Adversarial Attacks in image data and sensing: A threat to Autonomous Systems

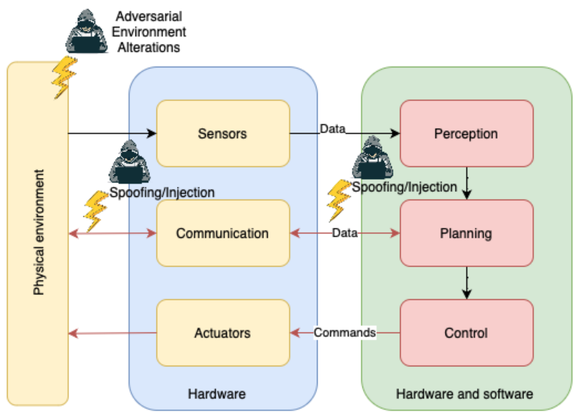
Lancaster University

Researcher: Alvaro Lopez Pellicer
Supervisors: Prof. Plamen Angelov, Prof. Neeraj Suri

Engineering and Physical Sciences Research Council

UKRI Trustworthy Autonomous Systems Hub

Attack Surfaces on Autonomous Systems Stack



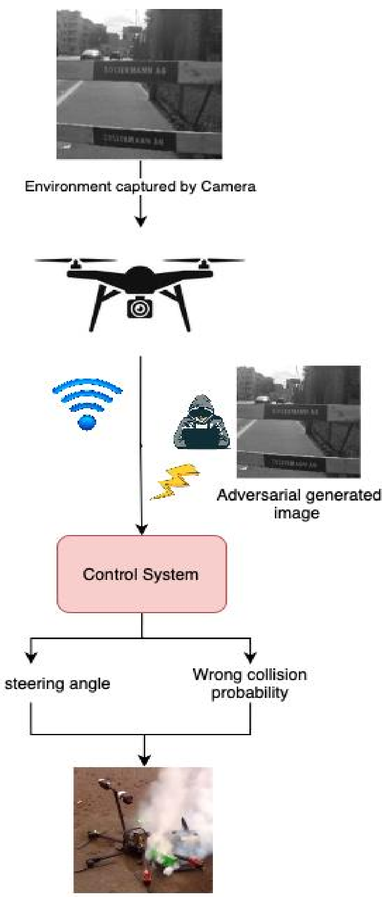
Autonomous systems face numerous challenges in their operation, due to the uncertain and dynamic multi-layer attack surfaces

Adversarial attacks undermine the security and trustworthiness of AS

These attacks can take various forms, such as data poisoning, model inversion, or evasion, and can have serious consequences for the safety, reliability, and privacy

Critical Impacts

- **Perception layer**: Adversarial attacks can manipulate the sensory input of an AS, causing the system to perceive incorrect or misleading information. For example, adversarial examples in computer vision can cause an AS to misclassify objects in the environment, leading to incorrect or unsafe actions.
- **Planning layer**: Adversarial attacks can also manipulate the AS's decision-making processes, leading to incorrect or suboptimal plans. For example, an attacker may introduce false information about the environment or other agents, leading to incorrect or unsafe plans.
- **Control layer**: Adversarial attacks can also affect the control layer of an AS, leading to incorrect or harmful actions. For example, an attacker may manipulate the control signals or inputs to the actuators, causing the AS to take actions that are not in line with its intended behaviour.

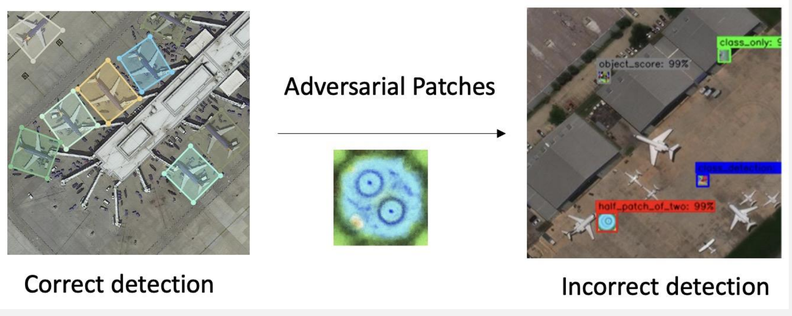


Attacks in the physical environment

Examples include:

Adversarial Stickers
✓ Target object classification

Adversarial patches
✓ Target object detection



Manipulate the environment in order to cause the system to behave in unintended or harmful ways.

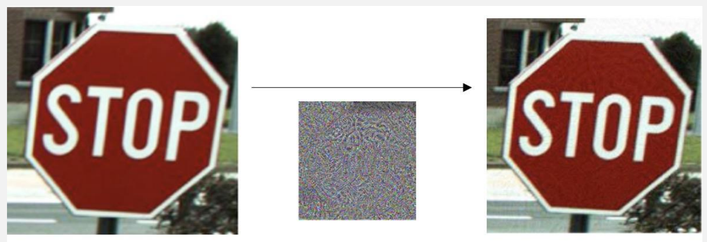
Attacks may be unnoticeable to humans when placed in the real world as they may be mistaken by decorations, urban art or vandalism and not seen as a bigger threat

Attacks on Digital Images

Example:

FGSM

$$\hat{x} = x - \epsilon \cdot sign(\nabla_x J(\theta, x, y))$$

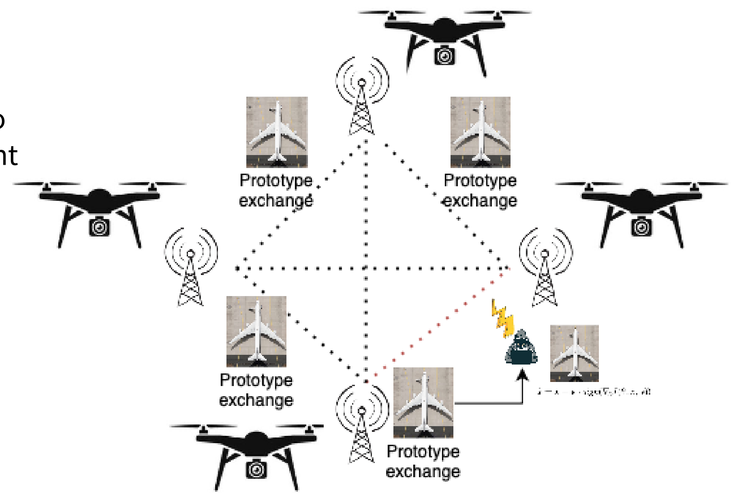


Creating small, carefully crafted perturbations to the input data in order to cause the machine learning model to produce incorrect or undesirable outputs.

Malicious perturbations in protype exchange in a FL environment

An attacker may use digital attacks to inject adversarial examples at different levels of a system such as in a distributed (Federated Learning) environment.

Given a Prototype based FL environment, threats may exist of spoofing in the prototype exchange stage with malicious images



Defence mechanisms

Different defence methods are being developed to tackle these challenges

They may be categorised as:

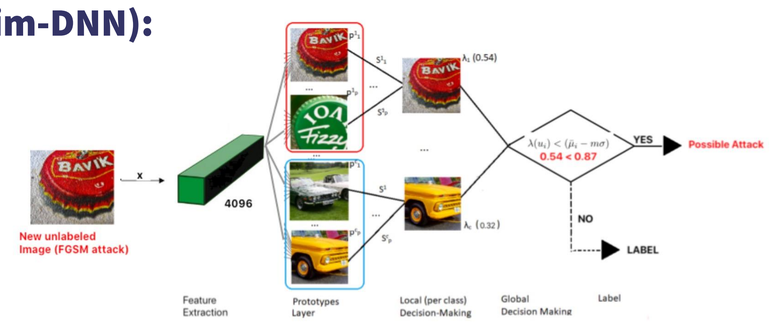
- Adversarial training
- Input data pre-processing
- Detection
- Provable
- Model ensemble
- Model distillation
- Hybrid defences

Requirements for robust to adversarial attacks systems in the context of AS:

- Able to detect attacks
- Able to react to detected attacks
- Evolve with new unknown types of attacks and situations

Similarity-based Deep Neural Network to Detect Imperceptible Adversarial Attacks (Sim-DNN):

Detect adversarial attacks through its inner defence mechanism that considers the degree of similarity between new data samples and autonomously chosen prototypes.



Future work

Robust to adversarial attacks evolving classification

- A prototype based framework able to detect and mitigate digital (noise based) adversarial attacks and learn from new classes
- Following the principle from Sim-DNN, this framework would be able to detect possible attacks or unseen classes.
- After detection, the flagged image will be inputted into a denoising framework which will remove adversarial perturbations (if any) and be able to determine whether the image was attacked or if it is a new unseen class and then create a new prototype for it

Advantages of the proposed framework

- Detect adversarial attacks with more confidence
- Mitigate detected adversarial attacks by removing the attack from the input and correctly reclassifying the image
- Evolving learning of new unseen classes

Disadvantages of the framework

- Potentially ineffective against physical attacks
- Will still have some of the drawbacks from Sim-DNN

Lancaster University

Cranfield University

TAS-S

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]