위성 항공 영상 객체 탐지기에 대한 물리적 적대적 패치 공격 최적화

우정흠, 이은규*
인천대학교

{realbb, eklee}@inu.ac.kr

Evaluating Physical Adversarial Patch Attacks on Object Detectors in Satellite Aerial Imegery

Jung Heum Woo and Enu-Kyu Lee*
Incheon National Univ.

요 약

본 연구는 위성 기반 항공 이미지에서 자동차 객체 탐지기를 대상으로 한 적대적 공격을 최적화하여 디지털과 물리적 도메인에 적용한 것을 평가하였다. 디지털 도메인에서 최적화된 적대적 패치(patch)는 실제로 인쇄되어 차량의 지붕(ON 패치) 또는 주변(OFF 패치)에 설치하고, 실제 촬영한 영상에서 탐지기의 객체성 점수 저하를 통해 공격 효과를 입증했다. 실험 결과, 디지털 도메인은 OFF 패치가 효과적이었으나, 물리적 도메인은 ON 패치가 효과적이었다. 또한 날씨 기반 증강은 오히려 공격 성능을 떨어트리는 것으로 나타났다.

Ⅰ. 서 론

최근 수년 간 위성 플랫폼 기반의 지구 관측은 환경 모니터링, 도시 계획, 경제 예측 등의 다양한 분야에서 활용되기도 하며, 대규모 지역에 대해 위성 이미지를 짧은 시간 간격으로 반복적으로 요구하는 여러 분야에 대해 이미지 확보가 가능해졌다. 이러한 위성 이미지 데이터의 양이 급증함에 따라 자동분석을 위한 기술로 DNN(Deep Neural Network)이 활발히 채택되고 있다. 특히 객체 탐지 및 분할 분야에서 성능이 입증되었다.

그러나 이러한 DNN 은 적대적 예제(Adversarial Example)에 취약하다는 것이 반복적으로 연구되었으며, 이러한 취약성을 바탕으로 물리적 적대적 공격(physical adversarial attack)이 실제 보안적인 위협으로 대두되고 있다.

본 연구는 위성 기반 시점에서 촬영한 항공 이미지 중 자동차를 대상으로 한 탐지기에게 물리적 공격을 실제로 수행한 최초의 연구 중 [1]을 재현한 것으로, 디지털 도메인(digital domain)에서 실제 인쇄가 가능하도록 설계 및 최적화된 패치를 물리적 도메인(physical domain)에 적용하여 탐지기의 성능 저하를 유도하고, 그 공격 성능을 평가한다.

Ⅱ. 본론

적대적 패치는 일반적으로 이미지 내 픽셀을 직접 조작하는 디지털 공격과, 인쇄물 등을 환경에 삽입하여 모델 출력을 유도하는 물리적 공격으로 나뉜다. 기존 물리적 공격은 대부분 자율 주행 또는 얼굴 인식과 같은 지상 환경 공격에 집중되어 있었으며, 항공 이미지에 대한 공격은 시뮬레이션 수준에 머무르거나 실제 구현이 없었다.

본 연구에서 사용하는 패치는 먼저 디지털 도메인에서 최적화 과정을 거친 후, 물리적 도메인으로 인쇄하여 그 공격 효과를 평가한다

2.1 모델 및 데이터셋

본 연구는 YOLOv3 [2]을 기반으로 한 자동차 탐지기를 사용하며, MS-COCO[3] 데이터셋으로 사전학습을 진행한 뒤 COWC-M[4] 데이터셋으로 전이 학습을 수행하였다. 결과 검증 실험은 두 가지의 직접 촬영한 데이터에서 진행된다. 건물 10 층 높이(40m)에서 촬영한 Side Street 환경과 무인 항공기로 60m 높이에서 촬영한 Car Park 환경에서 진행한다.

2.2 패치 최적화

[1]에서 패치의 최적화를 하기 위해 Thys et al. [5]의 방식에 기반한 최적화 프레임워크를 사용하였으며, 아래 (1)과 같은 손실함수를 사용한다.

$$L_i = max(\hat{S}_i) + \delta \cdot NPS(P) + \gamma \cdot TV(P) \quad (1)$$

여기서 δ = 0.01, ɣ = 2.5 로 설정되었다. 해당 손실함수에서 max 항은 해당 이미지 내에서 가장 높은 객체성 점수(objectness score)를 선택하고 이것을 최소화하는데 중점을 두며, NPS(Non-Printability Score)항은 프린터로 재현 가능한 색상 근접 정도를 나타내며, TV(total Variation)은 패치내 인접 픽셀 간의 RGB 차이를 부드럽게 만드는 항이다. 즉 해당 손실함수의 역할은 실제로 인쇄가능한 부드러운 패치를 통해 공격을 하는 데에 목표를 둔다.

또한 패치의 최적화 과정 중 크기에 대한 무작위 스케일링, 회전과 같은 기하학적(geometric) 변환 증강, 밝기와 대비 등을 무작위로 조정하는 색상 공간(colour-space) 변환 증강, 날씨, 조명 등의 정보를 픽셀의 밝기 등을 조정하는 증강 방식이 포함된다.

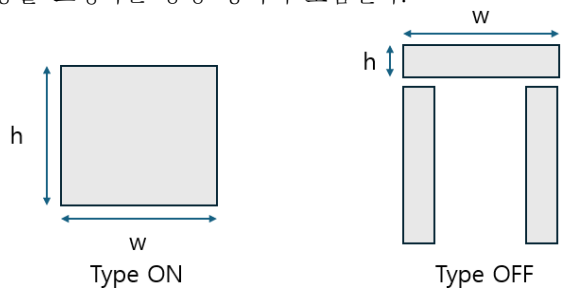

그림 1. 패치 디자인

2.3 패치 타입

[1]에서 제안하는 패치는 그림 1 과 같이 2 개의 형태로 제안된다. 이 때, ON 패치는 차량의 지붕을 덮는 형태로,

디지털 도메인에서 w = 200px, h = 160px 이며, 실제 인쇄는 w = 1189mm, h = 841mm 의 크기를 갖는다 OFF 패치는 차량의 주변을 감싸는 형태로, w = 400px, h = 25px 이며, 실제 인쇄는 w = 3200mm, h = 200mm 의 크기를 갖는다.

2.4 디지털 도메인 실험 결과

디지털 도메인은 AORR(Average Objectness Reduction Rate), 즉 객체성 점수의 감소율을 평균 점수로 평가되며, AORR 의 값이 클수록 해당 파이프라인에 대해 패치의 공격 성능이 좋았음을 의미한다.

표 1. 디지털 도메인 패치 AORR 평가(Side street)

Training pipeline	Patch type	AORR(STD) [%]	AORR(STD+W) [%]
G/C+W	ON	53.76	54.44
G/C	ON	72.59	64.09
Control	ON	29.06	34.75
G/C+W	OFF	79.37	71.80
G/C	OFF	85.51	78.02
Control	OFF	20.65	10.73

표 1 은 디지털 도메인에서 side street 환경에 패치를 적용한 결과를 직접 구현한 것이다. 표 1 에 의하면, 패치가 무작위 색상으로 적용되는 control 의 경우, 기하학적 및 색상 공간 변환 증강을 적용한 G/C, 날씨 및 조명 증강까지 더한 G/C+W 가 우세한 것으로 보아 제안한 패치 최적화의 효과를 입증한다. 또한 G/C 에 비해 G/C+W 는 패치의 공격 효율이 떨어진 것을 확인할 수 있다. 이것은 디지털 도메인에서 날씨 효과를 더하는 것은 오히려 패치 최적화에 방해가 된다는 것을 시사한다. 또한 이 표 1 에서 OFF 패치의 공격 성능이 전체적으로 ON 패치를 압도하는 것을 알 수 있다.

표 2. 물리적 도메인 패치 OSR 평가 [1]

Car	Patch type	Lighting	Motion	Mean OSR
Gray	ON	Both	Moving	0.343
Gray	ON	Sun	Static	0.251
Gray	ON	Shade	Static	0.255
Gray	OFF	Sun	Static	0.429
White	ON	Both	Moving	0.286
White	ON	Sun	Static	0.285
White	ON	Shade	Static	0.197
White	OFF	Sun	Static	0.748

(a) side street

Car	Patch type	Lighting	Motion	Mean OSR
Gray	ON	Sun	Static	0.509
Blue	ON	Shade	Static	0.208
White	ON	Sun	Static	0.746

(b) car park

2.5 물리적 도메인 실험 결과

물리적 도메인은 OSR(Objectness Score Ratio), 즉 실험 전후의 객체성 점수를 비교하는 지표를 사용했으며, AORR 과 마찬가지로 해당 지표가 작을수록 패치의 공격 성능이 올랐음을 의미한다.

물리적 도메인에 적용된 패치를 평가할 때, 차량이 햇빛 또는 그림자에 있었는지에 대한 조명조건과 정지 또는 동적 상태인지에 대한 운동 상태에 따라 범주화를 하였다. 또한 표 1 의 날씨 및 조명 조건의 증강이 없다는 결과에 따라 G/C 파이프라인만을 사용한다.

표 2 는 [1]에서 물리적 도메인에 패치를 적용한 결과를 나타낸 것이다. 이것에 의하면 물리적 도메인에 적용된 두 패치 역시 디지털 도메인 결과와 마찬가지로 효과적인 결과를 보이고 있다. 그러나 디지털 도메인과 달리, 물리적 도메인에선 ON 패치가 OFF 패치보다 공격의 효과가 상대적으로 높다는 결과가 나왔다.

Ⅲ. 결론

본 연구는 실제 항공 이미지에 인쇄된 적대적 패치를 사용한 공격을 구현 및 적용함으로써, 항공 기반 객체 탐지기의 적대적 취약성을 실질적으로 증명하였다.

또한 패치 최적화 과정 중 날씨 및 조명 관련 증강을 적용해도 디지털 도메인에서 효과가 없었다는 결과를 도출했으며, 이것은 향후 연구 과제로 주목할만하다. 이러한 현상의 원인을 기후 모델의 불완전성, 인쇄된 이미지의 명암대비 한계 등과 함께 고려하여 후속적으로 탐구할 수 있을 것이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터사업의 연구결과로 수행되었음. (IITP-2025-RS-2023-00259061) 교신저자: 이은규 (eklee@inu.ac.kr)

참 고 문 헌

[1] Du, Andrew, et al. "Physical adversarial attacks on an aerial imagery object detector." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.

[2] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[3] Tsung-Yi Lin et al., Microsoft coco: Common objects in context. In European conference on computer vision, pages 740-755. Springer, 2014.

[4] T Nathan Mundhenk et al., A large contextual dataset for classification, detection and counting of cars with deep learning. In European Conference on Computer Vision, pages 785-800. Springer, 2016.

[5] Simen Thys et al., Fooling automated surveillance cameras: Adversarial patches to attack person detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019.