DQfD 논문 리뷰

심층 강화학습은 결정 문제에서 좋은 성공을 이루었지만, 적당한 성능을 갖추려면 많은 사전 데이터가 필요하다.

DQfD 알고리즘은 전문가 데이터로부터 효율적으로 사전 학습, 실시간 학습을 할 수 있는 알고리즘으로, 논문 발표 시점에서 SOTA 성능을 보였다.

강화학습의 활용 분야는 deep model free q learning으로 아타리 게임 플레이, 로봇 모터의 end-to-end 정책 탐색 등에 많이 쓰인다.

그리고 또한 자율주행 자동차, 헬리콥터, 추천 시스템 등에 쓰이는데, 이런 분야에서는 학습에 쓰이는 연산이 많고 수백만회의 학습을 해야 만족할 만한 성능을 가지므로 시뮬레이션이 아닌 현실 세계의 문제를 해결하기에는 문제가 있다.

그런 문제점을 해결하기 위해 우리는 심층 강화학습 알고리즘인 Deep Q-learning from Demonstrations를 소개한다. 이 알고리즘은 적은 양의 사전 데이터도 상당하게 학습을 가속시킬 수 있다.

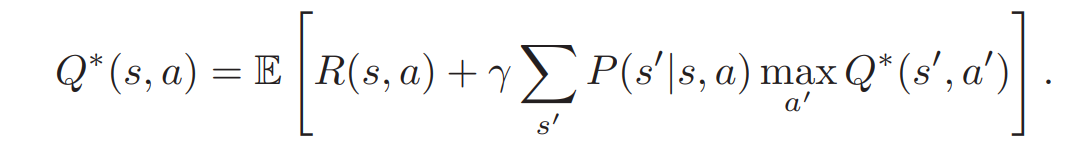
DQfD는 처음으로는 사전 데이터를 통해서 pretrain 과정을 거치는데, 이것은 지도 손실과 TD의 조합으로 이루어진다.

지도 손실은 전문가를 모방하게 만들고, TD 손실은 자신의 일관적인 가치 함수를 가지고 계속 학습되게 한다.

사전 학습이 끝난 뒤, 학습된 정책으로 도메인과 상호작용하면서 샘플된 데이터로 계속 학습하는데, PER 알고리즘은 미니배치 샘플링에 가중치를 두어서 효과적인 학습을 시킨다.

배경으로는 마르코프 결정 단계, MDP가 있다. MDP는 <S, A, R, T, 감마>로 이루어 지는데, 각각은 상태, 행동 집합, 보상 함수, 조건부 확률을 이용한 전이 함수, 손실 계수이다.

지역 최소값 함수인 Q-스타는 다음과 같은 벨만 방정식을 의미한다:



그러나 우리는 Q-스타를 계산하기 어려우니, 심층 신경망을 통해 Q 함수를 근사하는 방법인 심층 Q 신경망, DQN 방법이 2015년에 발표되었다.

DQN의 Q 함수는 상태 s와 네트워크의 파라미터 세타를 받는다.

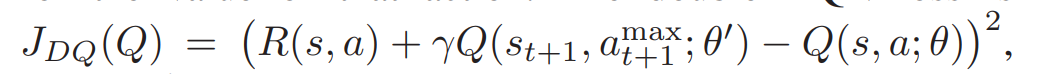
이중 Q 러닝 업데이트는 2016년에 발표되었는데 정책 네트워크와 타깃 네트워크 두가지를 통해 동작한다.

정책 네트워크는 학습되어서 액션을 추출하는 신경망이고, 타깃 네트워크는 주기에 맞춰 정책 네트워크의 가중치를 받아 또다른 Q값을 계산하는 네트워크이다.

이 설계의 장점은 좀 더 안정적으로 Q값이 구해진다는 것이다.

이중 Q 러닝 손실 함수는 다음과 같다.

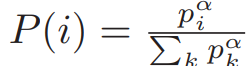
세타 프라임은 타깃 네트워크의 파라미터이다.



우선순위화된 경험 리플레이, PER 알고리즘은 리플레이 버퍼에서 더 중요한 트랜지션을 더 많이 샘플링하는 방법이다.

확률밀도함수는 다음과 같다.

알파는 임의의 스칼라 상수이다.



Pi는 다음과 같다.



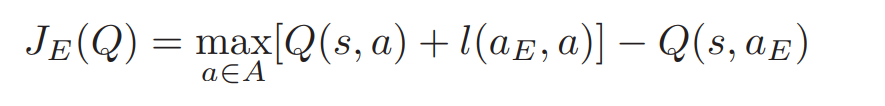
델타는 트랜지션에서의 마지막 TD 오류 값이고, 입실론은 모든 트랜지션이 샘플 될 수 있게 도와주는 작은 상수이다.

샘플링의 밀도는 이렇게 정한다.



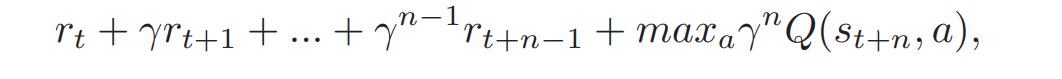
N은 리플레이 버퍼의 크기이고, 베타는 중요함의 계수이다.

그리고 우리는 큰 마진 분류 손실을 추가한다.

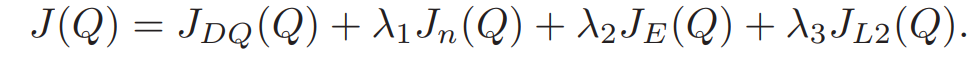


L 함수는 두 액션이 같으면 0, 다르면 양수값을 반환한다.

N회 반환도 추가한다.



총 손실은 다음과 같다.



람다들은 스칼라 계수이다.

알고리즘은 다음과 같다.

