Work Weekly Update 5

Syed Zaidi

September 2023

1 Code changes

The main change in the code was in the collect experience function. We introduced the idea of the FFM model and used it to predict the next states and use the loss as a reward encouraging the agents to explore new paths.

Main changes can be checked in the "algos/ppo.py".

2 Experiment Setup

We first ran the model for 600 epochs using all the default options with the exception of changing the **learning rate** in the base model from 0.0001 to 0.01, since using the previous took too much time to locate any rewards. We also used the "**pdacy**" value of 0.01 instead of 1.

Then, we ran the model with the modified collect experience function which uses the difference between the output of the FFM model (the predicted next state) and the next state (the real next state) as a reward for the agents.

2.1 Notes:

According to the papers we read at the beginning of our work, we should make several runs using different random seeds shared between the two cases and then compare the results. But currently, I can only run on my own laptop and each run takes a considerable amount of time. Therefore, I believe that currently, the experiment setup is more important than the current results.

3 Results

So, as seen from fig. 2, the FFM model is learning quickly to predict the next states. From fig. 1, we can observe that using the curiosity term helps the model find the rewards faster. This encourages the model to attempt "risky" moves which might lead to a decrease in the average reward. However, it can recover shortly, and sometimes, this "risky" move can lead to an increase in the average score.

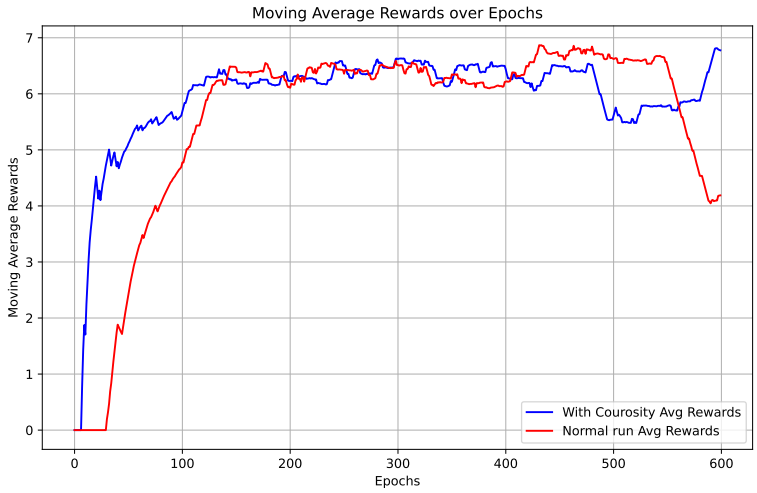

Figure 1: The collected reward from the agents (without the intrinsic reward)

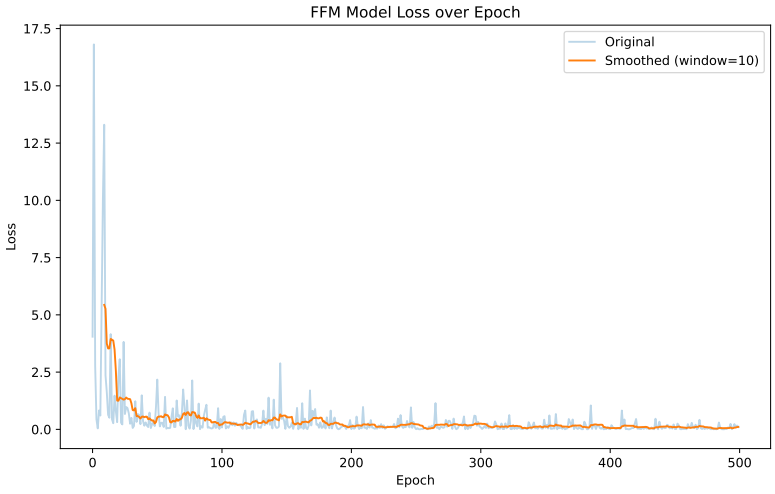

Figure 2: The FFM model loss over epochs