Instructions for Paper Artifacts

1 OVERVIEW

We use jupyter notebook to demo the codes and show the attack results on our pretrained models.

The folder contains three separate jupyter files for different datasets.

- **NeuGuard_ALEXNET_cifar10.ipynb** for CIFAR10 dataset
- **NeuGuard_ALEXNET_cifar100.ipynb** for CIFAR100 dataset
- **NeuGuard_Texas100.ipynb** for Texas100 dataset.

2 DATASETS

To run the attack evaluation, you will need to download the three datasets first.

For Texas100 dataset, the default way may take some time to load the data every time, we saved a npz file for faster loading.

- Name: texas100_data.npz download here. (https://drive.google.com/file/d/1G9-oWyLqiSTDuB2ku6xYY7MVWOur6OOA/view?usp=sharing).
- We load Texas100 data with a randomize order following the file random_r_texas100_prune. Please download it before running the code.

For CIFAR10 and CIFAR100 datasets, they will download autometically if you don't have them.

3 EVALUATION STEPS

For each code in jupyter notebook, we have similar running steps. To check the model used in the paper, please:

(1) Run all the cells above the **# start train** cell for initialization.
(2) Load model in the **# load saved model** cell and the following two cells to check the model accuracy.
(3) Load unsort NSH attack model in the **# load membership inference attack** cell and the following five cells to check the model accuracy. This step run the unsorted neural network based membership inference attack correspond to Table 4 in Section 6.1.
(4) Load sort NN attack model in the **# load NN attack model** cell and the following two cells to check the model accuracy. This step run the sorted neural network based membership inference attack correspond to Table 4 in Section 6.1.
(5) Run cell in **# load for metric base attack** and following cells to perform the metric based attack. This step run four metric based membership inference attacks correspond to Table 6 in Section 7.
(6) Run cell in **# load for c&w label-only attack** and following cells to perform the c&w label-only attack. This attack correspond to Table 8 in Section 7 for CIFAR10 and CIFAR100.

In the code we predefined the load models with name correspond to the specific pretrained models provided for inference and perform attack. We also provide the training code to train our defensive model and perform attack. You can modify the parameters and settings to train your own model.

4 PRETRAINED MODELS

We provide pretrianed models used for the work. Specifically, we upload the models trained using the proposed NeuGuard method for all three datasets, and we include both sorted and unsorted NN based attack models, respectively.

The pretrained models can be download here. (https://drive.google.com/drive/folders/1qjPOpicHpCoKcdmL2Iko5f7P6ho5MrIq?usp=sharing)