생명 점수 기반 가상 보상을 이용한 팩맨을 위한 심층 강화학습 성능 향상

박태화 O, 김경중*
세종대학교
dolavvv@naver.com, kimkj@sejong.ac.kr

Improvement of Deep Reinforcement Learning for Pac-Man using Virtual Rewards based on Life Score

Tae-Hwa Park O, Kyung-Joong Kim*
Department of Computer Science and Engineering, Sejong University, South Korea

요 약

최근, 심층강화학습을 이용하여 Atari 게임을 풀려는 시도가 있으며, 이때 점수만을 사용하여 보상을 정의한다. 하지만, 사람들의 경우 게임 화면으로부터 들어오는 다양한 정보를 해석하고, 이를 가상의 보상으로 활용한다. 본 연구에서는 생명 점수를 가상의 보상으로 활용하는 방안을 통해 심층 강화학습 성능을 향상해 본다. 기존 대표적인 심층강화학습 알고리즘인 A3C 가 잘 풀지 못하는 PacMan 게임을 대상으로 하였을 때, 개선된 성능을 보임을 확인하였다.

1. 서론

2015 년 Google DeepMind 는 게임 화면과 게임 점수만을 사용한 심층강화학습 DQN 을 이용하여 50 게 게임 중에서 [1] 절반 이상을 사람만큼 플레이함을 보였다[1]. 하지만, 일부 게임은 복잡도가 굉장히 높아서 단순한 화면 입력 만으론 풀기에 한계가 있었다. 복잡도가 높은 한 게임에 특화되어 푸는 모델[2]은 많이 연구 되었지만, DQN 보다 더 나은 성적을 보여주는 모델은 현재 연구 중에 있다[3]. 잘 풀리지 않는 게임 중 대표적으로 Pacman 이 있다. 이 게임은 스테이지 마다 달라지는 스테이지의 형태와 유령들의 움직임 때문에 상태가 매우 많이 나오게 된다. 또, 유령의 움직임은 항상 정해져 있지 않아 한 상태에서 최선의 선택이 달라 질 수 있다. PacMan 은 이미 Microsoft 사 에서 최고 점수까지 달성하는 심층 강화학습 방법을 발표했다. 하지만, 이 방법은 기존의 방식과 많이 다르며, 다른 게임에 적용하는데 어려움을 겪을 수 있다. 대부분의 게임에 적용 가능한 모델은 PacMan 을 잘 풀지 못하는데, 본 논문에서는 생명점수 개념을 이용해 이를 풀고자 한다. 생명점수는 게임의 끝을 정하는 구조 중 하나로, 게임과 밀접한 관련이 있다. 이 개념을 잘 풀리지 않는 게임인 PacMan 에 도입해 성능이 개선됨을 확인한다.

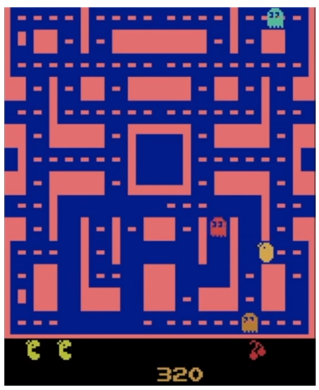

그림 1. 위 그림에서 생명점수는 2 점수는 320 이다.

2. 배경

DQN 은 화면을 상태입력으로 받고 상태 간의 강한 상관관계를 리플레이 메모리로 해결함으로, 한 개의 알고리즘으로 여러 게임을 풀면서 일부 게임은 사람보다 더 잘하는 에이전트를 학습시키는 데 성공하였다. 하지만, 리플레이가 많이 쌓여야 하므로 학습속도가 느리고, 그 리플레이들을 저장해놔야 하므로 많은 메모리를 사용하며, 가치함수 중 가장 높은 값을 가지는 액션을 선택하므로, 학습 과정이 불안하다.

A3C 는 성능은 학습이 충분히 된 DQN 과 비슷하지만,

1 * 교신저자

리플레이 메모리를 사용하지 않으며, 학습 시 다수의 Actor-Critic 에이전트를 비동기식으로 사용하기 때문에 DQN 보다 빠른 학습속도를 가져 새로운 baseline 으로 제시된다. [4]

2.1 생명점수

생명점수란 게임에서 정해진 플레이어의 목숨 수를 말한다. 한 번의 게임에서 주어진 목숨을 전부 사용하면 게임이 끝나게 된다. 이는 많은 게임이 채택하는 게임 구조 중 한 개이며, 다른 Atari 게임에서도 이를 확인 할 수 있다.

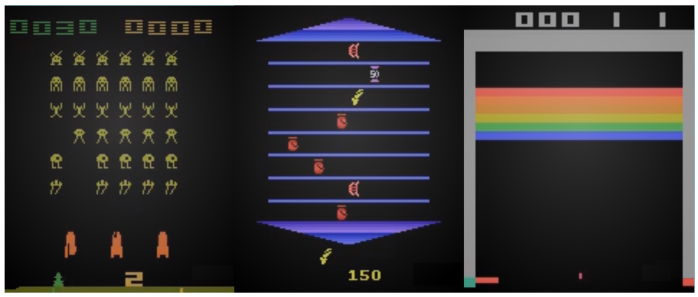

그림 2 다른 Atari 게임에서 확인되는 생명점수들

3. 제안하는 방법

환경이 주지 않는 학습에 필요한 보상을 전처리를 통해 가상의 보상 테이블 만든다. 이 테이블을 이용해 보상이 주어지는 순간 테이블을 동시에 참조해 추가적인 보상을 주는 것을 제시한다.

다만, 본 논문은 위의 가정에서 문제를 단순화해 가상의 보상을 사람이 직접 정의한다. 환경에서 주지 않는 추가적인 보상이 에이전트가 학습하는데 더 좋은 결과를 낼 수 있게 도움을 주는지 확인해본다. 이를 위해 환경에서 주는 보상의 형태에 변화가 생긴다.

$$Reward = R_{t+1}(S_t, A_t) + V_{t+1}(S_t, A_t)$$

위 수식에서 R 은 환경에서 제공하는 보상을 의미한다. V 는 사용자가 직접 만든 보상테이블에서 얻는 보상을 의미한다. 보상 테이블은 환경의 정보를 입력으로 받게 된다. 이 두 개의 보상의 합을 이용해 학습을 하게 된다.

그림 3 의 상태는 PacMan 이 pill 을 한 개 먹음과 동시에 유령한테 잡혔다. 이 상태에서 환경이 주는 R값은 10 이 되고 일반적인 환경은 10 의 보상을 받게 된다. 가상의 보상을 받는 환경에선 테이블을 참조해 V 값을 -1000 으로 준다. 이 둘을 더해 에이전트는 -990 의 보상을 받게된다.

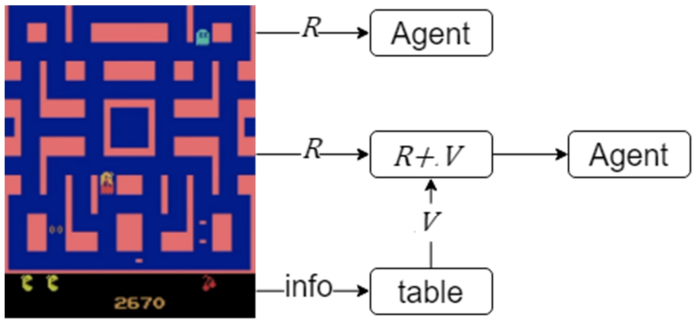

그림 3. 에이전트가 받는 보상의 흐름

state	R_t	$R_t + V_t$
Normal pill	10.0	10.0
Power pill	30.0	30.0
Bonus fruit	50.0	50.0
Weaked ghost	100*n	100*n
Lose lives	0.0	-1000.0

표 1. 각 환경에서 주는 보상 값

4. 실험 및 결과

본 논문에서는 환경으로 Open AI Gym 에서 일반적인 강화 학습으로 사람보다 나은 결과를 보여주지 않는 Ms. Pacman 게임을 선정하였다.

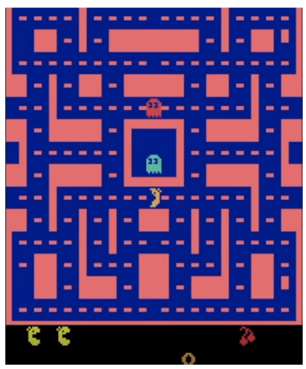

그림 4. Open ai gym 에서 Ms.PacMan-V0 의 input

화면을 입력으로 쓰지만 그림 4 에서 보이는 화면을 그대로 사용하지 않는다. 화면에서 검은색 하단부를 자르고 정규화를 거친 뒤 크기를 80*80*1 로 조정하여 신경망의 입력으로 사용하게 된다.

에이전트로 A3C 에 CNN 을 사용하였다. CNN 은 총 4 개의 층으로 구성하여 최종적으로 9 개의 출력을 가지는 형태로 구성하였다. 모델에는 변화를 주지 않은 상태로 16 개의 Actor 를 두 개의 환경을 만들어 실험했다.

그림 5 은 두 환경에서 PacMan 이 얼마나 오래 살아남았는지 알 수 있는 에피소드의 길이를 보여준다. 그림 6 은 PacMan 이 얻은 점수를 보여준다.

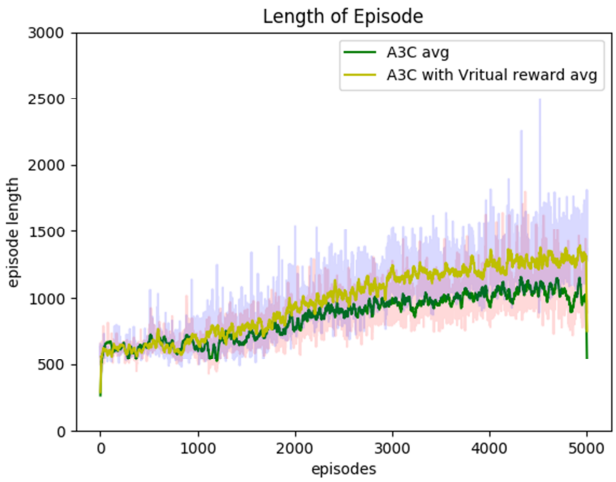

그림 5. 두 환경에서 에피소드의 길이의 평균값

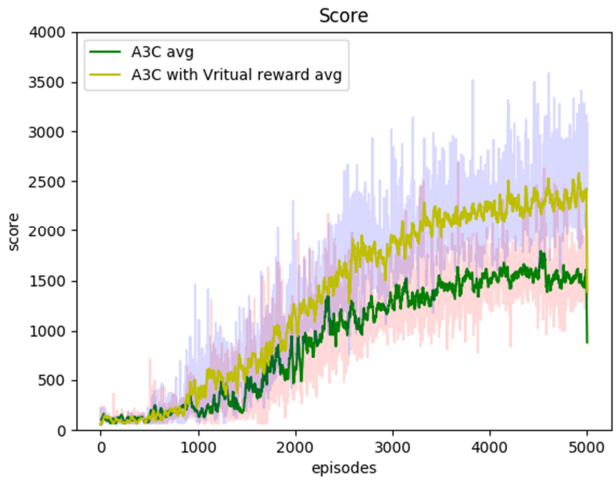

그림 6. 두 환경에서 점수의 평균값

두 그래프 그림 5 와 그림 6 을 비교해보면, 평균적인 점수와 에피소드의 길이 모두, 추가적인 보상을 주는 환경에서 평균적으로 더 높은 수치를 가지는 것을 확인할 수 있다.

이는 추가적인 보상이 주어지는 순간이 에이전트가 유령과 겹쳐져 생명점수가 줄어들고, 그로 인해 게임의 에피소드 길이가 짧아져 평균적으로 점수를 더 적게 얻는다는 것을 학습한 결과라고 볼 수 있다. 이렇게 학습이 끝난 모델은 유령과 겹칠 상황을 일부 피하려는 움직임을 보인다. 다만 Ms.Pacman 의 상태 개수가 매우 많기 때문에 항상 유령을 피하는 움직임을 보이는 것은 아니며, 학습 기간에 따라 차이가 있을 것이라 예상된다.

5. 결론

환경에서 주어지지 않은 사람이 판단하기에 보상을 받을 이유가 충분한 상태를 정의해서 따로 추가로 보상을 주었을 때, 그렇지 못한 환경보다 더 좋은 결과를 얻었다.

이는 이를 자동화, 일반화시키면 다른 여러 환경에 적용 가능한 모델이 나올 가능성을 보여준다. 또한 앞으로 이러한 모델을 만들기 위해 한가지의 상태에 뿐만 아니라 추가적으로 다양한 상태에 대해 보상을 주며 그 보상도 긍정적인 보상 과 부정적인 보상 모두에 대해 실험을 해보며 그 차이와 효율에 대해 비교 해보는 것이 해당 모델을 만들 때 필요할 것이다.

6. 감사의 글

2017 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2017R1A2B4002164).

참고문헌

[1] MNIH, Volodymyr, et al. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.

[2] VAN SEIJEN, Harm, et al. Hybrid reward architecture for reinforcement learning. In: Advances in Neural Information Processing Systems. p. 5392-5402. 2017.

[3] Hessel, Matteo, et al. "Rainbow: Combining Improvements in Deep Reinforcement Learning." arXiv preprint arXiv:1710.02298 2017.

[4] MNIH, Volodymyr, et al. Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning. p. 1928-1937. 2016.