Mastering the game of Go without human knowledge

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel & Demis Hassabis

서론

이 연구는 인간의 지식이나 데이터 없이 강화 학습만을 이용해 바둑을 플레이하는 인공 지능, 알파고 제로를 발전시킨 것에 대해 설명한다. 전문가 데이터를 사용한 학습 시스템은 비용이 많이 들고 신뢰성이 제한적일 수 있다. 그러나 강화 학습 시스템은 자신의 경험을 통해 학습하므로, 인간의 능력을 능가하고 인간의 지식이 부족한 영역에서도 작동할 수 있다. 알파고 제로는 인간의 데이터 없이 랜덤 플레이로부터 시작하여 강화 학습만으로 훈련되었으며, 단일 신경망과 몬테카를로 트리 검색(MCTS)을 통해 바둑판 상황을 평가하고 수를 선택하였다. 알파고 제로는 게임 보드의 흑백 돌 만을 입력으로 사용하였고, 새로운 강화 학습 알고리즘을 도입하여 이전 버전의 알파고보다 더욱 빠르고 정확하게 학습하였으며, 이로 인해 바둑 세계 챔피언을 이기는 등의 높은 성과를 보여주었다.

AlphaGo Zero 안의 강화학습

Deep Neural Network

이 신경망은 위치와 그 역사의 원시 보드 표현을 입력으로 취하여 이동확률 $(p, v) = f_\theta(s)$ 를 출력하고, 정책 네트워크와 가치 네트워크의 역할을 하나의 아키텍처로 결합한다. 이 신경망은 바둑판의 현재 상태를 분석하여 게임의 결과를 예측하는 가치 네트워크와 가능한 모든 다음 수를 평가하는 정책 네트워크 작업을 담당한다.

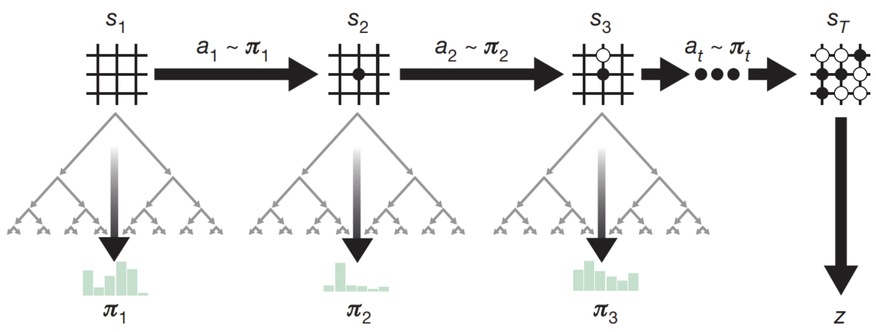
P : p(a) = Pr(a|s)를 선택할 이동 확률(a: 각 이동)

V : 현재 플레이어가 위치한 s에서 우승할 확률을 산정하는 스칼라 평가 값

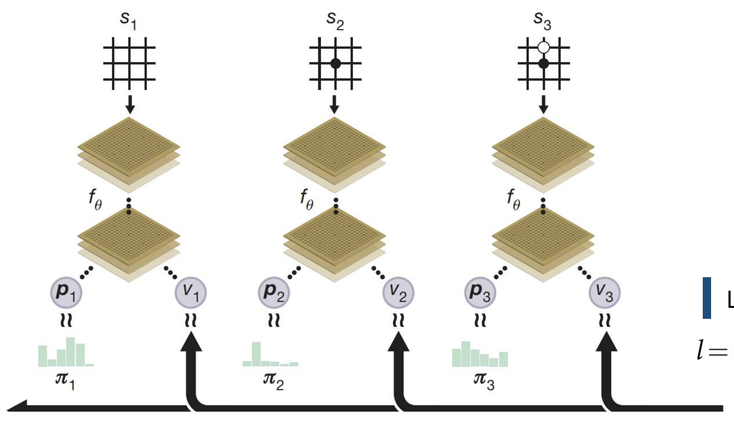
AlphaGo Zero의 신경망은 새로운 강화학습에 의한 self-play 게임에 대한 교육을 받는다. 각각의 위치 s에서 신경망에 의해 유도된 MCTS(Monte Carlo Tree Search)이 실행된다. MCTS 탐색은 각 이동한 확률 π를 출력한다. 이 탐색 확률은 일반적으로 $f_\theta(s)$ 의 p보다 훨씬 더 강한 이동을 선택하기에 MCTS는 강력한 정책개선 방법으로 간주된다. 향상된 MCTS 기반 정책을 사용하여 각 이동을 선택하고 게임 승자 z를 가치 샘플로 사용하여 검색을 통한 셀프 플레이를 강력한 정책 평가 연산자로 볼 수 있다.

Self-Play

자체적으로 번갈아 게임을 진행하며, State는 $s_1, s_{2,\dots,} s_t$가 있다. 이때 매 State마다 MCTS(Monte Carlo Tree Search)를 통해 π_t를 계산하고 Sampling 하여 a_t 를 얻는다. 게임이 끝나면 게임 승자 z값을 계산하게 되는데, 결국 각 state마다 (s_t, π_t, z_t)가 남는다.



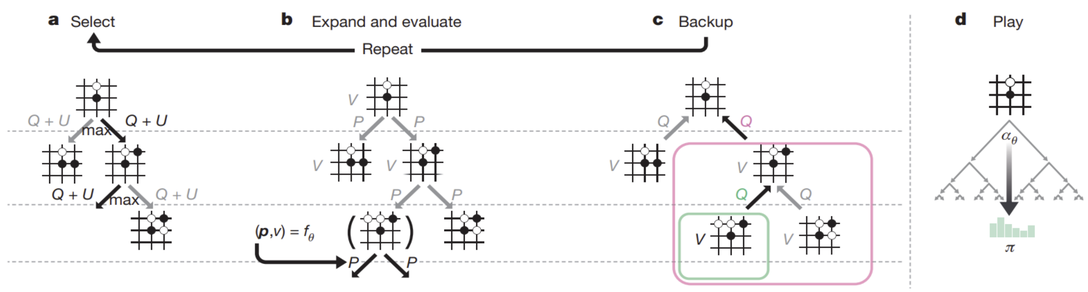
- $f_\theta(s) = (p, v)$
- p_t는 π_t와 같아지도록, v와 z와 같아지도록 θ를 업데이트한다. 데이터는 과거 게임에서 쌓은(s_t, π_t, z_t) 중 random sampling 하여 사용한다.



Loss 함수

$$l = (z - v)^2 - \boldsymbol{\pi}^{\mathrm{T}} \log \boldsymbol{p} + c\|\theta\|^2$$

AlphaGo Zero에서의 MCTS



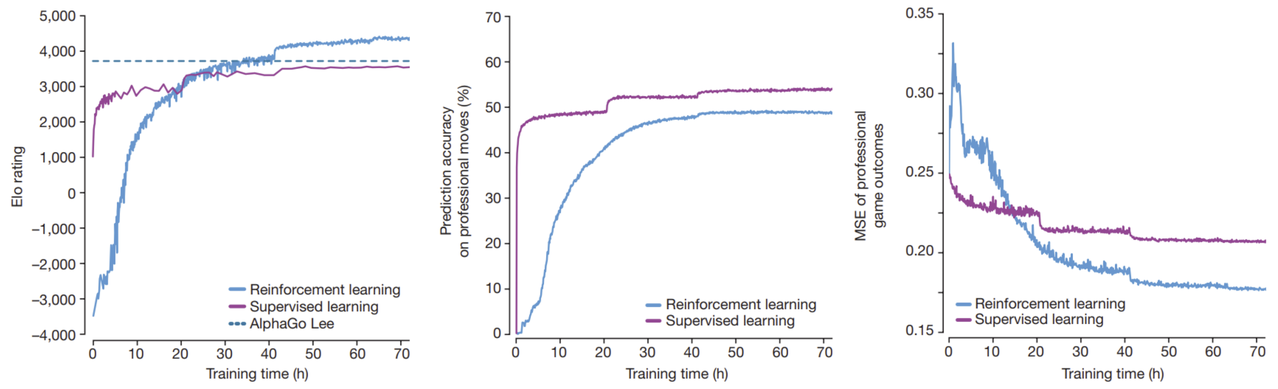
MCTS는 신경 네트워크 f_θ를 사용하여 시뮬레이션을 유도한다.

검색 트리의 각 edge(s,a)는 이전 확률 P (s,a), 방문횟수 N(s,a) 및 동작 값 Q(s,a)를 저장한다.

1. Leaf Node에 도달하면 $f_\theta(s) = (p, v)$ 를 evaluate 하여 p 와 동작 값 Q 를 계산한다.
2. 계산된 값을 이용해 Q 값을 Update 해준다.
3. 위 과정을 여러 번 반복하여 탐색이 종료되면 최종 확률 분포 π 가 반환된다.
4. $\pi_a \propto N(s, a)^{1/\tau}$ 방법으로 정책을 업데이트한다. (τ : temperature parameter)

AlphaGo Zero 교육의 경험적 분석

AlphaGo Zero의 경험적 평가



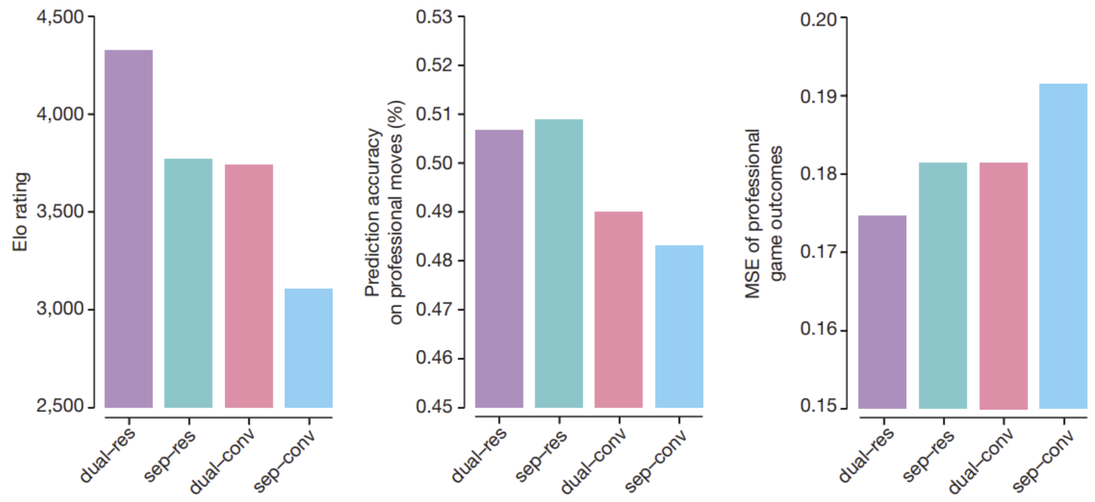
A. self-play 강화학습의 성능 B. 인간의 전문적인 움직임에 대한 예측 정확도 C. 인간 프로 게임 결과의 평균 제곱 오차

비교를 위해 같은 아키텍처를 사용한 네트워크로 KGS data-set에 대해 지도학습을 한다.
이때, ELO Rating은 0.4초의 Thinking Time을 사용하여 서로 다른 플레이어 간의 평가 게임에서 계산된다.

* Elo Rating : 바둑 경기의 승패를 기반으로 한 플레이어의 상대적인 실력을 수치화한 시스템으로, 게임에서 이길 경우 상승한다.

AlphaGo Zero와 AlphaGo Lee의 신경망 구조 비교

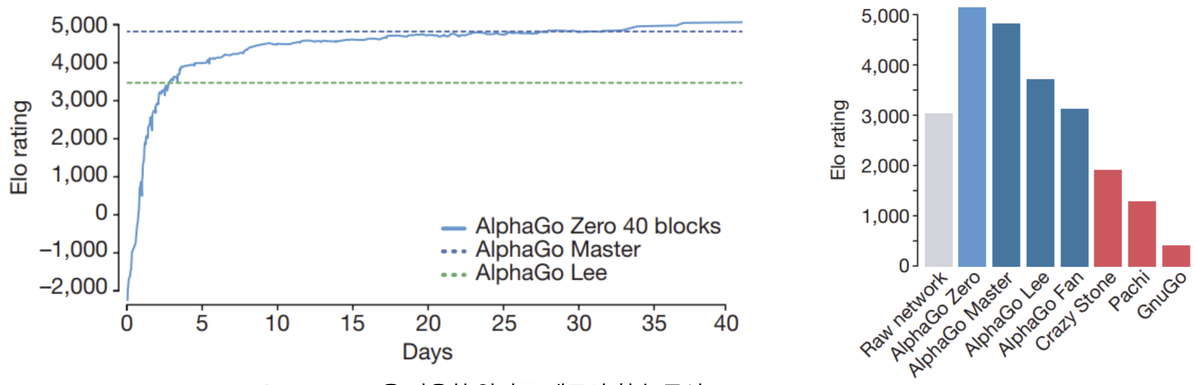
AlphaGo Zero의 신경 네트워크 아키텍처 성능을 AlphaGo Lee에서 사용된 이전 신경 네트워크 아키텍처와 비교하였고, 4개의 신경 회로망이 만들어진다.



- dual : Value와 Policy가 하나의 네트워크를 공유
- sep : Value와 Policy가 각각의 네트워크를 가지고 있음
- res : Residual network
- conv : Convolutional network
- dual-res : AlphaGo-Zero
- sep-conv : AlphaGo-Lee

AlphaGo Zero가 배운 지식

AlphaGo Zero의 성능



A. 40 residual block을 사용한 알파고 제로의 학습 곡선

AlphaGo Zero : 5,185
AlphaGo Master : 4,858
AlphaGo Lee : 3,739
AlphaGo Fan : 3,144

1. ELO 점수는 0.4초 THINKING TIME을 두고 평가
2. AlphaGo Zero 최종 성능
   - 기존 알파고와 다른 프로그램, Raw Network가 포함된 토너먼트의 결과
   - 5초의 thinking time을 두고 평가
   - Raw Network는 확률이 가장 높은 선택지를 바로 선택하는 방식
   - 200점 차이는 75% 확률로 이긴다는 뜻

결론

알파고 제로는 순수 강화학습만을 이용해 어려운 문제를 해결하는 능력을 입증하였다. 인간의 지침이나 사전 지식 없이도 바둑의 기본 규칙만으로 초인적 수준의 학습을 이루었고, 이는 약간 더 긴 학습 시간을 필요로 했지만 전문가 데이터로 학습한 것에 비해 훨씬 뛰어난 성능을 보였다. 인간은 수천 년 동안 바둑을 플레이하며 지식을 축적해 왔다. 하지만 알파고 제로는 며칠 동안의 학습으로 인간의 바둑 지식을 재발견하고 새로운 통찰력과 전략을 개발하는데 성공하였다.

숭실대학교 대학원 소프트웨어학과 석사과정 서영재