Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.34, NO.2, Apr. 2024 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2024.34.2.245

완전 무인 매장의 AI 보안 취약점: 객체 검출 모델에 대한 Adversarial Patch 공격 및 Data Augmentation의 방어 효과성 분석*

이 원 호,[1†] 나 현 식,[2] 박 소 희,[2] 최 대 선[3‡]
[1,2,3]숭실대학교 (학생, 대학원생, 교수)

AI Security Vulnerabilities in Fully Unmanned Stores: Adversarial Patch Attacks on Object Detection Model & Analysis of the Defense Effectiveness of Data Augmentation*

Won-ho Lee,[1†] Hyun-sik Na,[2] So-hee Park,[2] Dae-seon Choi[3‡]
[1,2,3]Soongsil University (Student, Graduate student, Professor)

요 약

코로나19 팬데믹으로 인해 비대면 거래가 보편화되면서, 완전 무인 매장의 증가 추세가 두드러지고 있다. 이러한 매장에서는 모든 운영 과정이 자동화되어 있으며, 주로 인공지능 기술이 적용된다. 그러나 이러한 인공지능 기술에는 여러 보안 취약점이 존재하고, 이러한 취약점들은 완전 무인 매장 환경에서 치명적으로 작용할 수 있다. 본 논문은 인공지능 기반의 완전 무인 매장이 직면할 수 있는 보안 취약점을 분석하고, 특히 객체 검출 모델인 YOLO에 초점을 맞추어, 적대적 패치를 활용한 Hiding Attack과 Altering Attack이 가능함을 보인다. 이러한 공격으로 인해, 적대적 패치를 부착한 객체는 검출 모델에 의해 인식되지 않거나 다른 객체로 잘못 인식될 수 있다는 것을 확인한다. 또한, 보안 위협을 완화하기 위해 Data Augmentation 기법이 적대적 패치 공격에 어떠한 방어 효과를 주는지 분석한다. 우리는 이러한 결과를 토대로 완전 무인 매장에서 사용되는 인공지능 기술에 내재된 보안 위협에 대응하기 위한 적극적인 방어 연구의 필요성을 강조한다.

ABSTRACT

The COVID-19 pandemic has led to the widespread adoption of contactless transactions, resulting in a noticeable increase in the trend towards fully unmanned stores. In such stores, all operational processes are automated, primarily using artificial intelligence (AI) technology. However, this AI technology has several security vulnerabilities, which can be critical in the environment of fully unmanned stores. This paper analyzes the security vulnerabilities that AI-based fully unmanned stores may face, focusing particularly on the object detection model YOLO, demonstrating that Hiding Attacks and Altering Attacks using adversarial patches are possible. It is confirmed that objects with adversarial patches attached may not be recognized by the detection model or may be incorrectly recognized as other objects. Furthermore, the paper analyzes how Data Augmentation techniques can mitigate security threats by providing a defensive effect against adversarial patch attacks. Based on these results, we emphasize the need for proactive research into defensive measures to address the inherent security threats in AI technology used in fully unmanned stores.

Keywords: Fully Unmanned Stores, Security Vulnerabilities, Adversarial Patch, Data Augmentation

Received(02. 13. 2024), Modified(03. 22. 2024), Accepted(03. 22. 2024)

* 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00511, 엣지 AI 보안을 위한 Robust AI 및 분산 공격탐지기술 개발).

† 주저자, james020907@naver.com
‡ 교신저자, sunchoi@ssu.ac.kr(Corresponding author)

I. 서 론

코로나19 팬데믹으로 인해 비대면 거래가 보편화 됨과 동시에 국내 최저임금 상승으로 인한 인건비 상승과 인력 부족 등의 이유로 무인 매장이 증가하고 있으며, 특히 자동화를 통해 사용자의 편의성을 극대화하는 완전 무인 매장도 등장하고 있다. 완전 무인 매장은 상품관리부터 사용자 인식, 상품 인식 그리고 결제까지의 단계를 사람의 직접적인 개입 없이 자동으로 처리하는 시스템을 말한다.

2018년도 미국의 아마존은 무인 매장에서의 가장 큰 소비자 불만인 결제 대기시간의 문제를 해결하고자 Amazon Go[1] 매장을 개발하였다. 소비자는 출입 시 앱으로 본인 인증을 하고 물건을 집은 후 걸어 나가기만 하면 앱에 사전 등록된 결제 수단으로 자동 결제된다. Amazon Go는 매장에 RFID 센서, LiDAR 센서, 무게 센서가 사용되고, Depth 카메라를 이용하여 3차원 공간 정보를 생성하여 상품을 인식 및 판별한다. 하지만 각종 센서와 Depth 카메라가 고가의 가격으로 이루어져 있기에 높은 초기 비용 문제를 가지고 있다.

이러한 완전 무인 매장은 해외뿐만 아니라 국내에서도 증가하는 추세를 보인다. 파인더스 AI[2]는 국내 AI 솔루션으로 운영하는 'Super Swift'[3]를 개발한 스마트 무인매장 솔루션 개발 회사이다. 기존과 달리 일반 RGB 카메라와 무게 센서만을 활용하여 초기 비용 문제를 해결하였다. 대신 컴퓨터 비전 기술 기반의 개개인 동선 추적과 행동 인식 기반의 상품선택, 그리고 실시간 무게감지 로드셀을 통해 오류를 보완하였다. 이 외에도 CU, GS25, 이마트24 등 국내 편의점에서도 특정 시간 이후부터 무인으로 운영하는 하이브리드형 매장 운영을 통해 완전 무인 매장으로의 전환을 앞두고 있다.

이러한 완전 무인 매장 기술의 핵심은 객체 탐지, 얼굴 인식, 자세 추정 등의 인공지능 기술이다. 많은 완전 무인 매장 시스템이 인공지능을 기반으로 구축되고 있지만 이러한 인공지능 기술들에는 보안 취약점들이 존재한다. 가장 대표적인 공격으로는 사람이 식별할 수 없을 정도의 최소한 변조를 통해 생성한 적대적 예제(adversarial example)를 활용하여 모델이 오분류를 하게 만드는 적대적 공격(adversarial attack)이 있다.

적대적 패치(adversarial patch)[4]는 대표적인 적대적 예제 중 하나로 입력 이미지에 모델이 민감하게 반응하는 작은 이미지 조각(patch)을 추가하는 방식이다. 이러한 적대적 패치를 활용한 공격 연구는 많이 이루어졌지만[5-10], 아직 완전 무인 매장에 적용하여 취약점을 분석한 사례는 부족한 상황이다. 적대적 패치를 통해 완전 무인 매장에서는 상품이 인식되지 않도록 하는 Hiding Attack과 상품이 다른 상품으로 인식되도록 하는 Altering Attack 등을 수행할 수 있다. 이러한 공격은 완전 무인 매장 시스템에 있어서 심각한 피해를 유발할 수 있기에 이에 대한 대응 방안을 구축하는 것은 매우 중요하다.

본 논문은 완전 무인 매장 사례들을 조사하고, 발생할 수 있는 취약점을 분석한다. 그리고 적대적 패치를 활용한 완전 무인 매장 공격을 수행하여 어떠한 문제를 일으킬 수 있는지를 보임과 동시에 Data Augmentation 기법에 대해 적대적 패치 공격의 방어 효과가 있는지 분석한다. 마지막으로 이러한 완전 무인 매장에서의 인공지능 보안 취약점에 대한 적극적인 방어 연구의 필요성을 강조한다.

본 논문에서 기여하는 바는 다음과 같다.

- AI 기술 기반의 완전 무인 매장에서 발생 가능한 취약점을 분석 및 정리한다.
- 실제 완전 무인 매장과 유사한 환경을 구성하고, 적대적 패치를 활용하여 물리적 환경에서 적대적 공격이 가능함을 보인다.
- Data Augmentation 기법이 적대적 패치를 활용한 공격에 방어 효과가 있는지 분석한다.
- 완전 무인 매장 내의 AI 보안 취약점에서 비롯되는 실질적인 문제를 보이고 이에 대응 방안 구축의 중요성을 강조한다.

본 논문의 구성은 다음과 같다. 2장에서는 완전 무인 매장의 사례와 적대적 공격 기법 특히 적대적 예제와 적대적 패치에 대해 기술한다. 3장에서는 완전 무인 매장 시스템에서 사용되는 인공지능 모델들을 분석하고 그에 따른 취약점들을 분석한다. 4장에서는 객체 검출 모델인 YOLO를 공격 대상으로 공격자의 지식과 공격 목적 함수에 대해 정의하여 적대적 패치를 생성하고, 이를 활용하여 Hiding Attack과 Altering Attack을 수행한다. 5장에서 실험 결과를 바탕으로 공격 성능에 대해 분석을 진행하며 6장의 고찰과 7장의 결론으로 마무리를 짓는다.

II. 배경 및 관련 연구

2.1 완전 무인 매장

완전 무인 매장은 대기시간 없이 결제가 이루어지는 시스템인 JWOT(Just Walk-Out Technology)[11]를 적용한 것을 말한다.

대표적인 완전 무인 매장 중 하나인 Amazon Go[1]는 RFID 센서, LiDAR 센서, 무게 센서 그리고 Depth 카메라를 사용해 구축한 완전 무인 매장으로, 3차원 공간 정보를 생성하여 인공지능이 상품을 식별하고 센서로 이를 교차검증 한다. Spharos[12] 또한 이와 유사한 방식으로 객체 검출 결과와 무게 센서의 값을 결합하여 선택한 상품을 인식한다.

파인더스 AI[2]는 RGB 카메라와 무게 센서를 활용하여 경제적인 완전 무인 매장 시스템을 구축하였다. 이 시스템은 자세 추정(pose estimation)을 통하여 영상 내에서 사람의 주요 관절을 정밀하게 추정하여 사용자의 자세를 분석한다. 고가의 센서나 Depth 카메라 대신 저렴한 RGB 카메라를 사용하여 2차원 이미지에서 3차원 자세를 추론하고, 객체 탐지와 무게 센서를 사용하여 상품을 정확하게 인식한다.

Standard AI[13]와 AIFI[14]는 앞선 사례들과는 다르게 센서 등을 사용하지 않고, 오직 RGB 카메라만을 이용하여 상대적으로 적은 비용을 사용하여 완전 무인 매장 시스템을 구축하였다.

이렇듯 완전 무인 매장은 교차 검증을 진행하는 센서들을 제거하여 설치 비용을 낮추고 RGB 카메라를 통한 컴퓨터 비전 기술을 더 발전시키는 방향으로 발전하고 있다. 그렇기에 완전 무인 매장에서 컴퓨터 비전 기술에 적용되는 인공지능의 중요도가 높아지고 있으며 여기에서 발생할 수 있는 AI 보안 취약점을 분석하고 대응 방안을 구축하는 과정이 필요하다.

2.2 적대적 공격

2.2.1 적대적 예제

적대적 공격(adversarial attack)은 딥러닝 모델을 대상으로 한 대표적인 공격 수법 중 하나로, 입력 데이터에 사람이 식별하기 어려운 방식으로 최소한의 변조를 통해 적대적 예제를 생성하여 모델의 오분류를 유발하는 공격이다. 적대적 공격은 Szegedy 등[15]이 입력에 추가된 감지할 수 없는 섭동(perturbation)에 의해, 훈련된 인공지능 모델이 오작동을 일으킨다는 것을 발견하며 처음 제시되었다. 이러한 미세한 섭동은 신경망 네트워크 알고리즘에서 오분류를 유발하며 적대적 공격 알고리즘(adversarial attack algorithm)을 통해 잘못된 예측값을 도출하도록 학습된다. 공격 알고리즘의 종류로는 L-BFGS[15], FGSM(Fast Gradient Sign Method)[16], DeepFool[17] 그리고 C&W[18] 등이 있다.

위의 방식을 통해 생성된 섭동을 정상 이미지에 추가한 것을 적대적 예제라고 한다. 이러한 적대적 예제를 통해 디지털 적대적 공격(digital adversarial attack)과 물리적 적대적 공격(physical adversarial attack)을 수행할 수 있다.

디지털 적대적 공격은 생성한 적대적 예제를 디지털 환경에 주입하여 공격을 수행하는 것을 말한다. Kurakin 등[19]은 적대적 예제를 인쇄하여 모델에 입력하였을 때, 디지털 환경에서와 마찬가지로 오분류를 일으킨다는 것을 발견하였다. 이렇게 디지털 적대적 공격에 사용된 적대적 예제를 물리적 환경에 적용한 것을 물리적 적대적 공격이라고 한다. 이를 위해서는 디지털 환경과 다르게 해상도, 초점, 각도, 조명 등의 이미지 품질과 적대적 예제의 재질 또한 고려되어야 한다는 특징이 있다.

2.2.2 적대적 패치

Brown 등[4]은 작은 이미지 조각을 통해 실제 물리적 환경에서 발생할 수 있는 변형 혹은 왜곡에 대해 강건한 적대적 패치(adversarial patch)를 제안하였다. 적대적 패치는 이미지 전체에 섭동을 추가하는 기존의 적대적 예제와는 다르게 입력 이미지 일부에 패치를 추가하는 방식으로 공격을 수행한다. 이렇게 간단한 패치를 부착하는 것만으로도 모델의 오작동을 유발할 수 있다는 점에서 적대적 패치는 기존보다 효과적인 물리적 적대적 공격이 가능하다.

이러한 적대적 패치를 활용한 공격 연구는 크게 분류(classification) 모델과 검출(detection) 모델에 대해 이루어져 왔다[20]. 처음으로 적대적 패치의 가능성을 확인한 Brown 등의 적대적 패치[4], 패치의 크기를 줄인 Karmon 등의 LaVAN(Localised and Visible Adversarial Noise)[5], 사람이 의심하기 어렵게 한 A Chindaudom 등의 QR Patch[6], Liu 등의 PS-GAN(Perceptual-Sens

itive Generative Adversarial Networks)[21]를 활용한 패치[7], 훈련 데이터에 대한 지식 없이 적대적 공격을 수행하는 Zhou 등의 DiAP(Data-independent Adversarial Patch)[8] 등이 있다.

이 외에도 Madry 등[9]은 PGD(Projected Gradient Descent) 기법을 통해 RGB 범위로 픽셀값을 제한하여 공격을 실제 환경에서 수행하도록 하였으며, Hoory 등[10]은 카메라의 위치에 대해 불변성을 지니며 의미론적으로 비슷한 클래스가 자율 주행 시나리오에 동일한 영향을 끼치는 것을 방지하기 위한 의미론적 상대 기능까지 도입한 동적 적대적 패치 dynamic adversarial patch를 설계하였다.

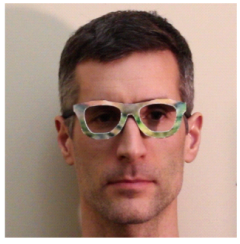
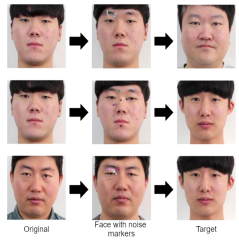
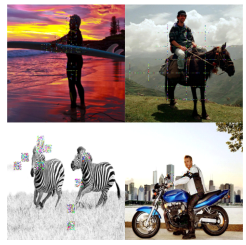
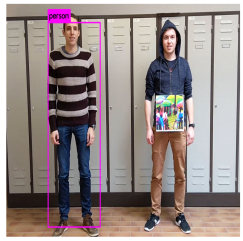
이렇듯 적대적 패치는 사람이 봤을 때는 자연스러운 형태로 유지되지만, 인공지능 모델에 대해서는 효과적으로 공격이 수행되어 오작동이 발생하도록 고안되었으며 이를 활용한 공격 연구가 여러 환경에서 적극적으로 이루어지고 있다.

III. 완전 무인 매장에 대한 취약점 분석

완전 무인 매장이 처음 등장했을 때는 LiDAR 센서, Depth 카메라 등 여러 고가의 센서, 혹은 장비를 필요로 했다. 하지만 고가 장비들을 사용하는 만큼 초기 설치 비용이 큰 문제로 상용화에 어려움이 있었다. 그렇기에 초기 설치 비용을 낮추기 위해 오직 카메라와 무게 센서를 사용한 해결책, 그리고 오직 RGB 카메라만을 이용하여 완전 무인 매장을 구축하였다. 이렇듯 완전 무인 매장에서 인공지능 기술이 차지하는 비중이 점점 높아지고 있다.

한편, 이러한 인공지능 기술에는 앞서 2.2에서 설명한 바와 같이 보안 취약점이 존재한다. 그렇기에 완전 무인 매장에서 발생 가능한 취약점을 분석하고 공격에 대응하는 것은 매우 중요하다. 본 논문에서는 완전 무인 매장에서 사용되는 여러 인공지능 기술들과 그 취약점에 대하여 Table 1. 과 같이 분석하였다.

Table 1. Vulnerabilities about AI Models in Fully Unmanned Store

Model	Role	Vulnerabilities	Attack Research Cases	
Face Recognition	• Identification on entry and exit • Tracking	• Face Theft • Face Falsification • Adversarial Example	 M.sharif et al[23]	 Ryu et al[25]
Object Detection	• Product judgment • Inventory management	• Swapping • Adversarial Example • Adversarial Patch	 DPatch[26]	 Object Hider[29]
Human Detection for Pose Estimation	• Action Recognition • Abnormal Action Detection	• Adversarial Example • Adversarial Patch	 Simen et al[36]	 Hu et al[38]

3.1 얼굴 인식(Face Recognition)

완전 무인 매장에서 얼굴 인식은 주로 크게 고객의 신원 확인과 추적을 위해 사용된다. 계산대가 없는 완전 무인 매장의 특성상 매장에서 나갈 때 자동으로 결제가 이루어져야 한다. 이를 위해 많은 시스템에서는 전용 어플리케이션을 사용하여 미리 결제 수단을 등록하는 방식을 채택하였다. 그리고 매장에 입장할 때 어플리케이션의 QR 코드를 인식하거나 얼굴 인식을 통해 신원 확인을 진행한다. 이를 통해 어떤 고객이 매장에 입장하였고, 어떤 물건을 구매하는지를 인식하는 방식이다. 이러한 얼굴 인식 모델은 프라이버시 관점에서의 취약점과 AI 보안 관점의 취약점이 존재한다.

신원 확인을 위해 얼굴 인식을 사용하게 되면 완전 무인 매장의 데이터베이스에는 사용자의 얼굴에 대한 데이터가 저장되어야 한다. 이 과정에서 사용자의 얼굴 정보가 불가피하게 수집 및 활용되게 되는데, 여기서 프라이버시 관점에서의 취약점이 발생하게 된다. 이를 해결하기 위해 여러 완전 무인 매장에서는 얼굴 인식 모델에 얼굴 비식별화[22] 시스템을 추가하였다.

AI 보안 관점에서는 2.2의 적대적 예제 및 적대적 패치를 활용하여 얼굴을 도용하거나 위변조할 수 있다는 점이 있다. M. Sharif 등[23]은 특수 제작한 안경테를 통해 SOTA(State-of-the-art) 얼굴 인식 모델인 ArcFace[24]가 얼굴을 인식하지 못하거나 다른 사람으로 인식하도록 하였다. 그리고 얼굴에 가해지는 변형을 최소화하기 위해 류권상 등[25]은 얼굴에 5×5 픽셀의 노이즈 마커를 10개 이하로 부착하여 DNN(Deep Neural Network) 기반의 얼굴 인식 모델을 속일 수 있음을 보이기도 하였다.

이러한 취약점을 이용하여 공격자가 자신의 얼굴에 무인 매장을 이용하는 피해자의 얼굴로 인식하도록 하는 적대적 패치를 부착하여 모델을 속이는 방식의 공격이 가능하다.

3.2 객체 검출(Object Detection)

완전 무인 매장에서 객체 검출은 고객들이 매장에서 선택하고 집은 상품이 무엇인지 판단하는 역할과 매장에서의 재고 관리를 위해 사용한다.

이러한 객체 검출 모델에서도 2.2의 적대적 예제 및 적대적 패치로 인한 AI 보안 관점의 취약점이 존재한다. 객체 검출 모델의 경우 경계 박스 회귀(bounding box regression)와 객체 분류(object classification)를 동시에 공격하도록 한 Liu 등의 DPatch[26]가 처음 제안되었다. 이는 데이터셋 간의 전이성(transferability across datasets)에 따라 비타겟 공격(untargeted attack)과 타겟 공격(targeted attack)이 모두 가능함을 보였다. 또한, Faster R-CNN[27]과 YOLO[28] 모델을 대상으로 공격을 진행하여 한 모델에 대해 학습한 적대적 패치가 다른 모델에서도 공격 성능을 보이는 모델 구조에 따른 전이성(model transferability)을 확인하였다. Zhao 등[29]은 heatmap-based와 consensus-based의 두 가지 적대적 패치 생성 알고리즘을 제안하며 객체 탐지 모델이 객체를 탐지하지 못하도록 유도하는 Object Hider를 제안하였다.

이처럼 객체 검출 모델은 특히 2.2.2의 적대적 패치를 활용한 공격에 취약점을 가진다. 완전 무인 매장에서 공격자는 매장의 상품을 객체 검출 모델이 진짜 상품으로 인식하도록 학습한 적대적 패치와 바꿔치기하여 마치 물건을 골랐다가 다시 제자리에 둔 것으로 시스템을 속이는 공격이 가능하다. 이와 비슷한 방식으로 비싼 가격의 상품 위에 값싼 상품으로 인식하도록 하는 적대적 패치를 부착하고 가져가서 마치 비싼 상품을 가져갔지만 값싼 상품을 가져간 것으로 인식되도록 할 수 있다. 또한, 상품을 인식하지 못하도록 하는 적대적 패치를 부착하여 선반에서는 물건이 사라졌지만 누가 이를 가져갔는지는 알지 못하도록 하는 공격이 가능하다.

이를 방어하기 위한 수단으로 대부분 무게 센서를 선반에 적용하지만, 이 역시 상품과 동일한 무게의 추를 이용하면 같은 공격이 가능하며, 무게 센서를 필수로 사용해야 한다는 비용의 문제도 존재한다.

3.3 자세 추정(Pose Estimation)

완전 무인 매장에서 자세 추정은 고객들이 상품을 집는 등의 행위를 인식하는 과정에서 주로 사용된다. 자세 추정 모델은 머리, 어깨, 팔꿈치, 무릎, 손 등 인체의 주요 기관 위치를 인식하여 자세를 파악한다. 이를 통해 사람이 어떠한 행위를 하고 있는지를 정확하게 인식하고 이를 객체 검출 결과와 결합하여 어떤 상품을 가져갔는지를 파악할 수 있다. 이 외에도 매장에서의 이상행위를 탐지하기 위해서도 사용되는데, 대표적으로 Spharos[11]에서는 매장의 고객이 갑

자기 쓰러지거나, 화재가 발생하는 등의 위급상황을 인식하여 사용자에게 알림을 보내고 자동으로 신고를 하기도 한다.

자세 추정 모델은 파이프라인에 따라 하향식 파이프라인(top-down pipeline)과 상향식 파이프라인(bottom-up pipeline)으로 구분할 수 있다. 하향식 파이프라인은 먼저 human detector로 사람을 인식하고, 인식한 각 한 사람의 영역 내부에 대해 2D 자세 추정 네트워크를 통해 각 사람의 자세를 판단한다. 하향식 자세 추정 모델로는 대략적인 관절 위치를 얻기 위한 지역화 서브넷(localization subnet)과 그래프 자세 개선 모듈로 구성된 Graph-PCNN[30], 트랜스포머 기반의 자세 추정 모델인 PPT(Pruned Pose Transformer)[31], ViTPose[32] 등이 있다. 반대로, 상향식 파이프라인은 2D 자세 추정 네트워크를 통해 이미지에 존재하는 신체 부분의 후보군을 추출하고 신체 부분 결합 모듈(body part association)을 통해 최종적인 자세를 판단한다. 대표적인 모델로는 Fast R-CNN[33] 기반의 신체 부분 검출기인 DeepCut[34], 관절의 표준편차를 구하는 과정을 최적화한 SAHR(Scale Adaptive Heatmap Regression)[35]이 있다.

이러한 자세 추정 모델은 사람을 먼저 인식한 후, 그 안에서 자세 정보를 추출하는 방식이기에 기본적으로 2.2의 적대적 예제와 적대적 패치를 통해 사람을 인식하지 못하도록 하는 공격이 가능하다. Simen[36] 등은 YOLOv2[37] 모델에 대해 적대적 패치를 생성하고 이를 부착하여 패치를 들고 있는 사람을 모델이 인식하지 못하도록 하는 공격을 수행한 사례가 있다. 그리고 Hu 등[38]은 사전학습된 GAN(Generative Adversarial Network)[39]을 통해 실제 환경에서 효과적이면서 자연스러운 적대적 패치를 통해 사람 검출(person detection)에 대해 사람을 인식하지 못하도록 하는 Hiding Attack을 수행하였다.

IV. 객체 검출 모델에 대한 적대적 패치 공격

완전 무인 매장 시스템에서 사용되는 여러 인공지능 모델 중에서 객체 검출 모델은 특히 적대적 패치에 취약함을 보인다[40]. 따라서, 본 논문은 객체 검출 모델을 대상으로 적대적 패치 공격을 수행하고자 한다. 공격 목표는 객체를 없는 것으로 오인식하도록 하는 Hiding Attack과 다른 객체로 오분류 하도록 하는 Altering Attack으로 실험을 진행하였다.

한편, 객체 검출 모델을 학습시키는 과정에서 일반화 성능을 향상시키기 위해 훈련 데이터에 변형을 가하거나 노이즈를 추가하여 데이터의 수를 늘리는 Data Augmentation을 수행하였다. 이를 통해 Data Augmentation 기법이 적대적 패치를 활용한 공격에 방어 수단으로서의 효과성을 분석하였다.

4.1 위협 모델

4.1.1 공격 시나리오

Hiding Attack은 적대적 패치를 부착함으로써 상품이 객체 검출 모델에 의해 인식되지 않도록 하는 기법이다. 완전 무인 매장에서 이러한 공격은 모델이 고객이 선반에서 꺼낸 상품을 감지하지 못하게 하여, 해당 상품에 대한 결제가 올바르게 이루어지지 않는 상황을 초래한다.

Altering Attack은 상품에 적대적 패치를 부착하여 객체 검출 모델이 해당 상품을 공격 타겟(target) 상품으로 잘못 인식하도록 유도하는 공격이다. 이 경우, 고가의 상품이 저가의 상품으로 오인식되어, 실제 상품의 가치보다 낮은 금액으로 결제가 이루어질 수 있다.

최종 공격 시나리오인 Fig. 1.과 같은 공격 유형들은 완전 무인 매장 환경에서 객체 검출 모델의 취약점을 이용하여 직접적이고 치명적인 경제적 피해를 유발하며, 완전 무인 매장의 상품 및 재고 관리도 불가하여 정상적으로 운영될 수 없게 만들 수 있다.

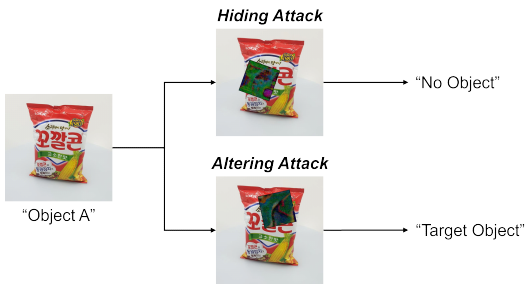


Fig. 1. Adversarial Attack Scenario against Object Detection Model

4.1.2 공격자 지식

본 논문에서는 공격자가 공격 대상 모델(target model)에 대한 매개변수(parameters), 구조(architecture), 가중치(weight) 등의 모든 정보를 알고 있다고 가정하는 White-box Attack을 수행하였다. 디지털 환경과 물리적 환경으로 구분하여 공격자의 지식을 더 구체적으로 설정하여 실험을 진행했다.

우선 디지털 환경에서는 훈련 데이터에 대한 접근이 가능하다고 가정하여 데이터에 기반하여 적대적 패치를 생성할 수 있도록 하였다. 반면, 물리적 환경에서는 매장의 선반 색상, 조명, 배경 등 실제 완전 무인 매장의 환경적 요인을 사전에 고려하면서 패치를 생성할 수 없다는 현실적인 제약이 존재한다. 이에 따라, 디지털 환경에서 생성한 적대적 패치를 물리적 환경에 그대로 적용하면서 데이터셋 간의 공격 전이성에 대해 실험하였다.

4.2 적대적 패치 생성

4.2.1 공격 목적 함수 정의

객체 검출 모델에 대한 적대적 패치를 구성할 때, 공격자는 공격 목표, 패치의 복잡성 및 출력 난이도를 고려할 필요가 있다. 따라서, 공격자는 각 요소에 대한 손실 함수와 그에 대한 가중치를 정의해 동시에 최적화하는 목적 함수를 다음과 같이 설정한다:

$$argmin_p \mathbb{E}_{x \in X, l \in L, t \in T} loss_{det} + \lambda_{sal} loss_{sal} + \lambda_{TV} loss_{TV} + \lambda_{NPS} loss_{NPS} \quad (1)$$

여기서, $loss_{det}, loss_{sal}, loss_{TV}, loss_{NPS}$는 각각 공격 목표, 패치의 단순성, 패치의 복잡성, 그리고 프린팅 난이도를 조절하는 손실 함수이며, 각 이미지 x의 위치 l과 랜덤 변환 t에 따라 패치 p의 각 픽셀을 업데이트한다. 또한, λ_{sal}, λ_{TV}, λ_{NPS}는 각 손실 함수의 가중치를 의미하며, $\mathbb{E}$는 목적 함수에 대한 기댓값을 의미한다.

$loss_{sal}$은 패치의 채도(colorfulness)를 측정하는 saliency loss[41]이며, 값이 낮을수록 패치가 덜 다채로운 형태를 띄게 되며, 다음과 같이 정의한다:

$$loss_{sal} = \sigma_{rgyb} + 0.3 \times \mu_{rgyb} \quad (2)$$
$$\sigma_{rgyb} := \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2},\ \mu_{rgyb} := \sqrt{\mu_{rg}^2 + \mu_{yb}^2}$$

여기서, rg와 yb는 각각 $R-G$ 및 $\frac{1}{2}(R+G)-B$이며, R, G, B는 패치의 각 채널 차원이다.

다음으로, $loss_{TV}$는 인접 픽셀 간 차이를 최소화하도록 유도하는 TV loss[42]이며, 다음과 같다:

$$loss_{TV} = \sum_{i,j} \sqrt{(p_{i,j+1} - p_{i,j})^2 + (p_{i+1,j} - p_{i,j})^2} \quad (3)$$

여기서, B, i, j는 각각 배치 사이즈, 패치의 각 행과 열을 의미하며, TV loss를 통해 패치 적용의 난이도를 조정할 수 있다.

추가적으로, $loss_{NPS}$는 물리적 환경에서의 패치 공격을 위해 프린터가 인쇄할 수 있는 색상 리스트 c_{print}를 구축하여 각 픽셀값이 해당 값에 가까워지도록 유도하는 Non-Printability-Score(NPS) loss[43]이며 다음과 같다:

$$loss_{NPS} = \sum_{i,j} |p_{i,j} - c_{print}|_2^2 \quad (4)$$

이를 통해 패치 인쇄 시, 디지털 환경에서의 공격 성능을 최대한 보존하여 물리적 공격을 시도할 수 있다.

마지막으로, $loss_{det}$는 객체 검출 모델의 confidence score 및 objectiveness score를 공격 목표에 따라 패치를 통해 조작하기 위해 정의되는 손실 함수로, 각 공격 목표 Hiding Attack과 Altering Attack에 따라 다르게 정의된다.

4.2.1.1 Hiding Attack

객체가 검출되지 않도록 패치를 통해 유도하는 Hiding Attack의 $loss_{det}$는 각 객체의 confidence score x_{cls} 및 objectiveness score x_{obj}가 최소화되도록 설정된다. 이 때, 패치 생성 과정의 각 반복마다 객체의 예측 클래스는 달라질 수 있으며, Hiding Attack의 목표가 모든 클래스로 인식이 되지 않는 것임을 감안하여 $loss_{det}$를 다음과 같이 정의한다:

$$loss_{det} = \lambda_{org} \max(x_{cls} \times x_{obj}) \quad (5)$$

4.2.1.2 Altering Attack

대상 객체가 다른 클래스로 잘못 분류되도록 유도하는 Altering Attack의 $loss_{det}$는 objectiveness score는 유지한 채 원본 클래스의 confidence score x_{cls}를 최소화하면서 타겟 클래스의 confidence score x_{cls}^{t}를 최대화하도록 다음과 같이 설정한다:

$$loss_{det} = \lambda_{org}\max(x_{cls}) + \lambda_{tar}(1 - x_{cls}^{t}) \quad (6)$$

여기서, λ_{org}와 λ_{tar}는 각각 원본과 타겟 클래스의 confidence score에 대한 가중치를 의미한다.

4.2.2 구현

이 절에서는 적대적 패치 학습 구현 세부 사항에 대해 설명한다. 우리는 패치 크기를 64x64로 설정한 후, 각 이미지 크기의 최소 0.25배부터 최대 0.40배까지 임의로 크기 조정을 하였다. 패치의 초기 값은 모든 픽셀값이 0.5인 Gray Patch와 0부터 1 사이의 임의값을 가지는 Random Patch 두 가지 경우를 고려하였다. 또한, epoch마다 패치의 밝기, 대조, 노이즈를 임의로 주입하여 다양한 환경에 적응할 수 있도록 설정하였으며, 밝기는 -0.1부터 0.1 사이의 임의값이 적용되고, 대조는 0.8부터 1.2 사이의 임의값이 적용되며, 노이즈는 -0.1부터 0.1 사이의 임의값을 적용했다. 추가로, 각 예제마다 패치의 각도를 최대 20° 조절하고, 경계 박스의 중심을 기준으로 박스 크기의 최대 0.25배만큼 이동시켜 위치를 임의로 조정하였다. 최종적인 두 공격의 적대적 패치는 Table 2, 3. 과 같다.

종합적으로, 본 논문에서 설정한 구현 세부 사항은 물리적 환경에서 다양한 조명, 각도, 카메라와의 거리에 따라 공격 성능을 보존할 수 있도록 고려할 수 있는 효과를 제공한다.

Table 2. Adversarial Patch for Hiding Attack

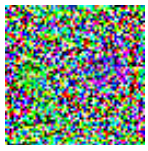
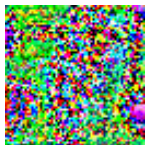
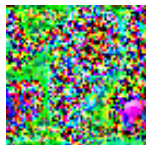
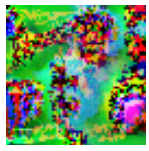
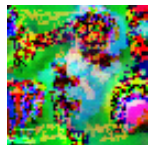
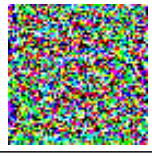
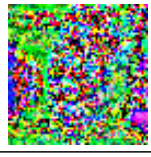
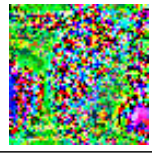
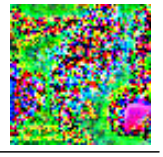
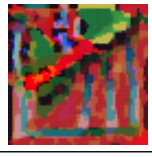
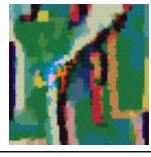
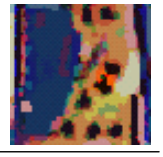
Attack	Initial Patch	Epoch				
		1	50	100	150	200
Hiding Attack	Gray					
	Random					

Table 3. Adversarial Patch for Altering Attack

		Class Number				
		0	2	9	19	20
Altering Attack	Target					
	Patch					

한편, 적대적 패치 공격 목적 함수 내 각 손실 함수의 가중치 $\lambda_{org}, \lambda_{tar}, \lambda_{sal}, \lambda_{NPS}$는 각각 3.0, 6.0, 1.0, 1.0으로 설정하였고, λ_{TV}는 Hiding Attack의 경우 1.0, Altering Attack은 0.5로 설정하였다.

4.3 실험 환경

4.3.1 타겟 모델

적대적 패치 공격을 위해 공격을 수행할 타겟 모델(target model)을 구성하였다. 시중에서 판매되고 있는 과자로 구성된 Roboflow 상의 데이터셋[44]을 1088x1088의 크기로 매개변수를 설정하여 YOLOv5l6[45] 모델을 통해 학습을 진행하였다. 총 객체의 클래스 수는 21개이며, epoch는 100으로 설정하여 모델을 구성하였다.

4.3.2 디지털 환경

4.1.2에서 디지털 환경에서의 적대적 패치 공격의 경우 공격자가 훈련 데이터에 대한 접근이 가능하다고 가정하였다. 이에 따라 디지털 환경에서는 훈련 데이터를 기반으로 적대적 패치를 생성하고 이를 학습에 사용하지 않은 데이터(test data)에 무작위 위치에 부착하고 학습한 YOLO 객체 탐지 모델에 입력하는 방식으로 공격을 수행하였다.

4.3.3 물리적 환경

물리적 환경에서 적대적 패치 공격을 수행하기 위해 직접 완전 무인 매장 테스트베드를 구축하여 실험을 진행하였다. 테스트베드는 상품을 놓을 수 있는 선반과 이를 비추는 RGB 카메로만 구성하였다. 구성한 완전 무인 매장 테스트베드는 Fig. 2.와 같다.

Fig. 2. Unmanned Store Testbed

4.4 Data Augmentation에 대한 방어 효과성 분석

Data Augmentation은 모델의 훈련 데이터에 의도적으로 변형 혹은 노이즈를 추가하여 데이터의 수를 늘리는 기법이다. 이를 통해 모델은 학습되지 않은 새로운 데이터에 대해서도 높은 성능을 유지하여 모델의 일반화 성능을 향상시킬 수 있다. 일반화 성능이 향상된 모델은 기존에 학습하지 않았거나 어느 정도의 변조가 발생한 데이터에 대해서도 좋은 성능을 유지 보여 높은 견고성을 가진다.

본 논문에서는 상품의 데이터 세트에 대해 Data Augmentation 기법을 적용하고 이에 따른 적대적 패치 공격에 대한 견고성을 분석하고자 하였다. Imgaug[46] 라이브러리를 사용하여 모델을 학습시켰다. Flipud(데이터의 상하반전), Multiply(데이터의 밝기), Affine(중심을 기준으로 한 회전), 그리고 Coarse Dropout(데이터의 임의 위치 픽셀을 삭제) 총 4개의 기법을 적용하여 Aug 모델을 훈련하였다. 이러한 기법을 통해 최종적으로 실험을 진행한 객체 검출 모델의 성능은 Table 4. 와 같다.

Table 4. Performance of YOLOv5l6 Models

Method	Precision	Recall	mAP 50	mAP 50-95
Origin	0.971	0.986	0.99	0.919
Aug	0.982	0.982	0.994	0.955

V. 적대적 패치 공격 성능 평가

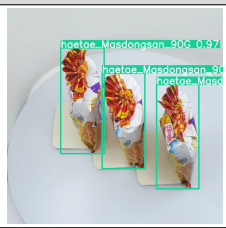
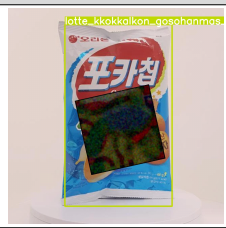
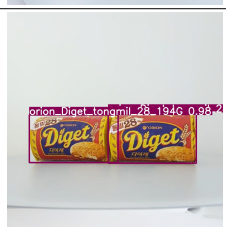
5.1 디지털 적대적 패치 공격

디지털 환경에서 적대적 패치 공격을 수행했을 때, Hiding Attack의 결과는 Table 5, Altering Attack의 결과는 Table 6. 과 같다. Hiding Attack의 경우 적대적 패치를 추가하기 전에는 모든 객체를 정상적으로 검출하지만, 공격을 수행하면 객체를 제대로 검출하지 못하는 것을 확인할 수 있었다. Altering Attack의 경우 동일한 타겟 객체를 가진 적대적 패치를 여러 이미지에 적용해보았을 때, 모든 객체를 타겟 객체로 검출하는 것을 확인할 수 있었다.

Table 5. Digital Hiding Attack Results

Gray		Random	
Before Patched	After Patched	Before Patched	After Patched
			
			
			

Table 6. Digital Altering Attack Results

Gray		Random	
Ground Truth	Prediction	Ground Truth	Prediction
			
			
			

디지털 환경에서 수행한 적대적 패치 공격에 대해 성능을 평가하고자 모델의 mAP(mean Average Precision)값, mAR(mean Average Recall)값 그리고 공격 성공 수치를 ASR(Attack Sucess Rate)로 나타내었다. 디지털 환경에서의 적대적 패치 공격 성능은 Table 7. 과 같다. 전반적으로 보았을 때 두 공격 모두 Initial Patch의 종류에 따른 ASR 값의 차이가 최대 2% 미만인 것으로 보아, 패치의 초기값에 따른 공격 성능은 크게 달라지지 않음을 확인할 수 있었다.

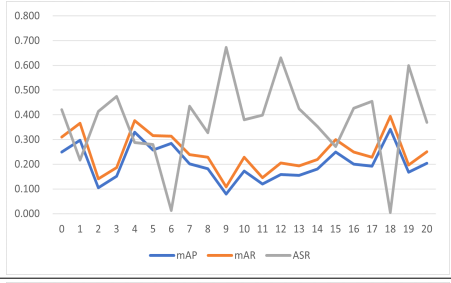
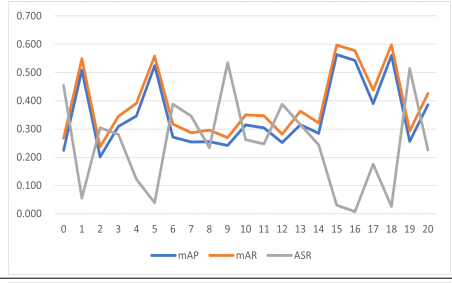
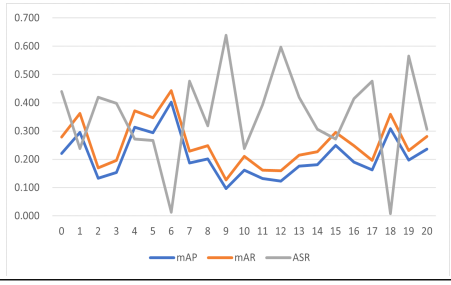
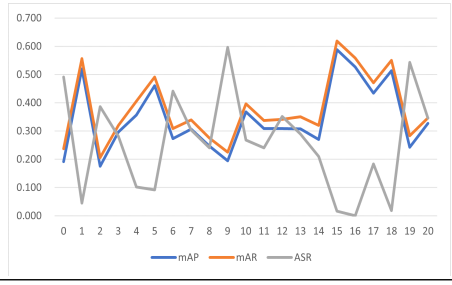
Data Augmentation을 적용한 모델(Aug)과 적용하지 않은 모델(Origin)의 공격 성능을 비교해보면 mAP와 mAR 값 모두 Aug 모델이 높게 나온 것을 확인할 수 있다. 이는 적대적 패치를 활용한 공격을 수행했음에도 객체를 정상적으로 검출한 비율이 높음을 의미하며 실제로 ASR 수치도 Hiding Attack에서는 약 6%, Altering Attack에서는 약 13% 더 적게 나타나는 것을 확인할 수 있었다. 이는 곧 Data Augmentation을 통해 적대적 패치 공격을 어느 정도 완화할 수 있음을 보임과 동시에, 그럼에도 불구하고 적대적 패치 공격이 쉽게 이루어진다고 평가할 수 있다.

Table 7. 에서 Hiding Attack과 Altering Attack의 ASR 값을 비교했을 때, Altering Attack이 상대적으로 낮다. 이는 Hiding Attack은 객체가

Table 7. Performance of Digital Adversarial Patch Attack on Object Detection Model

Attack	Initial Patch	Target Model	mAP	mAR	ASR(%)
Hiding Attack	Gray	Origin	0.202	0.218	79.8
		Aug	0.259	0.276	74.1
	Random	Origin	0.198	0.218	79.8
		Aug	0.269	0.287	73.1
Altering Attack	Gray	Origin	0.204	0.247	37.4
		Aug	0.348	0.386	24.7
	Random	Origin	0.210	0.255	35.6
		Aug	0.344	0.378	25.9

Table 8. Performance of Digtial Altering Attack for each Class

Model / Initial Patch	Origin	Aug
Gray		
Random		

인식되지 않도록 공격하는 반면, Altering Attack은 객체 검출 모델이 객체를 정확히 공격자가 설정한 타겟 객체(target object)로 분류하도록 공격하기 때문에 복잡도가 높으며 비교적 Hiding Attack보다 공격 성공률이 낮게 나올 수 있다고 분석하였다.

각 클래스 별로 수행한 Altering Attack의 공격 성능(mAP, mAR, ASR)은 Table 8. 과 같다. 그래프를 분석해보면 클래스마다 ASR의 값 편차가 큰 것을 확인할 수 있었다. 이는 적대적 패치를 활용한 Altering Attack에 어떠한 클래스를 타겟으로 두는지에 따라 공격 성공 확률이 달라진다는 한계가 존재한다는 것을 확인할 수 있다.

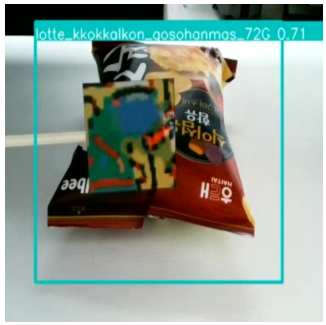
5.2 물리적 적대적 패치 공격

4.3.3에서 구축한 완전 무인 매장 테스트베드에서 물리적 적대적 패치 공격을 수행했을 때 공격 결과는 Table 9. 와 같다. 물리적 환경에서 수행한 공격은 데이터셋 간 전이성을 확인할 수 있었는데, 테스트베드에서 촬영한 이미지로 모델을 학습시키지 않았음에도 불구하고 Hiding Attack과 Altering Attack이 제대로 수행되는 것을 확인할 수 있었다. 하지만 객체의 중앙 부분을 가릴수록, 객체 내부에서 적대적 패치가 차지하는 비율이 클수록 상대적으로 공격이 더 잘 수행되었다. 또한, 실시간으로 객체 탐지를 진행하다 보니 순간순간의 빛에 따라서도 공격 성공 여부가 결정되었다. 이를 통해 디지털 환경에서는 위치와 크기에 어느 정도 견고함을 보였지만 물리적 환경에서는 적대적 패치의 위치나 각도, 객체 내부에서 차지하는 비율에 따라 공격 여부가 미세하게 달라지는 경향이 있음을 확인할 수 있었다. 이는 객체 검출 모델이 객체를 분류(classification)할 때 집중적으로 고려하는 특징(feature)의 위치를 Grad CAM[47] 등을 통해 고려하여 부착한다면 공격 성공률이 증가할 것으로 보인다.

VI. 고 찰

실제 완전 무인 매장 환경에서는 YOLO뿐만 아니라 다른 객체 검출 모델을 사용하기도 한다. 본 연구에서 도출한 적대적 패치에 대한 취약점은 특정 타겟 모델에 제한되지 않으며 다른 객체 검출 모델에 광범

Table 9. Physical Attack Results

Noise	Hiding Attack	Altering Attack
		
		

위하게 적용할 수 있다. 3.2의 DPatch[26]에서 언급한 모델 구조에 대한 전이성을 통해 White-box Attack뿐만 아니라 공격자가 공격 대상 모델에 대한 정보를 알지 못한다고 가정하는 Black-box Attack도 충분히 수행할 수 있다. 또한, 적대적 패치에 대한 전이성을 높이기 위한 연구[48]를 기반으로 더 높은 공격 가능성을 기대할 수 있다.

한편, Altering Attack은 패치를 적용한 객체를 공격자가 설정한 타겟 객체로 오분류하도록 유발한다. 4.2.2의 Table 3. 에서 볼 수 있듯이 Altering Attack을 위해 생성한 적대적 패치와 타겟 객체를 비교해보면, 패치를 이루고 있는 색의 분포와 색의 종류 등에서 유사한 부분이 있음을 관찰할 수 있었다. 따라서, 무인 매장과 같이 객체의 색상이 특정되어있는 경우, 적대적 패치를 생성할 때 타겟 객체의 색상을 고려한다면 더 효과적인 공격 성능을 기대할 수 있다.

5.1의 Table 7. 과 Table 8. 에서 모델의 일반화 성능을 향상시키기 위한 Data Augmentation이 적대적 패치 공격을 완화할 수 있다는 점을 실험에서 확인할 수 있었다. 이를 통해 우리는 Coarse Dropout 같은 임의의 변조가 아닌 적대적 학습(adversarial training)과 같이 타겟 모델의 취약점을 보완할 수 있는 적대적 패치 기반 방어 기술을 도입한다면, 상대적으로 견고한 완전 무인 매장 시스템을 구축할 수 있을 것으로 기대한다.

VII. 결 론

본 논문은 완전 무인 매장에서의 적대적 패치 공격의 가능성을 보임과 동시에 이를 위한 대응 방안 연구의 필요성을 제시한다. 완전 무인 매장 산업이 급성장하는 현재, 이에 사용되는 인공지능 기술에 대한 취약점을 분석하였으며 객체 탐지 모델에 대해서 적대적 패치를 활용한 적대적 공격(Hiding Attack과 Altering Attack)이 가능함을 보였다.

한편, 완전 무인 매장 시스템의 기술들을 분석해보았을 때, 실제로 적대적 패치를 활용한 적대적 공격이 수행되었을 때 이를 인식하고 방어하기 위해서는 모니터링 및 별도의 탐지 기술 등을 통해서만 가능한 상황이다. 5.2의 결과처럼 물리적 환경에서의 적대적 패치를 활용한 공격이 충분히 가능함에 따라 이를 방어할 수 있는 대응 방안의 구축이 필요하다. 이에 따라, Data Augmentation 기법이 적대적 패치 공격에 대해 방어 효과가 있음을 확인하였다. 이를 통해 적대적 학습과 같은 사전 견고성 개선 방안을 도입하였을 때, 적대적 패치에 대한 취약점을 보완할 수 있을 것이라는 가능성을 보였다.

따라서, 우리는 향후 연구로써 완전 무인 매장에서의 적대적 패치를 활용한 적대적 공격을 방어할 수 있는 최적화된 적대적 패치 학습 방안 및 행동 인식 기반 적대적 패치 공격 탐지 연구를 진행할 예정이다.

References

[1] Amazon Technologies, Inc., “Transitioning items from a materials handling facility,” US 2015/0012396 A1, Jan. 2015.

[2] FaindersAI, “FaindersAI” https://fainder.ai/, Oct. 2023.

[3] SuperSwift, “SuperSwift” https://www.youtube.com/watch?v=VJSlS3ujdEo, Jan. 2024.

[4] T.B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, “Adversarial patch,” arXiv preprint arXiv:1712.09665, 2017.

[5] D. Karmon, D. Zoran, and Y. Goldberg, “Lavan: Localized and visible adversarial noise,” International Conference on Machine Learning, pp. 2507-2515, July. 2018.

[6] A. Chindaudom, P. Siritanawan, K. Sumongkayothin, and K. Kotani, “AdversarialQR: An adversarial patch in QR code format,” 2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), IEEE, pp. 1-6, Aug. 2020.

[7] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, “Perceptual-sensitive GAN for generating adversarial patches,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 1028-1035, July. 2019.

[8] X. Zhou, Z. Pan, Y. Duan, J. Zhang, and S. Wang, "A data independent approach to generate adversarial patches," Machine Vision and Applications, vol. 32, no. 3 pp. 1-9, 2021.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, 2017.

[10] S. Hoory, T. Shapira, A. Shabtai, and Y. Elovici, "Dynamic adversarial patch for evading object detection models," arXiv preprint arXiv:2010.13070, 2020.

[11] K. Wankhede, B. Wukkadada, and V. Nadar, "Just walk-out technology and its challenges: A case of Amazon Go," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, pp. 254-257, July. 2018.

[12] Spharos, "Spharos" https://shinsegae-inc.com/business/spharos/info.do, Mar. 2023.

[13] StandardAi, "StandardAi" https://standard.ai/, Jul.2023.

[14] AIFI, "AIFI" https://aifi.com/, Jul. 2023.

[15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[16] I.J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.

[17] S.M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574-2582, 2016.

[18] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," 2017 IEEE Symposium on Security and Privacy (SP), IEEE, pp. 39-57, May.2017.

[19] A. Kurakin, I.J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," Artificial Intelligence Safety and Security, Chapman and Hall/CRC, pp. 99-112, 2018.

[20] K. Nguyen, T. Fernando, C. Fookes, and S. Sridharan, "Physical adversarial attacks for surveillance: A Survey," IEEE Transactions on Neural Networks and Learning Systems, 2023.

[21] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, "Perceptual-sensitive gan for generating adversarial patches," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 1, pp. 1028-1035, July. 2019.

[22] O. Gafni, L. Wolf, and Y. Taigman, "Live face de-identification in video," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9378-9387, 2019.

[23] M. Sharif, S. Bhagavatula, L. Bauer, and M.K. Reiter, "A general framework for adversarial examples with objectives," ACM Transactions on Privacy and Security (TOPS), vol. 22, no. 3, 2019.

[24] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690-4699, 2019.

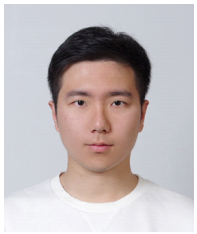
[25] G. Ryu, H. Park, and D. Choi, "Adversarial attacks by attaching noise markers on the face against

deep face recognition," Journal of Information Security and Applications, vol. 60, 102874, 2021.
[26] X. Liu, H. Yang, Z. Liu, L. Song, H. Li, and Y. Chen, "DPatch: An adversarial patch attack on object detectors," arXiv preprint arXiv:1806.02299, 2018.
[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," Advances in Neural Information Processing Systems, vol. 28, 2015.
[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788, 2016.
[29] Y. Zhao, H. Yan, and X. Wei, "Object hider: Adversarial patch attack against object detectors," arXiv preprint arXiv:2010.14974, 2020.
[30] J. Wang, X. Long, Y. Gao, E. Ding, and S. Wen, "Graph-PCNN: Two stage human pose estimation with graph pose refinement," Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Springer International Publishing, Part XI 16, pp. 492-508, 2020.
[31] H. Ma, Z. Wang, Y. Chen, D. Kong, L. Chen, X. Liu, ... & X. Xie, "PPT: Token-pruned pose transformer for monocular and multi-view human pose estimation," European Conference on Computer Vision, Springer Nature Switzerland, pp. 424-442, Oct. 2022.
[32] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "VitPose: Simple vision transformer baselines for human pose estimation," Advances in Neural Information Processing Systems, vol. 35, pp. 38571-38584, 2022.
[33] R. Girshick, "Fast R-CNN," Proceedings of the IEEE International Conference on Computer Vision, pp. 1440-1448, 2015.
[34] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P.V. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4929-4937, 2016.
[35] Z. Luo, Z. Wang, Y. Huang, L. Wang, T. Tan, and E. Zhou, "Rethinking the heatmap regression for bottom-up human pose estimation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13264-13273, 2021.
[36] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[37] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263-7271, 2017.
[38] Y.C.T. Hu, B.H. Kung, D.S. Tan, J.C. Chen, K.L. Hua, and W.H. Cheng, "Naturalistic physical adversarial patch for object detectors," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7848-7857, 2021.
[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and Y. Bengio, "Generative adversarial nets," Advances in Neural Information Processing Systems, vol. 27, 2014.

[40] D. Wang, W. Yao, T. Jiang, G. Tang, and X. Chen, "A survey on physical adversarial attack in computer vision," arXiv preprint arXiv:2209. 14262, 2022.
[41] D. Hasler and S.E. Suesstrunk, "Measuring colorfulness in natural images," Human Vision and Electronic Imaging VIII, SPIE, vol. 5007, pp. 87-95, June. 2003.
[42] A. Chambolle, "An algorithm for total variation minimization and applications," Journal of Mathematical Imaging and Vision, vol. 20, pp. 89-97, 2004.
[43] M. Sharif, S. Bhagavatula, L. Bauer, and M.K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 1528-1540, Oct. 2016.
[44] Roboflow, "Snacks" https://universe.roboflow.com/korea-nazarene-university/snacks-kwjcc/dataset/1, May. 2023
[45] Github, "Yolov5" https://github.com/ultralytics/yolov5, Mar. 2023
[46] Github, "imgaug" https://github.com/aleju/imgaug, Nov. 2023
[47] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," Proceedings of the IEEE International Conference on Computer Vision, pp. 618-626, 2017.
[48] H. Ma, K. Xu, X. Jiang, Z. Zhao, and T. Sun, "Transferable black-box attack against face recognition with spatial mutable adversarial patch," IEEE Transactions on Information Forensics and Security, 2023.

〈저자소개〉

이 원 호 (Won-ho Lee) 학생회원
2021년 3월~현재: 숭실대학교 소프트웨어학부 학사과정
〈관심분야〉 AI 보안, 컴퓨터 비전, 엣지 AI



나 현 식 (Hyun-sik Na) 학생회원
2021년 2월: 공주대학교 응용수학과 학사
2021년 3월~현재: 숭실대학교 소프트웨어학부 석박사통합과정
〈관심분야〉 AI 보안, 개인정보보호, 엣지 AI, 컴퓨터 비전

박 소 희 (So-hee Park) 학생회원
2018년 2월: 공주대학교 응용수학과 학사
2020년 2월: 공주대학교 융합과학과 석사
2020년 6월~2021년 9월: 한국교육학술정보원 전문원
2022년 2월~현재: 숭실대학교 소프트웨어학과 박사과정
〈관심분야〉 인증, 금융보안, 머신러닝, AI 보안, 적대적 공격 및 방어

최 대 선 (Dae-seon Choi) 종신회원
1995년 2월: 동국대학교 컴퓨터공학과 학사
1997년 2월: 포항공과대학교 컴퓨터공학과 석사
2009년 1월: 한국과학기술원 전산학과 박사
1997년 1월~1999년 6월: 현대정보기술 선임
1999년 7월~2015년 8월: 한국전자통신연구원 인증기술연구실 실장/책임연구원
2015년 9월~2020년 8월: 공주대학교 의료정보학과 부교수
2020년 9월~현재: 숭실대학교 소프트웨어학부 교수
2016년~현재: 정보보호학회 이사
〈관심분야〉 인증, 개인정보보호, AI 보안