Report

PPO solution for the Udacity Reinforcement Learning continuous control assignment.

Description of the learning algorithm

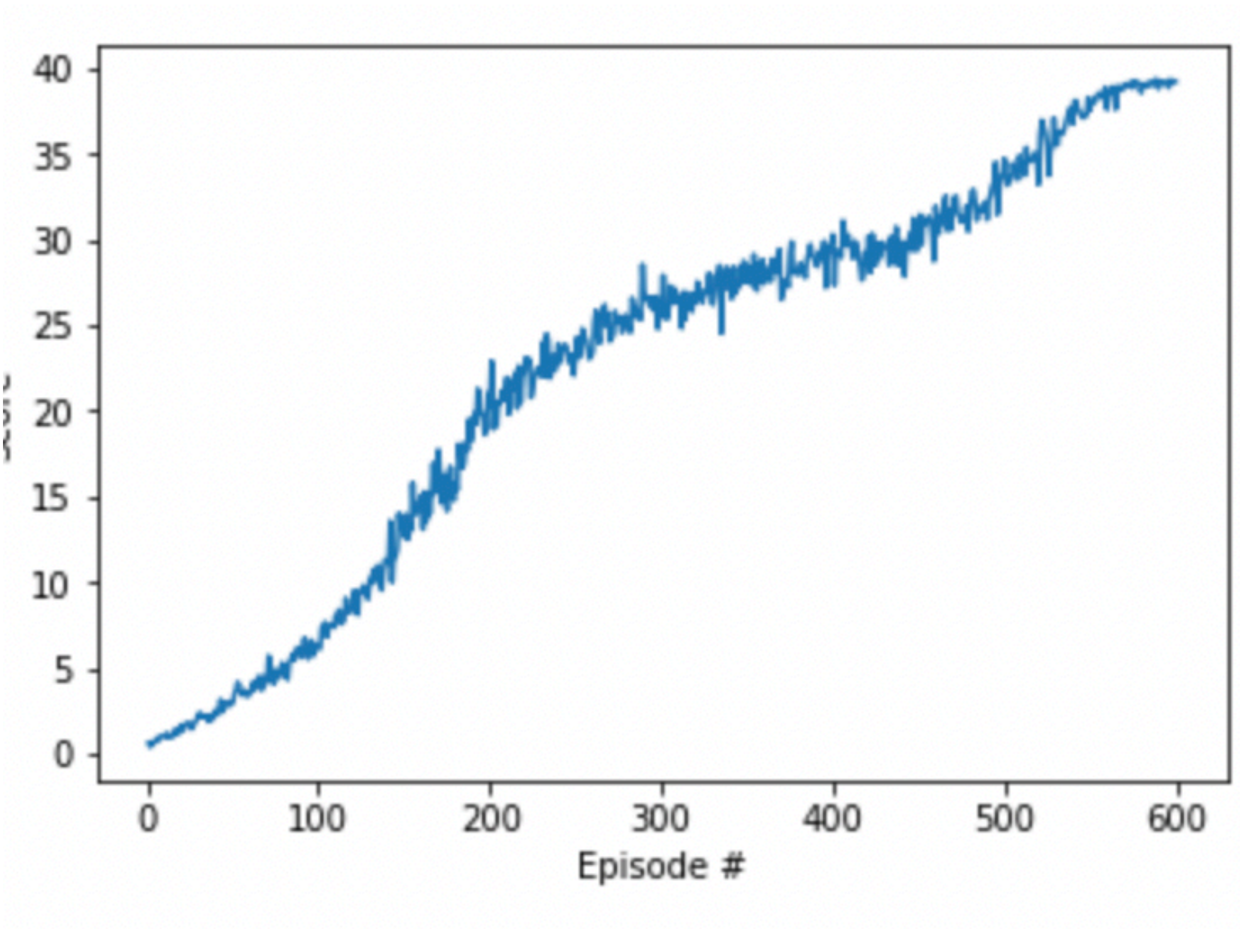
I chose to attempt to implement a version of the the PPO algorithm loosely based on the code described in the PPO section of this course.
The Udacity PPO code needed to be adapted for continuous observation and action spaces. In addition I add an actor critic model.

The PPO algorithm seems to be a good choice according to this paper https://arxiv.org/abs/1604.06778

The agent and critic models both use the same MLP with 2 hidden layers each with 128 units.

The output of the actor represents the mean value of the 4 actions of the agent. But the actions themselves are sampled randomly around the mean, using MultivariateNormal from torch's distributions package. This helps introduce a level of stochasticity into the learning process.

Results



As can be seen from the plot above, the algorithm solves the problem in around 400 epochs. The learning curve is relatively smooth and constantly increasing.
This is more episodes than the baseline example provided in the project explanation. However it runs fairly quickly and completes in under an hour. Since no time was provided for the benchmark it is not possible to compare.

The most significant parameter adjustment to improve performance seems to be the discount rate. Changing it from 0.99 to 0.95 increased performance drastically. Maybe, because rewards are given on every step, looking too far into the future isn’t so great.

I also increased the learning rate form 1e-4 to 1e-3. This helped speed up learning also.

Future Development

This implementation of PPO only uses a single mlp architecture for for both the actor and the critic, differing only in the size of the output layer. Having a separate architecture for each might provide opportunities for improvement since each would be able to have its own set of parameters.

Another thing to try would be to a duel head on the actor mlp, which could output both an action and log probabilities. This differs from the current implementation which selects both the action and log probabilities from a multivariate normal distribution.