Hessisches Kompetenzzentrum
für Hochleistungsrechnen

Hindsight Experience Replay for neural-guided Monte Carlo Tree Search



Project Manager
Jannis Brugger

Researchers
Prof. Dr. Kristian Kersting, Cedric Derstoff and Alexandros Vazaios

Principal Investigator
Prof. Dr.-Ing. Mira Mezini

Project Term
2024 - 2025

Clusters
Lichtenberg II Cluster Darmstadt

Software
PyTorch

Institute
Software Technology

University
Technische Universität Darmstadt

Introduction

Modern artificial intelligence often struggles with tasks where success is rare or difficult to define. In many scenarios, an autonomous agent only receives a signal of “success” when it perfectly completes a complex goal. If the agent fails, which it does frequently during the early stages of learning, it receives no feedback at all. This “sparse reward” problem makes learning incredibly slow, as the agent is essentially searching for a needle in a haystack without any hints. To overcome this, we utilize a technique called Hindsight Experience Replay. This method allows the agent to learn from its failures by pretending that whatever outcome it actually achieved was the goal all along. By “re-labeling” these past attempts, the agent learns the underlying mechanics of its environment much faster. However, combining this technique with advanced decision-making frameworks, specifically those that use look-ahead planning like the famous AlphaZero system, is a significant challenge. The primary motivation for this project was to bridge this gap. We required a High Performance Computer (HPC) because testing these combinations is computationally massive. Training a single agent requires thousands of simulations and constant updates to a complex neural network. To find the best settings, we had to run dozens of these intensive simulations simultaneously. Without the parallel processing power of an HPC, which allows us to run these tasks in “parallel” rather than one after the other, this research would have exceeded the timeframe of a bachelor thesis.

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN
TECHNISCHE UNIVERSITÄT DARMSTADT
UNIKASSEL VERSITÄT
Philipps Universität Marburg

Methods

The project centered on developing a configurable framework that integrates “look-ahead” planning with hindsight learning. The core method used is a search process that simulates many possible future moves before deciding on the best action. This search is guided by a neural network that predicts which moves are likely to lead to a win. When the agent fails to reach a specific target, our method looks back at the data and identifies a “virtual” goal that the agent did manage to reach. This data is then fed back into the neural network to improve its accuracy. Our research focused on how often these virtual goals should be used and how they should be selected to ensure the agent learns useful skills rather than getting distracted by irrelevant information.

Results

Our findings demonstrate that combining look-ahead search with hindsight learning creates a significantly more robust learner. In our initial phase, we tested the system on classic spatial puzzles where a “plain” version of the AI (without hindsight) failed to learn anything at all because it never accidentally stumbled upon the goal. In the second phase of the project, we benchmarked our configurable implementation against various settings. We discovered that by fine-tuning the ratio of “real” goals to “relabeled” goals, we could achieve success in environments that were previously considered unsolvable for this type of architecture. The HPC allowed us to verify these results across multiple different scenarios, ensuring that our success wasn't just a “lucky run” but a consistent improvement in the algorithm's intelligence.

Discussion

The results confirm that learning from failure through goal relabeling is a powerful tool when paired with strategic search algorithms. We have shown that the “AlphaZero” style of AI, which is world-class at games like Chess or Go, can be made much more flexible for general tasks where rewards are not easily found. The special challenge we overcame was balancing the complexity of the search with the overhead of the learning process.

Last Update: 2026-05-11 09:29