|  |  |
| --- | --- |
| 과목명 | 강화학습 (7-2) |
| 이름 | 이덕수 |
| 학번 | 20161600 |

이름과 학번을 적고, 답변을 적어주세요.

실습: 수업 내용을 자유롭게 요약하세요(A4 1 page 분량, 넘어도 상관없음)

- 지금까지 OpenAI(Retro-Game) UnityML(Lego,3DBall)등을 진행하였음

아마존 웹 서비스(이하 AWS) 에선 2018년 딥레이서(자율주행) 라는 강화학습 프레임워크 를 공개 이 프로그램을 사용하면 누구나 간단하게 자신만의 자율 주행 강화 학습 모델을 훈련시키고(PPO) 배포할수 있음 이를 적용하여 실제로 볼수있는 모형자동차 도 판매하고 있음 (AWS DeepRacer)

PPO 알고리즘은 2017년 OpenAI 팀이 발표한 논문 에서 소개된 최신 기법임

PPO 알고리즘은 로봇 제어, 게임 등 다양한 강화학습 등에서 우수성이 입증되었고 대표적 으로 DOTA2라는 게임에 활용됨 정책 경사 방법을 사용 하고 있고

특성 상태에서 취하는 행동의 확률 분포로 이루어진 정책 네트워크를 학습시킴

정책 네트워크는 신경망이 될수도 있고 다른 형태의 함수 근사 방법이될수도있음

상태를 입력으로 한다는 점에서 DQN의 신경망과 같지만 DQN의 출력이 Q함수

였다면 정책 네트워크의 출력은 행동의 확률 분포임. DQN과 비교할때 정책 경사의 장점 2가지는 DQN은 에이전트가 취할수있는 행동이 불연속한 경우에만 적용 가능 하지만 정책 경사는 출력인 '확률 분포'를 '정규 분포' 라고 가정하고 샘플링을 진행하면 에이전트의 연속된 행동을 추출할 수 있음 고로 정책 네트워크는

확률 분포를 기반으로 해서 행동을 선택함 무작위성을 부여하는 앱실론 같은 별도 개수가 필요하지 않고 에이전트는 기존 정책을 그대로 준수하며 에피소드를 진행해도 무방 그래서 정책 경사는 정책 이탈이 아닌 정책 밀착 기법에 속함

정책 경사의 순서론 정책을 무작위로 생성, 에이전트는 환경 내에서 정책에 따라 행동하며 확인한 일련의 상태와 보상을 샘플로 취함, 특정 행동으로 인한 보상을 확인 1.보상이 기댓값보다 높으면 해당행동 확률 증가 2. 기댓값 보다 낮으면 해당 행동 확률 감소 이를 반복하며 정책 파라미터를 조절함 그러나 단점으로

현재 정책을 최대한 개선하고싶지만 너무많은 에피소드가 추가되면 속도가 느려지고 또한 이를 개선하다 성능이 하락하면 안됨 이를 해결하기 위해 정책 갱신 전후의 비율을 클리핑 하는 방식으로 정책 업데이트의 양을 제한하여 위 문제

2가지를 해결함