Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.33, NO.6, Dec. 2023 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2023.33.6.907

적대적 회피 공격에 대응하는 안전한 자율주행 자동차 시스템*

이 승 열,[1†] 이 현 로,[2] 하 재 철[3‡]
[1,2,3]호서대학교 (학생, 대학원생, 교수)

Secure Self-Driving Car System Resistant to the Adversarial Evasion Attacks*

Seungyeol Lee,[1†] Hyunro Lee,[2] Jaecheol Ha[3‡]
[1,2,3]Hoseo University (Student, Graduate student, Professor)

요 약

최근 자율주행 자동차는 운전자 지원 시스템에 딥러닝 기술을 적용하여 운전자에게 편의성을 제공하고 있지만, 딥러닝 기술이 적대적 회피 공격(adversarial evasion attacks)에 취약함이 밝혀졌다. 본 논문에서는 객체 인식 알고리즘인 YOLOv5(You Only Look Once)를 대상으로 MI-FGSM (Momentum Iterative-Fast Gradient Sign Method)를 포함한 5가지 적대적 회피 공격을 수행하였으며 객체 탐지 성능을 mAP(mean Average Precision)로 측정하였다. 특히, 본 논문에서는 모폴로지 연산을 적용하여 적대적 공격으로부터 노이즈를 제거하고 경계선을 추출하여 YOLO가 객체를 정상적 탐지할 수 있는 방안을 제안하고 이를 실험을 통해 그 성능을 분석하였다. 실험 결과, 적대적 공격을 수행했을 때 YOLO의 mAP가 최소 7.9%까지 떨어져 YOLO가 객체를 정확하게 탐지하지 못하는 것을 87.3%까지 성능을 개선하였다.

ABSTRACT

Recently, a self-driving car have applied deep learning technology to advanced driver assistance system can provide convenience to drivers, but it is shown deep that learning technology is vulnerable to adversarial evasion attacks. In this paper, we performed five adversarial evasion attacks, including MI-FGSM(Momentum Iterative-Fast Gradient Sign Method), targeting the object detection algorithm YOLOv5 (You Only Look Once), and measured the object detection performance in terms of mAP(mean Average Precision). In particular, we present a method applying morphology operations for YOLO to detect objects normally by removing noise and extracting boundary. As a result of analyzing its performance through experiments, when an adversarial attack was performed, YOLO's mAP dropped by at least 7.9%. The YOLO applied our proposed method can detect objects up to 87.3% of mAP performance.

Keywords: Self-Driving Car, YOLO, Adversarial Evasion Attack, Morphology

I. 서 론

최근 자율주행 자동차에 대한 관심이 증대되면서 그 핵심 기술을 내재하고 있는 운전자 지원 시스템 ADAS(Advanced Driver Assistance System)에 관한 많은 연구가 이루어지고 있다. ADAS는 차

Received(09. 11. 2023), Modified(10. 17. 2023), Accepted(10. 18. 2023)
* 본 논문은 2021년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다(No.2021RIS-004).
* 본 논문은 2023년도 한국정보보호학회 하계학술대회에 발표한 우수논문을 개선 및 확장한 것임
† 주저자, stl990726@gmail.com
‡ 교신저자, jcha@hoseo.edu(Corresponding author)

량에 장착된 각종 센서와 카메라를 이용하여 외부 환경 정보를 감지하고 이를 통해 차량 스스로가 주행할 수 있도록 운전자에게 편의성을 제공하는 시스템이다.

이러한 자율주행 자동차 ADAS에도 딥러닝(deep learning) 기술이 적용되고 있는데 특히, 카메라에서 인식한 객체를 탐지하는데 우수한 성능을 보이고 있다. 그러나 일부 연구에서는 딥러닝 기반의 객체 탐지 알고리즘은 작은 섭동(perturbation)이 더해지면 그 모델의 성능이 떨어진다는 연구 결과가 있었다. 이와 같은 공격을 적대적 회피 공격(adversarial evasion attack)[1-3]이라고 하는데, 이 적대적 공격을 자율주행 자동차에 적용할 경우 자동차의 비정상적인 동작을 유발하게 되어 차량 안전을 크게 위협하는 요소가 된다.

자율주행 자동차에 대한 적대적 회피 공격을 방어하기 위해서 여러 대응 방안이 연구되고 있는데 대표적으로 적대적 학습(adversarial training)[4-6]과 잡음 제거(denoising)[7]방법 등이 있다. 적대적 학습은 기존 학습 데이터를 이용하여 적대적 회피 공격을 이용하여 추가적인 데이터 셋을 구성한 뒤 그 모델을 재학습하는 방법이다. 적대적 학습은 매우 간단한 방어 기법이지만 높은 방어 성공률을 보인다. 하지만 적대적 학습을 통해 학습한 딥러닝 모델은 적대적 이미지를 정상으로 제대로 탐지할수록 원본 이미지에 대한 탐지율은 떨어지게 되는 상호 완충(trade-off) 효과가 발생하게 된다. 즉, 적대적 이미지로부터 딥러닝 모델이 강건성을 가지게 될 때 적대적 공격으로부터의 방어 성공률은 증가하지만, 원본 이미지에 대한 정확도는 감소하는 현상이 나타나게 된다. 두 번째 대응 방법인 잡음 제거 방법은 적대적 회피 공격으로 생성된 섭동을 감소시켜 딥러닝 모델이 정상적으로 동작하게 만드는 방법으로서 입력 이미지에 대한 사전 신호 처리를 수행하여야 한다. 본 논문에서는 고속의 탐지 성능을 유지하면서도 카메라, 라이다 센서의 작동 지연 시간을 감소시킬수 있는 YOLO(You Only Look Once)v5[8-11]를 사용할 경우를 가정하여 적대적 회피 공격 성공 가능성을 진단한다. 구체적으로 적대적 회피 공격 알고리즘인 MI-FGSM (Momentum Iterative Fast Gradient Sign Method)[12], PGD(Projected Gradient Descent)[13], DeepFool[14], EOT-PGD (Expectation over Transformation PGD) [15], BPDA(Backward Pass Differentialble Approximation)[16]를 수행한 후 YOLO가 객체를 정상적으로 탐지하는지를 확인한다. 실험을 수행한 결과, 이 적대적 회피 공격들이 객체 탐지 성능을 크게 낮추게 됨을 확인한 후 대응하는 잡음 제거 알고리듬을 구현하였다. 본 논문에서는 입력 이미지에 모폴로지(morphology) 연산 기법을 사용하여 잡음 제거 및 경계선을 추출한 후 모델에 적용하였다. 본 논문에서 제안하는 모델은 데이터 셋을 추가하여 학습하거나 잡음 제거 딥러닝 모델을 추가하지 않은 환경에서도 92% 이상의 mAP(mean Average Precision) 성능을 보여 효율적으로 적대적 회피공격을 방어할 수 있음을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서 적대적 회피 공격과 관련한 내용을 기술하며, 3장에서는 실험 환경과 적대적 회피 공격을 구체적으로 설명한다. 4장에서는 공격에 대응할 수 있는 대응 기법을 제안하고, 5장에서 적대적 회피 공격 및 대응 방법이 적용된 경우의 성능을 평가한다. 마지막으로 6장에서 결론을 맺는다.

II. 자율주행 자동차에서의 적대적 회피 공격

2.1 객체 탐지 알고리즘 YOLO

자율주행 자동차에서 활용할 수 있는 객체 탐지 알고리즘인 YOLO는 영역 탐색(localization)과 분류(classification)를 동시에 수행하는 대표적인 one-stage detector이다. YOLO는 2016년 처음 제안된 이후 버전마다 모델 구조와 학습 방법이 수정되어 왔는데 이를 간략히 도시한 것이 Fig. 1.이다.

그림에서 보는 바와 같이 YOLO 알고리즘에서는 입력 데이터를 $S \times S$ 그리드로 나누어 각 그리드 셀에 사용자가 설정한 바운딩 박스를 통해 객체의 위치와 크기를 파악한다. 바운딩 박스는 해당 그리드 셀

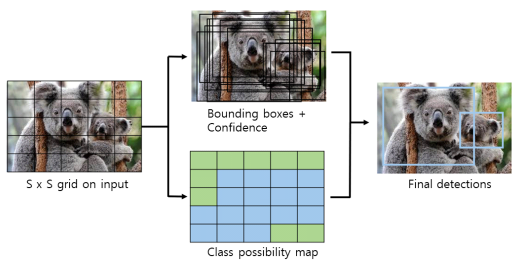

Fig. 1. Object detection using YOLO

에 위치하고 있는 객체의 거리와 크기에 대한 정보를 가지게 되며 해당 셀에 물체가 있을 확률인 신뢰 점수(confidence score) 정보를 가지게 된다. 이를 통해 클래스 분류와 영역 탐지를 동시에 수행할 수 있어 매우 빠른 객체 인식 속도를 보인다. 이러한 YOLO 알고리즘의 장점으로 인해 자율주행 자동차와 같이 실시간으로 객체를 인식해야 하는 상황에서 많이 사용된다.

특히, YOLOv5부터는 Focus 연산과 CSP (Cross Stage Partial) 모듈을 사용하게 되는데 YOLOv5에서 backbone의 첫 번째 계층인 Focus layer는 입력 데이터의 연산 과정을 효과적으로 처리하기 위해 사용되는 계층으로 기존 YOLOv3에서 3개의 계층을 거친 것과 동일한 연산량을 가지도록 설계되었다.

Focus layer에서는 입력 이미지를 작은 그룹으로 나눈 후 분할된 이미지를 연접(concatenation)하여 연산한다. 그래서 입력 데이터에 대한 정보 손실을 최소화하면서 이전 YOLO 알고리즘에 비해 더 빠른 속도로 데이터를 처리할 수 있다. Fig. 2.는 Focus layer 연산 수행 과정을 시각화한 것이다.

CSP 모듈은 기본 계층의 기능을 두 부분으로 분할한 다음 각 계층 구조를 통해 병합하여 연산을 수행한다. 최근의 딥러닝 모델들은 경량화 이후 기존 CNN의 성능이 크게 떨어지게 되는데 CSP 연산을 이용하게 되면 연산량을 20%까지 줄이면서 기존의 이미지 분류 작업을 수행할 때 정확도 측면에서 다른 CNN 모델보다 더 높은 성능을 보인다. Fig. 3.은 CSP 모듈의 연산 과정을 시각화한 것이다.

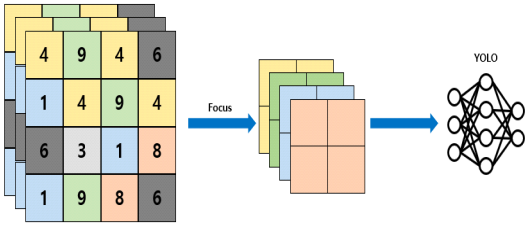

Fig. 2. The process of Focus layer

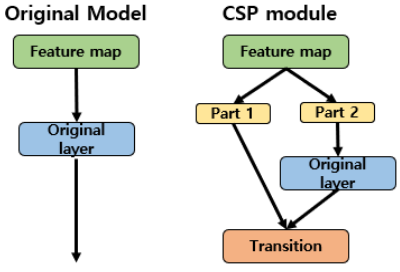

Fig. 3. CSP module structure

2.2 적대적 회피 공격

적대적 회피 공격은 카메라 등에 입력된 데이터를 분류하는 과정에서 다른 클래스로 유도하는 악의적 공격 방식이다. 다음 Fig. 4.에서 보는 바와 같이 입력 데이터에 육안으로는 식별하기 어려운 약간의 섭동을 추가하여 적대적 이미지를 만들어 딥러닝 모델의 분류 결과를 원래 클래스와 다르게 예측하도록 한다. 대표적인 적대적 회피 공격으로 FGSM(Fast Gradient Sign Method)[1]이 제안된 이후 이를 개선한 MI-FGSM[12], PGD[13], DeepFool[14], EOT-PGD[15], BPDA[16] 등과 같은 공격 방법들이 지속적으로 개발되었다.

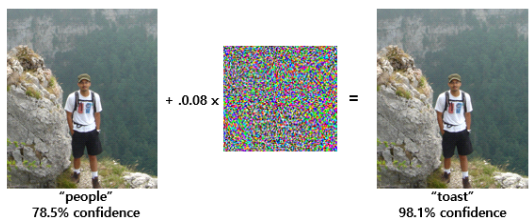

Fig. 4. An example of adversarial evasion attack

2.2.1 MI-FGSM

MI-FGSM 공격 방법은 FGSM을 여러 번 반복하여 섭동을 축적하여 공격하는 I-FGSM(Iterative FGSM)을 개선한 공격 방법이다. 즉, MI-FGSM 섭동을 축적하는 과정에서 손실 함수의 기울기만을 이용하던 것을 Momentum을 이용하는 방법을 추가한 것이다. 기존의 FGSM은 공격 대상 모델에 대해과소접합(under-fitting) 현상이 발생하며 I-FGSM은 과접합(over-fitting)이 발생하여 부적합한 국소 최대 값(poor local maxima)에 빠질 수 있는 단점이 있다.

반면, MI-FGSM은 모멘텀(momentum)을 추가하여 섭동을 계산하기 때문에 좀 더 글로벌한 최대 값를 찾을 수 있다. MI-FGSM을 이용하여 새로운 공격 이미지를 생성하는 수식은 다음과 같다.

$$g_{t+1} = \mu \cdot g_t + \frac{J(x_t^*, y^*)}{\| \nabla_x J(x_t^*, y^*) \|_l} \quad (1)$$

$$x_{adv} = x_{adv} + e^* sign(g_t) \quad (2)$$

수식에서 $\frac{J(x_t^* y^*)}{\| \nabla_x J(x_t^*, y^*) \|_l}$는 적대적 이미지 x_t와 타겟 클래스 y를 손실 함수 J를 통해 기울기를 계산한 것이다. 그 후 초기 모멘텀 값인 g_t에 가중치를 나타내는 μ와 기울기 값을 이용하여 최종적으로 섭동 g_t를 만들어 낸다. 최종적으로 만든 섭동 g_t를 이용하여 공격 강도인 Epsilon을 곱해줌으로써 적대적 이미지를 생성한다.

2.2.2 PGD

PGD는 FGSM를 반복하는 알고리즘으로 손실 함수의 기울기 값을 이용하여 각 step마다 섭동을 축적하여 적대적 이미지를 만들어 내는 알고리즘이다. 각 step마다 기울기 방향으로 이미지를 보정하여 적대적 이미지가 일정 범위를 넘어가지 않게 클램핑(clamping) 처리를 해서 원본 이미지와의 변화를 최소화한다. PGD에서 이미지 생성 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 손실 함수 J를 통해 입력 데이터 x와 타겟 클래스 y를 이용하여 기울기를 계산한 뒤 공격 강도를 나타내는 e를 곱해줌으로써 입력 데이터를 기울기 방향으로 이동시킨다. PGD는 반복적으로 기울기를 이용하여 적대적 이미지를 만들기 때문에 안정적이고 강력한 적대적 이미지를 만들 수 있지만 여러번 반복하여 만들기 때문에 계산비용이 높다는 단점이 있다.

2.2.3 DeepFool

DeepFool은 Euclidean distance를 최적화하는 공격 알고리즘이다. DeepFool은 여러 개의 point에서 딥러닝 모델의 결정 경계선을 근사화하여 최소한의 섭동으로 입력 데이터를 조작한다. DeepFool은 다른 적대적 공격과 달리 더 적은 반복 횟수로 적대적 이미지를 만들기 때문에 계산 비용이 낮고 라벨 값을 필요로 하지 않아 라벨 없는 공격에 유용하다. 그러나 비선형 구조를 가지고 있어 적대적 이미지를 만들어 내는데 시간이 오래 걸린다는 단점이 있다. DeepFool의 수식은 다음과 같다.

$$\overline{F_c} = F(x)_c - F(x)_{y'} \quad (4)$$

$$\widetilde{w_c} = \nabla_x F(x)_c - \nabla_x F(x)_{y'}. \quad (5)$$

수식 (4), (5)에서 $\overline{F_c}$는 $c \in [1,2,....,C]$번째 클래스의 모델의 예측값을 의미한다. 결정 경계선의 초평면 $c^* = \arg\min_{c \neq y} |\tilde{F_c}| / \| \tilde{w_c} \|_2$에 대해 작은 섭동을 계산한 뒤 각 단계 k마다 적대적 이미지 $\tilde{x}_{k+1} = \tilde{x}_k + \delta_k$를 생성한다.

2.2.4 EOT_PGD

EOT 기반 PGD는 일반적인 PGD와 비슷하지만 여러 가지 가능한 변환을 찾아 적대적 이미지를 최소화하는 방법을 추가한 알고리듬이다. EOT는 적대적 이미지와 원본 데이터와의 거리를 최소화하며 변환에 선택된 분포를 고려하여 두 값 사이의 예상 거리를 특정 임계값 아래로 유지하는 방법이다. EOT_PGD의 수식은 다음과 같다.

$$E_{t \sim T}[d(t(x'), t(x))] < \epsilon \text{ and } x \in [0,1]^d \quad (6)$$

수식 (6)에서 x는 원본 이미지, x'는 적대적 이미지를 나타낸다. T는 분포 전체에서 타겟 클래스를 나타내며 타겟 클래스에 대한 확률을 최대화하는 적대적 이미지 x'을 찾는다. EOT_PGD같은 경우에는 다양한 변환을 이용하기 때문에 매우 강력한 공격을 수행할 수 있지만 그에 따른 계산 비용이 많다는 단점이 있다.

2.2.5 BPDA

BPDA는 적대적 공격으로부터 방어하기 위해 기울기 값을 모호하게 만드는 적대적 방어 기술에 대한 공격 방식으로 딥러닝 모델이 역전파 단계에서 기울기를 근사화하여 적대적 이미지를 생성해 낸다. 수식

은 다음과 같다.

$$\nabla_x f(g(x))|_{x=\hat{x}} \approx \nabla_x f(x)|_{x=g(\hat{x})} \quad (7)$$

$f(\bullet)$는 딥러닝 모델, x는 입력 데이터이다. $\nabla_x f(x)$를 근사화하기 위해 $g(x)$를 수행하여 x에 대한 미분 가능한 근사치를 찾는다. 근사화한 값 $g(x)$은 $f(x)$를 대체하여 원본 입력 데이터와 가장 근사화된 적대적 이미지를 찾는다.

BPDA는 다른 적대적 공격과 달리 계산을 할 때 기울기 값이 손실될 수 있어 다른 적대적 공격보다 공격 성공률이 높지 않다는 단점이 있다.

2.3 적대적 회피 공격 방어 기법

적대적 회피 공격에 대응하는 방법은 크게 적대적 학습 방법, 잡음 제거, 적대적 이미지 탐지 방법이 있다. 먼저 적대적 학습 기법 같은 경우는 Goodfellow[17]가 처음으로 제안한 방법으로서 적대적 회피 공격에 대해 강건함을 증가시키기 위한 대응 방안이다. 적대적 학습은 생성한 적대적 이미지를 훈련 데이터셋에 추가하여 학습을 할 때 모델의 손실을 최소화 하며 모델을 업데이트하는 방법이다. 이러한 적대적 학습 기법은 적대적 공격으로부터 강건함을 보여주지만 기존에 학습된 모델보다 원본 이미지를 탐지할 때의 성능이 더 떨어진다는 단점이 있다.

적대적 이미지에 있는 잡음을 제거하는 방법은 Samangouei[18]가 Defense-GAN을 활용하여 설계한 기법이다. 이 방법은 생성 이미지와 원본 이미지의 차이를 최소화하여 모델을 추가 배치한 뒤 적대적 이미지의 잡음을 제거하여 깨끗한 이미지로 복구하는 기법이다.

적대적 이미지 탐지 방법은 원본 이미지와 적대적 이미지의 특성 차이를 계산하여 입력 이미지에 대해 feature를 추출한 후, 오토인코더와 같은 딥러닝 모델을 사용하는 회피 공격 탐지 기법이다[19].

2.4 모폴로지 연산

모폴로지(Morphology)는 영상 내부 객체의 형태와 구조를 분석하고 처리하는 기법으로서 객체의 경계선을 검출하거나 잡음을 제거하는 용도로 많이 사용하는 연산 기법이다[20]. 모폴로지는 팽창(dilatation) 연산과 침식(erosion) 연산을 하는데 팽창은 객체의 주변을 확장하는 연산으로 최외각 픽셀을 확장해 마스크의 유효 영역에 있는 픽셀이라면 두껍게 만드는 연산으로 픽셀이 끊어져 있으면 연결해 주는 역할을 수행한다. 팽창 연산을 통해 이미지 내에 있는 객체의 크기는 커지지만, 배경은 축소된다. Fig. 5.는 이미지 픽셀에 모폴로지의 팽창 연산을 수행한 그림이다.

수식(8)은 팽창 연산에 대한 수식으로 A는 이미지내의 객체 B는 구조요소, z는 B를 이동시킬 결과 이미지의 좌표를 나타내며 B위치 이동시 A와 겹쳐지는 부분이 생겨 A의 외곽 픽셀의 두께가 팽창되는 결과가 나타난다.

$$A \ominus B = z|(B)_z \subseteq A \quad (8)$$

모폴로지의 침식은 구조화 요소 커널을 이용해서 불필요한 영역을 깎아내는 연산이다. 침식 연산을 통해 이미지 내에 있는 객체는 감소하고 배경은 확대된다. 그렇기 때문에 팽창은 노이즈로부터 끊긴 픽셀을 연결하고 침식은 이미지 내의 노이즈를 제거하는 효과가 있다. 다음 Fig .6.은 이미지 픽셀에 모폴로지 침식 연산을 수행한 것을 나타낸 것이다.

수식 (9)는 침식 연산에 대한 수식으로 A는 이미지 내의 객체, B는 A를 침식하기 위해 사용되는 구조 요소, z는 B를 이동시킬 결과 이미지 내의 좌표를 나타낸다.

$$A \oplus B = z|(\hat{B})_z \cap A \neq \oslash \quad (9)$$

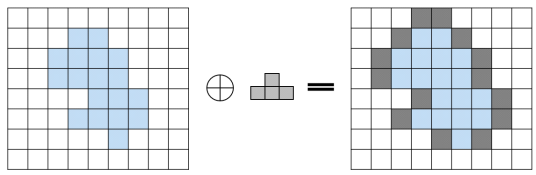

Fig. 5. Dilatation operation in Morphology

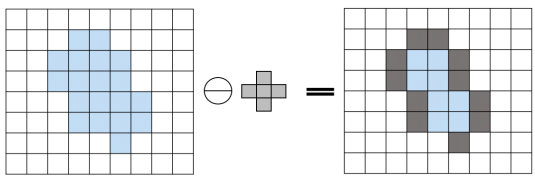

Fig. 6. Erosion operation in Morphology

팽창 연산을 적용한 이미지에서 침식 연산을 적용한 이미지를 빼면 경계 픽셀만 얻게 된다. 즉, 팽창을 수행한 이미지에 침식 연산을 수행한 이미지를 빼면 다음 Fig. 7.과 같이 이미지의 노이즈를 제거한 후 경계선을 추출하는 효과를 얻을 수 있다.

Fig. 7. Boundary images applied with Morphology

III. 적대적 회피 공격 실험 및 대응 분석

본 논문에서는 적대적 회피 공격의 정확성과 대응 방안의 강인성을 고찰하기 위해 다음과 같은 실험 환경을 구축하고 각 공격 모델에 따른 딥러닝 분류기의 성능을 분석하였다.

3.1 실험 환경

3.1.1 교통 표지판 탐지 알고리즘

본 논문에서는 자율주행 자동차용 교통 표지판 탐지 알고리즘으로 YOLO-v5를 사용하였다. YOLO-v5는 COCO 데이터 셋[21]을 이용하여 사전 학습(pre-trained)한 모델을 사용하였다. YOLO의 mAP를 계산에 필요한 confidence threshold의 값은 0.5로 설정하였고 IoU(Intersection Over Union)를 계산하기 위해 IoU threshold는 0.5로 설정하였다.

입력 데이터의 크기는 480x480으로 전처리 하였고, 학습 파라미터로 4 배치사이즈, 500 에폭, Adam optimizer, Binary Cross-Entropy Loss With Logits를 사용하였다. 또한 학습 최적화를 위해 Lambda Learning Rate Scheduler을 사용하여 학습에 도움을 주었다. 모델 개발은 파이토치 1.6.0과 Cuda 10.2 버전을 사용하여 실험 환경을 구성하였다.

3.1.2 교통 표지판 데이터 셋

교통 표지판 데이터 셋은 총 557개의 학습 데이터 셋과 89개의 테스트 데이터 셋으로 구성하였다. 데이터 셋의 클래스는 총 8개로 oneway_sign, speed_30_sign, speed_40_sign, speed_50_sign, speed_60_sign, speed_70_sign, speed_80_sign, stop_sign으로 구성되어 있다.

3.1.3 적대적 회피 공격

본 논문에서는 적대적 회피 공격 알고리즘으로 MI-FGSM, PGD, DeepFool, EOT_PGD, BPDA를 사용하였다. 각 공격의 반복 횟수는 10회로 설정하여 최종적인 적대적 섭동을 만들었고 섭동의 업데이트 크기를 제어하는 alpha와 Deep Fool의 overshoot은 0.02로 동일하게 설정하여 실험하였다.

3.1.4 모폴로지 연산

모폴로지 연산을 사용하기 위해 학습 데이터 셋을 gray scale 이미지로 변경한 후, 임계값 100을 기준으로 이진화하여 사용해 데이터 셋의 노이즈감소와 객체의 경계선 검출을 진행하였다.

3.1.5 성능 평가 방식

적대적 회피 공격에서 YOLO의 성능 평가와 모폴로지를 사용하여 적대적 회피 공격을 방어할 때의 성능을 평가하기 위해 Precision과 Recall을 계산한 후 IoU를 이용해 평균을 구하는 mAP를 사용하였다.

Recall은 데이터 값이 True인 것 중에서 YOLO가 True라고 예측한 비율이며 수식은 다음과 같다.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (10)$$

Precision은 YOLO가 True라고 분류한 것 중에서 실제로 True인 비율을 의미하며 다음과 같다.

$$\mathrm{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (11)$$

IoU은 객체 인식 분야에서 모델이 예측된 바운딩 박스 와 실제 Ground Truth가 일치하는 정도를 0~1 사이로 변환한 값을 나타낸 것이다.

본 논문에서는 IoU 값을 0.5로 지정하여 YOLO가 예측한 바운딩 박스 와 실제 Ground Truth의 IoU를 계산하여 0.5보다 높으면 객체 검출에 성공했다고 판단했다.

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{12}$$

3.2 YOLO에서의 적대적 회피 공격

YOLO 알고리즘을 이용한 객체 검출에서는 입력 데이터 내에 있는 객체의 바운딩 박스와 분류되는 클래스를 동시 예측한다. 따라서 다른 딥러닝 모델과는 다르게 손실 값을 lcls(Class Confidence Loss), lobj(Objectness Loss), lbox(Bounding Box Loss)로 정의된 3가지의 값을 합산한다. 여기서 lcls는 입력 데이터의 클래스와 YOLO가 예측한 클래스 손실값을 의미하고, lobj는 입력 데이터에 있는 객체의 존재 여부와 YOLO가 예측한 객체의 존재 여부에 대한 손실값을 의미한다. 또한, lbox는 입력 데이터 내에 객체의 바운딩 박스의 위치와 크기에 대한 손실 값이다. 상기한 바와 같이 YOLO는 총 3개의 손실 값을 가지기 때문에 기울기 기반의 적대적 회피 공격을 할 때 공격자는 필요한 손실 값을 선택해서 적대적 이미지를 만들어 낼 수 있다. 예를 들어 Table 1.은 각 손실값을 이용한 PGD 공격 기법을 사용하여 YOLO를 공격한 결과이다. Table 1에서 확인할 수 있듯이 Total Loss를 이용하여 공격하게 되면, mAP가 46.5%를 보이지만 바운딩 박스와 관련 있는 lbox를 이용하여 공격하게 되면 mAP가 37%로 가장 높은 공격 성공률을 보인다.

적대적 회피 공격에 어떤 손실 값을 사용하냐에 따라 다음 Fig. 8.에서 보는 바와 같이 YOLO의 예측 결과가 달라질 수 있다. 먼저 lcls를 이용하여 공격했을 때 YOLO가 예측한 클래스가 실제와 다른 것을 확인할 수 있다. 또한, lobj를 이용했을 때는 객체가 없는 위치도 바운딩 박스를 그린 것을 알 수 있다. 그리고 lbox를 이용하여 공격했을 때 바운딩 박스의 크기가 바뀐 것을 알 수 있고 Total Loss를 이용했을 때는 각 손실 값들의 공격을 합친 결과를 확인할수 있다.

Table 1. PGD results according to each loss

	Total Loss	lcls	lobj	lbox
mAP	46.5%	38%	53%	37%

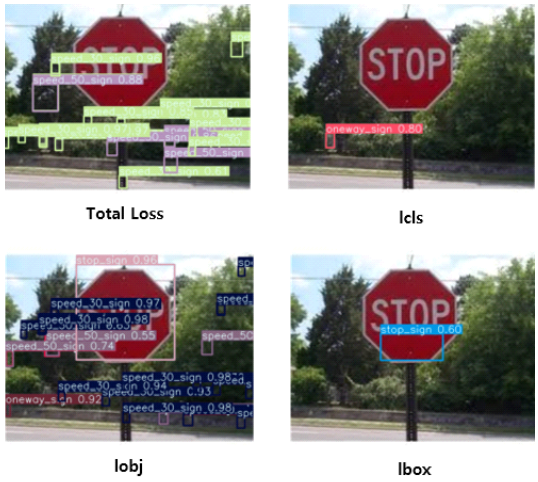

Fig. 8. PGD attack images according to each loss

논문에서는 총 3개의 손실 값들을 합친 Total loss와 lcls, lobj, lbox를 이용하여 섭동을 만들어 적대적 회피 공격을 수행하며 YOLO의 성능을 mAP로 평가를 수행한다[22].

IV. 제안 방법

본 논문에서 모폴로지 연산을 적용하여 YOLO로부터 적대적 회피 공격을 대응하는 기법을 제안한다. 모폴로지 연산을 활용한 적대적 회피 공격을 방어하는 수식은 다음과 같다.

$$X_{adv}(x,y) = \begin{cases} 1, X_{adv}(x,y) \geq T \\ 0, X_{adv}(x,y) \leq T \end{cases} \tag{13}$$

$$X_{adv} \oplus SE(x,y) = \max_{(i,j) \in SE}[X_{adv}(x-i, y-i)] \tag{14}$$

$$X_{adv} \ominus SE(x,y) = \max_{(i,j) \in SE}[X_{adv}(x+i, y+i)] \tag{15}$$

$$B(x,y) = (X_{adv} \oplus SE(x,y)) - (X_{adv} \ominus SE(x,y)) \tag{16}$$

$$output = f_\theta(B(x,y)) \quad (17)$$

X_{adv}는 적대적 이미지를 나타내고 $X_{adv}(x,y)$는 적대적 이미지의 각 픽셀을 나타낸다. 먼저 임계값을 나타내는 T를 기준으로 $X_{adv}(x,y)$를 0과 1로 바꾸어 이진화를 진행한 후 gray scale 이미지로 표현한다. 그 후에 팽창 연산을 수행하기 위해 구조 요소를 나타내는 SE와 X_{adv}와 합집합을 계산하여 적대적 이미지 내에 있는 객체를 확장시킨다.

침식 연산을 수행하기 위해 X_{adv}는 SE와의 교집합을 계산하여 이미지의 객체를 축소화한 뒤, 팽창 결과인 $X_{adv} \oplus SE(x,y)$와 침식 결과인 $X_{adv} \ominus SE(x,y)$의 차이를 구해 적대적 이미지의 경계선을 나타내는 $B(x,y)$를 구한다. 여기서, f_θ 는 YOLO를 나타내며 $f_\theta(B(x,y))$를 통해 바운딩박스, 클래스, 객체 점수를 나타내는 벡터값 $output$을 출력한다. 최종적으로 $output$을 이용하여 적대적 이미지에 대한 바운딩 박스를 그려 객체를 탐지함으로써 적대적 회피 공격을 방어할 수 있다.

V. 적대적 공격 실험 및 모델 성능 평가

먼저 MI-FGSM, PGD, DeepFool, EOT_PGD, BPDA를 사용하여 기존 교통 표지판 데이터 셋으로 학습한 YOLO 모델을 적대적 회피 공격을 수행해 보고자 한다. 다음 Table 2.는 적대적 회피 공격이 없는 YOLO 모델의 성능을 나타낸 것으로서 mAP가 96%에 이르는 성능을 보였다.

다음 Fig. 9.는 교통 표지판에 대한 적대적 회피 공격이 진행되는 과정과 모폴로지 기법을 이용한 방어 방법을 예시로 나타낸 것이다. 만약 속도가 50을 나타내는 표지판이 적대적 회피 공격을 적용하면 속도 80과 같이 오분류를 유도할 수 있으나 대응 기법인 모폴로지 연산이 적용된 경우는 적대적 노이즈가 삽입되었더라도 정상적인 속도로 인식함을 나타낸 것이다.

공격 실험에서 공격 강도 나타내는 파라미터인 Epsilon을 0.05, 0.1, 0.15까지 설정하여 실험하였다. Epsilon은 원본 이미지에 미세한 변화를 가하는 정도를 나타내므로 Epsilon의 값이 적을수록 변화는 미세하게 바뀌고 Epsilon값이 클수록 더 큰 변화가 일어난다. 따라서 Epsilon이 커질수록 공격 성공률은 높아지지만, Epsilon이 커질수록 노이즈가 커져 Fig. 10.과 같이 육안으로 식별할 수 있다는 단점이 있다. 그렇기 때문에 본 논문에서는 육안으로 식별하기 힘든 최소한의 변화를 주기 위해 Epsilon을 0.05, 0.1, 0.15를 설정하여 공격을 진행하였다.

다음 Table 3.은 각 적대적 회피 공격에 따른 모델의 성능을 mAP 값으로 나타낸 것이다. 특히, MI-FGSM에 의한 공격이 이루어진 경우는 탐지 성능에서 최대 10.6%까지 떨어지는 것을 확인할 수 있다. Epsilon을 0.05를 사용하여 약간의 섭동만

Table 2. Evaluation of original YOLO

Model	Precision	Recall	mAP
YOLO	88%	94%	96%

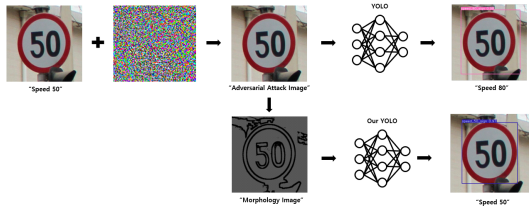

Fig. 9. Adversarial attack and countermeasure on road sign

Fig. 10. Perturbation according to Epsilon

Table 3. Evaluation of performance

Epsilon		0.05	0.1	0.15
MI-FGSM	YOLO	30.3%	16.9%	10.6%
	Our YOLO	92.4%	86.6%	80.7%
PGD	YOLO	46.5%	20.5%	20.5%
	Our YOLO	92.3%	92.1%	91.6%
DeepFool	YOLO	59.6%	46.6%	48.5%
	Our YOLO	92.4%	89.1%	81.6%
EOT_PGD	YOLO	43.2%	18.9%	7.9%
	Our YOLO	92.2%	91.6%	87.3%
BPDA	YOLO	57.4%	46.2%	33.2%
	Our YOLO	93.6%	82.5%	73.2%

추가해도 30.3%까지 mAP가 떨어진 것을 확인할 수 있었다. 여러 가지 공격 방법 중에서 Epsilon을 0.05를 적용했을 때는 MI-FGSM 공격이 우수함을 알 수 있다.

그럼에도 불구하고 본 논문에서는 모폴로지 기법을 사용하여 YOLO 모델에 적용하면 같은 조건하에서 약 92.4%까지 우수한 성능을 보임을 확인하였다. Epsilon을 0.15에서와 같이 공격의 강도가 큰 경우에도 mAP가 80.7%를 보이며 우수한 성능을 보이는 것을 확인하였다.

각 적대적 회피 공격을 수행해 본 결과, 본 논문에서 제안한 모폴로지를 이용한 경계선 추출 기법을 사용하면 BPDA 모델을 사용할 때 최소 73.2%에서 최고 93.6%까지의 mAP 성능을 보임을 확인하였다.

VI. 결 론

최근 자율주행 자동차의 ADAS 시스템 중 카메라를 사용한 이미지를 딥러닝 기법을 이용하여 판별하는 연구가 지속적으로 이루어지고 있다. 반면 악의적인 목적으로 이미지에 잡음을 추가하여 이미지 분류를 어렵게 함으로써 오동작을 유발하는 적대적 회피 공격 가능성이 현실화되고 있다.

본 논문에서는 카메라 이미지에 대한 객체 인식 알고리듬 YOLO에 섭동을 추가하는 방식으로 적대적 회피 공격을 시도하는 실험을 진행하였다. 실험 결과 대부분의 적대적 회피 공격 기법들은 낮은 mAP 값을 가짐으로써 오분류 가능성이 매우 크다는 것을 확인하였다. 또한, 모폴로지 기반의 잡음 제거 및 경계선 추출 알고리듬을 이용한 YOLO 모델에서는 원래 이미지를 정확히 판단할 수 있는 능력이 93.6%임을 실험을 통해 예측할 수 있었다. 따라서 제안하는 모폴로지 기반의 YOLO는 적대적 학습 방법론의 단점인 추가적인 데이터 셋 없이도 적대적 섭동을 제거하여 정상적으로 객체를 탐지할 수 있음을 확인하였다.

References

[1] I. Goodfellow, J. Shlens and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” In International Conference on Learning Representations(ICLR’15), pp. 7-9, Mar. 2015.

[2] N. Akhtar and A. Mian, “Threat of adversarial attacks on deep learning in computer vision: A survey,” IEEE Access, Vol. 6, pp. 14410-14430, Feb. 2018.

[3] N. Carlini and D. Wagner. ″Towards evaluating the robustness of neural networks,″ IEEE Symposium on Security and Privacy(SP’17), pp. 39-57, Jun. 2017.

[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), Aug. 2022.

[5] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, ″Distillation as a defense to adversarial perturbations against deep neural networks,″ IEEE Symposium on Security and Privacy (SP’16), pp. 582-597, Aug. 2016.

[6] 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

[7] W. Xu, D. Evans, and Y. Qi, “Feature squeezing:Detecting adversarial examples in deep neural networks,” Network and Distributed System Security Symposium(NDSS’18), pp. 1-15, Feb. 2018.

[8] Y. Lee and Y. Kim, ″Comparison of CNN and YOLO for Object Detection,″ Journal of the semiconductor & display technology, Vol. 19. No. 1, pp. 85-92, Mar. 2020.

[9] J. Redmon, S. Divvala, R. Cirshick and A. Farhadi, ″You only look once: Unified, real-time object detection,″

IEEE Conference on Computer Vision and Pattern Recognition(CVPR'16), pp. 779-788, 2016.

[10] Y. Zhang, Z. Guo, J. Wu, Y. Tian H. Tang and X. Guo, "Real-Time Vehicle Detection Based on Improved YOLOv5," J. of Sustainability, Vol. 14, Issue 19, pp. 1-19, Sep. 2022

[11] T. Mostafam, J. Chowdhury, K. Rhaman and R. Alam. "Occluded Object Detection for Autonomous Vehicles Employing YOLOv5, YOLOX and Faster R-CNN," Proceedings of the IEEE conference on Information Technology,Electronics and Mobile Communication Conference, pp. 0405-0410, Oct. 2022

[12] Y. Dong, F. Liao, T. Pang and H. su, J. Zhu, X. Hu and J. Li, "Boosting adversarial attacks with momentum," Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185-9193, Jun. 2018.

[13] 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(ICLR'18), 2018.

[14] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. "Deepfool: a simple and accurate method to fool deep neural networks," IEEE Conference on Computer Vision and Pattern Recognition(CVPR'16), pp. 2574-2582, Jun. 2016

[15] X. Liu, Y. LI, C. Wu and C.J. Hsieh, "ADV-BNN:Improved adversarial defense through robust Bayesian neural network," In International Conference on Learning Representations(ICLR'19), pp. 1-13, May. 2019.

[16] A. Athalye, N. Carilni and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," Proceedings of the 35th International Conference on Machine Learning (ICML'18), pp. 274-283, 2018.

[17] Akhtar, Naveed, and M. Ajmal. "Threat of adversarial attacks on deep learning in computer vision," Proceedings of the IEEE conference, Vol. 6, pp. 14410-14430, Feb. 2018

[18] P. Samangouei, M. Kabcab and R. Chellappa, "Defense-gan:Protecting classifiers against adversarial attacks using generative models," Proceedings of ICLR, pp. 1-17, May. 2018.

[19] M. Dongyu and C. Hao. "Magnet: two-pronged defense against adversarial examples," Proceedings of the ACM SIGSAC, pp. 135-147, Oct. 2017.

[20] A.N. Evans and X.U. Liu, "A morphological gradient approach to color edge detection," IEEE Trans. on Image Processing, Vol. 15, No. 6, pp. 1454-1463, May. 2006.

[21] T.U. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C.L. Zitnick, "Microsoft coco: Common objects in context," In European Conference on Computer Vision, Vol. 8693, pp. 740-755, Sep. 2014.

[22] J. Choi and A. Tian, "Adversarial attack and defense of YOLO detectors in autonomous driving scenarios," IEEE Intelligent Vehicles Symposium (IV), pp. 1011-1017, Jul. 2022.

〈저 자 소 개〉

이 승 열 (Seungyeol Lee) 학생회원
2018년 3월~현재: 호서대학교 컴퓨터공학부 학부과정
〈관심분야〉 자동차 보안, 양자내성 암호, 인공지능 보안

이 현 로 (Hyunro Lee) 학생회원
2022년 3월: 호서대학교 컴퓨터공학부 학사
2023년 3월~현재: 호서대학교 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
1993년 1월~현재: 한국정보보호학회 수석부회장
〈관심분야〉 암호학, 부채널 공격, 네트워크 보안, 정보보호