22.01.2024

Disentangling World Models with Reinforcement Learning

Tomasz Arczewski Tadeusz Dziarmaga

I. Introduction

In model-based reinforcement learning, the learning process involves acquiring knowledge of both the world model, which is a neural network representing the target environment, and a policy that operates within this world model. Typically, the policy learns through direct interaction with the data generated by the world model, as demonstrated in approaches such as Dreamer v2 [1]. However, a characteristic of conventional model-based RL is the absence of explicit control over the generation process. To enhance the performance of the final model, there is an opportunity to improve results by actively encouraging the model to generate more specific trajectories.

II. Main goals of the project

We focused on enhancing the degree of control within world models, utilizing methods such as PluGeN [2] to govern the trajectories generated by Dreamer v2 model, whose world model was trained utilizing the autoencoder VAE. The efficacy of our approach was assessed through experiments conducted on the Atari game Breakout. In the next phase of our project, we decided to simplify the problem by shifting our focus to a multi-task gridworld-based environment [3]. Our goal was to manage the generative process using more straightforward techniques, specifically relying on oversampling. The experimental environment used to evaluate this approach comprised two tasks that differed only in the size of the gridworld. Our strategy involved oversampling the more challenging (larger) task during policy training in the expectation of achieving a higher performance score on that specific task.

III. Results

During the evaluation of the PluGeN method on Breakout, three binary features were established for every frame of the game:

- horizontal position of the ball (indicating whether the ball is on the left or right)
- vertical position of the ball (indicating whether the ball is up or down)
- score feature (determining if the current score surpasses the median score for all frames in the dataset, set at 126)

To manipulate these features, we employed PluGen [2], a method that enables the alteration of features by disentangling latent space. In our initial approach, we randomly sampled frames from the dataset, encoded them into latent space, modified the predefined features, and then decoded them back into images for comparison with the original frames. However, we encountered challenges when applying this process to the first two features (vertical and horizontal position of the ball), as the ball consistently disappeared in all frames after passing through the decoder. Despite these difficulties, we achieved somewhat satisfactory results for the third feature, as depicted in Figure 1.

In the context of our exploration of oversampling strategies for gridworld-based environments, we established a setup comprising two tasks that differed solely in the size of the gridworld: a smaller (easier) task and a larger (harder) one. At the conclusion of each episode, the environment reset to the smaller task 99% of the time, with only a 1% probability of resetting to the harder task. This deliberate setup allowed us to diminish the performance of the larger task, providing an opportunity to assess oversampling strategies for subsequent testing and potential performance enhancement. The algorithm's learning process involves two primary components: learning the world model and learning the policy. Although we maintained the sampling approach for learning the world model unchanged, we implemented an oversampling strategy for the larger task during policy learning. We oversampled the larger task for policy learning so that it was sampled 50% of the time (given that the environment resets to the smaller task 99% of the time, initially, there is only a 1% probability of sampling the larger task from the replay buffer, which accumulates all the data from the training).

However, the oversampling strategy applied to policy learning did not yield the expected outcomes, as illustrated in Figure 2. We experimented with various oversampling percentages and alternative gridworld-based tasks, yet the results remained consistently similar. This discrepancy may be attributed to environmental noise within our gridworld-based experimental setup. Alternatively, another plausible explanation is that optimal performance in both the world model and policy learning is achieved when trained on similar datasets.

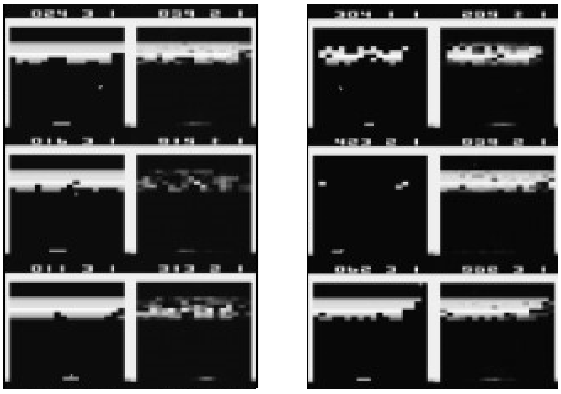

Figure 1: For each column, the left image displays samples from the dataset, while the right image showcases the same samples with a modified score feature. In the left image, the score feature was adjusted to ensure that the modified images have a score exceeding 126. In most cases, this adjustment results in a reduction in the number of blocks, indicative of a higher score. Conversely, in the right image, the score feature was modified to ensure that the modified images have a score less than 126. In most instances, this modification leads to an increase in the number of blocks, suggesting a lower score.

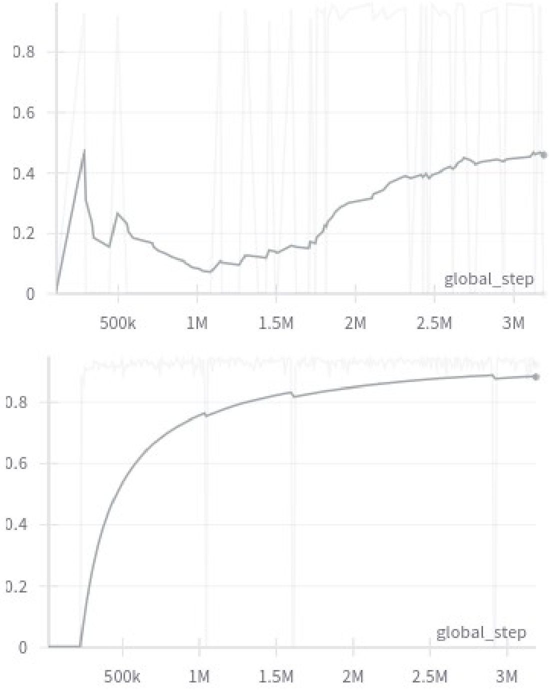

Figure 2: In the upper image, we assess performance on a larger task (of size 15x15), while in the lower image, we evaluate performance on an easier task (of size 9x9). These particular tasks are instances of a SimpleCrossing gridworld environment. Both tasks are integrated into a single environment, with an emphasis on oversampling the harder task. Despite a certain degree of knowledge transfer from the smaller to the larger task, and the oversampling of the larger task, the learning performance on the larger task remains significantly inferior to that of the smaller task.

IV. Division of work

The effort was distributed fairly evenly between us, with both collaborators contributing to every section of the project. The project was conducted between November 2023 and January 2024.

V. Bibliography

[1] Mastering Atari with Discrete World Models
[2] PluGeN: Multi-Label Conditional Generation From Pre-Trained Models
[3] https://minigrid.farma.org