[2022학년도 2학기 강화학습기초및응용]

Lunar Lander 환경 내
DEEP Reinforcement Learning 성능 평가

인공지능응용학부
202211504 강민정₩

---

* 기본 실험 환경: Lunar Lander (https://www.gymlibrary.dev/environments/box2d/)

*** 비교할 Deep RL 기법들: DQN/A2C/PPO**

* 비교할 성능 지표: 시간에 따른 학습 성능

---

1. 강화학습

강화학습의 최종 목표는 환경(Environment)과 상호작용하는 Agent를 학습시키는 것이다. Agent는 상태(State)라고 부르는 다양한 상황 안에서 행동(Action)을 취하면서 학습을 하게 된다. Agent가 취한 행동에 따라 양(+)이나 음(-), 또는 0의 값을 돌려받게 되는데 이를 보상(Reward)이라고 한다. Agent의 최종 목표는 처음 시작 시점부터 종료 시점까지 일어나는 모든 에피소드(Episode)에서 받을 수 있는 보상 값을 최대화하는 것이다. 이를 위해서 양의 보상 값을 받을 수 있도록 행동을 강화하게 된다. 이렇게 agent가 학습하는 과정에서 점점 발전하며 내리는 의사결정 전략을 정책(Policy)라고 한다. 아래 그림[1]은 agent와 환경 간 상호작용하는 예시를 볼 수 있다. 각 time step에서 agent는 해당 환경에서 정책에 따라 행동을 취하게 된다. 이후 환경은 agent에게 보상 값을 전달하고 새로운 state로 전환하게 된다

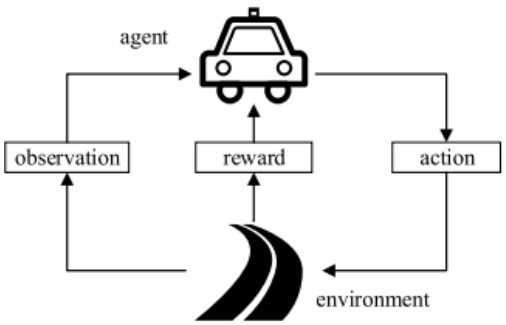


그림 1 환경과 에이전트의 상호작용

2. 알고리즘

2.1 Proximal Policy Optimization (PPO)

[강화학습기초및응용]

기말 과제: DEEP RL 성능 평가

Proximal Policy Optimization (PPO) 알고리즘[2]은 model-free based learning 방식으로 policy gradient learning의 단점을 보안하였다. 학습데이터를 재사용하는 모델로, Episode 단위로 반영하는 것이 아닌 step 단위로 학습 데이터를 만들어 내어 학습하는 방식으로 학습 효과를 높이는 방식을 취하고 있다.

2.2 Advantage Actor-Critic (A2C)

Advantage Actor-Critic (A2C) 알고리즘[3]은 한 입력층을 가지고, 은닉층에서 확률 분포를 반환하는 정책망과 상태의 가치를 반환하는 가치망으로 나뉘어 각각 판단하는 알고리즘이다. Actor(정책망)은 각 행동에 대한 확률 분포를 반환하게 되는데 정책망이 어떠한 행동을 취해야 하는지를 알려준다. 반면 Critic(가치망)은 상태의 가치를 반환하며 에이전트의 행동이 예상보다 얼마나 좋을지를 평가하게 된다.

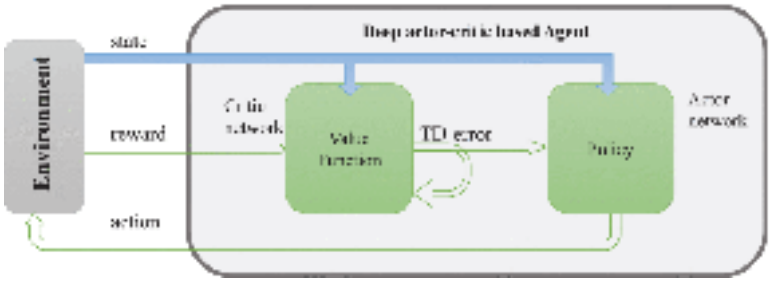


그림 2 Actor-critic 아키텍처

2.3 Deep Q-Network (DQN)

강화학습과 딥러닝을 결합하여 딥러닝을 통해 q-function을 근사하고자 하는 많은 연구가 진행되었으나 학습의 불안정성을 보이거나 알고리즘이 수렴되지 않는 한계점을 가지고 있었다. 이러한 학습 불안정성의 원인은 sample correlation, data distribution의 변화, 움직이는 target value로 볼 수 있다. Deep Q-Network(DQN)[4]는 experience replay와 target network라는 개념을 도입하여 기존 deep reinforcement learning의 한계점을 극복하고자 하였다.

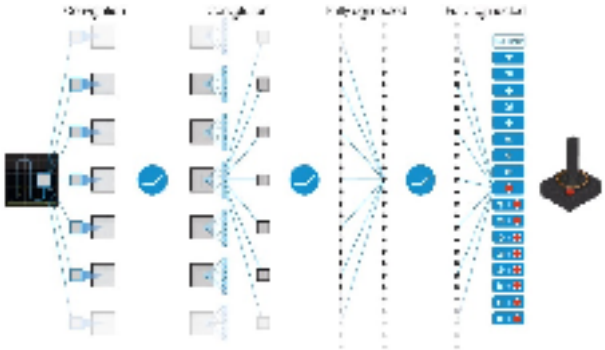


그림 3 DQN의 아키텍처

2.4 Double Deep Q-Network (DDQN)

기존 DQN 방식은 q-learning에서 max 연산자를 사용하기 때문에 q-value를 overestimate하는 문제자 존재한다. 따라서 q-value를 추정하는 과정에서 실제로는 다른 action이 최적임에도 불구하고 반드시 큰 값을 고르는 한계점이 있어 해당 문제를 해결하고자 Double Deep Q-Network (DDQN)[5]가 등장했다. DDQN은 각각 독립적으로 q-function을 추정하는 방식으로 학습이 진행된다. 두 개의 q-function 중 하나는 action을 선택하고, 하나는 action을 평가하는 방식이며 DQN 대비 더 효과적이다.

3. Lunar Lander 환경

Lunar lander 환경의 주요 목표는 AI 우주선(agent)이 OpenAi Gym에서 제공하는 시뮬레이션 환경에서 적절한 위치에 원활하게 착륙하는 방법을 스스로 학습하는 것이다. 아래의 그림[6]처럼 달 착륙선을 두 개의 깃발로 표시된 착륙 패드에 잘 착륙하는 것이 목표라고 할 수 있다. 해당 우주선의 연료는 무한히 존재한다고 가정하며, 총 4개의 이산적(discrete)인 행동이 존재한다. Lunar lander 환경 내 4개의 action space는 'do nothing', 'fire left orientation engine', 'fire main engine', 'fire right orientation engine'이다. 달 착륙선이 착륙 패드에 잘 도착한 후 휴식을 취하게 되면 100-140 사이의 보상을 받고, 착륙 패드에서 먼 곳에 도착하면 보상 값이 줄어든다. 만약 착륙선이 부서지면 추가적으로 -100의 보상을 받게 되며 착륙 후 휴식을 취하면 100의 추가적인 보상을 받게 된다. 200의 보상을 받게 되면 lunar lander 게임이 성공적으로 학습했다고 가정한다.

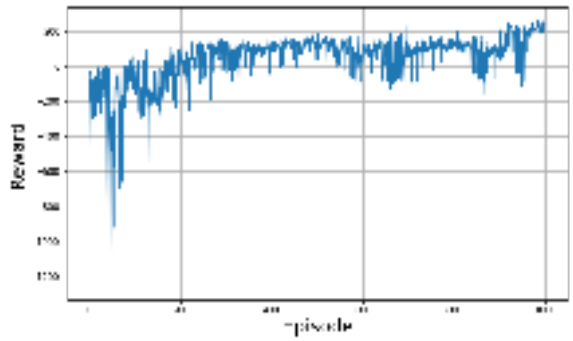
그림 4 Lunar lander 환경 시각화

4. 실험 결과 및 결론

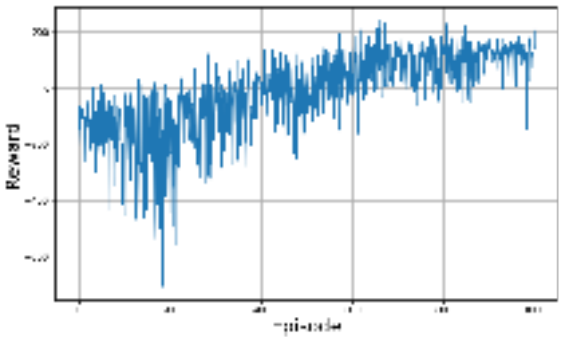
아래는 Lunar lander 환경 내 PPO, A2C, DQN, DDQN 총 4개의 알고리즘의 episode 별 reward의 값을 나타낸 그래프이다. Lunar lander 환경은 reward 값이 200이 되면 학습이 종료되는데 본 실험에서는 4개의 모델 모두 동일하게 종료 조건 없이 훈련되었다. Episode 값은 1000으로 설정

기말 과제: DEEP RL 성능 평가

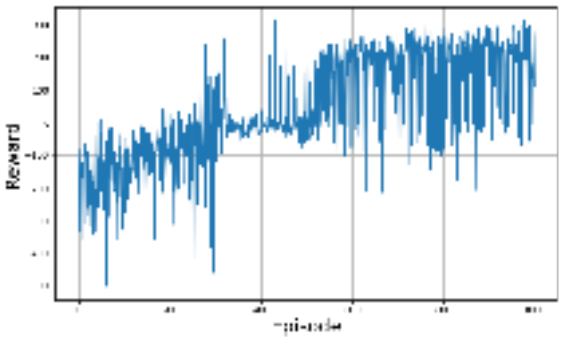
하였으며 threshold 값을 따로 설정하지 않아 early stopping 없이 훈련된다. 실험 결과, PPO 알고리즘이 다른 모델 대비 200 이상의 reward 값에 가장 먼저 도달하지는 않지만 도달 이후에는 가장 안정적으로 수렴되며 훈련되는 것을 확인할 수 있다.



PPO



A2C



DQN

참조문헌

[1] Wang, X., Wang, S., Liang, X., Zhao, D., Huang, J., Xu, X., ... & Miao, Q. (2022). Deep reinforcement learning: a survey. IEEE Transactions on Neural Networks and Learning Systems.

[2] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

[3] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K. (2016, June). Asynchronous methods for deep reinforcement learning. In International conference on machine learning (pp. 1928-1937). PMLR.

[4] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. nature, 518(7540), 529-533.

[5] Van Hasselt, H., Guez, A., & Silver, D. (2016, March). Deep reinforcement learning with double q-learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 30, No. 1).

[6] OpenAI gym Documentation