C-Day Abstract

With advancements in Artificial Intelligence(AI) and machine learning technology, intelligent machines have become a part of everyday life. This technology however is susceptible to disruption by a malicious attacker, whereby data input is carefully manipulated to defeat an AI system. One such attack is the development of an adversarial patch that can be applied to an image to defeat an AI object detection system. Protecting against these attacks is especially important in remote sensing applications such as satellite image analysis done by public sector government organizations such as the EPA, DoD, and NGA. In these applications, AI is trusted to automatically process larger volumes of data by identifying meaningful features more quickly than would be possible by humans. AI's defeat by malicious actors could have implications for the security of persons and property. Within this project we will explore the vulnerabilities of modern AI vision systems by developing an adversarial patch attack and identifying mitigation techniques in remote sensing.

Introduction

We are Senior Project group KI-AI-Sec in CS 4850 S03 for Fall 2023 at KSU. Our group consists of Babrah Koroma, Zeke Lipscomb, Kyle Bratcher, and Matthew Krupczak, advised by Dr. Kazi Aminul Islam and Sumaiya Tasneem. Mrs. Sharon Perry is the coordinator of the Senior Project program.

The advent of massively-parallel computer processors and Machine Learning has in the past decade allowed for the training of useful AI frameworks based on neural networks and transformers. These systems allow computers to understand and interpret a variety of information sources which do not adhere to rigid patterns, such as identifying the meaning of handwriting, speech, written words, and the type and placement of objects which appear in images or video.

These AI frameworks are now used in almost every part of daily life, from financial markets and physical security systems to automotive driver assistance and human computer interfaces. The benefits of such systems are myriad, however the security vulnerabilities such systems may be subject to are underexplored.

Our research concerns the discipline of object detection for remote sensing applications and their possible vulnerabilities. Neural network based object detection frameworks are used in many security-critical systems, such as for security cameras to recognize trespassers or analysis of satellite imagery by the U.S. Navy and Army. In addition, they are frequently used for commercial applications by firms such as Google, Uber, etc. for automatically interpreting vast amounts of data.

Malicious actors may desire to disrupt such detection, for example a robber might want to avoid being detected by a security camera or a foreign air force may wish to disguise its planes. To do so, an adversarial attack technique has been demonstrated by Thys et al. 2019. For this technique, an adversarial AI model can be trained to produce a pattern which, when placed on an object, fools an object detection framework and causes it to fail to recognize the object.

Our senior project group, under guidance from Dr. Islam, is adapting the research of Thys et al. 2019. Their adversarial model trainer is designed to train a patch generator which may disrupt a convolutional neural network (CNN) framework called You Only Look Once version 2 (YOLOv2). Their experiment targets a YOLOv2 model for person detection in imagery, whereas our adaptation allows it to target a newer object detection model trained with YOLOv8 on a remote sensing dataset for airplane object detection. Our adaptations add applicability to the remote sensing domain through analysis of satellite imagery for identification of airplanes. We are seeking to determine if this newer object detection framework is still vulnerable to this kind of attack scenario for remote sensing.

Dr. Islam's research group aims to use the attack our senior project group develops to create a defense system which can protect against adversarial attacks for remote sensing applications. Our attack implementation will be used to test the defense techniques of their research. Their research concerns new object detection techniques which are easier for human operators to interpret and are more resistant to such adversarial attacks.

Methodologies/Analysis

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). 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 the 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 meters^2, 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. 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.

Once we finished training this new object detection model, we modified the adversarial model trainer so it could target it in a 'whitebox' 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 real-world attack capabilities of the adversarial model. The inclusion of these scores incentivise the trainer to produce an adversarial model which 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.

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 which 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. 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 tunings 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.

Conclusion

Our group has dedicated ourselves to the topic of AI security to adversarial attacks in remote sensing applications. The advent of massively-parallel computer processors and Machine Learning have allowed computers to effectively process data in a much broader set of applications than was previously possible. As these systems are increasingly being used for interpreting vast amounts of remote sensing data such as satellite images or video feed, malicious actors are likewise increasingly looking to disrupt the effects of such systems. Our research is testing AI-based systems and informing future research on how to improve their resistance to such attacks.

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