Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.34, NO.5, Oct. 2024 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2024.34.5.1047

강화학습 모델에 대한 적대적 공격과 이미지 필터링 기법을 이용한 대응 방안*

이 승 열,[1†] 하 재 철[2‡]
[1,2]호서대학교 (대학원생, 교수)

Adversarial Attacks on Reinforce Learning Model and Countermeasures Using Image Filtering Method*

Seungyeol Lee,[1†] Jaecheol Ha[2‡]
[1,2]Hoseo University (Graduate student, Professor)

요 약

최근 심층 신경망을 이용한 강화학습 모델들이 자율주행, 스마트 팩토리, 홈 네트워크 등 다양한 첨단 산업 분야에 사용되고 있으나 적대적 공격(adversarial attacks)에 취약하다는 것이 밝혀졌다. 본 논문에서는 강화학습 기반의 딥러닝 모델인 DQN과 PPO를 자율주행 가상환경 HighwayEnv에 적용하여 FGSM(Fast Gradient Sign Method), BIM(Basic Iterative Method), PGD(Projected Gradient Descent) 그리고 CW(Carlini and Wagner)을 이용하여 적대적 공격을 수행하였다. 적대적 공격에 대응하기 위해 양방향 필터(bilateral filter) 알고리즘을 사용하여 적대적 이미지의 잡음을 제거함으로써 강화학습 기반의 딥러닝 모델들이 정상적으로 작동할 수 있는 방법을 제안하였다. 그리고 HighwayEnv 환경에서 에피소드 수행 길이(episode during)의 평균과 에이전트가 획득한 보상(episode reward)의 평균을 성능평가 지표로 사용하여 공격의 성능을 평가하였다. 실험 결과 양방향 필터를 통해 적대적 이미지의 잡음을 제거한 결과, 적대적 공격이 수행되기 이전의 성능을 유지할 수 있음을 보였다.

ABSTRACT

Recently, deep neural network-based reinforcement learning models have been applied in various advanced industrial fields such as autonomous driving, smart factories, and home networks, but it has been shown to be vulnerable to malicious adversarial attack. In this paper, we applied deep reinforcement learning models, DQN and PPO, to the autonomous driving simulation environment HighwayEnv and conducted three adversarial attacks: FGSM(Fast Gradient Sign Method), BIM(Basic Iterative Method), PGD(Projected Gradient Descent) and CW(Carlini and Wagner). In order to respond to adversarial attack, we proposed a method for deep learning models based on reinforcement learning to operate normally by removing noise from adversarial images using a bilateral filter algorithm. Furthermore, we analyzed performance of adversarial attacks using two popular metrics such as average of episode duration and the average of the reward obtained by the agent. In our experiments on a model that removes noise of adversarial images using a bilateral filter, we confirmed that the performance is maintained as good as when no adversarial attack was performed.

Keywords: Reinforcement Learning Model, Adversarial Attacks, Bilateral filter

Received(09. 03. 2024), Modified(10. 04. 2024), Accepted(10. 14. 2024)

* 본 논문은 2024년도 한국정보보호학회 하계학술대회에 발표한 우수논문을 개선 및 확장한 것임

* 본 논문은 2024년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력 기반 지역혁신 사업의 결과입니다. (No. 2021RIS-004)

† 주저자, stl990726@gmail.com

‡ 교신저자, jcha@hoseo.edu(Corresponding author)

I. 서 론

최근 스마트 팩토리, 홈 네트워크, 자율주행 자동차와 같은 많은 첨단 산업 분야에서 인공지능 기술들이 사용되면서 사용자들에게 편의성을 제공해주고 있다. 특히 실시간으로 연속적인 데이터를 처리가 필요한 로봇 산업이나 자율주행 자동차와 같이 예상치 못한 사고와 상황이 발생할 수 있는 환경에서 다양한 시행착오를 통해 학습하는 강화학습(reinforcement learning) [1]을 이용한 인공지능 기술들이 많이 사용되고 있다.

강화학습은 주어진 환경 정보가 부족할 때 반복적인 탐색을 통하여 의사결정 정책을 학습하고 최적의 의사결정 시 획득하는 보상(reward)을 최대화하는 학습 방법을 의미한다. 특히 차선 유지, 충돌 방지, 차선 변경 등 다양한 상황이 있는 자율주행 자동차의 경우 강화학습 기법이 중요한 역할을 한다. 그러나 최근 연구에서 강화학습 모델은 악의적인 적대적 공격(adversarial attacks)에 취약점이 있음이 밝혀졌다 [2].

적대적 공격은 사람의 눈으로는 식별하기 어려운 작은 섭동을 입력 이미지에 더함으로써 심층 신경망(Deep Neural Network, DNN) 모델들을 오동작하게 만드는 공격 방법으로 이러한 적대적 공격이 자율주행 자동차에 적용될 경우 비정상적인 동작을 유발할 수 있어 이에 대한 방어 대책이 필요하다.

적대적 공격을 방어하기 위해 다양한 방어 방법이 연구되고 있다. 대표적으로 적대적 학습(adversarial trainging) [3, 4]과 인공지능 모델을 추가로 사용하여 잡음을 제거하는 잡음 제거(denoising) [5, 6]이 있다. 적대적 학습은 인공지능 모델이 학습하는 과정에서 적대적 샘플을 학습 데이터에 추가하여 학습시킨 뒤 적대적 공격에 대응하는 강건한 모델을 만드는 방법이다. 하지만 이러한 적대적 학습은 적대적 샘플에 대한 강건성이

증가할수록 원본 이미지에 대한 정확도는 감소하게 된다는 단점이 있다.

인공지능 모델을 사용하여 잡음을 제거하는 방법에는 대표적으로 오토인코더(auto-encoder)가 있다. 오토인코더를 통해 적대적 샘플을 재구성하여 재구성된 이미지와 원본 이미지와의 차이를 줄여 인공지능 모델이 정상적으로 동작하도록 유도하는 방법이다. 그러나 이 방법은 인공지능 모델을 추가로 학습하기 때문에 추가적인 데이터셋과 학습 시간을 요구한다.

최근 Feng 등은 위와 같은 접근 방식과 달리 인공지능 모델의 출력을 암호화하여 공격자로부터 보호하는 가짜 그래디언트(fake gradient)기법을 제안하였다 [7]. 심층 신경망의 출력 레이어에 가짜 노드들을 추가하여 공격자가 실제 노드 출력 결과를 식별하게 어렵게 함으로써 적대적 공격으로부터 방어할 수 있음을 보였다.

본 논문에서는 고속도로 자율주행 시뮬레이션인 HighwayEnv [8]에 강화학습 모델인 DQN (Deep Q-Network)과 PPO (Proximal Policy Optimization)를 적용한 뒤 적대적 공격의 성공 가능성을 진단한다. 적대적 공격 알고리즘인 FGSM(Fast Gradient Sign) [9], BIM(Basic Iterative Method) [10], PGD(Projected Gradient Descent) [11] 그리고 CW(Carlini and Wagner) [12]을 수행한 결과 강화학습 모델들의 성능이 크게 저하되는 것을 확인하였다.

본 논문에서는 실험 결과를 바탕으로 적대적 공격에 대응하기 위해 이미지의 일반적인 잡음을 제거하면서 경계선을 보존하는데 사용되는 양방향 필터(bilateral filter) [13-15]을 이용하여 적대적 공격으로 발생하는 미세한 잡음을 제거하는 알고리즘을 구성하였다. 적대적 공격의 이미지 내 미세한 잡음을 추가하기 때문에 기존 양방향 필터 알고리즘에서 색상 차이에 따른 필터링 값을 높게 구성함으로써 적대적 공격으로부터 발생한 색상 변화를 효과적으로 감지할 수 있도록 하였다. 또한 적대적 공격을 통해 공간적으로 인접한 영역에서 발생하는 미세한 픽셀 변화를 효과적으로 처리하기 위해 공간적 거리에 따른 필터링 강도를 낮은 값으로 구성하여 기존 양방향 필터 알고리즘에 비해 더욱 국소적인 부분에서 잡음 제거를 수행하였다,

실험 결과 적대적 샘플에 양방향 필터(bilateral filter) 알고리즘을 적용하여 추가적인 학습 시간과 데이터셋 없이 적대적 공격이 수행되더라도 강화학습 알고리즘들이 정상적으로 수행하여 실시간으로 자율주행 차량이 동작하는 것을 확인하였다. 또한 적대적 공격이 수행되지 않은 환경에서도 큰 성능 저하 없이 주행이 가능한 것을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서 강화학습과 적대적 공격에 관련된 내용을 기술하며, 3장에서 제안하는 양방향 필터를 이용한 적대적 공격 대응 기법을 설명한다. 4장에서는 강화학습에 대한 적대적 공격 실험 환경과 성능을 평가한 뒤 5장에 결론을 맺

는다.

II. 관련 연구

2.1 고속도로 환경에서의 자율주행 자동차

HighwayEnv는 자율주행 자동차를 고속도로 환경에서 시뮬레이션 할 수 있는 환경을 제공해 주는 가상환경이다. HighwayEnv는 에이전트 차량(agent vehicle)을 제외한 다른 차량들의 개수와 움직이는 방식을 설정할 수 있다. 또한 시뮬레이션은 환경을 관측하는 방법으로 Kinematics, Grayscale image, Occupancy grid 그리고 Time to collision을 제공한다. Fig. 1은 HighwayEnv를 나타내며 시뮬레이션에서 동작하는 에이전트 차량은 주변 차량과의 충돌을 하지 않으며 최종 지점에 도착하는 것을 최종 목표로 한다. Fig. 1은 HighwayEnv에서 대상 차량의 주행을 시각화한 것이다.

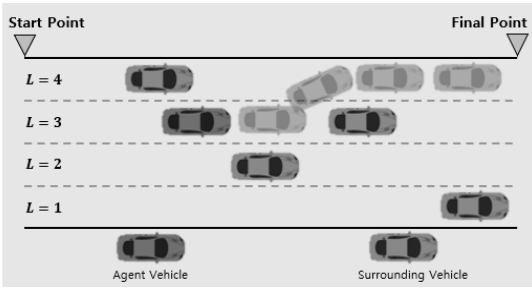

Fig. 1. Simulation in HighwayEnv environment

2.2 강화학습

강화학습은 최적의 의사결정을 탐색하기 위해 개발된 알고리즘으로 자율주행 환경이나 로봇 산업 분야와 같이 연속적인 의사결정이 필요한 환경에서 많이 사용되고 있다. 특히 심층 신경망을 이용한 기술들이 발전함에 따라 강화학습에 심층 신경망이 적용됨에 따라 성능 또한 크게 증가하고 있다. 대표적인 강화학습 알고리즘으로는 DQN(Deep Q-Network)[16]과 PPO(Proximal Policy Optimization)[17]이 있다.

먼저 DQN은 Q-learning에 심층 신경망을 적용한 모델로 기존의 Q-learning이 상태 공간과 행동 공간이 커지게 되면 모든 Q-value를 저장하기에 많은 메모리와 탐색시간을 필요하다는 단점을 심층 신경망을 통해 해결하기 위해 고안된 강화학습 모델이다. DQN은 CNN(Convolutional Neural Network), 경험 재생(experience replay) 그리고 타겟 네트워크(target network)로 구성되어 있으며 심층 신경망을 통해 입력받은 데이터를 통해 행동(action)을 출력하고 이를 통해 Q-value를 업데이트한다. 경험 재생은 에이전트가 과거에 경험한 샘플을 저장해두고 무작위로 선택하여 학습하는 방법으로 학습의 효율성을 높인다. 타겟 네트워크는 일정 간격으로만 업데이트가 되며 학습의 불안정성을 개선할 수 있다.

PPO는 에이전트가 환경과의 상호작용을 통해 데이터를 샘플링한 뒤 확률적 경사 상승법(stochastic gradient ascent)과 대리 목적 함수(surrogate objective function)의 최적화를 번갈아 수행하며 학습하는 모델이다. 특히 데이터를 샘플링할 때 기존의 방식과 달리 새로운 목적 함수를 사용하여 미니 배치의 업데이트를 여러 에포크에 걸쳐 가능하게 하여 안정적이고 효율적이며 다양한 상황에 적응할 수 있도록 한다는 장점이 있다. Fig. 2는 강화학습 모델이 적용된 에이전트가 환경과 상호작용하며 학습하는 과정을 나타낸다.

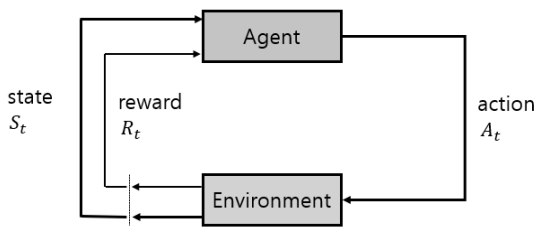

Fig. 2. Reinforcement learning process

2.3 적대적 공격

적대적 공격은 악의적인 공격자가 의도적으로 생성한 적대적 샘플을 이용하여 인공지능 모델의 오동작을 유도하는 공격 방식이다[18]. 적대적 샘플은 인간의 눈으로는 식별하기 어려운 작은 잡음을 의도적으로 입력 데이터에 추가하여 만들어진다. 이러한 적대적 샘플은 인공지능 모델이 잘못된 예측을 하도록 하거나 공격자가 원하는 특정 결과를 도출하도록 유도할 수 있다. Fig. 3은 딥러닝 모델에 적대적 공격을 수행하면 "팬더"가 "긴팔원숭이"로 오분류되는 것을

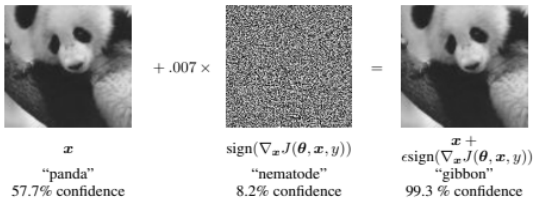

Fig. 3. Adversarial attack [18]

나타낸 것이다.

2.3.1 FGSM

Goodfellow 등은 입력 데이터에 약간의 잡음을 넣음으로써 심층 신경망이 오동작할 수 있는 FGSM(Fast Gradient Sign Method)을 제안하였다 [9]. FGSM은 계산이 간단하고 빠르기 때문에 적대적 샘플의 효과를 분석하고 모델의 취약성을 평가하는데 널리 사용된다. FGSM에서 적대적 샘플을 만드는 수식은 다음과 같다.

$$\acute{x} = x + \epsilon \cdot sign(\nabla L(x,y,\theta)) \quad (1)$$

수식 (1)에서 x는 원본 샘플을 나타내며 $\acute{x}$은 적대적 샘플을 나타낸다. y는 실제 라벨값을 나타내며 θ는 모델의 파라미터이다. x와 y를 통해 손실 함수 L을 이용하여 섭동을 생성할 수 있으며 섭동의 크기를 조절하는 ϵ를 통해 공격의 강도를 설정할 수 있다.

2.3.2 BIM

BIM(Basic Iterative Mehtod)는 FGSM에서 보다 높은 공격 성공률을 달성하기 위해 FGSM에 반복성을 추가하여 좀 더 강력한 적대적 샘플을 생성하는 알고리즘이다 [10]. 적대적 샘플을 생성할 때 각 단계에서의 섭동은 최대 반복 횟수 T에서 step size를 나타내는 $a = \epsilon / T$로 제한하여 더 세밀하게 섭동을 조정할 수 있다. BIM에서 적대적 샘플을 만드는 수식은 다음과 같다.

$$x_{t+1} = Clip_{[-\epsilon,\epsilon]} x_t + a \cdot sign(\nabla L(x_t,y,\theta)) \quad (2)$$

수식 (2)에서 a는 반복마다의 변조량을 나타내며 $Clip$은 변조량 a의 범위를 ϵ만큼으로 조정하는 함수이다.

2.3.3 PGD

PGD(Projected Gradient Descent)는 적대적 섭동을 생성할 때 step만큼 공격을 반복하여 정해진 ϵ범위에서 섭동의 최대화를 수행한다 [11]. BIM과의 차이는 PGD는 랜덤 위치에서 시작점을 초기화하여 특정 지역 최적점에 갇히는 현상을 방지함으로써 더 강력한 공격을 만들어 낼 수 있다. PGD에서 사용한 수식은 다음과 같다.

$$x^{t+1} = \Pi_{x+S}(x^t + e \times sign(\nabla_x J(\theta,x,y))) \quad (3)$$

수식 (3)에서 손실함수를 나타내는 J를 통해 입력 이미지 x와 타겟 클래스 y를 이용해 기울기(gradient)를 계산해 공격 강도를 나타내는 e를 곱해줌으로써 최종적인 적대적 샘플을 생성하게 된다.

2.3.4 CW

CW(Carlinin and Wagner)은 L_0, L_1 그리고 L_∞를 이용하여 적대적 데이터와 입력 데이터의 유사성을 측정한 뒤 반복적으로 잡음을 입력 데이터에 추가하여 최종적인 적대적 데이터를 생성한다 [12]. CW를 통해 적대적 데이터를 생성하는 수식은 다음과 같다.

$$\begin{aligned} &Minmize\, D(x, x+\delta) + c \cdot f(x+\delta) \\ &subjet\ to\ x+\delta \in [0,1]^n \end{aligned} \quad (4)$$

c는 공격 강도를 나타내며 f는 손실함수로 인공지능 모델이 입력 데이터에 대해 틀리게 예측하였을 때 0 이하의 값을 반환한다. 입력 데이터 x와 섭동 δ를 추가하여 생성한 적대적 데이터 $\acute{x}$를 L 규범을 이용하여 $D(x, x+\delta)$를 최소화하여 최종적인 적대적 데이터를 생성한다.

2.4 양방향 필터

양방향 필터 [13-15]는 객체의 윤곽선을 보존하면서 잡음을 제거할 수 있는 비선형 필터이다. 양방향 필터는 가우시안 분포를 이용하여 중심 픽셀에서 거리가 가까운 픽셀의 가중 평균과 중심 픽셀과 인접 픽셀값의 차이를 반영한 가우시안 형태의 필터값을

사용하여 잡음을 제거한다. 양방향 필터를 나타내는 수식은 다음과 같다.

$$BF[I]_p = \frac{1}{W_p}\sum_{q\in S} G_{\sigma_s}(\|p-q\|) \cdot G_{\sigma_\gamma}(|I_p - I_q|) \cdot I_q \quad (5)$$

여기서 W_p는 정규화 인수를 나타내며 수식 (6)와 같다.

$$W_p = \sum_{q\in S} G_{\sigma_s}(\|p-q\|) \cdot G_{\sigma_\gamma}(I_p - I_q) \quad (6)$$

수식 (6)에서 G_{σ_s}는 일반적인 가우시안 필터를 나타내며 픽셀값의 차이에 대한 필터 계수 G_{σ_γ}은 중심 픽셀과 인접 픽셀의 차이가 큰 윤곽선에 대해서는 작은 필터값을 가져 윤곽선을 보존하며, 비슷한 픽셀에서는 큰 필터값을 가져 가우시안 필터가 수행되도록 조절하는 값을 나타낸다. Fig. 4는 양방향 필터를 이용하여 이미지의 잡음을 제거한 그림을 비교한 것이다.

Original image Bilateral filter image

Fig. 4. Image applied with bilateral filter

2.5 적대적 공격 대응 기법

적대적 공격에 대응하기 위한 대표적인 방법으로 적대적 학습(adversarial training)과 인공지능을 이용한 잡음 제거(denoising) 방법이 있다. 적대적 학습은 적대적 샘플을 생성하여 데이터 세트를 구성한 뒤 딥러닝 모델을 학습 하기 때문에 적대적 공격에 강건하도록 학습이 되지만 기존에 학습된 딥러닝 모델보다 원본 이미지에 대한 성능이 감소 된다는 단점이 있다. 잡음 제거 방법으로는 오토인코더를 사용하는 방법이 있다. 오토인코더를 이용하여 적대적 샘플과 원본 이미지와의 거리 차이를 줄여 재구성된 이미지를 생성하여 딥러닝 모델이 올바르게 동작하도록 하는 방법이다. 그러나 이러한 방법은 딥러닝 모델을 추가로 사용해 비용을 추가로 요구한다는 단점이 있다.

최근 Feng 등 [7]은 기존의 방어 방법이 사전에 학습된 딥러닝 모델의 성능을 저하시킬 수 있다는 단점을 언급하며 가짜 기울기(fake gradient) 방법을 제안하였다. 딥러닝 모델의 출력 레이어에 가짜 클래스를 통해 가짜 기울기를 생성하여 실제 출력 정보를 숨기고 적대적 공격으로부터 방어하였다. 악의적인 공격자가 적대적 공격을 위해 획득한 기울기 값이 딥러닝 모델의 작동을 방해하지 않도록 가짜 노드들을 추가하게 되는데 가짜 노드와 진짜 노드는 랜덤 키를 이용하여 셔플링을 수행하게 된다. 그렇기 때문에 가짜 기울기 방법을 통해 공격자를 효과적으로 속여 적대적 공격을 방어할 수 있다. Fig. 5는 가짜 기울기 기법의 아키텍처를 나타낸다.

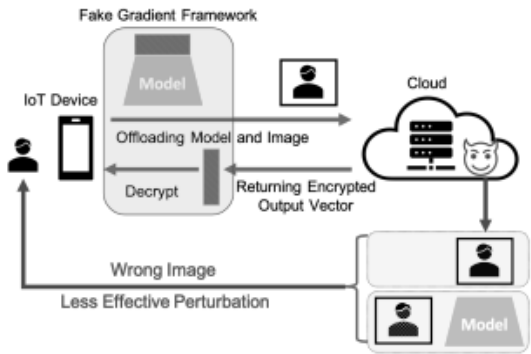

Fig. 5. Fake gradient system architecture [7]

III. 양방향 필터를 이용한 적대적 공격 대응

본 논문에서는 적대적 공격을 대응하기 위해 강화학습 모델에 공격이 수행되기 전에 양방향 필터를 이용하여 잡음을 제거하였다. 양방향 필터는 객체의 경계선을 유지하면서 이미지의 잡음을 제거하는 비선형 가우시안 필터이다. 먼저 적대적 샘플을 양방향 필터를 이용하여 잡음을 제거하는데 사용된 필터의 수식은 다음과 같다.

$$x^{adv}_{filtered}(p) = \frac{1}{W_p}\sum_{q\in S} G_{\sigma_s}(\|p-q\|) \cdot G_{\sigma_\gamma}(|x^{adv}_p - x^{adv}_q|) \cdot x^{adv}_q \quad (7)$$

여기서 x^{adv}는 적대적 샘플을 나타내며 S는 윈도우(window) p는 현재 픽셀을 나타낸다. q는 인접 픽셀을 나타내며 x^{adv}는 공간적 거리에 따른 가중치를 계산하는 함수, G_{σ_s}는 수식 (8)과 같이 정의된다.

$$G_{\sigma_s}(c) = \prod_{s \in x,y,z} \exp(-\frac{c_s^2}{2\sigma_s^2}) \quad (8)$$

G_{σ_s}는 공간적 거리와 관련된 가우시안 필터로 적대적 샘플 x^{adv}의 픽셀 위치 p와 q 사이의 공간적 거리로부터 계산된 가중치 σ_s를 이용하여 픽셀들을 필터링 하게 된다. σ_s의 값이 커질수록 더 넓은 범위의 픽셀 색상 범위를 고려할 수 있어 잡음 제거 효과가 강해진다는 장점이 있다.

픽셀의 밝기나 색상의 유사성에 대한 가중치를 계산하는 함수 G_{σ_r}의 수식은 다음과 같다.

$$G_{\sigma_r}(c) := \exp(-\frac{c^2}{2\sigma_r^2}) \quad (9)$$

G_{σ_r}는 픽셀의 유사성과 관련된 가우시안 필터로 x^{adv}의 밝기 또는 색상 차이를 보존하면서 잡음을 제거할 수 있게 되며 이미지의 경계 부분에서 픽셀의 값 변화를 부드럽게 스무딩(smoothing)할 수 있다. 이를 통해 이미지에서 구조적 특성은 유지하면서 경계 수준을 명확하게 하면서 잡음을 제거할 수 있게 된다. σ_r의 값이 커질수록 유사한 픽셀들 사이의 연결을 강화할 수 있어 이미지의 경계 부분을 더 잘 보존할 수 있게 된다.

다음 Fig. 6은 강화학습 모델에 대한 적대적 공격이 진행되는 과정과 양방향 필터를 이용한 방어 방법을 나타낸 것이다.

만약 전방에 차량이 존재하여 속도를 유지해야 하는 상황에서 입력되는 이미지에 적대적 공격을 적용하면 “가속”으로 강화학습 모델이 잘못된 예측을 하도록 유도할 수 있으나 대응 기법인 양방향 필터를 적용하였을 경우 적대적 섭동이 추가되더라도 “속도 유지”를 예측하는 것을 나타낸 것이다.

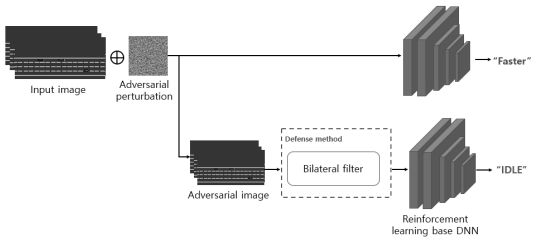

Fig. 6. Adversarial attack and countermeasure on autonomous driving

IV. 적대적 공격 실험 및 대응 방안

본 논문에서는 강화학습 모델인 DQN과 PPO를 대상으로 적대적 공격을 수행한 뒤 성능을 평가하였다. 이를 위해 다음과 같은 실험 환경을 구축한 뒤 각 공격 알고리즘에 따른 강화학습 모델의 성능을 분석하였다.

4.1 실험 환경

1) 고속도로 주행 환경에서의 행동 결정 시스템

본 논문에서는 자율주행 자동차의 행동 결정 시스템을 구현하기 위해 DQN과 PPO을 사용하였고 심층 신경망으로 CNN을 사용하였다. 강화학습 모델에 입력되는 입력 데이터는 128×64의 화면을 캡쳐한 회색 화면의 이미지 4장이 한 번에 입력이 된다. Table 1과 Table 2는 DQN과 PPO 모델의 파라미터 값을 나타낸 것이다.

Table 1. Parameters of the DQN model

Layer	Nodes	Filters	Padding	Activation function
Input	128x64	-		
Conv1	4	32	Same	ReLU
Conv2	32	64	Same	ReLU
Conv3	64	64	Same	ReLU
Flatten	-			
Dense1	3072	-		ReLU
Dense2	512	-		ReLU
output	5	-		

Table 2. Parameters of the PPO model

Layer	Nodes	Filters	Padding	Activation function
Input	128x64	-		
Conv1	4	32	Same	ReLU
Conv2	32	64	Same	ReLU
Conv3	64	64	Same	ReLU
Flatten	-			
Dense1	3072	-		ReLU
Dense2	512	-		ReLU
Dense3	256	-		ReLU
Dense4	256	-		ReLU
output	5	-		

HighwayEnv 시뮬레이션은 고속도로 환경을 구성하였고 차선은 총 4차선으로 구성하였다. 환경에서 주변 차량은 총 20대로 구성하였고 강화학습 모델들이 예측할 수 있는 행동은 "좌측 차선 변경, "속도 유지", "우측 차선 변경", "가속", "감속"으로 지정하였다.

2) 적대적 공격

본 논문에서는 적대적 공격 알고리즘으로 FGSM, PGD, BIM 그리고 CW를 사용하였다. PGD와 BIM에서 공격 강도를 나타내는 epsilon은 0.008와 0.015, CW의 공격 강도를 나타내는 c는 0.8과 1.5로 설정하여 실험을 수행하였다. epsilon 과 c 값이 커질수록 공격 성공률은 높아지지만 잡음이 커지기 때문에 육안으로 식별될 수 있다는 단점이 있다. PGD와 BIM의 반복 횟수는 10회, CW는 50회 반복하여 최종적인 적대적 섭동을 만들었다.

3) 양방향 필터

양방향 필터의 파라미터 중 중심 픽셀에서 고려되는 이웃 픽셀들의 영역을 나타내는 윈도우 S는 3으로 설정하였다. 이를 통해 적대적 공격의 국소적인 잡음을 제거할 수 있다. 픽셀 간의 색상 차이에 따른 필터링 민감도를 나타내는 γ는 20으로 설정하였다. 색상 차이가 γ 값 이내인 픽셀들만 필터링을 수행하여 적대적 공격을 통해 발생하는 미세한 픽셀의 색상 변화를 감지하여 효과적으로 제거할 수 있다.

4) 성능 평가 방식

HigwayEnv에서 동작하는 DQN과 PPO의 성능 평가와 양방향 필터를 이용한 적대적 공격 방어 성능을 평가하기 위해 HigwayEnv에서 100개의 에피소드를 통해 에피소드의 수행 길이(episode during) 평균과 대상 차량이 획득한 보상(episode reward)의 평균을 이용하여 평가를 진행하였다.

먼저 에피소드의 수행 길이 평균은 대상 차량이 고속도로 주행 환경에서 얼마나 오랫동안 주행했는지 나타내는 지표이다. 보상의 평균은 Table 3에서 확인할 수 있듯이 HigwayEnv에서 사용된 보상구조를 이용하였다. 대상 차량이 최고속도로 주행하거나 고속도로의 가장 오른쪽 차선을 주행하였을 때 보상을 받으며 주위 차량과 충돌했을 때 페널티를 받도록 구성하였다.

Table 3. Reward of HighwayEnv

Reward	Content
collision reward (default: -1)	The reward received when colliding with a vehicle
right lane reward (default: 0.1)	The reward received when driving on the right most lane
high speed reward (default: 0.4	The reward received when driving at full speed
lane change reward (default: 0)	The reward received at each lane change action

본 논문에서는 총 100개의 에피소드를 구성하여 각 에피소드마다 에이전트가 획득한 보상과 환경을 수행한 길이를 계산한 뒤 평균을 계산하여 성능을 평가하였다.

에피소드 수행 길이 평균을 나타내는 수식은 다음과 같다.

$$\bar{D}=\frac{1}{N}\sum_{i=1}^{N}D_i \tag{10}$$

D_i는 에피소드 i에서의 수행 길이를 나타내며 N은 총 에피소드 수를 나타낸다.

에피소드 보상 평균의 수식 (11)과 같다. R_i는 에피소드 i에서 에이전트가 획득한 총 보상을 나타낸다.

$$\bar{R}=\frac{1}{N}\sum_{i=1}^{N}R_i \tag{11}$$

4.2 강화학습 모델에서의 적대적 공격

본 논문에서는 먼저 강화학습 모델의 성능을 감소시켜 오동작을 유도하는 적대적 공격을 수행해 보고자 한다. 먼저 FGSM, BIM, 그리고 PGD를 사용하여 기존 강화학습 모델인 DQN과 PPO에 적대적 공격을 수행하였다. 다음 Table 4는 적대적 공격이 없는 DQN과 PPO의 성능을 나타낸 것으로 두 모델 모두 에피소드의 수행 길이와 보상 모두 평균 96%을 달성하였다.

다음 Table 5는 동일한 환경에서 각 강화학습 모

델에 대한 적대적 공격을 수행한 결과를 나타낸 것으로 공격 방법 중에서 epsilon을 0.008로 적용해서 DQN 모델을 공격했을 때는 에피소드 수행 길이가 23%, 보상 평균이 20.9%까지 감소하는 것을 확인할 수 있었다. epsilon을 0.008에서 세 가지 공격 방법 중 BIM 공격이 우수하였다. PPO 모델을 공격했을 때는 에피소드 수행 길이가 37.9%, 보상 평균이 36.3%까지 감소하는 것을 통해 PGD 공격이 가장 우수한 성능을 보인 것을 확인하였다.

또한 epsilon을 0.015에서와 같이 공격의 강도가 큰 경우 DQN 모델에서는 FGSM 공격에서 에피소드 수행 길이가 14.4%, 보상 평균이 12.8%까지 감소하는 것을 확인하였다. PPO 모델에서는 BIM 공격이 에피소드 수행 길이가 18.7%, 보상 평균이 17.2%를 보여 BIM 공격이 우수하였다.

CW 공격의 경우 c를 0.8로 지정하였을 때 DQN의 에피소드 수행 실이 평균은 46.9%, 보상 평균은 43%를 달성하였다. PPO의 경우 에피소드 수행 길이 평균이 37.4%를 달성하였고 보상 평균은 35.5%를 확인하였다. c가 1.5일 때 DQN의 에피소드 수행 길이 평균은 26.3%, 보상 평균은 24%를 확인하였고 PPO는 각각 19.2%, 18.2%를 달성하였다.

Table 4. Evaluation of reinforcement learning model

Model	During	Reward
DQN	96.86%	96.05%
PPO	96.72%	96.55%

Table 5. Evaluation results of adversarial attacks

Epsilon		0.008		0.015	
		During	Reward	During	Reward
FGSM	DQN	34.4%	31.8%	14.4%	12.8%
	PPO	43.5%	42.3%	21.8%	20.6%
BIM	DQN	23.0%	20.9%	18.7%	14.4%
	PPO	38.3%	36.6%	18.7%	17.2%
PGD	DQN	27.1%	23.9%	16.6%	12.8%
	PPO	37.9%	36.3%	20.1%	18.5%
C		0.8		1.5	
		During	Reward	During	Reward
CW	DQN	46.9%	43.0%	26.3%	24.0%
	PPO	37.4%	35.5%	19.2%	18.2%

4.3 대응 방법에 대한 성능 분석

본 논문에서 제안하는 양방향 필터를 이용한 필터링 기법을 이용하여 강화학습 모델의 성능을 분석하였다. 먼저 Table 6은 기존에 제시되었던 적대적 공격에 대한 대응 방법인 가짜 기울기(fake gradient) 기법과 비교 평가한 것이다. 양방향 필터를 적용한 DQN은 가짜 기울기 기법과 비교하였을 때 epsilon을 0.008로 설정한 BIM 공격에서 약 6% 성능이 향상된 것을 확인할 수 있었고 PPO는 PGD 공격에서 약 2% 성능이 향상된 것을 확인할 수 있다.

epsilon을 0.015로 설정한 FGSM에서 양방향 필터를 적용한 DQN은 가짜 기울기 기법보다 4%의 성능 증가를 확인하였고 PPO는 BIM에서 가짜 기울

Table 6. Performance evaluation of counter-measures

Epsilon		0.008		0.015	
		During	Reward	During	Reward
FGSM	DQN (proposed method)	94.4%	91.9%	83.3%	80.0%
	DQN (Fake gradient)	91.5%	89.6%	79.5%	77.4%
	PPO (proposed method)	91.9%	91.4%	82.7%	82.0%
	PPO (Fake gradient)	92.9%	92.5%	83.9%	83.3%
BIM	DQN (proposed method)	96.0%	93.9%	88.3%	85.6%
	DQN (Fake gradient)	90.6%	87.6%	81.5%	71.5%
	PPO (proposed method)	95.7%	95.5%	88.4%	87.8%
	PPO (Fake gradient)	92.5%	92.1%	83.8%	83.3%
PGD	DQN (proposed method)	94.1%	91.8%	91.7%	89.0%
	DQN (Fake gradient)	89.7%	85.9%	88.6%	75.0%
	PPO (proposed method)	94.9%	94.7%	93.7%	93.4%
	PPO (Fake gradient)	92.7%	92.3%	83.8%	83.3%
c		0.8		1.5	
		During	Reward	During	Reward
CW	DQN (proposed method)	97.2%	95.6%	96.4%	94.3%
	DQN (Fake gradient)	89.6%	87.8%	88.5%	86.5%
	PPO (proposed method)	96.0%	95.6%	95.9%	95.5%
	PPO (Fake gradient)	93.1%	92.8%	93.0%	92.7%

기 기법보다 5% 성능이 증가한 것을 확인하였다. CW를 이용한 공격에서 양방향 필터를 적용한 강화학습 모델들은 가짜 기울기 기법보다 높은 성능을 달성한 것을 확인하였다.

본 논문에서는 마지막으로 적대적 공격이 가해지지 않더라도 원본 이미지에 양방향 필터를 적용하면 강화학습 모델이 올바르게 주행할 수 있는지 확인하였다. 실험 결과 Table 7에서 확인할 수 있듯이 두 모델 모두 에피소드 수행 길이와 보상 평균이 95%이상을 보여 큰 성능의 저하 없이 주행할 수 있는 것을 확인하였다.

Fig. 7은 동일한 환경에서 t-SNE [19]을 이용하여 DQN 모델에서 입력되는 데이터에 대해 특징을 2차원 공간으로 시각화한 그림이다. t-SNE는 고차원 데이터를 저차원으로 표현할 수 있는 비선형 차원 축소 기법이다.

딥러닝 모델이 가지는 레이어의 특징은 고차원 영역의 데이터로 특징을 직관적으로 이해하기 어렵다는 단점이 있다. 이를 보완하기 위해 t-SNE를 사용하여 고차원 데이터를 저차원 데이터로 표현한 뒤 데이터에 나타낼 수 있는 군집으로 나타낼 수 있다. 이를 통해 데이터의 패턴이나 구조를 쉽게 나타낼 수 있으며 유사한 데이터들이 어떻게 군집화를 이루는지 직관적으로 이해할 수 있다.

먼저 Fig. 7의 (A)는 원본 데이터에 대해서 같은 클래스마다 군집화를 이루고 있는 것을 확인할 수 있었다. 하지만 (B)에서 BIM 알고리즘을 통해 생성된 적대적 샘플은 t-SNE 결과로 군집화가 정상적으로 이루어지지 않는 것을 확인하였다. 이는 적대적 공격을 통해 DQN이 정상적으로 입력 데이터를 정상적으로 예측하지 못한 것을 나타낸다.

Fig. 7의 (C)와 (D)는 각각 양방향 필터와 가짜 기울기 기법을 적대적 샘플에 적용한 후 저차원 데이터로 표현한 것이다. Fig. 7에서 (C)는 양방향 필터를 적용하였을 경우를 나타내며 일부 클래스를 제외

Table 7. Performance evaluation without adversarial attacks

Defense Method	DQN		PPO	
	During	Reward	During	Reward
Bilateral filter	96.8%	95.4%	96.1%	95.9%
Fake gradient	96.8%	96.0%	96.7%	96.5%

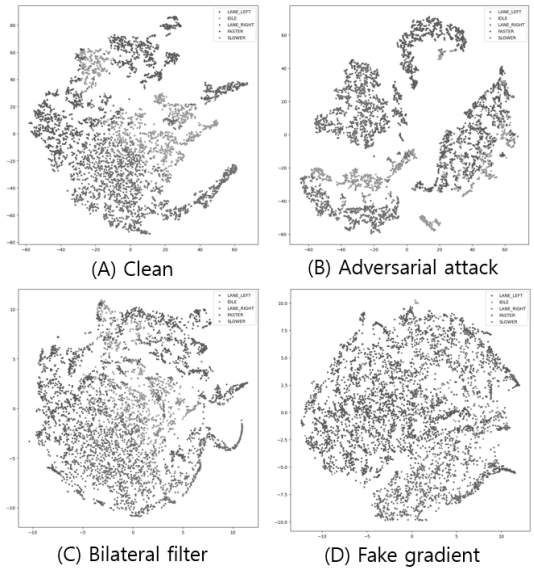

Fig. 7. 2-dimensional t-SNE of HighwayEnv data from DQN

하고 대부분의 클래스는 군집화를 이룬 것을 알 수 있다. 이를 통해 양방향 필터를 통해 잡음을 제거하여 적대적 공격을 대응할 수 있음을 확인하였다. 마지막으로 (D)는 가짜 기울기 기법을 적용한 결과로 클래스에 대한 군집화가 분산되어 있는 것을 확인할 수 있다. 이는 적대적 샘플에 대한 잡음을 완벽하게 제거하지 못했다는 것을 의미한다.

V. 결 론

최근 자율주행 자동차, 스마트 팩토리, 홈 네트워크에서 강화학습을 이용한 딥러닝 기술들이 많이 사용되고 있다. 반면 악의적인 공격자가 이미지에 잡음을 추가하여 딥러닝 기술이 적용된 강화학습의 예측을 어렵게 함으로써 오동작을 유발하는 적대적 공격에 취약하다는 단점이 있다.

본 논문에서는 고속도로 시뮬레이션 환경인 HighwayEnv에서 강화학습 모델 DQN과 PPO를 구현한 뒤 잡음을 추가하는 방식으로 적대적 공격을 수행하는 실험을 수행하였다. 실험 결과 적대적 공격 기법들을 통해 강화학습 모델이 적용된 에이전트는 올바르게 주행할 수 없는 것을 확인하였다. 또한 적대적 공격을 대응하기 위해 양방향 필터를 적용하는 방법을 강화학습에 적용하여 에이전트 차량이 정상적으로 주행하는지 실험을 통해 확인하였고 t-SNE를

통하여 분석하였다. 실험 결과, 제안하는 양방향 필터를 이용하여 적대적 공격이 수행되더라도 충분히 공격에 대응하게 되어 정상적인 주행을 할 수 있음을 확인하였다.

References

[1] R. S. Sutton and A. G. Barto. "Reinforcement Learning: An Introduction," 2nd ed. MIT Press, 2018.

[2] Y. C. Lin, Z.W. Hong, Y. H. Liao, M. L. Shih, M. Y. Liu, and M. Sun, "Tactics of adversarial attack on deep reinforcement learning agents," in Proc. Int. Joint Conf. Artif. Intell. (IJCAI), 2017, pp. 3756 - 3762

[3] S. S. Zheng, Y. Song, T. Leung and I. Goodfellow, "Improving the robustness of deep neural networks via stability training," IEEE Conference on Computer Vision and Pattern Recognition(CVPR'16), pp. 4480-4488, 2016.

[4] W. Zhao, S. Alwidian and Q. H. Mahmoud, "Adversarial Training Methods for Deep Learning: A Systematic Review," Journal of Algorithms, Vol. 15, Issue 8(283), 2022.

[5] W. Xu, D. Evans, and Y. Qi, "Feature squeezing:Detecting adversarial examples in deep neural networks," Network and Distributed System-Security Symposium(NDSS'18), 2018.

[6] S. L. Yin., X. L. Zhang, and L. Y. Zuo, "Defending against adversarial attacks using spherical sampling-based variational auto-ender", Neurocomputing, pp. 1-10, 2022.

[7] X. Feng, Y. Xie. , M. Ye, X. Tang, and B. Yuan. "Fake Gradient: A Security and Privacy Protection Framework for CNN-based Image Classification," In Proc ACM, pp. 5510 - 5518, 2021.

[8] Leurent, E.: An environment for autonomous driving decisionmaking,: https://github.com/ eleurent/highway-env, last accessed 2018.

[9] I. J. Goodfellow, J. Shelns, and C. zegedy, "Explaining and harnessing dversarial examples," In International Conference on Learning Representations, pp. 1-11, 2015.

[10] Kurakin, A., Goodfellow. I., and Bengio. S.: Adversarial examples in the physical world. In: Int. Conf. Learning Representations, pp. 1-14, 2018.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," In International Conference on Learning Representations, 2018.

[12] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," In IEEE symposium on scurity and privacy, pp. 39-57, 2017.

[13] S. Paris, F. Durand, "A fast approximation of the bilateral filter using a signal processing approach", International Journal of Computer Vision, Vol. 81, No. 1, pp. 24-52, Jan 2009.

[14] Xu. Long and N. H. Kim, "An Improved Adaptive Median Filter for Impulse Noise Removal", Journal of KIICE, Vol. 17, No. 4, pp. 989-995, April 2013.

[15] M. Elad, "On the origin of the bilateral filter and ways to improve it," IEEE Trans. Image Processing, Vol. 11, No. 10, pp. 1141-1151, October 2002.

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S.

Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Humanlevel control through deep reinforcement learning," Nature, vol. 581, no. 7549, pp. 529 - 533, 2015.

〔17〕 J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," Jul. 2017.

〔18〕 I. J. Goodfellow, J. Shlens, and C. Szegedy, " Explaining and harnessing adversarial examples," in Proc ICLR'15, pp. 1-11, 2015.

〔19〕 M. Laurens and H. Geoffrey, "Visualizing Data using t-SNE," Journal of Machine Learning Research pp. 2579 - 2605, 2018.

〈저자소개〉

이 승 열 (Seungyeol Lee) 학생회원
2024년 2월 : 호서대학교 컴퓨터공학부 졸업
2024년 3월~현재: 호서대학교 대학원 정보보호학과 석사과정
〈관심분야〉 자동차 보안, 양자내성 암호, 인공지능 보안

하 재 철 (Jaecheol Ha) 종신회원
1989년 2월: 경북대학교 전자공학과 학사
1993년 8월: 경북대학교 전자공학과 석사
1998년 2월: 경북대학교 전자공학과 박사
1998년 3월~2007년 2월: 나사렛대학교 정보통신학과 교수
2007년 3월~현재: 호서대학교 컴퓨터공학부 교수
2009년 1월~현재: 한국산학기술학회 이사
2022년 1월~현재: 국제차세대융합기술학회 부회장
1993년 1월~현재: 한국정보보호학회 회장
〈관심분야〉 암호학, 부채널 공격, 네트워크 보안, 정보보호