The Resource Problem of Using Linear Layer Leakage Attack in Federated Learning: *Supplementary Material*

	Sparse MANDRAKE	Robbing the Fed
CIFAR-100	77.5% (4957)	77.1% (4931)
MNIST	71.0% (4546)	75.1% (4803)
Tiny ImageNet	77.8% (4978)	77.7% (4970)

Table 1. Total leakage rate of sparse MANDRAKE and Robbing the Fed on various datasets. For all three datasets, 100 aggregated clients and batch size of 64 were used (6400 total images).

A. Leakage rate and reconstructions

Table 1 gives the leakage rate on the MNIST [6], CIFAR-100 [4], and Tiny ImageNet [5] datasets using sparse MANDRAKE and Robbing the Fed [3] as discussed in the main paper. We use a batch size of 64 with 100 clients, and the ratio of FC size to batch size is 4:1 (256 unit FC layer). The leakage rate on CIFAR-100 and Tiny ImageNet are roughly the same for both methods. Sparse MANDRAKE [7] has a slightly lower leakage rate than Robbing the Fed on MNIST.

Figure 1 shows the ground truth and reconstructions for a single, random client with a batch of 64 on Tiny ImageNet using the sparse MANDRAKE attack. 50 images were leaked from the client.

B. Trap weights under FL

We show the leakage rate using the trap weights attack [2] for different FC layer sizes on the downsampled Tiny ImageNet (32x32x3) dataset. We tune the scaling factor between 0.90 and 0.99 (step size of 0.01) to find the highest leakage rate, vary the FC layer ratio (FC layer size = batch size × num. clients × FC size ratio), and report the average over 10 runs. We apply this on several numbers of clients and Figure 2 shows the results compared to binning [3], which has roughly the same leakage rate regardless of the number of clients. Even while maintaining the same ratio between the FC layer size and total number of images, the leakage rate when using trap weights decreases as the number of clients increases.

We note that by using sparsity, trap weights are able to overcome this scalability problem. However, since the binning method of Robbing the Fed achieves a higher leakage rate for all FC layer size ratios, it is still a better choice.

C. Sparse variant of Robbing the Fed

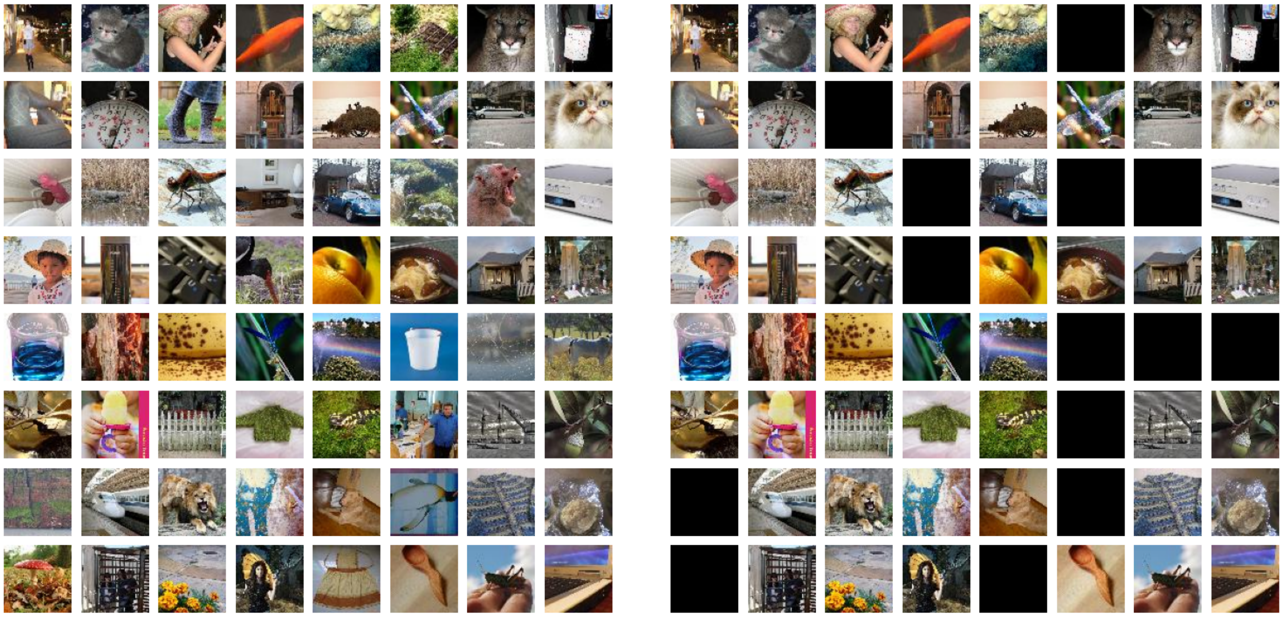
The sparse variant of Robbing the Fed (RtF) [3] is a method introduced in addition to their baseline in order to apply the attack in the FedAVG setting. The ”sparsity” mentioned in Section 4.3 of the RtF paper is discussing how to create activations in the fully-connected (FC) layer such that images should only activate a single neuron instead of a set of neurons. However, this does *not* reduce the resource usage added from the attack, which is what we address. With the main change being in the activation function, the same fundamental method as the baseline is used with aggregated updates and the FC layer size still needs to scale to compensate for the total number of images. These layers added to the model will still be fully dense with non-zero parameters.

D. Evaluating information leakage using mutual information

In practice, the amount of leaked information is typically quantified as the number of images a malicious server reconstructs (leaked). However, the reconstructions from the attack module can also leak some additional information that is not counted in the leakage rate. For example, while reconstructions of images can overlap, an observer can still obtain information about the training data (e.g., a malicious server who sees an overlap of digits 2, 3, and 8 might be able to identify that an 8 is in the reconstruction). In Section 4, we compared how much information was leaked to the server under a varying FC layer size using either the binning and trap weights method of linear layer leakage attacks, since MANDRAKE is able to use both.

We used the MNIST dataset for these experiments, and in order to measure the amount of information leaked into the gradient and the amount of information the server was able to reconstruct out of it, we compare the mutual information between: (1) the data batch x_k^{input} at user i and the aggregate gradient g of the attack at the server; (2) the data batch x_k^{input} and the reconstructions x_k at the server for user k. Note that by the data processing inequality, we have that:

$$\frac{I(x_k^{input}; x_k)}{I(x_k^{input}; g)} \leq 1. \tag{1}$$



(a) Ground truth

(b) Reconstruction

Figure 1. Reconstructed images from Tiny ImageNet from a random client with a batch size of 64 using MANDRAKE. The ground truth images (a) are shown on the left and the reconstructed images (b) are shown on the right. Any empty boxes within the reconstructed images indicate that reconstruction failed due to an overlap of image activations.

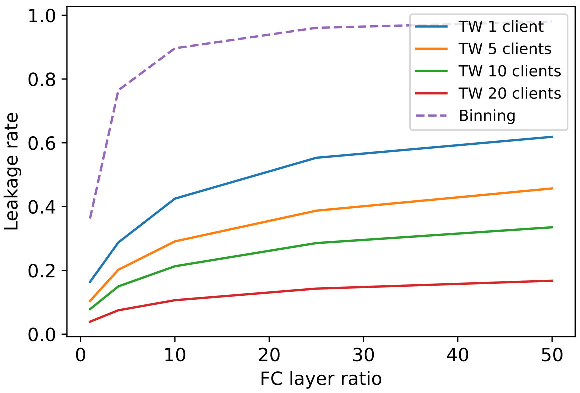


Figure 2. Leakage rate using trap weights (TW) for a batch size of 64 on Tiny ImageNet and varying the FC layer size ratio and number of clients. The leakage rate decreases with an increasing number of clients even if the ratio of FC size to total number of images remains the same. Binning (Robbing the Fed) has a higher leakage rate at all scales of FC size.

since the leaked images were reconstructed only using the gradient. In order to compute the mutual information terms in (1), we use the Mutual Information Neural Estimator (MINE) which is the SOTA method [1] to estimate the mutual information between two random vectors. For each FC layer size, we sampled 20,000 random batches of the users' data and used each to compute the aggregate gradient g and reconstructed images for a single user i. These 20,000 samples were used by MINE to estimate mutual information.

This same procedure was repeated multiple times in order to get multiple mutual information estimates and the average ratio was reported.

E. FedAVG

Unlike the gradients of the FC layers, the gradients of the convolutional layer are not necessary for the data reconstruction attack. A malicious server can then send a maliciously crafted model which would freeze the parameters of the convolutional layer to prevent changes from occurring over the local iterations of FedAVG.

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR, 10–15 Jul 2018. 2

[2] Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. When the curious abandon honesty: Federated learning is not private. *arXiv preprint arXiv:2112.02918*, 2021. 1

[3] Liam H Fowl, Jonas Geiping, Wojciech Czaja, Micah Goldblum, and Tom Goldstein. Robbing the fed: Directly obtaining private data in federated learning with modified models. In

International Conference on Learning Representations, 2022. 1

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009. 1

[5] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 1

[6] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 1

[7] Joshua C. Zhao, Atul Sharma, Ahmed Roushdy Elkordy, Yahya H. Ezzeldin, Salman Avestimehr, and Saurabh Bagchi. Secure aggregation in federated learning is not private: Leaking user data at large scale through model modification. *arXiv preprint arXiv:2303.12233*, 2023. 1