Case Study 4

BIOE 498/598 PJ

Spring 2021

Before completing this assignment, watch the free documentary on AlphaGo, a reinforcement learning system developed by DeepMind. The course website includes a link to the video. Some of the documentary is in Korean, so it's best to turn on subtitles.

1. AlphaGo uses a single reward at the end of the game (+1 for a win, 0 for a loss). Why did the engineers choose this strategy rather than reward good moves throughout the game? Is there any disadvantage to deferring all reward until the game finishes?

2. We could also defer all reward for our Gridworld games by giving +1 when the agent reaches the finish square and 0 for all other moves. Why did we choose to give a reward of -1 for each step instead? Why is this not a concern for AlphaGo?

3. Games like Go are called "perfect information" games. What does "perfect information" mean? Are perfect information games easier or harder for RL agents to solve? Do biological experiments have perfect information?

4. AlphaGo has three parts: 1.) policy neural network that recommends the next action for each state; 2.) a value neural network that predicts the probability of victory given each state; and 3.) a local search algorithm that plays ahead 50-60 moves using the policy and value networks. Our Gridworld agent did not include a local search feature. Why do games like Go require playing ahead to find good moves while Gridworld does not?

5. With Gridworld we demonstrated how policy improvement can start with a random policy and iteratively find the optimal policy. AlphaGo was "bootstrapped" by extracting a starting policy from online games. What is the advantage of starting with a bootstrapped policy?

6. DeepMind eventually built AlphaZero, an agent that learned to play Go (and chess and shogi) without bootstrapping. AlphaZero quickly learned to beat AlphaGo. Can you guess any reason why AlphaZero became superior to AlphaGo?