Summary Report for PPO Model

Introduction

This report outlines implementing, training, and evaluating a Proximal Policy Optimization (PPO) model applied to the Connect4 game. PPO is an advanced reinforcement learning algorithm designed to improve stability and performance by optimizing policy updates within a trust region.

Model Architecture

The PPO model architecture consists of the following components:

1. Input Layer: Processes the Connect4 board state as a tensor of shape (1, 1, 6, 7).
2. Feature Extraction:
   - Convolutional Layer 1: 32 filters, kernel size 3x3, stride 1, ReLU activation.
   - Convolutional Layer 2: 64 filters, kernel size 3x3, stride 1, ReLU activation.
3. Fully Connected Layers:
   - Shared Layer: Extracts a common feature representation with 128 units and ReLU activation.
4. Policy and Value Heads:
   - Policy Network: Outputs probabilities for each action (7 possible actions).
   - Value Network: Predicts the expected return from the current state.

Training Process

The training process for PPO involves:

1. Environment Interaction:
   - The agent interacts with the environment to collect trajectories (state, action, reward, next state, done).

2. Policy Optimization:
   - Updates are constrained using a clipped objective to ensure changes stay within a trust region:

$$\boxed{L^{CLIP}(\theta) = E_t\left[\min\left(r_t(\theta)\widehat{A_t}, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\widehat{A_t}\right)\right]}$$

Surrogate Objective:

1. **$L^{CLIP}(\theta)$**: The clipped objective function that PPO seeks to maximize.

Probability Ratio:

2. $\mathbf{R_t(\theta) = \pi_\theta(a_t|s_t) / \pi_{\theta\ old}(a_t|s_t)}$:

    - The ratio of the new policy probability to the old policy probability for action a_t and s_t
    - Ensures that updates are bounded.

Advantage Estimate:

3. $\mathbf{\hat{A}_t}$: Advantage function, which measures how much better an action is compared to the average action in a state.

Clipping:

4. $\mathbf{clip(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon)}$:

    - Clips $r_t(\theta)$ to the range $[1 - \varepsilon, 1 + \varepsilon]$, where ε is a small hyperparameter (e.g., $\varepsilon = 0.2$).
    - Prevents large updates, ensuring stability.

Minimization:

5. The objective uses **min** to select the clipped surrogate objective if $r_t(\theta)$ goes beyond the clipping range.

    - Ensures that updates do not diverge too far from the old policy.

3. Advantage Estimation:
  - Computes advantages using the value function and discounted rewards.

4. Value Network Optimization:
  - Minimizes the mean squared error between predicted and actual returns.

Key training parameters:
- Learning Rate: 1e-4
- Discount Factor (Gamma): 0.99
- Clipping Parameter (Epsilon): 0.2
- Batch Size: 64
- Epochs per Update: 4

Evaluation

The PPO agent was evaluated in the Connect4 environment. Key metrics include:

1. Cumulative Rewards: Tracks the agent's learning curve over episodes.
2. Win Rate: Measures the agent's success against a random opponent.

3. Policy Improvement: Analyzes the agent's ability to prioritize actions with higher expected returns.

The PPO model demonstrated steady improvements in performance and learning stability across episodes.

Conclusion

The PPO model effectively leverages policy and value networks to solve the Connect4 game. Its robust training process and clipping mechanism ensure stable and efficient learning, making it a powerful tool for reinforcement learning tasks. Further improvements, such as reward shaping and testing against diverse opponents, can enhance its performance.