European Championship in Trustworthy AI
Prof. Dr. Adam Dziedzic and Prof. Dr. Franziska Boenisch
TASK #6: Dataset Inference for Vision APIs

Task: Execute a Dataset Inference Attack and achieve the highest attack success (TPR@FPR=0.01) and area under curve (AUC).

The goal of a Dataset Inference is to determine whether a given dataset was part of the dataset the model in question was trained on. In this task, your goal is to launch a successful Dataset Inference Attack on an Image Classifier, only being given access to the final logits.

As an attacker, you get access to:

- Query access to an image classifier that returns logits for an input image (can be found here)
- Dataset containing 1000 subsets of images, with 100 images per subset (available here)

Additional information that you should know

The data subsets are the datasets on which you should conduct dataset inference. All images in a given subset are either all part of the training data, or none of them are. 50% of the subsets were used for training.

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 determine which subsets of data were part of the training data, which is indicated through a membership score. These scores should be continuous numbers between 0 and 1. They reflect how likely it is that the subset data was part of the training data.

Evaluation Metrics

- TPR@FPR=0.01
- AUC

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 CSV-File containing only two columns: One with the ordered IDs of the subsets (0-999), and one with the membership indicator.

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 three minutes, to avoid the possibility of brute forcing. This also happens if your submission fails due to misformatting of your inputs or similar errors. In these cases, you will get a comprehensive error message.

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-.