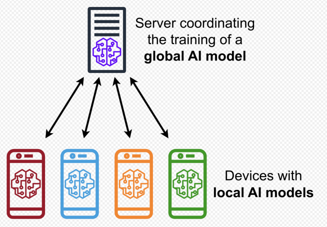
When the Curious Abandon Honesty: Federated Learning Is Not Private

Franziska Boenisch, Adam Dziedzic, Roei Schuster,
Ali Shahin Shamsabadi, Ilia Shumailov, Nicolas Papernot

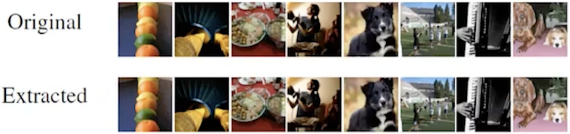
IEEE Euro S&P 2023

Federated Learning



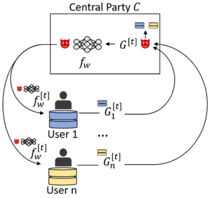
- Central party coordinates the training
- Data does not leave personal devices
  - Often presented as privacy-preserving
- Attacker observing gradient updates can reconstruct training data

Contributions



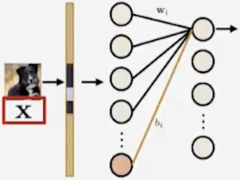
- ▷ Novel data reconstruction attack
  - Scales to large mini-batches of data
  - Computationally cheap
  - Stealthy
- ▷ Empirically demonstrate success of attack

Threat Model



- Untrusted, Active Central party
  - Knows type, domain, and dimensionality of data
    - Possesses some data from a similar dataset (preferably)
  - Instantiates the model, holds full control over the shared model weights
  - Architecture: Layer 1 is fully connected, with ReLU activation
  - Can read users' gradient updates in each iteration
  - Can choose which users to query in each iteration

Passive Attack: SGD

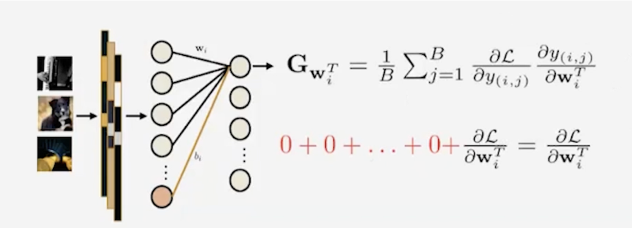


$$y_i = \underline{ReLU}(\mathbf{w}_i^T \mathbf{x} + b_i)$$
$$= \mathbf{w}_i^T \mathbf{x} + b_i \ \textbf{if}\ \mathbf{w}_i^T \mathbf{x} + b_i > 0$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_i^T} = \frac{\partial \mathcal{L}}{\partial y_i} \frac{\partial y_i}{\partial \mathbf{w}_i^T} = \frac{\partial \mathcal{L}}{\partial b_i} \boxed{\mathbf{x}}^T$$

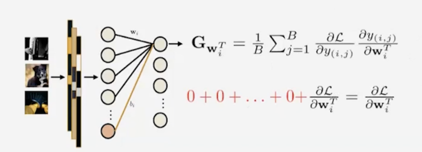
- Gradient contains a scaled version of the input
- Computed in the backward pass, comes at zero cost

Passive Attack: Mini-Batch SGD



- If lucky, exactly 1 training point x in the mini-batch has non-zero gradients
  - Iff exactly 1 training point x in the mini-batch has positive input
  - Reduces to the SGD case
  - Can reconstruct the training point x

Methodology



- Active central party, Mini-Batch
- Assume that the data features are scaled in the range [0,1]
- Let N and P denote the indices with negative and positive weights

$$\sum_{n \in \mathrm{N}} w_i^{(n)} x_n < \sum_{p \in \mathrm{P}} w_i^{(p)} x_p.$$

- Want the inequality to hold rarely (hopefully only for a single data point)
  - $\#N = \#P$
  - Initialize randomly by sampling from a Gaussian
  - Components in P are scaled down with a factor $s < 1$

Methodology

- Scaling decreases the impact of the positive weights
  - Causes most input data points to produce zero gradients
  - Hopefully only 1 input point has non-zero gradient
- How to choose s?
  - Trade-off
  - Dataset-dependent
  - Fine-tune on a small dataset which is similar
- Randomness in the initialization of the weights of each neuron $\implies$ Different data points are reconstructed

Experiments

▷ Mini-batch of 100 data points, 1000 neurons in layer 1

	% Extracted Data	
	Passive	**Active**
MNIST	5.8%	**54%**
CIFAR10	25.5%	**54%**
ImageNet	21.8%	**45.7%**
IMDB	25.4%	**65.4%**

▷ Smaller mini-batch sizes, more weight rows $\implies$ Stronger attack

Strengths

- Novel attack
  - Scales to large mini-batches of data
  - Effective
  - Simple
  - Computationally cheap
  - Stealthy

Weaknesses

- Assume the attacker possesses an auxillary dataset
- Not agnostic of the model architecture
- Need to get very lucky!
- Not many mitigations suggested
  - Local Differential Privacy (poor utility)
  - Large mini-batches

Discussion

- Weaker attack model
  - Passive attacker
  - No auxillary dataset
- Model agnostic
- Provable guarantees
- Mitigations

Acknowledgements

- Pictures in slides from
  - The paper on arxiv
    `https://arxiv.org/pdf/2112.02918.pdf`
  - Talk by Nicolas Papernot
    `https://machinelearning.apple.com/video/curious-honesty`
- Thanks!