동아대학교 DONG-A UNIVERSITY

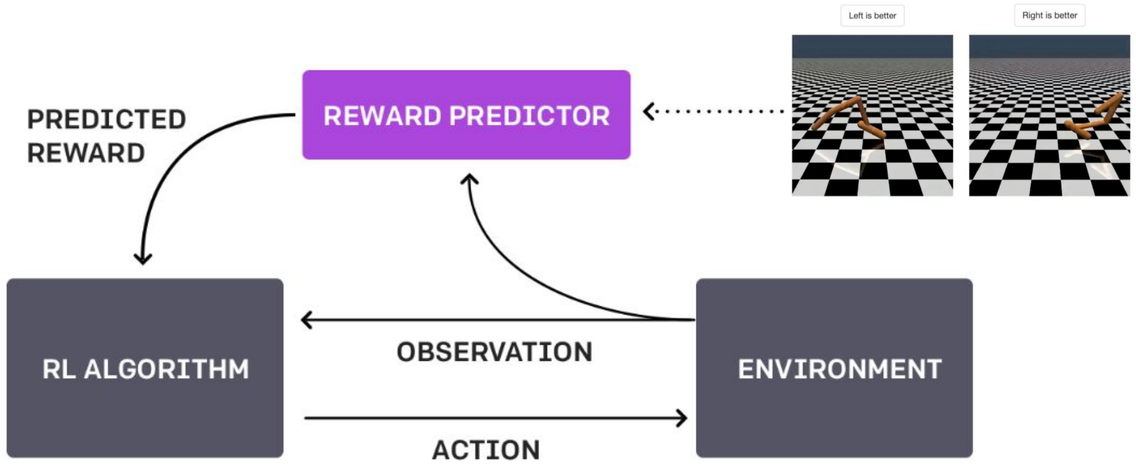
[실증적 SW개발프로젝트]

RLHF기반 로봇 팔 제어 프로그램 개발

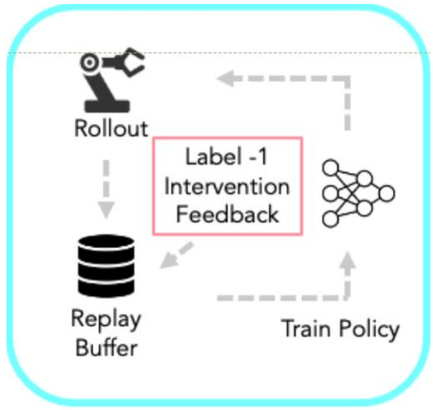
2143841 권은주
1824751 진현석
2051505 조현진

Contents

주제: 다양한 환경에서 강화학습 알고리즘 RLIF와 RLHF의 성능 비교 연구



1. RLHF 알고리즘 구조도



2. RLIF 알고리즘 구조도

3. 알고리즘 비교

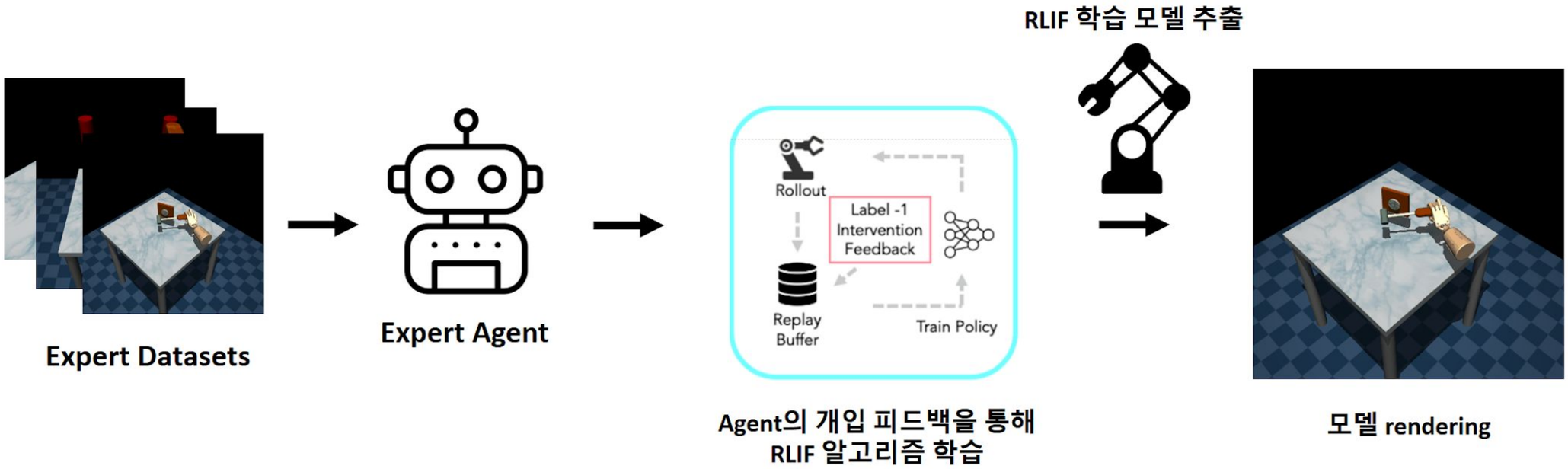
알고리즘	피드백 유형	Optimal policy	학습 효과
RLIF	Agent의 행동이 잘 못 된 경우 reward값 -1로 변경 (기본 0)	필요	복잡한 task 수행 가능 그러나, 인간이 행동에 대한 명확한 지식이 없을 때 학습 제한
RLHF	랜덤으로 주어지는 Trajectory 중 하나 선택	불필요	최적 행동을 몰라도 상대적으로 더 나은 trajectory 선택 가능 학습에 제한 X

목표: 다양한 환경에서 RLIF와 RLHF의 성능을 비교하여, 복잡한 작업 수행 시 RLHF의 상대적 이점을 증명

이유:

1. 알고리즘의 특성을 살펴보았을 때 **차원**이 **복잡**해질수록 **RLIF보다 RLHF가 더 좋은 성능**을 **나타낼 것으로 예상**
2. **RLHF**는 (RLIF 실험 환경 수준의) **복잡한 환경에서 실험 결과 부재**

1학기 목표: RLIF 알고리즘 학습 파이프라인 구축



실증적AI프로젝트 금주 활동내역

동아대학교 DONG-A UNIVERSITY

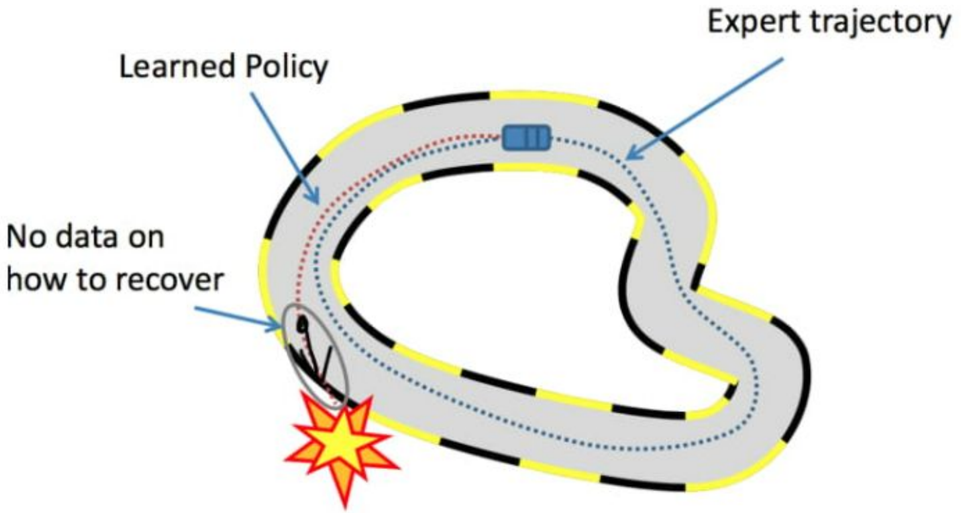
주제: RLHF를 이용한 협동 로봇 제어 프로그램 개발

금주 활동계획	1. Ardoit env expert RLIF 알고리즘 적용 2. 학습 성공 모델 그래프, 렌더링 3. 다양한 오프라인 강화학습, 모방학습 알고리즘으로 expert 생성		
	팀장 (권은주)	팀원 1 (조현진)	팀원 2 (진현석)
금주 개인별 활동내역	1. expert model RLIF 알고리즘 적용 2. 학습 성공 모델 그래프, 렌더링	1. 오프라인 강화학습 알고리즘 정리	1. 모방학습 알고리즘 정리
차주 활동계획	1. 실험 결과 정리 2. D4RL: Datasets for Deep Data-Driven Reinforcement Learning 리뷰		

1.Behavioral Cloning (행위 복제, BC)

정의

Expert의 행동을 직접 모방하는 모방학습의 기본적인 형태
Expert가 수행한 행동을 데이터로 수집하여 State-Action 쌍을 학습하



특징

- **학습 데이터**: State-Action 쌍
- **정책 학습**: State에 대해 적절한 행동을 예측하도록 학습

장점

1. **간단함**: 알고리즘이 단순하고 구현이 쉬움
2. **빠른 학습**: Expert-Data가 충분하면 빠르게 학습
3. **명확한 목표**: Expert의 행동을 그대로 모방하는 명확한 목표를 보

단점

1. **데이터 의존성**: Expert-Data 가 충분하지 않거나 불완전할 경우 성능이 저하될 우려
2. **오차 증폭 문제**: 학습된 정책이 전문가 데이터와 다른 행동을 할 경우, 그 차이가 누적되어 큰 오차로 이어짐
3. **일반화 어려움**: 새로운 상태에 대한 일반화가 어려울 수 있으며, 학습하지 않은 상태에서의 성능이 보장 X

2.Inverse Reinforcement Learning (역강화학습 , IRL)

정의

Expert의 행동을 통해 Reward 함수를 추정하고, 이를 기반으로 정책을 학습하는 방법
IRL은 Expert가 어떤 Reward 함수를 최대화하려고 하는지를 학습하는 것이 목표

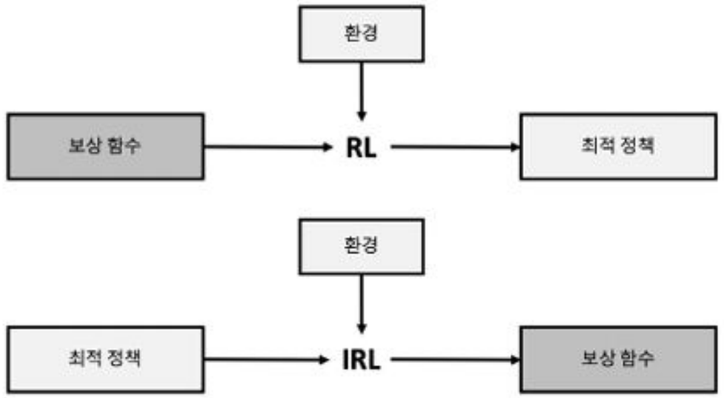


그림 1 RL과 IRL 개념 비교

특징

- **학습 데이터**: Expert의 State-Action 궤적
- **정책 학습**: 추정된 Reward 함수를 통해 최적의 정책을 학습

장점

1. **보상 함수 학습**: Reward 함수를 직접 학습하기 때문에, 학습된 정책이 보상 함수를 최대화하도록 유도
2. **일반화 능력**: 학습된 Reward 함수를 사용하여 다양한 환경에서 일반화된 정책 도출
3. **효율적 학습**: Expert-Data가 적을 때도 비교적 효율적으로 학습

단점

1. **복잡성**: Reward 함수를 추정하는 과정이 복잡하고 계산 비용이 많이 들 수 있다
2. **불확실성**: 추정된 Reward 함수를 Expert의 실제 의도를 정확히 반영하지 못할 수 있다
3. **데이터 요구**: Expert의 행동 궤적 데이터가 필요하며, 부정확하거나 불완전할 경우 성능에 영향을 미칠 수 있다

3. Generative Adversarial Imitation Learning (적대적 모방학습)

정의

생성적 적대 신경망(GAN)의 아이디어를 모방학습에 적용
두 개의 신경망, Generator(실제 데이터와 유사한 가짜 데이터를 생성하는 모델)와 Discriminator(주어진 데이터가 진짜인지 가짜인지를 판별하는 모델)가 서로 경쟁하면서 학습
GAIL에서는 Generator가 Expert의 행동을 모방하고, Discriminator가 Expert의 행동과 Generator의 행동을 구별하며 학습

특징

- **학습 데이터**: Expert의 State-Action 궤적
- **정책 학습**: Generator가 Expert와 구분되지 않도록 학습

장점

1. **표현력**: Generator는 Expert의 복잡한 행동 패턴을 학습할 수 있는 강력한 모델을 사용
2. **일반화 능력**: 다양한 상태에서도 Expert와 유사한 행동을 생성할 수 있는 능력을 학습
3. **자동 구별 학습**: 구분자를 통해 Expert와 생성자의 행동을 자동으로 구별하며 학습

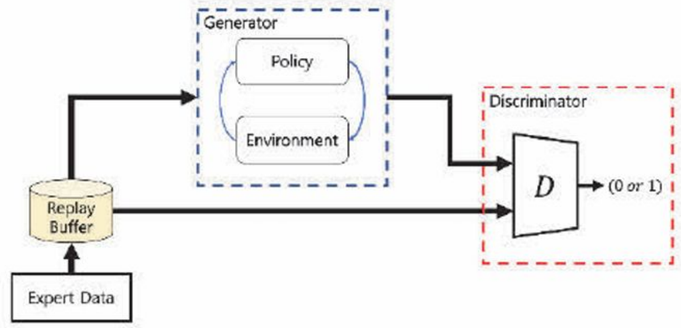


Fig. 3 Generative adversarial imitation learning

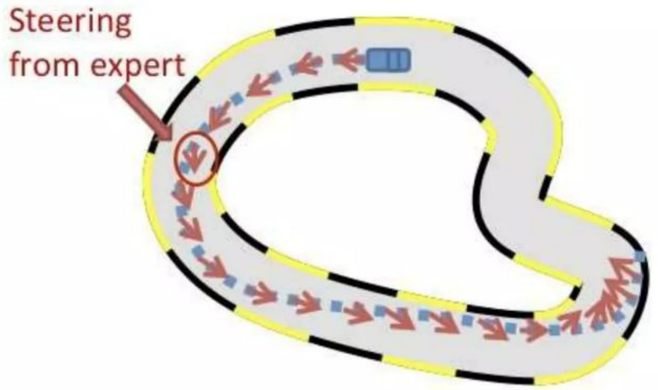
단점

1. **학습 불안정성**: GAN과 마찬가지로, 생성자와 구분자의 학습이 불안정할 수 있으며, 서로 균형을 맞추는 것이 어려울 수 있다
2. **복잡성**: 두 개의 네트워크를 동시에 학습시키는 것은 구현이 복잡하고 계산 비용이 많이 든다
3. **훈련 시간**: 생성자와 구분자 간의 상호 학습 과정이 시간이 오래 걸릴 수 있다

4.Dataset Aggregation (DAgger)

정의

Expert의 시연 데이터와 Agent의 경험을 반복적으로 결합하여 학습하는 모방학습 방법
초기엔 Expert의 정책을 따라가다가, Agent가 수행한 행동을 추가로 수집하여 데이터셋을 확장하고, 점점 더 나은 정책을 학습



특징

- **학습 데이터**: Expert의 State-Action 쌍과 Agent의 Action-Data를 결합
- **정책 학습**: Expert 와 Agent의 혼합 데이터를 통해 점진적으로 개선되는 정책을 학습

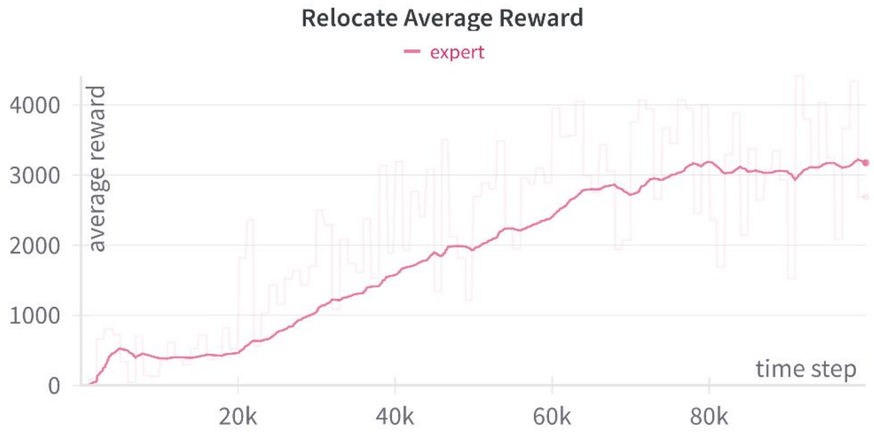
장점

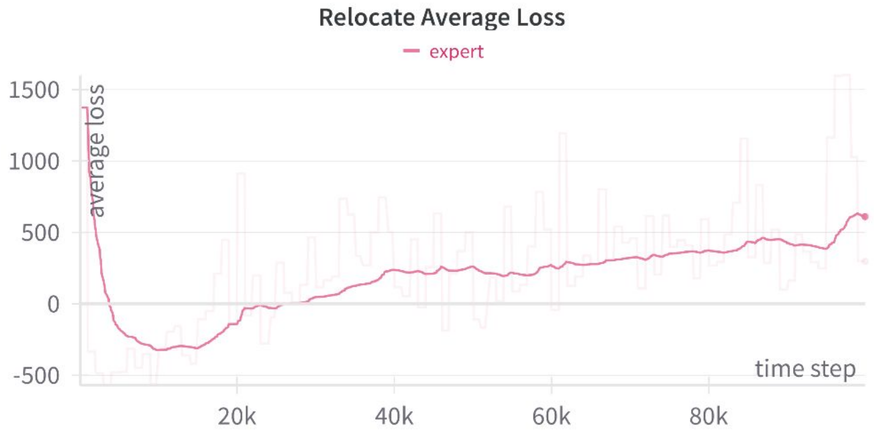
1. **오차 축소**: 전문가와 에이전트의 행동 데이터를 결합함으로써 오차 증폭 문제를 감소
2. **향상된 성능**: 반복적인 데이터셋 확장을 통해 에이전트의 정책이 점진적으로 개선
3. **데이터 효율성**: 전문가 데이터만으로 학습하는 것보다 더 적은 데이터로도 효율적으로 학습

단점

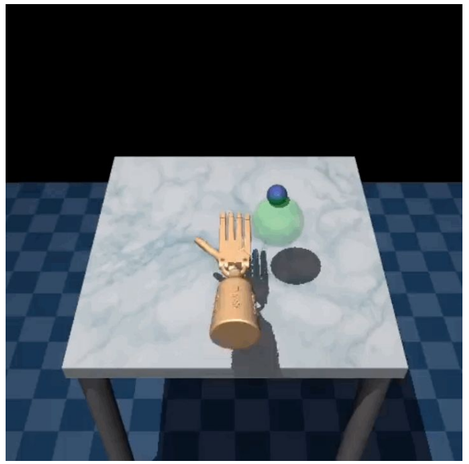
1. **복잡성 증가**: 반복적인 데이터셋 확장 과정이 추가적인 복잡성을 가져올 수 있다
2. **전문가 개입 필요**: 학습 과정 중 Expert의 개입이 필요할 수 있어, Expert의 노력이 많이 요구
3. **훈련 시간 증가**: 반복적인 데이터 수집과 학습 과정이 시간을 많이 소요

Expert Train Result





expert data



Agent

02. Expert Agent 학습 결과

- 직접 만든 **expert model**이 코드에서 작동하지 않아 제작자에게 요청
- 새로운 **expert file**이 생성 불가능 하다면 기존에 제공하는 **Pen, Walker2d, Hopper** 환경에서 실험할 예정

How to make an expert model? #1

Open **eunjuyummy** opened this issue yesterday · 0 comments

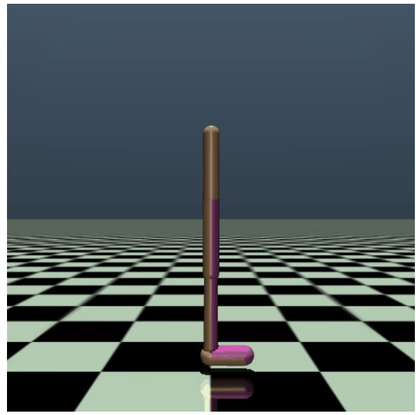
eunjuyummy commented yesterday

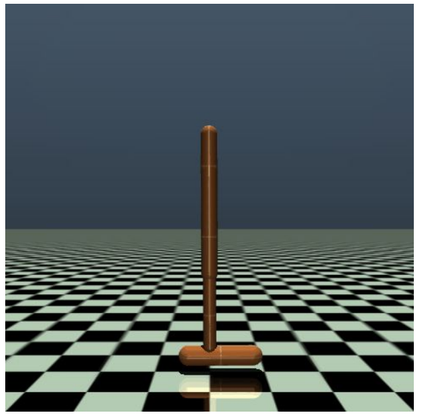
Hi, I'm a college student studying reinforcement learning.

I want to learn how to create an expert model (pkl) and run my code in a new environment.

Now, I tried to create a model using torchRL's IQL algorithm and run the code, but it didn't work well.

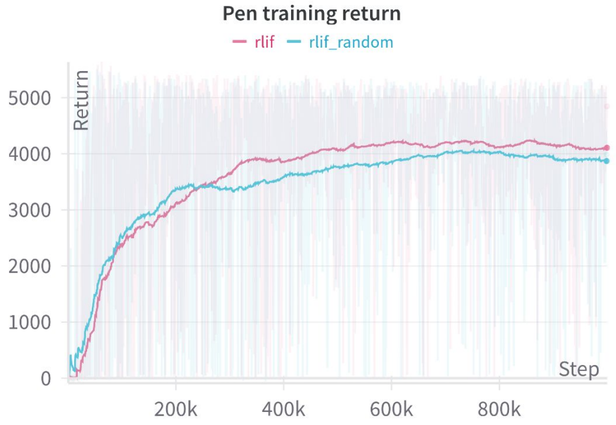
Your research is incredibly impressive. Thank you very much for sharing it.



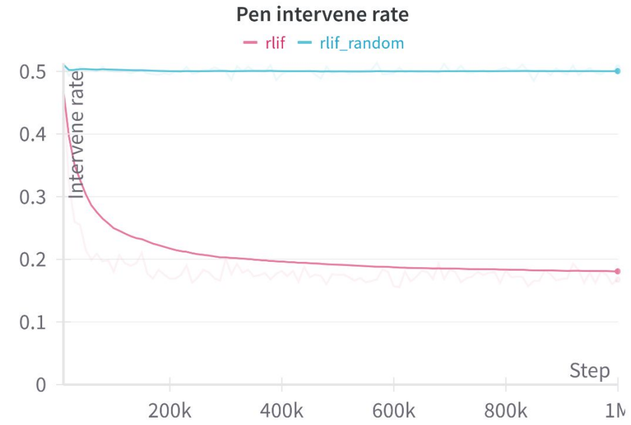


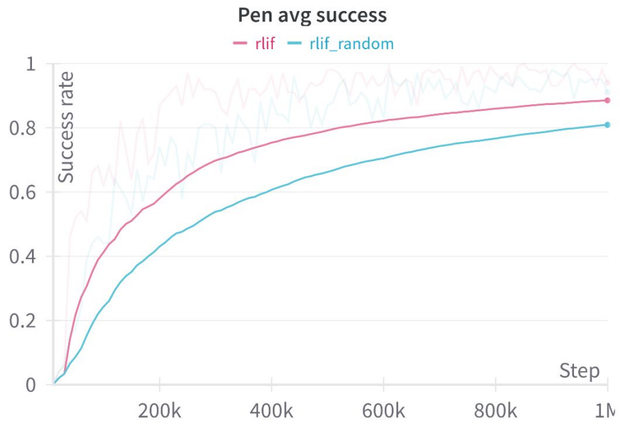
02. RLIF 학습 결과

- **rlif, rlif_intervene random가지 알고리즘 비교**
- **RLIF: 실제 policy의 q값과 Exeprt Agent의 q값을 비교하여 개입 여부를 결정**
- **RLIF_random: random으로 정해진 random 비율에 따라 개입 (50%)**



Pen Train Return: 학습 과정에서 받는 Return 값
*Return: 현재시점부터 미래보상까지 더해진 값





실증적AI프로젝트 금주 활동내역

동아대학교 DONG-A UNIVERSITY

주제: RLHF를 이용한 협동 로봇 제어 프로그램 개발

금주 활동계획	1. PureRL, IQL과 비교 그래프 2. D4RL: Datasets for Deep Data-Driven Reinforcement Learning 리뷰		
	팀장 (권은주)	팀원 1 (조현진)	팀원 2 (진현석)
금주 개인별 활동내역	1. PureRL, IQL과 비교 2. 그래프, 렌더링 결과 정리	1. 프로젝트 내용 총 정리	1. D4RL 리뷰
차주 활동계획			

Questions & Answers

Dept. of AI, Dong-A University

권은주 (kkkoj4284@donga.ac.kr)

진현석 (cpu132465@donga.ac.kr)

조현진 (gkfkgkdh@naver.com)

Github (https://github.com/eunjuyummy/AI_Project_CoRLHF)