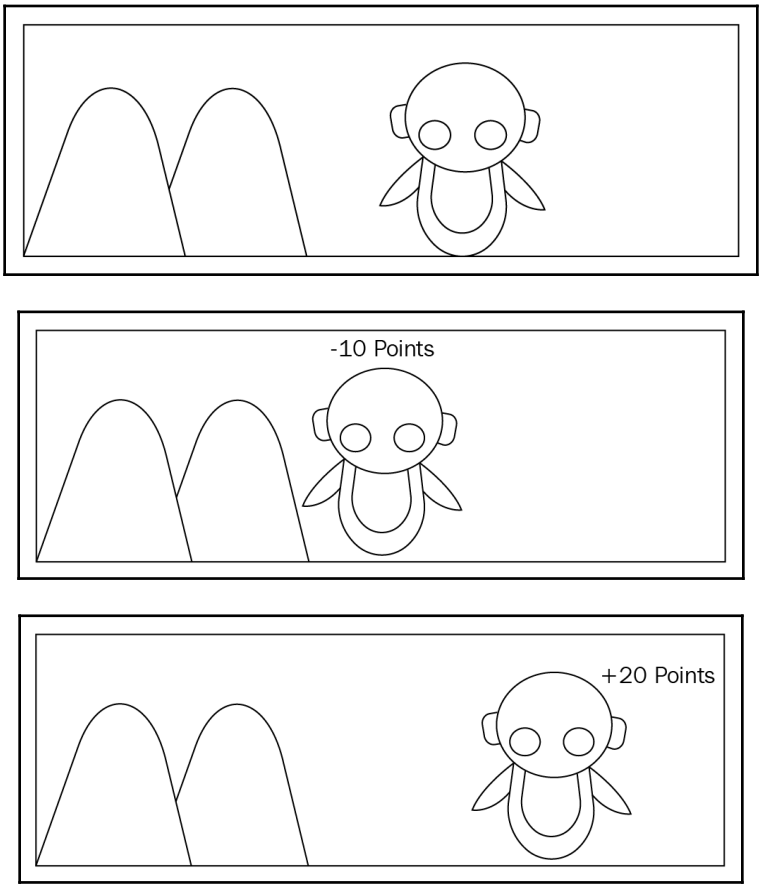
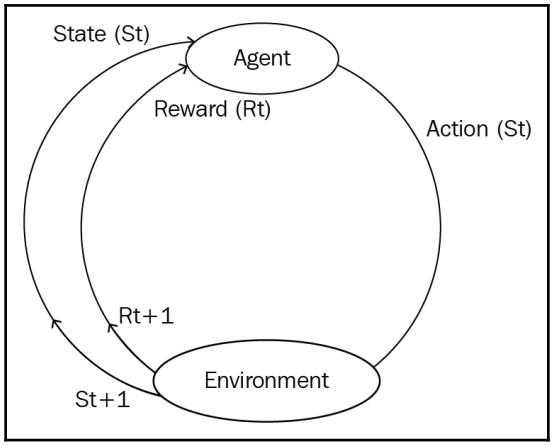
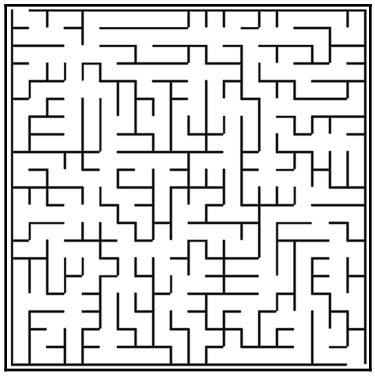
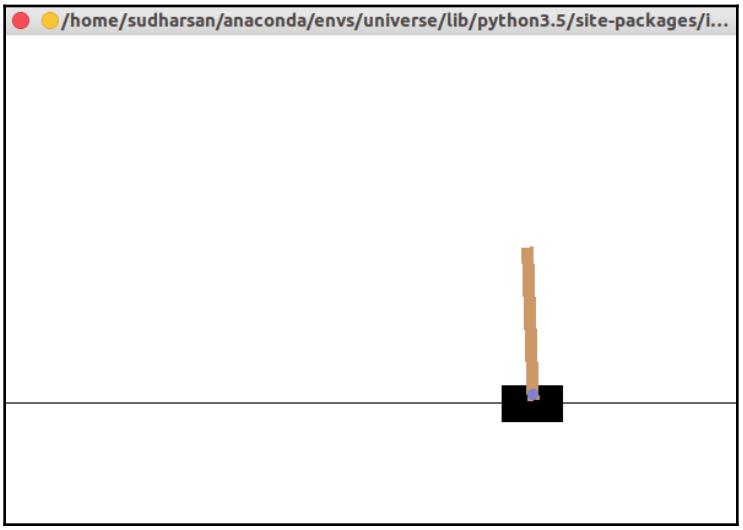
Chapter 1: Introduction to Reinforcement Learning

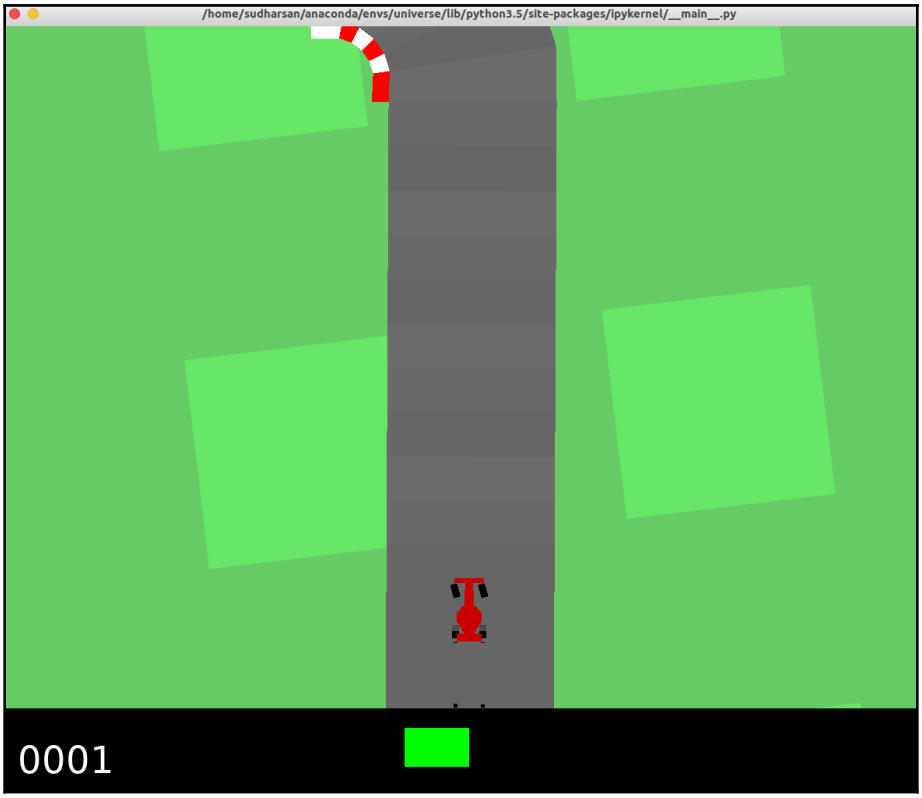


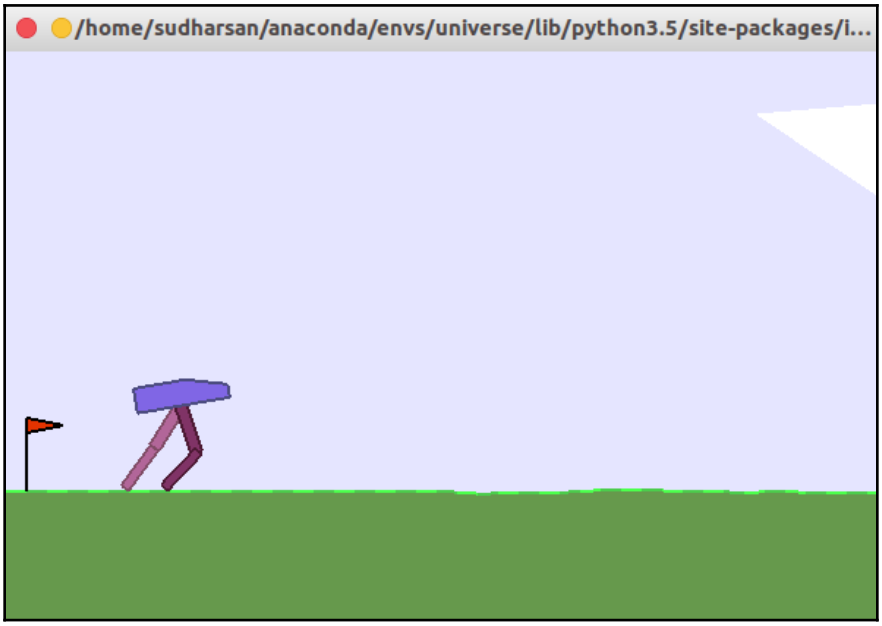

State (St)
Agent
Reward (Rt)
Action (St)
Rt+1
Environment
St+1



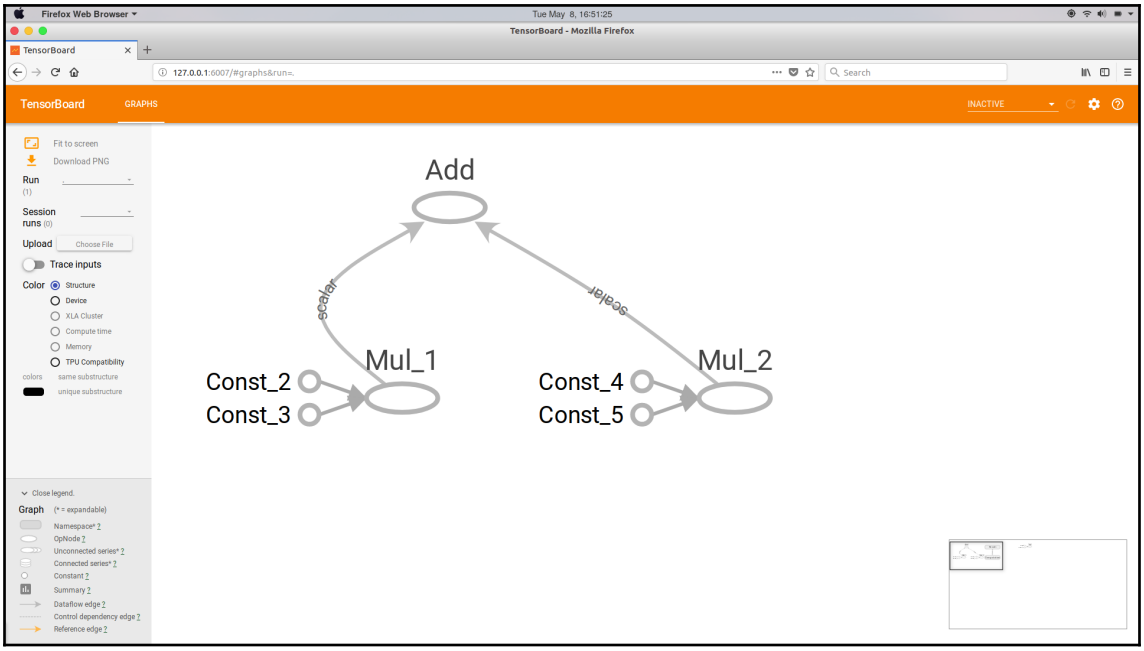
Chapter 2: Getting Started with OpenAI and TensorFlow

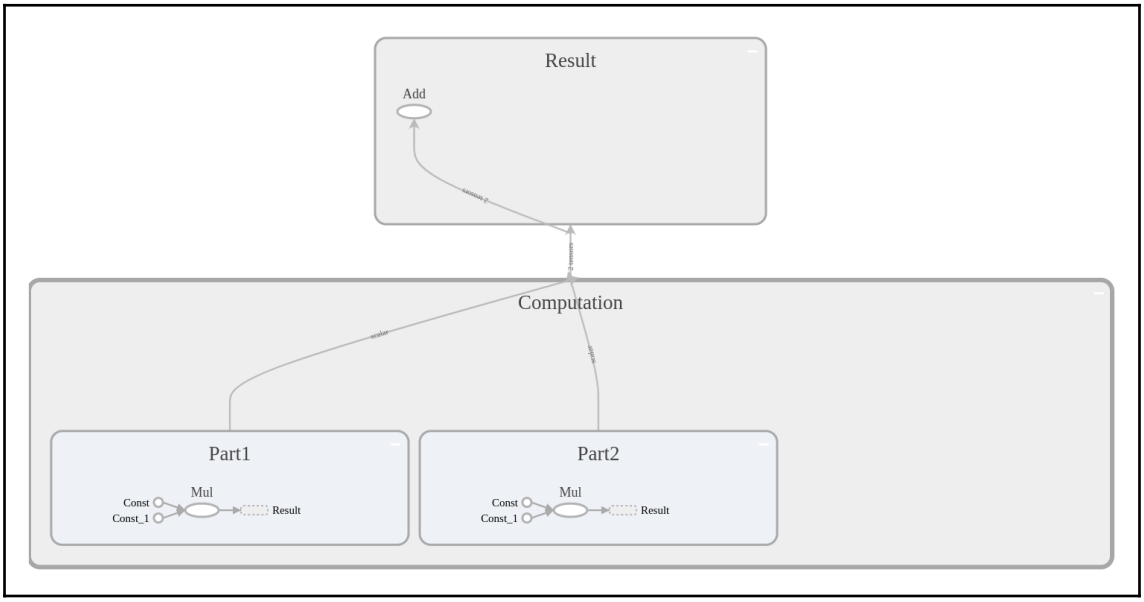



/home/sudharsan/anaconda/envs/universe/lib/python3.5/site-packages/ipykernel/__main__.py
0001

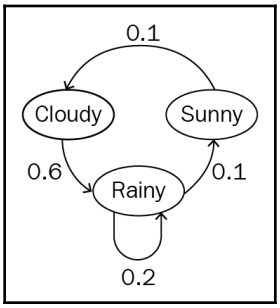

/home/sudharsan/anaconda/envs/universe/lib/python3.5/site-packages/i...

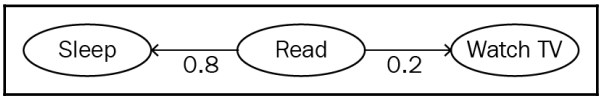

SCORE
16338
TIME
01:02:23
DISTANCE
TURBO
HELP
PAUSE
NEON RACER
LongAnimals Games 2010
MORE GAMES!

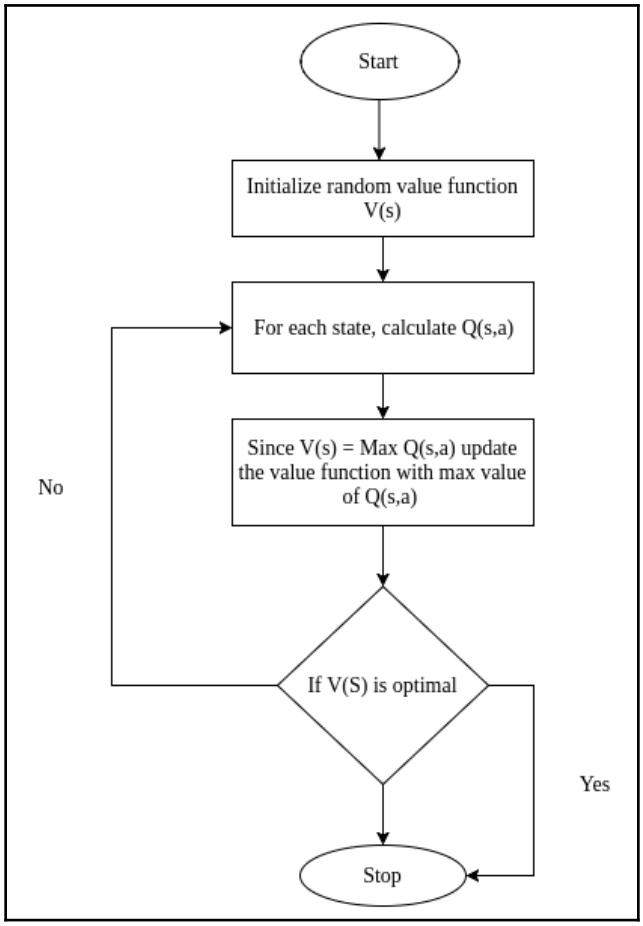

Firefox Web Browser
Tue May 8, 16:51:25
TensorBoard - Mozilla Firefox
TensorBoard
127.0.0.1:6007/#graphs&run=.
Search
TensorBoard
GRAPHS
INACTIVE
Fit to screen
Download PNG
Run
(1)
Session runs (0)
Upload
Choose File
Trace inputs
Color
Structure
Device
XLA Cluster
Compute time
Memory
TPU Compatibility
colors
same substructure
unique substructure
Close legend.
Graph
(* = expandable)
Namespace* ?
OpNode ?
Unconnected series* ?
Connected series* ?
Constant ?
Summary ?
Dataflow edge ?
Control dependency edge ?
Reference edge ?
Add
scalar
scalar
Mul_1
Mul_2
Const_2
Const_3
Const_4
Const_5


Result
Add
Computation
Part1
Mul
Const
Const_1
Result
Part2
Mul
Const
Const_1
Result

Chapter 3: The Markov Decision Process and Dynamic Programming





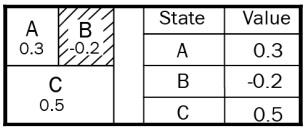


State	Value
A	0
B	0
C	0

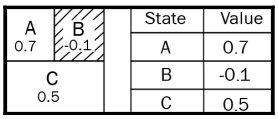
State (s)	Action (a)	Next State (s')	Transistion Probability ($P^a_{ss'}$)	Reward Probability ($R^a_{ss'}$)
A	0	A	0.1	0
A	0	B	0.4	-1.0
A	0	C	0.3	1.0
A	1	A	0.3	0
A	1	B	0.1	-2.0
A	1	C	0.5	1.0

State	Action	Value
A	0	-0.1
A	1	0.3
B	0	
B	1	
C	0	
C	1	

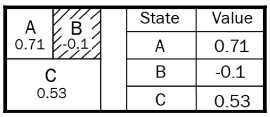
State	Value
A	0.3
B	
C	



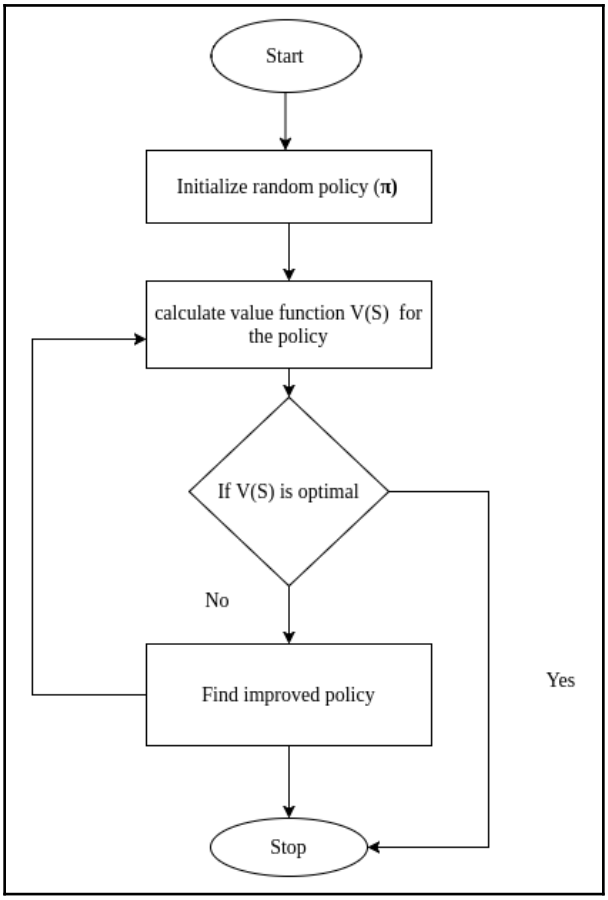
State	Value
A	0.3
B	-0.2
C	0.5

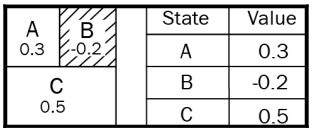


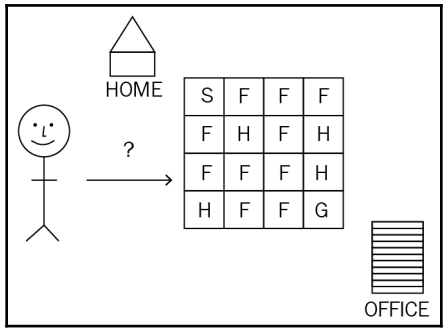
State	Value
A	0.7
B	-0.1
C	0.5



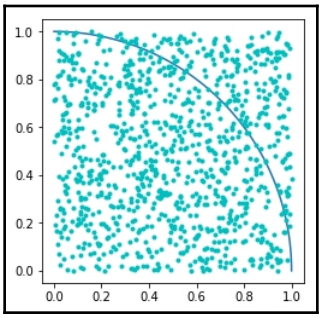
State	Value
A	0.71
B	-0.1
C	0.53

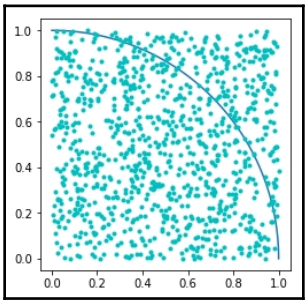

Start
Initialize random policy (π)
calculate value function V(S) for the policy
If V(S) is optimal
No
Yes
Find improved policy
Stop

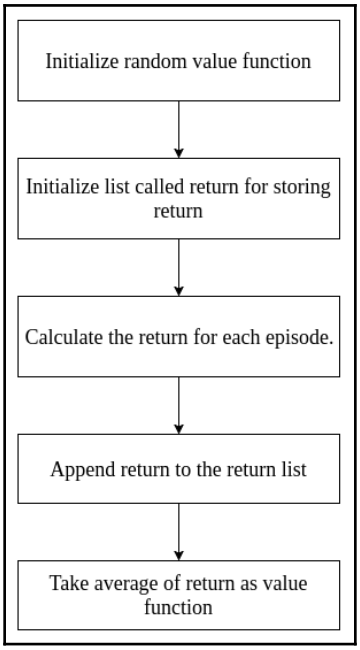

A
0.3
B
-0.2
C
0.5
State
Value
A
0.3
B
-0.2
C
0.5

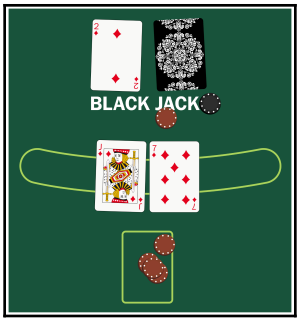

HOME
?
S F F F
F H F H
F F F H
H F F G
OFFICE

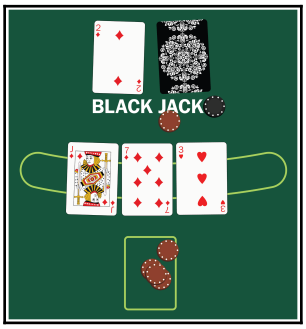
Chapter 4: Gaming with Monte Carlo Methods

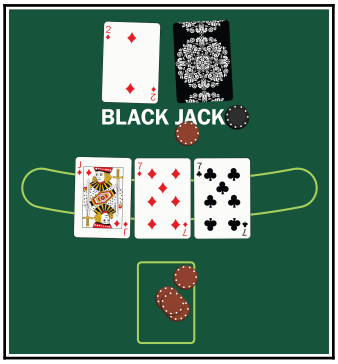


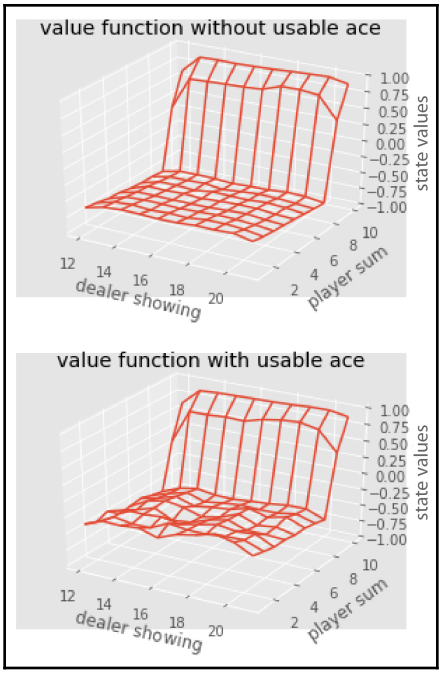


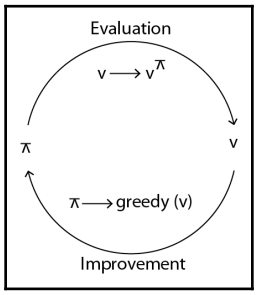

Initialize random value function
Initialize list called return for storing return
Calculate the return for each episode.
Append return to the return list
Take average of return as value function

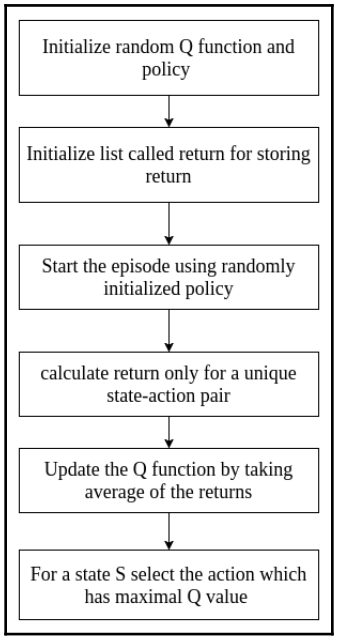

BLACK JACK

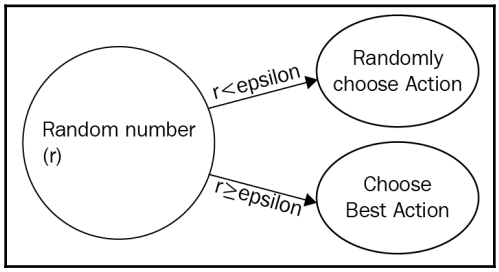

BLACK JACK


BLACK JACK

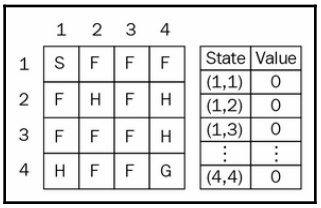

value function without usable ace
state values
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
dealer showing
12
14
16
18
20
player sum
2
4
6
8
10
value function with usable ace
state values
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
dealer showing
12
14
16
18
20
player sum
2
4
6
8
10


Evaluation
v ⟶ v^π
π
v
π ⟶ greedy (v)
Improvement


Initialize random Q function and policy
Initialize list called return for storing return
Start the episode using randomly initialized policy
calculate return only for a unique state-action pair
Update the Q function by taking average of the returns
For a state S select the action which has maximal Q value

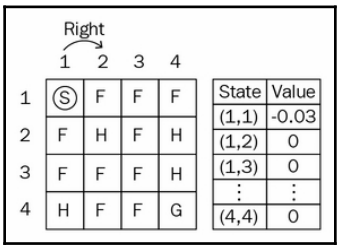

Random number (r)
r<epsilon
Randomly choose Action
r≥epsilon
Choose Best Action

Chapter 5: Temporal Difference Learning



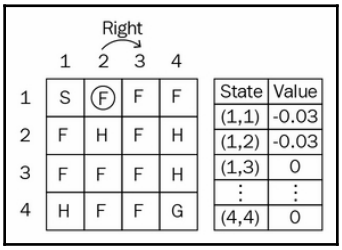
	1	2	3	4
1	S	F	F	F
2	F	H	F	H
3	F	F	F	H
4	H	F	F	G

State	Value
(1,1)	0
(1,2)	0
(1,3)	0
⋮	⋮
(4,4)	0



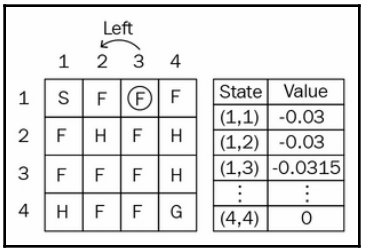
	1	2	3	4
1	Ⓢ	F	F	F
2	F	H	F	H
3	F	F	F	H
4	H	F	F	G

State	Value
(1,1)	-0.03
(1,2)	0
(1,3)	0
⋮	⋮
(4,4)	0



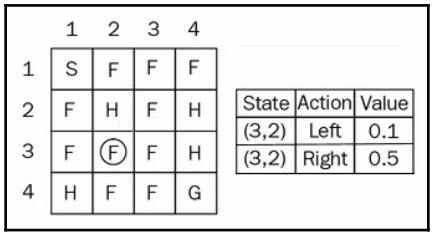
	1	2	3	4
1	S	Ⓕ	F	F
2	F	H	F	H
3	F	F	F	H
4	H	F	F	G

State	Value
(1,1)	-0.03
(1,2)	-0.03
(1,3)	0
⋮	⋮
(4,4)	0



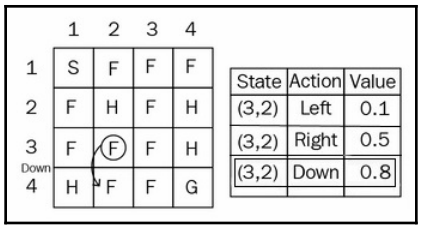
	1	2	3	4
1	S	F	Ⓕ	F
2	F	H	F	H
3	F	F	F	H
4	H	F	F	G

State	Value
(1,1)	-0.03
(1,2)	-0.03
(1,3)	-0.0315
⋮	⋮
(4,4)	0



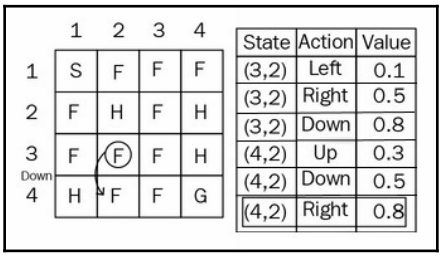
	1	2	3	4
1	S	F	F	F
2	F	H	F	H
3	F	(F)	F	H
4	H	F	F	G

State	Action	Value
(3,2)	Left	0.1
(3,2)	Right	0.5



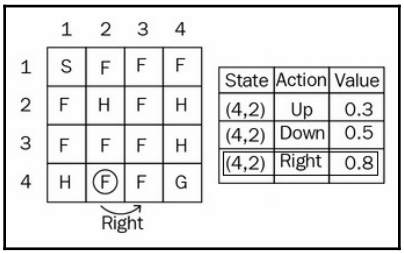
	1	2	3	4
1	S	F	F	F
2	F	H	F	H
3	F	(F)	F	H
Down				
4	H	F	F	G

State	Action	Value
(3,2)	Left	0.1
(3,2)	Right	0.5
(3,2)	Down	0.8



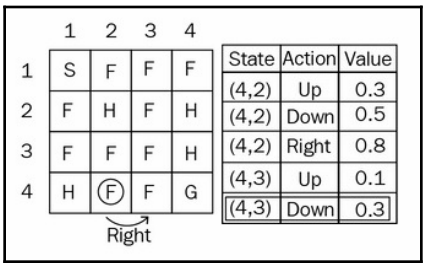
	1	2	3	4
1	S	F	F	F
2	F	H	F	H
3	F	(F)	F	H
Down				
4	H	F	F	G

State	Action	Value
(3,2)	Left	0.1
(3,2)	Right	0.5
(3,2)	Down	0.8
(4,2)	Up	0.3
(4,2)	Down	0.5
(4,2)	Right	0.8



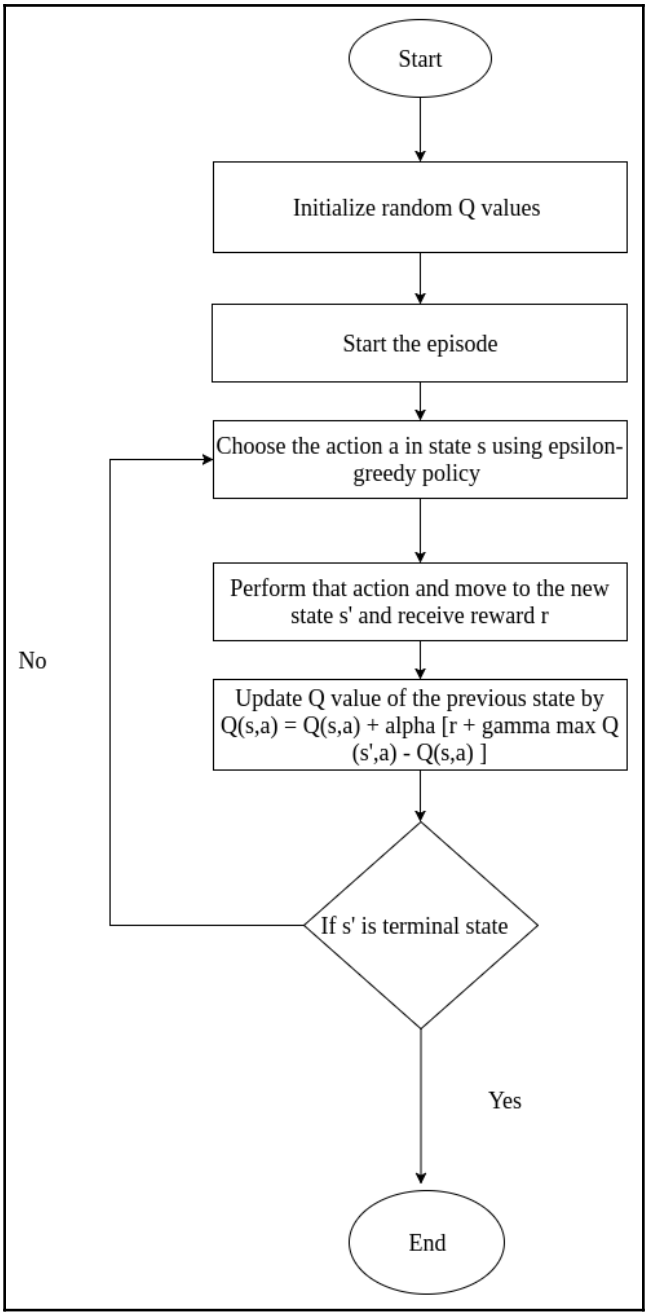
	1	2	3	4
1	S	F	F	F
2	F	H	F	H
3	F	F	F	H
4	H	(F)	F	G
		Right		

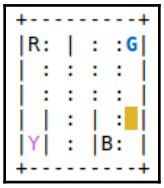
State	Action	Value
(4,2)	Up	0.3
(4,2)	Down	0.5
(4,2)	Right	0.8

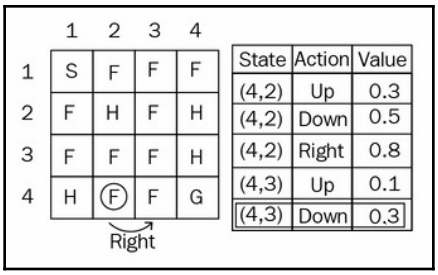


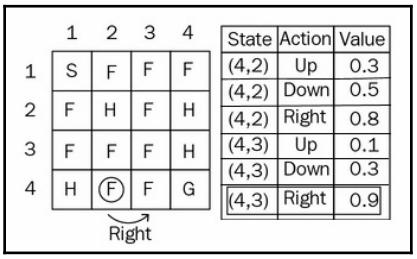
	1	2	3	4
1	S	F	F	F
2	F	H	F	H
3	F	F	F	H
4	H	(F)	F	G
		Right		

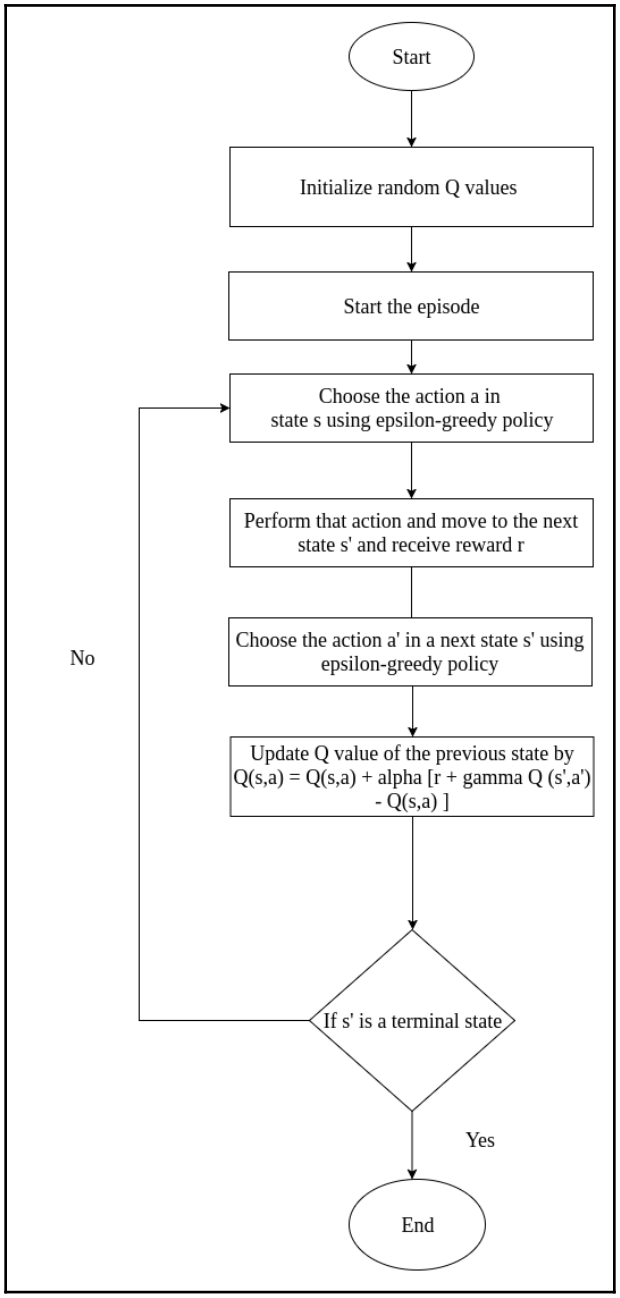
State	Action	Value
(4,2)	Up	0.3
(4,2)	Down	0.5
(4,2)	Right	0.8
(4,3)	Up	0.1
(4,3)	Down	0.3

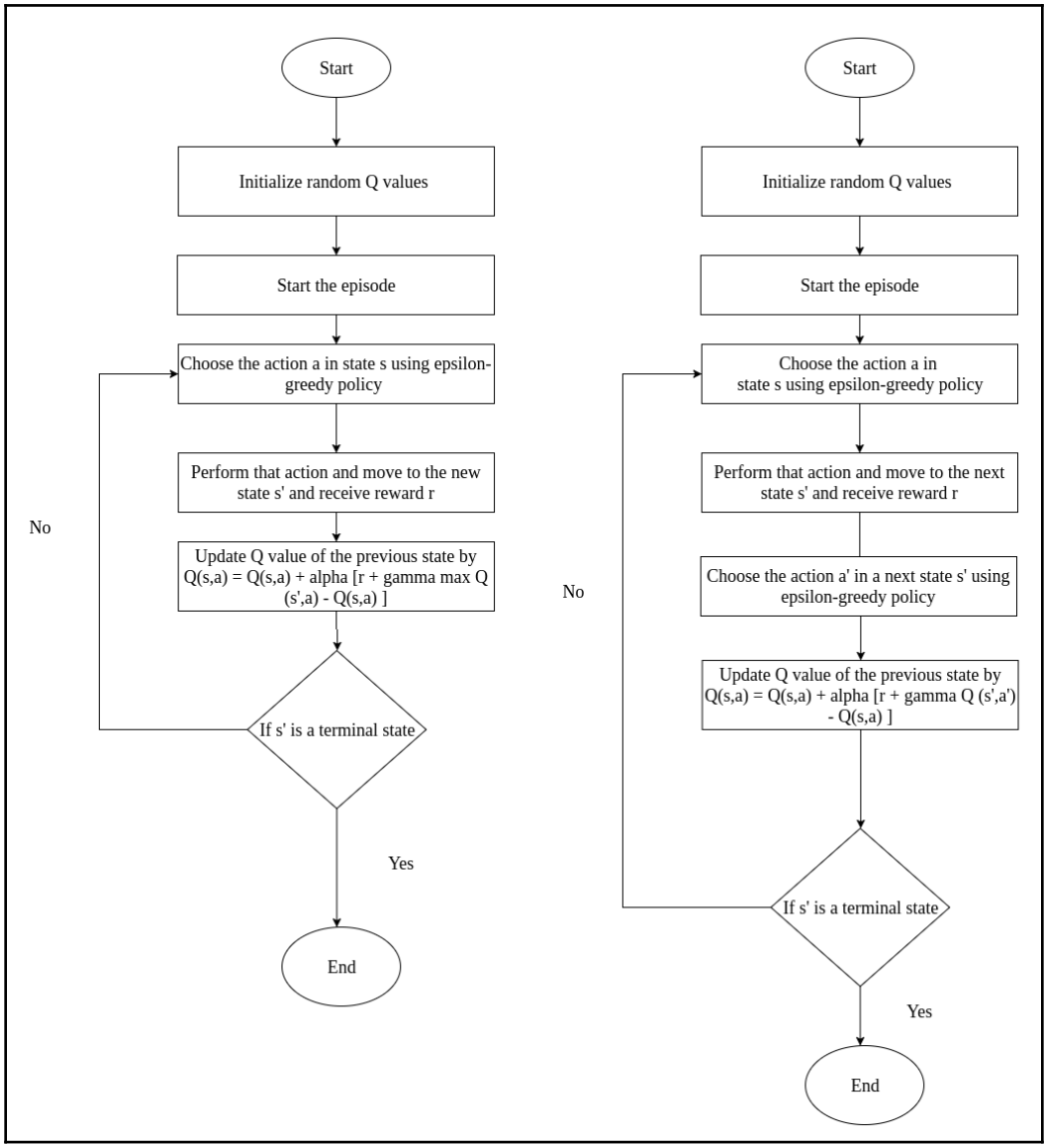

Start
Initialize random Q values
Start the episode
Choose the action a in state s using epsilon-greedy policy
Perform that action and move to the new state s' and receive reward r
Update Q value of the previous state by Q(s,a) = Q(s,a) + alpha [r + gamma max Q (s',a) - Q(s,a)]
If s' is terminal state
No
Yes
End


+---------+
|R: | : :G|
| : : : : |
| : : : : |
| | : | : |
|Y| : |B: |
+---------+

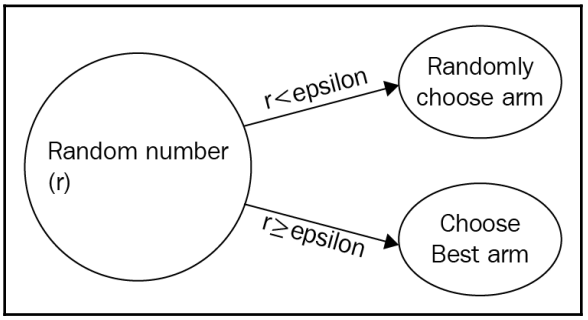

1 2 3 4
1 S F F F
2 F H F H
3 F F F H
4 H F F G
Right
State Action Value
(4,2) Up 0.3
(4,2) Down 0.5
(4,2) Right 0.8
(4,3) Up 0.1
(4,3) Down 0.3

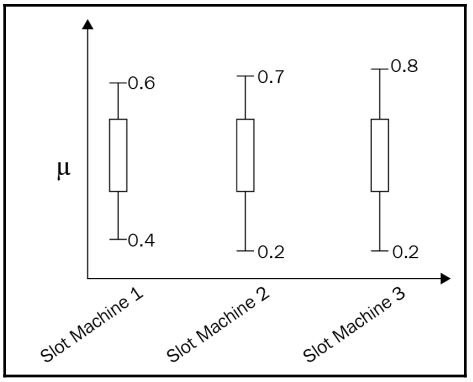

1 2 3 4
1 S F F F
2 F H F H
3 F F F H
4 H F F G
Right
State Action Value
(4,2) Up 0.3
(4,2) Down 0.5
(4,2) Right 0.8
(4,3) Up 0.1
(4,3) Down 0.3
(4,3) Right 0.9

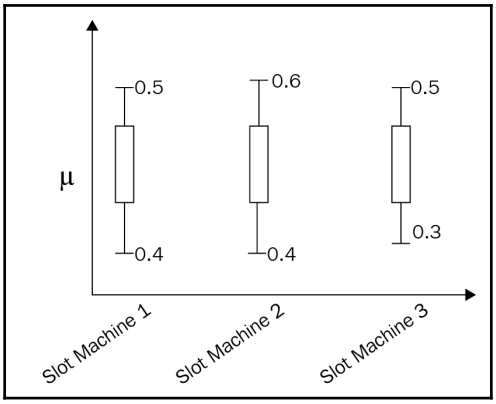

Start
Initialize random Q values
Start the episode
Choose the action a in
state s using epsilon-greedy policy
Perform that action and move to the next
state s' and receive reward r
Choose the action a' in a next state s' using
epsilon-greedy policy
Update Q value of the previous state by
Q(s,a) = Q(s,a) + alpha [r + gamma Q (s',a')
- Q(s,a)]
If s' is a terminal state
No
Yes
End

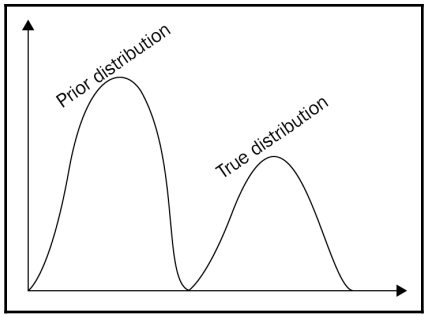

Start
Initialize random Q values
Start the episode
Choose the action a in state s using epsilon-greedy policy
Perform that action and move to the new state s' and receive reward r
Update Q value of the previous state by Q(s,a) = Q(s,a) + alpha [r + gamma max Q (s',a) - Q(s,a)]
If s' is a terminal state
No
Yes
End
Start
Initialize random Q values
Start the episode
Choose the action a in state s using epsilon-greedy policy
Perform that action and move to the next state s' and receive reward r
Choose the action a' in a next state s' using epsilon-greedy policy
Update Q value of the previous state by Q(s,a) = Q(s,a) + alpha [r + gamma Q (s',a') - Q(s,a)]
If s' is a terminal state
No
Yes
End

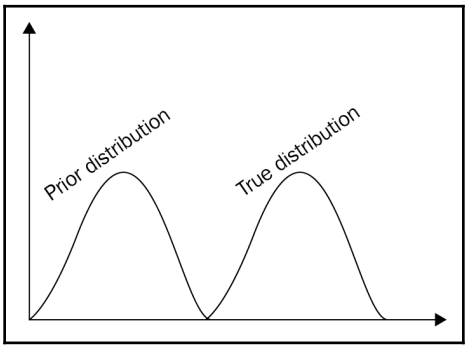
Chapter 6: Multi-Armed Bandit Problem



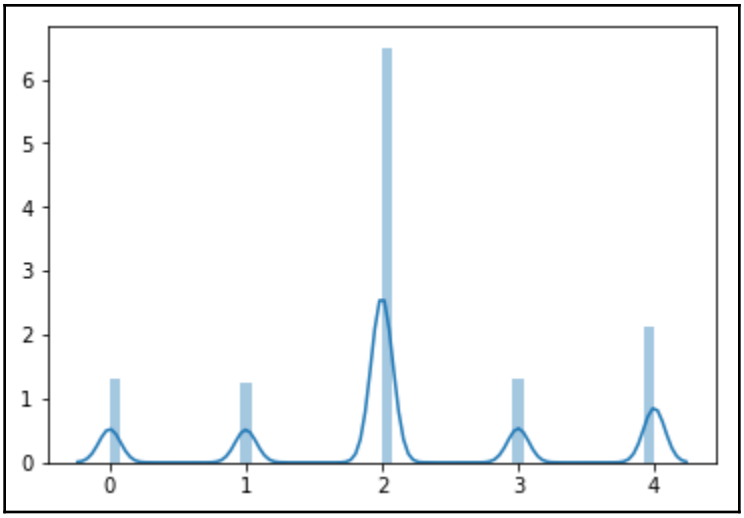



μ
0.5
0.4
0.6
0.4
0.5
0.3
Slot Machine 1
Slot Machine 2
Slot Machine 3

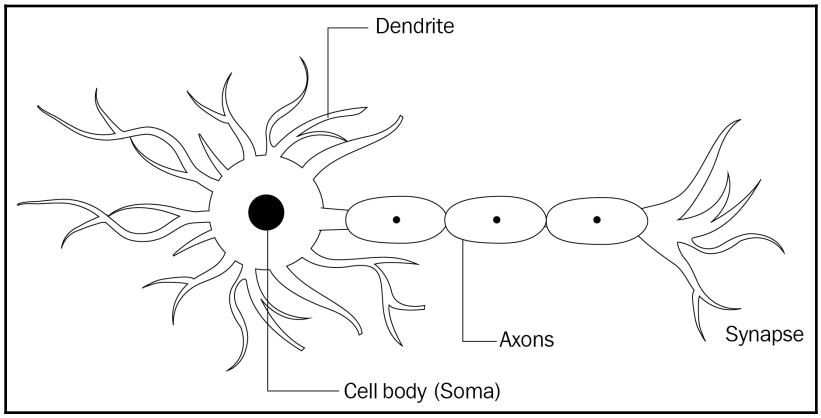

Prior distribution
True distribution

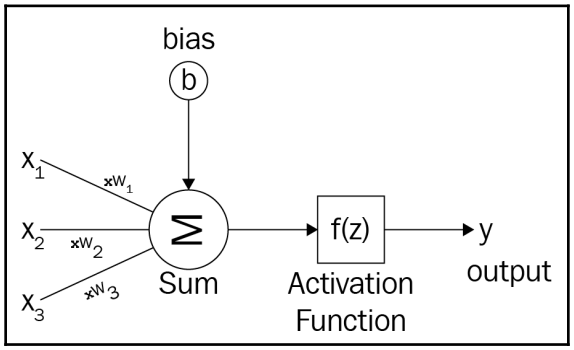

Prior distribution
True distribution

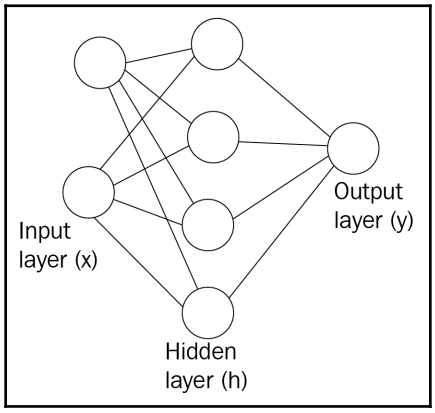
	Banner_type_0	Banner_type_1	Banner_type_2	Banner_type_3	Banner_type_4
0	1	1	0	1	1
1	0	1	1	1	0
2	1	1	0	0	1
3	0	0	0	0	1
4	0	1	1	1	1
5	0	1	1	0	1
6	1	0	0	1	1
7	0	1	1	0	1
8	0	0	1	0	1
9	0	0	0	1	0

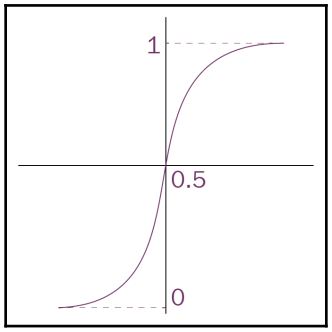


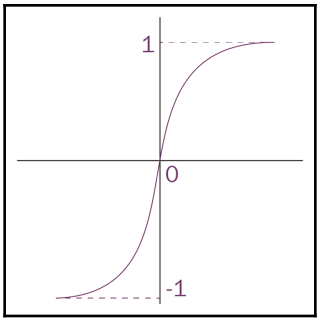
Chapter 7: Deep Learning Fundamentals

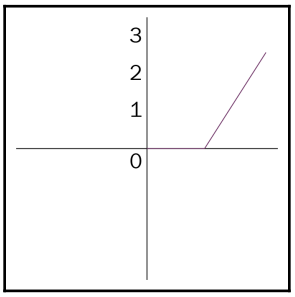


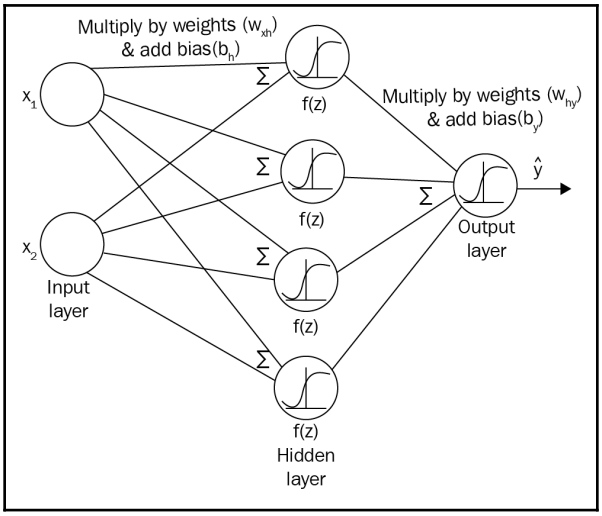


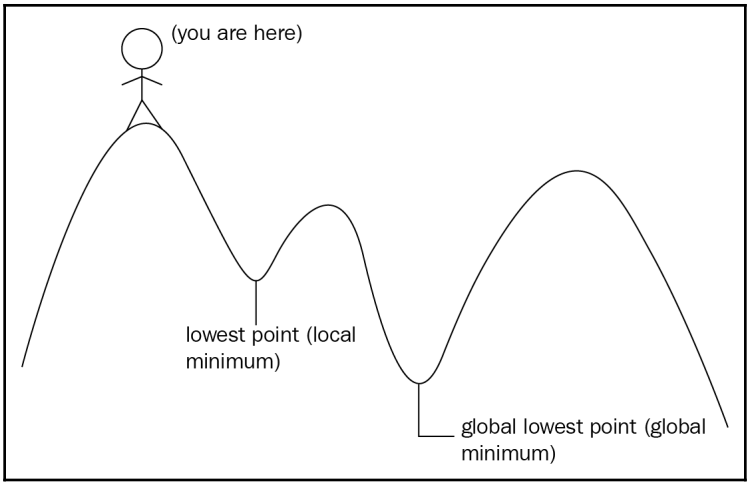


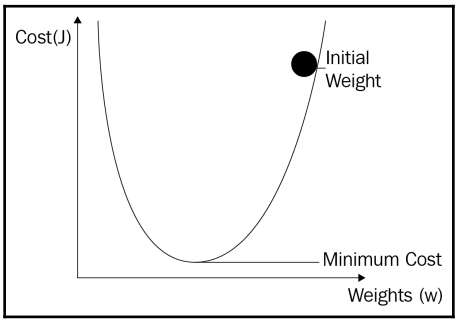

1
0.5
0

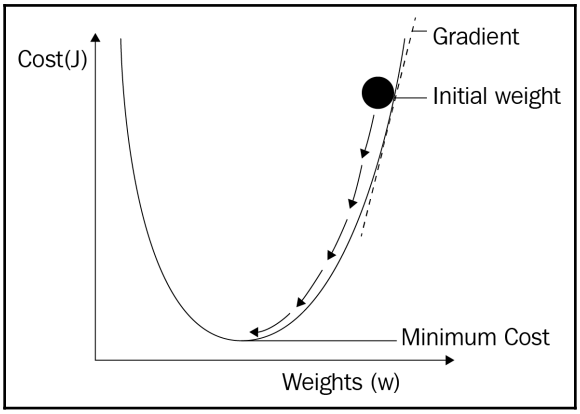

1
0
-1

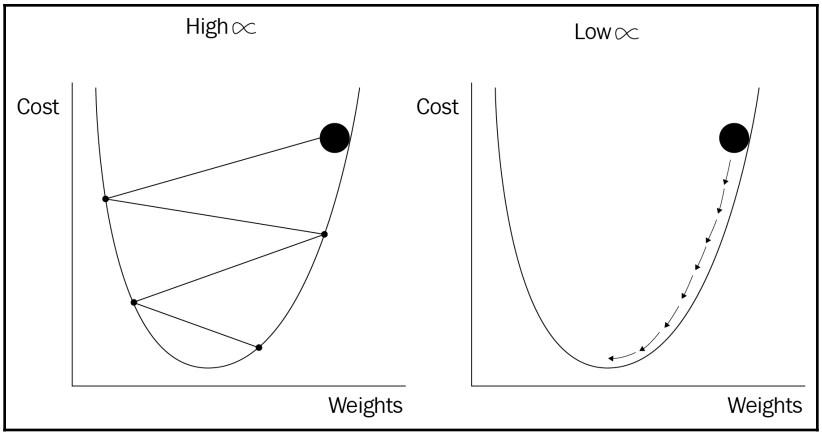

3
2
1
0

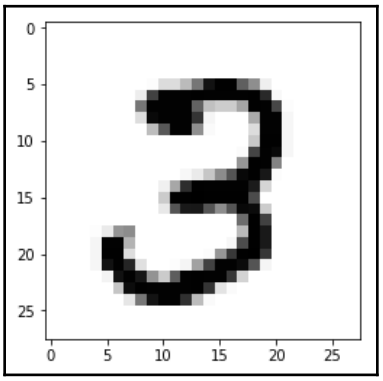

Multiply by weights (w_{xh})
& add bias(b_h)
x_1
x_2
Input
layer
Σ
f(z)
Hidden
layer
Multiply by weights (w_{hy})
& add bias(b_y)
Output
layer
$\hat{y}$

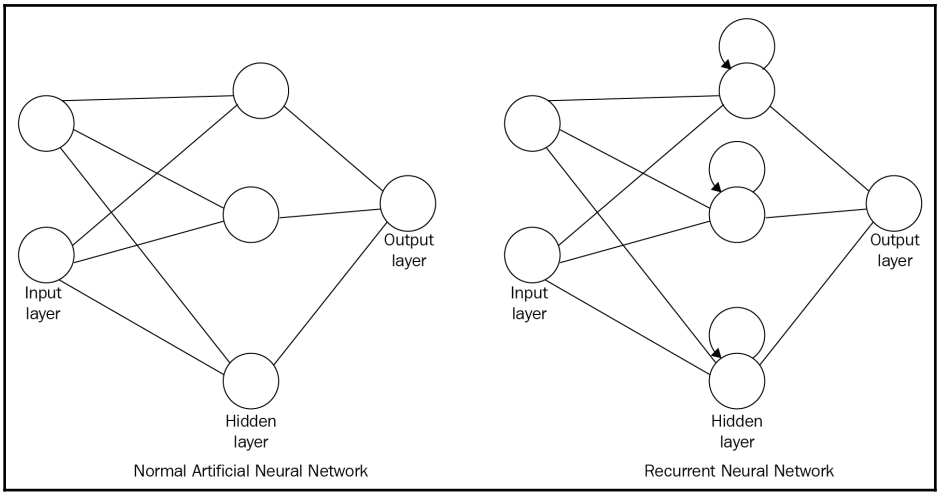

(you are here)
lowest point (local minimum)
global lowest point (global minimum)

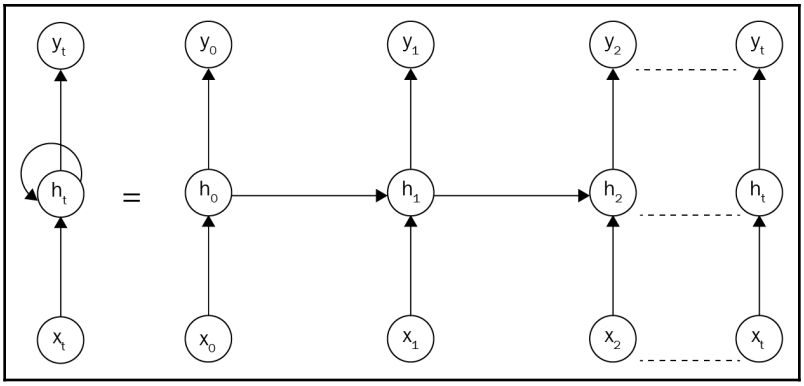

Cost(J)
Initial
Weight
Minimum Cost
Weights (w)

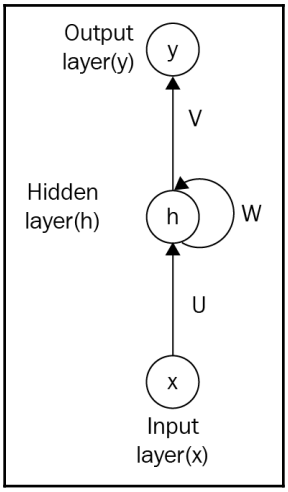

Cost(J)
Gradient
Initial weight
Minimum Cost
Weights (w)

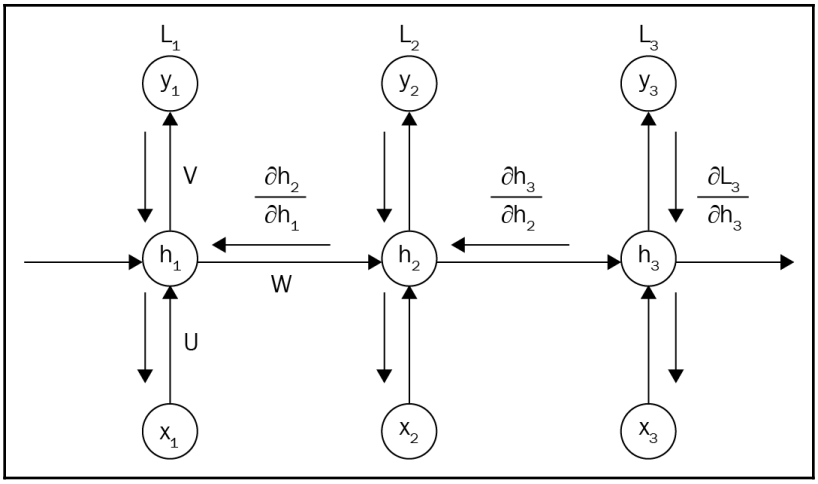

High ∝
Low ∝
Cost
Cost
Weights
Weights

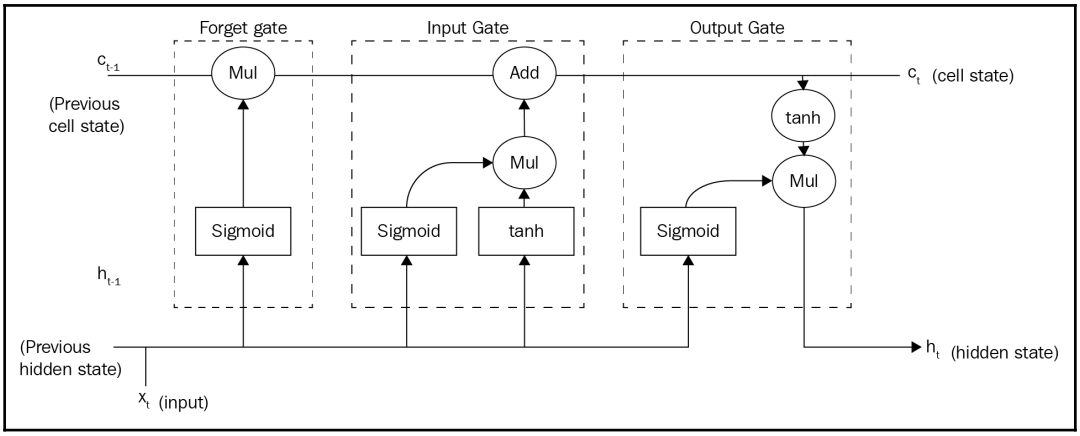

0
5
10
15
20
25
0
5
10
15
20
25


Input layer
Hidden layer
Output layer
Normal Artificial Neural Network
Input layer
Hidden layer
Output layer
Recurrent Neural Network

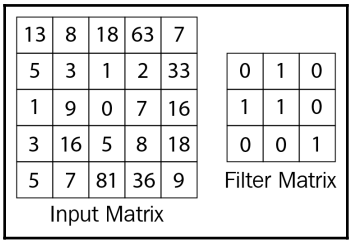

y_t
h_t
x_t
=
y_0
h_0
x_0
y_1
h_1
x_1
y_2
h_2
x_2
y_t
h_t
x_t

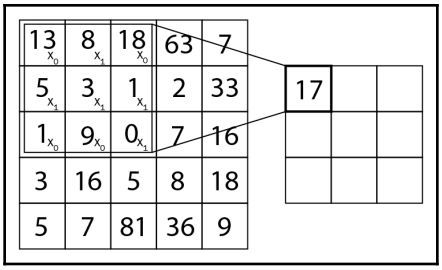

Output
layer(y)
y
V
Hidden
layer(h)
h
W
U
x
Input
layer(x)

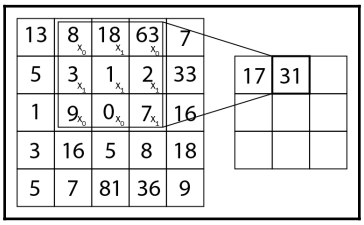

L1
L2
L3
y1
y2
y3
V
∂h2/∂h1
∂h3/∂h2
∂L3/∂h3
h1
h2
h3
W
U
x1
x2
x3

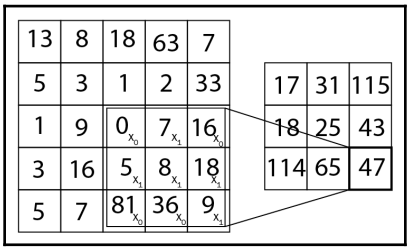


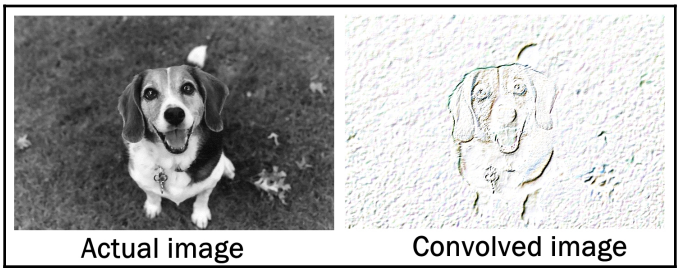
13	8	18	63	7
5	3	1	2	33
1	9	0	7	16
3	16	5	8	18
5	7	81	36	9

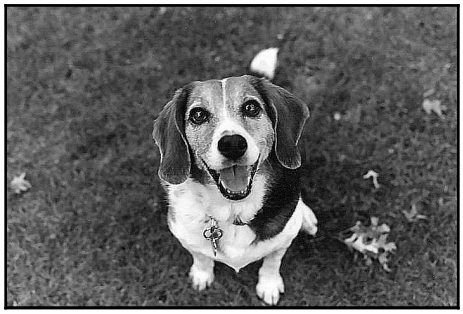

13 8 18 63 7
5 3 1 2 33
1 9 0 7 16
3 16 5 8 18
5 7 81 36 9
Input Matrix
0 1 0
1 1 0
0 0 1
Filter Matrix

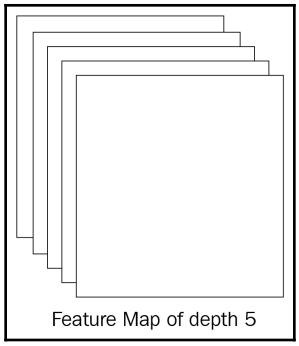

17

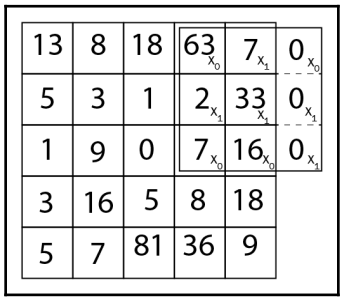

17 31

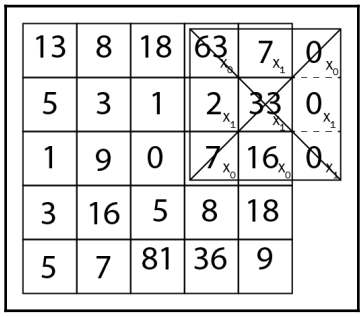

17 31 115
18 25 43
114 65 47

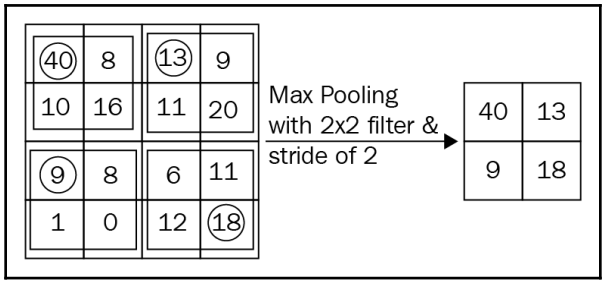

Actual image
Convolved image

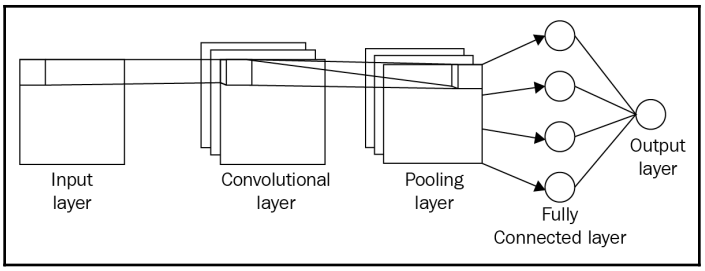


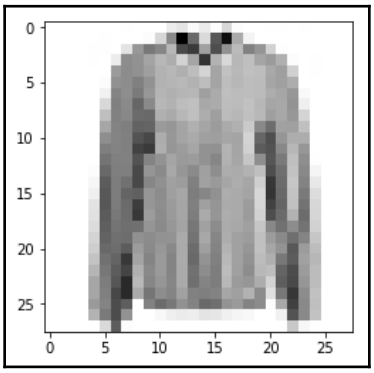

Feature Map of depth 5

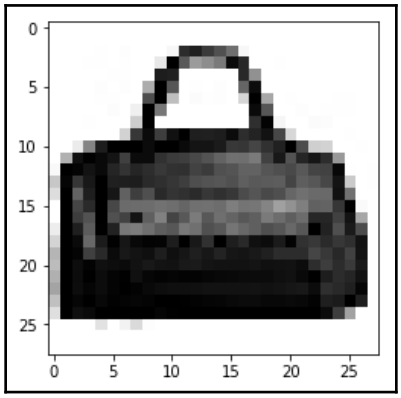

13 8 18 63 7 0
5 3 1 2 33 0
1 9 0 7 16 0
3 16 5 8 18
5 7 81 36 9


13 8 18 63 7 0
5 3 1 2 33 0
1 9 0 7 16 0
3 16 5 8 18
5 7 81 36 9

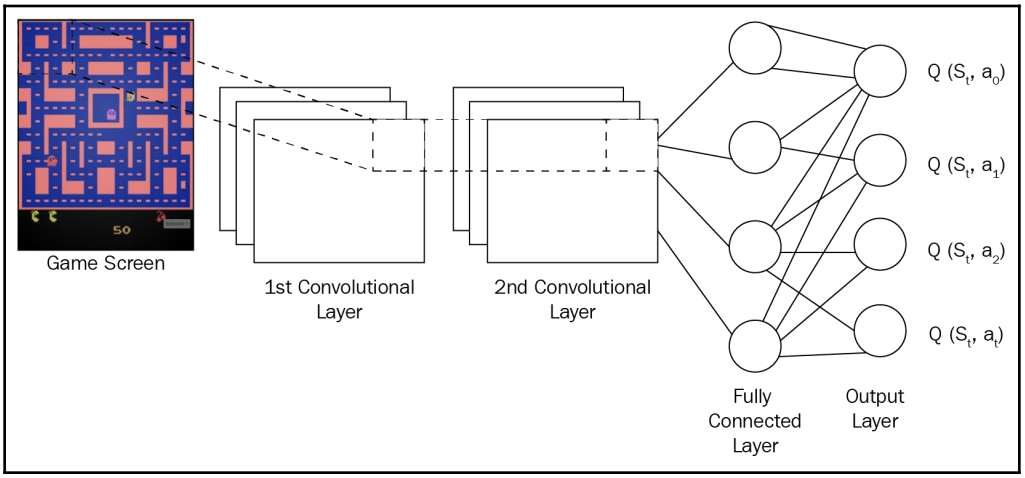

40
8
13
9
10
16
11
20
9
8
6
11
1
0
12
18
Max Pooling
with 2x2 filter &
stride of 2
40
13
9
18

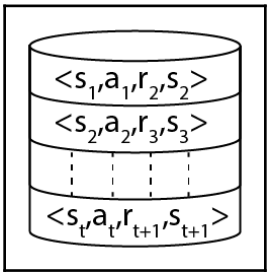

Input
layer
Convolutional
layer
Pooling
layer
Fully
Connected layer
Output
layer

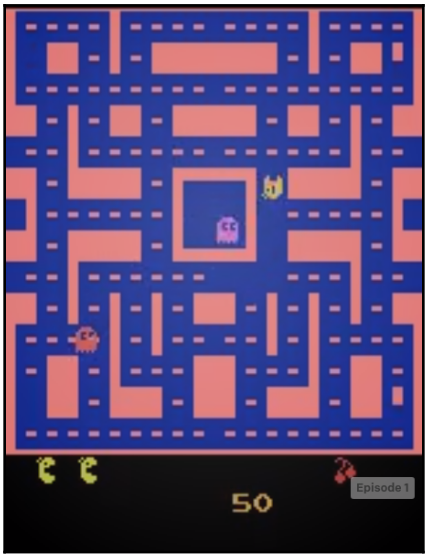

0
5
10
15
20
25
0
5
10
15
20
25

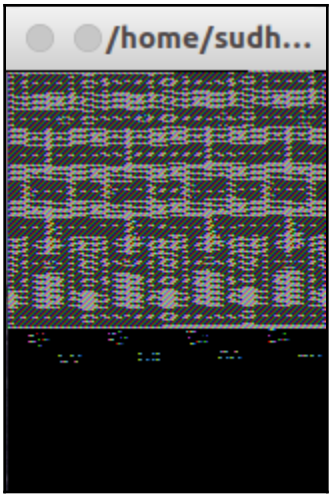

0
5
10
15
20
25
0
5
10
15
20
25

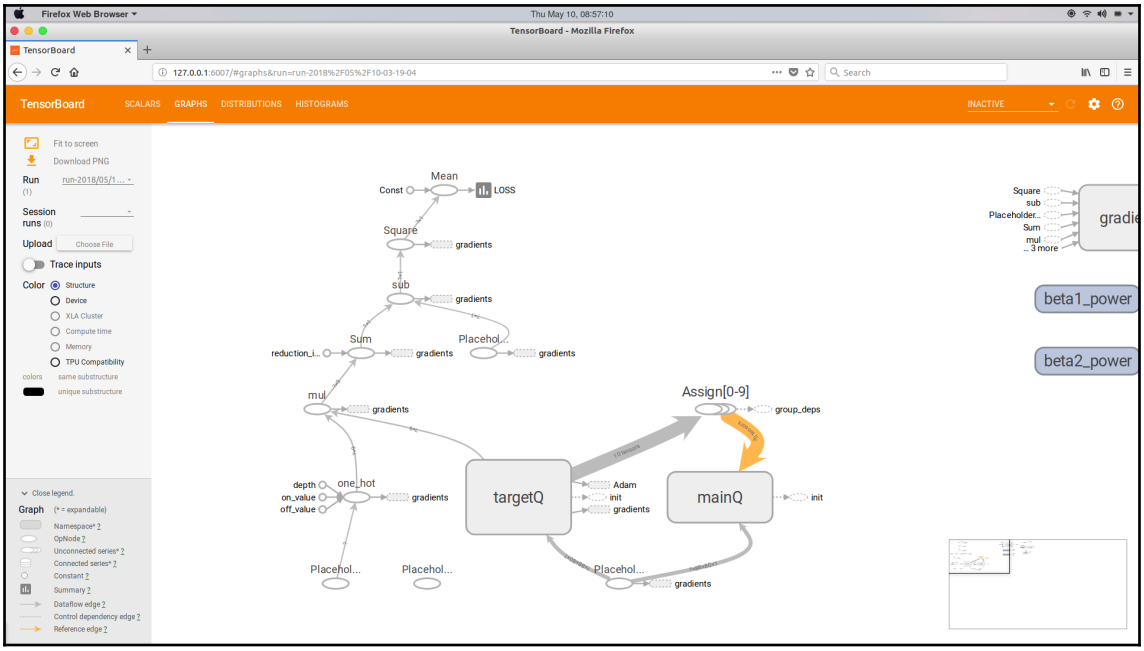
Chapter 8: Atari Games with Deep Q Network

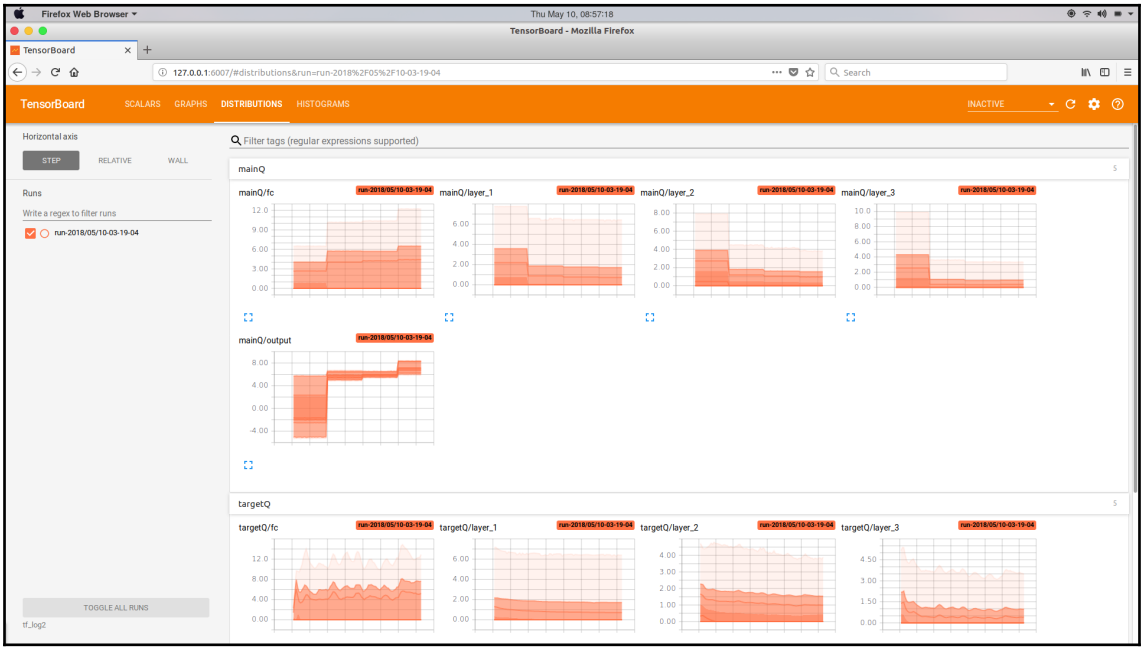


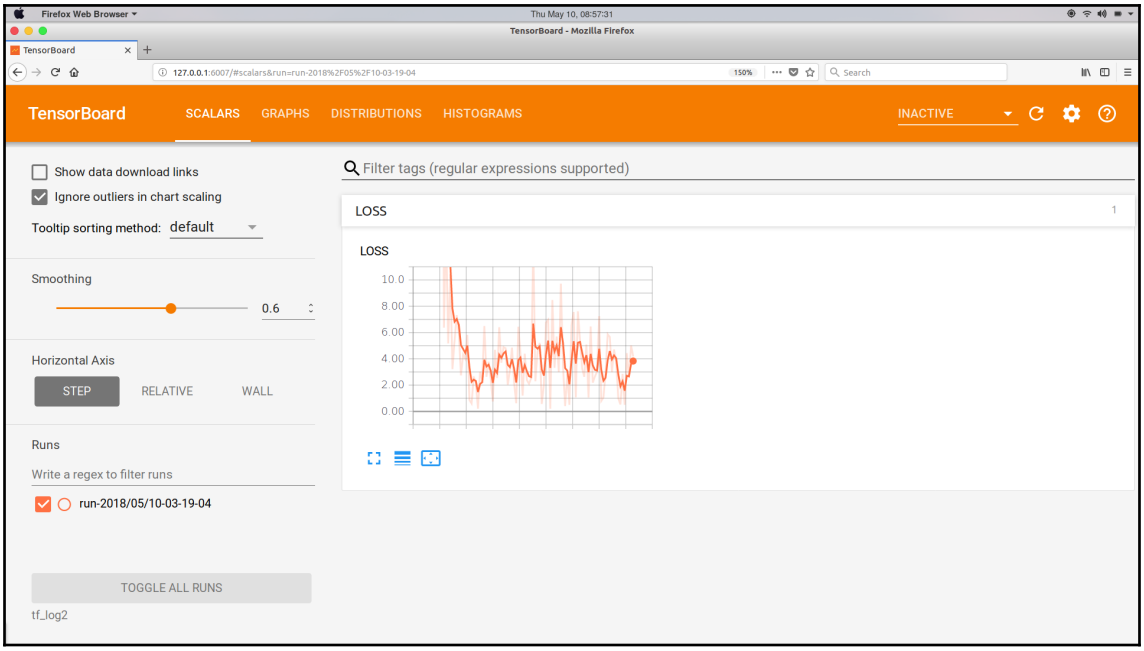


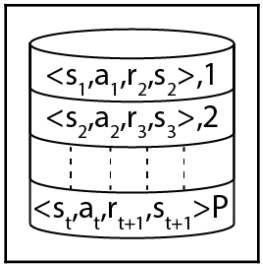

50
Episode 1

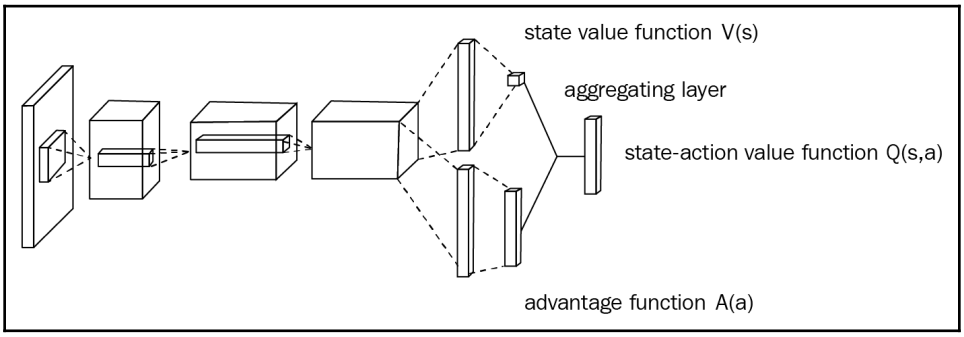

/home/sudh...


Firefox Web Browser
Thu May 10, 08:57:10
TensorBoard - Mozilla Firefox
TensorBoard
127.0.0.1:6007/#graphs&run=run-2018%2F05%2F10-03-19-04
TensorBoard
SCALARS
GRAPHS
DISTRIBUTIONS
HISTOGRAMS
INACTIVE
Fit to screen
Download PNG
Run
run-2018/05/1...
Session runs
Upload
Choose File
Trace inputs
Color
Structure
Device
XLA Cluster
Compute time
Memory
TPU Compatibility
same substructure
unique substructure
Close legend.
Graph (* = expandable)
Namespace* ?
OpNode ?
Unconnected series* ?
Connected series* ?
Constant ?
Summary ?
Dataflow edge ?
Control dependency edge ?
Reference edge ?
Mean
Const
LOSS
Square
gradients
sub
Sum
reduction_i...
Placehol...
mul
Assign[0-9]
group_deps
depth
one_hot
on_value
off_value
targetQ
Adam
init
mainQ
Square
sub
Placeholder...
Sum
mul
... 3 more
beta1_power
beta2_power

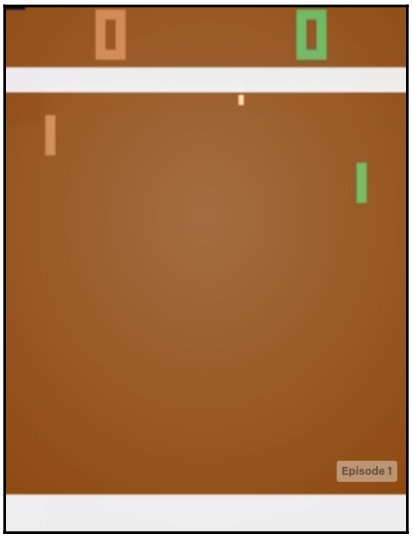

Firefox Web Browser
Thu May 10, 08:57:18
TensorBoard - Mozilla Firefox
TensorBoard
127.0.0.1:6007/#distributions&run=run-2018%2F05%2F10-03-19-04
TensorBoard
SCALARS
GRAPHS
DISTRIBUTIONS
HISTOGRAMS
INACTIVE
Horizontal axis
STEP
RELATIVE
WALL
Runs
Write a regex to filter runs
run-2018/05/10-03-19-04
TOGGLE ALL RUNS
tf_log2
Filter tags (regular expressions supported)
mainQ
mainQ/fc
mainQ/layer_1
mainQ/layer_2
mainQ/layer_3
mainQ/output
targetQ
targetQ/fc
targetQ/layer_1
targetQ/layer_2
targetQ/layer_3

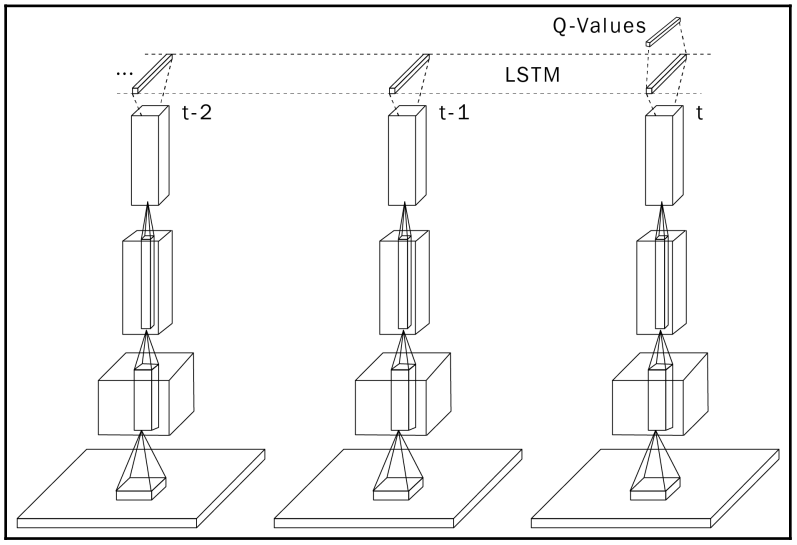

Firefox Web Browser
Thu May 10, 08:57:31
TensorBoard - Mozilla Firefox
TensorBoard
127.0.0.1:6007/#scalars&run=run-2018%2F05%2F10-03-19-04
150%
Search
TensorBoard
SCALARS
GRAPHS
DISTRIBUTIONS
HISTOGRAMS
INACTIVE
Show data download links
Ignore outliers in chart scaling
Tooltip sorting method: default
Smoothing
0.6
Horizontal Axis
STEP
RELATIVE
WALL
Runs
Write a regex to filter runs
run-2018/05/10-03-19-04
TOGGLE ALL RUNS
tf_log2
Filter tags (regular expressions supported)
LOSS
1
LOSS
10.0
8.00
6.00
4.00
2.00
0.00


<s1,a1,r2,s2>,1
<s2,a2,r3,s3>,2
<st,at,rt+1,st+1>P

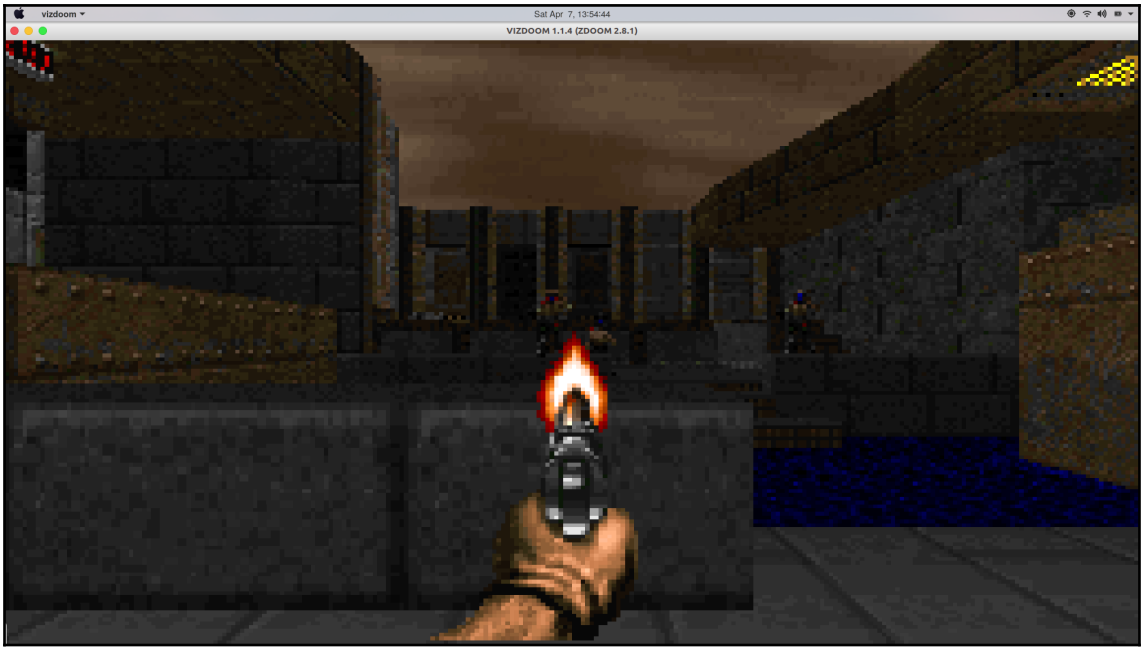

state value function V(s)
aggregating layer
state-action value function Q(s,a)
advantage function A(a)

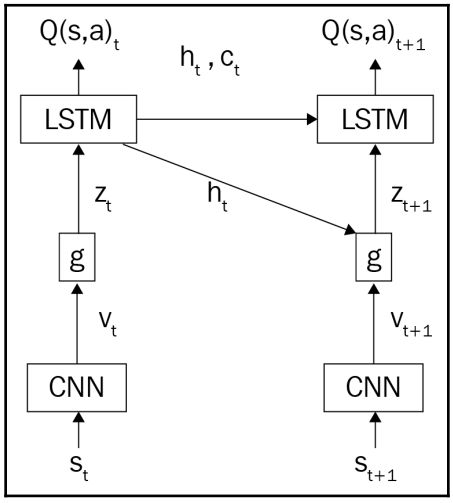
Chapter 9: Playing Doom with a Deep Recurrent Q Network



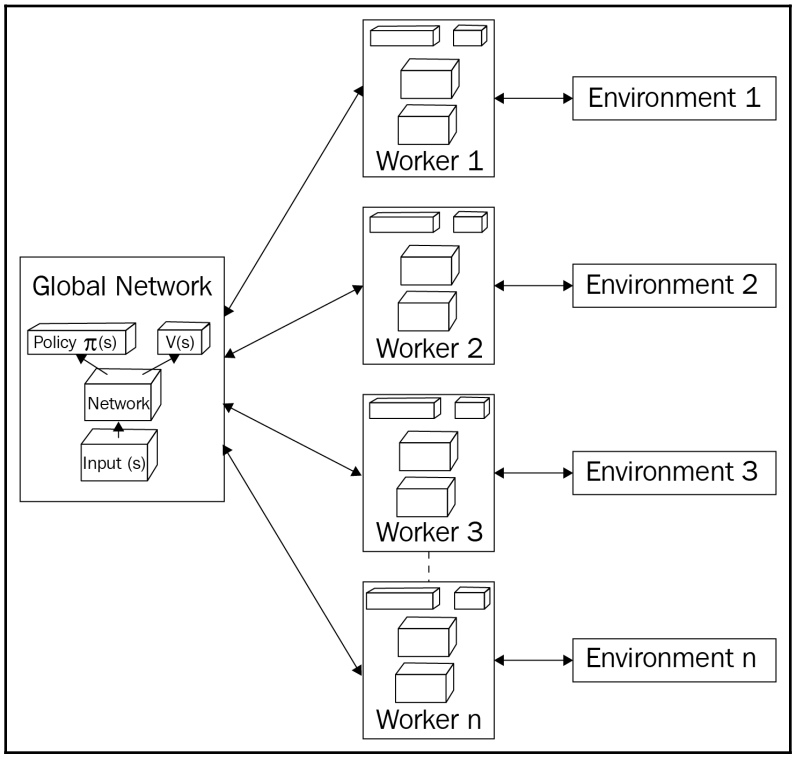


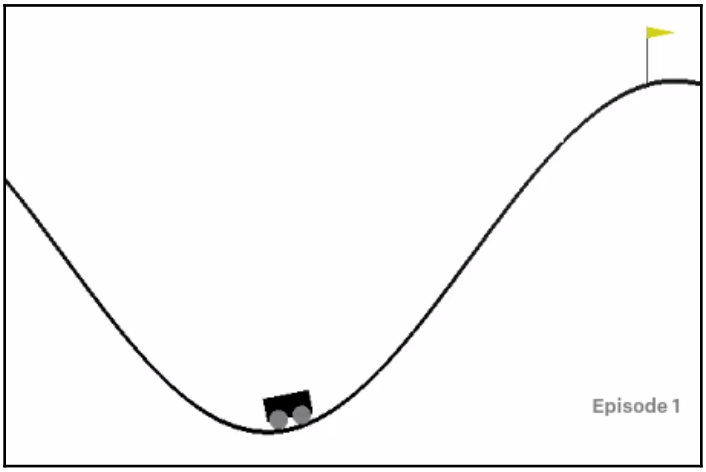

VIZDOOM 1.1.4 (ZDOOM 2.8.1)
39
100%
AMMO
HEALTH
ARMS
0%
ARMOR
39 / 200
0 / 50
0 / 50
0 / 300

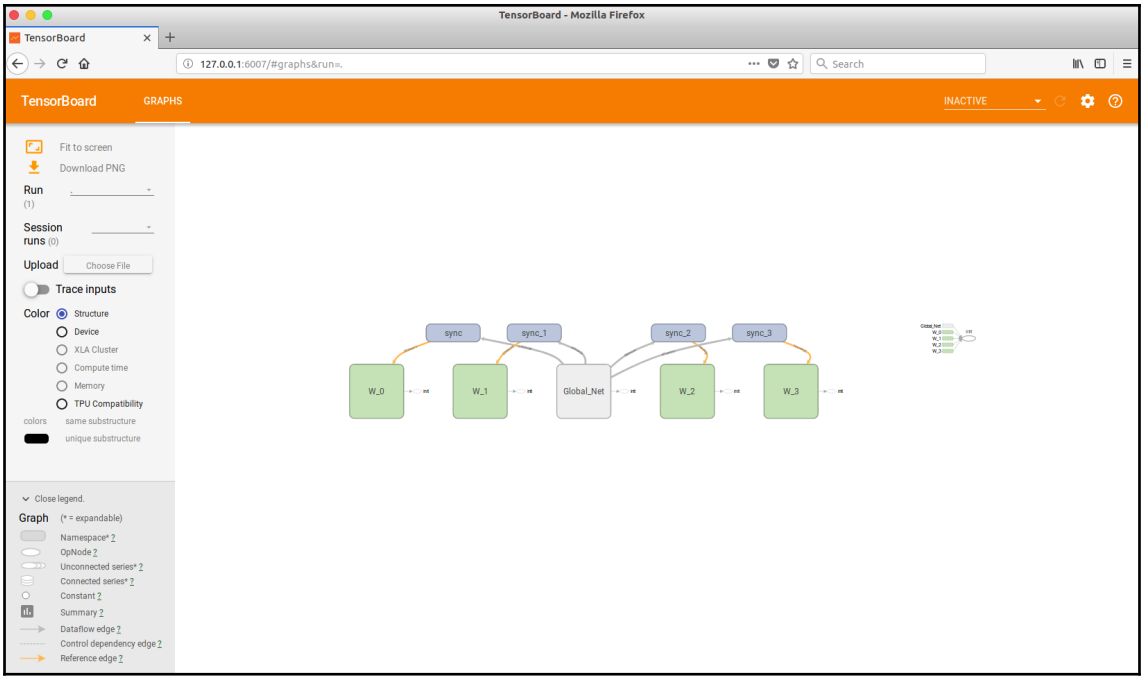

vizdoom
Sat Apr 7, 13:54:44
VIZDOOM 1.1.4 (ZDOOM 2.8.1)

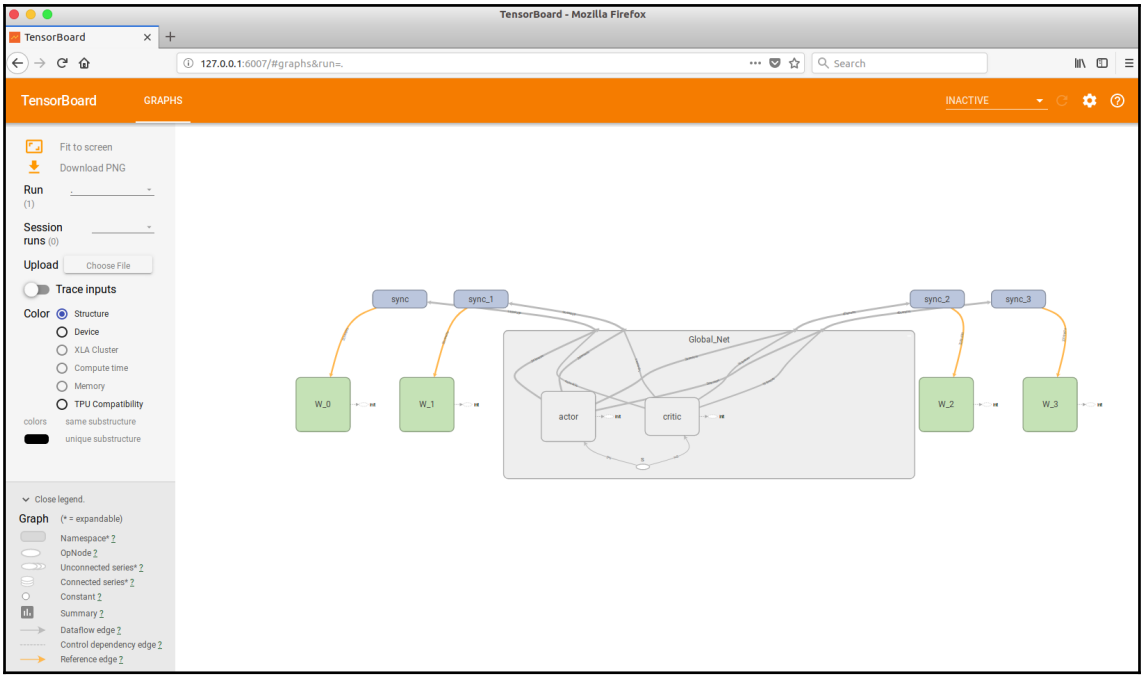

Q(s,a)t
ht , ct
Q(s,a)t+1
LSTM
LSTM
zt
ht
zt+1
g
g
vt
vt+1
CNN
CNN
st
st+1

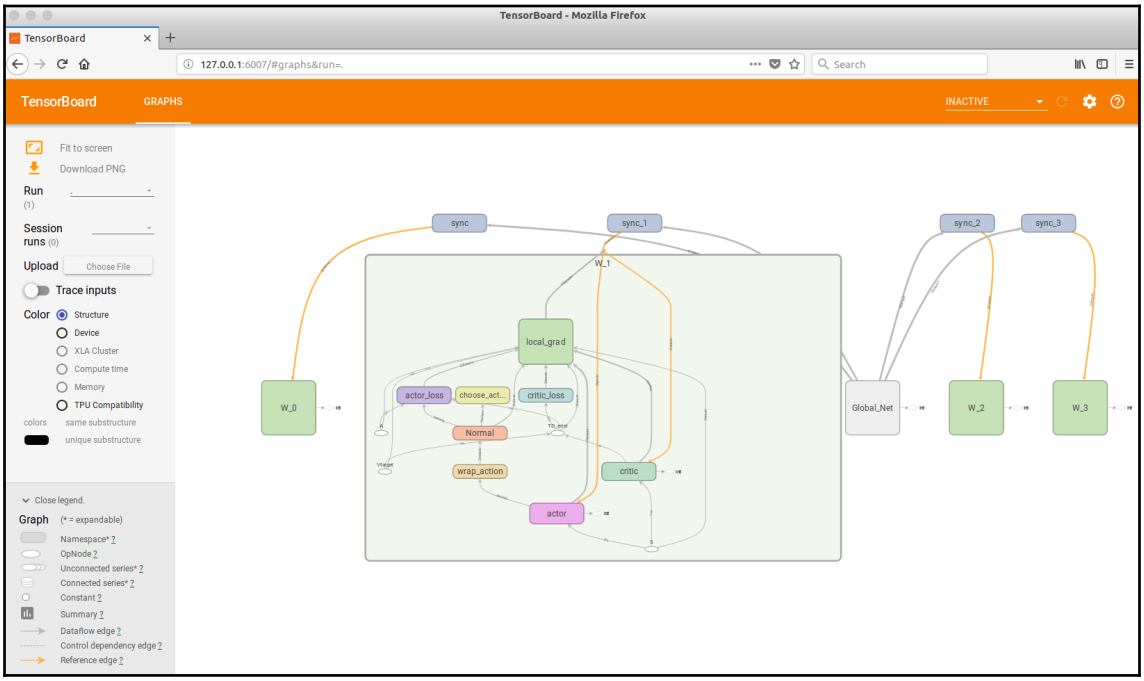
Chapter 10: The Asynchronous Advantage Actor Critic Network



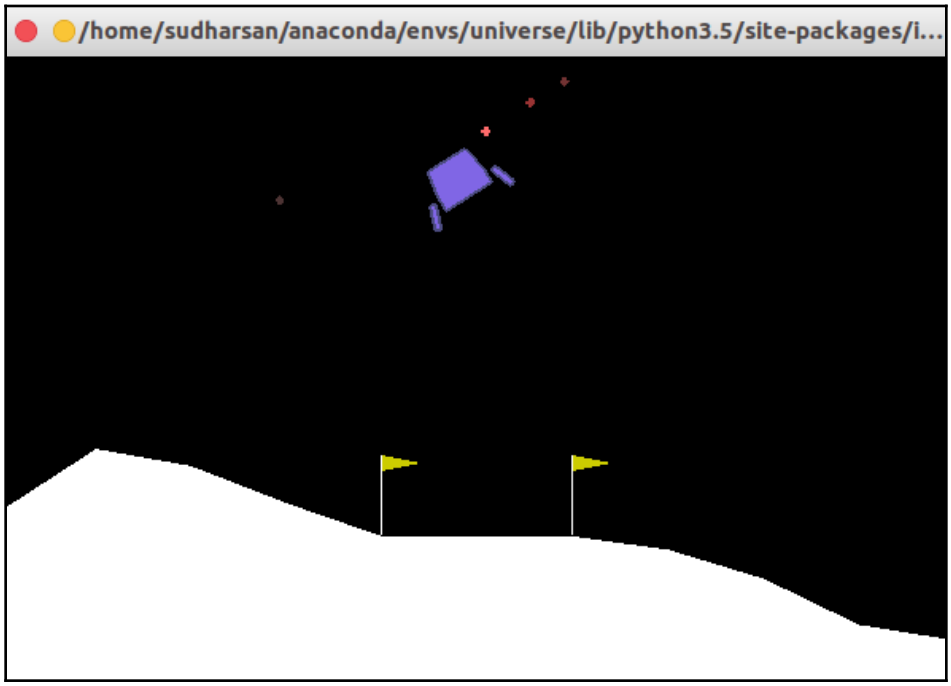

Episode 1

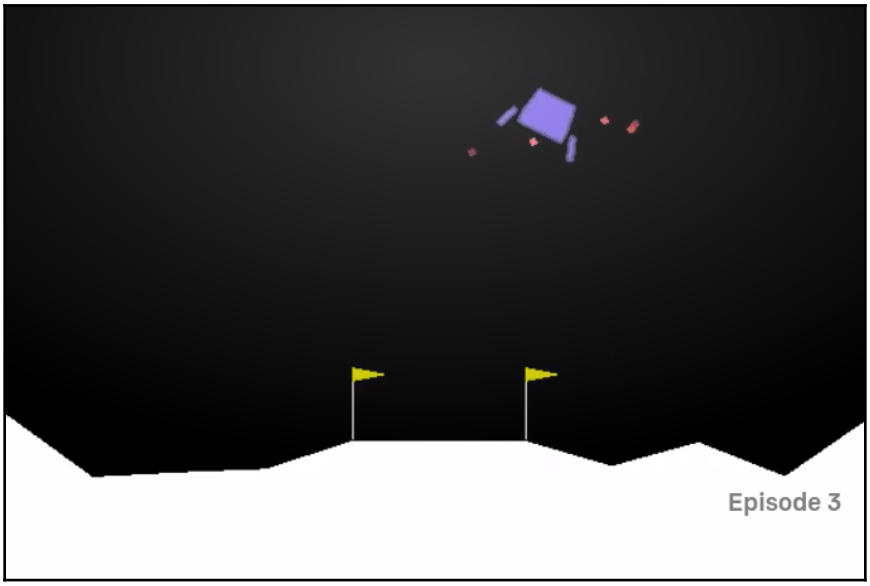

TensorBoard - Mozilla Firefox
TensorBoard
127.0.0.1:6007/#graphs&run=.
Search
TensorBoard
GRAPHS
INACTIVE
Fit to screen
Download PNG
Run
(1)
Session runs (0)
Upload
Choose File
Trace inputs
Color
Structure
Device
XLA Cluster
Compute time
Memory
TPU Compatibility
colors
same substructure
unique substructure
Close legend.
Graph
(* = expandable)
Namespace* ?
OpNode ?
Unconnected series* ?
Connected series* ?
Constant ?
Summary ?
Dataflow edge ?
Control dependency edge ?
Reference edge ?
sync
sync_1
sync_2
sync_3
W_0
W_1
Global_Net
W_2
W_3

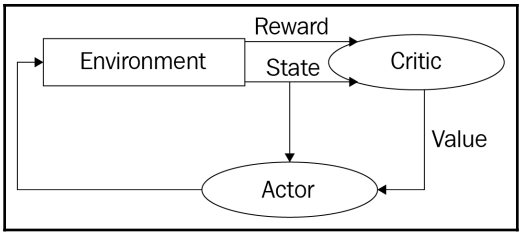

TensorBoard - Mozilla Firefox
TensorBoard
127.0.0.1:6007/#graphs&run=.
Search
TensorBoard
GRAPHS
INACTIVE
Fit to screen
Download PNG
Run (1)
Session runs (0)
Upload
Choose File
Trace inputs
Color
Structure
Device
XLA Cluster
Compute time
Memory
TPU Compatibility
colors
same substructure
unique substructure
Close legend.
Graph (* = expandable)
Namespace* ?
OpNode ?
Unconnected series* ?
Connected series* ?
Constant ?
Summary ?
Dataflow edge ?
Control dependency edge ?
Reference edge ?
sync
sync_1
sync_2
sync_3
Global_Net
W_0
W_1
W_2
W_3
actor
critic

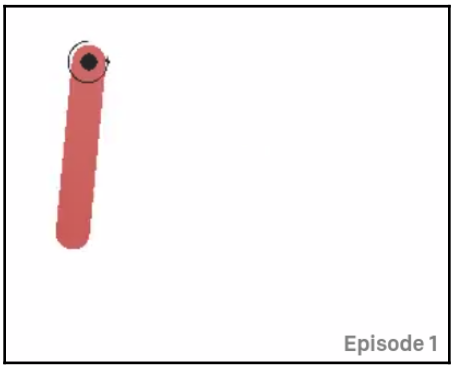

TensorBoard - Mozilla Firefox
TensorBoard
127.0.0.1:6007/#graphs&run=.
Search
TensorBoard
GRAPHS
INACTIVE
Fit to screen
Download PNG
Run (1)
Session runs (0)
Upload
Choose File
Trace inputs
Color
Structure
Device
XLA Cluster
Compute time
Memory
TPU Compatibility
colors
same substructure
unique substructure
Close legend.
Graph (* = expandable)
Namespace* ?
OpNode ?
Unconnected series* ?
Connected series* ?
Constant ?
Summary ?
Dataflow edge ?
Control dependency edge ?
Reference edge ?
sync
sync_1
sync_2
sync_3
W_1
local_grad
actor_loss
choose_act...
critic_loss
Normal
wrap_action
critic
actor
W_0
Global_Net
W_2
W_3

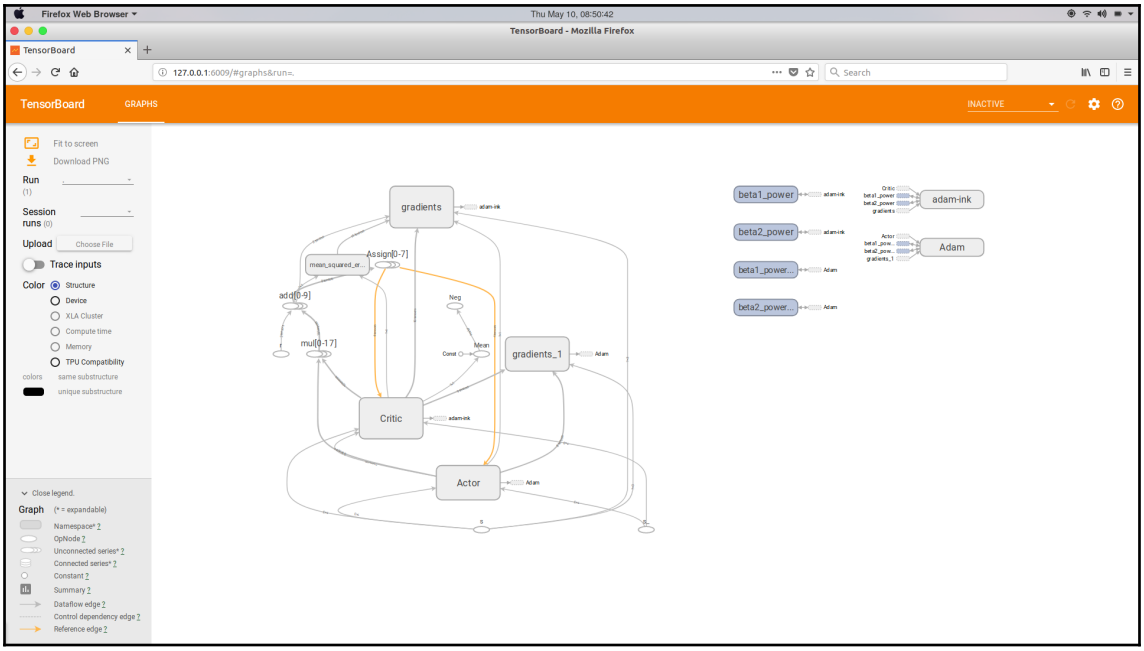
Chapter 11: Policy Gradients and Optimization

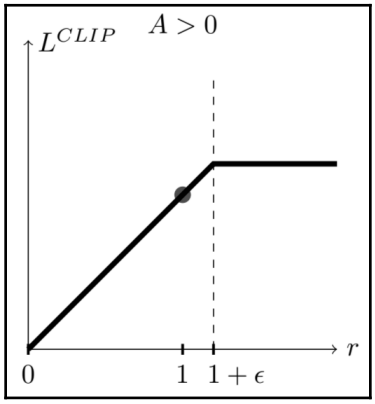


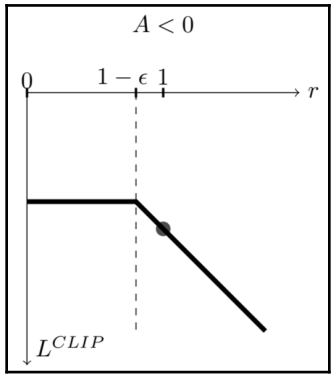

Episode 3


Reward
Environment
State
Critic
Value
Actor

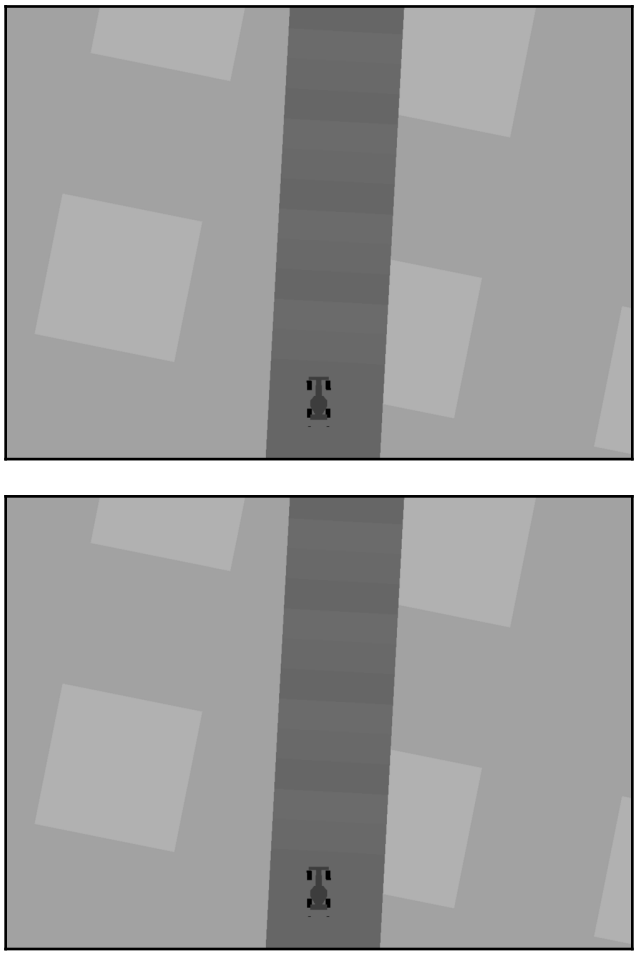

Episode 1


Firefox Web Browser
Thu May 10, 08:50:42
TensorBoard - Mozilla Firefox
TensorBoard
127.0.0.1:6009/#graphs&run=.
Search
TensorBoard
GRAPHS
INACTIVE
Fit to screen
Download PNG
Run
(1)
Session runs (0)
Upload
Choose File
Trace inputs
Color
Structure
Device
XLA Cluster
Compute time
Memory
TPU Compatibility
colors
same substructure
unique substructure
Close legend.
Graph (* = expandable)
Namespace* ?
OpNode ?
Unconnected series* ?
Connected series* ?
Constant ?
Summary ?
Dataflow edge ?
Control dependency edge ?
Reference edge ?
gradients
adam-ink
Assign[0-7]
mean_squared_er...
add[0-9]
mul[0-17]
Neg
Const
Mean
gradients_1
Adam
Critic
adam-ink
Actor
Adam
beta1_power
beta2_power
beta1_power...
beta2_power...
adam-ink
Adam

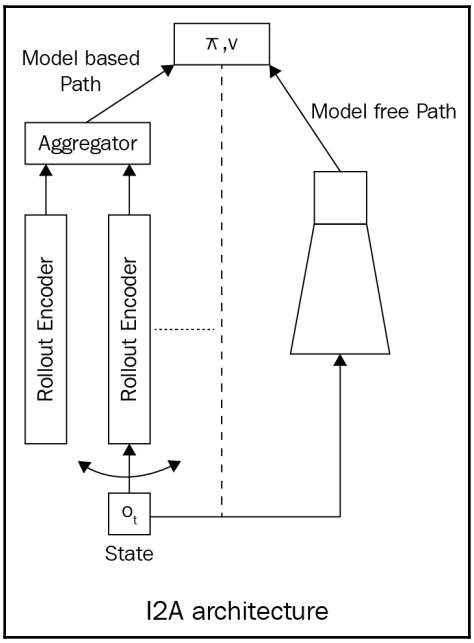

L^{CLIP}
$A > 0$
r
0
1
$1+\epsilon$


$A < 0$
0
$1-\epsilon$
1
r
L^{CLIP}

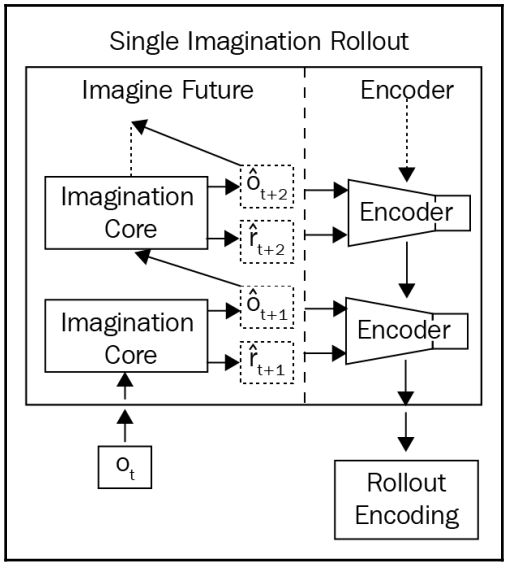
Chapter 12: Capstone Project – Car Racing Using DQN

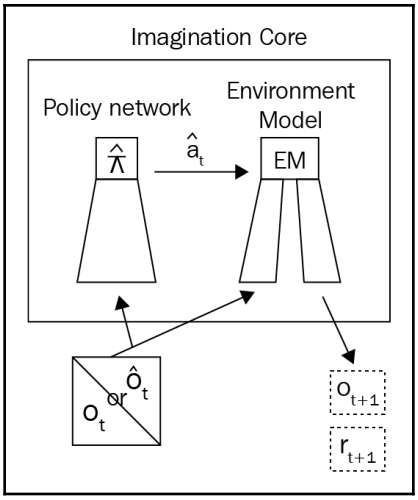


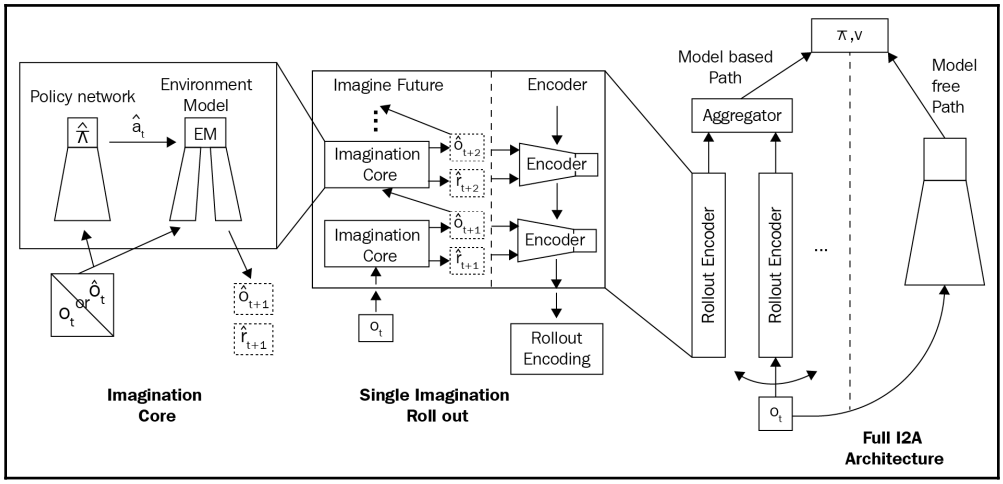
Chapter 13: Recent Advancements and Next Steps

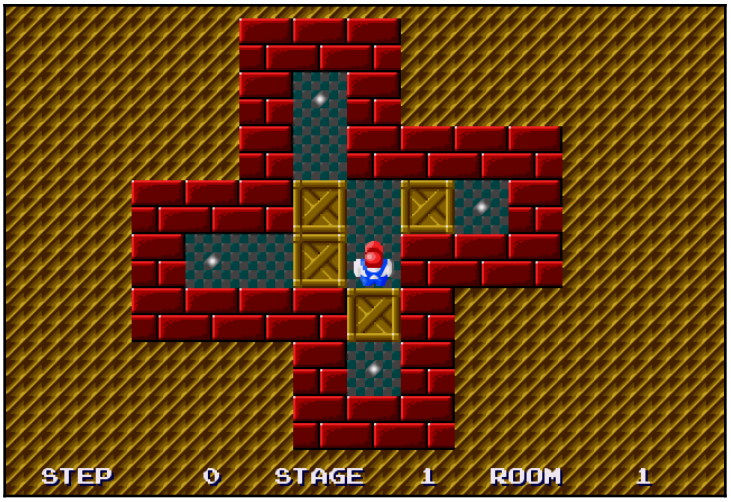


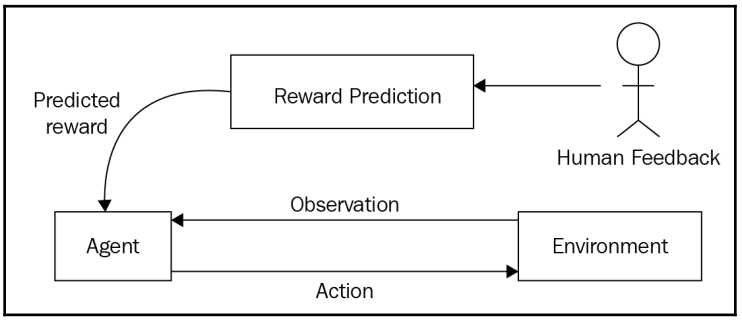
I2A architecture

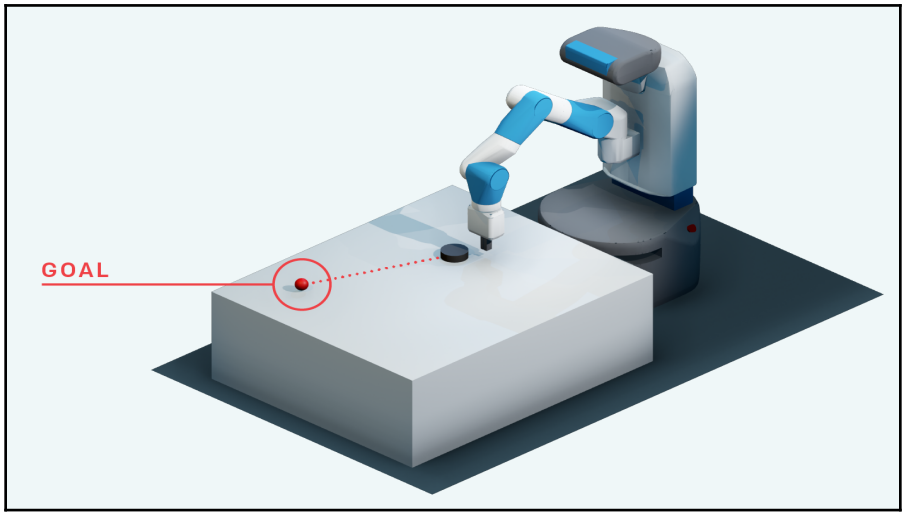

Single Imagination Rollout
Imagine Future
Encoder
Imagination Core
Imagination Core
Encoder
Encoder
Rollout Encoding

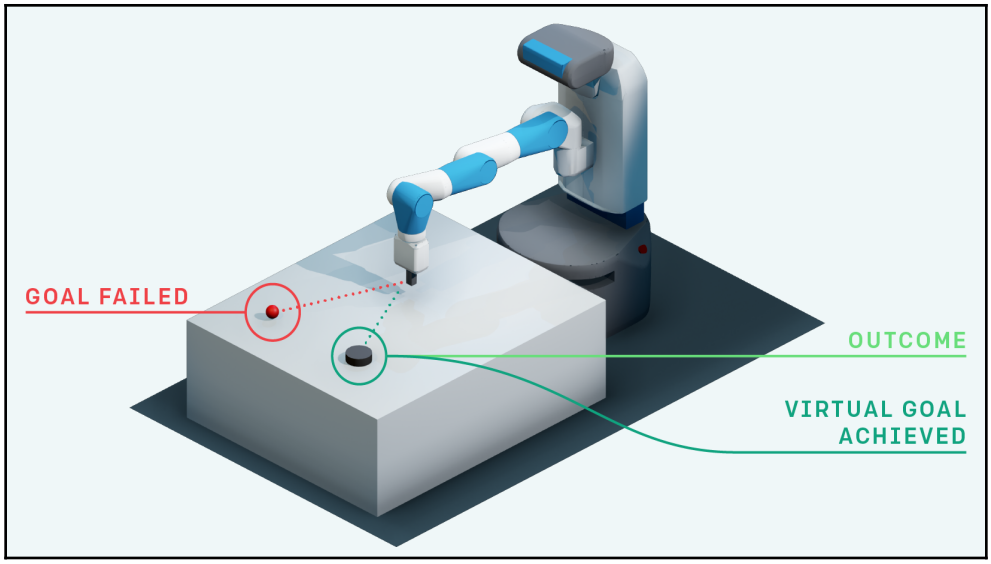

Imagination Core
Policy network
Environment Model
EM

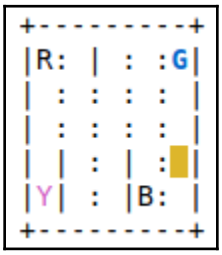

Policy network
Environment Model
EM
Imagination Core
or
Imagination Core
Imagine Future
Encoder
Encoder
Encoder
Rollout Encoding
Imagination Core
Single Imagination Roll out
π,v
Model based Path
Model free Path
Aggregator
Rollout Encoder
Rollout Encoder
...
Full I2A Architecture

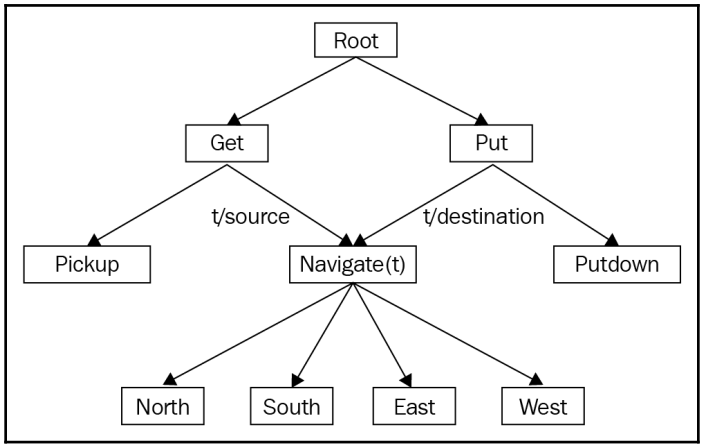

STEP 0 STAGE 1 ROOM 1

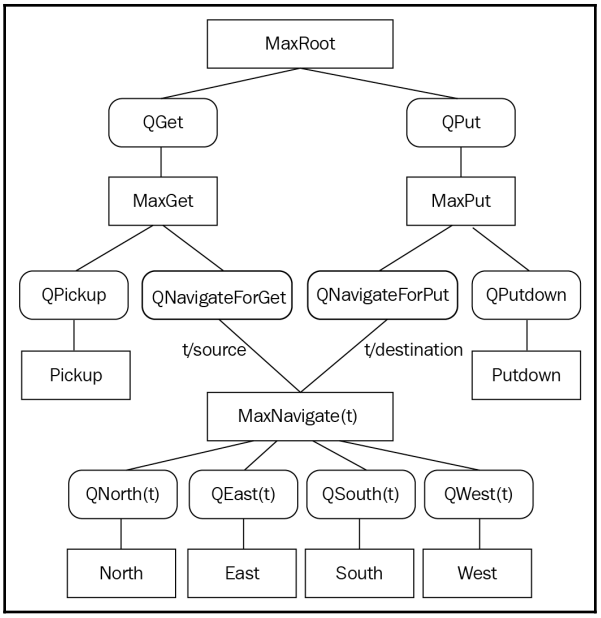

Predicted reward
Reward Prediction
Human Feedback
Observation
Agent
Environment
Action


GOAL


GOAL FAILED
OUTCOME
VIRTUAL GOAL
ACHIEVED


+---------+
|R: | : :G|
| : : : : |
| : : : : |
| | : | :|
|Y| : |B: |
+---------+


Root
Get
Put
t/source
t/destination
Pickup
Navigate(t)
Putdown
North
South
East
West


MaxRoot
QGet
QPut
MaxGet
MaxPut
QPickup
QNavigateForGet
QNavigateForPut
QPutdown
t/source
t/destination
Pickup
Putdown
MaxNavigate(t)
QNorth(t)
QEast(t)
QSouth(t)
QWest(t)
North
East
South
West