드론 AI 임무 소프트웨어에 대한 적대적 공격 탐지 및 대응 시스템 분석

윤여정*, 정일훈, 정승훈, 김진국, 최정완, 박정찬

국방과학연구소

*yjyoon.rufina@gmail.com

Analysis of adversarial attack detection and response system for drone AI mission software

Yoon Yeojeong, Jung Ilhoon, Jeong Seunghoon, Kim Jingoog, Choi Jeongwan, Park Jeongchan

Agency for Defense Development

요 약

최근 무인 항공기 기술이 빠르게 발전함에 따라 다양한 분야에서 UAV의 활용도가 높아지고 있다. 특히 군사 분야에서는 소형 드론을 이용하여 정찰, 유도, 감시, 통신/정보 중계 등의 임무를 수행하고 있으며, 그 중 드론 카메라를 활용한 정찰, 감시 임무는 전장에서 매우 유용하게 사용될 수 있다. 하지만 이러한 AI 임무 소프트웨어의 경우 특정 객체(object)에 적대적 패치(adversarial patch)를 붙이는 등의 방식을 통하여 임무 수행을 실패하게끔 하는 적대적 공격(adversarial attack)에 취약하다. 본 논문에서는 이에 대한 탐지 및 대응 시스템의 필요성을 분석하고 주요 기능을 식별한다.

Ⅰ. 서 론

드론이 점차 소형화, 경량화 됨에 따라 드론은 여러 산업군에서 다양하게 활용되고 있으며 군사 분야에서도 정찰, 유도, 감시, 통신/정보 중계 등의 임무를 수행하기 위해 소형 드론을 사용하고 있다. 드론 카메라를 이용한 정찰, 감시, 타켓팅 임무 등을 수행하는 경우 실시간 객체 인식을 위해서는 드론 자체적으로 AI 임무 모델을 운용해야 하며, 이는 적대적 공격(adversarial attack)의 위협을 내포한다.

적대적 공격의 주요 원리는 대상 이미지에 작은 변화를 주어 딥 네트워크가 판단할 때의 결정 경계(decision boundary)를 넘겨 잘못 인식하게 만드는 것이다.

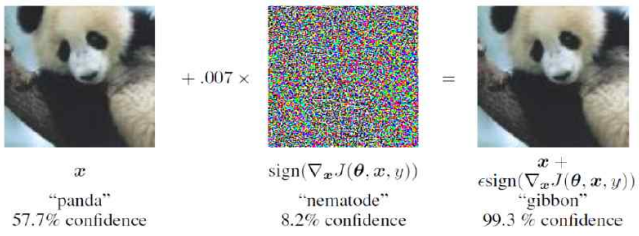

그림 1. 적대적 공격의 예[1]

적대적 공격은 교통 표지판을 대상으로 수행할 경우 자율 주행 자동차가 사고를 내도록 유도할 수 있는 등 보안, 안전이 중요한 여러 분야에서 큰 위협이 될 수 있다. 특히 주요 군사 작전을 수행해야 하는 AI 임무 소프트웨어를 탑재한 드론의 경우, 이와 같은 적대적 공격을 탐지하고 대응하는 시스템이 반드시 필요하다. 따라서 본 논문에서는 드론 AI 임무 소프트웨어를 대상으로 하는 적대적 공격을 탐지하고 대응하는 시스템에 대한 필요성을 분석하고 주요 기능을 식별한다.

Ⅱ. 본론

1. 드론 AI 임무 소프트웨어 정의

드론 AI 임무 소프트웨어에 대한 적대적 공격 및 탐지/대응 시스템의 주요 기능을 식별하기 위하여 먼저 드론 AI 임무 소프트웨어에 대한 정의가 필요하다. 본 논문에서는 드론 카메라를 통해 영상 및 이미지 정보를 수집하고, 이에 대한 객체 인식(Object detection) 및 분류(Classification)를 수행하는 AI 모델을 탑재한 소프트웨어를 드론 AI 임무 소프트웨어로 정의한다.

2. 적대적 공격 정의 및 탐지/대응 시스템 필요성

드론 AI 임무 소프트웨어의 기능을 객체 인식 및 분류로 정의하였으므로, 이에 대한 적대적 공격은 특정 객체에 적대적 패치를 붙임으로써 객체 인식을 불가능하게 하며, 오분류(Misclassification)를 수행하도록 하는 것으로 정의한다. 이를 각각 Object Detector 공격, Misclassification 공격으로 명명한다. Object Detector 공격은 드론 AI 임무 소프트웨어가 특정 객체(전차, 적군, 초소 등)를 인식하는 기능을 마비시키는 것으로 아래 그림과 같이 특정 객체에 적대적 패치를 붙임으로써 수행이 가능하다.

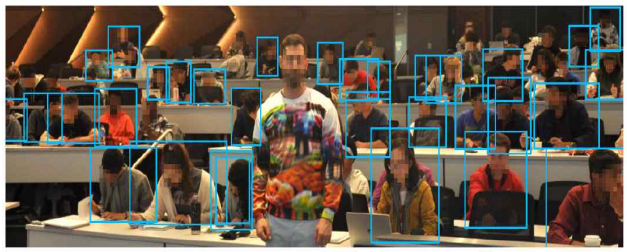

그림 2. Object Detector 적대적 공격의 예 [2]

또한 Misclassification 공격은 Object Detector 공격과 같이 공격 대상 객체에 대한 바운딩 박스(Bounding Box)를 유실시키는 것을 넘어서, 해당 객체를 다른 객체로 오분류 하게끔 적대적 패치를 생성하고 이를 이용하여 공격을 수행한다. 이는 아래 그림과 같이 멈춤 표지판을 스포츠 공과 같은 전혀 다른 객체로 오분류하게끔 AI 모델을 공격한다.

그림 3. Misclassification 적대적 공격의 예[3]

전장 중 드론 AI 임무 SW를 대상으로 위와 같은 공격을 수행할 경우 작전 임무 실패를 초래할 수 있다. 예를 들어 정찰 드론이 적대적 공격을 당할 경우 적군의 전력 파악에 오류가 생길 수 있으며, 자폭 드론에 적대적 공격을 수행할 경우 아군에 대한 공격을 유도할 수도 있다. 따라서 정확한 임무 수행과 임무 수행 결과 무결성 검증을 위하여 다양한 AI 임무 소프트웨어에 대한 적대적 위협을 식별하고, 이에 대한 탐지 및 대응 시스템을 반드시 구축할 필요가 있다.

3. 적대적 공격 탐지 요구사항 식별

위에서 서술한 대로 카메라 객체 인식 및 분류 모델은 적대적 공격에 취약하므로, 최근에는 카메라 RGB 이미지와 LiDAR Point Cloud 이미지를 퓨전하여 객체 인식 및 분류를 수행하는 센서퓨전 모델이 활발하게 연구개발되고 있다.

그림 4. 카메라-LiDAR 센서 퓨전 예[4]

위 그림과 같이 카메라와 LiDAR 데이터를 센서퓨전하여 AI 임무를 수행할 경우, 카메라에 대한 적대적 공격은 LiDAR 데이터에 의해 무효화될 수 있다. 하지만 일반 자율주행차량이나 군용 전차와는 다르게, 소형 드론의 경우 비행 시 사용 가능한 전력량에 제한이 있어 타 센서에 비해 상대적으로 무거운 LiDAR를 달고 모든 임무 비행을 수행하기에는 한계가 있다. 따라서 본 논문에서 정의한 드론 AI 임무 소프트웨어에 대한 적대적 공격을 탐지하기 위해 탐지를 위한 특정 LiDAR 드론을 활용하여 임무 수행 데이터를 수집하고, 이를 서버에 전송하여 카메라 임무 수행 결과와 카메라-LiDAR 센서 퓨전 임무 수행 결과를 비교함으로써 적대적 공격 여부를 탐지할 수 있다. 또한 이와 같은 탐지 기능을 통하여 임무 수행 결과에 대한 무결성을 검증할 수 있다.

4. 적대적 공격 대응 요구사항 식별

적대적 공격 탐지를 위해 서버에서 카메라-LiDAR 센서 융합을 수행할 경우 드론에서 임무를 수행하는 것에 비해 실시간성이 떨어질 수 있다. 또한 임무 수행 중 탐지를 위한 LiDAR 드론의 지속적 비행이 어려움에 따라 궁극적인 목표는 적대적 공격에 강건한 AI 임무 모델을 구축하고, 이를 카메라 임무 드론에서 운용하는 것이다. 이를 위하여 적대적 학습 기법을 활용할 수 있다.

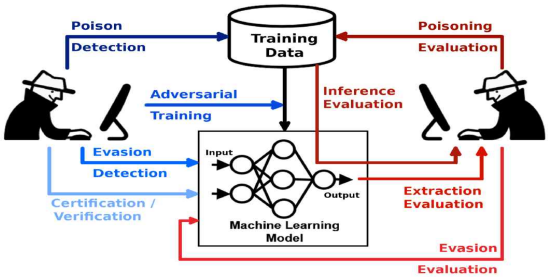

그림 5. 적대적 학습 기법[5]

적대적 학습 기법은 기존에 알려진 적대적 공격 알고리즘을 이용하여 Random Initialization Perturbation을 생성하고 학습하는 Multi-Perturbation 기법과 공격에 유효한 적대적 패치에 변형(밝기, 사이즈, 회전 등)을 적용하여 학습하는 Data Augmentation 기법 등이 있다. 적대적 공격에 대응하기 위하여 Multi-Perturbation 기법을 적용함으로써 기존에 알려진 적대적 공격 기법에 대한 모델 강건화를 수행하며, 탐지 시스템에서 탐지된 알려지지 않은 적대적 공격 패치를 활용하여 Data Augmentation 적대적 학습을 수행함으로써 알려지지 않은 적대적 공격에 대하여 대응할 수 있다. 이와 같은 모델 강건화를 통해 기존에 알려진 적대적 공격 기법과, 알려지지 않은 적대적 공격 기법에 모두 대응함으로써 보다 강건하고 신뢰도 있는 AI 임무 모델을 구축할 수 있다.

Ⅲ. 결론

본 논문에서는 드론 AI 임무 소프트웨어 대상의 적대적 공격에 대하여 분석하고 적대적 공격 탐지 및 대응 시스템의 필요성에 대하여 서술하였다. 또한 적대적 공격 탐지 및 대응 시스템에 대한 주요 기능을 식별하였다. 이를 통하여 카메라를 이용한 임무를 수행하는 소형 AI 임무 드론에 대한 위협을 탐지/대응함으로써 보다 강건한 AI 임무 모델을 구축할 수 있다.

참 고 문 헌

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[2] Z. Wu, S.-N. Lim, L. S. Davis, and T. Goldstein, “Making an invisibility cloak: Real world adversarial attacks on object detectors,” in Proc. Eur. Conf. Comput. Vis., in Lecture Notes in Computer Science, vol. 12349. Cham, Switzerland: Springer, 2020, pp. 1 - 17, doi: 10.1007/978-3-030-58548-8_1.

[3] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. 2018. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 52 - 68.

[4] G. A. Kumar, J. H. Lee, J. Hwang, J. Park, S. H. Youn, and S. Kwon, “LiDAR and camera fusion approach for object distance estimation in self-driving vehicles,” Symmetry, vol. 12, no. 2, p. 324, Feb. 2020.

[5] What are adversarial attacks inmachine learning and how toprevent them? https://www.labellerr.com/blog/what-are-adversarial-attacks-in-machine-learning-and-how-can-you-prevent-them/