Research Review
of
AlphaGo by the DeepMind Team

Arthur: Yuewei Wang

Date: 06/17/2016

Goals or Techniques Introduced

AlphaGo was developed by Google DeepMind team, for the purpose of training neural networks that can play Go intelligently. The goal is to develop a neural network that is smart and efficient enough to win human and other agents.

AlphaGo combines policy network and value network with Monte Carlo tree search(MCTS), where value network evaluates board position while policy network selects moves.

Firstly, a supervised learning policy network was trained on randomly sampled data using stochastic gradient ascent. Next, a reinforcement learning policy networks was trained to improve SL policy network, by playing current policy network with randomly selected previous policy network to improve weights. Then, value networks, which estimate a value function that predicts the outcome of board position, is introduced. Instead of outputting a probability distribution, value networks output single prediction. To solve overfitting problem, a new dataset was generated by letting the agent plays with itself. Finally, the policy networks and value networks are combined with MCTS. Since neural networks are strong but impractical while MCTS rollout is weaker by more efficient, combining them together can let them comprehend each other and help make a fast and high-accuracy decision.

The techniques that were applied to AlphaGo involves strategy algorithms like MCTS and machine learning models like neural networks. For search algorithm, an asynchronous policy and value MCTS algorithm was implemented to integrate with large neural networks efficiently. KGS data set containing the professional games were used to let the networks learn how to classify positions. Then the policy networks were further trained on policy gradient reinforcement learning. Regression

model, which is to minimize the mean squared error, was used to train the value networks to approximate the policy networks. Features like stone color, liberties (adjacent empty points of stone's chain), captures, legality, turns since stone was played, and outcome of a ladder search were introduced into the training process. In total, it was a 19×19×48 image stack consisting of 48 feature planes. In addition, a binary feature plane describing the current color to play is feature set for value networks.

Result

As a super spotlight all over the media, AlphaGo has achieved huge success. The outcome approved its intelligence: 99.8% winning rate against other Go programs, and defeated the human European Go champion by 5 games to 0.

An internal tournament and measuring the Elo rating of each program to evaluate AlphaGo's ability. Games were scored using Chinese rules with a komi of 7.5 points. In those game played with professional player Fan Hui, five formal and five informal games, AlphaGo won these games 5–0 and 3–2 respectively.