RL Training Log

Model_Setup-v0

```
# SOLUTION# We added some parameters to accelerate the traini
model = PPO(
    policy="MlpPolicy",
    env=env,
    n_steps=1024,
    batch_size=64,
    n_epochs=4,
    gamma=0.999,
    gae_lambda=0.98,
    ent_coef=0.01,
    verbose=1,
)
```

Model_Setup-v1

```
model = PPO(
policy = 'MlpPolicy',
env = env,
n_steps = 2048,
batch_size = 64,
n_epochs = 8,
gamma = 0.999,
gae_lambda = 0.999,
ent_coef = 0.01,
verbose=1
```

Output:

```
-----------------------------------------
| rollout/                |              |
```

```
|    ep_len_mean          | 918          |
|    ep_rew_mean          | 107          |
| time/                   |              |
|    fps                  | 733          |
|    iterations           | 28           |
|    time_elapsed         | 1251         |
|    total_timesteps      | 917504       |
| train/                  |              |
|    approx_kl            | 0.0049350252 |
|    clip_fraction        | 0.0291       |
|    clip_range           | 0.2          |
|    entropy_loss         | -1.23        |
|    explained_variance   | 0.907        |
|    learning_rate        | 0.0003       |
|    loss                 | 78           |
|    n_updates            | 216          |
|    policy_gradient_loss | -0.00066     |
|    value_loss           | 126          |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 935         |
|    ep_rew_mean          | 113         |
| time/                   |             |
|    fps                  | 727         |
|    iterations           | 29          |
|    time_elapsed         | 1306        |
|    total_timesteps      | 950272      |
| train/                  |             |
|    approx_kl            | 0.006356347 |
|    clip_fraction        | 0.039       |
|    clip_range           | 0.2         |
|    entropy_loss         | -1.22       |
|    explained_variance   | 0.977       |
|    learning_rate        | 0.0003      |
|    loss                 | 5.19        |
|    n_updates            | 224         |
|    policy_gradient_loss | -0.00218    |
```

```
|    value_loss           | 27.1        |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 940         |
|    ep_rew_mean          | 123         |
| time/                   |             |
|    fps                  | 721         |
|    iterations           | 30          |
|    time_elapsed         | 1361        |
|    total_timesteps      | 983040      |
| train/                  |             |
|    approx_kl            | 0.0070096166 |
|    clip_fraction        | 0.0421      |
|    clip_range           | 0.2         |
|    entropy_loss         | -1.2        |
|    explained_variance   | 0.964       |
|    learning_rate        | 0.0003      |
|    loss                 | 27.9        |
|    n_updates            | 232         |
|    policy_gradient_loss | -0.000613   |
|    value_loss           | 49.2        |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 932         |
|    ep_rew_mean          | 127         |
| time/                   |             |
|    fps                  | 717         |
|    iterations           | 31          |
|    time_elapsed         | 1416        |
|    total_timesteps      | 1015808     |
| train/                  |             |
|    approx_kl            | 0.0032689536 |
|    clip_fraction        | 0.0446      |
|    clip_range           | 0.2         |
|    entropy_loss         | -1.2        |
|    explained_variance   | 0.986       |
```

```
|    learning_rate        | 0.0003       |
|    loss                 | 9.55         |
|    n_updates            | 240          |
|    policy_gradient_loss | -0.00161     |
|    value_loss           | 22.5         |
-----------------------------------------
```

Model_Setup-v2

```
model = PPO(
policy = 'MlpPolicy',
env = env,
n_steps = 4096,
batch_size = 128,
n_epochs = 4,
gamma = 0.99999,
gae_lambda = 0.95,
ent_coef = 0.02,
verbose=1)
```

Output-v2:

```
-----------------------------------------
| rollout/                |              |
|    ep_len_mean          | 813          |
|    ep_rew_mean          | 96.6         |
| time/                   |              |
|    fps                  | 1194         |
|    iterations           | 14           |
|    time_elapsed         | 768          |
|    total_timesteps      | 917504       |
| train/                  |              |
|    approx_kl            | 0.0064315805 |
|    clip_fraction        | 0.045        |
|    clip_range           | 0.2          |
```

```
|    entropy_loss         | -1.18        |
|    explained_variance   | 0.984        |
|    learning_rate        | 0.0003       |
|    loss                 | 6.23         |
|    n_updates            | 52           |
|    policy_gradient_loss | -0.00356     |
|    value_loss           | 11           |
-----------------------------------------
-----------------------------------------
| rollout/                |              |
|    ep_len_mean          | 801          |
|    ep_rew_mean          | 106          |
| time/                   |              |
|    fps                  | 1156         |
|    iterations           | 15           |
|    time_elapsed         | 849          |
|    total_timesteps      | 983040       |
| train/                  |              |
|    approx_kl            | 0.0052980306 |
|    clip_fraction        | 0.0496       |
|    clip_range           | 0.2          |
|    entropy_loss         | -1.15        |
|    explained_variance   | 0.989        |
|    learning_rate        | 0.0003       |
|    loss                 | 3.26         |
|    n_updates            | 56           |
|    policy_gradient_loss | -0.00271     |
|    value_loss           | 8.28         |
-----------------------------------------
-----------------------------------------
| rollout/                |              |
|    ep_len_mean          | 803          |
|    ep_rew_mean          | 115          |
| time/                   |              |
|    fps                  | 1134         |
|    iterations           | 16           |
|    time_elapsed         | 924          |
|    total_timesteps      | 1048576      |
```

```
| train/                  |              |
|    approx_kl            | 0.0044425875 |
|    clip_fraction        | 0.0385       |
|    clip_range           | 0.2          |
|    entropy_loss         | -1.14        |
|    explained_variance   | 0.993        |
|    learning_rate        | 0.0003       |
|    loss                 | 1.72         |
|    n_updates            | 60           |
|    policy_gradient_loss | -0.00304     |
|    value_loss           | 5.16         |
-----------------------------------------
```

Evaluation-v2:

```
mean_reward=232.83 +/- 80.28094849842792
```

Model_Setup-v3.1

```
# Create the environment
env = make_vec_env('LunarLander-v2', n_envs=3)
```

```
model = PPO(
policy = 'MlpPolicy',
env = env,
n_steps = 4,
batch_size = 32,
n_epochs = 5,
gamma = 0.99999,
gae_lambda = 0.95,
```

```
ent_coef = 0.02,
verbose=1)
```

```
# SOLUTION
# Train it for 10,000 timesteps
model.learn(total_timesteps=10000)
# Save the model
model_name = "ppo-LunarLander-v2"
model.save(model_name)
```

Output:

```
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 206         |
|    ep_rew_mean          | -225        |
| time/                   |             |
|    fps                  | 246         |
|    iterations           | 834         |
|    time_elapsed         | 40          |
|    total_timesteps      | 10008       |
| train/                  |             |
|    approx_kl            | 0.00016328196 |
|    clip_fraction        | 0           |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.668      |
|    explained_variance   | 0.999       |
|    learning_rate        | 0.0003      |
|    loss                 | 1.7         |
|    n_updates            | 4165        |
|    policy_gradient_loss | -0.00233    |
|    value_loss           | 3.51        |
-----------------------------------------
```

mean_reward=-249.60 +/- 286.71541213927793

Model_Setup-v3.2

```
# Create the environment
env = make_vec_env('LunarLander-v2', n_envs=1)
```

```
# TODO: Define a PPO MlpPolicy architecture
# We use MultiLayerPerceptron (MLPPolicy) because the input i
# if we had frames as input we would use CnnPolicy
n_steps = 4
batch_size = 32
n_epochs = 5
model = PPO(
    policy = 'MlpPolicy',
    env = env,
    n_steps = n_steps,
    batch_size = batch_size,
    n_epochs = n_epochs,
    gamma = 0.99999,
    gae_lambda = 0.95,
    ent_coef = 0.02,
    verbose=1)
```

```
# SOLUTION
# Train it for 10,000 timesteps
model.learn(total_timesteps=10000)
# Save the model
model_name = "ppo-LunarLander-v2"
model.save(model_name)
```

Output:

mean_reward=-141.05 +/- 354.6015460046699

Model_Setup-v4.1

```
# Create the environment
env = make_vec_env('LunarLander-v2', n_envs=32)
```

```
# TODO: Define a PPO MlpPolicy architecture
# We use MultiLayerPerceptron (MLPPolicy) because the input i
# if we had frames as input we would use CnnPolicy
n_steps = 4096
batch_size = 128
n_epochs = 8
model = PPO(
    policy = 'MlpPolicy',
    env = env,
    n_steps = n_steps,
    batch_size = batch_size,
    n_epochs = n_epochs,
    gamma = 0.99999,
    gae_lambda = 0.95,
    ent_coef = 0.02,
    verbose=1)
```

```
# SOLUTION
# Train it for 10,000,000 timesteps
model.learn(total_timesteps=10000000)
# Save the model
model_name = "ppo-LunarLander-v2"
model.save(model_name)
```

Output:

```
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 293         |
|    ep_rew_mean          | 287         |
| time/                   |             |
```

```
|    fps                  | 1056         |
|    iterations           | 64           |
|    time_elapsed         | 7941         |
|    total_timesteps      | 8388608      |
| train/                  |              |
|    approx_kl            | 0.0026513708 |
|    clip_fraction        | 0.0317       |
|    clip_range           | 0.2          |
|    entropy_loss         | -0.53        |
|    explained_variance   | 1            |
|    learning_rate        | 0.0003       |
|    loss                 | 0.197        |
|    n_updates            | 504          |
|    policy_gradient_loss | -0.000409    |
|    value_loss           | 0.449        |
-----------------------------------------
-----------------------------------------
| rollout/                |              |
|    ep_len_mean          | 316          |
|    ep_rew_mean          | 285          |
| time/                   |              |
|    fps                  | 1060         |
|    iterations           | 65           |
|    time_elapsed         | 8036         |
|    total_timesteps      | 8519680      |
| train/                  |              |
|    approx_kl            | 0.0029160783 |
|    clip_fraction        | 0.0372       |
|    clip_range           | 0.2          |
|    entropy_loss         | -0.545       |
|    explained_variance   | 0.999        |
|    learning_rate        | 0.0003       |
|    loss                 | 0.211        |
|    n_updates            | 512          |
|    policy_gradient_loss | -0.000378    |
|    value_loss           | 4.04         |
-----------------------------------------
-----------------------------------------
```

```
| rollout/                |              |
|    ep_len_mean          | 328          |
|    ep_rew_mean          | 284          |
| time/                   |              |
|    fps                  | 1064         |
|    iterations           | 66           |
|    time_elapsed         | 8129         |
|    total_timesteps      | 8650752      |
| train/                  |              |
|    approx_kl            | 0.0028840213 |
|    clip_fraction        | 0.0313       |
|    clip_range           | 0.2          |
|    entropy_loss         | -0.537       |
|    explained_variance   | 0.998        |
|    learning_rate        | 0.0003       |
|    loss                 | 0.246        |
|    n_updates            | 520          |
|    policy_gradient_loss | 0.000147     |
|    value_loss           | 8.43         |
-----------------------------------------
-----------------------------------------
| rollout/                |              |
|    ep_len_mean          | 300          |
|    ep_rew_mean          | 289          |
| time/                   |              |
|    fps                  | 1068         |
|    iterations           | 67           |
|    time_elapsed         | 8217         |
|    total_timesteps      | 8781824      |
| train/                  |              |
|    approx_kl            | 0.0029235762 |
|    clip_fraction        | 0.0327       |
|    clip_range           | 0.2          |
|    entropy_loss         | -0.526       |
|    explained_variance   | 0.996        |
|    learning_rate        | 0.0003       |
|    loss                 | 16           |
|    n_updates            | 528          |
```

```
|    policy_gradient_loss | 0.000349     |
|    value_loss           | 23.5         |
-----------------------------------------
-----------------------------------------
| rollout/                |              |
|    ep_len_mean          | 324          |
|    ep_rew_mean          | 287          |
| time/                   |              |
|    fps                  | 1072         |
|    iterations           | 68           |
|    time_elapsed         | 8310         |
|    total_timesteps      | 8912896      |
| train/                  |              |
|    approx_kl            | 0.0039145024 |
|    clip_fraction        | 0.0581       |
|    clip_range           | 0.2          |
|    entropy_loss         | -0.563       |
|    explained_variance   | 0.999        |
|    learning_rate        | 0.0003       |
|    loss                 | 9.74         |
|    n_updates            | 536          |
|    policy_gradient_loss | 0.00022      |
|    value_loss           | 5.69         |
-----------------------------------------
```

Model_Setup-v4.2

```
# Create the environment
env = make_vec_env('LunarLander-v2', n_envs=32)
```

```
# TODO: Define a PPO MlpPolicy architecture
# We use MultiLayerPerceptron (MLPPolicy) because the input i
# if we had frames as input we would use CnnPolicy
```

```
n_steps = 4096
batch_size = 128
n_epochs = 8
model = PPO(
    policy = 'MlpPolicy',
    env = env,
    n_steps = n_steps,
    batch_size = batch_size,
    n_epochs = n_epochs,
    gamma = 0.99999,
    gae_lambda = 0.95,
    ent_coef = 0.02,
    verbose=1)
```

```
# SOLUTION
# Train it for 1,000,000 timesteps
model.learn(total_timesteps=1000000)
# Save the model
model_name = "ppo-LunarLander-v2"
model.save(model_name)
```

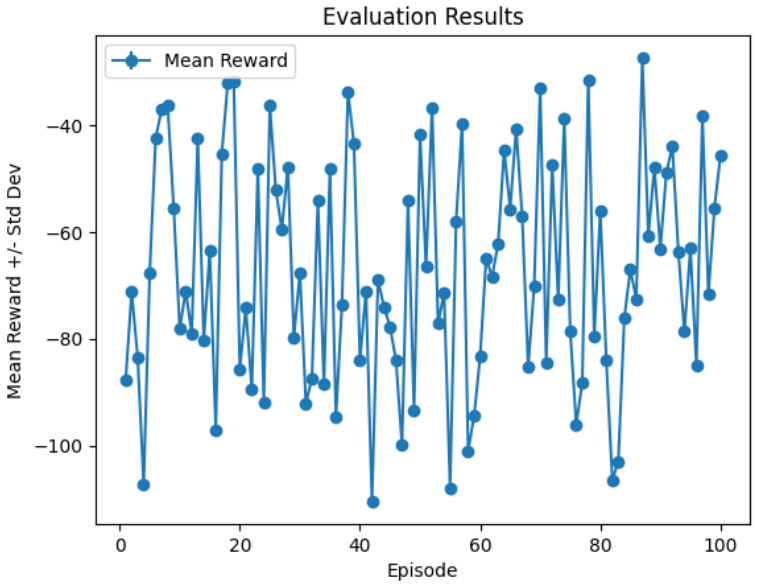
Output:

```
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 859         |
|    ep_rew_mean          | 24.4        |
| time/                   |             |
|    fps                  | 1069        |
|    iterations           | 8           |
|    time_elapsed         | 980         |
|    total_timesteps      | 1048576     |
| train/                  |             |
|    approx_kl            | 0.013855488 |
|    clip_fraction        | 0.145       |
|    clip_range           | 0.2         |
|    entropy_loss         | -1.18       |
|    explained_variance   | 0.771       |
```

```
|    learning_rate        | 0.0003      |
|    loss                 | 38.4        |
|    n_updates            | 56          |
|    policy_gradient_loss | -0.0102     |
|    value_loss           | 56.5        |
-----------------------------------------
```

mean_reward=-69.86 +/- 25.926564519913544

Trail Graph Evaluation :



Model_Setup-v4.3

```
# Create the environment
env = make_vec_env('LunarLander-v2', n_envs=32)
```

```
# TODO: Define a PPO MlpPolicy architecture
# We use MultiLayerPerceptron (MLPPolicy) because the input i
# if we had frames as input we would use CnnPolicy
```

```
n_steps = 4096
batch_size = 128
n_epochs = 8
model = PPO(
    policy = 'MlpPolicy',
    env = env,
    n_steps = n_steps,
    batch_size = batch_size,
    n_epochs = n_epochs,
    gamma = 0.99999,
    gae_lambda = 0.95,
    ent_coef = 0.02,
    verbose=1)
```

```
# SOLUTION
# Train it for 1,000,000 timesteps
model.learn(total_timesteps=1000000)
# Save the model
model_name = "ppo-LunarLander-v2"
model.save(model_name)
```

Output:

mean_reward=166.26 +/- 83.73979501428518

Model_Setup-v4.4

```
# Create the environment
env = make_vec_env('LunarLander-v2', n_envs=32)
```

```
# TODO: Define a PPO MlpPolicy architecture
# We use MultiLayerPerceptron (MLPPolicy) because the input i
# if we had frames as input we would use CnnPolicy
n_steps = 4096
```

```
batch_size = 128
n_epochs = 8
model = PPO(
    policy = 'MlpPolicy',
    env = env,
    n_steps = n_steps,
    batch_size = batch_size,
    n_epochs = n_epochs,
    gamma = 0.99999,
    gae_lambda = 0.95,
    ent_coef = 0.02,
    verbose=1)
```

```
# SOLUTION
# Train it for 5,000,000 timesteps
model.learn(total_timesteps=5000000)
# Save the model
model_name = "ppo-LunarLander-v2"
model.save(model_name)
```

Output:

```
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 283         |
|    ep_rew_mean          | 274         |
| time/                   |             |
|    fps                  | 2289        |
|    iterations           | 34          |
|    time_elapsed         | 1946        |
|    total_timesteps      | 4456448     |
| train/                  |             |
|    approx_kl            | 0.01099251  |
|    clip_fraction        | 0.0518      |
|    clip_range           | 0.2         |
```

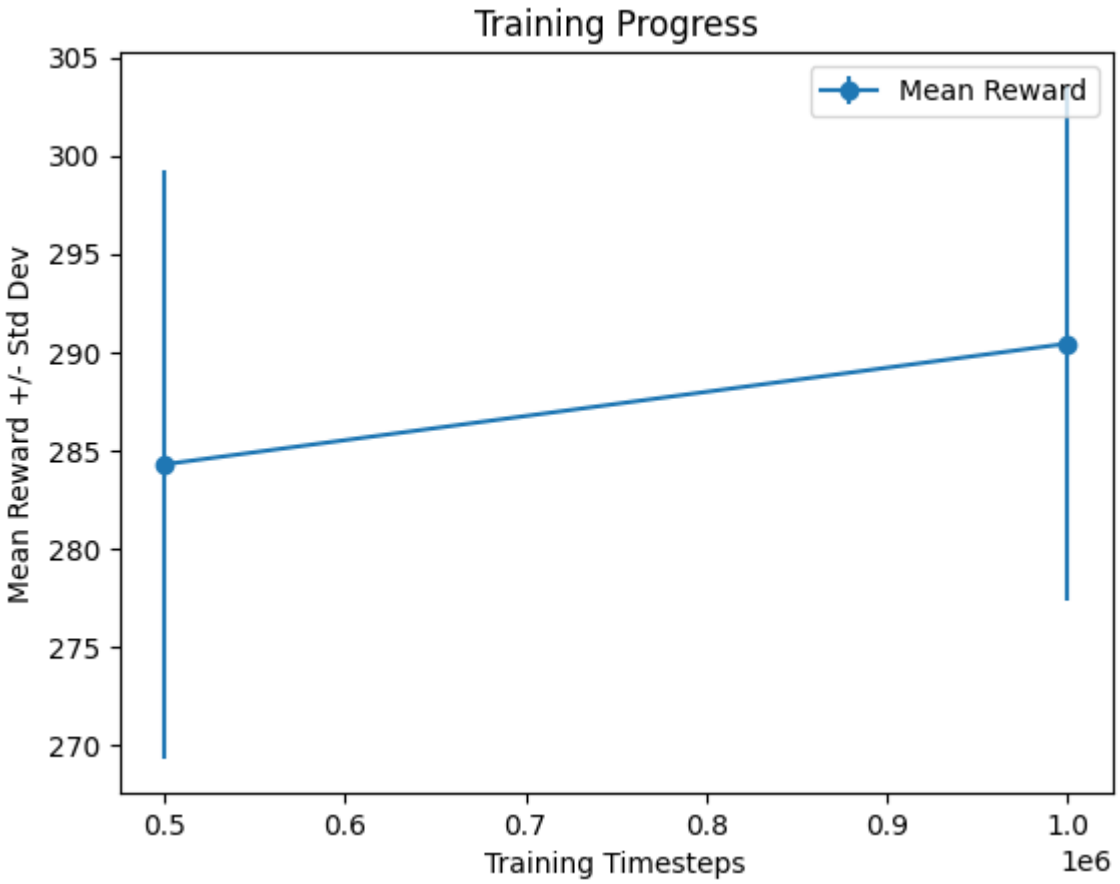
```
|    entropy_loss          | -0.726     |
|    explained_variance    | 0.893      |
|    learning_rate         | 0.0003     |
|    loss                  | 9.18       |
|    n_updates             | 264        |
|    policy_gradient_loss | -0.00126   |
|    value_loss            | 49         |
-----------------------------------------


-----------------------------------------
| rollout/                 |             |
|    ep_len_mean           | 906         |
|    ep_rew_mean           | 188         |
| time/                    |             |
|    fps                   | 2281        |
|    iterations            | 35          |
|    time_elapsed          | 2010        |
|    total_timesteps       | 4587520     |
| train/                   |             |
|    approx_kl             | 0.047576703 |
|    clip_fraction         | 0.0848      |
|    clip_range            | 0.2         |
|    entropy_loss          | -0.802      |
|    explained_variance    | 0.928       |
|    learning_rate         | 0.0003      |
|    loss                  | 5.02        |
|    n_updates             | 272         |
|    policy_gradient_loss | -0.000671   |
|    value_loss            | 46.3        |
-----------------------------------------
```

mean_reward=282.14 +/- 22.101280660194046

Graph Evaluation:

mean_reward=290.44 +/- 13.0963390313979


Training Progress
Mean Reward
Mean Reward +/- Std Dev
305
300
295
290
285
280
275
270
0.5
0.6
0.7
0.8
0.9
1.0
Training Timesteps
1e6