강화학습 변천사

덕근짱

강화학습이란.

- MDP(Markov Decision Process)기반의 최적화 개념과
  동물심리학을 결합한 인공지능 기반의 머신러닝 알고리즘이다.
- 환경을 중심으로 Agent가 환경에서 파생되는 여러 데이터와 상호작용하여 보상을 얻고,
  이 보상을 최대로 얻는 action을 선택하도록 개선하는 학습 방법이다.
- 또한 적절한 환경의 정의, 행동결정, 보상함수와 정책 설정의 최적화를 통해서 좋은 결과
  를 얻어내야 한다.
- 이 3번째 항목이 현재 내가 관심 있게 보는 것이다.

알고리즘 변천사

연도	알고리즘	특징
2015	Deep Q-Network(DQN)	- 신경망을 적용하여 Action 및 State에 대한 Q-function을 근사화 - Experience Replay Memory를 사용하여 Data Resource 절약
2015	Trust Region Policy Optimization(TRPO)	- Actor Critic(2000) 알고리즘 기반으로 반복되는 학습 정책을 정규화 - 규제(Penalty) 사용으로 불필요한 학습을 억제하여 빠르게 수렴
2016	Deep Deterministic Policy Gradient(DDPG)	DQN의 Replay Memory를 사용하여 연속적인 Batch 업데이트가 가능 - 빠른 수렴을 위해 정책제어를 사용함
2017	Asynchronous Advantage Actor Critic(A3C)	Actor Critic(2000) 알고리즘을 기반으로 A2C, A3C로 발전 - 멀티 A2C 환경을 동시에 학습하여 일관된 학습을 유도 - Discrete와 Continuous space에서 동시에 사용 가능
2017	Proximal Policy Optimization (PPO)	- TRPO와 유사한 알고리즘으로 대리 손실함수를 생성하여 더 빠른 수렴을 지원 - TRPO보다 구현이 간단하고 좋은 Sample Complexity를 사용할 수 있음
2018	Twin Delayed Deep Deterministic Policy Gradient (TD3)	DDPG 기반 다중 네트워크를 동시에 학습하여 정책의 과대평가(Overestimation)를 감소 - Target Action에 노이즈를 추가하여 Q-function의 오류를 방지
2018	Soft Actor Critic(SAC)	DDPG와 유사한 방식으로 Q-function의 근사값으로 탐색을 제어 - Off-policy 알고리즘을 사용하여 Sample Inefficiency를 해결할 수 있음 - 로봇 시뮬레이션 모델링을 위해 많이 사용됨

강화학습 접근 방식

- 시뮬레이션 시스템 환경으로부터 제공되는 외적 보상함수를 이용하여 환경변화에 대한 에이전트의 행동개선이 basic
- 하지만, 이 역시 사람이 직접 설정하는 것 이기 때문에 면밀하게 상호작용을 한다고 보기는 어렵다.
- 따라서 내적 보상함수를 사용하여 학습에 적용시키고자 한다.
- 에이전트가 활동하면서 환경에 '새롭게' 내재되어 있는 현상들을 학습하여 개선하는 방법이다.
- (각각의 task 케이스에 맞는 유기적 상호작용을 필두로 내세움)

Transferable Meta-skills

멀티 에이전트 강화학습 기술 동향 PDF 원문보기 인용
A Survey on Recent Advances in Multi-Agent Reinforcement Learning

- 일반적인 DNN은 성능 개선을 위해 파라미터(가중치)를 최적화 하는 방식을 택한다.
- 반면에 DQN같은 경우에는 내부의 정책을 개선하는 SGD(확률적 경사 하강법)또는 오차역전파 알고리즘을 통해 신경망을 개선하는 방식이다.
- 이와 같은 방법과 더불어 학습성능을 향상시키는 방법으로 다중에이전트가 있는데.
- 교통, 군사 목적의 강화학습 에이전트들은 기존의 환경 뿐만 아니라, 서로간의 상호작용 역시 고려해야 하기 때문에, 단일 에이전트 학습만으로는 효과를 보기 힘들다는 문제점에서 출발한 학습 기법.
- 서로의 유기적인 부분을 강조하면, 최적의 해를 찾기 힘들고, 분산형으로 각각 마이웨이를 시키면, 위에서 처럼 효과가 반감된다는 딜레마를 가지고 있음

Transferable Meta-skills(Learning)

- 이러한 분산 강화학습과 더불어, 메타기술이 접목된 강화학습 알고리즘 역시 부상하며
- 시스템 환경에서 발생하는 여러 요소에 따른 정책과 보상함수를, Meta-test 방식으로 분리하여 학습하는 방법이라고 하는데..

Meta-Learning

- 인간의 경우에는, 이전에 겪었던 여러 환경 덕분에, 새로운 task가 주어져도 빠르게 학습하는 경향이 있다. 이러한 원리를 차용하여, 다른 환경에서도 빠르게 학습 할 수 있도록 하는 것이 Meta-learning 이다.
- 따라서 meta 데이터라는 소규모의 데이터를 미리 학습 시킨 후에.
- 그것으로 feature를 찾아내어 벡터 방식인 knn을 통해서 접근하는 방식
- 그냥 그 상태로 집어던진 후에, context벡터와 같은 메모리를 얻게 하여 학습하는 방식인 model-based method라는 방식 역시 있으며,
- 그리고 마지막으로 meta dataset 으로 optimization을 하는 방식이 있다.

https://lilianweng.github.io/lil-log/2018/11/30/meta-learning.html

그래서 RL에는 어떻게 적용이 되는가.

- 보시다시피 초반에 제공되는 적은 데이터로 pretraining을 성공적으로 하는 것이 목표이기 때문에,
- 초반에 에피소드를 실행하되, 리턴 값들에 계수를 높게 잡아서 탐색 시키는, 그러한 방식으로의 접근이라고 볼 수 있다.
- 이 와 같은 부분은 초기값 설정을 랜덤으로 하지 않고 RMSE와 같은 방법을 사용하는 선형회귀와 비슷하다고 여겨진다.