Reinforcement Learning Lab Session 8

Dawid Miroyan

May 1, 2020

Introduction

In this document we will study the "Actor-Critic" Reinforcement Learning method on different environments. These methods are **A2C** and **PPO** on the Cartpole-v1 and LunarLander-v2 environments.

In these "Actor-Critic" methods two nueral networks will be used:

- **Actor:** Controls how our agent behaves (policy-based)
- **Critic:** Measures how good the state is (vlaue-based)

The Critic observes the taken actions and provides feedback. The Actor uses this feedback to update it's policy. Meanwhile the Critic will update their own way to provide feedback so it can be better next time.

Contents

1 Difference between A2C and PPO

The main difference between A2C and PPO (Proximal Policy Optimiziation) is that for PPO we will avoid having too large policy updates. This improves the stability of the Actor training.

2 Solving the LunarLander-v2 environment

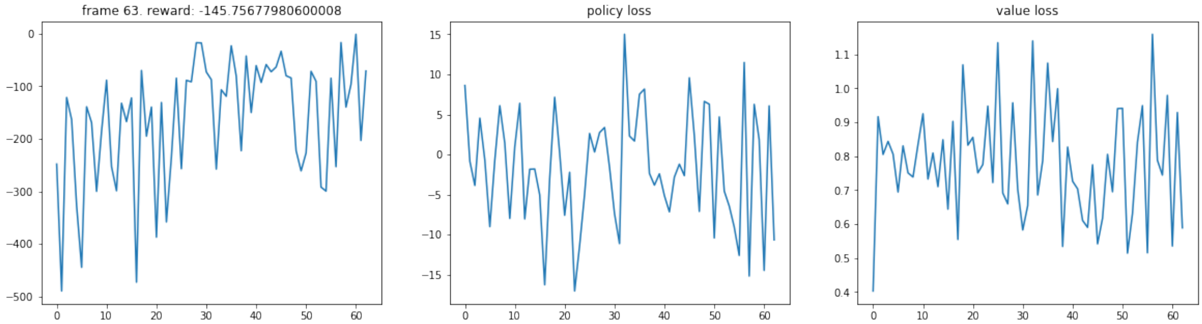

Figure 1: Reward, policy and value loss when training on the LunarLander environment using A2C.

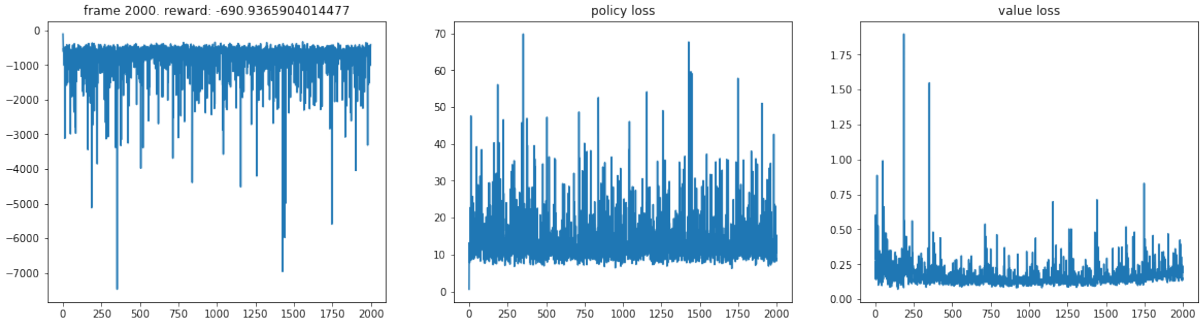

Figure 2: Reward, policy and value loss when training on the LunarLander environment using PPO.

3 Performance comparison

4 Potential improvements