Summary of: *Mastering the Game of GO with Deep Neural Networks and Tree Search*

https://storage.googleapis.com/deepmind-media/alphago/AlphaGoNaturePaper.pdf

Background and Goals: Go, though complex, is a deterministic game of perfect information. Such games have an optimal value function which determines outcomes and can be solved by computing this optimal value is a search tree of game states. Large games require search space reduction, since their breadth and depth make exhaustive search infeasible (computation time is prohibitive). Depth search space can be reduced by approximation of the optimal value whereas depth search space can be reduced by sampling (action, state) combinations assembled into policies (such as reinforcement learning q-tables). The complexity of GO renders depth approximation intractable with current computing limitations. Breadth reduction strategies, which have met with success in lower complexity games such as Backgammon, have produced weak, amateurish play in GO. To overcome these researchers applied deep convolutional neural networks (CNN) such as those successfully applied to image classification, facial recognition, and playing Breakout (ATARI), to the game of GO.

Workflow:

- Board positions are passed as 19 x 19 images and CNN constructs a representation
- CNN reduces breadth and depth of search tree by evaluating positions using a value network and sampling actions from a policy network

Pipeline: The pipeline consists of several stages of machine learning (ML)

- Train a supervised policy network from expert human move data
    - Train 13 layer policy network from 30 million positions
    - Predicted 55.7% of expert moves using only raw board positions and move history inputs (up from other attempts at 44.4% accuracy)
    - Small accuracy improvements led to large gains in playing strength

- Train a reinforcement learning policy network by optimizing outcomes of games self-play (Q learning) towards the goal of winning games rather than maximizing predictive accuracy.
    - Aimed at improving the supervised learning policy network
    - Games played between the current policy network and a randomly selected previous iteration of the policy network.
    - The reward function is zero for all non-terminal steps and the outcome reward (+1 win, -1 loss) .
    - Rewards are backpropogated to values throughout the network.

- Train a value network that predicts the winner of games played by the reinforcement learning policy network, against itself.

    - Focuses on position evaluation by estimating the value of a position by predicting the outcome of a state from games played by using a the same policy for both players (s,p for both players).
    - Suffered from overfitting due to the network memorizing game outcomes rather than generalizing to new positions.
    - Overfitting was mitigated by sampling 30 million distinct positions each samples from a separate game.

The extensive sampling and training routines require maximum efficiency from large computational resources but result is unprecedented success against both machine and human opponents.

Results: The search strategy along with clever use of computational resources resulted in:

- 99.8% win rate against other GO programs
- A 5 to 0 win record against a European GO champion (human). This is the first instance of a computer defeating a human in a full-sized game of GO.