Issue Report

2020. 02. 26. IS-091

알파스타(AlphaStar)의 AI 알고리즘

추형석 선임연구원†

(hchu@spri.kr)

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 AI정책연구팀 추형석 선임연구원(hchu@spri.kr)

《 Executive Summary 》

알파고 개발진으로 유명세를 탄 딥마인드(DeepMind)는 지난 2019년 10월 30일 세계 최고의 학술지 네이처에 스타크래프트2 인공지능(AI)인 알파스타(AlphaStar) 논문을 발표했다. 알파스타는 2019년 1월 처음으로 딥마인드 홈페이지를 통해 대중에 알려졌으며, 당시 정상급 프로게이머와 대결해 10승 1패를 거둬 큰 이슈가 됐다. 그러나 당시에는 알파스타의 세부적인 내용이 공개되지 않아 알파스타가 어떻게 프로게이머 수준에 등극했는지에 대해 제한적으로 이해할 수 있었다.

이 보고서는 네이처지에 실린 알파스타의 AI 알고리즘을 보다 쉽게 전달하고자 한다. 스타크래프트2라는 게임은 바둑과는 또 다른 차원의 지능적 행동을 요구한다. 서로 상성이 존재하는 전략, 불완전한 정보, 실시간 조작, 장기 계획 등 프로게이머 수준의 스타크래프트2 AI을 개발한다는 것은 AI 분야의 또 다른 그랜드 챌린지이기 때문이다. 알파스타의 개발 과정 역시 순탄치는 않았는데, 연구 초기에 공개한 결과는 일반인 수준정도였기 때문이다. 딥마인드는 성능을 향상시키기 위해 스타크래프트2 AI 학습을 수월하게 시킬 수 있는 학습 도구와 데이터를 대중에 공개하여 연구의 참여를 유도했다.

알파스타는 복잡한 스타크래프트2 AI를 구현하기 위해 10여 개의 AI 알고리즘을 활용했다. 알파스타의 학습 과정은 알파고와 매우 유사한데, 학습에 활용된 AI 모델은 매우 상이하다. 이 보고서는 쉬운 이해를 돕기 위해 바둑과 스타크래프트2, 알파고와 알파스타의 차이점을 바탕으로, 알파스타가 해결하고자 하는 도전과제에 대해 구체적으로 설명할 것이다. 이어서 알파스타 개괄적인 흐름을 지도학습-강화학습-리그학습으로 구분하여 설명하고, 각 단계가 어떠한 의미를 갖는지에 대해 소개하고자 한다.

알파스타는 91.7만 건의 리플레이 데이터를 활용해 44일간 학습하여 스타크래프트2 상위 0.2%에 도달했다. 알파스타를 접한 많은 전문가들은 알파스타가 AI 분야의 또 다른 그랜드 챌린지를 해결했다며 호평하는 한편, 바둑과 같이 정복한 것은 아니라고 밝혔다. 알파스타는 세계 챔피언을 노리기에 아직 보완이 필요하다는 것이 중론이지만, 딥마인드가 다시 한 번 우리나라를 방문해 세계 최고 수준의 국내 스타크래프트2 프로게이머와의 대결이 성사된다면, 다시 한 번 AI의 힘을 전 세계에 알리게 될 계기가 될 것이다.

《 Executive Summary 》

DeepMind, famous for AlphaGo, unveiled a study that would surprise the world once again on October 30, 2019. It was about AlphaStar, a StarCraft II AI. In fact, this result has already been foreseen. In January 2019, DeepMind presented its first StarCraft II AI, named AlphaStar, and announced that it had won 10-1 against professional gamers. At that time, however, the details of Alphastar were not disclosed, giving a limited understanding of how Alphastar became pro-gamer.

This report aims to provide a easier explanation of AlphaStar's AI algorithms published in Nature. A game called StarCraft II requires a different level of intelligence than Go. Developing StarCraft II AI at the pro-gamer level, including congruent strategies, incomplete information, real-time manipulation, and long-term planning, was another grand challenge in AI. The development process of AlphaStar was also not smooth, as the results of the initial research were not good enough. For this purpose, DeepMind encouraged participation in research by opening up learning tools and data to the public to facilitate StarCraft II AI learning.

AlphaStar has used over 10 AI algorithms to implement complex StarCraft II AI. AlphaStar's learning process is very similar to AlphaGo, but the AI models used for learning are very different. This report will elaborate on the challenges AlphaStar is trying to solve, based on the differences between Go and StarCraft II, AlphaGo and AlphaStar, for easy understanding. Next, I will explain the general flow of AlphaStar into supervised learning, reinforcement learning, and league learning, and introduce what each stage means.

AlphaStar learned about 44 days using 91.7 million replays to reach the top 0.2% of StarCraft II. Many experts who have encountered AlphaStar praised AlphaStar for solving another grand challenge in AI, but said it was not conquered as Go. Although AlphaStar is still in need of a supplement to become a world champion, if DeepMind visits Korea again and confronts the world's best Korean StarCraft II pro-gamer, the power of AI will be proven once again.

《 목 차 》

《 Contents 》

1. 서 론

구글 딥마인드는 지난 2019년 10월 30일 세계 최고의 학술지인 네이처(Nature)에 스타크래프트2(StarCraft Ⅱ) 게임 AI인 알파스타(AlphaStar)의 논문을 공개했다. 스타크래프트는 현 시대를 대표하는 게임으로 실시간 조작을 바탕으로 한 전략 시뮬레이션 게임이다. 특히 스타크래프트라는 게임은 1990년대 말 우리나라 PC방 확산에 큰 영향을 미쳤다. 이것은 e스포츠의 영역으로 이어져 우리나라는 스타크래프트를 중심으로 e스포츠가 확산되는 계기가 됐다. 우리나라에서 스타크래프트는 단순히 게임에서 벗어나 하나의 문화로 승화될 정도로 우리 사회에 미치는 영향이 컸다.

특히 스타크래프트는 배틀넷(Battle.net)이라는 시스템을 도입했는데, 이것은 전 세계에 있는 게이머들이 서로 온라인으로 대결할 수 있는 환경을 말한다. 스타크래프트 배틀넷은 인터넷의 발전에 힘입어 큰 성공을 거뒀는데, 이를 기점으로 게임을 직업으로 하는 프로 게이머의 등장과 이들이 우열을 경쟁하는 e스포츠가 탄생하는 계기가 됐다. 스타크래프트 프로 게이머는 빠르고 뛰어난 게임 플레이, 창의적이고 유연한 전략 변화 등으로 대중의 인기를 얻었다.

스타크래프트에 완벽한 전략이란 것은 없다. 스타크래프트는 가위-바위-보 게임과 같이 전략별로 상성이 존재하기 때문이다. 따라서 상대방의 전략에 따라 자신의 전략을 탄력적으로 변화시키는 것이 승리의 지름길이다. 또한 거시적인 전략부터 유닛 간 전투의 세밀한 전략까지 스타크래프트는 고도의 지능적 판단을 필요로 한다. 그 전략들을 실행시키기 위한 신속한 조작도 프로 게이머의 필수적인 자격 요건이다.

이러한 스타크래프트 게임의 특징은 스타크래프트 AI를 개발하는데 있어 많은 도전 과제를 부여한다. 스타크래프트는 바둑과 전혀 다른 장르의 게임으로, 알파고의 성공 방정식이 스타크래프트 AI로 이어진다는 것을 보장하지 않는다. 프로 게이머 수준의 스타크래프트 AI를 개발한다는 것은 인간의 지능적 행동을 모방하기 위한 단초를 잡을 수 있다는 의미에서 또 하나의 위대한 도전(Grand Challenge)의 영역이라 볼 수 있다.

스타크래프트 이미 게임 안에서도 AI 기능을 탑재하고 있다. 스타크래프트는 소위 컴퓨터와의 대결을 제공하기 때문이다. 이러한 스타크래프트 AI는 대부분 규칙 기반(rule-based) 접근을 취한다. ‘A라는 유닛은 B에 강하지만 C에 약하다’, ‘상대가 많은 유닛을 보유하고 있다면, 어딘가에 확장 기지가 있다’라는 식의 규칙으로 행동 방법을 정의한다. 그러나 이러한 규칙의 패턴을 플레이어가 알아차린다면, 플레이어는 이것을 역으로 이용해 컴퓨터와의 대결에서 쉽게 승리할 수 있다.

스타크래프트 AI[1]는 학술적으로도 중요한 주제다. 경진대회 형식으로 개최되는 스타크래프트 AI 개발은 2011년부터 지금까지 이어오고 있다. AI는 인간과 동일한 조건하에 개발되어야 하는 제약사항이 있기 때문에[2], 문제의 난도가 더욱 높다. 최근에는 심층학습(Deep Learning)의 부상으로 인공신경망 기반의 접근도 나오고 있지만, 그 수준은 여전히 프로 게이머를 넘어 설 수 없는 상황이다. 프로 게이머 수준의 스타크래프트 AI를 개발한다는 것은 바둑으로 보자면 프로급 바둑기사 수준의 AI를 개발하는 것과 유사하다. 바둑 AI 역시 알파고가 나오기 전에는 프로 바둑기사의 벽을 넘을 수 없었기 때문이다.

딥마인드는 알파고가 이세돌 9단과의 대국에서 승리한 뒤 스타크래프트2 AI 개발을 발표했다. 스타크래프트2는 1998년 처음 발매된 스타크래프트 게임의 계보를 잇는 후속작으로 지난 2010년 발매됐다. 스타크래프트2는 이전 작 보다 더욱 화려해진 그래픽과 다양화된 전략으로 전 세계 게이머의 주목을 받았다. 딥마인드의 새로운 도전은 그 성공 여부에 대해 많은 이슈가 있었지만, 가장 주목받은 것은 형평성이었다. 알파고의 경우, 일각에서 ‘인간과 슈퍼컴퓨터의 대결’이라며 형평성을 문제 삼았다. 스타크래프트2 AI 역시 인간과의 대결에서 형평성이 보장되어야 한다는 것이다. 스타크래프트2는 게임의 진행이 실시간으로 이루어진다. AI는 컴퓨터의 성능이 허락하는 한, 1초에 수 천 번의 행동도 명령할 수 있다. 혹자는 형평성을 위해 로봇 팔이 직접 게임을 수행해야 한다고도 주장했다. 또한 알파스타가 2019년 1월 프로게이머와의 대결에서 유일하게 패한 경기는 사람과 동일한 조작환경을 활용했다는 점에서 10승에 대한 형평성[3] 논란이 있었다.

1) 이 문단에서 언급하는 스타크래프트 AI는 1998년 출시된 스타크래프트 브루드워를 의미하며, 딥마인드가 AI를 개발한 대상 게임인 스타크래프트2와는 다르다.
2) 경진대회에서는 스타크래프트 AI가 시스템에 접근하여 정보를 얻는 방식은 활용할 수 없다.
3) 10승을 거둔 환경은 조작 화면이 카메라 뷰가 아닌 전체 화면을 활용한 결과였다.

딥마인드는 지난 2017년 8월 자사의 홈페이지에 스타크래프트2 AI 개발 경과를 공개했었다. 그 결론은 아직 프로 수준에 도달하지 못했다는 내용이었다. 그러나 딥마인드는 많은 연구자들을 스타크래프트2 AI 개발에 동참시키기 위해, 스타크래프트2 개발사인 블리자드(Blizzard)와 함께 학습 환경을 공개했다. 이것은 SC2LE(StarCraft 2 Learning Environment)라고 불리며, AI 개발자가 쉽게 게임 정보를 획득하고 활용할 수 있다. 이에 더하여 딥마인드는 약 6만 5천 건의 1대1 대결 리플레이 데이터를 공개했다. 리플레이는 특정 게임을 전지적 시점에서 모두 볼 수 있는 것으로, SC2LE를 통해 필요한 데이터를 추출할 수 있다.

딥마인드는 초기 결과 공개 이후 1년 6개월이 지난 2019년 1월 알파스타를 공개했다. 알파스타의 성적은 정상급 프로 게이머 2명과 상대해 10승 1패의 성적을 거뒀다. 초기 결과 발표 당시에는 1대1 매치의 성능이 일반인보다 조금 나은 수준이었기 때문에 아직 스타크래프트2 AI가 가야할 길은 멀다는 것이 중론이었다. 그러나 알파스타는 이러한 예측과는 정반대의 결과를 보여줬다. 특히 알파스타가 의미가 있는 점은 대결의 형평성을 어느 정도 확보했다는 것이다. 알파스타는 경기에서 특수한 권한을 갖지 않았고, 조작 속도의 평균은 프로 게이머 수준보다 낮았으며, 게이머용 컴퓨터 한 대를 활용해 대결했다. 또한 딥마인드는 철저하게 데이터 기반의 접근으로 알파스타를 구현했다.

알파스타에는 10가지가 넘는 심층학습 최신기술이 활용됐으며, 리플레이 데이터를 활용한 지도학습과 승률을 높이기 위한 강화학습, 그리고 학습된 AI 에이전트[4]가 서로 대결하는 알파스타 리그를 겪으며 성능이 향상됐다. 그 세부적인 내용이 담긴 논문이 바로 2019년 10월 네이처지에 실린 “Grandmaster level in StarCraft II using multi-agent reinforcement learning” 이다.

이 보고서는 알파스타의 네이처 논문을 분석하여 알파스타에 활용된 AI 알고리즘을 알기 쉽게 전달하기 위한 것이다. 이어지는 내용에서는 알파스타의 도전 과제와 AI 알고리즘에 대해서 살펴보겠다.

4) AI 에이전트는 하나의 독립된 시스템을 의미하며, 알파스타 AI 에이전트는 사람과 1대1 대결을 할 수 있는 AI로 볼 수 있다.

2. 알파스타의 도전 과제

스타크래프트2라는 게임의 장르는 실시간 전략 시뮬레이션이다. 여기서 시뮬레이션이라는 것은 가상의 공간에서 조작을 하는 것을 의미하며, 전략이라는 것은 상성이 존재하거나 행동에 대한 기회비용이 명확하다는 것을 의미한다. 특히 스타크래프트2의 전략은 어느 정도 추상화가 가능하지만 상황에 따라 유동적으로 변화시킬 수 있는 가능성도 고려해야 한다. 스타크래프트2의 전략을 한 마디로 표현하자면 복잡(complex)하다는 것인데, 그 복잡도는 단순한 규칙으로 표현하기가 어렵다. 따라서 스타크래프트2 AI의 성공여부는 복잡한 전략을 얼마나 잘 학습하고, 실제 경기에서 적절하게 쓰이느냐에 달려있다. 또한 AI 학습한 전략에 따라 상황을 판단하고 결정하는 것이 실시간으로 이루어져야 한다는 것도 AI 개발에 있어 중요한 도전이다.

딥마인드의 알파스타는 기본적으로 종단간(end-to-end) 학습을 목표로 한다. 종단간 학습이란 심층학습의 가장 일반적인 학습 방법으로 '데이터(입력)→AI알고리즘→결과(출력)'의 과정에서 사람의 개입 없이 입력값에 대한 출력을 얻는다는 의미다. 여기서 AI알고리즘은 주어진 입출력 데이터를 통해 학습된다고 볼 수 있다. 종단간 학습의 가장 쉬운 예는 기계 번역이다. 영어 문장을 한국어 문장으로 번역하는 임무를 가정해 보자. 일반적인 번역의 과정은 먼저 영어 문장을 품사별로 구분하고, 각 품사별로 대응되는 한국어 단어를 선정하는 것으로 시작한다. 여기서 이중적인 의미를 갖는 단어는 문맥에 따라 해석이 다르게 될 것이다. 과거에는 이 모든 것을 수작업으로, 즉, 규칙 기반의 접근으로 해결했었는데 심층학습으로 인해 그 방법론이 변화하게 된다. 기계 번역에서의 종단간 학습은 영어 문장과 이에 대응하는 한국어 번역 문장의 데이터 확보에서부터 출발한다. 충분히 많은 데이터가 모였다면, 종단간 학습은 앞서 언급한 문장의 품사별 구분과 문맥에 따른 단어의 선택이 인공신경망을 통해 학습된다는 것이다. 다시 말하자면, 데이터로 학습된 인공신경망이 품사 구분과 해석될 단어의 선택을 결정한다는 의미다.

종단간 학습에서는 입력과 출력 데이터의 정의, 그리고 이것을 학습할 AI 모델의 선택이 중요하다. 딥마인드는 최신 AI 알고리즘의 적극적인 활용을 통해 종단간 학습을 구현했다.

알파스타의 다섯 가지 도전과제[5)]

스타크래프트2 AI를 종단간 학습으로 구현하기 위해서는 입력, 출력, 그리고 AI 알고리즘의 선택이 필요하다. 특히 AI 알고리즘의 선택은 스타크래프트2가 가지고 있는 도전과제에서 자유로울 수 없다.

스타크래프트2의 대표적인 도전 과제는 크게 다섯 가지가 있다. **첫 번째는 게임 이론**을 따르는 특성을 갖고 있다. 스타크래프트2에서 하나의 완벽한 전략이란 없다. 서로 상성이 존재하고 물고 물리는 관계로 인해 스타크래프트2의 전략은 가위-바위-보 게임과 유사하다.

두 번째는 불완전한 정보다. 스타크래프트2를 구성하는 가장 큰 단위인 건물과 유닛은 각자의 시야를 가지고 있다. 플레이어는 유닛과 건물이 제공하는 시야에서만 정보를 얻을 수 있기 때문에, 상대의 정보를 얻기 위해서는 자신의 유닛을 적군 기지에 보내는 등의 행위가 필요하다. 상대에 대한 불완전한 정보는 어떠한 전략을 선택할 것인가에 대한 불확실성과 이어진다. 이러한 이유로 정찰이라는 행위가 필요하고, 이것을 통해 상대방의 전략을 추정할 수 있다. 그러나 전략의 추정은 확률적이고 경험적이기 때문에 주어진 정보를 통해 의도를 추론하는 것이 중요한 과제다.

세 번째는 장기 계획이다. 스타크래프트2 게임의 결과는 승-무-패의 세 가지 중 하나다. 그러나 결과까지 도달하기에는 소규모 전투에서부터 거시적인 전략까지 다양한 요인이 존재한다. 여기서 관건은 순간의 판단에 대한 보상이 즉각적으로 주어지지 않는다는 점이다. 예를 들어 유닛간의 전투에서는 많이 살아남는 쪽이 이득을 얻어간다. 그러나 특정 시점에서 특정 건물을 짓는다거나, 이를 통해 새로운 전투 유닛을 생산하는 등의 행위는 장기적으로 전략의 다양성을 확보하는 것이 가능하다. 그러나 여기에 투자되는 생산비용과 시간 그리고 이를 위해 포기해야할 수도 있는 기존의 전략 등은 모두 기회비용이기 때문에 이 행위의 가치를 즉각적으로 판단하기 어렵다. 따라서 스타크래프트2는 가능한 많은 전략적 선택지를 열어둘 수 있는 메커니즘이 필요하다.

5) AlphaStar: Mastering the Real-Time Strategy Game StarCraft II, Deepmind (2019.01.)에서 정리

네 번째는 실시간 제어다. 자신과 상대방이 번갈아가며 수를 두는 바둑과 달리 스타크래프트2는 실시간으로 이루어지는 연속적인 명령을 통해 게임이 진행된다. 여기서 명령은 마우스나 키보드의 입력 값으로 표현할 수 있다. 예를 들어 일꾼으로 A라는 건물을 짓는 행위는 ①일꾼이 있는 곳으로 화면의 이동, ②마우스로 일꾼을 선택, ③A라는 건물을 짓는 명령의 선택, ④A라는 건물을 지을 공간을 선택 하는 명령으로 표현할 수 있다. 스타크래프트2에서의 명령은 구조화하여 표현할 수 있는데, 무엇을(what), 누가(who), 어디에(where), 언제(when)로 구분 지을 수 있다. 특정 시점에 가능한 명령은 이론적으로 10^{26}가지이며, 통상적으로 하나의 게임에 수 만 번의 시점이 있기 때문에 모든 경우의 수를 고려한다는 것은 불가능하다.

마지막은 넓은 조작 공간이다. 스타크래프트2는 현재 게임을 플레이하고 있는 창(camera view)과 게임 공간 전체를 조망하는 미니맵으로 구성된다. 게임 플레이어는 대부분의 조작을 현재의 창에서만 할 수 있으며, 창 이외의 공간에 명령을 하기 위해서는 창의 이동이 필요하다. 이러한 조작 공간은 게임 자체의 복잡도를 높이는데 기여한다.

스타크래프트2 AI는 이 다섯 가지의 도전 과제를 해결해야 프로 게이머의 수준을 넘볼 수 있다. 게다가 이렇게 어려운 과제를 종단간 학습으로 구현한다는 것은 AI 모델이 상당히 많은 정보를 추론할 수 있어야 한다. 이것은 단순히 현대 AI 기술의 발전의 연장선상에 닿아 있다고 해석하기에는 무리가 있다. 알파고의 성공이 바로 스타크래프트2 AI의 성공으로 바로 이어지기가 매우 어렵다는 의미다. 스타크래프트2 AI는 바둑 AI와는 전혀 다른 도전과제를 해결해야 한다는 점에서 또 다른 그랜드 챌린지의 영역이라 볼 수 있다.

알파고와 알파스타

그렇다면 알파고와 알파스타는 얼마나 다를까? 먼저 바둑과 스타크래프트2는 게임 장르가 다르다. 알파고와 알파스타는 동일하게 종단간 학습의 접근을 취한다. 즉 학습을 위한 입력과 출력 데이터의 선정과 이 들의 상관관계를 통해 전략을 학습하는 AI 모델의 설계가 게임 AI를 구현하기 위한 출발점이다. 따라서 알파고와 알파스타는 학습을 위한 전체적인 과정이 서로 유사한 면이 있다. 그러나 그 세부적인 내용은 상이하다.

먼저 바둑과 스타크래프트2의 차이점은 <표 1>과 같이 설명할 수 있다.

<표 1> 바둑과 스타크래프트2의 차이점

구 분	바 둑	스타크래프트2
장 르	보드게임	실시간 전략 시뮬레이션
게임 진행	턴 방식	실시간 명령
게임 공간	19x19 격자 공간	최대 256x256 크기의 맵
경우의 수	약 10^{170} 개	매 스텝 당 10^{26} 개
소요 시간[주1)]	1 ~ 4시간	10분 ~ 1시간
상대방의 상황	모두 공개	정찰을 통해 습득

주1) 극단적인 상황을 제외한 일반적인 경우를 가정
자료 : 알파스타의 인공지능 알고리즘, SPRi(2019)에서 재구성

바둑과 스타크래프트2의 가장 큰 차이점은 게임 진행과 상대방에 대한 정보다. 바둑은 주어진 시간 안에 착수를 해야 하며, 주어진 시간을 모두 소비한 경우 착수에 일정 시간(초읽기)안에 착수를 해야 한다. 반면 스타크래프트2는 게임과 지속적으로 소통하며 실시간 명령을 내려야 한다. 바둑은 상대방의 정보가 모두 공개되어 있다. 그럼에도 바둑이 매우 어려운 문제로 분류됐던 것은 무한대에 가까운 경우의 수 때문이다. 스타크래프트2의 경우는 상대방의 정보가 공개되어 있지 않다. 상대방의 정보가 정찰이라는 행위로 얻어진다는 점은 전략의 다양화로 이어진다.

알파고와 알파스타는 큰 틀에서 유사한 학습 방법을 취한다. 두 AI 모두 기보와 리플레이 데이터를 바탕으로 지도학습을 수행한다. 이후 지도학습의 결과물인 AI 에이전트[6]는 자체 대결과 강화학습을 통해 성능을 개선하는 방식이다. 알파고와 알파스타의 차이점은 결국 종단간 학습을 구성하는 입력, 출력, AI 모델에서 발생한다. 특히 AI 모델의 경우 가장 큰 차이를 보이는데 이것은 게임의 장르가 서로 다르기 때문이다. 다음 <표 2>는 알파고와 알파스타를 비교를 나타낸다.

6) 에이전트(Agent)의 의미는 하나의 독립적인 AI 시스템이라고 볼 수 있다. 이 에이전트는 성능의 고하를 떠나 사람과 대결할 수 있는 AI로, 자체 대결을 통해 성능을 향상시킬 수 있다.

<표 2> 알파고와 알파스타의 비교

구 분	알파고 Lee[주1)]	알파스타
학습 데이터	16만 건의 기보	97.1만 건의 리플레이
입력 값	48가지 특성으로 나눠진 정보 (예, 흑돌, 백돌, 빈칸 위치 등)	최대 512개의 특성, 지도 정보, 플레이어 데이터, 게임 통계
출력 값	착수 가능 지점의 확률, 승리할 확률	5초 당 최대 22번의 명령
학습 방법	기보를 통한 지도학습과 자체 대국을 통한 강화학습	리플레이 데이터의 지도학습과 자체 대결을 통한 강화학습, 멀티 에이전트 기반의 리그학습
AI 모델	합성곱신경망, 심층Q학습, 몬테-카를로 트리 탐색(MCTS)	합성곱신경망, 장단기기억, 관계 신경망, 트랜스포머, 포인터 네트워크, 모방 학습 등
사람과의 대결에서 컴퓨팅 파워	48장의 TPU[주2)]	고성능 GPU 1장으로 구성된 컴퓨터

주1) 알파고는 다양한 버전이 있는데, 표에서의 비교 대상은 이세돌 9단과 대국에서 활용된 알파고 버전임

주2) TPU(Tensorflow Processing Unit)는 인공신경망 학습과 추론에 최적화된 계상 장치로 현대 심층학습에 많이 활용되는 GPU와 대비해 낮은 전력을 소모함

자료 : 알파스타의 인공지능 알고리즘, SPRi(2019)에서 재구성

3. 알파스타의 AI 알고리즘

알파스타를 전반적으로 이해하기 위해서는 크게 세 가지가 필요하다. <표 2>에서의 항목으로 보자면 ① 입력 값, ② 출력 값, ③ AI 모델과 학습방법이다.

(1) 알파스타의 입력과 출력

종단간 학습은 입력과 출력 데이터를 정의하는 것이 첫 번째 단계다. 앞서 예시로 든 기계 번역을 상기시켜보면, 입력은 영어 문장이고 출력은 이 영어 문장을 번역한 한국어 문장이다. 여기서 중요한 것은 입출력 데이터의 상관관계이다. 입력과 출력이 유의미한 관계가 있어야 AI 모델로 학습이 가능하기 때문이다. 알파스타 역시 이렇게 유의미한 입출력 데이터를 정의하는 것으로 출발한다. 입출력 데이터를 정의하는 방법에는 도메인 지식이 가장 중요한데, 이를 위해서는 스타크래프트2의 전문가와 AI 전문가의 협력이 필요하다. 실제로 딥마인드 연구진은 스타크래프트2 개발사인 블리자드와 더불어 정상급 프로게이머와 협력 연구를 통해 상관관계가 높은 입출력 데이터를 정의했다.

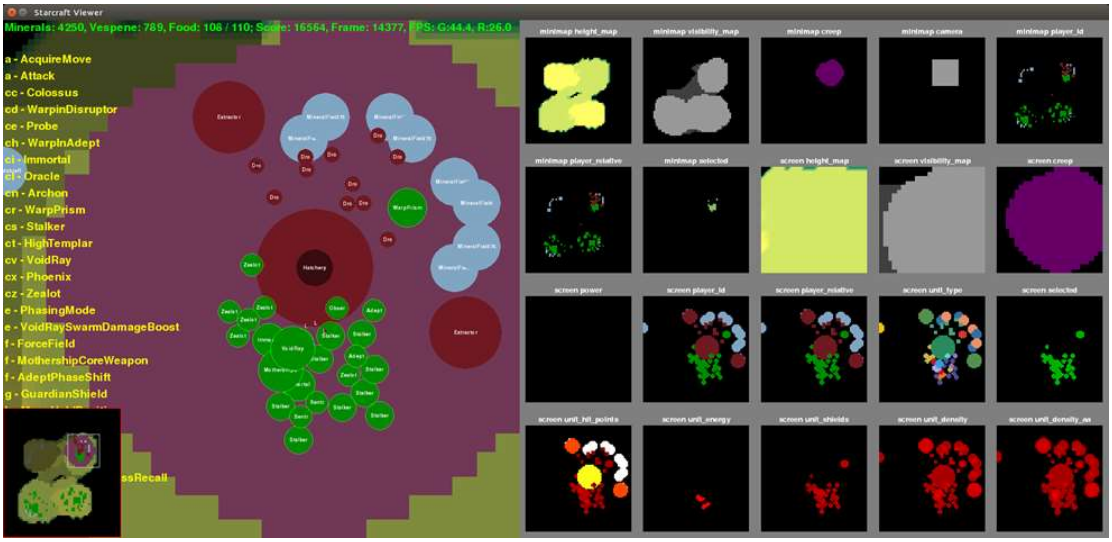
알파스타의 입력은 다음 <표 3>과 같이 구조화되어 표현할 수 있다. 입력 값은 크게 네 가지로 이해할 수 있는데 게임 상의 다양한 정보(Entity)와 시각적인 맵의 정보, 그리고 수치적인 정보로 구성된다.

<표 3> 알파스타의 입력

구 분	세 부 내 용
게임 상의 정보 (Entity)	- 유닛/건물 : 타입, 소유자, 상태, 시야의 내외 여부, 위치, 일꾼의 수, 특수 기능 - 보조정보 : 수송기 탑재 여부, 건물 상태, 자원 상태, 명령 순서 등
시각적 정보 (Map)	- 현재의 시야, 정찰된 시야 - 공격받은 유닛, 이동 가능한 영역, 건물을 지을 수 있는 영역
플레이어 고유 정보	- 선택한 종족, 유닛 업그레이드 수치 - 에이전트 통계 : 자원, 인구수 등과 관련된 정보
게임 통계	- 32x20 크기의 게임 조작화면 창(camera) - 게임 플레이 시간

자료 : Grandmaster level in StarCraft II using multi-agent reinforcement learning

알파스타의 시각적 정보의 예시는 [그림 1]과 같다. 이것은 앞서 소개한 스타크래프트2 AI 개발도구인 SC2LE에서 제공하는 기능으로 추출할 수 있다. [그림 1]은 현재의 조작화면과 미니맵 화면에서 서로 구분되는 정보를 추상화해 나타낸 것으로, 아군과 적군 유닛의 위치 및 체력, 현재의 시야, 카메라의 위치 등이 있다.



[그림 1] 알파스타의 시각적 정보의 예시

자료 : Starcraft II: A new challenge for reinforcement learning

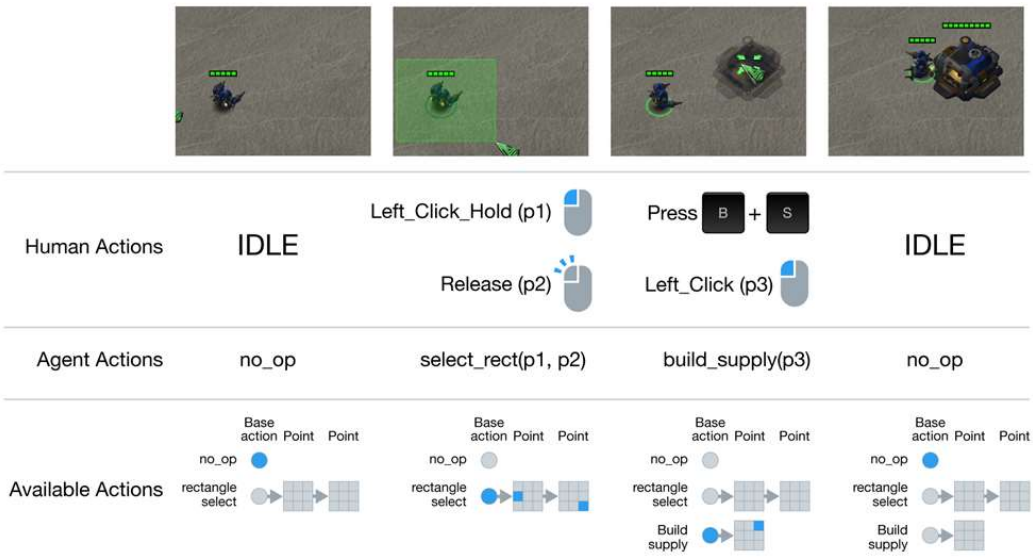
알파스타의 출력은 다음 <표 4>와 같이 행동(action)으로 구조화 된다.

<표 4> 알파스타의 출력

구 분	세 부 내 용
행동 타입	유닛을 움직이거나 건물을 지음, 유닛의 생산, 카메라의 이동 등 분류된 행동의 타입과 명령
선택된 유닛	행동을 수행하는 유닛
목표	행동을 수행하는 256x256의 한 지점
순서, 반복, 지연	명령의 실행 순서, 명령의 반복 여부, 다음 관측 정보가 들어오기 전 까지의 대기 시간

자료 : Grandmaster level in StarCraft II using multi-agent reinforcement learning

알파스타의 출력을 가시화해보면 아래 [그림 2]와 같이 이해할 수 있다. [그림 2]는 특정 건물을 만드는 행동에 대한 출력이다. 이것은 하나의 조작으로 이루어지기 어렵기 때문에 출력 자체가 일련의 연속된 명령으로 구성된다고 볼 수 있다. 출력의 형태에서 유추할 수 있듯이, 출력될 명령의 수는 상황에 따라 다를 수 있다. 또한 그 명령에 순서가 존재하기 때문에 이 부분까지 고려해야 한다. 다시 말하자면 알파스타는 다양한 정보를 입력 받아 출력 값으로 다양한 유닛에 지시를 내리는데, 이 지시에 대한 순서도 정해야 한다는 것이다.



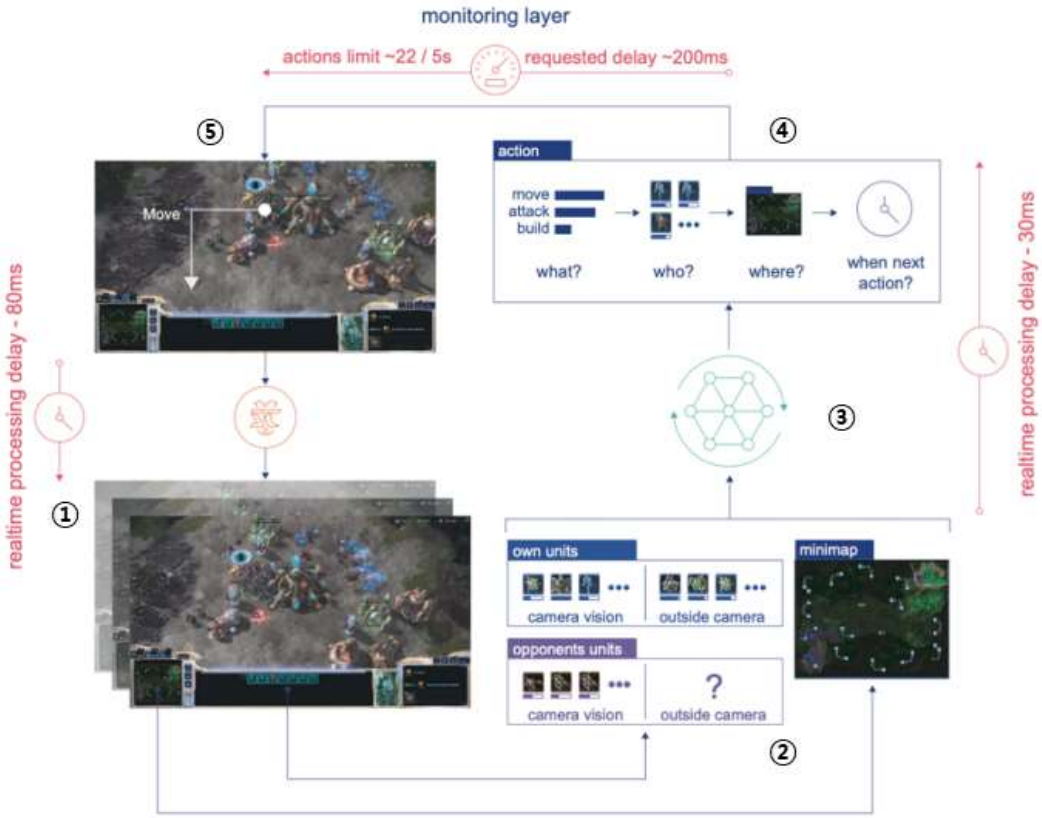
[그림 2] 알파스타의 출력(행동) 예시

자료 : AlphaStar: Mastering the Real-Time Strategy Game StarCraft II, Deepmind (2019.01.)
https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/

한편 이 출력 값의 수는 사람과의 대결에서 형평성을 갖추기 위해 단위 시간당 행동할 수 있는 상한을 정해야 한다. 스타크래프트2에서 얼마나 신속하게 명령을 처리하는지에 대한 지표로 분당 행동수(Actions Per Minute, APM)라는 것을 활용한다. 스타크래프트2 프로 게이머의 APM은 평균적으로 200~300대에 분포한다. 알파스타가 프로 게이머와 유사한 수준으로 조작에 개입하기 위해서는 이 분당 행동수에 제약이 있어야 하는데, 알파스타는 5초에 최대 22번 행동하는 것으로 제한했다.[7)]

7) 항상 최대치로 행동하면 알파스타의 APM 평균은 264

알파스타의 입출력의 조망은 다음 [그림 3]과 같다. ① 먼저 일정 시간의 게임 화면이 관측된다. ② 이 화면은 SC2LE의 API를 통해 다양한 정보를 추출하게 된다. 예를 들면 현재 게임 화면(camera) 안에 존재하는 아군과 적군의 유닛, 게임 화면 밖에 있는 아군의 정보, 미니맵 정보가 있다. ③ 이러한 정보는 알파스타의 에이전트에 입력으로 들어가고, ④ 출력으로 일련의 행동을 산출한다. 일련의 행동은 어떠한 행동을, 누가, 어디에, 언제 수행할 것인지로 구분되어 출력된다. ⑤ 이것은 실제 게임화면의 조작으로 이어진다.



[그림 3] 알파스타의 입출력

자료 : Grandmaster level in StarCraft II using multi-agent reinforcement learning

[그림 3]에서 ③은 알파스타 에이전트다. 이는 다음 절에서 자세히 살펴보겠다.

(2) 알파스타의 AI 모델과 학습 방법

알파스타의 AI 모델은 학습 방법에 따라 크게 3가지 단계로 나뉜다. 첫 번째는 **지도학습**으로 리플레이 데이터를 학습하는 과정이다. 지도학습의 결과는 하나의 알파스타 **에이전트**가 될 수 있는데, 이것을 고도화하기 위해 자체 대결 기반의 **강화학습**을 수행한다. 강화학습의 결과물은 이후 멀티-에이전트 강화학습을 거친다. 멀티-에이전트 강화학습은 에이전트에 서로 다른 목표를 부여해 대결하는 **리그학습**이다.

알파스타의 지도학습

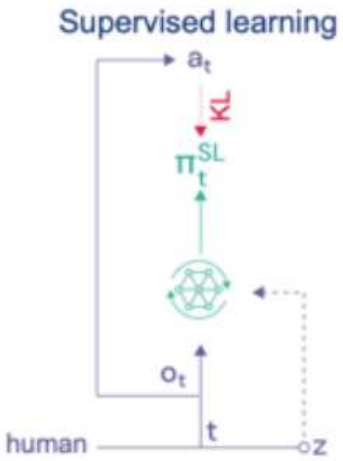
알파스타의 첫 단계는 지도학습(Supervised Learning)이다. 지도학습은 심층학습의 가장 기본적인 접근 방법으로, 시험지-답안의 관계를 통해 특정 AI 모델을 학습시키는 것이다. 알파스타에서 시험지-답안은 리플레이 데이터로부터 만들어진다. 알파스타는 97만 1천 건의 리플레이 데이터를 활용했는데, 이것은 스타크래프트2 배틀넷에서 상위 22%에 해당하는 경기다.[8] 리플레이 데이터에서 시험지는 현재 게임의 화면이고, 답안은 일련의 행동이다. 시험지-답안은 앞서 소개한 알파스타의 입력-출력의 관계를 갖는다. 딥마인드는 이 두 가지에 더하여 사람의 전략을 추출했다. 먼저 게임 시작부터 20개의 건물(유닛 생산 및 업그레이드 등)을 짓는 순서를 추출했다. 이것은 스타크래프트2에서 빌드 오더(build order)로 불리며, 여러 프로 게이머들에 의해 최적화된 길(route)이라고 볼 수 있다. 이에 더하여 해당 전략에 대한 통계적인 지표도 활용했다.

알파스타의 지도학습은 학습의 결과로 생성될 최초의 에이전트가 가능한 많은 전략을 보유하는 것을 목표로 한다. 이것은 스타크래프트2 게임의 특성상 선택 가능한 전략의 폭이 넓을수록 승리를 위한 다양한 전술과 대응이 가능하다는 이유에서다. 알파스타의 지도학습은 폭넓은 전략을 습득하는 것을 목표로 하기에, 성능 자체는 좋지 않다. 역설적으로 너무 많은 전략이 있다는 것은 어느 것이 적절한 전략 제대로 판단하기가 어려울 수 있기 때문이다.

8) MMR(Match Making Rate) 기준 3,500점 이상의 경기가 이 범위에 속한다. MMR은 플레이어의 실력을 가늠하는 지표로 높을수록 숙련자에 속한다. 통상적으로 MMR이 1,000점 이상 차이나면 압도적인 실의 차이가 존재한다는 것으로 해석할 수 있다.

알파스타의 지도학습은 많은 전략을 학습하기 위해 리플레이 데이터를 활용했는데, 과거 딥마인드가 발표했었던 알파고 초기 버전과 흐름이 유사하다. 알파고 역시 전문가의 바둑기보를 토대로 특정 바둑판 상태에서 전문가들이 선호하는 착수 지점을 학습했다. 마찬가지로 알파스타도 배틀넷 상에서 상위 22%의 게이머가 선호하는 빌드 오더와 전략을 학습한다고 볼 수 있다. 이렇게 사람의 데이터를 활용하는 방법을 모방 학습(imitation learning)이라고 한다. 바둑이나 스타크래프트2는 무한대에 가까운 경우의 수를 가지고 있다. 이로 인해 단순한 규칙만 주고 무작위로 행동하여 학습하는 것은 가능하다고 해도 그 결과가 좋지 않을 것이다. 그 기준점을 전문가의 시점으로 바라본다는 것이 알파스타 지도학습의 핵심이다.

알파스타의 지도학습을 개념화하여 표현한 것은 [그림 4] 같다.



- t : 주어진 시점
- o_t : t 시점에서의 관측된 입력 값 <표 3>
- a_t : t 시점에서의 행동(출력 값) <표 4>
- z : 사람의 전략
- π_t^{SL} : o_t와 z를 입력으로 받고 행동을 출력
- KL : 알파스타 에이전트의 출력 π_t^{SL}과 리플레이 데이터에서의 행동 a_t의 유사도를 측정 방법으로, 쿨백-라이블러 발산(KL divergence)이라 불림

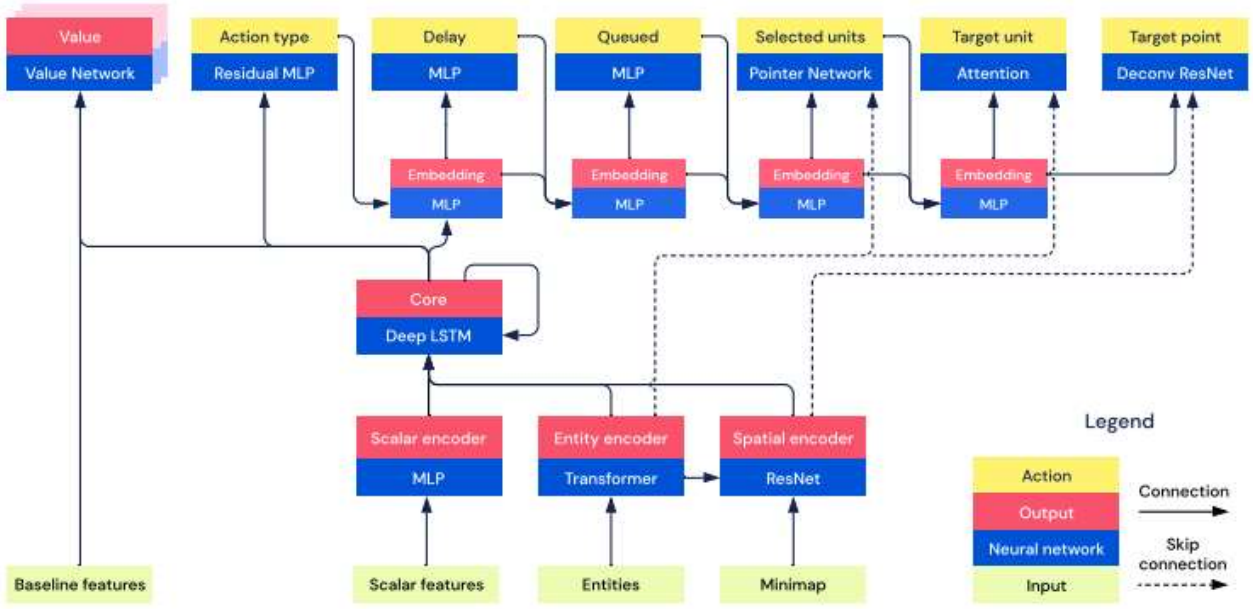
[그림 4] 알파스타의 지도학습의 개념도

자료 : Grandmaster level in StarCraft II using multi-agent reinforcement learning

알파스타의 지도학습에서는 다양한 방법을 시도했는데, 사람의 전략인 z와 함께 학습을 진행했을 때 가장 성능이 높았다. 이렇게 진행된 지도학습의 결과물이 사람과 직접 대결을 할 수 있는 알파스타의 첫 번째 에이전트가 된다. 이제 남은 과정은 스타크래프트2의 다양한 전략을 담은 에이전트가 상황별로 어떠한 전략을 선택할지에 대해 고도화를 시키는 작업이다. 이것은 이어질 강화학습과 멀티-에이전트 강화학습인 리그학습의 역할이다.

알파스타 에이전트

앞서 설명한 첫 알파스타 에이전트는 지도학습의 결과물이다. 이 에이전트의 세부구조는 일련의 명령을 산출하기 위해 다양한 기법을 활용했다. 알파스타 에이전트의 세부구조는 [그림 5]와 같다. [그림 5]에서의 입력(input)과 출력(action)은 각각 <표 3>과 <표 4>를 대비해서 보면 된다.



[그림 5] 알파스타 에이전트의 개념도

자료 : Grandmaster level in StarCraft II using multi-agent reinforcement learning

알파스타 에이전트는 상당히 복잡한데, [그림 5]에서 파란색으로 표기한 인공신경망(Neural Network)의 역할에 대해서 간략히 살펴보자. 먼저 다층퍼셉트론(Multi-Layer Perceptron, MLP)은 가장 간단한 인공신경망이다. MLP는 복잡한(비선형적인) 데이터에서 패턴을 인식하기 위해 특징을 추출하는 역할을 한다. 장단기기억(Long Short Term Memory, LSTM)은 시계열 데이터의 예측에서 많이 활용되는 방법이다. LSTM은 알파스타의 입력이 시간적으로 연속된 게임 화면이라는 점에서 활용된다고 볼 수 있다. 어텐션(Attention)은 단어 그대로 어디를 주목해야 하는지를 구현한 방법이다. 알파스타에서는 만약 공격을 할 때 어떤 유닛을 공격해야 할지를 나타낸다고 해석할 수 있다. 트랜스포머(Transformer)는 어텐션 기반의 방법론으로 입력 데이터에도 어텐션을 부여하는 자체 어텐션(Self

Attention)을 구현한 방법이다. 포인터 신경망(Pointer Network)은 가변적인 출력에 대비하기 위해 고안된 방법이다. 포인터 신경망 역시 어텐션의 개념을 담고 있으며, 알파스타에서는 다수의 유닛을 선택할 때 활용되었다. 잔차신경망(Residual Network, ResNet)은 2015년 이미지 인식 경진대회에서 우승을 차지한 신경망 구조로 이미지 패턴 인식에 최적화된 방법이다. 알파스타에서는 미니맵의 상황을 인식하는데 활용됐다. 가치 신경망(Value Network)은 현재의 상태가 얼마나 유리한지를 근사하는 것으로, 자세한 내용은 강화학습에서 다룰 것이다.

알파스타 에이전트의 구조는 현대 AI 기술의 정수가 모여 있다. 이 구조가 복잡할 수밖에 없는 것은 스타크래프트2가 그만큼 복잡하다는 사실을 반증한다. 이 에이전트의 구조에는 수많은 시행착오와 스타크래프트2의 도메인 지식이 녹아있다. 그러나 이 에이전트가 프로 게이머를 넘어선 에이전트로 변모되기까지 아직 강화학습의 영역이 남아있다.

알파스타의 강화학습

알파스타의 강화학습(Reinforcement Learning)은 지도학습으로 생성된 AI 에이전트를 출발점으로 삼는다. 강화학습의 기본은 정책(policy)과 가치(value)라는 개념이다. 정책은 주어진 상태(state)에 대해서 행동(action)하는 방법을 나타낸다. 가치는 현재의 상태가 자신에게 얼마나 유리한지를 나타내는 지표다. 알파고에서의 정책은 착수할 확률이고, 가치는 현재 상태에서 자신이 승리할 확률이다. 알파스타에서의 정책은 지금까지 설명한 알파스타 에이전트의 출력인 일련의 행동이다. 알파스타의 가치는 알파고와 마찬가지로 자신이 승리할 확률을 근사하는 역할을 한다. 알파스타의 강화학습은 기본적으로 에이전트 간의 자체 대결(Self-play)을 바탕으로 한다.

알파스타의 강화학습은 지도학습으로 생성된 에이전트의 승률을 높이는 역할을 한다. 승률을 높이기 위해서는 지도학습 에이전트의 정책이 얼마나 좋은지를 판단해야 한다. 이것이 바로 가치이며 가치는 인공신경망의 형태(Value Network)로 근사할 수 있다. 그러나 스타크래프트2 게임은 수 천 번의 행동이 지나야 게임 결과를 알 수 있다. 이것을 희소한 보상(sparse reward)이라고 한다. 이러한 경우 가치의 근사 값을 구하기가 매우 어렵다.

알파스타는 이것 문제를 해결하기 위해 모조 보상(pseudo reward)을 도입했다. 모조 보상은 알파스타 에이전트가 사람의 전략을 얼마나 따르는지를 판단하는 값이다. 모조 보상을 사용한다는 사실은 강화학습의 결과물인 에이전트가 사람의 전략을 따르게 하고자 하는 의도가 담겨있다. 지도학습 에이전트도 사람의 전략을 학습한 것이 사실이나, 폭넓은 전략의 확보에 치중했기 때문에 실제로 경기에 활용하면 사람의 전략과는 상이한 행동을 보인 것으로 분석된다.

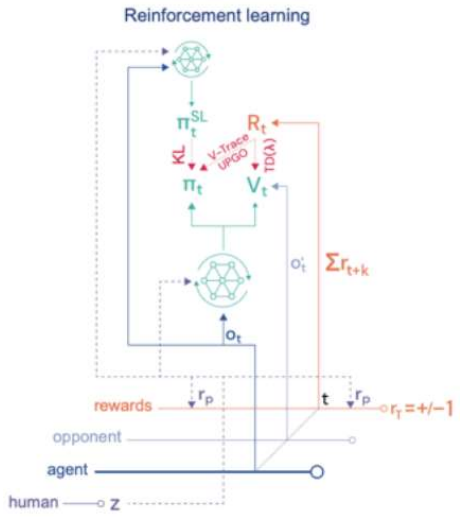
그러나 모조 보상을 통해 사람의 전략을 따르게 에이전트를 학습한다면, 필연적으로 지도학습이 확보한 전략의 다양성을 망각할 가능성이 높다. 알파스타는 이를 방지하기 위해 강화학습으로 학습된 정책과 기존 지도학습의 정책이 차이가 줄어드는 방향으로 학습을 유도했다. 이것은 **알파스타 에이전트가 강화학습을 거칠 때, 넓은 전략의 폭은 유지하면서 프로게이머의 전략을 잘 따라가도록 위한 것이다.** 이 부분에서 숨겨진 가정은 사람의 전략을 따르는 것이 승리할 확률이 높아진다는 것을 의미한다. 과거 알파고에서는 사람의 전략을 고수하는 것이 오히려 시야를 좁히는 역할을 했지만[9], 알파스타에서는 오히려 도움이 된다는 사실이다.

사람의 전략을 따라서 정책을 학습시키고 기존의 전략의 폭은 유지하는 한편, 승리할 확률을 예측하는 가치 신경망 역시 정책을 고도화시키는 역할을 한다. 알파스타의 강화학습의 구조는 큰 틀에서 off-policy A2C(Advantage Actor-Critic) 알고리즘을 활용했다. off-policy의 의미는 특정 정책으로 생성된 데이터가 동시에 그 특정 정책을 개선(update)시키는데 활용되지 않는다는 것이다. 즉, 특정 정책으로 종료된 게임 데이터가 일단 수집되고 난 뒤에 정책의 개선이 일어난다는 점이다. A2C는 대표적인 정책 기반의 강화학습으로 액터는 정책을, 크리틱은 이 정책이 얼마나 좋은지를 평가하는 가치로 이해할 수 있다. 이러한 기본구조 상에 알파스타의 강화학습은 정책과 가치를 개선하기 위해 3가지 알고리즘을 활용했다[10]. 이것은 시간차 학습(Temporal Difference, TD(λ)), off-policy의 학습 효율을 높이기 위한 V-trace, 새로운 자가 모방학습인 UPGO(Upgoing policy update)를 활용했다.

9) 알파고의 개선된 버전인 알파고 제로는 기보를 전혀 학습하지 않고 단순한 바둑의 규칙만으로 학습한 결과다. 그 결과 알파고 제로는 기존 알파고 버전들을 압도하는 실력을 갖추게 됐다. 알파고 제로의 함의는 결국 인간의 기보가 무수히 많은 경우의 수를 편향적으로 좁혔다는 의미다.

10) 강화학습 알고리즘은 그간 이론적인 연구가 활발히 이어져 대부분의 내용이 수학적인 논의로 이루어져 있다. 쉬운 이해를 돕기 위해 이 보고서에서는 이 3가지에 대한 상세내용을 다루지 않고 전반적으로 전개되는 알파스타의 강화학습을 중점적으로 논의했다.

[그림 5]는 알파스타의 강화학습을 개념화한 내용이다.



- [그림 4]와 동일한 모수는 생략
- o_t' : t 시점에서의 관측된 상대방의 입력 값
- r_p : 주어진 전략 z에 대한 모조 보상
- r_T : 종료 시점의 보상 (승리 : +1, 패배 : -1)
- R_t : 최종 보상 (종료 시점 보상과 모조 보상)
- π_t : 강화학습의 정책
- V_t : 강화학습의 가치

[그림 5] 알파스타의 강화학습의 개념도

자료 : Grandmaster level in StarCraft II using multi-agent reinforcement learning

알파스타의 리그학습

알파스타는 강화학습에 이어 멀티-에이전트 강화학습을 수행한다. 멀티-에이전트 강화학습의 목표는 일차적으로 에이전트의 성능 개선에 있고, 강화학습 자체 대결의 단점을 보완하는데 있다. 강화학습에서 수행한 자체 대결의 가장 큰 단점은 전략이 순환(cycle)을 이룬다는 점이다. A라는 전략이 B를 이기고, B라는 전략은 C를 이기는데, C는 다시 A를 이기는 현상이 발생한다. 이를 보완하고자 하는 방법이 바로 리그학습으로 명명된 멀티-에이전트 강화학습이다.

리그학습은 알파스타의 핵심이다. 알파스타의 성능은 리그학습을 거치면서 큰 폭으로 개선됐다. 리그학습의 방식은 자체대결이라는 점에서 강화학습과 유사하지만, 학습 방법은 강화학습과 다르다. 강화학습은 [그림 5]와 같이 지도학습에서 학습한 전략의 폭을 유지하면서 모조 보상을 통해 승률을 높이는 과정이다. 반면, 리그학습은 독립적인 에이전트 간의 대결 결과를 통해 에이전트를 구성하는 인공신경망의 가중치를 갱신하는 방식이다.

리그학습에는 세 가지 유형의 에이전트가 존재한다. 첫 번째는 메인 에이전트(Main Agent)로 지도학습과 강화학습을 거친 결과물을 시작점으로 하여 성능 개선을 목표로 한다. 두 번째는 메인 개척자(Main exploiter)로 메인 에이전트의 약점을 찾는 역할을 한다. 마지막은 리그 개척자(League exploiter)로 리그 자체의 약점을 찾아내고, 새로운 에이전트를 공급하는 역할을 한다.

세 종류의 에이전트 간의 대결은 다양성을 확보하기 위해 우선순위가 부여된 가상 자체 대결(Prioritized Fictitious Self-Play, 이하 PFSP)을 도입했다. PFSP의 직관적인 이해는 특정 에이전트가 상대 에이전트를 선정하는 과정에서 승리하기 어려운 상대를 고르거나 실력이 비슷한 상대를 고르는데 활용되는 일종의 확률이다. 예를 들어 메인 에이전트의 학습을 위한 상대는 메인 에이전트 간의 대결이 35%, 모든 에이전트에서 PFSP로 선정된 에이전트 50%, 나머지 15%는 제거된 에이전트로 구성된다. 리그학습은 세 가지 에이전트 유형 별로 32개의 TPU를 활용해 44일 동안 학습했으며, 그간 900여 개의 에이전트가 새로 생성됐다.

4. 결 론

지도학습과 강화학습, 그리고 리그학습을 거친 알파스타의 에이전트는 배틀넷에서 상위 약 0.2%안에 들어 그랜드마스터에 등극했다. 지난 2019년 1월에 프로게이머 MaNa와의 대결에서 유일하게 1패를 기록했는데, 이것은 알파스타 에이전트가 사람과 동등한 조작화면으로 경기에 임했던 결과였다. 나머지 10승은 맵 전체를 조작화면으로 활용했다는 점에서 형평성 문제가 제기되었다. 일반적으로 맵 전체를 동시에 조작화면으로 활용하는 것이 불가능하다. 사람이 보는 화면에 비춰지는 부분은 전체의 일부이기 때문이다. 그러나 이번 노문을 통하여 밝힌 바로는 모든 경기를 사람과 동등한 조건하에 펼쳐 승리를 거둠으로써 논란이 되었던 형평성 문제를 종식시켰다.

스타크래프트2라는 게임은 고도의 지적 능력을 요구한다. 특히 실시간 조작과 유연한 전략적 대응이라는 점에서 복잡도와 난도는 바둑과 비슷하다. 또한 스타크래프트2 AI 개발을 위한 수많은 도전과제를 해결했다는 점이 가장 큰 성과다.

게임을 기반으로 한 AI의 개발은 다양한 장점이 있다. 먼저 데이터를 충분히 확보할 수 있다는 점인데, 현대 AI의 핵심인 심층학습의 고질적인 문제인 데이터 부족을 해결할 수 있다. 게임의 보상은 현실세계의 보상보다 명확하다는 점에서 강화학습을 구현하는데 이점이 있다. 또한 대부분 컴퓨터를 활용해 게임이 가능하기 때문에 사람들과의 대결로 성능을 가늠하기 쉽다. 또한 바둑과 스타크래프트2와 같이 고도의 지능적 행동을 요구하는 게임의 경우 게임 AI를 성공적으로 개발한다면, AI 기술의 전반적인 발전에 기여할 것이다.

게임 AI 개발은 R&D 측면에서도 의미가 있다. 게임이라는 영역을 조금 더 확장시키면, 시뮬레이션 환경에서 지능적 행동이라고 볼 수 있다. R&D에서 개발하고자 하는 AI 기술을 시뮬레이션 환경으로 구현하는 사례는 쉽게 찾아볼 수 있으며[11], 개발 도구를 공개하여 많은 연구자의 참여를 유도하고 있다. 연구자 참여를 독려하는 이유는 원활한 데이터의 수급과 상호간의 대결을 통해 성능을 측정할 수 있기 때문이다. 현재 AI의 성능은 데이터의 절대량에 의존성이 강하

11) OpenAI의 고전 게임 AI 개발 환경인 Gym API, 숨바꼭질(Hide and Seek), 구글의 축구 전략 시뮬레이션 환경 등 AI 기술의 혁신이 시뮬레이션 환경을 통한 게임 AI 개발에서 일어나고 있다.

다. 이를 알고리즘적으로 극복하기 위한 연구는 연구실 수준의 실험적 결과가 대부분이라는 점에서 시뮬레이션 환경 기반의 AI 연구가 갖는 장점이 크다고 볼 수 있다. 우리나라는 지난 2019년 12월 『인공지능 국가전략』을 발표하여 AI 기술 경쟁력 확보를 위한 청사진을 제시했다. 특히 차세대 AI 기술의 선점을 위한 예비타당성조사를 추진한다고 밝혔는데, 알파스타에서 찾을 수 있는 AI R&D의 장점을 십분 반영하여 과감한 R&D 사업이 추진되어야 할 것이다.

알파스타는 과거 알파고의 행적에 빗대어 보자면 아직 개선의 여지가 남아있다. 알파스타의 논문에서는 전반적으로 사람의 전략을 십분 활용했는데, 이것은 사람의 전략이 스타크래프트2에서 이길 수 있는 성공 방정식이라는 것을 실험적으로 증명했다고 본다. 과거 알파고 제로는 사람의 개입이 전혀 없이 규칙만으로 자체 대결을 하여 학습된 AI로 바둑을 정복했다. 이것은 바둑기사의 기보를 학습했던 알파고가 오히려 사람으로 인한 편향을 갖게 됐다고도 볼 수 있다. 알파스타가 향할 미래는 어떠한 방향일지는 추측되지 않지만, AI의 가능성을 더욱 높일 수 있을 것이라 전망된다.

[참고문헌]

1. 국외문헌

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

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

Vinyals, Oriol, et al. “Grandmaster level in StarCraft II using multi-agent reinforcement learning.“ Nature (2019): 1-5.

Vinyals, Oriol, et al. “Starcraft ii: A new challenge for reinforcement learning.“ arXiv preprint arXiv:1708.04782 (2017).

2. 국내문헌

추형석, “알파스타의 인공지능 알고리즘”, SW산업동향, 소프트웨어정책연구소, 2019.

추형석, “AlphaGo Zero의 인공지능 알고리즘”, 이슈리포트 2017-009, 소프트웨어정책연구소, 2017.

추형석, “AlphaGo의 인공지능 알고리즘 분석”, 이슈리포트 2016-002, 소프트웨어정책연구소, 2016.

3. 기 타

AlphaStar: Mastering the Real-Time Strategy Game StarCraft II, Deepmind (2019.01.)
https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/
2019.10.31. 접속

주 의

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