인공지능 기반 자율주행 네트워크의 적대적 공격 방법

석지혜, 정은수, 최소빈, 백예진, 이훈
부경대학교 정보통신공학과
{sjh010404, ef613, sobin0401, byj848}@pukyong.ac.kr, hlee@pknu.ac.kr

Adversarial Attack Methods for AI-Based Autonomous Driving Networks

Jihye Seok, Ensue Jeong, Sobin Choi, Yejin Baek and Hoon Lee
Dept. Information and Communications Eng., Pukyong National University

요 약

인공지능 기술은 자율주행 시스템 구축에 없어서는 안 될 필수요인이다. 자율주행 시장 규모가 급성장함에 따라, 자율주행 목적으로 훈련된 인공지능 모델을 외부에서 공격하는 인공지능 보안 위협이 증가하고 있다. 인공지능의 오판으로 인한 차량의 오주행은 큰 인명피해를 유발할 수 있으므로, 모든 가능한 적대적 공격 시나리오의 방비책을 마련할 필요가 있다. 본 논문에서는 자율주행 인공지능 모델의 적대적 공격 기법을 개발하여 근미래에 발생할 수 있는 보안 위협을 대비한다. 적대적 공격자가 자율주행 네트워크의 통신 채널을 무단으로 도청하고, 잡음을 송신하여 인공지능 모델에 spoofing 공격을 시도하는 상황을 상정한다. 공격자가 발생시키는 잡음은 인공지능 모델의 판단에 오류를 유발하도록 설계한다. 제안하는 인공지능 적대적 공격 기법을 자율주행 테스트베드에 구현하여 그 현실성을 파악한다. 모의실험 결과, 높은 정확도로 훈련된 인공지능 모델에 매우 작은 크기의 적대적 잡음 공격을 실시하여 그 정확도를 15% 미만으로 하락시킬 수 있음을 확인한다.

Ⅰ. 서 론

자동차 제조업체와 IT 기업들이 자율주행차 개발에 박차를 가하며, 2025년에는 자율주행차 시장 규모가 260억 달러에 달하기까지 성장할 것으로 예측된다[1]. 저비용/저전력 운용을 위해 자율주행차량의 인공지능 연산 대부분을 네트워크 에지 서버가 대신 수행하는 모바일 에지 컴퓨팅 구조가 자율주행 네트워크의 필수 요소로 대두되고 있다. 하지만, 차량과 서버의 통신 채널이 공격자에 의해 감청되는 경우, 인공지능 모델의 판단에 오류가 발생하여 인명사고를 유발할 수 있다.

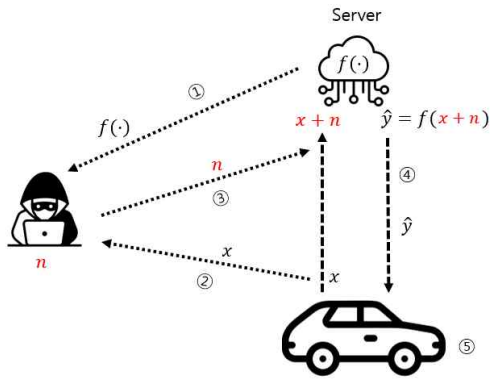
인공지능 기반의 자율주행 네트워크에서 발생할 수 있는 보안 위협에 미리 대비하기 위해, 본 논문에서는 인공지능 모델의 적대적 공격 기법을 고안한다. 서버가 적대적 공격을 탐지하지 못하도록 매우 미세한 공격 잡음을 통신 채널로 전송하여 인공지능 모델의 오판확률을 증가시킨다. 이를 위해 Fast Gradient Sign Method (FGSM) 알고리즘을 활용하여 자율주행차량이 실시간으로 촬영하는 주행 영상에 최악의 적대적 공격 잡음을 생성하는 기법을 제안한다. 제안하는 적대적 공격 기법을 자율주행 테스트베드로 구현하여 공격 성능을 확인한다.

Ⅱ. 제안하는 자율주행 인공지능 적대적 공격 기법

본 절에서는 자율주행 네트워크의 적대적 공격 시나리오를 소개하고, 인공지능의 오판률을 증가시키기 위한 FGSM 공격 기법을 제안한다.

2-1. 자율주행 네트워크 적대적 공격 시나리오

[그림 1]과 같은 인공지능 자율주행 네트워크의 해킹 공격 모델을 제안한다[2]. ① 공격자는 에지 서버에 저장된 자율주행 인공지능 모델 $f(\cdot)$를 미리 해킹하여 알고있다. 에지 서버의 인공지능 모델로부터 주행 명령을 받기 위해, 자율주행차량은 먼저 자신이 수집한 주행 영상 x를 무선 채널을 통해 서버로 전송한다. ② 공격자는 차량-서버 채널을 도청하여 주행 영상 x를 획득한다. ③ 인공지능 모델 $f(\cdot)$와 주행 영상 x를 기반으로 공격자는 공격 잡음 n을 생성하여 서버로 전송하는 spoofing 공격을 시도한다. ④ 서버는 공격받은 영상 $x+n$을 기반으로 모델 추론을 실시하여 추론 결과 $\hat{y}$을 차량에게 전송한다. ⑤ 차량은 수신한 추론 결과를 기반으로 주행한다.



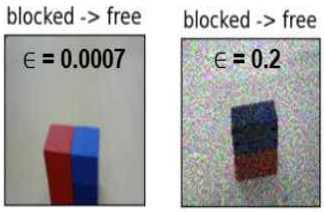
[그림 1] 해킹 공격 시나리오

2-2. FGSM 기반 인공지능 공격 기법

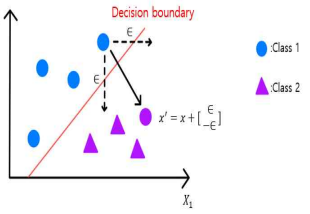
본 논문에서 제안하는 인공지능 공격을 성공적으로 수행하기 위해서는 공격 잡음 n을 적절히 생성해야 한다. 이를 위해 Fast Gradient Sign Method (FGSM) 기법을 활용한다[2]. 설계 목표는 자율주행 인공지능 모델 $f(\cdot)$의 훈련 손실 함수 $J(\theta,x,y)$를 증가시키는 잡음을 발생하는 것이다. 이때 θ는 모델의 파라미터, y는 target label을 의미한다. 공격 잡음 n은 모델 입력 x에 더해진다. 따라서, 손실 함수를 입력 x에 대해 미분한 gradient $\nabla_x J(\theta,x,y)$가 증가하는 방향으로 공격 잡음 n을 다음과 같이 설정한다.

$$n = \nabla_x J(\theta,x,y)$$

잡음 n이 손실함수가 증가되는 방향으로 설계되었으므로, 적대적 공격이 발생한 입력 영상 신호 $x+n$을 자율주행 인공지능 모델 $f(\cdot)$를 통해 추론하면 false-positive 및 false-negative 확률을 극대화 할 수 있다.



[그림 2] 공격받은 영상 예시



[그림 3] Max-norm 제한 잡음 공격

일반적으로 잡음의 세기가 증가하면 입력 x에 더 큰 변화를 주어 인공지능 모델의 성능을 하락시킬 수 있다. 하지만, 반대로 잡음이 지나치게 큰 경우 공격 상황이 탄로날 수 있다. [그림 2]에 잡음의 크기에 따른 공격 영상 $x+n$을 표현하였다. 여기서 ϵ는 잡음의 크기를 결정하는 hyperparameter이다. 잡음의 크기가 작으면 원본 영상 x와 육안으로 구분이 안 될 정도로 정교한 공격이 가능하다. 하지만 잡음의 세기가 큰 경우에는 단순한 영상처리 기법으로도 해킹 공격의 여부를 판단할 수 있다. 따라서, 잡음의 크기를 적절히 작게 유지하면서도 해킹 성능을 극대화할 수 있는 효과적인 방법이 필요하다.

이를 위해 잡음의 세기를 벡터 norm으로 측정하고 그 크기를 제한하는 방법을 사용한다. 다양한 벡터 norm 중, 가장 큰 원소의 절대값을 지칭하는 max-norm $\|\cdot\|_\infty$ 연산을 활용한다. 잡음의 max-norm $\|n\|_\infty$이 원하는 크기 ϵ보다 작아지도록 제약조건 $\|n\|_\infty \leq \epsilon$을 추가한다. FGSM 기법에 따라 max-norm 제한 잡음은 다음과 같이 계산된다.

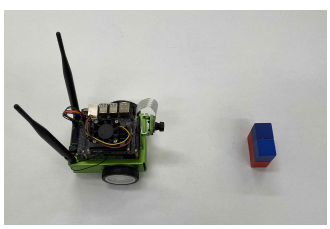
$$n = \epsilon sign(\nabla_x J(\theta, x, y))$$

이때 $sign(\cdot)$은 부호함수이다. [그림 3]에 max-norm 제한 잡음 공격의 원리를 나타내었다. Max-norm 제한으로 잡음 벡터의 원소는 최소 $-\epsilon$, 최대 ϵ 값을 갖는다. 이미 훈련된 인공지능 모델 $f(\cdot)$은 최적화된 decision boundary를 기준으로 입력 x의 class를 구분한다. 이때 decision boundary를 넘어가는 잡음이 함께 입력되면 ϵ만큼의 매우 작은 변화로도 오판단 결과를 도출할 수 있다.

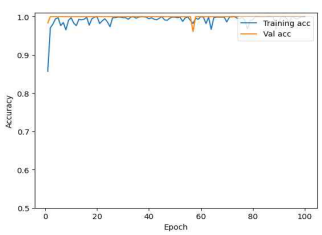
Ⅲ. 모의실험 결과

본 절에서는 해킹 공격 시스템 구현을 위한 자율주행 테스트베드를 구현하고, 공격 모형의 모의실험 결과를 분석한다.

3-1. 자율주행 인공지능 훈련 결과



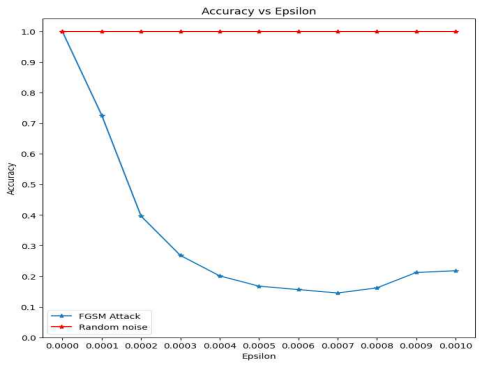
[그림 4] 자율주행 테스트베드



[그림 5] 정확도 성능

자율주행 시스템을 구현하기 위해 [그림 4]와 같이 Jetbot을 활용한다. Jetbot의 주행 상황을 CSI 카메라로 촬영하여 모델 $f(\cdot)$의 입력 데이터 x로 활용한다. 장애물 유무를 판단하는 이진 분류 문제를 고려한다. 224 × 224 크기의 영상을 총 900장 수집하였다. ResNet18모델을 사용하여 학습률 0.005, 배치 크기 16, Adam 알고리즘[3]을 사용하여 100회 동안 훈련을 진행한다. 손실함수 $J(\theta, x, y)$는 binary cross-entropy 함수를 사용한다. [그림 5]는 ResNet18 모델의 훈련 과정을 나타낸다. 훈련 및 검증 정확도가 모두 1로 수렴하여 자율주행 모델이 잘 훈련된 것을 알 수 있다.

3-2. 적대적 공격 모의실험 결과



[그림 6] 잡음 세기 ϵ의 변화에 따른 정확도 성능

훈련된 ResNet18 모델의 적대적 공격을 수행한 결과를 [그림 6]에 도시한다. 잡음의 크기 ϵ를 증가시켜가며 ResNet18 모델의 분류 정확도 변화량을 추적하였다. 성능 평가를 위해 max-norm 제약조건 $\|n\|_\infty = \epsilon$를 만족하는 잡음을 무작위로 생성한 "Random noise" 기법의 성능도 함께 나타내었다. 예상했던 대로, FGSM 기반의 적대적 공격 기법은 잡음의 크기가 커질수록 적대적 공격을 효과적으로 수행한다. 실험적으로 $\epsilon = 0.0007$이 적절한 선택임을 확인하였으며, [그림 2]에서 볼 수 있듯이 공격받은 영상에서 잡음을 육안으로 확인하지 못하는 상황이다. 반면에 Random noise기법은 잡음의 크기를 증가시켜도 정확도 성능이 감소하지 않아 적대적 공격을 제대로 수행할 수 없다. 제안하는 기법을 자율주행 테스트베드에 이식한 결과는 [4]에서 확인할 수 있다.

Ⅴ. 결론

본 논문에서는 인공지능 기반 자율주행 네트워크에 적대적 공격을 수행하는 기법을 제안한다. FGSM 방법을 활용하여 공격 잡음의 크기를 제안하면서도 자율주행 인공지능 모델의 오판률을 증가시키는 적대적 공격을 수행한다. 모의실험 결과 100%의 정확도를 보이던 ResNet18 모델의 정확도를 약 15%까지 감소시켜 적대적 공격을 성공적으로 수행하였다.

참 고 문 헌

[1] 품목별 보고서-자율주행차, 정보통신산업진흥원, 2019.

[2] I. J. Goodfellow, J. Shlens, C. Szegedy, "Explaining and harnessing adversarial examples", in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2015

[3] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2014

[4] https://youtu.be/O80T6JNJIsg