Complex MDP in online game

	논문명	journal	연구 주제	특징	Dataset
1	Mastering the game of Go without human knowledge		알파제로의 구동원리와 각종의의를 설명한다.	Self-played learning에 중점을 두었다는 점, ResNet 기용으로 깊이가 매우 깊은 신경망을 구성한 점.	KGS dataset
2	StarCraft II: A New Challenge for Reinforcement Learning	arxiv	딥마인드의 알파스타의 데이터 처리 방식, 구동원리 등등을 설명한다.	많은 수를 생각해야함과 동시에 좀 더 복잡한 환경이 고려되는 강화학습 모델이라는 점.	프로 선수들의 리플레이, Starcraft2 API
3	Towards Playing Full MOBA Games with Deep Reinforcement Learning	arXiv:2011.12692v4 [cs.AI] 31 Dec 2020		여전히 다양한 agent간의 상호작용 면에서는 아쉬운 면이 많다. 따라서 해당논문에서는 MOBA의 모든 플레이를 하도록 할 예정이다.	
4	Dota 2 with Large Scale Deep Reinforcement Learning	arXiv:1912.06680v1 [cs.LG] 13 Dec 2019		요약(불완전한 정보, 복잡하고 지속적인 상태- 행동의 mdp에서 AI 시스템의 수용량을 비약적으로 상승시키는데 주안점을 두었다 .	
5	The Arcade Learning Environment: An Evaluation Platform for General Agents				

모델	특징	한계점		사용 알고리즘
알파스타	Api를 활용하였다.	전략의 특이점을 찾는게 아니라 더 나은 컨트롤만을 추구함.		LSTM
Open AI Dota2		영웅제한, 여러 크립 조종제한. LSTM을 사용하여 최적화에 문제가 있다. -> transformer 등의 문맥 학습 알고리즘들을 사용하여 좀 더 다양하고 높은 학습을 고려.		LSTM
알파제로	최초로 self-play 를 적용함 (인간 개입 X)	패, 사활에 종종 오류를 발견 축에 대한 지식 부족.		Res_dual,MCTS
뮤제로				

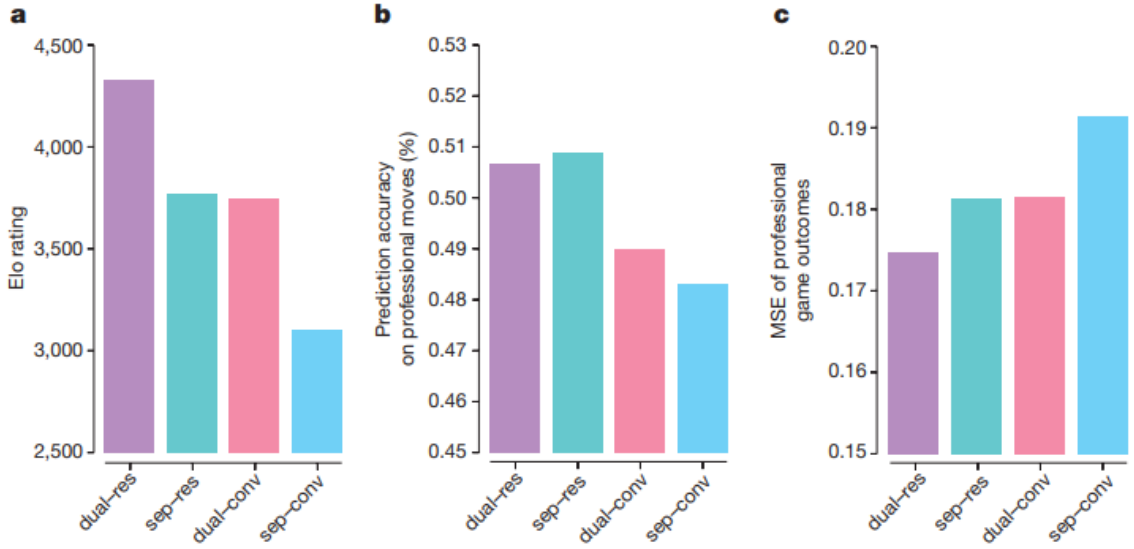
Figure 3. 성능차이

Alpha go Lee(이세돌과 대국하였던 강화학습 모델)

-> 여러 달 동안 학습했음에도 불구하고, 불과 36시간 만에 성능을 추월 당하였다.

심지어 Alpha Zero는 4텐서 유닛을 썼고, Alpha Lee는 48개의 유닛을 썼다.
-> 효율에서 엄청난 차이를 가지고 있다.

Figure 4. Lee와 zero의 신경망 차이



Conv - convolutional NN

Dual - value and policy

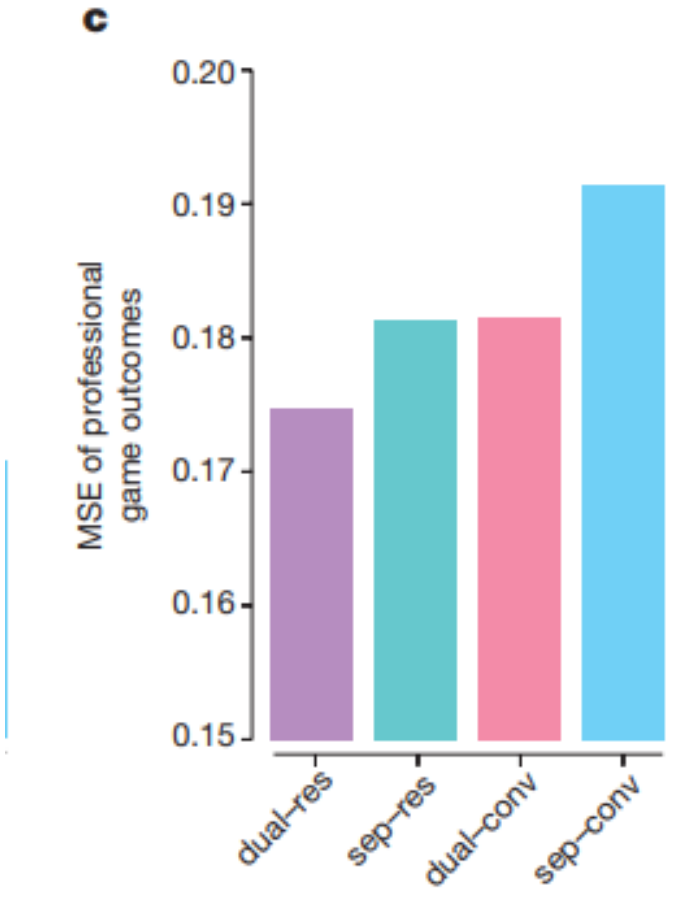
Sep - 단독으로 쓰인 것

Res - ResNet

특이점 발견

전반적인 게임 내에서의
예측 결과 지표로는

conv를 단독으로 사용한 것의 성능이
더 높다.

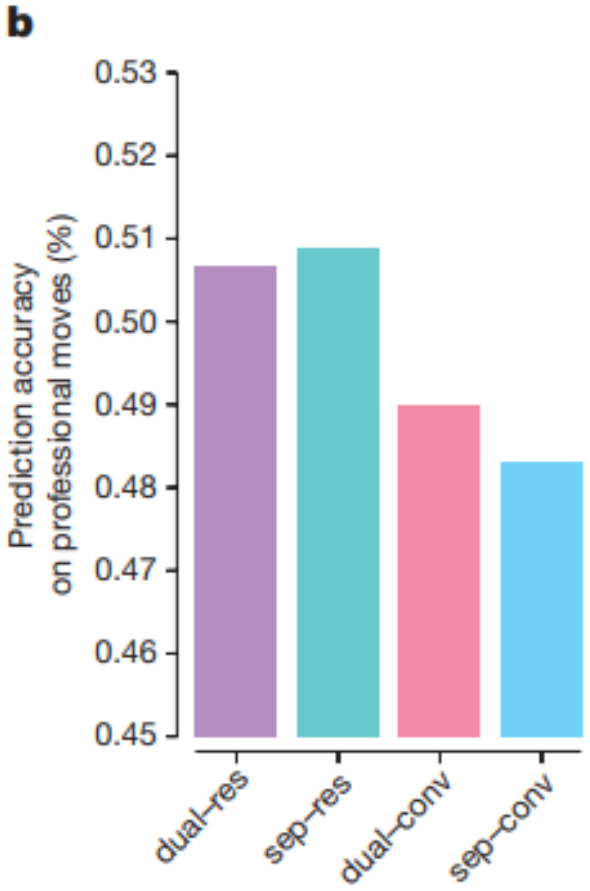


특이점 발견

그리고 또한
다음 착수에 대한 예측지표 역시

Res만 단독으로 사용한 NN이
성능이 가장 높은 것으로 나타났다.

->이 와중에 엄청난 경우의 수 임에도 불구하고 50%를 상회하는 예측 결과는
대단하다고 할 수 있다.



특이점 발견

그런데 정작 종합적 실력의 지표는
res-dual NN 가 압도적으로 높았다.

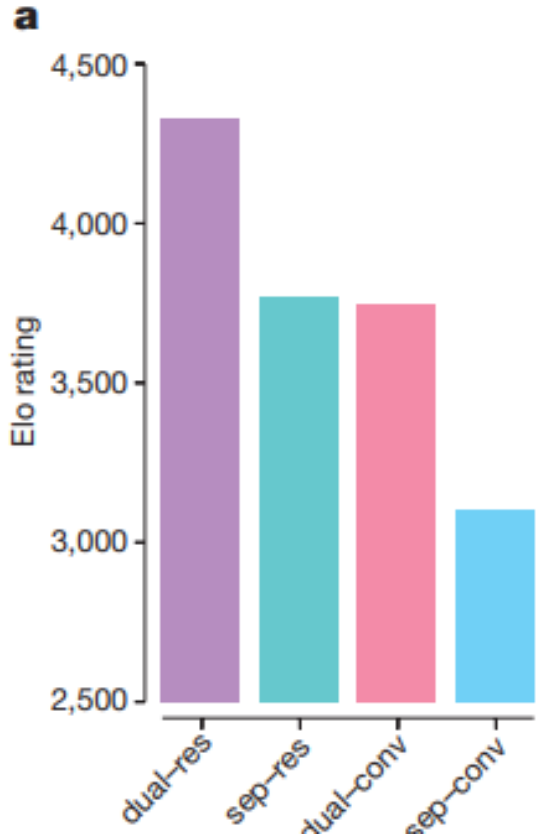
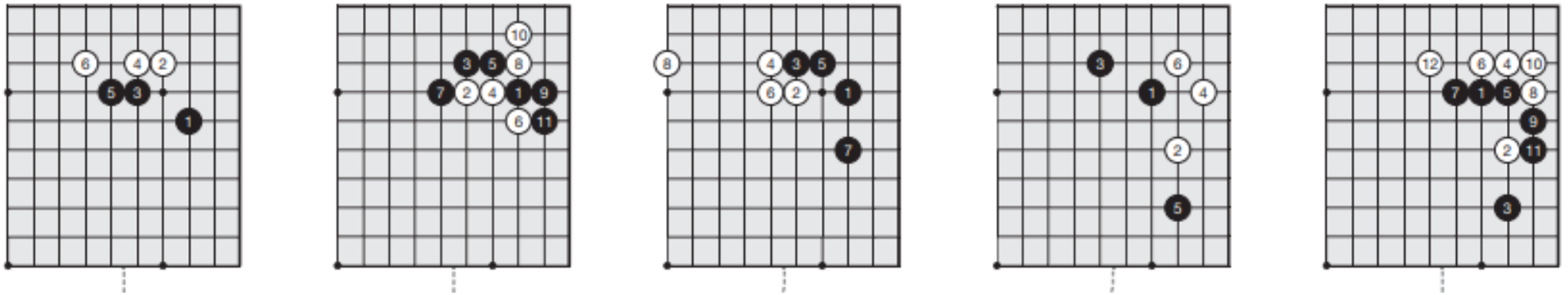


Figure 5. 알파제로가 학습한 바둑



학습과정에서 사람이 두는 5가지 정석(체스로 치면 오프닝)이 발견되었다.

알파제로의 바둑

초반에는 다소 인간이 두는 방식과 유사하게 하였으나,
나중에는 자신만의 강력한 전략을 만들어 내었다.
(기존의 3-3침투는 악수로 평가하였으나,

제로의 경우는 매우 자주 사용했다.)

처음 3시간에는 초보자 처럼 포로를 잡는 방식위주의 착수였다면,
19시간부터는 영향력과 영역싸움

70시간부터는 복합적으로 균형있는 착수를 시작하였다. (소규모 교전, 패 등등)

결론

더 적은 자원가지고도
ResNet + MCTS 의 시너지와

Self-play를 한다는 점에서 기존 알파고와 차이가 있었고,
개별 예측 지표는 뒤떨어져도, 전체적으로 판을 잡아내는 능력은

알파제로가 가장 강력했다.

분석한 한계점

축에서 굉장히 약한 모습을 보인다.

보통의 경우는 축머리만 확인하도록 하는데, 자가 대국의 경우는 끝까지 둬봐야
알 수 있기 때문, 별도의 예외 처리를 하지 않는다면 이 부분이 약점이 된다.

또한, 개별 예측 지표가 낮다는 것. 전체적으로 승률이 높은 플레이를 해내지만,
개별 예측지표가 낮은 것은 분명 장점은 아니다.

하지만, 개별 예측지표가 적음에도 궁극적인 승률이 높다는 것은 추가적인 분석이 필요하다.

Dota2 with Deep RL

abstract

불확실하고, 지속적인 상태-행동의 범위와, 많은 수를 고려해야하는
DOTA에 대한 도전을 하였다.

기존의 체스,바둑과의 차이점

- 행동과 관찰요소들이 고차원적이다.

  평균적으로 8000에서 80000가량의 선택가능한 행동이 있으며,
  룬, 나무, 미니언 등 고려해야할 요소가 고차원적이다.

- 한 화면에 잡히는 정보의 수준차이:

  한 시야 안에 모든 환경과 정보를 파악할 수 있는 아케이드, 바둑,체스와는 달리 DOTA2의 경우에는 지나치게 한정된 정보만 들어온다.
  따라서 상대방의 행동을 부정확한 정보들로 부터 정확하게 추론해내는 능력이 중요해진다.

- 길게 고려해야하는 게임이다.

  당장 체스만 하더라도 최대 80수, 바둑은 150수 정도면 승/패가 갈린다. 그리고 경우의 수도 그리 많지 않다. 하지만 DOTA2의 경우는 초당 30프레임으로 45분 가량이 소요된다.

한계점

- 일시적으로 여러 유닛을 컨트롤하게 하는 요소 배제(예시:환상의 룬, 만타 등등)
  너무 복잡한 계산은 피하도록 함.
- 117개 영웅 중 오직 17개의 상호작용가능한 영웅풀을 제공한다.
  자원부족이 원인이었던 것으로 보임.

기본적인 구조

1초에 60프레임이며 4프레임마다 행동을 하도록 함(이것을 타임스텝이라 한다.)

각 스텝마다 인간이 인지하는 수준의 정보를 관찰하도록함.
(위치, 체력, 스킬 쿨타임 등등)

현재 도타에서 제공하는 게임 인공지능들은 정책에 의한 행동이 아니라 프로그래밍 된 대로 구동되는것이 실재임. 따라서 본 논문은 또한 아이템과 기술을 구매하고 이를 활용함에 있어서

While we believe the agent could ultimately perform better if these actions were not scripted, we achieved superhuman performance before doing so. Full details of our action space and scripted actions are described in Appendix F 이 부분 해석이 난해함, F부분을 읽고 수정바람

세팅

환경의 몇 설정 값들은 훈련중에 랜덤으로 설정되었다. 게임 내에 영웅과 영웅들이 사는 아이템을 포함해서.

충분히 다양한 훈련의 게임들은 견고함에 있어서 필수이다 사람을 상대로한 넓고 다양한 전략과 상황에 있어서.

Game analytics

경기를 보고 느낀점

확실히 제한된 정보로 상대방의 정확한 위치까지 파악하는 것은 어려워 보였다.

하지만 도타의 경우는 순간이동이라는 기능을 "매우" 자주 쓰기 때문에
상대방의 위치를 더 예측하기 어려운 측면이 있다.

롤의 경우는 순간이동을 매우 제한되게 사용하기 때문에
(심지어는 아군의 오브젝트로만 이동이 가능하고 미니맵에 그 위치가 상대방에게도 드러남)
이러한 부분에서는 좀 더 예상의 적중률이 높을거라 예상된다.

경기를 보고 느낀점

경기 전반에 걸친 상대방의 '동선 ' 예측은 아무래도 성능이 떨어지더라도,

소규모 교전에서의 상대방의 스킬샷, 움직임에 대한 예측의 성능이 뛰어났으며,
아군과의 연계, 합류 등에서 높은 수준 플레이를 보여줌.

차후 계획

LSTM과 시퀀스 to 시퀀스 transformer에 대한 심층적인 이해

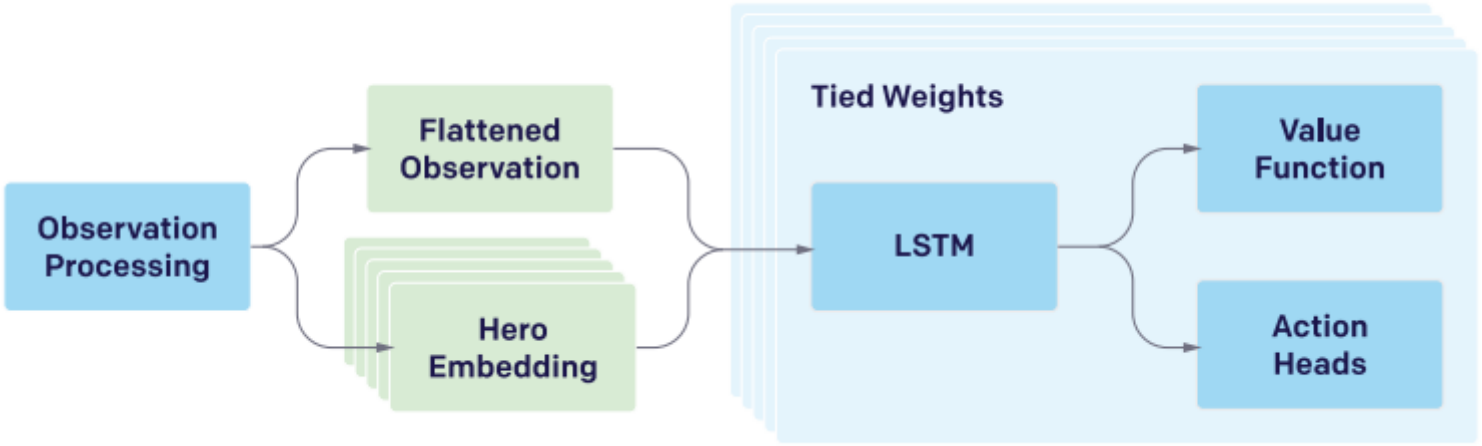
기존 논문들을 한번씩 읽어보고 비교테이블에 정리하면서
최근 알고리즘들에 맞게 최적화가능한 시도 방법 제시
게임들의 특징에 맞춰서 어느 부분에서 이득을 취할 수 있는지
획기적인 아이디어 제시

여러 게임 인플루언서들의 전략을 통해서 그들만의 전략을 분석해서

이를 녹여내는 것이 중요하다.
혹은 이들을 참고하게끔해서 인공지능에 국한된 논문에서

다소 게임에 치중된 논문을 진행해보거라.

Dota2 모델링 구조 -LSTM



복잡한 여러 배열의 관찰은 하나의 벡터로써 관리 된다. 그리고 이것은 4096개의 LSTM 모델을 지난다.
그리고 LSTM은 정책 아웃풋을 도출한다.
5개의 영웅들은 이 네트워크의 복제본이 적용되며, 각각의 인풋을 받고 활동한다. (각각의 hidden layer을 가지고)
네트워크는 프로세서