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

Tony Allen and Gavin Glenn

PURDUE UNIVERSITY

Chess, Shogi, and Go

Chess

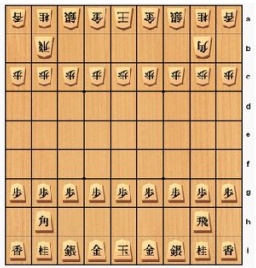
- Pieces have varying movement abilities, goal is to capture opponent's king

Shogi

- Similar to chess, but larger board and action space

Go

- Much larger board, but all pieces have same placement rules



Wikipedia.com

Board positions: 10^{50} 10^{120} 10^{170}

PURDUE UNIVERSITY

Properties of Chess, Shogi, Go

Rules	Go	Chess and Shogi
Translational Invariance	Yes	Partial (e.g. pawn moves, promotion)
Local	Yes (based on adjacency)	No (e.g. queen moves)
Symmetry	Yes (dihedral)	No (e.g. pawn moves, castling)
Action Space	Simple	Compound (move from/to)
Game Outcomes	Binary (look at prob. of winning)	Win / lose / draw (look at expected value)

2017 NIPS Keynote by DeepMind's David Silver

Go seems more amenable to CNNs

PURDUE UNIVERSITY

Computer Chess and Shogi: Prior Approaches

Computer chess is the most studied domain in AI

- Highly specialized approaches have been successful
- Deep Blue defeated World Champion G. Kasparov in 1997
- Before AlphaGo, state-of-the-art was Stockfish

Shogi only recently achieved human world champion level

- Previous state-of-the-art was Elmo

Engines like Stockfish and Elmo are based on alpha beta search

- Domain specific evaluation functions tuned by grandmasters

AlphaZero, starting from first principles

AlphaZero assumes no domain specific knowledge other than rules of the game

- Compare with Stockfish and Elmo's evaluation functions
- Previous version AlphaGo started by training against human games
    - It also exploited natural symmetries in Go both to augment data and regularize MCTS

Instead, AlphaZero relies solely on reinforcement learning

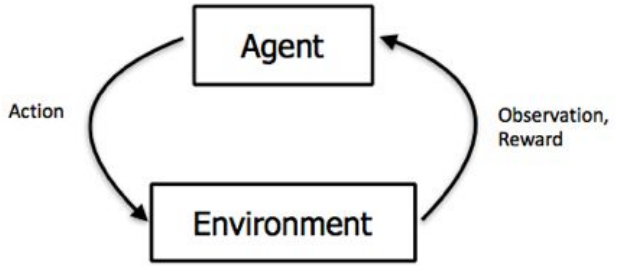
- Discovers tactics on its own

Reinforcement Learning

Agent receives information about its environment and learns to choose actions that will maximize some reward[1]

Compare with supervised learning, but instead of learning from a label, we learn from a time delayed label called reward

- Learn good actions by trial and error



rail.eecs.berkeley.edu

1. *Deep Learning with Python,* F. Chollet

PURDUE UNIVERSITY

Reinforcement Learning

This framework is better suited to games like GO

In theory a network could learn to play using supervised learning, taking recorded human Go games as input

- Large suitable datasets may not exist
- More importantly, the network will only learn to perform like a human expert, instead of learning true optimal strategy

RL can be done through self play

- Using current weights, play out an entire match against self
- Update weights according to results

PURDUE UNIVERSITY

AlphaZero

Deep Neural Network $\mathbf{f}_{\boldsymbol{\theta}}$ with parameters θ

Input:

- Raw board representation, **s**

Output:

- Policy vector, **p**
- Expected value of board position, **v**

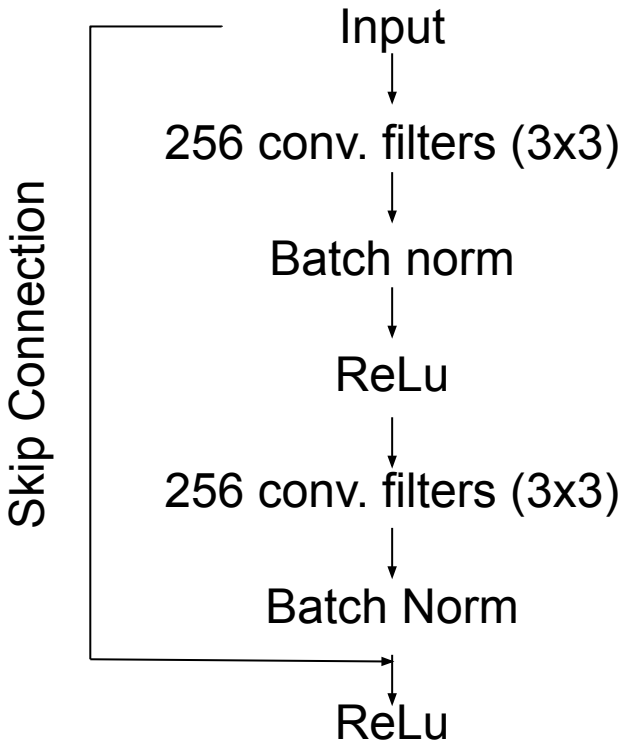
Loss:

$$\boldsymbol{Loss = (z - v)^2 - \pi^{T} \log(\mathbf{p}) + \mathbf{c}||\theta||^2}$$

Where $\boldsymbol{\pi}$, $\boldsymbol{z}$ come from tree search and self-play

PURDUE UNIVERSITY

AlphaZero



Architecture?

- Not stated explicitly (implied to be same as AlphaGo Zero)
- 20 (or 40) residual blocks of convolutions

PURDUE UNIVERSITY

AlphaZero

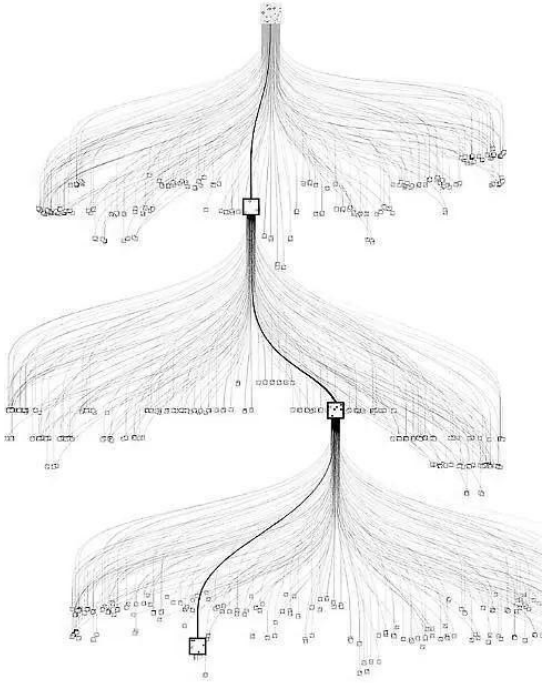
Monte Carlo Tree Search

Goal: Reduce depth and breadth of search

- Move distribution **p** helps reduce breadth
- Value **v** helps reduce depth

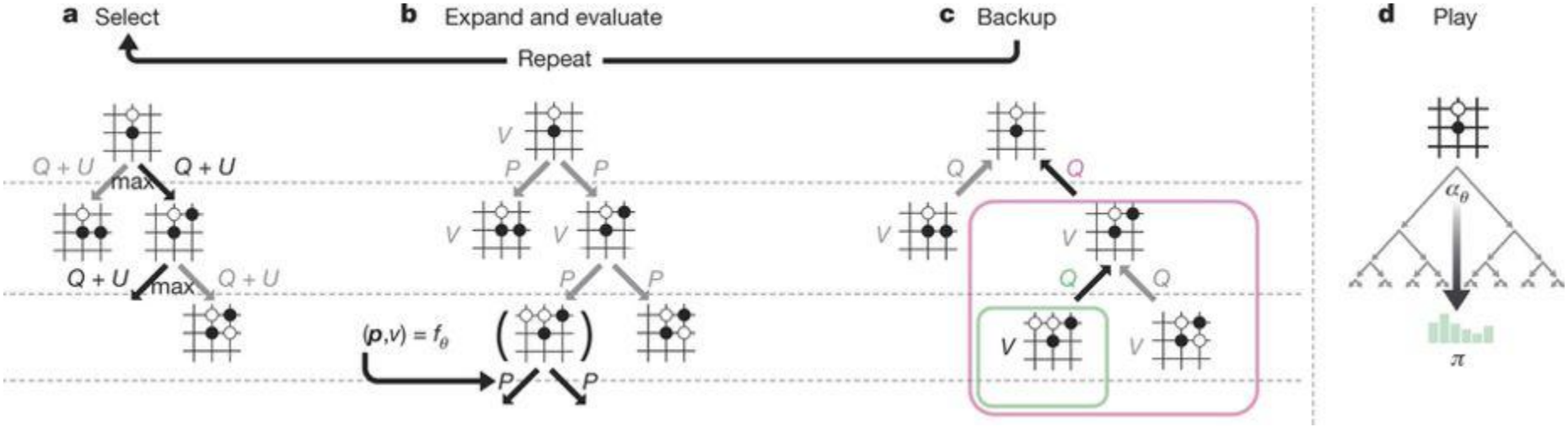
Not *exactly* what happens

It's best to see a picture!



2017 NIPS Keynote by DeepMind's David Silver

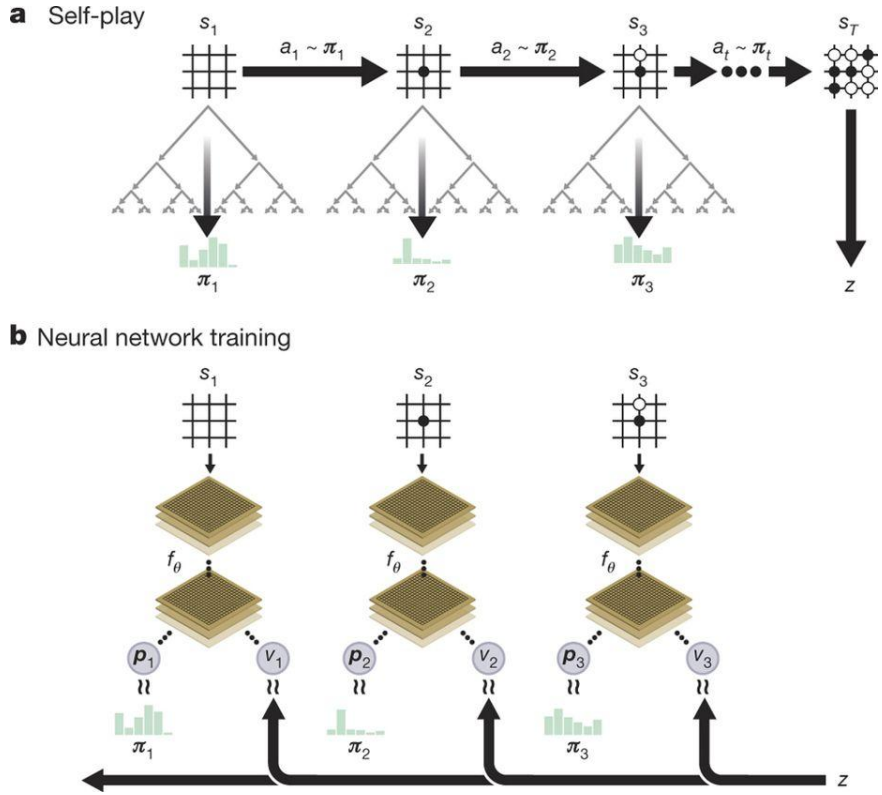
PURDUE UNIVERSITY



Monte Carlo Tree Search In a Picture

D Silver *et al. Nature* **550,** 354–359 (2017) doi:10.1038/nature24270

Training



Algorithm Pseudocode

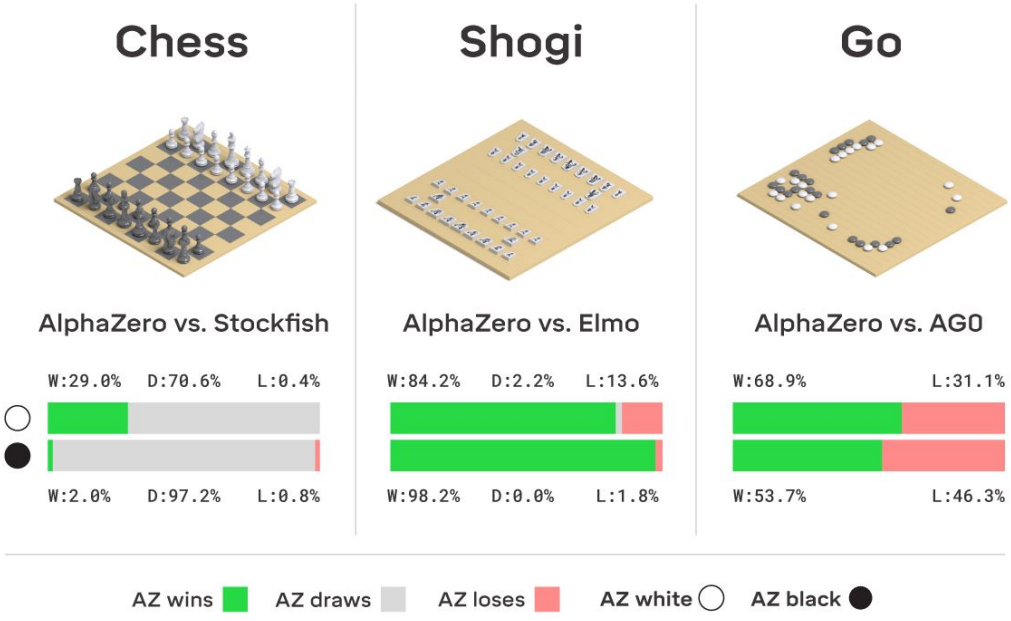
Initialize $\mathbf{f}_\theta$

Repeat:
  Play Game:
    While game not over:
      MCTS from state **s**
      Compute $\boldsymbol{\pi}$
      Pick new **s** from $\boldsymbol{\pi}$
  Update θ:
      For each (**s**, $\boldsymbol{\pi}$, ***z)***
        Back propagate

D Silver *et al. Nature* **550,** 354–359 (2017) doi:10.1038/nature24270

PURDUE UNIVERSITY

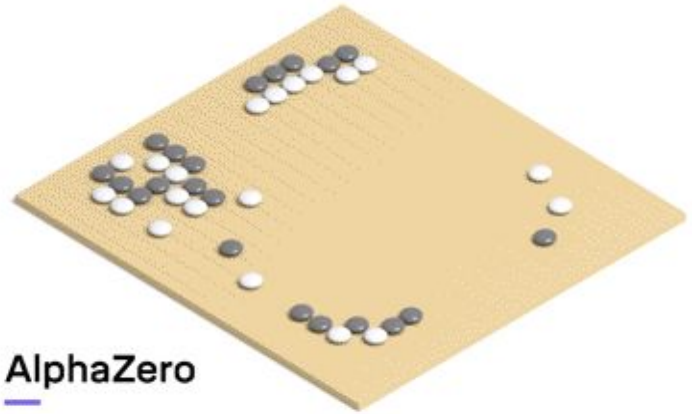
Results

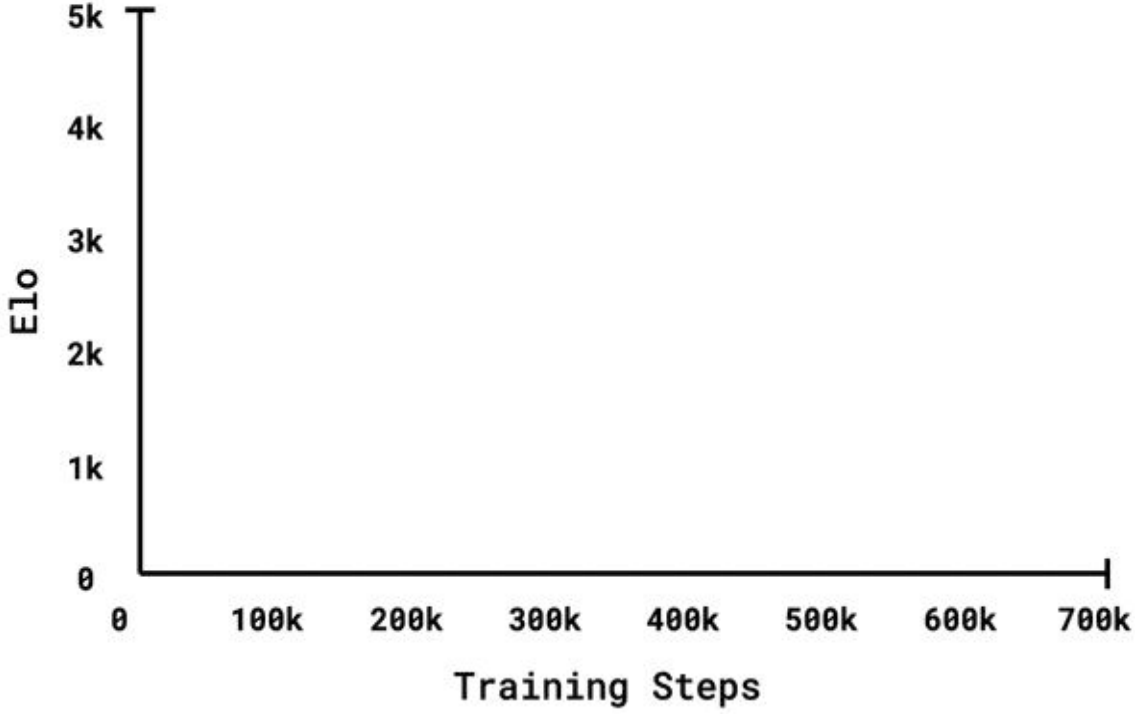


- Many game series versus previous state-of-the-art
- Convincingly beat Stockfish, Elmo
- Slight improvement over AG0

D Silver *et al. Nature* **550,** 354–359 (2017) doi:10.1038/nature24270

PURDUE UNIVERSITY





https://deepmind.com/blog/alphazero-shedding-new-light-grand-games-chess-shogi-and-go/

Concluding Remarks

Public Criticism

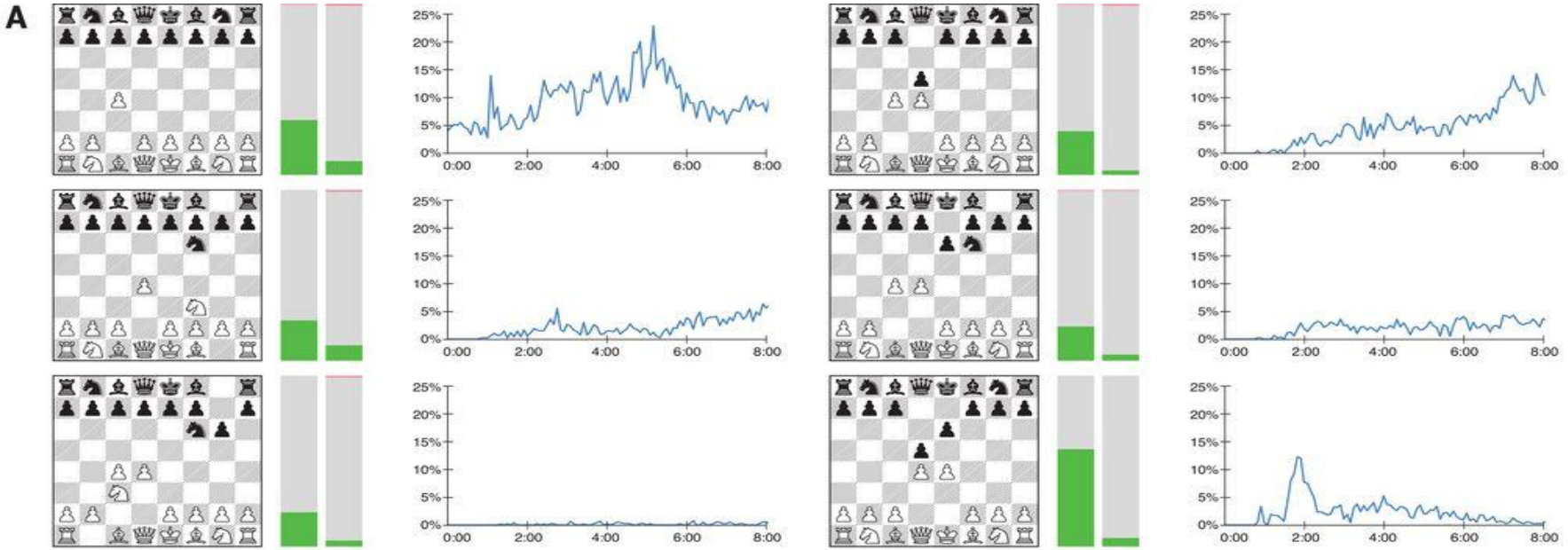
- Unfair advantages over competition
- Hard to replicate

Future Work

- Hybrid approach
- AlphaStar (Starcraft 2 AI)

PURDUE UNIVERSITY

Thank You



Bonus Slides

D. Silver et al. Science 07 Dec 2018: Vol. 362, Issue 6419, pp. 1140-1144 DOI: 10.1126/science.aar6404