Go in numbers

3,000 Years Old

40M Players

$10^{170}$ Positions
The Rules of Go

Capture

Territory
Why is Go hard for computers to play?

Brute force search intractable:

1. Search space is huge
2. “Impossible” for computers to evaluate who is winning

Game tree complexity = $b^d$
Convolutional neural network
Value network

Evaluation

\[ v_\theta(s) \]

Position

\( \theta \)

\( s \)
Policy network

Move probabilities

Position

\[ p_\sigma(a|s) \]

\( \sigma \)

\( s \)
Exhaustive search
Monte-Carlo rollouts
Reducing depth with value network
Reducing depth with value network
Reducing breadth with policy network
Deep reinforcement learning in AlphaGo

Human expert positions ➔ Supervised Learning policy network ➔ Reinforcement Learning policy network ➔ Self-play data ➔ Value network

- Classification
- Self Play
- Regression
Supervised learning of policy networks

**Policy network:** 12 layer convolutional neural network

**Training data:** 30M positions from human expert games (KGS 5+ dan)

**Training algorithm:** maximise likelihood by stochastic gradient descent

\[ \Delta \sigma \propto \frac{\partial \log p_\sigma(a|s)}{\partial \sigma} \]

**Training time:** 4 weeks on 50 GPUs using Google Cloud

**Results:** 57% accuracy on held out test data (state-of-the art was 44%)
Reinforcement learning of policy networks

**Policy network:** 12 layer convolutional neural network

**Training data:** games of self-play between policy network

**Training algorithm:** maximise wins $z$ by policy gradient reinforcement learning

$$\Delta \sigma \propto \frac{\partial \log p_{\sigma}(a|s)}{\partial \sigma} z$$

**Training time:** 1 week on 50 GPUs using Google Cloud

**Results:** 80% vs supervised learning. Raw network $\sim$3 amateur dan.
Reinforcement learning of value networks

**Value network:** 12 layer convolutional neural network

**Training data:** 30 million games of self-play

**Training algorithm:** minimise MSE by stochastic gradient descent

\[ \Delta \theta \propto \frac{\partial v_\theta(s)}{\partial \theta} (z - v_\theta(s)) \]

**Training time:** 1 week on 50 GPUs using Google Cloud

**Results:** First strong position evaluation function - previously thought impossible
Monte-Carlo tree search in AlphaGo: **selection**

\[ Q + u(P) \xrightarrow{\text{max}} Q + u(P) \]

\[ Q + u(P) \xrightarrow{\text{max}} Q + u(P) \]

\[ P \quad \text{prior probability} \]

\[ Q \quad \text{action value} \]

\[ u(P) \propto P/N \]
Monte-Carlo tree search in AlphaGo: expansion

$p_\sigma$  Policy network
$P$  prior probability
Monte-Carlo tree search in AlphaGo: evaluation

$v_\theta$  Value network
Monte-Carlo tree search in AlphaGo: rollout
Monte-Carlo tree search in AlphaGo: backup
### Deep Blue

- Handcrafted chess knowledge
- Alpha-beta search guided by heuristic evaluation function
- 200 million positions / second

### AlphaGo

- Knowledge learned from expert games and self-play
- Monte-Carlo search guided by policy and value networks
- 60,000 positions / second
Nature AlphaGo

Seoul AlphaGo
Evaluating Nature AlphaGo against computers

494/495 against computer opponents

>75% winning rate with 4 stone handicap

Even stronger using distributed machines

Go Programs

<table>
<thead>
<tr>
<th>Go Program</th>
<th>Professional dan (p)</th>
<th>Amateur dan (d)</th>
<th>Beginner kyu (k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlphaGo</td>
<td>9p</td>
<td>7p</td>
<td>5p</td>
</tr>
<tr>
<td></td>
<td>3p</td>
<td>1p</td>
<td>9d</td>
</tr>
<tr>
<td></td>
<td>5d</td>
<td>7d</td>
<td>3d</td>
</tr>
<tr>
<td></td>
<td>1d</td>
<td>5k</td>
<td>7k</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3k</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1k</td>
<td></td>
</tr>
</tbody>
</table>

Elo rating

Pachi
Fuego
Gnu

494/495 against computer opponents

>75% winning rate with 4 stone handicap

Even stronger using distributed machines
Evaluating Nature AlphaGo against humans


Match was played in October 2015

AlphaGo won the match 5-0

First program ever to beat a professional on a full size 19x19 in an even game
Seoul AlphaGo

Deep Reinforcement Learning (as Nature AlphaGo)

- Improved value network
- Improved policy network
- Improved search
- Improved hardware (TPU vs GPU)
Evaluating Seoul AlphaGo against computers

>50% against Nature AlphaGo with 3 to 4 stone handicap

CAUTION: ratings based on self-play results
Evaluating Seoul AlphaGo against humans

Lee Sedol (9p): winner of 18 world titles

Match was played in Seoul, March 2016

AlphaGo won the match 4-1
AlphaGo vs Lee Sedol: Game 1
AlphaGo vs Lee Sedol: Game 2
AlphaGo vs Lee Sedol: Game 3
AlphaGo vs Lee Sedol: Game 4
AlphaGo vs Lee Sedol: Game 5
Deep Reinforcement Learning: Beyond AlphaGo
What’s Next?
With thanks to: Lucas Baker, David Szepesvari, Malcolm Reynolds, Ziyu Wang, Nando De Freitas, Mike Johnson, Ilya Sutskever, Jeff Dean, Mike Marty, Sanjay Ghemawat.