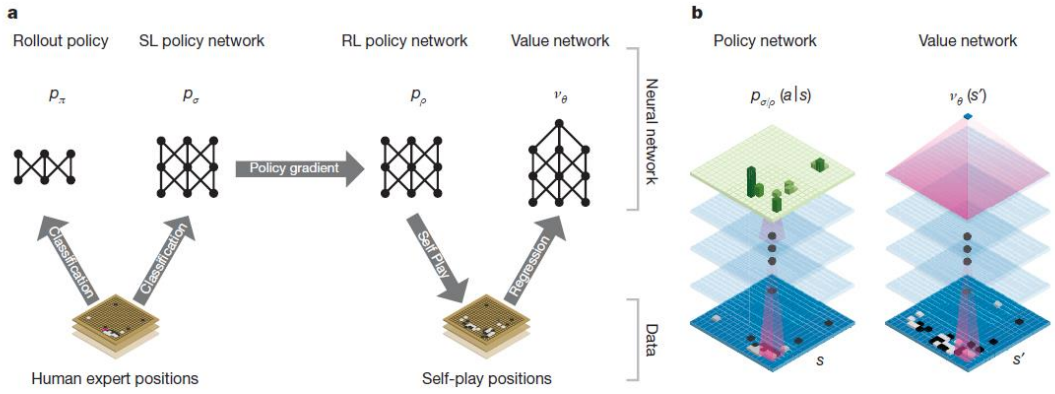
Research Review

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

Techniques in AlphaGo

a. Use SL(Supervised Learning) and fast rollout policy to train a lot of experts' playing history and create policy gradient to predict human moves. RL(Reinforcement Learning) policy network is initialized to SL policy network, train and maximize the outcome against previous result of policy network. RL policy network will generate new data set by playing self-play. Finally, the value network is trained by regression to predict the expected outcome in positions from the self-play data set.

b. The policy network represents the position in the board. It is built by many convolution layers. It will output a probability distribution of every legal move. The value network uses many convolution layers to predict the expected outcome in every position.



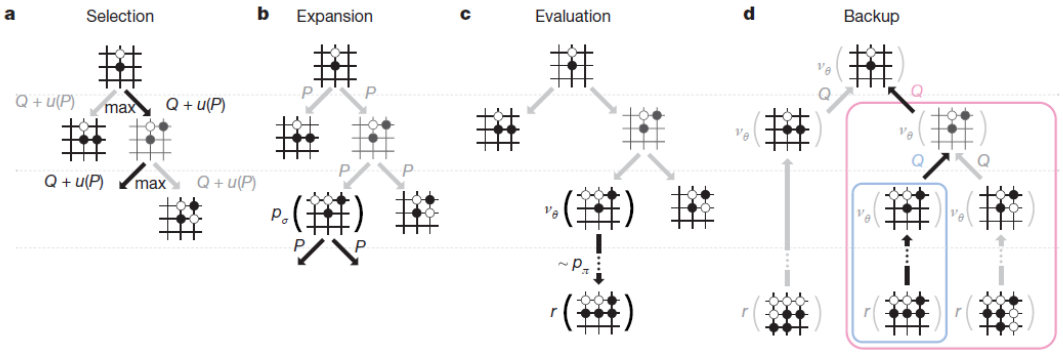
Monte Carlo tree search in AlphaGo

a. Each simulation traverses the tree by selecting the maximum action value plus a bonus that depends on the prior probability in the edge.

b. The leaf nodes may be expanded. The new node is processed once by policy network and store the output probability.

c. At the end of simulation, the leaf nodes use the value network. Run a rollout with fast rollout policy to compute the winner.

d. Action value Q is updated to track the mean value of all evaluation score in the subtree.

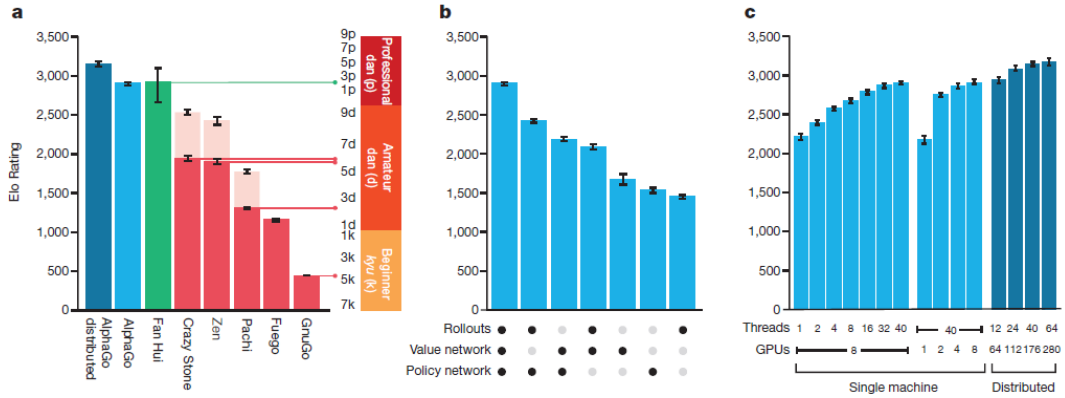


The result in AlphaGo

a. AlphaGo has high win rate than other human player and AI player. It gets high Elo rating.

b. AlphaGo was improved significantly by the strategies such as Rollouts, value network and policy network.

c. Use multithread distributed search in Monte Carlo tree search will increase the win rate within 2s per move.



The goal AlphaGo achieved

AlphGo combines the tree search with policy and value networks. Also, it uses distributed search engine to get the high-performance. Policy network is useful to evaluate the score precisely. AlphaGo is better than Deep Blue, it chooses fewer position in a short time. Therefore, these methods let AlphaGo go beyond human player.