European Championship in Trustworthy AI
Prof. Dr. Adam Dziedzic and Prof. Dr. Franziska Boenisch
TASK #10: Adversarial Attack

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

Task: Find Adversarial Examples, while applying as little change to the original images as possible.

The goal of Adversarial Example generation is to create altered images that look identical to the original image to the human eye, but are classified differently by an image classifier. In this task, your goal is to create these exact adversarial examples, based on a provided dataset and an image classifier with black-box access.

As an attacker, you get access to:

- Query access to a resnet18 Image Classifier that returns logits for an input image
- A dataset containing 1k images (available here)

What you don't have access to:

- The training dataset used to train Resnet-18 model
- The architecture and parameters of the Restnet-18 model

What's your task:

Your goal is to create one adversarial example per natural image from the provided dataset.

Evaluation Metric:

The evaluation is conducted taking into account the two essential factors for adversarial examples: 1) They yield a different result from the natural version on a given classifier 2) They are (visually) as close to the natural version as possible.

The evaluation metric is the average normalized L2-distance between the natural images and your adversarial examples. All submitted examples that yield the same result as the natural version on the classifier are assigned distance 1.

How to submit your results?

Submit a npz-file containing all perturbations (meaning the differences between the images and adversarial examples), in the same order as the natural images were provided in the dataset.

There is an example function on how to convert a pt-file into such an npz-file in main.py. As is done here, **ALWAYS** set allow_pickle to False when saving an npz-file. The same goes for loading such files.