

RUTGERS UNIVERSITY  
School of Engineering  
Department of Electrical & Computer Engineering  
ECE 472 – Robotics & Computer Vision– Fall 2022

## Final Project - Reinforcement Learning: Super Mario

Name (last, first) : Mehmet Ali Soner

netID : mas996

RUID: 196000499

Date: December 12, 2022

## Problem 1

For this section, I have followed the CartPole tutorial from pytorch's official website. I am going through the details of that specific tutorial in this section. However, the performance of this model turned out to be unstable and poor. In order to get better performance and video results, I switched over to the github commit that was provided one of our fellow students. The results you see in the Jupyter Notebook and the performance metric plots are obtained through the **github commit** code. The specific link for that github repo is: [https://github.com/SiftingSands/tutorials/blob/DQN\\_revise\\_training/intermediate\\_source/reinforcement\\_q\\_learning.py](https://github.com/SiftingSands/tutorials/blob/DQN_revise_training/intermediate_source/reinforcement_q_learning.py). Now, the detailed of the pytorch tutorial:

```
1 !apt-get update
2 !apt-get install -y xvfb x11-utils
3 %pip install pyvirtualdisplay==0.2.*
4 %pip install gym[classic_control]
```

I am calling apt-get update since I ran into issues with xvfb couple of times and this seem to resolve it. We install xvfb and pyvirtualdisplay for the video capturing. Then we import the gym package and create the environment for "CartPole":

```
1 import gym
2 if gym.__version__ < '0.26':
3     env = gym.make('CartPole-v0', new_step_api=True,
4         ↪ render_mode='single_rgb_array').unwrapped
5 else:
6     env = gym.make('CartPole-v0', render_mode='rgb_array').unwrapped
```

After this, I will skip few things such as the video capturing and importing other various packages. I want to focus on the DQN's linear layer. In this model, we have to have two possible outputs in our final/linear layer since we will either push the cart to the right or to the left. In order to add this layer, we have to know the output size of each convolutional layer so that we have the right input size for the linear layer. That's why we call conv2d\_size\_out three times since we have three convolutional layers. We then calculate the input size and add it with:

```
1 class DQN(nn.Module):
2     # Architecture code (3 conv layers, 3 batch norms)
3     ...
4     def conv2d_size_out(size, kernel_size = 5, stride = 2):
5         return (size - (kernel_size - 1) - 1) // stride + 1
6     convw = conv2d_size_out(conv2d_size_out(conv2d_size_out(w)))
7     convh = conv2d_size_out(conv2d_size_out(conv2d_size_out(h)))
8     linear_input_size = convw * convh * 32
9     self.head = nn.Linear(linear_input_size, outputs)
```

We then implement two functions that will get the cart's location on the screen and will get the screen itself:

```
1 def get_cart_location(screen_width):
2     world_width = env.x_threshold * 2
3     scale = screen_width / world_width
4     return int(env.state[0] * scale + screen_width / 2.0) # MIDDLE OF
5     ↪ CART
6
7 def get_screen():
8     screen = env.render().transpose((2, 0, 1))
```

```

8      # Cart is in the lower half, so strip off the top and bottom of the
      ↪ screen
9      _, screen_height, screen_width = screen.shape
10     ...
11     cart_location = get_cart_location(screen_width)
12     ...
13     screen = screen[:, :, slice_range]

```

I have mainly included the important lines of the two functions in the snippet. These two functions are crucial since we can describe the current state of the agent through the location of the cart on the screen. We achieve this with some algebra and the information we have about the environment's/simulation's dimensions.

Now we get to the part where we train the DQN. First we initialize the network, the optimization function and the memory.

```

1  n_actions = env.action_space.n
2  policy_net = DQN(screen_height, screen_width, n_actions).to(device)
3  target_net = DQN(screen_height, screen_width, n_actions).to(device)
4  target_net.load_state_dict(policy_net.state_dict())
5  target_net.eval()
6
7  optimizer = optim.RMSprop(policy_net.parameters())
8  memory = ReplayMemory(10000)

```

*n\_actions* is the amount of actions we can take in the environment. In this case, it will be two (move cart to right or left). Then, we initialize the policy and the target network. This is mainly for stability since it serves as a "back-up" of the policy network. The target network is basically a copy of the policy network with frozen parameters that get updated once in a while (we copy the parameters into *target\_net* on line 5 and put the network in eval mode on line 6). So, if overfitting or other errors were to happen in the policy network, the target network is there for stability. On line 6, we initialize the memory by creating a *ReplayMemory* object. The *ReplayMemory* class looks like this:

```

1  Transition = namedtuple('Transition',
2                          ('state', 'action', 'next_state', 'reward'))
3  class ReplayMemory(object):
4
5      def __init__(self, capacity):
6          self.memory = deque([], maxlen=capacity)
7
8      def push(self, *args):
9          """Save a transition"""
10         self.memory.append(Transition(*args))
11
12     def sample(self, batch_size):
13         return random.sample(self.memory, batch_size)
14
15     def __len__(self):
16         return len(self.memory)

```

The *ReplayMemory* class can hold *Transition* and can return random *Transitions* of a certain batch size. This is crucial since the policy network will learn from recent *Transition* tuples in the *ReplayMemory* object. It will use that information to optimize the model, to achieve a better policy. So, the memory object serves as the memory for the training.

Next, I will explain some details in the *select\_action()* and *optimize\_model()* function. The way we choose an action is the following:

```

1 def select_action(state):
2     ...
3     sample = random.random()
4     eps_threshold = EPS_END + (EPS_START - EPS_END) * \
5         math.exp(-1. * steps_done / EPS_DECAY)
6     steps_done += 1
7     if sample > eps_threshold:
8         with torch.no_grad():
9             return policy_net(state).max(1)[1].view(1, 1)
10    else:
11        return torch.tensor([[random.randrange(n_actions)]],
12                               ↪ device=device, dtype=torch.long)

```

Given an input state, we select the action. First, we get a random value and compare it to our epsilon threshold. If this random value is over the threshold, we go through the policy net with the input state and take the action with the highest reward. If the random value doesn't cross the threshold, we then select a random action out of the possible actions in this environment (right, left). With this function, we can now define the model optimization function.

The `optimize_model()` function is where the RL algorithm is implemented. First, I will explain how we "load" the necessary data.

```

1 def optimize_model():
2     if len(memory) < BATCH_SIZE:
3         return
4     transitions = memory.sample(BATCH_SIZE)
5     batch = Transition(*zip(*transitions))

```

We first try to get a batch of Transitions into the model. If there isn't enough Transitions in the memory, we return. If we have enough information stored in the memory, we load a sample from there. Then, we create a batch variable from the transitions variable. The batch holds all the information regarding the state, action, reward and next state. Then, we divide each information from the batch into individual variables.

```

1 non_final_mask = torch.tensor(tuple(map(lambda s: s is not None,
2                                           batch.next_state)),
3                                ↪ device=device,
4                                ↪ dtype=torch.bool)
5 non_final_next_states = torch.cat([s for s in batch.next_state
6                                     if s is not None])
7 state_batch = torch.cat(batch.state)
8 action_batch = torch.cat(batch.action)
9 reward_batch = torch.cat(batch.reward)

```

On line 1, we are extracting every next state that is not final, which is not None and storing them as a boolean. True for every non-final state and False for final states. This will be helpful afterwards when we will compute the values of the next states. On line 3, we do same thing, except we store the values of the states. Line 5-7, we extract information about states, rewards and actions from the batch and store them into individual variables. Afterwards, we compute the state action values like this:

```

1 state_action_values = policy_net(state_batch).gather(1, action_batch)

```

We pass through the policy net with the state batch as the input and observe, which action the policy net would have taken. We get these from our `action_batch` variable. Next step is to calculate the next state values:

```

1 next_state_values = torch.zeros(BATCH_SIZE, device=device)
2 next_state_values[non_final_mask] =
  → target_net(non_final_next_states).max(1)[0].detach()

```

On line 1, we allocate space for the results for our computation we are going to perform on line 2. On line 2, we pass through the target net with obtained next\_state\_values variable and extract the result/action that has the highest reward with the max function. We are using the target net for the improved stability that we mentioned before. The target net that is called on this line is the "old" target net that is constant. Another detail worth explaining is the significance of the non\_final\_mask variable. This is needed here since we will store a value in to next\_state\_values if the state was non-final and 0 if the state was final. This is determined by the values in the non\_final\_mask variable. The last important line of code in this function is the following:

```

1 expected_state_action_values = (next_state_values * GAMMA) +
  → reward_batch

```

This is where we actually compute which action to take in order to get to the next state for each state. We multiply the next\_state\_values by gamma in order to guarantee convergence and add the rewards to it. This equation can be found under the DQN algorithm page in the tutorial, where the policy function is described.

For the training loop, the main structure looks like this:

```

1 num_episodes = any_number
2 for i_episode in range(num_episodes):
3     # Initialize the environment and state
4     env.reset()
5     last_screen = get_screen()
6     current_screen = get_screen()
7     state = current_screen - last_screen
8     ...
9     for t in count():
10        # Select and perform an action
11        action = select_action(state)
12        _, reward, done, _, _ = env.step(action.item())
13        reward = torch.tensor([reward], device=device)
14
15        # Observe new state
16        last_screen = current_screen
17        current_screen = get_screen()
18        if not done:
19            next_state = current_screen - last_screen
20        else:
21            next_state = None
22
23        # Store the transition in memory
24        memory.push(state, action, next_state, reward)
25
26        # Move to the next state
27        state = next_state
28
29        # Perform one step of the optimization (on the policy network)
30        optimize_model()
31        if done:
32            break
33

```

```

34     # Update the target network, copying all weights and biases in
    ↪     DQN
35     if t % TARGET_UPDATE == 0:
36         target_net.load_state_dict(policy_net.state_dict())

```

We iterate through the model for `num_episodes` times. In each episode, we first reset the environment, initialize the screen, get the current and the last screen, which are identical at the start. The main reason why we reset the environment is because episodes are finished after a certain time period or they fail. So, we have to start again. Then on line 9, we start the episode. We call the `select_action()` that will return an action that we will input into the environment on line 12. This line of code steps through the environment and applies the inputted action in the environment. As the output, we get the reward and a boolean that indicates if this was a final state. We update the states by storing the `current_screen` into `last_screen`. This can be applied since we simulated the action in the environment on line 12, thus changed the actual current state. The `current_screen` variable holds the old screen after this. In order to update `current_screen`, we just call `get_screen()`. Based on the `done` variable, we either restart or we calculate the `next_state`, which is the difference between the current and the last screen, and transition into that. Before going to the next iteration in the loop, we save the necessary data from this transition and call `optimize_model()`. Every now and then we also update the target net's parameters as seen on line 35-36.

## Problem 2

In reinforcement learning, we want to create a policy function  $\pi$  that will maximize the reward in a given state  $s$  after performing some action  $a$  in a certain environment. The function that computes the reward from  $s$  and  $a$  is the  $Q$  function. In the CartPole case,

- the states:
  - the cart's x-coordinate
  - cart's velocity
  - the pole's angular velocity
  - the pole's angle
- the environment: the simulation that is simulating cart with the pole that is on a rail
- actions:
  - move cart to right
  - move cart to left
- reward:
  - reward = 1, when the pole stays within the specified angle ( $\pm 12^\circ$ ) and cart's x-coordinate is in range of (-2.4, 2.4)
  - reward = 0, when the above mentioned conditions are not met, then we fail and restart the environment

Essentially we want:

$$\pi^*(s) = \operatorname{argmax}_a Q^*(s, a)$$

However, we cannot achieve this since we don't know every possible state, action and reward in the world. What we do instead is, we try to approximate the  $Q$  function with the help of neural networks. We assume that the  $Q$  function is a Bellman equation and can write it like this:

$$Q^\pi(s, a) = r_0 + \gamma Q^\pi(s', \pi(s'))$$

This equation describes that the current  $Q$  function is the sum of the current reward  $r_0$  and the  $Q$  function of the future state  $s'$  and action  $a' = \pi(s')$  multiplied by a discount rate  $\gamma$ . This is a very powerful tool since we can describe the  $Q$  function recursively.

We can rewrite this equation into what will be inputted into the Huber loss. For the Huber loss function, we input what is called the temporal difference error,  $\delta$ :

$$\delta = Q(s, a) - \left( r_0 + \gamma \max_a Q(s', \pi(s')) \right)$$

The first term of the subtraction is the prediction, the values that we get when we take an action  $a$  in the state  $s$ . We get this value when we are inputting the state\_batch into the policy\_net to get the state\_action\_values. The second term of the subtraction is the target value that we obtain by passing through the target\_net with non\_final\_next\_states. The results of that pass through are the next\_state\_values. We then multiply the next\_state\_values by  $\gamma$  and add the reward batch to it.  $s'$  would be the future state,  $a'$  the future action. We want to minimize this error since we want to have our prediction and target values as close as possible to each other. We do that by inputting that into a Huber loss function.

### Problem 3

For performance metrics, we can observe the duration of each episode or the reward. This gives us basic understanding of how well the model performs. For every action where the pole is still upright, we add +1 to the reward variable we have or we add up every action that didn't result into a final state and include the terminating state as well. We do this with a for loop and the iteration variable in the loop condition:

```

1  for t in count():
2      if isRandom == True:
3          action = torch.tensor([env.action_space.sample()],
4                                  ↪ device=device, dtype=torch.long)
5      else:
6          action = select_action(state, policy_net)
7          observation, reward, terminated, truncated, _ =
8              ↪ env.step(action.item())
9          reward = torch.tensor([reward], device=device)
10         done = terminated or truncated
11
12         ...
13
14         if done:
15             episode_durations.append(t + 1)
16             break

```

This is how the loop for each episode looks like. When we have a terminal state, then we go into the if statement on line 12 and add how many steps we have taken, which equals to the duration of one episode. We then plot these duration with a helper plotting function. Here are the result of three scenarios, namely:

1. A random agent with num\_episodes = 600
2. A trained agent with num\_episodes = 100
3. A trained agent with num\_episodes = 600

Here are the results in that order:

The results are as expected. The worst performance out of the three is with the random agent, the second worst is the agent with the small training time (or less number of episodes) and the best performance comes from the trained agent with the highest training time. In Fig.1, we can observe that the episode duration on average stays constant and has a value of 25-30. This makes sense since the agent

doesn't update its policy net efficiently because it is selecting random action. There are some outliers that happen (peak at 79 and 70) because of the randomness of the action selection module. The module might take the all of the right actions in one episode. However, the probability of that happening is low because it has to randomly select all of the actions successively and it is reflected in this plot.

The second figure, Fig.2, shows more promising results. The performance in the first 70 episodes is in fact worse than the first 70 episodes with a random agent but there is sharp jump in performance around episode 80. Here, we can see that the performance significantly increases and stays at this level. We can also conclude that training an agent for 100 episodes isn't quite enough since it doesn't pass the "solved" threshold of 200 in this environment. The results seem promising however and we can predict that the duration of each episode is going to increase with more training. This agent outperforms the random agent even though it had 1/6 of episodes to train in.

The third figure, Fig.3, proves our prediction we had with Fig.2. Again, around 100 episodes there is a sharp peak that seems to be constant until episode = 200. From there, there is a linear increase in performance until episode = 450 from where the performance increase improves even further. The performance increase is a lot steeper. It is also around episode = 450, where the module passes the "solved" threshold of 200. After some time, the module cannot improve any further and it stagnates at around episode = 550. This is all expected and seems ordinary. What is surprising is that both train models had much worse performance in the first 60-80 episodes than the random agent. The performance in the first hundred episode also seem to decrease when we increase the number of episodes we want the model to simulate. This doesn't matter since we can arrive a higher episode duration or overall reward in the long run.

## Problem 4

Here are three other problems that can be solved with RL on open-AI gym:

### 1. Mountain Car (Classic Control):

- State: In this environment, there are only two states. The position of the car along the x-axis and the velocity of the car. The x-coordinate and the velocity of the car can from  $(-\infty, \infty)$ .
- Action: There are three actions in this environment. Accelerate to the right, accelerate to the left or don't accelerate at all.
- Environment: A car is at the bottom of a "sinusoidal valley" and has to get out of that valley by reaching the flag at the top of the hill. The starting position of the car is a random value in the range of  $[-0.6, -0.4]$ . The episode ends when the car's x-coordinate is greater than or equal to 0.5, which means it has reached the flag or when the episode length is over 200.
- Reward: In this environment, the reward is -1 for each timestep where the car hasn't reached the flag on top of the hill. So, the car has to reach the flag as quickly as possible.

### 2. Car Racing (Box2D):

- State: In this environment, the observation space consists of 96x96 pixels, which determine the track and the position of the racing car.
- Action: There are three actions if the environment is continuous and 5 if discrete. In the continuous scenario we have: steering (-1 full left, +1 full right), gas and break. In the discrete scenario, we have: do nothing, turn right, turn left, gas, brake.
- Environment: A racing car is on a racing track and is trying to complete it. It starts at the starting point of the track, in the middle of center of the road. The episode ends when all tiles have been visited. The racing car can go off the track but it will get rewarded with -100 and "die" per the Gym documentation.
- Reward: In this environment, the reward is -0.1 for every frame that the car took to get to the episode's end and  $+1000/N$ , where N is the amount of tiles that the car visited.



3. : Frozen Lake (Toy Text):

- State: In this environment, the observation space consists of one numeric value, which indicates the position of the agent on the lake. It can be calculated with  $agentposition = current\_row * n\_rows + current\_col$ .
- Action: There are four actions that the agent can take in this environment. Go up, left, down and right.
- Environment: An agent has to reach a goal  $G$  on a frozen lake with holes  $H$  from the starting point  $S$ . The main objective is to reach the goal **without** falling into one of the holes. The agent's movement may not be clear since it will not move into the direction of the goal all the time. This is simply because of the slippery nature of the frozen lake.
- Reward: In this environment, the only time you get a +1 is when you reach the goal. All other actions resulting into reaching a hole or any frozen tile, "pay out" a 0 as reward.