AlphaGo에서 AlphaStar까지, 그간의 기술 변화와 정책 동향

2019. 04. 12.
SPRi Spring Confernece

추 형 석

hchu@spri.kr

소프트웨어정책연구소 선임연구원

기술・공학 연구실

소프트웨어 중심사회의 Think Tank SPRi Software Policy & Research Institute

목 차

1. AlphaGo의 변천사

AlphaGo 버전 정리

- **AlphaGo Fan**
  - 2016년 1월 네이처 논문지에 게재된 AlphaGo 버전
  - 최대 1,920개의 CPU, 280장의 GPU를 활용한 결과
  - 합성곱신경망(지도학습 + 강화학습) + 몬테-카를로 트리 탐색 (MCTS)를 활용

- **AlphaGo Lee**
  - 2016년 3월 이세돌 9단과 대국한 버전
  - 실제 대국에는 48장의 TPU가 활용

- **AlphaGo Master**
  - 2017년 1월까지 Tygem 온라인 바둑에서 60전 전승
  - AlphaGo Lee의 개선된 버전
  - 2017년 5월 커제 9단과 대결

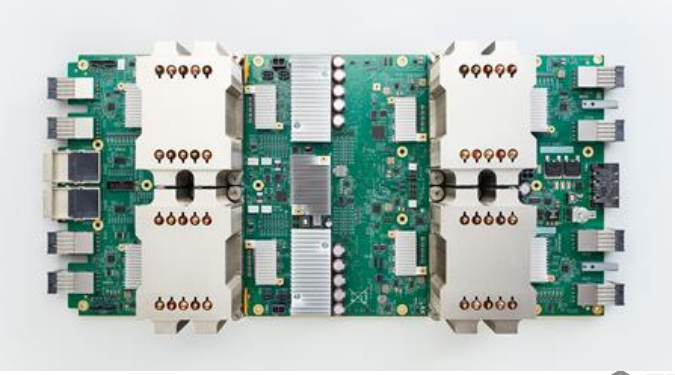
AlphaGo Zero의 등장

- 인간의 기보를 전혀 활용하지 않고 바둑 규칙만으로 학습한 결과

- 대국시 TPU 4장을 활용 (PC 한대 수준의 전력 소모)

※ TPU (Tensorflow Processing Units)

- TPU는 구글이 지난 2014년부터 개발한 인공지능 전용 HW로 학습(Training)과 추론(Inferencing)에 최적화
- 구글은 TPU의 성능을 분석한 논문을 발표하여 기존 연산처리장치 대비 최대 80배의 전력소비를 절감
- TPU의 핵심은 학습 기반의 인공지능에서 가장 빈번하게 발생하는 행렬 곱 연산에 최적화된 일종의 가속기

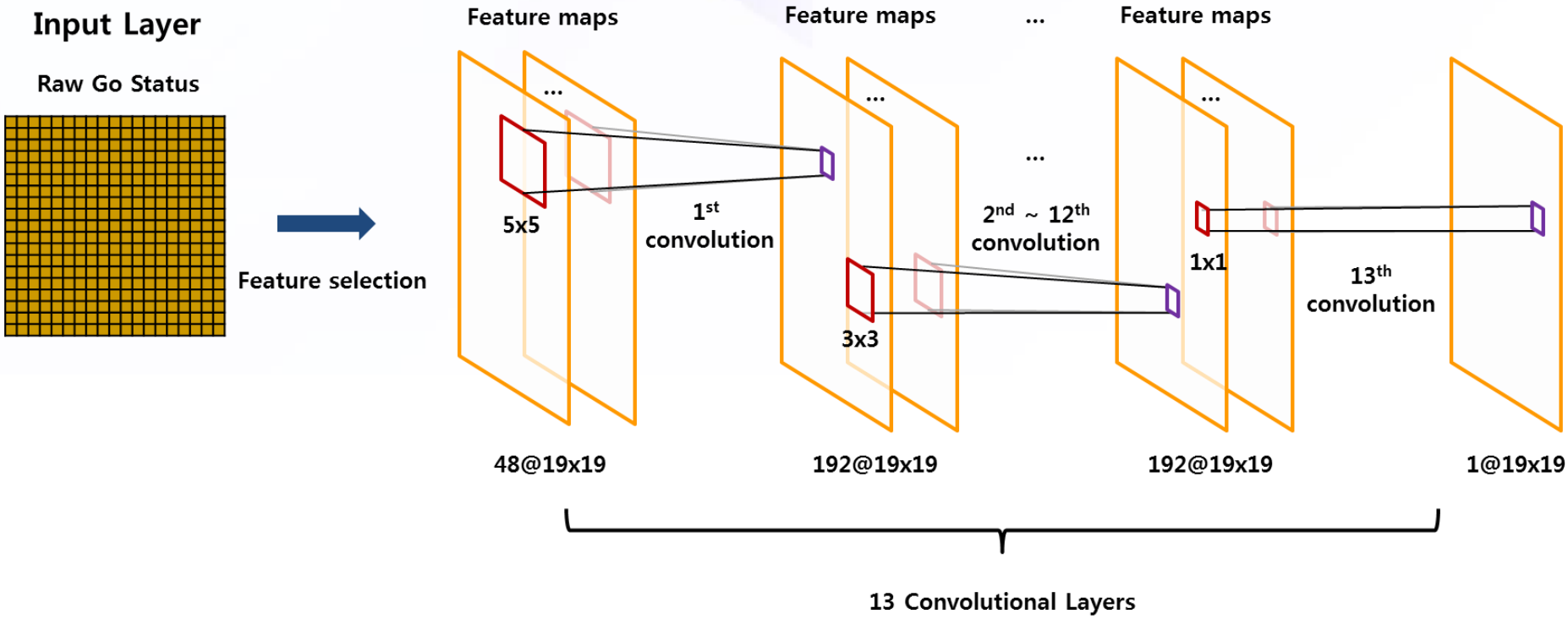


출처 : Build and train machine learning models on our new Google Cloud TPUs, Jeff Dean (2017) https://blog.google/topics/google-cloud/google-cloud-offer-tpus-machine-learning/

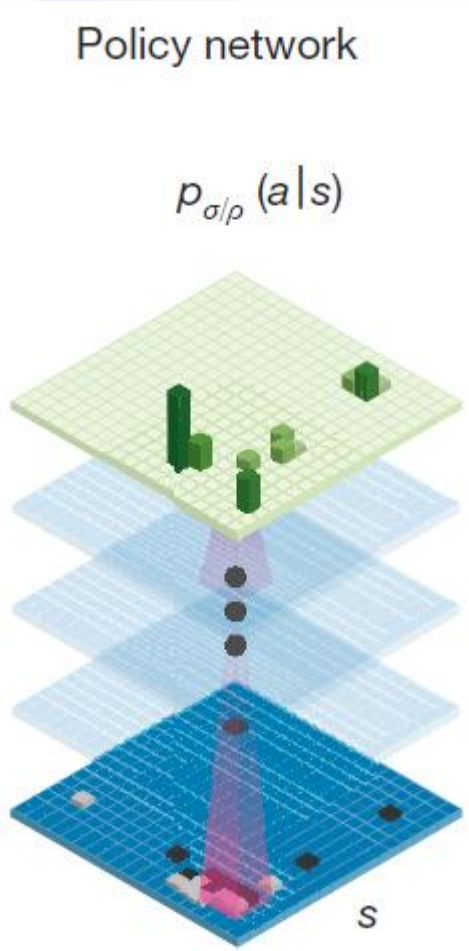
AlphaGo Zero의 네 가지 차별점

1. 무작위 방식(random play)의 자체 대국을 통해 강화학습 활용
2. 흑돌과 백돌 두 가지만 특징맵 활용
3. 인공신경망의 형태와 구조의 변화
4. 간단한 트리 탐색 기법 활용

(리뷰) AlphaGo Lee의 합성곱 신경망 구조

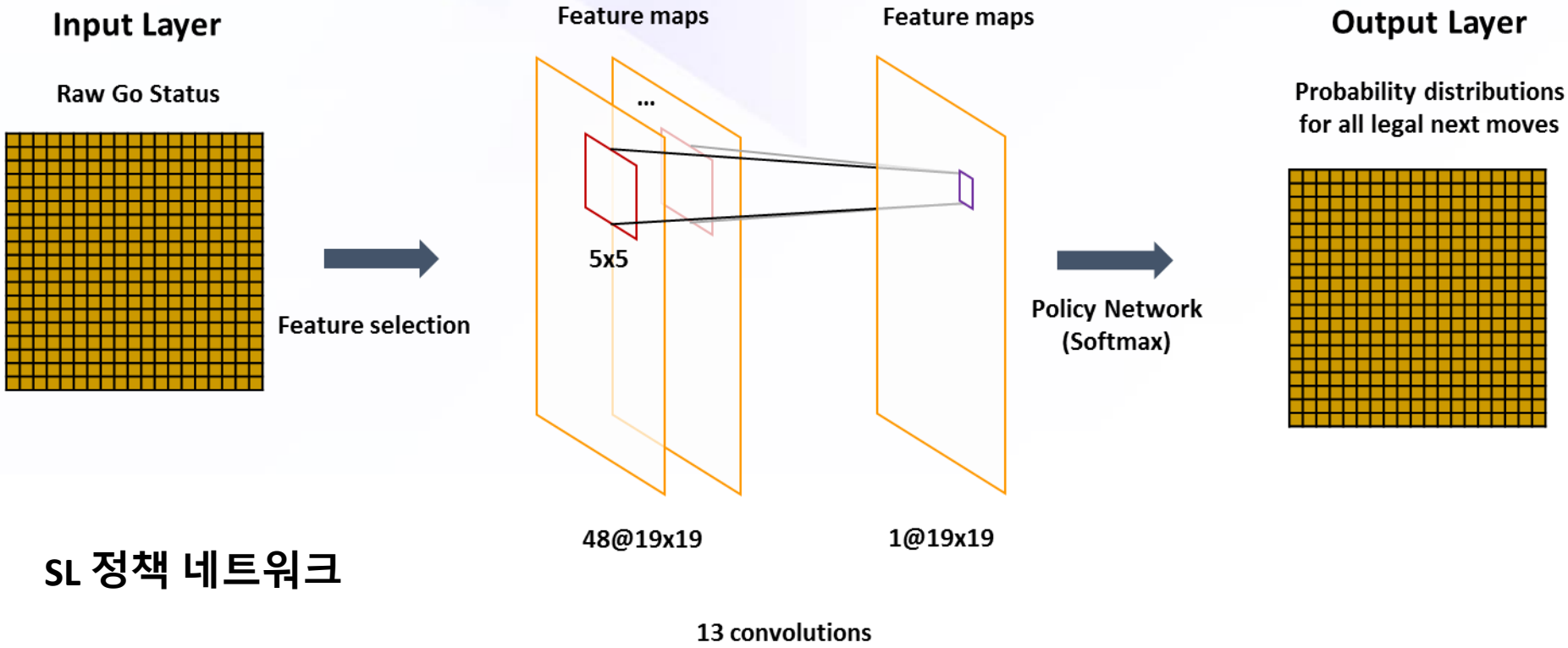


(리뷰) 정책망



Source: Mastering the game of Go with deep neural networks and tree search, Nature

(리뷰) 정책망 – 프로바둑기사의 선호도 학습

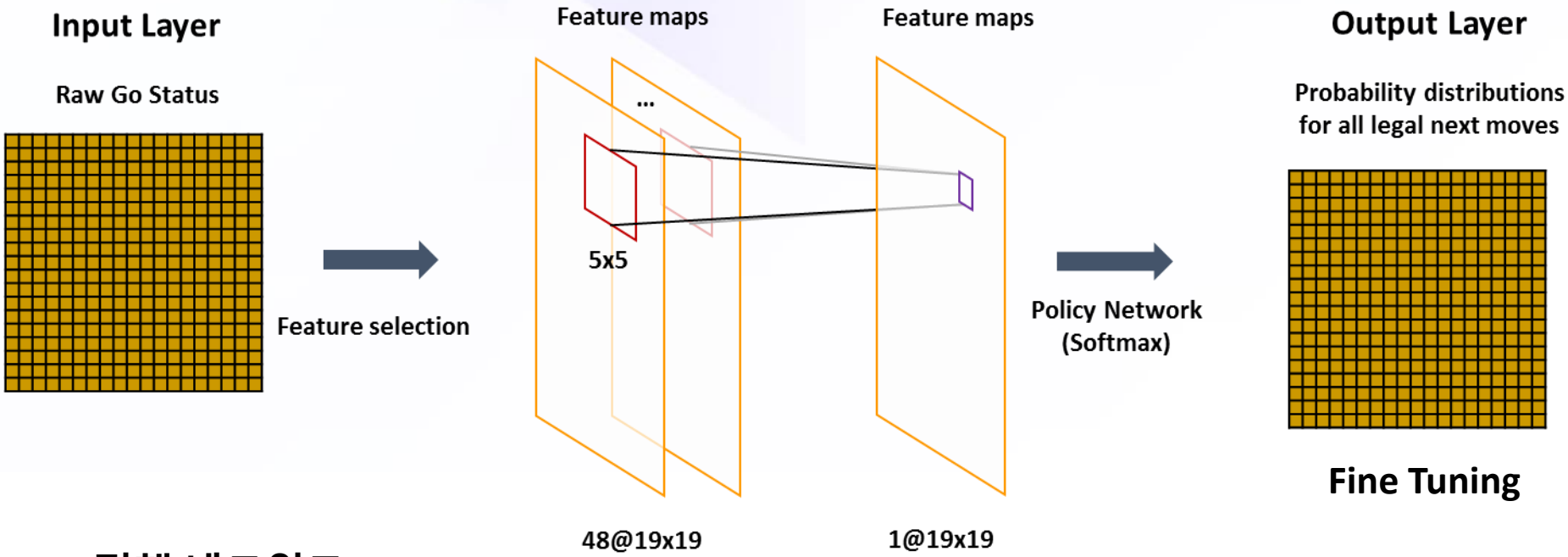


SL 정책 네트워크

데이터 : KGS 6~9단 기보 16만 개에서 바둑판 상태 약 3천만 개

계산비용 : 50 GPUs, 3 주

(리뷰) 정책망 – 스스로 경기하여 성능향상

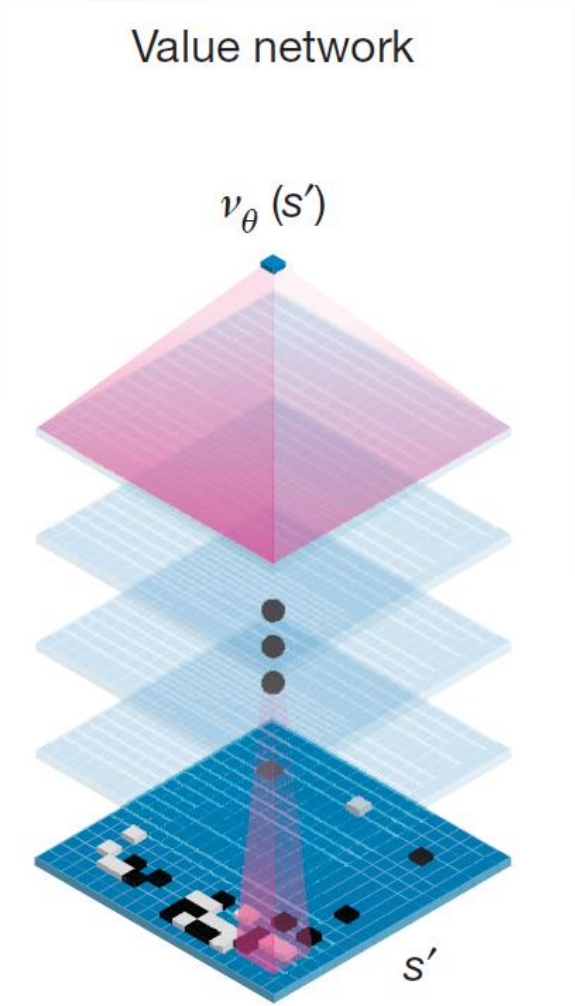


RL 정책 네트워크

데이터 : 128만 번의 게임을 수행하면서 착수전략을 강화함

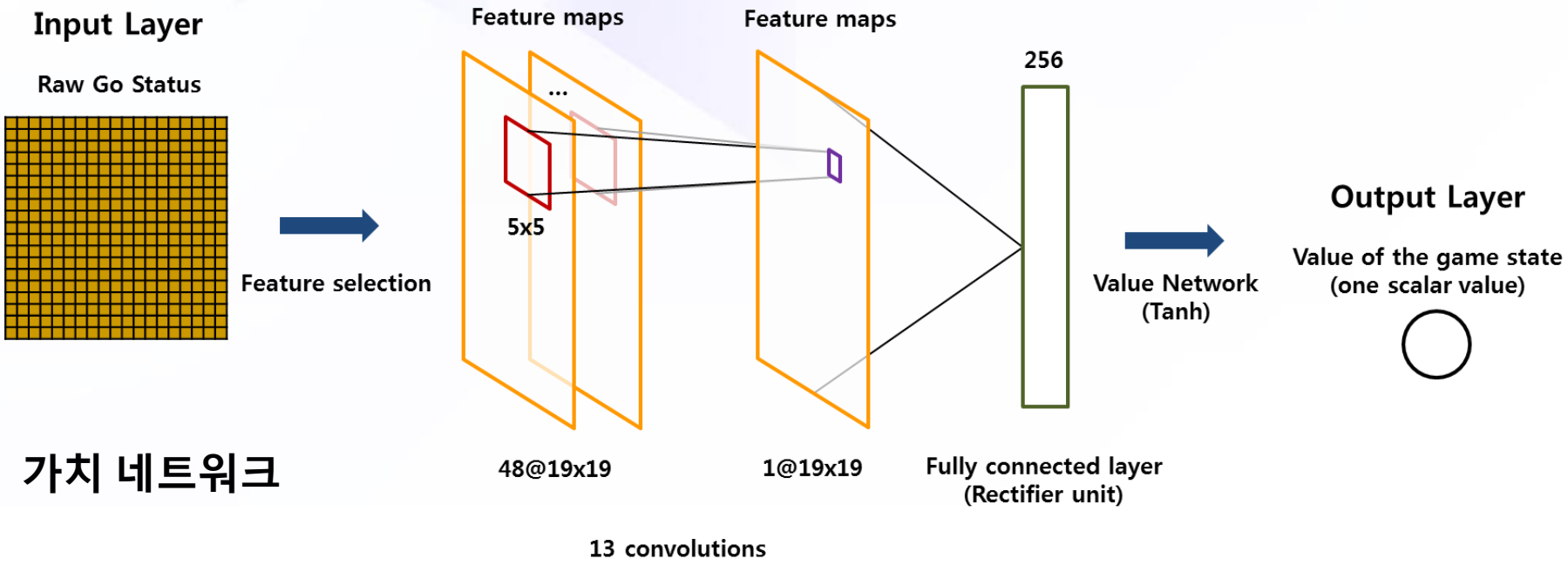
계산비용 : 50 GPUs, 1 일

(리뷰) 가치망



Source: Mastering the game of Go with deep neural networks and tree search, Nature

(리뷰) 가치망 – 바둑판 상태의 승률 계산



가치 네트워크

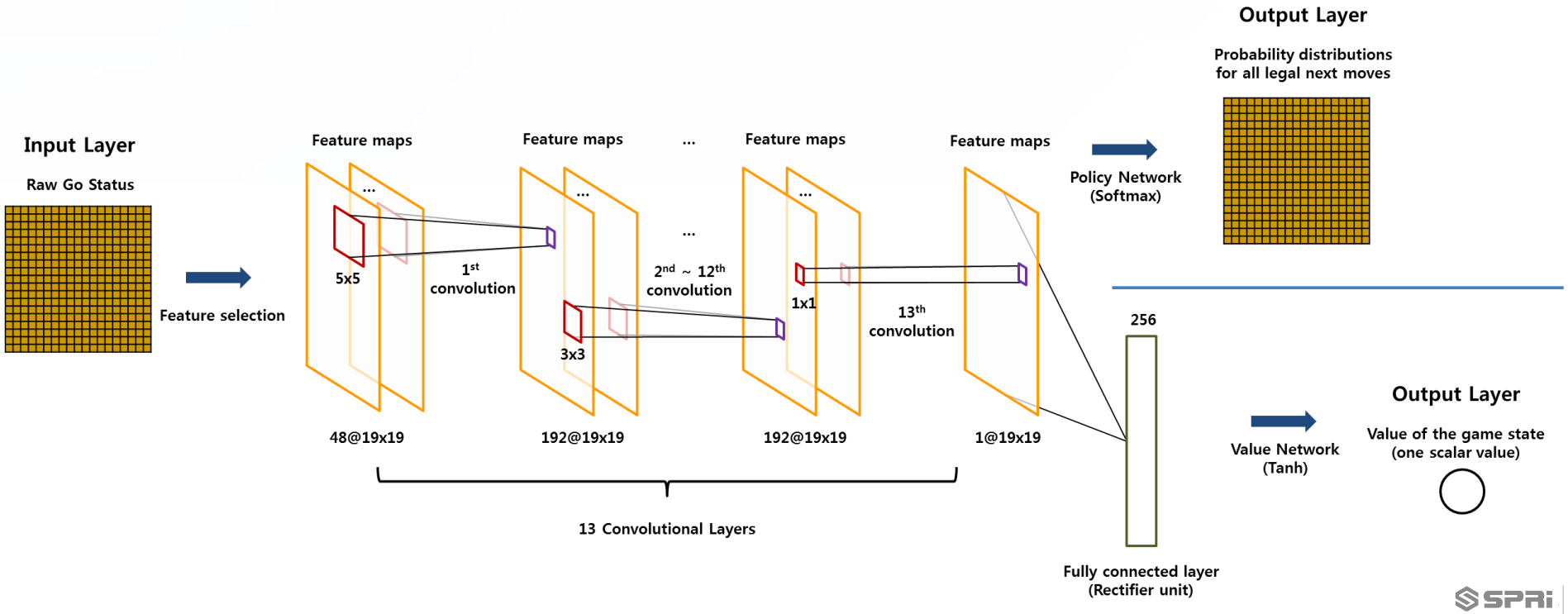
데이터 : 3천만 개의 모의전 (AlphaGo가 직접 생성)

계산비용 : 50 GPUs, 1 주

인공신경망 형태와 구조의 변화

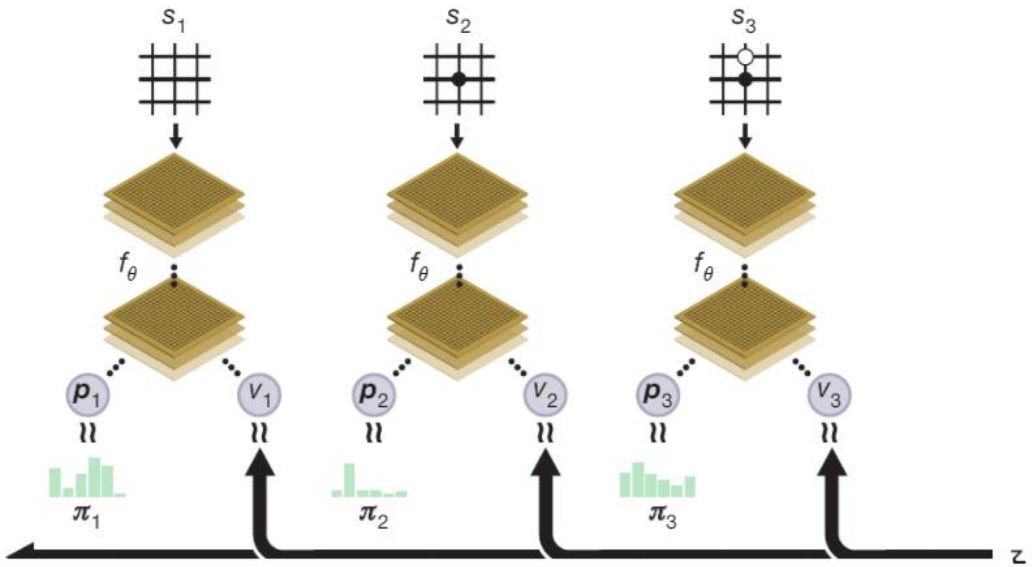
- **정책망과 가치망의 통합**
  - 정책망과 가치망의 역할은 궁극적으로 대국에서 이기기 위함

- **40층의 잔차신경망 활용 → 이미지 인식 성능의 개선**



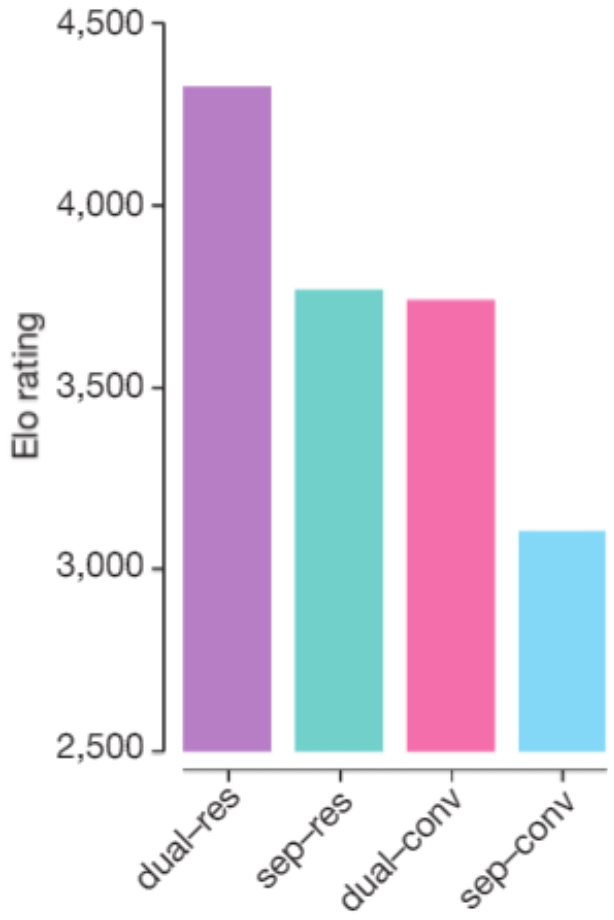
학습 방법의 개선

- **자체 대국의 결과를 정책망/가치망 학습에 활용**
  - **AlphaGo Master에서 시도됐으나, AlphaGo Master는 여전히 인간의 기보를 활용**



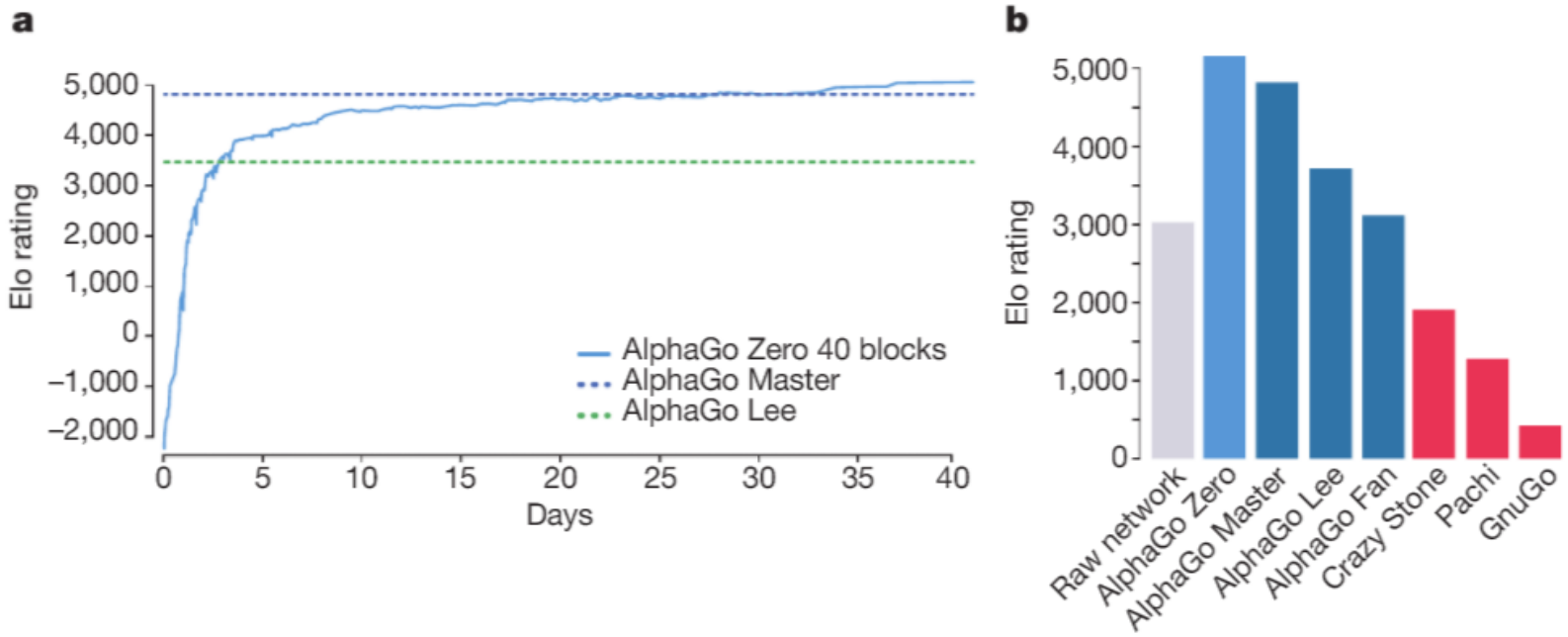
출처 : Silver, David, et al. "Mastering the game of go without human knowledge." *Nature* 550.7676 (2017): 354.

다양한 인공신경망 형태와 구조



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

AlphaGo Zero의 성능



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

AlphaZero

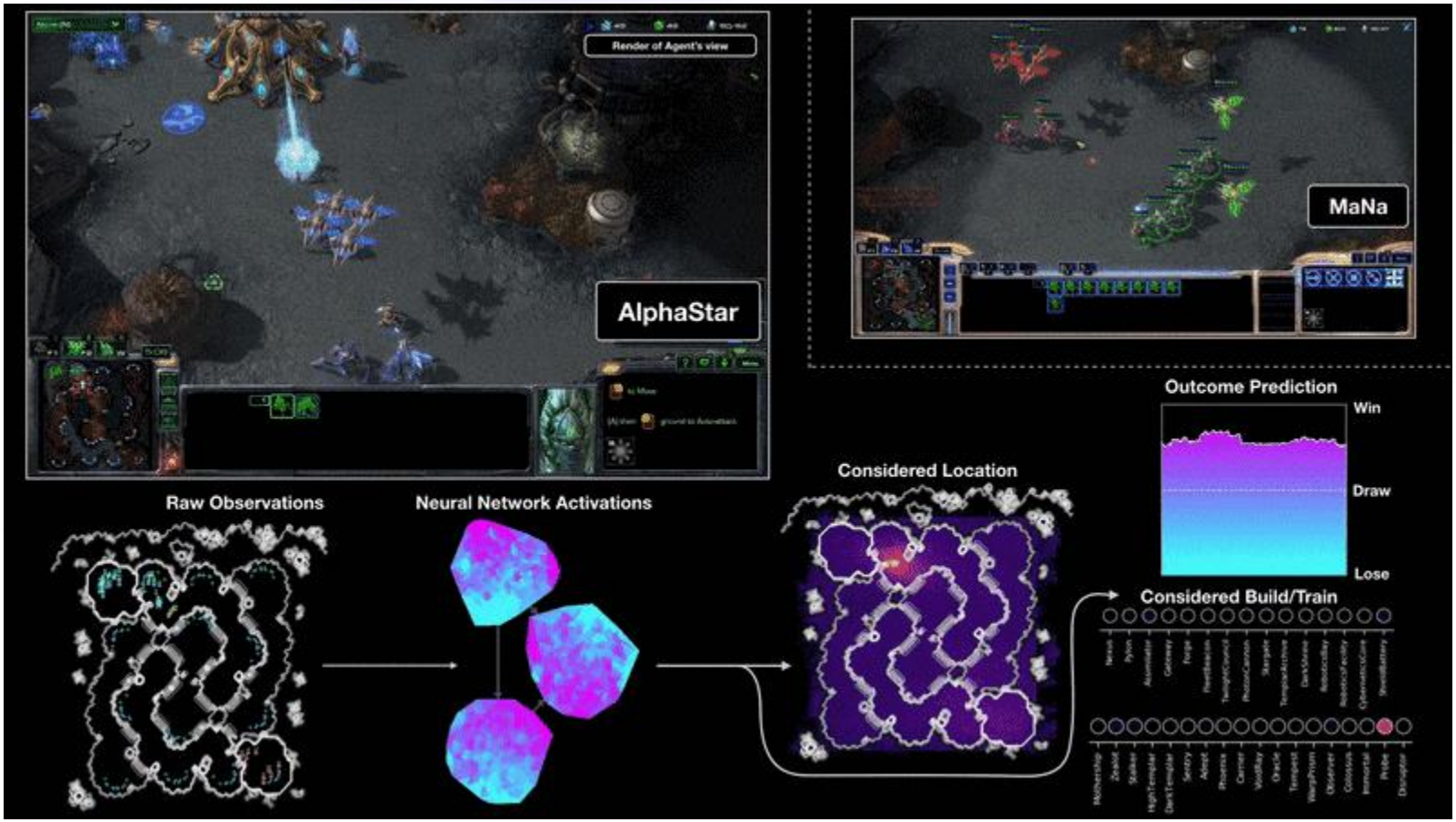
- AlphaGo Zero의 접근방법을 일반화
- 체스와 쇼기(일본식 장기)에 적용하여 압도적으로 승리
- 게임 규칙만으로 데이터를 스스로 생산하여 학습

2. AlphaStar의 부상

AlphaStar의 부상

- AlphaStar : 실시간 전략 시뮬레이션 게임 스타크래프트2(이하 SC2) 인공지능
- AlphaGo 개발진인 구글 딥마인드는 2016년 이세돌 9단과의 대국 이후, 후속 연구로 SC2 인공지능 개발을 천명 (2016.03.)
- SC2 인공지능 개발을 위한 데이터와 개발 도구 SC2LE(StarCraft II Learning Environment) 공개 (2017.08.)
  - 게임 리플레이 데이터 6만 5천 건, SC2 게임에 접근할 수 있는 API PySC2 공개
  - “StarCraft II: A New Challenge for Reinforcement Learning” 논문 공개
- SC2 프로게이머와의 대결에서 10승 1패의 성적을 거둠 (2019.01.)
  - Team Liquid의 “MaNa”와 대결 5승 1패, “TLO”와 대결 5승
  - 상세 내용이 담긴 논문은 리뷰 중으로 차후 공개 예정
- 현 시점까지 공개된 정보를 바탕으로 AlphaStar의 알고리즘 분석

AlphaStar의 인공지능



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

바둑 VS 스타크래프트2

구 분	바 둑	스타크래프트2
장르	보드 게임	실시간 전략 시뮬레이션
게임 진행	턴 방식	실시간 명령
게임 공간	19×19 격자 공간 (총 361개)	한 스텝의 행동 기준 10^8 가지의 조합 공간
소요 시간	1 ~ 4시간	10분 ~ 1시간
상대방의 상황	모두 공개	정찰을 통해 습득

AlphaStar의 도전과제

① 게임 이론 (Game theory)

- 가위-바위-보 게임과 같이 하나의 최상의 전략은 없음

② 불완전한 정보 (Imperfect information)

- 상대방의 정보는 정찰이라는 수단을 통해 획득 가능, SC2의 전략은 상대방의 정보를 통해 자신의 전략을 수정 및 고도화하는 방향으로 진행

③ 장기 계획 (Long term planning)

- 실 세계의 문제와 같이 원인과 결과가 즉각적으로 반영되지 않음

④ 실 시간 제어 (Real time control)

- 연속적인 동적 조작을 통해 게임이 진행됨

⑤ 넓은 조작 공간 (Large action space)

- 하나의 행동을 결정하기 위해 산술적으로 약 10^8 가지의 조합 공간을 가지며, 일반적으로 유효한 행동을 10 ~ 26개를 선정해야 함

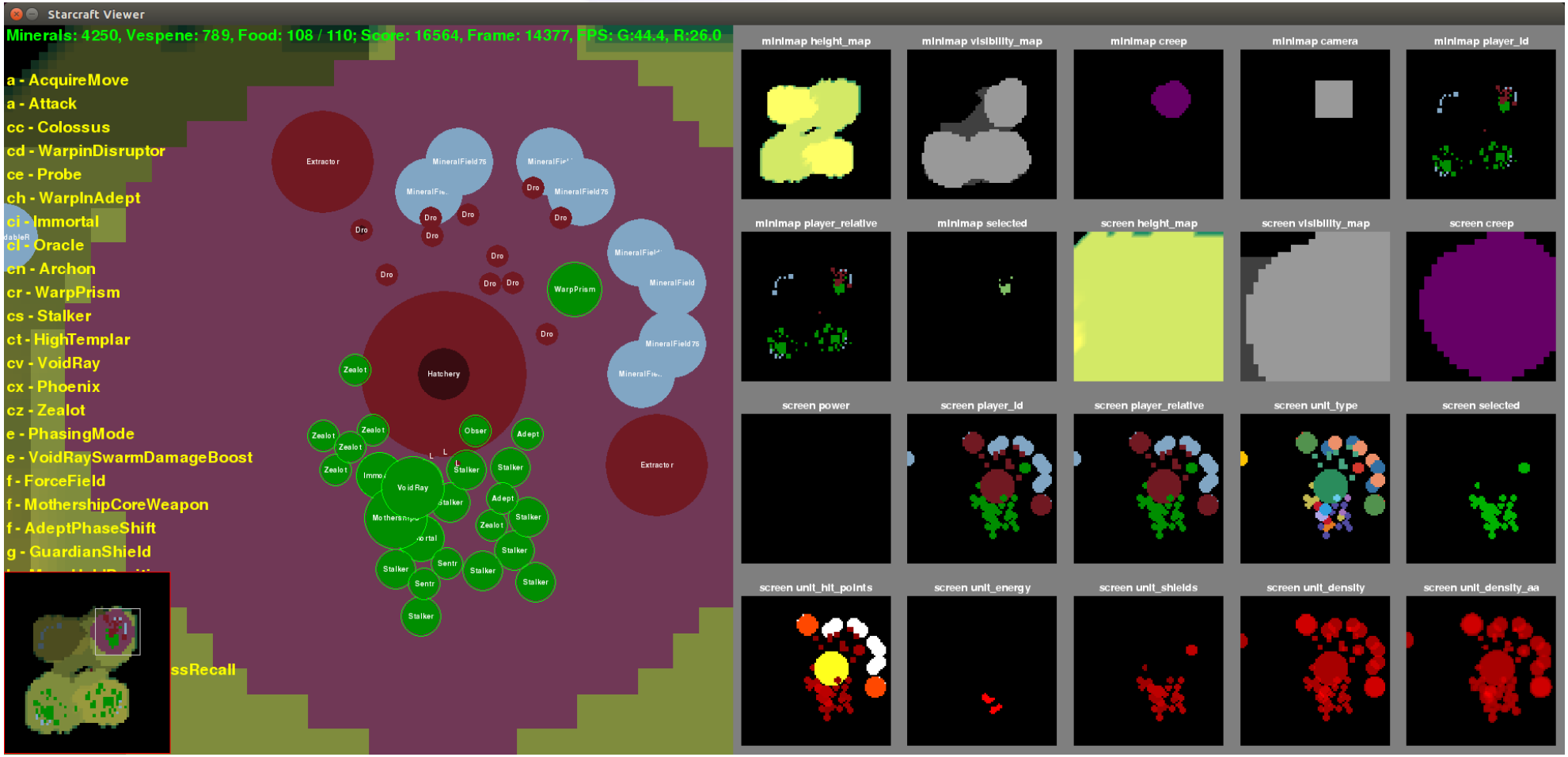
AlphaGo (Lee) VS AlphaStar

구 분	AlphaGo (Lee)	AlphaStar
학습 데이터	16만 건의 대국	80만 건 이상의 리플레이 데이터
입력 (inputs)	48개의 특성으로 나눠진 기보 (예 – 흑돌, 백돌, 빈칸 위치, 단수, 축, 꼬부림 등)	PySC2를 활용한 이미지 데이터 (예 – 현재 시야, 건물, 유닛 등)
출력 (outputs)	착수 가능 지점의 확률, 승리할 확률	일련의 행동 (10개 ~ 26개), 승리할 확률
학습 방법	기보를 통한 지도학습과 자체 대결을 통한 강화학습	리플레이 데이터의 지도학습과 자체 대결을 통한 강화학습
학습 구조 (architecture)	합성곱신경망, 몬테-카를로 트리 탐색(MCTS)	합성곱신경망, 장단기기억(LSTM), 어텐션, 포인트네트워크 등

AlphaStar의 차별점

- **리플레이 데이터를 활용한 학습**
  - SC2는 실시간 조작이기 때문에, 일련의 연속적인 행동(유닛 생산 및 이동, 자원 수집 등)이 출력값
  - 일련의 연속적인 행동은 일종의 문자열(sequence)로 표현할 수 있으므로, 기계 번역이나 언어 모델링에서 활용되는 sequence-to-sequence 기술이 활용
  - 또한 장기적인 전략을 구축하기 위해, 정보를 저장하는 메모리 모델이 활용
- **자체 대결을 통한 전략의 고도화**
  - 리플레이 데이터를 학습하여 생성된 인공지능 에이전트를 자체 대결시켜 성능 개선
  - AlphaStar League로 명명된 자체 대결은 새로운 강화학습 기술을 대거 채용하여 전략을 더욱 고도화한 새로운 에이전트 생성에 기여
  - 각 에이전트는 고유의 목적을 보유 (예 – 에이전트001은 에이전트002만을 이기는 것을 목표로 함, 에이전트003은 모든 에이전트에게 승리하는 것을 목표로 함)
  - 14일 간의 자체 대결 수행 (with 16 TPU v3), 200년 치에 해당하는 플레이 시간

AlphaStar의 입력

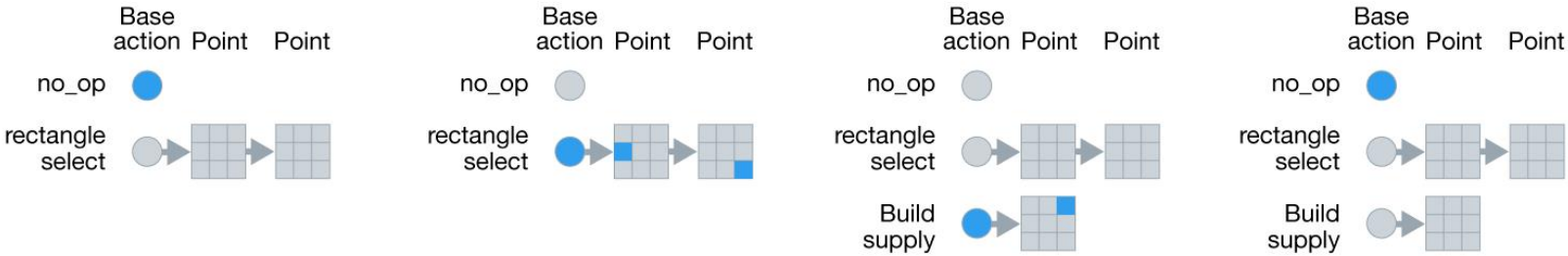


출처 : Vinyals, Oriol, et al. "Starcraft ii: A new challenge for reinforcement learning." *arXiv preprint arXiv:1708.04782* (2017).

AlphaStar의 출력

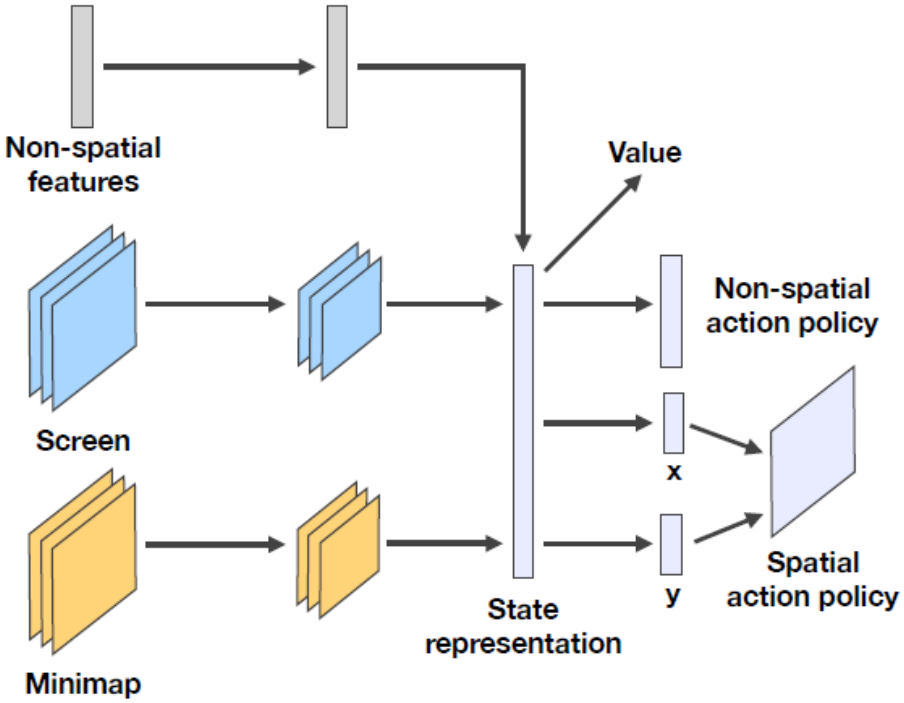
Human Actions	IDLE	Left_Click_Hold (p1) Release (p2)	Press B + S Left_Click (p3)	IDLE
Agent Actions	no_op	select_rect(p1, p2)	build_supply(p3)	no_op

Available Actions

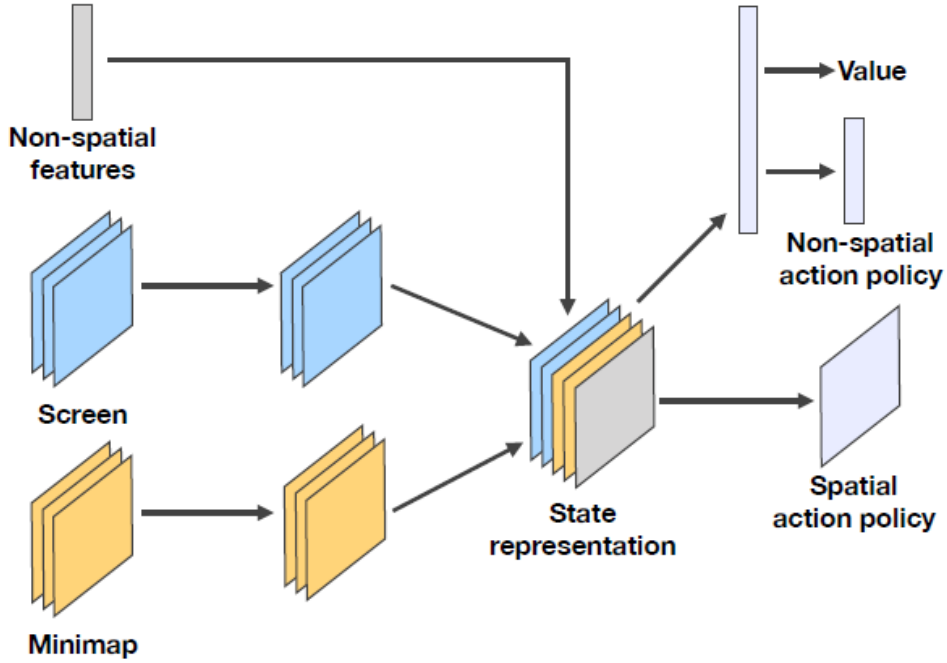


출처 : Vinyals, Oriol, et al. "Starcraft ii: A new challenge for reinforcement learning." *arXiv preprint arXiv:1708.04782* (2017).

AlphaStar의 인공신경망 (초기)



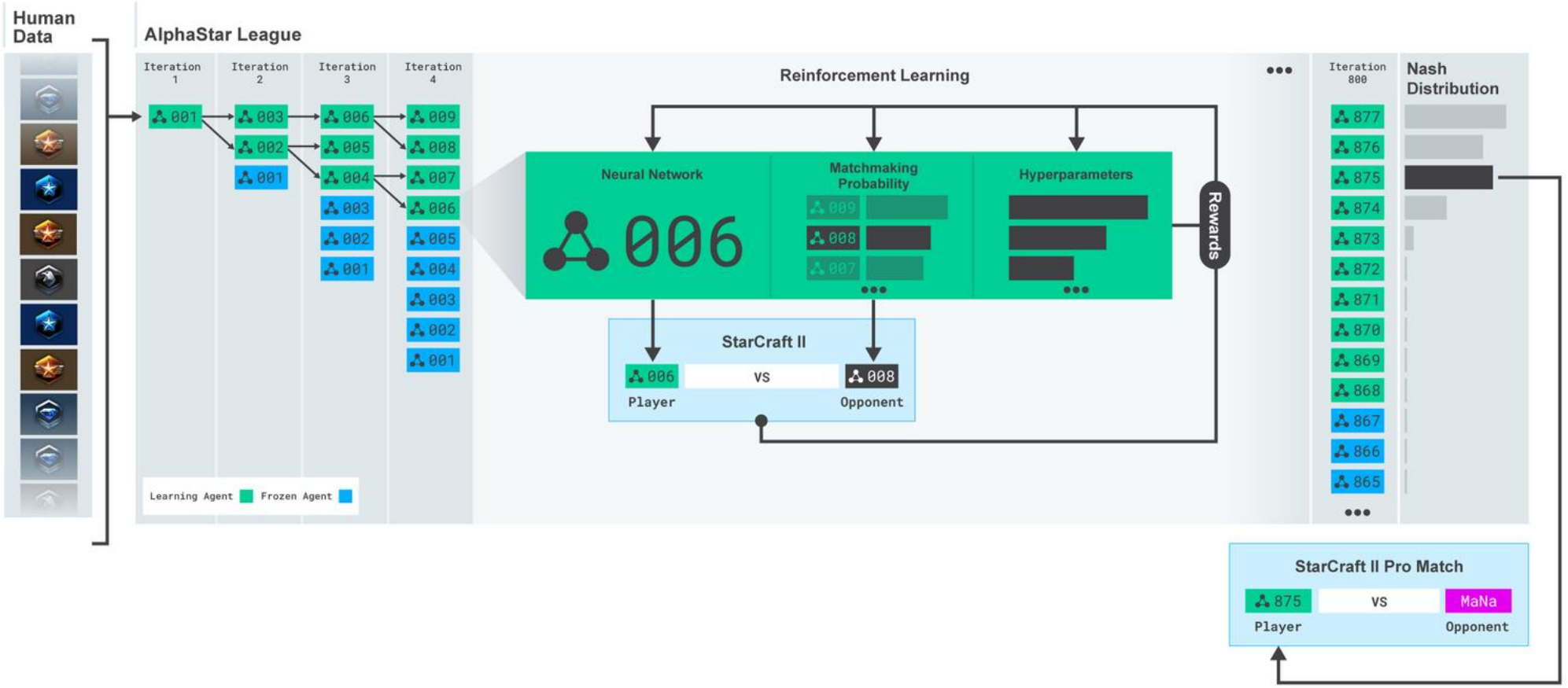
(a) Atari-net



(b) FullyConv

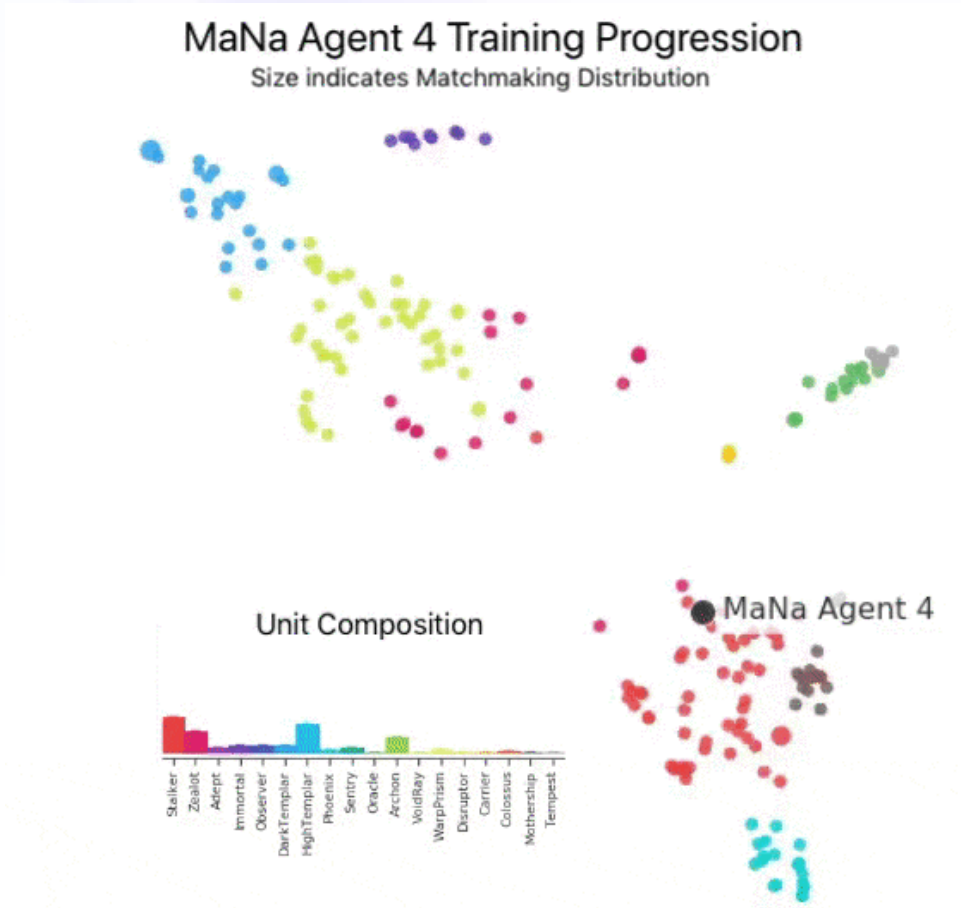
출처 : Vinyals, Oriol, et al. "Starcraft ii: A new challenge for reinforcement learning." *arXiv preprint arXiv:1708.04782* (2017).

AlphaStar 리그



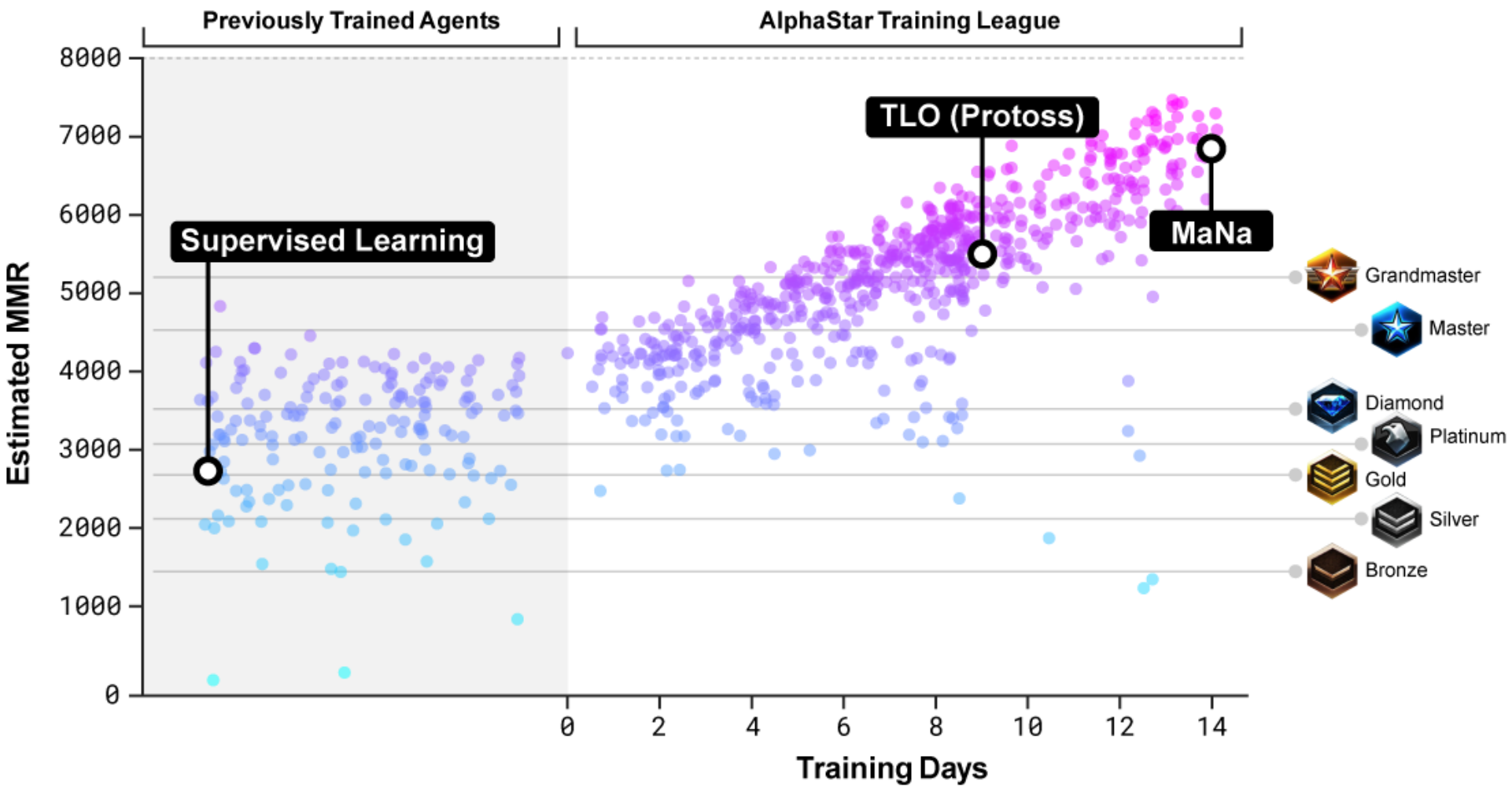
출처 : AlphaStar: Mastering the Real-Time Strategy Game StarCraft II, Deepmind, https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/

AlphaStar 리그의 학습과정



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

AlphaStar의 성능



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

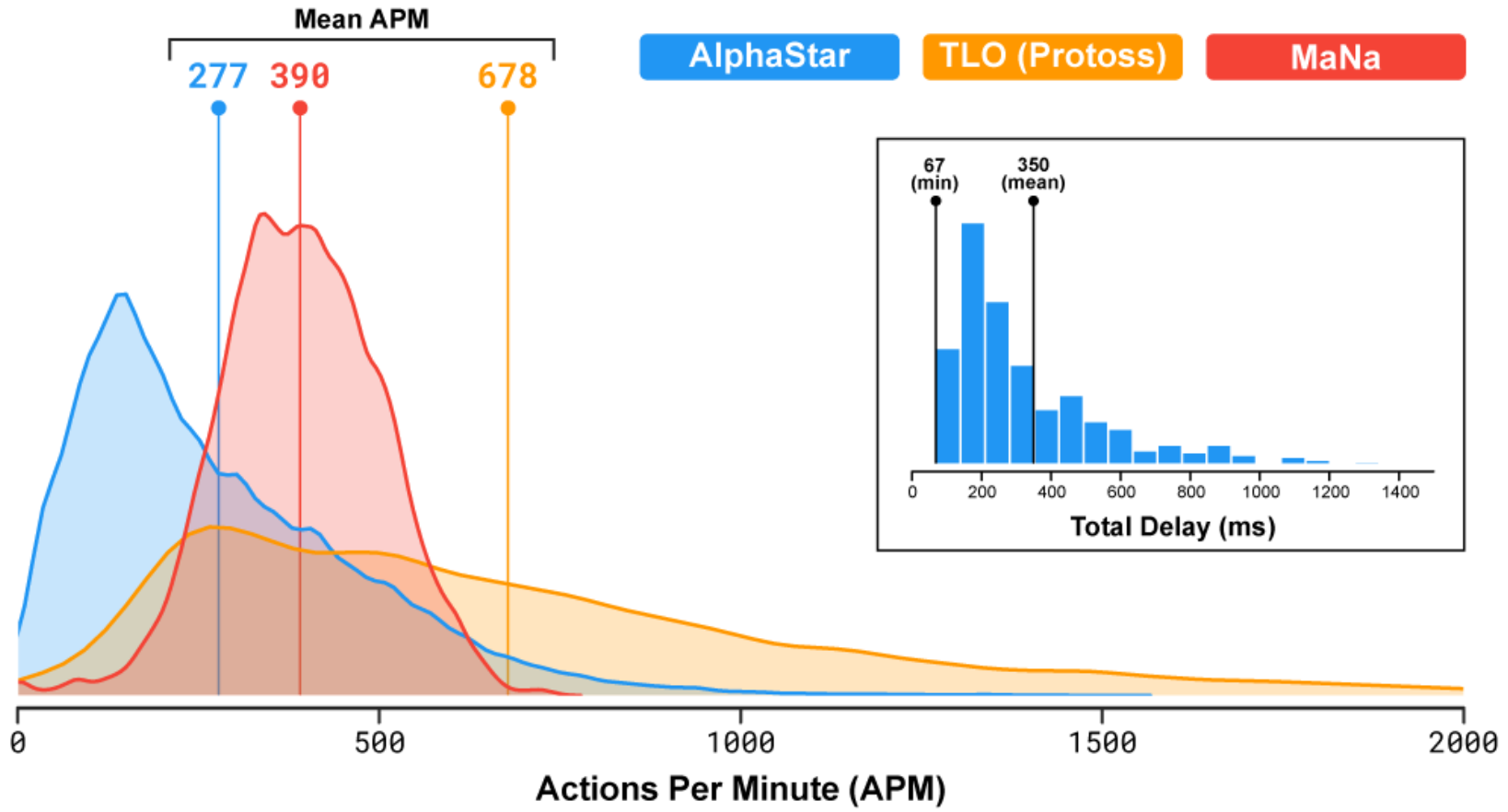
AlphaStar에 활용된 인공지능

- **리플레이 데이터 학습 → Gold 레벨**
  - **Transformer : 어텐션 메커니즘을 구현한 인공신경망 구조**
    - Relational deep reinforcement learning과 유사
  - **장단기기억(Long Short Term Memory, LSTM) : 시계열 데이터의 학습**
  - **Auto-regressive policy head : 계산을 줄이는 역할**
  - **Pointer network : 행동의 수를 유동적으로 산출 (10 ~ 26개)**
  - **Centralised value baseline : 유닛의 전투를 최적화하기 위한 기법**
- **알파스타 리그**
  - **Population-based reinforcement learning : 에이전트 모수(hyperparameter)의 최적화 기법**
  - **Multi-agent reinforcement learning : 서로 독립적인 다수의 에이전트의 장점을 통합하는 기법**
  - **Off-policy actor-critic, experience replay, self-imitation learning, policy distillation, nash distribution of the league**

AlphaStar에 활용된 인공지능

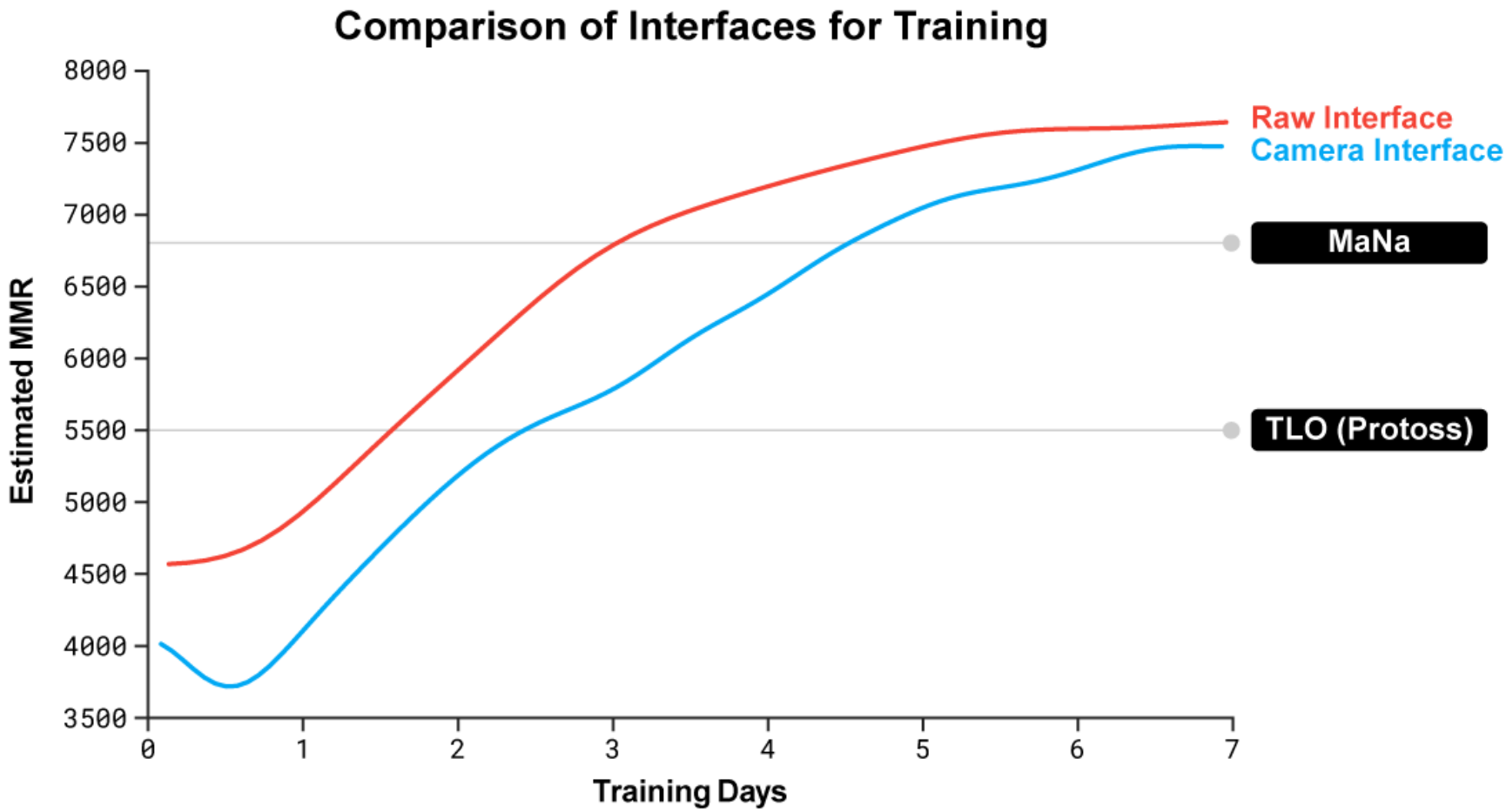
역 할	알고리즘
학습 성능 향상	Transformer, LSTM, Pointer network, Centralised value baseline
자체 대결 결과 고도화	Population-based reinforcement learning, Multi-agent reinforcement learning, Nash distribution of the league
계산 효율 향상	Auto-regressive policy head, Experience replay, Self-imitation learning, Policy distillation

AlphaStar와 프로게이머의 대결



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

AlphaStar와 프로게이머의 대결



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

AlphaStar의 의의

- **인공지능의 “Grand Challenge” 해결**
  - **바둑과는 다른 차원의 복잡성을 갖는 문제**
  - **기존 SC2 인공지능의 접근방법인 규칙 기반(rule-based)에서 탈피**
  - **프로게이머와의 대결에서 공정성 확보**
    - 프로게이머보다 낮은 분당 행동수
    - 실제 대결에서는 데스크톱 GPU 한 대 수준의 계산자원 활용

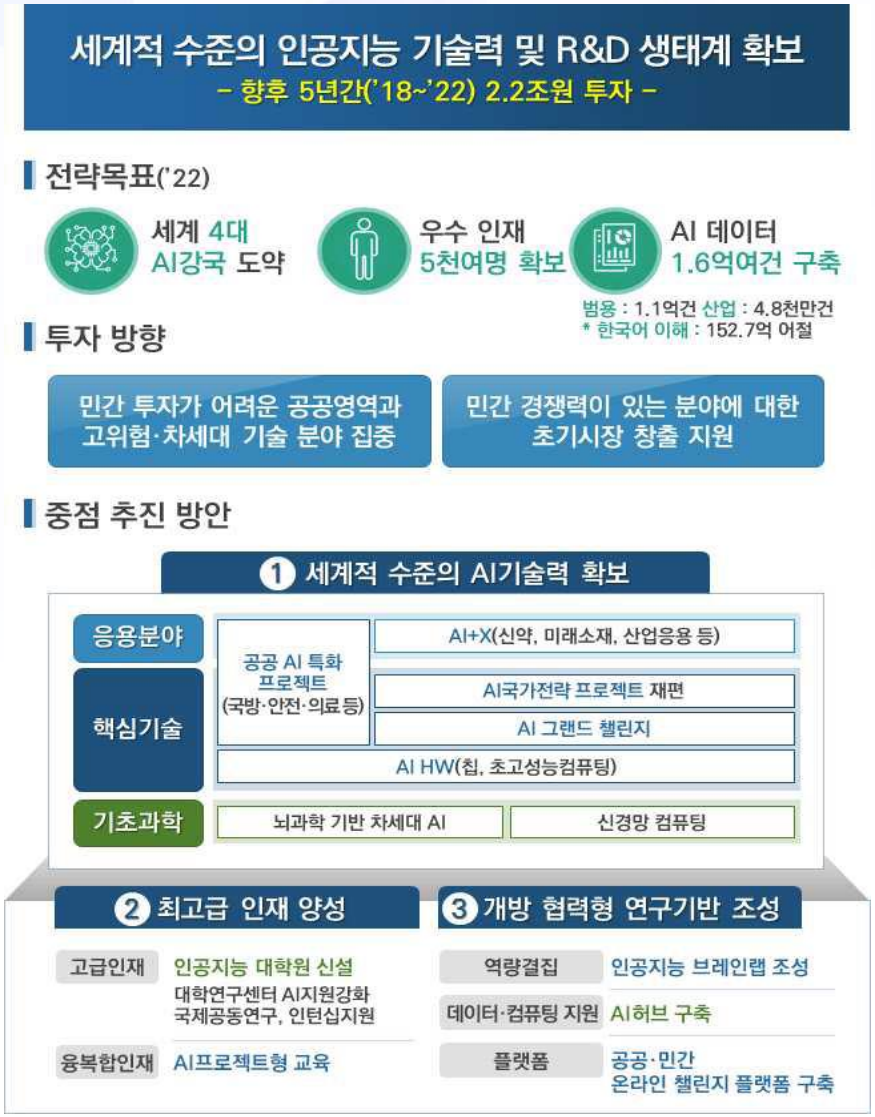
- **게임을 통한 인공지능 개발**
  - **승리하기 위한 목적이 분명하며, 풍부한 데이터를 심층학습에 활용 가능**
  - **어디서나 동일한 인터페이스를 활용하고 다른 연구자들과의 교류가 용이**
  - **프로게이머와의 대결로 성능 테스트 가능**
  - **게임은 일종의 시뮬레이션이므로 상세한 조작이 가능**

3. 국내 인공지능 정책 동향

AlphaGo 이후의 인공지능 정책 현황

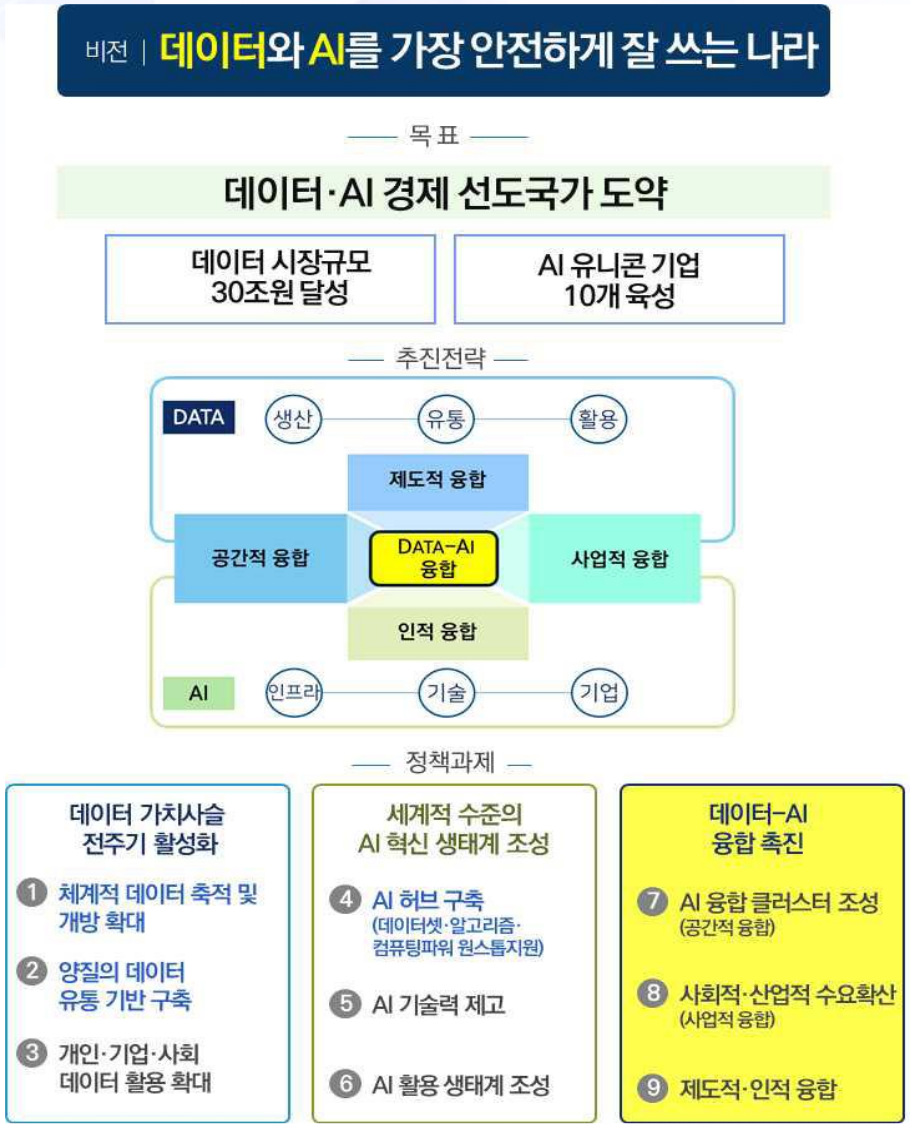
- 지능정보산업 발전전략, 청와대 (2016.03.)
- 지능정보사회추진단 출범 (2016.09.)
- 지능정보사회 중장기 종합대책 (2016.12.)
- 지능정보사회 선도 인공지능 프로젝트 (2017.01.)
- 대통령 직속 4차산업혁명위원회 출범 (2017.10.)
- 초연결지능형 네트워크 구축 전략 (2017.12.)
- 인공지능 R&D 전략 (2018.05.)
- 데이터 산업 활성화 전략 (2018.06.)
- 지능정보사회 윤리헌장 공표 (2018.06.)
- 데이터•AI경제 활성화 계획 (2019.01.)

인공지능 R&D 전략 (2018.05.)



출처 : I-Korea 4.0 실현을 위한 인공지능 R&D 전략, 과학기술정보통신부 (2018.05.)

데이터•AI 경제 활성화 계획 (2019.01.)



출처 : -혁신성장 전략투자- 데이터•AI경제 활성화 계획 (`19 ~ `23년), 관계부처합동 (2019.01.)

4. 결 론

결 론

- 인공지능의 특성 : Open Science → 지속적이고 빠른 발전

- 미국과 중국이 치열하게 경쟁하고 있는 가운데 우리의 전략은?
  - 공개된 인공지능을 활용한 신산업 창출 → 신산업 및 데이터 관련 규제 해소 필요
  - 인공지능 관련 원천기술 확보 → 고위험 도전형 과제
  - 인공지능 기술의 플랫폼화로 글로벌 경쟁력 취약

- 범용 인공지능의 출현에 대한 대비
  - 사람처럼 추론하고 자가 발전하는 인공지능의 출현은 우리의 삶을 송두리째 바꿀 가능성이 높음
  - 인공지능의 악의적인 활용에 대한 대비, 인공지능과 인류가 공존할 수 있는 사회적 합의 마련 등 인공지능의 사회적 파급효과에 대한 논의 지속 필요