ALE

-덕근짱-

abstract

딥마인드 사에서 처음으로 제시한 딥러닝 모델

- 높은 차원의 input을 받는 강화학습을 통해 control policies를 학습한다.
- 합성곱 신경망을 사용하였다. 또한 Q-learning도 사용하였다.
- Raw pixel을 입력으로 받고, 미래 보상의 가치를 매긴 것을 output으로

abstract

- 7개의 아타리 2600게임을 환경으로 연구함
- 알고리즘과 게임 자체에 대한 별도의 조정은 없이 연구

결과

- 6개의 게임에 접근을 성공하였고
- 3개의 게임은 전문가 수준을 상회하였다.

introduction

- 본 연구는 높은 수준의 데이터를 받고, 이를 강화학습에 적절히 학습 할 수 있도록 하는 것이며,
- 높은 수준이라 함은 인간이 인지하는 화면을 그대로 받는 것 (CNN) 그리고 이를 통해서 최대한 많은 게임에 적용가능한, 범용적인 NN을 만드는 것이다.

Back ground

1. 각각의 타임스탭에서 가능한 '합법적인' 행동을 취한다
2. 액션은 에뮬레이터를 통해서 상태를 수정하고 점수를 반영하는 방식
3. 일반적으로 앱실론(아타리 에뮬로 부터 형성된 환경)은 확률적이다.

task

- 에뮬레이터 내부의 상태는 agent로 하여금 관찰 당할 수 없다. 대신에, 에뮬레이터로 부터 사진의 형태로 정보를 얻게 된다.
- 이것은 raw pixel value로 현재 상태를 대표하는 screen이다.

알고리즘

$$Q^*(s,a) = \mathbb{E}_{s' \sim \mathcal{E}} \left[r + \gamma \max_{a'} Q^*(s',a') \middle| s,a \right]$$

전형적인 벨만 방정식을 베이스로 간다.

다만, 해당 논문에서는 정규화의 과정없이 각 결과에 이어서 estimate 하기 때문에, 비 실용적이라고 함.

- 아무래도 접근 방식이 한번에 알아볼 수 있는 것이 아니라, 직접 step을 밟아서 확인하는 방식이 맘에 들지 않았다고 판단함.

맘에 들지 않는 이유?

- 본래 좋은 Deep Learning 모델은 잘 정리된 데이터 셋을 가지고 있으나, basic한 RL 모델은 오직 reward를 통해 학습이 이루어지는 점
- 심지어 이 reward는 굉장히 느리며, (직접 읽어가기 때문에) 데이터가 불분명하다는 점을 지적하고 있다.
- 또한 데이터가 분포가 되어있는게 아닌, 그저 state간의 연관성으로 진행되기 때문에, data간의 관련성이 너무 크다는 점.

알고리즘

$$L_i(\theta_i) = \mathbb{E}_{s,a\sim\rho(\cdot)}\left[(y_i - Q(s,a;\theta_i))^2\right],$$

선형회귀의 폼을 가져와 '선형적인 구조'에서
근사시키는 방법론을 제시,

하지만, 비선형 구조의 정규화 데이터에서는 NN을 사용 할 것을 고려함.

알고리즘

- 이때에 사용하는 NN은 Q-network로 하고, 아래의 사진이 그 Q-network를 적용한 loss function 이다.
- 또한 아래의 함수는 이전 iteration에서부터 최적의 loss를 찾기 전까지 fix 시킨다.

$$L_i\left(\theta_i\right)=\mathbb{E}_{s,a\sim\rho(\cdot)}\left[\left(y_i-Q\left(s,a;\theta_i\right)\right)^2\right],$$

알고리즘

가장 중요한 것은, 이러한 function이 값들을 추적해 weight를 만들어 내는 것이 아니라, weight에 policy가 따라 간다는 점이다.

2021.07.16 update

TD-Gammon

- 인간의 전략에 대한 학습 없이 두 프로그램을 대련시키며,
- 오직 승패에 관한 피드백만 하는 방식을 얘기한다.

Target network

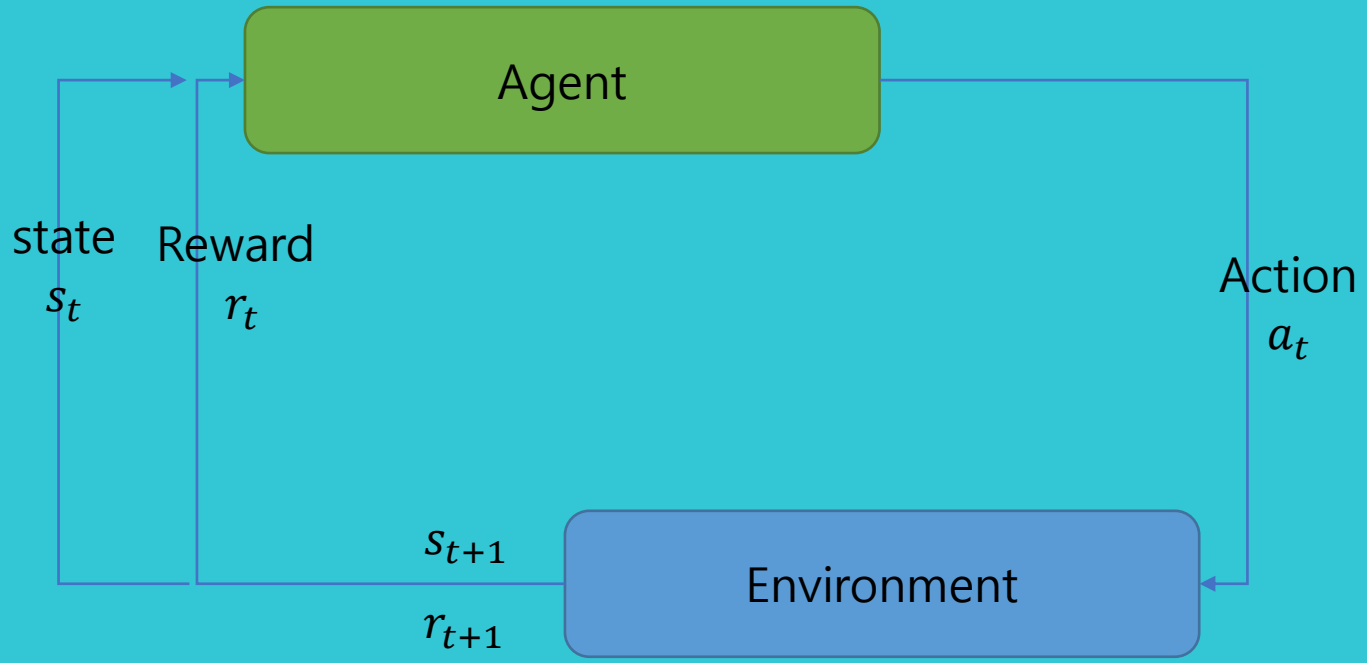
- 현재 상태 (a,s) 에대한 이터레이션 i의 최대 Q함수(정책)
- '행동분포'라고 부르기로 했다.
- 그렇다면 일일이 에피소드를 반복하면서 하는 방식보다,
- 이러한 분포를 관통하는 함수식을 근사하는 것으로, 정책의 방향을 잡는 방식이 DQN

의의가 무엇일까?

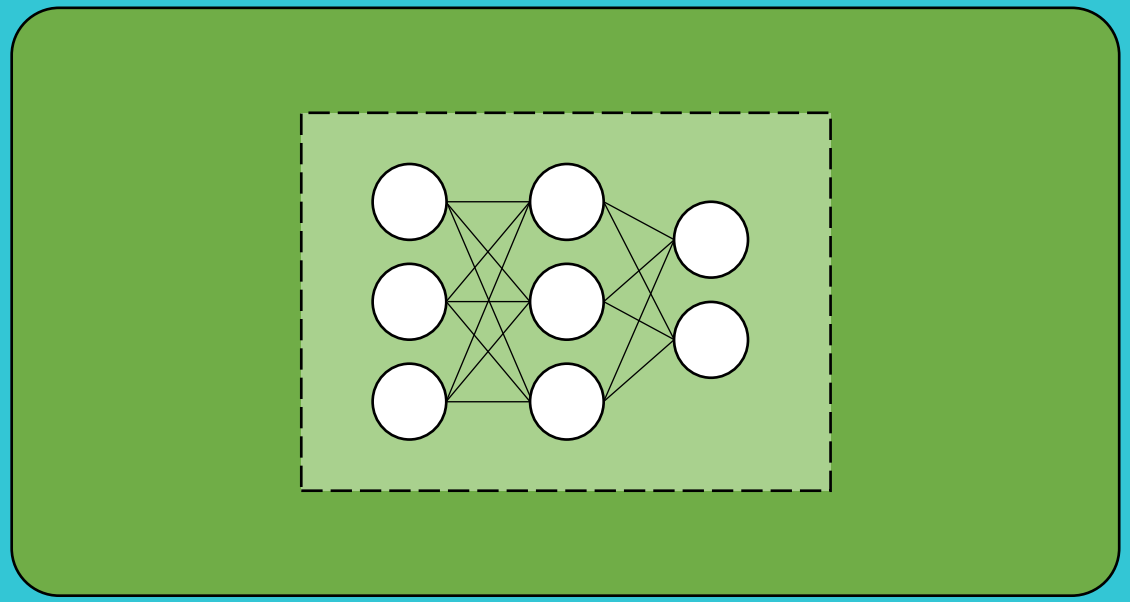
- 첫번째로, y(i)를 이끌어 내는 방법에서, 신경망을 학습 시키는 방법으로
- 현재상태, 행동, 보상 , 다음 상태로 구성된 데이터들로 셋을 만들어
  이를 통해 가중치를 수정하는 방식으로 구동된다.
- 그리하여 여러 번 Q table을 건들 필요 없이, 가중치에 현재 Q 함수를 비교하여 반영하는 식으로 학습을 시키면 된다.

2021.07.18

Q-networ와 벨만 최적 방정식의 큰 차이



Q-network



이번주 목표 task

- DQN 알고리즘 정리
- 알파스타 논문 정리

Related Work

DRL

- 컴퓨터 비전과 대화 인지에서 최근 돌파구는 의존하여 왔다. 매우 방대한 훈련 셋으로 부터 DNN을 효과적으로 훈련하는 방식으로.
- 가장 성공적인 접근은 raw pixels로 부터 훈련하는 것이다.
- SGD를 베이스로 lighweight를 업데이트 하는 방식.
- NN에 충분한 양의 데이터를 제공하는 방식은, handcrafted 데이터를 활용하는 것 보다 더 나은 대표성을 가진다.
- 이러한 성공은 영감을 주었다. 우리의 RL 접근에 대해서
- 우리의 목표는 SGD를 활용하여 raw pixel 데이터를 딥러닝에 적용하여 최적화 값을 찾는 것.

DRL