알파고 제로 vs 다른 알파고

세간에 알려진 알파고는 총 4가지 버전으로 존재한다. 지난 2015년 10월 천재 바둑 기사 판 후이(Fan Hui) 2단을 이기고 2016년 네이처(Nature)에 실린 버전인 알파고 '판(Fan)', 2016년 3월 이세돌 9단을 4대 1로 이긴 알파고 '리(Lee)', 커제 9단과 대결에서 3:0 완승을 거둔 알파고 '마스터(Master)', 그리고 2017년 네이처를 통해 공개된 알파고 '제로(Zero)'가 바로 그것이다. 참고로 알파고 리, 마스터, 제로의 구조와 학습법은 이번 논문에서 새롭게 소개됐다.

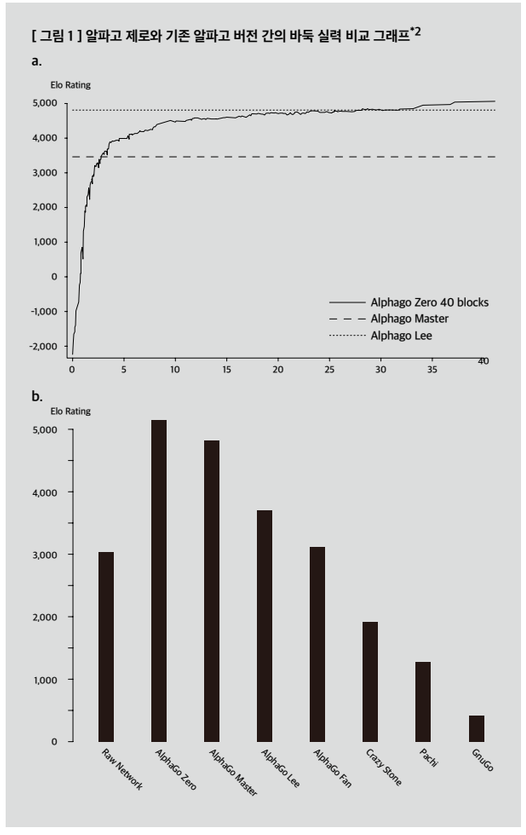
글 | 이수경 samantha.lee@kakaobrain.com

2016년 3월 알파고와 이세돌 9단이 펼치는 세기의 대결을 두눈으로 목도한 이후 인공지능을 제대로 공부해봐야겠다고 결심했다. 인공지능 본진이자 연구소인 카카오브레인으로 걸어들어온 이유다. 인공지능 기술과 이로 인해 바뀔 미래 사회를 다루는 글을 통해 사람들과 소통하고 싶다.

기술감수 | 천영재 yeongjae.cheon@kakaobrain.com

알파고 제로는 이전(前) 세대와 비교했을 때 월등한 성능을 자랑한다. 순위 산출에 사용되는 엘로(Elo) 점수[*1]를 기준으로 했을 때 알파고 제로는 5,185점을 보유하고 있다[그림 1]. 알파고 마스터(4,858점)는 327점, 알파고 리(3,739점)와는 1,446점, 알파고 판(3,144점)과는 2,041점의 격차가 있었다. 엘로 점수에서 800점 이상 차이 나면 승률이 100%라는 것을 고려했을 때, 알파고 제로가 현존하는 인공지능 바둑 컴퓨터로서 최정상급이라는 점을 부인하긴 어렵다.

[그림 1] 알파고 제로와 기존 알파고 버전 간의 바둑 실력 비교 그래프[*2]



이번 알파고 제로 논문이 시사하는 바에 대해 들어보고자 카카오브레인의 천영재 연구원과 감동근 아주대학교 교수로부터 자문을 구했다. 두 사람은 알파고 제로 이전과 알파고 제로 간 3가지 차이가 있다고 말했다.

첫 번째 : 신경망 통합

알파고 제로 전(前)세대들은 정책망(policy network)과 가치망(value network) 이라는 2가지 종류의 신경망을 갖췄다. 이 두 신경망을 구축한 이유는 앞으로 진행될 경기를 미리 여러 번 진행해보고, 승리할 가능성이 높은 수만을 효과적으로 탐색하기 위해서다.

실제 바둑 한 경기당 2×10^{170}이 넘는 경우의 수가 존재하는데, 이는 전세계에서 가장 큰 규모의 슈퍼컴퓨터로도 다 계산하기 어려운 규모다. 따라서 시뮬레이션 횟수를 줄이면서도(깊이), 승률이 높은 수(너비)를 찾는 탐색 알고리듬 구축이 관건이라고 볼 수 있다.

정책망은 바둑판 상태를 분석하여 361(=19×19)가지 경우의 수중에 가장 수읽기 해볼 만한 몇 가지 수를 선택한다. 가치망은 어떤 수를 두었을 때 그 후에 일어날 미래 대국을 시뮬레이션해본 뒤 그 결과로부터 승패를 예측한다. 보다 쉽게 이야기하자면 정책망은 '다음에 둘 수'를, 신경망은 '판세(승패)'를 예측한다.

이번 알파고 제로에서는 이 정책망과 가치망을 하나의 네트워크로 구현했다. 이 구조는 두 가지 의미를 내포한다. 하나는 자신만의 바둑 이론을 하나의 신경망으로 표현했다는 것이고, 또 하나는 성능을 높이는 방식을 선택했다는 것이다. 예측 정확도는 다소 낮아지나 값 오류(value error)는 낮추고 플레이 성능은 높일 수 있게 된다. 천영재 연구원은 "딥러닝 초기에 제안된 단순한 CNN 구조에서, 비교적 최근 제안된 레스넷(ResNet)[*3]으로 네트워크 구조를 변경해 성능 개선을 얻었다"며 "아울러 하나의 네트워크에서 정책망과 가치망을 한 번에 테스트함으로써 같은 시간 내 2배 더 많은 추론(inference)이 가능해졌고, 궁극적으로 트리 탐색에서 이득을 보았다"고 분석했다. 앞선 구조 변경은 엘로 점수를 대략 600점 올릴 수 있었던 원동력 중 하나로 간주된다.

두 번째 : 무(無)에서 유(有)로의 학습

전(前) 버전의 알파고에선 15만 건의 기보(棋譜, 한판의 바둑을 두어 나간 기록)로부터 3,000만개의 수를 입력받아 지도학습(supervised leaning) 방식으로 정책망을 학습해 나갔다. 이렇게 다음 수를 예측하는 정확도를 57%까지 끌어올린 이후, 알파고는 강화학습(reinforcement learning)을 통해서 정책망과 가치망을 다듬어 나갔다. 이 단계에선 스스로 새로운 전략을 발견하고, 바둑에서 이기는 법을 학습했다.

반면, 알파고 제로는 인간이 만든 기보나 수를 전혀 학습에 사용하지 않았다. 오로지 바둑 규칙만을 가지고 자가 대국을 두며 처음부터 끝까지 인간의 도움 없이, 스스로 바둑 이치를 터득해나갔다.

인간으로부터 전혀 배운 것이 없는 알파고 제로는 인간의 선입견과 한계로부터 자유를 얻었다. 그 덕분에 자신만의 독특한 정석(공격과 수비에 최선이라고 인정되는 수를 두는, 일련의 순서)을 개발했다. 사람이라면 바둑 세계에 입문하자마자 배우는 '축'의 개념을, 알파고 제로는 정작 학습이 상당히 진행된 다음에 발견하기도 했다.

감동근 교수는 "강화학습만으로 개발한 알파고 제로는 인간과는 전혀 다른 바둑을 둘지도 모른다고 생각했으나 오히려 인간이 지난 2,500년간 찾아낸 바둑의 수법이 아주 허황한 것이 아님을 보여줬다"고 평가했다.

다만 실전에서 인간 프로기사를 이길 수 있을지에 대해서는 의견이 분분하다. 알파고 제로는 가장 간단한 바둑 규칙(Tromp-Taylor rule)으로 개발됐다. 대표적으로 실전에서는 허용된 동형반복을 학습하지 못했다. 실전에서 삼패를 만들게 된다면 인간이 알파고 제로를 가지고 놀 수 있다는 것을 함의한다. 감 교수는 "이 때문에 구글 커제와 대결이 있었던 올해 5월까지도 구글 딥마인드팀이 알파고 제로에 대해 확신을 갖지 못한 것 같다"고 추측했다.

강화학습이라고 설명하는 부분은 다소 주의 깊게 볼 필요가 있다. 알파고 제로는 자가 대국한 결과를 가지고 네트워크를 지도학습을 반복, 최종적으로 높은 성능의 네트워크를 학습한다. 이는 일반적으로 보상(reward)만을 가지고 네트워크를 학습시키는 강화학습과는 다소 차이가 있다. 강화학습이 지도학습과 대비되는 가장 큰 특징은 학습 데이터가 주어지지 않는다는 점이다.

세 번째 : 효율적인 학습과 테스트

알파고 판과 알파고 리는 각각 1,202개의 CPU와 176개의 GPU를, 1,202개의 CPU와 48개의 TPU를 분산처리해 하나의 컴퓨터처럼 묶은 뒤 대국을 진행했다. 반면, 알파고 마스터와 알파고 제로는 4개의 TPU만을 가진 컴퓨터(싱글 머신)로 경기에 임했다.

이는 알파고 마스터와 제로가 기존 인공지능 바둑에서 당연하게 받아들였던 많은 부분을 제거, 속도 개선 효과를 얻었기에 가능했다. 대표적으로, 네트워크의 입력으로, 활로(liberty)의 수, 사석의 수, 불가능한 수의 위치 등 사람이 정의한 다양한 특징(hand-crafted features)은 사용하지 않고 단지 흰돌과 검은돌의 위치 정보만을 사용했고, 바둑을 끝까지 빠르게 두어보는 롤아웃(roll-out)을 제거했다. 또한, 네트워크의 고도화로 트리탐색의 효율을 높였다. 결과적으로 CPU 자원사용이 절대적으로 줄었고, 더 적은 수의 GPU(혹은 TPU)만으로도 이전 버전의 성능을 뛰어넘을 수 있었다.

다만 감 교수는 "TPU 몇 개를 갖춰서 학습시킨 지 불과 몇 시간 만에 이전 알파고 버전과 인간을 뛰어넘었다"며 접근 방식에 대해 우려를 표했다. 알파고 제로를 기준으로 대국에는 단일 머신(4 TPU)을 활용했지만, 학습에는 64 GPU와 19 CPU를 활용한 것으로 파악된다. 이는 하나의 실험 환경에서 이같은 컴퓨팅 자원을 활용했다는 의미로, 조작변인을 조금씩 바꿔가며 수십, 수백개의 실험을 병렬로 수행하려면 많은 양의 GPU(혹은 TPU) 자원이 더 필요할 수도 있다.

그저 단순히 원점(zero base)에서 학습을 시작한 지 수십 시간 만에 알파고 제로가 이전 버전을 뛰어넘은 것은 아니라는 의미다. 어마어마한 컴퓨팅 자원과 인력을 가지고도 최적의 인자(parameter)를 찾기 위해서는 최소 수개월이 필요할 수 있다.

추가로, 엘로 점수가 인간 프로 선수보다 1500점 정도 높다고 해서 5~6점 깔아야 한다는 시각은 근거가 약하다. 감 교수는 "아마 5단인 나는 호선(互先)*4으로 승률이 50%인 상대한테 2점을 접고도 10판을 둔다면 그 중 한판은 이기리라 기대할 수는 있어도, 세계 랭킹 1위인 커제(柯潔) 9단과 한국 1위 박정환 9단이 두면 2점이 아니라 정선(定先)*5으로도 10대 0이다"라고 설명했다.

알파고 제로의 공개로 가까운 미래에 인공지능을 탑재한 기계가 인간을 지배하는 것이 아니냐는 우려가 더 커졌다. 반면 이를 제대로만 활용한다면 인류가 당면한 각종 사회문제를 해결할 키가 될 것이라는 장밋빛 미래도 그려지고 있다. 분명한 건 알파고 제로가 19×19라는 작은 바둑판 내 문제를 푸는 최강자라는 점을 부인할 수 없다는 것이다. 다만 지구상에 존재하는 문제는 이보다 더 복잡한 경우의 수로 점철되어 있다는 점이다. 알파고 제로의 탄생에 환호하긴 아직 이르다. 아직 우리 인간이 가야 할 길은 멀고 풀어야 할 문제는 더 많다.

*1 참고 | 바둑 실력을 수치화 한 점수. 엘로 점수 차이가 200점 이상인 두 AI 맞붙는다면, 점수가 높은 AI가 이길 확률은 75%이다. 366점 차이라면 90%, 677점 차이는 99%, 800점 이상의 격차인 경우, 우위의 AI가 이길 확률은 사실상 100%가 된다. *2 논문 | Silver, D. et al. (2017). Mastering the Game of Go without Human Knowledge (p.13), doi:10.1038/nature24270. *3 참고 | CNN는 얀 레쿤 교수가 1989년 개발한 구조를 토대로 한다. 2012년 ILSVRC 이미지인식 대회에서 힌튼 교수팀의 알렉스넷(AlexNet)이 놀라운 성능 개선을 보이며 CNN에서 폭발적인 연구 성장이 이어져 왔다. 이후 딥러닝이 복잡한 문제를 해결하는 열쇠라는 게 밝혀지면서 이후로도 딥러닝 연구가 이어져오고 있다. VGGNet, 구글넷(GoogLeNet), 레스넷 등이 2011년 26% 수준의 인식오차율을 3.6%까지 낮춘 CNN 개량판이다. 그 중 레스넷은 마이크로소프트가 개발한 것으로, 이미지 인식 네트워크 중에서도 인기가 많다. *4 참고 | 호선은 바둑 플레이어 간 실력이 막상막하일 경우, 돌가리기를 통해 흑백을 정한 다음 시작하는 바둑을 뜻한다. *5 참고 | 정선은 두 사람 사이 다소 실력차이가 나서 실력이 다소 떨어지는 쪽이 흑으로 먼저 시작하는 바둑을 의미한다.