Limitations In Planning Ability In AlphaZero

Daisy Lin (xl1005@nyu.edu)
Center for Neural Science, 6 Washington Place
New York, NY 10003, USA
Brenden Lake (brenden@nyu.edu)
Center for Data Science and Department of Psychology, 6 Washington Place
New York, NY 10003, USA
Wei Ji Ma (weijima@nyu.edu)
Center for Neural Science and Department of Psychology, 6 Washington Place
New York, NY 10003, USA

Abstract

AlphaZero, a deep reinforcement learning algorithm, has achieved superhuman performance in complex games like Chess and Go. However, its strategic planning ability beyond winning games remains unclear. We investigated this using 4-in-a-row, a game used to study human planning. We analyzed AlphaZero's feature learning and puzzle-solving abilities. Despite strong gameplay, AlphaZero exhibited a 45% failure rate in puzzles. Feature analysis revealed limitations in its learned knowledge during self-play. We incorporated human-inspired features into its policy and value outputs, leading to a 13% improvement in puzzle-solving accuracy. Our findings highlight the potential for human insights to enhance AI strategic planning beyond self-play.

Keywords: Explainable AI; Deep Reinforcement Learning; Human-inspired AI

Introduction

While AlphaZero's mastery of complex games like Chess and Go is undeniable (Silver et al., 2017, 2018), a key question remains: what exactly does it learn through self-play, and are there limitations in its planning strategy? One crucial aspect of human intelligence is planning: the ability to simulate potential future states and actions. Humans excel at strategic planning in dynamic environments – a skill many AI models struggle with (Valmeekam, Olmo, Sreedharan, & Kambhampati, 2022). However, a comprehensive understanding of the difference between human and AI planning remains elusive.

This study bridges this gap by investigating AlphaZero planning mechanisms through the lens of human planning. We leverage the game of '4-in-a-row,' a task used to study human planning (van Opheusden et al., 2023). By training AlphaZero on '4-in-a-row' and comparing its performance with an established human cognitive model, we aim to uncover AlphaZero's concept learning and planning processes.

Methods

Task 4-in-a-row is a two-player game where players take turns placing pieces into a grid, aiming to connect four of their color horizontally, vertically, or diagonally.

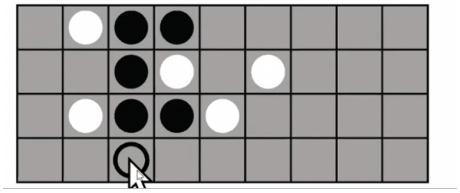

Figure 1: Examples board

AlphaZero Every agent consisted of a deep neural network (DNN) with Monte Carlo Tree Search (MCTS) (Silver et al., 2017, 2018). During training, the agents played 100 self-play games per iteration. Training examples consisted of state, MCTS output, and game outcome. The DNN was trained to predict both the value and policy using mean-squared error and cross-entropy loss functions respectively. ADAM optimizer updated the DNN with mini-batches of past training data. Network updates were accepted based on winning more than 50% games against the current best network.

Results

We investigated AlphaZero's planning ability in 4-in-a-row. We addressed two key questions: (1) Can AlphaZero acquire concepts similar to humans, and (2) Can we leverage human insights to improve its performance?

Playing strength

We first established a human performance benchmark to assess AlphaZero's playing strength. Top eight agents trained via self-play consistently surpassed the most skilled human player's Elo $(\text{mean Elo difference} = 90.4, \text{SD} = 17.7)$, a measure of playing strength (Glickman & Jones, 1999), demonstrating AlphaZero's effectiveness in winning games.

Feature Analysis

Probing To understand how AlphaZero became good at winning games, we employed feature probing techniques akin to concept activation vectors (Kim et al., 2018). This approach allowed the detection of features used by human players including 3-in-a-row and 2-in-a-row, identified by van Opheusden et al. (2023). We trained a classifier using the activations at a specific layer during a given training iteration to predict the presence of these human-used features.

This analysis revealed the network's acquisition of the crucial 3-in-a-row feature in both the value head and intermediate layers, even without exposure to human-generated data (Figure 2). However, this analysis did not identify the representation of another human-used feature, 2-in-a-row, within the network. This finding suggested potential limitations in AlphaZero's ability to learn the full spectrum of strategic features used by humans.

Feature representation with unsupervised methods We further explored what AlphaZero learned through self-play without using predefined concepts. To achieve this, we applied Nonnegative Matrix Factorization (NMF) to extract and visualize latent features from hidden layers (Lee, 2000; McGrath et al., 2022). We concatenated the activations from 14907(N) random game states into a matrix $Z \in \mathbb{R}^{36N \times 256}$, and approximated Z as the product of a weight matrix $F \in \mathbb{R}^{K \times 256}$ and feature matrix $\Omega \in \mathbb{R}^{36 \times K}$, minimizing the reconstruction error. The resulting factors gave insights into

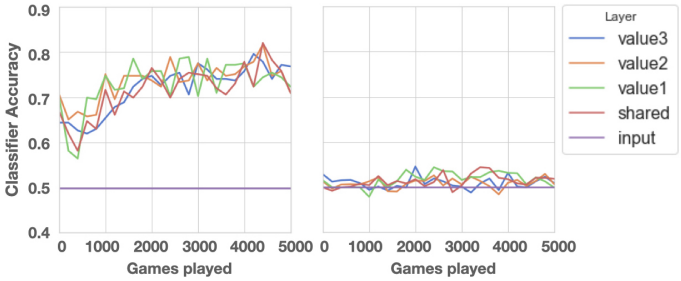

Figure 2: Feature Probing Analysis: '3-in-a-row' (left) versus '2-in-a-row' (right). Activations from the value head and a shared intermediate layer demonstrate learning of the 3-in-a-row and 2-in-a-row. Control inputs are included for reference.

the network's understanding of the game by highlighting important activation patterns.

NMF analysis revealed interpretable factors in the network's intermediate layers (Figure 3). These factors captured diagonal, vertical, and horizontal patterns, suggesting AlphaZero's ability to represent various game-relevant features.

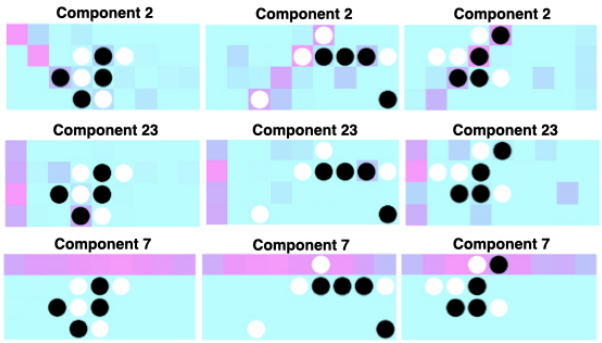

Figure 3: Visualization of NMF for selected factors. Panels show features captured by different residual blocks: diagonals, verticals, and horizontals.

Puzzle Testing

To test AlphaZero's problem-solving ability, we designed 92 puzzles derived from 4-in-a-row. Puzzles are particular game states that has a forced win for the current player within five moves (Figure 5). These puzzles demanded the construction of sequential threats to secure win.

Despite its strong playing strength, AlphaZero showed a surprising 45% failure rate in finding the forced win in these puzzles. In some instances, the agent showed overly defensive play, neglecting opportunities to build threats (Figure 5). This observation suggested a potential gap between AlphaZero's learned features and the specific features and reasoning used by humans.

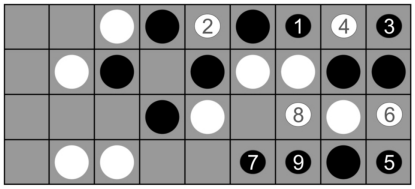

Figure 4: An example of a puzzle solved in 5 moves. Numbers indicate the order in which players placed their pieces.

We hypothesized that incorporating human-inspired features could enhance AlphaZero's in puzzle-solving performance. We incorporated a cognitive value function into both the policy and value output, using features not readily observed in the network's self-learned repertoire, including 2-in-a-row, and unconnected-2-in-a-row. This integration resulted in a significant 13% improvement in puzzle-solving accuracy. This finding highlighted the potential of incorporating human cognitive insights to augment AI performance in tasks requiring specific strategic reasoning patterns.

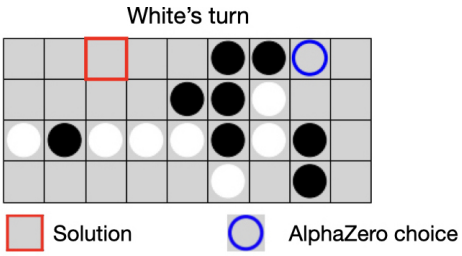

Figure 5: An example of AlphaZero's failure. AlphaZero (blue circle) chose to block opponent features instead of building threats.

Discussion

This study investigated AlphaZero's strategic planning in 4-in-a-row. Our findings offered insights into its potential limitations. Our feature analysis revealed a duality: AlphaZero learned human-interpretable features, but may not fully represent all human-used features. Despite superhuman playing strength, AlphaZero struggled with puzzles requiring a "logical sequence" reasoning (Steingrimsson, 2021). This suggested a gap between its learned strategies and human-like planning. To bridge this gap, we introduced human-inspired features to AlphaZero's policy and value estimations, which improved its puzzle-solving accuracy.

Our findings advocate for further exploration of human-inspired features in AI. This approach highlights the power of human insights in augmenting AI performance, and holds promise for expediting learning and improving adaptability in AI planning.

Acknowledgments

I am grateful for the support and guidance provided by Wei Ji Ma lab. Special appreciation is extended to Jeroen Olieslagers, Jordan Lei, and Nastaran Arfaei, whose comments improved the quality of this work. Additionally, I would like to thank Jieyu Li and Shucheng Li for their invaluable help in generating the puzzle bank.

References

Glickman, M. E., & Jones, A. C. (1999). Rating the chess rating system. *CHANCE-BERLIN THEN NEW YORK-*, *12*, 21–28.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning* (pp. 2668–2677).

Lee, S. (2000). Algorithms for non-negative matrix factorization. *NIPS*.

McGrath, T., Kapishnikov, A., Tomašev, N., Pearce, A., Wattenberg, M., Hassabis, D., ... Kramnik, V. (2022). Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, *119*(47), e2206625119.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, *362*(6419), 1140–1144.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... others (2017). Mastering the game of go without human knowledge. *nature*, *550*(7676), 354–359.

Steingrimsson, H. (2021). Chess fortresses, a causal test for state of the art symbolic [neuro] architectures. In *2021 ieee conference on games (cog)* (pp. 1–8).

Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2022). Large language models still can't plan (a benchmark for llms on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*.

van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, *618*(7967), 1000–1005.