A **Deep** Journey of Playing Games with RL

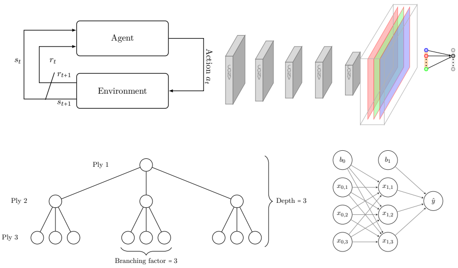
NSE Seminar

Kim Hammar
kimham@kth.se

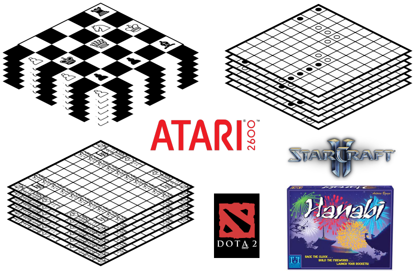
January 31, 2020

Why Games

AI & Machine Learning



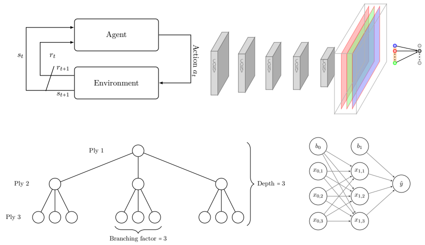
Games



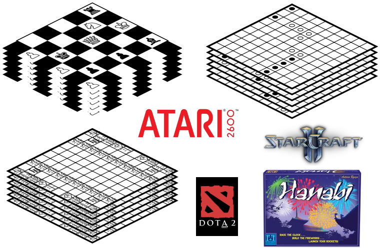
Why Combine the two?

Why Games

AI & Machine Learning



Games



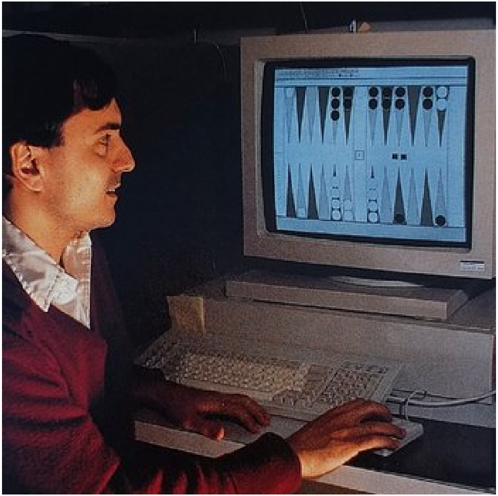
Why Combine the two?

- AI & Games have a long history (Turing '50& Minsky 60')
- Simple to evaluate, reproducible, controllable, quick feedback loop
- Common benchmark for the research community

1997: DeepBlue[1] vs Kasparov

2em1[1] Murray Campbell, A. Joseph Hoane, and Feng-hsiung Hsu. "Deep Blue". In: *Artif. Intell.* 134.1–2 (Jan. 2002), 57–83. ISSN: 0004-3702. DOI: 10.1016/S0004-3702(01)00129-1. URL: https://doi.org/10.1016/S0004-3702(01)00129-1.

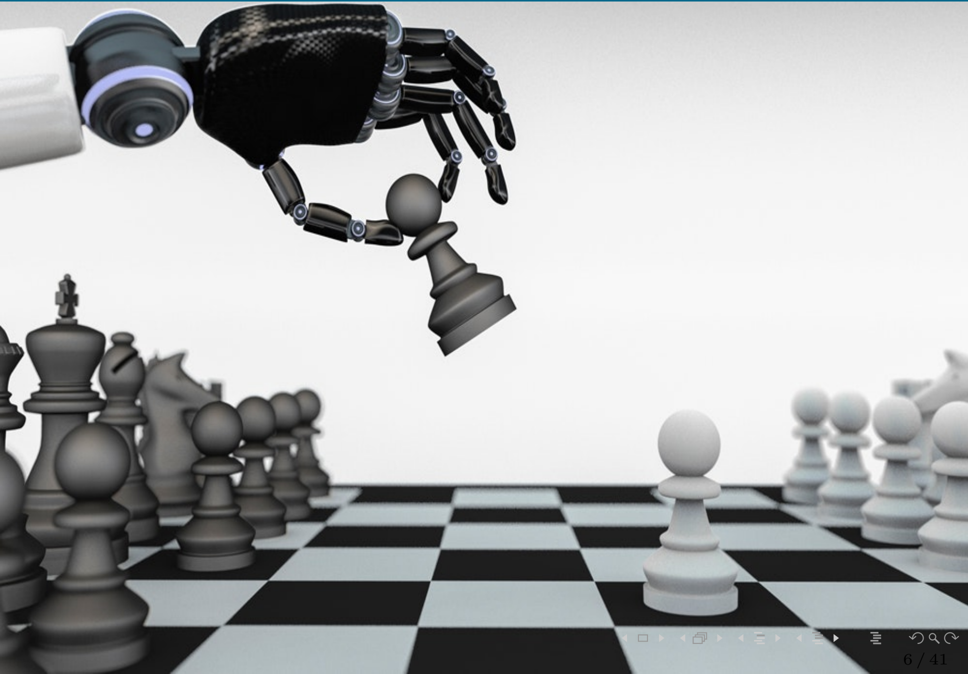
1992: Tesauro's TD-Gammon[2]



2em1[2] Gerald Tesauro. "TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play". In: *Neural Comput.* 6.2 (Mar. 1994), 215–219. ISSN: 0899-7667. DOI: 10.1162/neco.1994.6.2.215. URL: https://doi.org/10.1162/neco.1994.6.2.215.

1959: Arthur Samuel's Checkers Player[3]

2em1[3] A. L. Samuel. "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM J. Res. Dev.* 3.3 (July 1959), 210–229. ISSN: 0018-8646. DOI: 10.1147/rd.33.0210. URL: https://doi.org/10.1147/rd.33.0210, A. L. Samuel. "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM J. Res. Dev.* 3.3 (July 1959), 210–229. ISSN: 0018-8646. DOI: 10.1147/rd.33.0210. URL: https://doi.org/10.1147/rd.33.0210.



PAPERS IN FOCUS TODAY

- AlphaGo[4]
- AlphaGo Zero[5]
- AlphaZero[6]

Google DeepMind

AlphaGo	**AlphaGo Zero**	**Alpha Zero**
Nature, 6.5k citations	Nature, 2.5k citations	Science, 400 citations
2016	2017	2018

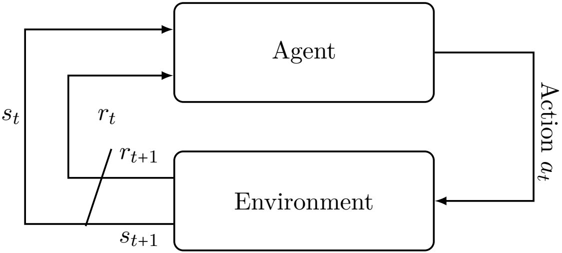
2em1[4] David Silver et al. "Mastering the Game of Go with Deep Neural Networks and Tree Search". In: *Nature* 529.7587 (Jan. 2016), pp. 484–489. DOI: 10.1038/nature16961.

2em1[5] David Silver et al. "Mastering the game of Go without human knowledge". In: *Nature* 550 (Oct. 2017), pp. 354–. URL: http://dx.doi.org/10.1038/nature24270.

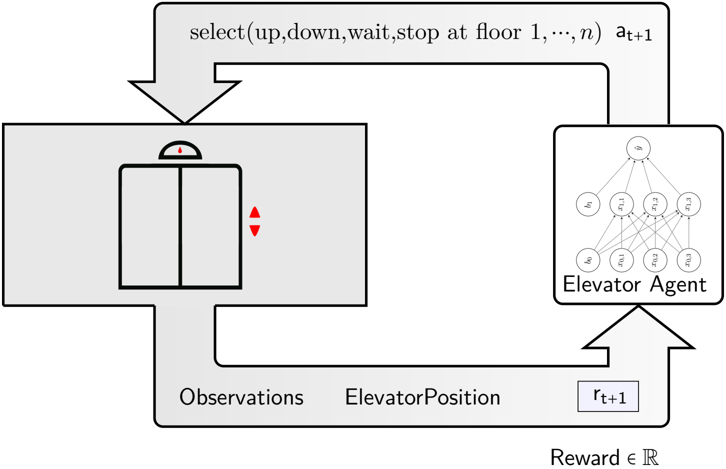
2em1[6] David Silver et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419 (2018), pp. 1140–1144. URL: http://science.sciencemag.org/content/362/6419/1140/tab-pdf.

THE REINFORCEMENT LEARNING PROBLEM

- Notation; **policy**: π, **state**: s, **reward**: r, **action**: a
- Agent's goal: maximize reward, $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \quad 0 \leq \gamma \leq 1$
- RL's goal, find optimal policy $\pi^* = \max_\pi \mathbb{E}[R|\pi]$

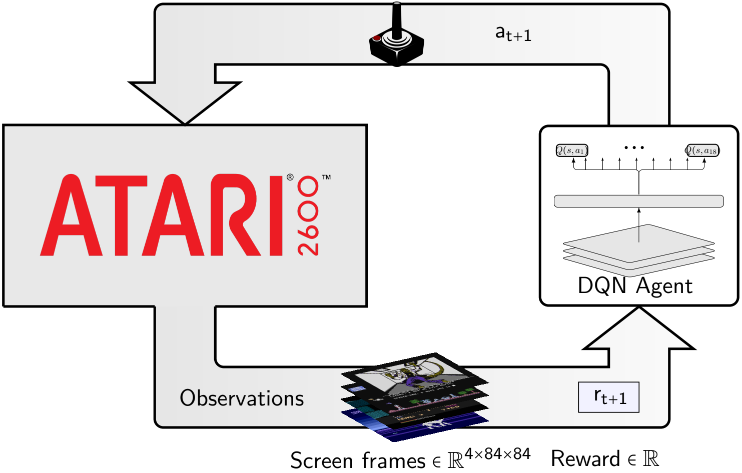


RL Examples: Elevator (Crites & Barto '95[7])



2em1[7] Robert H. Crites and Andrew G. Barto. "Improving Elevator Performance Using Reinforcement Learning". In: *Proceedings of the 8th International Conference on Neural Information Processing Systems*. NIPS'95. Denver, Colorado: MIT Press, 1995, 1017–1023.

RL Examples: Atari (Mnih ’15)[8]



2em1[8] Volodymyr Mnih et al. “Human-level control through deep reinforcement learning”. In: *Nature* 518.7540 (Feb. 2015), pp. 529–533. ISSN: 00280836. URL: `http://dx.doi.org/10.1038/nature14236`.

How to Act Optimally? (Bellman 57'[9])

$$optimal(s_t) = \max_{\pi} \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} \middle| s_t\right]$$

2em1[9] Richard Bellman. *Dynamic Programming.* Dover Publications, 1957. isbn: 9780486428093.

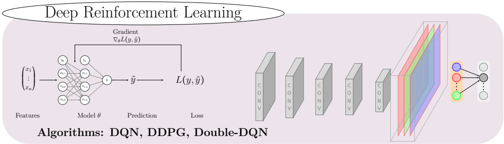
How to Act Optimally? (Bellman 57'[10])

$$
\begin{aligned}
optimal(s_t) &= \max_{\pi} \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} \middle| s_t\right] \\
&= \max_{\pi} \mathbb{E}\left[r_{t+1} \sum_{k=2}^{\infty} \gamma^{k-1} r_{t+k} \middle| s_t\right] \\
&= \max_{a_t} \mathbb{E}\left[r_{t+1} + \max_{\pi} \mathbb{E}\left[\sum_{k=2}^{\infty} \gamma^{k-1} r_{t+k} \middle| s_{t+1}\right] \middle| s_t\right] \\
&= \max_{a_t} \mathbb{E}\left[r_{t+1} + \gamma \max_{\pi} \mathbb{E}\left[\sum_{k=2}^{\infty} \gamma^{k-2} r_{t+k} \middle| s_{t+1}\right] \middle| s_t\right] \\
&= \max_{a_t} \mathbb{E}\left[r_{t+1} + \gamma \max_{\pi} \mathbb{E}\left[\sum_{k=2}^{\infty} \gamma^{k-2} r_{t+k} \middle| s_{t+1}\right] \middle| s_t\right] \\
&= \max_{a_t} \mathbb{E}\left[r_{t+1} + \gamma optimal(s_{t+1}) | s_t\right]
\end{aligned}
$$

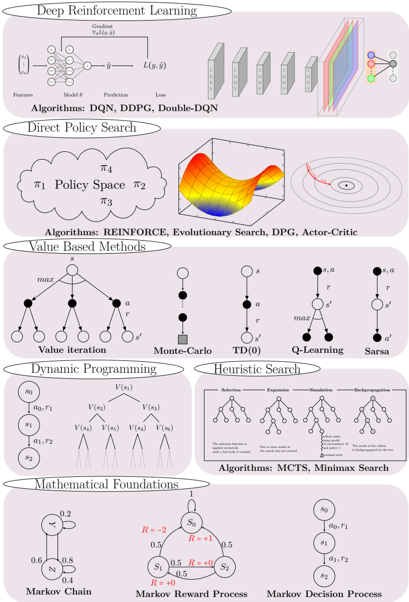
2em1[10] Richard Bellman. *Dynamic Programming.* Dover Publications, 1957. ISBN:

Reinforcement Learning: An Overview

Google DeepMind



Reinforcement Learning: An Overview



AlphaGo ’2016[11]

ARTICLE

doi:10.1038/nature16961

Mastering the game of Go with deep neural networks and tree search

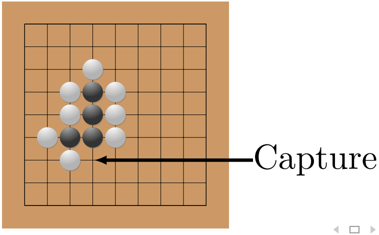
David Silver[1]*, Aja Huang[1]*, Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

2em1[11] David Silver et al. “Mastering the Game of Go with Deep Neural Networks and Tree Search”. In: *Nature* 529.7587 (Jan. 2016), pp. 484–489. DOI: 10.1038/nature16961.

The Mystery of Go, the Ancient Game That Computers Still Can't Win

The game of Go

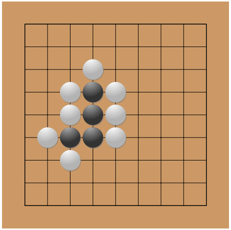
- The world's oldest game: 3000 years old, over 40M players world wide
- To win: **capture** the most territory on the board
  - Surrounded stones/areas are captured and removed
- Why is it so hard for computers? 10^{170} **unique states!!**, ≈ 250 **branching factor**
  - High branching factor, large board (19×19), hard to evaluate etc..



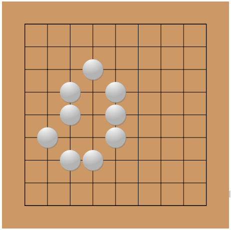
The game of Go

- The world's oldest game: 3000 years old, over 40M players world wide
- To win: **capture** the most territory on the board
  - Surrounded stones are captured and removed
- Why is it so hard for computers? 10^{170} **unique states!!** ≈ 250 **branching factor**
  - High branching factor, large board (19×19), hard to evaluate etc..

(1)

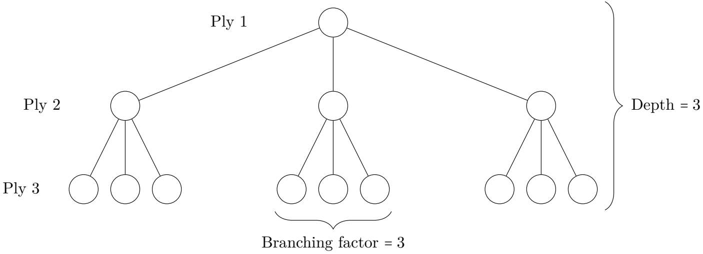


(2)

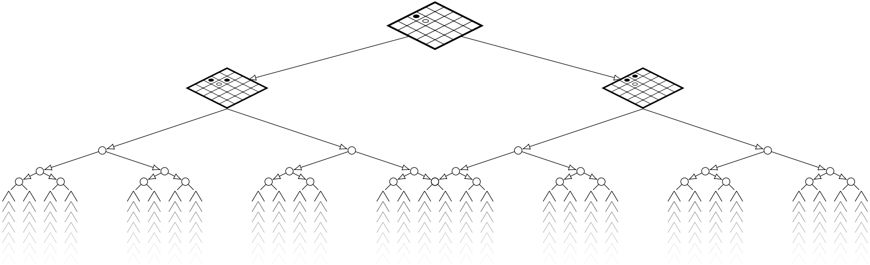


Game Trees

- How do you program a computer to play a board game?
- Simplest approach:
  - (1) Program a game tree; (2) Assume opponent think like you; (3) Look-ahead and evaluate each move
  - Requires Knowledge of game rules and evaluation function

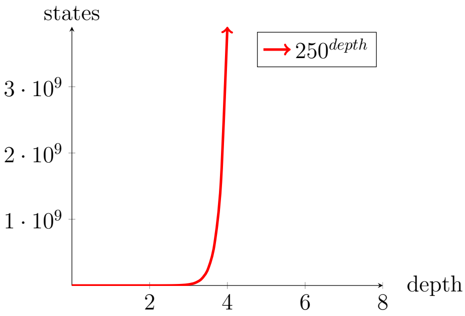


Search + Go = 💣

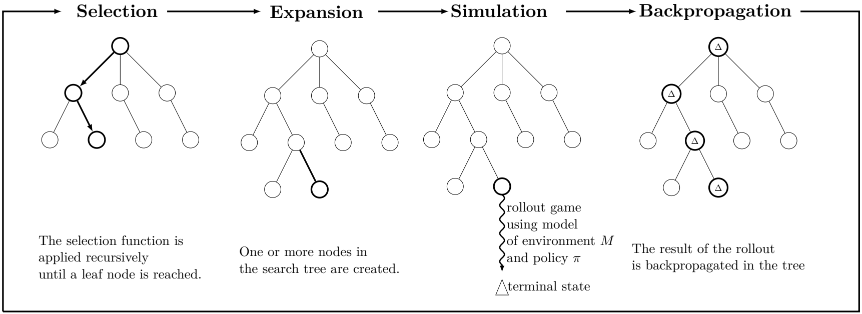


Some Numbers

- **Atoms in the universe**
  - $\approx 10^{80}$
- **States**
  - Go: 10^{170}, Chess: 10^{47}
- **Game tree complexity**
  - Go: 10^{360}, Chess: 10^{123}
- **Average branching factor**
  - Go: 250, Chess: 35
- **Board size (positions)**
  - Go: 361, Chess: 64



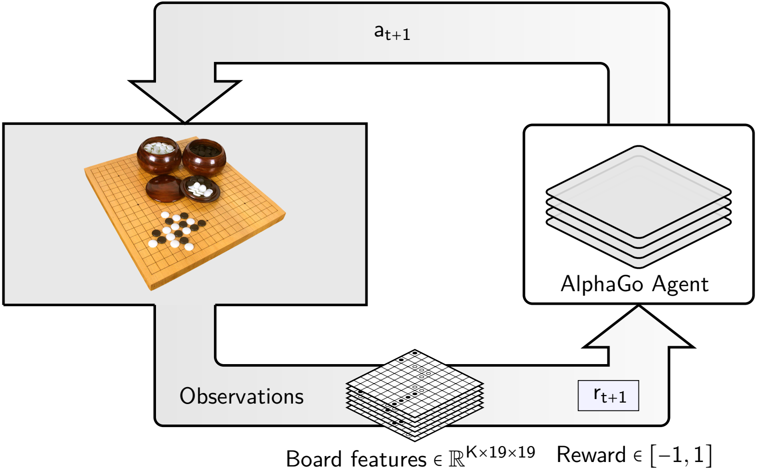
Monte-Carlo Tree Search



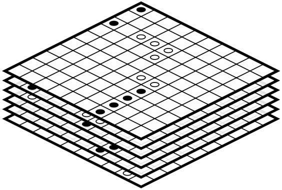
AlphaGo's Approach

- Brute-Force Search does not work
  - At least not until hardware has improved **a lot**.
- Human Go professionals rely on small search guided by intuition/experience
- **AlphaGo's Approach:** Complement MCTS with **"artifical intuition"**
  - Artificial intuition provided by two neural networks: value network and policy network

Computer Go as an RL Problem



AlphaGo Training Pipeline (1/2)

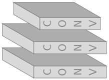


Human Expert Moves D_1

AlphaGo Training Pipeline (1/2)

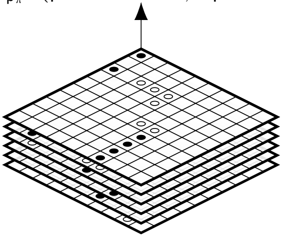
Supervised Rollout Policy Network

$p_\pi(a|s)$



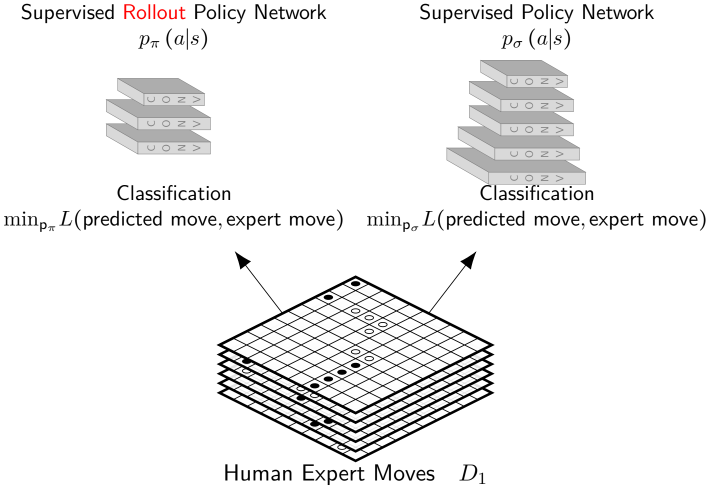
Classification

$\min_{p_\pi} L(\text{predicted move}, \text{expert move})$



Human Expert Moves D_1

AlphaGo Training Pipeline (1/2)

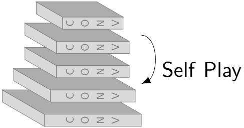


AlphaGo Training Pipeline (2/2)

Reinforcement Learning Policy Network

$p_\rho(a|s)$

Initialize with p_σ weights



PolicyGradient

$$\mathsf{J}(\mathsf{p}_\rho) = \mathbb{E}_{\mathsf{p}_\rho}\left[\sum_{\mathsf{t}=0}^{\infty} \mathsf{r}_\mathsf{t}\right]$$

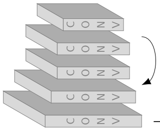
$$\rho \leftarrow \rho + \alpha \nabla_\rho \mathsf{J}(\mathsf{p}_\rho)$$

AlphaGo Training Pipeline (2/2)

Reinforcement Learning Policy Network

$p_\rho(a|s)$

Initialize with p_σ weights

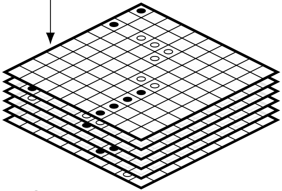


Self Play

PolicyGradient

$J(p_\rho) = \mathbb{E}_{p_\rho}\left[\sum_{t=0}^{\infty} r_t\right]$

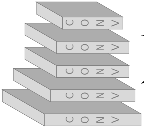
$\rho \leftarrow \rho + \alpha \nabla_\rho J(p_\rho)$



Self Play Dataset D_2

AlphaGo Training Pipeline (2/2)

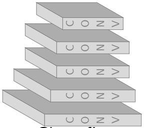
Reinforcement Learning Policy Network
$p_\rho(a|s)$
Initialize with p_σ weights



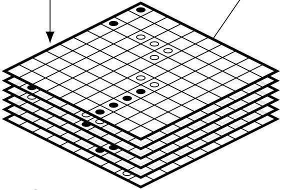
Self Play

PolicyGradient
$J(p_\rho) = \mathbb{E}_{p_\rho}\left[\sum_{t=0}^{\infty} r_t\right]$
$\rho \leftarrow \rho + \alpha \nabla_\rho J(p_\rho)$

Supervised Value Network
$v_\theta(s')$

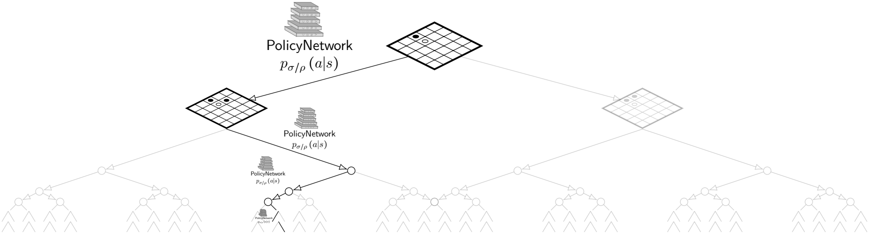


Classification
$\min_{v_\theta} L(\text{predicted outcome, actual outcome})$

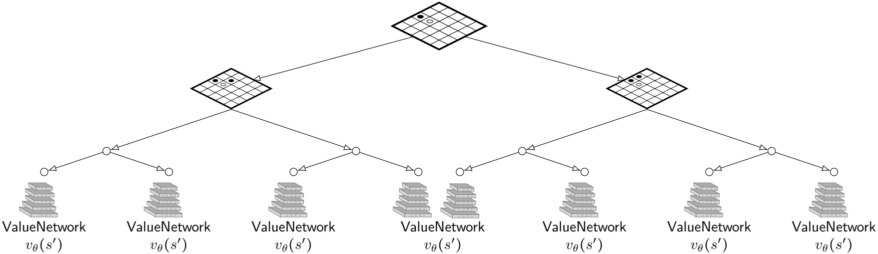


Self Play Dataset D_2

Guided Search Search Using the Policy Network

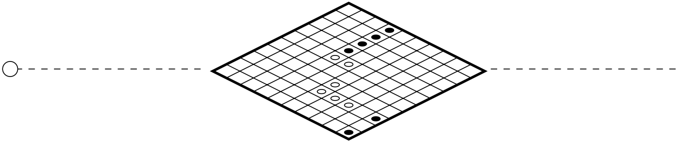


Depth-Limited Search Using the Value Network



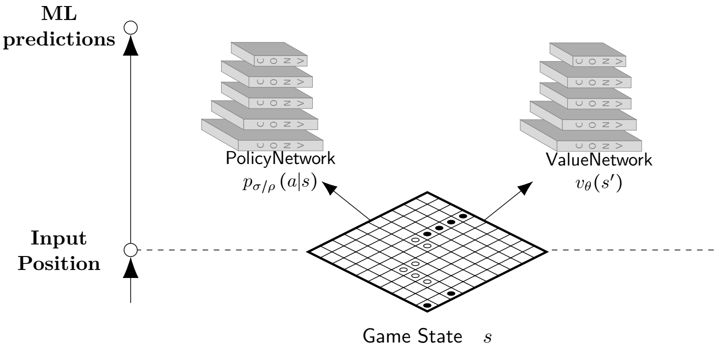
AlphaGo Prediction Pipeline

Input Position

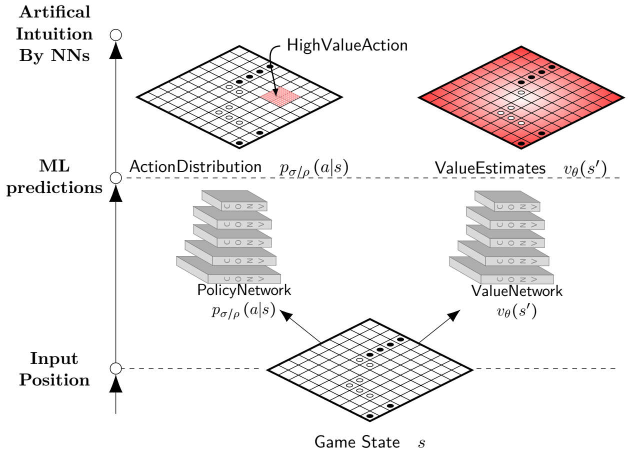


Game State s

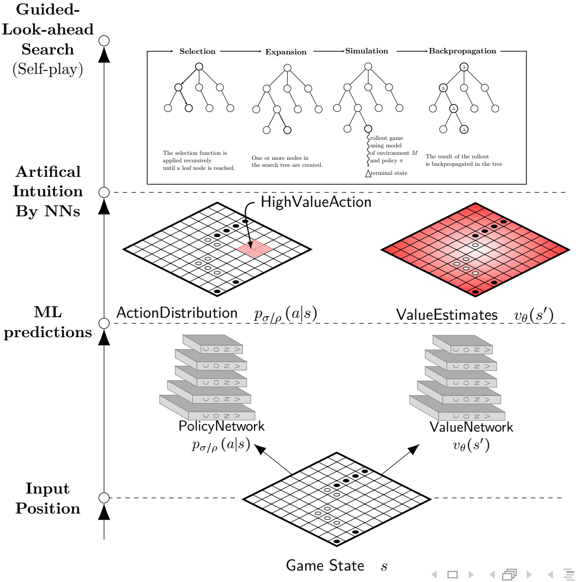
AlphaGo Prediction Pipeline



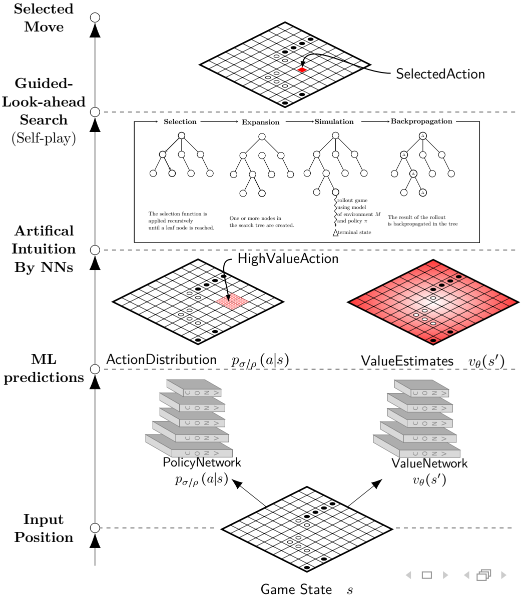
AlphaGo Prediction Pipeline



AlphaGo Prediction Pipeline



AlphaGo Prediction Pipeline



AlphaGo Vs Lee Sedol

- In March 2016, Alpha Go won against Lee Sedol 4-1
- Lee Sedol was 18-time World Champion prior to the game
- Two famous moves: Move 37 by AlphaGo and Move 78 by Sedol

AlphaGo: Key Takeaways

1. Different AI techniques can be complementary
   - Supervised learning
   - Reinforcement learning
   - Search
   - Rules/Domain Knowledge
   - **What ever it takes to win!!** ☺
2. Self-play
3. Vast computation still required for training and inference
   - AlphaGo used 1200 CPUs and 176 GPUs

AlphaGo Zero ’2017[12]

nature

Article | Published: 19 October 2017

Mastering the game of Go without human knowledge

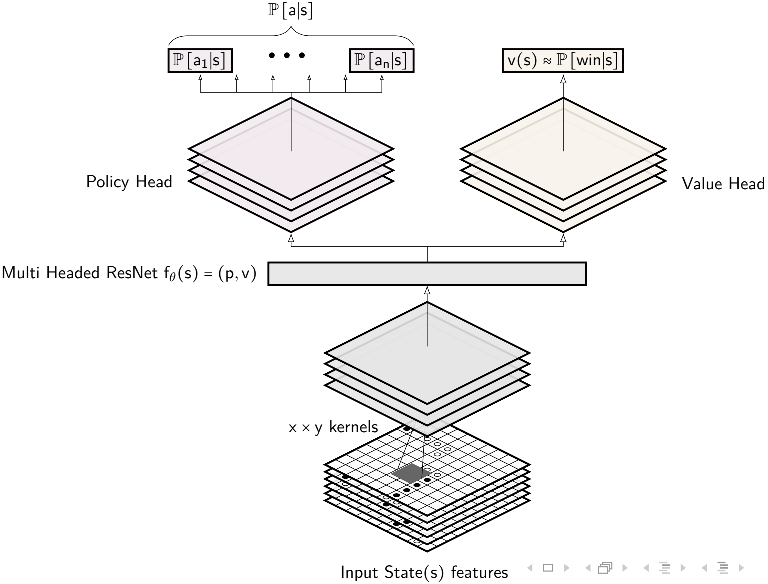
David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel & Demis Hassabis

2em1[12] David Silver et al. “Mastering the game of Go without human knowledge”. In: *Nature* 550 (Oct. 2017), pp. 354–. URL: http://dx.doi.org/10.1038/nature24270.

AlphaGo Zero

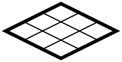
- AlphaGo Zero is a successor to AlphaGo
- AlphaGo Zero is **simpler** and **stronger** than AlphaGo
  - AlphaGo Zero beats AlphaGo 100 – 0 in matches
- AlphaGo Zero starts from **Zero** domain knowledge
  - Uses a single neural network (compared to 4 NNs in AlphaGo)
  - Learns by Self-Play only (No supervised learning like in AlphaGo)

AlphaGo Zero Neural Network Architecture

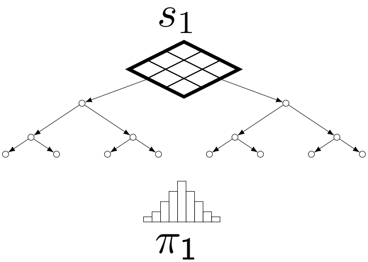


AlphaGo Zero Self-Play Training Algorithm

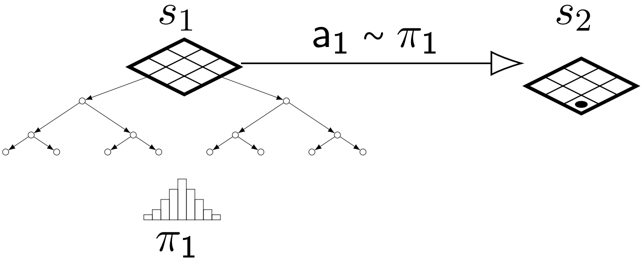
s_1



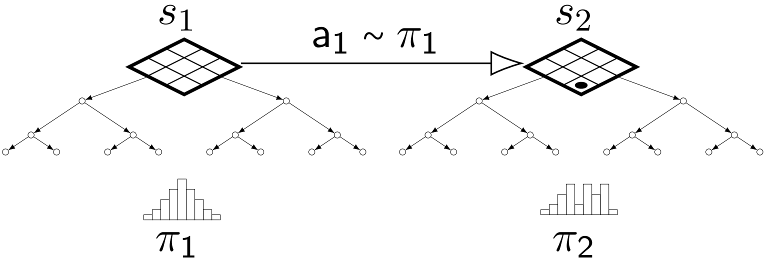
AlphaGo Zero Self-Play Training Algorithm



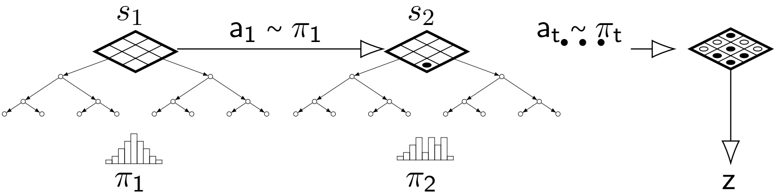
AlphaGo Zero Self-Play Training Algorithm



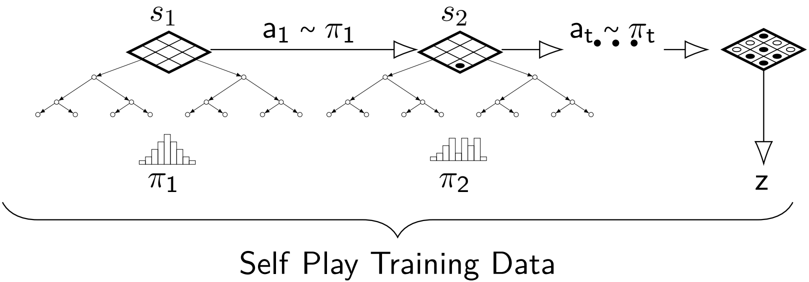
AlphaGo Zero Self-Play Training Algorithm



AlphaGo Zero Self-Play Training Algorithm

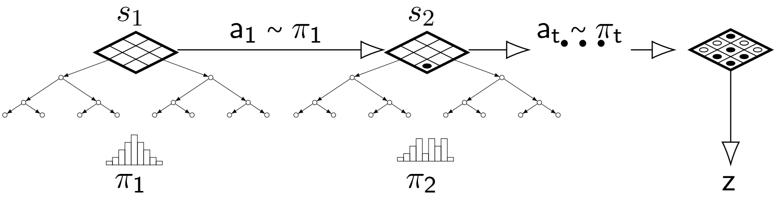


AlphaGo Zero Self-Play Training Algorithm



$f_\theta(s_t) = (p_t, v_t)$ $\theta' = \theta - \alpha \nabla_\theta L((p_t, v_t), (\pi_t, z))$

AlphaGo Zero Self-Play Training Algorithm

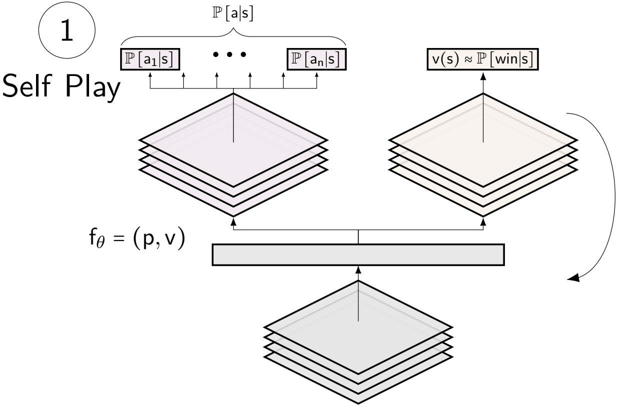


Self Play Training Data

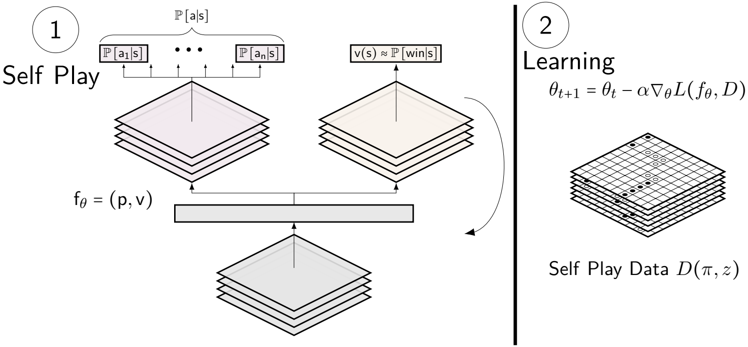
$$f_\theta(s_t) = (p_t, v_t) \qquad \theta' = \theta - \alpha \nabla_\theta L((p_t, v_t), (\pi_t, z))$$

$$L(f_\theta(s_t), (\pi_t, z)) = \underbrace{(z - v_t)^2}_{\text{MSE}} - \underbrace{\pi_t^T \log p_t + c\|\theta\|^2}_{\text{Cross-entropy loss}}$$

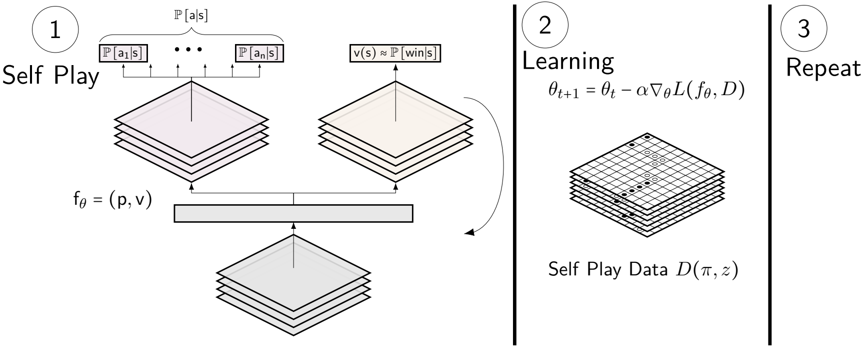
AlphaGo Zero Self-Play Training Pipeline



AlphaGo Zero Self-Play Training Pipeline



AlphaGo Zero Self-Play Training Pipeline



AlphaGo Zero: Key Takeaways

1. A **simpler** system can be more powerful than a complex one (AlphaGo Zero vs AlphaGo)

2. Neural networks can be combined like LEGO blocks

3. **ResNet** is better than traditional ConvNets

4. **Self-play**

AlphaZero ’2018[13]

RESEARCH

COMPUTER SCIENCE

A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play

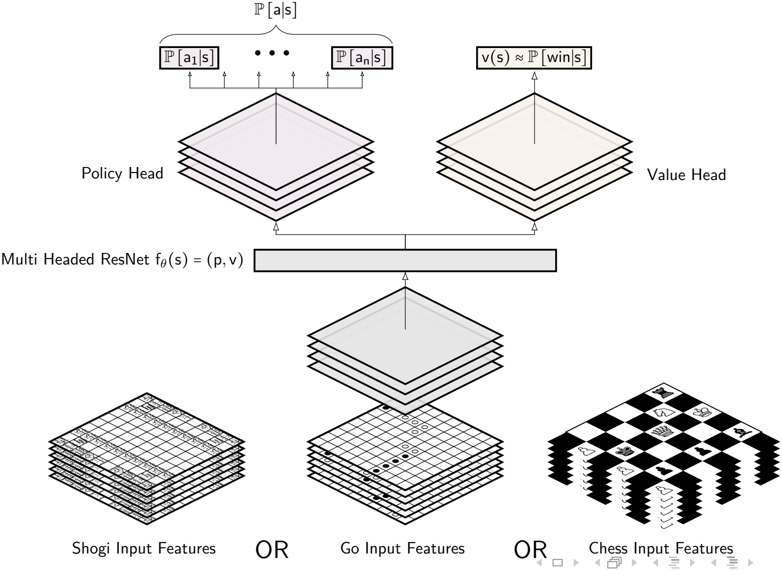
David Silver[1,2*†], Thomas Hubert[1*], Julian Schrittwieser[1*], Ioannis Antonoglou[1], Matthew Lai[1], Arthur Guez[1], Marc Lanctot[1], Laurent Sifre[1], Dharshan Kumaran[1], Thore Graepel[1], Timothy Lillicrap[1], Karen Simonyan[1], Demis Hassabis[1†]

2em1[13] David Silver et al. “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play”. In: *Science* 362.6419 (2018), pp. 1140–1144. URL: `http://science.sciencemag.org/content/362/6419/1140/tab-pdf`.

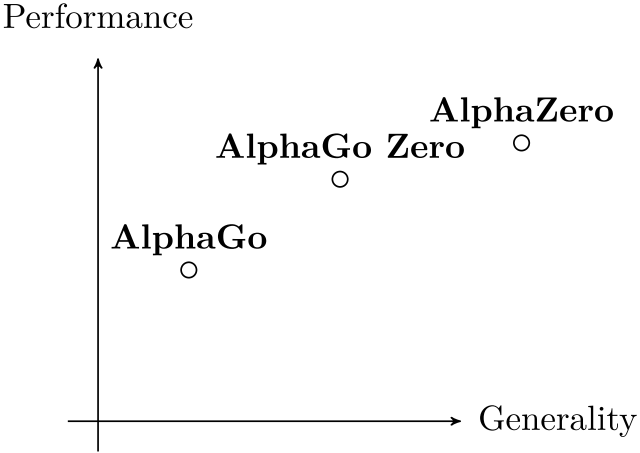
AlphaZero

- AlphaGo Zero is able to reach superhuman level at Go without any domain knowledge...
- As AlphaGo Zero is not dependent on Go, **can the same algorithm play other games?**
- AlphaZero extends AlphaGo to play not only Go but also Chess and Shogi
    - The same algorithm achieves superhuman performance on all three games

AlphaZero


$\mathbb{P}[a|s]$
$\mathbb{P}[a_1|s]$
$\mathbb{P}[a_n|s]$
$v(s) \approx \mathbb{P}[win|s]$
Policy Head
Value Head
Multi Headed ResNet $f_\theta(s) = (p, v)$
Shogi Input Features
OR
Go Input Features
OR
Chess Input Features

AlphaZero is much Simpler than AlphaGo, Yet More Powerful



AlphaZero: Key Takeaways

1. Being able to play three games at a superhuman level, AlphaZero is one step closer to **general** AI
2. Massive compute power still required for training AlphaZero
   - 5000 TPUs
3. Self-play

Next Challenge..

Presentation Summary

- Sometimes a simpler system can be more powerful than a complex one
- Universal research principle: **strive for generality**, simplicity, Occam's Razor
- Self-play: no human bias, learn from first principles
- **Deep RL is still in its infancy**, a lot to more to be expected in the next few years
- Open challenges: Sample efficiency, data efficiency
  - Yes, AlphaGo can learn to play Go after hundreds of game years, but a human can reach a decent level of play in only a couple of hours
  - How can we make reinforcement learning more efficient? Model-based learning is a research area with increasing attention

References[19]

- DQN[14]
- AlphaGo[15]
- AlphaGo Zero[16]
- AlphaZero[17]
- AlphaStar[18]

2em1[14] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". In: *Nature* 518.7540 (Feb. 2015), pp. 529–533. ISSN: 00280836. URL: http://dx.doi.org/10.1038/nature14236.

2em1[15] David Silver et al. "Mastering the game of Go without human knowledge". In: *Nature* 550 (Oct. 2017), pp. 354–. URL: http://dx.doi.org/10.1038/nature24270.

2em1[16] David Silver et al. "Mastering the game of Go without human knowledge". In: *Nature* 550 (Oct. 2017), pp. 354–. URL: http://dx.doi.org/10.1038/nature24270.

2em1[17] David Silver et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419 (2018), pp. 1140–1144. URL: http://science.sciencemag.org/content/362/6419/1140/tab-pdf.

2em1[18] Oriol Vinyals et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning". In: *Nature* 575 (Nov. 2019). DOI: 10.1038/s41586-019-1724-z.

2em1[19] Thanks to Rolf Stadler for Reviewing and discussing drafts of this presentation