**ABSTRACT**

I present a comparison of computational efficiency between training an agent for playing Pong in the Atari Environment in a sequential manner and in a parallelized manner. The agent is trained using a Deep Reinforcement Learning algorithm called Advantage Actor Critic (A2C). The parallel version of the algorithm Asynchronous A2C (A3C) is analyzed using multiple CPU cores on a single system instead of special hardware like Graphical Processing Units (GPU) or Tensor Processing Units (TPU).

**KEYWORDS**

**INTRODUCTION**

Reinforcement Learning algorithms are a class of algorithms that have proven to be quite promising in the field of decision making. That makes games a suitable area for their application. Combine them with Deep Learning which is known to produce rich representations of input data (visual data in this case) and we get the ability to play games at a human or even better-than-human level (1,2).The game chosen here is Pong which is an atari game and is provided as an environment in OpenAI (3).

**REINFORCEMENT LEARNING**

The setup consists of an agent A that interacts with the environment E . The agent interprets the environment by moving from one state st to another state st+1 at every time step t. In each state it receives an observation from the environment which might contain all or partial information about the environment in that state. It chooses an action at among a set of predefined actions A = {a1,...an} based on some computation and as a result lands in another state and receives a signal called the reward Rt. The agent keeps accumulating these rewards till it reaches the goal or the episode terminates. An episode runs for a predefined max tmax number of time steps.

The action at each time step is taken by an agent in accordance with a policy π that is either learned as in Policy Gradient methods or a predefined as in Value Based methods.

A policy is a function that given a state s, gives a probability distribution over A.

The value function of a state s under a policy π is the expected return when starting in s and following π thereafter. Similarly the value of taking action a in state s under policy π is the expected return starting from s, taking the action a and thereafter following policy π (4).

**ADVANTAGE ACTOR CRITIC (A2C)**

The agent here consists of two neural networks. One of them, the Actor, spits out a probability distribution over a set of actions A at each time step and learns the optimal policy based on a loss function, more about which has been described later. The other, the Critic, calculates the action-value at each time step taken telling how good the action taken is and learns to give better estimates of values.

References

[1] Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Graves, Alex, Antonoglou,Ioannis, Wierstra, Daan, Riedmiller, Martin

Playing atari with deep reinforcement learning. NIPS Deep Learning Workshop 2013.

[2] DeepMind

Human-level control through deep reinforcement learning.

Nature, 518(7540):529-533, 02 2015

1. Brockman, G., Cheung, V., Petersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.

OpenAI Gym. CoRR abs/1606.01540 (2016)

[4] Sutton, RIchard, Barto, Andrew

Reinforcement Learning: An Introduction

MIT Press, 1998