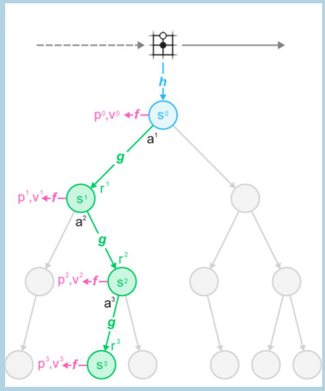
Smarter Moves: Enhancing the Exploration Method of MuZero

Action Selection in Model-Based Reinforcement Learning

Francisco Ruas Vaz
Delft University of Technology

MuZero Architecture



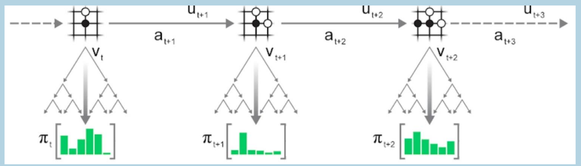


Image source: Julian Schrittwieser

Introduction and Background

- MuZero is a state-of-the-art model-based reinforcement learning agent that achieves superhuman performance by learning its own model of the environment and planning with Monte Carlo Tree Search (MCTS)
- At each turn, it observes the current environment maps it to its own internal model and then uses this model to perform MCTS, simulating future trajectories to identify the most promising moves. It then picks one of these actions to take in the actual environment
- A critical but unexplored component is its action selection strategy, how it chooses an action at the end of planning, at the root node of tree

Methodology

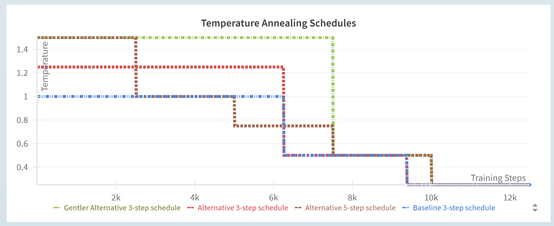
Research Goal: To investigate how modifying this action selection strategy impacts MuZero's performance across 3 different Atari games, which vary in complexity (increasing from left to right)

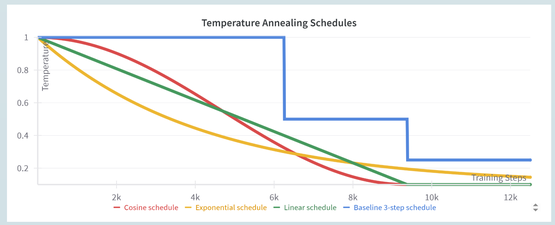

Pong


Breakout

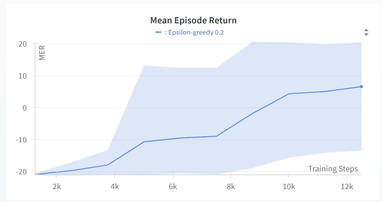

Ms. Pac-Man

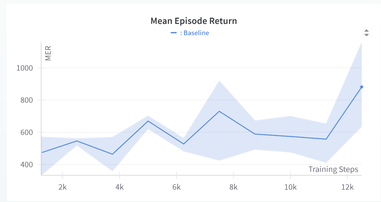
- We first used the **baseline**, a standard 3-step temperature schedule.
- Then for a simple comparison we tried the less sophisticated **epsilon-greedy**, with epsilon value of 0.2 and 0.5
- **Alternative step schedules**, some with higher initial exploration and one with a more granular schedule
- Different **temperature annealing functions** were also implemented: Linear, Exponential and Cosine
- Finally a more advanced strategy was used, **Thompson Sampling** (TS), using a 5-model ensemble

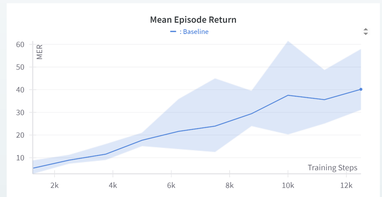


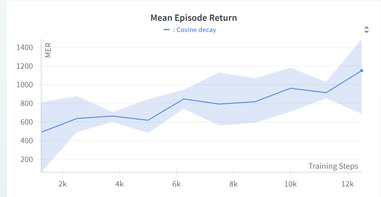


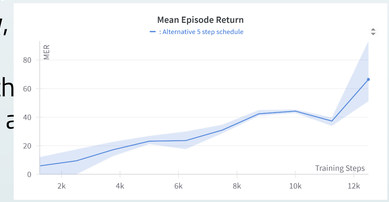
Results

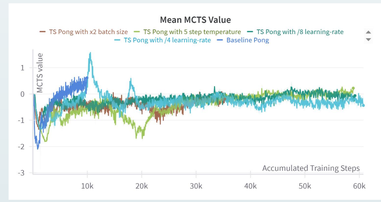












- Epsilon-greedy performs worse than the baseline even in Pong, with much higher variance error
- The smooth Cosine annealing schedule was the best performing strategy overall, especially in the harder games
- As task difficulty increases, the choice of exploration strategy becomes critical
- For Pong there was not much difference between temperature methods, annealing exploration pulls ahead in Breakout and in Ms. Pac-Man performance different is significant
- Thompson Sampling completely failed to achieve competitive returns, stalling at a low score. Interestingly, however, its internal MCTS values were normal and similar to the successful baseline agent

Conclusion

- The best strategy is task-dependent. Simple methods work for simple games, but fail on complex ones
- For harder tasks, smooth and sustained exploration like Cosine annealing, consistently outperforms the simpler random methods as well as the harsher step schedules
- Computational Constraints: Training MuZero is extremely time-consuming (4-5 hours per run). This limited the number of runs and the depth of hyperparameter tuning.
- The TS experiments were only completed for Pong due to the long training times, and its failure prevented a full comparison on harder games.

- Future Work
  - Debug Thompson Sampling, potentially by making it pick a model per episode instead of action
  - Implement methods that dynamically adjust exploration during training
  - Conduct a large-scale study to find the true optimal parameters for each strategy and environment
  - Test on on other environments, especially on harder exploration games like Montezuma's Revenge

References

- MuZero: Model-based planning with MCTS [Schrittwieser et al., 2020]
- MuZero research: What model does MuZero learn? [He et al., 2024]
- Thompson Sampling: Posterior sampling for exploration [Osband et al., 2013]