동아대학교 DONG-A UNIVERSITY

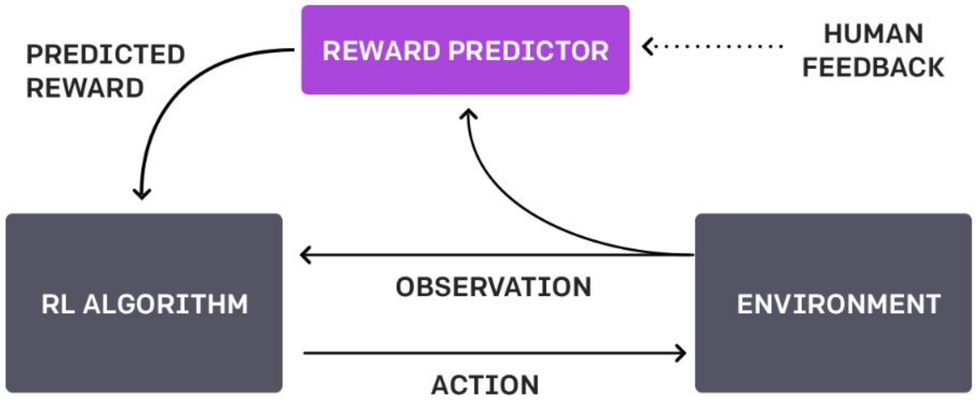
[실증적SW개발프로젝트]

RLHF기반 로봇 팔 제어 프로그램 개발

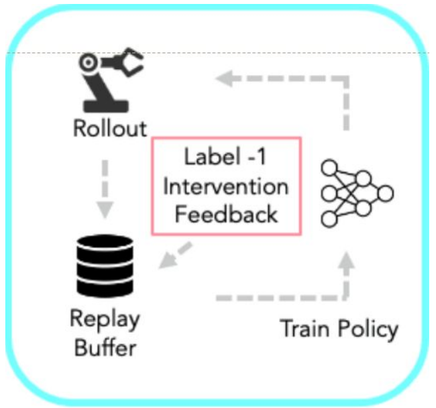
2143841 권은주
1824751 진현석
2051505 조현진

Contents

주제: 고차원 환경에서 강화학습 알고리즘 RLIF와 RLHF의 성능 비교 연구



1. RLHF 알고리즘 구조도



2. RLIF 알고리즘 구조도

3. 알고리즘 비교

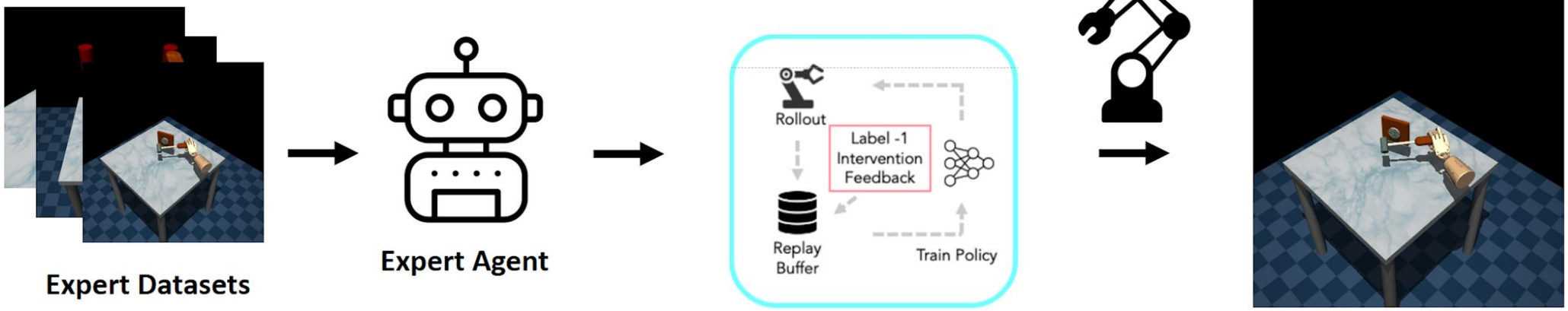
알고리즘	피드백 유형	Optimal policy	학습 효과
RLIF	Agent의 행동이 잘 못 된 경우 reward값 -1로 변경 (기본 0)	필요	복잡한 task (로봇 팔 작업)수행 가능 그러나, 인간이 행동에 대한 명확한 지식이 없을 때 학습 제한
RLHF	랜덤으로 주어지는 Trajectory 중 하나 선택	불필요	최적 행동을 몰라도 상대적으로 더 나은 trajectory 선택 가능 학습에 제한 X

목표: 고차원 환경에서 RLIF와 RLHF의 성능을 비교하여, 복잡한 작업(로봇 팔 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 **복잡한 환경**(로봇 팔 작업)**에서 실험 결과 부재**

1학기 목표: RLIF알고리즘 학습 파이프라인 구축



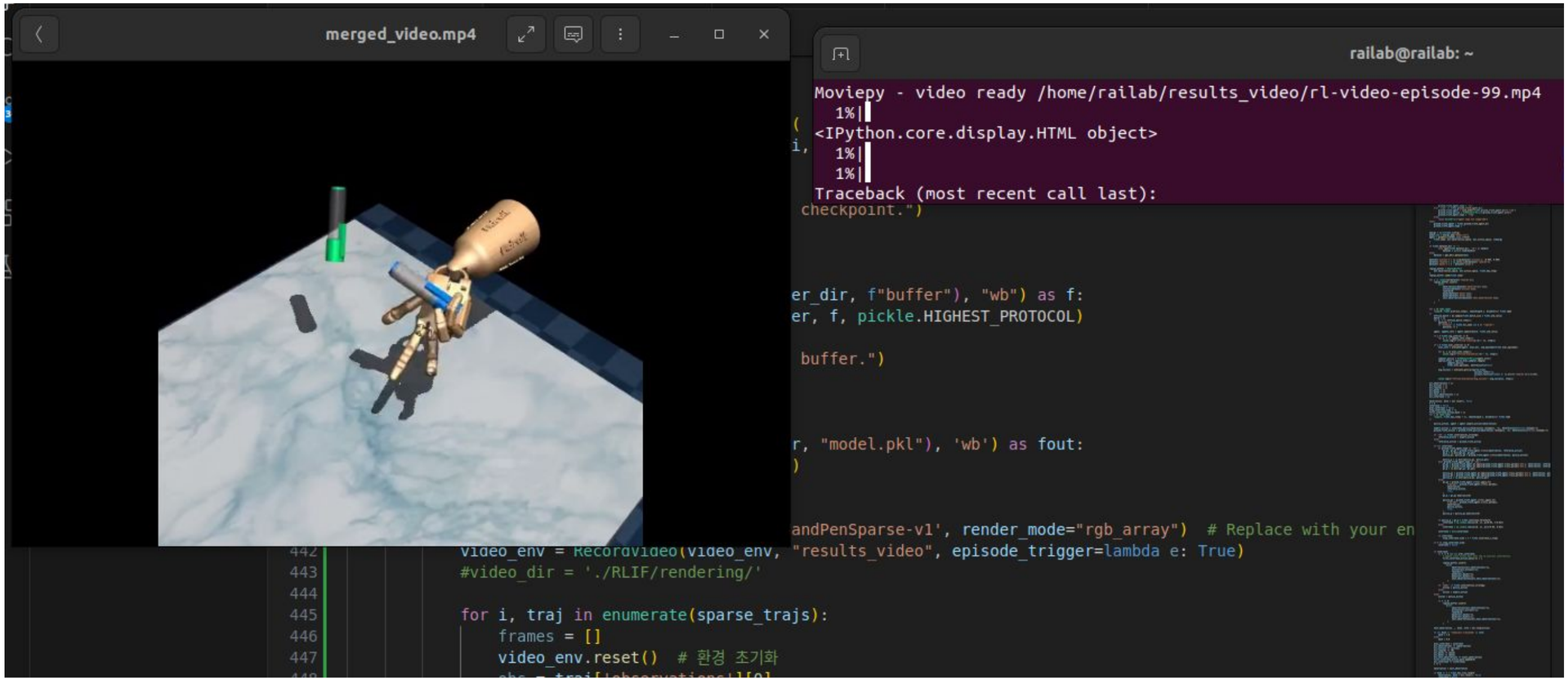
실증적AI프로젝트 금주 활동내역

주제: RLHF를 이용한 협동 로봇 제어 프로그램 개발

금주 활동계획	1. Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations 논문 정리 2. DAPG project github 내용 정리 3. Hammer RLIF 알고리즘 학습		
	팀장 (권은주)	팀원 1 (조현진)	팀원 2 (진현석)
금주 개인별 활동내역	1. Hammer expert model RLIF 알고리즘 적용 a. GPU 에러 해결 2. 학습 결과 GUI 생성	1. DAPG 논문 정리	1. DAPG project github 내용 정리
차주 활동계획	1. Ardoit hammer expert RLIF 알고리즘 적용 2. 학습 성공 모델 그래프, 렌더링 3. 다양한 오프라인 강화학습, 모방학습 알고리즘으로 expert 생성		

02. RLIF 학습 영상 생성

- 총 100만 번 학습 중 1e4 단위로 학습한 모델의 평가 결과를 영상으로 확인



기존 방식

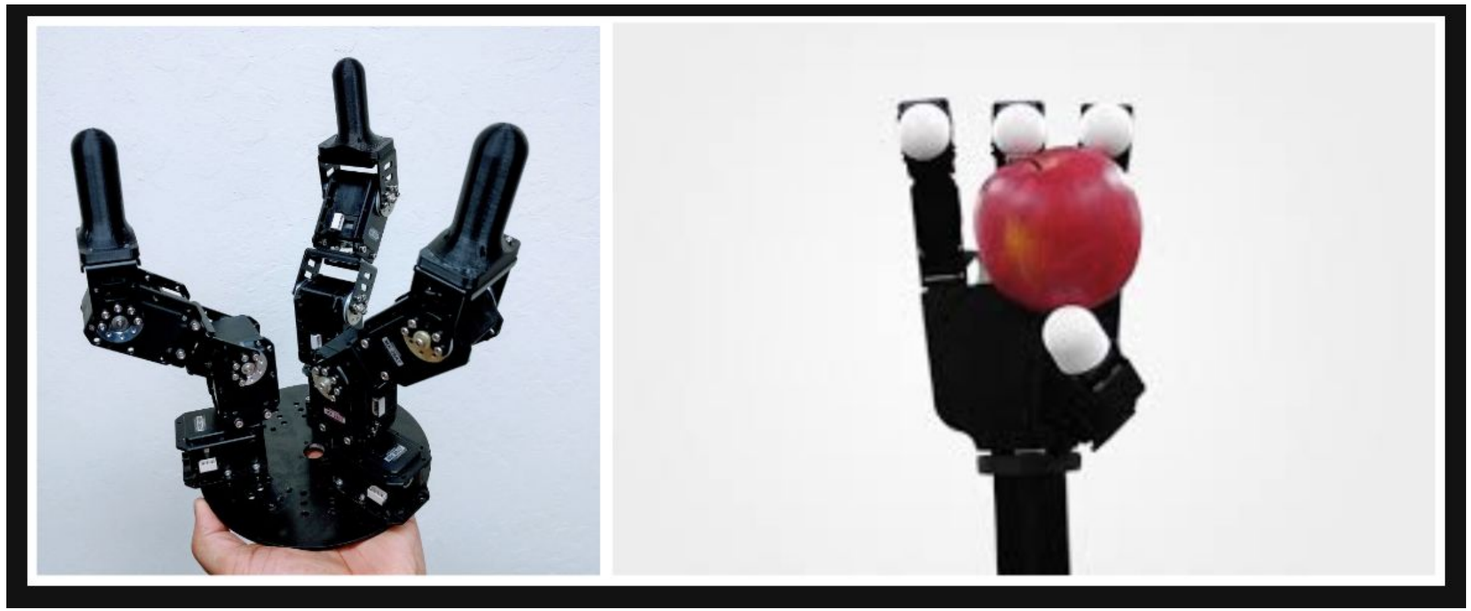
- 학습된 모델을 불러와서 **render_mode를 human**으로 변경해서 사용

RLIF에서 영상을 생성하기 위해 사용한 방식

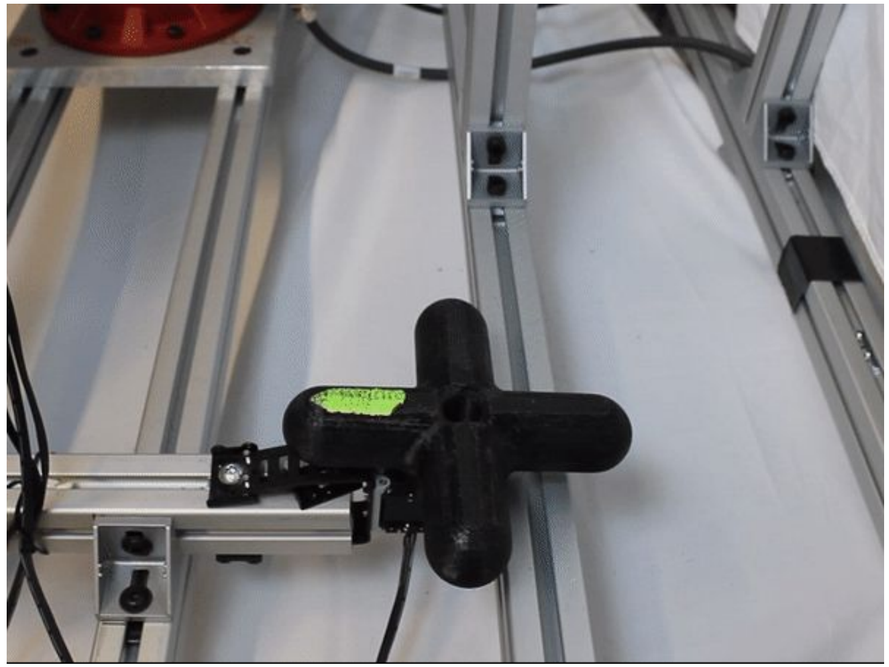
- 학습된 모델의 일부 **sample trajectory**를 불러와 환경에서 순서대로 불러올 수 있도록 함.

simple parallel jaw grippers

- 높은 가격
- 방대한 양의 시뮬레이션 데이터 필요
  □ 비용 증가



Dynamixel Claw($2500)　　*Allegro Hand*($15,000)



실제 환경에서 RL학습의 장점

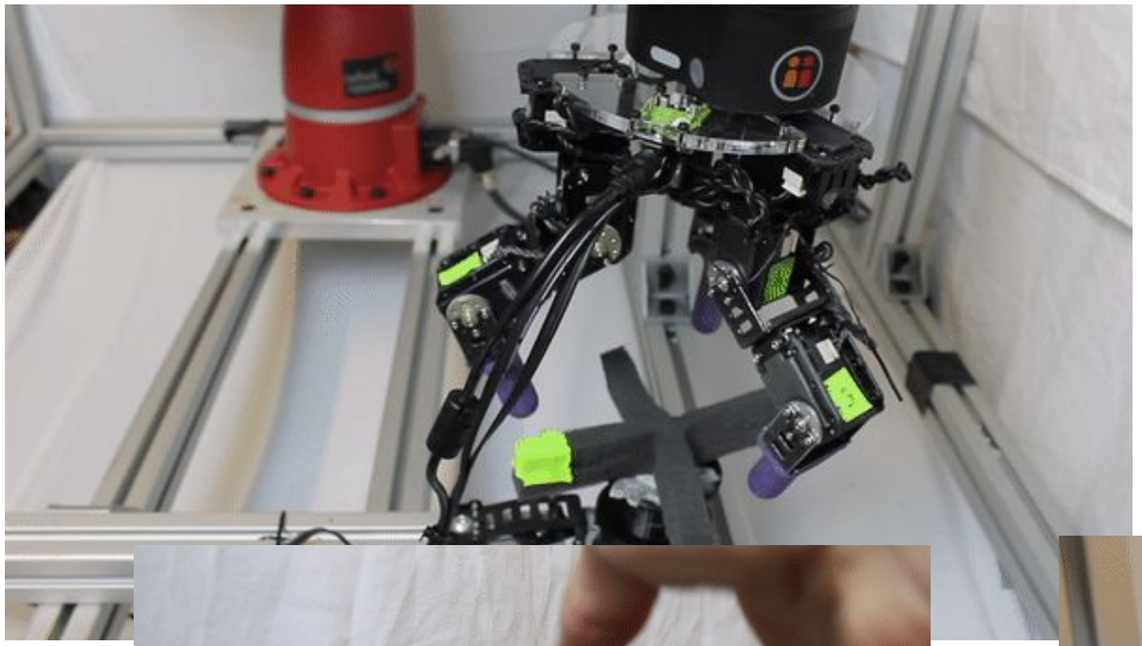
- 최소한의 가정을 필요
- 다양한 기술 레퍼토리를 자율적으로 습득하는 데 적합
- 말랑한 재질의 폼 벨브의 회전을 학습 가능
- 저예산 + 데스크탑

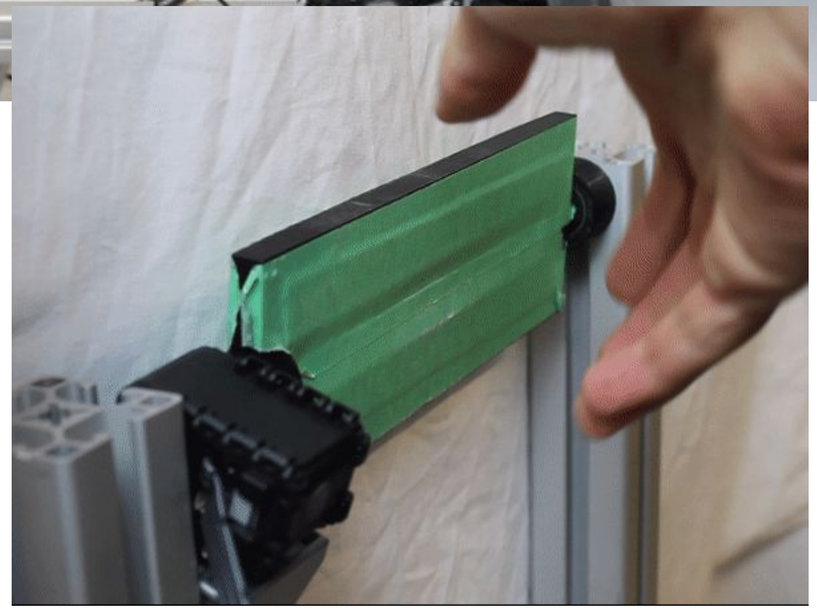
실험 환경의 목표

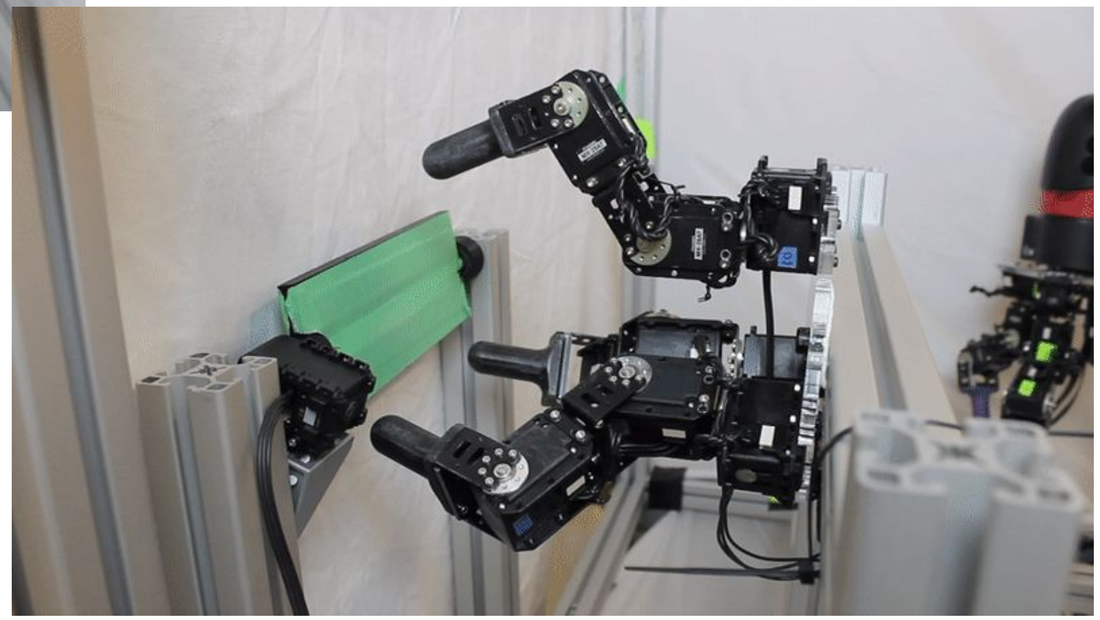
- 밸브나 수도꼭지를 180도 회전

실험 조건

- 리워드 함수에 대한 접근 외에는 다른 정보는 없다고 가정
- 다른 물체나, 핸드를 사용하는 등 다른 환경에서
- 기술을 쉽게 다시 학습 가능







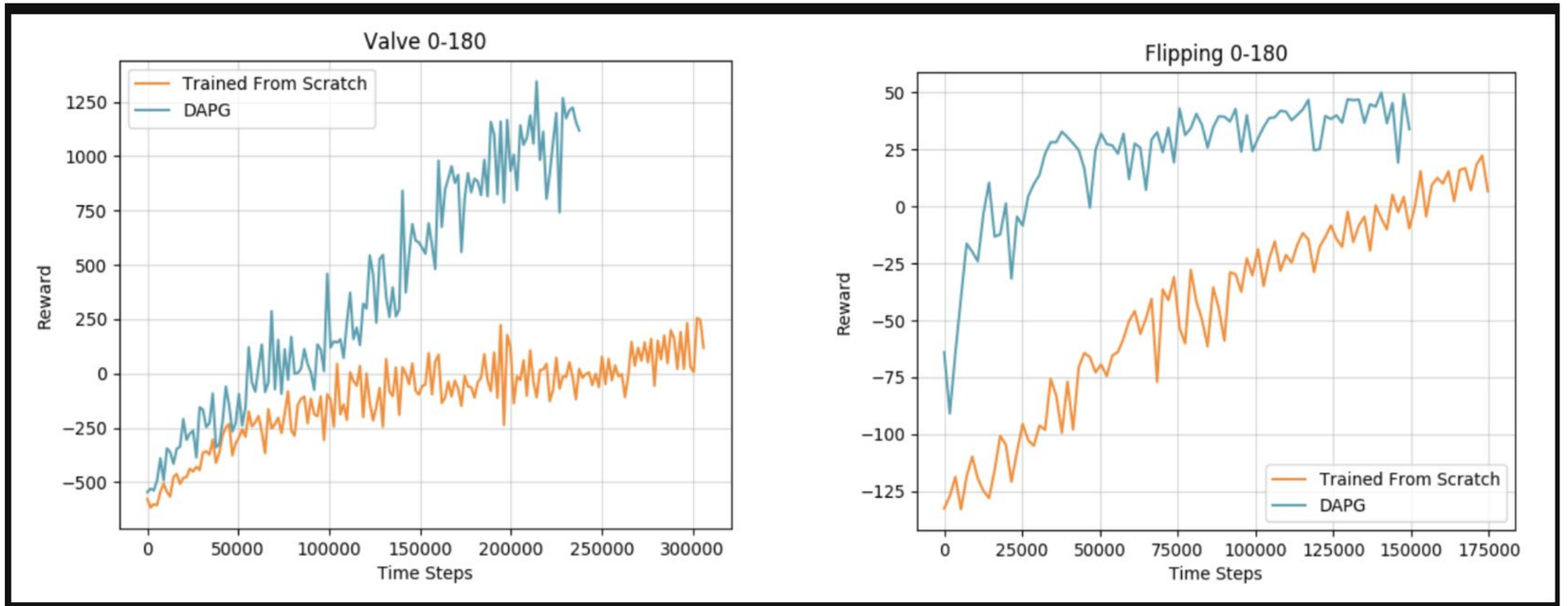
추가 학습 결과

180도 회전을 포함한 수평 방향에도 적용되는 것 확인

인간의 Demo를 활용하여 학습 시키기

1.행동 복제를 통해 좋은 초기 Policy를 제공해주는 것

2.학습 과정 전반에 걸쳐 궤적 추적 보조 보상을 사용하여 탐색을 안내하는 보조 학습 신호를 제공

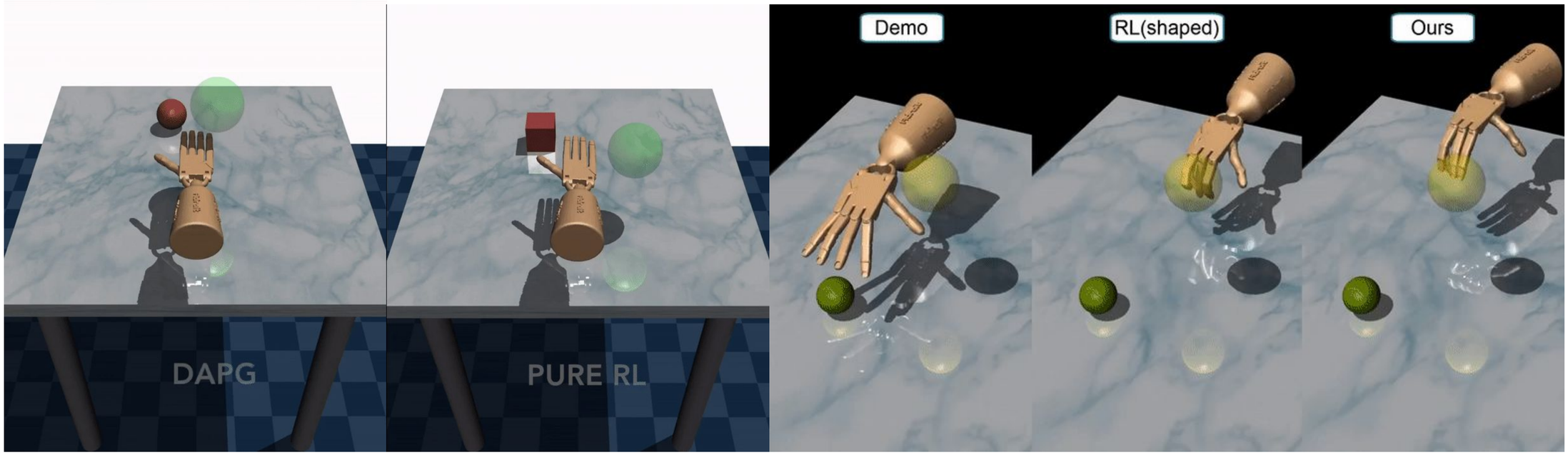


Learning Curves of RL from scratch on hardware vs DAPG.

인간의 Demo를 활용하여 학습 시키기

1.행동 복제를 통해 좋은 초기 Policy를 제공해주는 것

2.학습 과정 전반에 걸쳐 궤적 추적 보조 보상을 사용하여 탐색을 안내하는 보조 학습 신호를 제공



DAPG의 구성 요소

1. Demonstration Learning(시범학습) :

- 초기 정책을 설정하는 데 사용
- expert data (즉, 전문가의 행동을 기록한 데이터)를 이용하여 초기 정책을 학습
- 초기 탐색 단계를 빠르게 진행

2. Policy Gradient RL:

- 시범 학습을 통해 초기화된 정책을 기반으로, 강화 학습을 통해 정책을 Fine-tuning하는 방식
- 실제 과제 목표를 최적화

3. 시범을 통한 학습 신호 보강:

- 학습 과정 중 시범 데이터를 활용하여 보조 학습 신호를 제공
- 정책이 시범 데이터와 크게 벗어나지 않도록 유도

DAPG와 NPG 비교

1.초기화 단계:

1. **DAPG**
   - Behavior Cloning을 사용하여 전문가 시범을 모방
   - 초기 탐색 과정을 단축
2. **NPG**:
   - 강화 학습만으로 학습
   - 초기 탐색 단계에서 많은 샘플이 필요, 학습 속도 감소

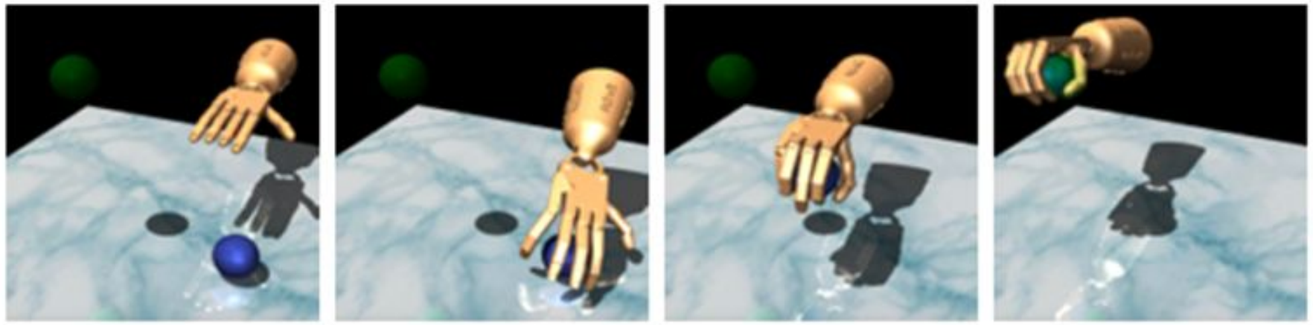
2.샘플 복잡도:

1. **DAPG**
   - 모방으로 샘플 복잡도를 크게 줄임
   - DAPG가 NPG보다 최대 30배 더 빠르게 학습 가능
2. **NPG**
   - 강화 학습만으로 학습하기 때문에 더 많은 샘플이 필요
   - 이로 인해 학습 시간 증가

4.성능 및 안정성:

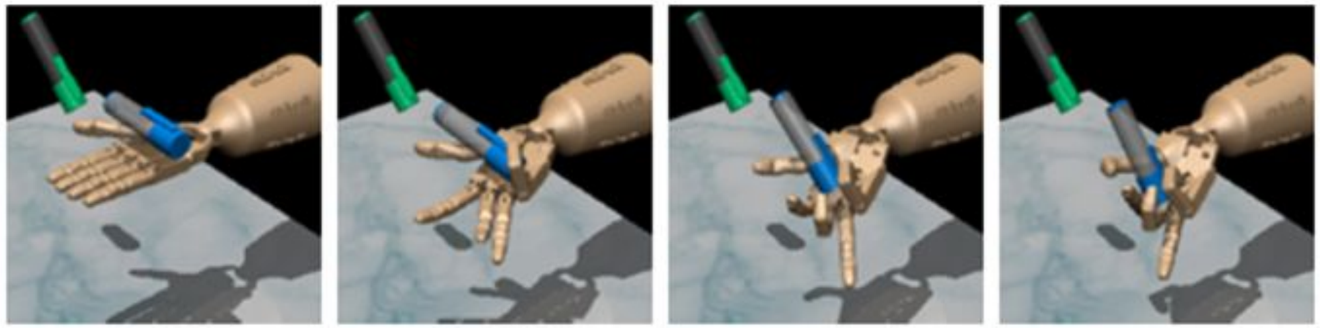
1. **DAPG**
   - Behavior Cloning 을 통해 초기 정책을 설정한 후, 강화 학습을 통해 개선
   - 이 과정에서 정책의 안정성과 성능 증가
2. **NPG**
   - 샘플 복잡도 문제와 초기 탐색 단계에서의 어려움
   - DAPG에 비해 성능, 안정성 감소

실험 환경 소개



실험 환경 1.Object Relocation(객체이동)

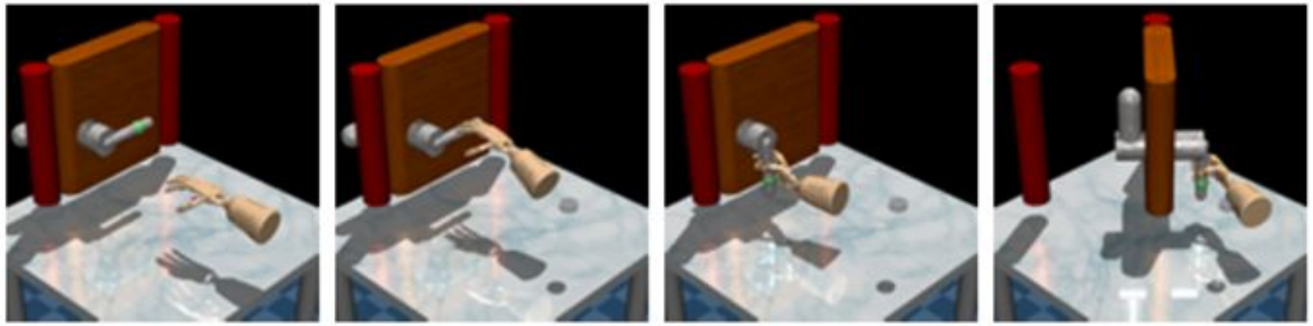
- 파란 공을 초록 목표 지점으로 이동시키는 작업
- 공과 목표의 위치는 작업 공간 전체에서 무작위로 배치
- 객체가 목표의 epsilon-ball 안에 있으면 작업이 성공한 것으로 간주



실험 환경 2.In-Hand Manipulation (손 안의 조작)

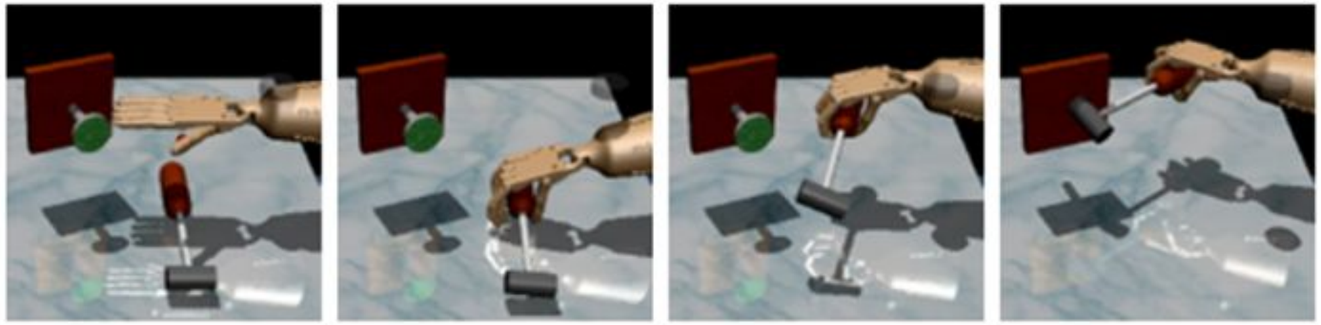
- 파란 펜을 초록 목표의 방향과 일치하도록 재배치하는 작업
- 손의 베이스는 고정되어 있으며, 목표는 모든 구성으로 무작위화
- 방향이 허용 오차 내에서 일치하면 작업이 성공한 것으로 간주

실험 환경 소개



실험 환경 3: Door Opening(문 열기)

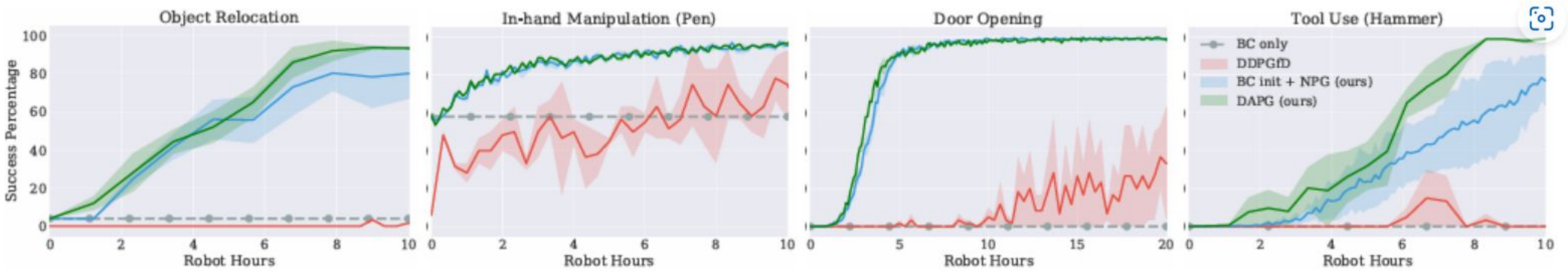
- 걸쇠를 풀고 문을 여는 작업
- Agent는 환경 상호작용을 활용하여 걸쇠에 대한 이해를 발전



실험 환경 4: Tool Use (도구 사용)

- 망치를 집어 들어 힘껏 사용하여 못을 판자에 박는 작업
- 못의 위치는 무작위로 배치, 최대 15N의 힘을 흡수할 수 있는 건조 마찰
- 못의 전체 길이가 판자 안에 들어가면 작업이 성공한 것으로 간주

실험 결과:



1.Object Relocation (물체 재배치):
- DAPG (녹색)와 NPG (파란색)을 비교했을 때, DAPG가 더 빠르게 높은 성공률에 도달
- 이는 Demonstration Learning을 통해 초기 정책을 설정한 덕분

2.In-Hand Manipulation (손 안에서 물체 조작):
- 다양한 위치와 방향에서 물체를 조작하는 과제에서도 DAPG가 가장 높은 성공률을 유지하며 안정적으로 높은 성능을 보임.
- NPG 역시 높은 성공률을 보이지만 DAPG에 비해 약간 뒤처짐

3.Manipulating Environmental Props (환경 소품 조작):
- 문 열기와 같은 복잡한 환경 조작 작업에서도 DAPG는 매우 빠르게 높은 성공률에 도달하여 유지
- NPG는 조금 느리지만 결국 높은 성공률에 도달

4.Tool Use (도구 사용):
- 망치를 사용하여 못을 박는 작업에서 DAPG는 높은 성공률에 도달하며 NPG보다 더 빠르게 학습
- NPG는 중간 정도의 성공률을 보임

실증적AI프로젝트 금주 활동내역

주제: RLHF를 이용한 협동 로봇 제어 프로그램 개발

금주 활동계획	1. Ardoit hammer 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 2. 오프라인 강화학습 알고리즘 expert 생성	1. 모방학습 알고리즘 정리 2. 모방학습 알고리즘 expert 생성
차주 활동계획	1. 실험 결과 정리 2. D4RL: Datasets for Deep Data-Driven Reinforcement Learning 리뷰		

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)