भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Soteria: Provable Defence against Privacy Leakage in Federated Learning from Representation Perspective

Sun, Jingwei, Ang Li, Binghui Wang, Huanrui Yang, Hai Li, and Yiran Chen. "Soteria: Provable defense against privacy leakage in federated learning from representation perspective." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9311-9319. 2021.

Guided by:
Prof. C Krishna Mohan

Presenter:
Yash Shukla
CS23MTECH14018

Teaching Assistant:
K Naveen Kumar
PhD Research Scholar

Index

- Introduction
- Problem statement(Recap)
- Limitations
- Motivation
- Proposed method
- Datasets
- Research gaps..
- Datasets / model..
- Experimentation/results..
- Research Challenges faced / gaps..
- Novelty..
- Conclusion..
- Future work
- References..
- Reproduced results .. In that hyperparameters

Introduction

- Federated Learning(FL) decentralizes model training across devices.
- preserving privacy by locally training models and **sharing only updates**
- However, recent concerns have arisen regarding **privacy leakage** in FL, particularly due to the sharing of **model updates** among participating devices
- Essential cause of privacy leakage in FL **remains incompletely understood.** Hence hindering the development of robust defence.

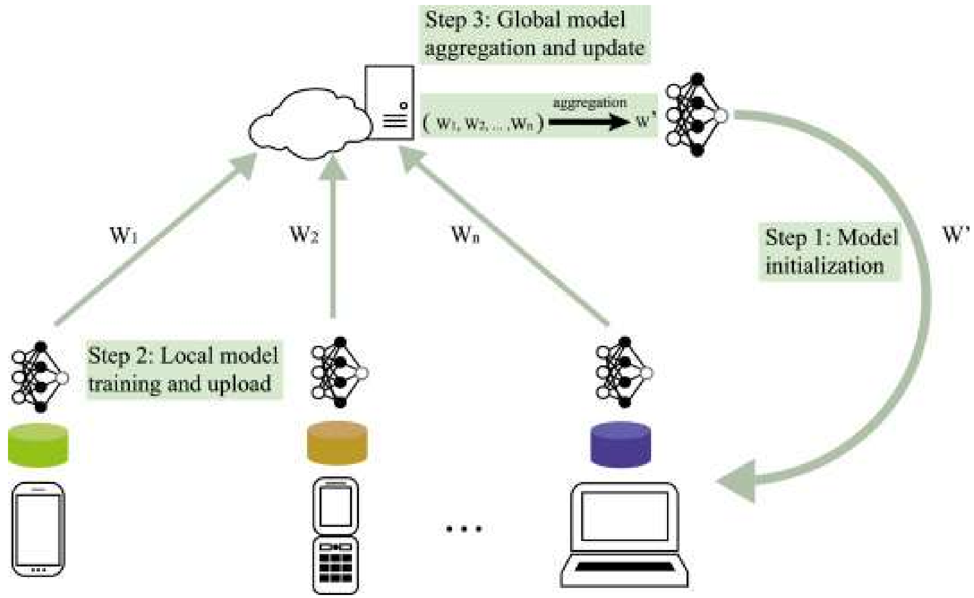


Image Reference

Problem statement

- Current defense strategies have been presented to prevent privacy leakage like differential privacy, secure multi-party computation, and data compression.
- But this approaches incur **either significant computational overhead** or **unignorable accuracy loss**.
- Sharing model updates makes vulnerable to **inference attacks** like **property inference attack** and **model inversion attack**.
- The essential cause of privacy leakage in FL, specifically concerning **data representation leakage** from model updates, has not been thoroughly explored.

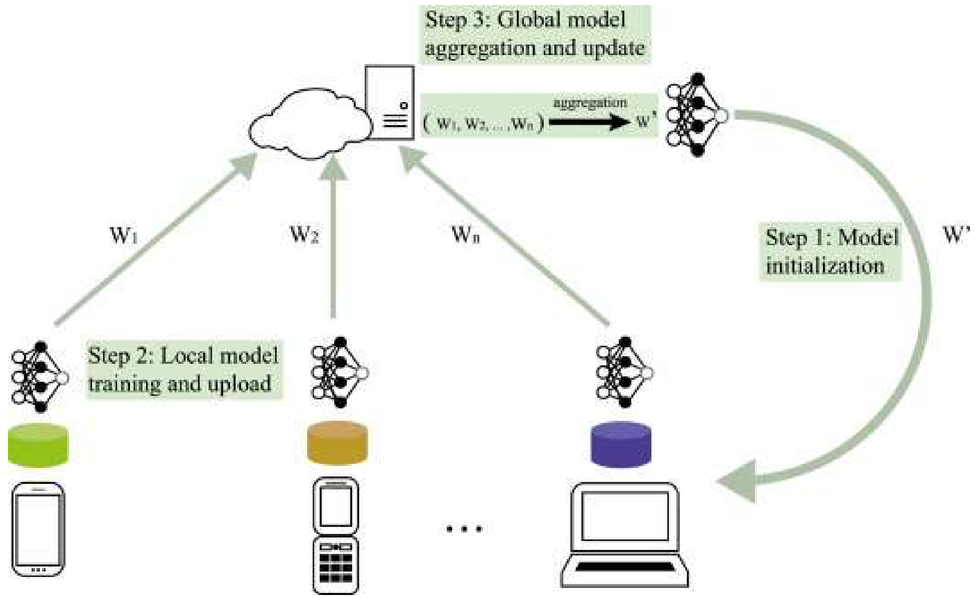


Image Reference

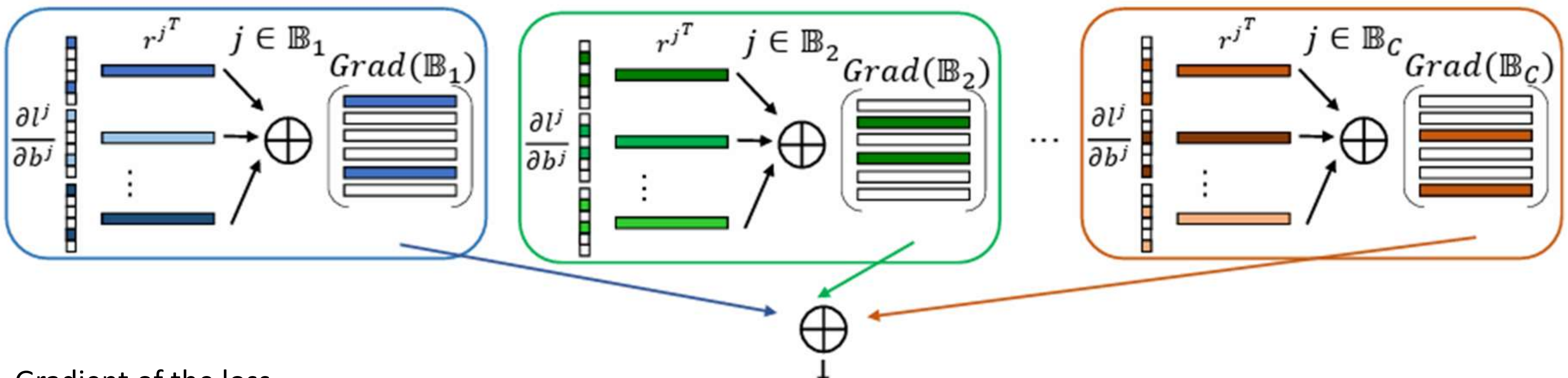
Limitations (of previous work) and Motivation

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- **Non-IID** data **characteristics exacerbate representation leakage**, further compromising privacy.
- Key observation is that the **data representation leakage** from gradients serves as the **essential cause** of privacy leakage in FL.
- The class-wise **data representations are embedded** in **shared local model updates**, and such data representations can be inferred to perform **model inversion attacks** like DLG (**Deep Leakage from Gradients**) and GS(**Gradient Similarity**)
- Therefore, the information can be severely leaked through the model updates.

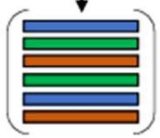
Limitations (of previous work) and Motivation (Contd.)

Data representations tend to be embedded in different rows of gradient(intuition behind the equation)



Gradient of the loss function (l) is calculated with respect to the weight matrix of the FC layer

$$\frac{\frac{1}{B}\sum_{k=1}^{B}\partial l^k}{\partial W}$$



Change in loss with respect to the output for each data sample in batch

Change in loss with respect to the output for each **data sample in batch**

Represents the gradient with respect to the data samples in Bi

Provides an overall gradients that indicates how the weights in the FC layer should be adjusted to minimize the loss function across the entire batch of data

$$= \sum_{i=1}^{C}\left(\frac{1}{|\mathbb{B}_i|}\sum_{j\in\mathbb{B}_i}\frac{\partial l^j}{\partial \boldsymbol{b}^j}(\boldsymbol{r}^j)^{\mathrm{T}}\right) \triangleq \sum_{i=1}^{C} Grad(\mathbb{B}_i)$$

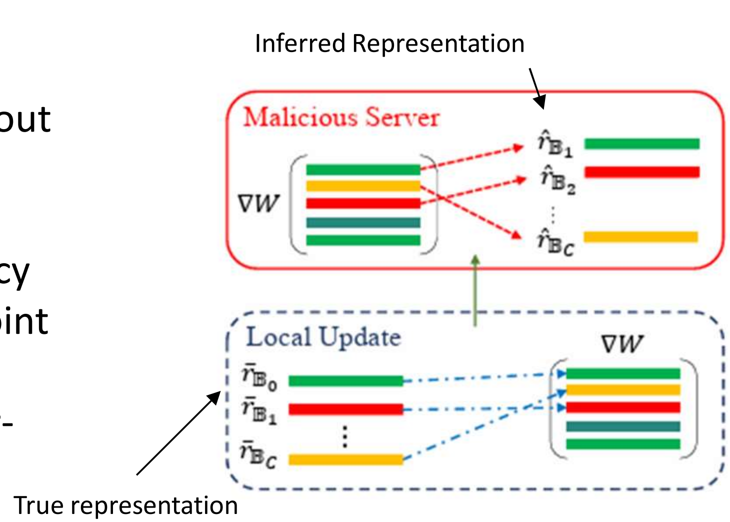
Figure : Illustration of the gradient updates of class-wise data in a batch

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Limitations (of previous work) and Motivation(Contd)

Embedded in gradients(Inferring the class-wise representations) :

- Different classes have different representations, the information about representations gets reflected in different rows of the overall gradients.
- Data stays more separate (**less entagled in gradients**) which is privacy concern ,because its easier for attacker to understand which datapoint belongs to which classes if they access gradients.
- Colored bars represents **the magnitude of values** in gradients ,color-coded indicates different classes



Inferring the class-wise data representations :

- Evaluation metric : **Correlation co-efficient** (cor) between true data representation and inferred representation for each class on each participating device.
- Cor is much lower compared to non-IID settings. Because having more diverse data on each device makes representation harder to isolate.

Table 1. Average *cor* across 200 communication rounds for different layers under different settings.

Local Training Configurations	FC1	FC2	FC3
E=1, B=32	0.98	0.99	0.99
E=5, B=32	0.82	0.90	0.92
E=10, B=32	0.70	0.78	0.82
E=1, B=16	0.82	0.93	0.99
E=1, B=8	0.85	0.89	0.92
E=1, B=32 (IID)	0.48	0.31	0.18

Proposed Defence method (Soteria)

- Soteria aims to perturb the data representations in a specific layer with **goals**:
  - **Reduce Privacy leakage** : Perturbed Representations should make it difficult to reconstruct the original input data
  - **Maintain Model performance** : Perturbed Representations should remain similar to the original representation to avoid impacting model accuracy.

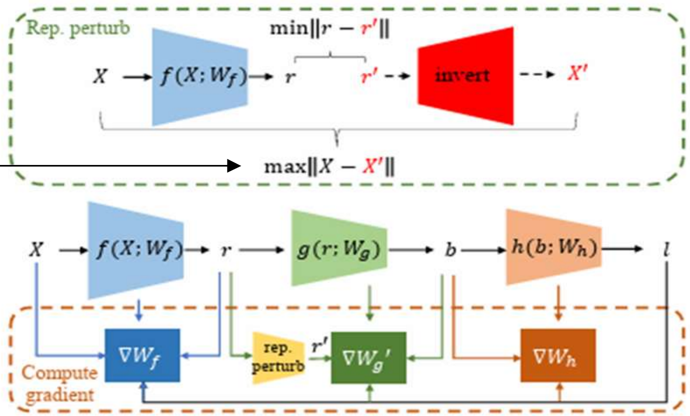


Figure 4. Illustration of our representation perturbation defense.

Formalization of Goals :

X : Raw Input Data

r': Reconstructed Representation after Perturbation

r : Clean Data Representation (without Perturbation)

X' : Reconstructed Input Data using the Perturbed Representation

Proposed Defence method (Soteria)(Contd)

- Optimization problem : Finding optimal perturbed data representation (r') that satisfies the two goals mentioned earlier.

Objective : **Achieving Goal 1:** $\max_{r'} ||X - X'||_p,$

Constraint : **Achieving Goal 2:** s.t., $||r - r'||_q \leq \epsilon,$

- Flow of Information : (X)(Original Data) → Feature Extractor(f) → (Cleaned Representation)(r)
- This Algorithm is Identifying the largest elements in a set derived From the data representation & gradients. These elements are use To create perturbed representations.
- Lp Norms measures the distance between two points (reconstructed input vs original input) larger norms – more Information content.
- P=2 : This corresponds to MSE between reconstructed & Original Input,which defence aims to maximize (**increase dissimilarity**)
- Q=0 : choice simplifies the solution & improves communication efficiency

Algorithm 1 Learning perturbed representation r' with $q = 0$ and $p = 2$.

Input: Training data $X \in \mathbb{R}^{M \times N}$; Feature extractor $f : \mathbb{R}^{M \times N} \to \mathbb{R}^L$ before the defended layer; Clean data representation $r \in \mathbb{R}^L$; Perturbation bound: ϵ;
Output: Perturbed data representation $r' \in \mathbb{R}^L$;

function PERTURB_REP(X, f, r, ϵ)
  Compute $||r_i(\nabla_X f(r_i))^{-1}||_2$ for $i = 0, 1, ..., L-1$;
  Find the set $\mathbb{S}$ which contains the indices of ϵ largest elements in $\{||r_i(\nabla_X f(r_i))^{-1}||_2\}_{i=1}^{L}$;
  $r' \leftarrow r$;
  Set $r'_i = 0$ for $i \in \mathbb{S}$;
  return r';
end function

$$r' = \arg\max_{r'} ||X - X'||_p, \;\; s.t. ||r - r'||_q \leq \epsilon$$

Proposed Defence method (Soteria)(Contd):

- **Certified Robustness Guarantee** :

$$||X - X'||_p \geq \frac{||r - r'||_p}{||\nabla_X f||_p}.$$

Distance between original data(X) & Reconstructed data(X')

Specific value of p determine how the distance is calculated.
L2(Euclidean distance)
p=2(In this case)

Magnitude of the function applied to input(how certain model was)

Represents the distance between original representation and perturbed representation(smaller the distance ensures both are nearly similar.(minimizing the impact on model performance

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Proposed Defence method (Soteria)(Contd)

This algorithm trains a local model on a device while incorporating a defence mechanism(g)

- Calculates loss & feature representation
- Calculates feature representation by applying feature extractor
- Calculates output(b) of the defended layer by applying it to the feature representation(r)
- Calculates feature representation(l) after defended layer (h) to the output (b) from the defended layer.
- Computes the gradients of the loss function with respect to model parameters
- Perturbing the feature representation and update model parameters.

Algorithm 2 Local training process with our defense on a local device.

Input: Training data $\boldsymbol{X} \in \mathbb{R}^{M\times N}$; Local objective function $F : \mathbb{R}^{M\times N} \rightarrow \mathbb{R}$; Feature extractor $f : \boldsymbol{W}_f \in \mathbb{R}^{M\times N} \rightarrow \mathbb{R}^L$ before the defended layer; The defended layer $g : \boldsymbol{W}_g \in \mathbb{R}^L \rightarrow \mathbb{R}^K$; Feature extractor after the defended layer $h : \boldsymbol{W}_h \in \mathbb{R}^K \rightarrow \mathbb{R}$; Local model parameters $\boldsymbol{W} = \{\boldsymbol{W}_f, \boldsymbol{W}_g, \boldsymbol{W}_h\}$; Learning rate η.

Output: Learnt model parameter $\boldsymbol{W}$ with our defense.

Initialize $\boldsymbol{W}$;
for $\mathbb{B}$ in local training batches **do**
  for $\boldsymbol{X} \in \mathbb{B}$ **do**
    $l \leftarrow F(\boldsymbol{X}; \boldsymbol{W})$;
    $\boldsymbol{r} \leftarrow f(\boldsymbol{X}; \boldsymbol{W}_f)$;
    $\boldsymbol{b} \leftarrow g(\boldsymbol{r}; \boldsymbol{W}_g)$; // e.g., $\boldsymbol{b} = \boldsymbol{W}_g \boldsymbol{r}$ for FC layers
    $l \leftarrow h(\boldsymbol{b}; \boldsymbol{W}_h)$;
    $\{\nabla \boldsymbol{W}_f, \nabla \boldsymbol{W}_g, \nabla \boldsymbol{W}_h\} \leftarrow \nabla_{\boldsymbol{W}} F(\boldsymbol{X}; \boldsymbol{W})$;
    $\boldsymbol{r}' \leftarrow Perturb\_rep(\boldsymbol{X}, f(; \boldsymbol{W}_f), \boldsymbol{r}, \epsilon)$;
    $\nabla \boldsymbol{W}_g' \leftarrow \tau(l, b, r', \boldsymbol{W}_g)$; // e.g., $\nabla \boldsymbol{W}_g' = \frac{\partial l}{\partial \boldsymbol{b}} \boldsymbol{r}'^T$ in FC
    $\nabla \boldsymbol{W} = \{\nabla \boldsymbol{W}_f, \nabla \boldsymbol{W}_g', \nabla \boldsymbol{W}_h\}$;
    $\boldsymbol{W} \leftarrow \boldsymbol{W} - \eta \nabla \boldsymbol{W}$;
  end for
end for

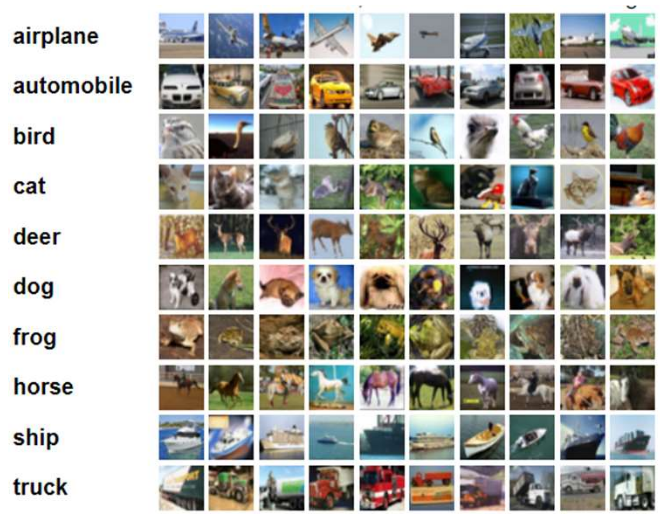
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Dataset used

- MNIST (Handwritten Digits)

- CIFAR-10 : consist of 10 classes

- Non – IID (Non – Independent & Identically Distributed) Distribution are created for both datasets, 100 devices and each device holds 2 Random Classes with 100 Samples



Experimentation/results

Attacks : Two different model inversion attacks

1) **DLG (Deep Leakage from Gradients)** : Aims to reconstruct devices' data using their uploaded gradients.It optimizes reconstructed data to minimize the Euclidean distance as a measure of similarity between raw gradients and reconstructed.

2) **GS (Gradient Similarity)** : Similar to DLG ,It utilized cosine similarity between raw gradients and dummy gradients to optimize the reconstructed data.

Baseline defenses :

1) **GC (Gradient Compression)** : reduces the communication cost by discarding gradients with magnitudes below a certain threshold.

2) **DP (Differential Privacy)** : injects noise into gradients uploaded th the server to achieve a theoretical privacy guarantee.

Experimentation/results

Utility and Privacy Trade-off :

MSE : Algorithm iterates over each pixel in both reconstruted image and raw image. Calculates pixel difference and averages the squared difference.

Lower MSE: implies the Reconstructed image is more similar to the Original Image,Hence High Risk of Privacy Leakage

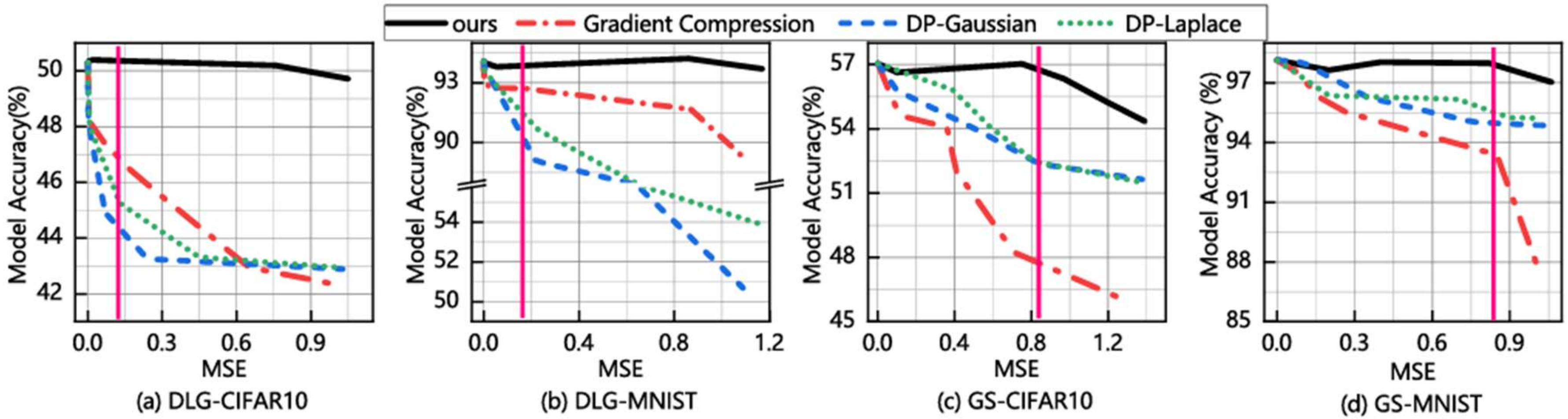


Figure : Compared defenses on model accuracy and MSE between reconstructed image and raw image for different attack baselines and datasets. The pink vertical line is the boundary that there constructed image is unrecognizable by human eyes if MSE is higher.

Experimentation/results

- Accuracy = (correctly classified data / total no of the data points) * 100
- A **higher accuracy** value signifies **good utility**.
- A **lower Accuracy** value signifies that defence mechanism might be **the impacting the model learning ability**.
- **Goal** is to find a defence mechanism that offers a good balance between privacy protection **(low MSE)** & Model Utility **(high Accuracy)**

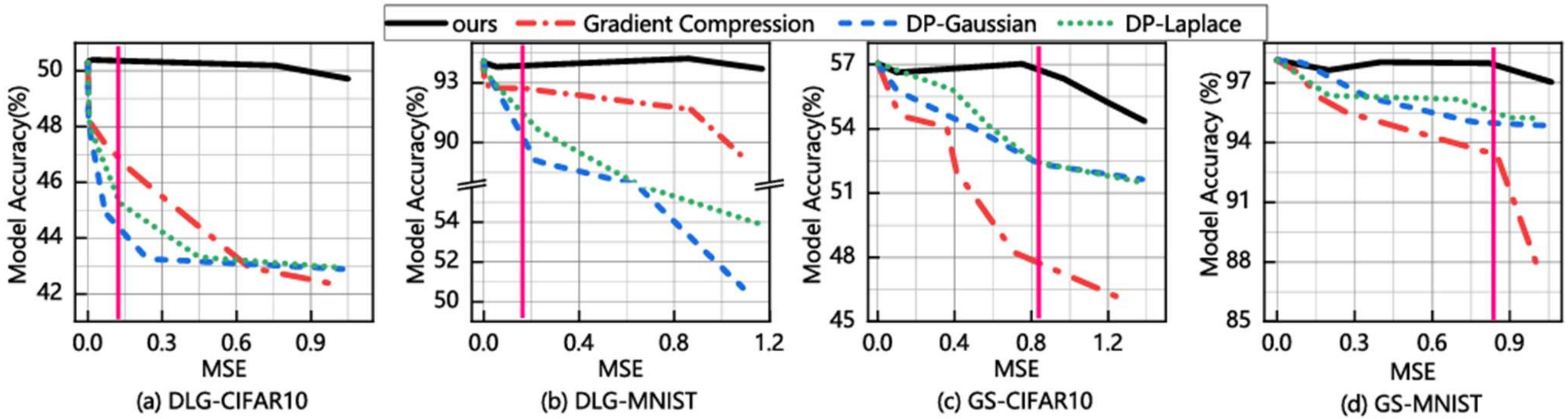


Figure : Compared defenses on model accuracy and MSE between reconstructed image and raw image for different attack baselines and datasets. The pink vertical line is the boundary that there constructed image is unrecognizable by human eyes if MSE is higher.

Summary

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- Data representation is the essential cause of privacy leakage.
- data representations are embedded in gradients.
- Inferred class-wise data representations
- Perturb the data representations with two goals : reducing the privacy leakage, maintain FL performance

Conclusion

- Results demonstate our defence is 160 times better than baseline defence.
- Defence learned to perturb data representation in such a way that quality of the reconstructed data is severely degraded,while maintaining the performance.
- Derived the robustness guarantee

Future Work

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- Reproduce the Results mentioned in the paper.
- Investigate if they still contain some residual information about the original data.
- Investigate the impact of various p-norm and q-norm
- Try to tweek other configurations and extend analysis of data representation leakage to have more comprehensive understanding of privacy in FL.

Privacy Assessment on Reconstructed Images: Are Existing Evaluation Metrics Faithful to Human Perception?

Xiaoxiao Sun[†] Australian National University xiaoxiao.sun@anu.edu.au
Nidham Gazagnadou[‡] Sony AI nidham.gazagnadou@sony.com
Vivek Sharma[‡] Sony AI viveksharma@sony.com
Lingjuan Lyu[‡] Sony AI lingjuan.lv@sony.com
Hongdong Li[†] Australian National University hongdong.li@anu.edu.au
Liang Zheng[†] Australian National University liang.zheng@anu.edu.au

Auditing Privacy Defenses in Federated Learning via Generative Gradient Leakage

Zhuohang Li[1] Jiaxin Zhang[2] Luyang Liu[3] Jian Liu[1]
[1]University of Tennessee, Knoxville [2]Oak Ridge National Laboratory [3]Google Research
zli96@vols.utk.edu, zhangj@ornl.gov, luyangliu@google.com, jliu@utk.edu

References

- Milad Nasr, Reza Shokri, and Amir Houmansadr. "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning." In 2019 *IEEE Symposium on Security and Privacy (SP)*, 2019. 2
- Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. "Beyond inferring class representatives: User-level privacy leakage from federated learning." In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications. IEEE*, 2019. 1, 2, 3
- Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. "On the convergence of fedavg on non-iid data" In *International Conference on Learning Representations*, 2019. 6, 8, 11, 12
- Ligeng Zhu, Zhijian Liu, and Song Han. "Deep leakage from gradients" *In Advances in Neural Information Processing Systems*, 2019. 1, 2, 3, 4, 7
- Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." In Proceedings of the 22nd *ACM SIGSAC Conference on Computer and Communications Security*, 2015. 1, 2

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Any Questions?

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Thank you