Harmful Bias: A General Label-Leakage Attack on Federated Learning from Bias Gradients

Nadav Gat and Mahmood Sharif (Tel Aviv University)

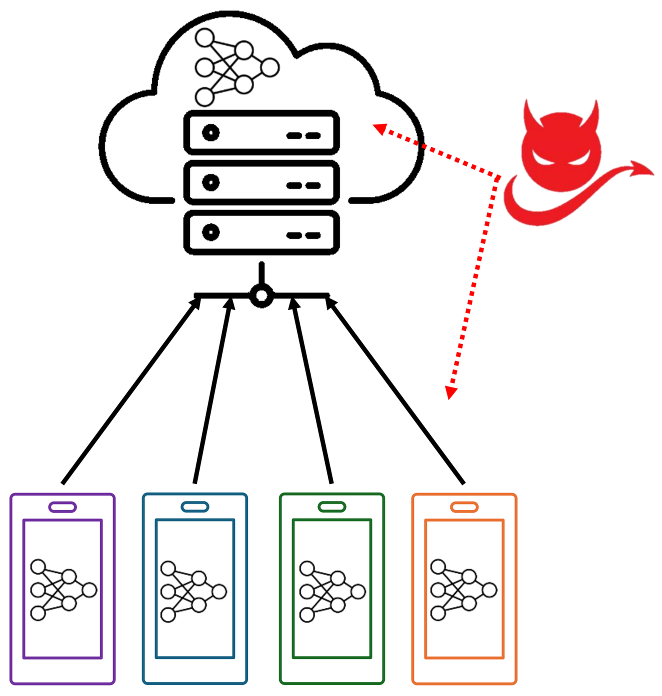
Code

Paper

Motivation

Federated Learning (FL) allows collaborative model training w/o sharing raw data.

Yet, FL carries **no privacy guarantees**.



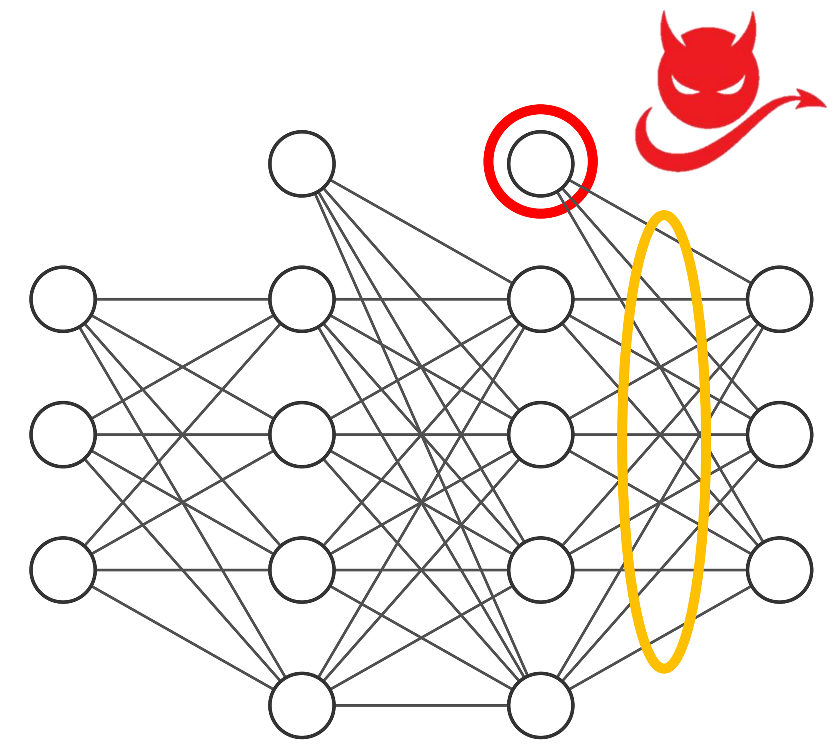
We study **label reconstruction:** extracting the batch labels from updates in classification tasks.

Labels may be **very sensitive** – personal text in Gboard, medical conditions, etc.

Labels also necessary for **data reconstruction**, users' raw samples may be revealed.

Background and Prior Work

SOTA label reconstruction: **LLG** (Wainakh et al. 2022) uses analysis of the last weight gradient. Assumes **non-negative activation** functions.



DLG (Zhu et al. 2019) randomly initializes labels and optimizes with data. Less accurate than LLG and heavier in compute.

Data reconstruction attacks (Geiping et al. 2020, Yin et al. 2021) rely on **correct labels** to achieve best results.

Different defenses suggested, most common is **Differential Privacy** (DP) – adding noise and clipping gradients (Abadi et al. 2016, El Ouadrhiri and Abdelhadi 2022).

Our Approach

We analyze the gradient of the last bias and find:

$$\nabla b_L^i = -\frac{\lambda_i}{B} + \frac{1}{B}\sum_{k=1}^{B} p_i^{(k)}$$

Batch size

# of occurrences of label i

Prob. of class i for sample k

Untrained models: $p_i^{(k)} \approx 0$

Trained models: $p_i^{(k)} \approx v_i$, average class confidence

$\rightarrow \lambda_i$ can be inferred given ∇b_L^i

$LLBG_\gamma - v_i$ guessed to be a constant

$LLBG_{aux} - v_i$ estimated per class with auxiliary data

Algorithm: LLBG attack against trained models

Input: $\beta = \nabla b_L$, batch size B, average confidence of model per class $v = (v_1, \ldots, v_n)$

$m \leftarrow -\frac{1}{B}$;

$C' \leftarrow [\,]$; // Initialize labels list

for $i \leftarrow 1$ **to** n **do** // Guaranteed labels

  if $\beta_i < 0$ **then**

    $C' \leftarrow C' + [i]$;

    $\beta_i \leftarrow \beta_i - m \cdot (1 - v_i)$; // Sample impact

while $|C'| < B$ **do** // Heuristic for rest

  $l \leftarrow \arg\min\{\beta\}$;

  $C' \leftarrow C' + [l]$;

  $\beta_l \leftarrow \beta_l - m \cdot (1 - v_l)$

return C'

Results

Baselines: **LLG**, **EBI** – bias gradient with empirical estimation.

2 vision datasets, 9 different models (untrained and trained), several defenses.

Label Reconstruction success vs. MLPs w/ diff. activations

Activation	LLG	EBI	LLBG
ReLU	81.93 ± 1.94	79.11 ± 1.38	$\mathbf{99.56 \pm 0.39}$
LeakyReLU	81.95 ± 1.95	79.11 ± 1.38	$\mathbf{99.56 \pm 0.39}$
Sigmoid	82.88 ± 1.61	82.80 ± 1.62	$\mathbf{97.62 \pm 0.94}$
Tanh	36.72 ± 21.10	79.16 ± 1.34	$\mathbf{99.48 \pm 0.40}$

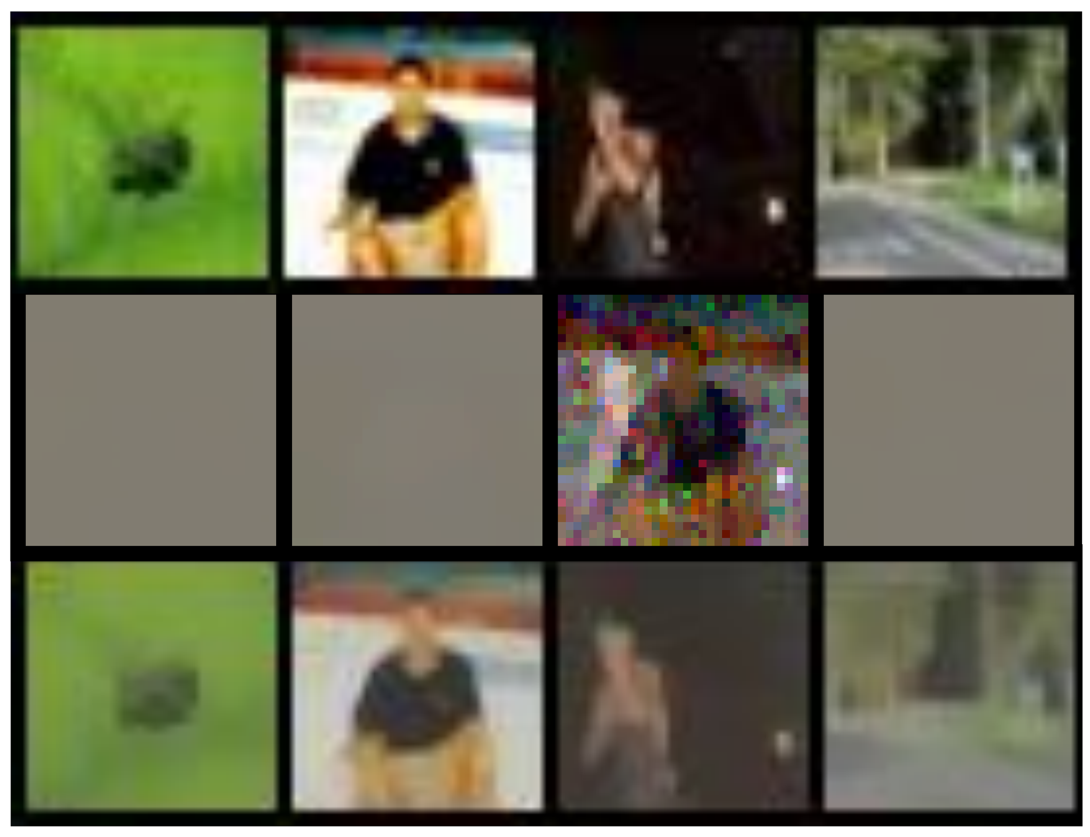
LLBG – highest success in **45 out of 52 cases.**

Data reconstructed using labels reconstructed w/ LLG is of higher fidelity, both **empirically and visually**.

Original batch

Batch reconstructed w/ LLG

Batch reconstructed w/ LLBG



LLBG was more robust against most defenses, except for a defense tailored for it – **removing the bias** parameters.