Exposing AlphaGo(Zero)'s Weaknesses

Authors: Kien Nguyen and Khiem Nguyen
Advisors: Dr. Ze-Li Dou and Dr. Liran Ma

TCU College of Science & Engineering
reSEaRCh
Science and Engineering Research Center

Overview

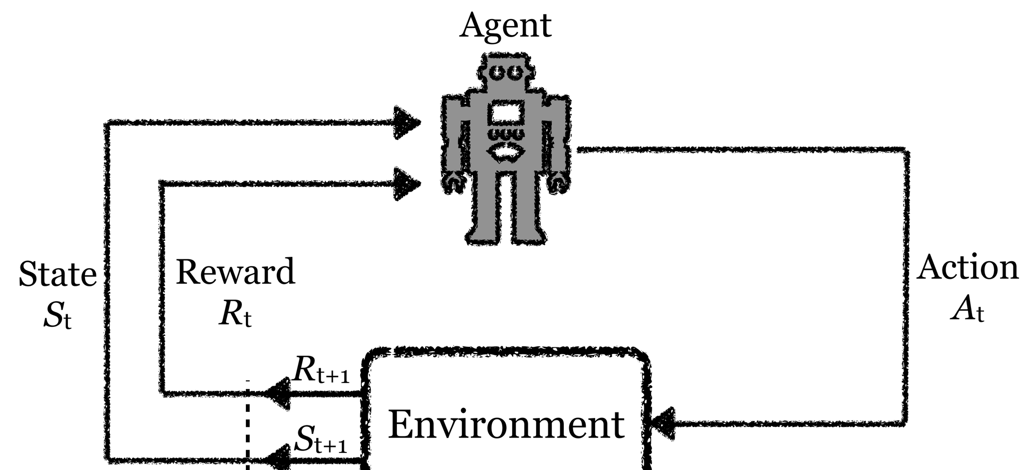
SciCom Let's Talk Science

AlphaGo(Zero), though being the state-of-the-art computer Go program in recent years, exhibits weaknesses upon careful analysis. They are caused by the binary reward signal, win or loss, of the environment setting during the training phase. This characteristic impedes the agent's ability to discover the optimal lines of play, despite its superhuman level of play.

Go

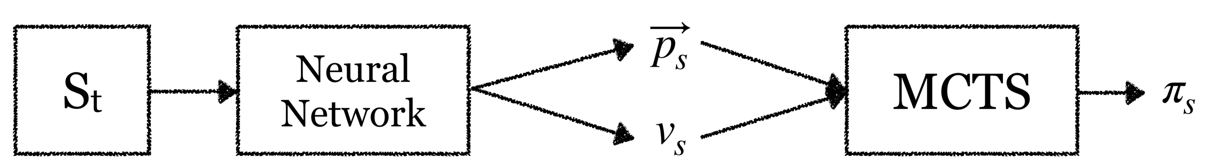
- Board game for 2 players first invented in China.
- Goal is to take over more territory.
- One plays Black stones, one White. Black plays first.
- Stones are captured when a group of stones lose all their liberties.
- A player's **score** is the number of stones plus the number of intersections surrounded by the stones, on the board.
- **Komi** is a compensation added to White's final score, due to disadvantages for going second.
- Standard board size is 19×19.
- Number of board positions is around 10^{172}.

Reinforcement Learning



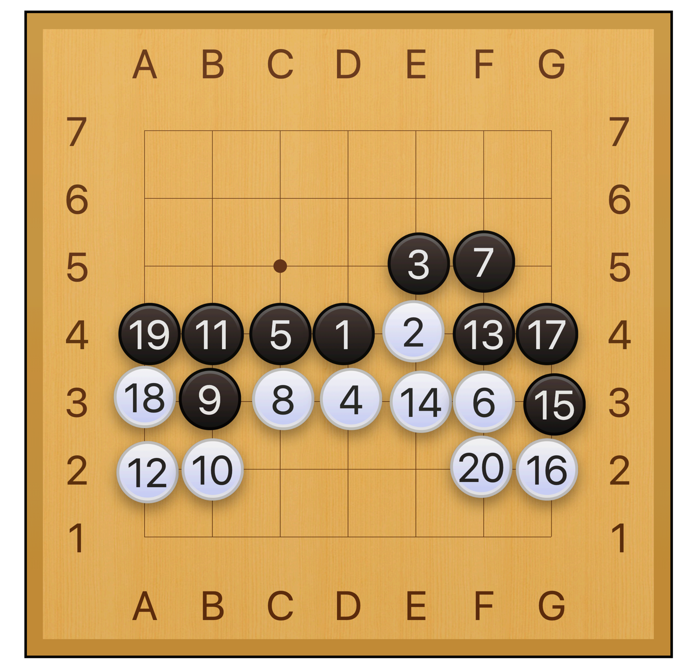
- Agent interacts with the environment by taking an action A_t. The environment emits a reward signal R_{t+1} and transits to state S_{t+1}. The goal is to maximize the reward in the long run.
- In AlphaGo(Zero)
  - The **environment**:
    - the Go board and rules
    - reward function ($r \in \{-1, 1\}$)
  - The **action**: placing stones on the board

AlphaGo(Zero)'s Components



For each turn,
- **S_t**: the state at time-step t; contains a history of board snapshots and who to play.
- **Neural Network**: takes in S_t and outputs a move probability vector $\vec{p_s}$ and win percentage scalar v_s.
- **MCTS**: $\vec{p_s}$ and v_s are utilized the Monte Carlo Tree Search that outputs the final policy.
- π_s: the final policy from which the next action is chosen.

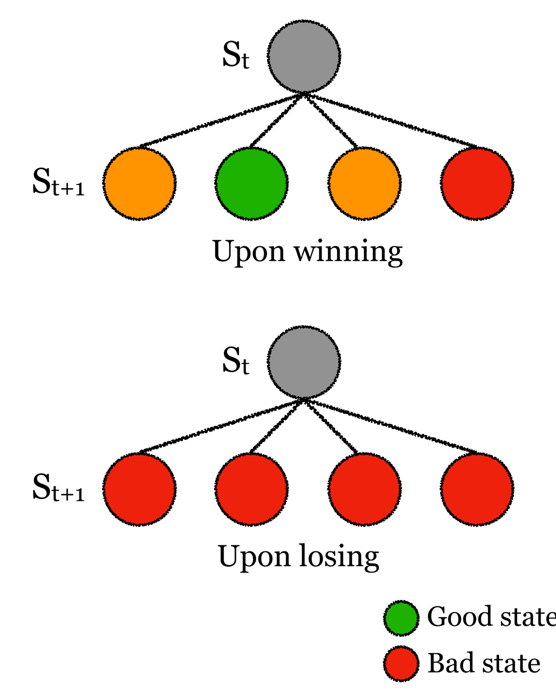
Optimal Play



- An optimal play happens if an agent either maximizes the winning margin or minimizes the losing margin at the end of the game.
- Currently, Go is not solved for the standard board size of 19×19, or more precisely any board size greater than 7.
- In our experiment, we scale the board down to 7×7 for computation feasibility, and because optimal sequences of play are known on this board size.

Weakness

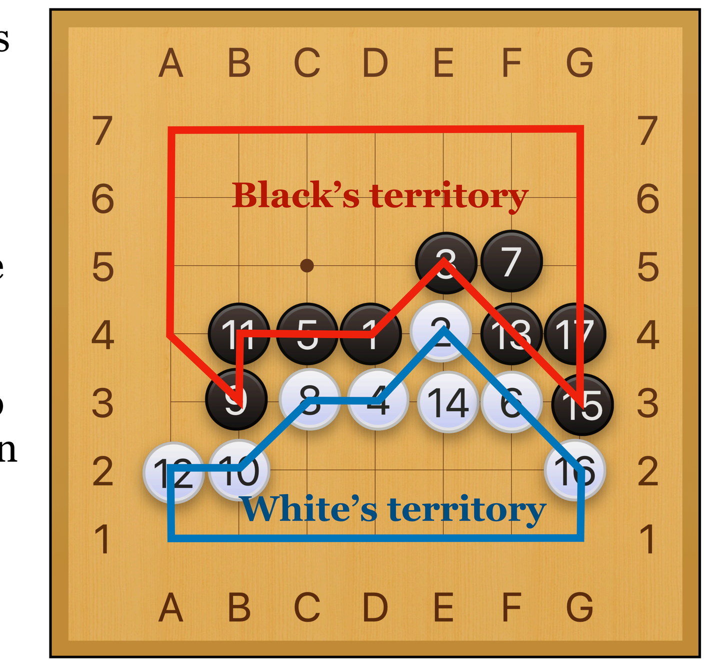
- In AlphaGo(Zero), the reward signal is $r \in \{-1, 1\}$ with -1 being a loss and 1 being a win from the current player's perspective.
- Due to this characteristic, the agent tends to play to win, rather than to obtain the *optimal* sequence(s). We dub this behavior *strategic*.
- When winning, the agent tends to be more conservative to protect its victory. When losing, it becomes more aggressive because losing by 0.5 or 5.5 points earns the same reward signal.



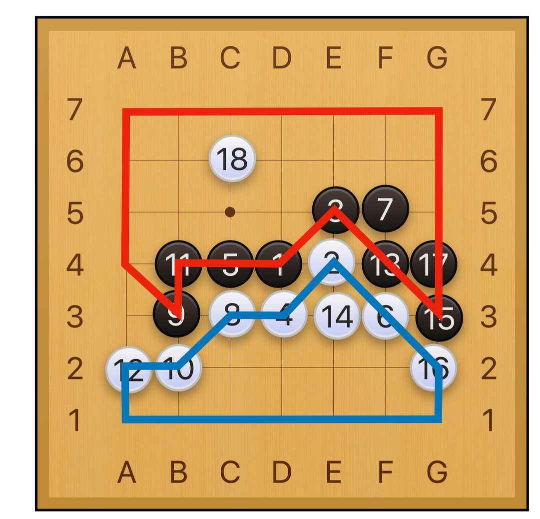
- The value of the next states S_t seem to be the same for the losing side.

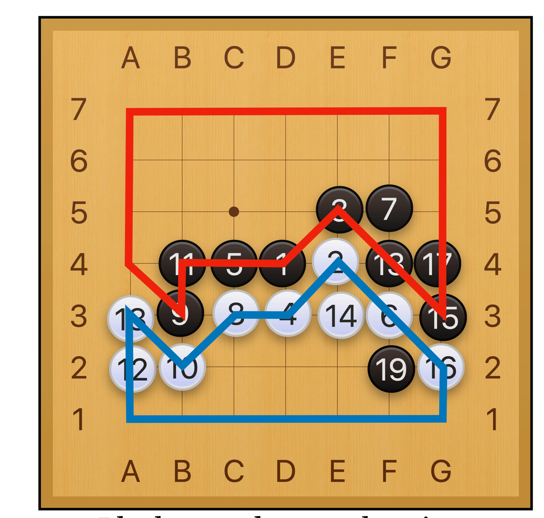
Experiment

- To confirm our hypothesis described in Section *Weakness*, we train two agents, one with komi 8.5 and one with 9.5. At 8.5, Black has more advantage to win the game, and vice versa at 9.5. We pick a solution proposed by a Go professional Li Zhe, shown in Section *Optimal Play*, and remove 3 last moves (18, 19, 20) in the sequence.



E1. Overplay


White overplay, 8.5-komi


Black overplay, 9.5-komi

- The graphics illustrate Black's and White's overplays when losing that result in aggressiveness urging them to recklessly play into the opposing territory.

E2. Optimal Play Reproduction

- When we pit a 8.5-komi Black agent against a 9.5-komi White agent, we are able to reproduce the line of optimal play mentioned earlier because both players perceive to be winning, thus playing less aggressive.

Conclusion and Future Work

- In this work, we show that AlphaGo(Zero)'s behaves sub-optimally in sensitive situations via well-designed experiments.
- Our future work aims to quantify the true strength of AlphaGo(Zero). Using the Go game, as a medium, to further understand the state-of-the-art artificial intelligence, we hope to develop a method to improve the model following such quantifications.

Technologies Used

TensorFlow
Deep Learning framework

Write training script

Implement Go game

References

- D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of Go without human knowledge," Nature, vol.550, pp. 354–359, Oct. 2017.
- Z. Li, "Qi Lu Qi Pan Zui You Jie (7x7 Go Optimal Solutions), "The World of Weiqi, vol. 20, Oct 2015.
- G.-C. Pascutto and contributors, "Leela Zero," 2019. [Online]. Available: https://zero.sjeng.org
- Ze-Li Dou, Liran Ma, Khiem Nguyen, and Kien X. Nguyen, "Paradox of AlphaZero: Strategic vs. Optimal", IPCCC, 2020.

Acknowledgements

The team would like to thank the following individuals:
- **Dr. Ze-Li Dou:** for providing insights of the Go game and guiding the research direction.
- **Dr. Liran Ma:** for eliciting new ideas and addressing numerous technical problems.