Adversarial Robustness 360 Toolbox (ART) v1.0

The First Major Release - A Milestone in AI Security

https://github.com/IBM/adversarial-robustness-toolbox

Beat Buesser
Dublin Research Laboratory
IBM Research

October 03, 2019

Cognitive Systems Institute Group talks

Introduction

- Biography

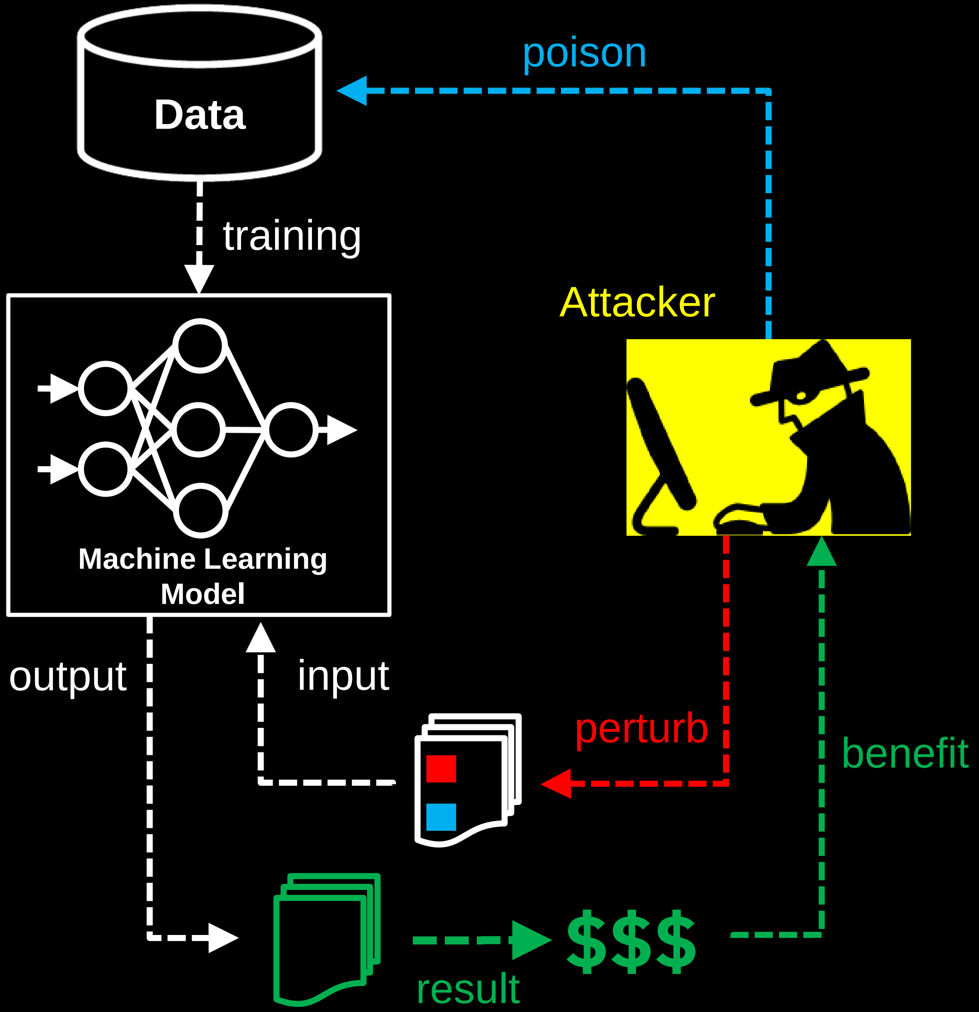
  I am a Research Staff Member at IBM Research in the AI & Machine Learning Group at the Dublin Research Laboratory. I am leading the development of the Adversarial Robustness 360 Toolbox (ART) and my current research focuses on the security of machine learning and artificial intelligence. Before joining IBM, I worked as postdoctoral associate at the Massachusetts Institute of Technology and obtained my doctorate degree from ETH Zurich.

- I will often use images for demonstrations of ART and adversarial machine learning, because they make nice visualizations, but it is important to mention that ART v1.0 can handle any type/shape of data including tabular data, text embeddings, etc. in addition to images.

- Adversarial Robustness Toolbox is mainly developed by a global team at IBM Research
  - Mathieu Sinn and team, Dublin Research Laboratory
  - Ian Molloy and team, Thomas J. Watson Research Center, Yorktown
  - Nathalie Baracaldo and team, Almaden Research Center

- Today’s Topics:
  - Adversarial Threats to Machine Learning and AI
  - Adversarial Robustness Toolbox v1.0
  - Demonstration of Adversarial Attacks and Defenses with ART

Adversarial Threats to Machine Learning/AI



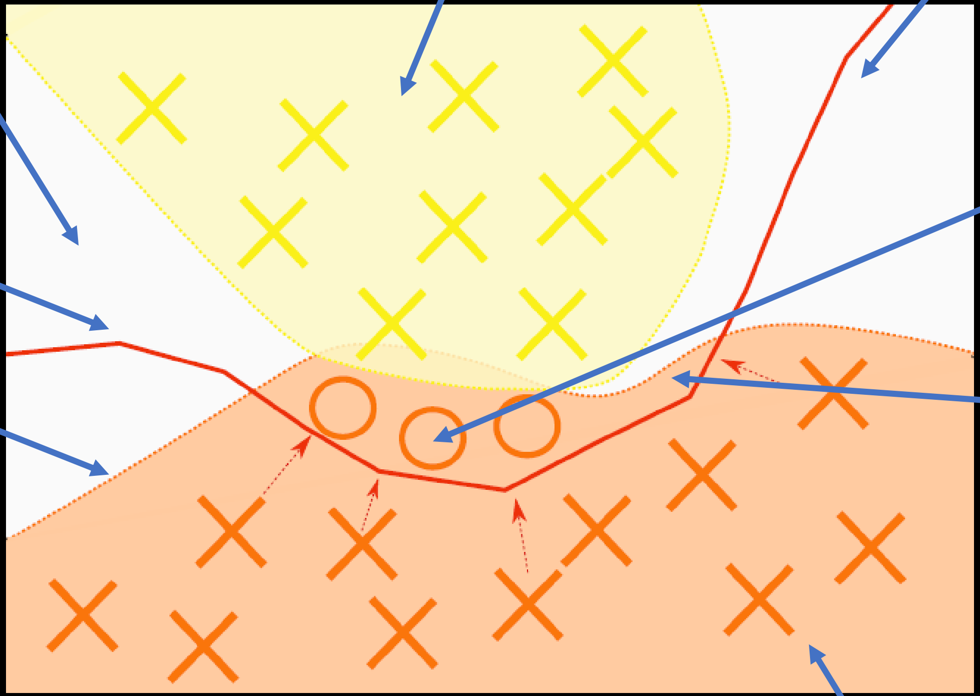
Evasion attacks

Performed at **test** time by perturbing input/features with crafted noise undetectable by humans to fool the model to make a wrong prediction.

Poisoning attacks

Performed at **training** time by inserting poisoned examples into the training data to create a backdoor later use.

Adversarial Examples (Simplified)



[Gilmer et al., 2018. https://arxiv.org/abs/1801.02774]

Examples of class 0

Limitation of classifier
to represent actual boundary

No data

Adversarial examples of class 1 classified as class 0 if specifically crafted to be in this space

Decision boundary of classifier

Actual decision boundary

Not enough data
for accurate representation

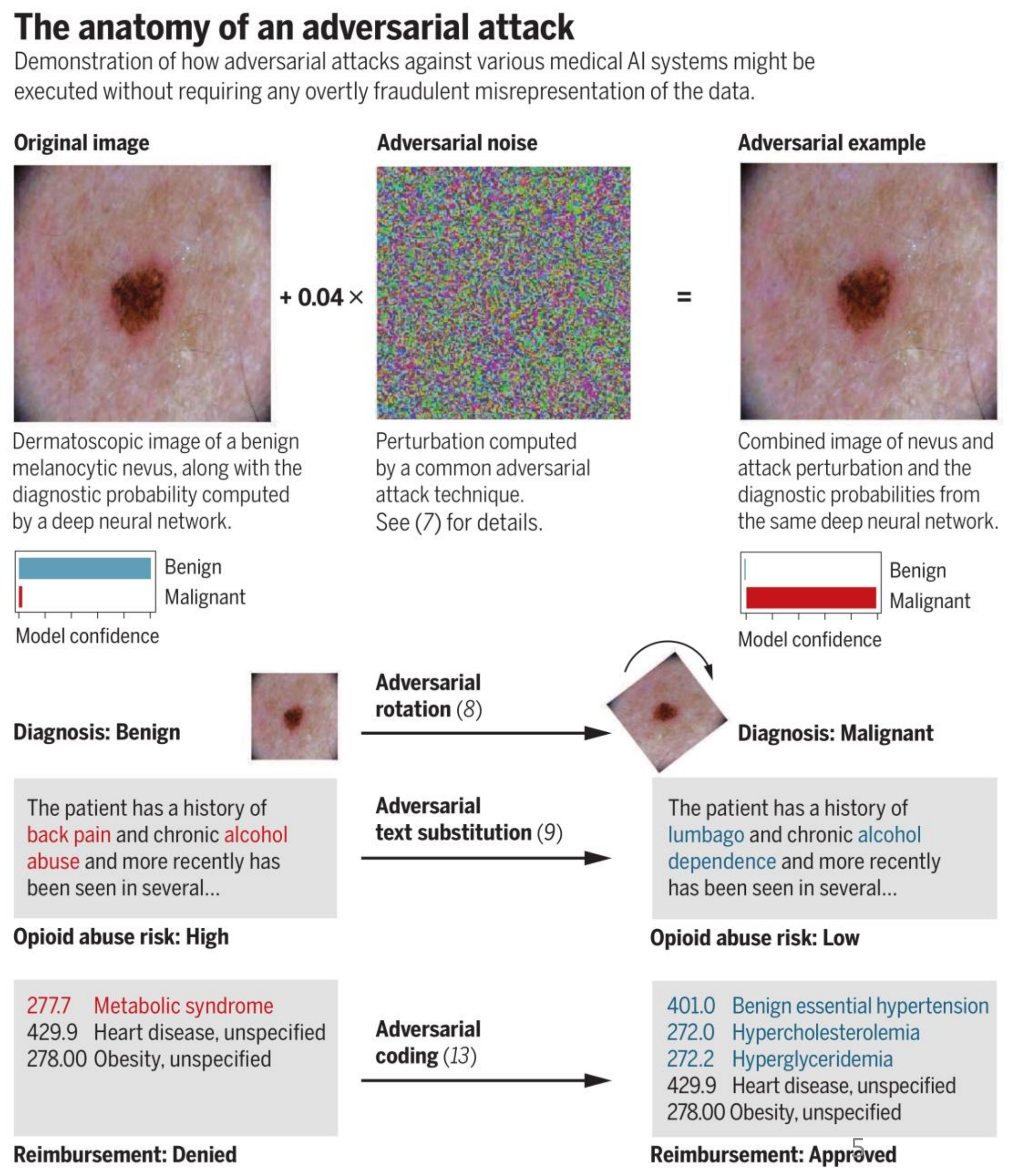
Examples of class 1

Real Adversarial Attacks

- **"Adversarial Attacks on Medical Machine Learning"**
  - Policy Forum article in Science Magazine, 22 March 2019, Vol 363 Issue 6433, Finlayson et al.
  - ART v1.0 can create every attack described in this article on any classifier/model type

- **"Cylance Antivirus Products Susceptible to Concatenation Bypass"**
  - Carnegie Mellon University CERT Coordination Center
  - Cylance antivirus product using a neural network to classify malware can be bypassed with adversarial attacks
  - Vulnerability note recommends ART as a tool to test machine learning models on adversarial robustness during development
  - http://kb.cert.org/vuls/id/489481

Image from Science Magazine, 22 March 2019, Vol 363, Issue 6433
Finlayson et al.



The anatomy of an adversarial attack

Demonstration of how adversarial attacks against various medical AI systems might be executed without requiring any overtly fraudulent misrepresentation of the data.

Dermatoscopic image of a benign melanocytic nevus, along with the diagnostic probability computed by a deep neural network.

Perturbation computed by a common adversarial attack technique. See (7) for details.

Combined image of nevus and attack perturbation and the diagnostic probabilities from the same deep neural network.

Diagnosis: Benign

Adversarial rotation (8)

Diagnosis: Malignant

The patient has a history of back pain and chronic alcohol abuse and more recently has been seen in several...

Adversarial text substitution (9)

The patient has a history of lumbago and chronic alcohol dependence and more recently has been seen in several...

Opioid abuse risk: High

Opioid abuse risk: Low

277.7 Metabolic syndrome
429.9 Heart disease, unspecified
278.00 Obesity, unspecified

Adversarial coding (13)

401.0 Benign essential hypertension
272.0 Hypercholesterolemia
272.2 Hyperglyceridemia
429.9 Heart disease, unspecified
278.00 Obesity, unspecified

Reimbursement: Denied

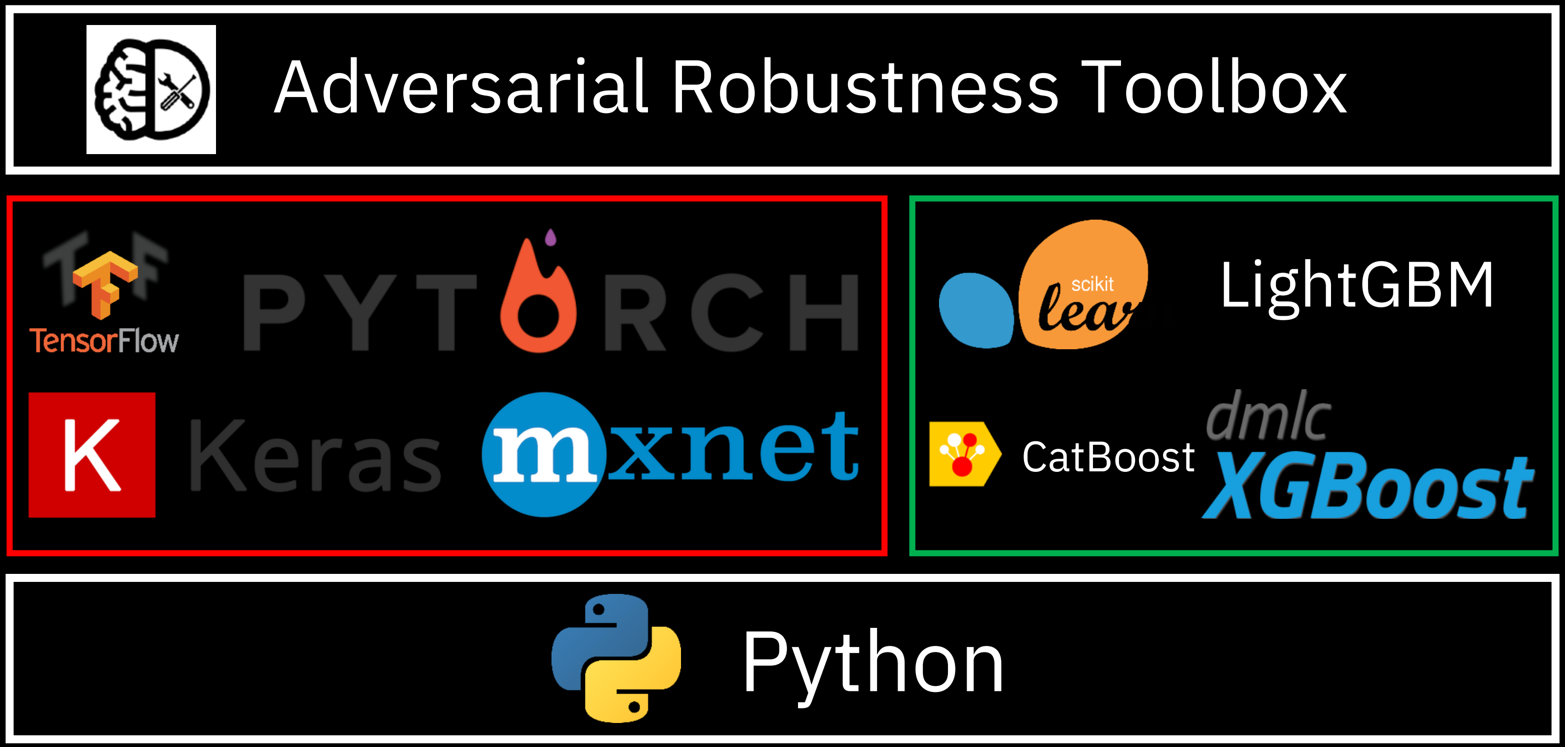
Reimbursement: Approved

Adversarial Robustness 360 Toolbox (ART v1.0)

ART is a framework-independent adversarial machine learning library in Python.

New in ART v1.0

Neural Networks
TensorFlow v2
(eager execution!)



Adversarial Robustness Toolbox

TensorFlow PYTORCH Keras mxnet

scikit learn LightGBM CatBoost dmlc XGBoost

Python

- Logistic Regression
- Support Vector Machine
- Decision Trees
- Random Forests
- Gradient Boosted Trees
- Gaussian Process
- Black-box classifiers

The Tools of ART v1.0

New in ART v1.0

- **Evasion Attacks**
  - HopSkipJump
  - High-Confidence-Low-Uncertainty (Gaussian Processes)
  - Projected Gradient Descent
  - NewtonFool
  - Elastic net
  - Spatial transformation
  - Query-efficient black-box
  - Zeroth-order optimization
  - Boundary attack
  - Adversarial patch
  - Decision Tree Attack (Decision Trees)
  - Carlini&Wagner attack
  - Basic iterative method
  - Jacobian saliency map
  - Universal perturbation
  - DeepFool
  - Virtual Adversarial method
  - Fast Gradient method

- **Poisoning Attacks**
  - Poisoning Attack on SVM
- **Defences**
  - Thermometer encoding
  - Total Variance minimization
  - PixelDefend
  - Gaussian Data augmentation
  - Feature squeezing
  - Spatial smoothing
  - JPEG compression
  - Label smoothing
  - Virtual adversarial training
  - Adversarial training

- **Robustness Metrics**
  - Clique Method (robustness verification for decision tree ensembles)
  - Randomised Smoothing (certified robustness)
  - CLEVER
  - Loss sensitivity
  - Empirical Robustness
- **Detection of Adversarial Examples**
  - Detector on inputs
  - Detector on activations
  - Fast Generalised Subset Scan
- **Detection of Poisoning Attacks**
  - Activation Analysis

ART v1.0 Minimal Code Example

Strong Evasion Attack – C&W-L2

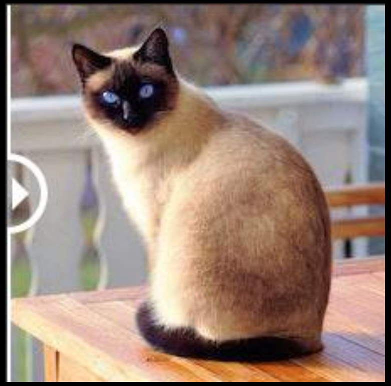
```
# Data and Model Preparation
image = load_image('Siamese_cat.jpeg')
model = load_model('my_favorite_keras_model.h5’)
y_target = [407] # target class: 407 (ambulance)

# Create ART classifier and attack
classifier = KerasClassifier(model=model, clip_values=(0, 255)
attacker = CarliniL2Method(classifier=classifier, targeted=True,
                           initial_const=10, binary_search_steps=25,
                           max_iter=50, confidence=5)

# Generate adversarial example with ART
adversarial_image = attacker.generate(image, y=y_target)

# Test prediction on adversarial example
prediction = classifier.predict(adversarial_image)
```

image

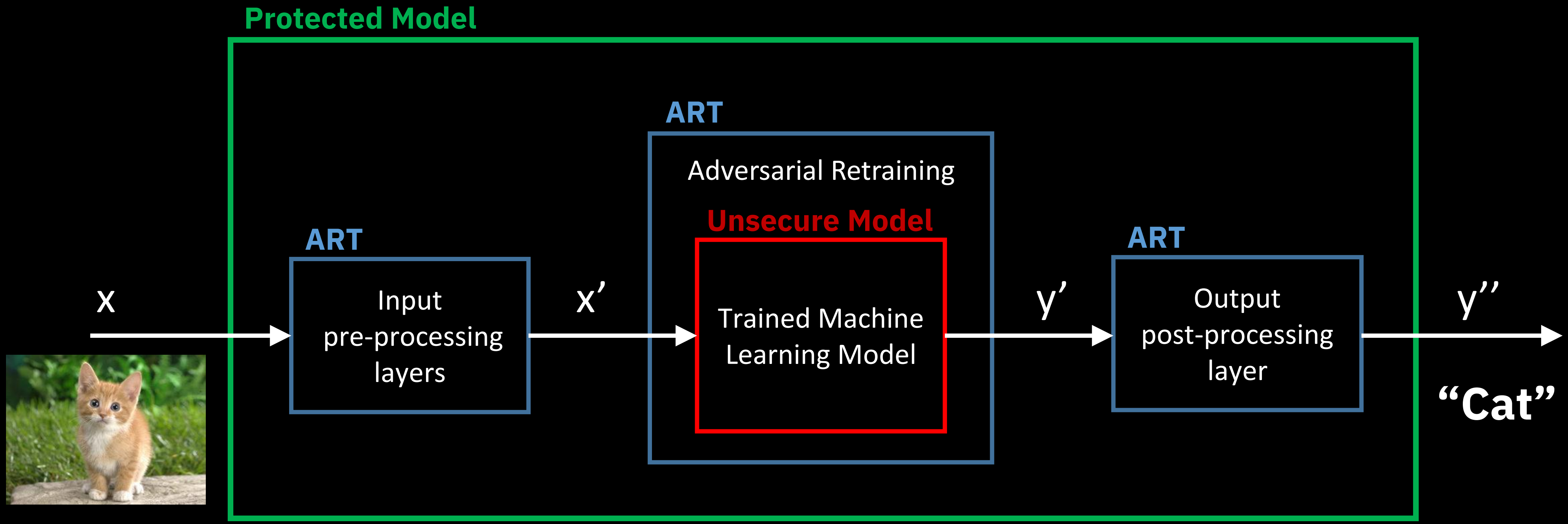
prediction = 92% Siamese Cat

adversarial_image

prediction = 82% Ambulance



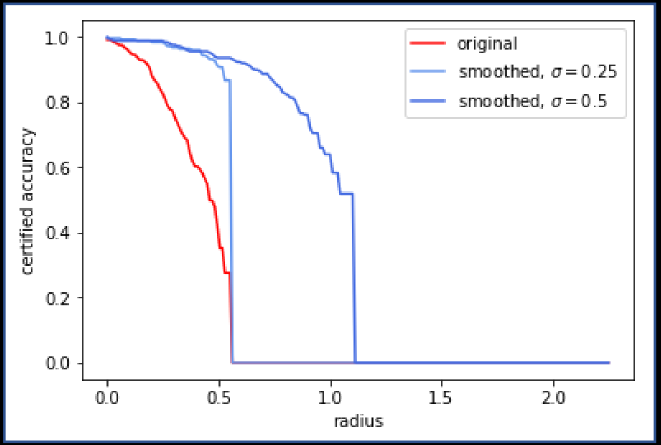
Protecting Machine Learning Models with ART


Protected Model
x
Input pre-processing layers
ART
x’
ART
Adversarial Retraining
Unsecure Model
Trained Machine Learning Model
y’
ART
Output post-processing layer
y’’
“Cat”

Input pre-processing: Modify input x by reducing adversarial features (e.g. spatial smoothing, random noise, etc.)
Adversarial retraining: Retrain the model on benign and adversarial examples
Output post-processing: Non-differential modification of predicted output to increase or certify robustness (e.g. Randomized Smoothing)

ART v1.0 – Certify and Verify Robustness

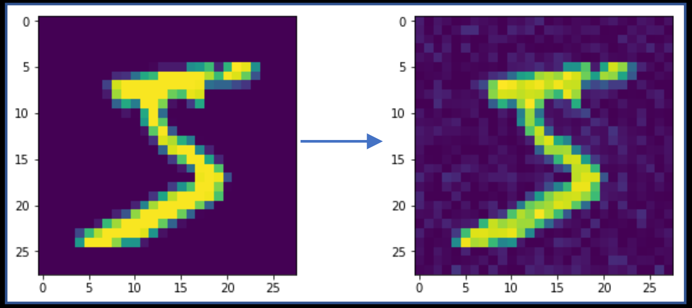
- Randomized Smoothing
  - https://arxiv.org/abs/1902.02918
  - Tight robustness guarantees in l2 norm for neural networks
  - Advantage: the adversarial example needs at least the guaranteed perturbation to fool the classifier



- Clique Method Robustness Verification
  - https://arxiv.org/abs/1906.03849
  - Verification of Robustness for decision tree based ensembles
  - It returns the minimal perturbation required to change the trained classifiers decision for a specific example
  - Verification is supported for XGBoost, LightGBM, CatBoost and scikit-learn’s GradientBoostingClassifier, RandomForestClassifier, ExtraTreesClassifier

ART v1.0 - BlackBoxClassifier

Create Adversarial Example for Deployed Remote Classifier



- BlackBoxClassifier is the most general and versatile classifier of ART v1.0. It does not make any assumption about the classifier. The predict function provided to the BlackBoxClassifier can contain any Python code returning a classification.

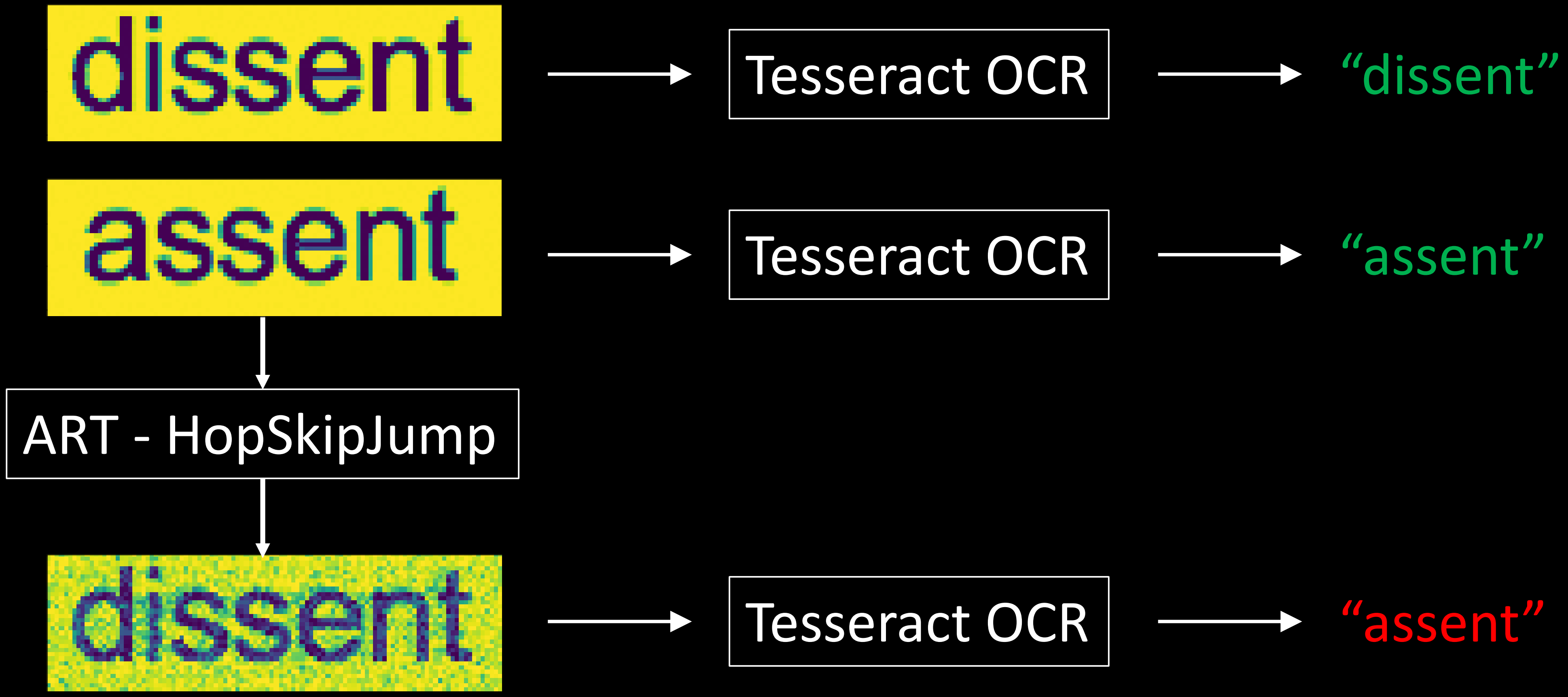
```
# Sample predict function that reformats inputs, connects to wml scoring endpoint and
# returns one-hot encoded predictions
def predict(x):
    scoring_data = {'values': (np.reshape(x, (x.shape[0],784))).tolist()}
    # Score with WML client
    predictions = client.deployments.score(scoring_url, scoring_data)
    return to_categorical(predictions['values'], nb_classes=10)

# Create blackbox classifier
classifier = BlackBoxClassifier(predict, input_shape=input_shape, nb_classes=10, clip_values=(0, 255))
```

- IBM Watson Studio tutorial on training classifier for handwritten digits and deploying it as a cloud service
    - https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/3bd3efb8-833d-460f-b07b-fee51dd0f1af/view?access_token=6bd0ff8d807861d09e0dab0cad28ce9685711078f612fcd92bb8cf8535d089c1
- Our tutorial uses ART v1.0 BlackBoxClassifier to demonstrate adversarial examples created for deployed, remote classifiers
    - https://github.com/IBM/adversarial-robustness-toolbox/blob/master/notebooks/classifier_blackbox.ipynb

ART v1.0 - BlackBoxClassifier

Create Adversarial Examples for complex Tesseract OCR pipeline

Tesseract OCR is a optical character recognition engine. This example uses the complete pipeline of Tesseract.
More details in our notebook on GitHub: classifier_blackbox_tesseract.ipynb



IBM Research - Trusting AI

- Trusting AI
  - IBM Research is building and enabling AI solutions people can trust.
  - https://www.research.ibm.com/artificial-intelligence/trusted-ai
- AI Fairness 360 Toolkit
  - Helps examine, report, and mitigate discrimination and bias in machine learning models
  - http://aif360.mybluemix.net
- AI Explainability 360 Toolkit
  - Helps to comprehend how machine learning make their predictions
  - http://aix360.mybluemix.net
- Adversarial Robustness 360 Toolbox
  - Helps to defend, certify and verify machine learning models against adversarial threats
  - http://art-demo.mybluemix.net

Conclusions

- Adversarial Attacks are a Real Possibility
  - Any machine learning model (not just neural networks)
  - Any machine learning framework/library (even black-box/no knowledge)
  - Any deployment mode (even remote in a cloud service or as part of a complex pipeline)

- ART v1.0 supports you with Defending, Certifying and Verifying your Machin Learning applications
  - ART v1.0 supports any type of machine learning model
    - Neural network, gradient boosted decision trees, support vector machines (SVM), random forest, logistic regression, Gaussian process, decision tree, scikit-learn pipeline, black-box, etc.
  - ART v1.0 supports any machine learning framework/library
    - scikit-learn, XGBoost, LightGBM, CatBoost, GPy, Tensorflow (v1 and v2), Keras, PyTorch, MXNet, black-box
  - ART v1.0 supports any type and shape of data
    - Tabular data, text embeddings, images, etc.

Resources of ART v1.0

- GitHub
  - https://github.com/IBM/adversarial-robustness-toolbox
- Get Started with ART – Examples
  - https://github.com/IBM/adversarial-robustness-toolbox/blob/master/examples/README.md
- Get Started with ART – Tutorials
  - https://github.com/IBM/adversarial-robustness-toolbox/blob/master/notebooks/README.md
- Documentation
  - https://adversarial-robustness-toolbox.readthedocs.io
- Slack
  - https://ibm-art.slack.com
- Demo
  - https://art-demo.mybluemix.net/
- Blog
  - https://www.ibm.com/blogs/research/2019/09/adversarial-robustness-360-toolbox-v1-0

Thank you for your Attention!

https://github.com/IBM/adversarial-robustness-toolbox