서로 다른 학습 기반을 적용한 강화학습에서의 보상 방법에 관한 연구

A Study On Reward Methods in Reinforcement Learning Using Different Learning Bases

가톨릭대학교 THE CATHOLIC UNIVERSITY OF KOREA

김 광운, 양 정진 / 가톨릭대학교
kgo0748@gmail.com jungjin@catholic.ac.kr

1. 서론

연구의 요약

기계학습의 한 종류인 강화학습은 행동을 통해 얻게 되는 보상을 근거로 학습을 진행한다. 따라서 보상을 최대로 얻을 수 있는 방향으로 행동을 혹은 행동의 순서를 선택하게 된다. 이때 행동을 선택하게 되는 근거에 따라서 달라지는 보상 방법에 관하여 장단점을 비교 분석했다.

연구의 목적

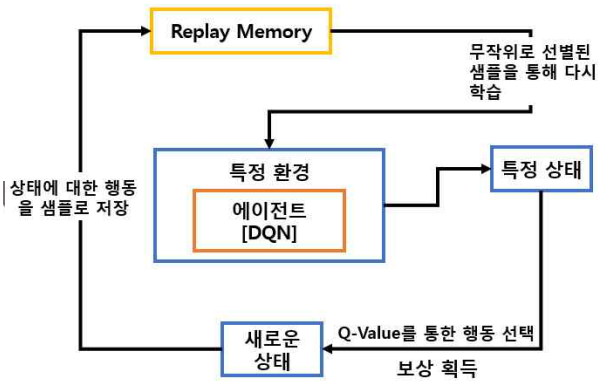
- 게임 환경에서 두 가지의 다른 학습 기반을 적용하여 강화학습의 진행
- 시행횟수에 따라 주어지는 보상에 따라서 학습 기반에 대한 비교

2. 연구 방법

[그림 1] - 연구 진행 과정

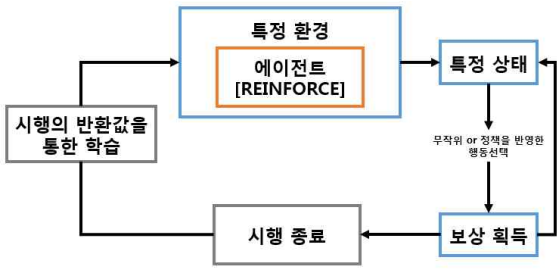
가. 사용 방법론

Deep Q-Networks



[그림 2] - DQN 학습 진행과정

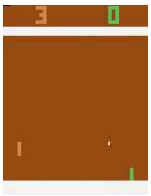
REINFORCE



[그림 3] - REINFORCE 학습 진행 과정

나. 학습 환경선정 및 전처리

학습 환경

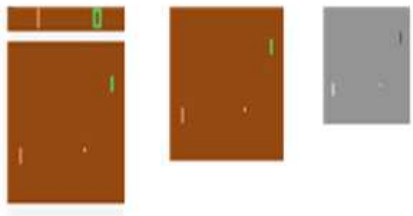


Pong-v0

- Pyhton 모듈의 Gym Atari에 속해있는 Pong-v0 환경설정
- 에이전트 승패에 따라 ±1 의 보상 획득
- 패배시에도 -1점의 보상 획득으로 학습용이

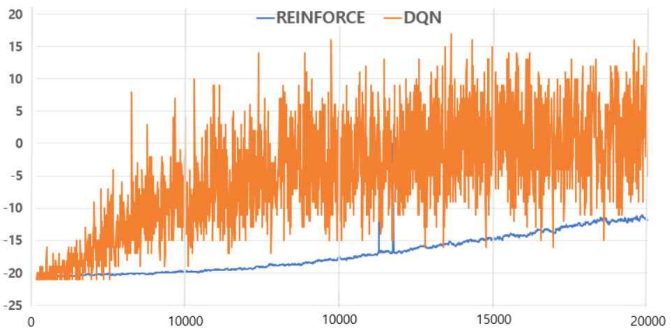
[그림 4] - Pong-v0

학습 환경 전처리



[그림 - 5] 학습 환경 데이터 가공 과정

3. 연구 결과



[그림 6] - 강화학습 방법론 간 학습 결과 비교 도표

- 시행초기 방법론 간 점수 격차 미미
- 10,000번을 기준으로 DQN을 이용한 학습에서 ϵ이 약 0.5에 가까워졌으며 -13점 획득. REINFORCE 방법론은 -17.8점 획득으로 격차가 벌어지는 것 가시적으로 확인
- 20,000번을 기준으로 DQN을 이용한 학습에서 ϵ이 약 0.15에 가까워졌으며, 획득 점수는 14점의 보상획득. REINFORCE 방법론의 경우 -11.8점 획득

4. 결론 및 향후계획

가. 결론

- 두 방법론은 시행 횟수의 증가에 따라 점차 학습이 진행되는 공통점 존재한다.
- DQN을 이용한 학습이 REINFORCE를 이용한 학습보다 진행속도가 빠르다.
- REINFORCE를 이용한 학습이 DQN을 이용한 학습보다 보상의 편차가 고르게 분포되어 안정적인 학습이 진행된다.

나. 향후 계획

- Double Q-Learning과 Actor-Critic Policy Gradient의 방법론으로 확장 및 적용하여 기존의 결과와 비교 계획