AlphaZero: Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm

The game of chess is the most widely-studied domain in the history of artificial intelligence. Recently, the AlphaGo Zero algorithm achieved superhuman performance in the game of Go, by representing Go knowledge using deep convolutional neural networks, trained solely by reinforcement learning from games of self-play.

AlphaZero is a general-purpose reinforcement learning algorithm that can achieve superhuman performance across many challenging domains without any additional domain knowledge, except the rules of the game. AlphaZero achieved superhuman level of play within 24 hours in the games of chess and shogi (Japanese chess) as well as Go, and convincingly defeated a world-champion program, Stockfish, Elmo, and 3-day version of AlphaGo Zero in each case.

Instead of a handcrafted evaluation function and move ordering heuristics, AlphaZero utilises a deep neural network. This neural network takes the board positions as an input and outputs a vector of move probabilities for each action and a scalar value estimating the expected outcome. AlphaZero learns these move probabilities and value estimates entirely from self-play and these are then used to guide its search. Instead of an alpha-beta search with domain-specific enhancements, AlphaZero uses a general-purpose Monte-Carlo tree search (MCTS) algorithm.

Comparison of Alpha Zero vs AlphaGoZero

AlphaZero	AlphaGo Zero
Estimates and optimizes the expected outcome by taking account of draws or potential other outcomes	Estimates and optimizes the probability of winning, assuming binary win/loss outcomes
The rules of chess and shogi are asymmetric. It does not augment the training data and does not transform the board position during MCTS	The rules of Go are invariant to rotation and reflection. It augments the training data and transforms the board positions during MCTS
It simply maintains a single neural network that is updated continually, rather than waiting for an iteration to complete.	After each iteration of training, the performance of the new player was measured against the best player and if it won by a margin then it replaced the best player.
Reuse the same hyper-parameters for all games without game-specific tuning	Tuned the hyper-parameter of its search by Bayesian optimization
Board state and actions are encoded by spatial planes or a flat vector based only on the basic rules for each game.	Board state and actions are encoded by spatial planes or a flat vector based only on the basic rules for each game.
Hardware: 4 TPUs [v2], single machine	Hardware: 4 TPUs [v2], single machine

Results

Game	Win	Draw	Loss
Chess	28	72	0
Shogi	90	2	8
Go	60	-	40

Tournament evaluation of AlphaZero in chess, shogi, and Go, as games won, drawn or lost from AlphaZero's perspective, in 100 game matches against Stockfish, Elmo, and the recent AlphaGo Zero after 3 days of training respectively. Each program was given 1 min of thinking time per move.