LEARNING VIA SELF-PLAY: AN ALPHAGO/ALPHAGO ZERO STORY

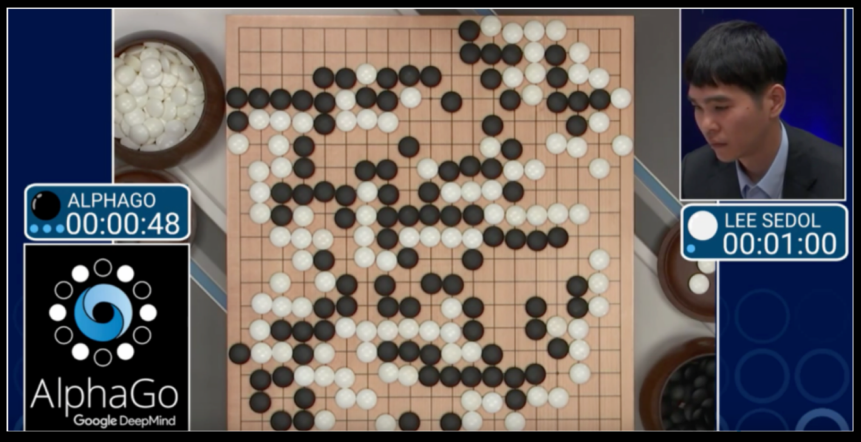
John Tan Chong Min
21 Jun 2022

DeepLearning.AI

A meetup hosted by our community

PIE & AI

ALPHAGO



EXPLORE-EXPLOIT

How to choose an optimal action?

DeepLearning.AI

A meetup hosted by our community

PIE & AI

HOW TO DECIDE?
CHICKEN RICE VS CHICKEN CHOP

VS

EXPLOIT-EXPLORE

VS

We want to keep exploiting a good choice to reap its benefits,
But we must also ensure we explore other options sufficiently in case they are better.

MONTE CARLO

Random Simulations

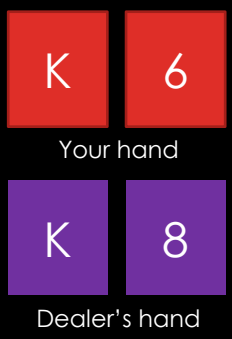
DeepLearning.AI

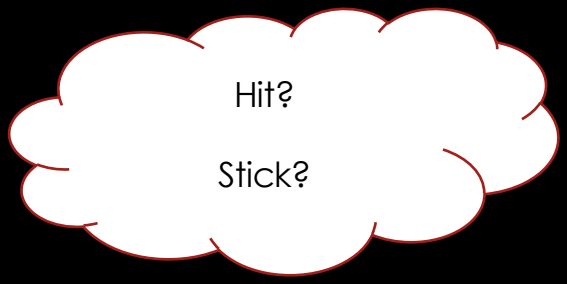
A meetup hosted by our community

PIE & AI

MONTE CARLO

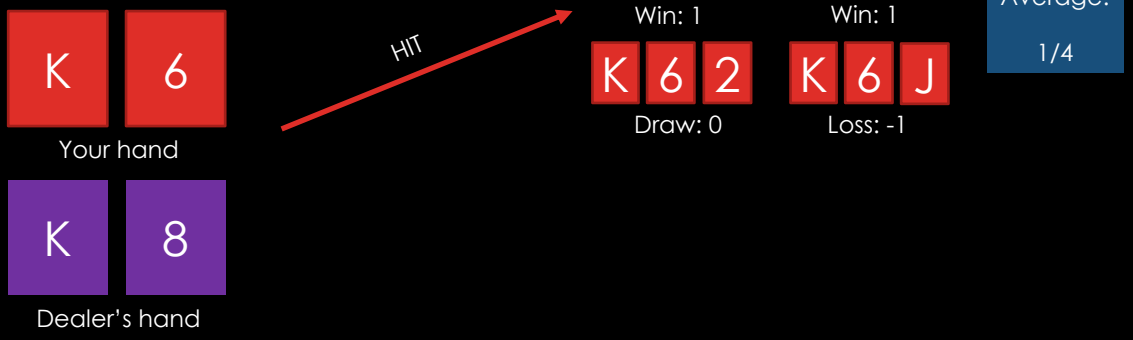
- Random simulations to derive a value

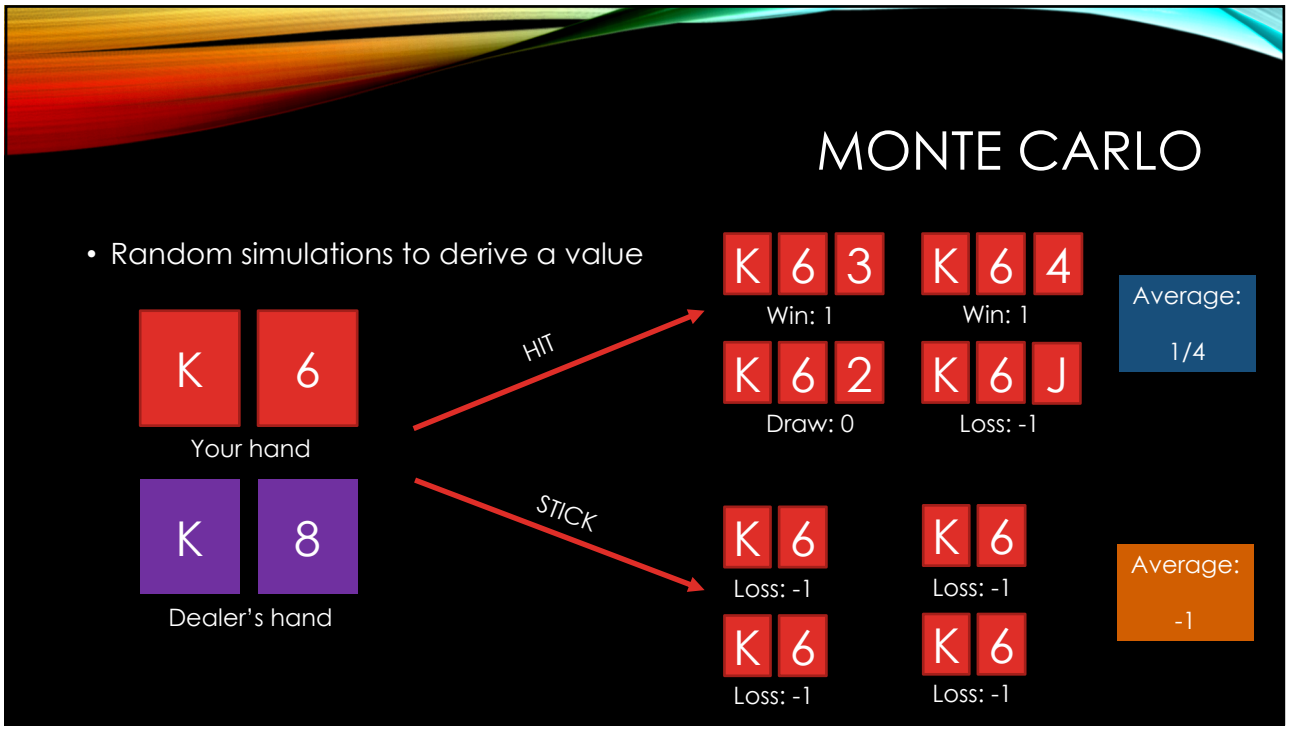


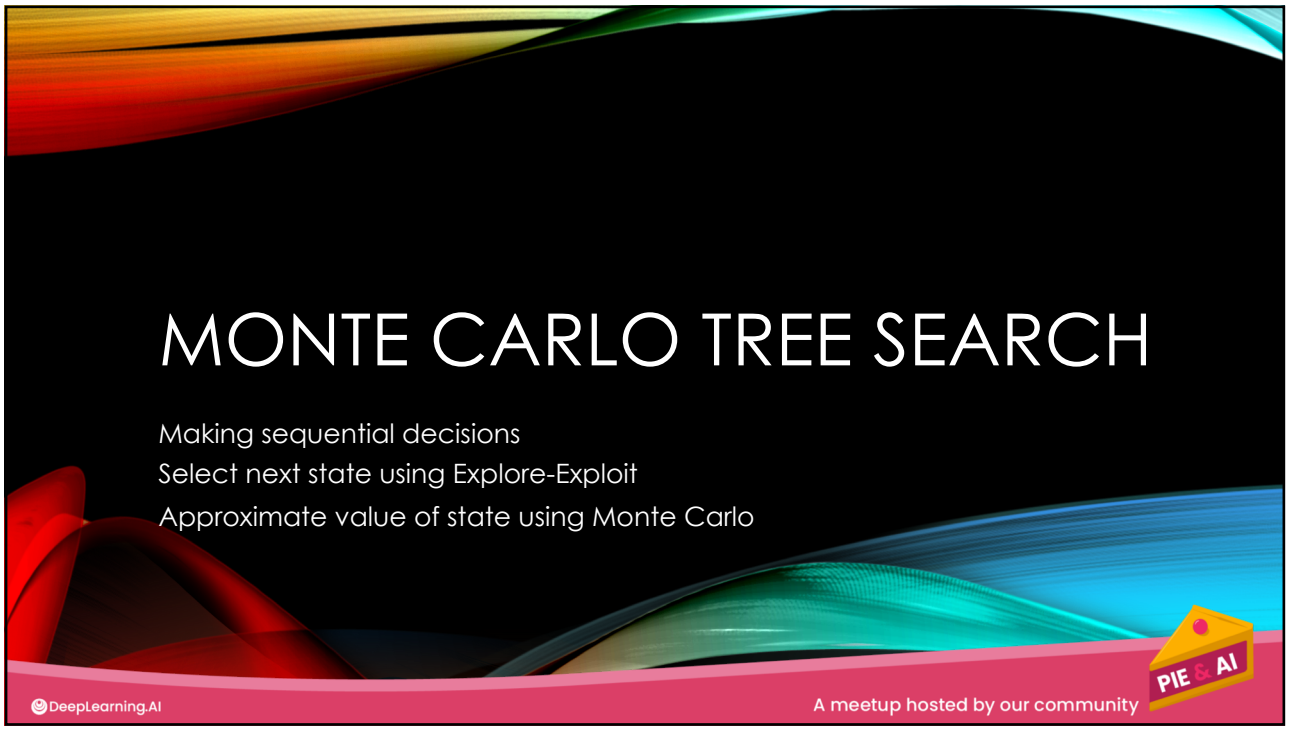


MONTE CARLO

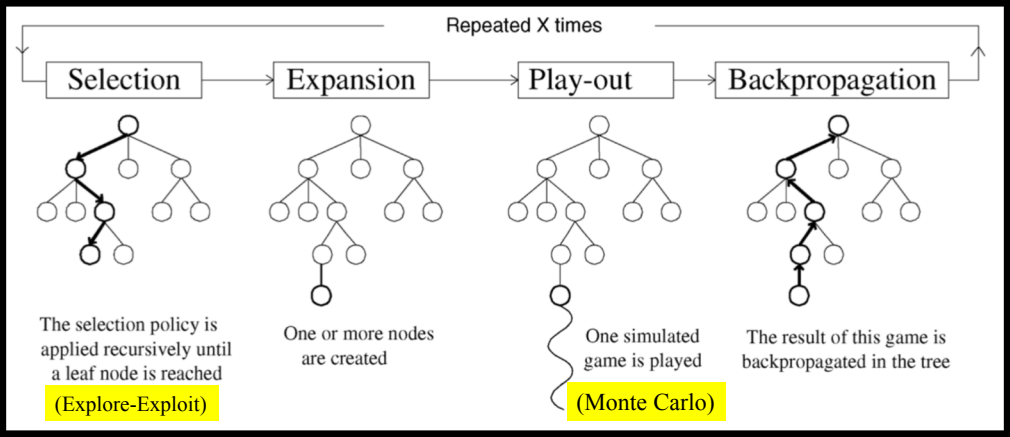
- Random simulations to derive a value




MONTE CARLO
• Random simulations to derive a value
K 6
Your hand
K 8
Dealer's hand
HIT
STICK
K 6 3
Win: 1
K 6 4
Win: 1
K 6 2
Draw: 0
K 6 J
Loss: -1
K 6
Loss: -1
K 6
Loss: -1
K 6
Loss: -1
K 6
Loss: -1
Average: 1/4
Average: -1


MONTE CARLO TREE SEARCH
Making sequential decisions
Select next state using Explore-Exploit
Approximate value of state using Monte Carlo
DeepLearning.AI
A meetup hosted by our community
PIE & AI

MCTS PROCEDURE



LIMITATIONS

- Monte Carlo rollouts may not be an accurate gauge of state value
  - Long exact sequence to victory/loss may not be found
- Large amounts of simulations needed for larger state spaces
- How to solve?
  - Functional approximation via Neural Networks!

ALPHAGO

Using both Neural Networks and MCTS

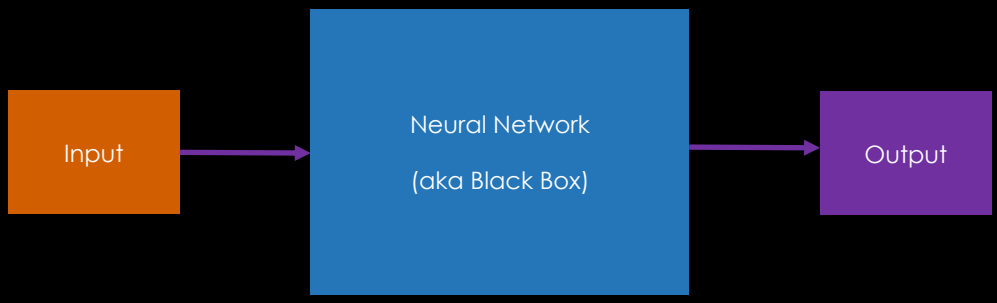
DeepLearning.AI

A meetup hosted by our community

PIE & AI

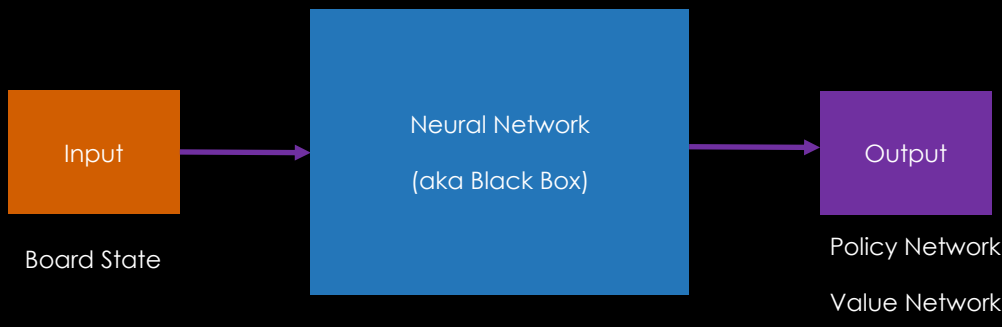
NEURAL NETWORK

- Black box approximation



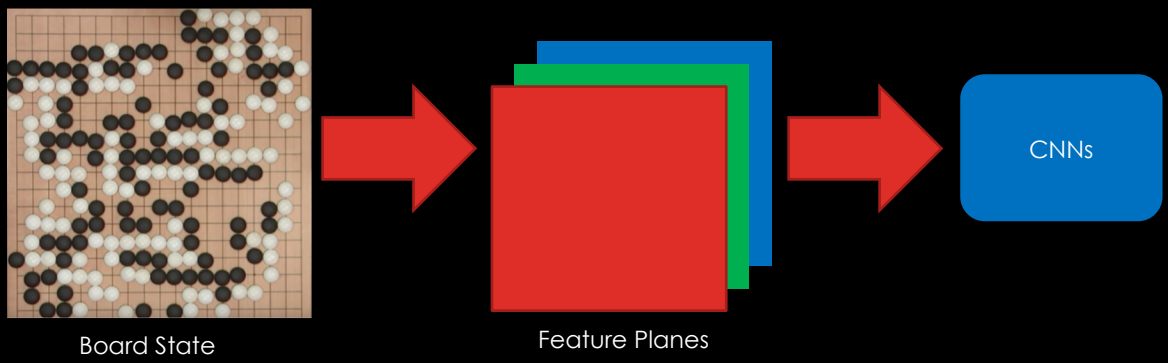
NEURAL NETWORK

- Black box approximation



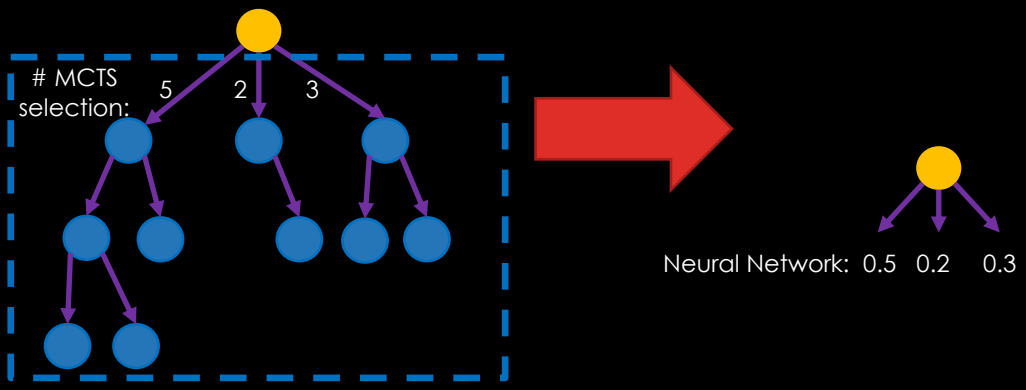
WHAT NEURAL NETWORK?

- Input is a grid
- Can use grid-like networks (CNN: Convolutional Neural Networks) to process



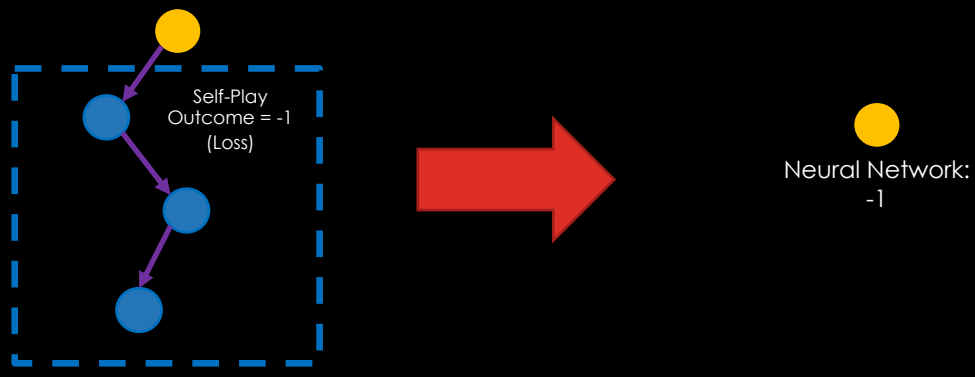
POLICY NETWORK (BREADTH)

- Focus on moves with higher win rate!



VALUE NETWORK (DEPTH)

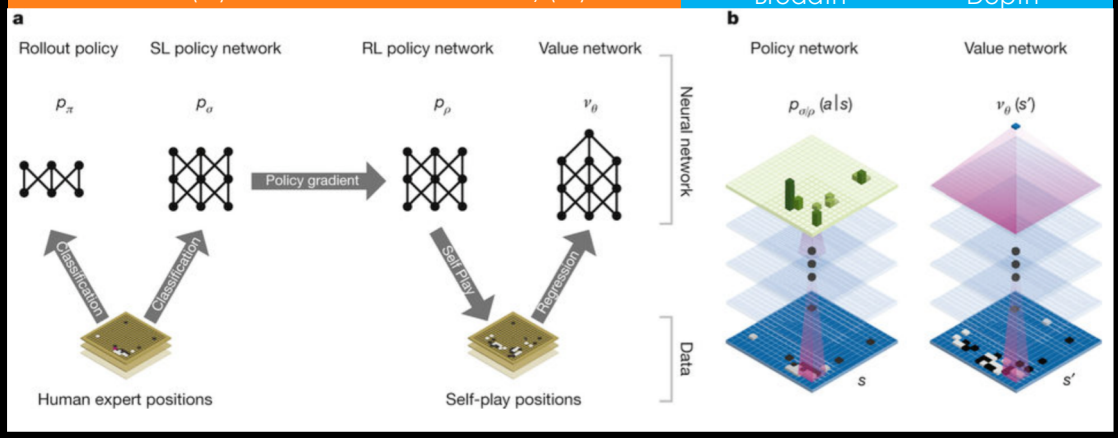
- Look ahead without simulation!
- Even more accurate than Monte Carlo rollouts



ALPHAGO: AN OVERVIEW

Find winning moves from
Humans (SL) Self-Play (RL)

Neural Network
Breadth Depth

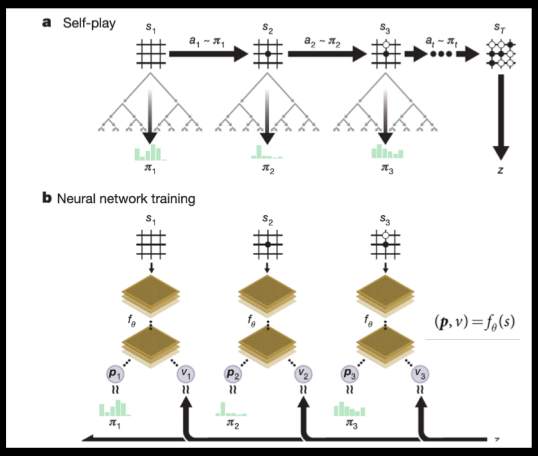


ALPHAGO ZERO

No more human expert knowledge
Learns *tabula rasa* (from a clean slate)
Just MCTS and reinforcement learning through self-play

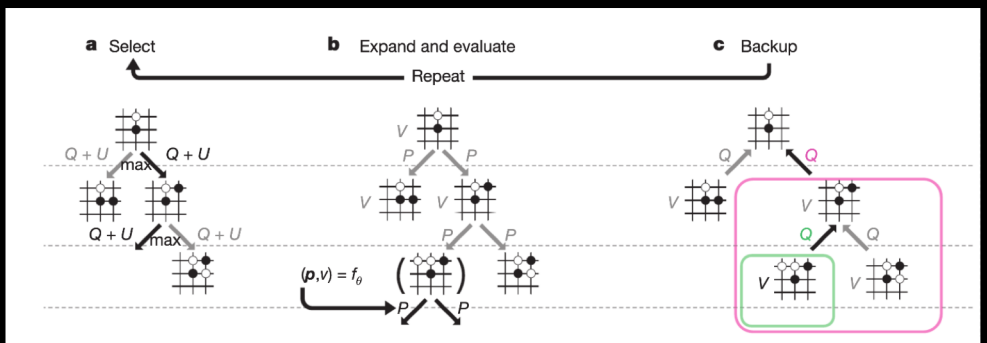
DeepLearning.AI
A meetup hosted by our community
PIE & AI

ALPHAGO ZERO: AN OVERVIEW



- Generate data using self-play games
- Train so that policy network **p** mimics the improved policy distribution π generated by MCTS
- Train so that value network output ***v*** matches the game outcome **z** for all time steps

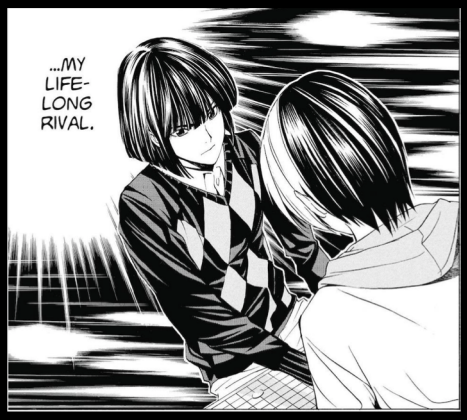
MCTS IN ALPHAGO ZERO



Exploit (Q) + Explore (U: Policy Network) for move selection	Instead of random rollouts, Value Network ***V*** gives state evaluation	State-action value (Q) updated to average of values ***V*** below it

SELF-PLAY: SPARRING AT RIGHT LEVEL

1. Initialise **"student"** and **"teacher"** networks with random weights
2. The **"student"** plays against the **"teacher"**
3. Self-play games train the **"student"** network
4. Once the **"student"** surpasses the **"teacher"**, **"teacher"** takes on **"student"** values
5. Repeat Steps 2-4 again



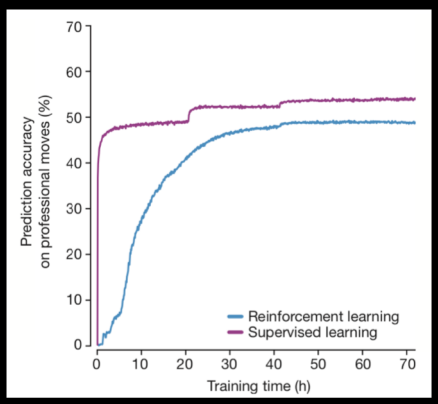
Hikaru no Go

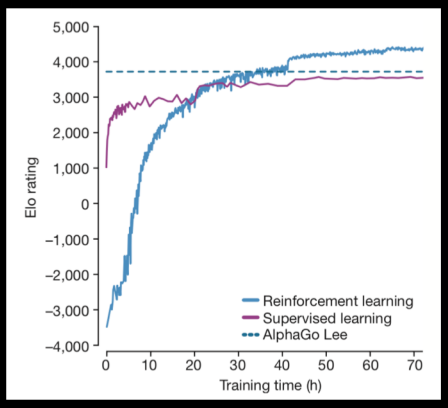
SIMPLICITY IS BETTER: HUMAN FEATURES CAN BE DISTRACTING

Feature	# of planes	Description
Stone colour	3	Player stone / opponent stone / empty
Ones	1	A constant plane filled with 1
Turns since	8	How many turns since a move was played
Liberties	8	Number of liberties (empty adjacent points)
Capture size	8	How many opponent stones would be captured
Self-atari size	8	How many of own stones would be captured
Liberties after move	8	Number of liberties after this move is played
Ladder capture	1	Whether a move at this point is a successful ladder capture
Ladder escape	1	Whether a move at this point is a successful ladder escape
Sensibleness	1	Whether a move is legal and does not fill its own eyes
Zeros	1	A constant plane filled with 0
Player color	1	Whether current player is black

AlphaGo Zero only uses a subset of features of AlphaGo

ALPHAGO ZERO PERFORMANCE





- Poorer prediction of professional moves, but higher playing performance!

HOW TO ACHIEVE SUPERHUMAN PERFORMANCE?

1. Clear objective
2. Basic rules of the environment
3. A continual way to self-improve with respect to the objective

THANK YOU

Presented by: John Tan Chong Min

Blog: https://delvingintotech.wordpress.com/

Twitch: https://www.twitch.tv/johncm99

DeepLearning.AI

A meetup hosted by our community

PIE & AI