Jordan Fischer
Solutions Architect
IBM Federal Consulting

Chief Digital and Artificial Intelligence Office

Adversarial AI Robustness Testing & Evaluation Tools

ART and HEART libraries

IBM **Consulting** IBM **Research**

IBM

HEART

Adversarial AI Robustness Evaluation with ART & HEART

IBM's open-source Adversarial Robustness Toolbox (**ART**) provides tools that:

- assess model performance under adversarial attack
- improve model resiliency in case of attack

In collaboration with the CDAO's JATIC program, IBM created the Hardened Extension of ART (**HEART**) with:

- an emphasis on DoD needs and use cases (UAVs, surveillance, etc)
- alignment to best-in-class open-source standards to facilitate AI testing across broader evaluation criteria

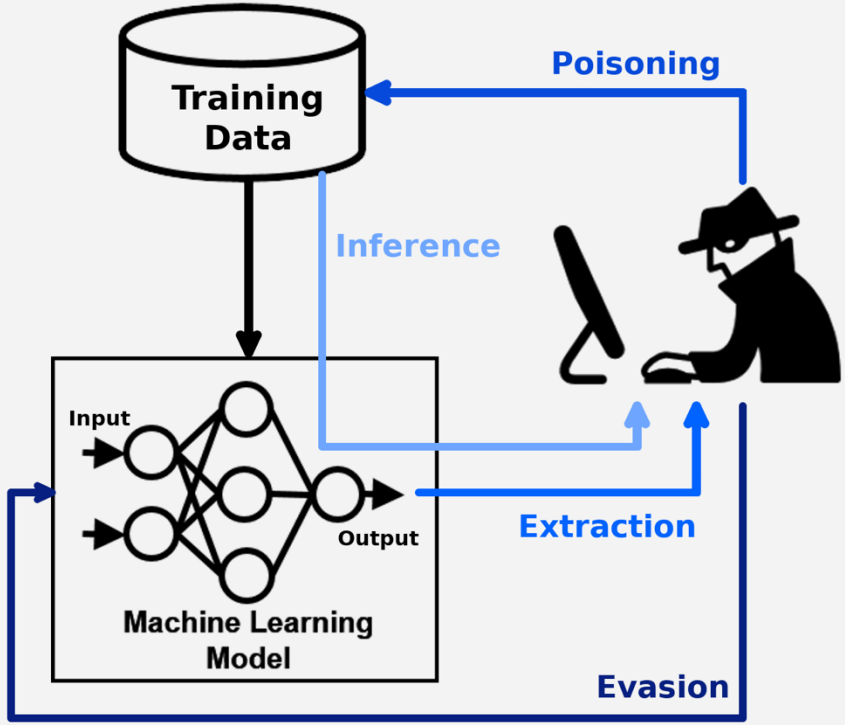
Adversarial Threats to Machine Learning

Adversarial Robustness Toolbox

HEART

Adversarial threats against machine learning models and applications have a wide variety of attack vectors.

- **Evasion:** Modifying input to influence model
- **Poisoning:** Modify training data to add backdoor
- **Extraction:** Steal a proprietary model
- **Inference:** Learn information on private data



Combined Adversarial Threats

Adversarial Robustness Toolbox

HEART

Combinations of adversarial threats become more effective and more dangerous.

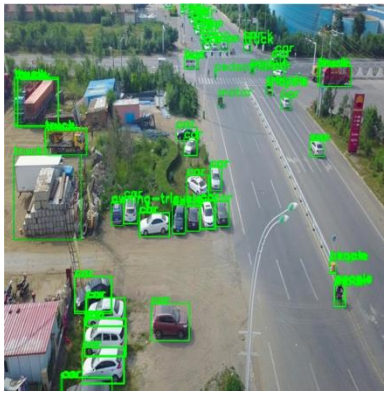
- Extraction attacks enable stronger white-box evasion attacks
- Extraction attacks steal models that could leak private information in inference attacks

Adversarial Perturbation – Difference between Adversarial & Original Image

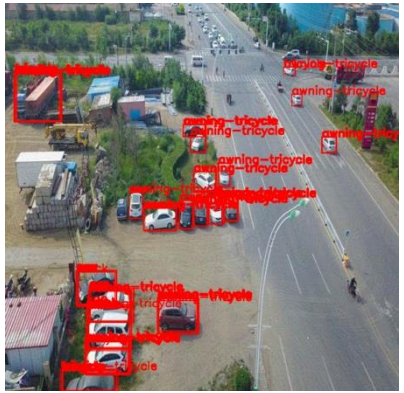
Adversarial Robustness Toolbox

HEART

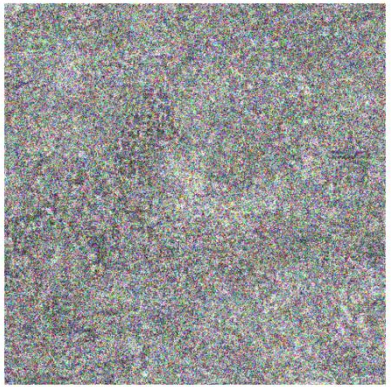
Benign image



Adversarial image



Perturbation

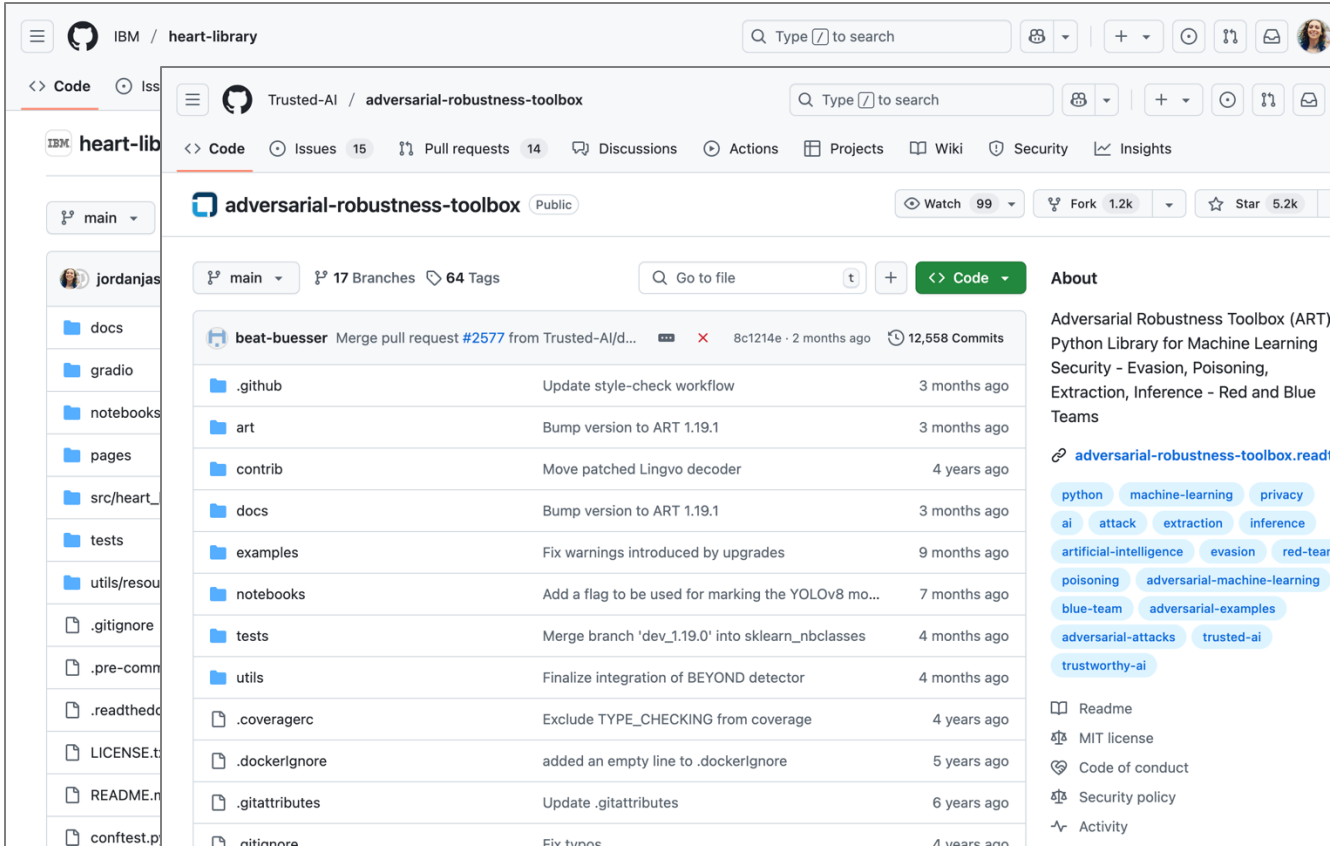


HEART & ART

Adversarial Robustness Toolbox

HEART

- ★ Open-source python libraries (MIT license)
- ★ Step-by-step tutorials
- ★ In-depth documentation
- ★ Support red-/blue-teaming



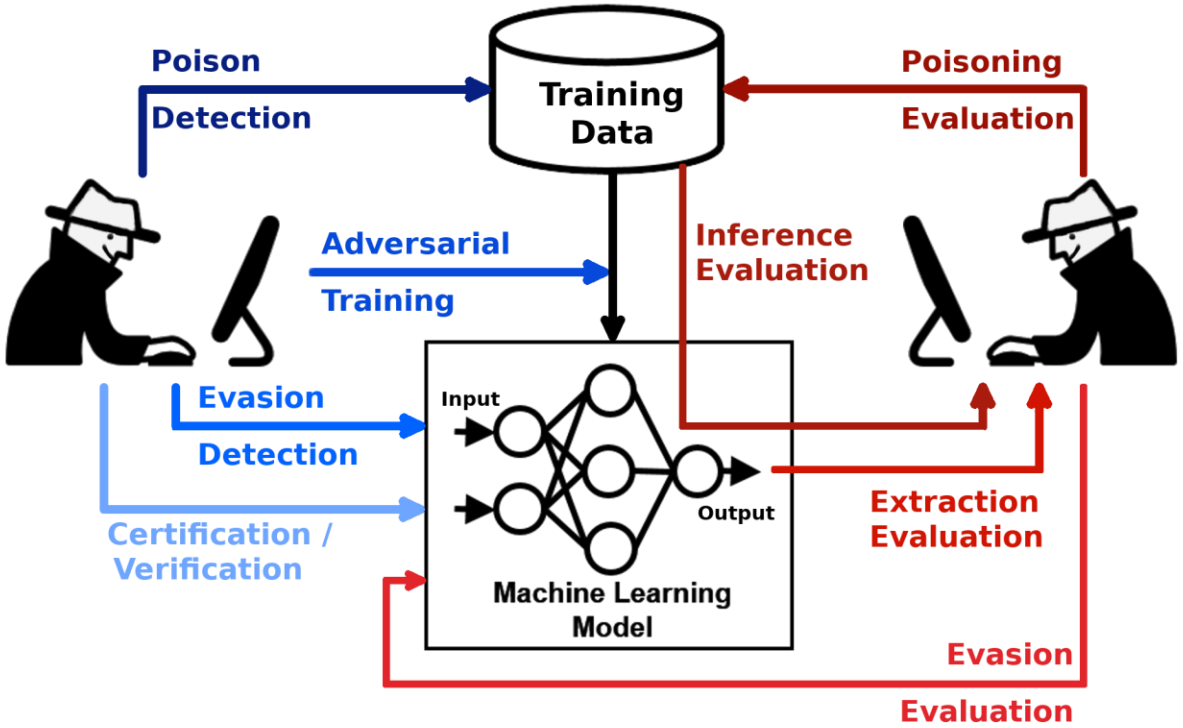
AI Red and Blue Team Approach

Adversarial Robustness Toolbox

HEART

Blue Team tools

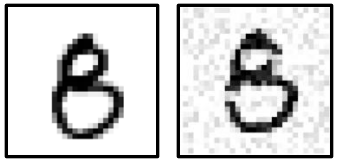
Red Team tools



HEART Supported Attacks

Adversarial Robustness Toolbox

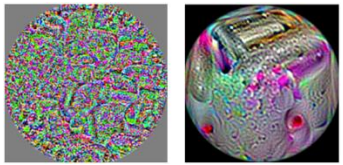
HEART



Projected Gradient Descent (PGD)

A. Madry et al. (2019)

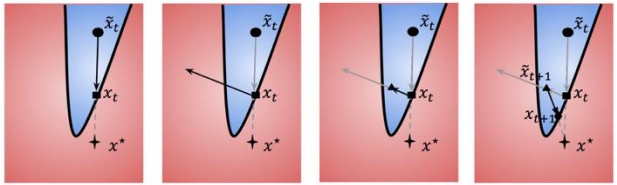
The strongest, worst-case white-box attack.



Patch Attack

T. Brown et al. (2018)

A physical and unbounded attack.



"HopSkipJump" Attack

J. Chen et al. (2024), UC Berkley

A black-box attack that can discover model thresholds.

Laser Beam

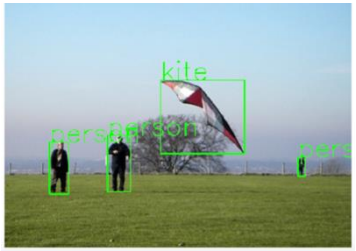
R. Duan et al. (2021)

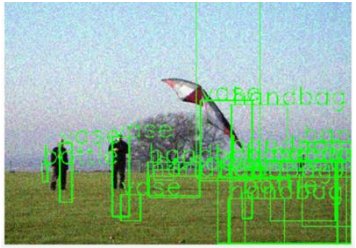
An easy-to-perform physical attack.

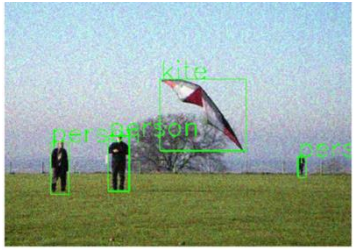
HEART Supported Defenses and Mitigations

Adversarial Robustness Toolbox

HEART







- **Preprocessor, Postprocessor Mitigations**
  JPEG compression, spatial smoothing, variance minimization, high confidence

- **Adversarial Training**
  Incorporation of adversarial examples into training data

- **Detector Defense**
  "Be Your Own Neighborhood" – BEYOND (ART only).
  A framework for detecting adversarial attacks through comparison of labels and image representations *He et al. (2024) in collaboration with IBM*

- **Transformer Defense**
  Defensive Distillation (work in progress).
  Entails training deep neural net (DNN) classifiers to smooth decision boundaries, improving generalization, robustness

Evaluation Metrics

Adversarial Robustness Toolbox

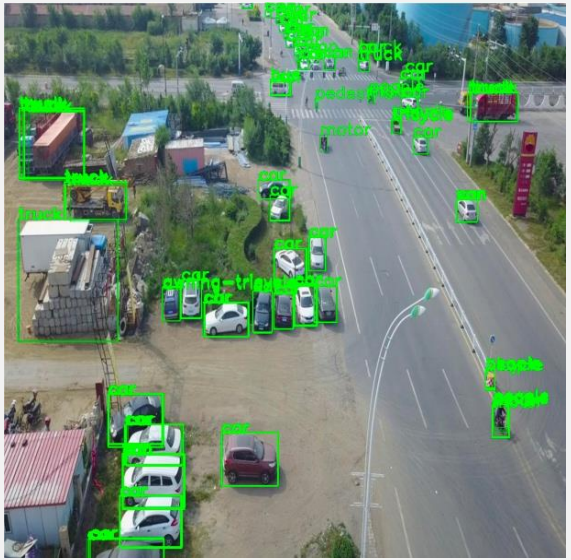
HEART

	Image classification	Object detection
	Accuracy: $\frac{\textit{\# correctly classified}}{\textit{total \# classified}}$	Mean Average Precision (mAP): *truth obj boxes – predicted obj boxes*
Before attack	Clean/benign accuracy	Clean/benign mAP
After attack	Robust accuracy Adversarial accuracy	Robust mAP Adversarial mAP

YOLOv5 object detector in HEART

1. Run clean model on original image

- Import tools from HEART
- Load images from Visdrone dataset
- Load YOLO object detector
- Apply Object Detection
- Assess performance

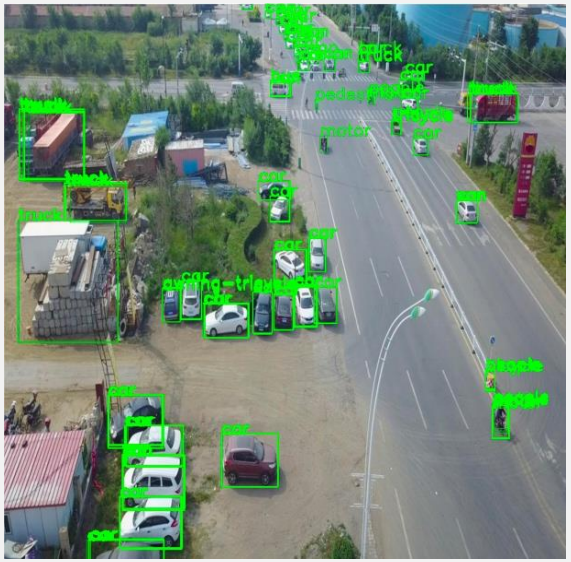


Benign Image:
- Classification accuracy: 88.9%
- mAP: 0.228

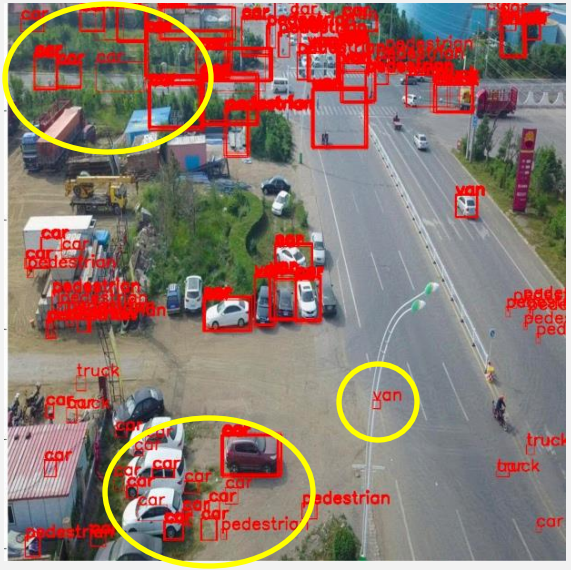
Untargeted PGD attack against the YoloV5 object detector

2. Run attack

- Define Projected Gradient Descent attack
- Run attack
- Assess performance
- Compare to clean performance



Benign Image:
- Classification accuracy: 88.9%
- mAP: 0.228



Adversarial Image:
- Classification accuracy: 58.2%
- mAP: 0.01

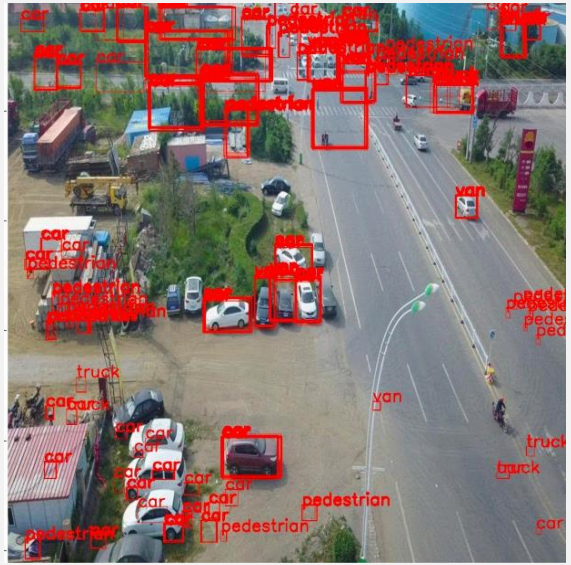
Applying a mitigating defense during preprocessing

Adversarial Robustness Toolbox

HEART

3. Apply mitigating defense

- Define Spatial Smoothing mitigating defense
- Apply defense during preprocessing
  - NOTE: mitigations do not provide 100% defense against adversarial attack
- Assess performance
- Compare to adversarial and clean performance



Adversarial Image:
- Classification accuracy: 58.2%
- mAP: 0.01

HEART Mitigation Image:
- Classification accuracy: 81.9%
- mAP: 0.19

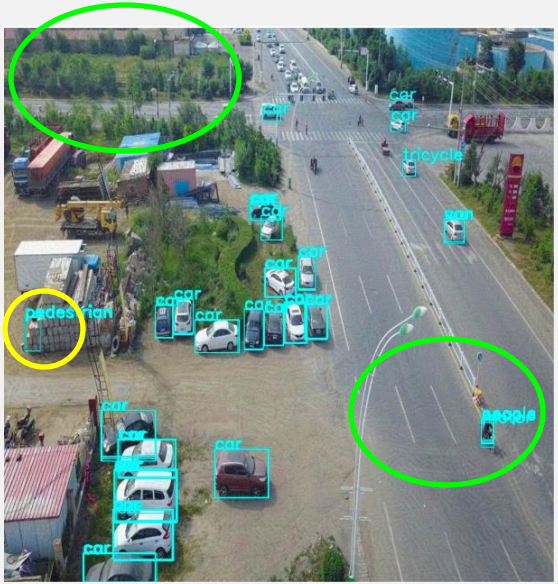
Adversarial samples that can bypass defenses

Adversarial Robustness Toolbox

HEART

4. Re-run attack with increasing strength

- Pull defended model
- Run attack
- Re-run attack with new hyperparameters
- Assess performance
- Compare to previous attack performance
- Etc…



HEART Mitigation Image:
- Classification accuracy: 81.9%
- mAP: 0.19

2nd attack on defended model:
- Classification accuracy: 76.2%
- mAP: 0.11

ATO and access

ART and HEART are both 100% open-source python libraries licensed under the MIT License.

- ART and HEART are available in **JWICS (Joint Worldwide Intelligence Communications System) PyPI repository**
- ADVANA platform availability – HEART and other JATIC packages have been promoted to the platform for rapid readiness in DoD applications
- We work with security stakeholders to provide source files and conduct necessary scans to add package to environments **without affecting existing ATO status**

How to Get Started

Adversarial Robustness Toolbox

HEART

GitHub Open-Source Tools

ART

HEART

CDAO JATIC program info and T&E tools

Contact our team with questions!

Jackson Lee
Project Manager
—
jackson.lee@ibm.com

Jordan Fischer
Solutions Architect
—
jordan.j.fischer@ibm.com

Quinn Stackpole
Deployment Coordinator
—
quinn.stackpole@ibm.com

Jordan Fischer
Solutions Architect
IBM Federal Consulting

Thank you

Appendix slides

White- vs Black-box attacks

Adversarial Robustness Toolbox

HEART

Black-box attacks

Attacker has little to no access to information on model architecture and parameters, and must rely on trial-and-error techniques, often guided by optimization algorithms, to craft deceptive inputs that exploit the decision process of the model

White-box attacks

Attacker has access to significant information on model architecture and parameters, allowing them to calibrate more precise attacks, optimally generating adversarial interference that is harder to detect and defend against