Playing Atari with Deep Reinforcement Learning 논문 리뷰

- 논문 이해를 위한 용어, 개념들

: CNN, RL, DL, Q-learning, SGD, Replay Memory, Loss function

- 논문 소개 요약

Vision이나 speech와 같은 high-dimensional sensory inputs로부터 agent를 학습시키는 것은 RL의 오랜 과제. 딥러닝이 발전함에 따라 Vision, Speech와 같은 고차원의 데이터들을 추출하는 것이 가능하다. 딥러닝을 강화학습에 도입해보는 것!

- 논문에서 제기한 문제

1. 딥러닝에서는 라벨링된 train data가 필요하지만 강화학습에서는 어떠한 행위에 대한 시도의 결과로 학습하기에 비교적 시간이 더 걸림.

2. 딥러닝의 데이터들은 각각 독립적, 그러나 강화학습에서는 corr 가 높음.

-> 이를 해결하기 위에 replay memory 사용!

3. 강화학습에서 새로운 행동마다 데이터의 분포가 변하는데 이는 딥러닝의 가정과 충돌하여 문제가 될 수 있음.

- 배경지식

Agent는 현재 상황만 보고있기 때문에 전체적인 상황을 이해하기 힘듬. Agent의 목표는 미래 보상을 극대화시키는 방향으로 행동을 선택하는 것. 시간이 지날수록 reward의 가치가 낮아지는것은 할인율 (r).

The optimal action-value function obeys an important identity known as the Bellman equation

Bellman equation에 따라 모든 가능한 행동에 대해 다음 단계의 reward를 안다면 현재에서 다음 reward를 더하면서? 이 값을 최대화하는 것.

In practice, the behaviour distribution is often selected by an -greedy strategy that follows the greedy strategy with probability 1 − and selects a random action with probability . ( 네모는 입실론 계수인듯)

Greedy strategy를 따른다는 말인듯. 강화학습에서 데이터들 간의 연관성을 깨고 일정한 확률로는 random 행동을 하고 남는 확률로는 greedy stratedy를 따라 행동하기로함.

- Deep Reinforcement Learning

(DQN 참고 사이트 - <https://sumniya.tistory.com/18>)

DQN은 Atari Game의 raw pixel들을 input으로해서 CNN을 function approximator로 이용하여 value function이 output. 결국 그 value function은 future reward를 추정하는데 이용되는 재료.

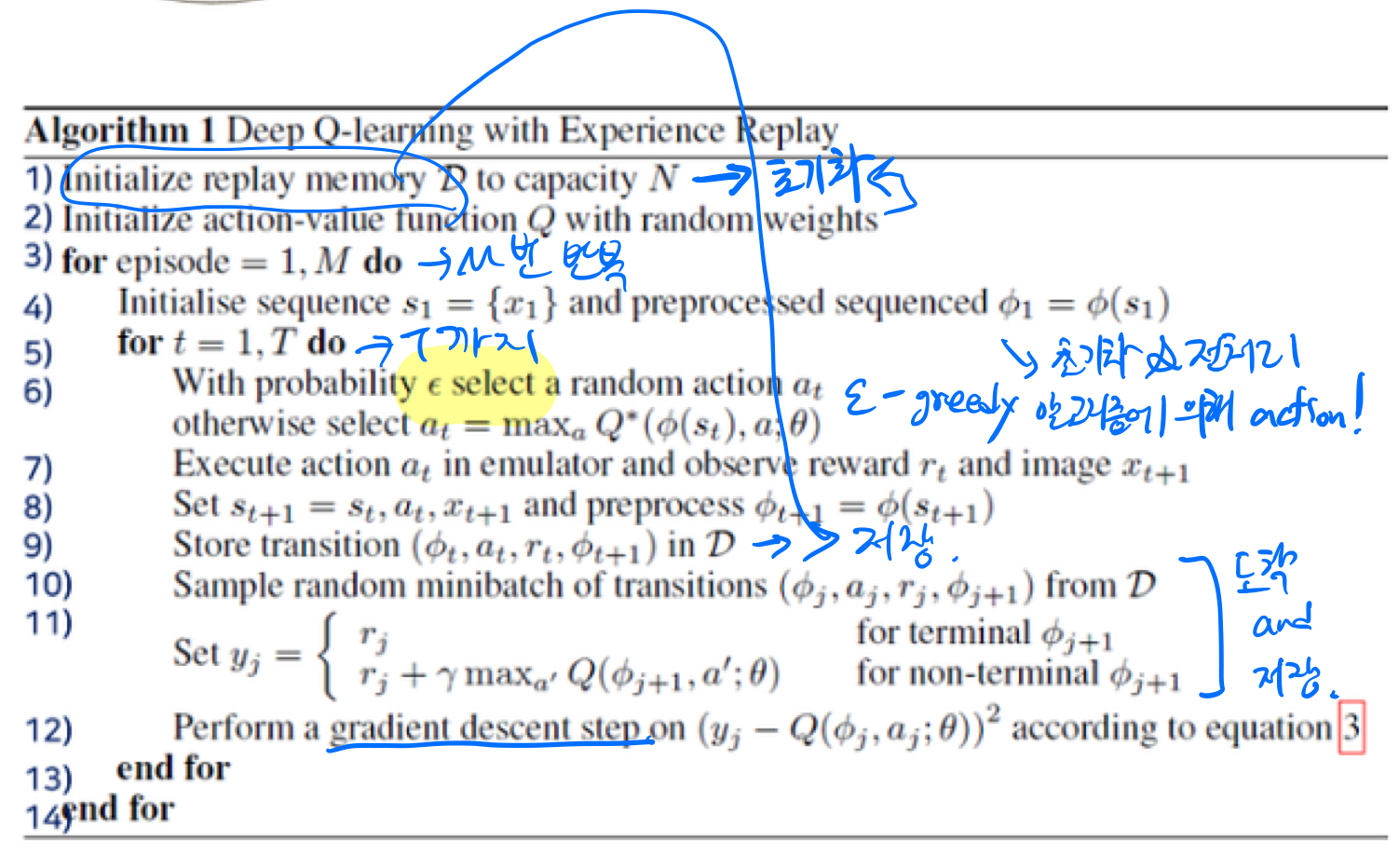
1. raw pixel을 받아와 directly input data로 다룬 것

2. CNN을 function approximator로 이용한 것

3. 하나의 agent가 여러 종류의 Atari game을 학습할 수 있는 능력을 갖춘 것

4. Experience replay를 사용하여 data efficiency를 향상한 것

(저는 논문의 수도 코드를 보고 이해를 했습니다.)



- 전처리와 모델 아키텍쳐

Raw Atari frames : 128 color palette를 가진 210 \* 160 pixel images

Input의 차원을 줄이고 전처리를 진행. 논문에서는 RGB를 gray-scale로 변환, 이미지의 차원을 110 X 84로 변환한다고 설명했다. 그리고 전처리된 이미지는 NN의 인풋으로 들어가게 되는데 이는 84 X 84의 정사각형으로 잘라낸 이미지가 들어간다.

- Reward structure

While we evaluated our agents on the real and unmodified games, we made one change to the reward structure of the games during training only. Since the scale of scores varies greatly from game to game, we fixed all positive rewards to be 1 and all negative rewards to be −1, leaving 0 rewards unchanged.

양의 보상은 1, 음의 보상은 -1, 변화없음은 0.

- RMSProp

In these experiments, we used the RMSProp algorithm with minibatches of size 32. The behavior policy during training was -greedy with annealed linearly from 1 to 0.1 over the first million frames, and fixed at 0.1 thereafter. (네모상자는 e)

RMSProp 알고리즘을 크기가 32인 미니배치에 적용.

- Frame Skipping Technique

- 모든 프레임을 보고 행동을 하는게 아니라 특정 프레임을 보고 행동을 선택하게 함.

- 결론

이 논문에서는 강화학습을 위한 새로운 딥러닝 모델을 소개! 적용하기 위하여 SGD, replay memory 등을 적용한 Q-learning 변형 기법을 소개.  
2번 과제

신호등 프로젝트를 듣고 생각난 프로젝트인데 이와 같이 지하철에도 똑같이 적용해보면 좋을 것 같다고 생각했습니다.

주제 : 출퇴근 교통체증 해결 및 지하철 편리성 향상을 위한 지하철 배차간격 제어

환경 : 종점과 출발점에서 지하철들의 배차 간격

상태 : 현 호선의 지하철에 탑승 혹은 대기 인원을 대상.

행동 : 지하철 출발 혹은 종료.

보상 : 이 시점에 지하철을 내보냈을 때 -> 탑승인원 비율이 대기인원보다 높아질 수 있는가?

\* 고려사항 : 그렇다면 다 탈 수 있게 매번 차를 보내면 되겠지만 지하철의 수는 제한, 그리고 운행 시 지하철 탑승 인원이 하차하는 비율도 역마다 고려를 해서 출발시켜야함.