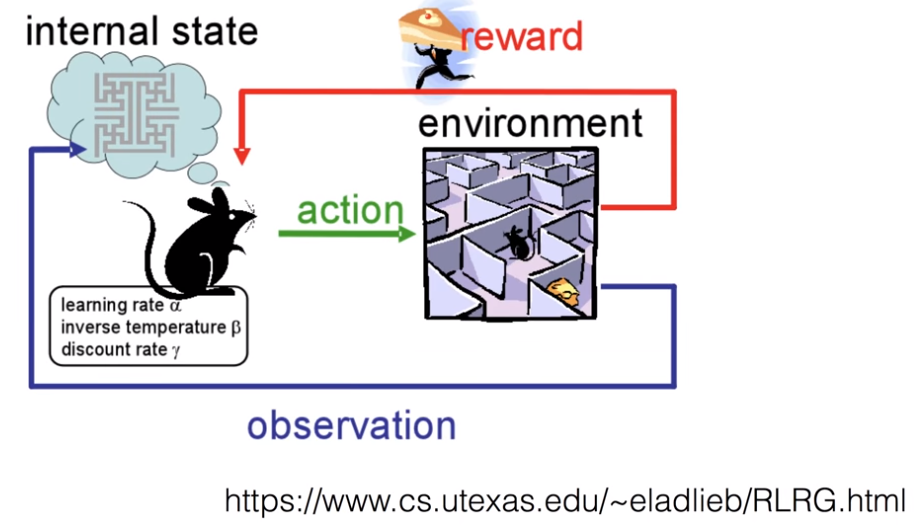
참고자료

홍콩과기대 김성훈 교수님 Reinforcement Learning with TensorFlow&OpenAI Gym강의

<https://www.youtube.com/watch?v=dZ4vw6v3LcA&feature=youtu.be>

강아지 훈련에도 쓰이는 용어. positive reinforcement training. 보상을 바탕으로 한 학습

칭찬과 꾸중을 이용한 자연스러운 학습, 현실에서도 메이저한 학습법



Actor가 Environment에서 Action을 취하면 State가 update되고 잘하면 Reward가 주어진다.

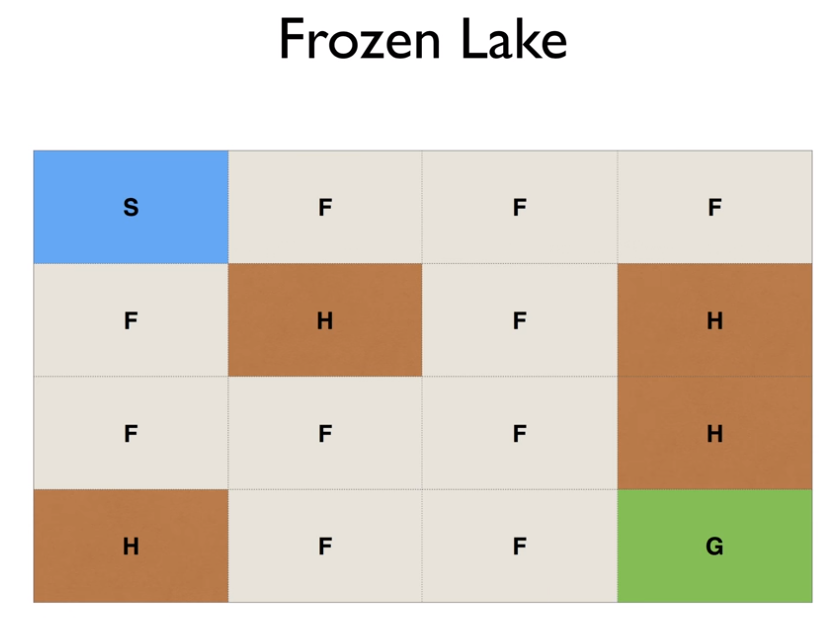
컴퓨터 RL은 1997년도 ML책에도 나올 정도로 오래된 개념. 재평가되어 주목받기 시작한 것은 Deepmind가 등장하면서부터. 한가지 게임이 아니라 모든 아타리 게임에 적용 가능한 학습법!

논문 명 “Human-level control through deep reinforcement learning”, Nature 26 Feb.2015

DeepMind는 게임 뿐만 아니라 구글데이터센터에도 적용되어 냉각에 드는 에너지를 40%절약.

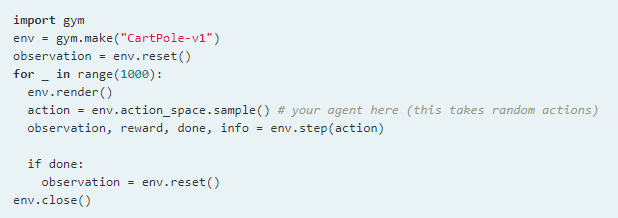
RL의 적용 분야 – 로봇 관절제어, 비즈니스 재고관리, 주식투자, Ecommerce/미디어 광고.

OpenAI GYM: environment를 구현해주는 API이다



예시 : Start, Frozen, Hole, Goal 등의 환경.

Agent는 Action(위,아래,좌,우)를 취하고 Environment는 state와 reward를 주게 된다. 목표는 Goal.



만약 끝났으면 리셋하고 닫는다

gym 불러오고

environment 만들고

첫번째 초기화 한 상태를 가져오고

1000번 반복하는데

render 하고, action을 한 뒤, step하면 상태, 리워드, 끝났는지 여부, 정보 등등등이 들어 오며

궁금한 점: 그래서 RL과 NN의 정확한 차이는 뭐지?

NN은 supervised learning, 즉 입력값과 그에 따른 결과가 이미 알려져있다. 이를 이용해 weight을 정하고, 나름의 loss function을 정하면 이를 바탕으로 loss를 최소화 하는 방향으로 NN은 학습 가능하다. 그러나 RL은 Unsupervised learning으로, loss function이라는 것이 존재하지 않는다. loss를 도출해낼 정답이 애초에 없기 때문이다.

알파고는 CNN을 사용했다고 했는데, RL도 사용했다고 한다. 구체적으로 어떻게 적용된걸까?

딥블루, 왓슨, 알파고(딥마인드)에 대한 글<https://fehani.tistory.com/8> 이것을 보면 알파고는 정책망과 가치망으로 나누어 학습을 하였다고 나와있다. 이때 cnn으로는 문제에 대한 표현 방법을 찾는다고 하는데, 아직은 어려운 것 같다.

<https://trendw.kr/2016-03112466.t1m> 알파고원리에 대한 또다른 글

여기에는 상당히 자세히 나와있다. 19\*19 바둑판을 그 자체로 cnn으로 구현하고, 다음 수를 어디에 놓아야 하는가 (정책망)을 3000만개의 기보를 통해 – 이 기보는 정답이 있다! – 학습시켜서, 돌을 어디에 놓아야 하는 가에 대한 학습을 시켰다. 거기에 MCTS의 Playout을 위해 Rollout policy또한 학습시켰다. 여기에 그치지 않고, “과거의 자신”과의 대결을 시켜서 이기면 현 사용 네트워크 weight을 강화하고 지면 반대방향으로 가게 해서 정책망을 더 강화시켰다. 그리고, MTCS에 필요한 v(s)값 – 현재 상태에서 플레이어의 승률 – 을 찾기 위해 value network 가치망 역시 만들었다. 이것 역시 cnn으로 만든 것이다. 아까 자신과의 대결을 통해 얻은 기보를 저장해 놓고 이기는 상태와 지는 상태로 구분해 학습시켰다. 이후, 이들 결과들을 MTCS의 S – E – E - B과정에 적절히 적용하면서 V(s)를 구하는 것을 반복한다. 이러면 일정 시간동안 갱신하게 되면 최종적으로 선택할 수 있는 돌들의 위치 중 가장 많이 방문했던 위치를 선택하여 돌을 놓는다.

여기서 MCTS가 RL에 속하는 것인가? 그게 아니라 과거의 자신과 대결하는 것이 강화학습인 것 같다.

OpenGym으로 돌아와서, 실습을 시작해 보겠다.

문제점 : 방향키를 입력 받을 때 필요한 헤더인 termios가 unix전용이라는 것 같다. 윈도우 환경인 나는 어떻게 해야 하는가?

msvcrt를 임포트 하면 된다고 한다. msvcrt.getch()를 시도했다.

그런데 문제는 getch()는 1byte만 받는데, 방향키는 2byte라는 것이다.

강의 댓글에 있는 ord()를 사용한 예시를 내가 약간 변형해서 적용해보기로 했다.

잘 작동한다!