# 역강화학습 기술 동향

## Research Trends on Inverse Reinforcement Learning

**이상광** (S.K. Lee, sklee@etri.re.kr) 지능형지식콘텐츠연구실 책임연구원
**김대욱** (D.W. Kim, dooroomie@etri.re.kr) 지능형지식콘텐츠연구실 연구원
**장시환** (S.H. Jang, jjangshan@etri.re.kr) 지능형지식콘텐츠연구실 연구원
**양성일** (S.I. Yang, siyang@etri.re.kr) 지능형지식콘텐츠연구실 책임연구원

**ABSTRACT**

Recently, reinforcement learning (RL) has expanded from the research phase of the virtual simulation environment to a wide range of applications, such as autonomous driving, natural language processing, recommendation systems, and disease diagnosis. However, RL is less likely to be used in these complex real-world environments. In contrast, inverse reinforcement learning (IRL) can obtain optimal policies in various situations; furthermore, it can use expert demonstration data to achieve its target task. In particular, IRL is expected to be a key technology for artificial general intelligence research that can successfully perform human intellectual tasks. In this report, we briefly summarize various IRL techniques and research directions.

**KEYWORDS** 역강화학습, 모방학습, 견습학습, 강화학습

## Ⅰ. 서론

강화학습(RL: Reinforcement Learning)에서는 에이전트(agent)가 어떤 상태(state)에서 행동(action)을 수행할 때마다 그 성능에 대한 피드백을 제공하는 보상 함수(Reward Function)가 주어진다. 이 보상 함수는 최적 정책(Optimal Policy)을 구하는 데 이용되며, 이때 예상되는 미래 보상 값이 최대가 된다. 예를 들어 게임 플레이 에이전트 생성을 위해 RL을 이용하는 경우, 플레이를 통해 얻게 되는 점수, 승패 결과, 플레이어 체력 등 에이전트 성능을 평가할 수 있는 보상의 요건들이 주어지며, 이를 통해 최대 기대 보상을 획득할 수 있는 최적 정책이 계산된다. 게임 플레이 에이전트는 이렇게 계산된 정책에 따라 주어진 상태에 대해 최적의 행동을 수행하게 된다.

* DOI: https://doi.org/10.22648/ETRI.2019.J.340609
* 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2019년도 문화기술연구개발 지원사업으로 수행되었음[19CS1710, 메타 플레이 인식 기반 지능형 게임 서비스 플랫폼 개발].

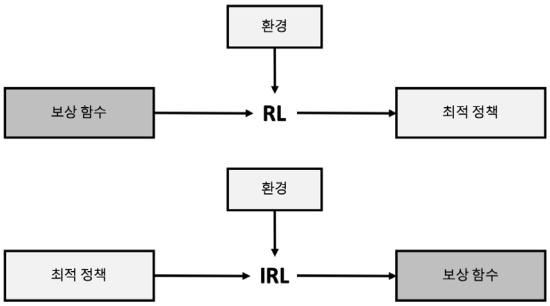


그림 1 RL과 IRL 개념 비교

표 1 RL과 IL 개념 비교

| Type | RL | IL |
|---|---|---|
| Learn | Learn through rewards | Learn through demonstrations |
| Policy | Trial and error | No rewards necessary |
| Speed | Super speed simulation | Real time interaction |
| Style | Agent becomes “Optimal” at task | Agent becomes “Human–like” at task |

하지만, 실세계에서 특정 모델에 대한 보상 함수를 구하는 것은 매우 복잡한 문제이다. 예를 들어, 보상 함수는 단일 속성이 아닌 다속성으로 구성되는 경우가 대부분이다. 즉, 보상 함수를 정의할 때 미지의 보상 속성까지 추가적으로 고려해야 한다.

역강화학습(IRL: Inverse Reinforcement Learning)은 에이전트의 정책이나 행동 이력을 통하여 그 행동을 설명하는 보상 함수를 구하는 알고리즘이다. 즉 주어진 설정이 RL의 역이 되며, 에이전트가 최선의 행동을 선택했다는 가정하에 이 행동에 대한 보상 함수를 추정하는 학습 방식이다. 따라서 RL과 달리 복잡한 상황에서 다양한 보상 요소를 반영하여 최적의 정책을 찾는 데 용이하다.

RL과 IRL의 개념적 비교는 그림 1과 같다. RL은 주어진 보상 함수를 통해 최적 정책을 계산하는 반면, IRL은 최선의 행동 이력(최적 정책)을 입력으로 보상 함수를 찾는다.

본 고에서는 IRL 기술들을 설명하고, 최근의 연구 동향을 살펴보고자 한다. 먼저 IRL의 주요 알고리즘에 대해 설명한 후, IRL 응용과 연구 동향에 대한 요약으로 결론을 맺는다.

## Ⅱ. 모방학습

실세계에서는 RL에 필요한 보상을 구하기 어렵기 때문에 비교적 간단한 시뮬레이션이나 게임과 같이 명확한 보상을 구할 수 있는 환경을 기반으로 한 연구가 활발히 진행되고 있는 추세이다.

다른 방식으로 접근하는 모방학습(IL: Imitation Learning)은 학습자로 하여금 최상의 성능을 달성하기 위해 전문가의 행동을 모방하려고 하는 순차적 작업이다[1].

IL은 직접적인 보상이 요구되지 않으며, 정책을 직접적으로 설계하여 전문가가 원하는 행동을 보다 쉽게 발현시킬 수 있다는 장점이 있다.

### 1. 행동복제

그림 2와 같이 행동복제(BC: Behavior Cloning) 기술은 전문가를 통해 쌍으로 이루어진 상태($o_t$) 및 동작($u_t$) 시퀀스 시연 궤적(Demonstration Trajectory)을 수집하여 정책($\pi_\theta$)을 지도학습한다[2].

일반적으로 BC 방식은 직관적으로 이루어지는 단순 작업의 경우 학습 효율성이 높지만, 문제해결을 위한 복잡도가 높아질수록 학습을 위한 데이터

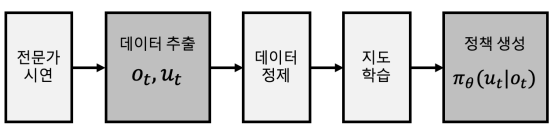


그림 2 지도학습 기반 IL 알고리즘

가 많아야 하며 테스트가 누적될수록 오차가 커지는 문제가 발생할 수도 있다[3,4].

또한, 전문가 시연 궤적도 사람이 수동적으로 행동한 데이터를 기반으로 수집되기 때문에 양적, 질적 한계가 존재하며, 결측 구간에서 성능이 현저하게 저하되는 문제가 발생한다.

### 2. 견습학습

견습학습(AL: Apprenticeship Learning)은 전문가의 시연으로부터 보상 함수를 만들고 계산된 보상 함수를 통해 최적의 정책을 학습하는 알고리즘이다[5]. 보상 함수가 수집된 특성에 대해 학습이 가능한 선형 조합(Linear Combination)이라 가정하고, 전문가 시연 궤적으로부터 보상 함수를 추정하는 방식으로 앞서 설명한 BC 기법과 의미적으로 유사하다. 하지만 IRL 방식을 연계 활용하면 전문가의 시연을 무작정 따라하는 감독학습 방법 대비 적은 데이터로 학습이 가능하고 예상치 못한 환경 대응에 강인하기 때문에 보다 효과적이고 향상된 성능을 보인다.

그림 3과 같이 전문가와 학습자의 기대치($u_e$, $u_o$) 집합으로부터 계산된 성능차이($t$)를 최소화하는 과정을 통해 보상값($R$)을 찾고, 이를 RL에 적용하여 최적 정책을 업데이트함과 동시에 수집되는 전문가의 기대치를 기존 집합에 증강함을 반복 수행하다가 성능차이가 임계치 이하로 수렴하면 학습을 종료한다.

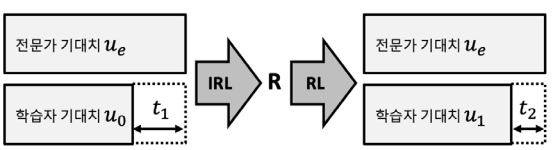


그림 3 IRL 기반 AL 알고리즘

## III. IRL 알고리즘

### 1. ALIRL

앞서 설명한 것처럼 AL은 전문가의 시연에 기반하여 진행되는 학습 방법으로, 마코프 결정 과정(MDP: Markov Decision Process) 환경에서 문제해결 임무(task)를 수행하는 전문가 시연을 관찰할 수 있을 때, 보상이 주어지지 않았거나 보상 방법을 결정하기 어려운 문제를 효과적으로 해결할 수 있는 방법으로 활용할 수 있다. 하지만, 단지 전문가의 행동을 그대로 따라하는 모방(mimic) 문제에만 집중하여 정책을 학습하는 BC 방식은 경로가 길어질수록 시시각각 다르게 주어지는 상태 정보에 잘 대응하지 못하는 단점이 있다.

이를 효과적으로 처리하는 정책을 배우기 위해서는, 전문가 행동을 그대로 따라 하는 것보다 전문가 행동에서 목적하는 의도를 학습하는 것이 더 효과적이다. 따라서 관찰되는 전문가의 시연 속에는 보상 함수를 최대화하기 위한 행동이 내재되어 있다고 가정하고, 숨겨진 보상 함수를 복구하기 위해 IRL을 사용하는 ALIRL(Apprenticeship Learning via IRL) 알고리즘[5]이 제안되었다. 즉 ALIRL은 보상 함수를 우리가 알 수 있는 feature들의 선형 조합으로 표현하고, 이를 IRL 방법으로 학습하는 알고리즘이다.

이를 위한 학습 진행은 크게 4단계로, feature expectation 집합으로부터 계산한 전문가와 학습자의 성과(performance) 차이를 통해 보상 함수를 찾는 IRL 단계와 IRL 단계에서 얻은 보상 함수에 대한 최적 정책을 찾는 RL 단계, 그리고 RL 단계에서 구한 정책으로부터 몬테 카를로(Monte Carlo) 수행을 통해 새로운 feature expectation을 구하는 단계를 반복하다가, 마지막으로 충분히 가까워지는 feature expectation이 확보되었을 때 학습을 종료하는 단계

로 나누어 볼 수 있다. 결과적으로, 이를 통한 실험에서 상대적으로 적은 횟수로도 학습이 가능하면서 전문가 시연과 비슷한 성능을 얻을 수 있는데, 특히 이러한 방법론은 자동차 운전을 학습하는 문제와 같이 매 순간의 상황 속에서 전문가의 행동 특성(Driving Style)을 잘 습득하는 특징을 확인할 수 있다.

## 2. MaxEnt IRL

앞서 기술한 ALIRL은 전문가와 학습자 행동 간 feature expectation의 계산을 통해 보상 함수를 구하는 데 있어 모호한 면이 있다. 즉 각 정책은 수많은 보상 함수에 대해 최적일 수 있으며, 수많은 정책이 동일한 feature count를 발생시킬 수도 있다. 이러한 모호성 문제를 원칙적으로 해결하기 위해서는 단일 확률적 정책을 제공해야 한다.

최대 엔트로피 IRL(Maximum Entropy IRL, Max-Ent IRL)[6]은 확률 분포 선택에 대한 모호성을 해결하기 최대 엔트로피 원리(Principle of Maximum Entropy)[6]를 이용하였다. 즉 MaxEnt IRL은 관측된 데이터로부터 feature expectation 정합 제약 조건을 따르는 경로를 통해 확률 분포를 구하고, 그 분포의 엔트로피를 최대화하는 알고리즘이며, 그 과정은 다음과 같다.

1) 파라미터 초기화
2) 보상에 관한 최적의 정책 계산
3) 상태 방문 빈도 계산
4) 확률 분포 엔트로피 최대화를 위한 그래디언트 계산
5) 계산된 그래디언트를 이용하여 파라미터 업데이트
6) 2)번부터 반복

MaxEnt IRL의 보상 함수는 사람이 수동으로 선택한 feature에 대한 선형 조합으로 정의된다. 이 방식은 실제 보상을 선형 모델을 통해 정확하게 근사화할 수 없는 경우 최적이 아닌 차선책이 된다. 이러한 선형 모델의 한계점을 극복하기 위해 가우시안 프로세스(GP: Gaussian Process) 프레임워크를 이용하는 비모수(non-parametric) 방식인 GPIRL[7]이 대안책으로 제안되었다. 하지만, 이 알고리즘은 전문가 시연 개수에 비례하여 계산 복잡성이 증가하는 심각한 단점을 갖고 있다.

심층 최대 엔트로피 IRL(Maximum Entropy Deep IRL, DeepIRL)[8]은 IRL에 대한 최대 엔트로피 패러다임을 기본으로 복잡하고 비선형적인 보상 함수를 효과적으로 근사화하는 심층 신경망을 통해 feature를 선택해야 하는 전처리 과정 없이도 우수한 성능을 보여주었다. 또한, DeepIRL의 알고리즘 복잡도는 시연 샘플 개수와 무관하다는 장점도 포함하고 있다.

## 3. GCL

비용 함수(Cost Function)를 얻기 위한 학습 방법(Cost Learning)의 어려운 점은 효과적으로 학습이 진행될 수 있는 feature를 찾아야 하는 것과 더불어, 반복적인 비용 최적화(Iterative Cost Optimization) 과정 속에서 현재 결정된 비용에 대해 최적의 정책을 찾는 연산[6]이 동시에 요구된다는 것이며, 이는 계산 복잡도 문제를 발생시킨다. 특히, 로보틱스와 같은 고차원 연속 시스템(High-Dimensional Continuous System)을 위한 역학 문제를 대상으로 비용 함수를 얻을 때 그 문제점이 두드러지게 된다.

이러한 문제점을 해결하기 위해 샘플 기반 근사화(Sample-Based Approximation)를 통해 분배 함수(Partition Function)를 계산하는 방식이 적용되고

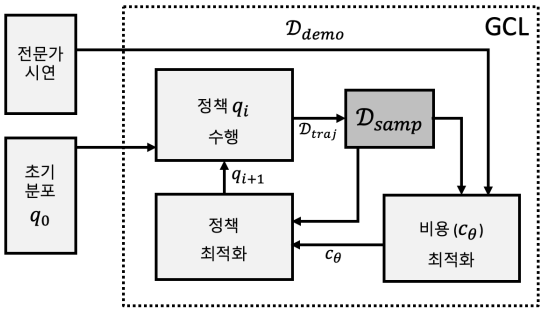


그림 4 GCL 개요도

있는데, 최근에는 신경망(Neural Network) 기반의 비선형 함수 근사화와 정규화 기법(Regularization Technique)을 이용하는 GCL(Guided Cost Learning) 알고리즘[9]이 제안되었다(그림 4).

참고문헌 [6]에서 비용 학습의 내부 루프를 통해 정책 탐색(Policy Search)이 이루어졌던 것과 달리, GCL에서는 정책 탐색의 내부 루프에서 비용값이 업데이트된다. 결국 정책 최적화 과정이 비용 함수를 '가이드'하는 형태를 취하고 있다.

학습 알고리즘은 궤적에 대한 비용 분포(Cost Distribution)의 엔트로피가 최대가 되도록 샘플링 분포(Sampling Distribution)를 조정해나가며, 샘플 기반 근사화 방식의 역최적화 제어(Inverse Optimal Control) 목적 함수를 위해 중요도 샘플링(Importance Sampling) 방법을 취하고 있다.

또한 전문가 시연을 통해 얻어지는 샘플 궤적은 정책 업데이트와 분배 함수 추정을 위해 사용되는데, 참고문헌 [10]의 실험을 통해 복잡한 관절과 3차원 연속공간에서 GCL으로 얻어지는 로봇 행동 학습은 우수한 결과를 보여주었다.

## 4. GAIL

GAIL(Generative Adversarial Imitation Learning)[2]은 앞서 설명된 GCL[9]의 구조에서 더 효율적인 계산을 위해 적대적 생성 방법(GAN: Generative Adversarial Net)[11]을 도입한 것이다. 기존의 역강화학습을 이용한 모방학습 방법론은 다음과 같은 과정을 따른다.

1) 정책 초기화
2) 현재 정책과 전문가의 정책을 비교하여 그래디언트 계산
3) 계산된 그래디언트를 이용하여 보상 함수 업데이트
4) 현재 보상 함수로부터 최적의 정책 계산
5) 2)번부터 반복

이러한 과정에서 가장 문제가 되는 부분이 바로 4)번 단계이다. 현재 보상 함수로부터 최적의 정책을 계산하기 위해 RL을 처음부터 끝까지 진행해야 하며, 이는 많은 시간과 계산 자원을 필요로 한다. 따라서 GCL은 이러한 계산을 줄이기 위해 RL으로 최적의 정책을 한 번에 찾지 않고, IRL과 번갈아가면서 서로 동시에 업데이트가 이루어지도록 하였다. IRL을 통해 보상 함수가 일부 수정되고, 이러한 불완전한 보상 함수를 통해 정책이 일부분 개선되며 최적화가 진행된다. 이는 GAN의 관점으로 해석할 수 있는데, RL 과정에서 업데이트되는 정책은 전문가의 정책을 모방하기 위해 계속해서 유사한 정책을 만들려 하는 생성기(generator)의 역할을 하고, IRL 과정에서 업데이트되는 보상 함수는 전문가가 수행한 행동인지 아닌지를 구별하려 하는 판별기(discriminator)의 역할을 한다.

GAIL의 수식 도출과정은 다음과 같다. 먼저 기존의 IL에 사용되는 RL과 IRL의 과정을 하나의 식으로 합성하고, 정책을 조금 더 쉽게 다루기 위해 일대일 대응관계에 있는 occupancy measure를 도입하여 변수를 치환한다. Occupancy measure란 특정한 정책 아래에서 발생하는 상태-행동 쌍의 정규

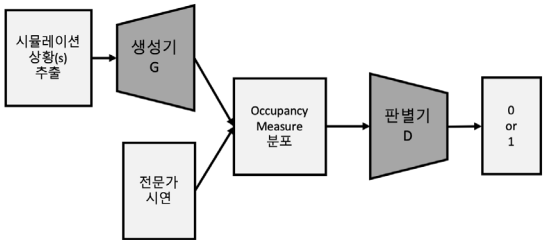


그림 5 GAIL 알고리즘 개요도

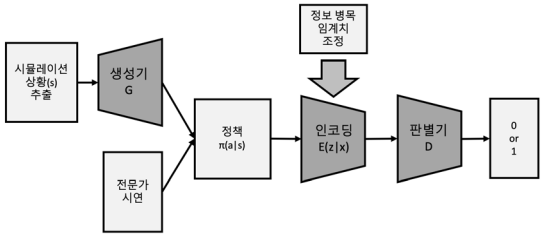


그림 6 VAIL 알고리즘 개요도

화되지 않은 분포를 의미한다. 정규화되지 않았기 때문에 이 분포는 총합이 1을 넘을 수 있다. 이로부터 IL 문제에서 분포를 매칭시키는 문제로 바뀌게 된다. 치환된 수식에는 보상 정규화 함수 $\psi$를 추가한 뒤, convex conjugate 함수($\psi^*$)로 변형하여 수식을 단순화하고, 정규화 함수를 GAN의 형태로 정의하여 GAN과 유사한 목적 함수를 만들었다.

그림 5는 GAIL의 알고리즘을 도식화한 것이다. 실제 알고리즘에서는 RL 과정에서 정책이 급하게 변하는 것을 방지하기 위해 TRPO(Trust Region Policy Optimization)[12]를 적용하였다. GAIL은 GCL에서 비효율적이었던 정책 샘플링 과정을 occupancy measure와 GAN을 도입함으로써 효율적인 IL을 가능하게 하였다.

## 5. VAIL

GAIL을 실제 문제에 적용하려면 내부에서 핵심적인 역할을 하는 GAN을 빼놓을 수 없다. 하지만 GAN의 경우 생성기와 판별기 사이의 학습 균형을 맞추기가 어렵다는 단점으로 인해 IL의 성능을 보장할 수 없다는 문제가 발생된다. 따라서 VAIL (Variational Adversarial Imitating Learning)[13]에서는 이러한 문제점을 해결하기 위해 그림 6과 같이 판별기에 정보 병목(Information Bottleneck)을 가하는 방법을 제안하였다.

기본적인 구조는 GAIL과 유사하나, 모방하고자 하는 전문가의 데이터와 생성기로부터 출력되는 데이터 $x$는 판별기에 입력되기 전에 인코딩 과정을 거쳐 데이터 $z$가 들어가게 된다. 인코딩을 통해 정보를 제한하는 과정에서는 상호 정보량(Mutual Information)을 이용하였다. 즉 상호 정보량 값이 크면 $x$의 정보가 고스란히 $z$로 들어갔음을 의미하며, 값이 작으면 $x$의 정보가 $z$로 전달이 제대로 안 되었음을 의미한다. 상호 정보량 계산 과정에서 $z$의 확률분포를 모르기 때문에 값을 직접 계산할 수 없다. 따라서 KL-divergence(Kullback-Leibler divergence)형태로 변형하여 근사 값을 구하였다.

VAIL에서 제안하는 정보 병목은 판별기의 학습 속도를 직접적으로 늦추는 역할을 한다. 즉 생성기가 불완전하게 학습된 상태에서 판별기가 너무 빠르게 학습되는 현상을 방지해 주는 동시에, 생성기에게는 학습이 잘 될 수 있도록 도움이 되는 그래디언트 정보를 제공해 준다. 또한, 정보를 압축하는 과정에서 불필요한 정보를 무시할 수 있도록 해준다.

## 6. InfoGAIL

앞서 기술한 GAIL은 전문가 시연 궤적 내에서 상이한 행동 유형을 구별하지 못하며, 고차원 입

력의 문제를 다루는 데 있어 그 성능이 보장되지 않는다는 한계를 갖고 있다. InfoGAIL[14]은 InfoGAN[15]과 유사하게, 잠재 공간(Latent Space)과 궤적 간의 상호 정보량을 근사적으로 최대화하는 요소를 GAIL에 확장 적용하여, 좀 더 추상적이고 높은 레벨의 잠재 변수(Latent Variable)를 통해 낮은 레벨의 행동을 제어할 수 있는 정책을 구성할 수 있다. 또한 InfoGAIL은 선행 지식을 통해 전문가 시연에서 추론할 수 없는 보상 요소를 추가하여 IL과 RL이 혼합된 형태의 보상 함수를 정의하였으며, 최적화 과정을 개선시키기 위해 WGAN(Wasserstein GAN)[16] 기술을 도입하였다.

InfoGAIL은 화소를 입력으로 사용하여 복잡한 고차원 동적 환경에서 인간과 유사한 행동 생성이 가능하다. TORCS 경주 시뮬레이터(The Open Source Racing Car Simulator)[17] 실험을 통해 InfoGAIL이 잠재 코드에 따라 서로 다른 운전 유형을 생성한다는 것을 보여준다.

## IV. 결론

본 고에서는 IRL 방식들을 살펴보고, 각 알고리즘의 특성을 분석해 보았다. 최근 RL은 가상 시뮬레이션 환경의 연구 단계에서 자율 주행, 자연어 처리, 추천 시스템, 질병 진단 등 광범위한 응용 단계로 확장되고 있다. 하지만, RL은 이러한 복잡한 실세계 환경에서 활용 가능성이 낮아진다. IRL은 다양한 상황에서 최적의 정책을 찾을 수 있으며, 전문가의 시연 데이터를 통해 좀 더 정확하고 세밀하게 목표 임무를 수행할 수 있다. 특히, IRL은 인간의 지적 업무를 성공적으로 수행할 수 있는 인공 일반 지능(AGI: Artificial General Intelligence) 연구의 주요 핵심 기술이 될 것으로 기대된다.

### 약어 정리

| | |
|---|---|
| AGI | Artificial General Intelligence |
| AL | Apprenticeship Learning |
| ALIRL | Apprenticeship Learning via IRL |
| BC | Behavior Cloning |
| GAIL | Generative Adversarial Imitation Learning |
| GAN | Generative Adversarial Network |
| GCL | Guided Cost Learning |
| IL | Imitation Learning |
| IRL | Inverse Reinforcement Learning |
| MDP | Markov Decision Process |
| RL | Reinforcement Learning |
| TORCS | The Open Source Racing Car Simulator |
| TRPO | Trust Region Policy Optimization |
| VAIL | Variational Adversarial Imitation Learning |
| WGAN | Wasserstein GAN |

### 참고문헌

[1] A. Attia et al., "Global overview of imitation learning," arXiv:1801.06503, 2018.

[2] J. Ho et al., "Generative adversarial imitation learning," Advances in Neural Information Processing Systems, 2016.

[3] S. Ross et al., "Efficient reductions for imitation learning," Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010.

[4] S. Ross et al., "A reduction of imitation learning and structured prediction to no-regret online learning," Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011.

[5] P. Abbeel et al., "Apprenticeship learning via inverse reinforcement learning," Proceedings of the Twenty-First International Conference on Machine Learning, 2004.

[6] B. Ziebart et al., "Maximum Entropy Inverse Reinforcement Learning," Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, 2008.

[7] S. Levine et al., "Nonlinear inverse reinforcement learning with gaussian processes," Advances in Neural Information Processing Systems, 2011.

[8] M. Wulfmeier et al., “Maximum entropy deep inverse reinforcement learning,” arXiv:1507.04888, 2015.

[9] C. Finn et al., “Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization,” Proceedings of the Thirty-Third International Conference on International Conference on Machine Learning, 2016.

[10] http://rll.berkeley.edu/gcl

[11] I. Goodfellow et al., “Generative adversarial nets,” Advances in Neural Information Processing Systems, 2014.

[12] J. Schuman et al., “Trust region policy optimization,” Proceedings of the Thirty-Second International Conference on International Conference on Machine Learning, 2015.

[13] X. Peng et al., “Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow,” Proceddings of the International Conference on Learning Representations, 2019.

[14] Y. Li et al., “InfoGAIL: Interpretable Imitation Learning from Visual Demonstrations,” Advances in Neural Information Processing Systems, 2017.

[15] X. Chen et al., “InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets,” Advances in Neural Information Processing Systems, 2016.

[16] M. Arjovsky et al., “Wasserstein Generative Adversarial Networks,” Proceedings of the Thirty-Fourth International Conference on International Conference on Machine Learning, 2017.

[17] http://torcs.sourceforge.net/