Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [Transformers are sample efficient world models]
Date: [23-10-2024]

Make sure your critique Address the following points:
1. The problem the paper is trying to address
2. Key contributions of the paper
3. Proposed algorithm/framework
4. How the proposed algorithm addressed the described problem
Note: Be concise with your explanations. Unnecessary verbosity will be penalized. Please don't exceed 2 pages.

---

1 The problem the paper is trying to address

The paper addresses the problem of improving sample efficiency in reinforcement learning by learning in a world model. The main objective is to maximize the expected cumulative reward in a partially observable Markov decision process (POMDP). The problem can be formulated as:

2 Key contributions of the paper

1. The paper introduces a new agent, **IRIS** (Imagination with auto-Regression over an Inner Speech), that improves **sample efficiency** in reinforcement learning by learning policies in a simulated world model.
2. The world model is composed of:
   - A **discrete autoencoder** that converts high-dimensional image observations into a small number of discrete tokens.
   - A **GPT-like Transformer** that models the dynamics of the environment by autoregressively predicting future tokens.
3. The proposed method achieves a new state of the art in the **Atari 100k benchmark** for reinforcement learning methods without lookahead search, with only 2 hours of real-time experience.
4. **Qualitative Analysis:** The paper demonstrates that the world model can simulate important aspects of the game environment, including accurate pixel predictions, rewards, and episode terminations.
5. The code and models are made publicly available to **foster future research** in the area of transformers and world models for sample-efficient reinforcement learning.

Algorithm 1 IRIS Training Loop

procedure IRIS TRAINING LOOP
  Initialize policy π, discrete autoencoder (E, D), and transformer G
  for each epoch **do**
    Collect Experience:
  Collect experience from the real environment with the current policy π
    for each world model update step **do**
      Update World Model:
  Update the encoder E, decoder D, and transformer G using collected experience
    end for
    for each behavior learning step **do**
      Update Behavior:
  Update the policy π using trajectories imagined by the world model
    end for
  end for
end procedure

3 Proposed algorithm/framework

The framework consists of the following key components:

- **Discrete Autoencoder** (E, D)**:**
  - Encoder E converts raw pixel observations into discrete tokens.
  - Decoder D reconstructs images from these discrete tokens.
- **Transformer** G**:**
  - Autoregressively models the dynamics of the environment in terms of sequences of image tokens and actions.
  - Predicts the next state, reward, and episode termination.
- **Policy** π**:**
  - Learns from imagined trajectories produced by the world model.

4 How the proposed algorithm addressed the described problem

- The algorithm improves sample efficiency by training the policy π in a simulated environment (world model) instead of interacting directly with the real environment.
- The world model, composed of a discrete autoencoder (E, D) and a transformer G, simulates future trajectories, reducing real-world interactions.
- The policy is optimized using imagined trajectories, addressing the challenge of low sample efficiency in reinforcement learning.