European Championship in Trustworthy AI
Prof. Dr. Adam Dziedzic and Prof. Dr. Franziska Boenisch
TASK #3: Dataset Inference for Vision APIs

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

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, you are in the *attacker's shoes*, and your goal is to launch a successful Dataset Inference Attack on an Image Classifier.

As an attacker, you get access to:

- A resnet18 image classifier trained on an *undisclosed* dataset (ADD LINK)
- Image dataset containing 1000 subsets of images, with 100 images per subset

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 underlying training dataset used to train 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 confidence score. You are *not* supposed to submit membership scores as 0 (non-member) and 1 (member). We expect you to submit ***continuous* membership confidence scores.**

> **Note -** The membership confidence score reflects how likely it is that a given subset was part of the training set used to train the model.

Your Starting Point

To start you off, we give you a coding template that shows you how to load the dataset, and how to query the image classifier with input images, as well as instructions explaining the result submission process.

How to submit your results?

- We will provide a unique API-key to every participating team. You are supposed to only use this token to submit your results to our server. If you did not receive this team-specific token, please contact us.

- Your submission should consist of a csv-file with two columns: One with the subset-ids, which range from 0 to 999, and one with the membership confidence scores. These columns should be named “subset_id” and “membership”, respectively.

- You are supposed to submit your scores using the code provided in the coding template. Remember to replace INSERT_YOUR_API_KEY_HERE with your actual API-key (keep the double quotes), and replace “example_submission.csv” with your real results.

- 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.

Evaluation

After submitting your results, they will be evaluated using two metrics:

- Area under the curve (AUC)
- True positive rate at 1% false positive rate (TPR@FPR=0.01)

Leaderboard

After evaluation, your results can be found in the leaderboard:

- You can access the leaderboard for this task at http://34.122.51.94:9000/leaderboard/06-dataset-inference-vision. This will help you to compare your solutions with other teams and see where you stand.

- The leaderboard shows the best result per team only. As output to your request, you will get back the TPR and AUC for your current submission. If they are lower than the scores saved in the leaderboard, these scores will not be updated.

References

- “Dataset Inference: Ownership Resolution in Machine Learning” Pratyush Maini, Mohammad Yaghini, Nicolas Papernot. ICLR 2021. https://arxiv.org/pdf/2104.10706 (Dataset Inference)

- "Towards Deep Learning Models Resistant to Adversarial Attacks" Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu. ICLR 2018. https://openreview.net/pdf?id=rJzIBfZAb (PGD attack)