Model Overview

This submission features a reinforcement learning (RL) agent trained using Proximal Policy Optimization (PPO) with recurrent layers. The agent is designed to excel in 1v1 platform fighting scenarios in Brawlhalla, utilizing self-play and a curated reward function to optimize its strategy. The model was trained for approximately 744,186 timesteps and demonstrates strong performance in both offensive and defensive play.

Training Process

The agent was trained from scratch using Recurrent PPO (R-PPO) to enable adaptive decision-making and improved long-term action planning. Self-play was incorporated with different opponent types, including the BasedAgent, to ensure diverse and challenging training conditions. The final model was selected from a checkpoint at timestep 744,186 based on its performance metrics and stability.

Reward Function Design

The reward function was carefully structured to balance offensive aggression with survival-oriented strategies. The primary reward components included:

- **Damage Interaction Reward (weight: 7.0)**: Encourages the agent to successfully land attacks and deal damage.
- **Danger Zone Reward (weight: 3.5)**: Penalizes the agent for being in high-risk areas, promoting stage control.
- **Head to Opponent (weight: 0.08)**: Encourages movement towards the opponent for active engagement.
- **Target Height Reward (weight: 0.05)**: Maintains a preferred height on the platform.
- **Penalize Attack Reward (weight: -0.5)**: Discourages unnecessary taunting and inefficient attack strategies.
- **Taunt Reward (weight: 0.1)**: Minor incentive for taunting, allowing for psychological play.

Additional reward signals were included for critical game events:

- **Win Signal (weight: 100)**: Large reward for match victories.
- **Knockout Signal (weight: 15)**: Incentivizes successfully eliminating opponents.
- **Combo Reward (weight: 8)**: Rewards executing attacks on stunned opponents.

Training Performance

The final model exhibited the following key training statistics:

- **Episode Length Mean**: 494 steps
- **Episode Reward Mean**: 614

- **Entropy Loss**: -16.8 (indicating stable policy exploration-exploitation balance)
- **Explained Variance**: 0.981 (demonstrating strong predictive capability of the value function)
- **Policy Gradient Loss**: -0.0092
- **Learning Rate**: 0.0003

These metrics suggest that the agent successfully learned an optimal balance between aggression and survival, achieving strong generalization within the training environment.

Opponent Selection & Self-Play

To refine its competitive performance, the agent trained against various opponents, with the primary opponent pool including:

- **Self-Play Agent (8 instances)**: Facilitated self-improvement through iterative learning.
- **BasedAgent (2 instances)**: Provided a structured baseline opponent.

This setup allowed the agent to develop robust strategies adaptable to different playstyles.

Conclusion

The final trained model is a well-rounded competitor, effectively balancing aggression and defense while maintaining stage control. Through structured reinforcement learning, self-play, and a well-designed reward system, the agent has achieved a high level of proficiency in the game environment. The model is expected to perform competitively in tournament settings, leveraging its adaptive strategies and learned behaviors to outmaneuver opponents.