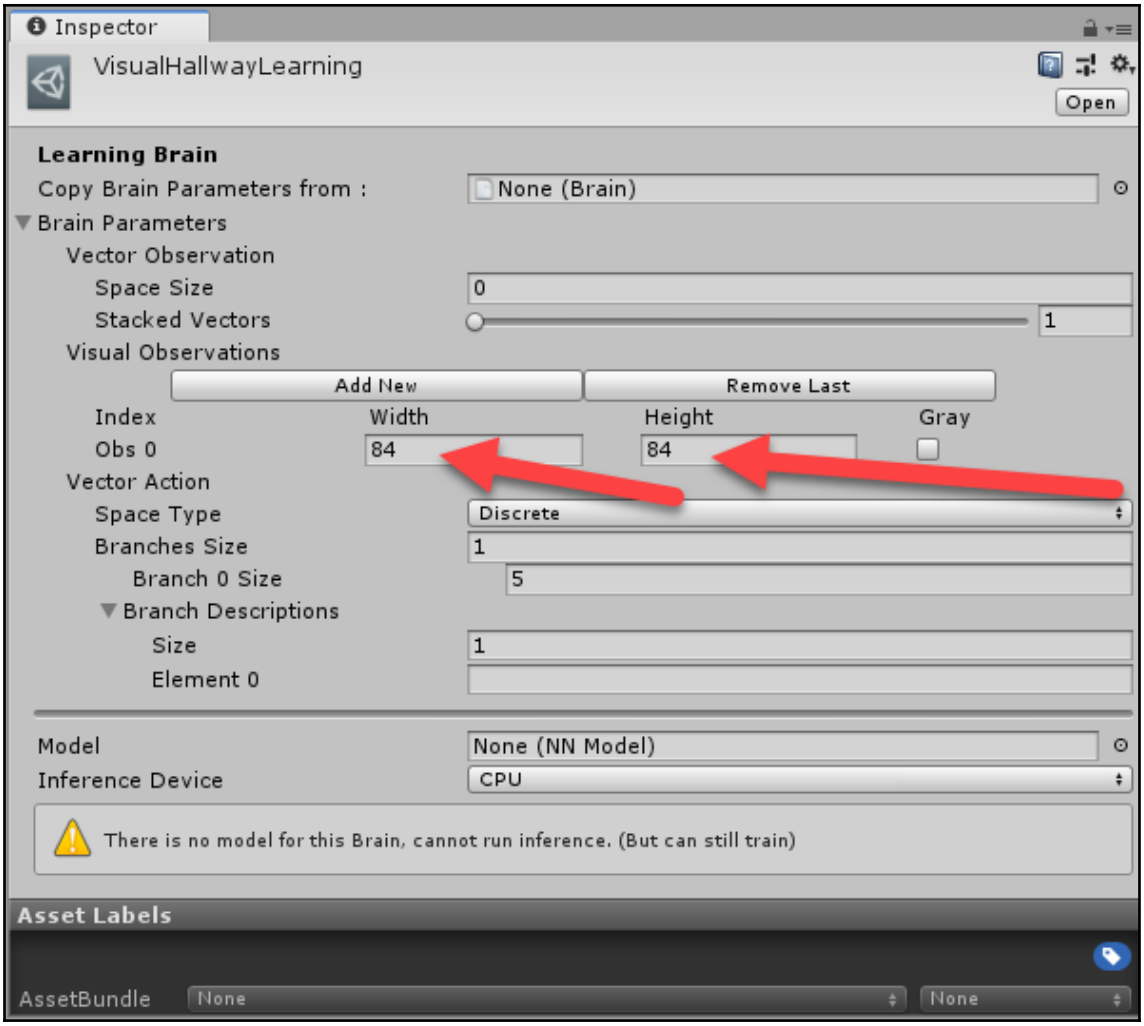
+ Code + Text

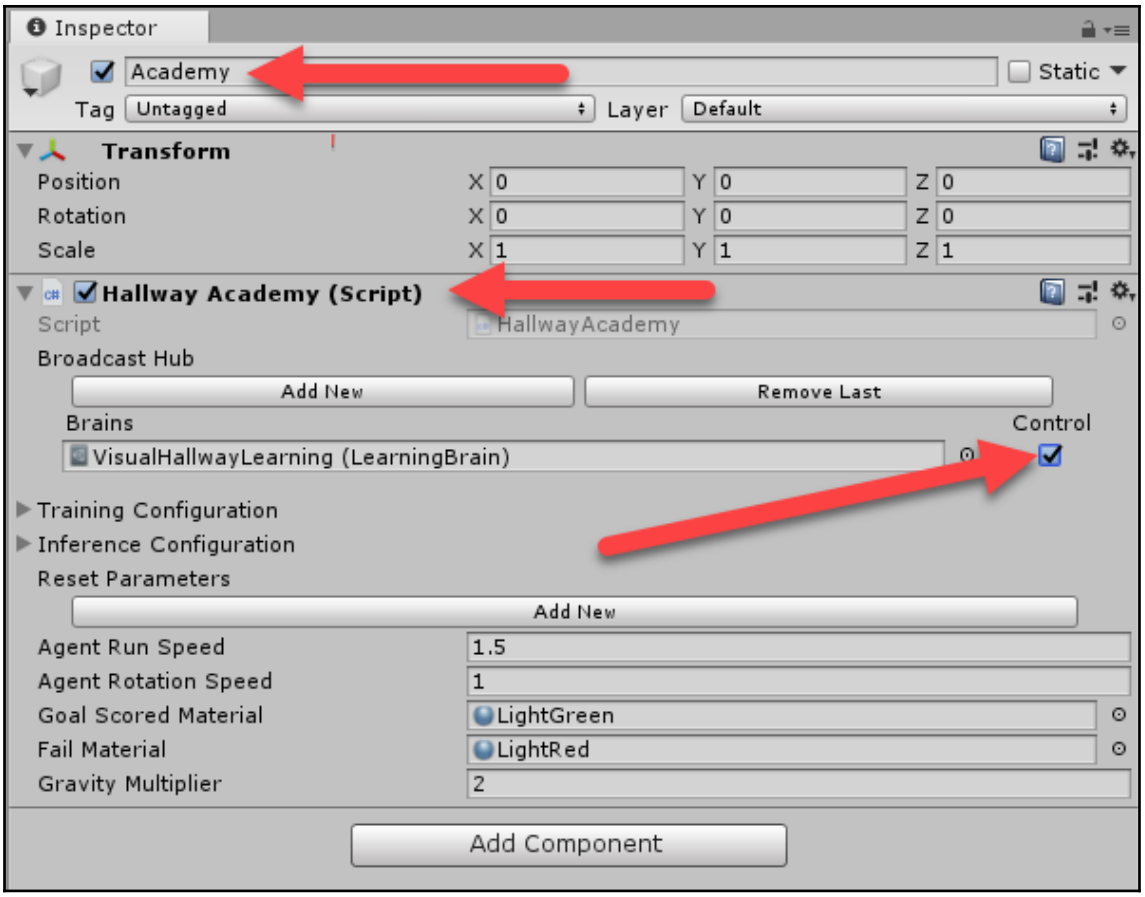
```
(pid=491)   'type': 'MultiAgentBatch'}
(pid=491)
(pid=491) 2019-11-11 00:36:55,819       INFO tf_policy.py:358 -- Optimizing variable <tf.Variable 'default_policy/fc_1/kernel:0' shape=(4, 256) dtype=float32>
(pid=491) 2019-11-11 00:36:55,819       INFO tf_policy.py:358 -- Optimizing variable <tf.Variable 'default_policy/fc_1/bias:0' shape=(256,) dtype=float32>
(pid=491) 2019-11-11 00:36:55,819       INFO tf_policy.py:358 -- Optimizing variable <tf.Variable 'default_policy/fc_out/kernel:0' shape=(256, 256) dtype=float32>
(pid=491) 2019-11-11 00:36:55,819       INFO tf_policy.py:358 -- Optimizing variable <tf.Variable 'default_policy/fc_out/bias:0' shape=(256,) dtype=float32>
(pid=491) 2019-11-11 00:36:55,819       INFO tf_policy.py:358 -- Optimizing variable <tf.Variable 'default_policy/q_func/action_value/hidden_0/kernel:0' shape=(256, 256) dtype=float32_ref>
(pid=491) 2019-11-11 00:36:55,819       INFO tf_policy.py:358 -- Optimizing variable <tf.Variable 'default_policy/q_func/action_value/hidden_0/bias:0' shape=(256,) dtype=float32_ref>
(pid=491) 2019-11-11 00:36:55,819       INFO tf_policy.py:358 -- Optimizing variable <tf.Variable 'default_policy/q_func/action_value/dense/kernel:0' shape=(256, 2) dtype=float32_ref>
(pid=491) 2019-11-11 00:36:55,819       INFO tf_policy.py:358 -- Optimizing variable <tf.Variable 'default_policy/q_func/action_value/dense/bias:0' shape=(2,) dtype=float32_ref>
(pid=491) 2019-11-11 00:36:55,819       INFO tf_policy.py:358 -- Optimizing variable <tf.Variable 'default_policy/q_func/state_value/dense/kernel:0' shape=(256, 256) dtype=float32_ref>
(pid=491) 2019-11-11 00:36:55,819       INFO tf_policy.py:358 -- Optimizing variable <tf.Variable 'default_policy/q_func/state_value/dense/bias:0' shape=(256,) dtype=float32_ref>
(pid=491) 2019-11-11 00:36:55,819       INFO tf_policy.py:358 -- Optimizing variable <tf.Variable 'default_policy/q_func/state_value/dense_1/kernel:0' shape=(256, 1) dtype=float32_ref>
(pid=491) 2019-11-11 00:36:55,819       INFO tf_policy.py:358 -- Optimizing variable <tf.Variable 'default_policy/q_func/state_value/dense_1/bias:0' shape=(1,) dtype=float32_ref>
(pid=491) 2019-11-11 00:36:55,819       INFO tf_run_builder.py:92 -- Executing TF run without tracing. To dump TF timeline traces to disk, set the TF_TIMELINE_DIR environment variable.
Result for DQN_CartPole-v0_1_lr=0.001:
  custom_metrics: {}
  date: 2019-11-11_00-37-05
  done: false
  episode_len_mean: 18.3
  episode_reward_max: 40.0
  episode_reward_mean: 18.3
  episode_reward_min: 8.0
  episodes_this_iter: 59
  episodes_total: 107
  experiment_id: c939f9681a9c4a46a46f4dd24c6b7b40
  hostname: 74d868c92a77
  info:
    grad_time_ms: 9.305
    learner:
      default_policy:
        cur_lr: 0.0010000000474974513
        max_q: 2.99139404296875
        mean_q: 2.078482151031494
        mean_td_error: 0.18657320737838745
        min_q: 0.9282804131507874
        model: {}
    max_exploration: 0.902
    min_exploration: 0.902
    num_steps_sampled: 2000
    num_steps_trained: 8000
    num_target_updates: 3
    opt_peak_throughput: 3438.853
    opt_samples: 32.0
    replay_time_ms: 2.895
    sample_time_ms: 14.544
    update_time_ms: 11.078
  iterations_since_restore: 2
```

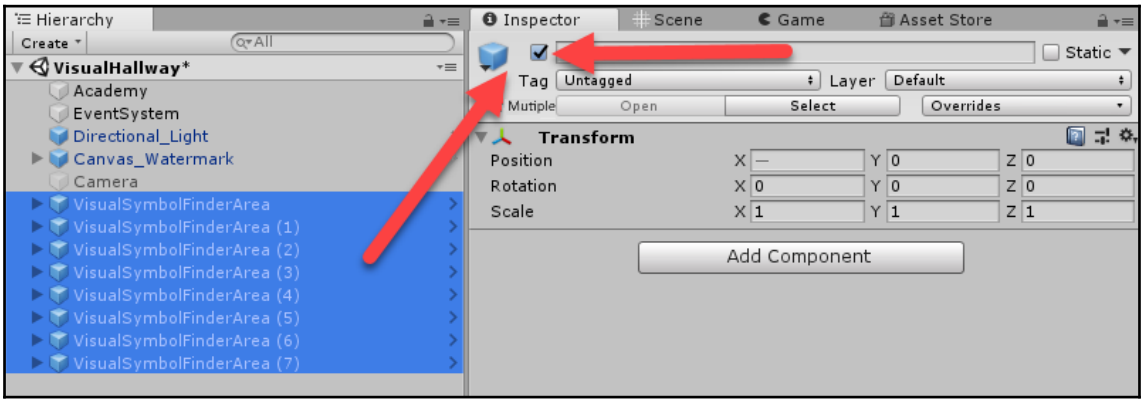
Chapter 13: 3D Worlds




CNN
CNN
flatten
Linear
(FC)
Linear
(FC)
Action
1 value
Action
2 value
Action
3 value

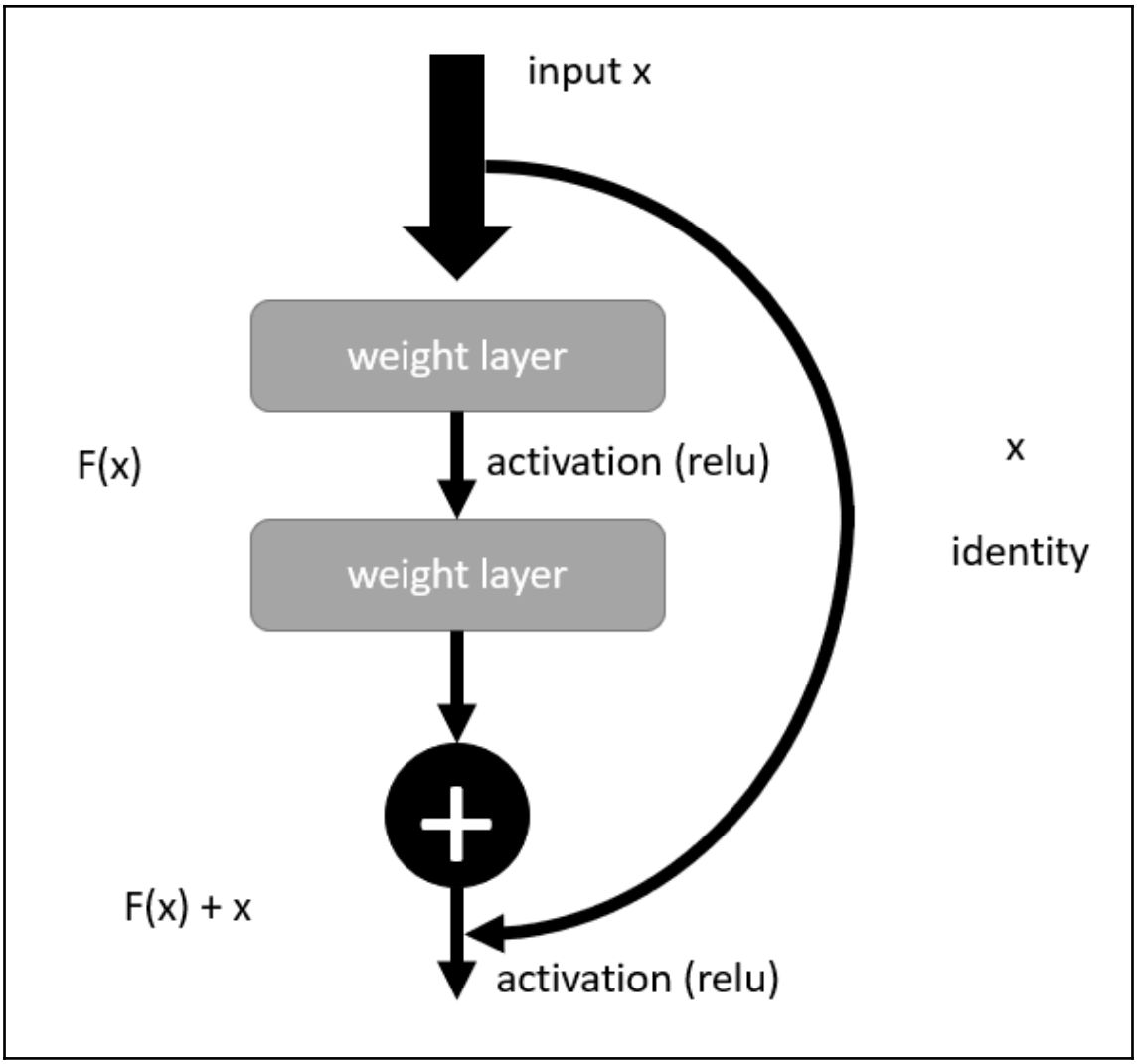

Inspector
VisualHallwayLearning
Open
Learning Brain
Copy Brain Parameters from :
None (Brain)
Brain Parameters
Vector Observation
Space Size
0
Stacked Vectors
1
Visual Observations
Add New
Remove Last
Index
Width
Height
Gray
Obs 0
84
84
Vector Action
Space Type
Discrete
Branches Size
1
Branch 0 Size
5
Branch Descriptions
Size
1
Element 0
Model
None (NN Model)
Inference Device
CPU
There is no model for this Brain, cannot run inference. (But can still train)
Asset Labels
AssetBundle
None
None

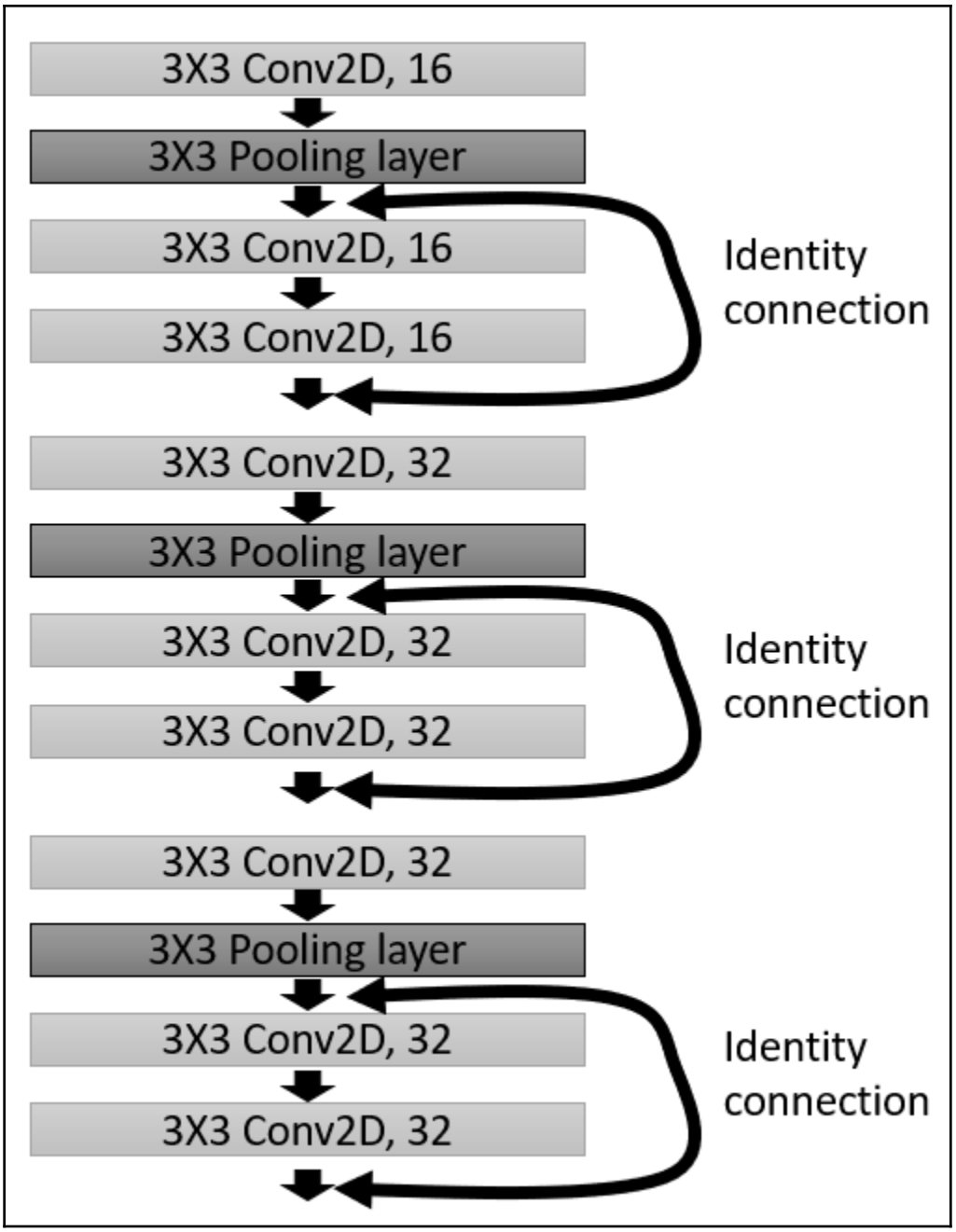

Inspector
Academy
Static
Tag
Untagged
Layer
Default
Transform
Position
Rotation
Scale
Hallway Academy (Script)
Script
HallwayAcademy
Broadcast Hub
Add New
Remove Last
Brains
Control
VisualHallwayLearning (LearningBrain)
Training Configuration
Inference Configuration
Reset Parameters
Add New
Agent Run Speed
1.5
Agent Rotation Speed
1
Goal Scored Material
LightGreen
Fail Material
LightRed
Gravity Multiplier
2
Add Component

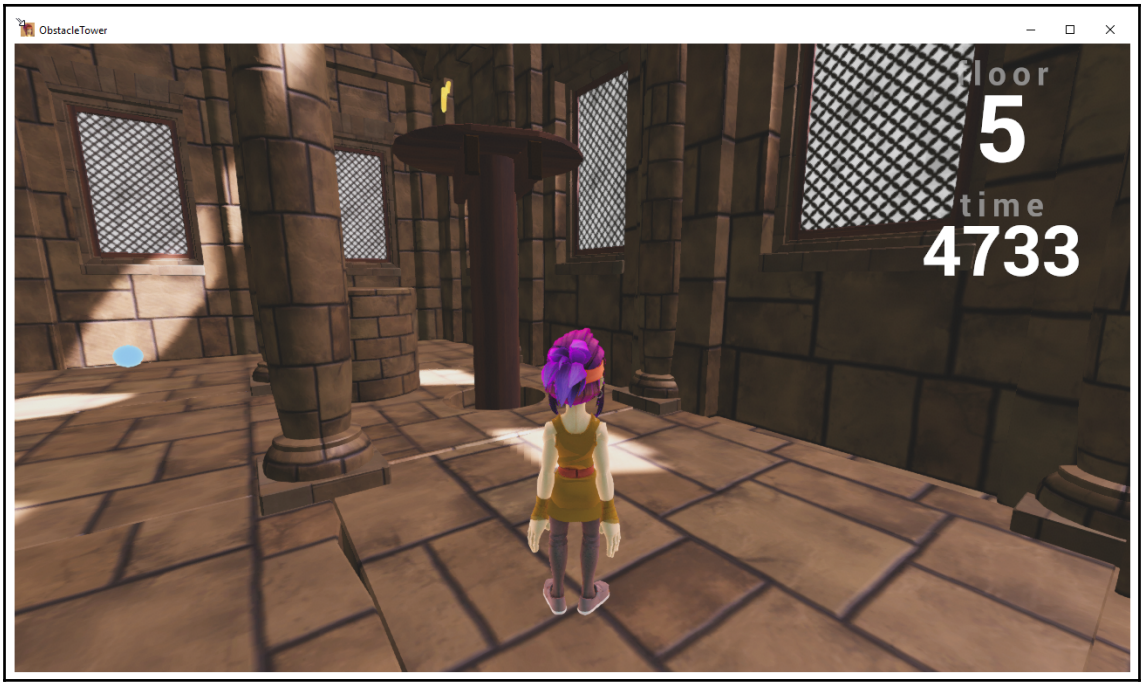

Hierarchy
Create
All
VisualHallway*
Academy
EventSystem
Directional_Light
Canvas_Watermark
Camera
VisualSymbolFinderArea
VisualSymbolFinderArea (1)
VisualSymbolFinderArea (2)
VisualSymbolFinderArea (3)
VisualSymbolFinderArea (4)
VisualSymbolFinderArea (5)
VisualSymbolFinderArea (6)
VisualSymbolFinderArea (7)
Inspector
Scene
Game
Asset Store
Static
Tag
Untagged
Layer
Default
Mutiple
Open
Select
Overrides
Transform
Position
Rotation
Scale
Add Component

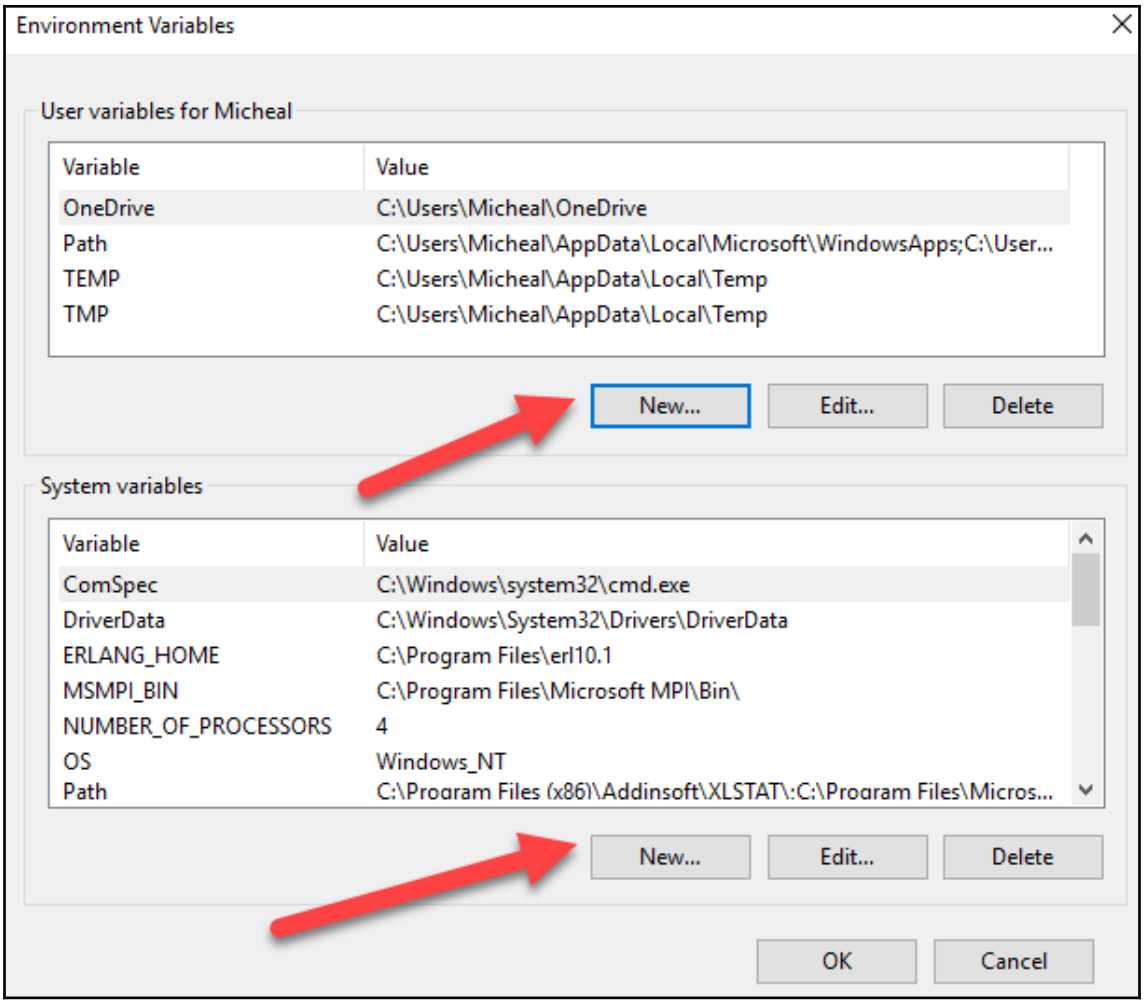
```
2019-11-16 15:25:19.214278: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror if you do not have a GPU
 set up on your machine.
WARNING:tensorflow:
The TensorFlow contrib module will not be included in TensorFlow 2.0.
For more information, please see:
  * https://github.com/tensorflow/community/blob/master/rfcs/20180907-contrib-sunset.md
  * https://github.com/tensorflow/addons
  * https://github.com/tensorflow/io (for I/O related ops)
If you depend on functionality not listed there, please file an issue.
```

```
INFO:mlagents.trainers:CommandLineOptions(debug=False, num_runs=1, seed=-1, env_path=None, run_id='cob_1', load_model=False, train_
del=True, save_freq=50000, keep_checkpoints=5, base_port=5005, num_envs=1, curriculum_folder=None, lesson=0, slow=False, no_graphics
False, multi_gpu=False, trainer_config_path='config/trainer_config.yaml', sampler_file_path=None, docker_target_name=None, env_args=
one, cpu=False)
INFO:mlagents.envs:Start training by pressing the Play button in the Unity Editor.
```


input x
weight layer
F(x)
activation (relu)
x
identity
weight layer
F(x) + x
activation (relu)


3X3 Conv2D, 16
3X3 Pooling layer
3X3 Conv2D, 16
3X3 Conv2D, 16
Identity connection
3X3 Conv2D, 32
3X3 Pooling layer
3X3 Conv2D, 32
3X3 Conv2D, 32
Identity connection
3X3 Conv2D, 32
3X3 Pooling layer
3X3 Conv2D, 32
3X3 Conv2D, 32
Identity connection


ObstacleTower
floor
5
time
4733


Environment Variables
User variables for Micheal
Variable
Value
OneDrive
C:\Users\Micheal\OneDrive
Path
C:\Users\Micheal\AppData\Local\Microsoft\WindowsApps;C:\User...
TEMP
C:\Users\Micheal\AppData\Local\Temp
TMP
C:\Users\Micheal\AppData\Local\Temp
New...
Edit...
Delete
System variables
Variable
Value
ComSpec
C:\Windows\system32\cmd.exe
DriverData
C:\Windows\System32\Drivers\DriverData
ERLANG_HOME
C:\Program Files\erl10.1
MSMPI_BIN
C:\Program Files\Microsoft MPI\Bin\
NUMBER_OF_PROCESSORS
4
OS
Windows_NT
Path
C:\Program Files (x86)\Addinsoft\XLSTAT\;C:\Program Files\Micros...
New...
Edit...
Delete
OK
Cancel

Locked Door

- [ ] Closed door
- [x] Locked door
- [ ] Boxed door
- [ ] Open door
- [ ] Key
- [ ] Box
- [ ] Hurtle
- [ ] Orb
- [ ] Goal
- [ ] Box target
- [ ] Box undo

Save

highest uncertainty/entropy

higher uncertainty/entropy

high uncertainty/entropy

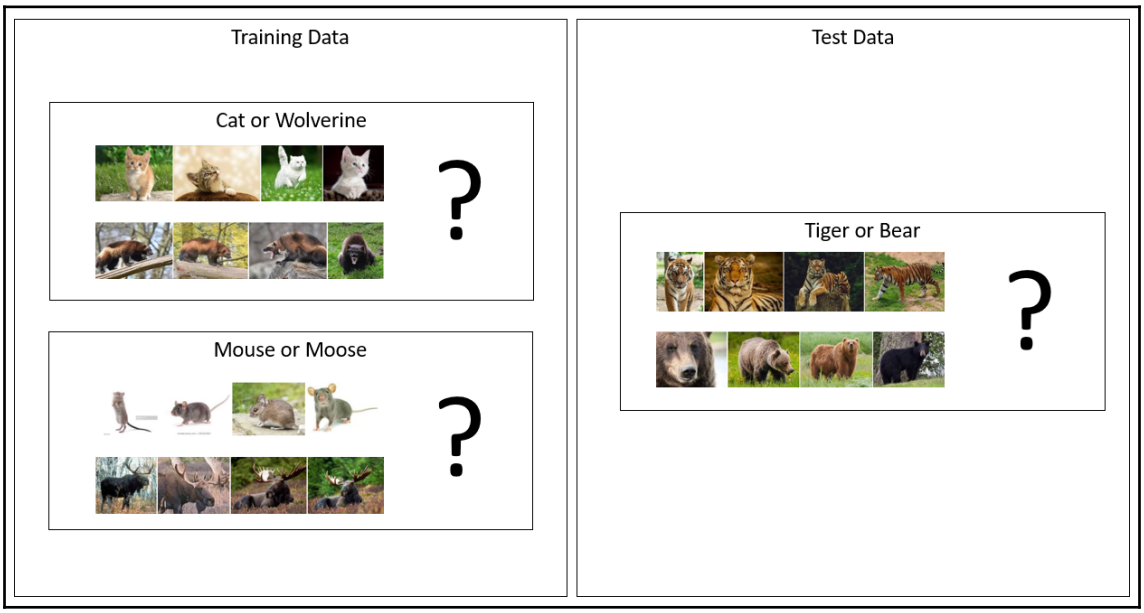
low uncertainty/entropy

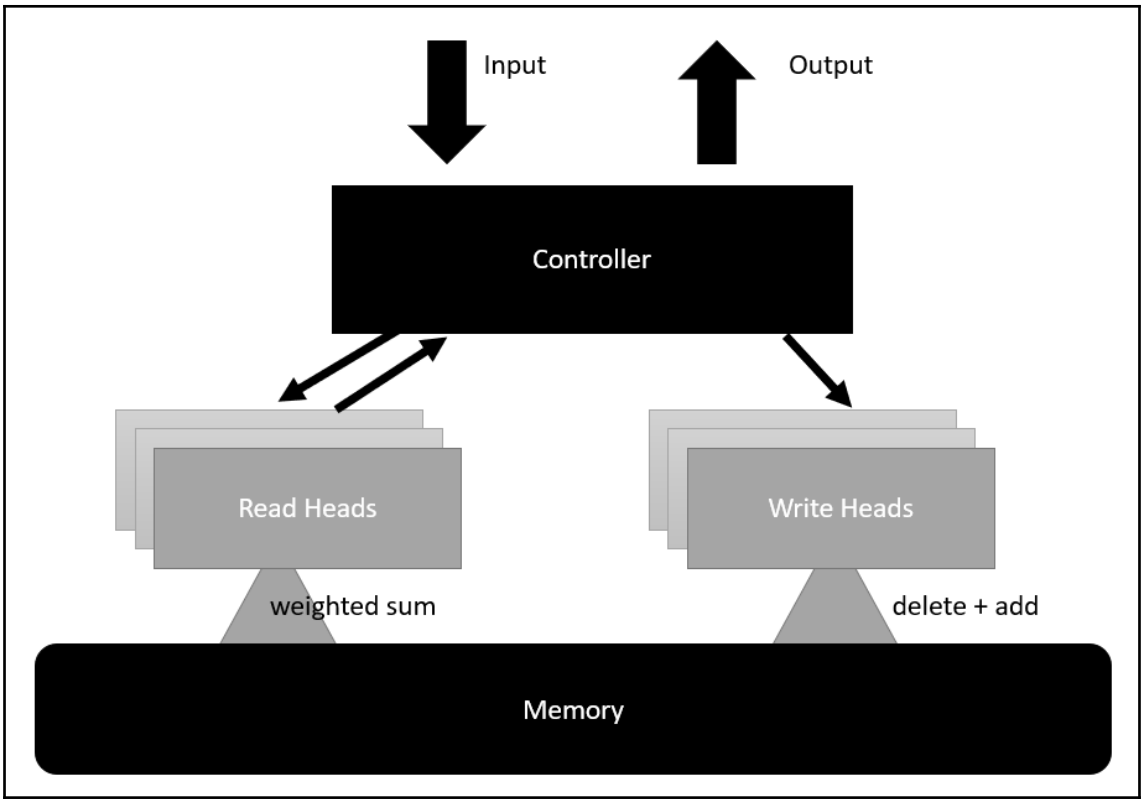
lower uncertainty/entropy

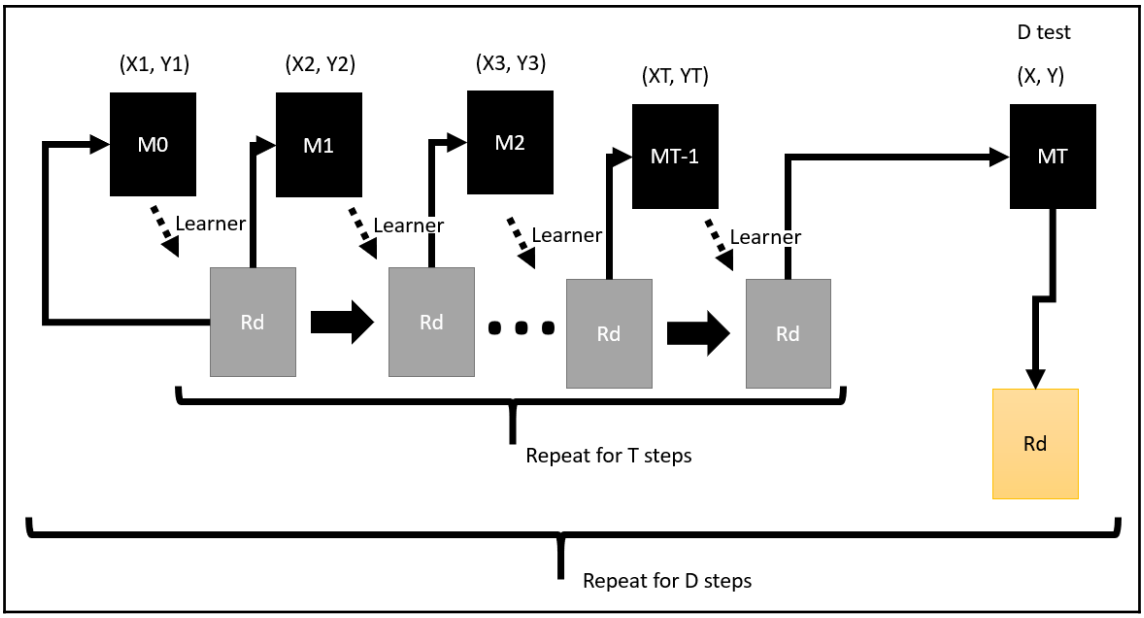
lowest uncertainty/entropy

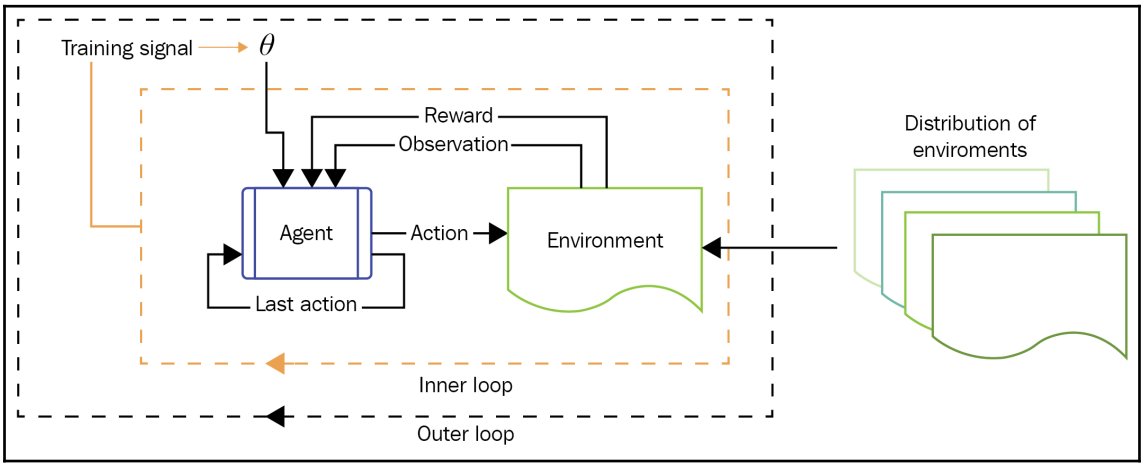
Trajectory

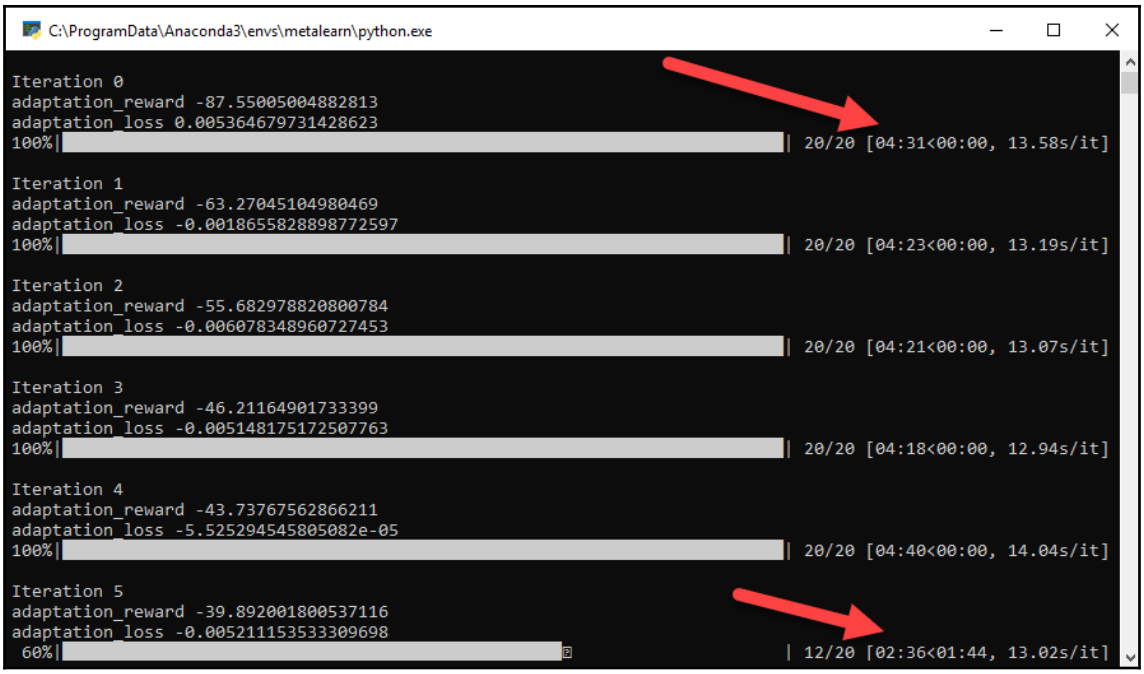
Chapter 14: From DRL to AGI

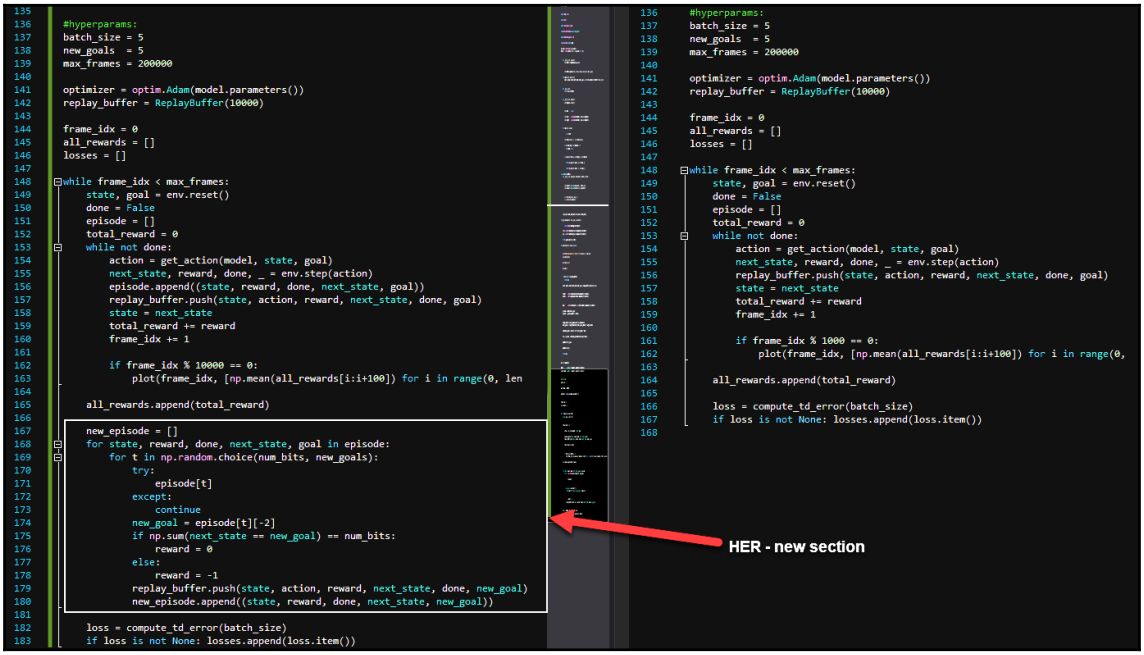


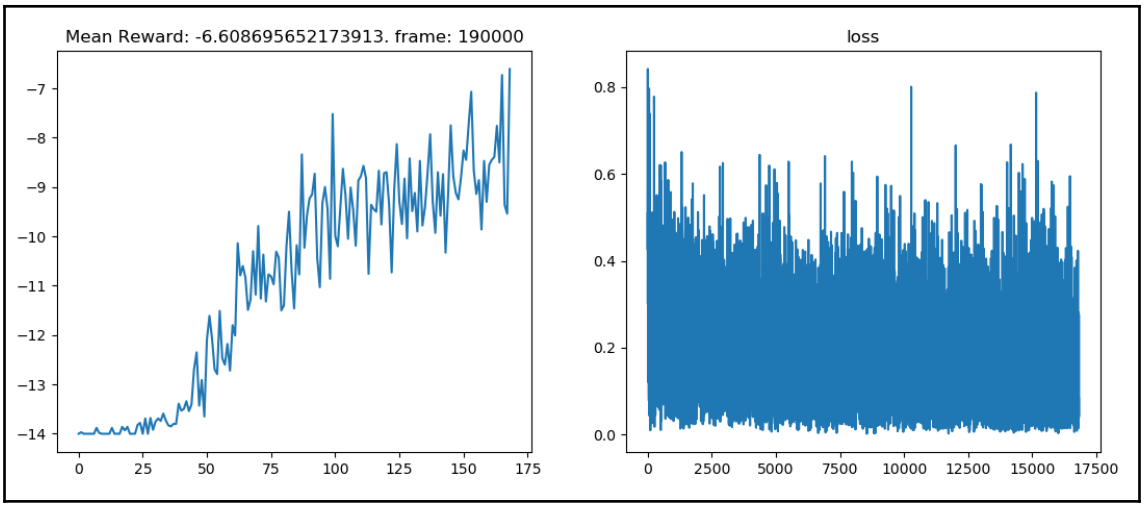

Input
Output
Controller
Read Heads
Write Heads
weighted sum
delete + add
Memory

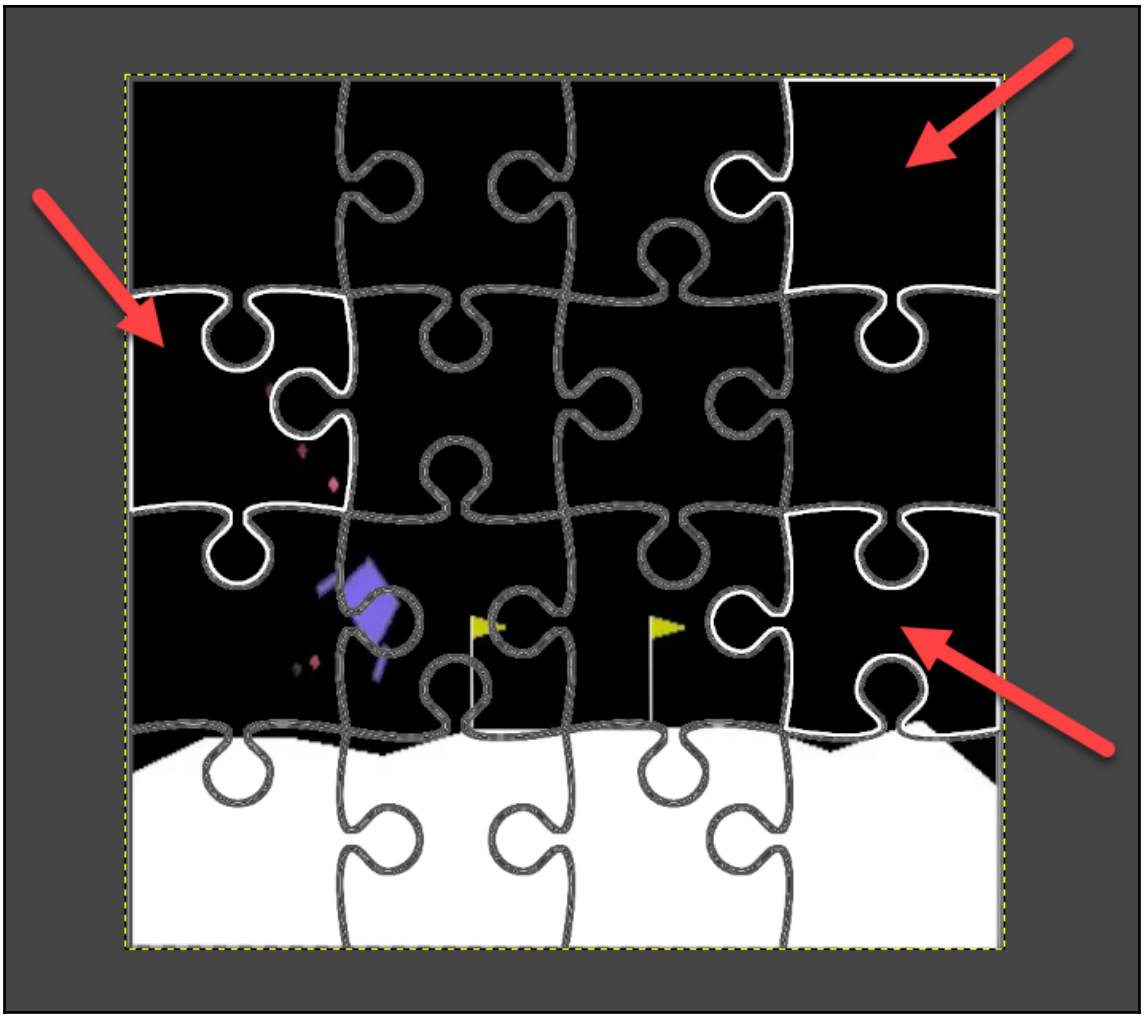

(X1, Y1)
(X2, Y2)
(X3, Y3)
(XT, YT)
D test
(X, Y)
M0
M1
M2
MT-1
MT
Learner
Learner
Learner
Learner
Rd
Rd
Rd
Rd
Rd
Repeat for T steps
Repeat for D steps

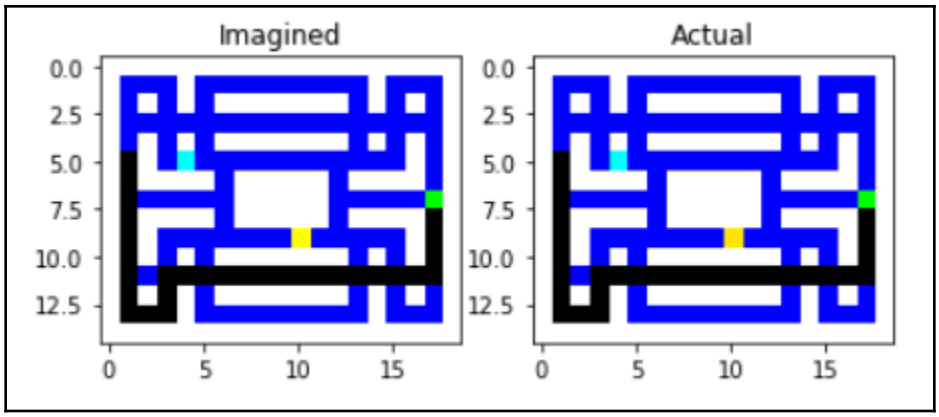

Training signal
θ
Reward
Observation
Agent
Action
Environment
Last action
Distribution of enviroments
Inner loop
Outer loop

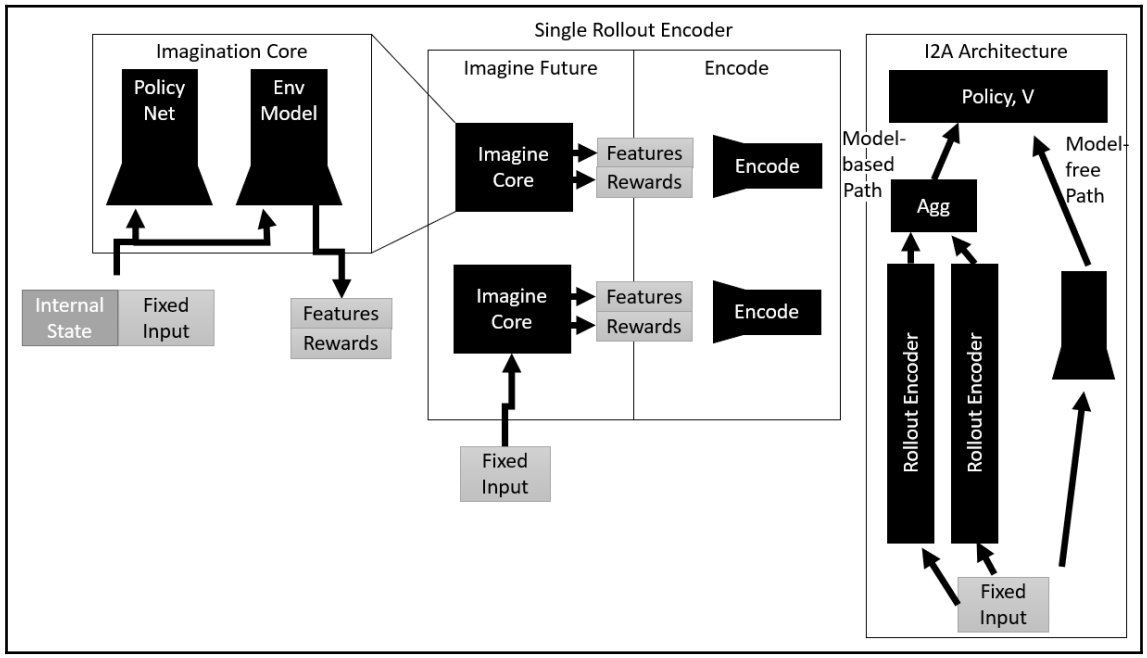

C:\ProgramData\Anaconda3\envs\metalearn\python.exe
Iteration 0
adaptation_reward -87.55005004882813
adaptation_loss 0.005364679731428623
100%| | 20/20 [04:31<00:00, 13.58s/it]
Iteration 1
adaptation_reward -63.27045104980469
adaptation_loss -0.0018655828898772597
100%| | 20/20 [04:23<00:00, 13.19s/it]
Iteration 2
adaptation_reward -55.682978820800784
adaptation_loss -0.006078348960727453
100%| | 20/20 [04:21<00:00, 13.07s/it]
Iteration 3
adaptation_reward -46.21164901733399
adaptation_loss -0.005148175172507763
100%| | 20/20 [04:18<00:00, 12.94s/it]
Iteration 4
adaptation_reward -43.73767562866211
adaptation_loss -5.525294545805082e-05
100%| | 20/20 [04:40<00:00, 14.04s/it]
Iteration 5
adaptation_reward -39.892001800537116
adaptation_loss -0.005211153533309698
60%| | 12/20 [02:36<01:44, 13.02s/it]

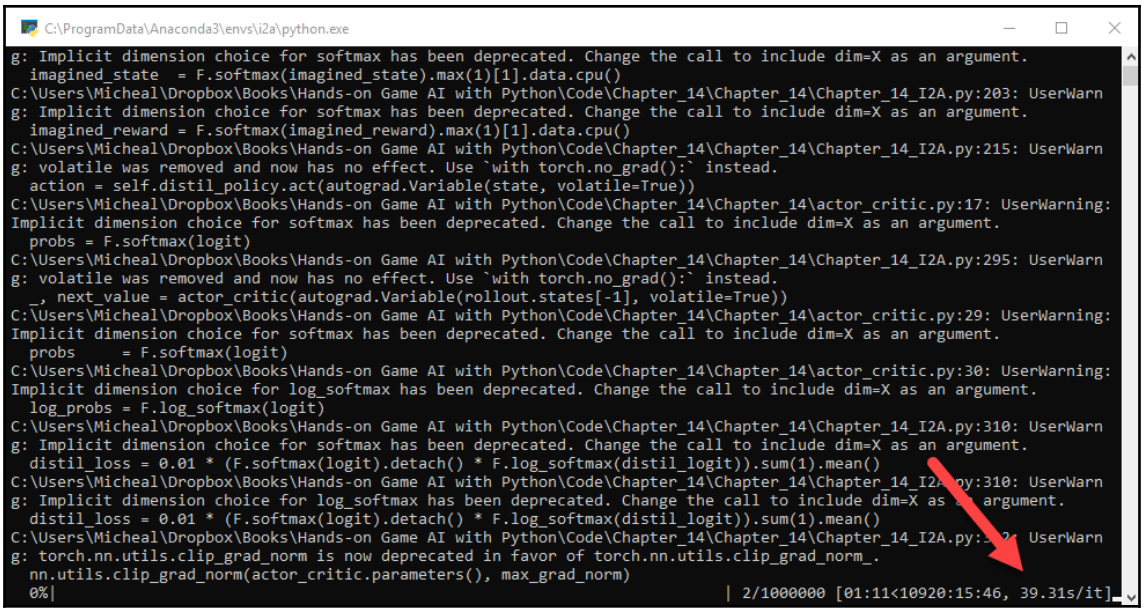

#hyperparams:
batch_size = 5
new_goals = 5
max_frames = 200000
optimizer = optim.Adam(model.parameters())
replay_buffer = ReplayBuffer(10000)
frame_idx = 0
all_rewards = []
losses = []
while frame_idx < max_frames:
state, goal = env.reset()
done = False
episode = []
total_reward = 0
while not done:
action = get_action(model, state, goal)
next_state, reward, done, _ = env.step(action)
episode.append((state, reward, done, next_state, goal))
replay_buffer.push(state, action, reward, next_state, done, goal)
state = next_state
total_reward += reward
frame_idx += 1
if frame_idx % 10000 == 0:
plot(frame_idx, [np.mean(all_rewards[i:i+100]) for i in range(0, len
all_rewards.append(total_reward)
new_episode = []
for state, reward, done, next_state, goal in episode:
for t in np.random.choice(num_bits, new_goals):
try:
episode[t]
except:
continue
new_goal = episode[t][-2]
if np.sum(next_state == new_goal) == num_bits:
reward = 0
else:
reward = -1
replay_buffer.push(state, action, reward, next_state, done, new_goal)
new_episode.append((state, reward, done, next_state, new_goal))
loss = compute_td_error(batch_size)
if loss is not None: losses.append(loss.item())
HER - new section


Mean Reward: -6.608695652173913. frame: 190000
−7
−8
−9
−10
−11
−12
−13
−14
0
25
50
75
100
125
150
175
loss
0.8
0.6
0.4
0.2
0.0
0
2500
5000
7500
10000
12500
15000
17500




Imagined
Actual
0.0
2.5
5.0
7.5
10.0
12.5
0
5
10
15


Imagination Core
Policy Net
Env Model
Internal State
Fixed Input
Features
Rewards
Single Rollout Encoder
Imagine Future
Encode
Imagine Core
Features
Rewards
Encode
Imagine Core
Features
Rewards
Encode
Fixed Input
I2A Architecture
Policy, V
Model-based Path
Model-free Path
Agg
Rollout Encoder
Rollout Encoder
Fixed Input


C:\ProgramData\Anaconda3\envs\i2a\python.exe
g: Implicit dimension choice for softmax has been deprecated. Change the call to include dim=X as an argument.
imagined_state = F.softmax(imagined_state).max(1)[1].data.cpu()
C:\Users\Micheal\Dropbox\Books\Hands-on Game AI with Python\Code\Chapter_14\Chapter_14\Chapter_14_I2A.py:203: UserWarn
g: Implicit dimension choice for softmax has been deprecated. Change the call to include dim=X as an argument.
imagined_reward = F.softmax(imagined_reward).max(1)[1].data.cpu()
C:\Users\Micheal\Dropbox\Books\Hands-on Game AI with Python\Code\Chapter_14\Chapter_14\Chapter_14_I2A.py:215: UserWarn
g: volatile was removed and now has no effect. Use `with torch.no_grad():` instead.
action = self.distil_policy.act(autograd.Variable(state, volatile=True))
C:\Users\Micheal\Dropbox\Books\Hands-on Game AI with Python\Code\Chapter_14\Chapter_14\actor_critic.py:17: UserWarning:
Implicit dimension choice for softmax has been deprecated. Change the call to include dim=X as an argument.
probs = F.softmax(logit)
C:\Users\Micheal\Dropbox\Books\Hands-on Game AI with Python\Code\Chapter_14\Chapter_14\Chapter_14_I2A.py:295: UserWarn
g: volatile was removed and now has no effect. Use `with torch.no_grad():` instead.
_, next_value = actor_critic(autograd.Variable(rollout.states[-1], volatile=True))
C:\Users\Micheal\Dropbox\Books\Hands-on Game AI with Python\Code\Chapter_14\Chapter_14\actor_critic.py:29: UserWarning:
Implicit dimension choice for softmax has been deprecated. Change the call to include dim=X as an argument.
probs = F.softmax(logit)
C:\Users\Micheal\Dropbox\Books\Hands-on Game AI with Python\Code\Chapter_14\Chapter_14\actor_critic.py:30: UserWarning:
Implicit dimension choice for log_softmax has been deprecated. Change the call to include dim=X as an argument.
log_probs = F.log_softmax(logit)
C:\Users\Micheal\Dropbox\Books\Hands-on Game AI with Python\Code\Chapter_14\Chapter_14\Chapter_14_I2A.py:310: UserWarn
g: Implicit dimension choice for softmax has been deprecated. Change the call to include dim=X as an argument.
distil_loss = 0.01 * (F.softmax(logit).detach() * F.log_softmax(distil_logit)).sum(1).mean()
C:\Users\Micheal\Dropbox\Books\Hands-on Game AI with Python\Code\Chapter_14\Chapter_14\Chapter_14_I2A.py:310: UserWarn
g: Implicit dimension choice for log_softmax has been deprecated. Change the call to include dim=X as an argument.
distil_loss = 0.01 * (F.softmax(logit).detach() * F.log_softmax(distil_logit)).sum(1).mean()
C:\Users\Micheal\Dropbox\Books\Hands-on Game AI with Python\Code\Chapter_14\Chapter_14\Chapter_14_I2A.py:
g: torch.nn.utils.clip_grad_norm is now deprecated in favor of torch.nn.utils.clip_grad_norm_.
nn.utils.clip_grad_norm(actor_critic.parameters(), max_grad_norm)
0%| | 2/1000000 [01:11<10920:15:46, 39.31s/it]