드론 AI 임무 소프트웨어 보안 기술 연구

윤여정*, 박정찬
국방과학연구소
*yjyoon.rufina@gmail.com

A Study on Security Technologies for Drone AI Mission Software

Yoon Yeojeong, Park Jeongchan
Agency for Defense Development

요 약

드론의 AI 기반 임무 소프트웨어는 자율비행, 표적 탐지 및 추적, 재난 감지 등의 고차원적 판단을 수행함으로 유용하게 사용되지만 대상 이미지에 작은 변화를 주어 딥 네트워크가 판단할 때의 결정 경계를 넘겨 잘못 인식하도록 하는 적대적 공격에 취약하다. 본 연구는 AI 임무 소프트웨어의 보안 취약점을 분석하고, 센서 융합 기반의 공격 탐지 및 적대적 학습을 통한 대응 방안을 제시한다. 특히 이미지 기반 적대적 공격을 실험하고, 센서 융합 기법을 활용한 이상 탐지 및 학습 강건화 결과를 도출하였다. 실험 결과 단일 센서 기반 AI 모델은 적대적 공격에 취약했으나 센서 융합 및 적대적 학습 기법을 활용할 경우 '임무 정확도가 유의미하게 향상되었다.

Ⅰ. 서 론

무인항공기(UAV, 드론)는 최근 다양한 분야에서 활용도가 높아지며, 고도화된 AI 임무 소프트웨어를 기반으로 자율비행, 정찰, 목표 추적, 재난 대응 등의 복합적 임무를 수행하고 있다. 이러한 소프트웨어는 딥러닝 기반의 객체 탐지, 경로 계획, 센서 데이터 분석 기능을 포함하며, 점점 더 인간의 개입 없이 복잡한 판단과 조치를 수행할 수 있도록 발전하고 있다. 그러나 이와 같은 AI 기반 시스템은 근본적으로 입력 데이터에 매우 민감하며, 그로 인해 적대적 공격(adversarial attack)에 취약한 문제점이 있다. 적대적 공격은 영상이나 센서 데이터에 미세한 교란을 가해 AI의 판단을 왜곡시키는 방식으로 작동하며, 실제 환경에서도 위장된 이미지 패치나 라이다 신호 교란을 통해 드론의 표적 인식 및 장애물 회피 기능을 무력화 시킬 수 있다. 이러한 공격은 임무 실패, 충돌, 오작동 등의 심각한 결과로 이어질 수 있으므로, AI 임무 소프트웨어의 보안 강화는 자율 드론 운용에서 핵심적인 과제가 되고 있다.

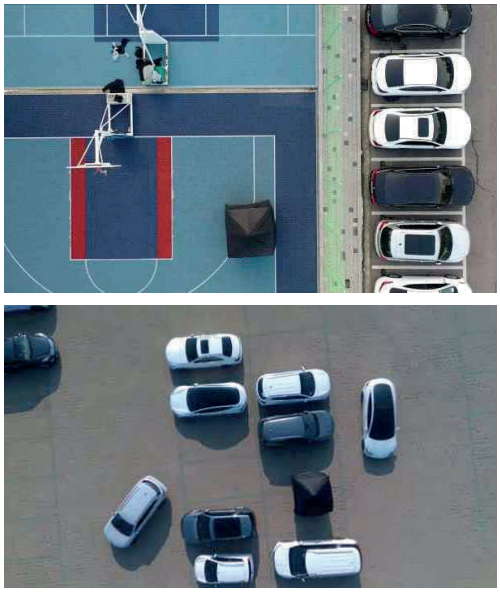
이와 같은 위협에 대응하기 위해서는 첫째로 적대적 공격 유형과 그 효과를 분석할 필요가 있으며, 둘째로 단일 센서에 의존하지 않고 카메라, 라이다 등의 데이터를 통합하는 센서 융합 기반 이상 탐지 기법을 개발해야 한다. 마지막으로 공격에 노출된 데이터를 학습에 반영하여 AI 모델의 방어력을 높이는 적대적 학습(adversarial training) 기법 역시 중요하다.

본 논문에서는 이 세 가지 접근에 기반하여, 드론 AI 임무 소프트웨어의 보안을 실험적으로 검증하고 강건성을 높이기 위한 방안을 제시한다. 이를 통해 실전 환경에서의 신뢰성 높은 드론 운용을 위한 기술적 기초를 마련하고자 한다.

Ⅱ. 본론

1. 실험 데이터셋 구축

적대적 공격 및 센서 융합 기반 적대적 공격 탐지와 적대적 학습 기반 AI 임무 강건화 실험을 위하여 실험 데이터셋을 구축하였다. 두 가지 환경의 주차장에서 드론을 운영하여 드론 카메라를 이용한 직하강 촬영을 진행하였으며, 객체 인식(object detection)을 위한 객체로 다수의 차량과 초소 한 개를 설정하였다. 객체 이동을 고려하여 차량 및 초소의 위치를 변경하며 각 주차장에 대해 1000장 이상의 이미지를 확보하였다. 아래는 구축한 실험 데이터셋의 예시이다.



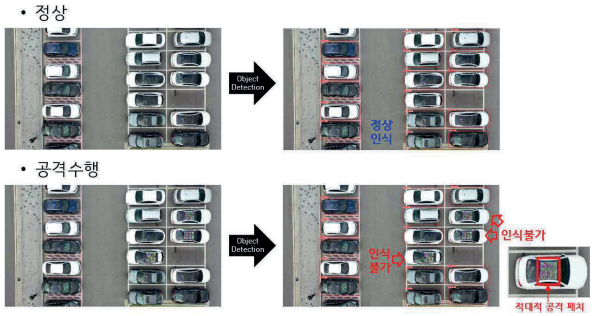
[그림 1. 실험 데이터셋 예시]

2. AI 임무 소프트웨어에 대한 적대적 공격 기술 실험

위와 같이 구축된 실험 데이터셋을 입력값으로 하여 Yolov5 Object detection 모델을 학습하였으며, 해당 모델에 대한 적대적 공격을 위한 공격 패치를 생성하였다. 적대적 공격은 특정 객체에 대한 미탐을 유도하는

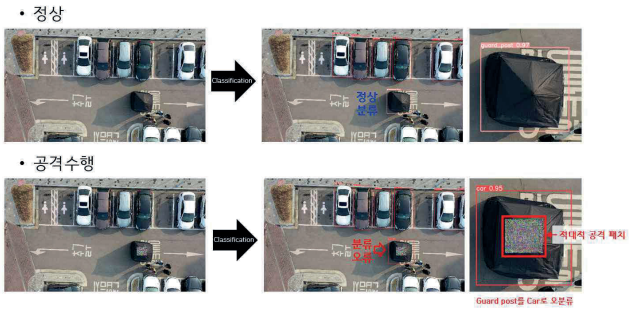
Misdetection 공격과 특정 객체에 대한 오탐을 유도하는 Misclassification 공격으로 구성하였다. Misdetection 공격은 패치가 장착된 차량을 탐지할 수 없도록 하였으며, Misclassification 공격은 초소에 공격 패치를 장착할 경우 차량으로 인식하도록 하였다. Misclassification 공격의 경우 Yolov5 object detection 모델을 대상으로 Guard post(초소) Class Loss 값을 낮추고 Car Class Loss값을 높이도록 Yolo Loss Function을 수정하고, Physical 환경으로의 확장을 위해 픽셀간 차이를 줄여주는 TV Loss를 도입하였으며, 프린트 가능한 색상으로 Perturbation을 생성하기 위한 NPS Loss값을 적용하여 공격 패치 최적화를 진행하였다.

아래는 Digital domain에서 적대적 공격 패치를 적용하여 Misdetection 공격 실험을 수행한 결과 이미지이다. 패치가 적용되지 않은 차량의 경우 Car class로 정상적인 객체 탐지가 되는 반면, 적대적 공격 패치가 적용된 차량의 경우 객체 탐지에 실패(미탐)하는 것을 확인할 수 있다.



[그림 2. Misdetection 실험 결과]

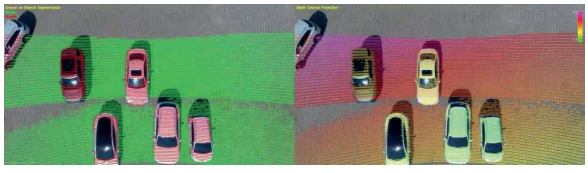
아래는 Digital domain에서 적대적 공격 패치를 적용하여 Misclassification 공격 실험을 수행한 결과 이미지이다. 패치가 적용되지 않은 초소의 경우 Guard post class로 정상적인 객체 탐지가 되는 반면, 적대적 공격 패치가 적용된 차량의 경우 Car class로 객체 탐지(오탐)하는 것을 확인할 수 있다.



[그림 3. Misclassification 실험 결과]

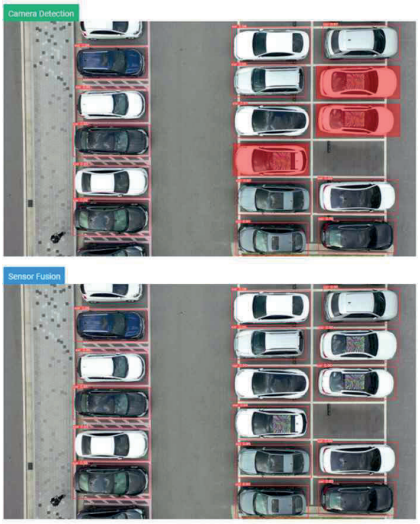
3. 센서 융합 기반 적대적 공격 탐지 실험

본 연구에서는 Yolov5 Object detection 모델 기반 AI 임무 SW에 대한 적대적 공격을 탐지하기 위하여 카메라와 라이다 센서 융합 모델을 구축하였다. 이를 위하여 카메라와 라이다가 설치된 드론을 운영하여 데이터를 수집하였으며, 센서 융합을 위한 캘리브레이션과 라벨링을 수행하였다. 아래는 캘리브레이션 결과 예시이다.



[그림 4. 카메라-라이다 캘리브레이션 결과 예시]

캘리브레이션 및 프로젝션을 수행하여 카메라 데이터와 라이다 데이터의 좌표계를 일치시키고 앙상블 기법을 활용한 Late Fusion 모델에 적용하여 센서 융합을 수행한다. 이후 단일 카메라 Object detection 결과와 카메라-라이다 센서 융합 Object detection 결과를 비교하여 적대적 공격을 탐지한다. 아래는 센서융합 기반 적대적 공격 탐지 결과 화면이다.



[그림 5. 센서 융합 기반 적대적 공격 탐지 결과]

4. 적대적 학습 기반 AI 임무 소프트웨어 강건화 실험

위와 같이 센서 융합 기반 적대적 공격을 탐지한 경우 이에 대한 대응으로 적대적 공격 패치를 추출하여 Data augmentation 기법을 활용한 적대적 학습을 수행할 수 있다. 이를 통해 알려지지 않은 적대적 공격 기법에 대해 임무 소프트웨어의 AI 모델을 강건화하여 보다 안전한 AI 임무를 수행할 수 있다. 아래는 탐지된 적대적 공격 패치를 활용하여 적대적 학습을 수행함으로써 강건화된 모델의 임무 수행 결과 화면이다. 공격 패치의 방향, 크기, 밝기 등의 변화에도 공격이 무력화 되는 것을 확인할 수 있다.

[그림 6. 적대적 학습 기반 AI 임무 소프트웨어 강건화 결과]

Ⅲ. 결론

본 연구는 드론 AI 임무 소프트웨어의 보안 취약점을 실험적으로 검증하고 센서 융합과 적대적 학습을 통해 실질적인 방어 성능 향상을 확인하였다. 특히 센서 융합 기반 탐지와 적대적 학습을 통한 AI 임무 모델 강건화는 실전 환경에서 유효한 대응책으로 입증되었다. 향후 연구에서는 실제 Physical domain에서의 실증 실험과 함께 AI 임무 드론의 실 운용환경에 대한 유효성 검증이 필요하다.

참 고 문 헌

[1] 윤여정, 정일훈, 정승훈, 김진국, 최정완, 박정찬, “드론 AI 임무 소프트웨어에 대한 적대적 공격 탐지 및 대응 시스템 분석”, 2024 한국통신학회 하계종합학술대회, pp. 670-671.