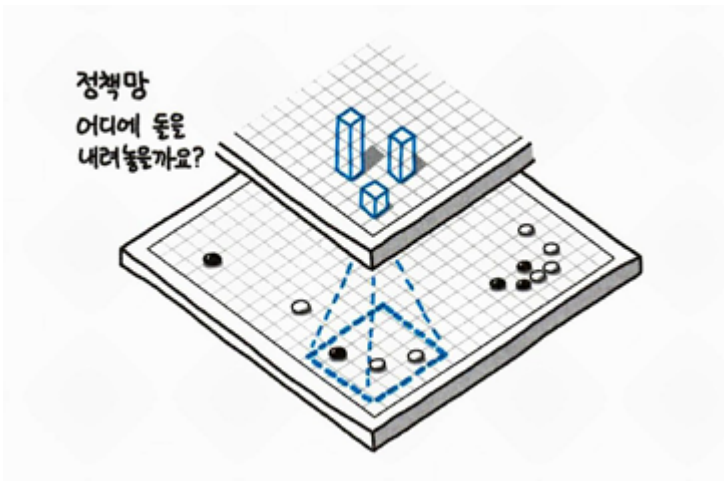
2장. 알파고

알파고, 두 종류의 인공 신경망

- 정책망 Policy Network
- 가치망 Value Network

정책망(Policy Network)



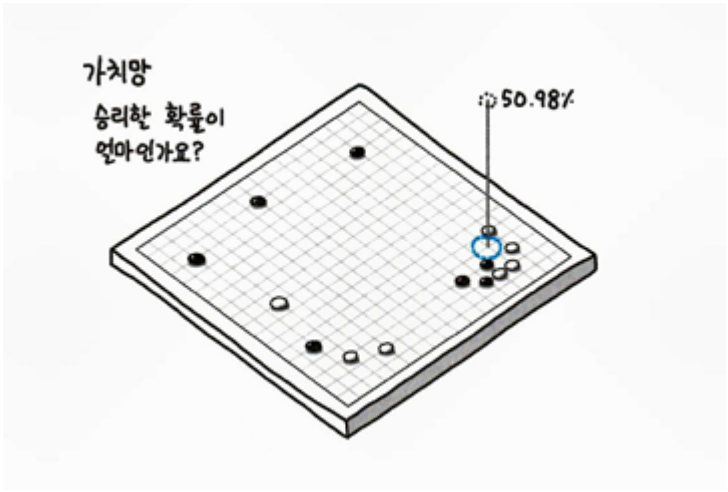
- 사람이 만든 기보를 이용해 학습
  - 그러나 바둑 기사들 마다 스타일이 다르기 때문에 바둑의 모든 국면을 학습할 수는 없음
- 따라서 정책망을 크게 3가지로 나눔
  - 사람의 기보를 이용한 기보학습 정책망
  - 롤아웃 정책망
    - 기보학습 정책망과 비슷하지만 훨씬 작고 가벼운 망
  - [핵심] 강화학습 정책망
    - 스스로 대국하며 강화학습을 수행
    - 강화학습 정책망 이용 시 기보학습 정책망과 대국을 둘 경우 80% 확률로 강화학습 정책망이 승리
-

	정책망	특징
1	사람의 기보를 이용해 학습한 정책망(기보학습 정책망)	57% 정확도. 5단 수준
2	롤아웃 정책망	24% 정확도. 1,500배 빠름
3	스스로 대국하며 강화학습한 정책망(강화학습 정책망)	기보학습 정책망과 대전하면 80% 확률로 승리

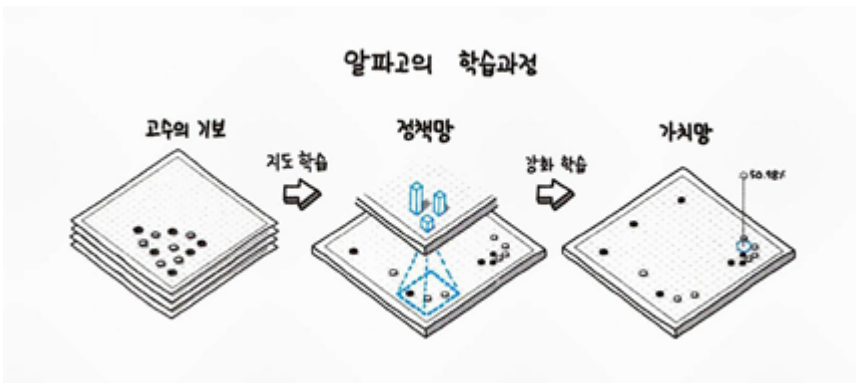
가치망(Value Network)

- 현재 국면에서 승패 여부를 예측하는 망
  - 승리할 가능성이 높은지, 패배할 가능성이 높은지를 확률로 표현

- 정책망의 경우 361개의 바둑 칸 수 중 수를 둬야할 한 지점을 골라내는 신경망이라면, 가치망은 오로지 승리할 가능성만 계산

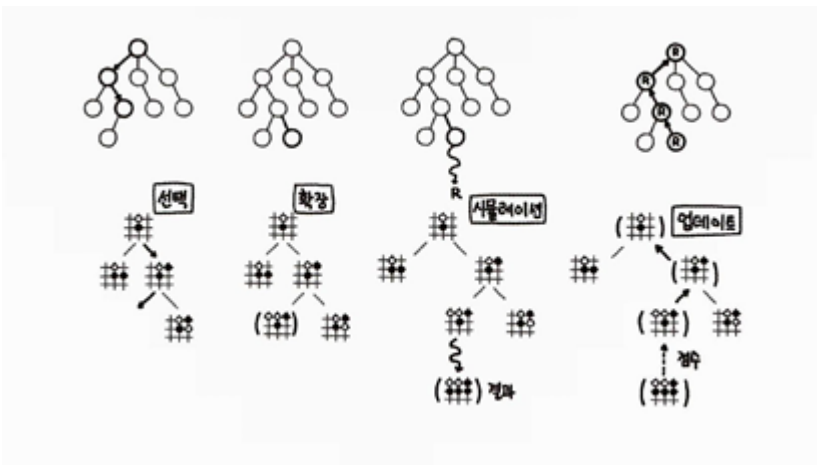


- 알파고의 학습 과정

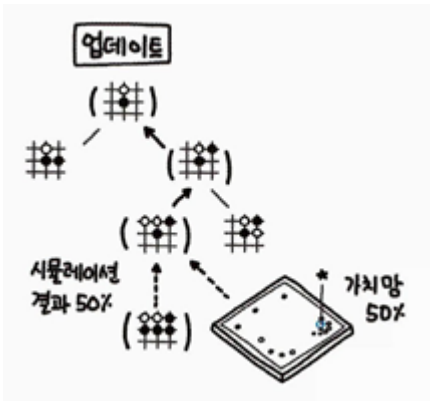


알파고가 수를 두는 방법

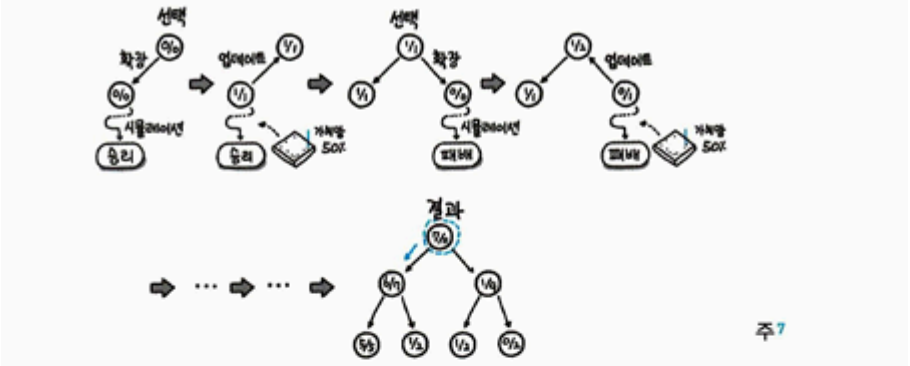
- 몬테카를로 트리 탐색 활용
  - 무작위로 샘플링하여 정답을 찾는 방식
- 몬테카를로 트리 탐색 과정
-



- 승리할 가능성이 높아보이는 수부터 선택하여 트리 탐색 시작
- 그후 기보학습 정책망으로 다음 수를 어디에 둘지 확장
- 게임이 끝날때 까지 시뮬레이션
  - 빠른 시뮬레이션을 위해 롤아웃 정책망을 사용
- 시뮬레이션을 통해 만든 가치망의 점수를 승리 여부에 함께 반영
-



- 업데이트 처리
  - 모든 수에 점수를 업데이트 한 뒤 가장 많이 진행한 수를 다음 수로 선택
    - 이유: 신뢰도 때문
- 정리
-



주7

1. 어떤 수에서 시작할지 승리할 가능성이 높아 보이는 수를 선택합니다.
2. 기보학습 정책망을 이용해 다음 수를 어디에 둘지 확장합니다.
3. 롤아웃 정책망을 이용해 끝까지 시뮬레이션합니다.
4. 가치망의 점수를 50% 반영한 승패 여부를 모든 수에 업데이트합니다.
5. 1~4의 과정을 계속해서 반복한 후 최종적으로 가장 많이 진행한 수를 다음 수로 결정합니다.

여기까지가 바로 이세돌과 승부를 겨룬 알파고가 수를 결정하는 원리입니다.