Tobigs 15기 Week8 강화학습 – 15기 이성범

Playing Atari with Deep Reinforcement Learning 리뷰

[Abstract]

Atari(미국의 콘솔 게임)의 게임들을 강화학습을 활용한 Deep Learning Model을 만들어 약 2600개의 게임 중 3개의 게임에서 사람보다 더 높은 성능을 보이는 결과를 가져왔다. Deep Learning Model은 CNN을 사용한 변형된 Q-learning을 사용하여 만들어졌다. 이 모델은 Atari 게임 속의 이미지 데이터를 입력 받아 어떠한 행위를 하고(움직임) 그 행위에 대한 점수를 얻고(공 튀기기면 공으로 벽을 부수고 얻은 점수) 그 행동에 대한 결과 값을 함수를 통해 얻게 된다.

리뷰

1. 대부분의 딥러닝 모델은 사람이 직접 라벨링을 한 많은 양의 학습 데이터가 필요하다. 하지만 강화 학습은 행동에 대한 보상을 통해서 최고의 보상을 찾는 방식의 사이클적인 학습 방법이기 때문에 보상에 의존하는 학습 방식에 의하여 데이터 자체가 sparse 하고 noise 하며 delayed 된다.

2. 딥러닝 모델의 데이터는 각각 독립적이다. 하지만 강화학습에서는 한 행동을 한 후 다음 행동을 하기 때문에 데이터 간의 연관성이 높다. 데이터 간의 상호 연관성이 높으면 지역 최적점에 빠질 확률이 높다. 예를 들어 첫번째 행동이 계속 50점을 주는 행동이었고 이 행동이 최고인줄 알고 학습을 했는데 처음에는 0인 행동이었지만 10번째 행동에 갑자기 1000점을 주는 행동이 발생할 수 있기 때문에 데이터 간의 상호연관성이 높아서 이러한 지역 최적점에 빠질 문제가 크다는 것이다.

3. 딥러닝 모델의 데이터는 이미 정해진 데이터를 가지고 학습하기 때문에 데이터의 분포가 변하지 않는다. 하지만 강화학습의 경우 새로운 행동을 배울 때마다 데이터의 분포가 변하게 된다. 이는 기존 딥러닝 모델의 가정을 배반하는 일이기 때문에 기존의 딥러닝 모델의 가정들을 사용할 수 없게 만든다.

따라서 이 논문은 위의 강화학습에 딥러닝 모델을 활용할 수 없었던 3가지 문제점을 해결하고자 변형된 Q-Learning와 Experience Replay Memory를 사용했다.

우선 모델은 input으로 게임의 이미지의 픽셀을 받는다. 그런데 여기서의 이미지는 정지된 형태가 아닌 4개의 연속적인 이미지 데이터 이다.(예를 들어 점프를 한다고 가정할 때 이미지가 정지되어 있다면 현재 캐릭터가 점프를 하는지 점프에서 내려오는 구분할 수 없을 것이다. 하지만 연속되어 있다면 위에서 아래로, 아래에서 위 등의 방향을 구분할 수 있게 된다.) 이 input 데이터를 CNN 모델을 통해서 처리한 후(사람이 다루기 까다로웠던 데이터를 CNN을 통해서 해결) ouput으로 게임의 모든 행동에 대한 Q-value를 출력한다. 그 후 유의한 행동들을 다음 행동들로 결정한

다. 이러한 방식을 통하여 계속 진행되어 하나의 게임이 종료된다면 하나의 episode가 끝난 것이다. 이러한 종료된 episode들을 ERM에 저장한다. 그런데 기존의 대부분의 딥러닝 모델들은 보통 하나의 학습이 끝나면 학습 결과를 바탕으로 손실함수를 구해 파라미터를 수정하는 방식을 거칠 것이다. 하지만 강화학습의 경우 데이터 간의 상관관계가 높고 학습때마다 데이터의 분포가 다르기 때문에 학습마다 파라미터를 수청한다면 손실함수의 수렴이 진동이 발생하여 제대로 수렴하지 못할 것이다. 따라서 이러한 방식을 해결하기 위해 업데이트 전까지 모든 학습에 대한 파라미터를 고정시킨다. 그리고 모든 에피소드가 끝나면 저장된 ERM에서 일부를 uniformly random하게 sampling하여 mini-batch를 구성한 후 역전파를 통하여 파라미터를 업데이트 한다. 이러한 ERM 방식을 통한 파라미터 업데이트를 통하여 데이터 간의 상관 관계를 없애고 데이터 분포가 일정하다고 볼 수는 없지만 그래도 이전의 강화학습보다는 더 일정한 분포의 데이터를 얻을 수 있다. 그리고 과거에는 파라미터를 한번 업데이트 하면 파라미터를 업데이트 하기 위해 수집되었던 방대한 데이터를 모두 버렸지만 이 논문에서는 ERM을 통하여 과거의 데이터도 계속해서 학습에 활용할 수 있도록 하여 조금 더 데이터를 효율적으로 활용할 수 있게 했다. 그리고 이러한 방식의 파라미터 업데이트 방식은 다음 파라미터에서의 훈련 데이터를 조금 더 효율적으로 결정할 수 있게 해준다. 예를 들어 현재 행동이 점프라면 다음 행동에 대한 학습 데이터를 어느정도 알기 때문에 현재 점프라는 행동을 고려하여 다음 행동을 효율적으로 선택할 수 있게 된다. 이러한 학습 방식을 통하여 서로 다른 Atari 게임 7개에 대하여 단 하나의 NN만을 사용하여 그 결과가 기존 결과를 능가하게 되었다.

위 논문은 이미지 sequence 만을 사용하여 CNN으로 feature을 만들어 action을 결정한다는 것, 과거의 에피소드를 계속해서 저장하는 ERM 방식을 통하여 데이터를 효율적으로 활용한다는 것, 학습 시에 파라미터를 고정시키고 모든 학습이 끝나면 파라미터를 업데이트 한다는 점 등에서 과적의 강화학습과 큰 차이점을 가져오고 그 성능 또한 과거의 결과를 능가했다는 점에서 큰 의의를 주며 메모리 성능이 중요한 ERM 방식을(과거의 데이터를 계속 저장하는 방식이기 때문에) 활용한 논문의 실험을 오늘날에 다시 실험한다면 더 높은 성과를 보일 것 이다.

참고자료

- https://brunch.co.kr/@kakao-it/161
- https://velog.io/@kth811/1.-Playing-Atari-with-Deep-Reinforcement-Learning
- https://mangkyu.tistory.com/60
- http://sanghyukchun.github.io/90/
- https://sumniya.tistory.com/18