CERIAS

The Center for Education and Research in Information Assurance and Security

LOKI: Large-scale Data Reconstruction Attack against Federated Learning (S&P 24, CVPR 24)

Joshua C. Zhao, Ahaan Dabholkar, Atul Sharma, Saurabh Bagchi

Motivation

Federated learning (FL) is a machine learning paradigm originally introduced to allow for collaborative model training while simultaneously allowing user client data to stay private.

An example of industry production FL is G-Board where the goal of the model is to improve next word prediction. Here, Google is the FL server, and the clients consist of cellphones.

Instead of sending client data to the server, the clients receive a model and train on their local data before sending their updates to the server (gradients or model parameters). Compared to traditional centralized training, this allows the user data to stay local.

Google Research Philosophy Research Areas Publications People Resources Outreach Careers Blog

BLOG ›

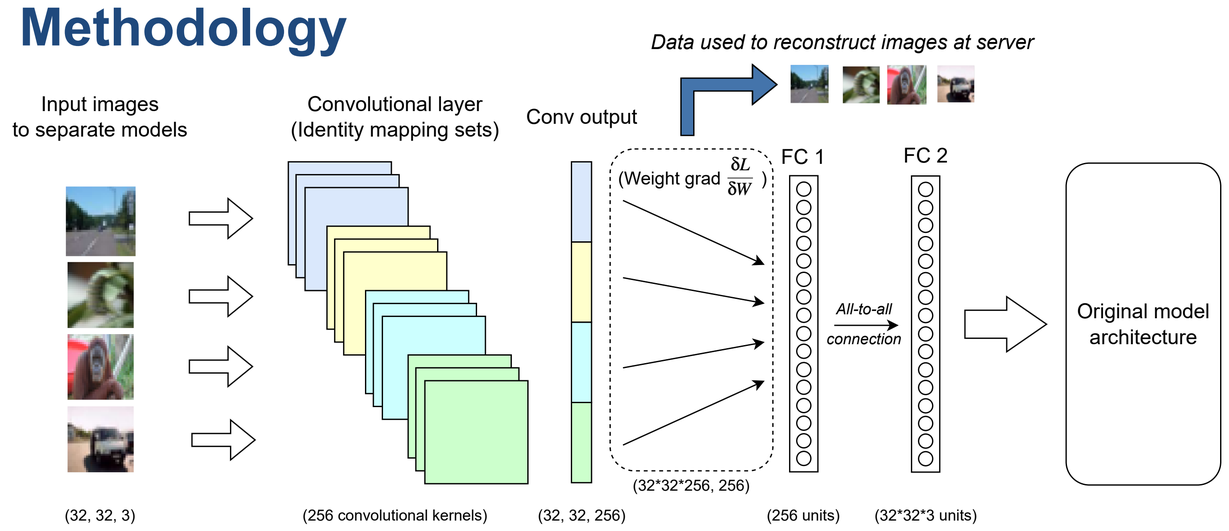
Federated Learning: Collaborative Machine Learning without Centralized Training Data

THURSDAY, APRIL 06, 2017

Posted by Brendan McMahan and Daniel Ramage, Research Scientists

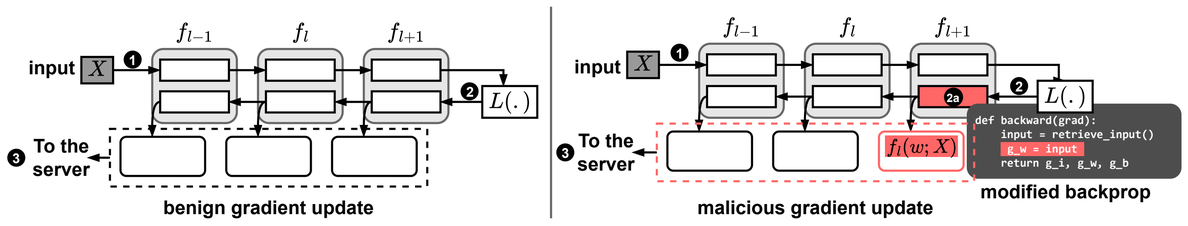
Key question: Can a malicious server still reconstruct the private training data using the client updates?

Methodology



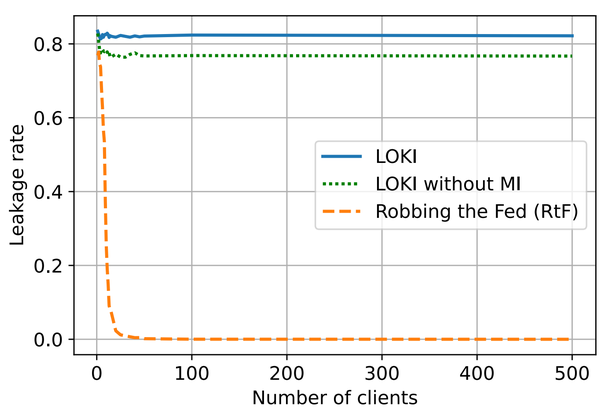
We introduce **LOKI**, a large-scale data reconstruction attack against FL. The attack can be

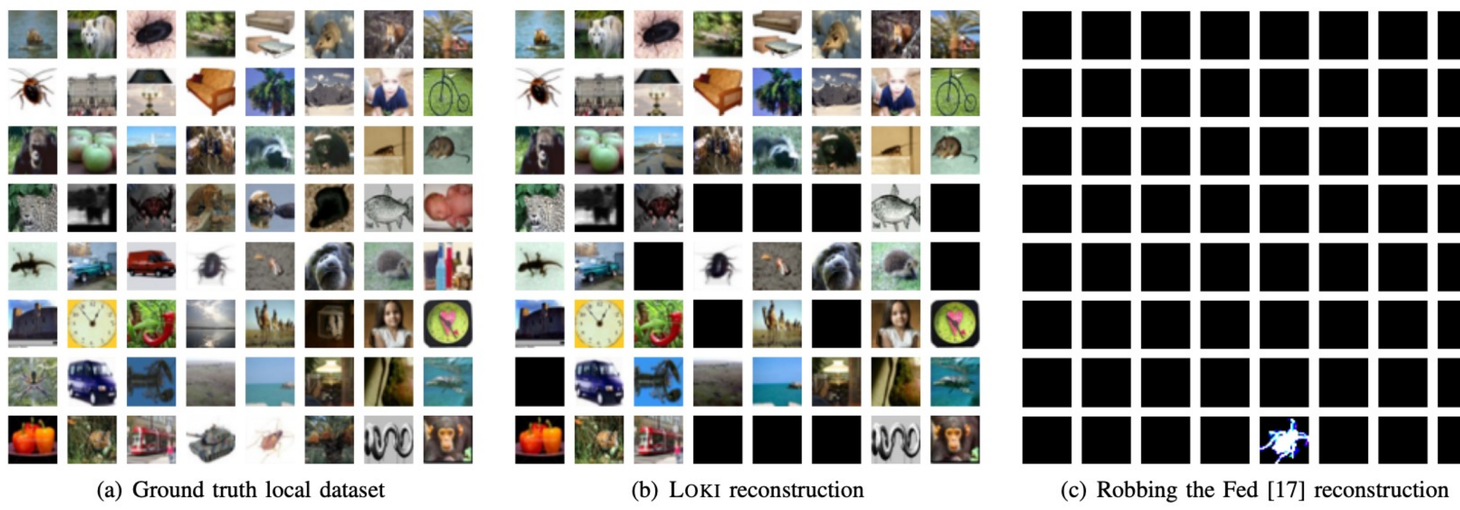
- By modifying the layers in front of the network, the gradient updates include information that can be used to directly recover the training images.
- Carefully crafted parameters allow LOKI to scale the attack even in the presence of defense methods such as secure aggregation and hundreds of clients.
- Reconstructed images are nearly identical to the private client training data.



Results

LOKI achieves higher leakage rate than prior SOTA. In the presence of secure aggregation, a standard defense mechanism used in FL, LOKI achieves scalability with no diminishing returns in the number of leaked images.





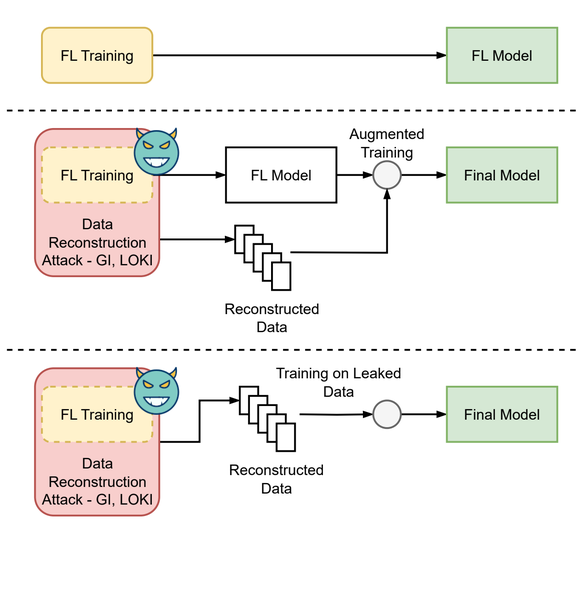
(a) Ground truth local dataset (b) LOKI reconstruction (c) Robbing the Fed [17] reconstruction

Can the malicious attacker use leaked data to train a more effective machine learning model than just FL alone?

When attacking 100 clients with a local dataset size of 64 images, LOKI can leak 5,290 images out of 6,400 in a single iteration (82.66%) leakage rate.

Dataset	Metrics	LOKI	RtF [17] + MI [20]
CIFAR-100	Leaked imgs	5290	50
	Total imgs	6400	6400
	Leakage rate	**82.66%**	**0.78%**
Tiny ImageNet	Leaked imgs	5202	49
	Total imgs	6400	6400
	Leakage rate	**81.28%**	**0.77%**
MNIST	Leaked imgs	4907	49
	Total imgs	6400	6400
	Leakage rate	**76.67%**	**0.77%**

TABLE 2. LEAKAGE RATE FOR FEDAVG AGGREGATED UPDATE WITH 100 PARTICIPATING CLIENTS. $\alpha = 1e-4$, $CSF = 100$ AND 5 LOCAL EPOCHS OF 8 ITERATIONS OF MINI BATCH SIZE 8 USED.



	Num. clients / Batch size / FC factor	Test accuracy
Centralized	-	94.38
FedAvg	10	72.76
	50	68.71
Gradient inversion (Inverting grads [6])	4	90.34
	8	86.16
	16	76.83
Linear layer leakage (LOKI [30])	**8**	**93.16**
	4	92.94
	2	91.90
	1	88.86

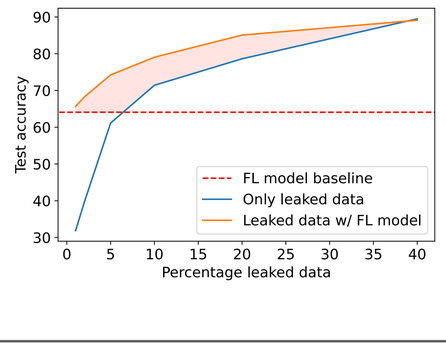
(a) CIFAR-10

	Num. clients / Batch size / FC factor	Test accuracy
Centralized	-	98.89
FedAvg	10	96.17
	50	96.18
Gradient inversion (Inverting grads [6])	8	95.96
	16	95.50
	32	94.29
Linear layer leakage (LOKI [30])	**8**	**98.82**
	4	98.70
	2	98.72
	1	98.46

(b) MNIST

	Num. clients / FC factor	Validation accuracy
Centralized	-	48.55
FedAvg	10	37.00
	50	35.06
Linear layer leakage (LOKI [30])	**8**	**46.70**
	4	45.09
	2	41.15
	1	35.20

(c) Tiny ImageNet



LOKI has higher leakage rate, but the high-quality leaked data also allows for training of machine learning models with higher accuracy than prior work.

PURDUE UNIVERSITY

CERIAS