Given the enormous search Space and the inability to design an accurate evaluation function, the game of GO has long been seen as the most challenging games for artificial intelligence. Google Deep Mind presents an AI Agent named AlphaGo that uses deep neural networks, expert supervised learning and reinforcement learning to effectively learn both policy and evaluation functions from actual gameplay, reaching super human level in tournament Gameplay.

Implementation

'Value Networks' are used to evaluate the state of the current board for long term likelihood of success, and 'policy networks' enforce the rules of the game in selecting possible moves.

The Networks are learned in a series of steps, moving from:

1) Learning the 'policy network' from available game play records using supervised learning (SL)
2) Training a 'reinforcement learning policy network' (RL) that continues to improve the (SL) Network in adversarial gameplay against itself, learning from its own experience.
3) Training an 'evaluation network' on its own database of moves to effectively predict the winner from a given state in the gameplay

The **policy network** is trained on human expert moves using an alternating network of Convolutional and rectified Nonlinear Units with a final Softmax Layer for probability prediction over all legal moves. This initial network was then improved by policy gradient learning on maximizing wins against older versions of itself in a large number of parallel gameplays. The final network architecture incorporated 13 Layers trained on over 30 million positions from a public GoServer, which maximized accuracy under the constraint of evaluation time during search.

A KNN is used to learn a **representation of the 19 x 19 board** state to reduce the depth and breadth of the search space.

The **evaluation network** was used to approximate the evaluation score for the strongest policy using the RL policy network, following the same Architecture as the policy network, but outputs a single evaluation metric. It was trained on state-outcome-pairs using stochastic gradient descend, minimizing MSE loss, sampled from board states from parallel games to strong correlation of successive board positions.

Search combined MCTS with policy and evaluation networks in a lookahead beam search, traversing by simulation and maximizing action value plus bonus to encourage explorative gameplay. Leaf Nodes are evaluated first by the policy netork and the output propabilities are stored as priors for each action at this stage. The simulation terminates by evaluating all leaf nodes using the value network and the outcome of random roll out played till termination, updating action values and visit counts on all traversed edges.

Results

The previously available Programs where beaten in 99.8% of games played, and at the time of the Article Alpha Go had already beaten the European Go Champion 5/0. Unlike the implementation of DeepBlue, AlphaGo does not involve any handcrafted evaluation functions, but relies solemnly on general purpose machine learning Algorithms, which is it's overall biggest technological leap.