

Adversarial Robustness Toolbox Lab

Introduction

ART is a library developed by "IBM Research" which is dedicated to adversarial machine learning. Its purpose is to allow rapid crafting and analysis of attacks and defense methods for machine learning models. ART provides an implementation for many state-of-the-art methods for attacking and defending classifiers. See below links for more information on ART.

ART Demo: <https://art-demo.mybluemix.net>

ART Blog: <https://www.ibm.com/blogs/research/2018/04/ai-adversarial-robustness-toolbox/>

ART Github: <https://github.com/IBM/adversarial-robustness-toolbox>

In this lab, you will work with the Adversarial Robustness Toolbox (ART) and implement an adversarial attack and its defense on a trained model. You will work with a model trained on the German Traffic Signs dataset (see Citation below) and get the model to misclassify a stop sign.

Dataset Citation

J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In Proceedings of the IEEE International Joint Conference on Neural Networks, pages 1453–1460. 2011.

@inproceedings{Stallkamp-IJCNN-2011,

```
author = {Johannes Stallkamp and Marc Schlipsing and Jan Salmen and Christian Igel},
booktitle = {IEEE International Joint Conference on Neural Networks},
title = {The {G}erman {T}raffic {S}ign {R}ecognition {B}enchmark: A multi-class
classification competition},
year = {2011},
pages = {1453--1460}
```

```
}
```

Objectives

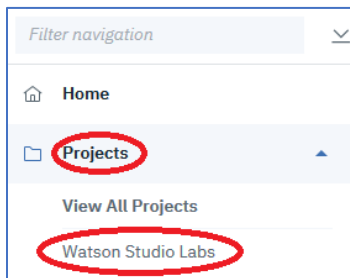
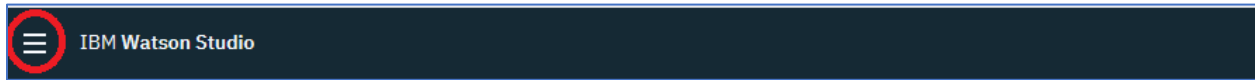
Upon completing the lab, you will learn how to:

1. load a Tensorflow trained model
2. create an ART classifier object using the loaded model
3. perform an adversarial attack
4. perform a defense to make sure manipulated images can still be classified correctly

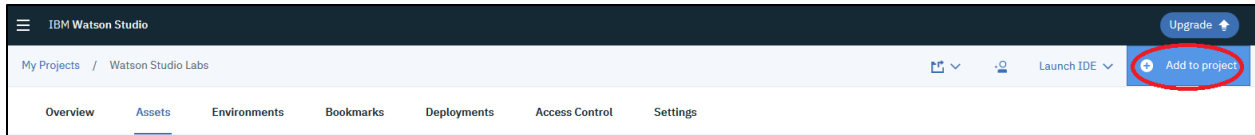
Lab Steps

Step 1 - Create a Jupyter Notebook

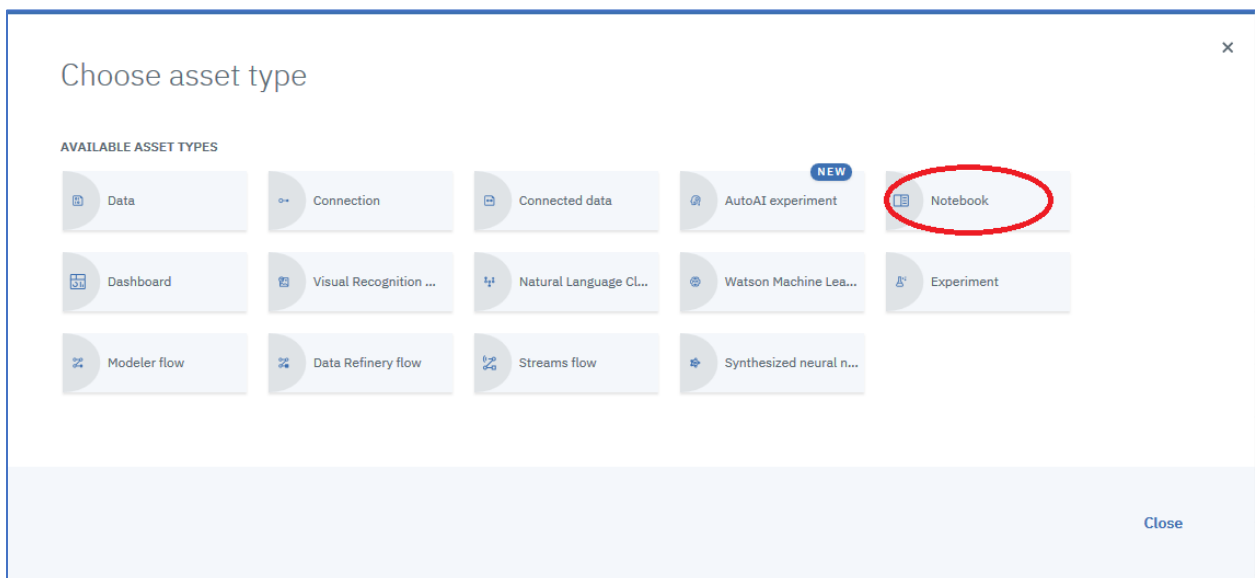
1. Click on the hamburger icon , then click on **Projects**, and then **Watson Studio Labs** (or whatever you named the project)



2. We are now going to create a notebook in our project. This notebook will be created from a url that points to the ART notebook in the github repository. Click the **Add to project** link.



3. Click on **Notebook**



- Click on **From URL** under **New Notebook**, enter **ART Demo** for the **Name**, optionally enter a **Description**, leave the default for the **runtime**, and cut and paste the following url into the **Notebook URL** field.

https://github.com/bleonardb3/ML_POT_07-23/blob/master/Lab-2/ART%20Demo.ipynb

Click **Create Notebook**.

The screenshot shows the 'New notebook' interface with the following fields and values:

- Blank From file From URL** (selected)
- Name:** ART_Demo (32 characters remaining)
- Description (optional):** (empty, 500 characters remaining)
- Select runtime:** Default Python 3.6 XS (2 vCPU and 8 GB RAM)
- Notebook URL:** https://github.com/bleonardb3/ML_POT_07-23/blob/master/Lab-4/Lab4%20-%20ART%20Demo.ipynb
- Buttons:** Cancel, Create Notebook

- Follow the instructions in the notebook.