Issue Report

2018. 1. 23. 제2017-009호

AlphaGo Zero의 인공지능 알고리즘

추형석 선임연구원†

SPRi 소프트웨어정책연구소

○ 본 보고서는 「과학기술정보통신부 정보통신진흥기금」을 지원받아 제작한 것으로 과학기술정보통신부의 공식의견과 다를 수 있습니다.

○ 본 보고서의 내용은 연구진의 개인 견해이며, 본 보고서와 관련한 의문사항 또는 수정・보완할 필요가 있는 경우에는 아래 연락처로 연락해 주시기 바랍니다.

- 소프트웨어정책연구소 기술·공학연구실 추형석 선임연구원(hchu@spri.kr)

《 Executive Summary 》

구글 딥마인드가 개발한 인공지능 바둑 프로그램 AlphaGo는 지난 2017년 5월 중국의 바둑신성 커제 9단과 대결에서 완승한 뒤 바둑계에서 화려하게 은퇴했다. 커제 9단과 대결했던 AlphaGo는 과거 이세돌 9단과 대결했던 AlphaGo보다 완벽에 가까울 정도로 개선됐다. 그렇다면 어떻게 개선된 것일까? 딥마인드는 커제 9단의 대국이후 개선된 형태의 AlphaGo에 대해 구체적인 내용을 공개한다고 밝혔다. 딥마인드의 최고경영자인 데미스 하사비스는 특히 개선된 AlphaGo가 인간의 기보를 전혀 학습하지 않았고, 컴퓨터 1대 수준에서 경기에 임했다는 사실이 기존과의 차별점이라고 밝히면서 대중의 궁금증을 자아냈다.

2017년 10월 세계 최고의 학술지 네이처에는 “Mastering the Game of Go without Human Knowledge”라는 제목의 논문이 게재됐다. 바로 개선된 AlphaGo의 세부 내용을 담은 AlphaGo Zero에 관한 논문이다. 사실 AlphaGo를 개선한다는 것은 매우 도전적인 영역으로 인식됐다. 그 이유는 역설적으로 AlphaGo가 사용한 인공지능 알고리즘 때문이다. 과거 AlphaGo는 전문 바둑기사의 착수 선호도 예측과 바둑판 상태의 승률을 계산하기 위해 심층학습(Deep Learning)을 활용했다. 심층학습의 가장 큰 한계는 예측한 결과에 대한 인과관계를 설명할 수 없다는 점이다. 다시 말하면, 과거 AlphaGo가 실수했던 측면의 어떤 부분이 잘못됐는지를 전혀 알 수 없다는 것이다. 그러나 AlphaGo Zero는 이러한 우려를 불식시키며 개선에 성공하고 바둑계의 최정상 자리를 차지했다.

이번 보고서에서는 AlphaGo Zero의 인공지능 알고리즘을 분석해보고자 한다. 특히 과거 AlphaGo와의 어떠한 차별점이 있는지에 대해 집중적으로 다룰 것이다. 결론적으로 AlphaGo Zero는 인간의 기보를 전혀 학습하지 않았고, 자체 대국 결과를 학습 데이터로 활용하는 방법을 시도했다. 그 결과 AlphaGo Zero는 최정상 바둑 실력을 입증했다. 수 천 년을 이어온 바둑이 약 40일 간 학습한 인공지능에 정상을 내준 것이다.

《 Executive Summary 》

AlphaGo, an artificial intelligence Go program developed by Google's Deep Mind, retired brilliantly from Go community after winning against Ke Jie in May, 2017. AlphaGo, which confronted Ke Jie, was improved to be closer to perfection than the AlphaGo, which confronted Lee Se-dol. So how did it improve? Deep Mind announced that it will release specific details of the improved version of AlphaGo since competition with Ke Jie. Deep Mind CEO Chief Executive Demis Hassabis said that the fact that the improved AlphaGo did not learn human knowledge at all and that the game was played at the level of one computer was a distinction from the past.

In October, 2017, the world's leading journal Nature published a paper entitled "Mastering the Game of Go without Human Knowledge." It is an article on AlphaGo Zero which contains details of the improved AlphaGo. In fact, improving AlphaGo was seen as a very challenging area. The reason is paradoxically because of the AI algorithm used by AlphaGo. In the past, AlphaGo used deep learning to calculate the winning rate and preference positions of professional Go player. The greatest limitation of deep learning is that it can not account for the causal relationship between predicted results. In other words, in the past, AlphaGo never knew what went wrong in order to make up the mistake. However, AlphaGo Zero succeeded in improving the situation by eliminating these concerns and took the top spot in the Go.

In this report, I try to analyze AlphaGo Zero's artificial intelligence algorithm. In particular, I will focus on what differentiates AlphaGo from the past. In conclusion, AlphaGo Zero has not trained human Go data, but tried to use self-play data for training. AlphaGo Zero proved to be the best player in the league with its excellent results. Thousands of years passed Go history gave the summit to the learned artificial intelligence for about 40 days.

《 목 차 》

《 Contents 》

1. 배 경

지난 2016년 3월 이세돌 9단과 대결하여 4:1로 승리한 AlphaGo는 인류 역사의 큰 획을 긋는 성과로 기억됐다. 당시 AlphaGo는 완벽하지 않은 프로그램이었다. 의미 없는 수를 두거나, 계가 시점에서 실수를 하는 등 개선의 여지가 남아있었기 때문이다. 또한 슈퍼컴퓨터 급의 계산 장비를 활용했다는 점에서 형평성의 논란도 있었다. 그러나 AlphaGo가 보여준 바둑 실력은 이미 최정상임을 입증한 상황이었기 때문에, 더 성능을 높인다는 것은 도전의 영역이었다.

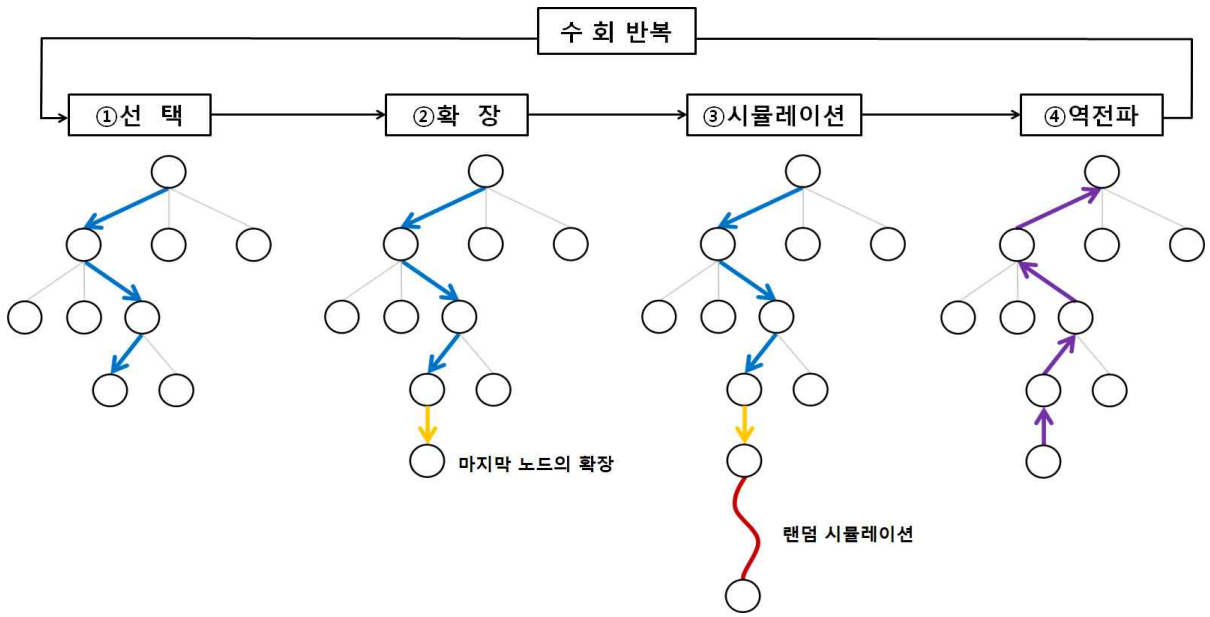
이세돌 9단과의 대결이후 AlphaGo의 성능을 향상시킨다는 점은 요원할 것으로 예측됐다. 가장 큰 이유는 역설적으로 AlphaGo의 알고리즘 때문이다. AlphaGo는 전문 바둑기사의 착수 선호도 예측과 현재 바둑판 상태의 승리할 확률을 예측하기 위해 인공신경망(Artificial Neural Network)기술을 활용했다. AlphaGo의 인공신경망은 이미지 분류에 최적화된 합성곱신경망(Convolutional Neural Network)으로, 현재 바둑판을 48가지 특징(Feature map)으로 세분화하여 학습에 활용했다. 인공신경망 기술은 신경계의 정보전달 기능을 모사한 방법론으로 현대 심층신경망(Deep Neural Network)의 모태가 된다. 인공신경세포(Artificial Neuron)를 여러 층 쌓은 구조인 심층신경망은 복잡한 비선형적인 현상을 분류하고 예측하는데 활용된다. 그러나 심층신경망에는 가장 큰 단점이 존재한다. 바로 입력과 출력의 상관관계를 설명할 수 없다는 점이다. 다시 말하자면 AlphaGo가 범했던 실수에 대한 원인이 무엇인지 규명해내기 어렵다는 것이다.

AlphaGo 개발 회사인 딥마인드(DeepMind)는 이러한 상황을 개선하기 위해 탐색 알고리즘인 몬테-카를로 트리 탐색(Monte-Carlo Tree Search, MCTS)을 향상시켰다. MCTS의 개념은 탐색 알고리즘으로부터 출발한다.[1] 탐색 알고리즘은 기존 바둑 인공지능에서 가장 널리 활용되는 방법이다. 탐색 알고리즘의 기능은 수를 읽는 행위와 비슷하게 볼 수 있다. 바로 현재 바둑판 상태에서 승리할 수 있는 수를 예측하는 것이다. 바둑 게임을 비롯한 보드 게임(board game)[2]은 구조상 게임의 진행상황을 트리(Tree)형태로 표현할 수 있다. 따라서 승리할 확률이 매우 높은 수를 찾는 것은 트리를 탐색하여 최적의 경로를 찾는 행위로 볼 수 있다.

1) 자세한 내용은『AlphaGo의 인공지능 알고리즘』소프트웨어정책연구소(2016) 이슈리포트 참고
2) 경기자 두 명이 서로 번갈아 가며 행위를 하는 게임으로 바둑, 체스, 장기 등이 해당함.

그러나 현재 바둑판 상태에서 승리할 확률을 정확하게 아는 것은 매우 어렵다. 바둑 경기는 경우의 수가 거의 무한대에 가깝기 때문이다. AlphaGo는 특정 바둑판 상태의 승률을 정확하게 근사하기 위해 인공신경망 알고리즘을 활용했다는 점이 기존 접근법과 차별점이다. 인공신경망 기술 역시 만능은 아니기 때문에, 간단한 규칙에 의거한 임의(random) 시뮬레이션과 그 결과를 되먹임(feedback)하여 승률 예측 성능을 향상시켰다. 이 접근 방법이 MCTS로 일반적인 트리 탐색(Tree Search)과 구분된다. 몬테-카를로는 일종의 임의 시뮬레이션을 나타내는 것으로 이해할 수 있다.

몬테-카를로 트리 탐색(MCTS) 알고리즘



[그림 1] MCTS 알고리즘의 개념도

<바둑 게임에서의 MCTS의 진행 방식>

①선택 : 현재 바둑판 상태에서 특정 경로로 수읽기를 진행

②확장 : 일정 수 이상 수읽기가 진행되면 그 지점에서 한 단계 더 착수 지점을 예측(게임 트리의 확장)

③시뮬레이션 : ②에서 선택한 노드에서 바둑이 종료될 때까지 고속 시뮬레이션 수행. 속도가 빠르기 때문에 여러 번 수행할 수 있으나 착수의 적정성은 떨어짐

④역전파 : ③의 결과를 종합하여 확장한 노드의 가치(②에서 한 단계 더 착수한 것의 승산)를 역전파하여 해당경로의 승산 가능성을 갱신

딥마인드는 이세돌 9단과의 대결 이후에 AlphaGo의 MCTS를 개선시키는 데 집중했다. 이세돌 9단과 대결한 AlphaGo(이하 AlphaGo Lee)는 16만 개의 전문 바둑 기보에서 추출한 2,940만 개의 바둑판 상태를 학습하여 착수 선호도를 예측하고, MCTS를 활용해 대국을 진행했다. AlphaGo Master는 AlphaGo Lee와 유사하게 전문 바둑기사의 기보를 바탕으로 학습한 인공신경망을 활용한다. AlphaGo Master는 AlphaGo Lee의 인공신경망을 개선하기 위해 자체 대국의 MCTS에서 산출한 착수선호도를 추가적으로 학습한 것이 차별점이다. 다시 말하자면 AlphaGo Lee에서 임의 시뮬레이션은 단순히 트리 탐색에 활용됐으나, AlphaGo Master에서는 시뮬레이션 결과를 학습하는 부분에도 활용했다.

AlphaGo Master는 이세돌 9단과의 대국 이후 온라인 바둑 사이트 Tygem에 'Master'라는 아이디로 60전 전승(2017년 1월 기준)을 거뒀다. 전 세계의 쟁쟁한 9단 기사들과의 대국에서 모두 승리했다는 사실은 앞서 기술한 알고리즘의 개선이 상당한 성능향상을 달성했다는 것이다. 대국에 참여했던 바둑 기사들은 AlphaGo Master가 인간이 바둑을 두는 틀에서 벗어났고, 인공지능의 바둑 실력이 인간을 완전히 뛰어넘을 가능성이 높다고 평했다. 또한 지난 2017년 5월 개최된 'Future of Go Summit'에서 AlphaGo Master는 중국의 바둑 신성 커제 9단과 대결하여 3:0으로 승리했다. 딥마인드는 AlphaGo Lee에서 제기됐던 문제를 해결하기 위해 인공신경망 기술에서 해결책을 찾기보다 전체적인 알고리즘의 변화를 모색했다. 인과관계가 불분명한 인공신경망을 개선하는 것은 모래사막에서 바늘을 찾는 것만큼 어려운 일이다. 인공신경망의 개선은 수많은 모수[3]를 미세조정하며 경험적으로 추정하는 것이 일반적이기 때문이다.

그러나 AlphaGo Master 역시 전문 바둑기사의 기보에 의존한 결과였다. AlphaGo Master는 MCTS의 자체 대국 결과를 인공신경망에 학습시켰다는 점이 차별점이나, 인공신경망은 과거 인간의 기보를 학습한 정보로 구축된 것이기 때문이다. 그러나 딥마인드의 최고경영자 데미스 하사비스는 2017년 1월 한 국제학회에서 인간의 기보를 전혀 학습하지 않은 AlphaGo를 개발 중이라고 밝혔다. AlphaGo Master를 뛰어넘는 AlphaGo의 내용은 지난 2017년 10월 세계 최고의 학술지인 네이처에 소개됐다. 이것이 이번 보고서에서 다루고자 하는 AlphaGo Zero다. 2장에서는 AlphaGo Zero의 특징을 살펴보고, 3장에서는 AlphaGo Zero의 인공지능 알고리즘에 대해서 소개한다.

3) 학습률, 은닉층의 개수, 합성곱신경망의 크기, 활성함수의 선택 등 많은 모수가 존재

2. AlphaGo Zero의 특징

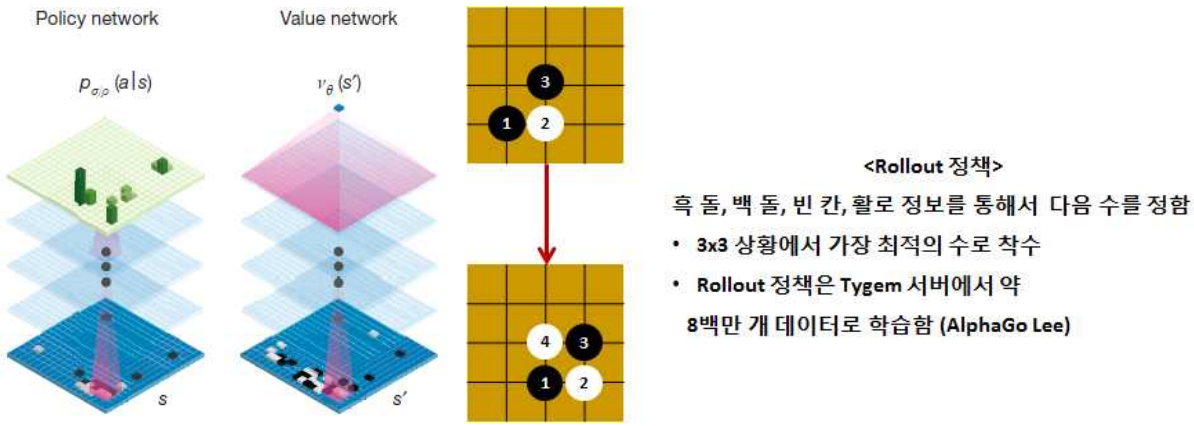
AlphaGo Zero의 상세 내용은 2017년 10월 네이처에 발간된 ‘Mastering the Game of Go without Human Knowledge’에서 확인할 수 있다. 먼저 제목에서도 알아낼 수 있듯이 AlphaGo Zero는 인간의 바둑기보를 전혀 활용하지 않고 개발한 바둑 인공지능이다.

AlphaGo Zero가 기존 AlphaGo Lee(이세돌 9단과 대결), 혹은 AlphaGo Master(커제 9단과 대결)와의 **차별점은 크게 네 가지**가 있다. 먼저 ①무작위 방식(random play)의 자체 대국을 통해 강화학습(reinforcement learning)을 활용한 점이다. 과거 두 가지 버전의 AlphaGo는 약 16만 개의 전문 바둑 기보를 학습한 반면, AlphaGo Zero는 스스로 바둑을 두면서 학습하는 접근법을 선택했다. ②두 번째 차별점은 인공신경망 학습에 활용된 특징 맵(feature map)이다. AlphaGo Zero는 흑돌과 백돌 두 가지만을 활용했다. 기존 AlphaGo는 흑돌, 백돌을 포함한 눈, 활로, 꼬부림 등 48가지 특징을 활용했다는 점을 상기하면, AlphaGo Zero는 오히려 더 단순한 접근방법을 취했다.

③다음으로 인공신경망의 형태와 구조의 변화다. AlphaGo Lee는 [그림 2 (좌)]와 같이 두 가지 형태의 인공신경망을 활용했다. 정책망(Policy Network)은 전문 바둑기사의 착수 선호도를 학습한 인공신경망이다. 특정 바둑판 상태가 정책 네트워크에 입력으로 들어가게 되면, 출력으로 규칙상 착수 가능한 모든 수에 대해서 착수 선호도(확률 값)가 산출된다. 이 착수 선호도는 MCTS에서 트리를 확장하는 시점에 활용된다. 정책망의 기능은 게임 트리의 폭을 줄이는 용도로, 전문 바둑기사의 관점에서 착수할 지점을 선별하는 역할을 한다. 가치망(Value Network)은 현재 바둑판 상태에서 승리할 확률을 찾아내는 역할을 한다. 앞서 기술했듯이 바둑은 거의 무한대 경우의 수를 가지고 있기 때문에, 특정 바둑판 상태의 승률을 정확히 예측하기는 어렵다. 따라서 승률을 예측하기 위한 접근을 취할 수밖에 없다. AlphaGo Lee는 가치망을 학습하기 위해 자체 대국으로 생성된 3,000만 개의 바둑 판 상태를 활용했다. 가치망의 기능은 게임 트리의 깊이를 줄이는 역할을 한다. 정리하자면, AlphaGo Lee는 정책망과 가치망이라는 두 가지 인공신경망을 활용했다. 반면, AlphaGo Zero는 두 가지 신경망을 하나로 합쳤다. AlphaGo Lee는 두 가지 신경망을 따로 학습하는 전략을 택한 반면, 정책

망과 가치망의 역할이 궁극적으로 게임에서 승리하기 위한 방법이기 때문에 AlphaGo Zero는 하나로 합쳐진 형태를 활용한 것으로 추정된다. 인공신경망의 구조 역시 변경됐다. AlphaGo Zero는 ILSVRC[4] 2015의 우승팀인 마이크로소프트가 개발한 잔차신경망(Residual Network)을 활용했다. 잔차신경망은 이미지 분류 문제를 95%의 정확도로 해결한 것으로 이미지에서 패턴을 인식하는데 가장 성능이 좋은 것으로 알려져 있다.

④마지막 차별점은 트리 탐색 기법을 개선했다. 기존 AlphaGo의 MCTS는 랜덤 시뮬레이션 단계에서 롤아웃(rollout) 정책[5]을 활용했다. 롤아웃은 정책망과 유사한 역할을 하나, 빠른 속도로 시뮬레이션하기 위해 고안된 간단한 정책이다. 롤아웃은 [그림 2 (우)] 과 같이 바로 직전 착수한 지점을 중심으로 3x3 바둑판에 간단한 규칙에 의해 바둑돌을 신속하게 착수하는 과정으로 진행된다. 이후 게임의 결과를 바둑 트리의 노드 값에 되먹임하여, 착수 전략을 향상시킨다. 이 롤아웃 정책은 복잡한 바둑게임을 지나치게 단순화한 경향이 있었기 때문에, 이 부분이 기존 AlphaGo의 실수의 원인 중 하나라고도 볼 수 있다. 따라서 AlphaGo Zero에서는 롤아웃을 활용하지 않고, 랜덤 시뮬레이션의 정책으로 앞서 기술한 잔차신경망을 적용했다. 인공신경망 기술 기반의 정책은 롤아웃보다 훨씬 많은 계산을 요구하지만 정확도가 향상되는 장점이 있다.



[그림 2] 정책망과 가치망의 개념(좌)과 롤아웃 정책(우)

자료 : Mastering the game of Go with Deep neural networks and tree search, Nature(2016)

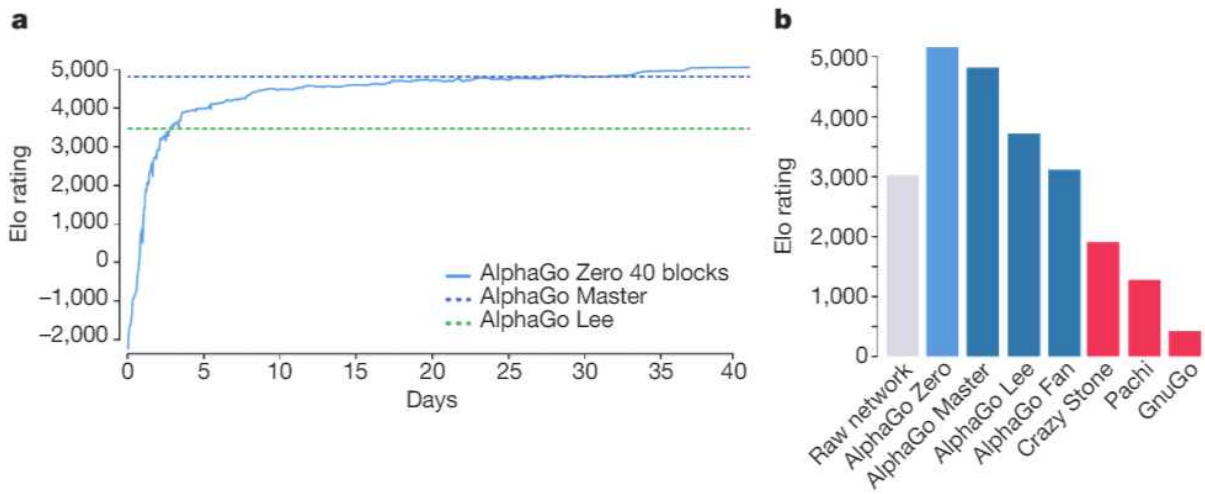
4) Imagenet Large Scale Visual Recognition Challenge는 지난 2010년부터 개최된 경진대회로 수 천만장의 이미지(Imagenet 데이터베이스)에서 객체를 적절하게 분류하는 임무를 수행함.
5) 여기서 정책의 의미는 착수를 결정하는 규칙으로 해석 가능함.

잔차신경망(Residual Network)

- 잔차신경망은 2015년 ILSVRC(Imagenet Large Scale Visual Recognition Challenge)의 우승팀인 마이크로소프트 연구진이 개발한 인공신경망으로 이미지 인식에 특화
  - 이미지 분류에서 3.57%의 오차율로 우승
  - 152층의 합성곱신경망 층을 활용
- 인공신경망의 은닉층의 개수가 많아질수록 성능이 개선될 여지가 있었으나, 기술적으로 30층 이상에서는 성능이 정체되는 현상이 발견되어, 이를 개선하기 위한 방법으로 잔차신경망이 소개됨
- 잔차신경망은 구조적으로 동일한 성능을 갖는 신경망을 중복하여 연결하고, 그 사이에 얕은 신경망을 추가함으로써 성능향상을 도모함

이것으로 AlphaGo Zero의 네 가지 차별점을 살펴봤다. 요약하자면 AlphaGo Zero는 인간의 기보를 전혀 사용하지 않고 무작위 대국을 통해 학습했다. 대국 상황의 분류 성능을 극대화하기 위해 최신 잔차신경망 구조를 활용했으며, 과거 정책망과 가치망의 기능을 하나로 합쳤다. 또한 일원화된 탐색 전략으로 성능향상을 도모했다. 그러나 지금까지의 내용을 토대로 보면, 아직도 AlphaGo Zero가 '왜 인간을 넘어서는 성능을 갖게 됐는지' 쉽게 이해되지는 않는다. 기술적인 차별점은 있으나 왜 이러한 차별점으로 인해 성능이 향상됐는가는 여전히 의문으로 남는다. 그 근거에 대해서는 3장에서 더 세부적으로 살펴볼 것이다.

AlphaGo Zero의 성능은 언론에도 많이 보도됐듯이 무작위 대국을 토대로 학습하여 3일 만에 학습했던 AlphaGo Lee의 성능을 뛰어넘었다. 3일 간 AlphaGo Zero는 약 490만 자체 대국을 수행하여 바둑 지식을 학습했다. AlphaGo Zero의 자체 대국 착수는 약 0.4초 마다 이루어졌으며, 이 시간동안 1,600회의 MCTS를 계산했다. 최종적으로 AlphaGo Zero는 40일 간의 학습을 통해 AlphaGo Master를 능가했다. 이 기간 동안 AlphaGo Zero는 2,900만 번의 자체 대국을 수행했다. AlphaGo Zero와 기존 AlphaGo와의 성능은 다음 [그림 3]과 같고, AlphaGo Lee, Master, Zero의 차별점은 <표 1>과 같다.



[그림 3] AlphaGo Zero의 학습 기간과 성능

a. 40일간 자체 대국으로 학습한 AlphaGo Zero의 성능

b. 다양한 바둑 인공지능 프로그램 간의 성능 비교

자료 : Mastering the game of Go without human knowledge, Nature(2017)

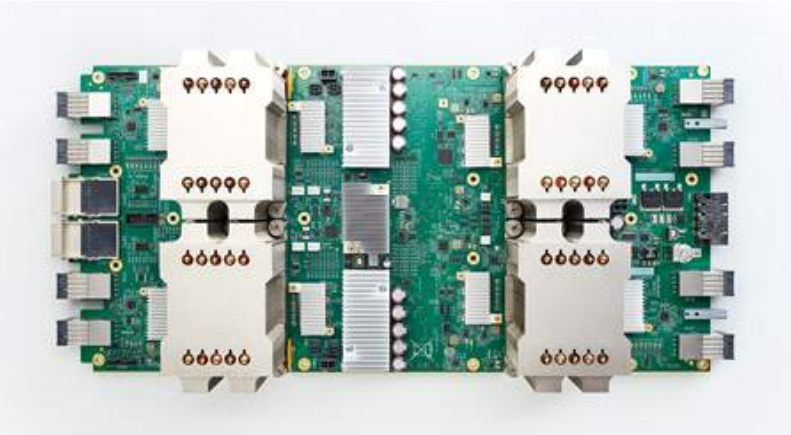
<표 1> AlphaGo의 버전별 차별점

버 전	차 별 점
AlphaGo Lee	- 최초로 프로 9단 바둑기사와 대결하여 승리 - 정책망과 가치망 학습에 인간의 기보 16만 개 활용
AlphaGo Master	- 바둑 온라인 경기 사이트 Tygem에서 60전 전승 - MCTS의 자체대국 결과를 정책망과 가치망 학습에 활용 (AlphaGo Lee에서 활용된 신경망을 초기값으로 활용)
AlphaGo Zero	- AlphaGo Master와 대국하여 89:11로 압도적 승리 - AlphaGo Master와 같은 방식으로 정책망과 가치망을 학습했으나, 인간의 기보를 전혀 활용하지 않음

특히 AlphaGo Zero는 TPU(Tensorflow Processing Unit, 그림 4) 네 장을 활용하여 달성했다는 점이 주목할 부분이다. TPU는 구글이 개발한 연산처리장치로, 인공신경망 연산에 최적화되어 있다. 현대 인공지능 컴퓨팅 인프라로 주목받고 있는 연산처리장치는 GPU(Graphical Processing Unit)가 대표적이나, 전력 소비가 상대적으로 높은 편이다. TPU는 인공신경망에 필요한 연산만을 처리하기 위한 구조를 구현하여 GPU대비 최대 80배의 전력 절감 효과를 달성했다.[6] 또한 TPU

6) Jouppi, Norman P., et al. "In-datacenter performance analysis of a tensor processing unit." arXiv preprint arXiv:1704.04760 (2017).

네 장의 구성은 고성능 PC 1대 수준의 전력으로도 구동하기 때문에, 슈퍼컴퓨터급 자원을 활용한 AlphaGo Lee와의 형평성 문제도 해결했다. 이 부분 역시 지난 2017년 1월 데미스 하사비스가 차세대 AlphaGo를 개발하기 위해 중점을 두었던 사항이다. 저전력과 고성능의 두 마리 토끼를 모두 잡기에는 매우 요원한 일이라 여겨졌지만, 두 가지를 모두 해결한 딥마인드의 저력이 놀라울 따름이다.



[그림 4] Tensorflow Processing Unit (TPU)

자료: Build and train machine learning models on our new Google Cloud TPUs, Jeff Dean (2017)

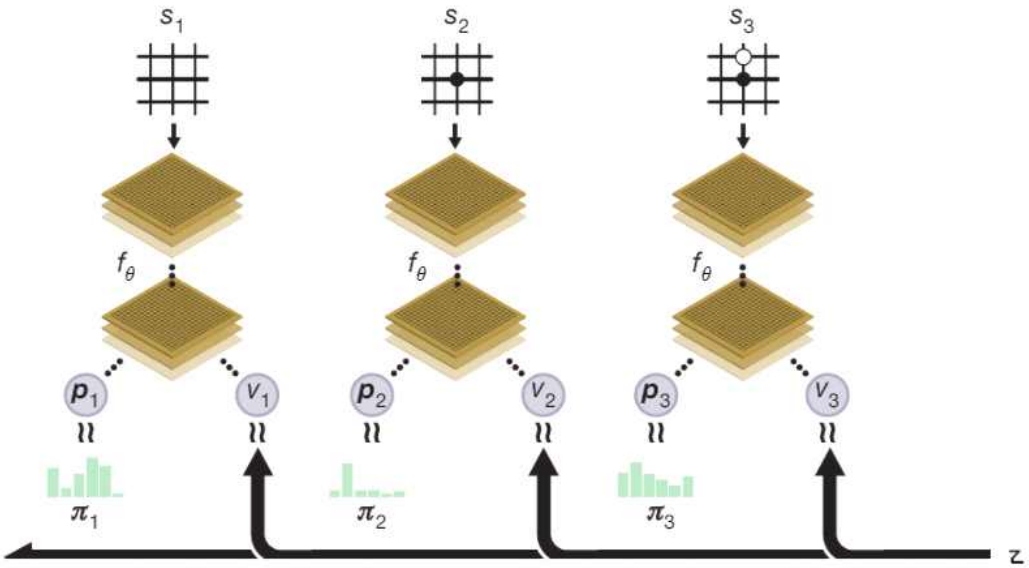
3. AlphaGo Zero의 인공지능 알고리즘

AlphaGo Zero는 자체 대국 결과를 기반으로 학습했다. 자체 대국 알고리즘에는 인간의 기보가 전혀 활용되지 않았고, 전적으로 바둑규칙만을 적용했다. AlphaGo Zero에는 여전히 심층학습 기술이 활용됐다. 일반적으로 심층학습은 많은 데이터를 요구하는데, AlphaGo Zero는 이 데이터를 자체 대국으로 생산하여 학습했다.

AlphaGo Zero의 자체 대국은 이미 알려졌다시피 무작위 접근으로부터 출발한다. 자체 대국에서 착수를 결정하는 인공신경망 역시 무작위로 초기화 된다. AlphaGo Zero는 자체 대국에서 한 번의 착수를 위해 1,600번의 MCTS 탐색을 실시한다. 바둑 게임 트리를 확장하는 방법은 앞서 기술한 잔차신경망이 활용되고, 가장 초기 단계에는 잔차신경망을 구성하는 모수가 모두 무작위로 초기화된다. 1,600번의 MCTS 탐색은 1,600번의 수를 읽는 것으로 이해할 수 있으며, 인공신경망의 학습이 성공적으로 이루어질수록 1,600번의 수읽기가 더 정교해진다는 것을 의미한다. 수읽기가 정교해질수록 AlphaGo Zero의 바둑 실력은 향상된다고 볼 수 있다.

그렇다면 AlphaGo Zero의 인공신경망은 어떠한 방법으로 학습된 것일까? 그 해답은 1장에서 기술했던 AlphaGo Master에서 찾아볼 수 있다. AlphaGo Master는 MCTS를 통한 자체 대국의 결과를 인공신경망 학습에 활용했던 것이 AlphaGo Lee와의 차별점이었다. MCTS를 통한 자체 대국은 일종의 시뮬레이션과 같기 때문에 결과물로 기보가 나온다. AlphaGo Master는 MCTS 자체 대국 기보를 학습하는 접근으로 성능향상에 성공했다. **AlphaGo Zero는 MCTS 자체 대국 결과를 학습했다는 점에서 AlphaGo Master와 동일한 접근을 취하지만, AlphaGo Zero는 MCTS 시뮬레이션에서 바둑규칙 이외에 어떠한 인간의 지식이 포함되지 않았다는 것이다.** AlphaGo Zero는 약 3일 간 490만 자체 대국을 학습하여 진화했다. 490만 자체 대국은 각각의 바둑판 상태로 분할되어, 인공신경망 학습은 2,048개 바둑판 상태를 단위로 총 70만 번을 수행했다.[7] 학습에 활용된 하드웨어는 GPU 64개와 CPU 19개다. 4장의 TPU를 활용한 것은 학습이 완료된 버전(inference)을 토대로 대국을 수행하는데 활용됐다. 지금까지 기술한 AlphaGo Zero 알고리즘의 학습 과정을 표현하면 [그림 5]와 같다.

7) 490만 자체 대국이 평균적으로 약 292수 까지 진행된 것으로 분석됨



[그림 5] AlphaGo Zero의 학습 과정

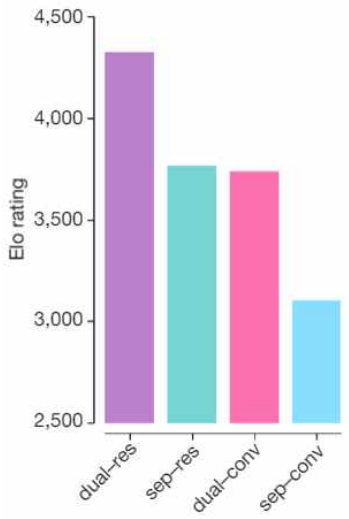
s_t : 바둑판의 상태, f_θ : AlphaGo Zero의 인공신경망

p : 인공신경망의 결과값 - 착수선호도, π : MCTS의 착수선호도

v : 인공신경망의 결과값 - 승리 확률, z : MCTS 시뮬레이션 결과 (승패여부)

자료 : Mastering the game of Go without human knowledge, Nature(2017)

인공신경망의 형태와 구조 변화도 AlphaGo Zero의 바둑실력 향상에 기여했다. 정책망과 가치망을 합친 인공신경망 형태는 이미 AlphaGo Master에서 구현한 결과였다. 정책망과 가치망의 역할은 궁극적으로 바둑 게임에서 승리하기 위한 도구이기 때문에, 이를 합치는 방향은 적절한 접근이라고 분석된다. 그러나 AlphaGo Master의 인공신경망의 초기값은 무작위로 초기화 된 것이 아닌, 인간의 기보를 통해 학습된 것을 활용했다는 점이 AlphaGo Zero와의 차이점이다. 이 부분이 시사하는 바는 인간의 바둑 지식이 편향되어 있다고도 추론할 수 있다. 수 천년을 이어온 바둑 격언이 오히려 무한대의 경우의 수에서 오는 바둑의 다양성을 제한했다고도 볼 수 있기 때문이다. 또한 AlphaGo Zero는 이미지 인식에 가장 좋은 성능을 보유한 잔차신경망을 사용하여 바둑 실력을 더 향상시켰다. 딥마인드는 다음 [그림 6]과 같이 다양한 형태와 구조의 시뮬레이션을 통해 궁극적으로 앞서 소개한 AlphaGo Zero의 인공신경망 형태를 활용했다.



[그림 6] 다양한 인공신경망 형태와 구조의 비교

dual-res : 정책망과 가치망이 하나로 합쳐진 잔차신경망 (AlphaGo Zero)
sep-res : 정책망과 가치망을 따로 분리한 잔차신경망
dual-conv : 정책망과 가치망이 하나로 합쳐진 합성곱신경망
sep-conv : 정책망과 가치망을 따로 분리한 합성곱신경망 (AlphaGo Lee)
자료 : Mastering the game of Go without human knowledge, Nature(2017)

지금까지 AlphaGo Zero의 인공지능 알고리즘에 대해 살펴봤다. 그러나 여전히 AlphaGo Zero가 기존 AlphaGo를 뛰어넘는 성능을 갖게 된지에 대한 명쾌한 해답은 여전히 내리기 힘들다. MCTS와 인공신경망 알고리즘의 변화 등 기술적인 차별점은 존재하나, 알고리즘의 변화로 인한 성능 향상은 직관적이고 경험적인 결과라는 점이기 때문이다. AlphaGo Zero의 알고리즘은 대부분 AlphaGo Master에서 적용된 것을 차용했다. 다른 점은 단지 인간의 기보를 전혀 사용하지 않았다는 것이다. 딥마인드는 AlphaGo Zero의 접근을 일반화하기 위해 지난 2017년 12월 5일 'AlphaZero'에 대한 논문을 arXiv에 게재했다.[8] AlphaZero는 바둑과 유사한 보드 게임인 체스와 쇼기(일본식 장기)에서 AlphaGo Zero와 동일하게 인간의 대국을 전혀 활용하지 않는 방법을 적용하여 학습했다. AlphaZero는 기존 인공지능 프로그램과 대결하여 압도적으로 승리하여 그 성능을 입증했다.

8) Silver, David, et al. "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm." arXiv preprint arXiv:1712.01815 (2017).

4. 결 론

AlphaGo Zero는 인간의 기보를 학습하지 않았다는 점에서 가장 큰 관심을 받았다. 최근 인공지능의 큰 흐름을 주도하고 있는 심층학습(Deep Learning)은 필연적으로 대규모 데이터를 요구하기 때문이다. 심층신경망을 수학적으로 표현하자면 수백만 개의 미지수를 갖는 비선형 함수다. 수백만 개의 미지수를 적절하게 추정하기 위해 필요한 데이터는 상식적으로 미지수보다는 많아야 한다. 결국 성공적인 심층학습을 위해서는 양질의 데이터 공급이 우선적이라고 볼 수 있다. AlphaGo Zero는 양질의 데이터 공급처를 자체 대국에서 찾은 것이 지금까지 일반화된 접근법과의 차별점이다. 딥마인드는 AlphaGo Zero에 이어지는 AlphaZero를 통해 규칙으로 데이터를 생산하여 학습한다는 메커니즘을 증명했다.

또한 무작위 접근이 오히려 성능향상에 긍정적인 역할을 한 것으로 추정된다. 과거 AlphaGo가 학습했던 인간의 기보는 수 천 년의 바둑 격언을 바탕으로 한다. 그러나 바둑의 경우의 수는 인간이 인지할 수 있는 것보다 훨씬 많다. 그간 누적된 인류의 바둑 지식은 분명 바둑의 정수가 담겨있으나, 편향이 존재할 가능성이 높다고 볼 수 있다. 학습 방법과 MCTS 구조가 모두 동일한 AlphaGo Zero와 AlphaGo Master의 차이점은 인간의 기보를 학습했는지에 대한 여부다. AlphaGo Zero가 AlphaGo Master를 89대 11로 물리쳤다는 점에서, 전문 바둑기사의 기보가 편향이 있다는 사실을 간접적으로 추론할 수 있다.

그러나 AlphaGo Zero의 접근법이 성공했는지에 대한 논리적인 근거는 여전히 불분명하다. 기술적인 알고리즘의 개선을 통해 성능이 향상됐다는 점일 뿐이다. 결국 AlphaGo Zero의 성능은 직관적이고 경험적인 결과에 의거한 것이라고 볼 수 있다. 이것은 인공신경망의 특성과도 부합하는 것이다. 인공신경망 역시 현상과 패턴을 분류하는데 높은 성능을 가지고 있으나, 왜 잘하는지에 대한 인과관계가 명확하지 않기 때문이다. 이를 해결하기 위해 설명 가능한 인공지능(Explainable AI)에 대한 연구가 활발히 추진 중이다. 미국의 방위고등연구계획국(DARPA)는 2018년 설명 가능한 인공지능 연구에 수 천만 달러의 연구비를 투입할 예정이다.

AlphaGo Zero가 보여준 성과는 지난 이세돌 9단과의 대국만큼 충격적인 결과다. AlphaGo가 공개된 지 채 2년이 되기 전에, 딥마인드는 인공지능 역사에 다시 한 번 큰 획을 그었다. 인공지능의 진화 속도는 우리가 생각했던 것 보다 매우 빠를수도 있다.

[참고문헌]

1. 국외문헌

Silver, David, et al. “Mastering the game of Go with deep neural networks and tree search.” Nature 529.7587 (2016): 484-489.

Silver, David, et al. “Mastering the game of go without human knowledge.” Nature 550.7676 (2017): 354.

He, Kaiming, et al. “Deep residual learning for image recognition.“ Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

2. 국내문헌

AlphaGo의 인공지능 알고리즘 분석, 이슈리포트 2016-002, 소프트웨어정책연구소 (2016)

게임인공지능 동향, SPRi SW산업 동향, 소프트웨어정책연구소(2016)

알파고 세계 바둑계를 정복하다, SPRi SW산업 동향, 소프트웨어정책연구소(2017)

알파고 제로, 인공지능의 새 길을 열다, SPRi 칼럼, 소프트웨어정책연구소(2017)

주 의

1. 이 보고서는 소프트웨어정책연구소에서 수행한 연구보고서입니다.
2. 이 보고서의 내용을 발표할 때에는 반드시 소프트웨어정책연구소에서 수행한 연구결과임을 밝혀야 합니다.