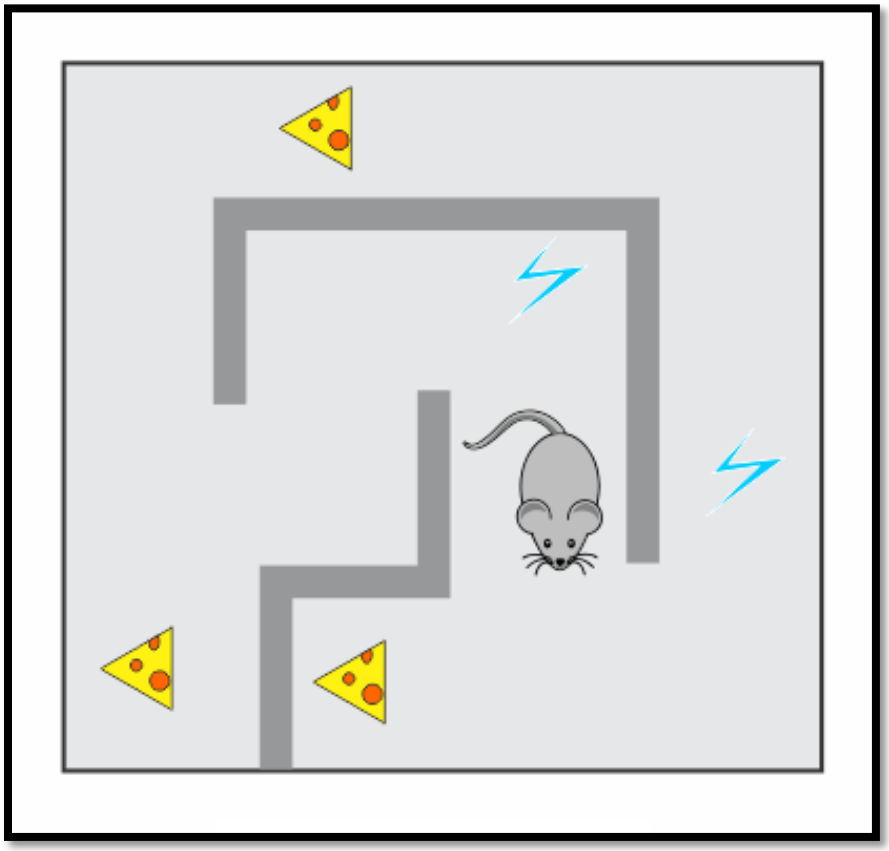
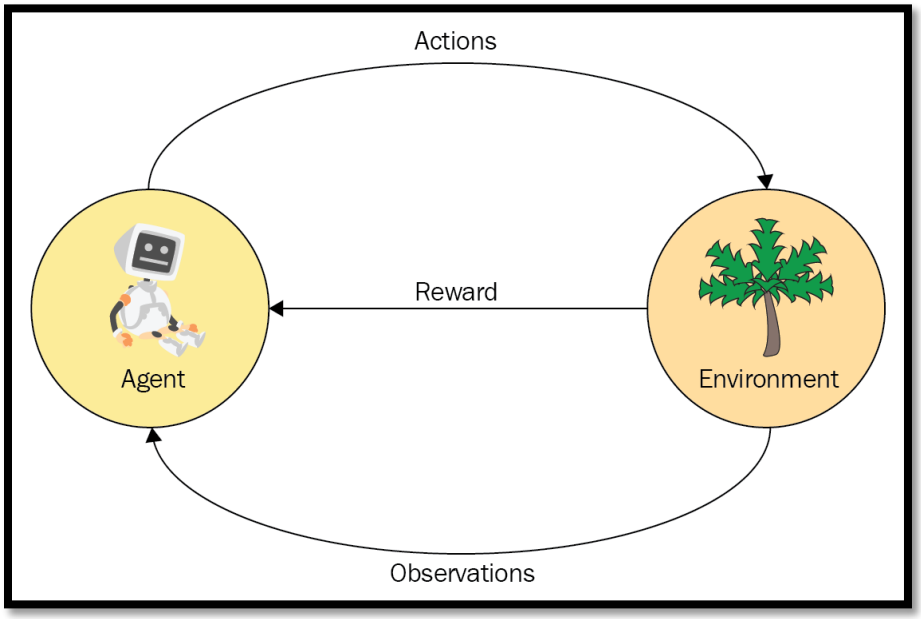
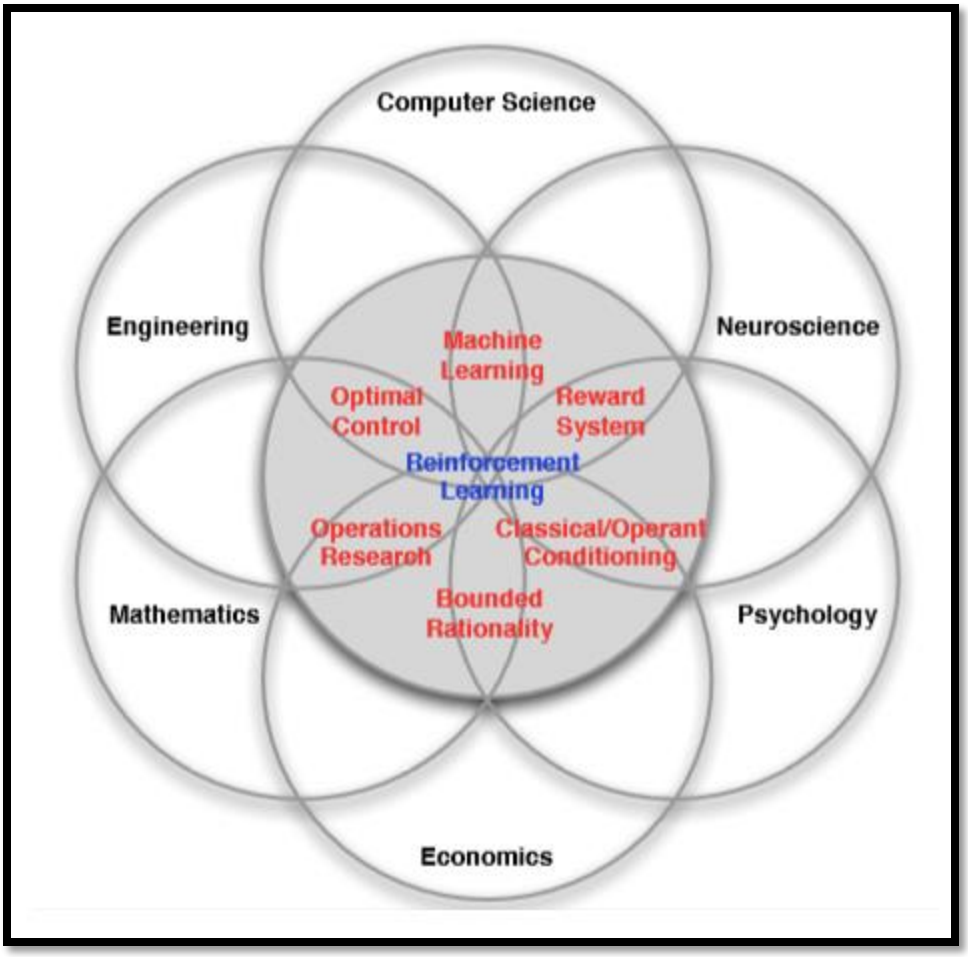
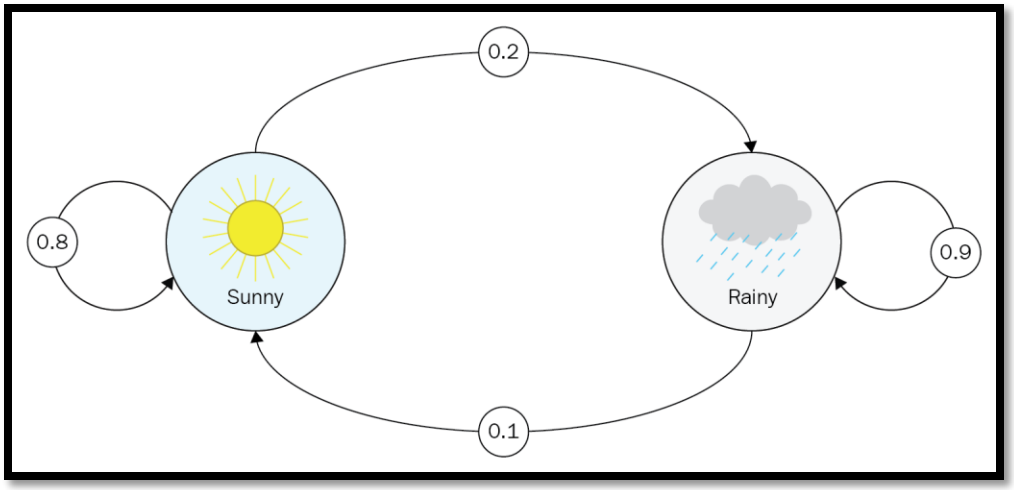
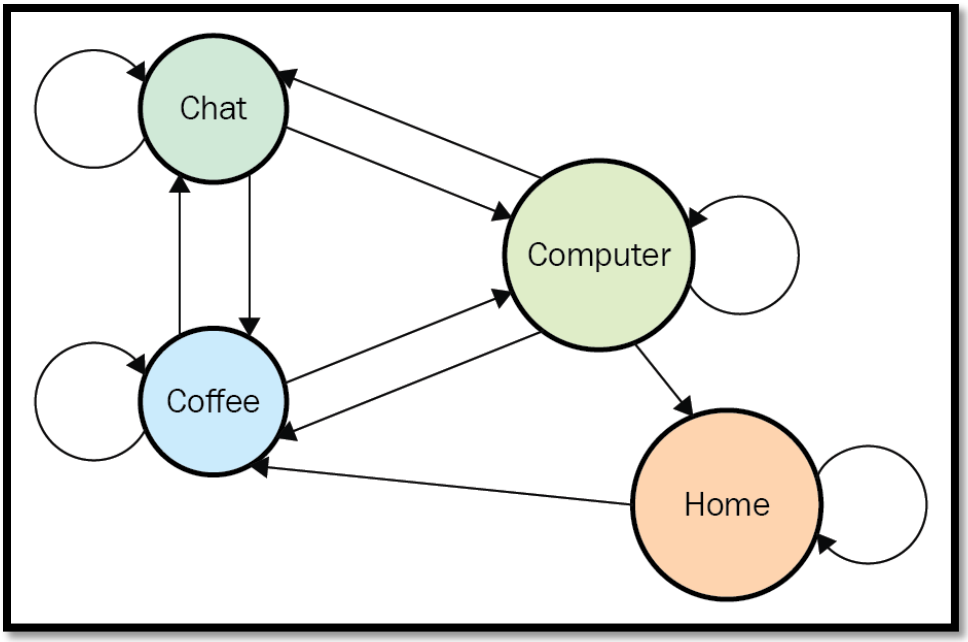
Chapter 1: What is Reinforcement Learning?

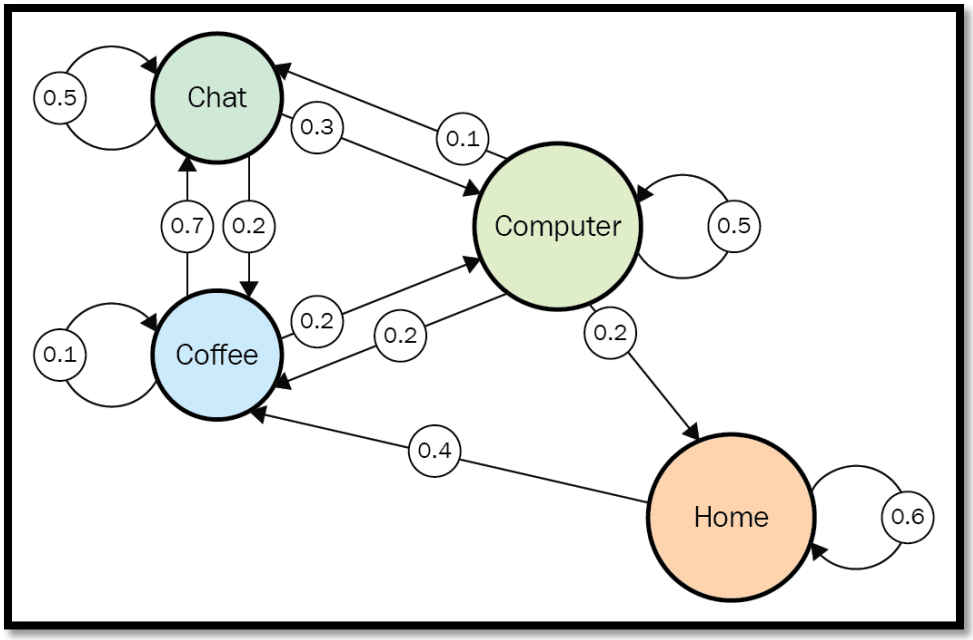


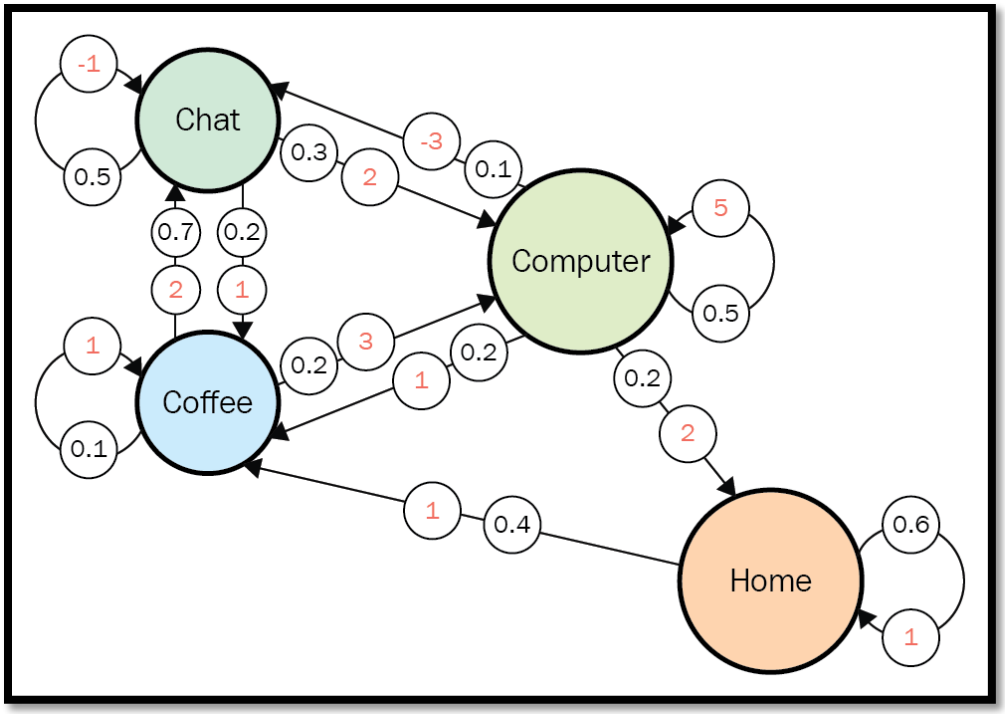


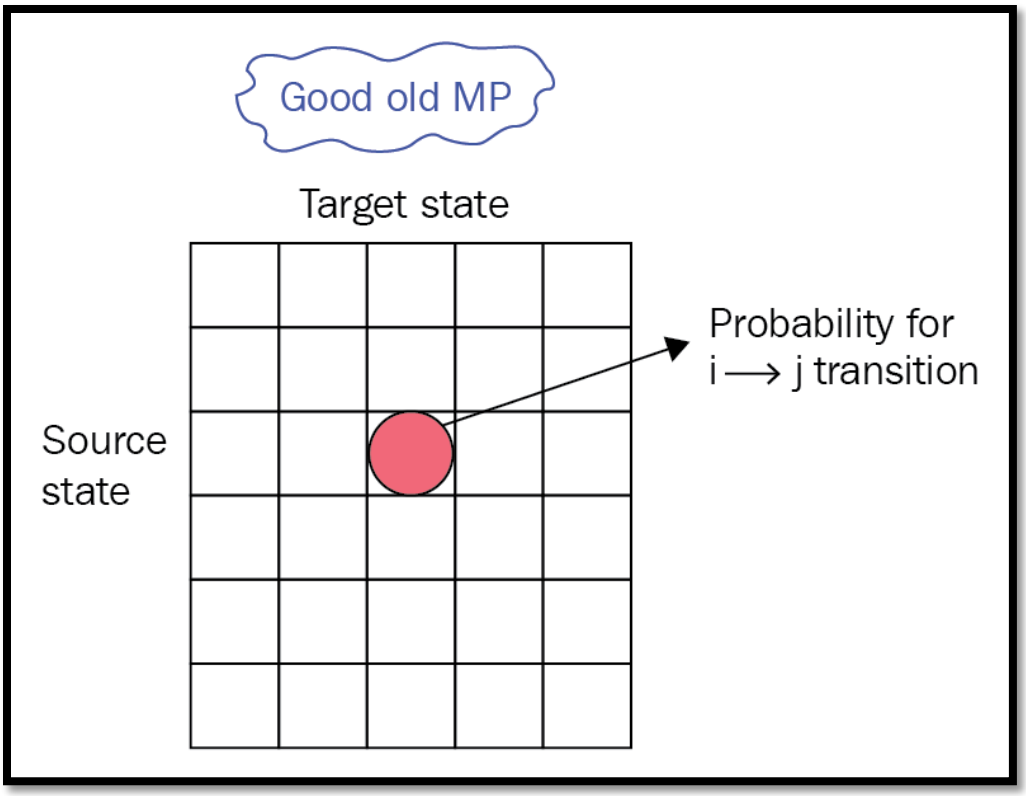

Computer Science
Engineering
Neuroscience
Machine Learning
Optimal Control
Reward System
Reinforcement Learning
Operations Research
Classical/Operant Conditioning
Bounded Rationality
Mathematics
Psychology
Economics

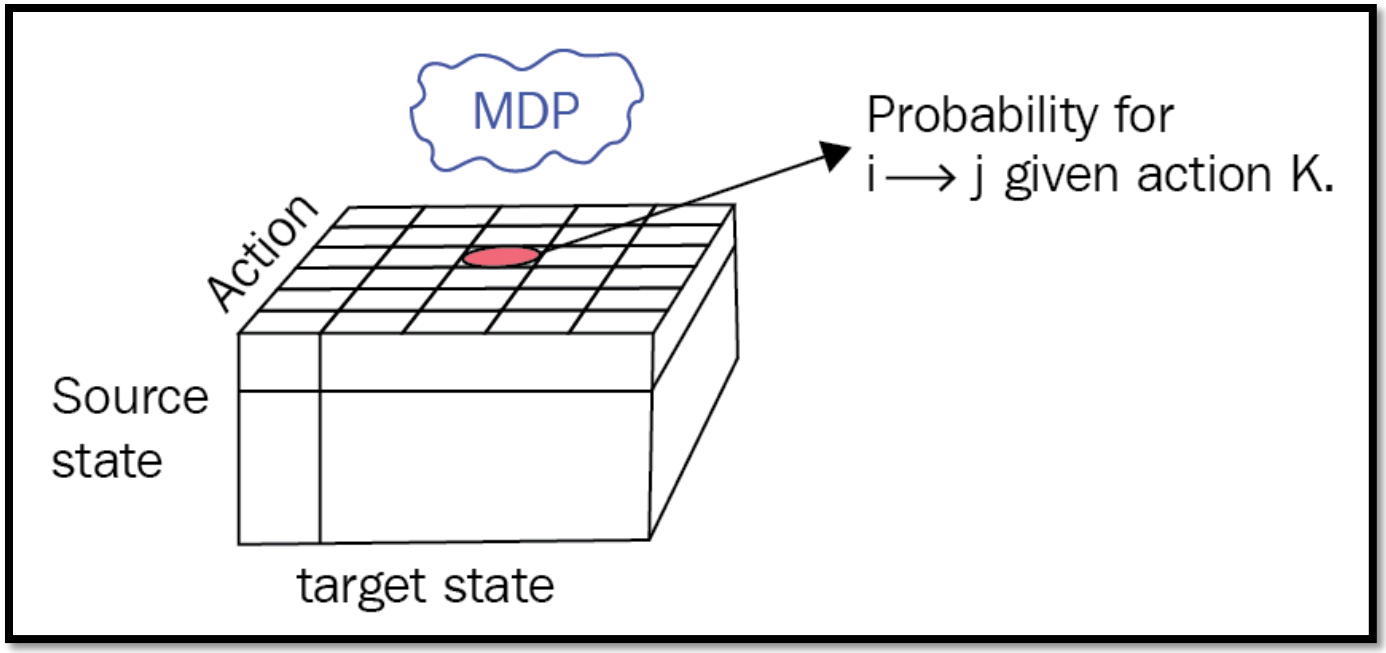

0.2
0.8
Sunny
Rainy
0.9
0.1

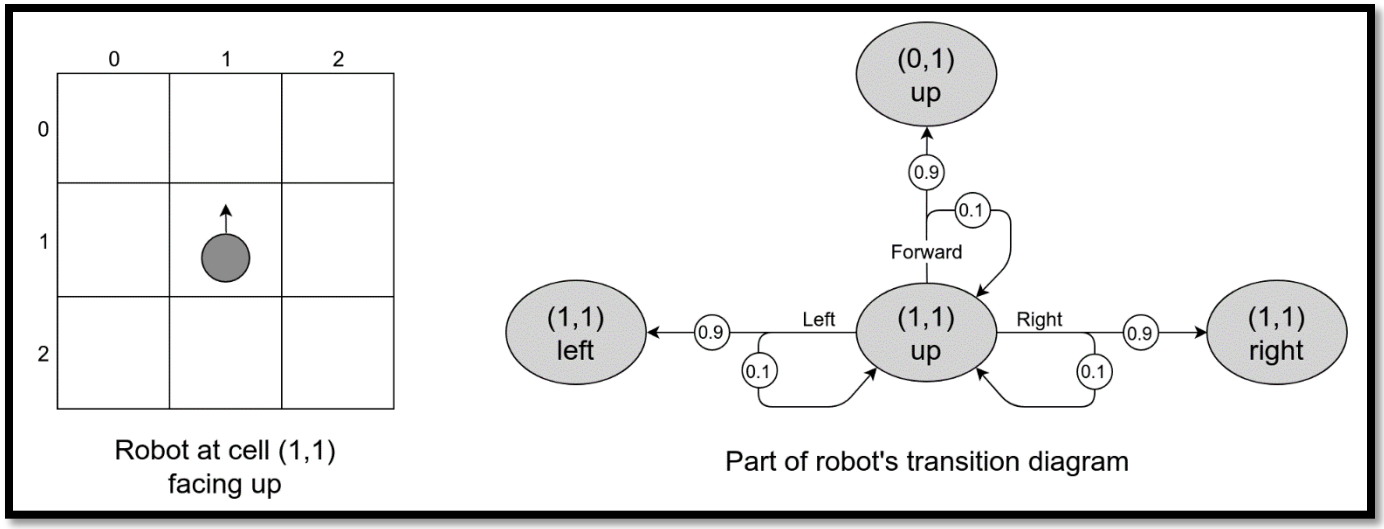

Chat
Computer
Coffee
Home


0.5
Chat
0.3
0.1
0.7
0.2
Computer
0.5
0.1
Coffee
0.2
0.2
0.2
0.4
Home
0.6

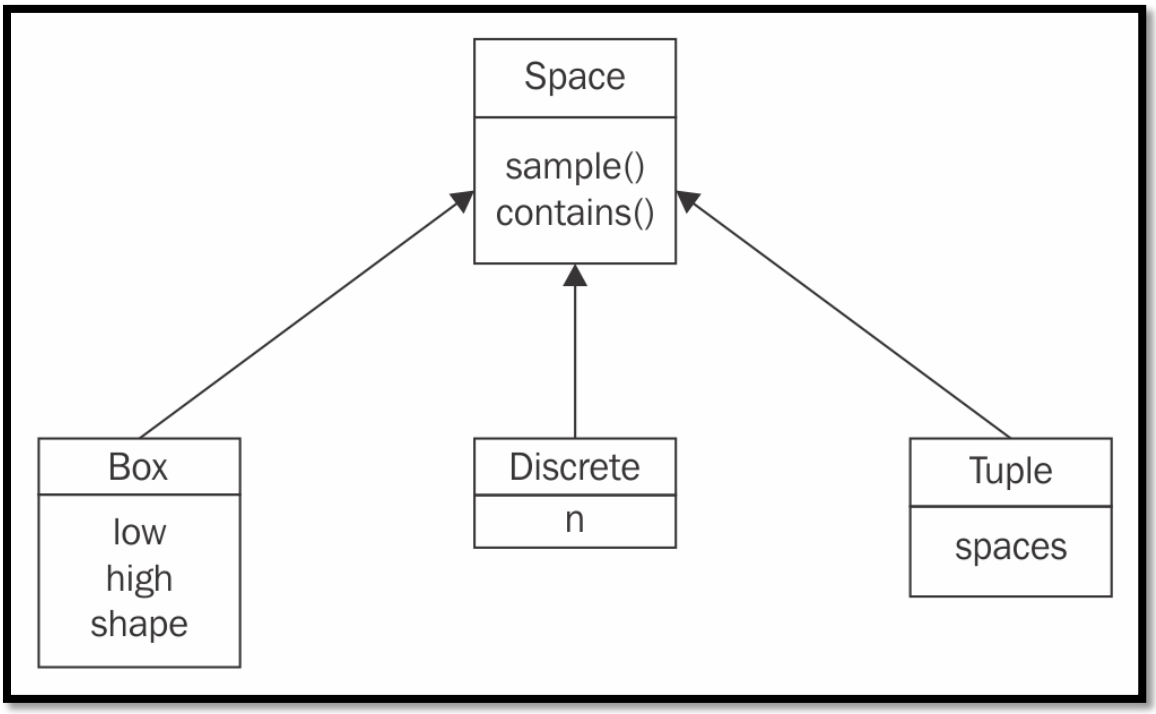

Chat
Computer
Coffee
Home
-1
0.5
0.3
2
-3
0.1
5
0.5
0.7
2
0.2
1
1
0.1
0.2
3
1
0.2
0.2
2
1
0.4
0.6
1

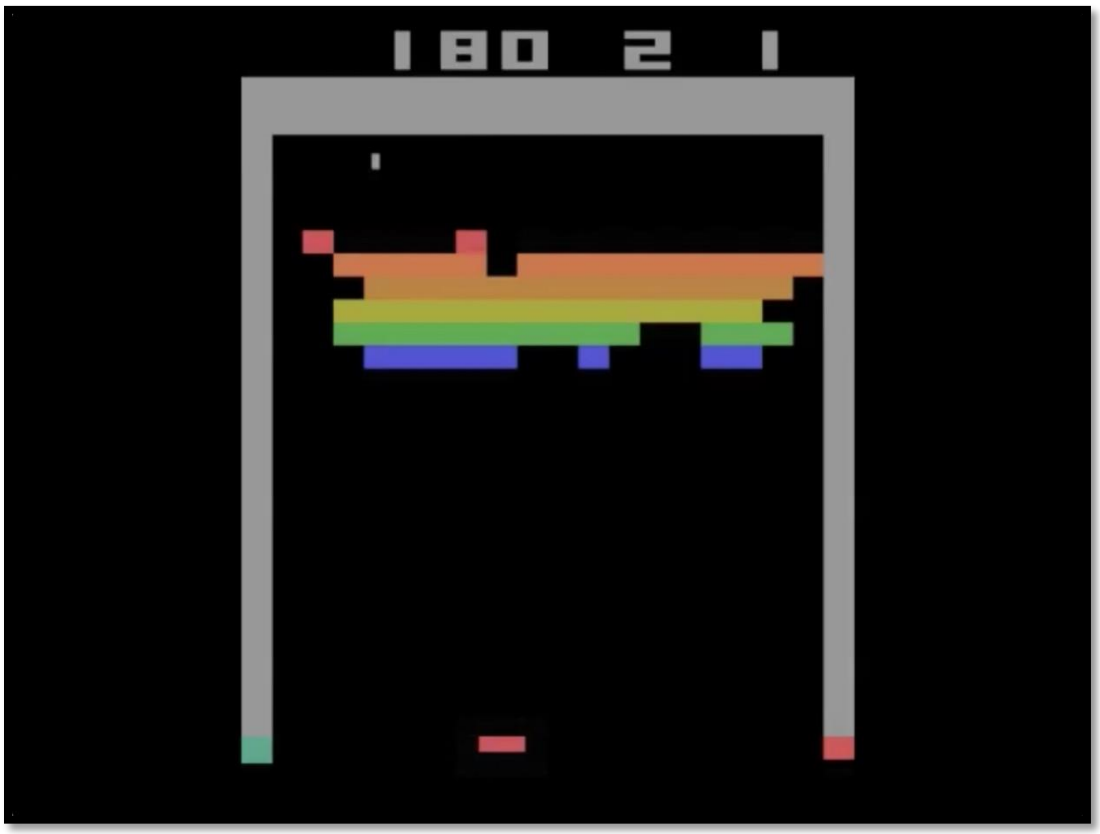

Good old MP
Target state
Source state
Probability for i ⟶ j transition

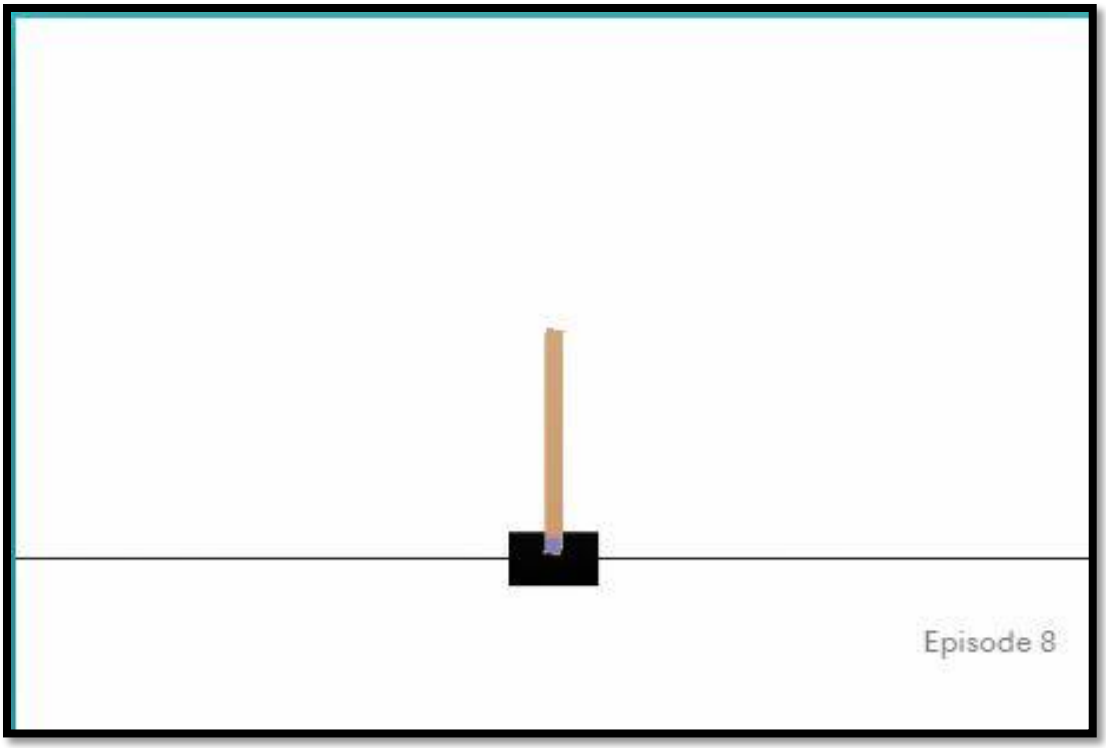

MDP
Probability for
i ⟶ j given action K.
Action
Source
state
target state

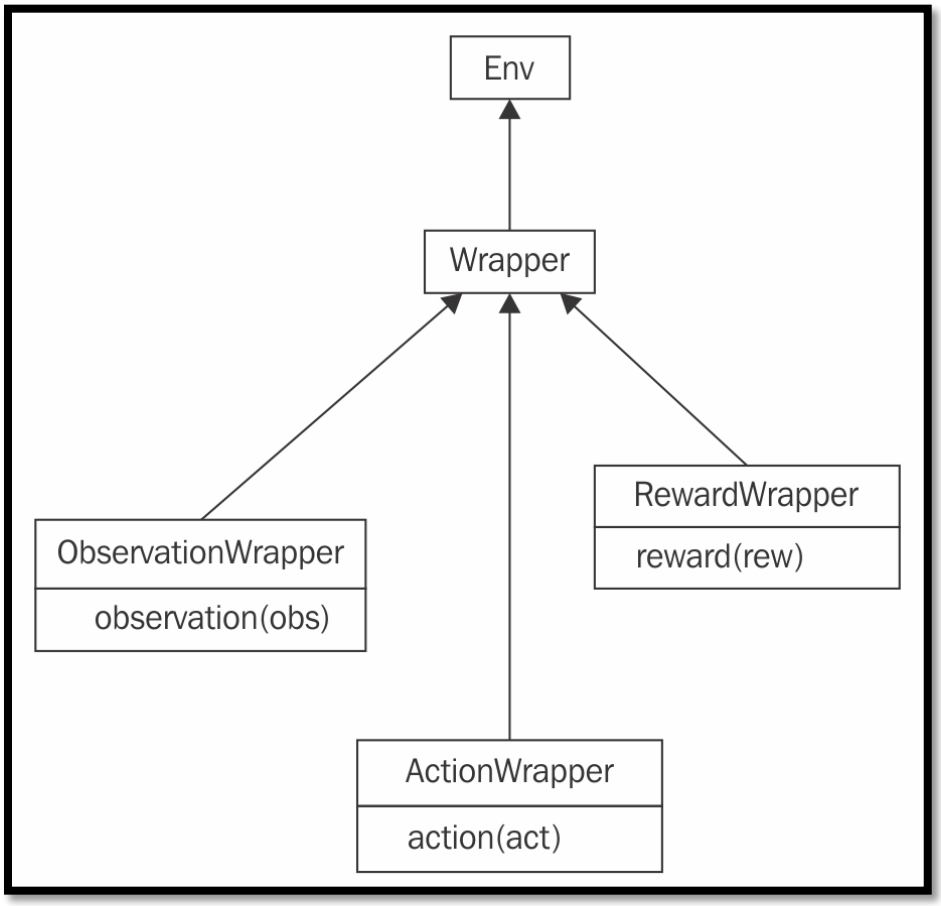

0
1
2
0
1
2
Robot at cell (1,1)
facing up
(0,1)
up
0.9
0.1
Forward
(1,1)
left
0.9
Left
(1,1)
up
Right
0.9
(1,1)
right
0.1
0.1
Part of robot's transition diagram

Chapter 2: OpenAI Gym






Episode 8

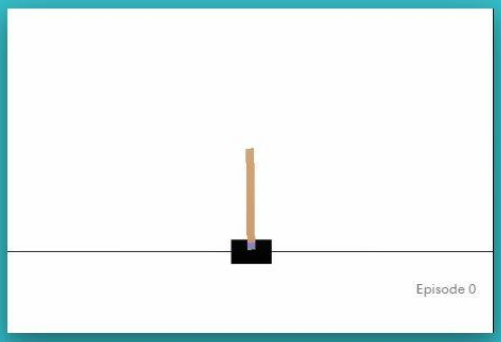

Env
Wrapper
ObservationWrapper
observation(obs)
RewardWrapper
reward(rew)
ActionWrapper
action(act)

CartPole-v0

A pole is attached by an un-actuated joint to a cart, which moves along a frictionless track. The system is controlled by applying a force of +1 or -1 to the cart. The pendulum starts upright, and the goal is to prevent it from falling over. A reward of +1 is provided for every timestep that the pole remains upright. The episode ends when the pole is more than 15 degrees from vertical, or the cart moves more than 2.4 units from the center.

CartPole-v0 defines "solving" as getting average reward of 195.0 over 100 consecutive trials.

This environment corresponds to the version of the cart-pole problem described by Barto, Sutton, and Anderson [Barto83].



karpathy's algorithm, Took 211 episodes to solve the environment. 195.27 ± 1.57 a year ago

CartPole-v0 Evaluations

ALGORITHM	EPISODES BEFORE SOLVE	SUBMITTED
n1try's algorithm writeup	85.0	11 days ago
mbalunovic's algorithm writeup	306.0	11 days ago
ruippeixotog's algorithm writeup	933.0	12 days ago
ruippeixotog's algorithm writeup	961.0	13 days ago

< DoomDefendLine-v0

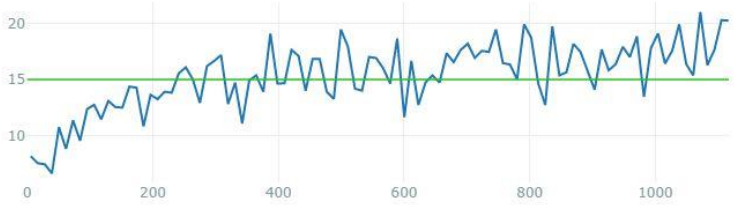
algorithm on DoomDefendLine-v0

Writeup · 5 months ago

Shmuma

Learning performance

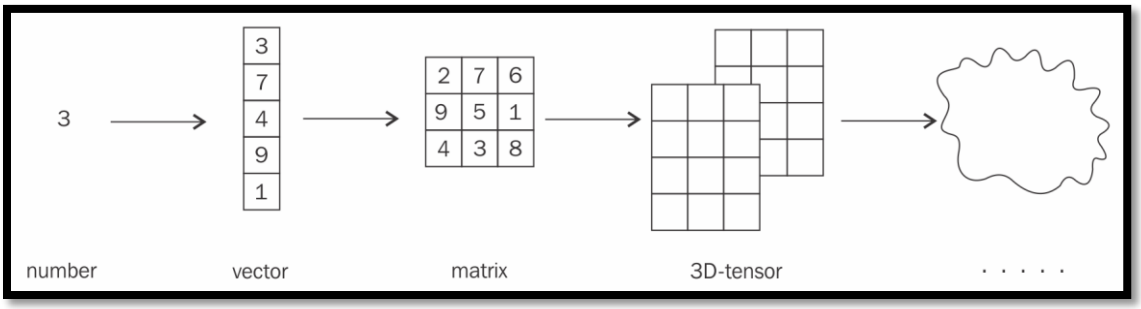
Episode Total Reward ▾ Episode ▾

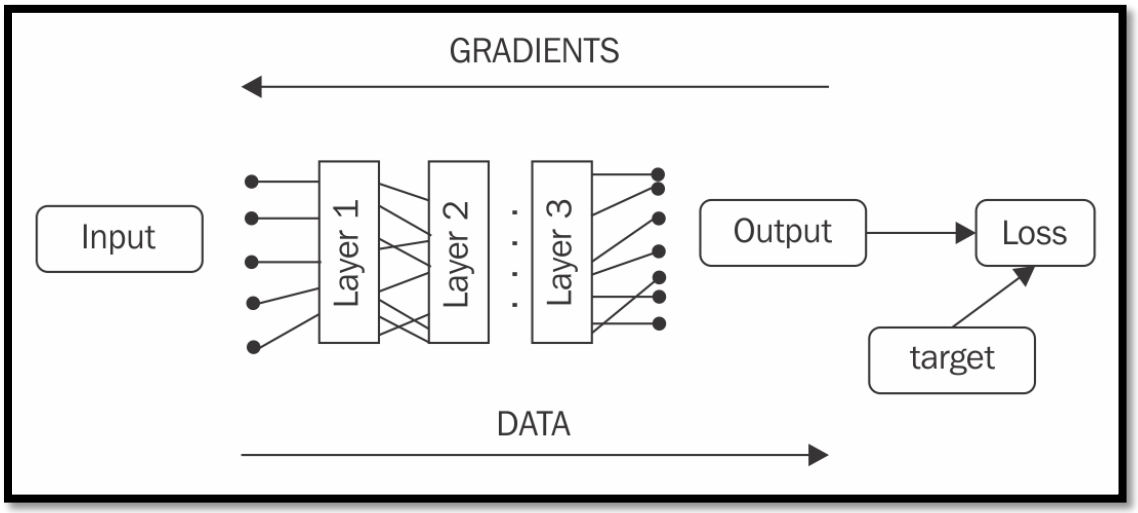


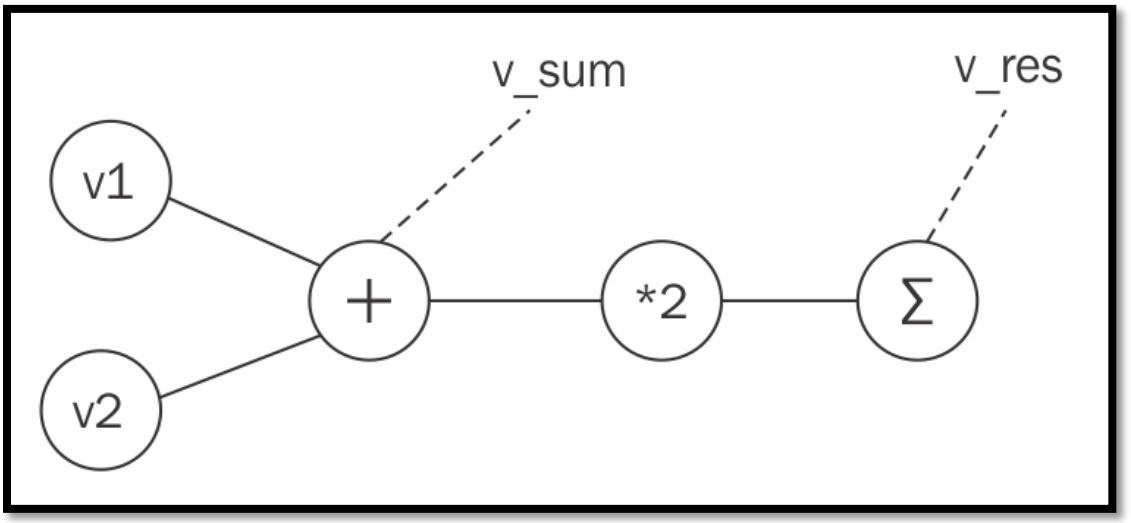
Solved after 205 episodes. Best 100-episode average reward was 18.42 ± 0.76.
(DoomDefendLine-v0 is considered "solved" when the agent obtains an average reward of at least 15.0 over 100 consecutive episodes.)

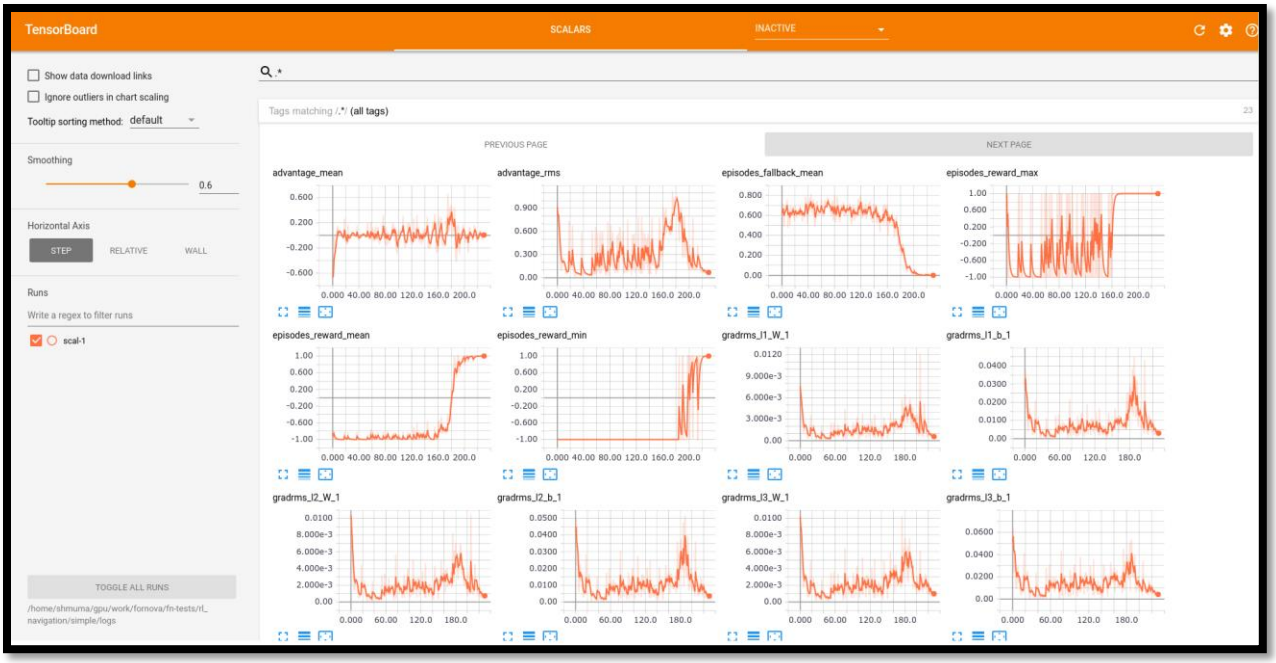
205	1123	✓	40m		
Episodes to solve	Total episodes	Solved	Time to solve	Download	Tweet

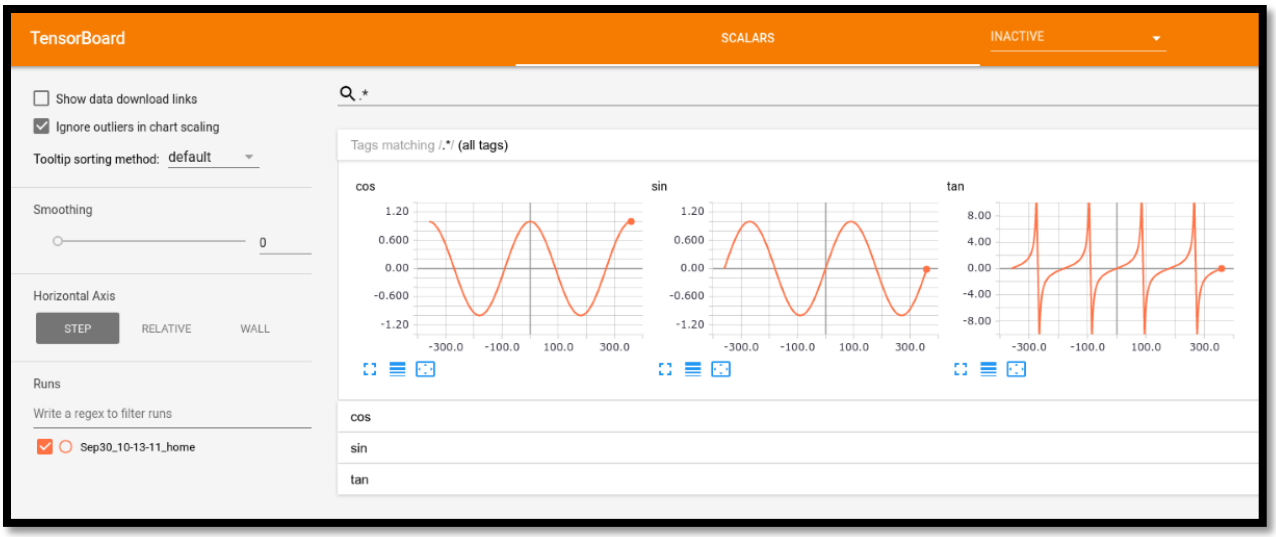
Chapter 3: Deep Learning with PyTorch

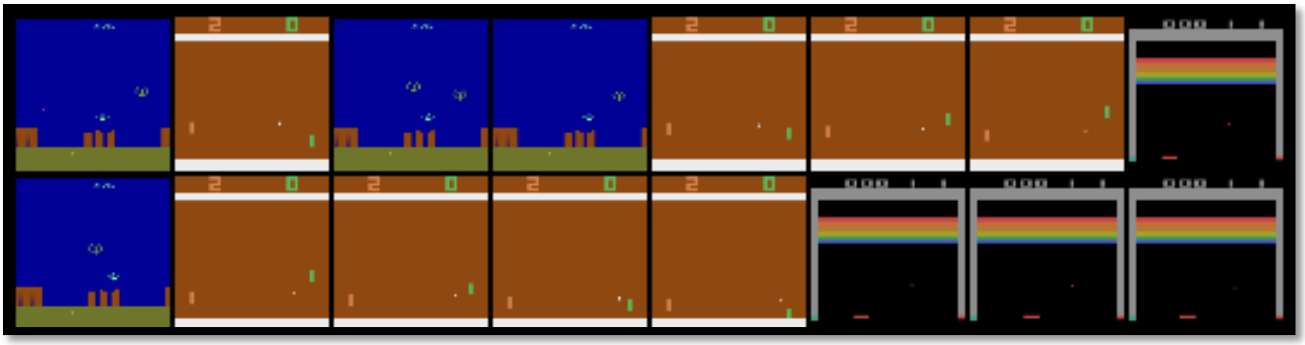


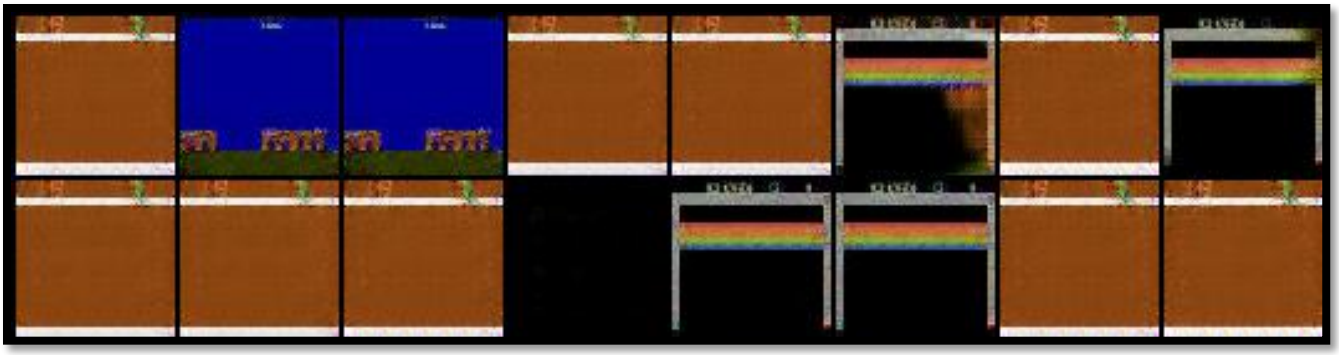




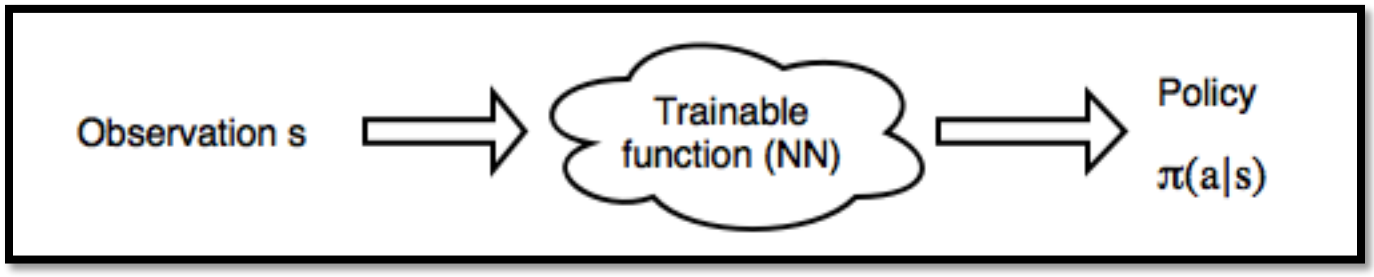

TensorBoard
SCALARS
INACTIVE
Show data download links
Ignore outliers in chart scaling
Tooltip sorting method: default
Smoothing
0.6
Horizontal Axis
STEP
RELATIVE
WALL
Runs
Write a regex to filter runs
scal-1
TOGGLE ALL RUNS
/home/shmuma/gpu/work/fornova/fn-tests/rl_navigation/simple/logs
.*
Tags matching /.*/ (all tags)
23
PREVIOUS PAGE
NEXT PAGE
advantage_mean
advantage_rms
episodes_fallback_mean
episodes_reward_max
episodes_reward_mean
episodes_reward_min
gradrms_l1_W_1
gradrms_l1_b_1
gradrms_l2_W_1
gradrms_l2_b_1
gradrms_l3_W_1
gradrms_l3_b_1

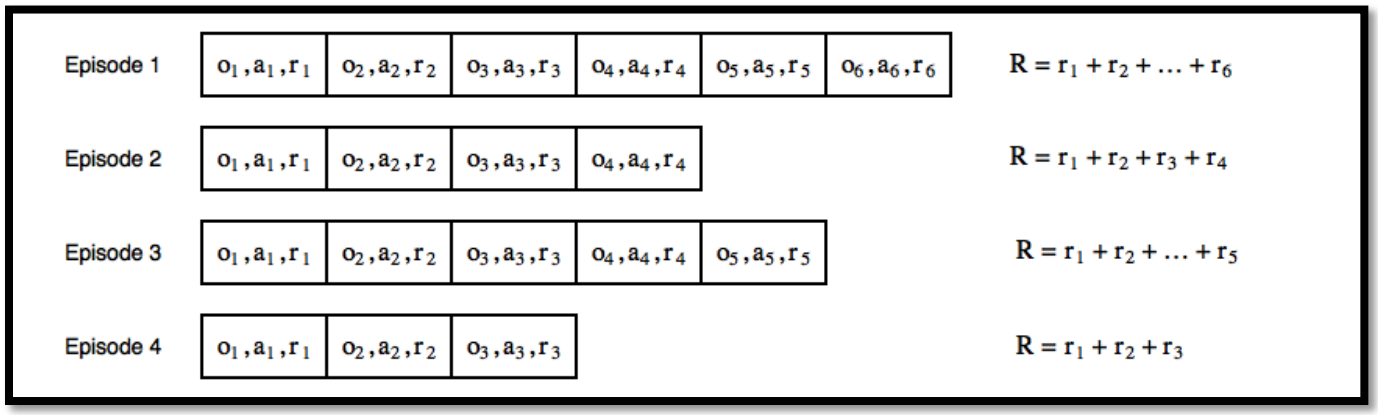

TensorBoard
SCALARS
INACTIVE
Show data download links
Ignore outliers in chart scaling
Tooltip sorting method: default
Smoothing
0
Horizontal Axis
STEP
RELATIVE
WALL
Runs
Write a regex to filter runs
Sep30_10-13-11_home
.*
Tags matching /.*/ (all tags)
cos
sin
tan

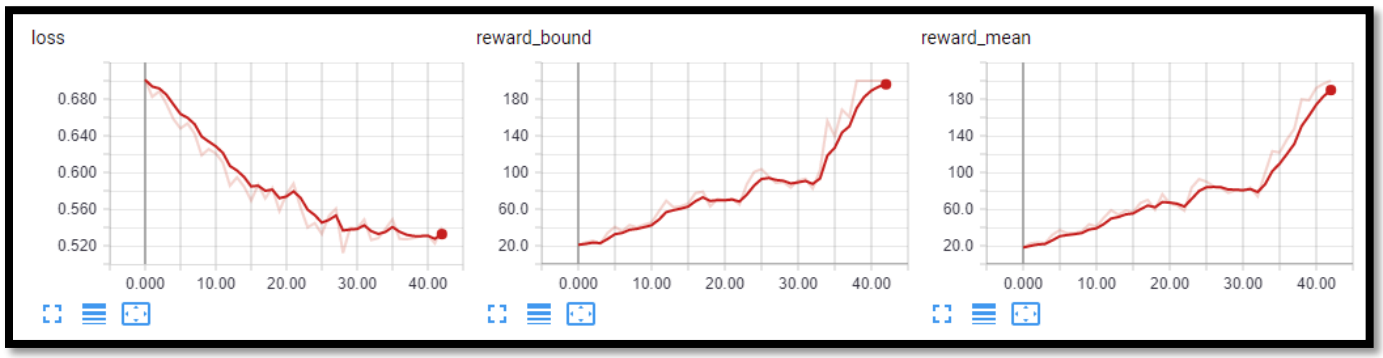


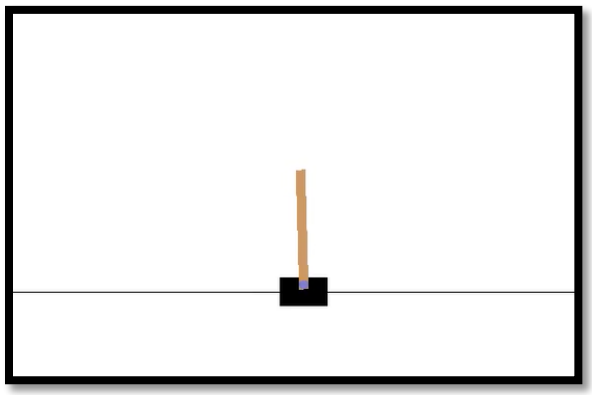


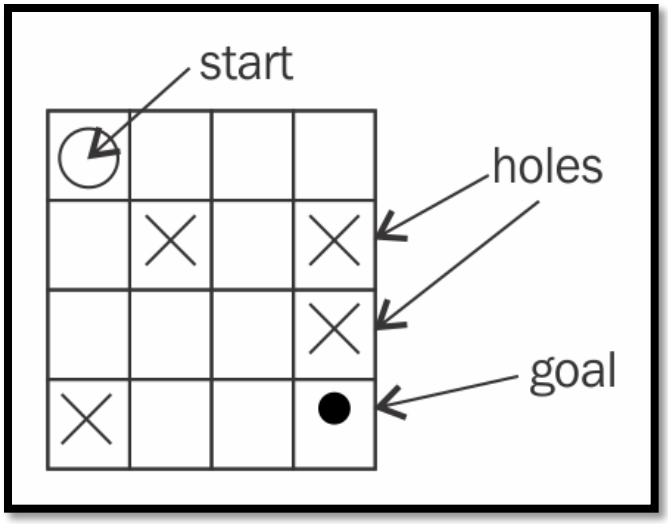
Chapter 4: The Cross-Entropy Method

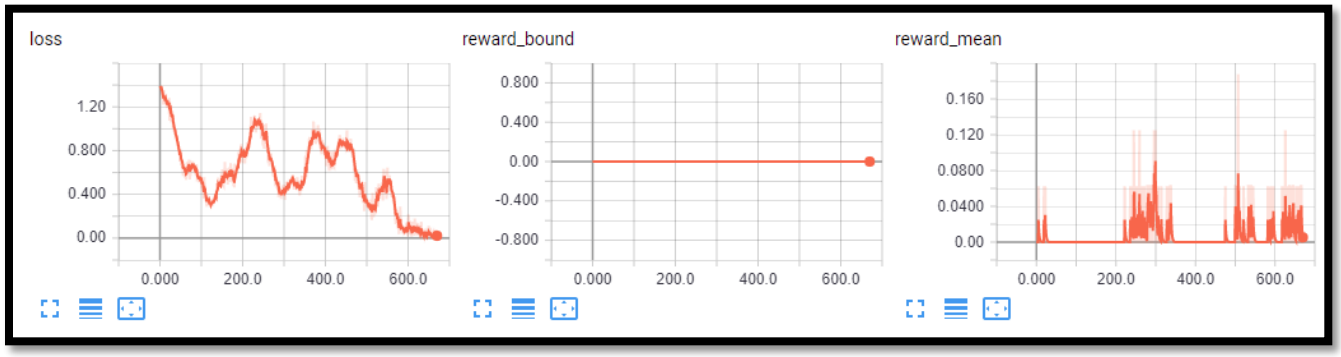


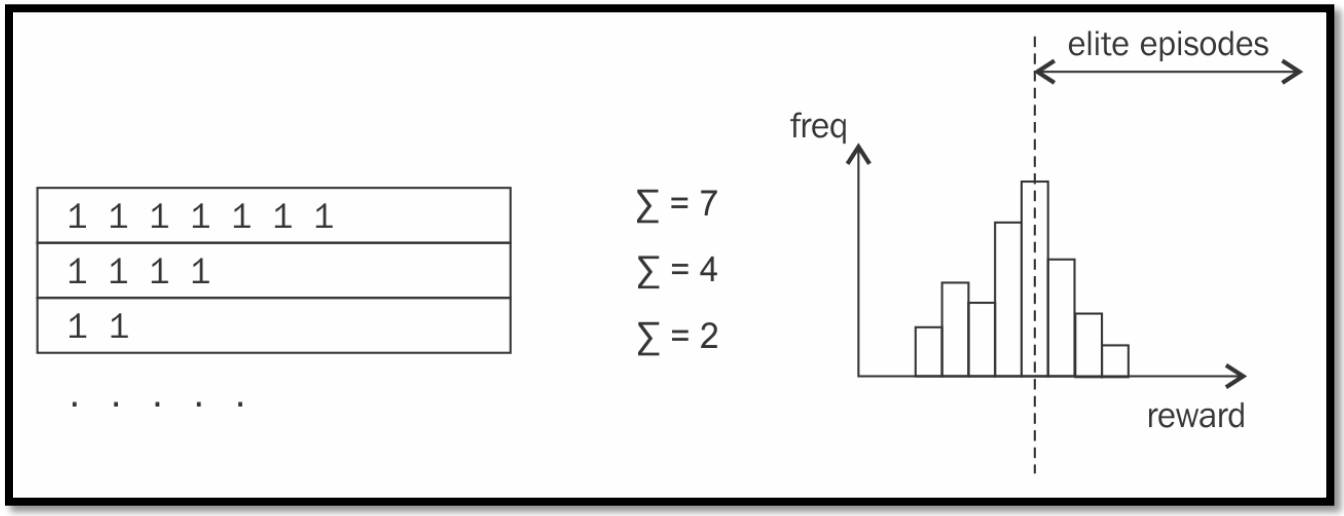


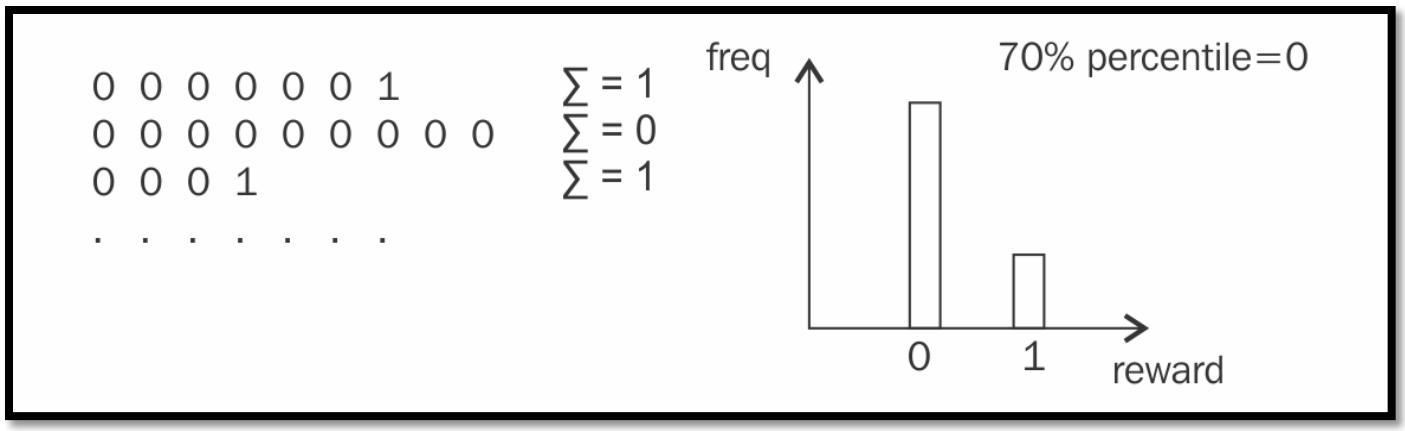


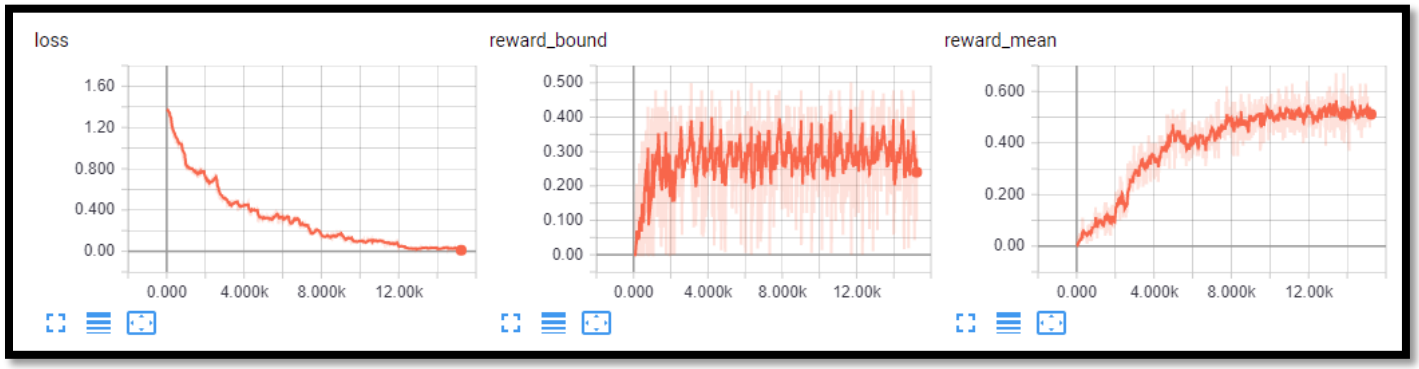


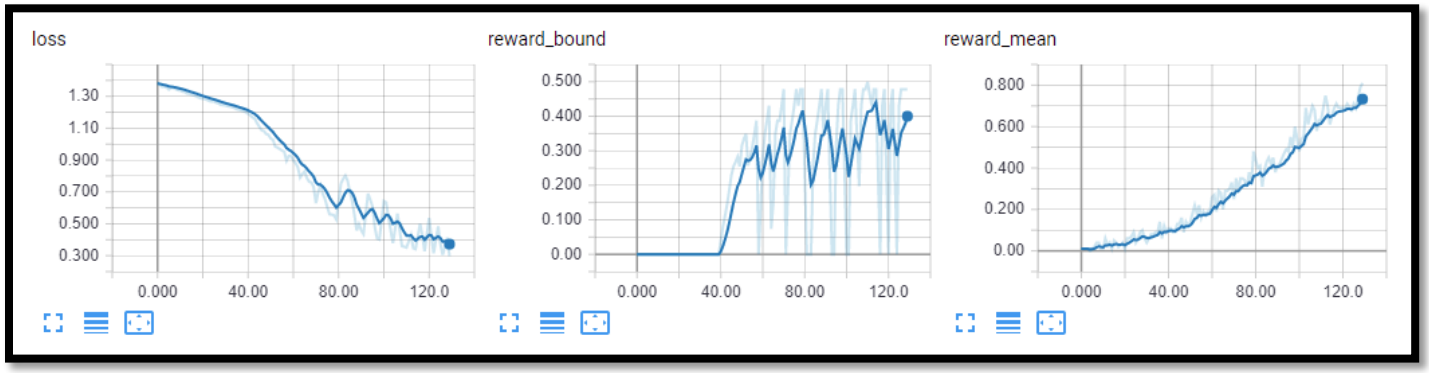

start
holes
goal

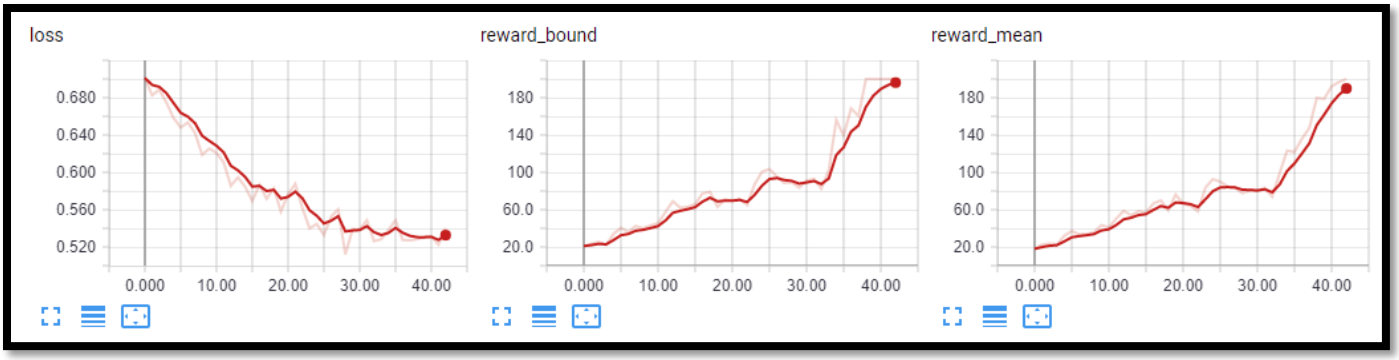

loss
1.20
0.800
0.400
0.00
0.000
200.0
400.0
600.0
reward_bound
0.800
0.400
0.00
-0.400
-0.800
0.000
200.0
400.0
600.0
reward_mean
0.160
0.120
0.0800
0.0400
0.00
0.000
200.0
400.0
600.0

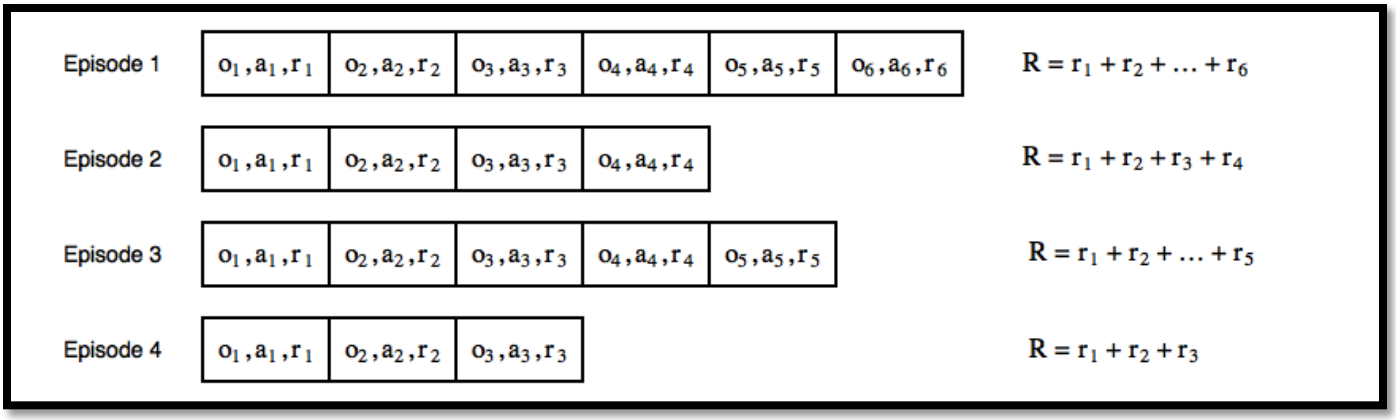

1 1 1 1 1 1 1
1 1 1 1
1 1
. . . . .
∑ = 7
∑ = 4
∑ = 2
elite episodes
freq
reward


0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0
0 0 0 1
. . . . . .
∑ = 1
∑ = 0
∑ = 1
freq
70% percentile=0
0
1
reward

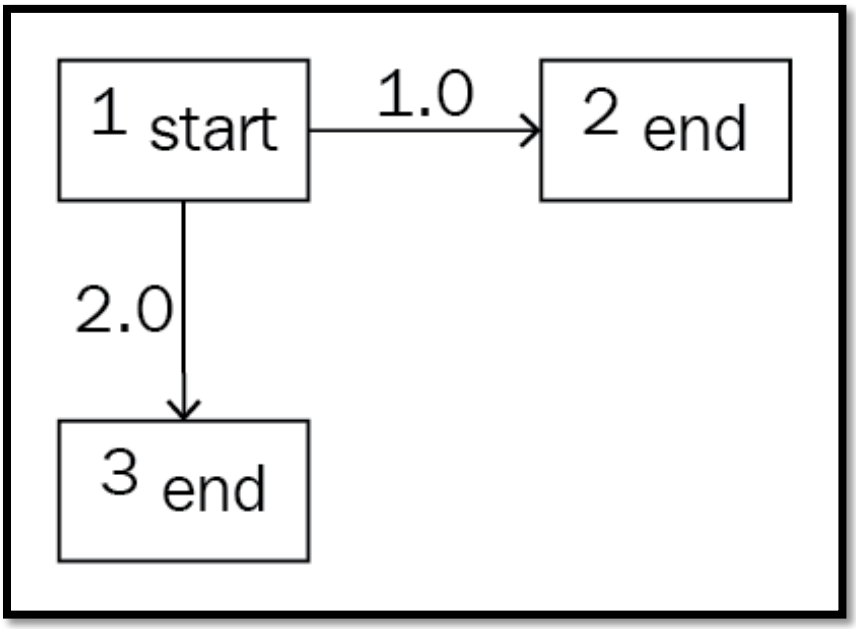

loss
1.60
1.20
0.800
0.400
0.00
0.000
4.000k
8.000k
12.00k
reward_bound
0.500
0.400
0.300
0.200
0.100
0.00
0.000
4.000k
8.000k
12.00k
reward_mean
0.600
0.400
0.200
0.00
0.000
4.000k
8.000k
12.00k

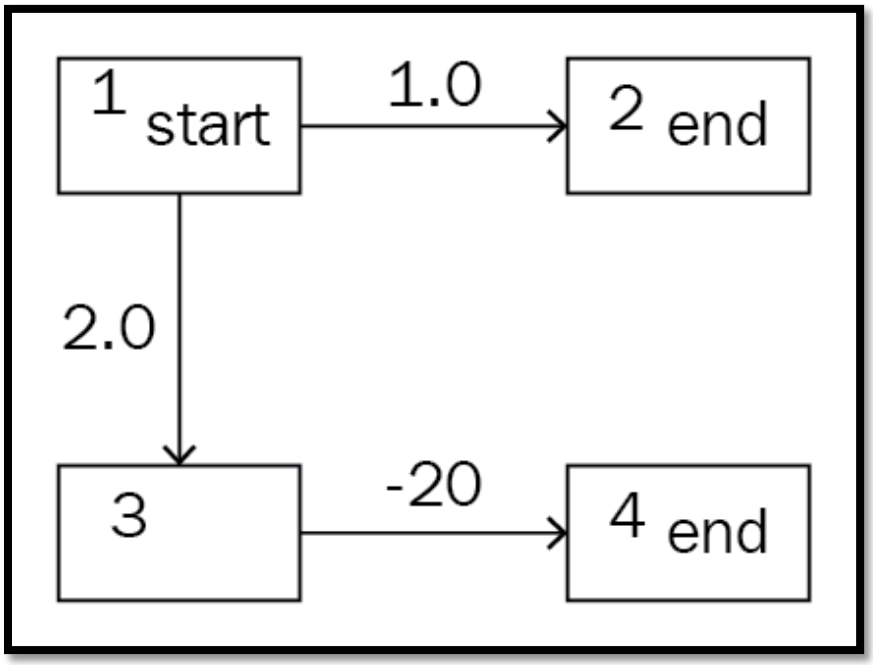

loss
1.30
1.10
0.900
0.700
0.500
0.300
0.000
40.00
80.00
120.0
reward_bound
0.500
0.400
0.300
0.200
0.100
0.00
0.000
40.00
80.00
120.0
reward_mean
0.800
0.600
0.400
0.200
0.00
0.000
40.00
80.00
120.0

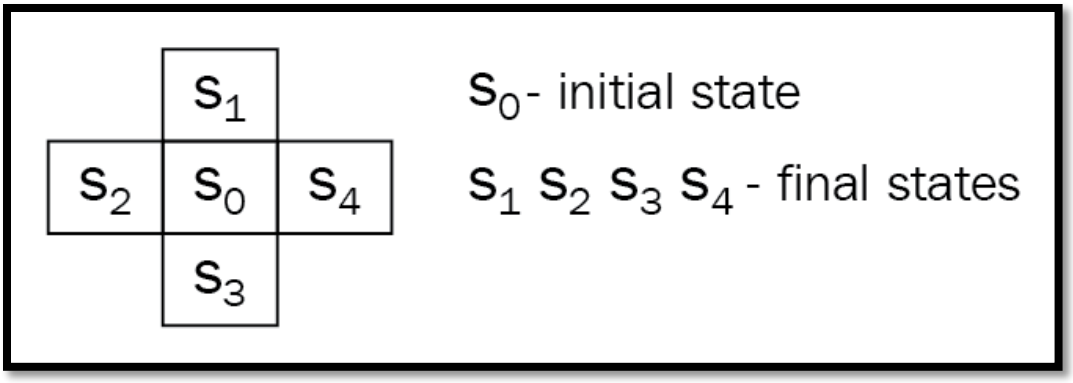

loss
0.680
0.640
0.600
0.560
0.520
0.000
10.00
20.00
30.00
40.00
reward_bound
180
140
100
60.0
20.0
0.000
10.00
20.00
30.00
40.00
reward_mean
180
140
100
60.0
20.0
0.000
10.00
20.00
30.00
40.00

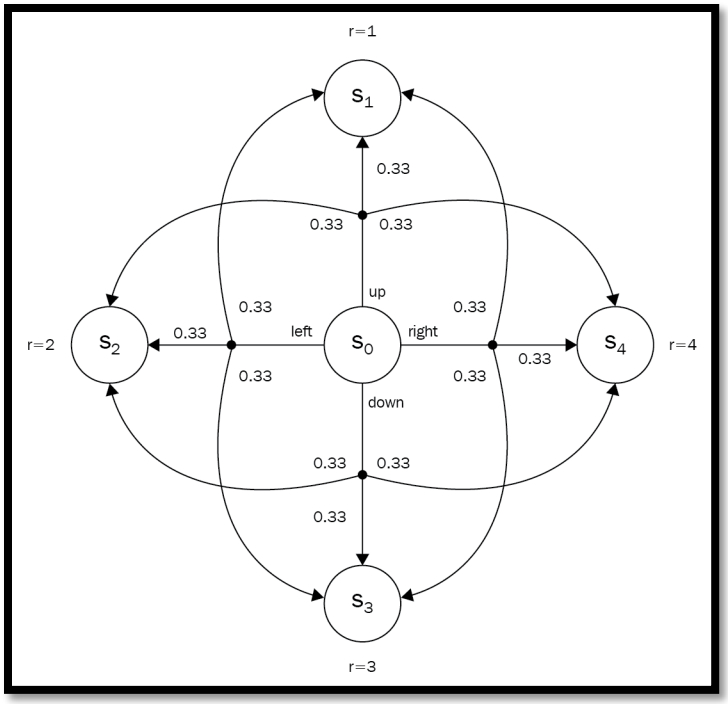

Episode 1
o_1, a_1, r_1 | o_2, a_2, r_2 | o_3, a_3, r_3 | o_4, a_4, r_4 | o_5, a_5, r_5 | o_6, a_6, r_6
$R = r_1 + r_2 + \ldots + r_6$
Episode 2
o_1, a_1, r_1 | o_2, a_2, r_2 | o_3, a_3, r_3 | o_4, a_4, r_4
$R = r_1 + r_2 + r_3 + r_4$
Episode 3
o_1, a_1, r_1 | o_2, a_2, r_2 | o_3, a_3, r_3 | o_4, a_4, r_4 | o_5, a_5, r_5
$R = r_1 + r_2 + \ldots + r_5$
Episode 4
o_1, a_1, r_1 | o_2, a_2, r_2 | o_3, a_3, r_3
$R = r_1 + r_2 + r_3$

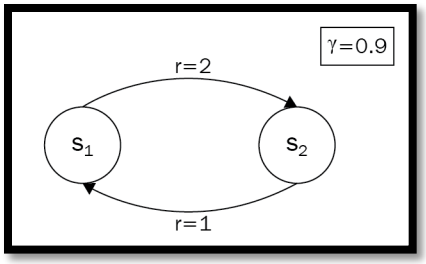
Chapter 5: Tabular Learning and the Bellman Equation

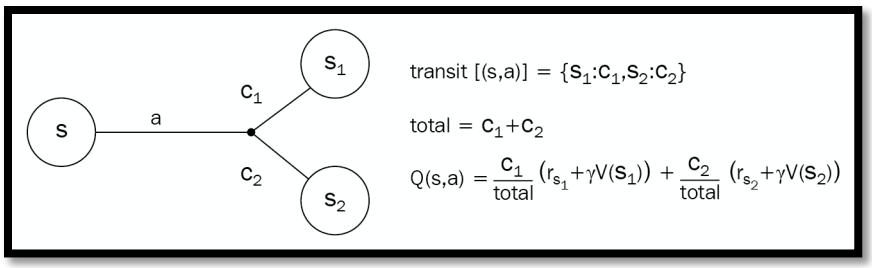


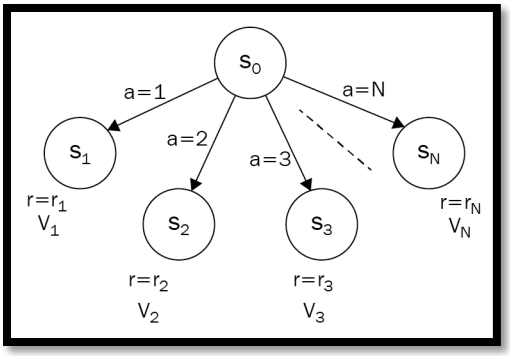


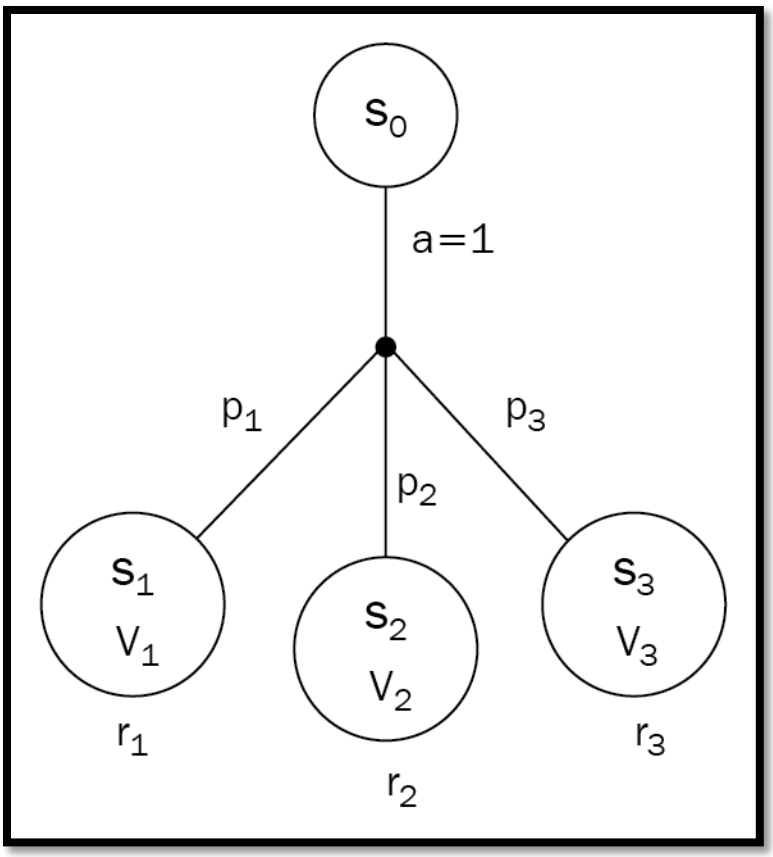

s_1
s_2 s_0 s_4
s_3
s_0 - initial state
s_1 s_2 s_3 s_4 - final states

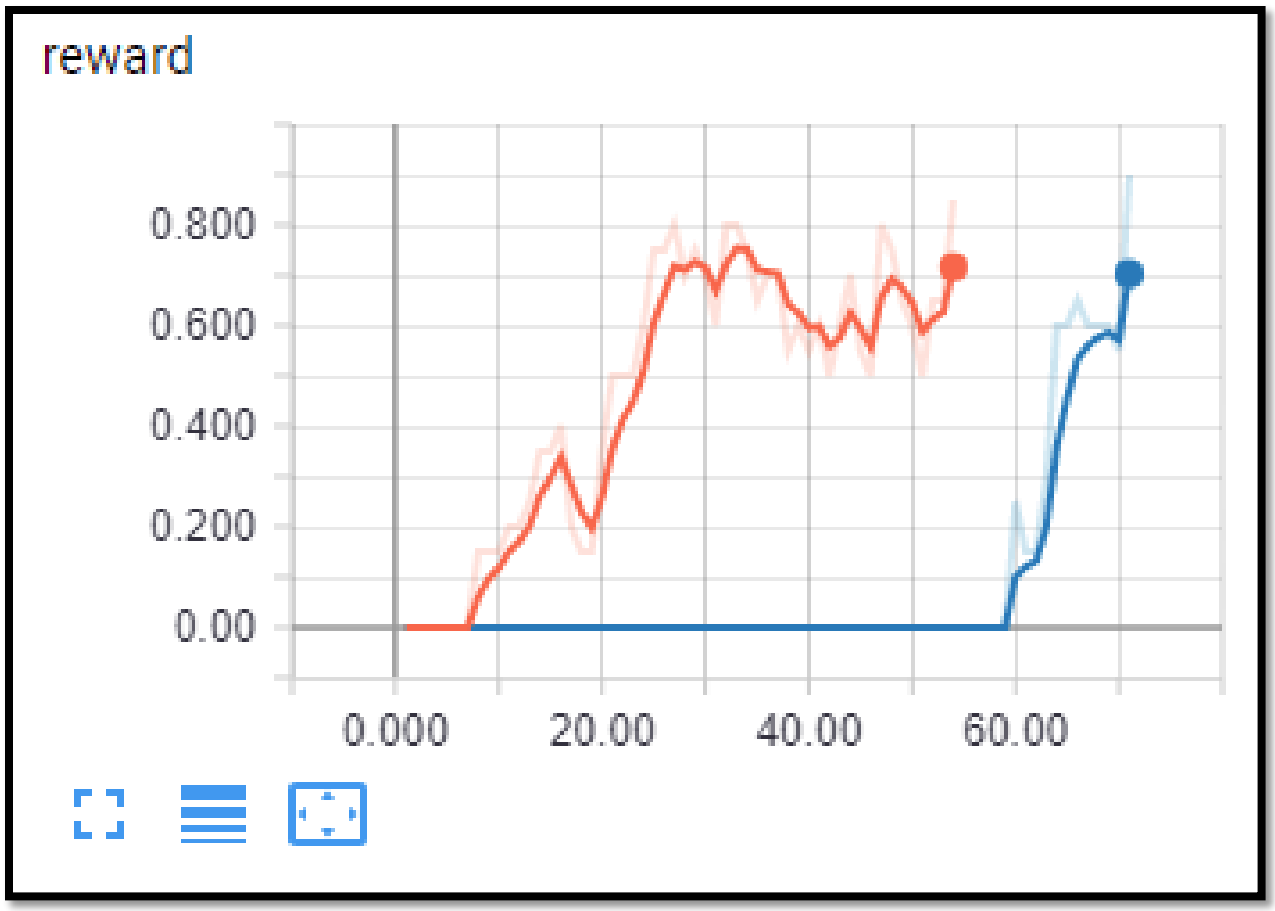

r=1
s_1
0.33
0.33
0.33
up
0.33
0.33
r=2
s_2
0.33
left
s_0
right
s_4
r=4
0.33
0.33
0.33
down
0.33
0.33
0.33
s_3
r=3


γ=0.9
r=2
s_1
s_2
r=1

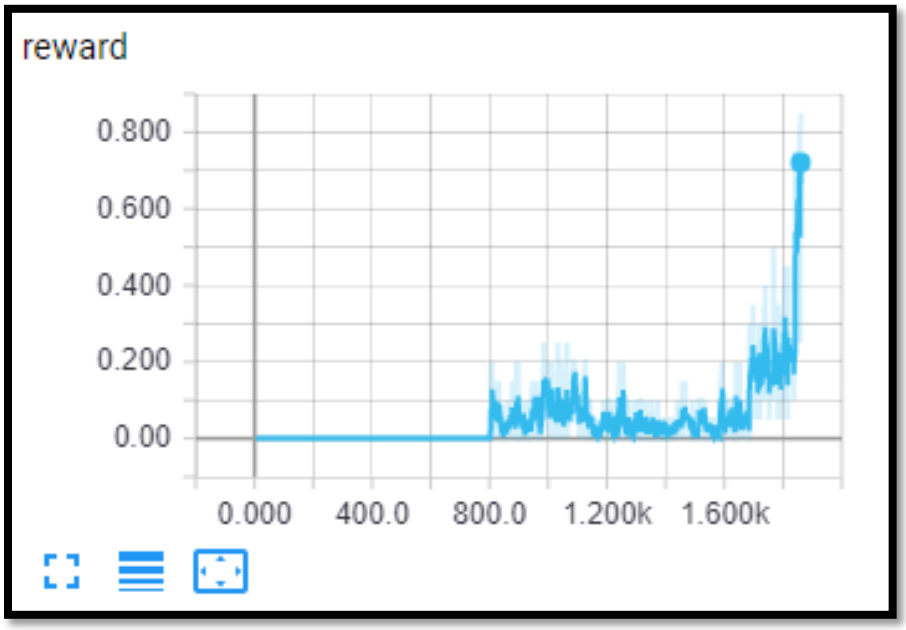

s
a
s_1
c_1
c_2
s_2
transit [(s,a)] = {s_1:c_1,s_2:c_2}
total = c_1+c_2
$Q(s,a) = \frac{c_1}{total}(r_{s_1}+\gamma V(s_1)) + \frac{c_2}{total}(r_{s_2}+\gamma V(s_2))$

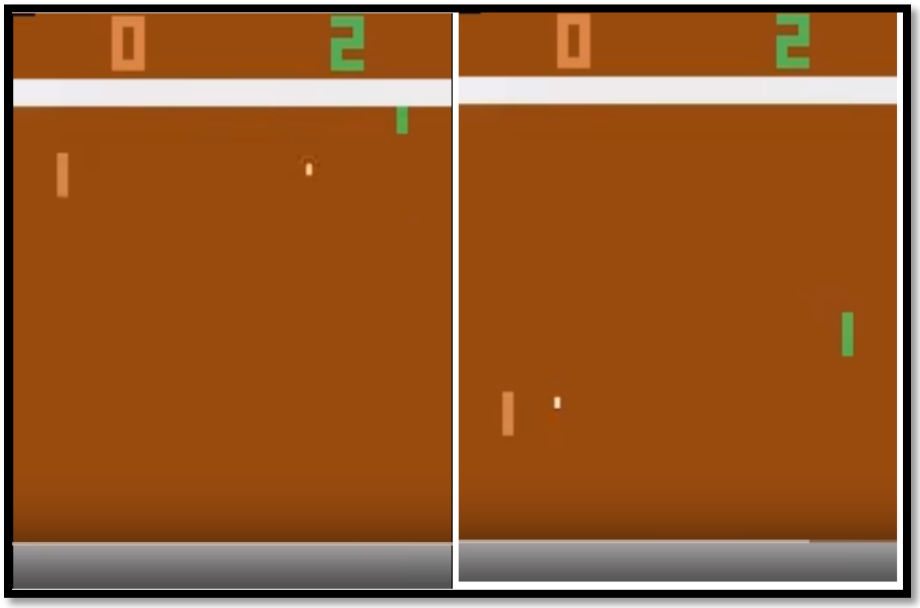

s_0
a=1
a=2
a=3
a=N
s_1
s_2
s_3
s_N
$r=r_1$
V_1
$r=r_2$
V_2
$r=r_3$
V_3
$r=r_N$
V_N

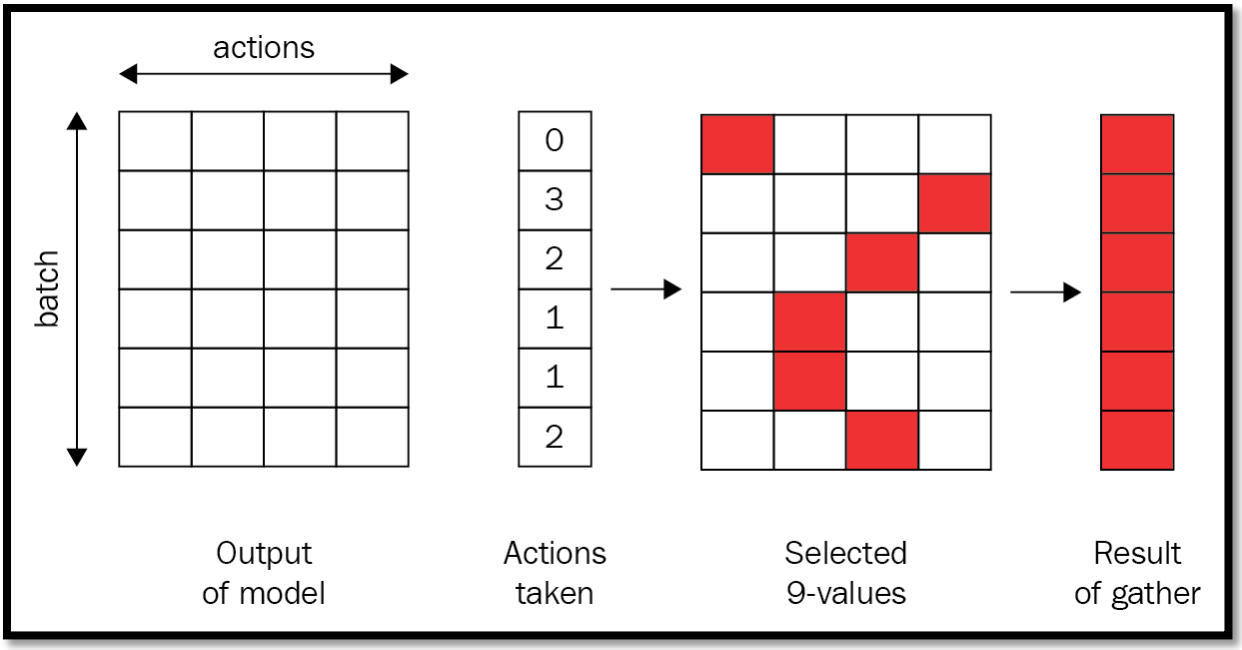

s_0
a=1
p_1
p_2
p_3
s_1
V_1
r_1
s_2
V_2
r_2
s_3
V_3
r_3

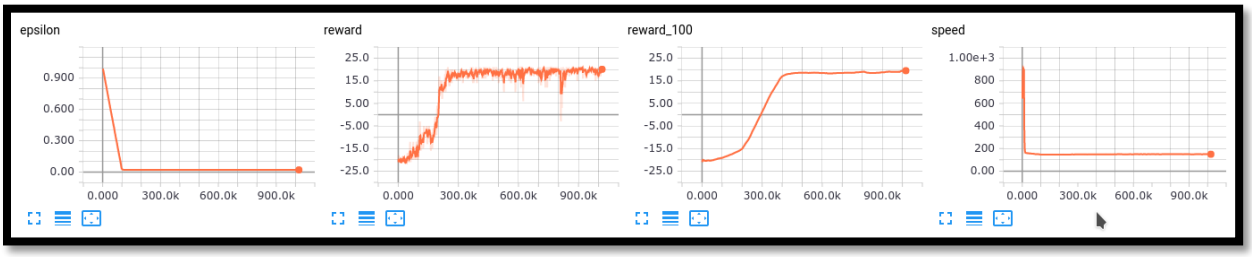

reward
0.800
0.600
0.400
0.200
0.00
0.000
20.00
40.00
60.00

Chapter 6: Deep Q-Networks

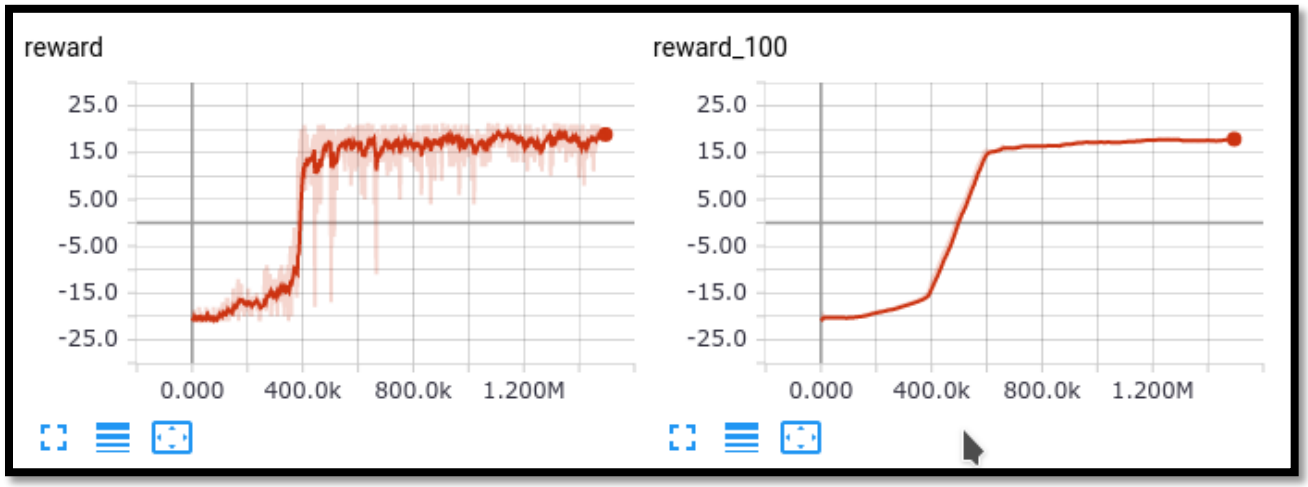


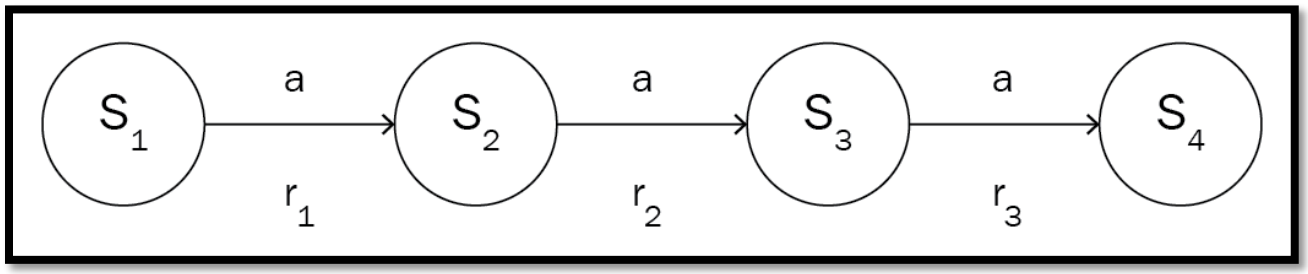


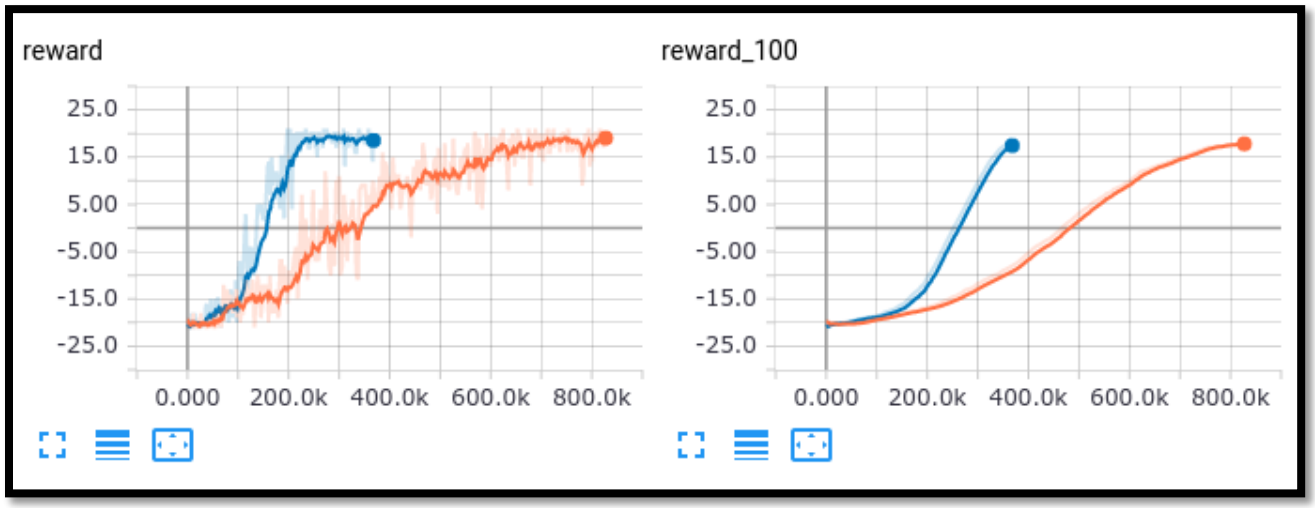

actions
batch
0
3
2
1
1
2
Output of model
Actions taken
Selected 9-values
Result of gather

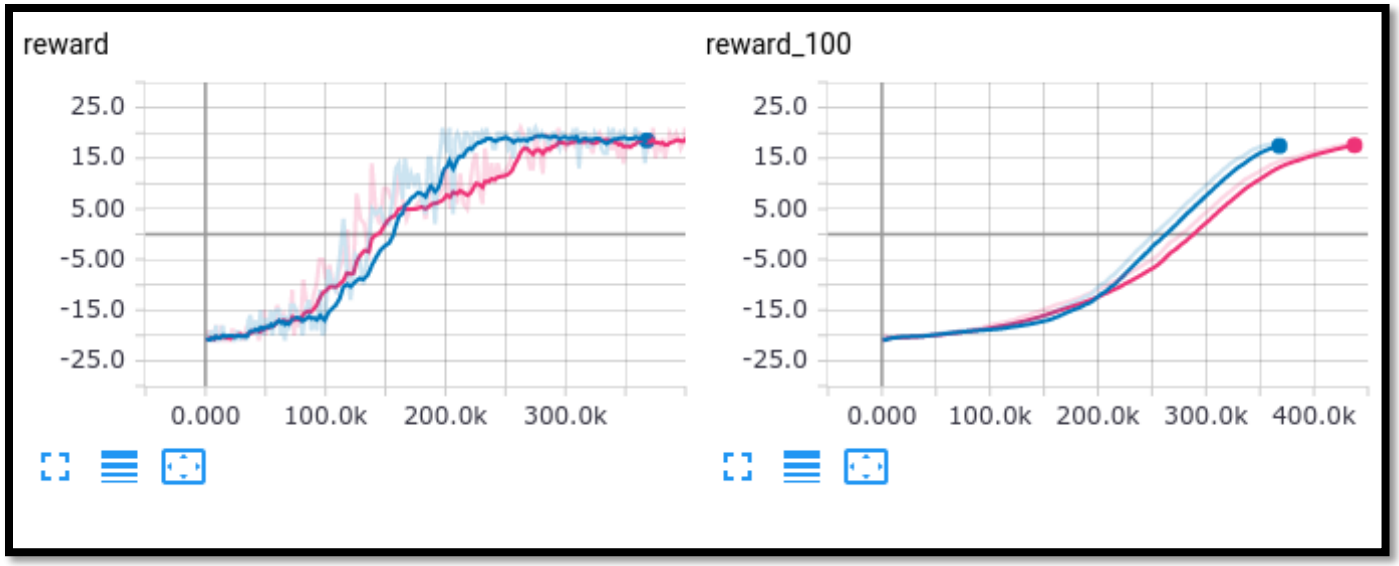

epsilon
0.900
0.600
0.300
0.00
0.000
300.0k
600.0k
900.0k
reward
25.0
15.0
5.00
-5.00
-15.0
-25.0
reward_100
speed
1.00e+3
800
600
400
200
0.00

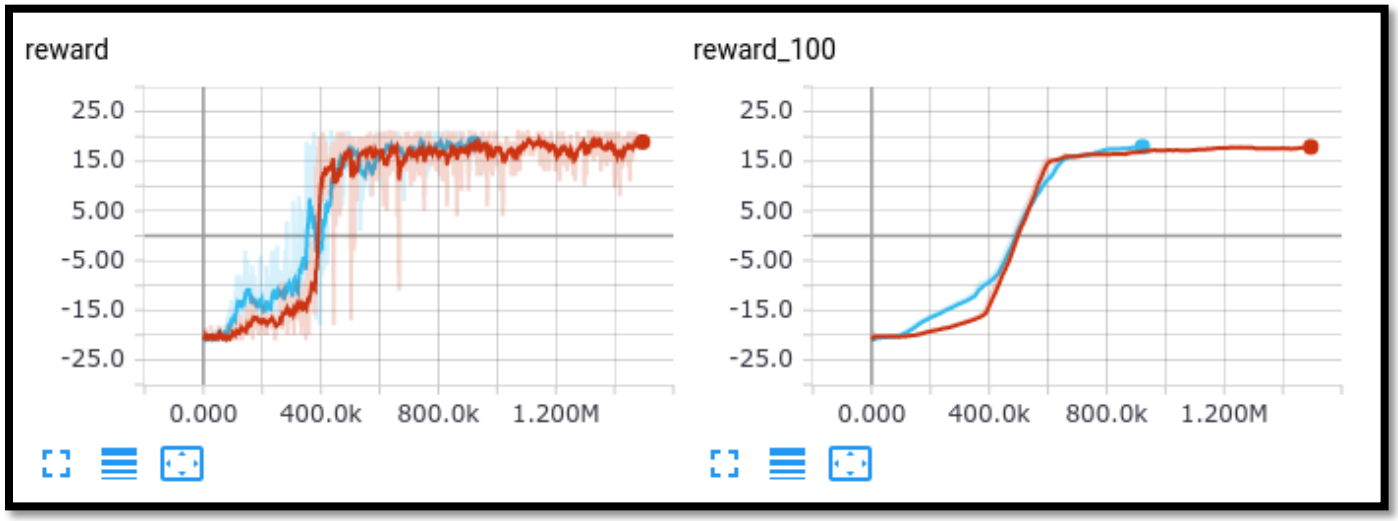
Chapter 7: DQN Extensions

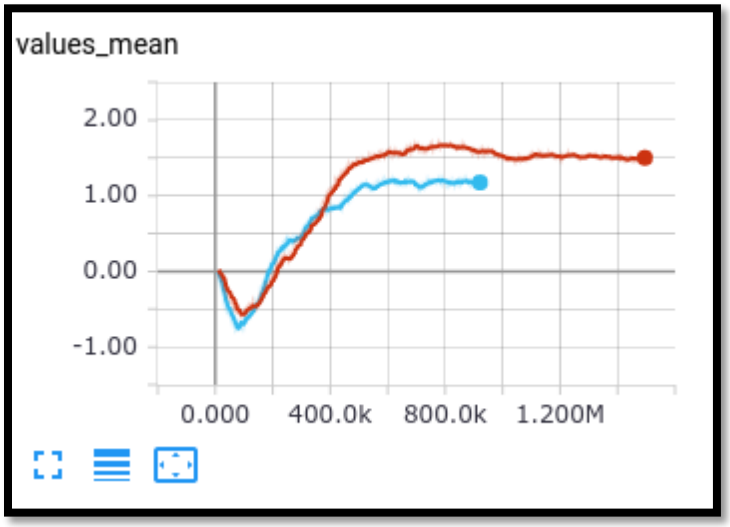


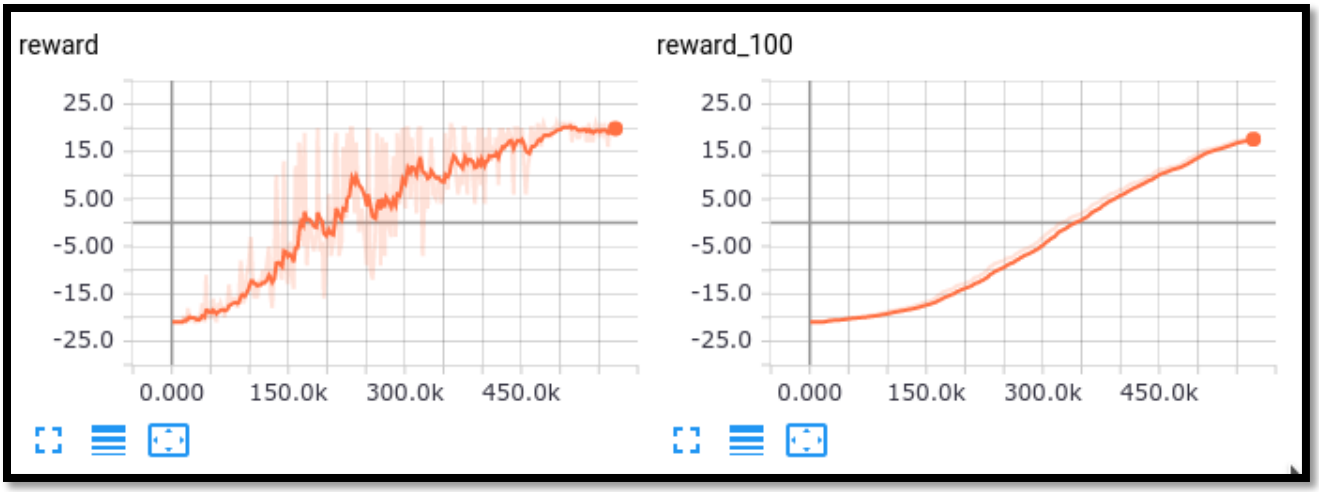


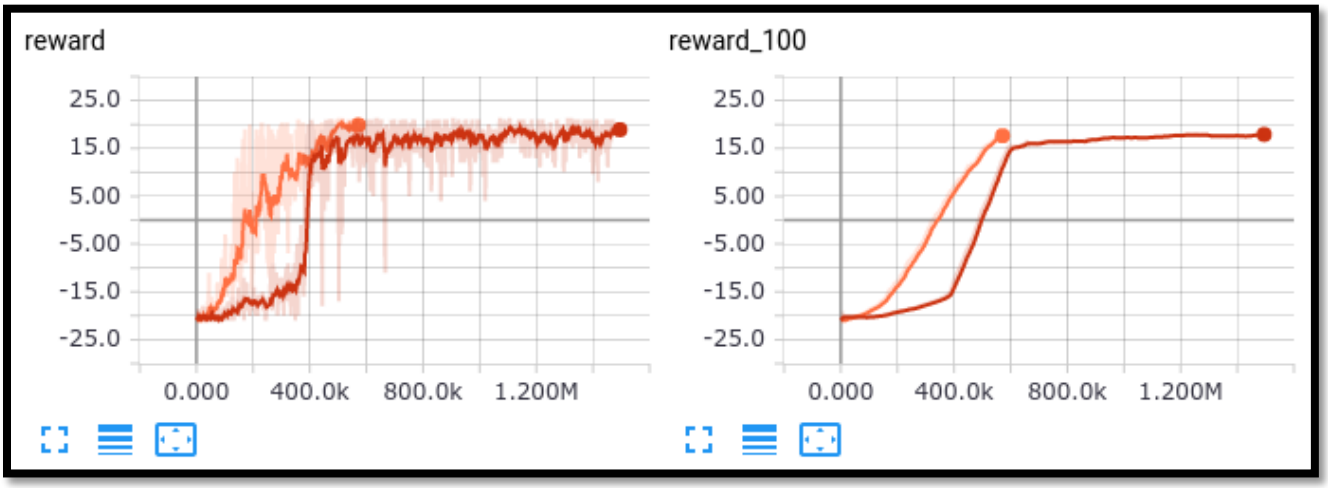


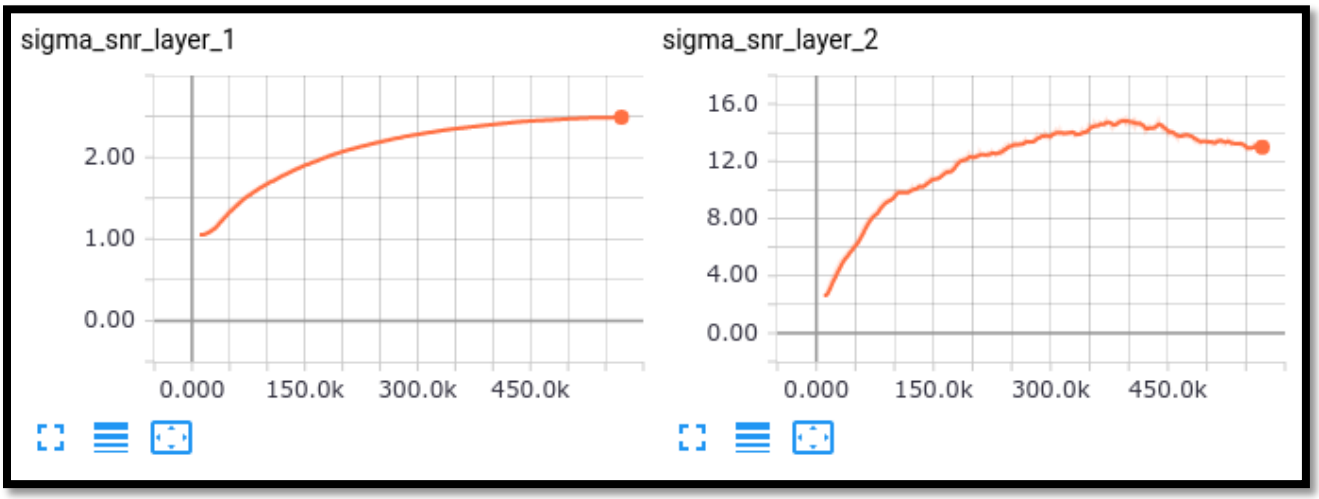

reward
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
100.0k
200.0k
300.0k
reward_100
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
100.0k
200.0k
300.0k
400.0k

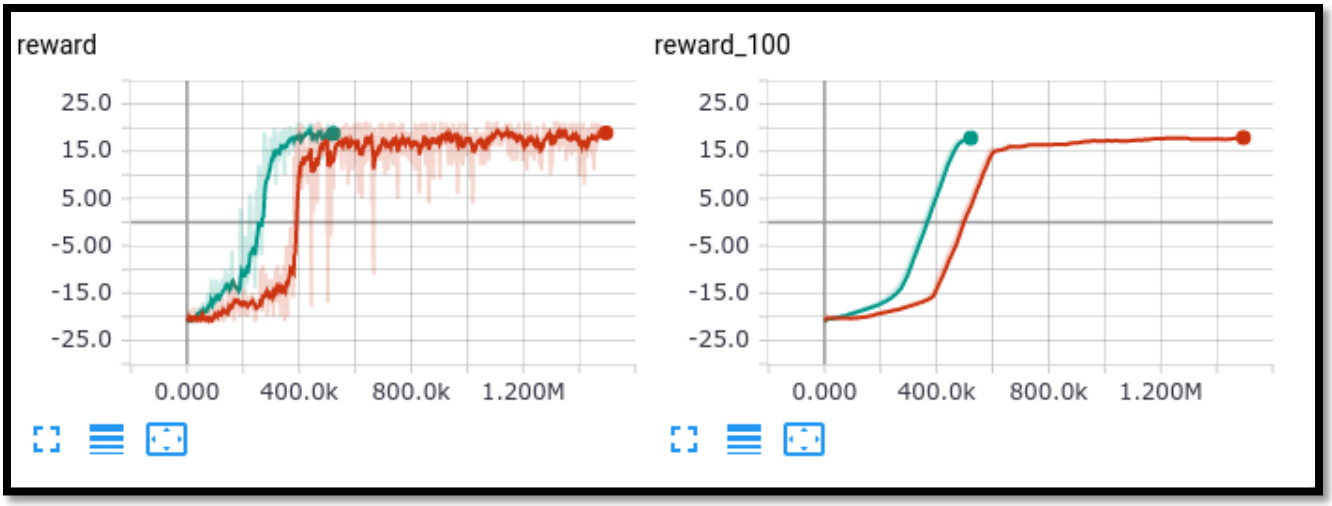

reward
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
400.0k
800.0k
1.200M
reward_100
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
400.0k
800.0k
1.200M

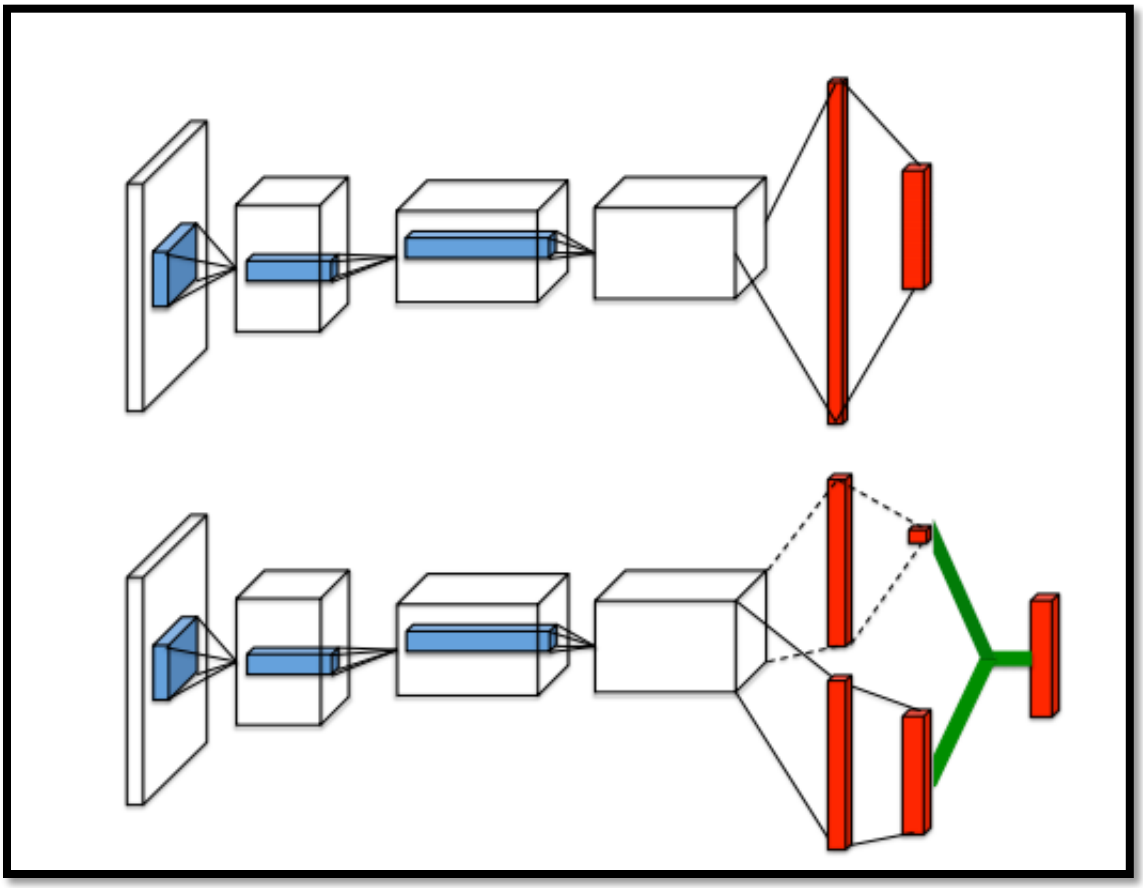

values_mean
2.00
1.00
0.00
-1.00
0.000
400.0k
800.0k
1.200M

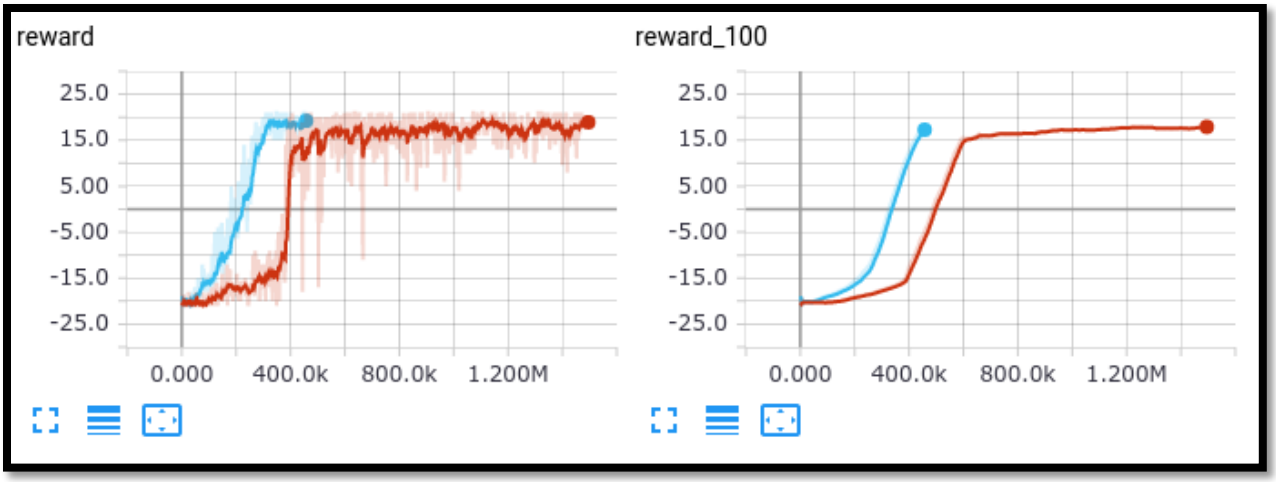

reward
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
150.0k
300.0k
450.0k
reward_100
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
150.0k
300.0k
450.0k

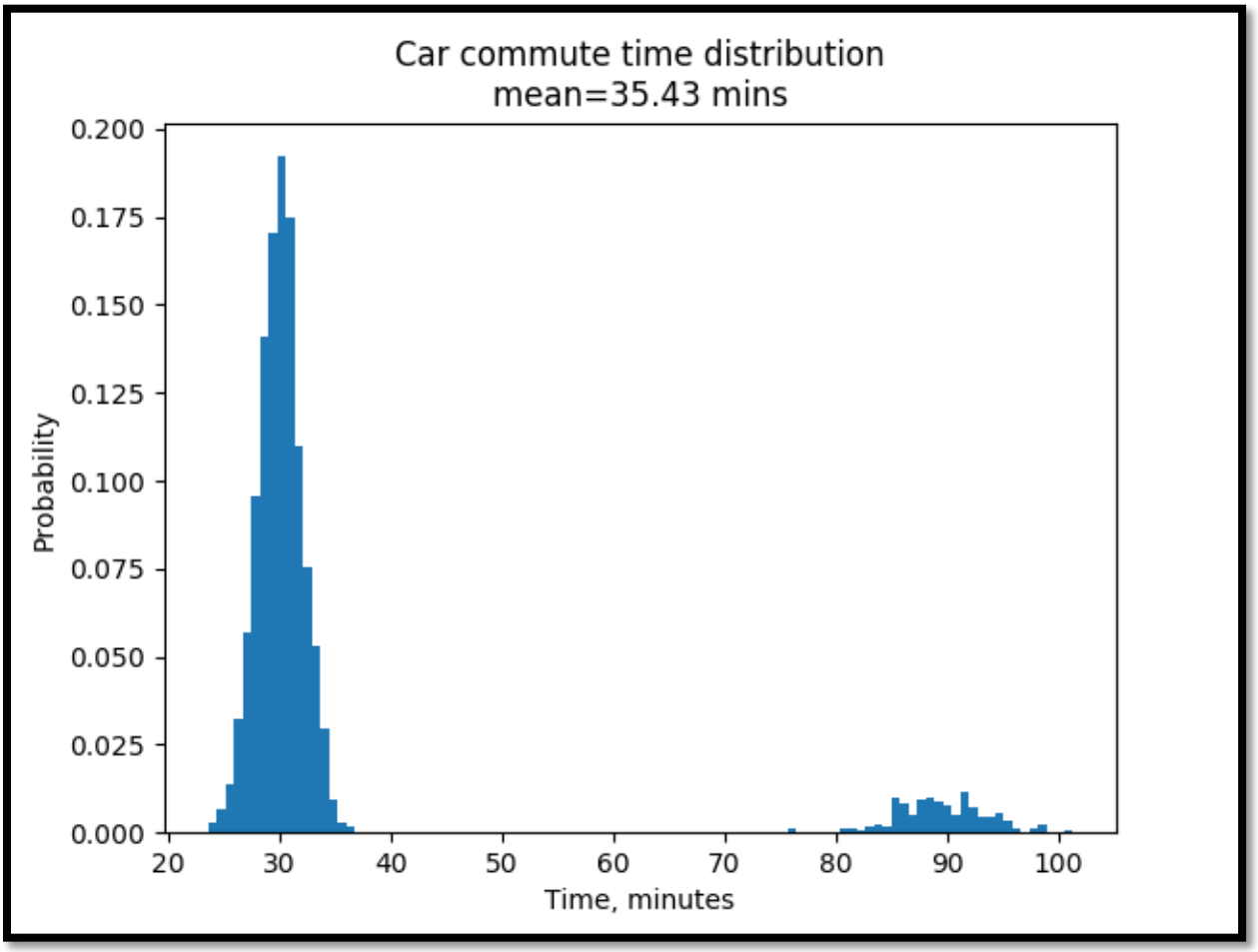

reward
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
400.0k
800.0k
1.200M
reward_100
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
400.0k
800.0k
1.200M

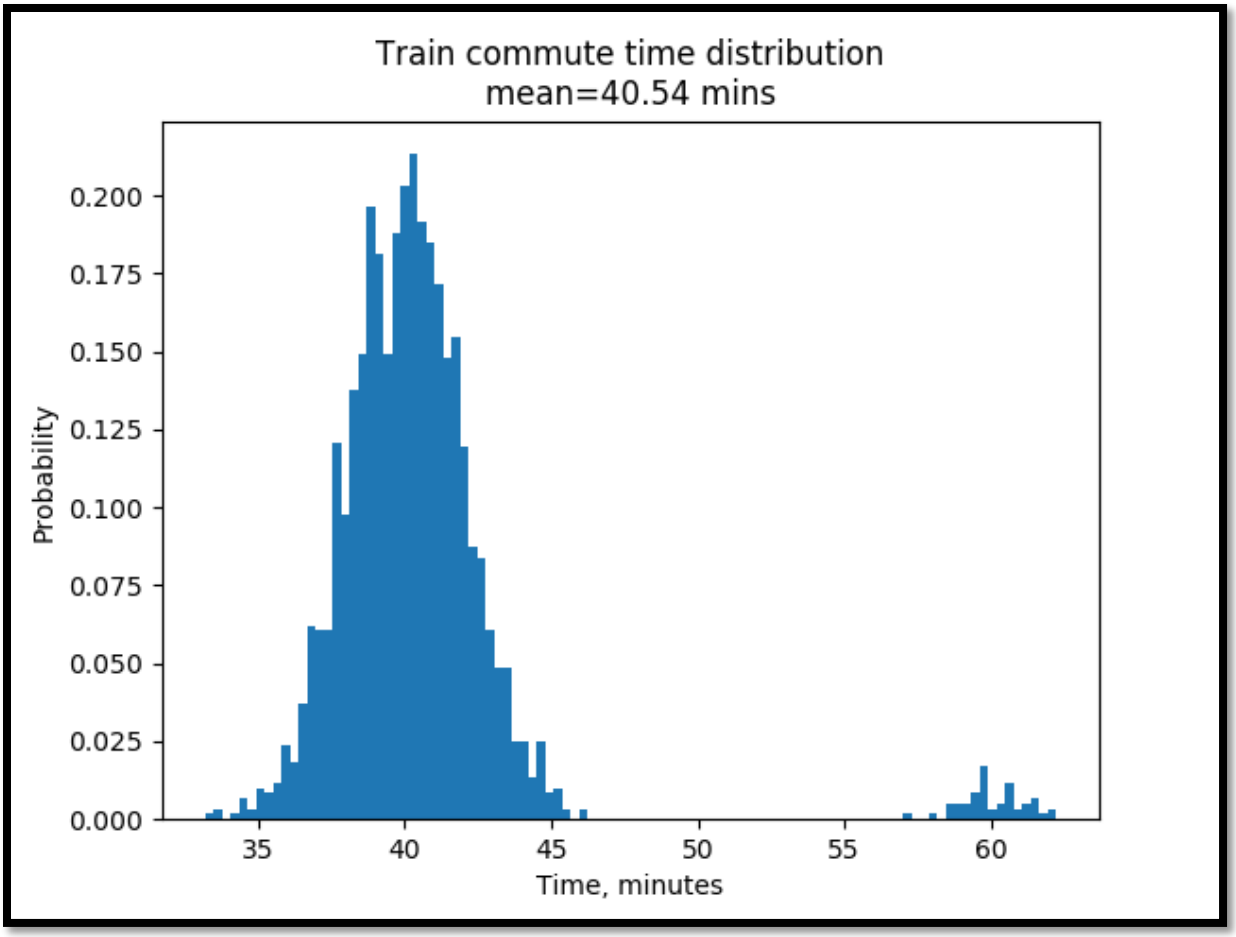

sigma_snr_layer_1
2.00
1.00
0.00
0.000
150.0k
300.0k
450.0k
sigma_snr_layer_2
16.0
12.0
8.00
4.00
0.00
0.000
150.0k
300.0k
450.0k

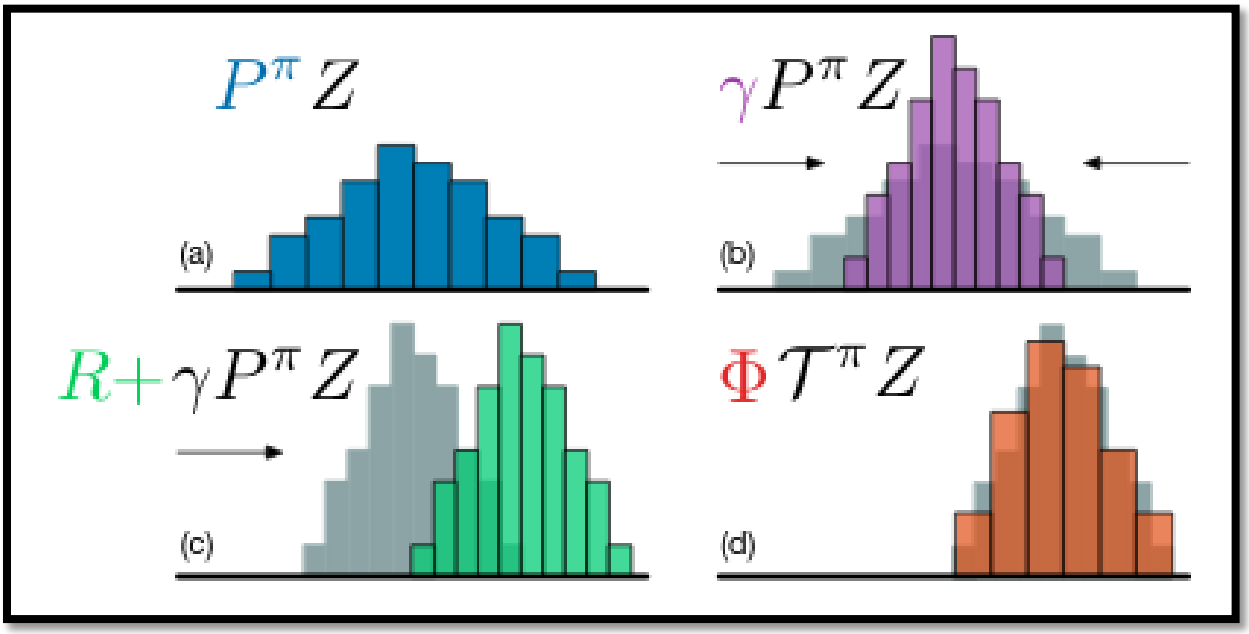

reward
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
400.0k
800.0k
1.200M
reward_100
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
400.0k
800.0k
1.200M

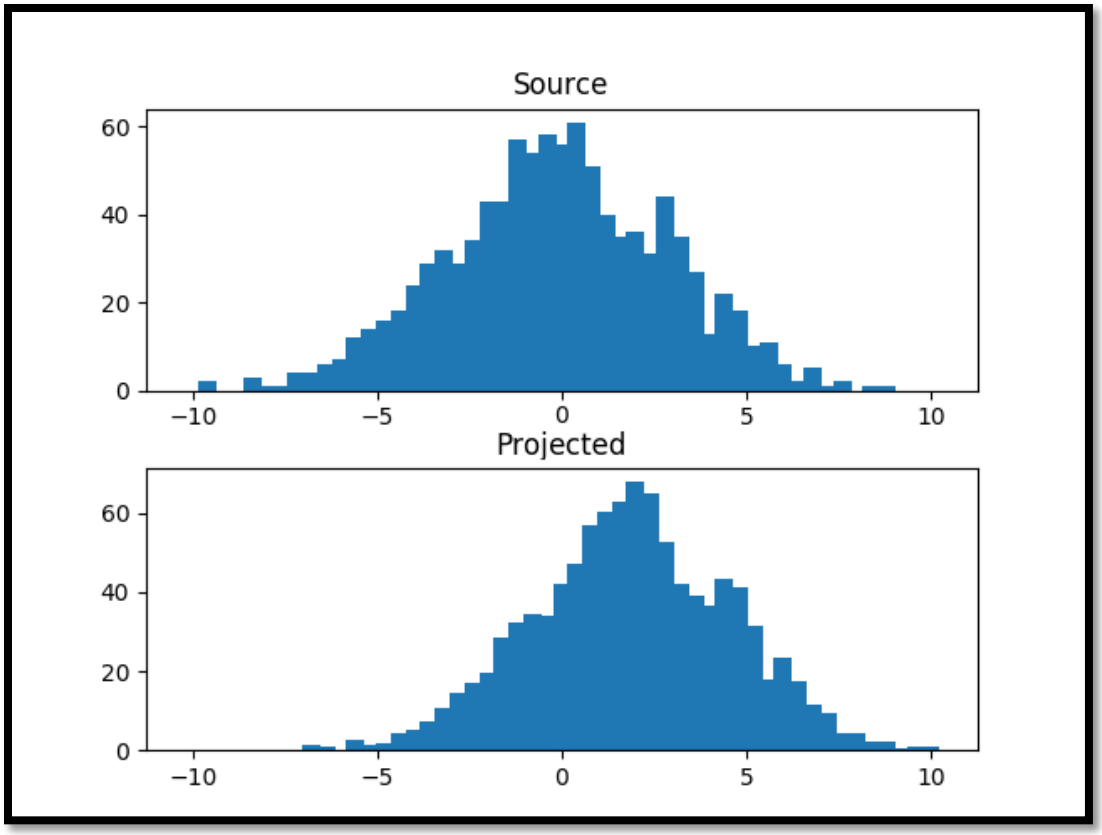


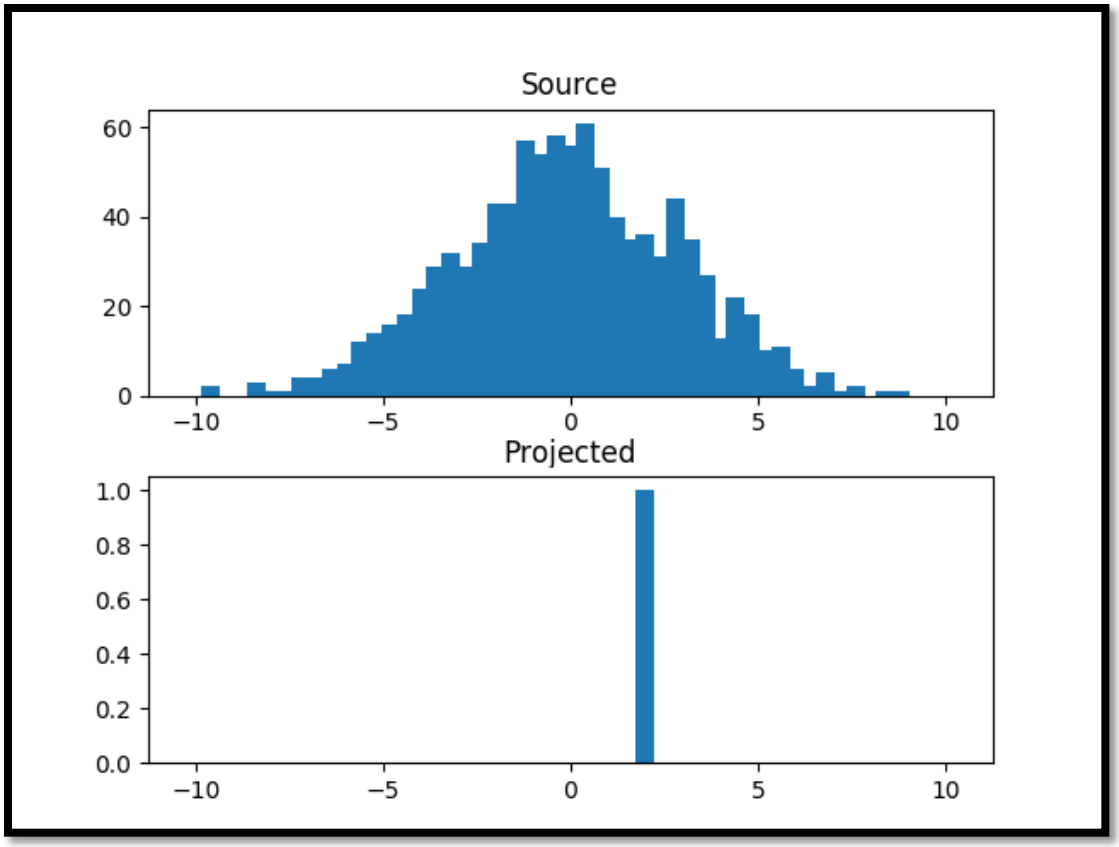

reward
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
400.0k
800.0k
1.200M
reward_100
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
400.0k
800.0k
1.200M

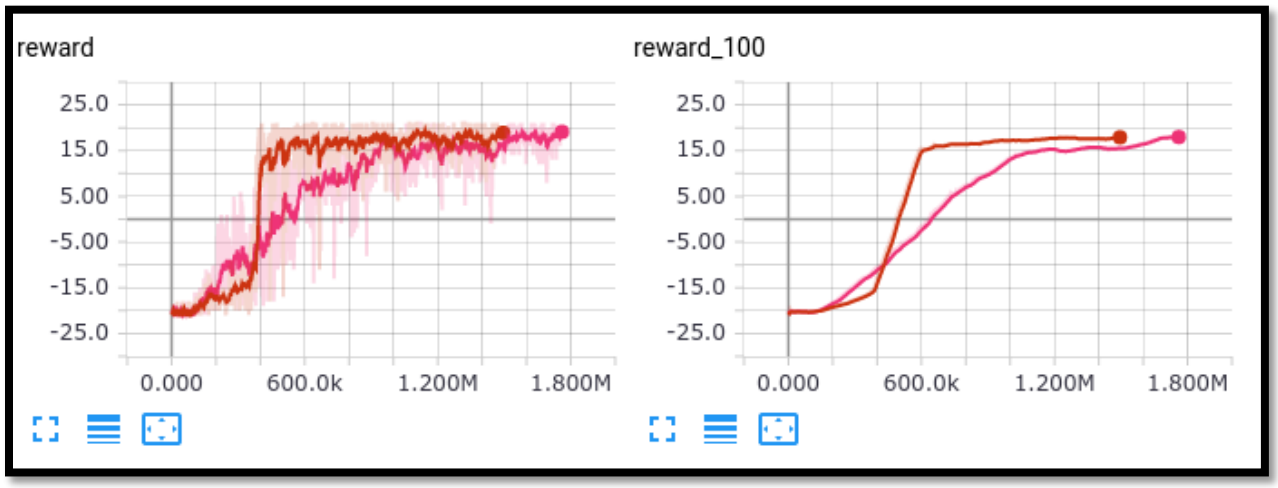

Car commute time distribution
mean=35.43 mins
Probability
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
20
30
40
50
60
70
80
90
100
Time, minutes

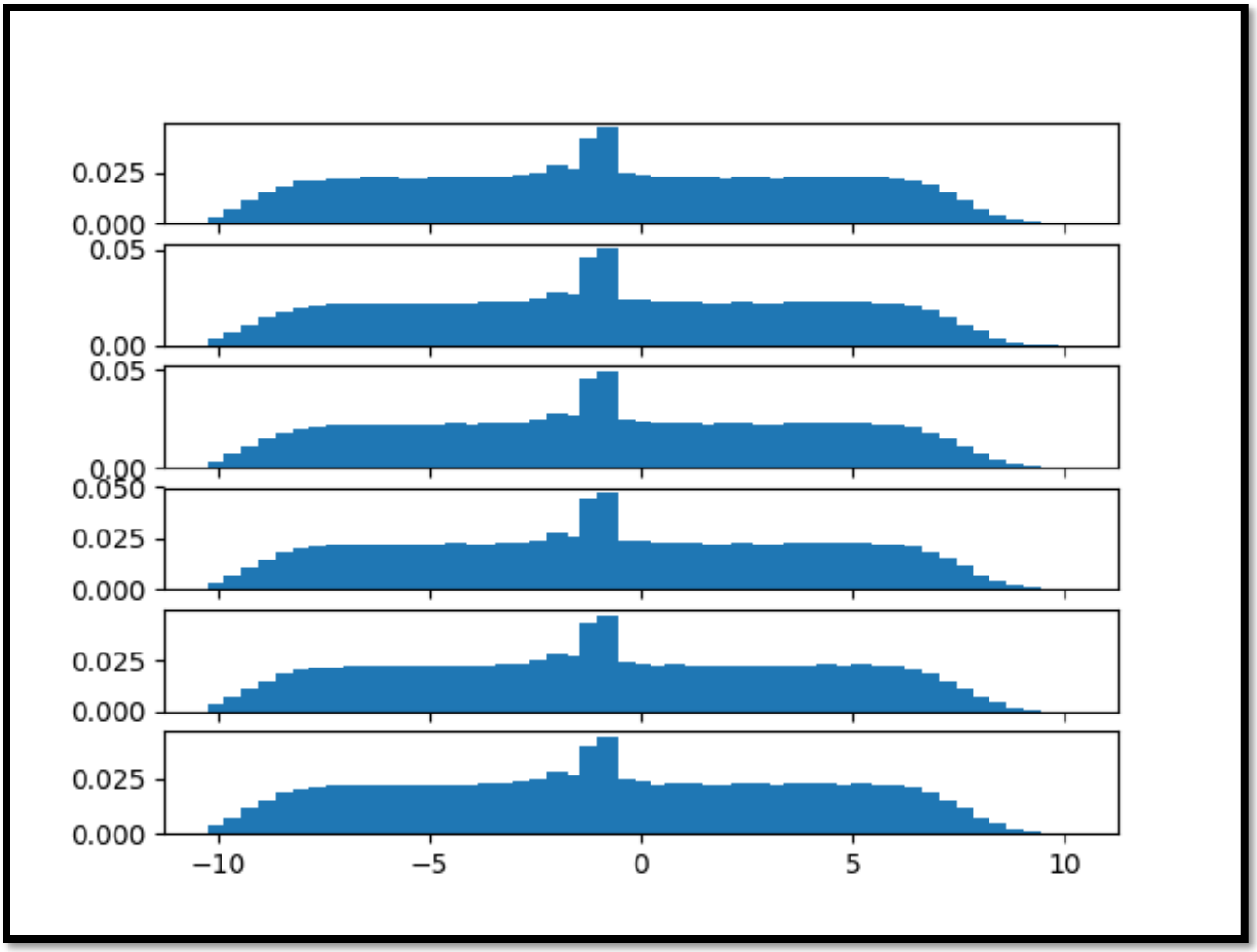

Train commute time distribution
mean=40.54 mins
Probability
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
35
40
45
50
55
60
Time, minutes

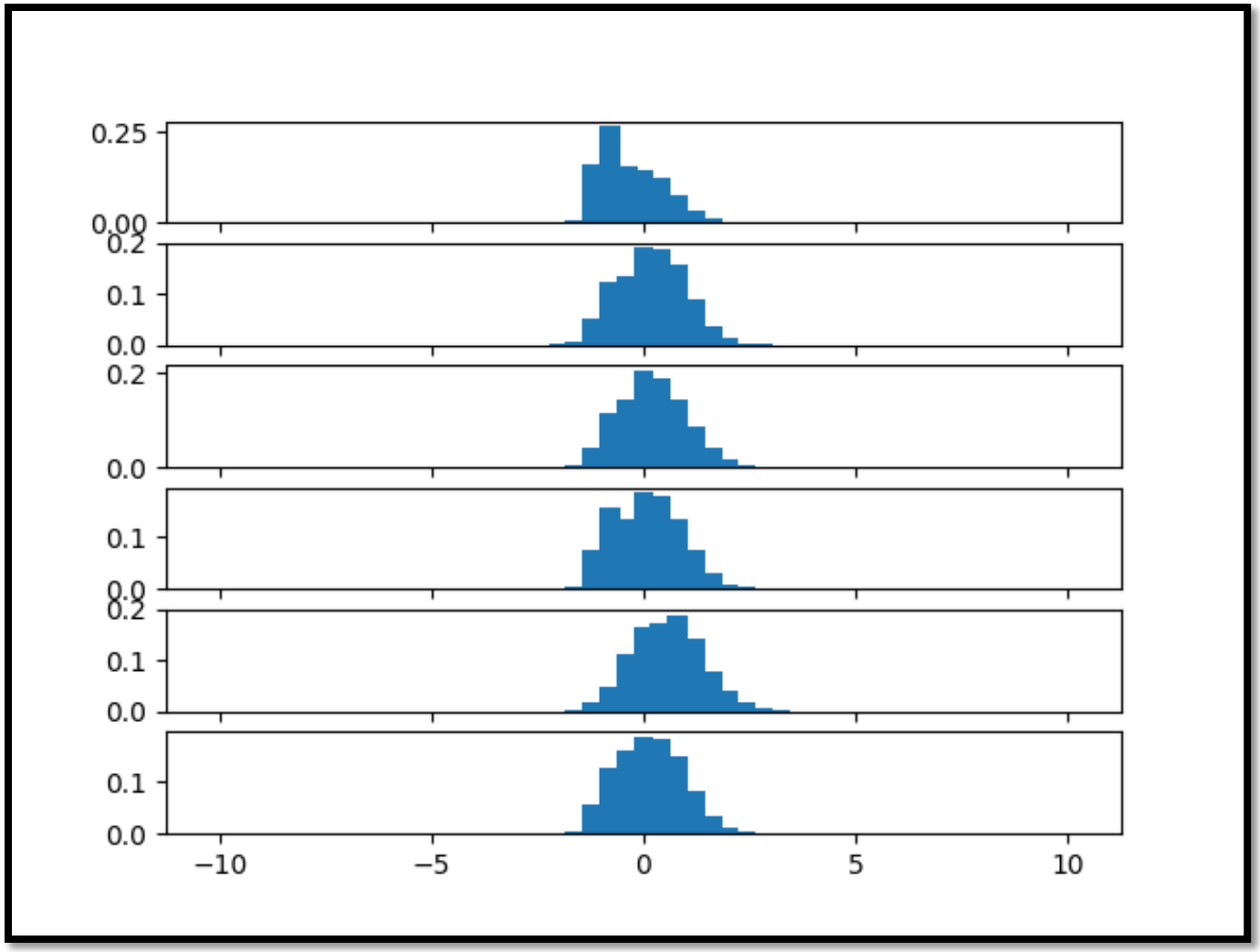

$P^\pi Z$
(a)
$\gamma P^\pi Z$
(b)
$R+\gamma P^\pi Z$
(c)
$\Phi \mathcal{T}^\pi Z$
(d)

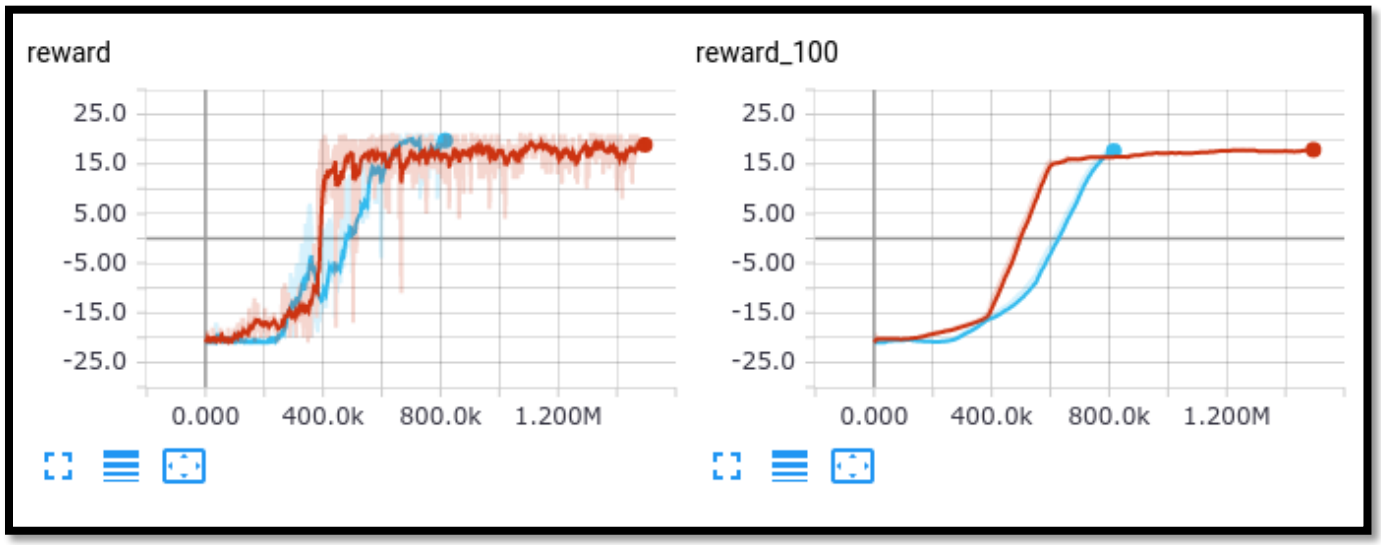

Source
60
40
20
0
−10
−5
0
5
10
Projected
60
40
20
0
−10
−5
0
5
10


Source
60
40
20
0
−10
−5
0
5
10
Projected
1.0
0.8
0.6
0.4
0.2
0.0
−10
−5
0
5
10


reward
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
600.0k
1.200M
1.800M
reward_100
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
600.0k
1.200M
1.800M


0.025
0.000
0.05
0.00
0.05
0.00
0.050
0.025
0.000
0.025
0.000
0.025
0.000
−10
−5
0
5
10


0.25
0.00
0.2
0.1
0.0
0.2
0.0
0.1
0.0
0.2
0.1
0.0
0.1
0.0
−10
−5
0
5
10


reward
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
400.0k
800.0k
1.200M
reward_100
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
400.0k
800.0k
1.200M

Chapter 8: Stocks Trading Using RL

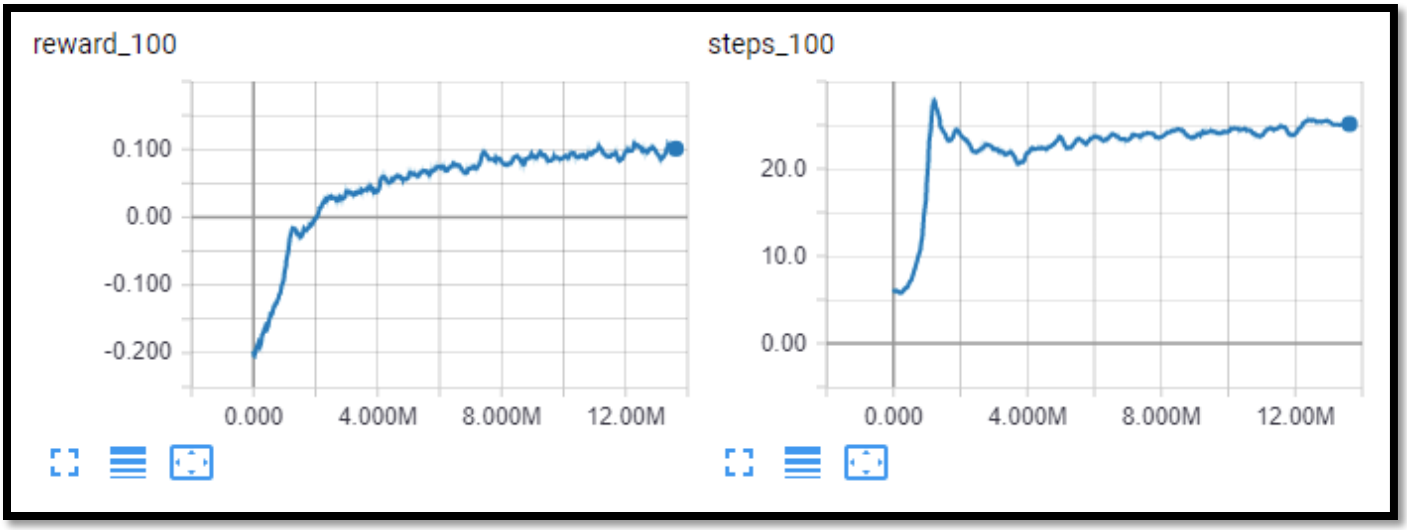
Data enconding: conv_1d=False, vector

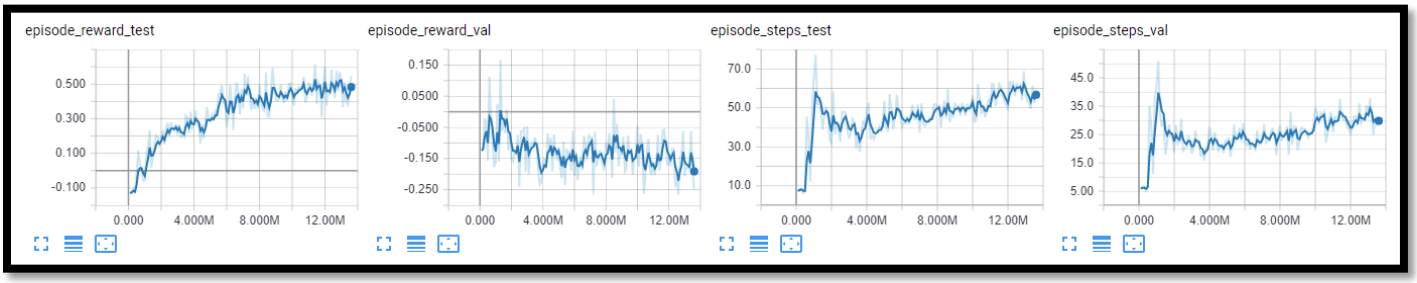
High	Low	Close	Volume	H	L	C	V	...

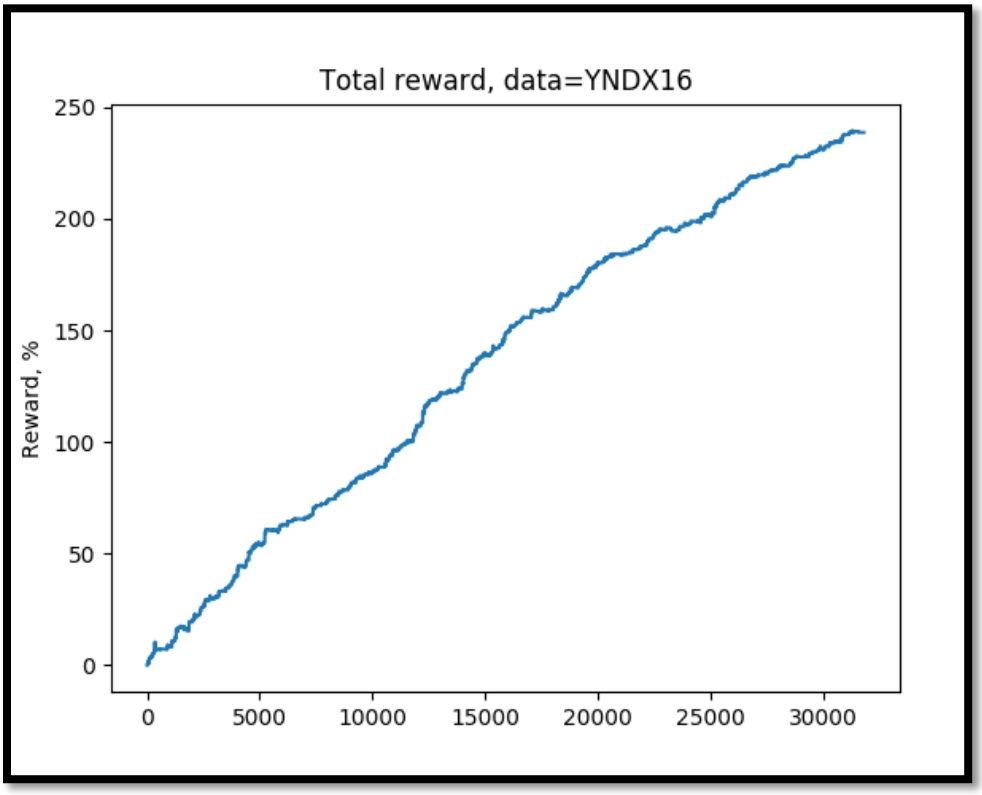
Bar 1 Bar 2

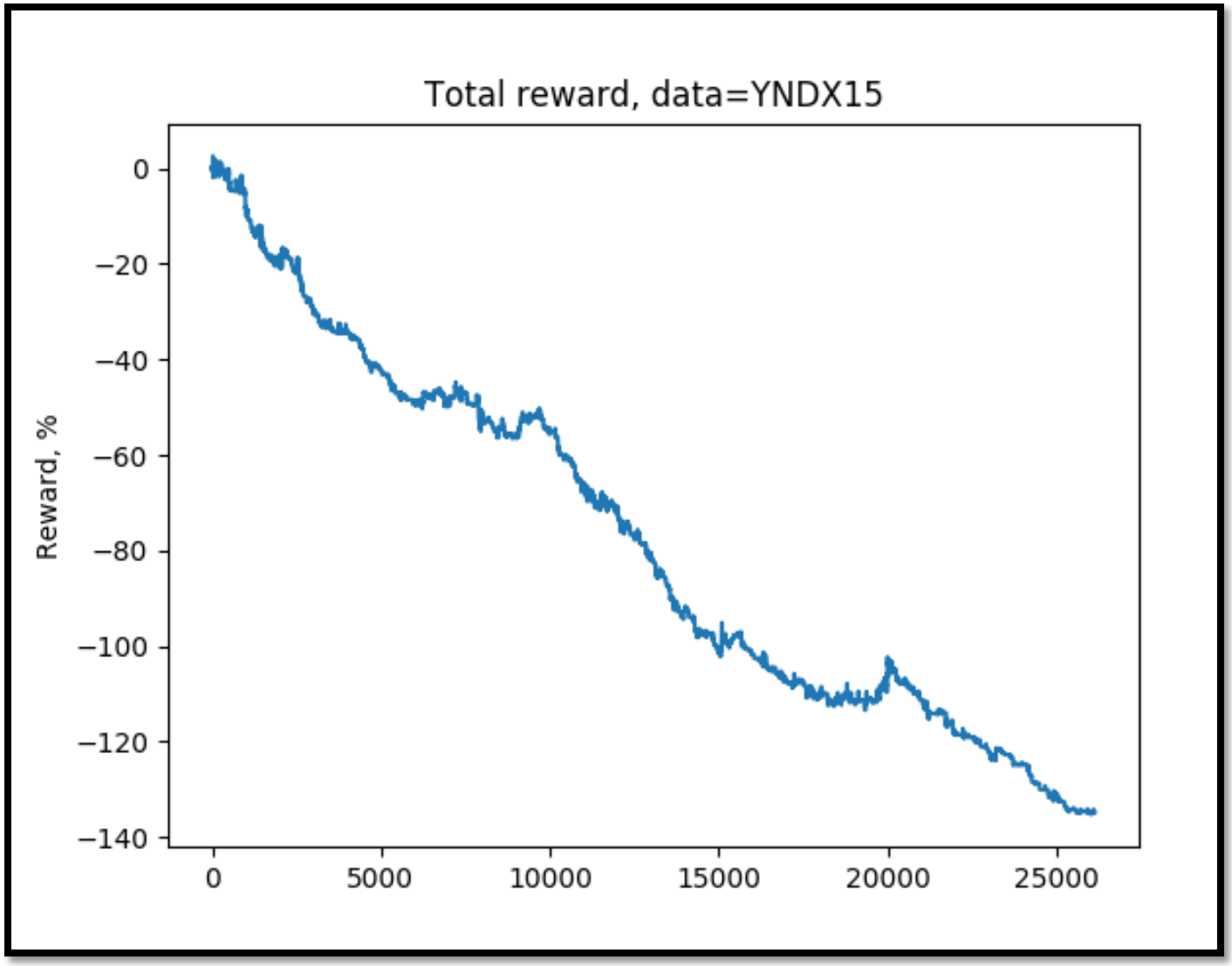
Data enconding: conv_1d=True, matrix

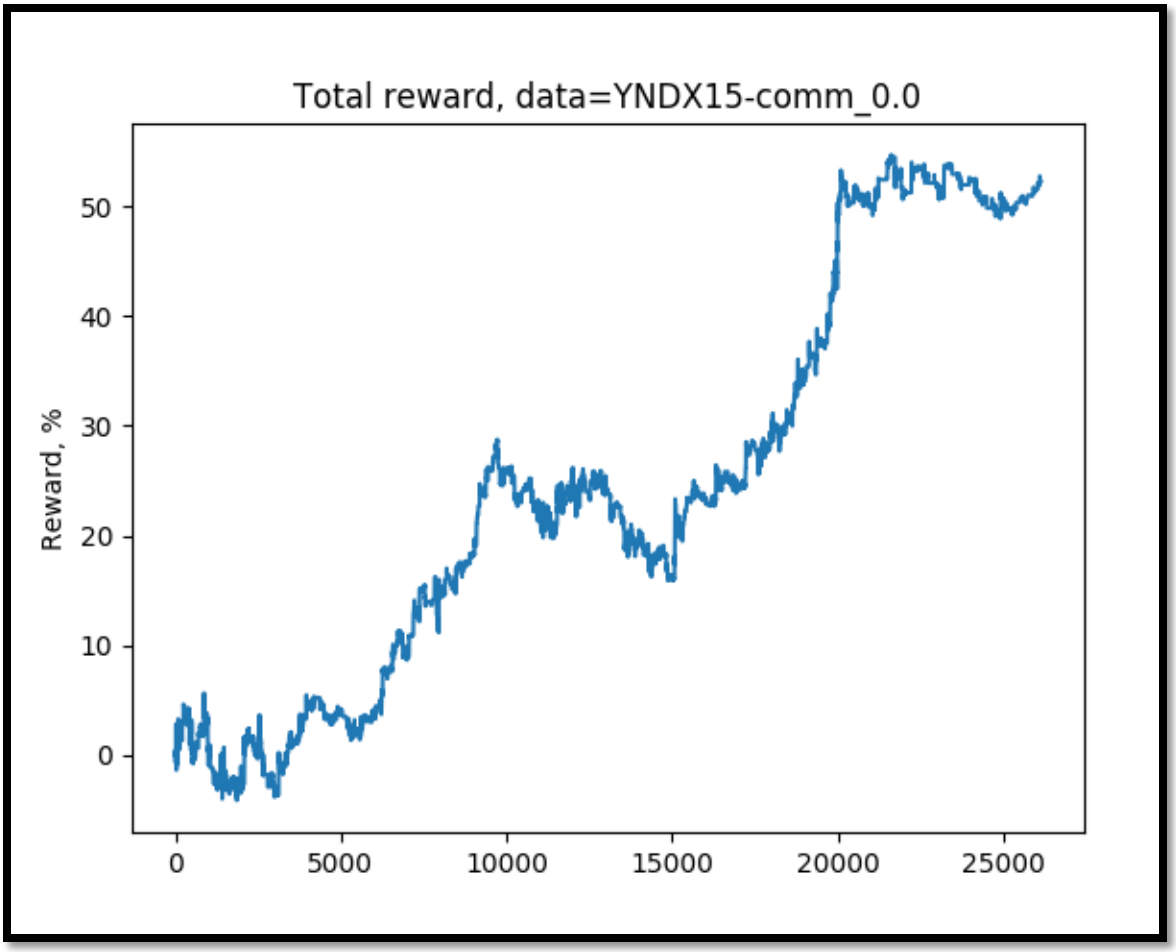
High prices for all bars
Low prices
Close prices
Volumes

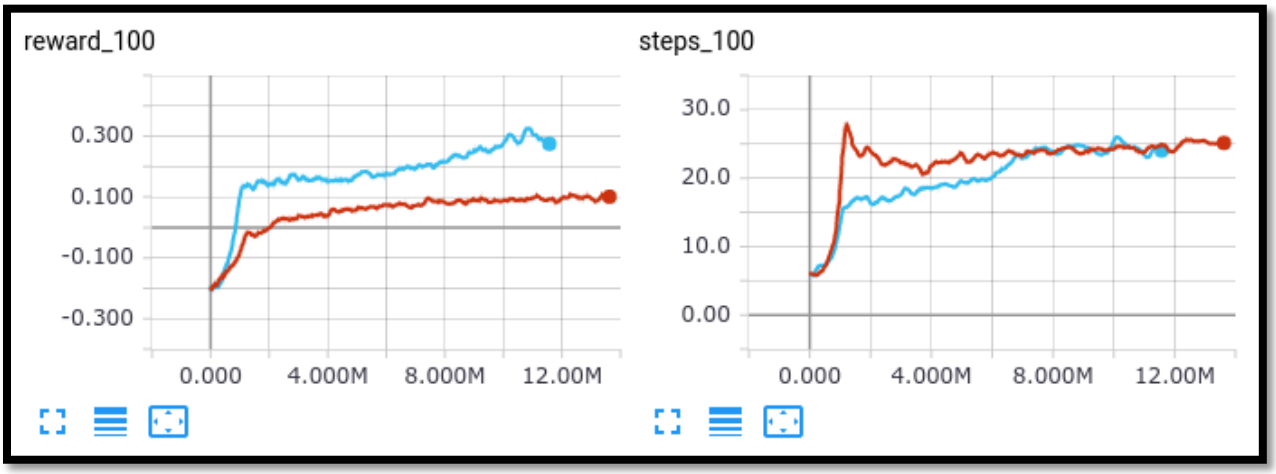

reward_100
0.100
0.00
-0.100
-0.200
0.000
4.000M
8.000M
12.00M
steps_100
20.0
10.0
0.00
0.000
4.000M
8.000M
12.00M

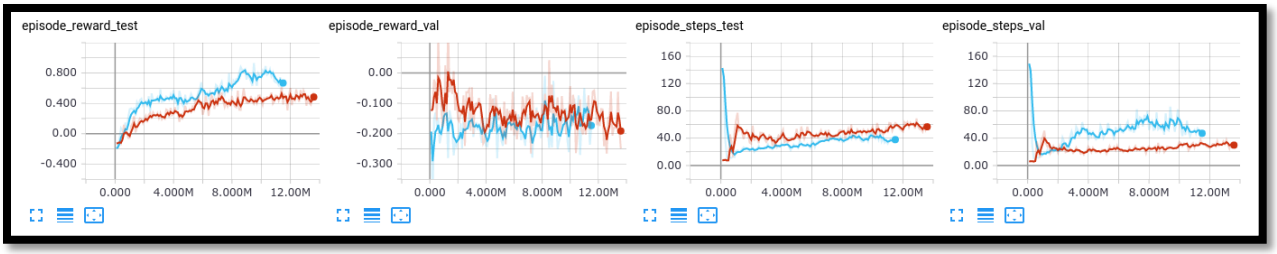

episode_reward_test
episode_reward_val
episode_steps_test
episode_steps_val

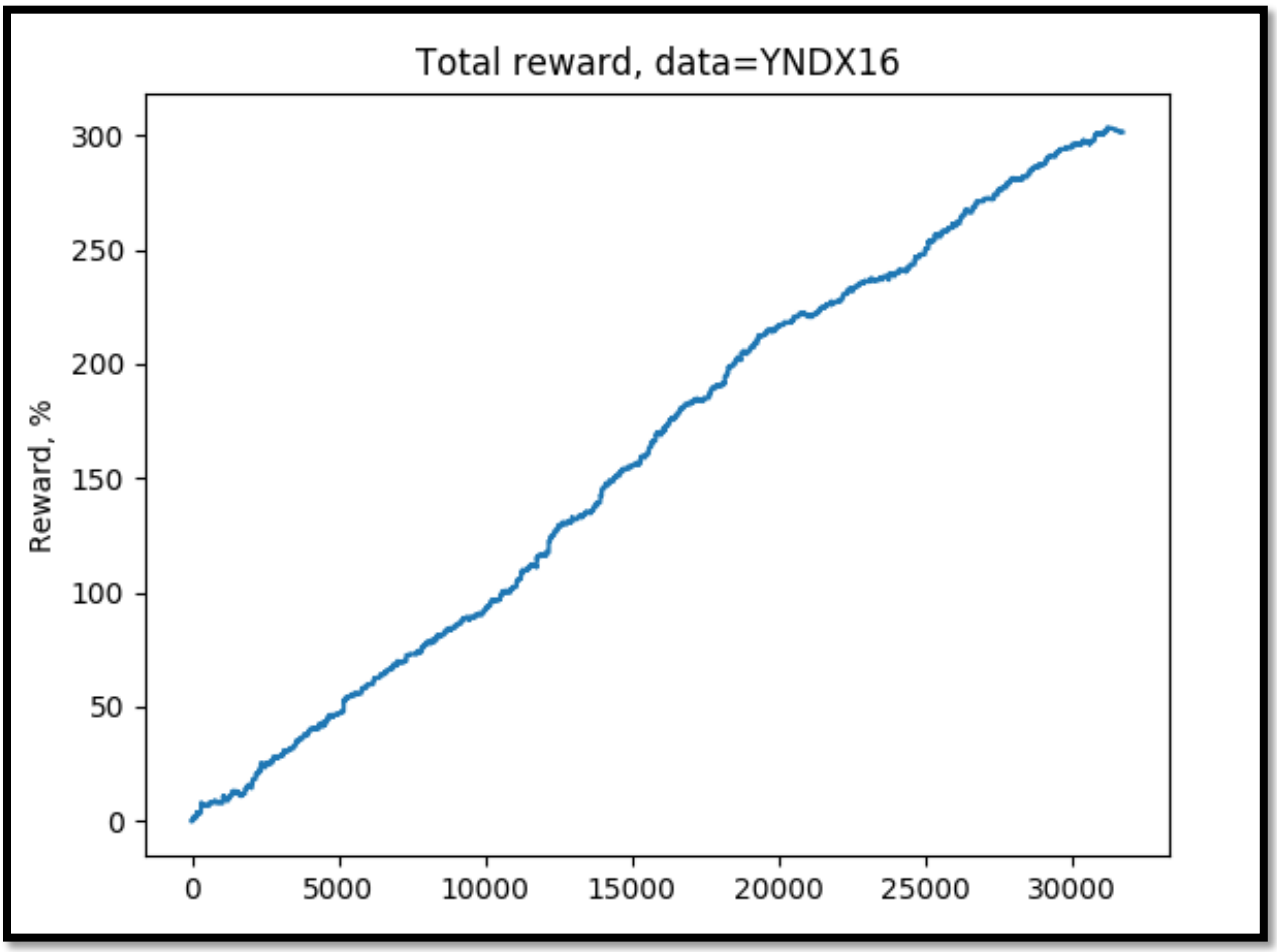

Total reward, data=YNDX16
Reward, %
250
200
150
100
50
0
0
5000
10000
15000
20000
25000
30000

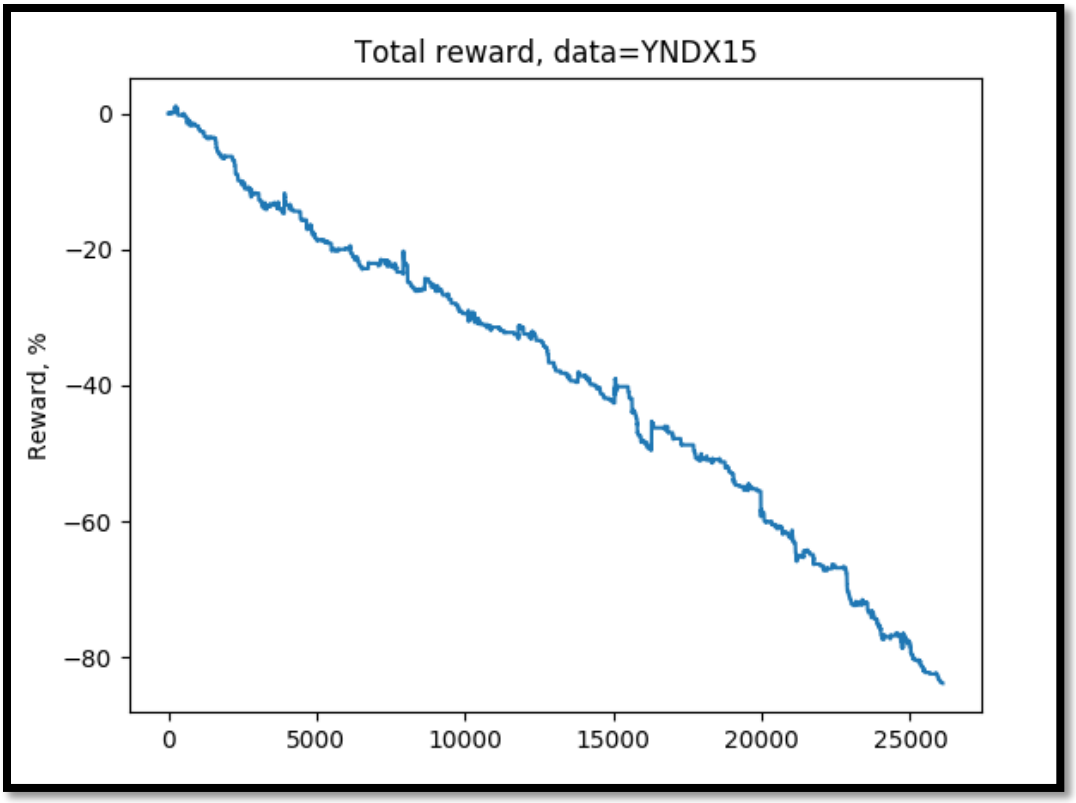

Total reward, data=YNDX15
Reward, %
0
−20
−40
−60
−80
−100
−120
−140
0
5000
10000
15000
20000
25000

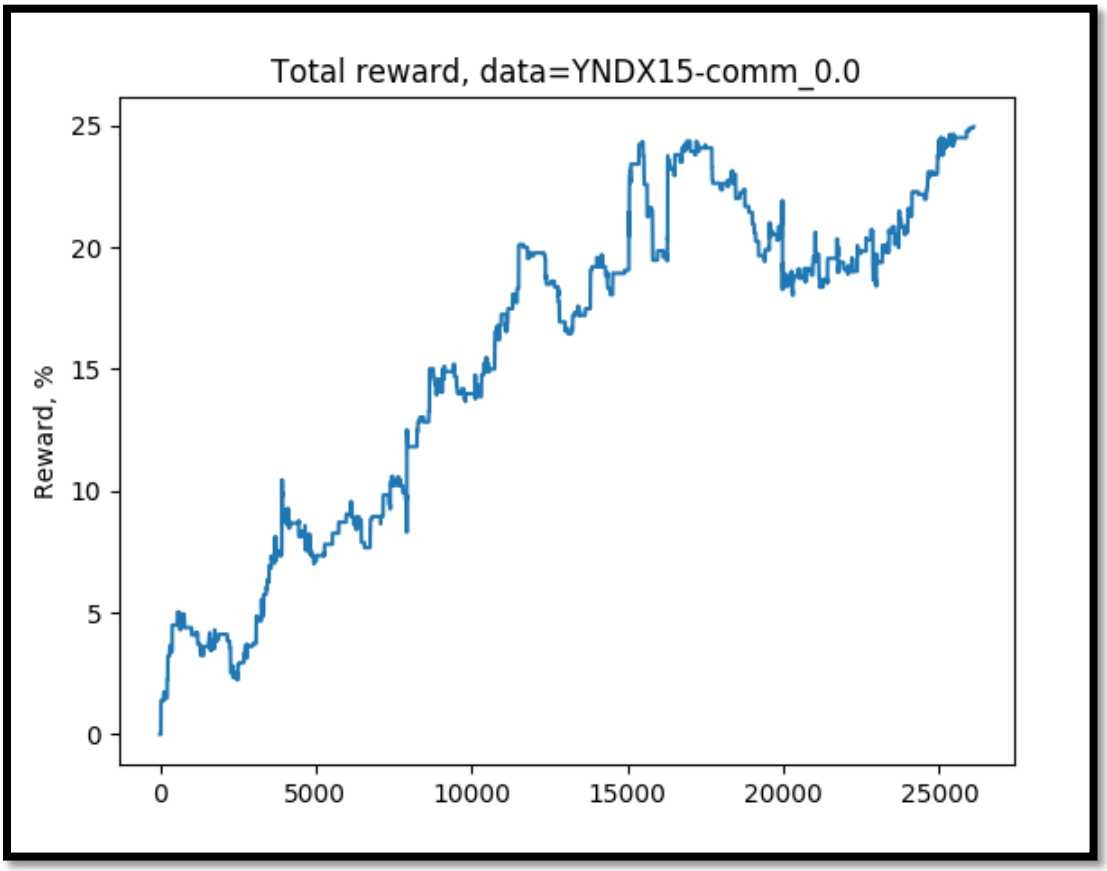

Total reward, data=YNDX15-comm_0.0
Reward, %
50
40
30
20
10
0
0
5000
10000
15000
20000
25000


reward_100
0.300
0.100
-0.100
-0.300
0.000
4.000M
8.000M
12.00M
steps_100
30.0
20.0
10.0
0.00
0.000
4.000M
8.000M
12.00M

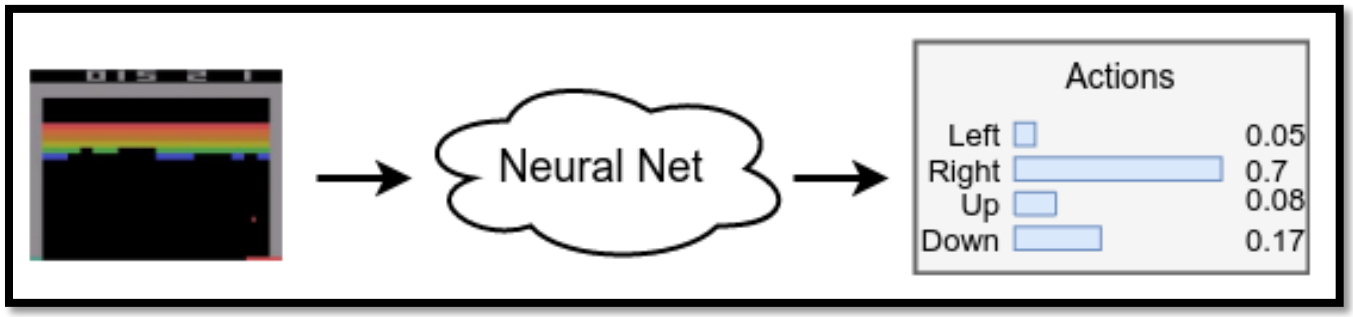

episode_reward_test
0.800
0.400
0.00
-0.400
0.000
4.000M
8.000M
12.00M
episode_reward_val
0.00
-0.100
-0.200
-0.300
0.000
4.000M
8.000M
12.00M
episode_steps_test
160
120
80.0
40.0
0.00
0.000
4.000M
8.000M
12.00M
episode_steps_val
160
120
80.0
40.0
0.00
0.000
4.000M
8.000M
12.00M

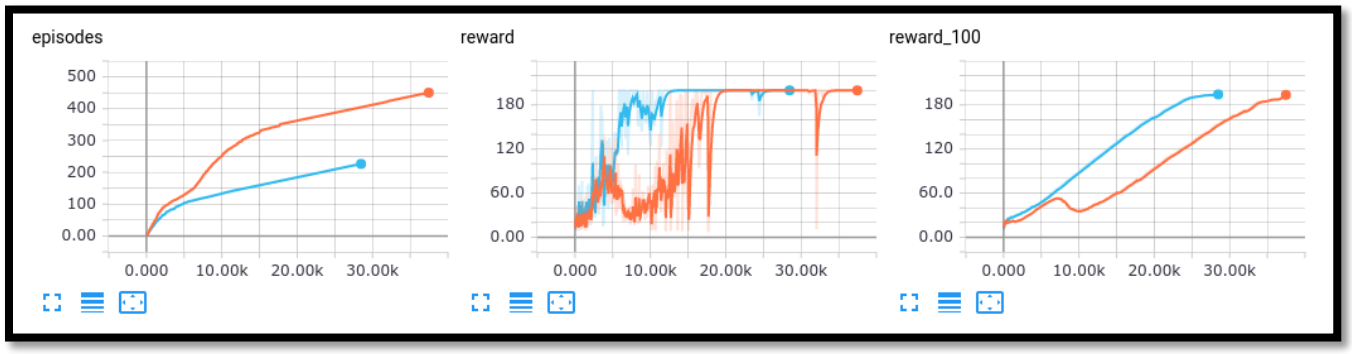

Total reward, data=YNDX16
Reward, %
300
250
200
150
100
50
0
0
5000
10000
15000
20000
25000
30000

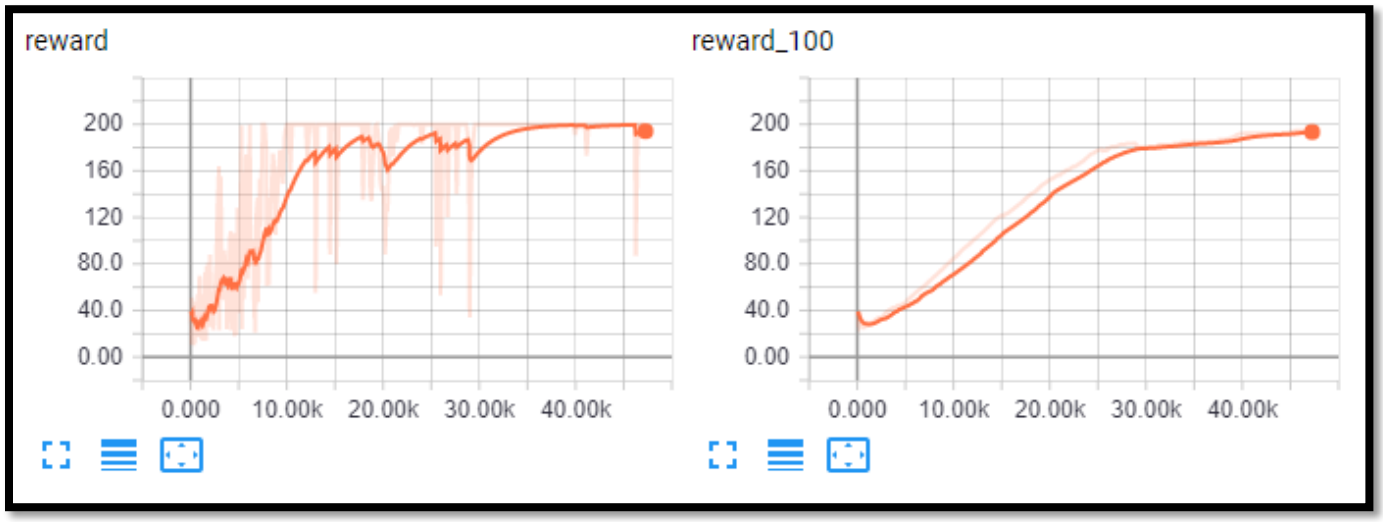

Total reward, data=YNDX15
Reward, %
0
−20
−40
−60
−80
0
5000
10000
15000
20000
25000

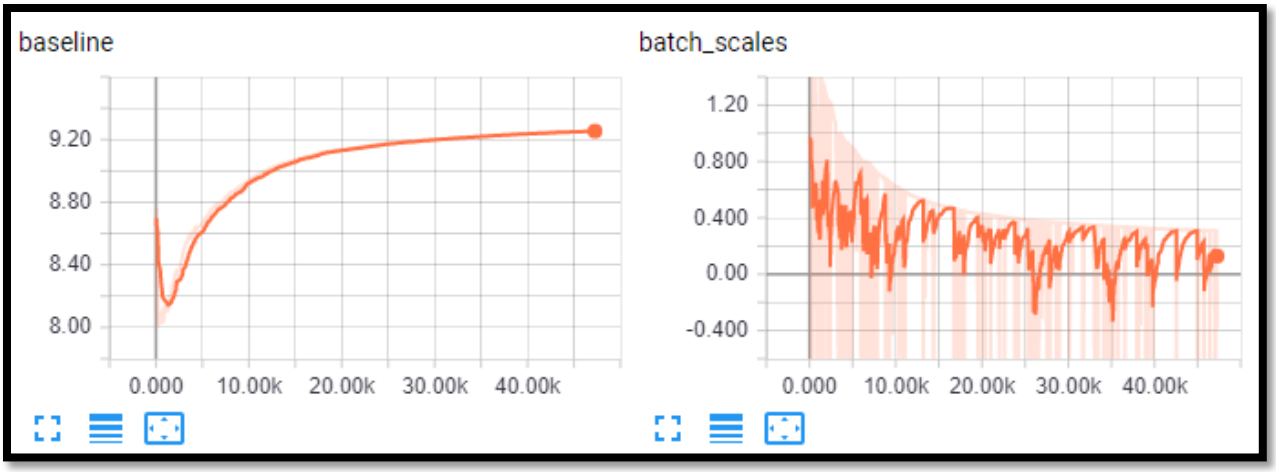

Total reward, data=YNDX15-comm_0.0
Reward, %
25
20
15
10
5
0
0
5000
10000
15000
20000
25000

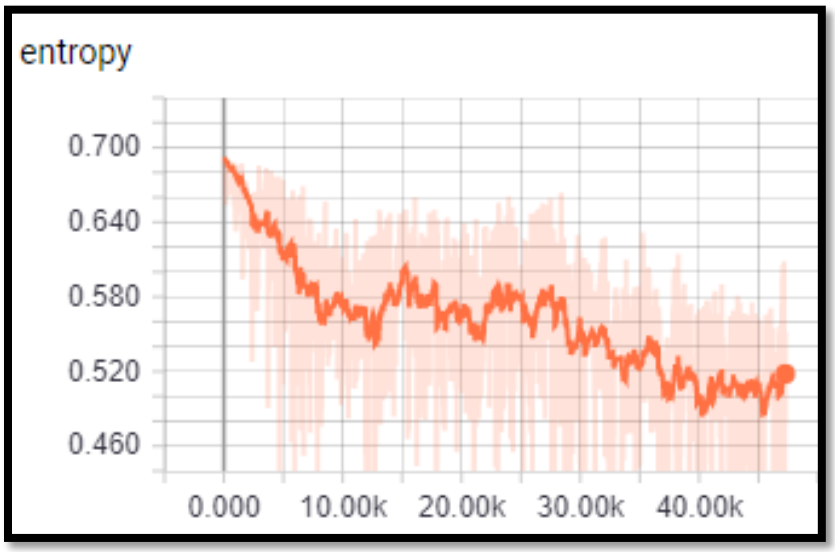
Chapter 9: Policy Gradients – An Alternative

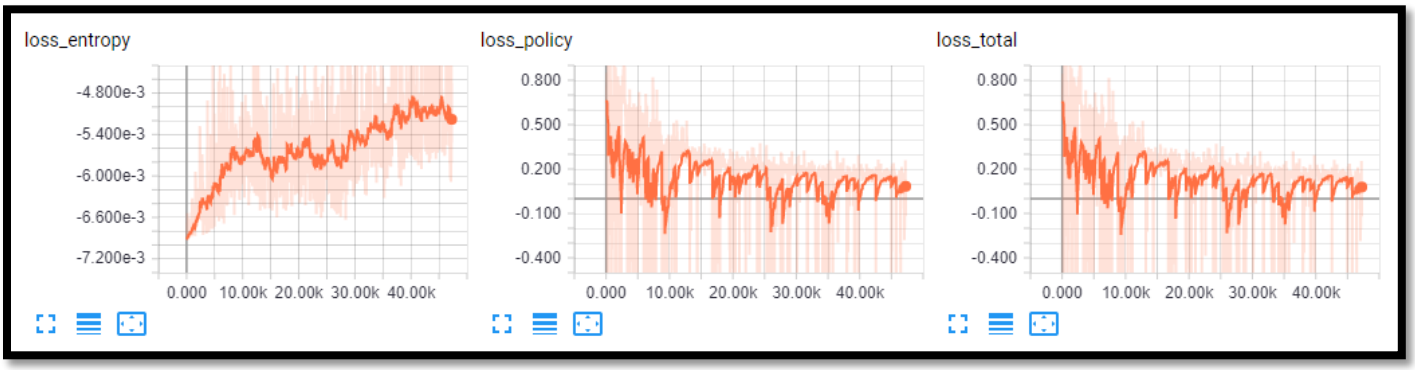


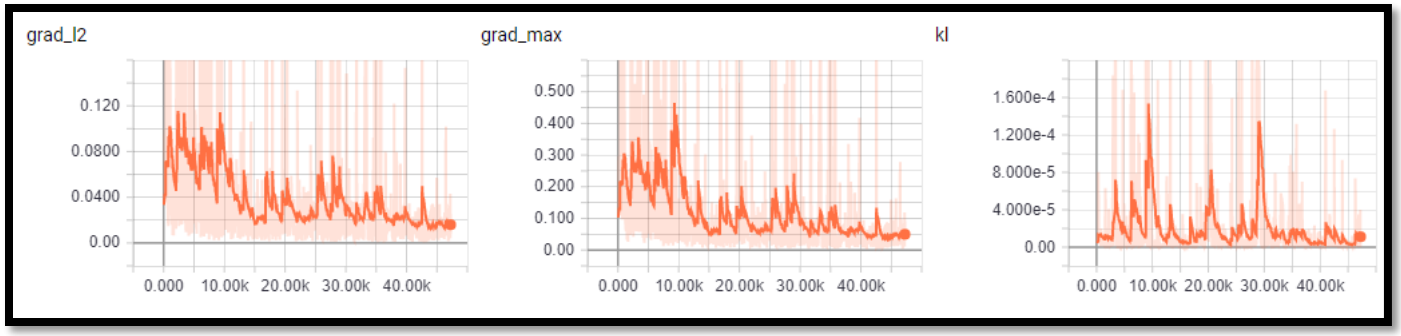




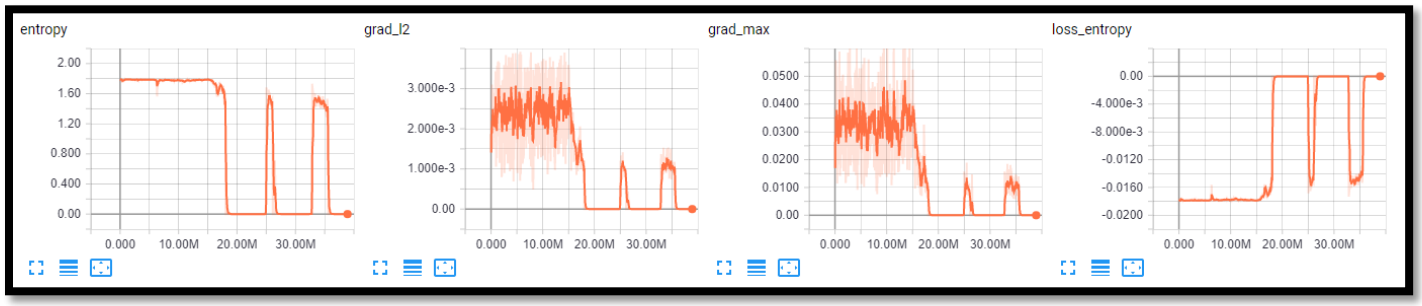

baseline
9.20
8.80
8.40
8.00
0.000
10.00k
20.00k
30.00k
40.00k
batch_scales
1.20
0.800
0.400
0.00
-0.400
0.000
10.00k
20.00k
30.00k
40.00k

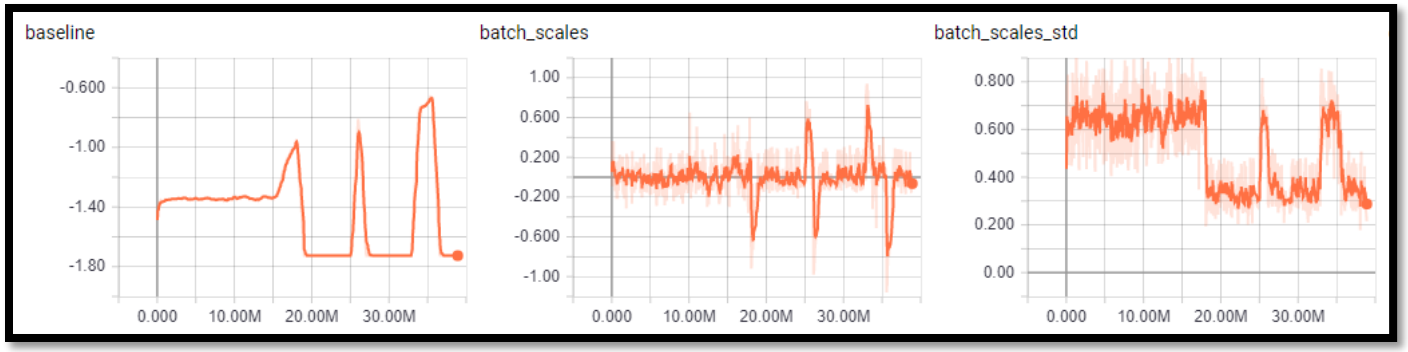

entropy
0.700
0.640
0.580
0.520
0.460
0.000
10.00k
20.00k
30.00k
40.00k

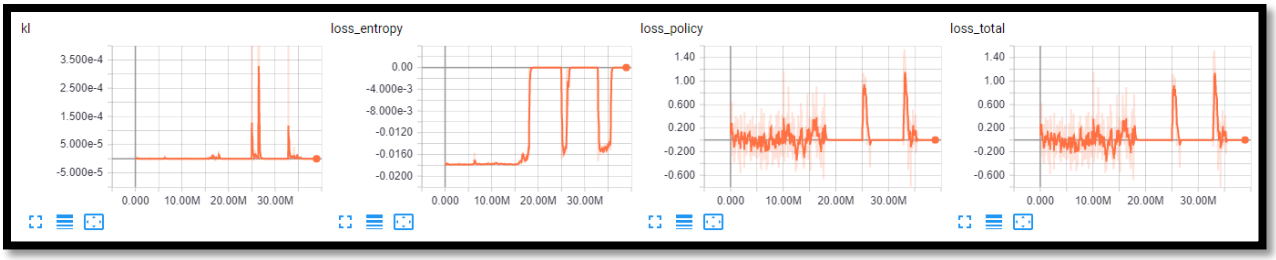

loss_entropy
-4.800e-3
-5.400e-3
-6.000e-3
-6.600e-3
-7.200e-3
0.000
10.00k
20.00k
30.00k
40.00k
loss_policy
0.800
0.500
0.200
-0.100
-0.400
0.000
10.00k
20.00k
30.00k
40.00k
loss_total
0.800
0.500
0.200
-0.100
-0.400
0.000
10.00k
20.00k
30.00k
40.00k


grad_l2
0.120
0.0800
0.0400
0.00
0.000 10.00k 20.00k 30.00k 40.00k
grad_max
0.500
0.400
0.300
0.200
0.100
0.00
0.000 10.00k 20.00k 30.00k 40.00k
kl
1.600e-4
1.200e-4
8.000e-5
4.000e-5
0.00
0.000 10.00k 20.00k 30.00k 40.00k

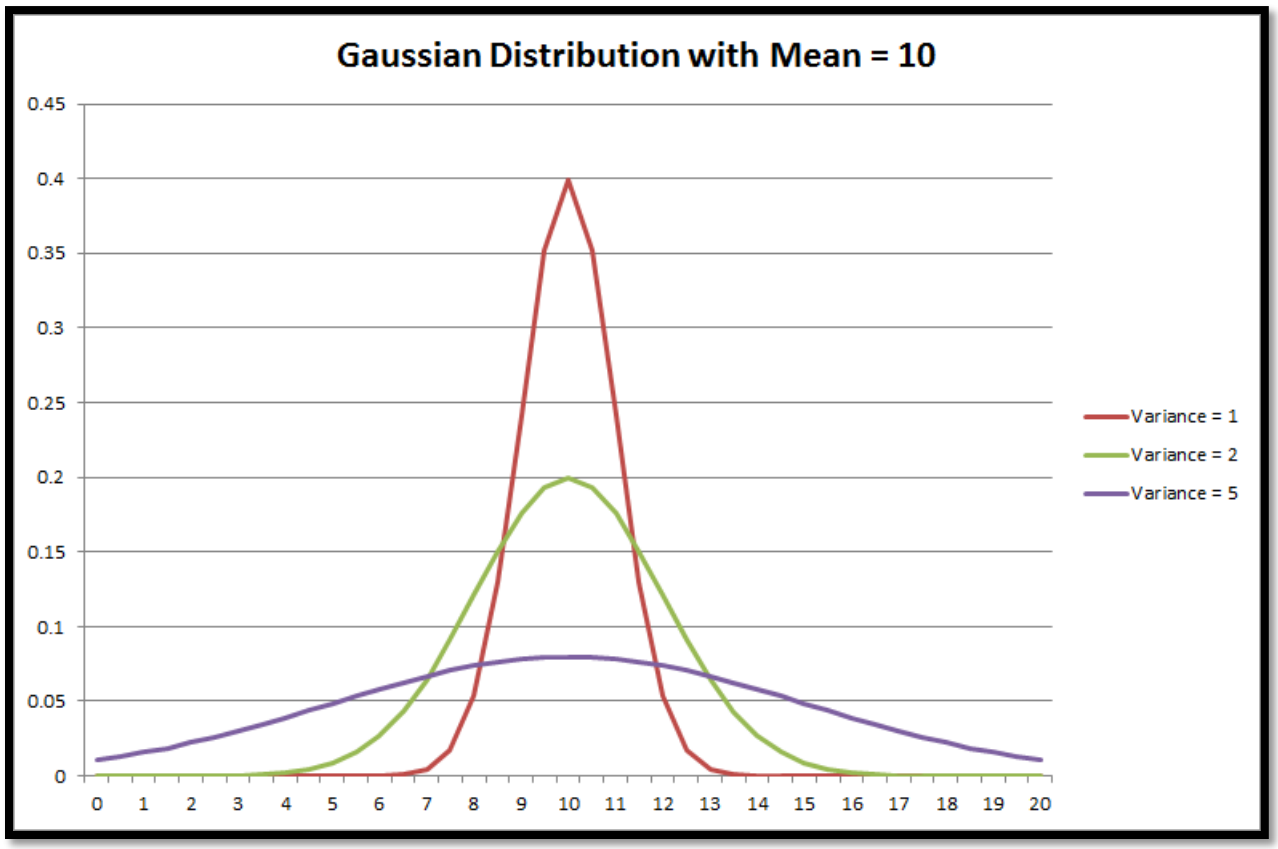

reward
-16.0
-18.0
-20.0
-22.0
0.000 10.00M 20.00M 30.00M
reward_100
-17.5
-18.5
-19.5
-20.5
-21.5
0.000 10.00M 20.00M 30.00M

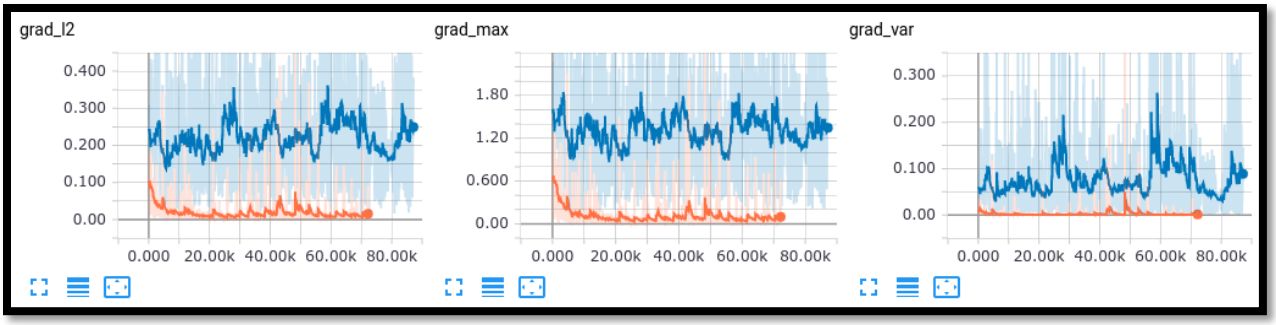

entropy
2.00
1.60
1.20
0.800
0.400
0.00
0.000 10.00M 20.00M 30.00M
grad_l2
3.000e-3
2.000e-3
1.000e-3
0.00
0.000 10.00M 20.00M 30.00M
grad_max
0.0500
0.0400
0.0300
0.0200
0.0100
0.00
0.000 10.00M 20.00M 30.00M
loss_entropy
0.00
-4.000e-3
-8.000e-3
-0.0120
-0.0160
-0.0200
0.000 10.00M 20.00M 30.00M

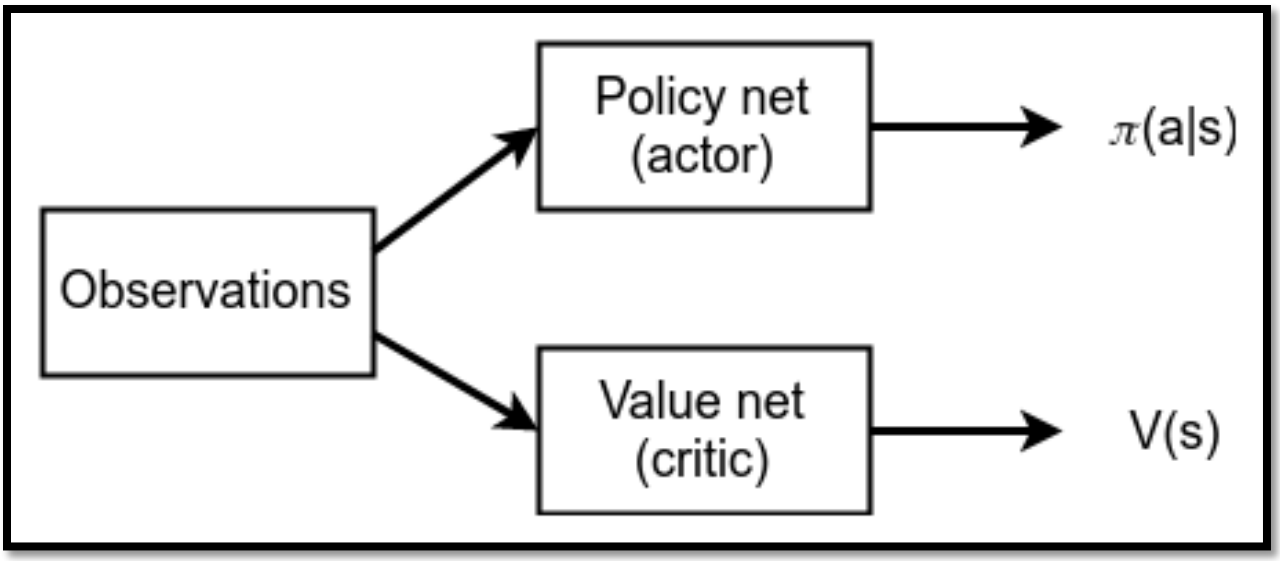

baseline
-0.600
-1.00
-1.40
-1.80
0.000 10.00M 20.00M 30.00M
batch_scales
1.00
0.600
0.200
-0.200
-0.600
-1.00
0.000 10.00M 20.00M 30.00M
batch_scales_std
0.800
0.600
0.400
0.200
0.00
0.000 10.00M 20.00M 30.00M

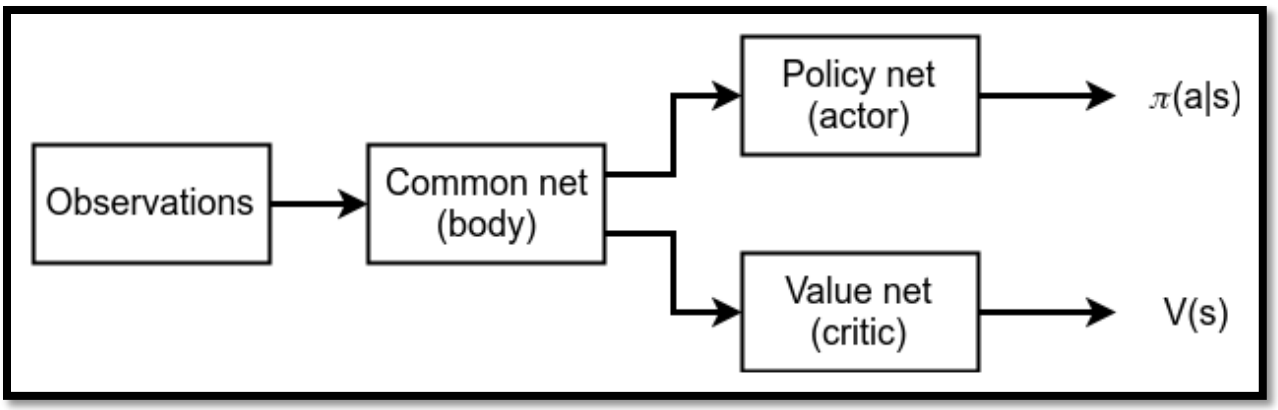

kl
3.500e-4
2.500e-4
1.500e-4
5.000e-5
-5.000e-5
0.000
10.00M
20.00M
30.00M
loss_entropy
0.00
-4.000e-3
-8.000e-3
-0.0120
-0.0160
-0.0200
0.000
10.00M
20.00M
30.00M
loss_policy
1.40
1.00
0.600
0.200
-0.200
-0.600
0.000
10.00M
20.00M
30.00M
loss_total
1.40
1.00
0.600
0.200
-0.200
-0.600
0.000
10.00M
20.00M
30.00M

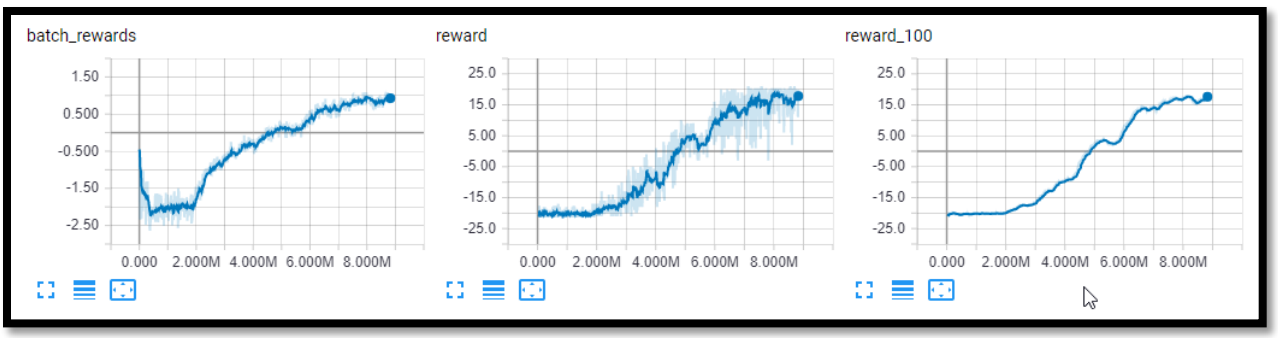
Chapter 10: The Actor-Critic Method

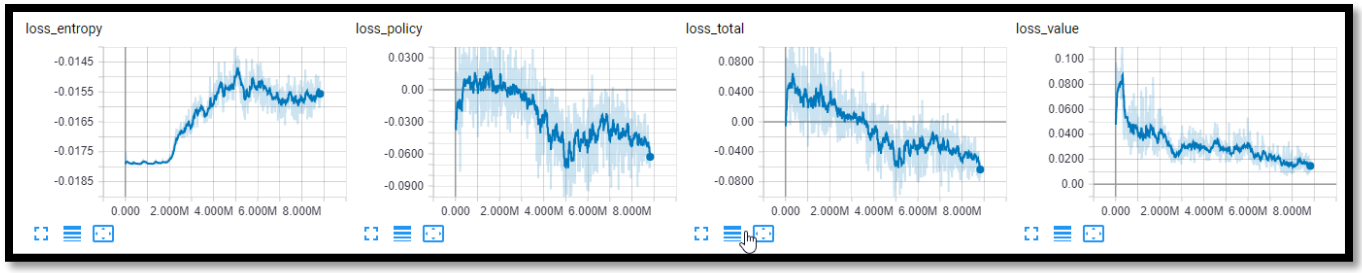


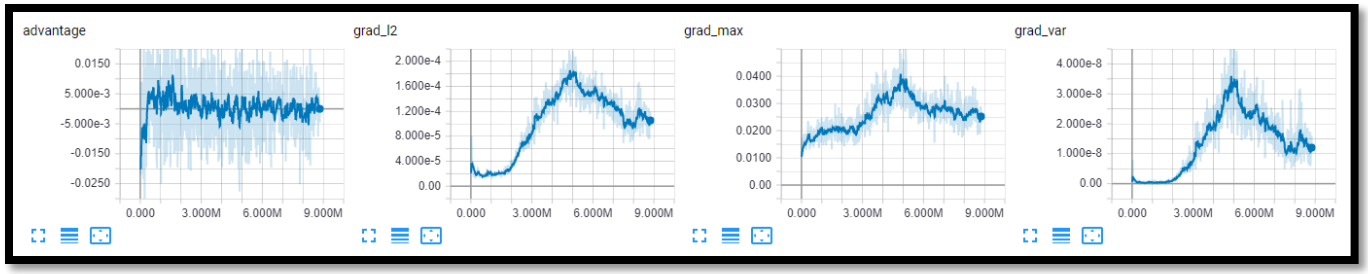

grad_l2
0.400
0.300
0.200
0.100
0.00
0.000 20.00k 40.00k 60.00k 80.00k
grad_max
1.80
1.20
0.600
0.00
0.000 20.00k 40.00k 60.00k 80.00k
grad_var
0.300
0.200
0.100
0.00
0.000 20.00k 40.00k 60.00k 80.00k

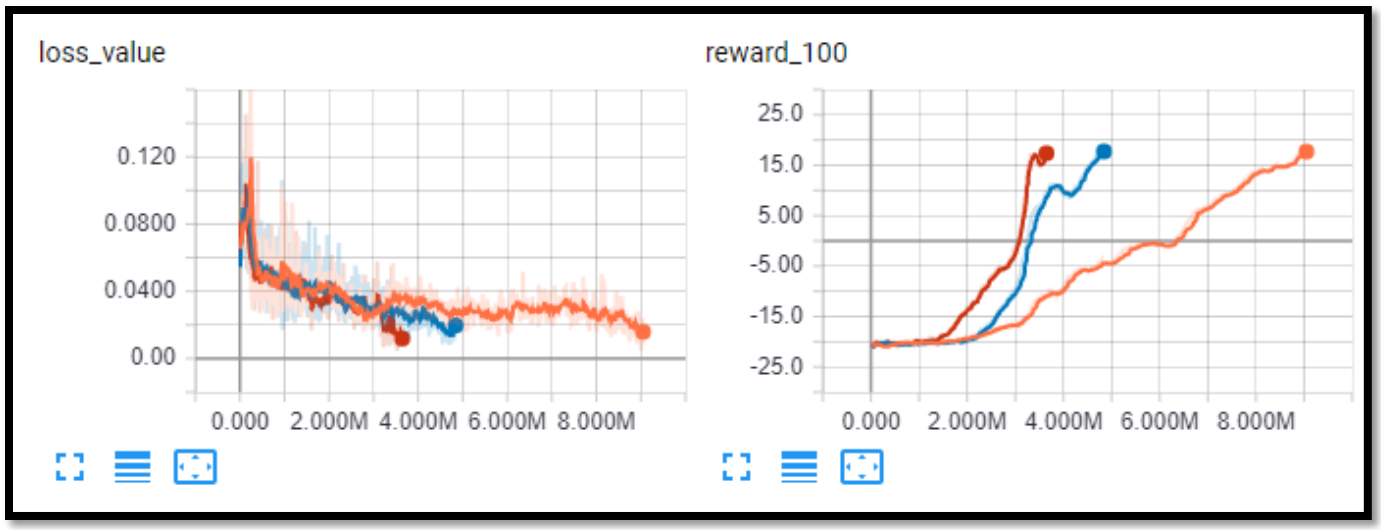

Observations
Policy net (actor)
Value net (critic)
π(a|s)
V(s)


Observations
Common net (body)
Policy net (actor)
Value net (critic)
π(a|s)
V(s)

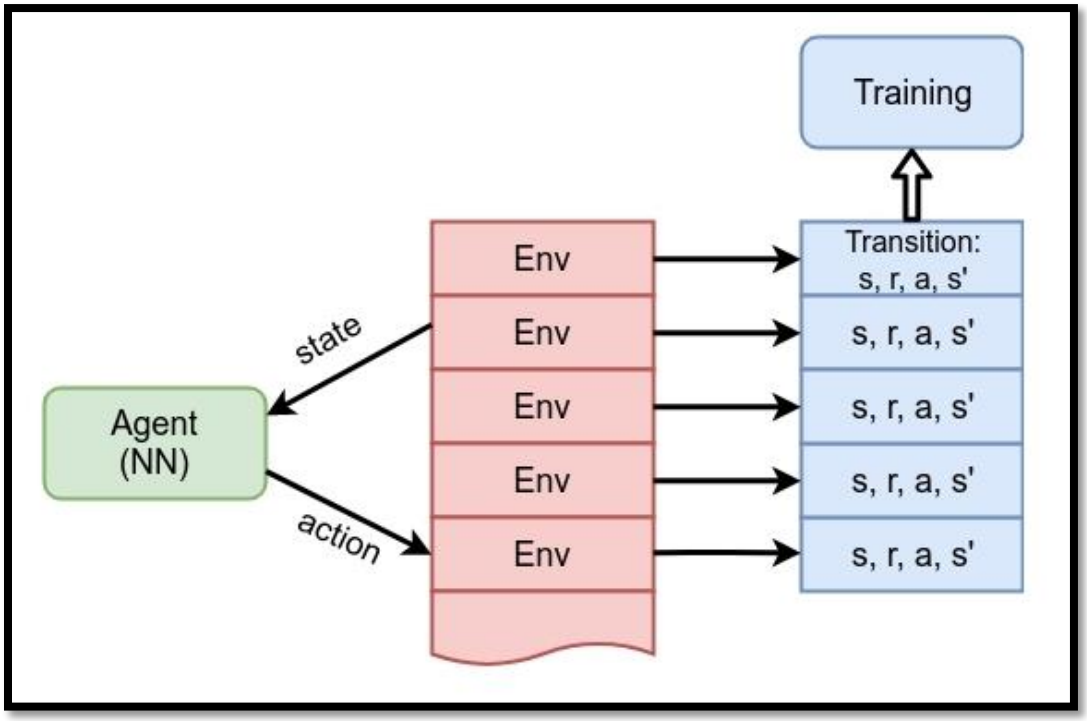

batch_rewards
1.50
0.500
-0.500
-1.50
-2.50
0.000 2.000M 4.000M 6.000M 8.000M
reward
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000 2.000M 4.000M 6.000M 8.000M
reward_100
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000 2.000M 4.000M 6.000M 8.000M

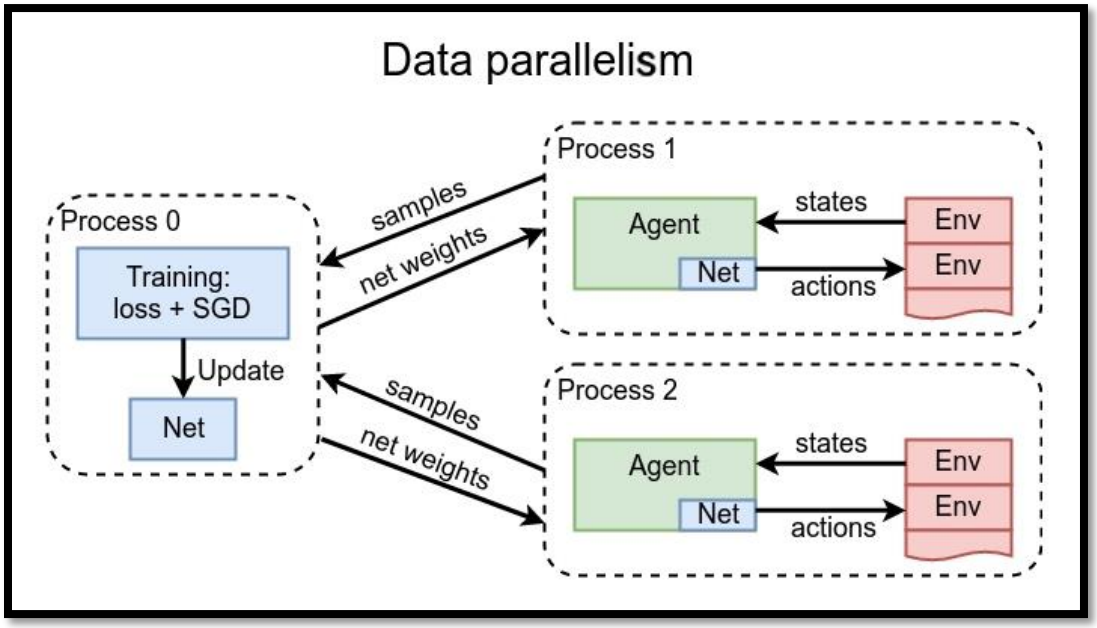

loss_entropy
-0.0145
-0.0155
-0.0165
-0.0175
-0.0185
0.000 2.000M 4.000M 6.000M 8.000M
loss_policy
0.0300
0.00
-0.0300
-0.0600
-0.0900
0.000 2.000M 4.000M 6.000M 8.000M
loss_total
0.0800
0.0400
0.00
-0.0400
-0.0800
0.000 2.000M 4.000M 6.000M 8.000M
loss_value
0.100
0.0800
0.0600
0.0400
0.0200
0.00
0.000 2.000M 4.000M 6.000M 8.000M

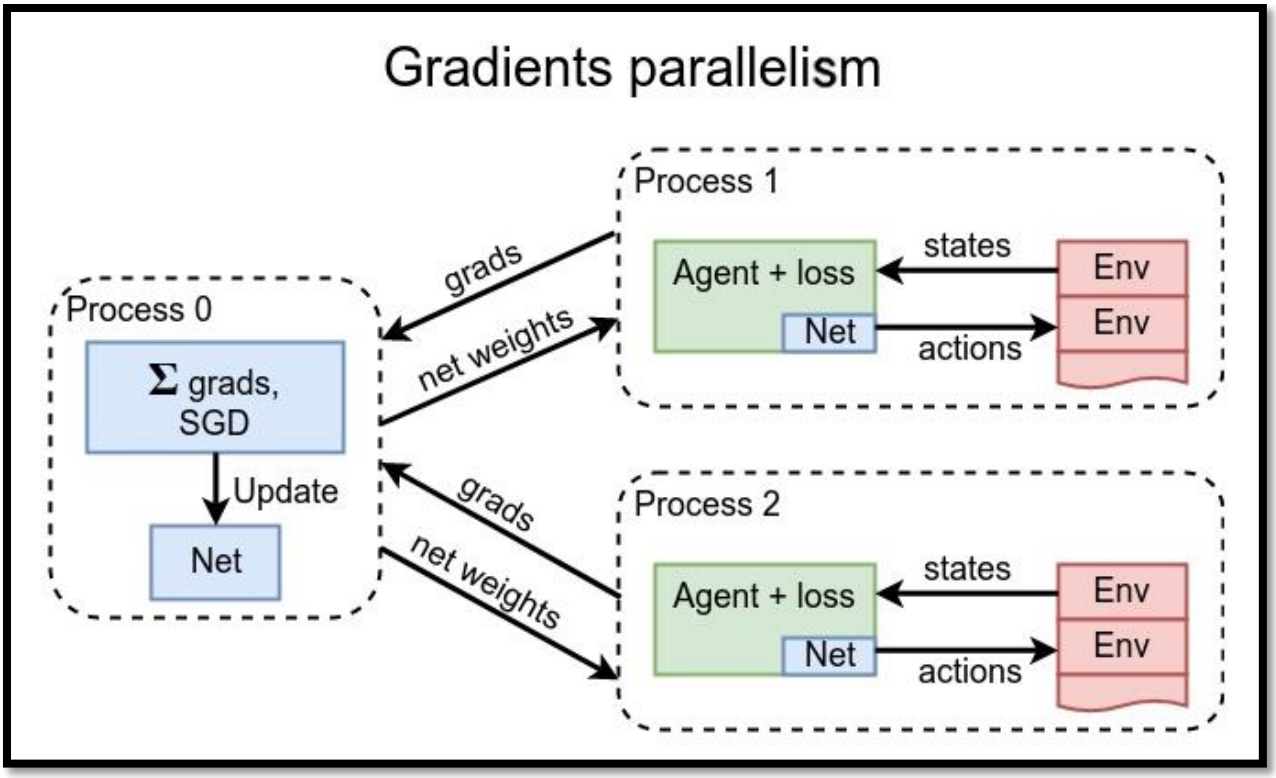

advantage
0.0150
5.000e-3
-5.000e-3
-0.0150
-0.0250
0.000 3.000M 6.000M 9.000M
grad_l2
2.000e-4
1.600e-4
1.200e-4
8.000e-5
4.000e-5
0.00
0.000 3.000M 6.000M 9.000M
grad_max
0.0400
0.0300
0.0200
0.0100
0.00
0.000 3.000M 6.000M 9.000M
grad_var
4.000e-8
3.000e-8
2.000e-8
1.000e-8
0.00
0.000 3.000M 6.000M 9.000M

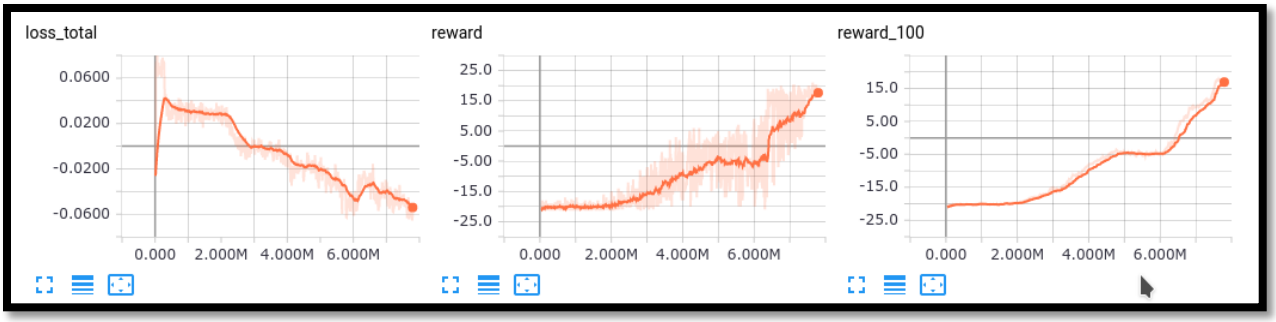

loss_value
0.120
0.0800
0.0400
0.00
0.000 2.000M 4.000M 6.000M 8.000M
reward_100
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000 2.000M 4.000M 6.000M 8.000M

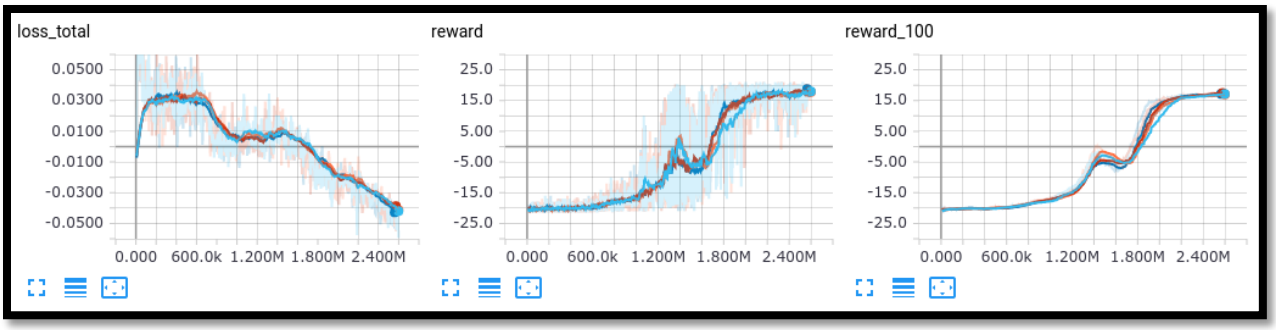
Chapter 11: Asynchronous Advantage Actor-Critic



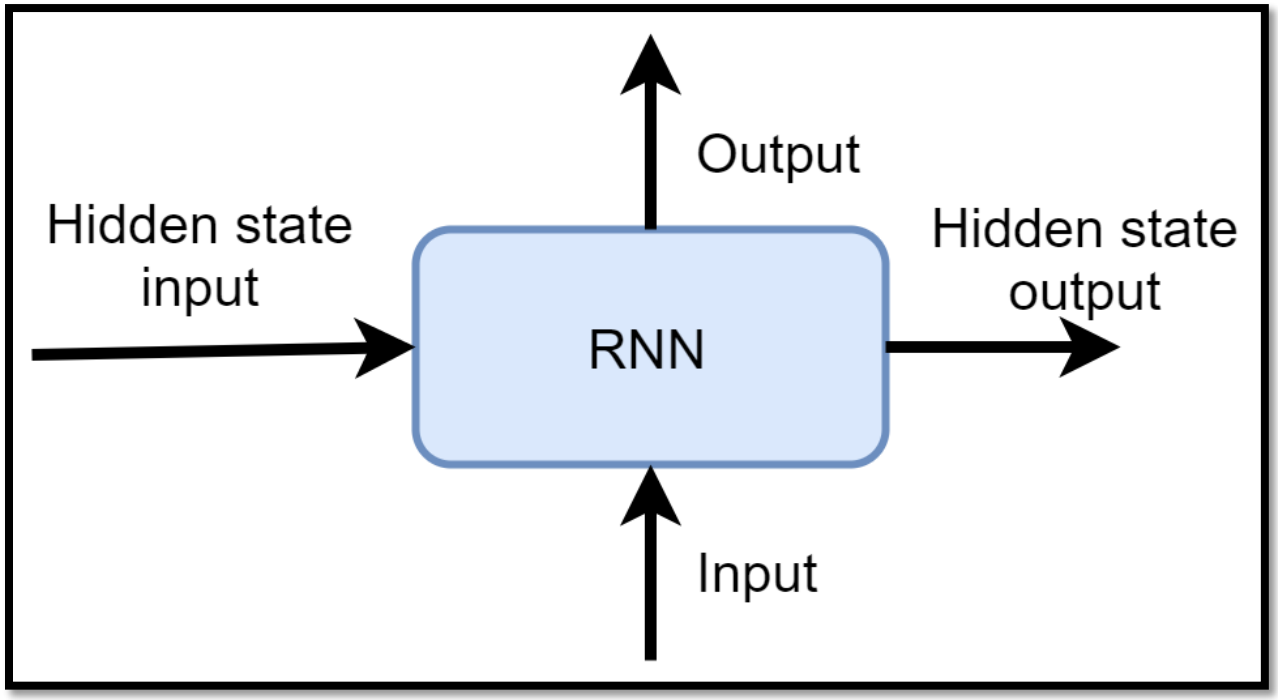


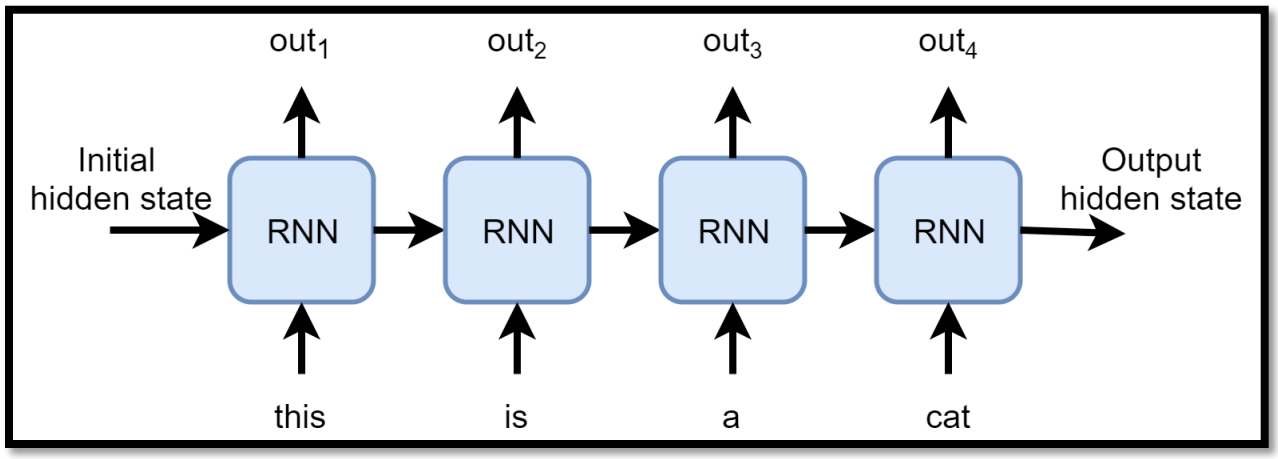

Gradients parallelism
Process 1
Agent + loss
Net
states
actions
Env
Env
grads
net weights
Process 0
Σ grads, SGD
Update
Net
grads
net weights
Process 2
Agent + loss
Net
states
actions
Env
Env

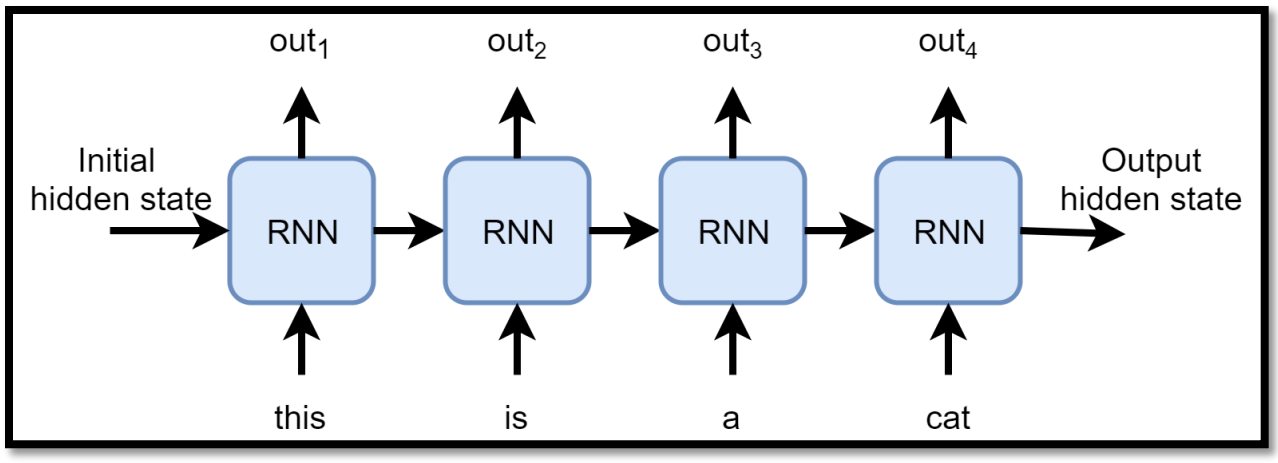

loss_total
0.0600
0.0200
-0.0200
-0.0600
0.000
2.000M
4.000M
6.000M
reward
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
2.000M
4.000M
6.000M
reward_100
15.0
5.00
-5.00
-15.0
-25.0
0.000
2.000M
4.000M
6.000M

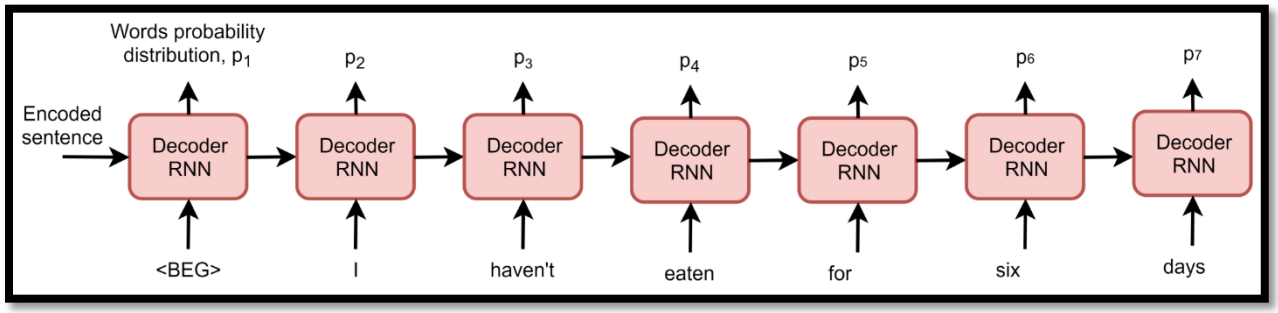

loss_total
0.0500
0.0300
0.0100
-0.0100
-0.0300
-0.0500
0.000
600.0k
1.200M
1.800M
2.400M
reward
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
600.0k
1.200M
1.800M
2.400M
reward_100
25.0
15.0
5.00
-5.00
-15.0
-25.0
0.000
600.0k
1.200M
1.800M
2.400M

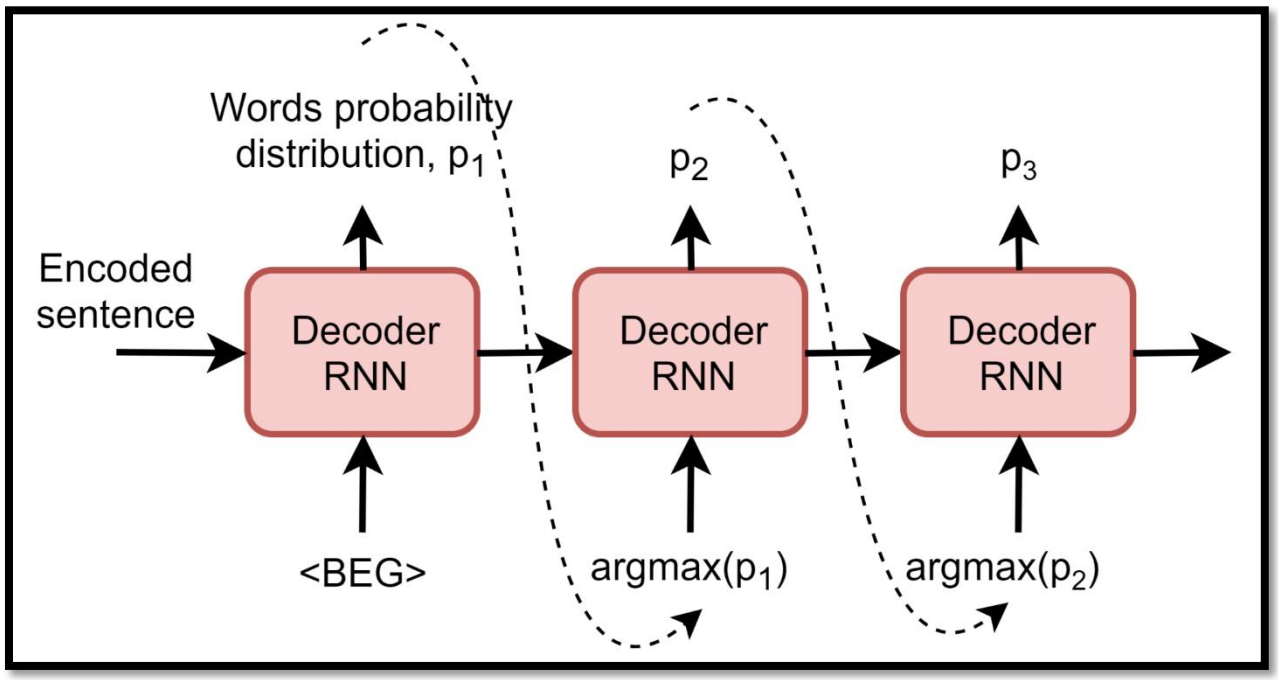
Chapter 12: Chatbots Training with RL

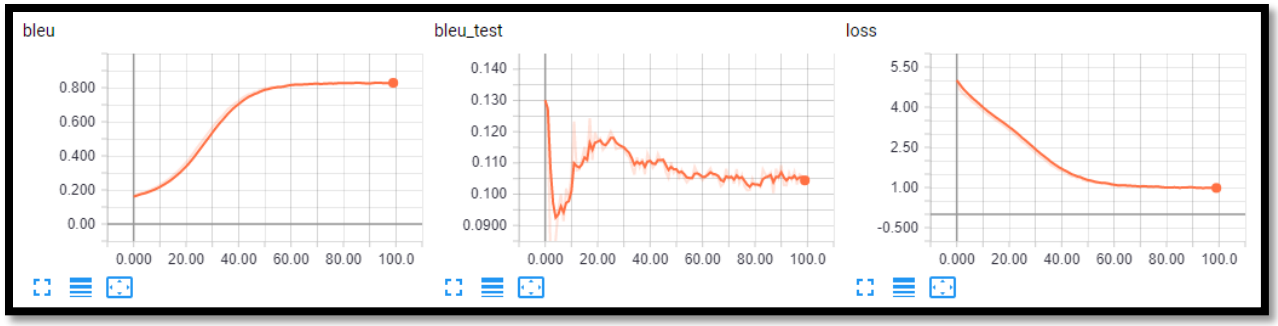


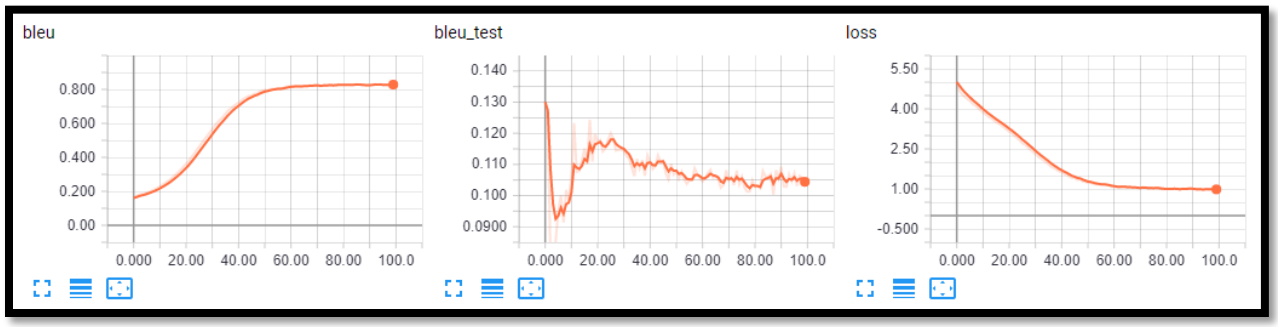


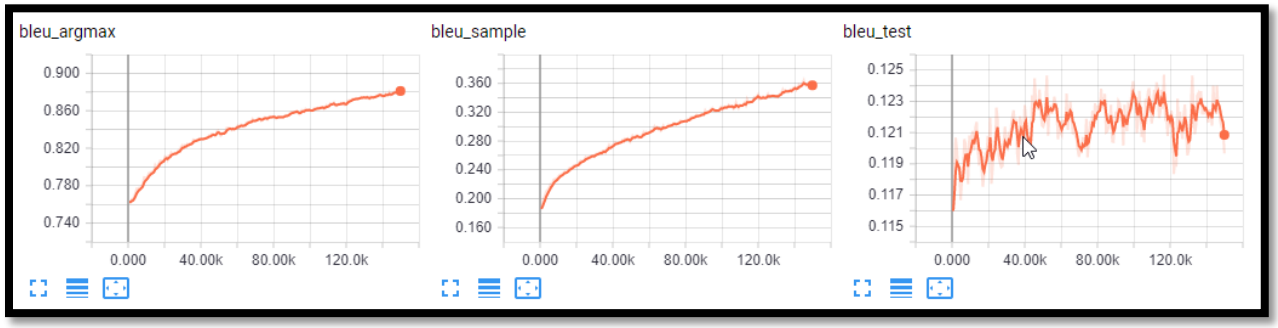

out1
out2
out3
out4
Initial hidden state
RNN
RNN
RNN
RNN
Output hidden state
this
is
a
cat

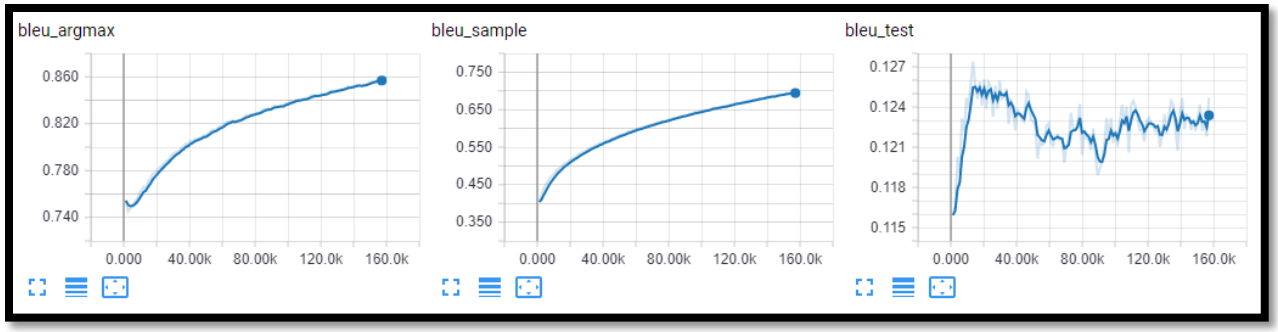

Words probability distribution, p1
p2
p3
p4
p5
p6
p7
Encoded sentence
Decoder RNN
Decoder RNN
Decoder RNN
Decoder RNN
Decoder RNN
Decoder RNN
Decoder RNN
<BEG>
I
haven't
eaten
for
six
days

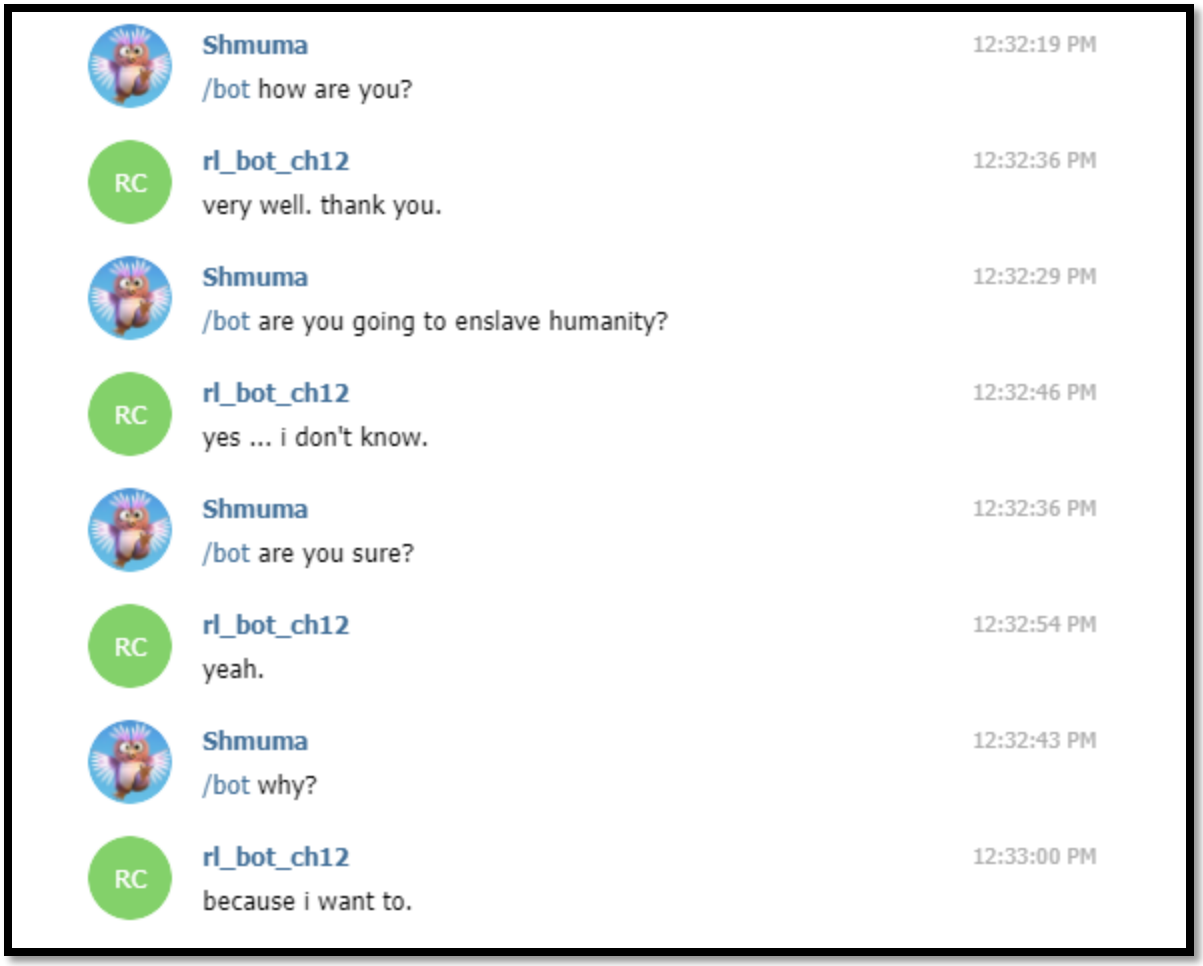

Words probability distribution, p1
p2
p3
Encoded sentence
Decoder RNN
Decoder RNN
Decoder RNN
<BEG>
argmax(p1)
argmax(p2)


bleu
0.800
0.600
0.400
0.200
0.00
0.000 20.00 40.00 60.00 80.00 100.0
bleu_test
0.140
0.130
0.120
0.110
0.100
0.0900
0.000 20.00 40.00 60.00 80.00 100.0
loss
5.50
4.00
2.50
1.00
-0.500
0.000 20.00 40.00 60.00 80.00 100.0

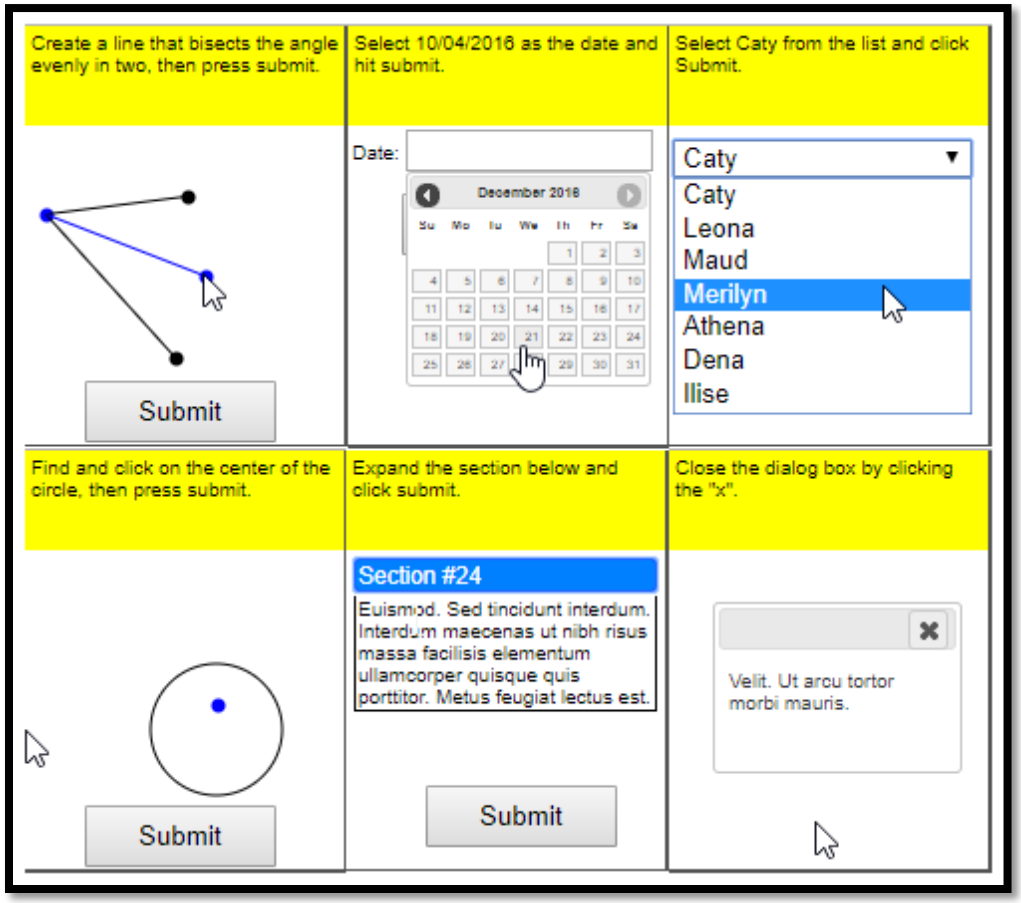

bleu
0.800
0.600
0.400
0.200
0.00
0.000 20.00 40.00 60.00 80.00 100.0
bleu_test
0.140
0.130
0.120
0.110
0.100
0.0900
0.000 20.00 40.00 60.00 80.00 100.0
loss
5.50
4.00
2.50
1.00
-0.500
0.000 20.00 40.00 60.00 80.00 100.0

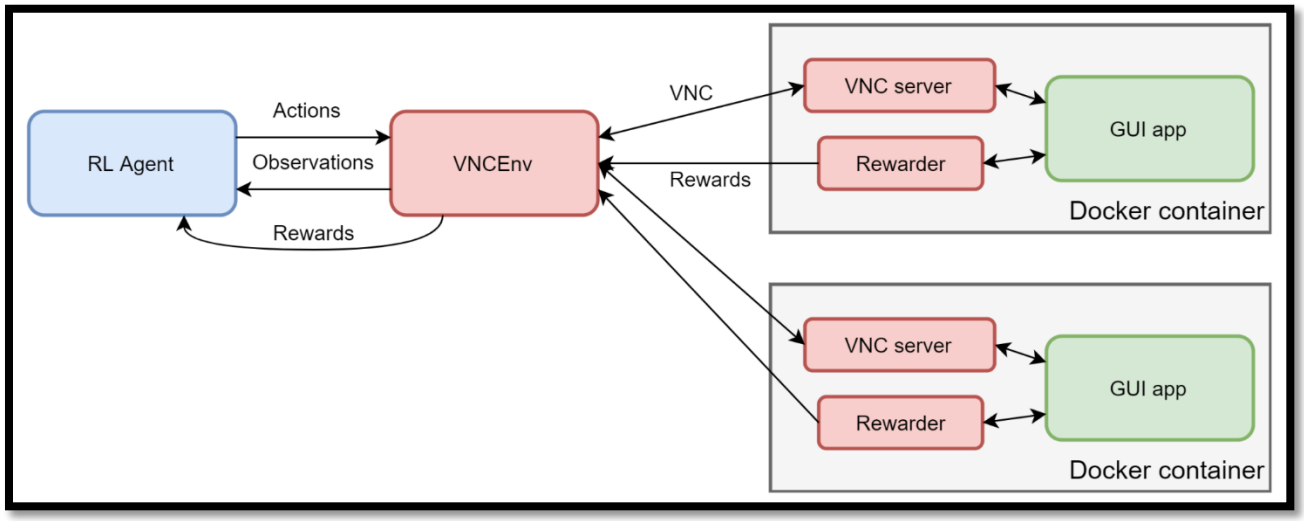

bleu_argmax
0.900
0.860
0.820
0.780
0.740
0.000 40.00k 80.00k 120.0k
bleu_sample
0.360
0.320
0.280
0.240
0.200
0.160
0.000 40.00k 80.00k 120.0k
bleu_test
0.125
0.123
0.121
0.119
0.117
0.115
0.000 40.00k 80.00k 120.0k

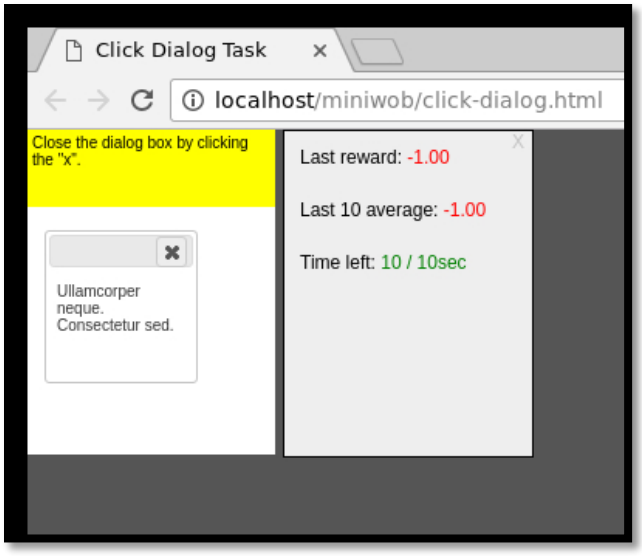

bleu_argmax
0.860
0.820
0.780
0.740
0.000 40.00k 80.00k 120.0k 160.0k
bleu_sample
0.750
0.650
0.550
0.450
0.350
0.000 40.00k 80.00k 120.0k 160.0k
bleu_test
0.127
0.124
0.121
0.118
0.115
0.000 40.00k 80.00k 120.0k 160.0k

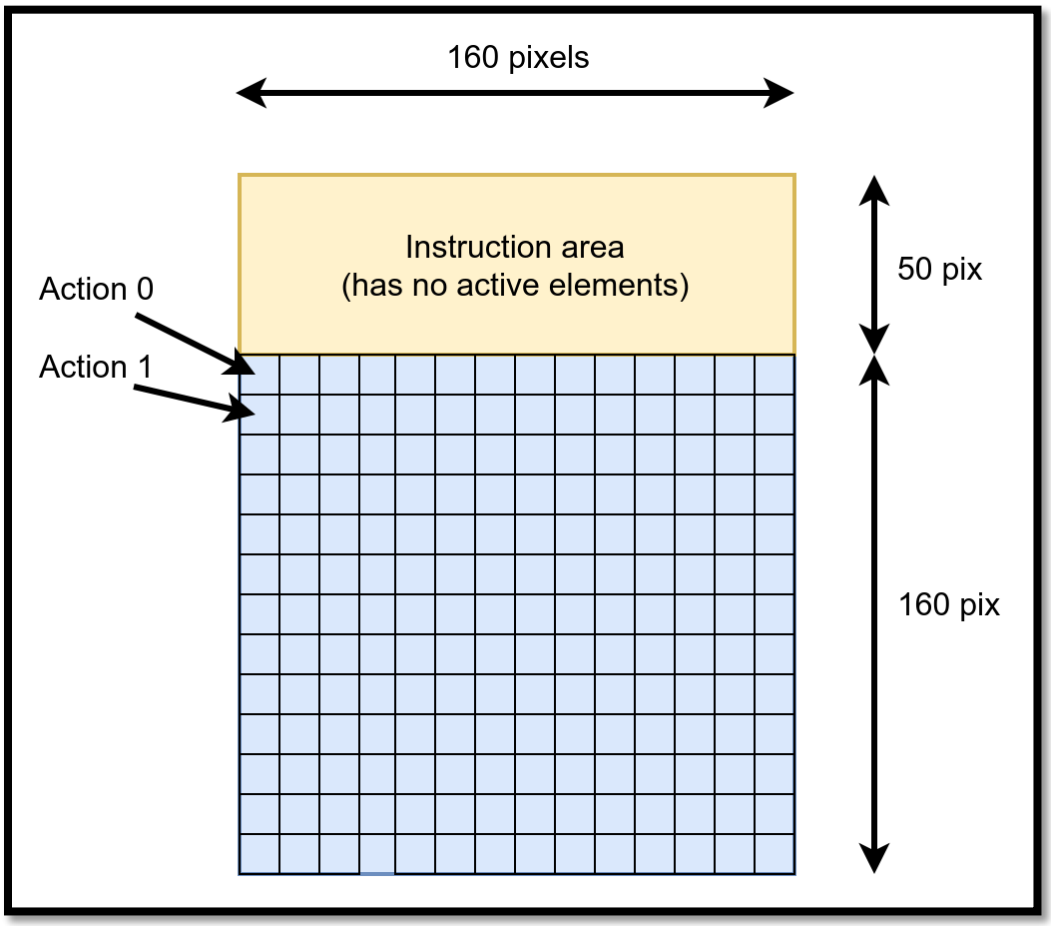

Shmuma
12:32:19 PM
/bot how are you?
RC
rl_bot_ch12
12:32:36 PM
very well. thank you.
Shmuma
12:32:29 PM
/bot are you going to enslave humanity?
RC
rl_bot_ch12
12:32:46 PM
yes ... i don't know.
Shmuma
12:32:36 PM
/bot are you sure?
RC
rl_bot_ch12
12:32:54 PM
yeah.
Shmuma
12:32:43 PM
/bot why?
RC
rl_bot_ch12
12:33:00 PM
because i want to.

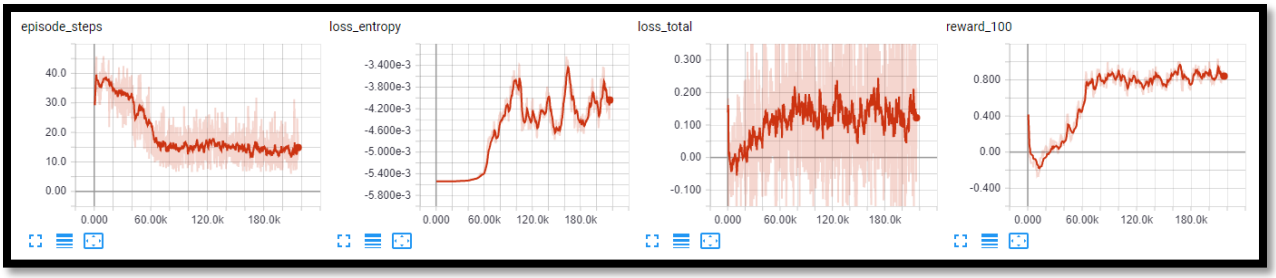
Chapter 13: Web Navigation

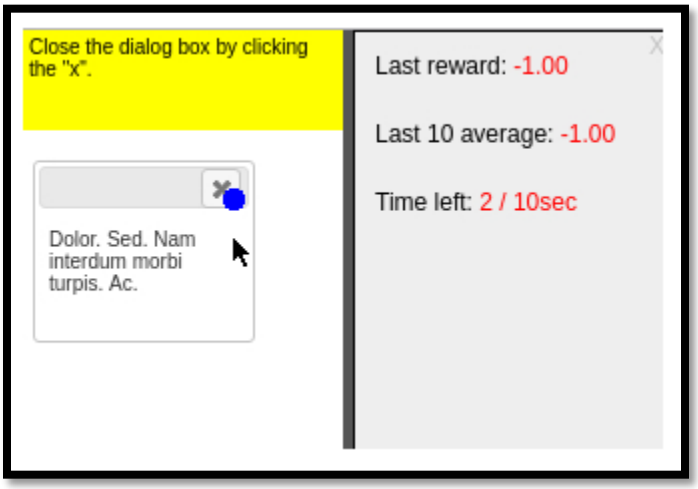


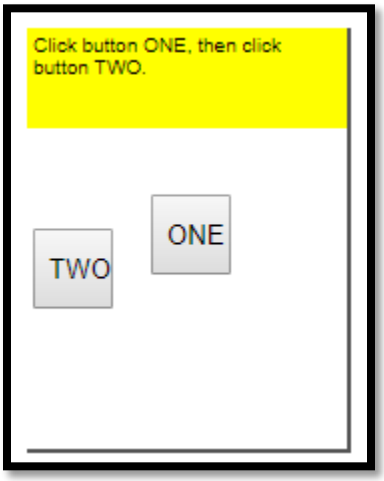


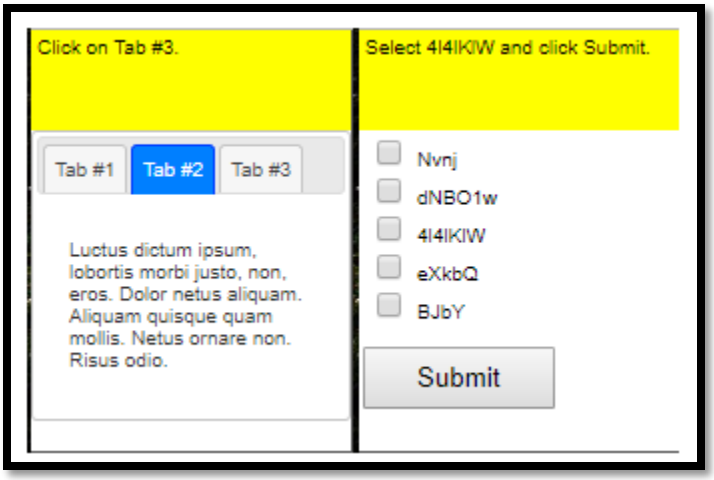

Click Dialog Task
localhost/miniwob/click-dialog.html
Close the dialog box by clicking the "x".
Ullamcorper neque. Consectetur sed.
Last reward: -1.00
Last 10 average: -1.00
Time left: 10 / 10sec

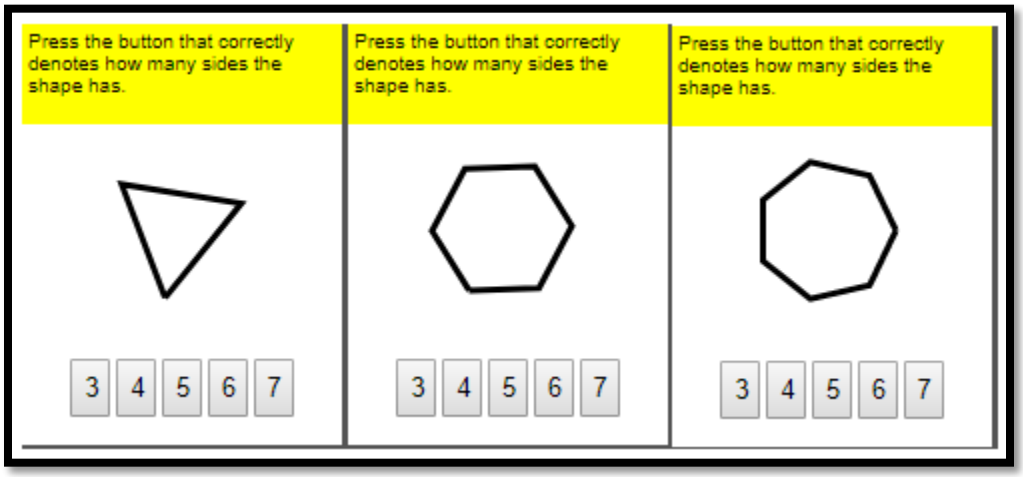

160 pixels
Instruction area
(has no active elements)
50 pix
Action 0
Action 1
160 pix

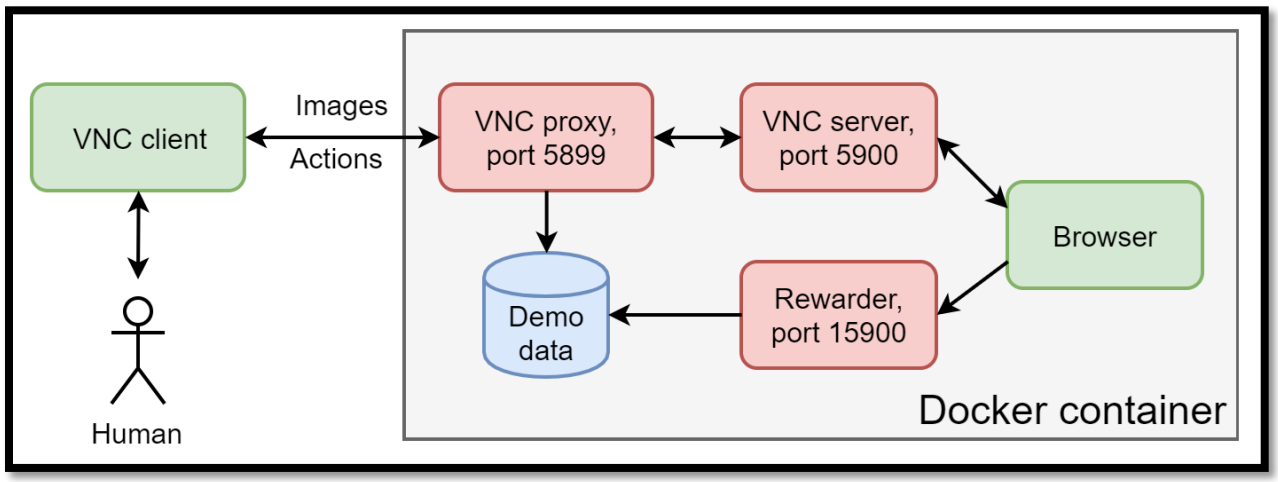

episode_steps
40.0
30.0
20.0
10.0
0.00
0.000
60.00k
120.0k
180.0k
loss_entropy
-3.400e-3
-3.800e-3
-4.200e-3
-4.600e-3
-5.000e-3
-5.400e-3
-5.800e-3
0.000
60.00k
120.0k
180.0k
loss_total
0.300
0.200
0.100
0.00
-0.100
0.000
60.00k
120.0k
180.0k
reward_100
0.800
0.400
0.00
-0.400
0.000
60.00k
120.0k
180.0k

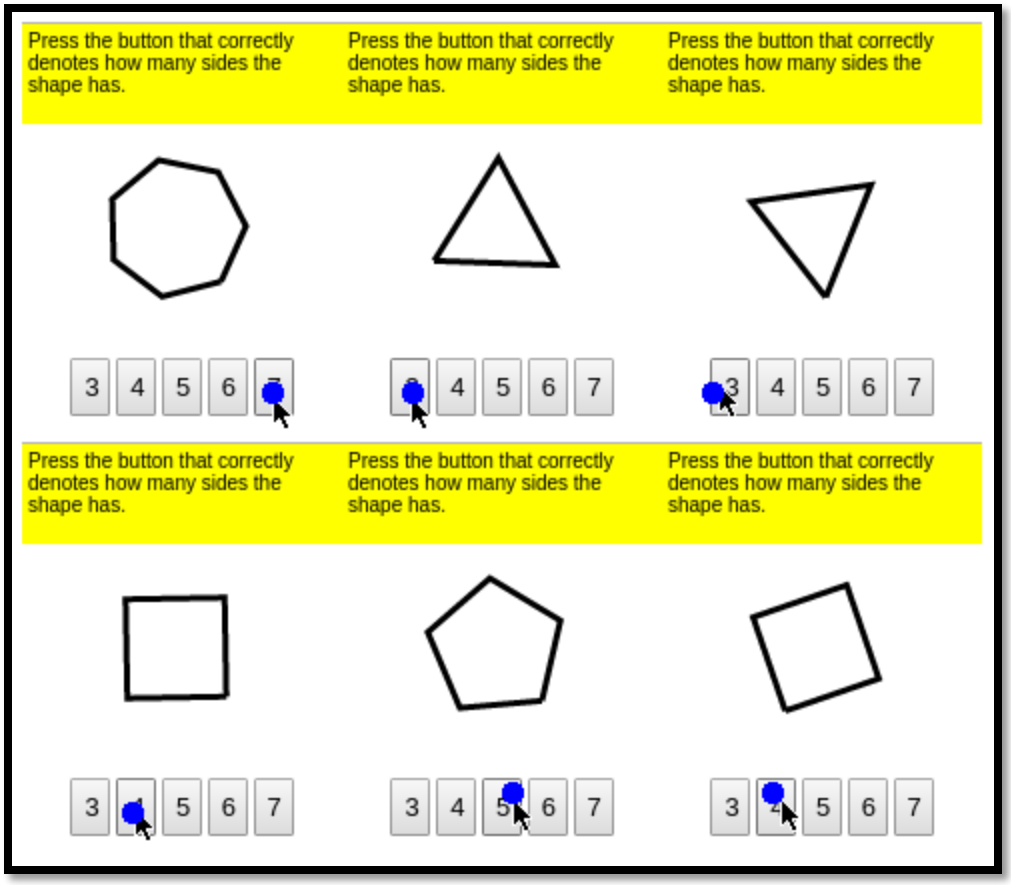

Close the dialog box by clicking the "x".
Dolor. Sed. Nam interdum morbi turpis. Ac.
Last reward: -1.00
Last 10 average: -1.00
Time left: 2 / 10sec

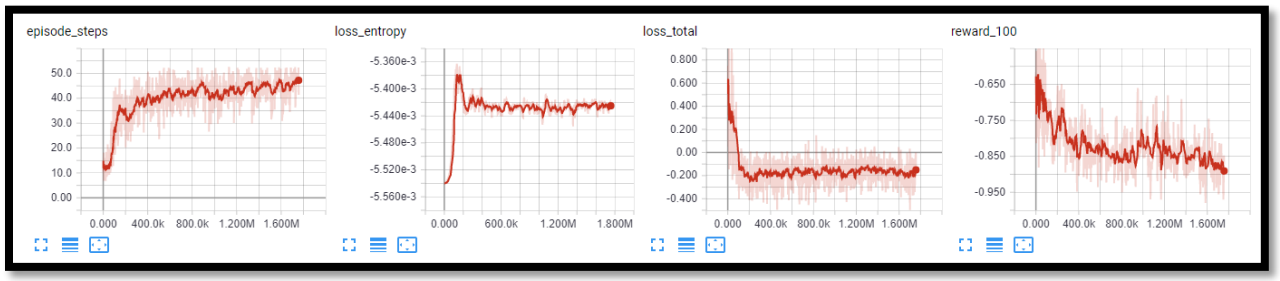

Click button ONE, then click button TWO.
TWO
ONE

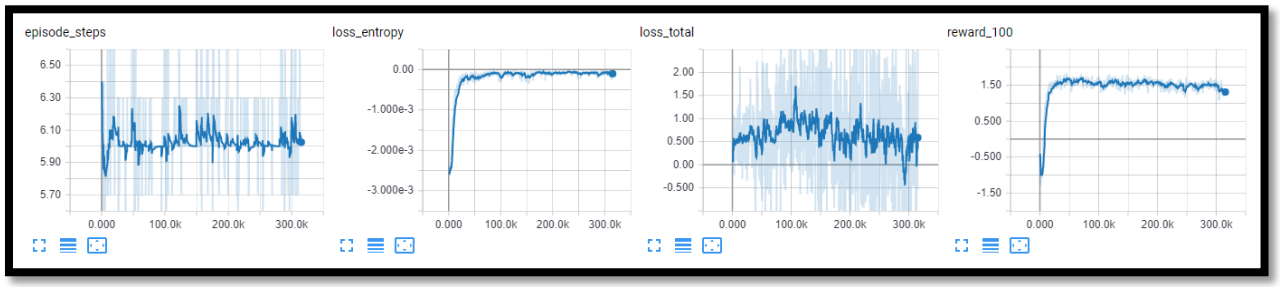

Click on Tab #3.
Select 4l4lKlW and click Submit.
Tab #1
Tab #2
Tab #3
Luctus dictum ipsum, lobortis morbi justo, non, eros. Dolor netus aliquam. Aliquam quisque quam mollis. Netus ornare non. Risus odio.
Nvnj
dNBO1w
4l4lKlW
eXkbQ
BJbY
Submit

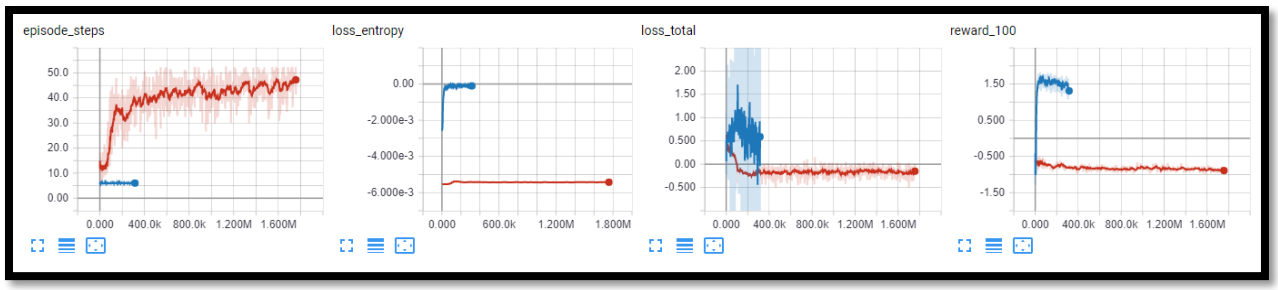

Press the button that correctly denotes how many sides the shape has.
3 4 5 6 7
Press the button that correctly denotes how many sides the shape has.
3 4 5 6 7
Press the button that correctly denotes how many sides the shape has.
3 4 5 6 7

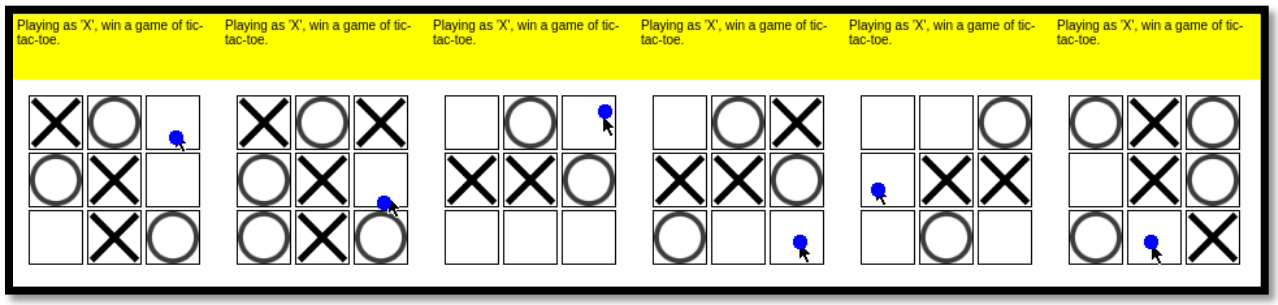

VNC client
Images
Actions
VNC proxy, port 5899
VNC server, port 5900
Browser
Demo data
Rewarder, port 15900
Human
Docker container

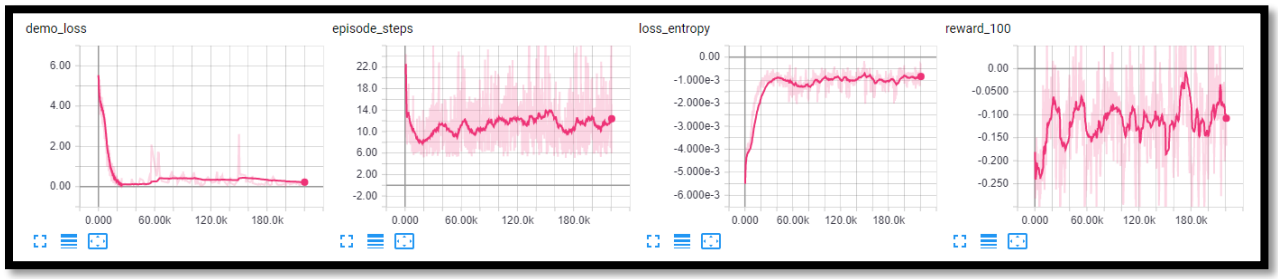

Press the button that correctly denotes how many sides the shape has.
Press the button that correctly denotes how many sides the shape has.
Press the button that correctly denotes how many sides the shape has.
3 4 5 6 7
3 4 5 6 7
3 4 5 6 7
Press the button that correctly denotes how many sides the shape has.
Press the button that correctly denotes how many sides the shape has.
Press the button that correctly denotes how many sides the shape has.
3 4 5 6 7
3 4 5 6 7
3 4 5 6 7

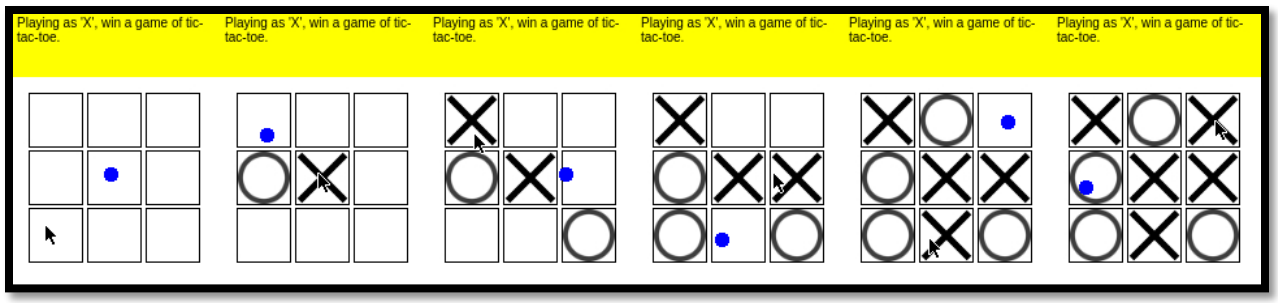

episode_steps
loss_entropy
loss_total
reward_100

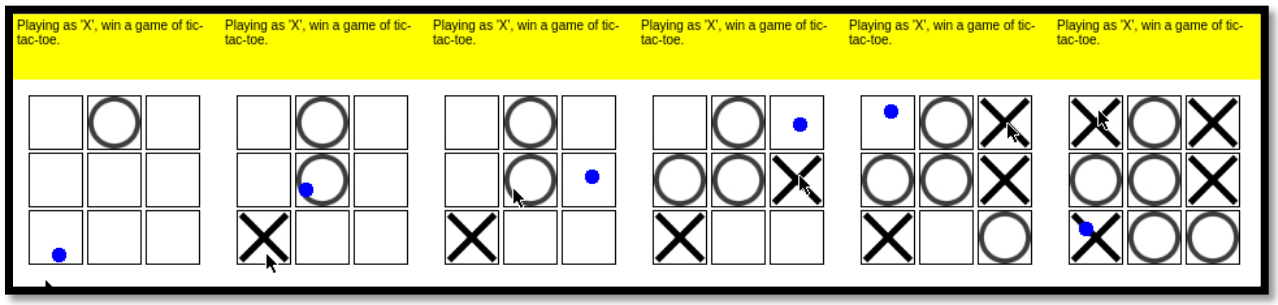

episode_steps
loss_entropy
loss_total
reward_100

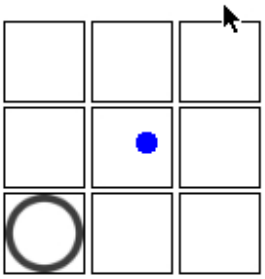
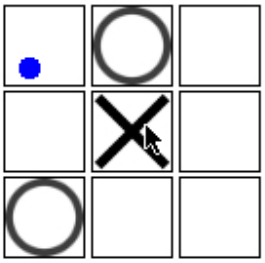
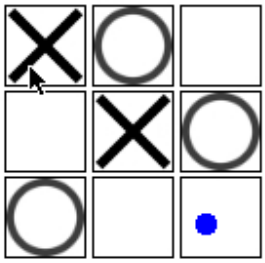
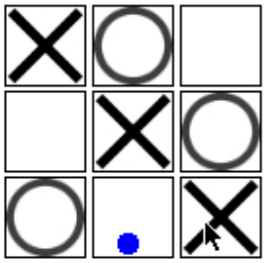
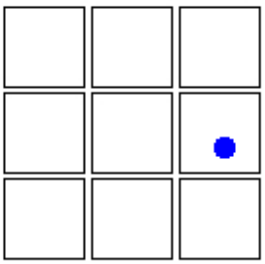
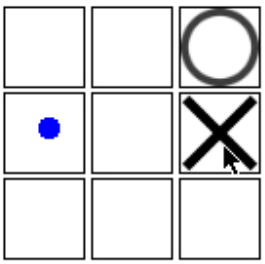
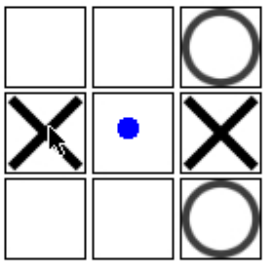
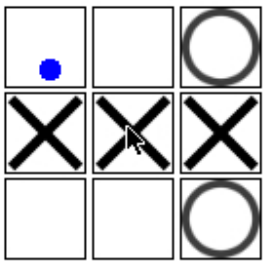

episode_steps
50.0
40.0
30.0
20.0
10.0
0.00
0.000
400.0k
800.0k
1.200M
1.600M
loss_entropy
0.00
-2.000e-3
-4.000e-3
-6.000e-3
0.000
600.0k
1.200M
1.800M
loss_total
2.00
1.50
1.00
0.500
0.00
-0.500
0.000
400.0k
800.0k
1.200M
1.600M
reward_100
1.50
0.500
-0.500
-1.50
0.000
400.0k
800.0k
1.200M
1.600M

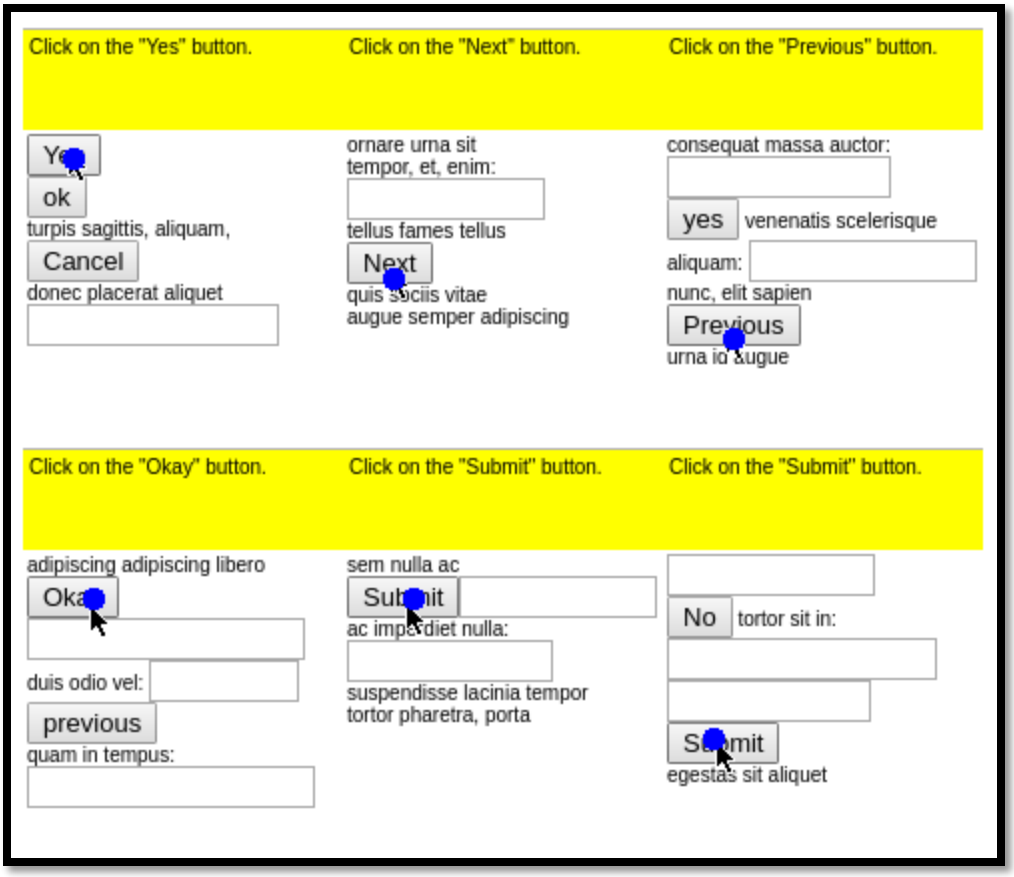

Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.


demo_loss
6.00
4.00
2.00
0.00
0.000
60.00k
120.0k
180.0k
episode_steps
22.0
18.0
14.0
10.0
6.00
2.00
-2.00
0.000
60.00k
120.0k
180.0k
loss_entropy
0.00
-1.000e-3
-2.000e-3
-3.000e-3
-4.000e-3
-5.000e-3
-6.000e-3
0.000
60.00k
120.0k
180.0k
reward_100
0.00
-0.0500
-0.100
-0.150
-0.200
-0.250
0.000
60.00k
120.0k
180.0k

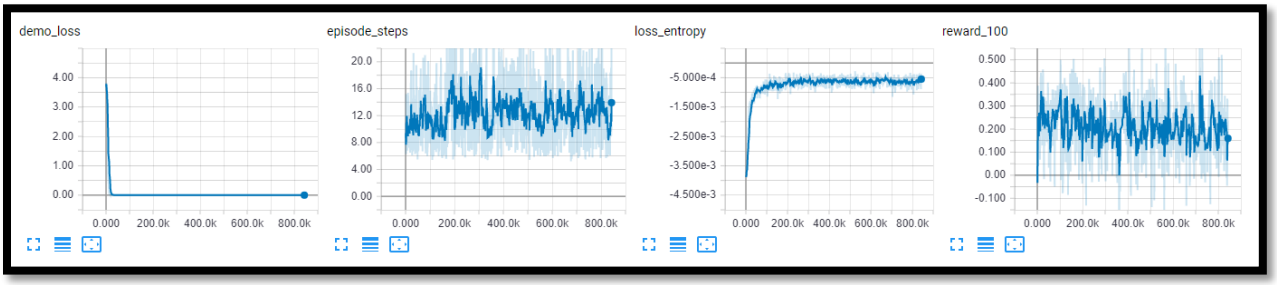

Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.


Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.
Playing as 'X', win a game of tic-tac-toe.

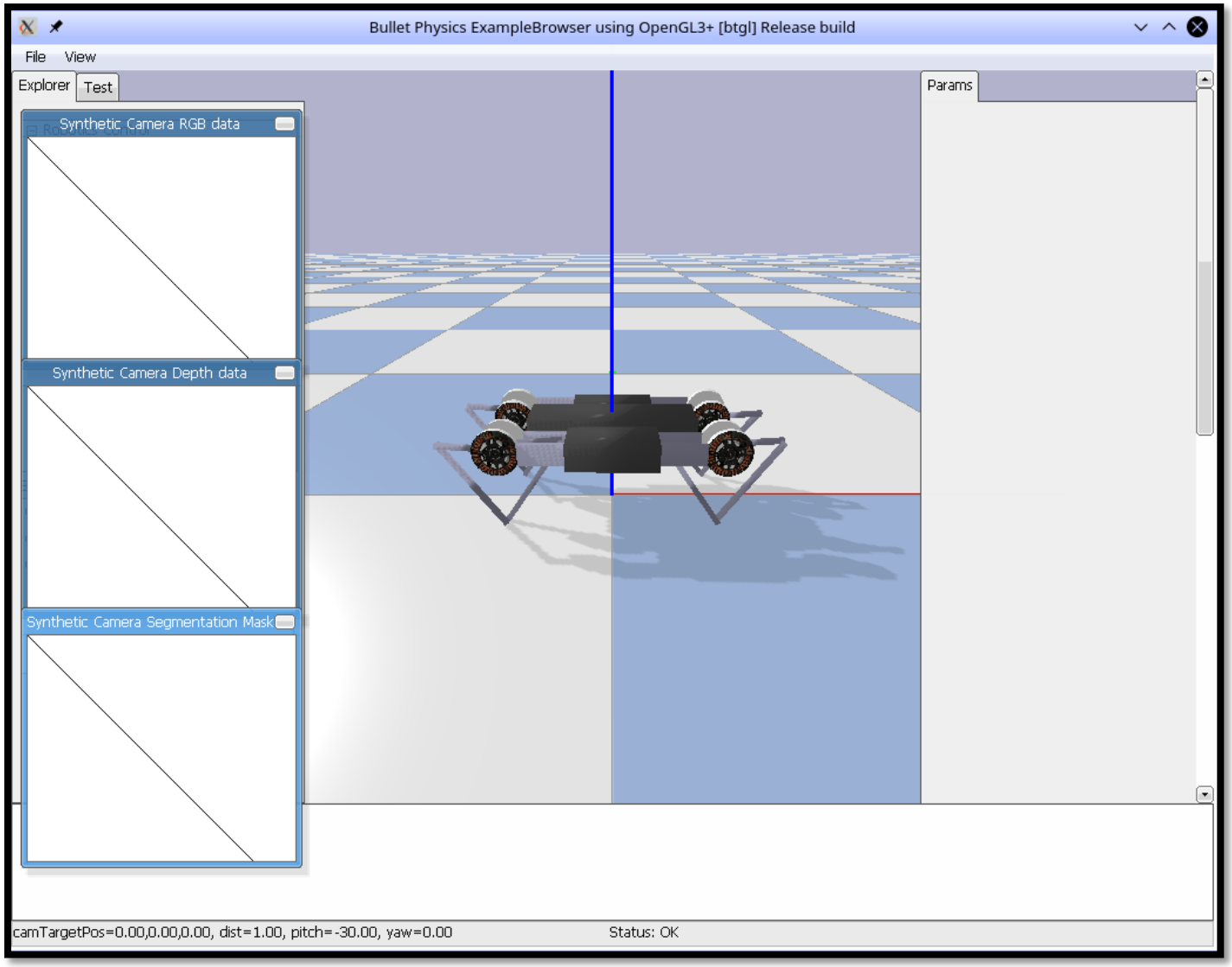
Playing as 'X', win a game of tic-tac-toe.	Playing as 'X', win a game of tic-tac-toe.	Playing as 'X', win a game of tic-tac-toe.	Playing as 'X', win a game of tic-tac-toe.
			
Playing as 'X', win a game of tic-tac-toe.	Playing as 'X', win a game of tic-tac-toe.	Playing as 'X', win a game of tic-tac-toe.	Playing as 'X', win a game of tic-tac-toe.
			

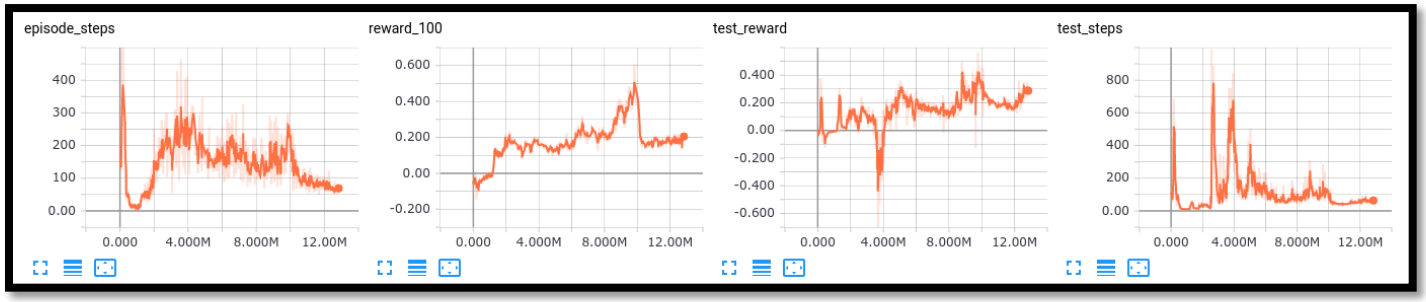

Click on the "Yes" button.
Click on the "Next" button.
Click on the "Previous" button.
ok
turpis sagittis, aliquam,
Cancel
donec placerat aliquet
ornare urna sit
tempor, et, enim:
tellus fames tellus
Next
quis sociis vitae
augue semper adipiscing
consequat massa auctor:
yes
venenatis scelerisque
aliquam:
nunc, elit sapien
Previous
urna id augue
Click on the "Okay" button.
Click on the "Submit" button.
Click on the "Submit" button.
adipiscing adipiscing libero
duis odio vel:
previous
quam in tempus:
sem nulla ac
ac imperdiet nulla:
suspendisse lacinia tempor
tortor pharetra, porta
No
tortor sit in:
egestas sit aliquet

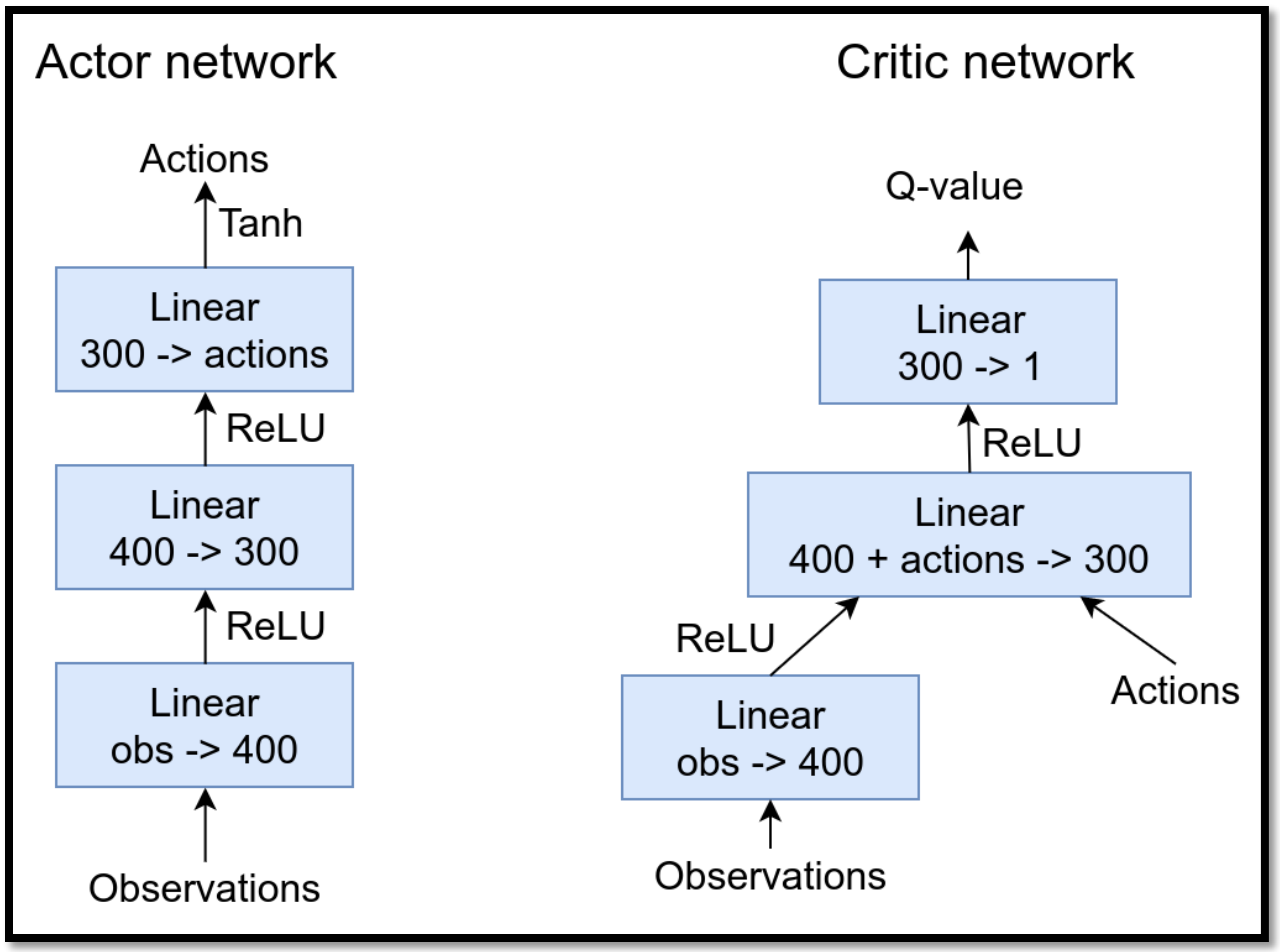

demo_loss
episode_steps
loss_entropy
reward_100

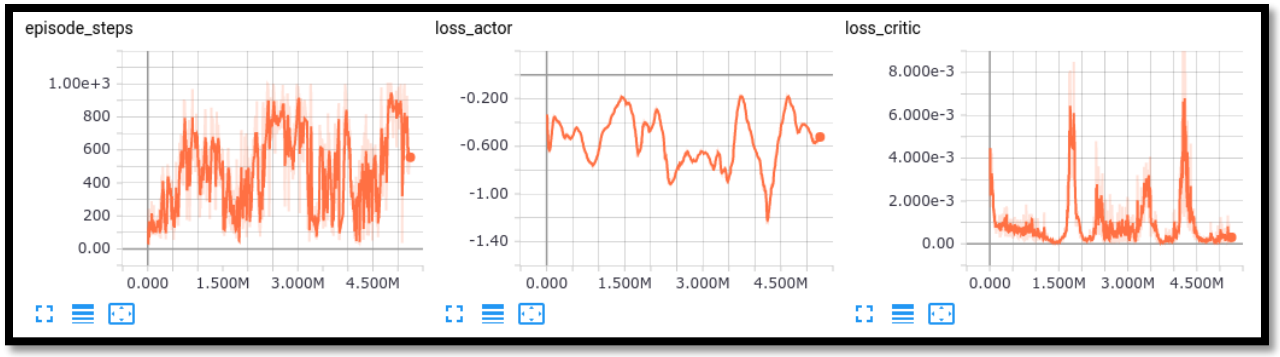

demo_loss
episode_steps
loss_entropy
reward_100

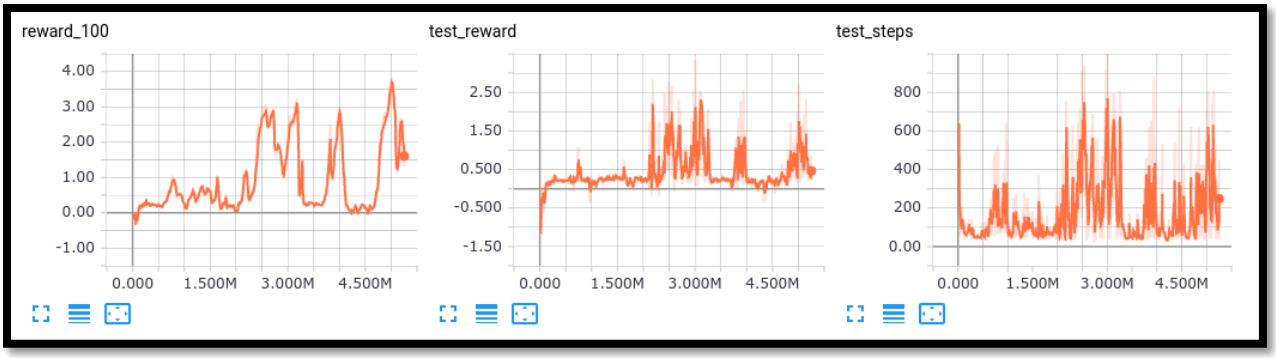
Chapter 14: Continuous Action Space

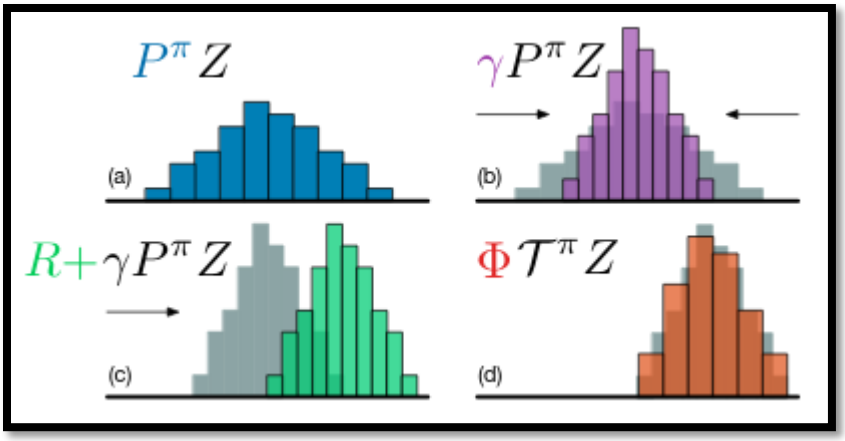


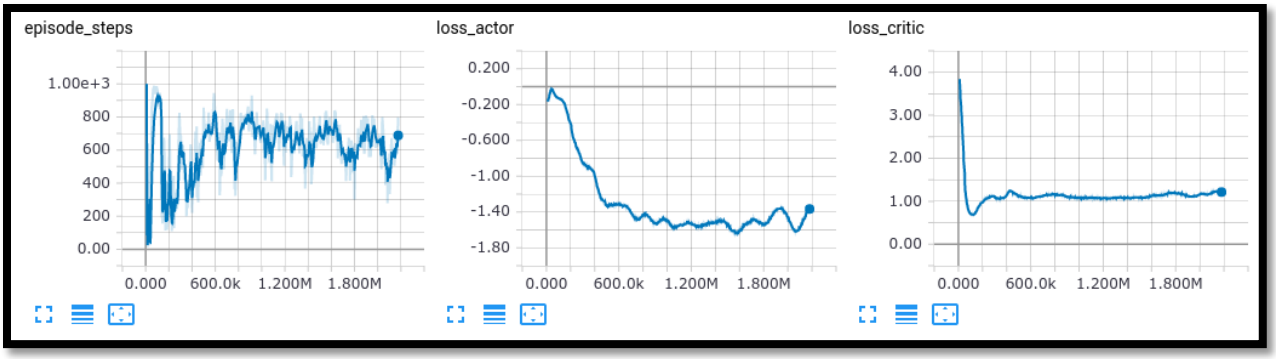


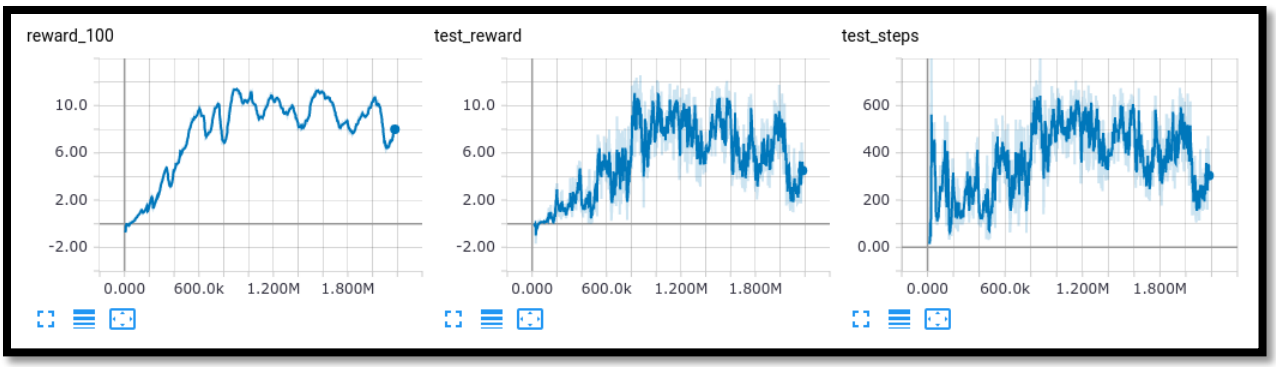

Actor network
Actions
Tanh
Linear
300 -> actions
ReLU
Linear
400 -> 300
ReLU
Linear
obs -> 400
Observations
Critic network
Q-value
Linear
300 -> 1
ReLU
Linear
400 + actions -> 300
ReLU
Actions
Linear
obs -> 400
Observations


episode_steps
1.00e+3
800
600
400
200
0.00
0.000
1.500M
3.000M
4.500M
loss_actor
-0.200
-0.600
-1.00
-1.40
0.000
1.500M
3.000M
4.500M
loss_critic
8.000e-3
6.000e-3
4.000e-3
2.000e-3
0.00
0.000
1.500M
3.000M
4.500M

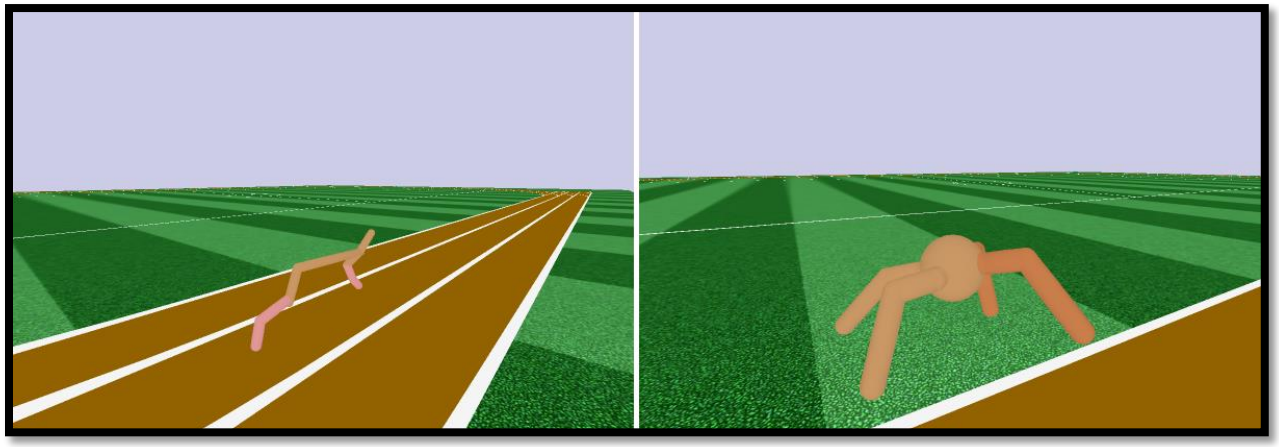

reward_100
4.00
3.00
2.00
1.00
0.00
-1.00
0.000
1.500M
3.000M
4.500M
test_reward
2.50
1.50
0.500
-0.500
-1.50
0.000
1.500M
3.000M
4.500M
test_steps
800
600
400
200
0.00
0.000
1.500M
3.000M
4.500M

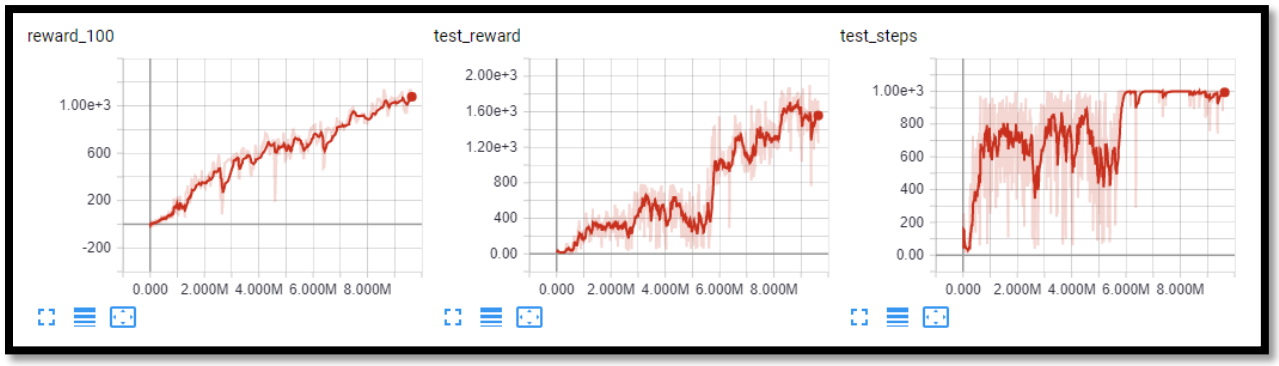

$P^\pi Z$
(a)
$\gamma P^\pi Z$
(b)
$R+\gamma P^\pi Z$
(c)
$\Phi \mathcal{T}^\pi Z$
(d)

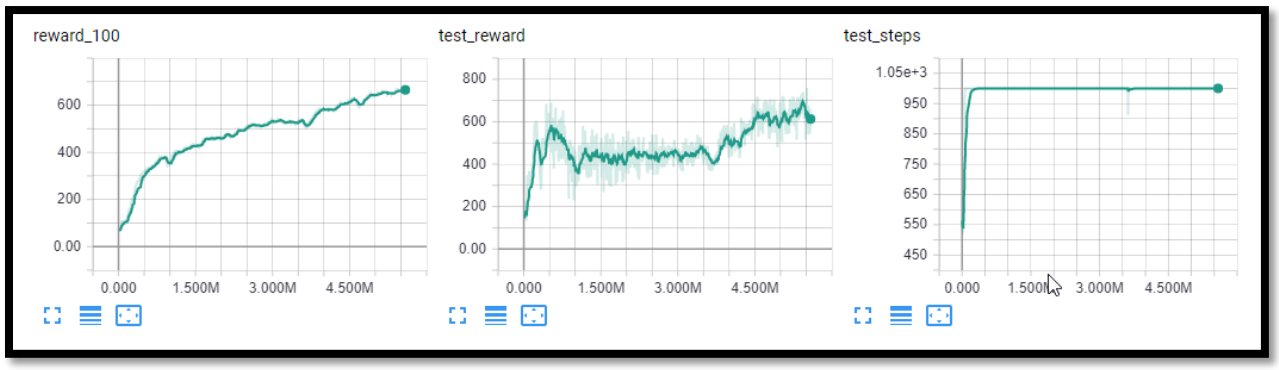

episode_steps
1.00e+3
800
600
400
200
0.00
0.000
600.0k
1.200M
1.800M
loss_actor
0.200
-0.200
-0.600
-1.00
-1.40
-1.80
0.000
600.0k
1.200M
1.800M
loss_critic
4.00
3.00
2.00
1.00
0.00
0.000
600.0k
1.200M
1.800M

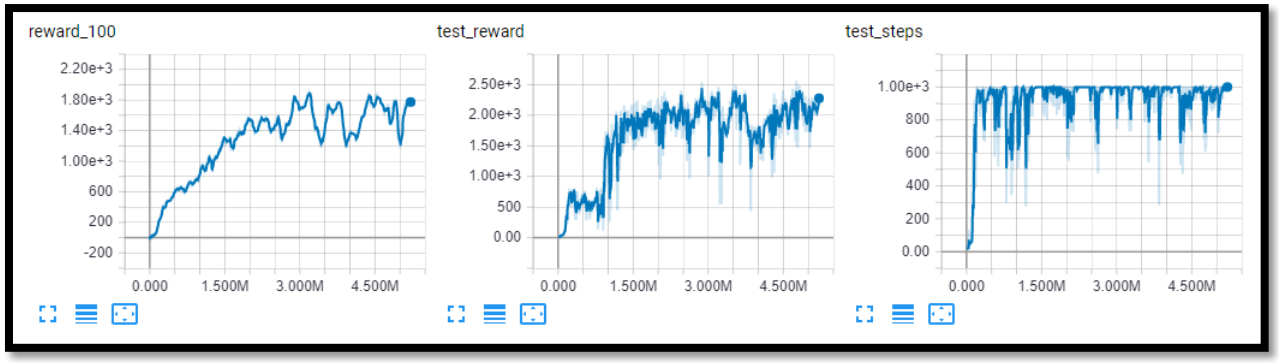

reward_100
10.0
6.00
2.00
-2.00
0.000
600.0k
1.200M
1.800M
test_reward
10.0
6.00
2.00
-2.00
0.000
600.0k
1.200M
1.800M
test_steps
600
400
200
0.00
0.000
600.0k
1.200M
1.800M

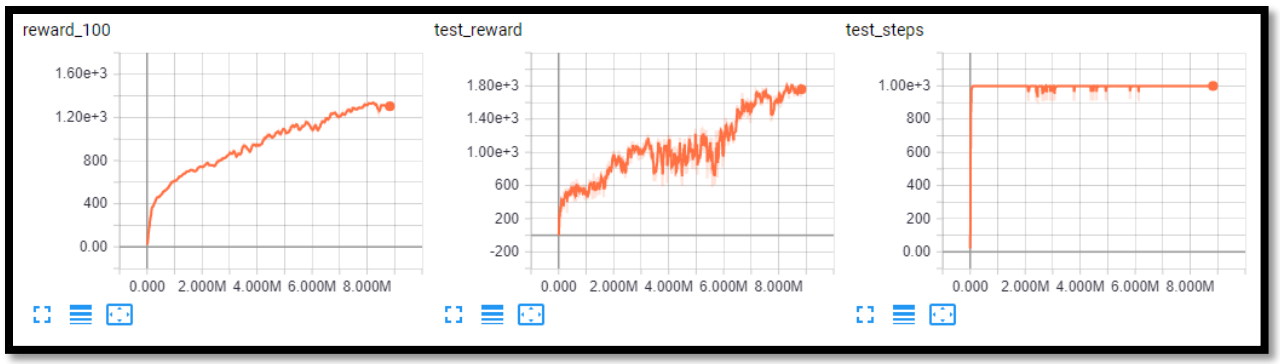
Chapter 15: Trust Regions – TRPO, PPO, and ACKTR

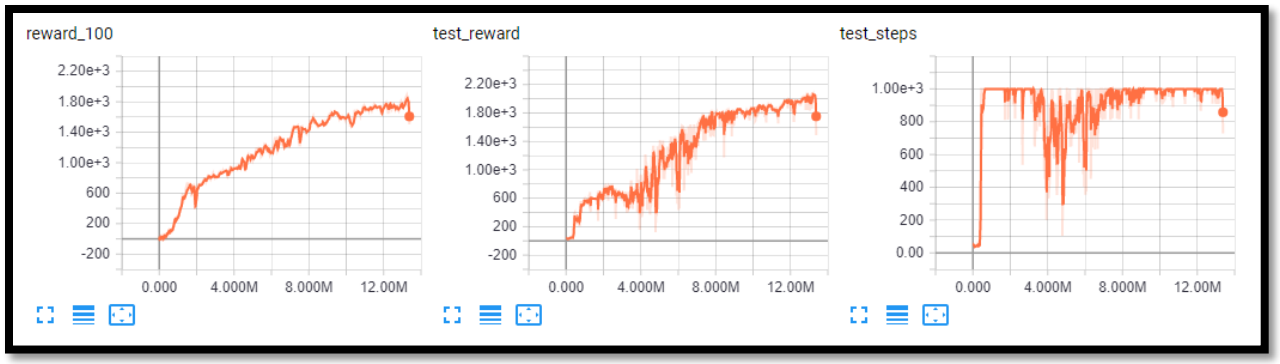


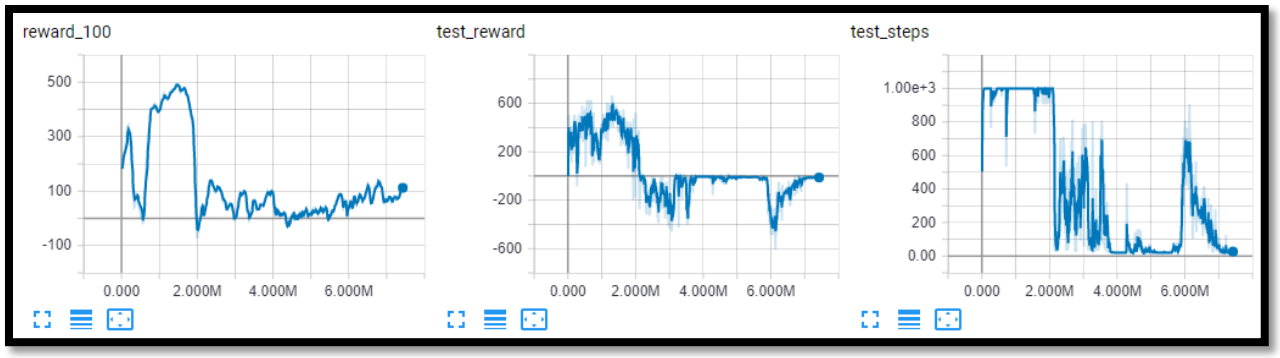


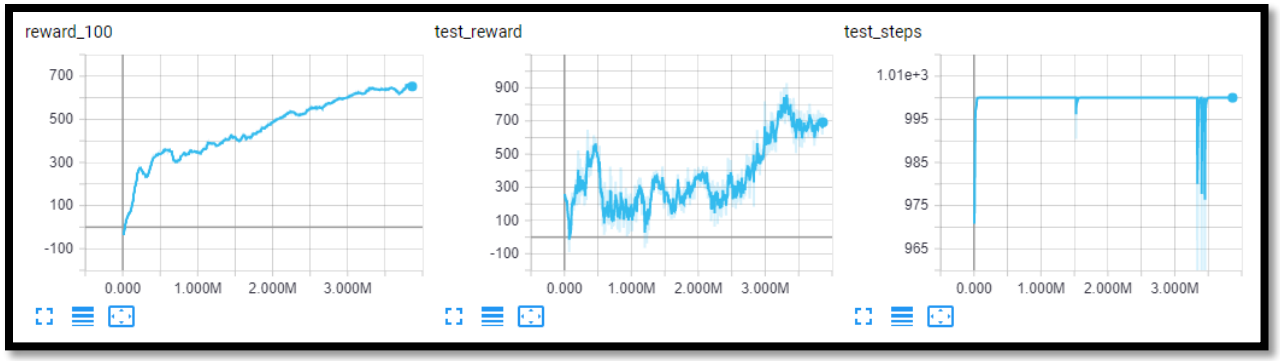


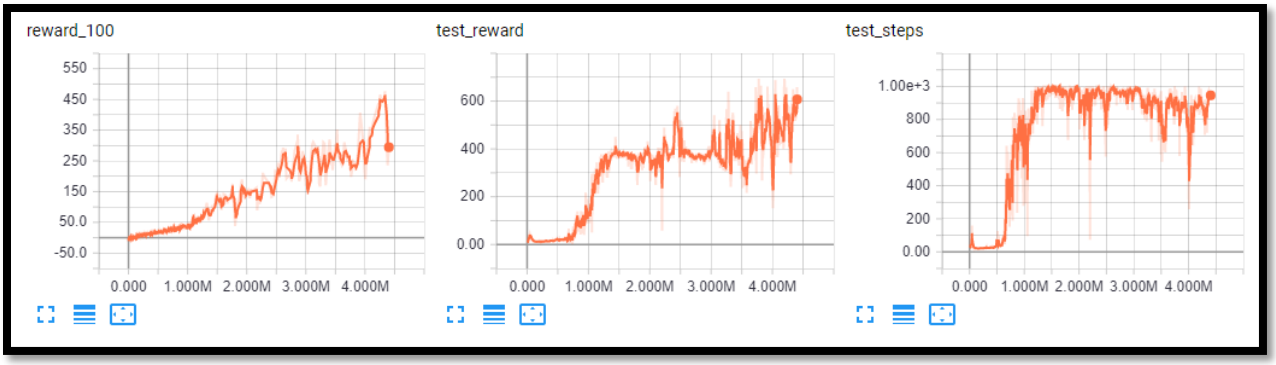



reward_100
1.60e+3
1.20e+3
800
400
0.00
0.000 2.000M 4.000M 6.000M 8.000M
test_reward
1.80e+3
1.40e+3
1.00e+3
600
200
-200
0.000 2.000M 4.000M 6.000M 8.000M
test_steps
1.00e+3
800
600
400
200
0.00
0.000 2.000M 4.000M 6.000M 8.000M

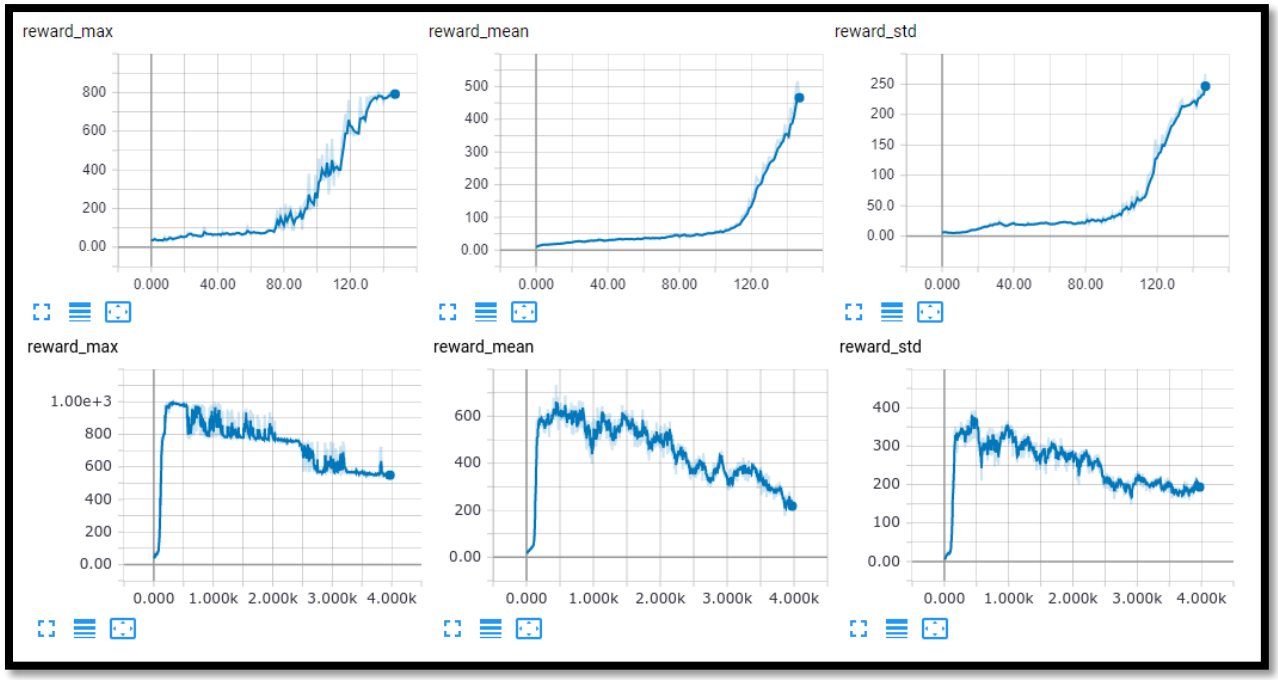

reward_100
2.20e+3
1.80e+3
1.40e+3
1.00e+3
600
200
-200
0.000 4.000M 8.000M 12.00M
test_reward
2.20e+3
1.80e+3
1.40e+3
1.00e+3
600
200
-200
0.000 4.000M 8.000M 12.00M
test_steps
1.00e+3
800
600
400
200
0.00
0.000 4.000M 8.000M 12.00M

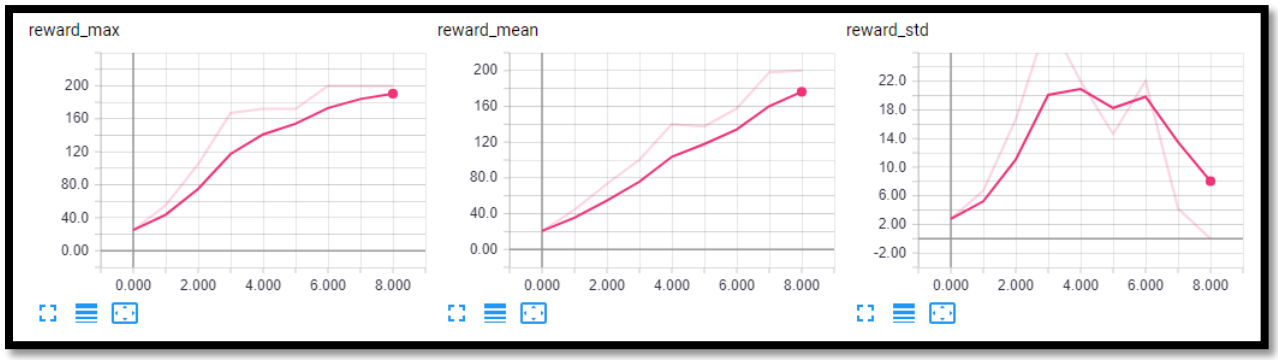

reward_100
500
300
100
-100
0.000 2.000M 4.000M 6.000M
test_reward
600
200
-200
-600
0.000 2.000M 4.000M 6.000M
test_steps
1.00e+3
800
600
400
200
0.00
0.000 2.000M 4.000M 6.000M

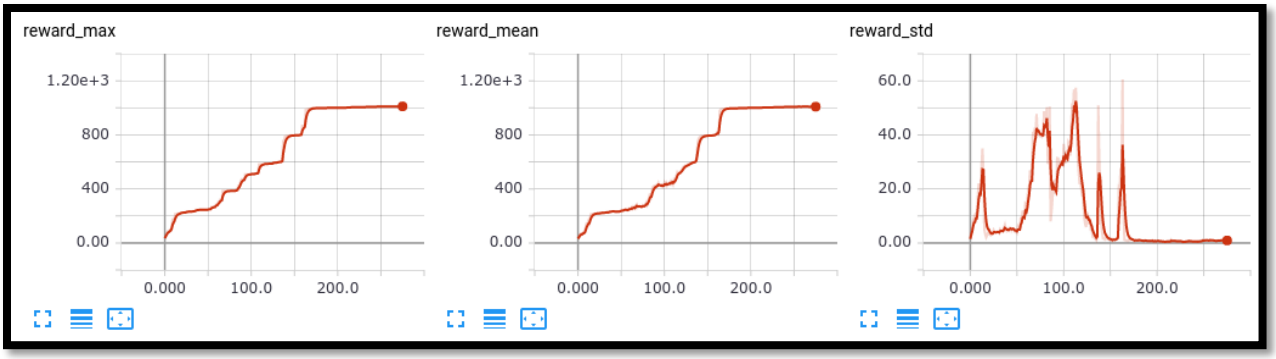

reward_100
700
500
300
100
-100
0.000 1.000M 2.000M 3.000M
test_reward
900
700
500
300
100
-100
0.000 1.000M 2.000M 3.000M
test_steps
1.01e+3
995
985
975
965
0.000 1.000M 2.000M 3.000M


reward_100
550
450
350
250
150
50.0
-50.0
0.000
1.000M
2.000M
3.000M
4.000M
test_reward
600
400
200
0.00
0.000
1.000M
2.000M
3.000M
4.000M
test_steps
1.00e+3
800
600
400
200
0.00
0.000
1.000M
2.000M
3.000M
4.000M

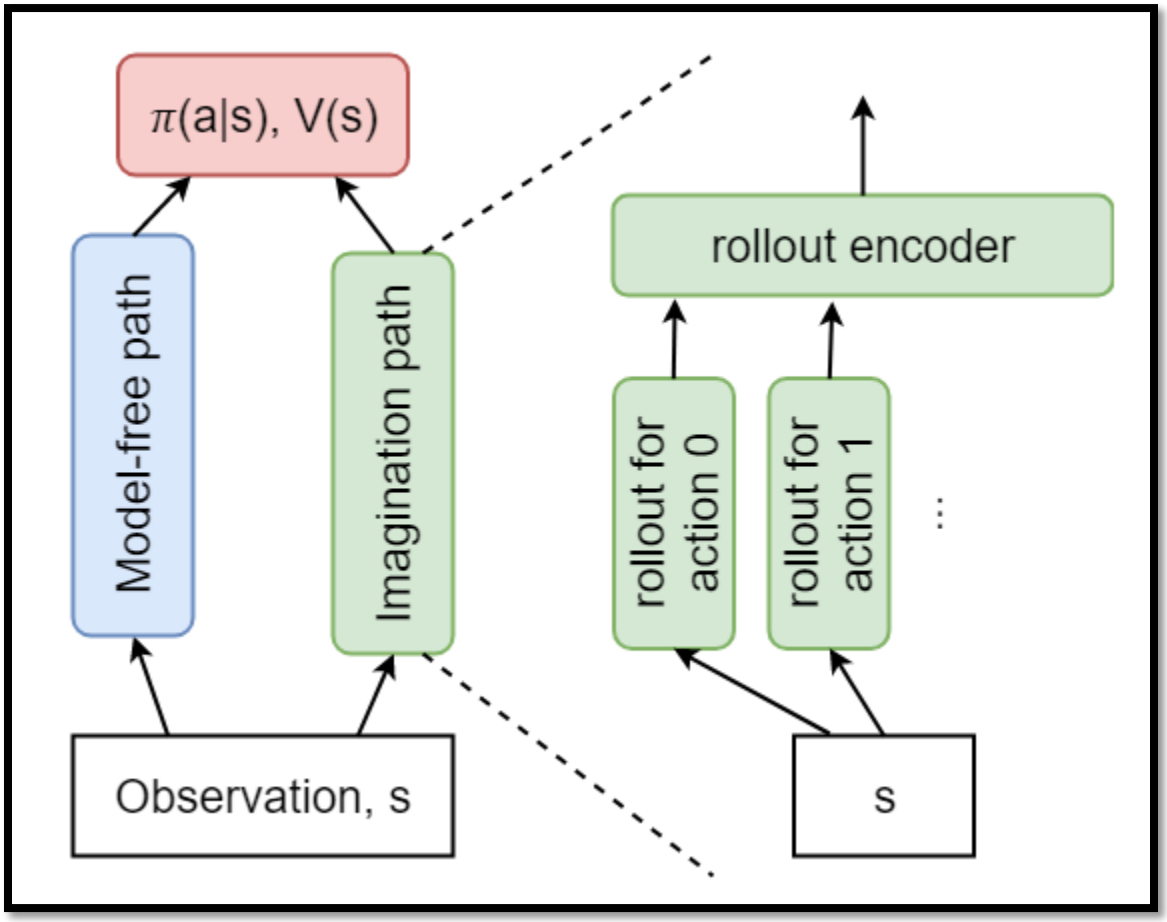
Chapter 16: Black-Box Optimization in RL

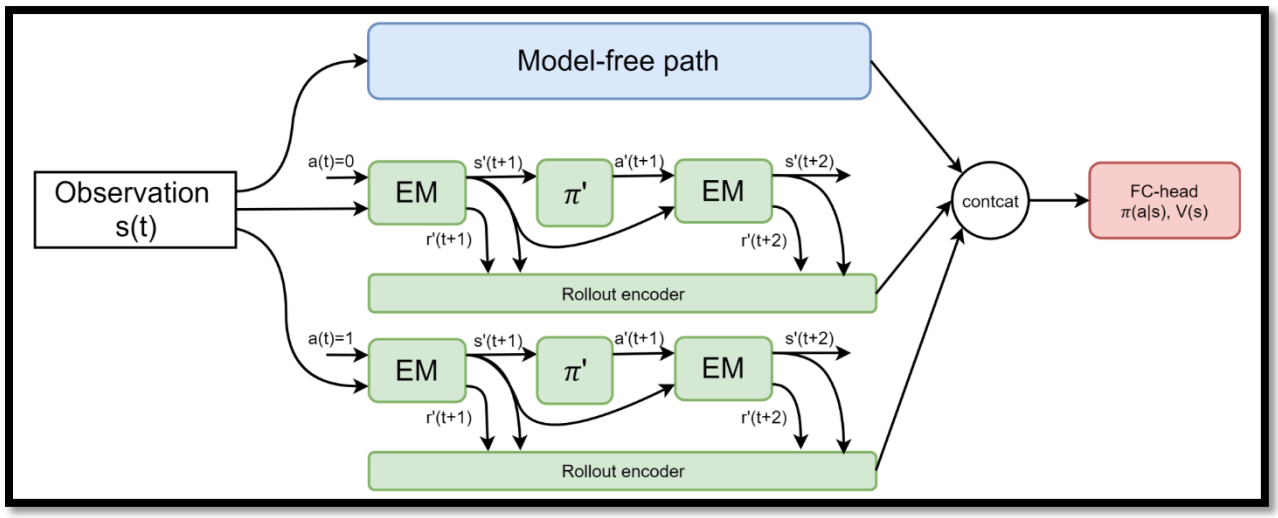


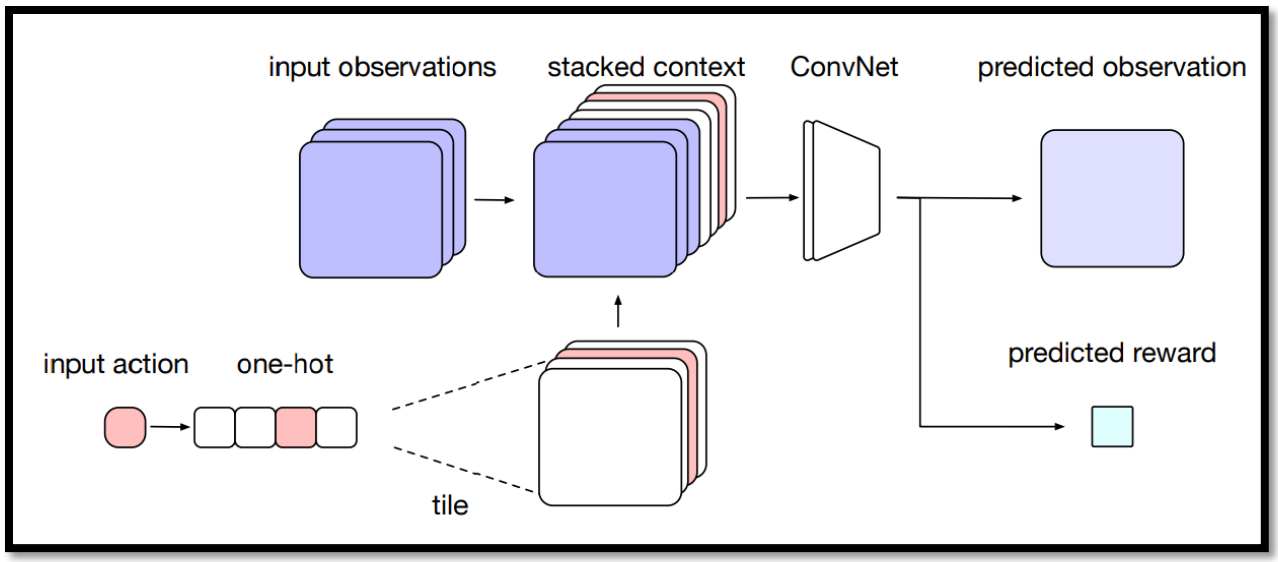


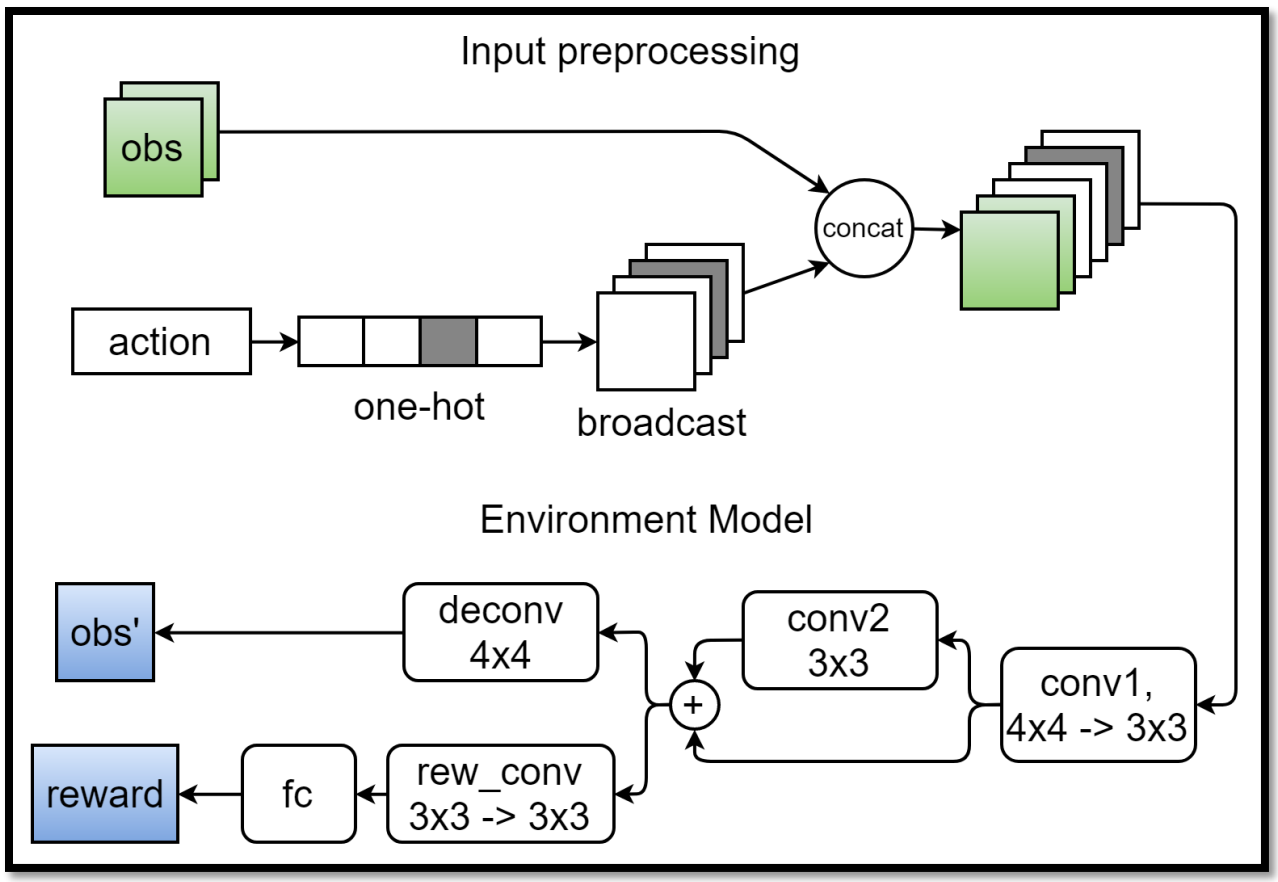

reward_max
1.20e+3
800
400
0.00
0.000
100.0
200.0
reward_mean
1.20e+3
800
400
0.00
0.000
100.0
200.0
reward_std
60.0
40.0
20.0
0.00
0.000
100.0
200.0

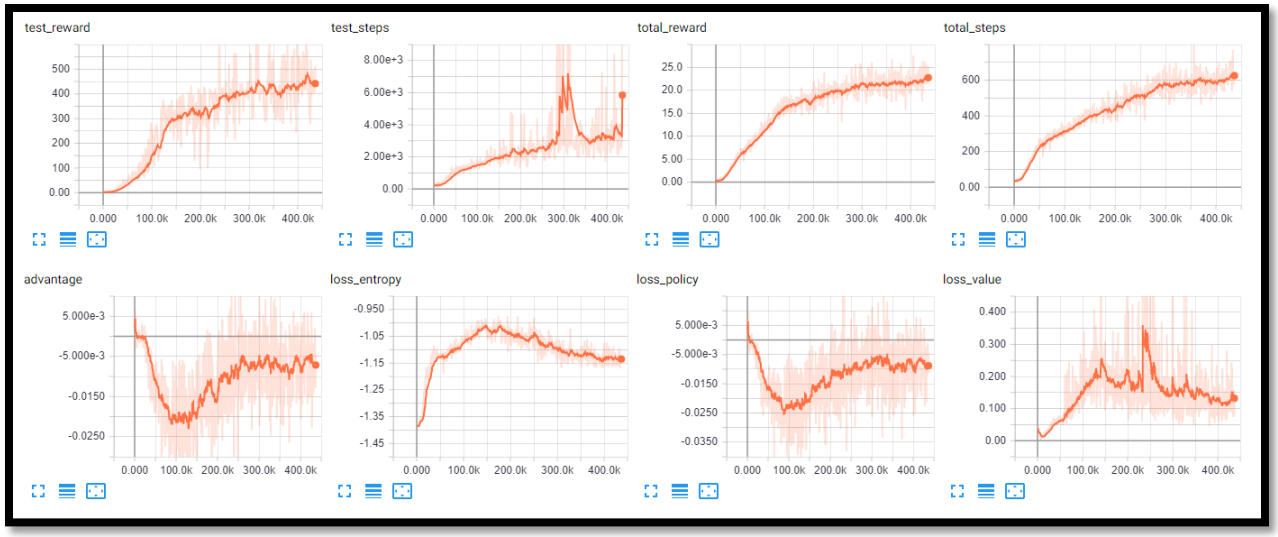
Chapter 17: Beyond Model-Free – Imagination

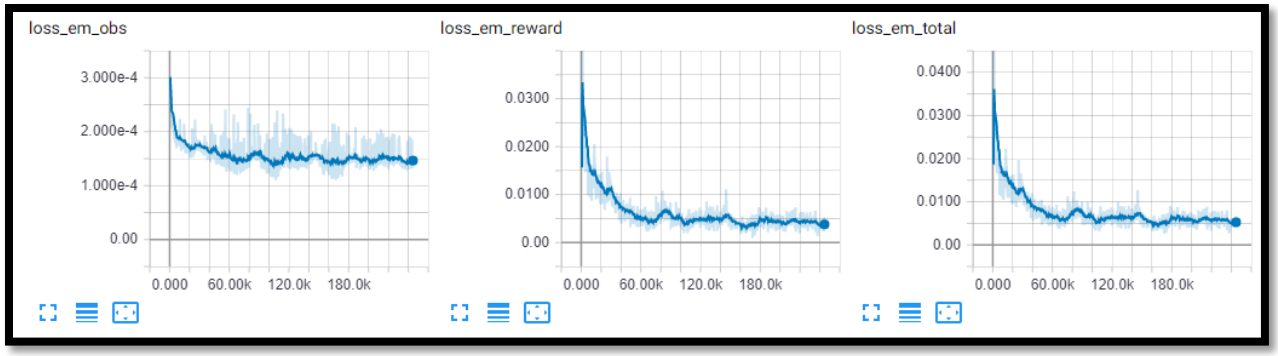


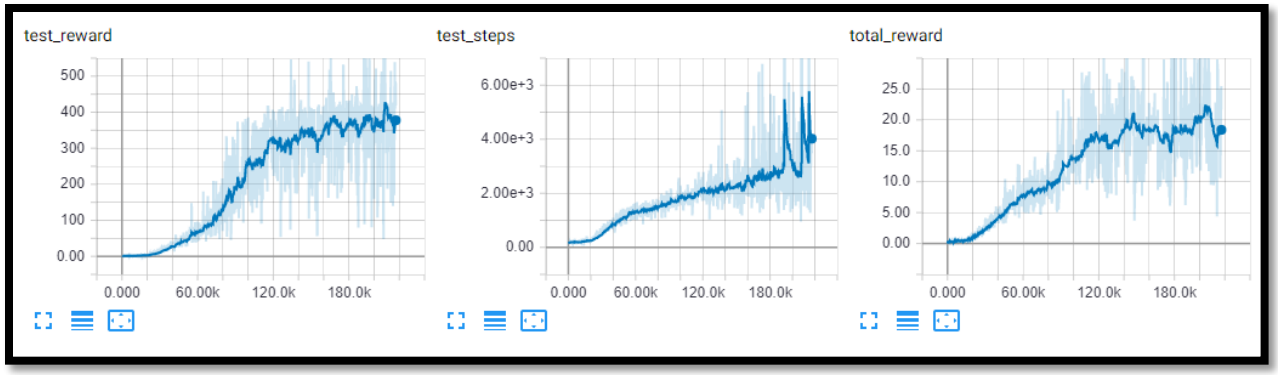


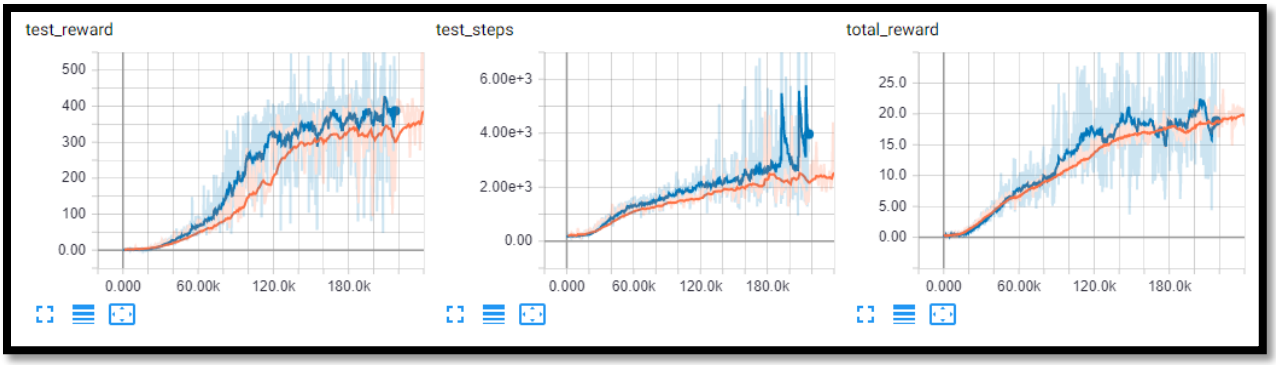

input observations
stacked context
ConvNet
predicted observation
input action
one-hot
predicted reward
tile


Input preprocessing
obs
concat
action
one-hot
broadcast
Environment Model
obs'
deconv
4x4
conv2
3x3
conv1,
4x4 -> 3x3
reward
fc
rew_conv
3x3 -> 3x3

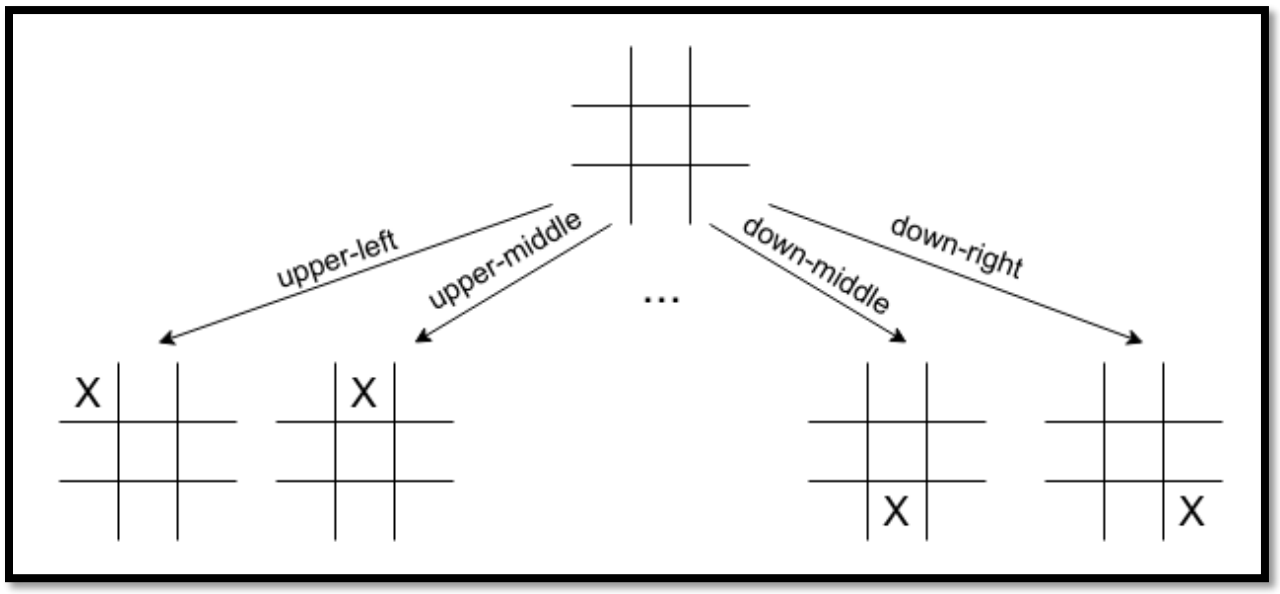

test_reward
test_steps
total_reward
total_steps
advantage
loss_entropy
loss_policy
loss_value

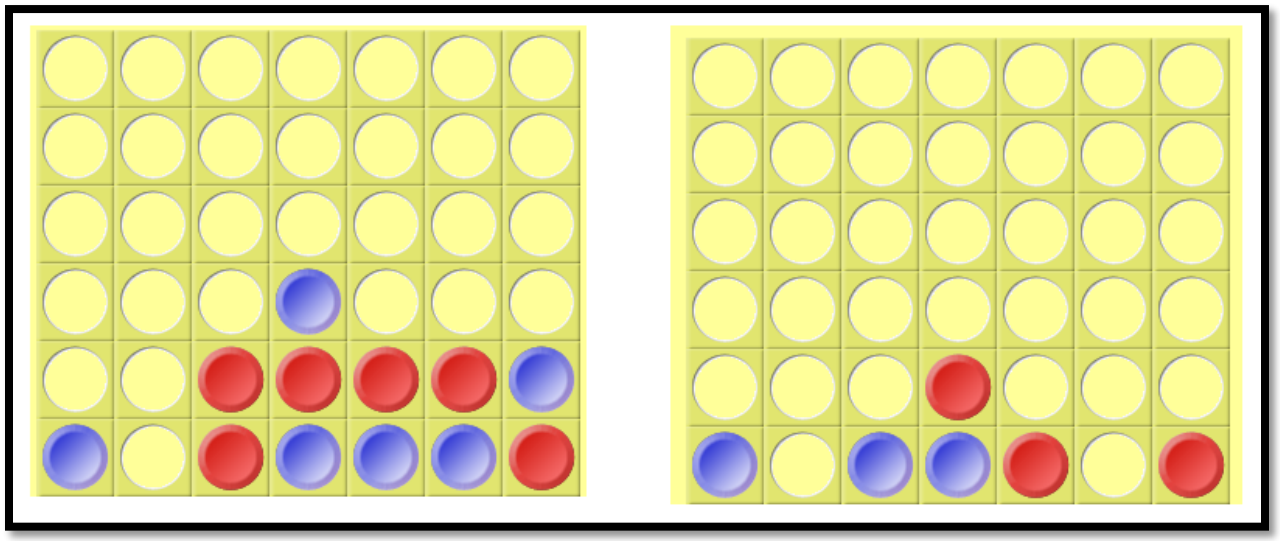

loss_em_obs
loss_em_reward
loss_em_total

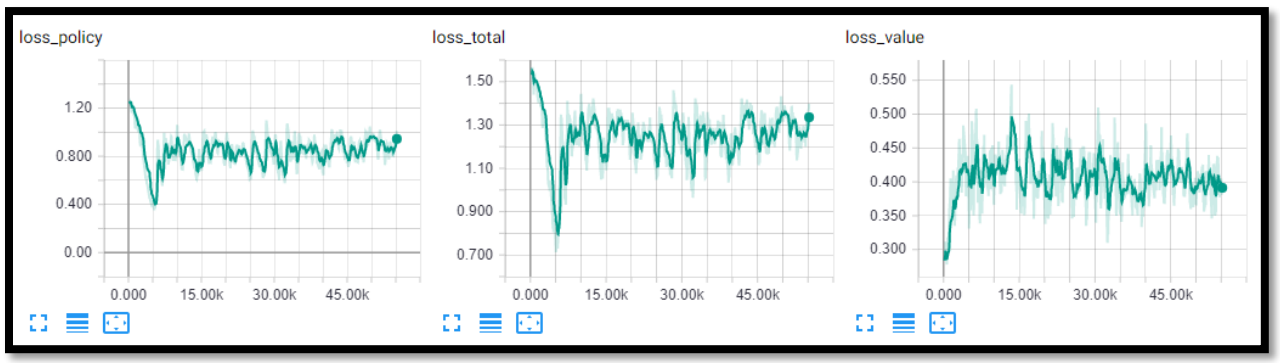

test_reward
test_steps
total_reward

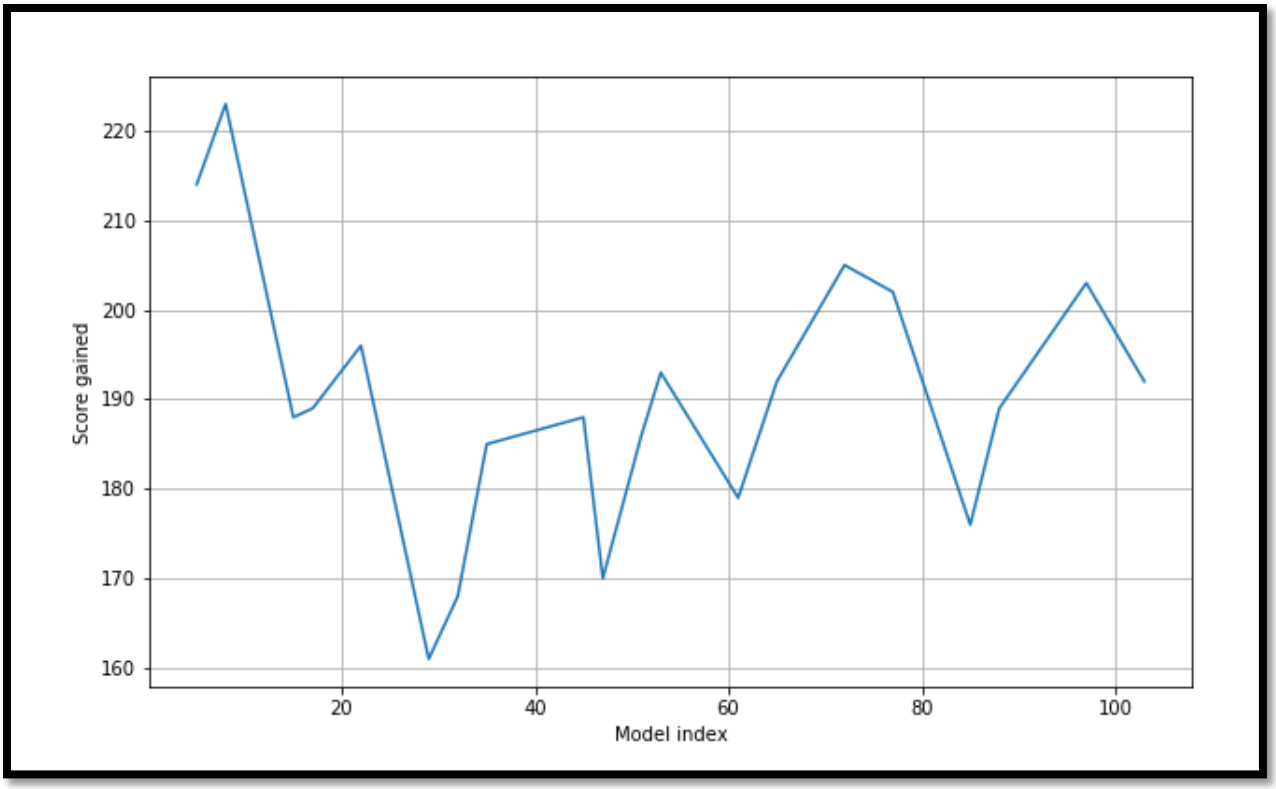

test_reward
500
400
300
200
100
0.00
0.000
60.00k
120.0k
180.0k
test_steps
6.00e+3
4.00e+3
2.00e+3
0.00
0.000
60.00k
120.0k
180.0k
total_reward
25.0
20.0
15.0
10.0
5.00
0.00
0.000
60.00k
120.0k
180.0k

Chapter 18: AlphaGo Zero








Score gained
220
210
200
190
180
170
160
20
40
60
80
100
Model index

Shmuma

/top

RB

rl_ch18_bot

```
Leader board
            Luklusik: won=31, lost=9, draw=0
  best_008_02500.dat: won=8, lost=8, draw=0
              Shmuma: won=6, lost=6, draw=0
  best_019_11300.dat: won=2, lost=4, draw=0
  best_015_09200.dat: won=2, lost=4, draw=0
  best_005_01900.dat: won=2, lost=6, draw=0
              mbilan: won=2, lost=3, draw=0
  best_007_02400.dat: won=1, lost=0, draw=0
             Roman8i: won=1, lost=3, draw=0
  best_010_05900.dat: won=1, lost=0, draw=0
  best_001_00800.dat: won=1, lost=2, draw=0
  best_014_08600.dat: won=1, lost=0, draw=0
  best_004_01600.dat: won=1, lost=0, draw=0
  best_009_05700.dat: won=1, lost=2, draw=0
  best_002_00900.dat: won=1, lost=5, draw=0
  best_022_12200.dat: won=1, lost=5, draw=0
  best_061_34100.dat: won=0, lost=2, draw=0
                None: won=0, lost=1, draw=0
  best_047_28000.dat: won=0, lost=2, draw=0
```