UNIVERSITÉ CÔTE D'AZUR

MSc. Data Science & Artificial Intelligence

Inverse Problems in Image Processing

Chuan XU & Angelo RODIO

Lab 4

Author: Joris LIMONIER
joris.limonier@gmail.com
Date: March 21, 2023

Contents

1 Thread model: Deploy models **1**
1.1 Understanding the attack 1
1.1.1 Level of information of the attacker 1
1.1.2 What is a model inversion attack? 1
1.2 Getting familiar with the dataset 1
1.2.1 Plot one image 1
1.2.2 Name of the data heterogeneity 1
1.3 Evaluating the performance of the attack 1
1.3.1 Precise notation for the SSIM metric. 1
1.4 Exercise 1 2

2 Thread model: Honest-but-curious server **2**
2.1 Exercise 2 2

1 Thread model: Deploy models

1.1 Understanding the attack

1.1.1 Level of information of the attacker

The attacker has access to the global model only, as can be seen in the function:

```
attacker.py
→Attacker
→model_inversion_attack
```

which only takes as argument the global model (`model`) and the number of rounds (`nb_rounds`).

1.1.2 What is a model inversion attack?

A model inversion attack, also called gradient inversion attack, is an attack that aims at recovering the training dataset from the global model. This is the attack that the attacker performs in our case.

1.2 Getting familiar with the dataset

1.2.1 Plot one image

1.2.2 Name of the data heterogeneity

In this case, each of the $N = 10$ client holds one image of each of the 40 subjects. The data distribution for each client is **homogeneous**. This is to be opposed to the case where the data distribution would be splitted by labels, *e.g.* the case $N = 40$ where each client holds 10 image of only one subject.

1.3 Evaluating the performance of the attack

1.3.1 Precise notation for the SSIM metric.

The SSIM metric represents the similarity between two images. It uses the following elements:

1. μ_x is the mean of the image x.
2. μ_y is the mean of the image y.
3. σ_x is the standard deviation of the image x.
4. σ_y is the standard deviation of the image y.
5. σ_{xy} is the covariance between the images x and y.
6. L is the dynamic range of the images.
7. k_1 and k_2 are two constants that stabilize the division when the denominator is close to zero. They are set to $k_1 = 0.01$ and $k_2 = 0.03$ from the original paper

8. $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are two constants that depend on the dynamic range of the images. They stabilize the division when the denominator is close to zero.

The SSIM metric is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{1}$$

The SSIM ranges between -1 and 1. The closer to 1, the more similar the two images are. The closer to -1, the more different the two images are. It can be interpreted as a correlation coefficient.

1.4 Exercise 1

Table 1 shows the SSIM for various values of α and local steps E. α represent how much the local model is IID. Larger values of α means that the local model is more IID. We see

	$\alpha = 0.1$	0.5	0.7
$E = 1$	0.018	0.017	0.017
10	0.025	0.022	0.023
20	0.025	0.023	0.024
50	0.027	0.024	0.026

Table 1: SSIM for various values of the non-IID coefficient α and the local steps E.

that for a given number of local steps E, the SSIM increases as α increases. This means that the more the local model is IID, the better the global model is.

2 Thread model: Honest-but-curious server

2.1 Exercise 2

We run the experiment with the attack occurring at various number of rounds. The results are shown in Figure 1. We see that the SSIM increases as time passes and the attack loss

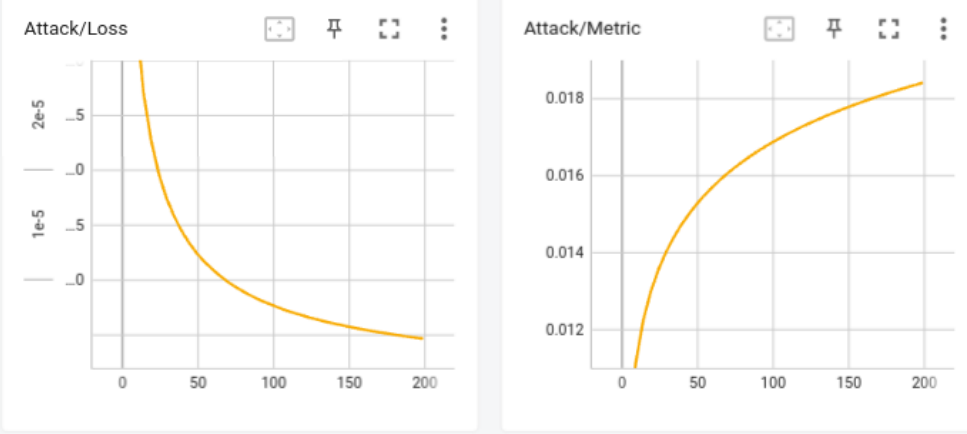

Figure 1: SSIM and attack loss for various values number of rounds

decreases. This means that it is better for the attacker to perform the attack at the end

of the training. This is because the global model is better at the end of the training. It has already learnt a well-performing model so it is easier for the attacker to recover the training dataset.
This conclusion actually makes sense as the initial model is not trained at all (pretty much random) so it is harder for the attacker to recover the training dataset.