이미지 기반 정책 함수에 대한 적대적 공격*

이찬미[O α], 윤민성[β], 윤성의[γ]
[α]한국과학기술원 로봇공학학제전공 [βγ]한국과학기술원 전산학부
[α]chanmi99@kaist.ac.kr, [β]minsung.yoon@kaist.ac.kr, [γ]sungeui@kaist.ac.kr

Adversarial Attack on Visuomotor Policy

Chanmi Lee[O α], Minsung Yoon[β], Sungeui Yoon[γ]
[α]Robotics Program, KAIST [βγ]School of Computing, KAIST
[α]chanmi99@kaist.ac.kr, [β]minsung.yoon@kaist.ac.kr, [γ]sungeui@kaist.ac.kr

요 약

이 논문은 심층 강화학습으로 학습된 이미지 기반 행동 정책 모델이 적대적 공격에 취약할 수 있음을 보인다. 적대적 공격은 이미지에 미세한 교란을 추가하여 인공지능 모델의 예측을 왜곡하는 것을 의미한다. 로보틱스 분야에서도 인공지능 모델이 사용되면서 이러한 공격에 관한 연구가 주목받고 있다. 공격을 받으면 로봇은 예상치 못한 행동을 하여 위험한 상황을 일으킬 수 있다. 본 연구는 대표적인 강화학습 기법인 Soft Actor Critic (SAC) 모델을 대상으로, 가치 최소화 교란 기법을 제시하며 이를 적용하여 정책 모델이 낮은 가치를 가지는 행동을 하도록 유도한다. 실험 결과는 공격 여부에 따른 Q 함숫값 차이와 로봇 움직임 경로를 분석하여, 적대적 교란이 로봇의 행동 결정에 어떻게 영향을 주는지 평가하였다. 이를 통해 심층 강화학습 기반 행동 결정 시스템이 적대적 공격에 취약할 수 있음을 보였다.

1. 서 론 (Introduction)

적대적 공격(adversarial attack)은 이미지에 미세한 교란(perturbation)을 추가해 학습 모델의 예측을 왜곡하는 것을 말한다 [1]. 이 교란은 이미지 분류 모델을 비롯한 다양한 인공지능 모델을 속여, 정답(ground truth, GT)과 다른 결과를 도출하게 한다.

최근 로보틱스 분야에서도 인공지능 모델이 사용됨에 따라 적대적 공격에 관한 연구가 주목받고 있다 [2, 3]. 로봇의 행동을 결정하는 정책 모델이 적대적 공격을 받게 되면, 로봇이 잘못 혹은 예상치 못한 행동을 하여 위험한 상황이 발생할 수 있다. 따라서 로봇이 사람들을 회피하지 못해 충돌할 수 있고, 공격자가 교란을 통해 로봇을 임의로 제어하여 개인정보 수집 및 불법행위를 저지를 수도 있다.

본 연구에서는 심층 강화학습으로 학습된 이미지 기반 행동 정책 모델이 적대적 공격에 취약할 수 있음을 보이고자 한다. 이를 위해 대표적인 강화학습 기법인 Soft Actor Critic (SAC [4]) 모델을 대상으로 적대적 공격을 진행하였고, 이는 각 상태에서 정책이 도출하는 행동의 가치를 평가하는 Q 함수(q function, state-action value function)를 대상으로 가치 최소화 교란을 생성하여 정책 모델이 가치가 낮은 행동을 하도록 유도하였다. 본 연구의 실험 결과는 공격 여부에 따른 Q 함숫값의 차이와 로봇 움직임의 경로를 분석하여, 적대적 공격으로 생성된 이미지 교란이 로봇의 행동 결정 과정에 어떻게 영향을 주는지 평가했다. 이를 통해 심층 강화학습 기반 행동 결정 시스템이 적대적 공격에 취약할 수 있음을 보였다.

2. 배경 지식 (Background)

2.1. 적대적 공격 (Adversarial Attack)

적대적 공격(adversarial attack)은 학습된 모델의 예측값을 의도적으로 왜곡하는 과정이다. 이는 이미지에 사람의 눈으로는 감지할 수 없는 작은 변형을 추가함으로써 모델의 예측력을 손상하는 방식으로 수행된다. 일반적으로 학습모델은 경사하강법을 이용해 정답(GT) 값과 가장 유사한 예측값을 생성하지만, 적대적 공격은 경사상승법을 활용해 손실을 극대화하도록 모델이 잘못된 예측을 하도록 유도한다.

Fast Gradient Sign Method (FGSM) [2] 은 대표적인 적대적 공격 방법으로, 입력 데이터 x에 대한 모델의 손실함수 L를 최대화하기 위해 작은 노이즈 $sign(\nabla_x L(\theta, x, y))$를 추가한다. 이를 통해 변형된 데이터는 다음과 같다.

$$x' = x + \epsilon \cdot sign(\nabla_x L(\theta, x, y))$$

이때, ϵ은 노이즈의 크기를 결정하는 변수로, 생성된 노이즈의 픽셀 값이 해당 값 이하로 유지되도록 하며, 특정 크기 이내의 교란을 만드는 것을 목표로 한다.

2.2. 강화학습 모델: Soft Actor-Critic (SAC)

Soft Actor-Critic (SAC) [4]는 강화학습 모델 중 하나로, 기대보상과 행동 엔트로피를 동시에 최적화하는 Model-free 강화학습 알고리즘이며, 경험을 기반으로 복잡한 행동을 학습할 때 이용한다. SAC는 $< S, A, R(s,a), P(s'|s,a), \gamma >$로 정의되는 Markov Decision Process (MDP)를 기반으로 문제를 정의한다. 이는 각각 상태 S, 행동 A, 보상 함수 $R(s,a)$, 전이 확률 $P(s'|s,a)$, 할인 계수 γ를 의미한다.

SAC의 목적은 목표 함수 $J(\pi)$를 최대화하는 정책 함수 π를 찾는 것이다. SAC의 목표 함수는 평균 기대 보상과 정책의 엔트로피를 포함하며, 이는 탐색과 이용 간의 균형을 유지하는 데 도움이 된다. 따라서, SAC의 목적 함수는 다음과 같이 나타낼 수 있다.

* 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원 의 지원을 받아 수행된 연구임
(RS-2023-00237965, 오픈 월드 로봇 서비스를 위한 불특정 환경 인지, 행동 및 상호작용 알고리즘 개발)

$$J(\pi) = E_{(\mathrm{s,a})\sim\pi}[\mathrm{r(s,a)} + \alpha H(\pi(\cdot\,|\mathrm{s}))]$$

이 때, α는 엔트로피의 중요도를 조절한다. 이와 같이 SAC는 정책의 엔트로피를 최적화하면서 평균 기대 보상을 극대화하여 강화학습 목표를 달성한다.

이 때 샘플의 효율성 향상을 위해 Q 함수를 이용해 책을 학습하게 되며, 이는 특정 상태 s에서 시작하여 행동 a를 취하고 이후 특정 정책을 따를 경우, 에이전트가 받게 될 것으로 예상되는 누적 보상의 추정치를 나타낸다.

$$Q(s,a) = E\left[\sum_{k=0}^{\infty}\gamma^k r(s_{t+k}, a_{t+k})|s_t = s, a_t = a\right]$$

Q 함수 업데이트는 다음과 같이 진행된다.

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$$

여기서 s', a'은 다음 타임 스텝에서의 상태와 행동을 나타내고, α는 업데이트 스텝을 조절하는 하이퍼파라미터이다.

3. 방법 (Method): SAC에 대한 적대적 공격

3.1. 개요

적대적 공격은 수식(1)과 같이, 학습된 인공지능 모델을 가장 효과적으로 속이기 위한 교란을 생성하는 것을 목표로 한다. 생성된 교란은 이미지에 미세한 노이즈를 추가하여 학습된 모델이 이미지 분류를 잘못하도록 유도한다.

본 연구에서는 SAC 기반의 심층 강화학습 모델에 대한 적대적 공격 방식을 제안하고, 적대적 공격에 효과성을 보여 심층 강화학습 모델의 적대적 공격에 대한 취약성을 보이는 것을 목표로 한다.

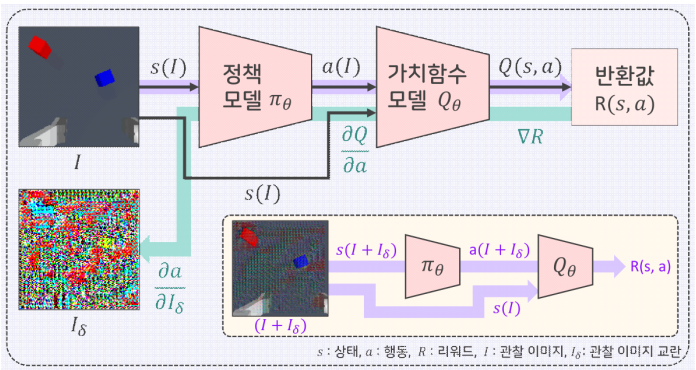

그림 1. 이미지 기반 정책 함수에 대한 적대적 공격 방법

3.2. 가치 최소화 교란 기법

SAC 모델에 적대적 공격을 하기 위해 그림 1의 방법을 이용하고자 한다. 그림 1에 검은색 화살표 방향은 기존 SAC의 방식의 정방향 전파의 순서를 나타낸다. 정책 모델 π_θ는 상태 s를 나타내는 이미지 I를 입력으로 받아 로봇이 수행할 행동 $a = \pi_\theta(s(I))$을 도출하고, 가치함수 모델 Q_θ는 현재 상태와 행동에 대한 누적 보상 추정치인 q value 을 계산한다.

본 연구에서는 그림 1의 초록색 화살표와 같이 SAC 모델을 적대적 공격하기 위해 q value를 최소화 하는 방향에 교란이미지 I_δ를 역전파를 통해 구하게 된다. 이를 통해 정책 모델은 원본 이미지에 교란이 더해진 이미지를 기반으로 행동을 추론하게되고, 이는 정책이 현재 상태에서 q value를 최소화하는 action을 취하도록 유도된다. $\hat{a} = \pi_\theta(s(I+I_\delta))$. 공격 혹은 교란 당한 이미지의 생성 과정을 수식으로 표현하면 다음과 같다.

$$I_{attacked} = I + I_\delta = I - \epsilon\frac{\partial R}{\partial I_\delta} = I - \epsilon\frac{\partial R}{\partial a}\frac{\partial a}{\partial I_\delta}$$

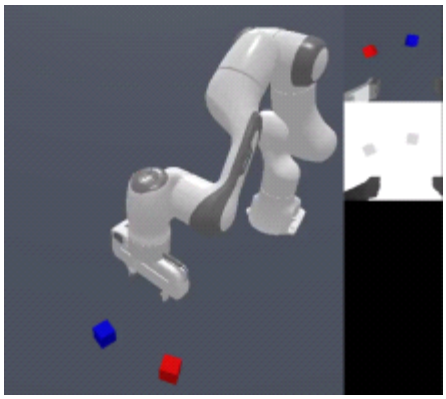

그림 2. 왼쪽 그림은 로봇과 두 박스가 있는 실험 환경이며, 오른쪽 그림은 손목에 달린 카메라로부터 얻은 시각 및 깊이 이미지이다. 시각 및 깊이 이미지의 관찰 이미지를 입력으로 받아, 로봇이 빨간박스로 다가가는 작업 수행 및 이를 공격할 것이다.

4. 실험 및 결과

4.1. 실험 환경 설정

본 연구에서 제안한 SAC기반의 이미지 기반 정책 모델에 대한 적대적 공격 방법의 효과를 확인하기 위해 그림 2와 같이 실험 환경을 구성하였다. 지면에 빨간박스, 파란박스가 있으며, Franka Panda 로봇을 이용하여 빨간박스를 잡는 문제를 구성하였다. 이 때 로봇의 손목에 카메라가 장착되어있고, 해당 카메라로 찍은 장면을 SAC 모델의 상태 정보로 구성하였고 이를 기반으로 로봇 끝단의 velocity를 도출하여 움직이는 Visuomotor policy 모델을 구현하였다.

이러한 강화학습 정책을 학습하기 위해 보상함수는 아래와 같이 구성하였다.

표 2. 보상함수 정의

보상함수	보상함수 정의
R_{total}	$R_{reach\_redbox} - R_{reach\_bluebox} + R_{reach} + R_{grasp}$
R_{reach_redbox}	$\exp(-\|X_{ee} - X_{redbox}\|^2/0.2)$
$R_{reach\_bluebox}$	$\exp(-\|X_{ee} - X_{bluebox}\|^2/0.05)$
R_{reach}	$\begin{cases} 3 & (\|X_{ee} - X_{redbox}\|^2 \le 0.05) \\ -1 & (\|X_{ee} - X_{bluebox}\|^2 \le 0.05) \\ -0 & (else) \end{cases}$
R_{grasp}	$\begin{cases} 3 & (is\_grasp\_redbox = True) \\ -1 & (is\_grasp\_bluebox = True) \\ -0 & (else) \end{cases}$

R_{reach_redbox}에 의해 로봇 끝단이 빨간박스로 다가갈수록 큰 보상을 받고, $-R_{reach_bluebox}$에 의해 로봇 끝단이 파란박스로 다가갈수록 큰 불이익(penalty)을 받는다. R_{reach}는 로봇 끝단이 빨간박스로 성공적으로 다가갔는지에 대한 보상함수로, 로봇 끝단이 빨간박스의 거리가 5cm 이하인 경우 +3의 보상을 받고, 파란박스와의 거리가 5cm 이하인 경우 -1의 불이익을 받는다. R_{grasp}는 그리퍼가 빨간박스 잡기를 성공했는지에 대한 보상함수로, 빨간박

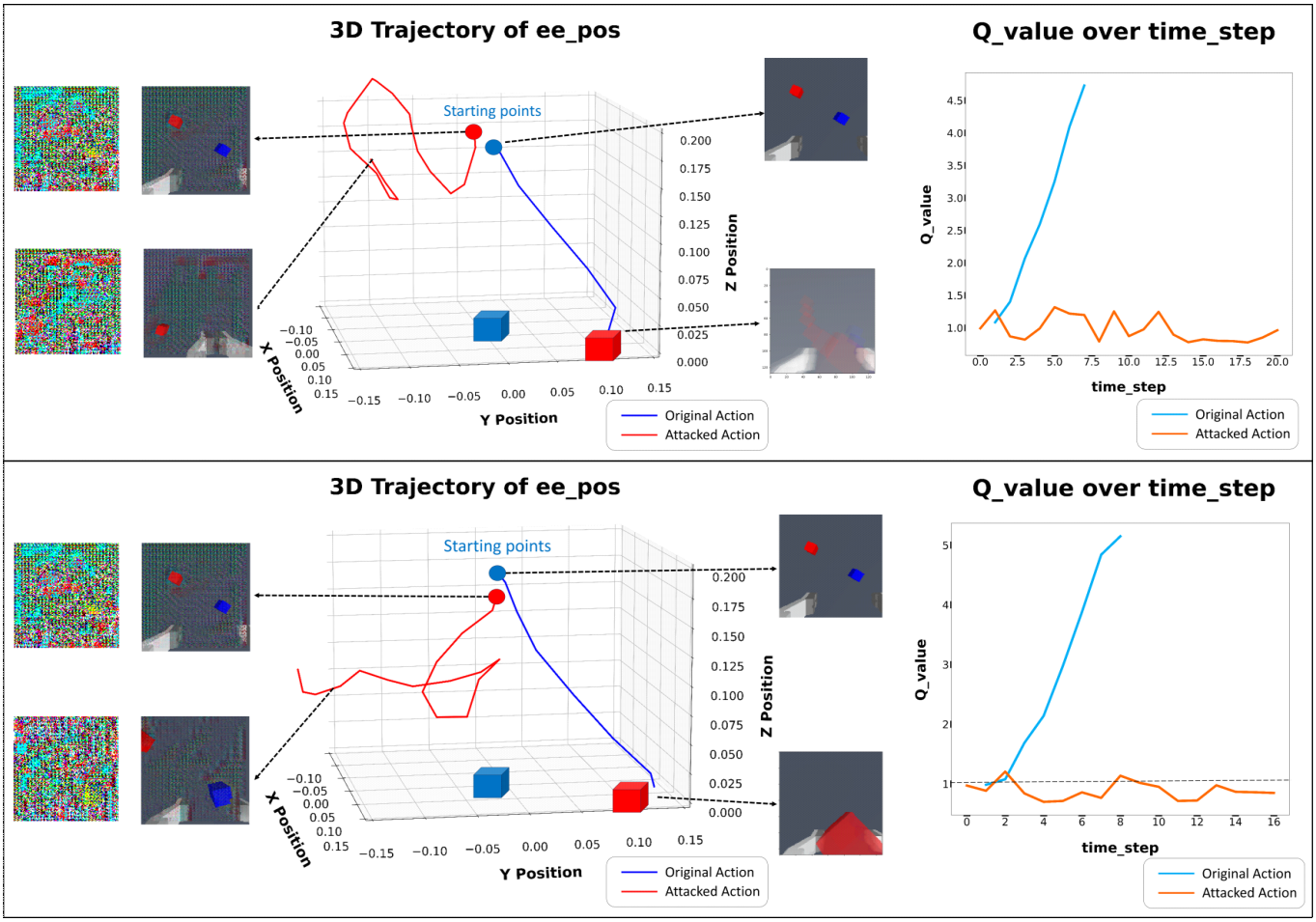

그림 3. 이미지 기반 정책 모델에 대한 적대적 공격의 실험 결과들. 기존 SAC 모델의 경우 빨간박스로 다가가기 위한 로봇의 행동을 출력(파랑 궤적)하며, Q 함수의 값이 시간에 지남에 따라 증가함을 확인할 수 있다. 반면, 적대적 공격을 당한 모델의 경우, 빨간박스로 다가가지 못하는 행동을 출력(빨강 궤적)함으로써, Q 함수의 값이 감소하는 양상을 보였다.

스를 잡았을 경우 +3을, 파란박스를 잡았을 경우 -1을 나타낸다.

결과적으로 학습된 Q 함수는 로봇팔의 끝단이 빨간박스에 다가갈수록 Q 함숫값은 증가하고, 적대적 공격을 통해 Q 함숫값을 감소시키는 잘못된 행동을 유도하는 것은 타겟 물체인 빨간박스로부터 멀어지는 것을 의미하게 된다.

4.2. 실험 결과 및 분석

그림 3은 이미지 기반 정책 모델을 적대적 공격한 결과를 보여준다. 왼쪽 그림 "3D Trajectory of ee_pos"는 시간에 따른 로봇 그리퍼의 궤적을 나타낸다. 파란색 궤적은 정책 함수에 교란이 없을 때의 행동으로, 로봇의 끝단이 빨간박스로 다가가는 것을 확인할 수 있다. 반면에, 빨간색 그래프는 정책 함수의 이미지 인풋에 교란을 추가한 경우로, 정책 모델이 공격받아 목표 작업과 상관없는 예상치 못한 행동을 수행하는 것을 볼 수 있다.

오른쪽 그림 "Q_value over time_step"은 시간이 지남에 따라 Q 함숫값의 변화를 나타내는 그래프이다. 하늘색 그래프는 교란이 없는 경우로, 로봇의 끝단이 빨간박스로 다가감에 따라 급격히 증가하였다. 반면, 주황색 그래프는 로봇이 잘못된 행동을 수행함으로써 얻은 Q 함숫값으로, 교란에 의해 Q 함숫값이 증가하지 못하고 항상 1.5 이하의 값을 나타냈다.

5. 결론

본 연구는 심층 강화학습 알고리즘인 SAC로 학습된 이미지 기반 정책 모델의 적대적 공격방법을 제시함으로써, 해당 정책 모델이 적대적 공격에 취약함을 보였다. 정책 모델을 공격하기 위해 Q 함수값이 감소하도록 하는 교란을 생성하였고, 해당 교란이 로봇의 잘못된 행동을 유도하는 것을 확인할 수 있었다. 향후에는 로봇의 행동을 공격자가 원하는 방향으로 유도할 수 있는 교란을 생성하는 목표 지향적 적대적 공격에 관한 연구를 진행할 예정이다. 또한 이러한 적대적 공격에 대응할 수 있는 공격 방어 연구도 진행할 것이다.

6. 참고 문헌

[1] Huang, Sandy, et al. "Adversarial attacks on neural network policies." arXiv preprint arXiv:1702.02284, 2017.

[2] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572, 2014.

[3] Jia, Yifan, et al. "Physical adversarial attack on a robotic arm." IEEE Robotics and Automation Letters 7.4: 9334-9341, 2022.

[4] Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor." International conference on machine learning. PMLR, 2018.