15기 황보진경

Reinforcement Learning Assignment: Paper Review

March 22, 2021

Playing Atari with Deep Reinforcement Learning

DeepMind Technologies

고차원의 input data로부터 agent를 컨트롤하는 방법은 강화학습의 큰 과제였다. 대부분의 성공적인 강화학습 모델은 사람이 수동으로 feature들을 조합해주는 식이었기 때문에 feature를 어떻게 만드느냐에 따라 성능이 달라졌다. 딥러닝이 발전하면서 고차원 데이터를 처리할 수 있게 되었지만, 강화학습은 여전히 해결되지 못한 과제가 남아있었다. 첫째, 방대한 양의 라벨링된 트레이닝 데이터가 필요하다. 둘째, 현재의 상태와 다음상태의 보상이 독립적이지 않다. 셋째, 행동에 따라 데이터의 분포가 변한다. 따라서 본 논문에서는 복잡한 강화학습 환경 내에서 비디오 영상을 input으로 받아 contril policy를 학습시키는 것을 보여주고자 한다.네트워크는 Q-Learning algorithm을 통해 학습되었다. 특히 연관된 데이터와 일정하지 않은 분포와 관련된 문제들을 해결하기 위해, experience relplay mechanism을 통해 과거 행동에 의한 학습 분포를 완화시켰다.

agent는 타임 스텝마다 액션을 하나 선택한다. 에뮬레이터(환경)는 internal state(agent에 의해 관측되지 않는다)와 점수(reward)를 업데이트한다. 이때, agent는 스크린에 나타난 정보만을 보고 판단을 해야 한다. 이렇게 점수와 액션으로 이뤄진 시퀀스는 유한 마르코브 프로세스로 나타낼 수 있다. 즉, 표준적인 RL method for MDP를 사용할 수 있다. agent의 목표는 미래의 보상을 최대화하는 액션을 취하는 것이다. 이를 위해 Q-function을 정의하고, 그것의 기댓값을 최대화하도록 목적함수를 설계한다. 이 때, 기본적으로 Bellman equation을 사용하여 업데이트를 반복하면 결국 최적의 action-value function을 얻을 수 있다. 그러나 이것은 실용성이 떨어지기 때문에 본 논문에서는 이 함수를 Neural Net으로 대체하고자 하였다. 이 알고리즘은 에뮬레이터와 별개로 작동하는 model-free, off-policy algorithm이다.

TD-Gammon의 접근법과는 대조적으로 본 논문에서는 Experience Replay를 사용하여 agent가 타임스텝마다 진행한 시퀀스를 저장하였다. 그리고 임의로 하나를 골라 학습을 진행하였다. 이러한 형태는 기존의 방법들보다 많은 이점을 가지고 있다. 첫째, 각 스텝에서의 경험이 많은 weight update에 이용되기 때문에 data effcient하다. 둘째, 연속적인 샘플들은 연관성이 높기 때문에 학습이 비효율적이다. 셋째, 행동 분포를 균형있게 선택하여 local min으로 수렴하거나 발산하는 현상을 피하고 매끄럽게 학습을 진행할 수 있다.

input은 128개의 컬러파레트로 이뤄진 210*160사이즈의 이미지 프레임이다. 이를 흑백으로 바꾸고 110*84로 다운샘플링한, 실제 플레이 공간인 84*84 부분만 자름다.

Q-value를 구하는 방법은 두가지가 있다. 첫째, history와 action을 input으로 하고 예측된 Q-value를 구한다. 단점은 action에 따라 연산량이 Linear하게 늘어난다는 것이다. 둘째, history를 input으로 하여 모든 행동에 대해 Q-value를 구한다. 장점은 연산량이 줄어든다.

모델의 Input은 84*84*4 이미지이다. 첫번째 레이어는 16개의 8*8필터(stride=4)이다. 두번째 레이어는 32개의 4*4필터(stride=2)이다. 마지막 히든 레이어는 Fully-Connected(256)이다. Ouput 레이언는 각 행동에 대한 output을 내는 FC layer이다.

본 논문에서는 딥러닝을 적용한 강화학습 모델을 소개하였으며, 2600개의 Atari Game에 대해 Control Policy를 학습하였다. 그리고 모델이나 하이퍼파라미터 조정 없이 7개의 게임 중 6개에 대해서 SOTA 성능을 보여주었다.