RL-Squared Final Report

Anshuman Dash, Phillip Peng, Matthew Shen
Department of Computer Science
University of California, Santa Barbara
Email: {anshumandash, phillippeng, matthewshen}@ucsb.edu

***Abstract*—We aim to investigate the use of Reinforcement Learning in the context of the game of Rocket League. Through the uses of PPO and Curriculum Learning through self-play, we train bots on a variety of reward structures and hyperparameters. To do an empirical comparison of the different models, we evaluate the bots against each other in a tournament setting of 1v1 matches to determine model ELO. We analyze the results of this tournament as well as individual models and discuss the implications of our findings.**

I. Introduction

Reinforcement Learning (RL) is a powerful tool for training agents to perform tasks in a variety of environments. We aim to investigate the use of RL in the context of the game of Rocket League. Rocket League is a popular video game that combines soccer with rocket-powered cars. The game is played in a 3D environment, and players control their cars to hit a ball into the opposing team's goal. The game itself holds a large number of states making it a challenging environment for RL agents to learn in.

II. Related Work

The use of RL to play games has been a popular research topic in recent years, most famously with AlphaGo. The application of RL in other games such as Dota 2 and Starcraft II has also been explored.

A. PPO

III. Model and Methods

Through this work we largely explore how a change in reward structure as well as model hyperparameters affect the performance of trained agents. All models are trained using the Proximal Policy Optimization (PPO) algorithm.

A. Reward Scaling

B. Reward Structures

C. Curriculum Learning

Curriculum learning is a training strategy where the learning process is organized in a way that gradually increases in complexity.

We find that this doesn't perform well for a variety of reasons. The increasing sparsification of rewards makes it difficult for the agents to learn effectively. In order to address this, we propose a modified approach that balances exploration and exploitation.

IV. Results

We evaluate the performance of the trained agents by having them play against each other in a Placket Luce tournament [**?**]

V. Conclusion

VI. Future Work

References

[1] A. Sigal, H.-C. Lin, and A. Moon, "Improving generalization in reinforcement learning training regimes for social robot navigation," *arXiv preprint arXiv:2308.14947*, 2023.

Algorithm 1 PPO Algorithm

Input: initial policy parameters θ_0, initial value function parameters ϕ_0
for k = 0,1,2, ... **do**
    Collect set of trajectories $\mathcal{D}_k$ by running policy $\pi_k = \pi(\theta_k)$ in the environment
    Compute rewards-to-go $\hat{R}_t$
    Compute advantage estimates $\hat{A}_t$ for current value function V_{ϕ_k}
    Update policy by maximizing PPO-Clip objective:
        $\theta_{k+1} = \arg\max_{\theta} \theta \mathbb{E}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} \hat{A}_t \right]$
end for