Reinforcement Learning Sustainability Benchmark

Luca Strefezza

June 28, 2024

Contents

1 Context

This project addresses the energy consumption of deep reinforcement learning (DRL) solutions and their impact on the environment and business costs.

Beginning with the resurgence of the field following the development of *Deep Q-Networks* (DQN) by DeepMind in the early 2010s [1], there have been a number of algorithm proposals over time that with minor modifications to DQN or using a completely different paradigm (such as policy gradient methods) sought to improve the performance achieved by the learning agent.

Although the performance of the various solutions has been extensively studied and tracked, little effort has been directed toward understanding how the tweaks to the DQN introduced to improve performance impacted energy consumption, or what the cost of the alternative approaches developed was, per se and in comparison with previous solutions.

The motivation behind this project is to fill this gap by evaluating the trade-offs between performance and energy consumption for several widely used deep reinforcement learning (DRL) algorithms. Understanding these trade-offs is crucial for businesses and researchers who aim to optimize both performance and sustainability in their applications. This project aims to provide valuable insights into the energy efficiency of different DRL approaches, enabling informed decisions about their use in various contexts.

To reach this goal we train various reinforcement learning algorithms on the same task, the choice of which is discussed in section 3.2 on page 4. Section 3.1 on page 3 describes the selected algorithms, whose choice was made taking into account that DRL algorithms can be devided in two main categories: *value based* (i.e. algorithms based on the approximation of a value function, be it the state-value function or the action-value function) and

policy gradient. The latter are methods that approximate directly the policy, and includes as a special case the *actor-critic methods*, which approximate simultaneously a policy (said actor) and a value function (said critic).

2 Goals

The primary goal of this project is to benchmark the energy consumption and performance of various deep reinforcement learning algorithms. Specifically, we aim to:

1. evaluate the energy consumption of different DRL algorithms when trained on the same task;
2. compare the performance of these algorithms in terms of their ability to achieve high scores on the given task;
3. analyze the trade-offs between performance and energy consumption to identify the most efficient algorithms;
4. provide a comprehensive report that can guide practitioners in selecting the appropriate DRL algorithms based on specific use-case requirements.

By achieving these goals, the project will contribute to the broader understanding of the sustainability implications of deep reinforcement learning technologies.

3 Methodological Steps to Conduct to Address the Goals

The methodology used follows from the basic idea of this benchmark: to execute all the algorithms for the same number of environment interactions, so that we can compare the score they achieve and the energy consumption of each one of them. Additionally, a good comparison would be to take the score obtained by the lowest performer in this initial trial and re-train all the algorithms until they reach that score. This would allow us to compare how much time and energy each algorithm requires to achieve the same performance level. Unfortunately, time and resources constraints make retraining all algorithms unfeasible, so we will approximate this second comparison by using the returns from the logging of the training during the first trial. This logging includes the *global_step*, indicating the environment interaction we

are at, and the *episodic_return*, which is the return of the episode (i.e., the score on which to compare), as well as all performance and power consumption data up to that point. By analyzing these logs, we will estimate how much time and energy each algorithm would take to reach the score obtained by the lowest performer in the initial trial.

The following sections outline the several key steps involved in the methodology adopted for this project.

3.1 Algorithm Selection

In our benchmark we will consider, regardind the first cathegory, the DQN, which constitutes the first example of success of deep reinforcement learning (so that we have a sort of baseline), and RAINBOW, a method that involves a lot of the tweaks and improvement made to the original DQN. In addition to these two, we will test various of the single tweaks to assess their individual contribution to energy consumption and performance, and more advanced methods like SPR (Self-Predictive Representations, introduced in the fifth work cited).

Regarding policy gradient and actor critic methods, we will start with a basic one like REINFORCE and/or REINFORCE with baseline (chapter 13 of the first cited work) or the very similar Vanilla Policy Gradient (VPG). We will then move on to Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradient (DDPG) and its evolutions Twin Delayed DDPG (TD3, seventh work cited) and DRQ (Data-regularized Q, fourth cited work).

We will benchmark both value-based and policy gradient methods. The selected algorithms are:

- **Value-Based Methods:**
  - Deep Q-Network (DQN)
  - RAINBOW
  - Individual tweaks to DQN (Double DQN, Dueling DQN, etc.)
  - Self-Predictive Representations (SPR)
- **Policy Gradient Methods:**
  - REINFORCE
  - Proximal Policy Optimization (PPO)
  - Deep Deterministic Policy Gradient (DDPG)
  - Twin Delayed DDPG (TD3)
  - Data-Regularized Q (DRQ)

3.2 Task Selection

Regarding the task on which to compare the algorithms, there are several suitable candidates: Atari 100k, one of the continuous control task of the DeepMind Control Suite, or one of the many other task (besides Atari) included in OpenAI Gymnasium (formerly Gym), and so on. We will probably opt for the Atari 100k, a discrete task that consists of playing one of the Atari games for 100000 interactions. The reason for this choice is that this is a widely used benchmark, well suited for running almost all popular DRL algorithms; however, we reserve the right to change the number of interactions (obviously keeping it consistent for all algorithms) or the task altogether, should hardware and time constraints dictate it.

The primary task for benchmarking will be the Atari 100k, which involves training the algorithms to play Atari games for 100,000 interactions. This benchmark is widely used and well-suited for evaluating the performance of popular DRL algorithms. However, we may adjust the number of interactions or select a different task based on hardware and time constraints.

3.3 Experiment Setup

1. **Environment Setup:** All algorithms will be implemented and run in a controlled environment to ensure consistent comparison.

2. **Training Procedure:** Each algorithm will be trained for the same number of environment interactions to allow for a fair comparison of energy consumption and performance.

3. **Energy Measurement:** The energy consumption of each algorithm will be measured using appropriate tools and methodologies.

4. **Performance Evaluation:** The performance of each algorithm will be evaluated based on the scores achieved in the Atari 100k benchmark.

5. **Additional Benchmark:** As an additional comparison, all algorithms will be re-trained until they reach the score obtained by the lowest performer in the initial benchmark, and their energy consumption and time to achieve this score will be recorded.

3.4 Data Collection and Analysis

1. Collect data on energy consumption and performance for each algorithm.

2. Analyze the data to identify trade-offs between performance and energy consumption.

3. Generate visualizations and statistical analyses to present the findings.

References

[1] V. Mnih *et al.*, *Playing atari with deep reinforcement learning*, 2013. arXiv: 1312.5602 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1312.5602.