Auditing Privacy Defenses in Federated Learning via Generative Gradient Leakage
CVPR 2022

Presenter : M Anand Krishna
CS22MTECH14003
IIT Hyderabad

Under Supervision of : Prof C Krishna Mohan,
TA : Zarka Bashir

February 23, 2023

Overview

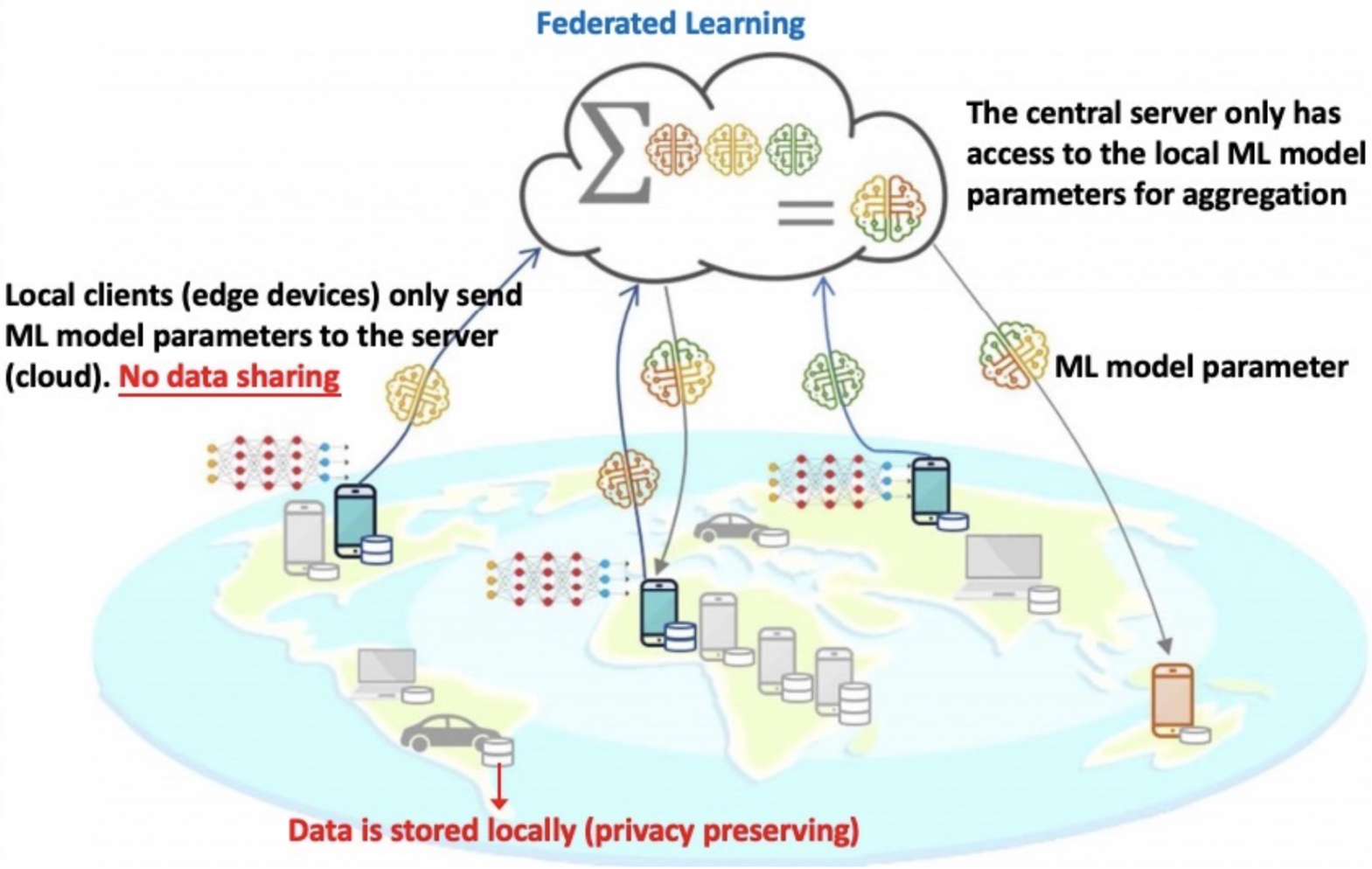
Federated Learning

- Multiple clients collaboratively learn a shared model under coordination of central server

- Clients **do not share private data** instead, they just share the gradient information to central server to update the model

Federated Learning Contd.

- Collaborative learning without centralized training data



Federated Learning Applications.

- Mobile computing
  - e.g. text prediction in Google Gboard
- Telemedicine
  - e.g. multi institutional collaborations for learning medical diagnosis model without sharing patient's private data

Privacy Risks in FL

- Even though only model parameters/gradient information is shared from client, it is found that private information can still be extracted from **exchanged gradients**

Privacy Risks in FL contd..

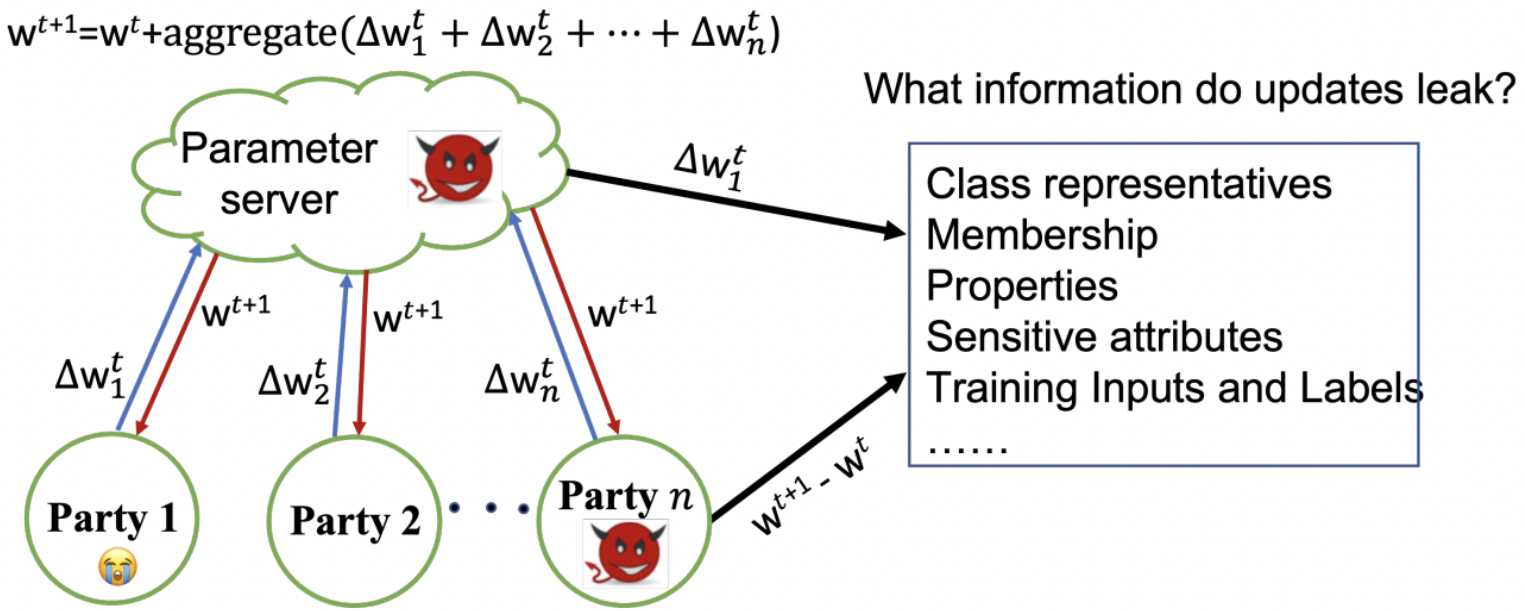


Figure: A demo of privacy leakage in FL. Attacker can infer various private information about the victim participant from the received gradients or the snapshot of the FL model parameters. [2]

[2] Image Credits : **Privacy and Robustness in Federated Learning: Attacks and Defenses**

How to prevent privacy leakage? - Currently Available Privacy Defenses

- Crypto-based solutions
  - e.g. Secure Multi-party computation, Homomorphic encryption. Remains vulnerable to the inference over the output
- Gradient Degradation based solutions
  - Differential Privacy:
    - Use a randomised technique to distort the gradients before sharing it to central server.
    - But there is trade-off in differential privacy is between privacy and utility.
- Gradient Sparsification
  - Compress the gradients by pruning small values

Are these defenses sufficient?

- In this work, author studies the challenging scenario, where clients applies local defenses before sharing the gradients
- And author tries to reconstruct a high fidelity images from the shared degraded gradients.

Are these defenses sufficient? Contd..

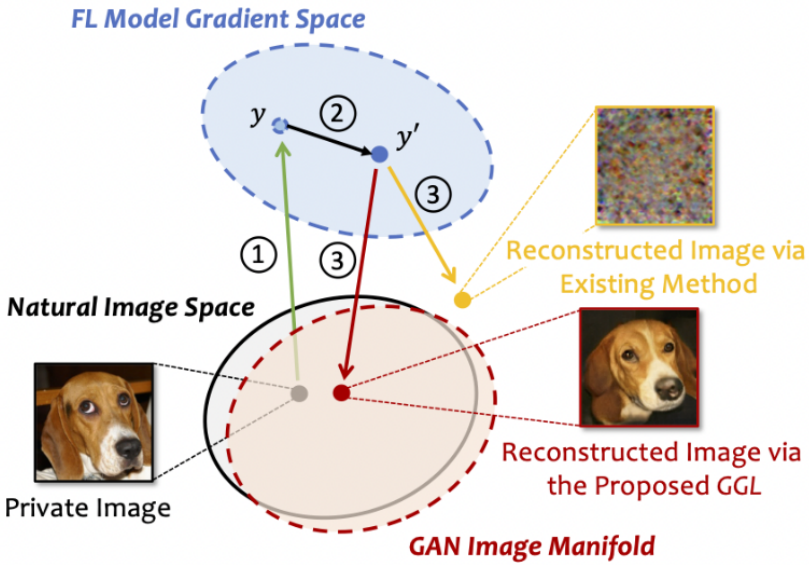


Figure: Illustration of data leakage via gradient.[4]
(1) Client computes gradients on its private data
(2) Client applies defense to degrade the computed gradients y
(3) Adversary attempts to reconstruct the private image from the shared gradients y'

[4] Image Credits : **Auditing Privacy Defenses in Federated Learning via Generative Gradient Leakage**

Problem formulation

The task of reconstructing a training image $x \in R^d$ from its gradients $y \in R^m$ can be formulated as a non-linear inverse problem

$$\mathbf{y} = F(\mathbf{x}),$$

where $F(\mathbf{x}) = \nabla_{\boldsymbol{\theta}} \mathcal{L}\left(f_{\boldsymbol{\theta}}(\mathbf{x}), c\right)$ is forward operator that calculates the gradients of the loss, provided with label c and FL model f_θ
When defense is applied at the client's side, the problem becomes:

$$\mathbf{y} = \mathcal{T}(F(\mathbf{x})) + \boldsymbol{\varepsilon},$$

where $\mathcal{T}(\cdot)$ is referred to as the lossy transformation (e.g., sparsification) and ε is the additive noise (e.g., DP)

Current Approach and Its limitation

- Existing methods aim to solve this inverse problem by using image priors in a penalty form

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathbb{R}^d}{\operatorname{argmin}} \mathcal{D}(\mathbf{y}, F(\mathbf{x})) + \lambda\omega(\mathbf{x}),$$

where $\mathcal{D}(\cdot)$ is a distance metric, $\omega(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ is the standard image prior (e.g., total variation)

Current Approach and Its limitation

- Only effective for reconstructing images from the actual gradients
- When reconstructing from a set of low-fidelity and noisy gradients, it would suffer from the limited identification ability of hand-crafted priors
- Result: return false solutions that are not valid natural images

Generative Gradient Leakage

- Insight
  - Leverage a generative model trained on public image datasets as a learned natural image prior to ensure that reconstructed image is of good image quality
- Given a well-trained generator $G(\cdot)$, we solve:

$$\mathbf{z}^* = \underset{\mathbf{z}\in\mathbb{R}^k}{\operatorname{argmin}} \underbrace{\mathcal{D}(\mathbf{y}, \mathcal{T}(F(G(\mathbf{z}))))}_{\text{gradient matching loss}} + \lambda \underbrace{\mathcal{R}(G;\mathbf{z})}_{\text{regularization}} ,$$

where $z \in R^k$ is the latent space of the generative model, and $\mathcal{R}(G;\mathbf{z})$ is a regularization term that penalizes latent vectors which deviate from the prior distribution

Generative Gradient Leakage

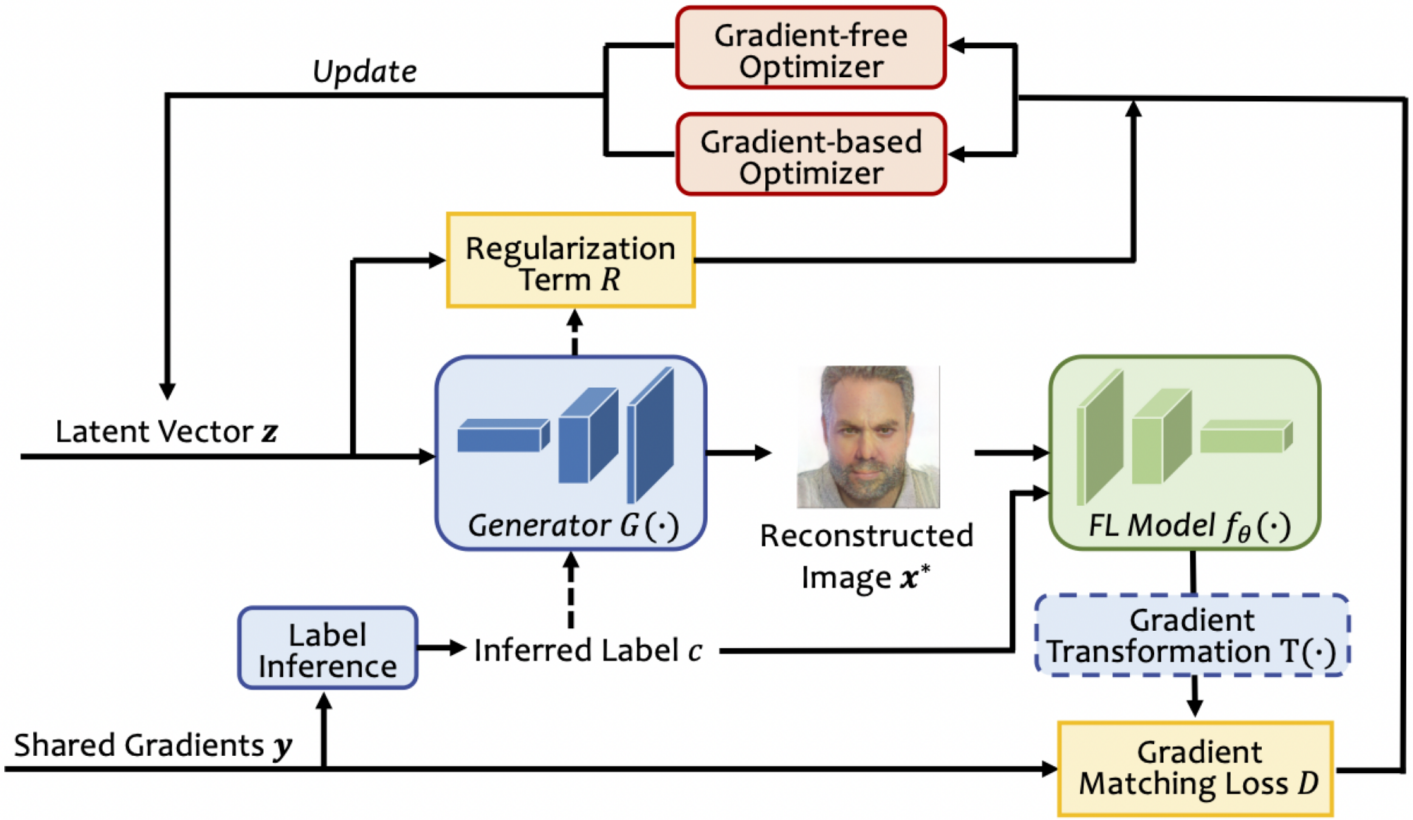


Figure: Overview of author's approach [5]

[5] Image Credits : **Auditing Privacy Defenses in Federated Learning via Generative Gradient Leakage**

Optimization strategies

- Challenges
  - The target inverse problem is highly *non convex and non linear*.
  - The existing data reconstruction attacks are based on the **gradient based optimiser** like $L - BFGS$ [6] and *Adam*.

- Issues with gradient based optimiser
  - Outcome is highly based on choice of initialization
  - For complex models, often high chance that cost function will converge to sub optimal minima

[6] Limited-memory BFGS (L-BFGS or LM-BFGS) is an optimization algorithm in the family of quasi-Newton methods that approximates the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS)

Proposed optimisation strategy

- Author explores two gradient-free optimisation strategies
  - Bayesian optimisation(BO)
  - Covariance Matrix Adaptation Evolution Strategy (CMA-ES)

Datasets

- CelebA
  - The dataset was collected from the Internet and contains images of celebrities with a diverse range of ethnicities, ages, and facial features.
- ImageNet
  - ImageNet is a large image database with over 14 million images in 20,000 categories for computer vision tasks.

ImageNet and CelebA has become a popular benchmark dataset for a variety of computer vision tasks, including face recognition, attribute prediction, and generative models.

Experiment Results

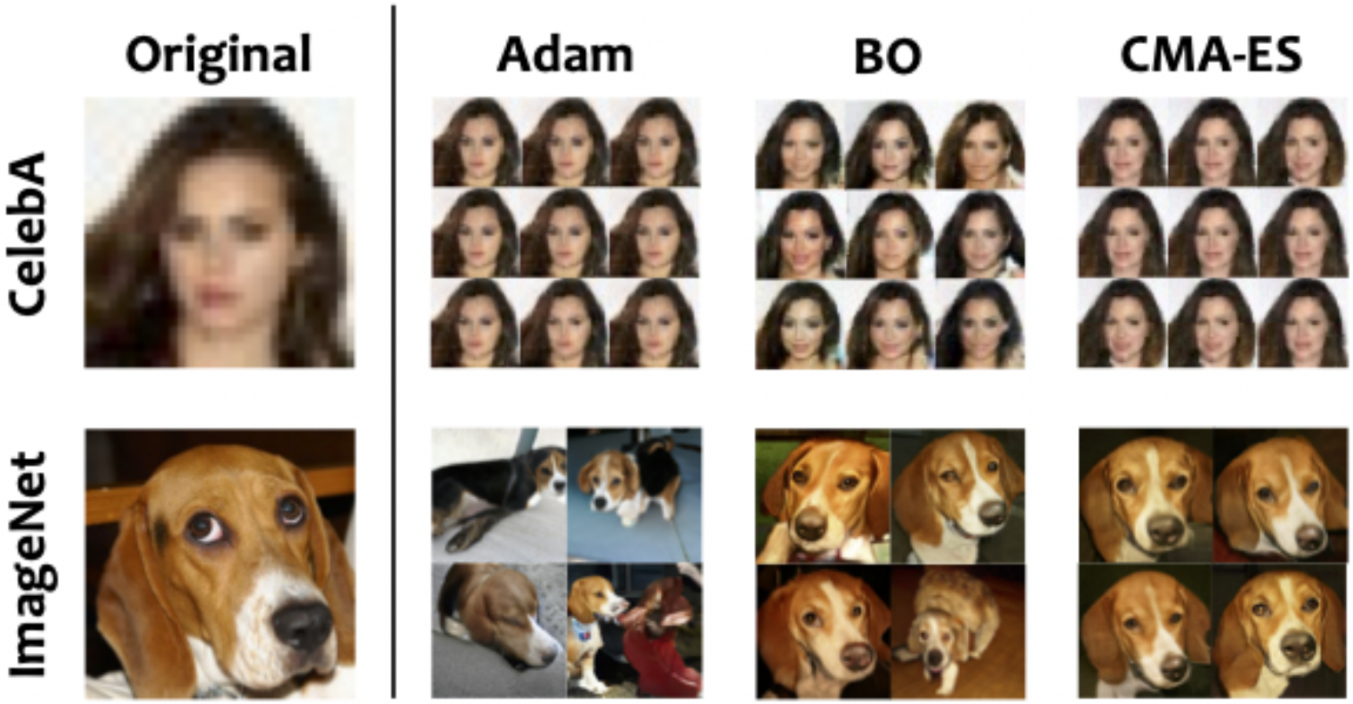


Figure: Visual comparison of different optimizers. The images on the right are the reconstruction samples produced by three types of optimizers with different random seeds[8]

[8]Image Credits : **Auditing Privacy Defenses in Federated Learning via Generative Gradient Leakage**

My observation

- On CelebA, gradient-based and gradient-free optimizers show similar performance
- On ImageNet, the gradient-based Adam fails to recover any useful information
- Gradient-free optimizers **(BO and CMA-ES)** are still able to find samples that resemble the original private image and are more resilient to different initialization conditions

Experiment Results Contd.

Dataset	Metric	Adam		BO		CMA-ES	
		Mean	Std.	Mean	Std.	Mean	Std.
CelebA	MSE-I ↓	**0.0427**	0.0025	0.0813	0.0131	0.0708	**0.0008**
	PSNR ↑	**13.6965**	0.2593	10.9455	0.6816	11.4989	**0.0533**
	LPIPS ↓	**0.1435**	**0.0083**	0.2162	0.0328	0.2136	0.0133
	MSE-R ↓	**0.0003**	**0.0001**	0.0012	0.0003	0.0015	0.0022
ImageNet	MSE-I ↓	0.5918	0.1955	**0.2648**	0.0181	0.2667	**0.0119**
	PSNR ↑	2.4433	1.3565	**5.7783**	0.2992	5.7420	**0.1988**
	LPIPS ↓	0.7983	0.0280	0.6166	0.0590	**0.5736**	**0.0209**
	MSE-R ↓	0.1051	0.0703	0.0035	0.0005	**0.0018**	**0.0002**

Figure: Quantitative comparison of different optimizers.[9]

- MSE-I: the mean squared error (MSE) between the generated and true gradients of the model at iteration i.
- PSNR: the peak signal-to-noise ratio (PSNR) between the generated and true gradients of the model at iteration i.
- LPIPS: the learned perceptual image patch similarity (LPIPS) between the generated and true gradients of the model at iteration i
- MSE-R: the MSE between the generated and true gradients of the model over all iterations

[9] Image Credits : **Auditing Privacy Defenses in Federated Learning via Generative Gradient Leakage**

Experiment Results Contd.

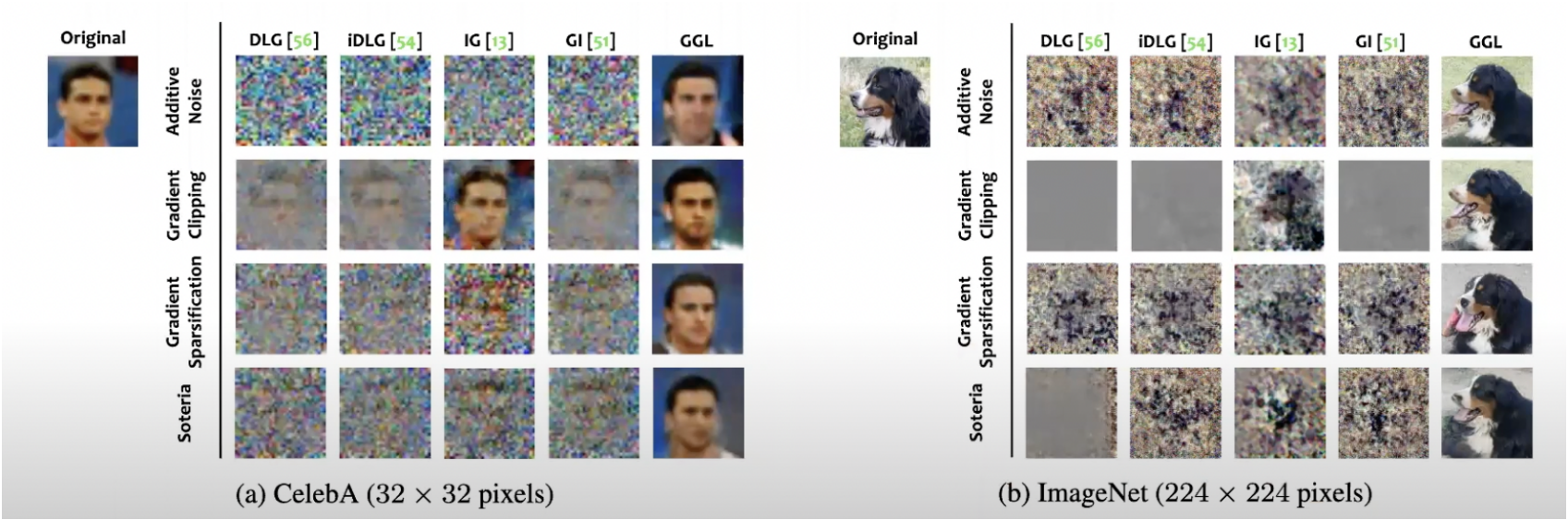


Figure: Comparison with other existing attacks .[10]

[10] Image Credits : **Auditing Privacy Defenses in Federated Learning via Generative Gradient Leakage**

Experiment Results Contd.

Dataset	Attack	Additive Noise [44, 56]				Gradient Clipping [14, 48]				Gradient Sparsification [56]				Soteria [44]			
		MSE-I ↓	PSNR ↑	LPIPS ↓	MSE-R ↓	MSE-I ↓	PSNR ↑	LPIPS ↓	MSE-R ↓	MSE-I ↓	PSNR ↑	LPIPS ↓	MSE-R ↓	MSE-I ↓	PSNR ↑	LPIPS ↓	MSE-R ↓
CelebA	DLG [56]	0.6479	1.8843	0.8197	0.0021	0.2097	6.7831	0.7375	0.0326	0.3335	4.7679	0.7986	0.0155	0.3624	4.4069	0.8007	0.0285
	iDLG [54]	0.6261	2.0329	0.8209	0.0025	0.1960	7.0762	0.7280	0.0326	0.3301	4.8124	0.8035	0.0162	0.3269	4.8553	0.8036	0.0396
	IG [13]	0.4880	3.1151	0.8260	0.0097	**0.0543**	**12.6517**	0.2998	**0.0003**	0.4103	3.8687	0.7975	0.0113	0.3441	4.6326	0.8008	0.0316
	GI [51]	0.5738	2.4116	0.8302	0.0023	0.1790	7.4701	0.7142	0.0322	0.2958	5.2888	0.7775	0.0163	0.3179	4.9768	0.7991	0.0409
	GGL	**0.0780**	**11.0766**	**0.1906**	**0.0010**	0.0760	11.1902	**0.1670**	0.0015	**0.0768**	**11.1466**	**0.1620**	**0.0007**	**0.0968**	**10.1434**	**0.2561**	**0.0007**
ImageNet	DLG [56]	0.7438	1.2852	0.9353	0.0049	0.3809	4.1912	0.9798	2.1610	0.4432	3.5336	0.8907	0.0075	0.5990	2.2253	0.9195	0.5415
	iDLG [54]	0.7352	1.3359	0.9392	0.0041	0.3699	4.3190	0.9473	1.8810	0.4357	3.6077	0.8935	0.0077	0.6089	2.1542	0.9198	0.5425
	IG [13]	0.3081	5.1120	0.8677	0.4490	**0.1432**	**8.4386**	0.7476	0.0214	0.2993	5.2376	0.8805	0.0501	0.3683	4.3373	0.8700	0.5057
	GI [51]	0.6593	1.8090	0.9448	0.0031	0.3702	4.3154	0.9451	1.8807	0.4404	3.5611	0.8889	0.0072	0.6235	2.0511	0.9169	0.5792
	GGL	**0.2686**	**5.7089**	**0.5915**	**0.0018**	0.2230	6.5163	**0.5592**	**0.0015**	**0.2141**	**6.6920**	**0.5170**	**0.0017**	**0.2484**	**6.0477**	**0.5685**	**0.0022**

Figure: Observation : Existing attacks [12] struggle to reconstruct a realistic image with the present of any privacy defense mechanism. The proposed GGL is still able to synthesize high quality images that are similar to the original ones

12

1. Deep Leakage from Gradients (DLG) : gradient leakage attack with l2 gradient matching loss and L-BFGS optimizer;
2. Improved Deep Leakage from Gradients (iDLG): improved DLG attack with label inference;
3. Inverting Gradients (IG) : gradient leakage attack with cosine distance as loss and total variation as prior, optimized using Adam;
4. GradInversion (GI): gradient leakage attack with l2 gradient matching loss and Adam optimizer.

Future Work / Novelty idea

- Applying the gradient leakage attack to other types of federated learning, such as horizontal federated learning or federated transfer learning.
- Developing novel privacy defenses specifically designed to defend against the gradient leakage attack.

References

Nitin Agrawal, Ali Shahin Shamsabadi, Matt J Kusner, and Adria‘ Gasco n

Quotient: two-party secure neural network training and prediction.

Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (2019)

Amir Beck and Marc Teboulle

”Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems”.

IEEE transactions on image processing (2009)

THANK YOU


भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad