Reinforcement Learning
Exercise 05

Finn Hülsbusch & Dennis Heinz

Task 1. PPO

PPO is a policy optimization algorithm that uses a family of policy iteration methods to improve the agent's policy iteratively. It operates by iteratively sampling trajectories from the environment and updating the policy based on the collected data. PPO employs a trust region approach to ensure that policy updates remain within a reasonable range and prevent drastic policy changes. This allows for stable and effective learning in RL tasks. For the PPO runs, we used the following hyperparameters illustrated in Table 1.

Hyperparameter	**In Code**	**Value**
Total Timesteps of the Experiment	total_timesteps	Cheetah: 1000000 Ant: 2000000
Learning Rate	learning_rate	$3e^{-4}$
No. Steps per policy	num_steps	2048
Learning Rate Annealing	anneal_lr	True
Gamma	gamma	0.99
Lambda (GAE)	gae_lambda	0.95
No. Minibatches	num_minibatches	32
Epochs to update policy	update_epochs	10
Suggorate clipping coef.	clip_coef	0.2
Entropy coef.	ent_coef	0.0
Value function coef.	vf_coef	0.5
Max. Grad Clipping Norm	max_grad_norm	0.5
Target KL threshold	target_kl	None
seed	seed	1

Table 1: Used Hyperparameters

Cheetah:

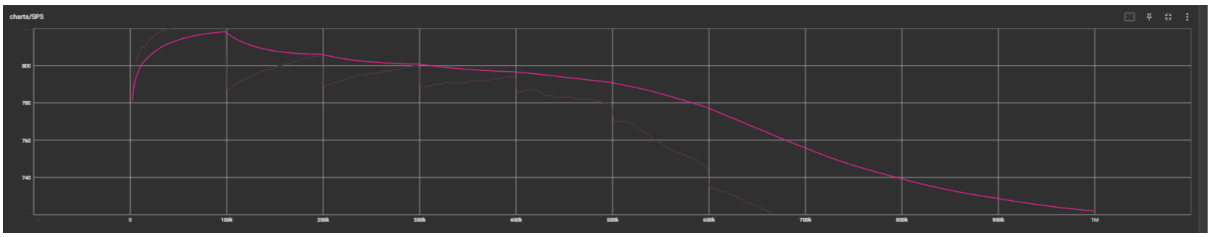

Figure 1. SPS (Cheetah)

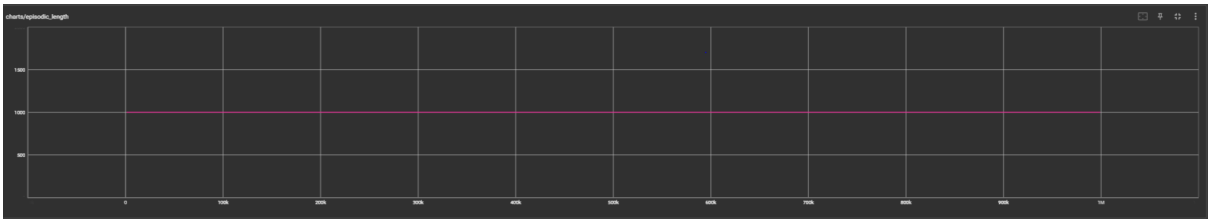

Figure 2. Eps. Length (Cheetah)

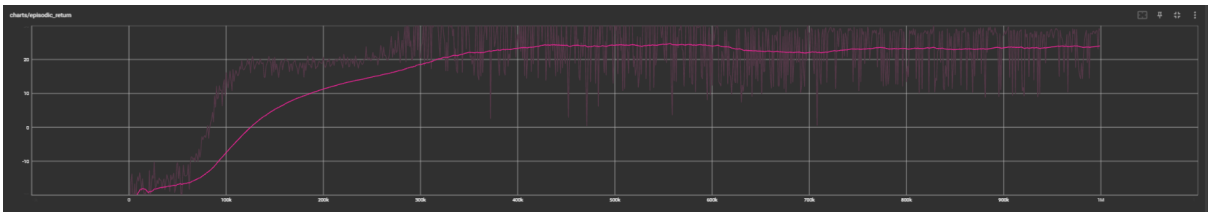

Figure 3. Return (Cheetah)

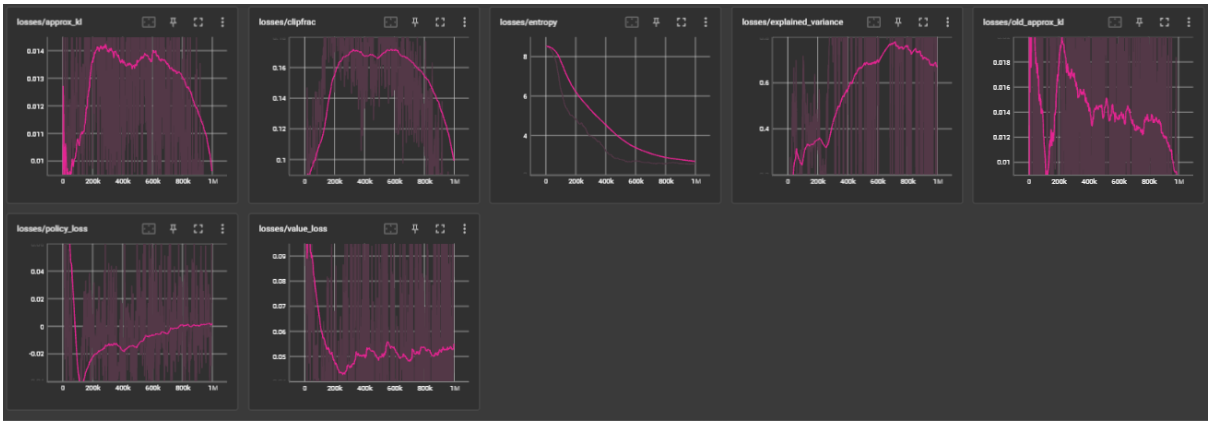

Figure 4. Loss (Cheetah)

For the cheetah environment, the best results for the tested hyperparameter configurations are depicted in Figures 1 to 5 and the SPS, Eps. Length and Return (smoothed) are illustrated in Table 2.

SPS	Eps. Len	Return
712	1000	24

Table 2: Overall Best Results for Cheetah

Ant:

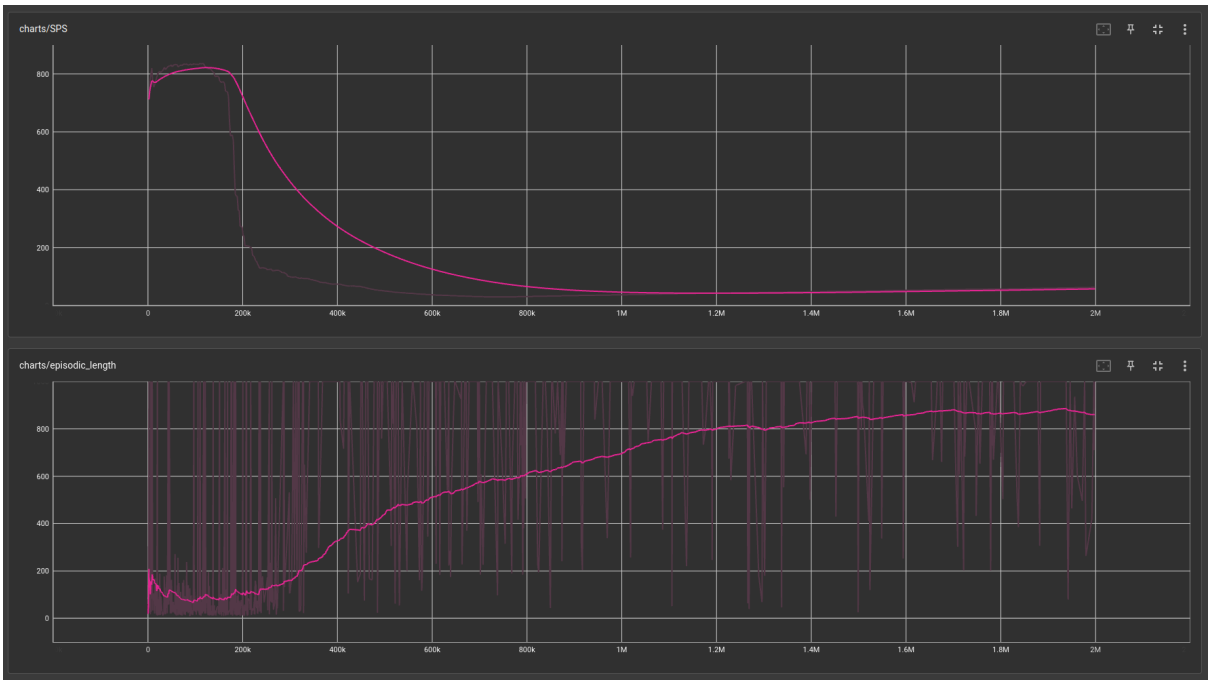

Figure 5. SPS and Eps. Length (Ant)

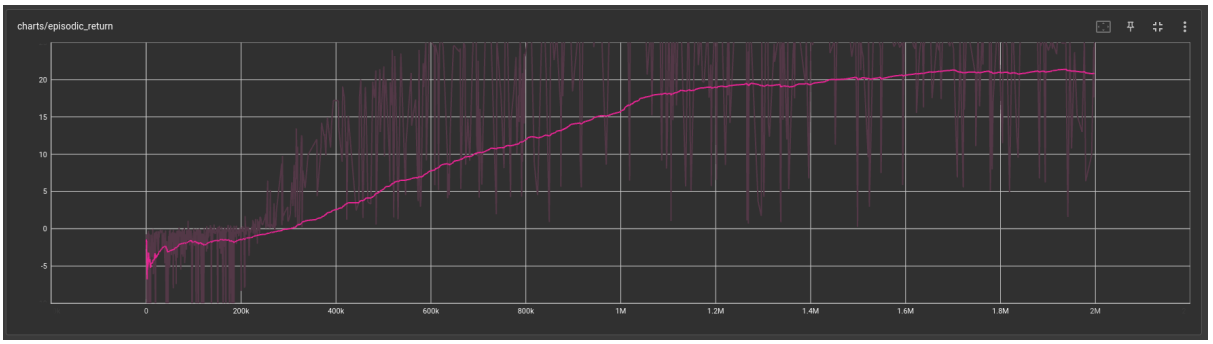

Figure 6. Return (Ant)

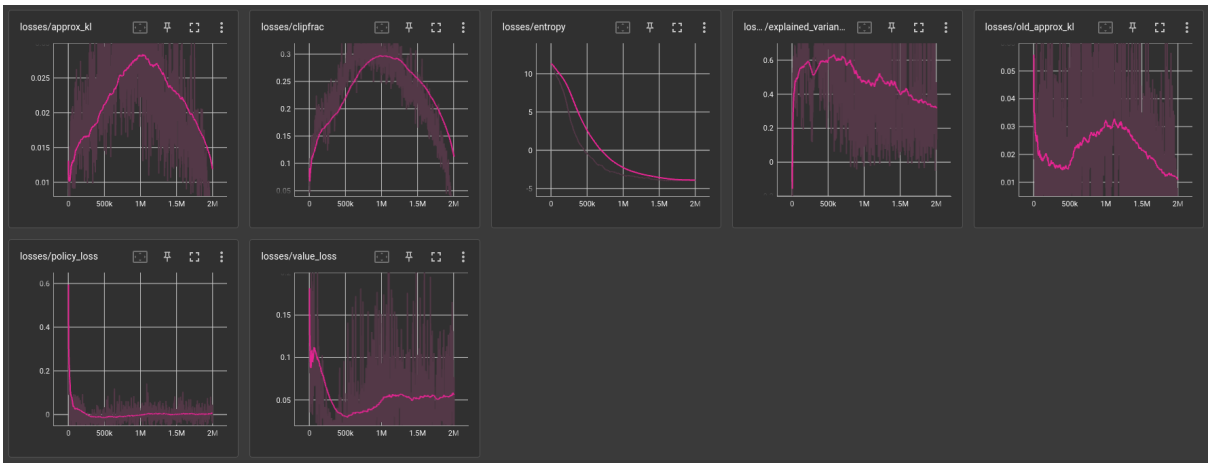

Figure 7. Loss (Ant)

For the ant environment, the best results for the tested hyperparameter configurations are depicted in Figures 5 to 7 and the SPS, Eps. Length and Return (smoothed) are illustrated in Table 3.

SPS	Eps. Len	Return
63	861	20

Table 3: Overall Best Results for Ant

Generally, for both environments, the SPS measure is not accurate as the machine running this env was 100% loaded.