R E P O R T

[인공지능특강] Homework2

과목명	인공지능특강
분반	1 분반
교수	한연희
학번	2331036013
이름	조 재 민
제출일	2024년 06월 07일 금요일

서론) Homework의 내용과 목적

본 과제의 목적은 Deep Q-Network(DQN) 알고리즘의 파생 모델인 Double DQN과 Dueling DQN을 실제 코드로 구현하고, 이들의 성능을 비교하는 데 있다. 이러한 비교를 통해 각 알고리즘의 장단점을 명확히 파악하고, 다양한 강화 학습 문제에서의 적용 가능성을 평가하고자 한다. 각 알고리즘의 공정한 평가를 위하여 각각 학습을 5회 수행하며, 하이퍼파라미터는 모두 동일하게 설정하였다.

```
config = {
    "env_name": ENV_NAME,                          # 환경의 이름
    "max_num_episodes": 1_500,                     # 훈련을 위한 최대 에피소드 횟수
    "batch_size": 32,                              # 훈련시 배치에서 한번에 가져오는 랜덤 배치 사이즈
    "learning_rate": 0.0001,                       # 학습율
    "gamma": 0.99,                                 # 감가율
    "steps_between_train": 1,                      # 훈련 사이의 환경 스텝 수
    "target_sync_step_interval": 500,              # 기존 Q 모델을 타깃 Q 모델로 동기화시키는 step 간격
    "replay_buffer_size": 300_000,                  # 리플레이 버퍼 사이즈
    "epsilon_start": 0.95,                         # Epsilon 초기 값
    "epsilon_end": 0.01,                           # Epsilon 최종 값
    "epsilon_final_scheduled_percent": 0.75,       # Epsilon 최종 값으로 스케줄되는 마지막 에피소드 비율
    "print_episode_interval": 10,                  # Episode 통계 출력에 관한 에피소드 간격
    "train_num_episodes_before_next_test": 50,                  # 검증 사이 마다 각 훈련 episode 간격
    "validation_num_episodes": 3,                  # 검증에 수행하는 에피소드 횟수
    "episode_reward_avg_solved": 200,              # 훈련 종료를 위한 검증 에피소드 리워드의 Average
}
```

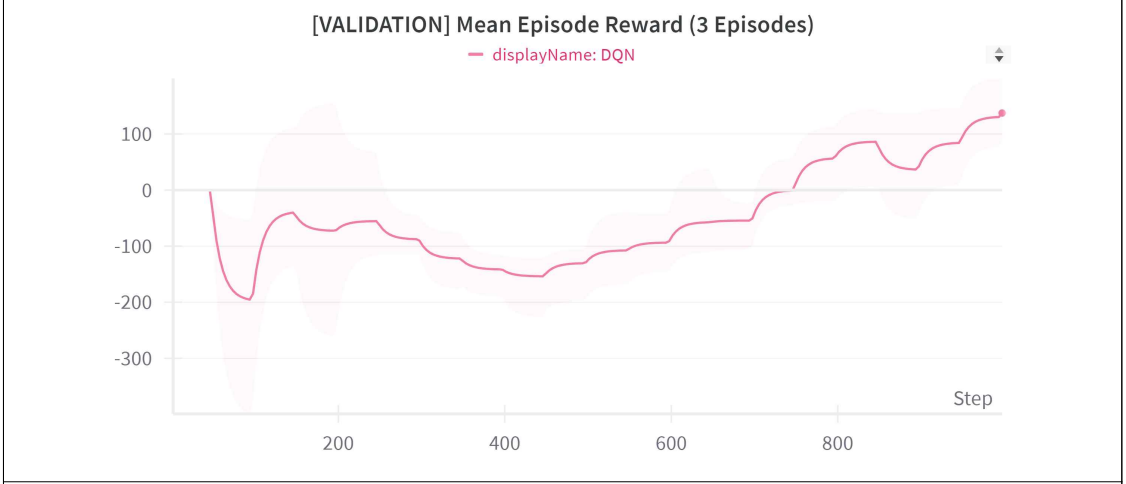
<하이퍼파라미터 설정값>

본론)

본론1_DQN)

```
[Episode  10, Time Steps    956] Episode Reward: -103.57379534683007, Replay buffer:    956, Loss:  9.079, Epsilon: 0.94, Training Steps:   924, Elapsed Time: 00:00:01
[Episode  20, Time Steps  1,943] Episode Reward: -282.82604940242925, Replay buffer:  1,943, Loss: 223.810, Epsilon: 0.93, Training Steps: 1,911, Elapsed Time: 00:00:04
[Episode  30, Time Steps  2,916] Episode Reward: -55.209659140035654, Replay buffer:  2,916, Loss:  4.520, Epsilon: 0.92, Training Steps: 2,884, Elapsed Time: 00:00:06
[Episode  40, Time Steps  3,926] Episode Reward: -228.76018763781292, Replay buffer:  3,926, Loss:  3.967, Epsilon: 0.92, Training Steps: 3,894, Elapsed Time: 00:00:08
[Episode  50, Time Steps  4,939] Episode Reward: -239.78065746191814, Replay buffer:  4,939, Loss:  2.109, Epsilon: 0.91, Training Steps: 4,907, Elapsed Time: 00:00:11
[Validation Episode Reward: [-482.36869113 -259.74303945 -350.34145356]] Average: -364.151
[Episode  60, Time Steps  5,829] Episode Reward: -300.87579215155284, Replay buffer:  5,829, Loss:  9.626, Epsilon: 0.90, Training Steps: 5,797, Elapsed Time: 00:00:13
[Episode  70, Time Steps  7,053] Episode Reward: -190.41808179387618, Replay buffer:  7,053, Loss: 36.218, Epsilon: 0.89, Training Steps: 7,021, Elapsed Time: 00:00:16
[Episode  80, Time Steps  7,959] Episode Reward: -217.7814053400475, Replay buffer:  7,959, Loss:  2.388, Epsilon: 0.88, Training Steps: 7,927, Elapsed Time: 00:00:18
[Episode  90, Time Steps  9,013] Episode Reward: -138.69177794395617, Replay buffer:  9,013, Loss:  4.154, Epsilon: 0.87, Training Steps: 8,981, Elapsed Time: 00:00:23
[Episode 100, Time Steps 10,049] Episode Reward: -18.04327681865459, Replay buffer: 10,049, Loss: 46.814, Epsilon: 0.87, Training Steps: 10,017, Elapsed Time: 00:00:25
[Validation Episode Reward: [-171.96301358 -197.07853398 -103.16955719]] Average: -157.404
[Episode 110, Time Steps 11,248] Episode Reward: -75.26637019305156, Replay buffer: 11,248, Loss: 14.533, Epsilon: 0.86, Training Steps: 11,216, Elapsed Time: 00:00:29
[Episode 120, Time Steps 12,360] Episode Reward: -146.1801745509864, Replay buffer: 12,360, Loss:  7.109, Epsilon: 0.85, Training Steps: 12,328, Elapsed Time: 00:00:32
[Episode 130, Time Steps 13,389] Episode Reward: -133.57626115089187, Replay buffer: 13,389, Loss: 15.370, Epsilon: 0.84, Training Steps: 13,357, Elapsed Time: 00:00:35
[Episode 140, Time Steps 14,365] Episode Reward: -116.37772649677348, Replay buffer: 14,365, Loss: 19.962, Epsilon: 0.83, Training Steps: 14,333, Elapsed Time: 00:00:38
[Episode 150, Time Steps 15,332] Episode Reward: -130.86590326515844, Replay buffer: 15,332, Loss:  7.482, Epsilon: 0.82, Training Steps: 15,300, Elapsed Time: 00:00:40
[Validation Episode Reward: [ -95.89984867 -120.56384624  -86.26824984]] Average: -100.911
```

Train - Terminal image captured



[VALIDATION] Mean Episode Reward (3 Episodes)
학습 5회 수행, 스무딩 강도 : 0.5

DQN 학습 과정은 Homework1에서 수행했던 코드와 동일하게 진행했으므로 설명은 생략함

본론2_Double_DQN)

$$Q_\theta(s,a) \leftarrow Q_\theta(s,a) + \alpha(r + \gamma \max_{a'} Q_{\bar{\theta}}(s',a') - Q_\theta(s,a))$$

```
with torch.no_grad():
    q_prime_out = self.target_q(next_observations)
    # next_state_values.shape: torch.Size([32, 1])
    max_q_prime = q_prime_out.max(dim=1, keepdim=True).values
    max_q_prime[dones] = 0.0

    # target_state_action_values.shape: torch.Size([32, 1])
    targets = rewards + self.gamma * max_q_prime
```

<DQN의 Q값 추정 방식>

↓

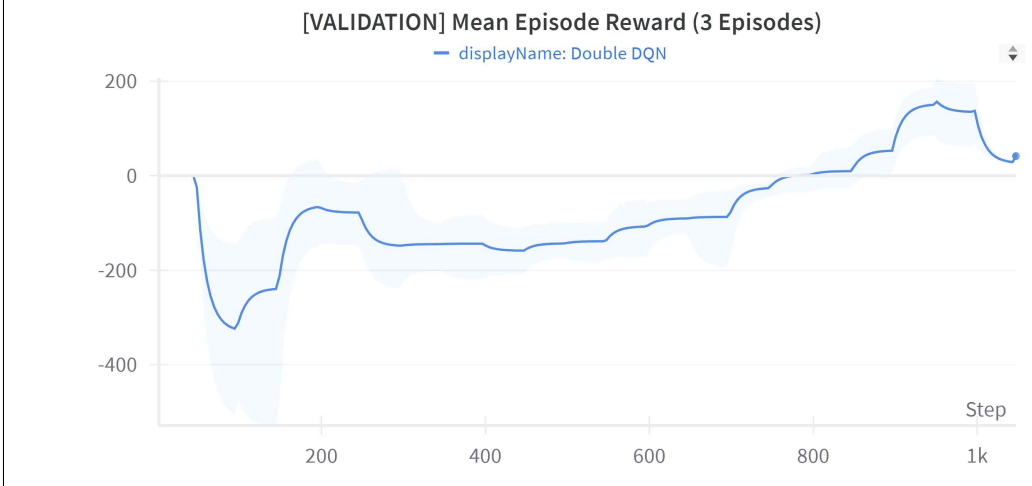
$$Q_\theta(s,a) \leftarrow Q_\theta(s,a) + \alpha(r + \gamma Q_{\bar{\theta}}\left(s', arg\max_{a'} Q_\theta(s',a')\right) - Q_\theta(s,a))$$

```
with torch.no_grad():
    q_prime_out = self.target_q(next_observations)
    # next_state_values.shape: torch.Size([32, 1])
    next_actions = q_prime_out.argmax(dim=1, keepdim=True)

    q_target_out = self.target_q(next_observations)

    max_q_prime = q_target_out.gather(dim=-1, index=next_actions)
    max_q_prime[dones] = 0.0

    # target_state_action_values.shape: torch.Size([32, 1])
    targets = rewards + self.gamma * max_q_prime
```

<Double DQN의 Q값 추정 방식>

```
[Episode 310, Time Steps 31,757] Episode Reward: -112.31874559541262, Replay buffer: 31,757, Loss:  1.700, Epsilon: 0.69, Training Steps: 31,725, Elapsed Time: 00:01:20
[Episode 320, Time Steps 32,949] Episode Reward: -91.09771109327644, Replay buffer: 32,949, Loss:  1.532, Epsilon: 0.68, Training Steps: 32,917, Elapsed Time: 00:01:23
[Episode 330, Time Steps 34,146] Episode Reward: -34.12640168205945, Replay buffer: 34,146, Loss:  2.551, Epsilon: 0.67, Training Steps: 34,114, Elapsed Time: 00:01:27
[Episode 340, Time Steps 35,368] Episode Reward: -7.525960889716757, Replay buffer: 35,368, Loss:  4.007, Epsilon: 0.67, Training Steps: 35,336, Elapsed Time: 00:01:31
[Episode 350, Time Steps 36,482] Episode Reward: -54.66872752976951, Replay buffer: 36,482, Loss: 20.195, Epsilon: 0.66, Training Steps: 36,450, Elapsed Time: 00:01:34
[Validation Episode Reward: [-236.86762065 -153.31313749 -179.91960232]] Average: -190.033
[Episode 360, Time Steps 37,535] Episode Reward: -63.495640758061995, Replay buffer: 37,535, Loss: 230.088, Epsilon: 0.65, Training Steps: 37,503, Elapsed Time: 00:01:37
[Episode 370, Time Steps 38,692] Episode Reward: -12.736665209207132, Replay buffer: 38,692, Loss: 426.089, Epsilon: 0.64, Training Steps: 38,660, Elapsed Time: 00:01:41
[Episode 380, Time Steps 39,954] Episode Reward: -25.470234643545155, Replay buffer: 39,954, Loss:  2.945, Epsilon: 0.63, Training Steps: 39,922, Elapsed Time: 00:01:45
[Episode 390, Time Steps 41,171] Episode Reward: -96.63995299096165, Replay buffer: 41,171, Loss: 19.894, Epsilon: 0.62, Training Steps: 41,139, Elapsed Time: 00:01:48
[Episode 400, Time Steps 43,288] Episode Reward: -78.97359434531762, Replay buffer: 43,288, Loss:  2.911, Epsilon: 0.62, Training Steps: 43,256, Elapsed Time: 00:01:55
[Validation Episode Reward: [-457.02455815  -62.45792508 -161.13281831]] Average: -226.872
[Episode 410, Time Steps 44,558] Episode Reward: -123.19755167629275, Replay buffer: 44,558, Loss:  2.077, Epsilon: 0.61, Training Steps: 44,526, Elapsed Time: 00:01:59
[Episode 420, Time Steps 45,677] Episode Reward: 18.173268356580962, Replay buffer: 45,677, Loss:  1.303, Epsilon: 0.60, Training Steps: 45,645, Elapsed Time: 00:02:02
[Episode 430, Time Steps 47,565] Episode Reward: -142.0541297796347, Replay buffer: 47,565, Loss: 53.279, Epsilon: 0.59, Training Steps: 47,533, Elapsed Time: 00:02:08
[Episode 440, Time Steps 49,731] Episode Reward: -14.172893264340843, Replay buffer: 49,731, Loss:  2.253, Epsilon: 0.58, Training Steps: 49,699, Elapsed Time: 00:02:15
[Episode 450, Time Steps 50,987] Episode Reward: 11.672309477640894, Replay buffer: 50,987, Loss: 14.623, Epsilon: 0.57, Training Steps: 50,955, Elapsed Time: 00:02:19
[Validation Episode Reward: [-383.32934801  138.12188764  -87.82774697]] Average: -111.012
```

Train - Terminal image captured



[VALIDATION] Mean Episode Reward (3 Episodes)

학습 5회 수행, 스무딩 강도 : 0.5

본론3_Dueling_DQN)

$$Q(s,a) = V(s) + (A(s,a) - \frac{1}{|A|}\sum_{a'} A(s,a'))$$

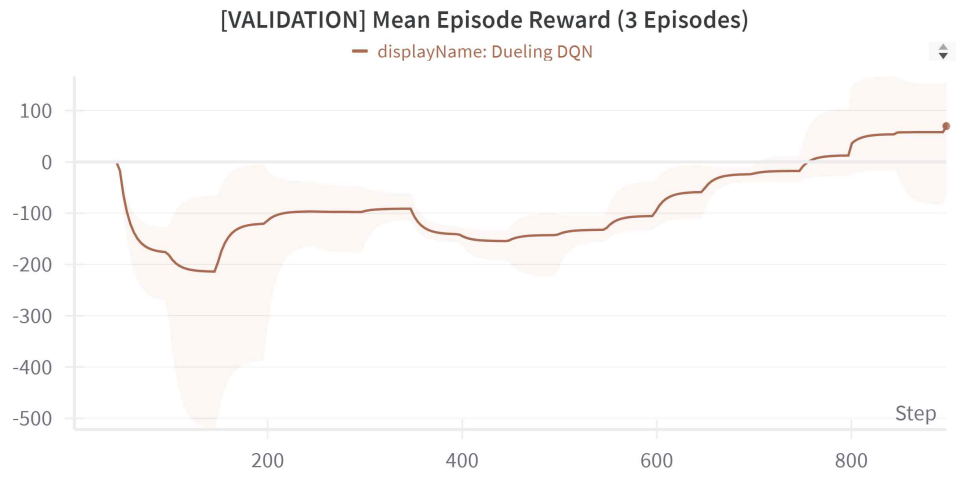
<Dueling DQN의 Q값 추정 방식>

```python
class QNet(nn.Module):
    def __init__(self, n_features, n_actions):
        super(QNet, self).__init__()
        self.n_features = n_features
        self.n_actions = n_actions

        # 기존 DQN과 달리, Dueling DQN을 위해 두 개의 스트림을 추가
        self.fc1 = nn.Linear(n_features, 128)
        self.fc2 = nn.Linear(128, 128)

        self.value_stream = nn.Linear(128, 1)
        self.advantage_stream = nn.Linear(128, n_actions)

    def forward(self, x):
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))

        # Dueling DQN: Value와 Advantage를 계산하여 Q 값을 산정
        value = self.value_stream(x)
        advantage = self.advantage_stream(x)

        # Dueling DQN의 수식을 반영
        q_values = value + (advantage - advantage.mean(dim=1, keepdim=True))
        return q_values

    def get_action(self, state, epsilon):
        if np.random.rand() < epsilon:
            return np.random.randint(self.n_actions)
        else:
            state = torch.FloatTensor(state).unsqueeze(0)
            q_values = self.forward(state)
            return q_values.argmax().item()
```

<Dueling DQN의 Q값 추정을 위해 Q-Network 및 산식을 수정함>

```
[Episode 360, Time Steps 38,517] Episode Reward: -56.39239649285044, Replay buffer: 38,517, Loss: 31.655, Epsilon: 0.65, Training Steps: 38,485, Elapsed Time: 00:01:37
[Episode 370, Time Steps 39,925] Episode Reward: -83.39834795050082, Replay buffer: 39,925, Loss:  2.450, Epsilon: 0.64, Training Steps: 39,893, Elapsed Time: 00:01:41
[Episode 380, Time Steps 41,063] Episode Reward: -62.50433906050277, Replay buffer: 41,063, Loss:  6.075, Epsilon: 0.63, Training Steps: 41,031, Elapsed Time: 00:01:44
[Episode 390, Time Steps 42,481] Episode Reward: -10.236863050631655, Replay buffer: 42,481, Loss: 26.649, Epsilon: 0.62, Training Steps: 42,449, Elapsed Time: 00:01:48
[Episode 400, Time Steps 43,747] Episode Reward: -31.29324626657359, Replay buffer: 43,747, Loss:  5.906, Epsilon: 0.62, Training Steps: 43,715, Elapsed Time: 00:01:52
[Validation Episode Reward: [-220.89116498 -118.74622329  -82.18126252]] Average: -140.606
[Episode 410, Time Steps 45,145] Episode Reward: -89.66181348508796, Replay buffer: 45,145, Loss: 34.387, Epsilon: 0.61, Training Steps: 45,113, Elapsed Time: 00:01:57
[Episode 420, Time Steps 46,495] Episode Reward: -36.97955753844687, Replay buffer: 46,495, Loss: 41.826, Epsilon: 0.60, Training Steps: 46,463, Elapsed Time: 00:02:01
[Episode 430, Time Steps 48,976] Episode Reward: -106.08432467146287, Replay buffer: 48,976, Loss: 18.843, Epsilon: 0.59, Training Steps: 48,944, Elapsed Time: 00:02:08
[Episode 440, Time Steps 50,669] Episode Reward: -89.30410971243134, Replay buffer: 50,669, Loss:  5.130, Epsilon: 0.58, Training Steps: 50,637, Elapsed Time: 00:02:13
[Episode 450, Time Steps 52,098] Episode Reward: -32.47834130166552, Replay buffer: 52,098, Loss:  3.110, Epsilon: 0.57, Training Steps: 52,066, Elapsed Time: 00:02:18
[Validation Episode Reward: [-189.14844298  -64.95010986 -154.24503685]] Average: -136.115
[Episode 460, Time Steps 54,329] Episode Reward: -24.383120964071693, Replay buffer: 54,329, Loss:  6.478, Epsilon: 0.57, Training Steps: 54,297, Elapsed Time: 00:02:24
[Episode 470, Time Steps 56,112] Episode Reward: -104.73293265934402, Replay buffer: 56,112, Loss:  5.551, Epsilon: 0.56, Training Steps: 56,080, Elapsed Time: 00:02:30
[Episode 480, Time Steps 58,607] Episode Reward: 14.236292861513945, Replay buffer: 58,607, Loss: 13.683, Epsilon: 0.55, Training Steps: 58,575, Elapsed Time: 00:02:37
[Episode 490, Time Steps 60,498] Episode Reward: -49.332549057956705, Replay buffer: 60,498, Loss:  1.420, Epsilon: 0.54, Training Steps: 60,466, Elapsed Time: 00:02:43
[Episode 500, Time Steps 63,835] Episode Reward: -89.46609751911095, Replay buffer: 63,835, Loss: 22.243, Epsilon: 0.53, Training Steps: 63,803, Elapsed Time: 00:02:53
[Validation Episode Reward: [-153.687681   -148.61692057 -131.14651679]] Average: -144.484
```

Train - Terminal image captured



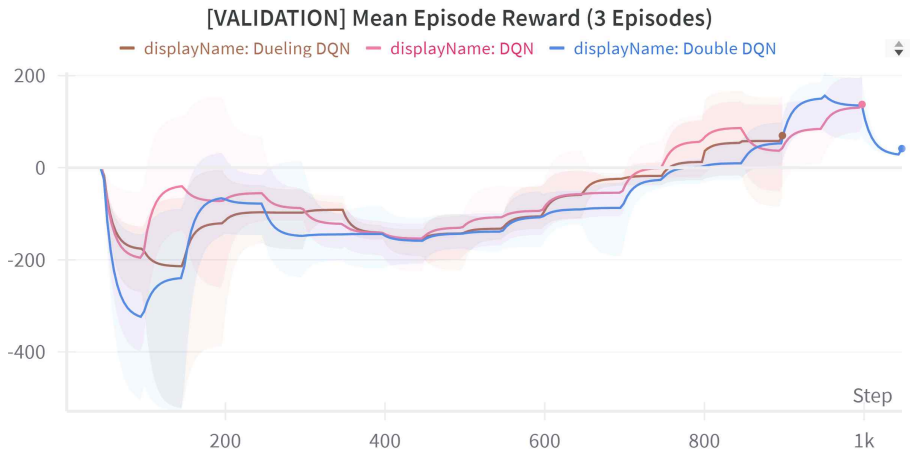
[VALIDATION] Mean Episode Reward (3 Episodes)

학습 5회 수행, 스무딩 강도 : 0.5

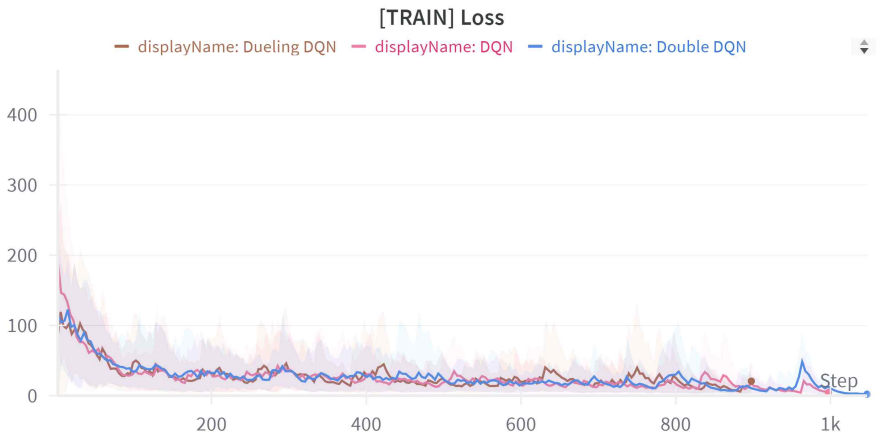
본론4_test_model)

DQN	dqn_LunarLander-v2_200.7_2024-06-05_23-21-21.pth [EPISODE: 0] EPISODE_STEPS: 251, EPISODE REWARD: 245.8 [EPISODE: 1] EPISODE_STEPS: 355, EPISODE REWARD: 265.5 [EPISODE: 2] EPISODE_STEPS: 259, EPISODE REWARD: 237.0 dqn_LunarLander-v2_208.8_2024-06-05_21-21-34.pth [EPISODE: 0] EPISODE_STEPS: 400, EPISODE REWARD: 205.9 [EPISODE: 1] EPISODE_STEPS: 459, EPISODE REWARD: 269.9 [EPISODE: 2] EPISODE_STEPS: 472, EPISODE REWARD: 217.0 dqn_LunarLander-v2_238.9_2024-06-05_22-04-45.pth [EPISODE: 0] EPISODE_STEPS: 556, EPISODE REWARD: 209.2 [EPISODE: 1] EPISODE_STEPS: 612, EPISODE REWARD: 179.2 [EPISODE: 2] EPISODE_STEPS: 365, EPISODE REWARD: 241.5 dqn_LunarLander-v2_256.2_2024-06-05_22-44-47.pth [EPISODE: 0] EPISODE_STEPS: 450, EPISODE REWARD: 228.4 [EPISODE: 1] EPISODE_STEPS: 403, EPISODE REWARD: 185.3 [EPISODE: 2] EPISODE_STEPS: 268, EPISODE REWARD: 278.8 dqn_LunarLander-v2_256.9_2024-06-05_23-57-27.pth [EPISODE: 0] EPISODE_STEPS: 314, EPISODE REWARD: 263.5 [EPISODE: 1] EPISODE_STEPS: 527, EPISODE REWARD: 231.6 [EPISODE: 2] EPISODE_STEPS: 358, EPISODE REWARD: 236.7 DQN : [Test] Mean of Episode rewards: 226.7008349781355 DQN : [Test] Standard Dev. of Episode rewards: 38.79233141255183
Double DQN	dqn_LunarLander-v2_203.6_2024-06-06_02-11-46.pth [EPISODE: 0] EPISODE_STEPS: 448, EPISODE REWARD: 258.3 [EPISODE: 1] EPISODE_STEPS: 278, EPISODE REWARD: 242.5 [EPISODE: 2] EPISODE_STEPS: 344, EPISODE REWARD: 254.0 dqn_LunarLander-v2_214.8_2024-06-06_00-29-39.pth [EPISODE: 0] EPISODE_STEPS: 511, EPISODE REWARD: 159.2 [EPISODE: 1] EPISODE_STEPS: 579, EPISODE REWARD: 193.7 [EPISODE: 2] EPISODE_STEPS: 463, EPISODE REWARD: 205.6 dqn_LunarLander-v2_219.7_2024-06-06_02-48-36.pth [EPISODE: 0] EPISODE_STEPS: 483, EPISODE REWARD: 223.3 [EPISODE: 1] EPISODE_STEPS: 661, EPISODE REWARD: 202.1 [EPISODE: 2] EPISODE_STEPS: 1000, EPISODE REWARD: 111.0 dqn_LunarLander-v2_226.4_2024-06-06_01-36-43.pth [EPISODE: 0] EPISODE_STEPS: 377, EPISODE REWARD: 249.0 [EPISODE: 1] EPISODE_STEPS: 420, EPISODE REWARD: 265.7 [EPISODE: 2] EPISODE_STEPS: 357, EPISODE REWARD: 238.6 dqn_LunarLander-v2_245.2_2024-06-06_01-00-32.pth [EPISODE: 0] EPISODE_STEPS: 371, EPISODE REWARD: 295.0 [EPISODE: 1] EPISODE_STEPS: 185, EPISODE REWARD: 264.8 [EPISODE: 2] EPISODE_STEPS: 236, EPISODE REWARD: 249.8 Double DQN : [Test] Mean of Episode rewards: 225.75732537560125 Double DQN : [Test] Standard Dev. of Episode rewards: 41.62504050340304
Dueling DQN	dqn_LunarLander-v2_203.6_2024-06-06_02-11-46.pth [EPISODE: 0] EPISODE_STEPS: 315, EPISODE REWARD: 219.4 [EPISODE: 1] EPISODE_STEPS: 347, EPISODE REWARD: 239.3 [EPISODE: 2] EPISODE_STEPS: 1000, EPISODE REWARD: 130.9 dqn_LunarLander-v2_210.3_2024-06-06_19-33-40.pth [EPISODE: 0] EPISODE_STEPS: 497, EPISODE REWARD: 234.2 [EPISODE: 1] EPISODE_STEPS: 480, EPISODE REWARD: 202.5 [EPISODE: 2] EPISODE_STEPS: 425, EPISODE REWARD: 243.2 dqn_LunarLander-v2_216.5_2024-06-06_21-14-43.pth [EPISODE: 0] EPISODE_STEPS: 504, EPISODE REWARD: 224.6 [EPISODE: 1] EPISODE_STEPS: 406, EPISODE REWARD: 261.8 [EPISODE: 2] EPISODE_STEPS: 373, EPISODE REWARD: 251.7 dqn_LunarLander-v2_228.1_2024-06-06_20-45-20.pth [EPISODE: 0] EPISODE_STEPS: 649, EPISODE REWARD: 221.9 [EPISODE: 1] EPISODE_STEPS: 309, EPISODE REWARD: 233.8 [EPISODE: 2] EPISODE_STEPS: 602, EPISODE REWARD: 203.2 dqn_LunarLander-v2_229.8_2024-06-06_20-21-47.pth [EPISODE: 0] EPISODE_STEPS: 450, EPISODE REWARD: 223.5 [EPISODE: 1] EPISODE_STEPS: 445, EPISODE REWARD: 250.5 [EPISODE: 2] EPISODE_STEPS: 421, EPISODE REWARD: 239.7 Dueling DQN : [Test] Mean of Episode rewards: 227.9304231843386 Dueling DQN : [Test] Standard Dev. of Episode rewards: 28.362524804067316

결론)



<DQN, Double DQN, Dueling DQN의 Mean episode reward>



<DQN, Double DQN, Dueling DQN의 Loss>

DQN, Double DQN, Dueling DQN의 에피소드 보상 추이를 비교 분석한 결과, DQN과 Double DQN은 학습 성능 면에서 통계적으로 유의미한 차이를 보이지 않았습니다. 이러한 결과는 Double DQN이 DQN의 overestimation bias를 줄이는 데 효과적이지만, 전반적인 보상 추이에서는 큰 개선을 가져오지 못했음을 시사합니다.

반면, Dueling DQN은 다른 두 기법에 비해 약 10% 더 빠른 학습 속도를 나타냈으며, 이는 state-value function과 action-advantage function을 분리하여 네트워크의 학습 효율을 극대화한 결과로 볼 수 있습니다. 특히, 모델 테스트 시 Dueling DQN은 보상의 분산이 더 낮게 나타나, 모델의 안정성이 상대적으로 높다는 것을 확인할 수 있었습니다.

한편, 본 과제에서 사용한 환경(Lunar Lander)과 같이 state와 action이 비교적 단순한 환경이었기 때문에, Double DQN과 Dueling DQN의 구조적 장점이 인상적인 효과를 발휘하지 못한 것으로 판단됩니다. 복잡한 행동 공간에서는 이러한 기법들이 더 큰 성능 향상을 가져올 수 있을 것으로 예상됩니다.

따라서, Dueling DQN이 Q-learning 기반 알고리즘 중 더 효율적이고 안정적인 학습을 제공할 수 있는 가능성이 높을 것 같습니다. 특히, 고차원 행동 공간에서는 Dueling DQN의 장점이 더욱 두드러질 것으로 기대됩니다.