자율주행을 위한 적대적 공격 및 방어 딥러닝 모델 연구

김채현[1], 이진규[2], 정은[3], 정재호[2], 이현정[4], 이규영[5]
[1]고려대학교 바이오의공학부
[2]고려대학교 건축사회환경공학부
[3]명지대학교 소프트웨어융합학과
[4]성신여자대학교 융합보안공학과
[5]한국과학기술원 정보보호대학원
kchaehyeon01@gmail.com, jink.lee0070@gmail.com, wjddms0926@gmail.com,
brian7479@naver.com, lhj020731@gmail.com, leeahn1223@kaist.ac.kr

Study of Adversarial Attack and Defense Deep Learning Model for Autonomous Driving

Chae-Hyeon Kim[1], Jin-Kyu Lee[2], Eun Jung[3], Jae-Ho Jung[2], Hyun-Jung Lee[4], Gyu-Young Lee[5]
[1]School of Biomedical Engineering, Korea University
[2]Dept. of Civil, Environmental and Architectural Engineering, Korea University
[3]Dept. of Convergence Software, Myongji University
[4]Dept. of Convergence Security Engineering, Sungshin Women's University
[5]Graduate School of Information Security, KAIST

요 약

자율주행의 시대가 도래함에 따라, 딥러닝 모델에 대한 적대적 공격 위험이 함께 증가하고 있다. 카메라 기반 자율주행차량이 공격받을 경우 보행자나 표지판 등에 대한 오분류로 인해 심각한 사고로 이어질 수 있어, 자율주행 시스템에서의 적대적 공격에 대한 방어 및 보안 기술 연구가 필수적이다. 이에 본 논문에서는 GTSRB 표지판 데이터를 이용하여 각종 공격 및 방어 기법을 개발하고 제안한다. 시간 및 정확도 측면에서 성능을 비교함으로써, 자율주행에 최적인 모델을 탐구하고 더 나아가 해당 모델들의 완전자율주행을 위한 발전 방향을 제안한다.

1. 서론

자율주행의 시대가 열리고 있다. 자율주행차 세계 시장 규모는 2035년 1조 달러(약 1300조원)로, 연평균 41% 성장할 것으로 전망되며, 2030년 레벨3(시스템의 요청 있을 경우에만 운전자 개입 필요한 수준; 시스템 스스로 교통 신호를 파악하여, 차량 제어와 주행 주도권이 인간에서 시스템으로 전환됨) 이상의 자율주행 신차 보급률이 50% 이상일 것으로 예측된다.

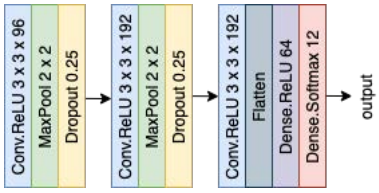
카메라를 통한 주행 환경 인식은 자율주행의 핵심 기술이다. 자율주행차는 카메라를 통해 얻은 이미지를 딥러닝 객체 인식 모델 등으로 분석하여 실시간 주행 상황을 판단하는데, 이때 딥러닝 모델이 적대적 공격에 취약하다는 심각한 문제점이 존재한다. 카메라 기반 자율주행차에 적대적 공격이 가해질 경우, 인식 객체인 보행자, 표지판 등에 대해 오분류가 발생하여 생명과 직결된 심각한 사고로 이어질 수 있다. 따라서 자율주행 시스템에서의 적대적 공격에 대한 방어 및 보안 기술 연구는 필수적이다.

본 연구는 표지판 데이터셋으로 적대적 공격 이미지를 생성하고, 그에 따른 방어 모델을 개발한다. 개발한 모델에 대해 정확도와 방어 동작 시간을 고려한 성능을 비교하여, 정확한 실시간 판단이 이루어져야 하는 자율주행에 최적화된 모델을 제시한다. 또한 개발한 모델을 완전자율주행에 적용하기 위한 발전 방향을 제시한다.

2. 자율주행을 위한 적대적 공격 및 방어 모델 개발

2.1. 실험 환경

본 연구는 약 4만개의 표지판 이미지로 이루어진 GTSRB 데이터셋[1]을 사용한다. 43개 클래스 중 이미지가 1,000개 이상인 12종을 선정해, 총 12,000개의 훈련데이터와 6,180개의 테스트데이터를 사용한다. 사용한 CNN 분류기 구조는 (그림 1)과 같으며, 훈련 데이터셋에 대해 99.6%, 테스트 데이터셋에 대해 97.5%의 정확도를 보였다.



(그림 1) CNN 분류기 구조

2.2. 적대적 공격 모델 개발

정상 이미지에 육안으로 구분이 불가한 미세한 변동(perturbation)을 의도적으로 추가하면, 딥러닝 모델은 높은 확신(confidence)을 가지고 오분류를 야기한다. 이렇게 조작된 이미지가 적대적 예제이고, 해당 모델에 적대적인 예제를 찾아 제작하는 것이 적대적 공격(Adversarial attack)이다.

본 연구에서는 FGSM(식1)과 PGD(식2) 공격 기법을 각각 GTSRB에 맞게 구현하여 공격용 데이터셋을 제작하였고, 이를 개발한 방어 모델의 성능 확인에 사용하였다. 정상 이미지와 적대적 예제의 예시는 (그림 2)와 같으며, 공격 기법 별 CNN 분류기에 대한 정확도는 <표 1>과 같다.

$$\tilde{x} = x + \epsilon sign(\nabla_x J(\theta, x, y)) \quad \text{(식1) [2]}$$

$$x^{t+1} = \Pi_{x+S}(x^t + \alpha\, sgn(\nabla_x L(\theta, x, y)) \quad \text{(식2) [3]}$$

원본 이미지	FGSM (ε = 0.03)	PGD (ε = 0.03)
		

(그림 2) 정상 이미지 및 적대적 예제 예시

(%)	perturbation (ϵ)					
	0(정상)	0.02	0.03	0.05	0.08	0.10
FGSM	96.90	64.56	53.36	40.38	20.59	20.11
PGD		55.74	43.65	29.15	12.84	10.24

<표 1> 공격 데이터에 대한 CNN accuracy

2.3. 적대적 방어 모델 개발 및 제안

2.3.1. 화이트박스 및 블랙박스 방어

적대적 공격에 대한 방어 모델은 크게 화이트박스, 블랙박스 두 종류로 나뉜다. 화이트박스 방어는 적대적 공격에 대한 정보를 알고 있는 경우에 해당하며, 공격 데이터를 모델 학습에 사용하는 적대적 학습이 대표적이다. 블랙박스는 적대적 공격의 종류와 무관하게 방어를 수행한다. 화이트박스는 모델의 성능을 높이기 위해, 블랙박스는 어떤 공격이 가해질지 알 수 없는 실제 자율주행 상황에서 주로 사용된다.

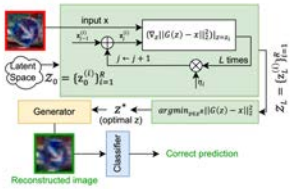
본 연구는 미리 알려진 공격에 대해 높은 성능을 보이지만 공격에 대한 모든 정보를 알고 있어야 하는 화이트박스 모델(적대적 학습)과, 공격에 대한 정보와 학습 없이도 방어를 수행하는 블랙박스 모델(MagNet, Defense-GAN, DefPCA)을 각각 구현하고 비교·융합하여 최적 모델을 개발한다.

2.3.2. MagNet 모델 개발

MagNet은 오토인코더로 구성된 Detector와 Reformer로 이루어진 방어 모델이다. Detector로 재구성 오류를 통해 적대적 이미지를 탐지하고 Reformer로 적대적 이미지를 재구성함으로써 방어를 수행한다. [4] 본 연구에서 MagNet을 GTSRB에 맞게 개발 및 학습하였으며, 2가지 공격을 통해 방어성능을 확인하였다. 사용한 오토인코더 구조와 훈련 파라미터

Structure		parameters	
Detector Ⅱ	Detector Ⅰ & Reformer		
Conv.Sigmoid 3×3×3	Conv.Sigmoid 3×3×3	Optim. Method	Sigmoid
Conv.Sigmoid 3×3×3	AveragePooling 2×2	Learning Rate	0.01
Conv.Sigmoid 3×3×1	Conv.Sigmoid 3×3×3	Batch Size	32
	Conv.Sigmoid 3×3×3	Epochs	350
	Upsampling 2×2	Regularization	L2
	Conv.Sigmoid 3×3×3		
	Conv.Sigmoid 3×3×1		

<표 2> Detector와 Reformer 구조 및 훈련 파라미터 [4]

Structure	parameters	
	Batch Size	256
	Epochs	20,000
	Noise Size	(1,100)
	L (Noise Update)	200
	R (Sample Extract)	10

<표 3> Defense-GAN 구조 및 훈련 파라미터 [5]

는 <표 2>와 같다. 학습 시 입력 이미지에 0.1의 가우시안 잡음을 추가한 후 [0,1] 범위로 절삭하여 사용하였다.

2.3.3. Defense-GAN 모델 개발

Defense-GAN은 생성기와 분류기의 적대적 훈련 과정을 통해 무작위 잡음을 원본 이미지 분포에 수렴시키는 GAN을 이용한 방어 모델이다. 원본 이미지 분포로 학습된 생성기에, 적대적 공격 이미지에 최적화된 잡음을 넣어, 공격 시 가해진 변동을 완화함으로써 방어를 수행한다.[5]

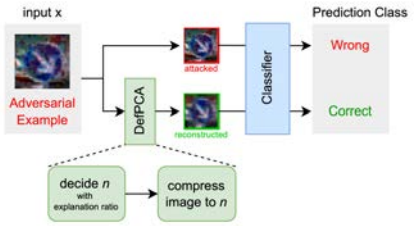
본 연구에서 잡음이 생성되는 잠재공간을 최적화하는 DCGAN 기반 Defense-GAN을 GTSRB에 맞게 개발 및 학습하고, 2가지 공격을 적용해 방어 성능을 확인하였다. 모델 구조[5]와 훈련 파라미터는 <표 3>과 같다.

2.3.4. DefPCA 모델 제안 및 개발

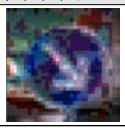
PCA(Principal Component Analysis)를 통해 추출된 데이터의 주성분만으로 설명력이 유지된다는 개념에서 착안하여, 적대적 공격 이미지를 주성분으로 재구성해 적대적 공격을 완화 및 방어하는 DefPCA 모델을 새롭게 제안 및 개발하였다. 공격 이미지에 가해진 변동은 재구성된 이미지에 반영되지 않거나 적게 반영되어 방어를 수행한다. 본 연구에서는 DefPCA를 개발하고 GTSRB 데이터셋에 맞게 학습하였으며, 구현한 FGSM 및 PGD 공격에 대해 방어 성능을 확인하였다. PCA 주성분 개수(n)에 따른 설명력을 분석해, 주성분 개수를 n = 5, n = 10으로 설정하여 이미지를 재구성하였다. DefPCA 모델 구조는 (그림 3)과 같다.

3. 실험 결과 : 최적 모델 및 발전 방향 제안

Defense-GAN과 DefPCA를 통해 재구성된 이미지 예시는 (그림 4)와 같다. 모델 별로 적대적 학습 없이 방어를 진행했을 때 평균 정확도는 MagNet은 17.92%p, Defense-GAN은 26%p, DefPCA는 1.13%p 증가하였다. 정상 데이터 분포를 학습하여 이미지를 재구성 및 생성하는 MagNet



(그림 3) DefPCA 모델 구조

공격이미지 (ε = 0.1)	Defense-GAN	DefPCA (n = 5)
		

(그림 4) 공격 및 방어를 통해 재구성한 이미지 예시

FGSM acc		eps				
Defense	AdvTr	0.02	0.03	0.05	0.08	0.10
No defense	X	64.56	53.36	40.38	20.59	20.11
AdvTr	O	94.26	88.64	86.50	85.11	84.59
MagNet	X	71.47	67.73	55.89	50.49	42.98
	O	68.06	64.78	59.43	56.92	62.84
Defense GAN	X	82.00	76.00	70.00	53.00	51.00
	O	67.00	64.00	49.00	47.00	38.00
Def PCA n=5	X	62.60	57.75	39.35	26.45	18.50
	O	91.20	94.00	94.35	95.35	95.20
Def PCA n=10	X	68.20	56.10	37.50	25.80	15.30
	O	93.90	98.75	98.60	98.40	98.80

<표 4> FGSM에 따른 방어기법 별 분류 정확도

PGD acc		eps				
Defense	AdvTr	0.02	0.03	0.05	0.08	0.10
No defense	X	55.74	43.65	29.15	12.84	10.24
AdvTr	O	96.29	95.15	90.02	88.60	71.01
MagNet	X	70.19	61.88	56.49	40.06	30.89
	O	90.30	86.36	79.23	80.06	62.00
Defense GAN	X	82.00	74.00	62.00	48.00	50.00
	O	83.00	79.00	72.00	80.00	55.00
Def PCA n=5	X	64.90	50.95	27.10	18.95	10.15
	O	97.65	97.45	96.70	95.75	14.25
Def PCA n=10	X	62.15	47.85	25.05	17.30	8.90
	O	98.75	98.70	97.55	97.35	13.55

<표 5> PGD에 따른 방어기법 별 분류 정확도

과 Defense-GAN이 높은 방어율을 보였다. 적대적 학습을 적용해 방어를 진행했을 때 평균 정확도는 MagNet은 35.93%p, Defense-GAN은 28.33%p, DefPCA 53.24%p 증가하였다. 적대적 학습과 방어모델을 결합한 결과 세 경우 모두 성능 향상을 보였으며, 데이터 분포에 대한 학습 없이, 이미지를 재구성하는 DefPCA가 적대적 학습과의 융합에 가장 적합했다. (표 4, 5)

방어 수행 시간 측면에서는 MagNet, DefPCA, Defense-GAN 순으로 속도가 빨랐다. MagNet은 0.11ms/이미지 정도의 매우 빠른 처리 속도를 보였으며, Defense-GAN은 실시간 주행 판단에 적용되기에는 다소 느린 22,228ms/이미지 속도를 보였다. (표 6)

방어 성능과 수행 시간을 고려했을 때, 현재 자율주행을 위한 방어 모델로 MagNet이 가장 적합하다. 적대적 학습이 완전히 가능한 자율주행 환경에서는, 높은 방어율과 3ms 수준의 수행 시간을 보인 DefPCA와 적대적 학습 결합 모델이 가장 적합하다. Defense-GAN의 경우, 적대적 학습이 불가한 자율주행상황에서 다양한 변동에 대해 일관적으로 높은 성능을 보인다. 또한 잠재공간을 최적화해 생성기로 새로운 가상의 이미지를 만들어 주는 모델이므로, 높은 변동에도 방어율 유지가 가능하다. 그러나 실시간 처리가 불가한 문제가 있다. 이에 PCA, 다운샘플링, 주파수 분석 등의 기법을 결합하여 계산 및 생성 과정을 간소화하는 모델 경량화의 개선 방향을 제안한다. 또한 CNN 모델에 정상 및 공격 이미지를 푸리에 분석한 데이터를 이진 분류로 학습한 경우 약 93.62%의 정확도로 분류가 가능함을 확인하였다. 이를 공격 Detector로써 적용하여, 선택적 방어를 통한 모델 최적화/실용화에 기여할 수 있을 것이다.

(ms)	MagNet	Defense-GAN	DefPCA
t	0.1104	22,228	2.83

<표 6> 이미지 1장 당 방어 평균 소요 시간

4. 결론 및 향후 연구

본 연구에서는 표지판 데이터셋을 이용하여 적대적 공격에 대한 방어 모델을 개발하고, 방어 모델들에 대해 정확도 및 방어 시간 성능을 비교 분석하여 자율주행에 적합한 모델을 제시하였다. 또한 개발된 방어 모델에 모델 경량화 아이디어를 적용해 자율주행의 실시간 정보 처리에 기여할 수 있는 방안을 제시하였다. 본 연구에서 제시한 모델을 통해 딥러닝 보안모델의 실시간 처리 및 완전자율주행 시스템의 구현을 기대한다.

※ 본 프로젝트는 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] J.Stallkamp, M.Schlipsing, J.Salmen, C.Igel, Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition, Neural Networks, Available online 20 Feb 2012, ISSN 0893-6080, 10.1016/j.neunet.2012.02.016.

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

[3] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).

[4] Dongyu Meng and Hao Chen. MagNet: a Two-Pronged Defense against Adversarial Examples. In ACM Conference on Computer and Communications Security (CCS), 2017.

[5] Samangouei, P., Kabkab, M., and Chellappa, R. Defense-gan: Protecting classifiers against adversarial attacks us- ing generative models. In International Conference on Learning Representations, 2018.