Federated Learning & Data Privacy, 2022-2023

Submitted by: **Marina Diaz Uzquiano and Bhargav Ramudu Manam**

21 March 2023

Privacy in FL

1 Thread model: Deployed models

1.1 Understanding the attack

1. What level of information does the attacker have?

   Solution: In deployed models, the attacker has access to the final global model i.e., all the model parameters (weights) of the predictor function.

2. What is a model inversion attack?

   Solution: Reconstructing the training data of the clients using the deployed model (parameters) by exploiting information leakage from the model's output.

1.2 Getting familiar with the dataset

The Database of Faces, (formerly "The ORL Database of Faces"), contains a set of face images taken between April 1992 and April 1994 at the AT&T Laboratories Cambridge. The database was used in the context of a face recognition project carried out in collaboration with the Speech, Vision and Robotics Group of the Cambridge University Engineering Department.

There are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). A preview image of the Database of Faces is available.

The files are in PGM format, and can conveniently be viewed on UNIX (TM) systems using the 'xv' program. The size of each image is 92x112 pixels, with 256 grey levels per pixel. The images are organised in 40 directories (one for each subject), which have names of the form sX, where X indicates the subject number (between 1 and 40). In each of these directories, there are ten different images of that subject, which have names of the form Y.pgm, where Y is the image number for that subject (between 1 and 10).

1. Plot one image from the dataset.

   Solution:

```
plt.imshow(self.data.squeeze().numpy(), cmap='gray')
plt.show()
```

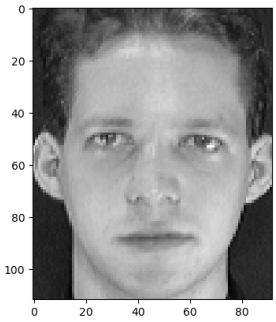


Figure 1: An image from Faces dataset

2. Consider a FL system with $N = 10$ clients. We wrote a code snippet that splits the dataset among the clients, such that each client holds a photo of all the subjects. Using the FL terminology, how would you refer to such data heterogeneity?

   Solution:

   Such data heterogeneity in FL can be termed as homogeneous or i.i.d data distribution. Since, all the clients hold exactly one photo of all subjects (uniform distribution), hence homogeneous. All the photos across the clients are made up of same subjects (identical) but distinct photos (independent), hence i.i.d distribution.

1.3 Evaluating the performance of the attack

The attack performance is evaluated by the average SSIM over 40 people. The structural similarity index measure (SSIM) is a metric used for measuring the similarity between two images.

1. Give the precise notation for the SSIM metric.

   Solution:

   $SSIM(x, y) = (2\mu_x * \mu_y + c_1)(2\sigma_{xy} + c_2)/({\mu_x}^2 + {\mu_y}^2 + c_1)({\sigma_x}^2 + {\sigma_y}^2 + c_2)$

   where:

   (a) x and y are two image patches to be compared,

(b) μ_x and μ_y are the mean values of x and y,

(c) σ_x and σ_y are the standard deviations of x and y,

(d) σ_{xy} is the covariance between x and y,

(e) c_1 and c_2 are constants to avoid instability when the denominator is close to zero.

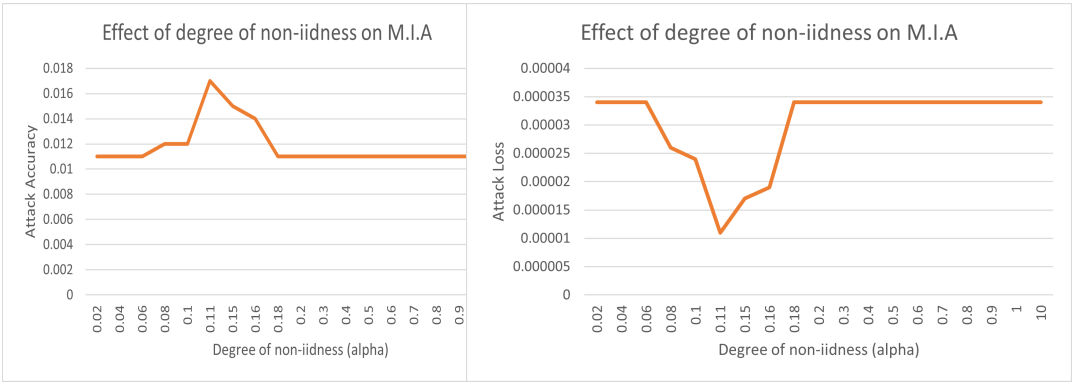
1.4 Exercise 1

We ran the FL training process (weighted by the local dataset size) for facial recognition, assuming an IID data distribution and local epoch $E = 1$. The adversary attacks on the final deployed model by doing model inversion attack.

- Fixed the number of local epoch $E = 1$, run the code for different degree of non-iidness. Analyze the effect of the degree of non-iidness (e.g., $\alpha = \{0.1, 0.5, 0.7\}$ on the model inversion attack performance in one figure. Give an explanation to your observation and conclusion.

  Solution:

  As can be seen from the figure (2), the attack metric stays invariant when α changes except for a small range of α values with maximum accuracy for $\alpha = 0.11$ (Here, I have chosen 10 server communication rounds while training the FL model). This can be due to the fact that the final global model parameters (weights) are good enough to reconstruct the training data irrespective of the individual client data heterogeneity.



(a) Accuracy (b) Loss

Figure 2: Effect of the degree of non-iidness (α) on the model inversion attack performance

- Fixed the degree of non-iidness $\alpha = 0.5$, run the code for different number of local epoches. Analyze the effect of the number of local epoches (e.g., E

$= \{1, 2, 3, 4, 5\}$) on the model inversion attack performance in one figure. Give an explanation to your observation and conclusion.

Solution:

As can be seen from the figure (3), the attack accuracy increases with increase in local steps at each client as this results in a better global model. However, after a certain local steps (80 in our case) there is no improvement in the attack accuracy and in fact is observed to be decreasing (this is not shown in the figure below but was observed in my experiments for a higher local steps). This kind of phenomenon is common in a FL setting as discussed during the course.

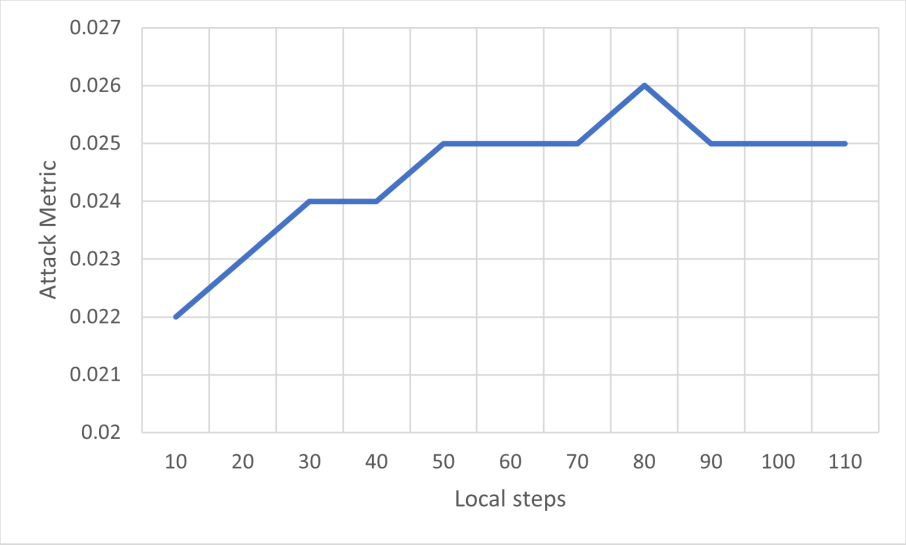


Figure 3: Evolution of the attack metric (accuracy) with respect to the number of local steps

2 Thread model: Honest-but-curious server

2.1 Exercise 2

We enhance the capacity of the attacker to an honest-but-curious server administrator. He can has raw access to the message exchanged in FedAvg.

Assuming that the attacker executes the model inversion attack on the local returned model by a client c at time t, denoted by $\theta^c(t)$.

- Fixed the number of local epoch E = 1 and the degree of non-iidness $\alpha = 0.5$, run the attack on the returned model $\theta^c(t)$ on different communication round t. Analyze the effect of eavesdropped communication rounds (e.g.,

t = $\{1, T/4, T/3, T/2, T/1.5, T/1.25, T\}$ where T is the total number of communication rounds) on the model inversion attack performance in one figure. Give an explanation to your observation and conclusion.

- Compare the performances with the one obtained in the previous thread model. What is your observation? Try to give your explanation.

  Solution:

  As you can see from the figure (4), the Model Inversion Attack performs the best when we attack at the last possible communication round. This is due to the fact that the global model at the last communication round has the best parameters representing the predictor function and hence gives us the best chance to infer the training data.

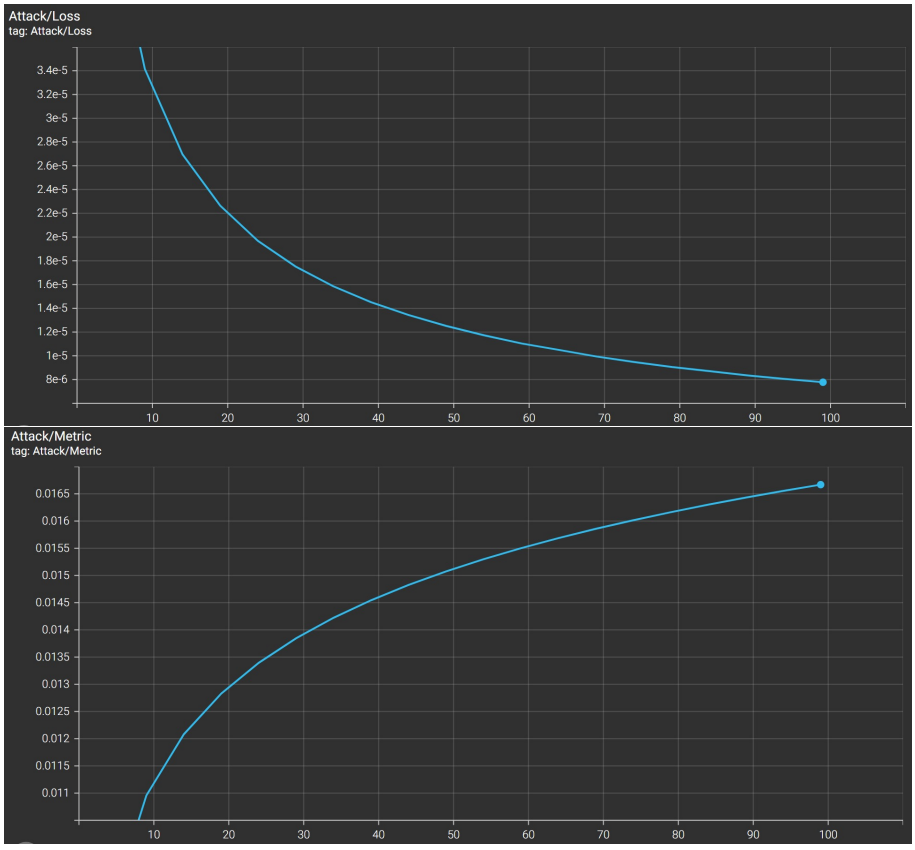


Figure 4: Performance of a Model Inversion Attack after different communication rounds