Leak and Learn: An Attacker's Cookbook to Train Using Leaked Data from Federated Learning

Supplementary Material

CIFAR-10			
	Number of clients	IID (I) or Non-IID (N)	Test Acc.
FedAvg	10	I	**75.13**
	10	N	72.76
	50	I	71.45
	50	N	68.71
FedSGD	10	I	71.24
	10	N	68.78
	50	I	65.95
	50	N	60.88
MNIST			
	Number of clients	IID (I) or Non-IID (N)	Test Acc.
FedAvg	10	I	96.62
	10	N	96.17
	50	I	96.68
	50	N	96.18
FedSGD	10	I	96.68
	10	N	96.76
	50	I	**96.84**
	50	N	96.83
Tiny ImageNet			
	Number of clients	IID (I) or Non-IID (N)	Test Acc.
FedAvg	10	I	37.18
	10	N	37.00
	50	I	**38.84**
	50	N	35.06
FedSGD	10	I	35.56
	10	N	34.27
	50	I	32.77
	50	N	26.56

Table 7. Federated learning test accuracy on CIFAR-10, MNIST, and Tiny ImageNet. A bias of 0.5 is used for the non-IID training. The same settings are used between FedSGD and FedAvg outside of the number of rounds. The number of rounds in FedSGD is 3× the number of rounds in FedAvg (3 local iterations in FedAvg).

7. Additional federated learning results

Table 7 shows additional test accuracy in federated learning on CIFAR-10, MNIST, and Tiny ImageNet. We include results for both IID and non-IID (with bias$= 0.5$). For FedSGD training, we use 3× the number of rounds compared to FedAvg (so the models have seen the same amount data in both cases, as we have 3 local iterations in FedAvg). All other settings are the same. An (expected) observed trend is that IID training outperforms non-IID. Both CIFAR-10 and Tiny ImageNet in FedAvg perform better than FedSGD in all settings. For MNIST, the performance is similar regardless of FedAvg or FedSGD, IID or non-IID, achieving around 96% accuracy across the board.

8. Sample reconstructions

For Inverting Gradients, we use a learning rate of $= 0.01$ and total variation of 1e-6 on CIFAR-10. For MNIST, we use a learning rate of $= 0.01$ and a total variation of 0. These parameters achieved the best reconstruction quality for us.

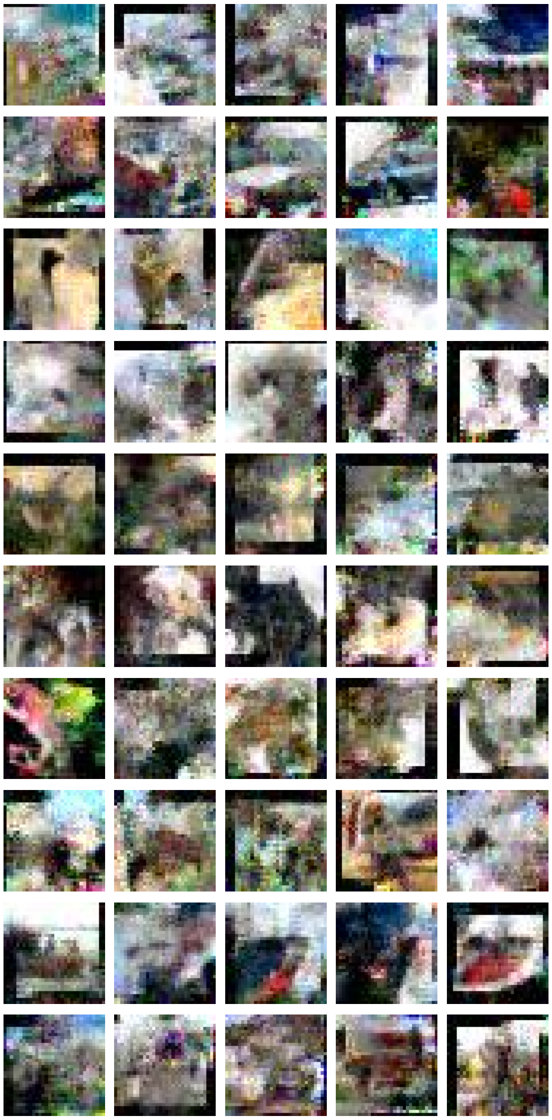

Figure 7. CIFAR-10 sample reconstructions from Inverting Gradients batch size 16 with a PSNR < 12. Each row is a different class. While the images are very noisy, using a set of them for training achieves a model with 45.05% accuracy.

Figure 7 shows sample reconstructions from Inverting Gradients [6] batch size 16 with PSNR < 12. Each row shows 5 images from each of the classes in CIFAR-10. The rows correspond to airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks respectively. While the im-

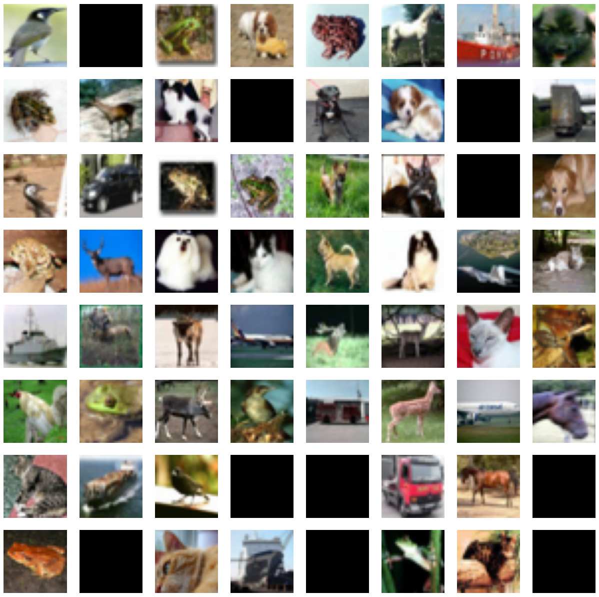

Figure 8. LOKI CIFAR-10 reconstructions using CSF= 500 in FedAvg. 54 images out of 64 images are leaked.

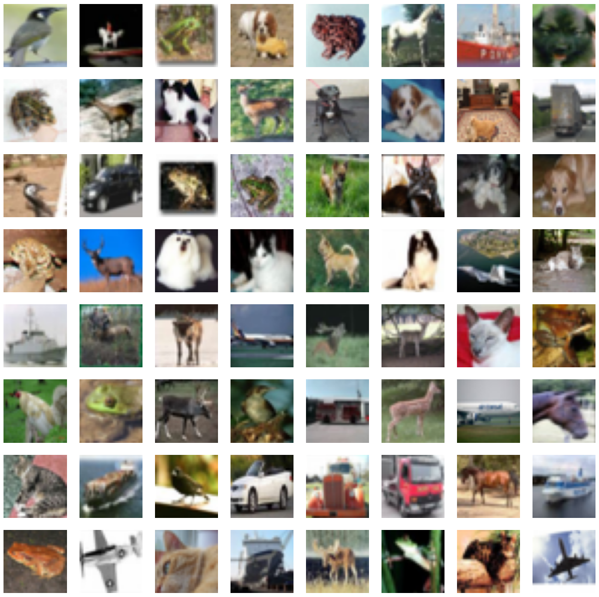

Figure 9. Ground truth CIFAR-10 images.

ages are noisy, there is some contextual information that can still be observed in the reconstructions. As discussed in Section 4.7, removing these images from the training set results in a small decrease in model performance from 76.83% to 76.16%. Training on only the set of images with PSNR < 12 results in a 45.05% test accuracy.

Figure 8 shows CIFAR-10 reconstructions from LOKI in FedAvg using CSF= 500. Figure 9 shows the corresponding ground truth images. For this particular set of images, 54 images out of 64 are leaked (84.38% leakage rate).

	FC size factor	Leakage rate	Test accuracy
LOKI	8	87.58	**93.16**
	4	78.93	92.94
	2	59.76	91.90
Robbing the Fed	8	87.50	93.10
	4	78.97	**93.02**
	2	59.72	**92.12**
Trap Weights	8	58.11	91.84
	4	45.92	90.09
	2	30.46	86.38

Table 8. Leakage rate and test accuracy on CIFAR-10 for LOKI, Robbing the Fed, and Trap Weights in FedSGD. FC layer size factors of 8, 4, and 2 used with a batch size of 64. Models trained from scratch on leaked data.

SSIM	% imgs kept	Test accuracy
> 0.7	17.25	72.32
> 0.6	30.26	75.68
> 0.5	44.76	75.94
> 0.4	61.71	76.61
> 0.3	80.56	77.02
> 0.2	95.55	77.31

(a) PSNR above threshold

SSIM	% imgs kept	Test accuracy
< 0.7	82.75	70.94
< 0.6	69.74	61.66
< 0.5	55.24	57.34
< 0.4	38.29	51.56
< 0.3	19.44	47.00
< 0.2	4.45	32.36

(b) PSNR below threshold

Table 9. Training models using CIFAR-10 leaked data from inverting gradients batch size 16. Only the reconstructions with an SSIM (a) above and (b) below the threshold are used in training.

9. Linear layer leakage method comparison

Table 8 shows the leakage rate and test accuracy on CIFAR-10 for LOKI [30], Robbing the Fed [5], and trap weights [1]. Attacks are done in FedSGD with a batch size of 64. LOKI and Robbing the Fed have no additional parameters besides the FC size factor (FC layer size = FC size factor × batch size). For trap weights, in addition to the FC size factor, a scaling factor of 0.96 achieves the highest leakage rate for each FC size factor (checked by 0.1 increments). LOKI and Robbing the Fed achieve very similar leakage rates and model performances. Trap weights has lower leakage rate than both other methods and, as a result, lower model performance for the same FC size factors.

10. SSIM threshold

Table 9 shows the test accuracy of models trained while removing images based on the SSIM. Table 9a shows accuracy when only images *above* an SSIM threshold are used. Table 9b shows accuracy when images *below* an SSIM threshold are used. For SSIM, removing a set of the worst images with SSIM < 0.2 or < 0.3 results in a small model performance increase compared to when all images are included (which achieves 76.83%). Similar to PSNR, training on a set of the worst quality reconstructions (SSIM < 0.2) achieves 32.36% accuracy, a higher accuracy than random guessing, but much lower performance compared to the baseline.