Learning Optimal Strategies in Stochastic Grid-Based Puzzles via Policy Optimization

Ananya Singla, Arnav Mehta, Mohil Ahuja, Vaibhav Dabas
ananya.singla.ug23@plaksha.edu.in, arnav.mehta.ug23@plaksha.edu.in,
mohil.ahuja.ug23@plaksha.edu.in, vaibhav.dabas.ug23@plaksha.edu.in

Abstract

This project addresses the challenge of learning optimal strategies in stochastic, grid-based puzzle environments. We propose a reinforcement learning (RL) framework where an agent interacts with a dynamic tile-matching grid to maximize long-term scores under limited moves. Unlike traditional single-step reward optimization, the environment incorporates delayed rewards from chain reactions and board objectives. We evaluate **Proximal Policy Optimization (PPO)**, **Group Relative Policy Optimization (GRPO)**, and additional PPO variants to determine their efficiency and stability. All experiments are designed to run on standard consumer laptops using lightweight Gymnasium-based simulations.

Keywords: Reinforcement Learning, PPO, GRPO, Grid-Based Games, Puzzle Optimization.

I. Introduction and Related Work

Grid-based match-three puzzle games present a significant RL challenge due to large combinatorial state spaces, stochastic board generation, and finite action budgets. Traditional algorithms often struggle because greedy reward maximization fails to capture long-term cascade values and board control strategies.

Recent advances in policy-gradient methods, specifically PPO and its variants, provide stable learning through clipped policy updates and efficient on-policy training without massive compute requirements. This project extends prior research on board games to match-three environments, focusing on PPO baseline performance and GRPO for improved relative reward estimation. Unlike compute-heavy systems, this work targets CPU-friendly simulations suitable for undergraduate research.

II. Problem Statement and Proposed Work

A. Grid-Based Puzzle Environment

We design a lightweight environment featuring a discrete N×N grid of colored tiles. The action space involves swapping adjacent tiles, with invalid moves penalized. The episode ends after a fixed move limit or goal completion.

The reward structure is divided into:

- **Immediate Reward:** Points gained from initial matches.
- **Delayed Reward:** Bonuses from chain reactions (cascades) and special-tile interactions.
- **Objective Reward:** Clearing specific targets such as score thresholds or level-specific obstacles.

B. Proposed Reinforcement Learning Methods

Following the course's empirical evaluation track, we compare three distinct approaches:

1. **PPO Baseline:** Agents learn a stochastic policy using a clipped surrogate objective and a value-function baseline.
2. **GRPO Extension:** We implement Group Relative Policy Optimization to compare relative advantages within trajectory groups, aimed at reducing variance from sparse, high-reward cascade events.
3. **PPO Variants:** We evaluate lightweight modifications, including **Entropy-annealed PPO** and **Reward-normalized PPO**, chosen specifically for CPU-efficient training.

The implementation utilizes **Python + Gymnasium** for the environment and **PyTorch** for small policy networks (MLP or CNN), ensuring feasibility on standard laptops.

III. Proposed Evaluation

Agents will be evaluated across 10,000 simulated episodes to ensure a thorough showcase of effectiveness. The metrics include:

- **Score Performance:** Average final score and consistency across random seeds.
- **Move Efficiency:** Score generated per move and success rates in completing level objectives.
- **Learning Stability:** A comparison of convergence speeds between PPO and GRPO and an analysis of variance in training curves.
- **Generalization:** Performance on unseen board layouts and transferability to larger grid sizes.

IV. References

- [1] https://github.com/karan1149/candy-crush-rl
- [2] R. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.
- [3] https://ieeexplore.ieee.org/document/10069395/
- [4] K. He, B. Banerjee, and P. Doshi, "Cooperative-Competitive Reinforcement Learning with History-Dependent Rewards," *arXiv:2010.08030v1*, 2020.