JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 10, pp. 2987-2994, Oct. 2024

Check for updates

# 저편향·고분산된 보편적 데이터를 이용한 효율적인 오프라인 강화학습 방법

권 은 주[1] · 김 현 석[2*]
[1]동아대학교 AI학과 학사과정
[2]동아대학교 컴퓨터·AI공학부 교수

# Effective Offline Reinforcement Learning with Low-Bias, High-Variance Practical Data

**Eunju Kwon[1] · Hyunseok Kim[2*]**

[1]Bachelor Course, Department of AI, Dong-A University, Busan 49315, Korea
[2]Assistant Professor, Division of Computer and AI, Dong-A University, Busan 49315, Korea

**[요 약]**

대규모 언어 모델(Large Language Model) 기반 Robotic Transformer 활용으로 로봇은 복잡한 시퀀스 문제도 스스로 해결할 수 있게 되었으나, 여전히 단위 행동은 사전에 획득된 고품질 전문가 데이터를 이용해 학습되고 있다. 전문가 데이터는 특정 행동을 완벽하게 수행하는 예시로 수집하는 데 드는 시간과 비용이 많이 발생하며, 모든 상황을 고려할 수 없어 편향된 특성이 있다. 따라서, 본 논문에서는 저편향·고분산된 보편적인 데이터를 이용하여 복잡한 문제 환경에서 강화학습이 효율적으로 학습하는 방법을 제안한다. 또한, RLIF(Reinforcement Learning via Intervention Feedback) 알고리즘을 사용한 학습 평가와 환경의 노이즈 차이에 따른 도메인 랜덤화 실험을 진행한다. 본 논문에서는 저편향·고분산된 보편적 데이터를 사용하여 기존의 전문가 데이터만으로 학습했을 때보다 보상이 증가하고 새로운 환경에서도 높은 점수를 유지하는 모델을 만들 수 있다는 결과를 제시한다. 본 연구는 복잡한 환경변화에도 효율적인 강화학습을 실현하는 데 기여할 것으로 기대된다.

**[Abstract]**

The Robotic Transformer, based on the Large Language Model (LLM), allows robots to independently solve complex sequence problems. However, the primitive actions are still derived from high-quality expert data obtained in advance. Expert data consists of flawless execution of specific actions, but it is time-consuming and expensive to collect, and it is biased because it does not account for all situations. In this paper, we propose a method based on reinforcement learning to efficiently learn in complex problem environments using low-bias and high-variance practical data. Additionally, we evaluate the learning using the Reinforcement Learning via Intervention Feedback (RLIF) algorithm and conduct a domain randomization experiment to assess the impact of environmental noise. We demonstrate that using low-bias, high-variance practical data can help create a model that boosts rewards and maintains high scores in new environments compared to training with only expert data. This study contributes to efficient reinforcement learning in complex and changing environments.

**색인어 :** 모방학습, 강화학습, 오프라인 강화학습, 전문가 데이터, 범용 인공지능
**Keyword :** Imitation Learning, Reinforcement Learning (RL), Offline RL, Expert Demo, Artificial General Intelligence

**http://dx.doi.org/10.9728/dcs.2024.25.10.2987**



**Received** 08 August 2024; **Revised** 06 September 2024
**Accepted** 26 September 2024

***Corresponding Author; Hyunseok Kim**

**Tel:** +82-51-200-7928
**E-mail:** hertzkim@dau.ac.kr

## Ⅰ. 서 론

최근, 거대 언어 모델(Large language Model) 기반 Robotic Transformer를 활용한 로봇 제어 연구가 활발히 진행되고 있다[1],[2]. Robotic Transformer는 로봇이 수행할 작업에 대해 순차적인 행동을 생성하지만, 세부적인 단위 동작은 여전히 강화학습(Reinforcement Learning; RL)을 활용해 구현하고 있다. 강화학습은 지도학습에서 정답 데이터가 필요한 것과 달리 환경과 상호작용을 통해 학습하는 방법으로, Atari[3], AlphaGo[4]와 같이 상태 변화를 예측할 수 있는 환경에서 우수한 성능을 보여주고 있다. 하지만, 스마트 팩토리[5], 자율 주행 시스템[6], 무인 비행 시스템[7] 등 실세계 환경은 훨씬 더 복잡하고 불확실성이 많아 강화학습을 적용하는데 한계가 있다. 이를 해결하기 위해 미리 획득된 데이터 샘플로부터 학습을 진행하고 다시 실제 환경과 상호작용하는 오프라인 강화학습[8] 방식이 활용되고 있다. 하지만, 지도학습에서 겪는 불균형한 데이터 분포 문제와 비슷하게, 오프라인 강화학습에서도 고편향·저분산된 데이터 분포 문제로 인해 학습 효율이 떨어지는 문제가 있다[9].

전문가 데모 행동 데이터를 활용한 모방학습(Imitation Learning)은 행동을 복제하는 Behavior Cloning, 역으로 보상함수를 추정하는 Inverse Reinforcement Learning 등이 있다. 최적의 전문가 데이터는 분포가 특정 영역에 몰려 있어 다양한 상황을 반영하지 못하며, 수집과 생성 과정에서 시간과 비용이 많이 든다는 단점을 가지고 있다[10]. 이를 해결하기 위해 오프라인 강화학습에 모방학습을 적용한 DAgger (Data Aggregation)[11]는 정답에 해당하는 행동을 알려주는 방식으로 학습이 진행된다. 그러나, 모든 상황에서 정확한 행동을 알려주는 것은 불가능하므로, 필요한 경우에만 전문가 데이터를 개입시키는 RLIF(Reinforcement Learning via Intervention Feedback)[12]가 등장했다. RLIF는 전문가 데이터를 활용하여 전문가 에이전트를 학습시키고, 다시 전문가 에이전트가 새로운 데모 데이터를 생성하는 방식이다. 하지만, 동일한 데이터를 반복하여 사용하므로 학습 과정에서 편향이 발생할 수 있으며, 이로 인해 학습의 효율이 떨어지는 문제가 있다.

본 논문에서는 RLIF에서 사용하던 최적의 전문가 데이터 대신 저편향·고분산된 보편적인 데이터를 사용하여 효율적으로 학습하는 방법을 제안한다. 또한, 도메인 분포 변화(Domain Distribution Shift)에도 대응할 수 있는지 검증하기 위해 도메인 랜덤화(Domain Randomization) 실험을 추가로 진행한다. 본 논문은 제안된 모델의 성능을 평가하기 위해 에이전트가 학습 과정에서 수행한 행동에 따른 보상 분포를 분석한다. 또한, 본 논문은 실험을 통해 고편향·저분산된 기존의 데이터를 사용했을 때 보상이 적은 경우와 큰 경우 두 극단으로 나뉘어 있지만, 저편향·고분산된 데이터를 사용할 때는 보상이 넓은 범위에 고르게 분포되는 것을 보인다. 학습 평가 그래프에서도 저편향·고분산 데이터를 사용할 때 보상이 지속해서 증가하였으며, 도메인 랜덤화 실험에서도 높은 점수를 유지하는 것을 확인하였다.

본 논문은 이러한 분석을 통해 에이전트와 데이터의 특성에 따른 적절한 데이터 선택 방법을 제안하며, 저편향·고분산된 보편적인 데이터로도 높은 성능을 유지하며 다양한 상황에서 더 안정적이고 효율적인 학습을 제공하고, 동일한 데이터를 계속해서 사용했을 때 발생하는 편향을 줄일 수 있음을 보여준다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 모방학습과 RLIF 차이를 분석하고, 모방학습에서 전문가 중심 데이터 방식의 한계점을 설명한다. 3장과 4장에서는 저편향·고분산된 보편적인 데이터를 활용한 강화학습 방법을 제안하며, 5장에서는 Gymnasium의 Hopper 환경에서 비교 실험 결과를 제시한다. 그리고, 6장에서는 결론 및 향후 연구에 관해 기술한다.

## Ⅱ. 관련 연구

### 2-1 모방학습(Imitation Learning)

강화학습은 데이터 기반의 학습에서 벗어나 실시간으로 환경과 상호작용하며 지도학습으로는 해결할 수 없는 명시적인 정답이 존재하지 않는 문제를 해결할 수 있다. 그러나, 무작위 시행착오 학습 방식과 복잡한 환경에서 적절한 보상함수 설계의 어려움으로 인해, 강화학습은 실세계에 적용하기 어렵다는 한계가 있다. 이러한 문제를 해결하기 위해 전문가 행동을 모방하여 학습을 진행하는 모방학습 방법이 제안되었다[13]. 모방학습의 학습 주체인 에이전트는 숙련된 전문가의 행동을 관찰하고 모방함으로써 더욱 효율적으로 학습한다.

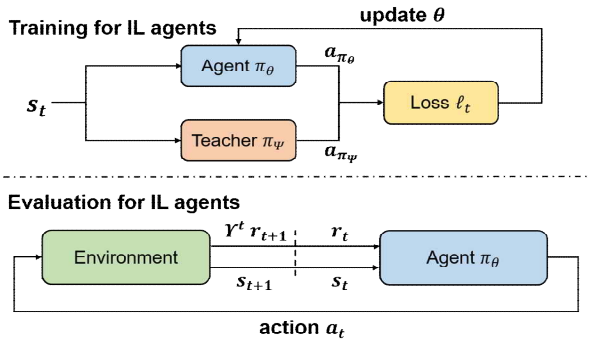


**그림 1.** 일반적인 모방학습의 작동 방식
**Fig. 1.** General imitation learning workflow

그림 1은 일반적인 모방학습의 작동 방식을 보여준다. 모방학습의 훈련 단계에서 에이전트 $\pi_\theta$와 교사 $\pi_\psi$는 동일한 상태 $s_t$에서 각 행동 $a_{\pi_\theta}$와 $a_{\pi_\psi}$을 생성하고, 두 행동 사이의 손실 $\ell_t$을 계산하여 에이전트의 정책 $\theta$를 업데이트한다. 평가

단계에서 환경이 에이전트 $\pi_\theta$에게 상태 $s_t$를 제공하면 에이전트는 행동 $a_t$를 선택한다. 이후, 환경은 새로운 상태 $s_{t+1}$와 보상 $r_{t+1}$을 제공하는 방식으로 학습이 진행된다. 그러나, 모방학습 역시 인간이 시연할 수 있는 상황에서만 사용할 수 있어, 인간이 시연할 수 없는 범위의 문제나 인간을 뛰어넘는 동작을 학습하는 데는 한계가 있다.

### 2-2 모방학습 방법

#### 1) Behavior Cloning

모방학습의 초기 방법인 행동 복제(Behavior Cloning)는 지도학습을 통해 전문가의 행동 데이터를 학습한 후 주어진 상태에서 가장 가능성 있는 행동을 예측하는 방식으로 학습한다. 자율 주행 자동차에서 운전자의 조작을 모방하여 자율 주행 기능을 구현하는 연구[14]와 같이, 강화학습의 보상함수 설계 어려움에 대한 한계를 극복할 수 있다. 그러나 인간의 시연이 잘못되었을 때 제대로 학습할 수 없으며 새로운 상황에 대처하기 어렵다는 문제가 있다.

#### 2) Interactive Imitation Learning

상호작용 모방학습(Interactive Imitation Learning)은 행동 복제의 한계를 해결하기 위해 등장했다. 에이전트는 학습 과정에서 전문가와 실시간으로 상호작용하며 피드백을 받아 정책을 개선한다. 잘못된 행동을 할 때만 피드백을 받기 때문에 일반화 성능이 높아지며 새로운 상황에 더 잘 적응할 수 있다. DAgger와 RLHF(Reinforcement Learning with Human Feedback)[15] 등, 다양한 방식의 실시간 피드백 기반 모방학습은 에이전트의 성능을 향상하며, 행동 복제로는 해결할 수 없었던 새로운 상황에 대한 적응 문제를 해결할 수 있다. 그러나, 학습 과정에서 전문가의 지속적인 참여가 필요하며 여전히 인간의 한계를 뛰어넘는 학습이 불가능하다는 문제점이 있다.

#### 3) Inverse Reinforcement Learning

역강화학습(Inverse Reinforcement Learning)은 에이전트의 정책 또는 관찰된 행동을 바탕으로 에이전트가 따르는 보상함수를 추정한다. 이 방법을 통해 작업 수행 시 기록된 데이터를 사용하여 작업 수행에 개입하지 않고도 다른 에이전트를 모델링할 수 있는 자율에이전트를 구축할 수 있다[16]. 대표적인 방법으로는 GAN 구조를 활용하여 보상함수를 추정하는 AIRL(Adversarial Inverse Reinforcement Learning)[17]이 있으며 보상 설계가 복잡한 문제에서 특히 유용하다. 전문가의 시연 데이터를 사용하여 학습하기 때문에 인간의 지속적인 개입이 필요하지 않으며, 행동을 그대로 따라 하지 않고 데이터를 바탕으로 보상함수를 추정하기 때문에 일반화에 더 강하다. 하지만, 여전히 품질이 좋은 전문가 데이터를 필요로 하며 예측 불가능한 요소가 있는 경우 보상함수를 추정하기 어렵다.

#### 4) Generative Adversarial Imitation Learning

적대적 생성 모방학습(Generative Adversarial Imitation Learning)은 생성적 적대 신경망(GAN, Generative Adversarial Network)의 개념을 모방학습에 적용한 방법이다. 여기서 생성자(Generator)는 전문가의 행동을 모방하고, 판별자(Discriminator)는 전문가의 행동과 생성자의 행동을 구별하려 한다. 이 과정을 통해 생성자는 점점 더 전문가의 행동을 잘 모방하며 판별자를 혼동시키도록 훈련된다. 판별자가 생성된 데이터와 실제 데이터를 구별하지 못하면 생성 모델이 성공적으로 모방한 것이 된다[18]. 이 방법은 생성자와 판별자의 대립을 통해 더 높은 일반화 능력을 갖추게 되어, 역강화학습에 비해 명시적인 보상함수 설계 없이 암묵적으로 보상 구조를 학습한다는 장점이 있다. 하지만, 여전히 고품질의 전문가 데이터에 의존한다는 한계를 가지고 있다.

본 논문에서는 모방학습의 한계를 극복하기 위해 설계된 RLIF 알고리즘을 사용한다. RLIF는 가상의 전문가 에이전트를 학습시키고 이를 사용해 학습하는 방식으로, 고품질 전문가 데이터가 없어도 효과적으로 학습하는 방법이다. 또한, 본 논문에서는 저편향·고분산된 보편적인 데이터를 사용하여 효율적으로 학습하는 방법을 제안하며 실험을 통해 성능을 입증하고자 한다.

## Ⅲ. 연구 방법

### 3-1 데모 데이터로 학습한 전문가 에이전트

본 논문에서는 RLIF[12]에서 제공하는 표 1의 전문가 에이전트를 사용하여 학습을 진행한다. 최적의 전문가 데이터에서 50개의 궤적을 샘플링하여 각 알고리즘으로 전문가 에이전트를 만든다. 이때, Expert Level이 90%란 뜻은 특정 작업의 최고 전문가 점수 대비 90%를 달성할 수 있음을 의미한다. 첫 번째 전문가는 RLPD(Reinforcement Learning with Prior Data)[19]를 통해 학습한 것으로, Expert Level이 110%이고, 두 번째 전문가 에이전트는 BC를 통해 학습한 것으로 Expert Level이 40%이다.

**표 1.** 전문가 에이전트
**Table 1.** Expert agents

| Expert Agent | Training Algorithm | Expert Level |
|---|---|---|
| 110EA | RLPD | 110% |
| 40EA | BC | 40% |

**표 2.** 오프라인 데이터 세트
**Table 2.** Offline datasets

| Dataset | Num | Min | Mean | Max |
|---|---|---|---|---|
| hopper-expert-2 | 1027 | 1646 | 3511 | 3759 |
| hopper-medium-2 | 2186 | 315 | 1422 | 3222 |

### 3-2 Expert, Medium 데이터의 특성과 차이

본 논문에서는 D4RL[20]에서 제공하는 6가지 형태의 데이터 중, 표 2에 명시된 Expert 데이터와 저편향·고분산된 데이터를 대변할 수 있는 Medium 데이터를 사용한다. 데이터는 observation, action, reward, terminal, timeout, info로 구성된다. Expert 데이터는 최적의 전문가 데이터를 기반으로 하며, Medium 데이터는 정책을 온라인으로 훈련하고 훈련을 조기 중단한 후, 부분적으로 훈련된 정책에서 샘플을 수집하여 생성된 비최적 데이터이다.

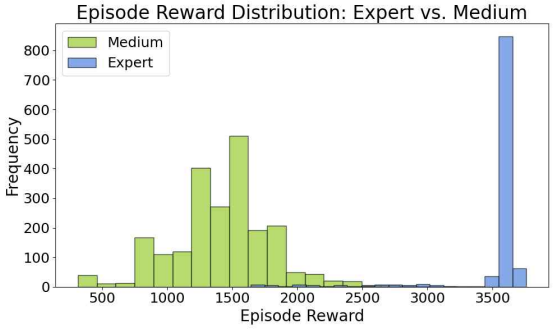


**그림 2.** 데이터 세트 에피소드 보상 분포
**Fig. 2.** Distribution of datasets episode reward

**• 에피소드 보상 분포에 따른 데이터 특성 비교**

그림 2는 두 데이터의 에피소드 보상 분포를 시각화한 것이다. Medium 데이터는 보상이 다양한 값에 걸쳐 분포된 저편향·고분산 형태의 데이터로, 다양한 상황에 대한 예시를 포함하고 있어 여러 시나리오에 적응하고 더 넓은 범위의 문제를 해결하는 데 유리하다. 반면, Expert 데이터는 특정 동작에 최적화되어 있어 높은 보상을 일관되게 달성하지만, 다양성 부족으로 여러 상황에 적응하기 어렵다는 특징을 가지고 있다.

### 3-3 RLIF 알고리즘

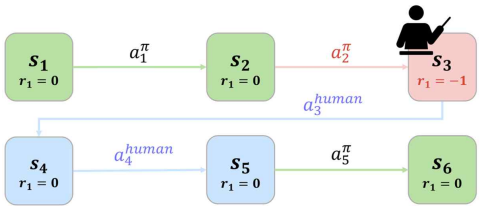


**그림 3.** RLIF의 작동 과정
**Fig. 3.** Working process of the RLIF

기존의 대화형 모방학습 DAgger에서는 인간 전문가가 에이전트의 행동을 관찰하고, 바람직하지 않은 행동을 직접 교정한다. 이는 최적에 가까운 개입이 요구되며, 전문가의 능력을 넘어서지 못하는 한계를 가진다. 반면, RLIF 알고리즘은 그림 3과 같이 개입 결정을 강화학습의 보상 신호로 사용한다. 전문가가 개입하지 않으면 보상은 0으로 유지되고, 전문가가 개입하면 −1로 변경된다. 이를 통해 에이전트가 잘못된 행동을 스스로 판단하고 더 나은 결정을 내릴 수 있도록 유도한다. 따라서 직접적인 행동 지시 없이도 학습이 가능하므로 최적의 전문가가 아니더라도 사용 가능하다.

**1) RLIF 알고리즘의 개입 전략**

RLIF는 수식 1의 가치 기반 전략을 사용하여 전문가의 개입 시점을 결정한다. 학습자의 성능을 평가하고 필요한 경우 전문가의 개입을 통해 성능을 높인다. 상태 $s$에서 에이전트의 행동을 기반으로 state-action의 누적 보상을 나타내는 $Q$값을 계산한다. 전문가 정책 $\pi^{\exp}(s)$으로 계산된 $Q$값 $Q^{\pi^{ref}}(s,\pi^{\exp}(s))$과 에이전트 정책 $\pi(s)$으로부터 계산된 $Q$값 $Q^{\pi^{ref}}(s,\pi(s))$을 비교한다. 이때 전문가의 $Q$값이 에이전트의 $Q$값보다 크다면, 전문가의 개입 확률은 $\beta$이다. 본 논문에서는 $\beta$값을 0.95로 설정하여 진행했다.

$$P(Intervention|s) = \begin{cases} \beta, & \text{if } Q^{\pi_{ref}}(s,\pi_{\exp}(s)) > Q^{\pi_{ref}}(s,\pi(s)) + \delta \\ 1-\beta & others \end{cases} \quad (1)$$

**2) RLIF 알고리즘의 학습 정보 저장 방식**

학습 과정에서 오프라인 데이터를 활용하는 비율을 나타내는 'offline_ratio' 매개변수를 설정할 수 있다. 본 논문에서는 'offline_ratio'를 0.5로 설정하여, 절반은 오프라인 데모 데이터를 선택하고, 절반은 개입을 통해 에이전트가 학습한 정보를 선택하여 Replay Buffer에 저장한다. 에이전트 스스로 학습한 정보와 오프라인 데이터의 비율을 동일하게 설정함으로써, 두 가지 데이터가 학습 과정에서 미치는 영향을 균등하게 평가할 수 있도록 한다.

## Ⅳ. 실험 방법

### 4-1 실험 환경

본 논문은 그림 4에 제시된 Gymnasium[21]의 Hopper에서 실험을 진행했다. Hopper는 다양한 관절과 신체 부위 구성을 가진 로봇 시뮬레이션이다. 환경의 목표는 네 개의 신체 부위를 연결하는 세 개의 경첩에 토크(torque)를 가하여 앞(오른쪽) 방향으로 움직이는 홉을 만드는 것이다.

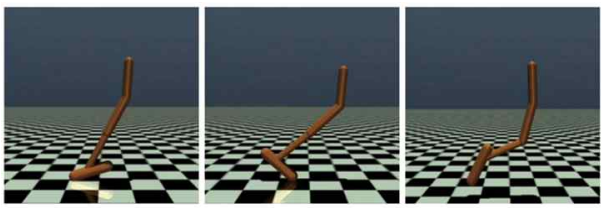

**그림 4.** 실험 환경
**Fig. 4.** Experimental environment

• **실험 환경 MDP 정의**

강화학습은 마르코프 결정 프로세스(Markov Decision Process, MDP)를 통해 순차적으로 행동을 결정한다. MDP는$(S, A, P, R, \gamma)$ 다섯 가지 요소로 구성되며, Hopper의 MDP는 표 3과 같다. 상태 집합 $S$는 로봇의 다양한 각도 및 속도 등 11개의 관찰값으로 구성된다. 행동 집합 $A$는 로봇의 세 개의 관절에 적용되는 토크를 나타낸다. 보상함수 $R$은 에이전트의 state-action에 따른 보상을 나타낸다.

**표 3.** 환경의 MDP 구성 요소
**Table 3.** MDP components of environment.

| Element | Definition |
|---|---|
| $S$ | Set of all possible states<br>{height of the hopper, angle of the (top, thigh joint, leg joint, foot joint), velocity of the (torso x-coord, torso z-coord)of the top,<br>angular velocity of the (top, thigh hinge, leg hinge, foot hinge)} |
| $A$ | Set of all possible actions<br>{Torque applied on the (thigh rotor, leg rotor, foot rotor)} |
| $P$ | probability from one state to the next given an action |
| $R$ | Value of action<br>{healty_reward + forward_reward − ctrl_cost} |
| $\gamma$ | Reward value paid |

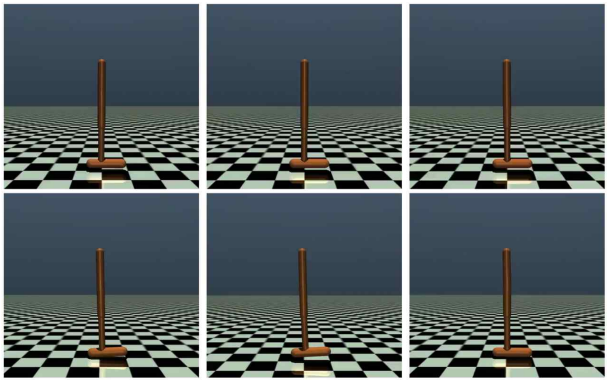

**그림 5.** 노이즈 변화에 따른 에이전트 초기 자세 차이
**Fig. 5.** Agent initial posture due to noise variation

### 4-2 모델 학습

본 논문의 학습 모델은 표 4와 같다. 예를 들어 110EA-E 모델의 경우, 학습 데이터의 절반은 표 2의 오프라인 데이터 'hopper-expert-2'를 나머지 절반은 110EA를 사용하여 RLIF 알고리즘을 통해 실시간으로 수집된 데이터를 사용한다. 각 모델은 1,000,000 스텝 동안 학습하며, 에피소드 종료 후 Replay Buffer에 데이터를 저장한다. 학습 진행 중에는 10,000 스텝마다 성능을 평가한다. 학습이 완료된 후, 모델의 도메인 랜덤화 실험을 통해 일반화에 적합한 모델 여부를 평가한다.

**표 4.** 학습되는 모델 종류
**Table 4.** Types of models being trained

| Model | Expert Agent | Dataset |
|---|---|---|
| 110EA-E | 110EA | hopper-expert-2 |
| 110EA-M | 110EA | hopper-medium-2 |
| 40EA-E | 40EA | hopper-expert-2 |
| 40EA-M | 40EA | hopper-medium-2 |

### 4-3 도메인 랜덤화 실험

Hopper는 'reset_noise_scale' 파라미터를 사용하여 에피소드를 초기화할 때 초기 위치와 속도에 무작위 정도를 조절할 수 있으며, 본 실험에서는 기본값 0.005로 설정하였다. 노이즈는 음의 'reset_noise_scale'에서 양의 'reset_noise_scale' 범위 내에서 균일하게 분포된 무작위 값으로 추가하였다. 본 논문에서는 표 4의 모델에 노이즈를 기본값과 5배로 설정한 사잇값으로 설정하여 실험을 진행했다. 그림 5는 무작위 정도에 따른 초기 자세를 보여주며, 무작위 정도가 높아질수록 초기 시작 범위가 커진다.

## V. 실험 결과 및 결과 분석

### 5-1 실험 결과

#### 1) 에이전트의 행동 보상에 대한 그래프 분석

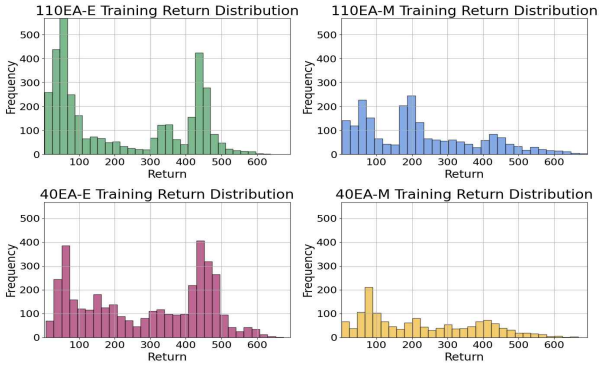


**그림 6.** 에이전트의 선택된 행동에 따른 보상 분포
**Fig. 6.** Distribution of returns according to agent's actions

그림 6은 오프라인 데이터를 제외한 에이전트가 받은 Return 값의 빈도를 나타내는 그래프로 에이전트가 학습하는 과정에서 선택된 행동에 따라 얻는 보상의 분포를 보여준다. x축은 학습 중 에이전트가 받은 Return 값을 나타내며, y축은 학습 과정에서 받은 보상의 빈도를 의미한다. Medium 데이터를 사용한 경우, 에이전트는 전체의 1/3 수준의 보상을 주로 받는다. 반면, Expert 데이터를 사용하였을 때는 중간 보상을 건너뛰고 높은 보상을 받는 것을 확인할 수 있다.

Medium 데이터는 다양한 분포를 포함하고 있어 학습 초반에 탐험을 중점으로 다양한 경험을 쌓을 수 있어서 최고점(많이 나온 보상 값)의 위치가 작지만, 분산은 크다. 반면, Expert 데이터의 경우 최적의 동작만을 포함하고 있어 학습 과정에서 탐험 과정 없이 바로 높은 보상을 얻는 것을 확인할 수 있다. Medium 데이터는 다양한 분포를 포함하고 있어 변화가 있는 활동에 대비해 학습을 가능하게 하며, Expert 데이터는 주어진 문제에 대한 최적의 동작만을 포함하고 있어 더 빠르고 높은 성능을 목표로 하는 경우 활용이 가능하다. 똑똑하지 않은 에이전트의 경우, 다양한 경험을 쌓으면서 점진적으로 학습하여야 하므로 Medium 데이터를 사용하는 것이 더 유용하다. 반면, 이미 충분한 지식을 가진 에이전트의 경우 Expert 데이터를 사용하여 빠르게 목표에 도달할 수 있다. 결론적으로 데이터의 특성에 따라 학습 성능과 보상 분포가 크게 달라짐을 알 수 있으며, 학습하고자 하는 과제와 에이전트의 수준에 따라 적절한 데이터를 사용하여야 한다.

### 2) 에이전트의 학습 평가 그래프 분석

그림 7은 Hopper 환경에서 학습 과정 중 모델의 정책을 10,000 스텝마다 평가한 그래프로 x축은 학습 중 진행된 스텝 수를 나타내며, y축은 학습 과정에서 받은 보상의 빈도를 의미한다. 시간의 흐름에 따라 모델이 학습하면서 받는 보상의 전체적인 변화를 관찰할 수 있다. 각 모델의 학습 성능은 다음과 같이 해석할 수 있다. Medium 데이터를 사용한 모델(40EA-M, 110EA-M)은 학습 초기에 보상이 빠르게 상승하여 효율적으로 학습하는 모습을 보인다.

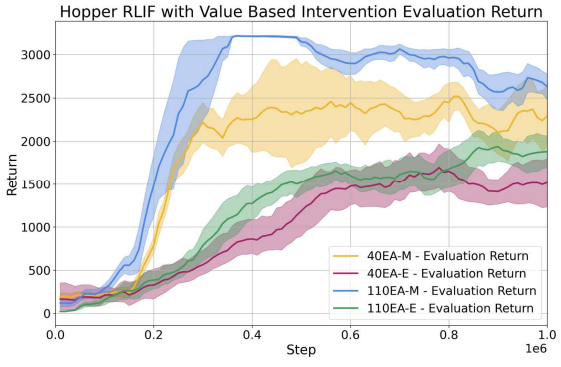


**그림 7.** 에이전트의 학습 성능 평가
**Fig. 7.** Evaluation of agent's performance

Medium 데이터를 사용한 110EA-M 모델은 높은 보상을 지속해서 유지한다. 표 4의 노이즈 랜덤화 실험에서도 Medium 데이터를 사용하면 노이즈와 관계없이 높은 성능이 유지되는 것을 확인할 수 있다. 따라서, 전문가 수준의 모델은 복잡한 데이터와 결합함으로써 에이전트가 다양한 상황에 대응할 수 있는 능력을 키울 수 있다는 것을 확인했다. 또한, 저편향·고분산된 보편적인 데이터가 모델이 다양한 상황에서 편향 없이 학습하고, 더 폭넓은 경험을 통해 더욱 일반화된 성능을 발휘하도록 돕는다는 것을 알 수 있다.

**표 4.** 실험 결과: 도메인 랜덤화에 따른 리턴값
**Table 4.** Experimental results: Return based on domain randomization

| Noise | Model | Return |
|---|---|---|
| 0.005 (default) | 110EA-E | 1875 |
| | **110EA-M** | **2550** |
| | 40EA-E | 1810 |
| | **40EA-M** | **2542** |
| 0.025(x5) | 110EA-E | 1707 |
| | **110EA-M** | **2568** |
| | 40EA-E | 1844 |
| | **40EA-M** | **2093** |

## Ⅵ. 결 론

본 논문에서는 기존의 전문가 데이터 중심 학습 방법 대신 저편향·고분산 데이터로 학습하는 방법을 제안하며, 모델의 일반화 성능을 개선하였다. 실험 결과, Expert 데이터 중심의 기존 연구에서는 에이전트의 수준에 상관없이 동일한 학습 성능을 보여주었지만, Medium 데이터를 사용하였을 때는 전문가 에이전트의 수준에 상관없이 학습 성능을 더 높일 수 있음을 확인하였다. 또한, 도메인 랜덤화 실험에서도 Medium 데이터를 사용한 모델이 노이즈에 덜 민감하게 반응하는 것을 확인할 수 있었다. 본 연구를 통해 전문가 에이전트가 다양한 상황을 반영하고 복잡한 데이터와 결합하는 것이 에이전트의 성능 향상에 도움이 됨을 확인하였다. 본 연구는 기존 Expert 위주의 학습 방식에서 저편향·고분산된 보편적인 데이터를 사용한 새로운 연구 방향을 제시하며, 실세계와 같이 환경 변화가 많은 상황에서도 에이전트가 작업을 잘 수행하도록, 강화학습이 보편적인 데이터를 활용하여 효율적으로 학습하는데 기여할 것이다.

## 감사의 글

"이 논문은 2023년도부터 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임"(No.2023-0-00076, SW 중심대학(동아대학교))

## 참고문헌

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, ... and B. Zitkovich, "Rt-1: Robotics Transformer for Real-World Control at Scale," arXiv:2212.06817, August 2023. https://doi.org/10.48550/arXiv.2212.06817S

[2] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K.

Choromanski, ... and B. Zitkovich, “Rt-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control,” arXiv:2307.15818, July 2023. https://doi.org/10.48550/arXiv.2307.15818

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing Atari with Deep Reinforcement Learning,” arXiv:1312.5602, December 2013. https://doi.org/10.48550/arXiv.1312.5602

[4] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, ... and D. Hassabis, “Mastering the Game of Go with Deep Neural Networks and Tree Search,” *Nature*, Vol. 529, No. 7587, pp. 484-489, January 2016. https://doi.org/10.1038/nature16961

[5] H. B. Choi, J. B. Kim, G. Y. Hwang, K. H. Kim, Y. G. Hong, and Y. H. Han, “Cooperative Multi-Agent Reinforcement Learning-Based Behavior Control of Grid Sortation Systems in Smart Factory,” *KIPS Transactions on Computer and Communication Systems*, Vol. 9, No. 8, pp. 171-180, August 2020. https://doi.org/10.3745/KTCCS.2020.9.8.171

[6] H. Eom, J. Kim, S. Ji, and H. Choi, “Autonomous Parking Simulator for Reinforcement Learning,” *Journal of Digital Contents Society*, Vol. 21, No. 2, pp. 381-386, February 2020. https://doi.org/10.9728/dcs.2020.21.2.381

[7] S. G. Park and D. H. Kim, “Autonomous Flying of Drone Based on PPO Reinforcement Learning Algorithm,” *Journal of Institute of Control, Robotics and Systems*, Vol. 26, No. 11, pp. 955-963, November 2020. https://doi.org/10.5302/J.ICROS.2020.20.0125

[8] A. Kumar, J. Hong, A. Singh, and S. Levine, “When Should We Prefer Offline Reinforcement Learning over Behavioral Cloning?,” arXiv:2204.05618, April 2022. https://doi.org/10.48550/arXiv.2204.05618

[9] D. J. Foster, A. Krishnamurthy, D. Simchi-Levi, and Y. Xu, “Offline Reinforcement Learning: Fundamental Barriers for Value Function Approximation,” arXiv:2111.10919, August 2022. https://doi.org/10.48550/arXiv.2111.10919

[10] N. E. Corrado, Y. Qu, J. U. Balis, A. Labiosa, and J. P. Hanna, “Guided Data Augmentation for Offline Reinforcement Learning and Imitation Learning,” arXiv:2310.18247, March 2024. https://doi.org/10.48550/arXiv.2310.18247

[11] S. Ross, G. J. Gordon, and J. A. Bagnell, “A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning,” in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, Fort Lauderdale: FL, pp. 627-635, April 2011.

[12] J. Luo, P. Dong, Y. Zhai, Y. Ma, and S. Levine, “RLIF: Interactive Imitation Learning as Reinforcement Learning,” arXiv:2311.12996, March 2024. https://doi.org/10.48550/arXiv.2311.12996

[13] N. Gavenski, O. Rodrigues, and M. Luck, “Imitation Learning: A Survey of Learning Methods, Environments and Metrics,” arXiv:2404.19456, April 2024. https://doi.org/10.48550/arXiv.2404.19456

[14] D. A. Pomerleau, “Efficient Training of Artificial Neural Networks for Autonomous Navigation,” *Neural Computation*, Vol. 3, No. 1, pp. 88-97, March 1991. https://doi.org/10.1162/neco.1991.3.1.88

[15] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, “Deep Reinforcement Learning from Human Preferences,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS ’17)*, Long Beach: CA, pp. 4302-4310, December 2017.

[16] S. Arora and P. Doshi, “A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress,” *Artificial Intelligence*, Vol. 297, 103500, August 2021. https://doi.org/10.1016/j.artint.2021.103500

[17] J. Fu, K. Luo, and S. Levine, “Learning Robust Rewards with Adversarial Inverse Reinforcement Learning,” arXiv:1710.11248, August 2018. https://doi.org/10.48550/arXiv.1710.11248

[18] J. Ho and S. Ermon, “Generative Adversarial Imitation Learning,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS ’16)*, Barcelona, Spain, pp. 4572-4580, December 2016.

[19] P. J. Ball, L. Smith, I. Kostrikov, and S. Levine, “Efficient Online Reinforcement Learning with Offline Data,” in *Proceedings of the 40th International Conference on Machine Learning (ICML ’23)*, Honolulu: HI, pp. 1577-1594, July 2023.

[20] J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine, “D4RL: Datasets for Deep Data-Driven Reinforcement Learning,” arXiv:2004.07219, February 2021. https://doi.org/10.48550/arXiv.2004.07219

[21] Gymnasium Documentation. Hopper [Internet]. Available: https://gymnasium.farama.org/environments/mujoco/hopper.

**권은주(Eunju Kwon)**

2022년～현　재: 동아대학교 AI학과 학사과정
※관심분야 : 강화학습, 인공지능, 멀티모달, 생성형 AI

**김현석(Hyunseok Kim)**

2001년 : 동아대학교 전자공학과 (공학학사)
2005년 : 한국과학기술원 대학원 (공학석사)
2014년 : 한국과학기술원 대학원 (공학박사)

2001년～2003년: 삼성전자 연구원
2005년～2009년: LG전자 선임연구원
2011년～2022년: 한국전자통신연구원 책임연구원
2022년～현　재: 동아대학교 컴퓨터·AI공학부 조교수
※관심분야 : 강화학습, 로봇, 인공지능, 군집지능