

영상탐지체계와의 각도를 고려한 적대적 패치 공격 기법의 성능 향상 연구

김구경*, 조영호(교신저자)**

*국방대학교 국방관리대학원 사이버컴퓨터공학과

**국방대학교 국방관리대학원 사이버컴퓨터공학과 교수

email: rlarnrud11@korea.kr, younghocho@korea.kr

A Study on Improving the Performance of Adversarial Patch Attack Technique Considering the Angle with Visual Detection Systems

Kukyeong Kim*, Youngho Cho(Corresponding Author)**

*Master's Course, Department of Cyber Security and Computer Engineering,
Korea National Defense University

**Professor, Department of Cyber Security and Computer Engineering,
Korea National Defense University

요 약

YOLO와 같은 객체 탐지 모델은 다양한 산업 분야에서 실시간 인공지능 탐지 기술로 활용되고 있으나, 적대적 패치 공격(adversarial patch attack)에 취약하다. 특히, 기존 연구는 정적인 환경과 정면 중심 이미지에 초점을 맞추어 진행되었으며, 실제 물리적 현실에서 발생하는 시점 변화(viewpoint variation)에 따른 탐지 회피 성능 저하 문제는 충분히 다뤄지지 않았다. 따라서 본 논문은 시점 변화에 따라 적대적 패치의 탐지 회피 효과가 달라지는 현상을 실험적으로 확인하고, 변환 기법을 적용하여 변형된 이미지를 정면 시점과 유사한 형태로 변환함으로써 공격 성능을 향상시킬 수 있음을 보인다. 이를 위해 MNIST 데이터셋을 기반으로 한 CNN 모델을 활용해 시점에 따른 분류 결과를 비교하고, 변환 후 회피 성능이 회복되는 과정을 정량적으로 분석하였다. 이를 통해 향후에는 물리적 조건 변화에 강건한 적대적 패치를 설계하고, 실시간으로 변화하는 외부 환경의 조건값을 이용하여 공격 최적화 알고리즘을 개발할 것이다.

1. 서론

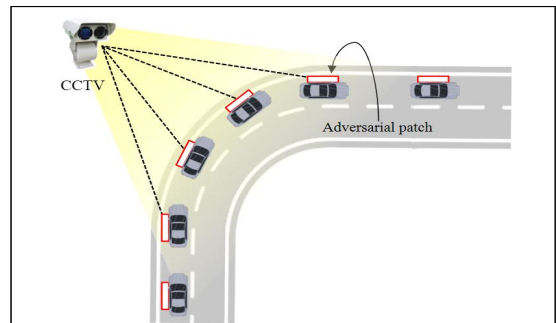
현재 객체 인식 및 탐지(Object Recognition and Detection) 기술은 자율주행, CCTV, 자동차 시스템 등 다양한 분야에서 핵심적으로 사용되고 있다. 특히, YOLO(You Only Look Once)[1]와 같은 객체 탐지 모델은 실시간 탐지에 최적화된 구조로 빠른 연산 속도와 높은 정확도를 동시에 제공하여 산업 전반에 널리 활용되고 있다.

그러나 이러한 객체 탐지 기술의 발전과 보급에도 불구하고, 최근 인공지능 모델의 보안성과 신뢰성 문제가 지속적으로 대두되고 있다. 인공지능 모델의 결정 과정에 미세한 노이즈를 가하여 의도적으로 잘못된 판단을 유도하는 적대적 예제 공격(Adversarial Examples Attack)의 위협은 가상 환경을 넘어 물리적 공간으로까지 확장되고 있다[2]. 특히, 적대적 패치(Adversarial Patch)는 객체 탐지 시스템의 오작동을 유도할 수 있는 대표적인 적대적 예제 공격이지만, 현재까지 대부분의 연구는 정적인 환경에서 정면에 촬영되거나 그와 가까운 이미지 중심

으로 공격 효과를 측정하고 있다[3, 4].

그에 따라, 현실 세계에서 빈번히 발생하는 시점(viewpoint), 조명, 거리 등 다양한 외부 조건 변화를 고려한 적대적 패치 공격의 효과와 성능 분석에 관한 연구들은 부족하다.

실제 물리적 환경에서는 [그림 1]과 같이, 영상 객체 탐지체계가 움직이는 객체(예: 자동차)를 촬영할 경우 다양한 시점의 변화가 생기고 적대적 패치와의 각도(angle) 역시 변할 수 있다. 이로 인해 탐지를 회피하려는 적대적 패치 공격 효과는 물리적 현실에서는 제한적일 수 있다.

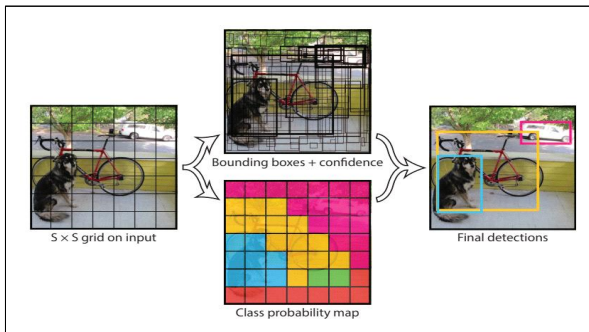


[그림 1] 도로 위에서의 자동차의 각도에 따른 적대적 패치의 시점 변화

본 연구는 물리적 환경에서 객체 탐지체계와 적대적 패치와의 각도 차이에 따른 적대적 패치 공격 성공률 영향을 분석하고, 각도 변화에도 적대적 패치 공격의 효과를 유지할 수 있도록 하는 강건한 동적 적대적 패치 이미지 변환 방법을 제안한다. 이를 통해, YOLO 기반 객체 탐지 시스템에 대한 발전된 적대적 패치 공격을 소개하여 위험성을 알리고 그에 대한 방어 방법을 함께 소개하고자 한다.

2. 배경지식 및 관련 연구

2.1 객체 탐지: YOLO 모델



[그림 2] YOLO 객체 탐지 방법[1]

YOLO는 실시간 객체 탐지를 목표로 개발된 대표적인 딥러닝 기반 탐지 모델로, 이미지 전체를 한 번에 처리하여 객체의 위치와 클래스를 동시에 예측한다. 즉, 원본 이미지를 동일한 크기의 $s \times s$ 개의 영역으로 분할한 후, 각 영역 중앙을 중심으로 미리 정의된 형태로 지정된 경계 박스의 수를 예측하고 이에 대한 신뢰도를 계산하여 최종 객체 탐지를 하는 방식이다[그림 2]. 기존의 R-CNN 계열 모델들이 탐지 속도가 느리다는 한계를 극복하고, 단일 네트워크 구조를 통해 빠른 연산과 높은 정확도를 제공하면서 객체 탐지 분야에서 빠르게 확산되었다.

2.2 적대적 공격: 적대적 예제와 적대적 패치

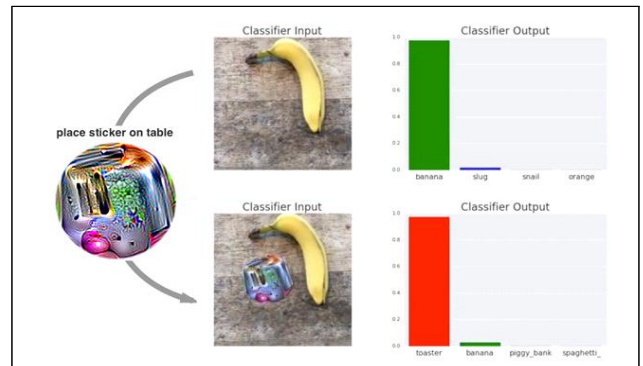
2.2.1 적대적 예제 공격(Adversarial Example Attack)

적대적 예제는 머신러닝, 특히 딥러닝 모델이 처리하는 입력 데이터에 아주 작은 변화를 주어 모델이 잘못된 출력을 내도록 유도하는 공격 기법이다[2]. 즉, 기존 이미지에 오분류를 유도할 수 있는 섭동(Perturbation)을 추가하여 이미지를 생성한 뒤 교란하는 공격 기법이다. 최근에는 이러한 공격이 디지털 환경을 넘어 물리적 환경에서도 유효함이 입증되면서 그 위험이 현실화되고 있다. 예를 들어, Eykholt 등[5]은 도로 표지판에 적대적 스티커를 부착하여 탐지 모델의 객체 탐지를 교란시켰고, McAfee[4]는 35mph 속도제한 표지판에 2인치 테이프를 부착하여 85mph로 오인식을 유발시켜 차량의 과속을 유도할 수 있음을 보였다. 이러한 공격은 자율주행 차량, 감시 시스템, 경계용 드론

과 같은 실시간 객체 인식 기반 시스템에 심각한 영향을 미칠 수 있어 이러한 공격에 대한 신뢰성과 안전성을 갖춘 AI 시스템을 구축하는 것은 중대한 도전 과제이다.

2.2.1 적대적 패치 공격(Adversarial Patch Attack)

적대적 예제 공격이 주로 사람의 눈에 띄지 않는 미세한 노이즈를 가해 모델 출력을 교란하는 방식이었다면, 적대적 패치(Adversarial Patch)는 이미지의 일부분에 시각적으로 명확한 패턴을 삽입하여 모델의 판단을 의도적으로 왜곡시키는 물리적 공격 기법이다[6]. 즉, 노이즈 기반 공격과 달리 사람의 눈에도 쉽게 인식될 수 있는 시각적 패턴을 사용하며, 물리적으로 인쇄하거나 디지털 화면에 띄우는 방식으로도 적용이 가능하다는 점에서 실효성이 높다. 또한 탐지하려는 모델을 대상으로 원본 이미지 내 특정 위치에 패치를 삽입할 경우, 모델이 전혀 엉뚱한 클래스로 분류하거나, 아예 탐지를 실패하도록 유도할 수 있다. 아래 [그림 3]과 같이 첫 번째 이미지의 바나나를 잘 인식하는 반면, 적대적 패치와 함께 있는 두 번째 이미지는 toaster로 오분류하게 된다.



[그림 3] 적대적 패치를 통한 탐지 우회[6]

2.3 기존 연구

Evtimov 등[7]은 물리적 환경에서의 적대적 공격은 관측되는 거리와 각도가 달라질 수 있으며, 단순한 객체에 섭동을 숨기기 힘들다는 문제점을 해결하기 위해 다양한 환경 조건에서 강건한 적대적 공격 생성 알고리즘을 제시하였다. 특히 탐지 모델이 정지 신호 등의 교통 표지판들을 다른 신호로 인식되도록 패치를 생성하였으며, 스티커 형태로 부착 가능한 최초의 패치 공격 위장 공격(camouflage attack)을 제시하였다. 하지만 이는 거리와 각도와 같은 외부 요인을 고려하여 패치를 생성한 것이 아닌 단순히 객체와의 거리와 각도에도 효과적이고, 강건한 패치를 생성한 연구였다.

또한, H.J. Jeong 등[8]은 물리적 적대적 패치의 크기, 회전 각도, 패치 부착 위치 등의 조건들의 적대적 효과에 대해 연구하였다. 하지만 이는 다양한 외부 조건들에 대해 디지털 패치와 물리적 패치의 차이와 물리적 패치의 강건성에 대해 연구하였다. 이

기존 연구들은 다양한 외부 환경에 대한 연관성과 이를 막기 위한 여러 알고리즘을 제안하였지만, 객체와의 거리나 시점과 같은 객체 및 배경들의 외부 환경이나 조건들을 고려하여, 적대적 패치를 최적화하는 연구는 없었다. 따라서 본 연구는 다양한 시점과 각도에서 강건한 탐지 회피 성능을 유지하는 적대적 패치의 생성과 공격 효과에 초점을 두었다.

3. 초도 실험

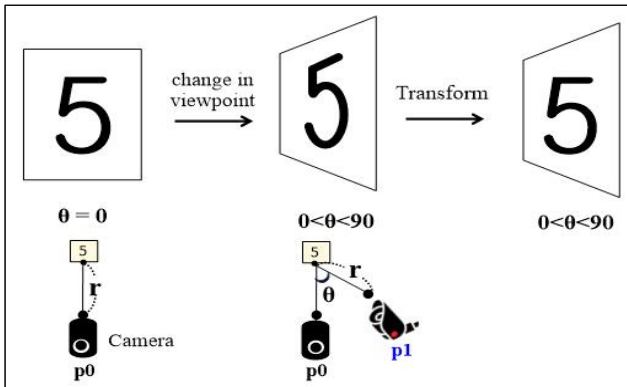
3.1 실험 목적

초도 실험의 목적은 다음 두 가지이다.

첫째, 적대적 패치와 같은 평면화된 이미지를 인식할 때 시점(viewpoint)이나 각도(angle)의 변화에 따라 적대적 패치의 인식률이 떨어지며 그에 따라 오분류 될 수 있음을 실제 실험을 통해 보여 연구 수행의 타당성을 확인하는 것이다.

둘째, 적대적 패치 이미지를 영상탐지체계와의 각도(angle)를 고려하여 해당 각도에서 최대한 정면에서 보이는 것과 같이 촬영되도록 변환(Transform)함으로써 적대적 패치의 인식률과 오분류를 높일 수 있다는 것을 실험을 통해 보이는 것이다. 이를 통해, 영상탐지체계와의 시점이나 각도에 의해 왜곡된 적대적 패치의 공격 성공률(우회율)을 높일 수 있는 아이디어의 활용 가능성을 확인하는 것이다.

3.2 실험 방법



[그림 4] 초도실험 구성도

3.2.1 실험 환경

본 실험에서는 숫자를 인식할 수 있는 MNIST 데이터셋[9]을 통해 학습된 CNN 모델을 사용하였고, 해당 모델의 정확도(Accuracy)는 98.87%이다. 정사각형의 종이에 숫자 '5'를 프린트하여, 삼성 스마트폰(Galaxy S24) 기본 카메라를 통하여 정면, 측면에서 촬영하였다. 즉, 숫자 '5'를 적대적 패치라고 가정하고, [그림 4]에서와 같이 정면(카메라와 적대적 패치와의 각도 = 0°), 측면(카메라와 적대적 패치와의 각도 = 45°)에서 사진을 찍고, 측면에서 찍은 사진을 정면과 유사하게 보일 수 있도록 아래에서

설명한 변환(Transform) 알고리즘[10]을 적용하였다.

3.2.2 변환(Transform) 방법

시점 변화에 의한 왜곡이 존재하는 이미지로부터 정면 시점의 정보를 복원하기 위해, 본 연구에서는 OpenCV 기반의 시점 변환(Perspective Transformation)[9] 기법을 적용하였다. 이 기법은 4개의 기준점으로 원본 이미지의 평면 좌표계에서 정규화된 직사각형 영역으로 투영하는 변환행렬을 계산하였다.

3.3 실험결과 및 분석

[표 1] 실험결과

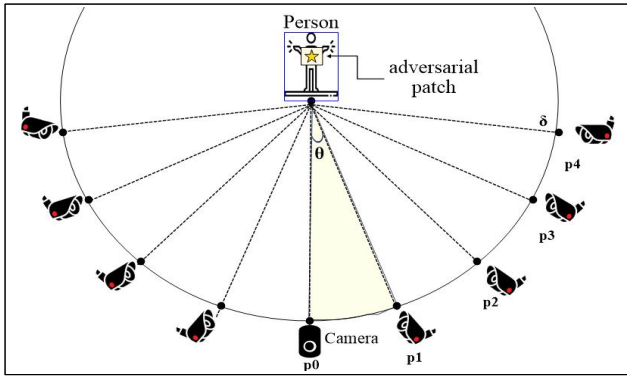
촬영(변환) 사진			
각도(θ)	0° (정면)	45° (측면)	45° (측면)
판별 숫자	5	9	5
신뢰도	0.7276	0.7241	0.6709

첫째, 정면(0°)에서 찍은 숫자는 0.7 이상의 신뢰도로 숫자 '5'로 판별하였으나(적대적 패치 공격 성공), 45°의 측면에서 찍은 숫자는 숫자 '9'로 판별하였다(적대적 패치 공격 실패). 따라서, 이 실험결과를 통해, 영상탐지체계와 적대적 패치 사이의 각도 차이에 따라 적대적 패치의 인식률이 저하되어 오분류될 가능성이 있다는 것을 간접적으로 확인하였다.

둘째, 측면에서 찍은 숫자를 정사각형의 종이의 4개의 꼭지점을 기준으로 시점 변환 기법을 통해 변환하고, CNN 기반의 모델을 통해 숫자를 판별하였을 때 숫자 '5'가 신뢰도 약 0.6 이상으로 판별하는 것을 알 수 있다. 따라서, 제안 아이디어의 적용을 통해 숫자의 인식률이 오르고 본래의 클래스로 인식되었으므로 그에 따른 적대적 패치의 효과도 유효할 것임을 간접적으로 확인하였다.

4. 결론 및 향후 연구 계획

측면 시점에서 본 평면 이미지는 시점이나 각도에 따라 왜곡되는 데, 이 왜곡된 이미지를 변환을 통하여, 정면에서 보는 이미지처럼 만들어 해당 이미지의 인식률을 높일 수 있다는 것을 초도 실험을 통하여 보여주었다. 이와 같은 실험 결과를 적용하여 [그림 5]와 같이 2차원의 적대적 공격 패치(이미지)가 시점 변화에 따라 YOLO와 같은 객체 탐지 모델에 대한 공격 성공률(우회율)이 달라지고, 시점 변환 기법 등을 통하여 정면에서의 공격 성공률만큼 증가시킬 수 있음을 보여준다.



[그림 5] 객체의 탐지 각도에 따른 우회율 실험

따라서 향후 연구 계획은 다음과 같다.

우선, 측면과 같은 왜곡된 시점에서 정면과 같은 적대적 패치 공격 성공률을 높일 수 있는 방법을 연구하여, 적대적 패치가 포함된 객체가 실시간으로 이동할 때 객체와 영상탐지체계(카메라)의 각도를 자동 인식하여 적대적 패치 이미지를 동적 변환하여 적대적 패치 공격 성공률을 극대화시킬 수 있는 기법을 설계할 계획이다.

다음으로, 각도(angle) 이외에도 객체와의 거리(distance), 조명(illumination) 등의 외부 환경 변수를 종합적으로 고려하여 적대적 공격 성공률 저하 가능성 문제를 심도있게 연구하고 적대적 패치의 공격 성공율을 증가시킬 수 있는 방안을 고안하는 연구를 수행할 계획이다.

참고문헌

- [1] J. Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2016.
- [2] I. Goodfellow, et al., "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.
- [3] B Miao et al., "AdvLogo: Adversarial Patch Attack against Object Detectors based on Diffusion Models," arXiv preprint arXiv:2409.07002v2, 2025.
- [4] McAfee, "Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles," 2020.
- [5] K. Eykholt, et al., "Robust physical-world attacks on deep learning visual classification," Proceedings of the IEEE conference on CVPR. 2018.
- [6] T. Brown, et al., "Adversarial patch," arXiv preprint arXiv:1712.09665. 2017
- [7] I. Evtimov, et al., "Robust Physical-World Attacks on Machine Learning Models," arXiv preprint, arXiv:1707.08945 v2, 2017.
- [8] H. Jeong, et al., "Attack Success Rate Analysis of

Adversarial Patch in Physical Environment," Journal of KIISE Vol.50 No.2, 2023.

- [9] Y. Lecun, et al., "Gradient-based Learning Applied To Document Recognition - Proceedings of the IEEE," Proceedings of the IEEE, vol. 86, no. 11, 1998, doi: 10.1109/5.726791.
- [10] G. Bradski et al., "Learning OpenCV: Computer Vision with the OpenCV Library," O'Reilly Media, 2008.