KI-1-AI-Sec — Adversarial-Attacks-in-ML-based Remote-Sensing-Applications
CS 4850 - Section 02 – Fall 2023 2023-11-26
Sharon Perry
ki-1-ai-sec.github.io
github.com/KI-1-AI-Sec/adversarial-yolo

 Kyle Bratcher	 Zeke Lipscomb (Team Lead)
 Babrah Koroma	 Matthew Krupczak

Team Members:

Name	Cell Phone	Email
Dr. Kazi Aminul Islam (Client)	470.578.6005 (Office phone)	kislam4@kennesaw.edu
Kyle Bratcher	678.521.9164	kylesbratcher@gmail.com
Zeke Lipscomb (Team Lead)	770.408.8145	Zekeclass@gmail.com
Babrah Koroma	404.952.0637	Babrah.6766@gmail.com
Matthew Krupczak	678.778.7335	matthew@krupczak.org
Sharon Perry (Advisor)	770.329.3895	Sperry46@kennesaw.edu

Table of Contents:

Introduction	**3**
Background/Significance	**4**
Goals and Objectives	**6**
Methodologies	**6**
Challenges and Risks	**8**
Preliminary Results	**10**
Project Management and Team Dynamics	**10**
Conclusion and Future Direction	**10**
Reference	**11**

Introduction

Adapting Advanced AI for Enhanced Object Detection Security

In the rapidly evolving domain of artificial intelligence (AI) and machine learning, our Senior Project group, KI-AI-Sec, in CS 4850 S03 for Fall 2023 at Kennesaw State University (KSU), takes on the challenge of enhancing the security of AI-driven object detection systems. Comprising team members Babrah Koroma, Zeke Lipscomb, Kyle Bratcher, and Matthew Krupczak, under the guidance of Dr. Kazi Aminul Islam and with support from Sumaiya Tasneem, our project navigates the complexities of adapting adversarial AI techniques to newer generation object detection frameworks. Mrs. Sharon Perry, as the coordinator of the Senior Project program, oversees our progress and provides invaluable academic support.

The advent of massively parallel computer processors and advanced machine learning algorithms has revolutionized the way AI frameworks understand and interpret diverse information sources. These technologies empower computers to recognize and process non-rigid patterns, such as handwriting, speech, written words, and notably, the identification of objects in images or video feeds. Our daily lives are increasingly intertwined with these AI systems, as they find applications in various sectors ranging from financial markets to physical security systems, automotive driver assistance, and human-computer interfaces.

Our project zeroes in on the realm of object detection for remote sensing applications, a critical component in security-sensitive systems. These systems are pivotal in scenarios ranging from trespasser recognition in security cameras to the analysis of satellite imagery for military and commercial purposes. However, despite their widespread utility, the security vulnerabilities inherent in these systems remain underexplored and pose significant risks.

The crux of our research focuses on addressing these vulnerabilities. We aim to adapt and refine an existing adversarial AI attack pattern generator, initially designed to work with older AI models, for compatibility with a newer generation object detection framework, namely YOLOv8. The goal is to evaluate whether these advanced AI frameworks remain susceptible to adversarial attacks, a phenomenon where malicious entities manipulate AI models to fail in correctly identifying objects.

Our project is not just about identifying vulnerabilities but also about contributing to the development of more secure AI systems. The insights and outcomes from our project are intended to assist Dr. Islam's research group in creating a defense mechanism that can safeguard remote sensing applications from such adversarial attacks. This endeavor places our team at the forefront of research that aims to bolster the security of AI systems against emerging threats, ensuring their reliability and integrity in critical applications.

Background/Significance

The emergence and advancement of Convolutional Neural Networks (CNNs) have marked a significant milestone in the field of computer vision. This type of object classifier is inspired by biological processes in the connectivity pattern between neurons in animals' visual cortex of the brain. Digital CNNs take an image as input, where it is fed through the entire network. These networks, with their data-driven, end-to-end learning process, have shown exceptional results in a wide range of computer vision tasks. A typical CNN learns to identify basic filters at the lower layers and abstract high-level features at the upper layers, containing millions of parameters (which are analogous to individual neurons). While this architecture yields highly accurate models, it also leads to decreased interpretability, making it challenging to understand the rationale behind certain classifications by the network, such as identifying a person in an image.

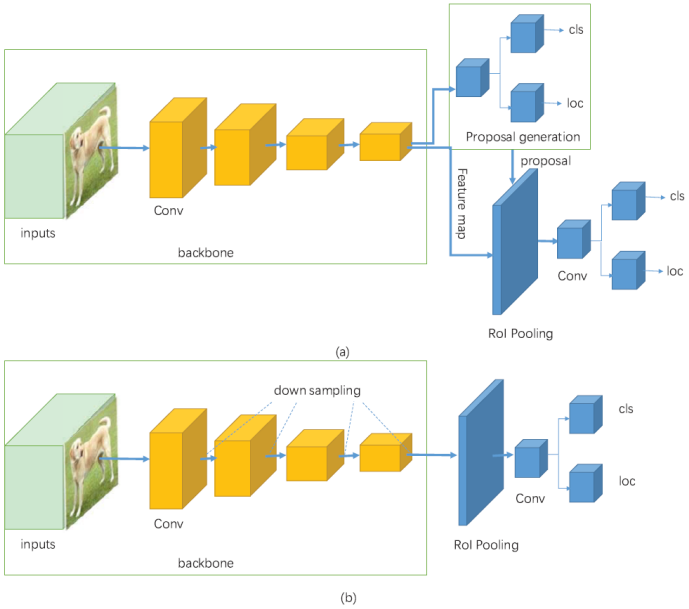

Figure 1: A Convolutional Neural Network contains multiple hidden layers which allow it to model the complexity involved in analyzing visual images. While this technique produces good object detection accuracy, it is difficult to explain how the "black box" arrives at its conclusions.

CNNs have allowed machine vision systems to perform much faster and more accurate classification than previous approaches. However, neural networks are vulnerable to adversarial attacks.

Adversarial attacks on machine learning models, particularly classification tasks, have become a significant concern alongside the growth of CNNs. The early stages of these attacks, as noted by Bigio et al. in 2014 and later expanded by Szegedy et al., involved subtle manipulations of image pixel values to mislead the network into incorrect classifications. The development of these attacks evolved rapidly, with methods like the fast gradient sign method by Goodfellow et al. making it more practical to create adversarial images. These methods demonstrated that even minute changes in an image could lead to drastically different outcomes from the AI model.

By modifying the image with a specifically targeted pattern of noise, the object detector can be defeated:

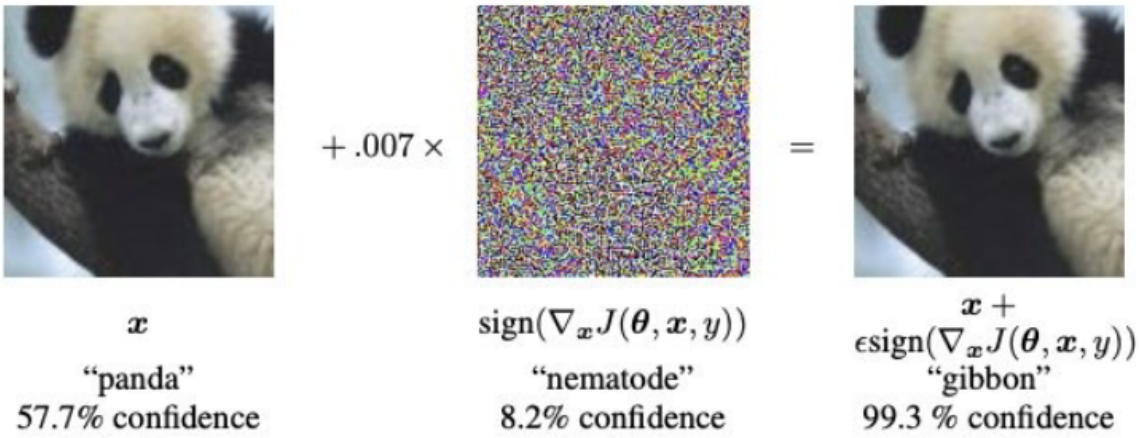

Figure 2: The addition of a specially-crafted noise pattern generated by an adversarial model confuses a detector CNN into thinking an image of a panda is actually a gibbon. The difference to the original image is imperceptible to the human eye.

To a human, the images look identical, however the noise is enough to disrupt an image classifier to recategorize a panda as a gibbon. This approach can also be used to disrupt a CNN's inference that an image contains an object at all. Thys et al. 2019 developed a method to bring this attack into the real world. By placing a "patch" within an image, and carefully tuning the noise within the patch, an object detection framework can be defeated in a real world setting.

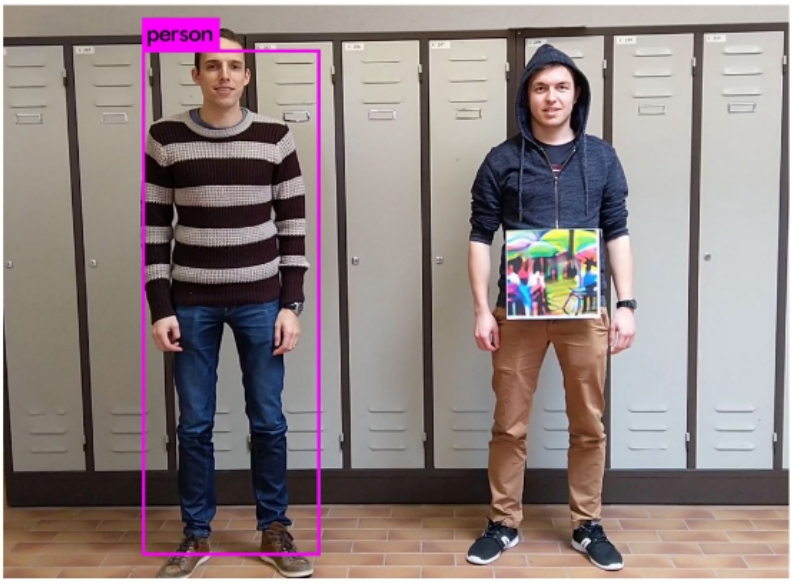

Figure 3: A printout of an adversarial pattern allows an attack to become invisible to a security camera

In the context of object detection, adversarial attacks have been demonstrated in real-world scenarios. For instance, Chen et al. focused on fooling object detectors like Faster R-CNN in detecting stop signs by applying patterns across the entire area of the signs. Similarly, Eykholt et al. used the YOLOv2 detector for a white box attack on stop signs, proving that these attacks could transfer to other detectors. These examples underline that objects with fixed visual patterns, like traffic signs, are vulnerable to such attacks.

The primary goal of our work is to generate adversarial patches that can fool an airplane detection system for satellite imagery in real-world scenarios. This is a more complex challenge compared to previous work on uniform objects like stop signs, due to the highly variable appearance of airplanes in different contexts and orientations. The approach involves optimizing an image to minimize probabilities related to the appearance of an airplane as detected by the model. This optimization process includes various transformations to make the patches robust for real-world applications, such as rotation, scaling, and adjusting brightness and contrast. In addition, a Non-Printability Score (NPS) and Total Variation (TV) score are included in the optimization function. The inclusion of these metrics helps to ensure the end result is capable of being printed using a normal consumer color printer and will be effective under a variety of viewing angles and lighting conditions.

Goals and Objectives

Thys et al conducted their research using a CNN called You Only Look Once Version 2 (YOLOv2). Our project seeks to use a newer version, YOLOv8, to investigate vulnerabilities within the field of remote sensing. To this end, our project has three major goals. Firstly, a new model trained on a remote-sensing dataset with a newer version of YOLO. This model serves as a target to develop a new adversarial attack. Secondly, the inference pipeline of Thys. et al is modified to target this newer model. Thirdly, the new inference pipeline is used to generate adversarial patches targeting this newer model. This entire process is documented extensively, to build a new pipeline for developing an adversarial attack. Future students under Dr. Islam will be able to use the documentation provided to both advance our current research and develop new adversarial attacks on other frameworks.

Methodologies

Our group started with the adversarial agent code published by Thys et al. 2019 for their experiment which attacks a person detector trained with the INRIA dataset (which contains labeled persons within pictures).

Figure 4: Patch pattern generated by our group in replicating Thys et al.'s experiment.

Our group trained a new object detection model to be the target of our attack for remote sensing. We used the YOLOv8 framework to train a new model on Aircraft Detection from Airbus High-Resolution Satellite Imagery Dataset. This dataset is published by Airbus Defense and Space from imagery captured by their twin Pleiades satellites at 50cm resolution. The dataset contains 103 2560x2560 pixel images covering 1280 meters², each of various airports worldwide.

Airplanes within each image had been hand-labeled in the dataset with a polygon mask. Our group adapted a Jupyter notebook on Kaggle to a Python script which converted each polygon mask for an airplane into a bounding box rectangle in YOLO format.

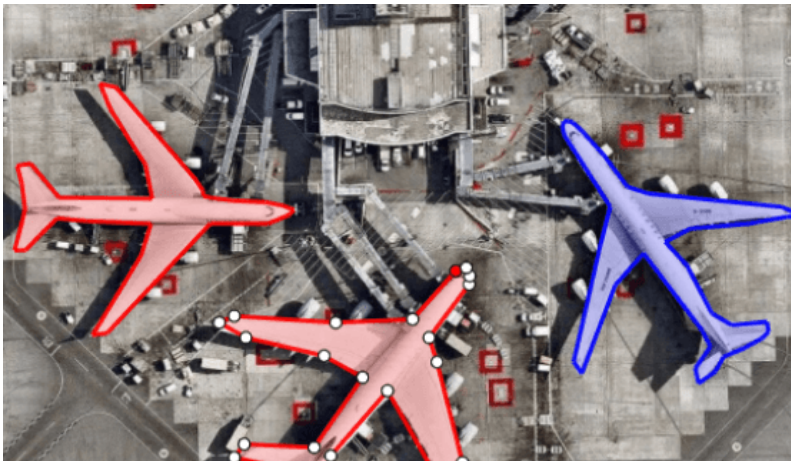

Figure 5: a visualization of polygon masks as appear in the unmodified dataset

The script also broke the images up into 2952 512x512 tiles with 64 pixels overlap. This resultant data was used to train our YOLOv8 airplane object detection model.

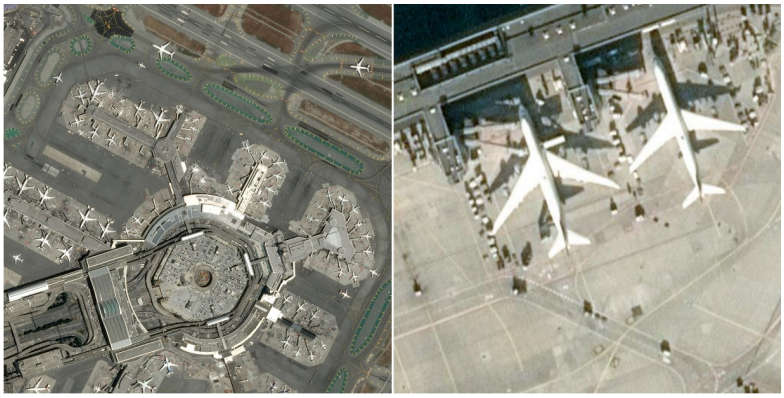

Figure 6: An unmodified 2560x2560 satellite image from the original dataset (left)
One of 2952 512x512 tiles randomly generated by our script (right)
Label data is not shown in either image

Once we finished training this new object detection model, we modified the adversarial model trainer so it could target it in a 'white box' attack scenario. In such a scenario, it is assumed that an attacker has access to a copy of the object detection model they wish to disrupt with the placement of an adversarial patch pattern. Our modifications involved replacing the trainer's YOLOv2 image inference pipeline with the newer YOLOv8 framework and writing our own custom data loader and training configuration. The adversarial model is trained iteratively by the trainer. The adversarial model generates a patch (initially random) and the trainer places it within the bounding box of every target object within an image. Then, the model's loss function is evaluated by the trainer. The weights of this model are modified over subsequent iterations by the adversarial trainer to minimize its loss function using a technique called stochastic gradient descent. Training the adversarial model in such fashion affects the adversarial model's patch output rather than the victim detection model. Object Detection confidence loss, Non-printability score, and a total variation score are included in the trainer's loss function so as to improve the real-world attack capabilities of the adversarial model. The inclusion of these scores incentivizes the trainer to produce an adversarial model that generates patches that prevent AI models from recognizing the target object yet can be printed and deployed in an attack in a real-world setting.

Challenges and Risks

Adapting to YOLOv8: Architectural Challenges and Hardware Limitations

The process of adapting an adversarial AI attack pattern generator, originally designed for the YOLOv2 framework, to the newer YOLOv8 presented a series of significant challenges and risks, both in terms of technical architecture and hardware requirements.

Architectural Differences between YOLOv2 and YOLOv8:

One of the major challenges we faced was the substantial architectural differences between YOLOv2 and YOLOv8. These differences are not merely incremental updates but represent fundamental shifts in how these models process and interpret visual data. YOLOv8, being a more advanced version, incorporates newer methodologies and algorithms that were not present in YOLOv2. This required our team to deeply understand these architectural changes and develop novel approaches to adapt the adversarial attack model accordingly. The task was akin to retrofitting a new engine into a different make and model of a car, demanding significant modifications and tuning to ensure compatibility and optimal performance.

Hardware Limitations and Computational Challenges:

Our project's success heavily relied on the ability to train the adversarial pattern generation model, a process that is computationally intensive and requires robust hardware support. We encountered several hurdles related to the specific hardware needed for this training. Initially, we considered using Google Colab for its access to NVIDIA GPUs, but the free tier offered limited time, making it impractical for our extensive training requirements. One of our team members had a consumer GPU at home, but its memory capacity was severely constrained, posing a bottleneck for the kind of large-scale data processing we needed.

Kennesaw State University's High-Performance Computer:

In search of a solution, we turned to Kennesaw State University's High-Performance Computer (HPC). This resource proved invaluable, offering the computational power necessary to train our model effectively. However, gaining access to the HPC was not without its challenges. We had to navigate the process of obtaining authorization, which was time-consuming. Once access was granted, we faced technical challenges in setting up and running jobs reliably on the system. The use of TORQUE scripts, essential for managing our tasks on the HPC, presented difficulty because it provided no useful debugging output during development. We spent considerable time debugging and adjusting these scripts to ensure they worked correctly and efficiently.

Moral risks and ethical considerations

The primary motivation for this project is to produce an authentic attack scenario that may be used by researchers to build more robust object detection models for remote sensing that are harder to fool. We needed to consider the ethical implications of developing an attack scenario carefully to ensure that its creation or publication would pose no undue risk. We analyzed whether such an attack scenario would be of greater help to researchers developing defense techniques than the potential harm that could be imposed on system operators and their bystanders by a malicious actor. The reasons for our conclusion that the work should proceed are two-fold: First, the attack scenario, while requiring significant adaptation for the newer framework, was not novel. State actors, or even malicious actors of far lesser resources could perform the same type of work as our project but without making it available to researchers for

the development of defense techniques (akin to a “zero-day” vulnerability, which cybersecurity researchers prefer to discover and disclose responsibly rather than allow to be hoarded privately). Secondly, the fields of adversarial AI attacks and even AI object detection, in general, are relatively new and our work can influence the course of development of object detection technology to be safer and more reliable. This is not a conclusion we came to lightly, as it merited significant thought to ensure our work furthered the public good and was in line with the ethical norms of cybersecurity.

Preliminary Results

We have trained a YOLOv8 object detection model to recognize airplanes using the Airbus Satellite Imagery dataset. This model had a Mean Average Precision 50(mAP50) score of 0.91 with Precision and Recall scores of 0.96 and 0.85, indicating good object detection performance on the aircraft dataset. We were then able to modify the inference pipeline for the adversarial trainer to target this YOLOv8-based remote sensing airplane object detection model rather than the YOLOv2-based person detection one as in the original experiment.

Currently, our adversarial model may generate a rudimentary patch that targets our YOLOv8 aircraft detection model. However: this patch only slightly reduces the detection capabilities of the model, not enough for an effective attack. Application of the patch currently has a negligible effect on the targeted model’s mAP score, whereas an effective attack would be expected to lower this score considerably. We are refining our implementation to allow the attack portion to process the intermediate tensors of the detection model transparently. In other words, the architecture of YOLOv8 is different enough from YOLOv2 to present significant challenges in adapting the attack technique. It requires not only the outputs of the model itself but also some internal information from the detection model itself (known as a “whitebox” attack scenario) to generate an effective adversarial pattern. An ongoing concern of our research is to determine whether an object detection model created with the newer YOLOv8 framework is still vulnerable to this kind of adversarial attack. We are adjusting our methodology, including our code modifications to the trainer and tuning for the trainer’s hyperparameters. Whether or not this method of attack is viable against YOLOv8 as it was for YOLOv2 will be an important result for the topic of security against adversarial attacks in remote sensing and will inform future research for creating more attack-resistant models.

Project Management and Team Dynamics

Conclusion and Future Direction

As our Senior Project culminates, we reflect on the strides we have made in constructing an adversarial patch attack against a remote sensing YOLOv8 model. Building upon the foundational work of Thys et al., we have ventured into uncharted territories, tackling the complex challenges presented by advanced machine vision systems. While our current outcomes have not reached the level of performance we initially envisioned, the journey has been a rich learning experience, paving the way for future advancements in this field.

Our iterative process, though met with hurdles, has led to a comprehensive understanding of the nuances involved in creating effective adversarial attacks for newer AI frameworks. Dr. Islam, who has been closely involved in guiding and mentoring our efforts, acknowledges the value of our work so far. The insights we have gathered, coupled with the modifications and strategies we have developed, will serve as a robust foundation for his research group. This foundational work is poised to facilitate further explorations into the vulnerabilities present in remote sensing applications and the development of strategies to mitigate these vulnerabilities.

Looking ahead, the project's trajectory is set to encompass a broader scope beyond YOLOv8 to include a variety of object detection frameworks. This expansion is critical, as it aligns with the evolving nature of AI technologies and the need for comprehensive security measures that keep pace with these advancements. The continuation of this project under Dr. Islam's guidance ensures that future researchers will be well-equipped with the tools and knowledge to fortify object detection frameworks against sophisticated adversarial attacks.

The contribution of our project extends beyond the confines of our university lab. It plays a pivotal role in shaping the future of AI security, particularly in the domain of remote sensing. As AI continues to integrate deeper into critical systems, the importance of ensuring its resilience against adversarial threats becomes increasingly paramount. Our research, though at its nascent stage, is a step towards creating a safer, more secure digital world where AI systems can be trusted to perform with integrity and reliability.

Reference

Thys, S., Van Ranst, W., & Goedemé, T. (2019). Fooling automated surveillance cameras: adversarial patches to attack person detection. In CVPRW: Workshop on The Bright and Dark Sides of Computer Vision: Challenges and Opportunities for Privacy and Security.

Airbus Defense and Space. (n.d.). Sample Aircraft Detection Dataset from Airbus High-Resolution Satellite Imagery [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/airbusgeo/airbus-aircrafts-sample-dataset