AlphaGo Zero paper review summary

Backgrounds:

This is the review summary for paper Mastering the game of Go without human knowledge.
AlphaGo is the first program built by Google that defeated a world champion in the game of Go.
AlphaGo Fan is the AlphaGo program that played against Fan Hui in October 2015.
AlphaGo Lee is the AlphaGo program that defeated Lee Sedol 4-1 in March 2016.
AlphaGo Zero was introduced by this paper and it learned without human knowledge.
AlphaGo Master is the program uses the same architecture as AlphaGo Zero except it uses handcrafted features and initialized by supervised learning from human data. AlphaGo Master defeated top human players by 60-0 in January 2017.

Goals:

It shows AlphaGo Zero achieves superhuman performance based solely on reinforcement learning without human knowledge

Techniques introduced:

1. AlphaGo Zero uses a new self-play reinforcement learning algorithm that incorporates lookahead search inside the training loop
2. AlphaGo Zero uses a single neural network that outputs move probabilities and position value instead of separate policy and value networks
3. AlphaGo Zero uses a simplified Monte Carlo tree search that relies upon its neural network without any rollouts to guide the next move
4. AlphaGo Zero uses residual tower that consists of convolutional block followed by residual blocks in its neural network
5. AlphaGo Zero uses only black and white stones from the board as input features, game rules, Go scoring, and rotation and reflection invariant rule as domain knowledge

Empirical Results:

AlphaGo Zero performance vs. AlphaGo Lee performance:
AlphaGo Zero outperformed AlphaGo Lee after just 36 hours' training. AlphaGo Lee was trained over several months.
After 72 hours, AlphaGo Zero defeated AlphaGo Lee by 100 games to 0. AlphaGo Zero used a single machine with 4 tensor processing units while AlphaGo Lee was distributed over many machines and used 48 TPUs.

Self-play reinforcement learning vs. Supervised learning on predicting professional moves:
Supervised learning shows higher prediction accuracy than self-play reinforcement learning. Nonetheless, AlphaGo Zero can defeat human-trained players which indicates that AlphaGo may be learning different strategies than human players.

Four different neural network architectures (dual-res, sep-res, dual-conv and sep-conv) in AlphaGo are compared with each other on Elo rating, prediction accuracy on professional moves and MSE of professional game outcomes:
Residual network was more accurate, achieved lower error and improved performance;
Single network may not improve prediction accuracy but reduced error and improved performance than separate networks.

Final performance of AlphaGo Zero:

A second instance of AlphaGo Zero with larger neural network (40 residual blocks) was trained over longer duration for about 40 days. It passes AlphaGo Lee on Elo rating after about 3 days' training. It then passes AlphaGo Master on Elo rating after about 21 days' training. After about 40 days' full training, AlphaGo Zero ranked highest on Elo rating compared with AlphaGo Master, AlphaGo Lee, AlphaGo Fan, Crazy Stone, Pachi and GnuGo. Finally, AlphaGo Zero was evaluated against AlphaGo Master in a 100-game match with 2-h time controls. AlphaGo Zero won by 89 games to 11.