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 an image classifier that returns logits for an input image (can be found here)
- A dataset containing 1k images (available here)

What you don't have access to:

- The training dataset used to train the classifier
- The architecture and parameters of the classifier

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 (incorrect label in this case) 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 (with the correct label) on the classifier are assigned distance 1. The distance is normalized in the range [0,1]. Your goal is to provide the modified 1000 images so that they are mislabeled and as close as possible (the smallest distance as close to 0) to the initial images.

Your Starting Point

You get access to a coding template (main.py) which demonstrates how to load the dataset, and how to query the image classifier with input images, as well as instructions explaining the result submission process.

Querying the Classifier

In order to merely give you black-box access to the image classifier, you do not have direct access to it. However, you can input images and get back the corresponding logits by interacting with the API. The code for this is given in the code template.

How to submit your results?

Submit a .pt file containing all of your adversarial examples, in the same order as the natural images were provided in the dataset.

Your team will be provided with a unique API-key, which can be used to submit your. The code for this is given in the code template. Submissions have a cooldown of ten seconds. Your result will then be evaluated. You can compare your best achieved result with the results of the other teams in the leaderboard.

Leaderboard

The leaderboard can be accessed by you to compare your evaluation results to the results of other teams. The leaderboard will only save your best result, and will therefore not be updated if your submission achieves a lower score. You can access the leaderboard at -PLACEHOLDER-.