Indian Institute of Technology Hyderabad

Multimedia Content Analysis (CS6880)

SOTERIA | Provable Defense against Privacy Leakage in Federated Learning from Representation Perspective

Authors: Jingwei Sun[1], Ang Li[2], Binghui Wang[3], Huanrui Yang[4], Hai Li[5],Yiran Chen[6]

Paper Venue: CVPR 2021

Name	Roll Number
Yash Shukla	CS23MTECH14018

Guided By:
Prof. C Krishna Mohan (IIT-Hyderabad)

Teaching Assistant:
K Naveen Kumar (PhD Research Scholar IIT-Hyderabad)

Contents

1 Motivation **i**

2 Problem Statement. **i**

3 Challenges Faced **i**
3.1 Existing Methods i
4.2 Proposed Methodology Overview ii

5 Reproduced results **iv**
5.1 Dataset v
5.2 Original Results vi
5.3 Original Results vii
5.4 Reproduced Results vii

6 Limitations (Research Gap) in the author's proposal. **viii**

7 Novelty and its implementation **viii**
7.1 Novel Idea viii
7.2 Implementation Flow ix

8 Future work direction **x**

9 Conclusion **x**

10 References **xi**

1 | Motivation

- Federated Learning (FL) is a widely-used distributed learning framework that addresses privacy concerns by not directly sharing private data. However, recent studies have revealed that sharing model updates in FL can expose it to inference attacks. In this study, we highlight our key finding that leakage of data representation from gradients is the main source of privacy breaches in FL. We provide an analysis to elucidate how this leakage occurs. Based on this insight, we propose a defense mechanism called Soteria to counter model inversion attacks in FL. Our defense strategy involves learning to perturb data representation to significantly degrade the quality of reconstructed data, while maintaining FL performance. Additionally, we establish certified robustness guarantees for FL and convergence guarantees for FedAvg after implementing our defense. To assess the effectiveness of our defense, we conduct experiments on MNIST and CIFAR10 datasets to defend against the DLG and GS attacks. Our results show that, without sacrificing accuracy, our defense approach can increase the mean squared error between reconstructed and raw data by up to 160 times for both attacks, compared to baseline defense methods.

2 | Problem Statement.

- Representation leakage can compromise the privacy of individuals whose data is used for training in Federated learning

  Soteria aims to perturb the data representations in a specific layer (e.g , a fully connected layer) to achieve two goals.

  1) Reduce privacy leakage : The perturbed representations should make it difficult to reconstruct the original input data.

  2) Maintain model performance : The perturbed representations should remain similar to the original representations to avoid impacting the model accuracy

- 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.

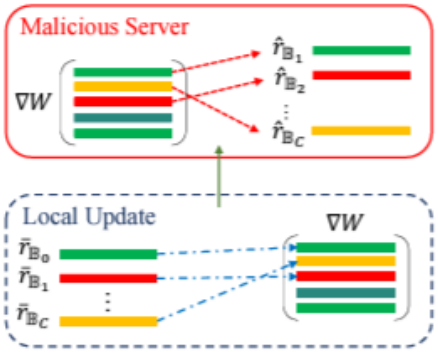


Figure 2.1: Inferred Representations

3 | Challenges Faced

3.1 | Existing Methods

- 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.

4.2 | Proposed Methodology Overview

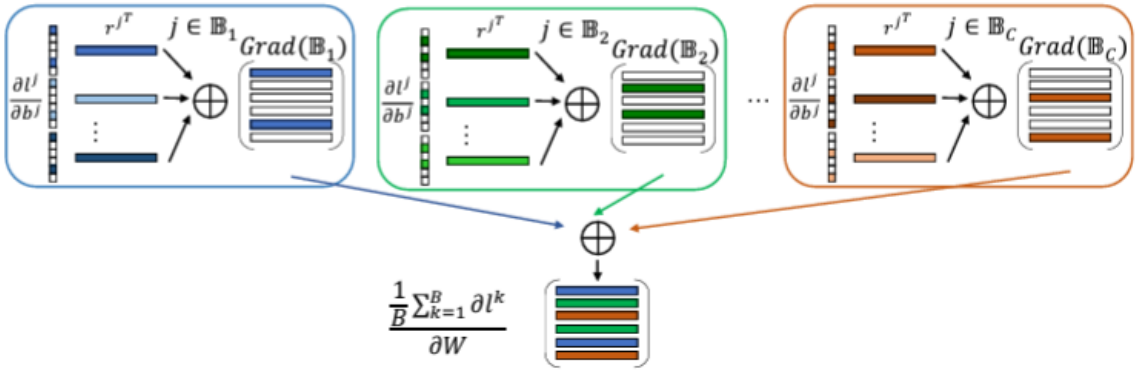


Figure 4.2: Data representations tend to be embedded in different rows of gradient(intuition behind the equation)

1. **Framework Overview**
   Soteria aims to perturb the data representations in a specific layer with goals
2. **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
3. **Formulation 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
4. **Mathematical Formulations**
   - **Mathematical formulation of Goals**

     Achieving Goal 1: maximize the p-norm of the difference between X and X':

     $$\max_{r'} \|X - X'\|_p \tag{4.1}$$

     Achieving Goal 2: subject to the constraint on the q-norm of the difference between r and r' being less than or equal to ϵ:

     $$\|r - r'\|_q \leq \epsilon \tag{4.2}$$

   - **Objective Function**

$$\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\frac{\partial l^i}{\partial W} = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\frac{\partial l^i}{\partial b}\frac{\partial b}{\partial W} = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\frac{\partial l^i}{\partial b^i}(r^i)^T \tag{4.3}$$

- **Optimization Problem**

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

  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 used 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)

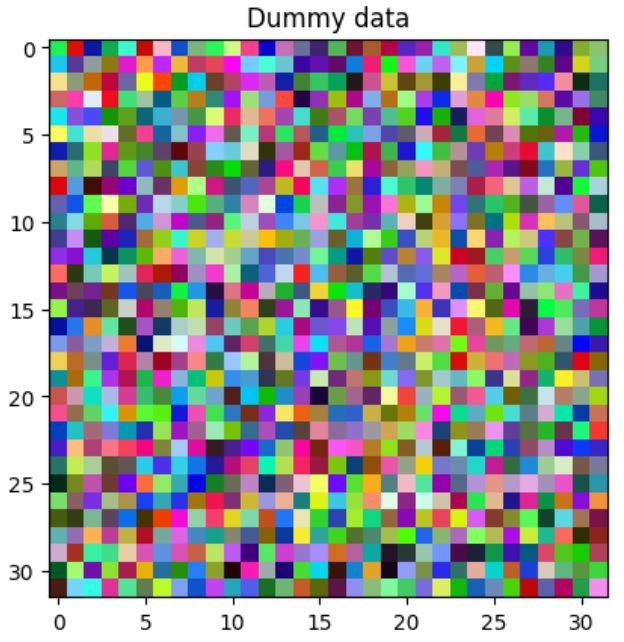
  Q=0 : choice simplifies the solution improves communication efficiency

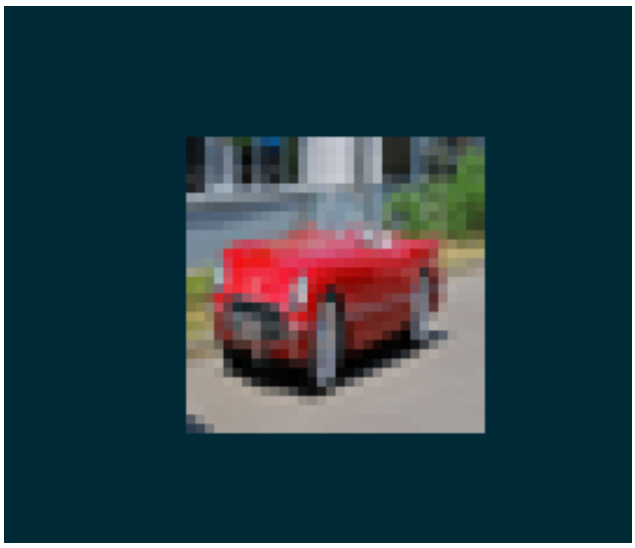
- **Evaluation and Metrics**

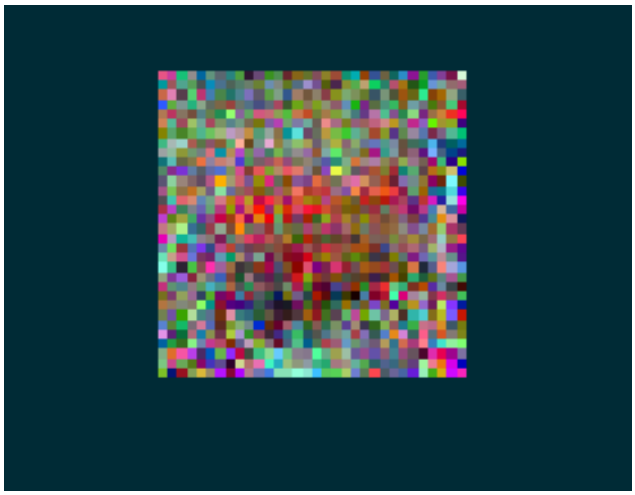
  Evaluation metric : Correlation co-efficient (cor) between true data representation and inferred representation for each class on each participating device.

  Cor is much lower compared to non-IID settings. Because having more diverse data on each device makes representation harder to isolate.

5 | Reproduced results

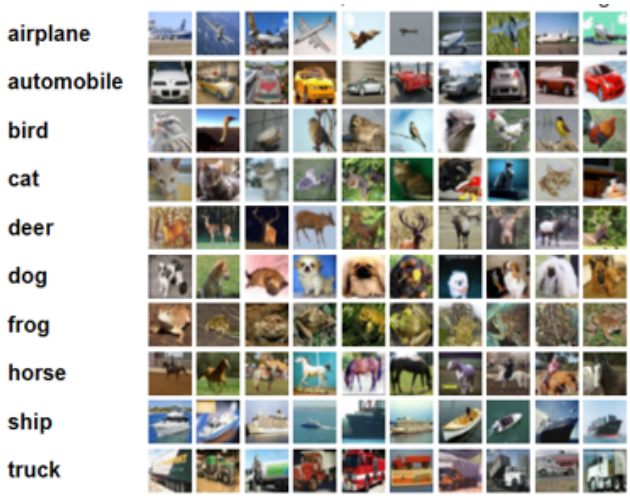




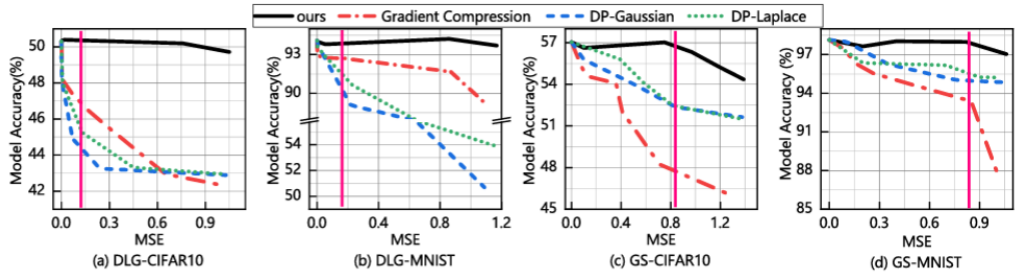


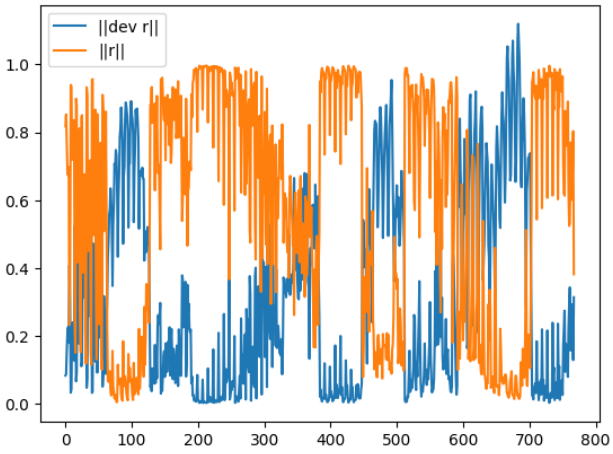

iter=0
iter=10
iter=20
iter=30
iter=40
iter=50
iter=60
iter=70
iter=80
iter=90
iter=100
iter=110
iter=120
iter=130
iter=140
iter=150
iter=160
iter=170
iter=180
iter=190
iter=200
iter=210
iter=220
iter=230
iter=240
iter=250
iter=260
iter=270
iter=280
iter=290

5.1 | Dataset

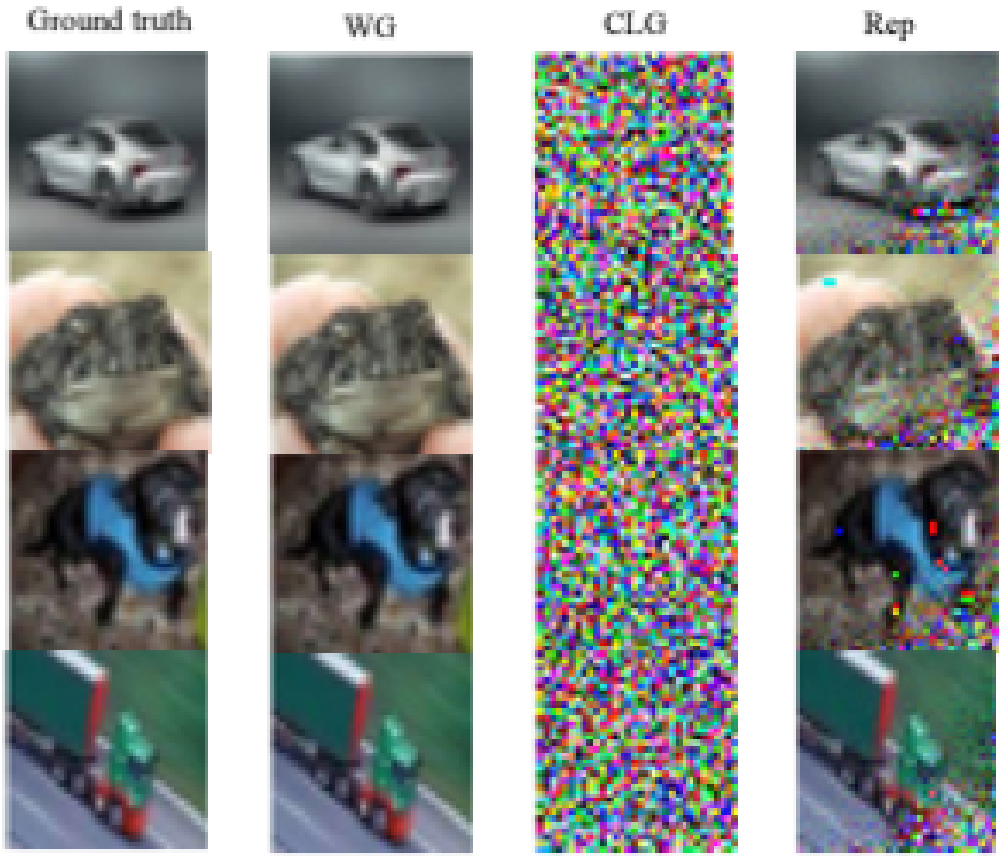

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

5.2 | Original Results

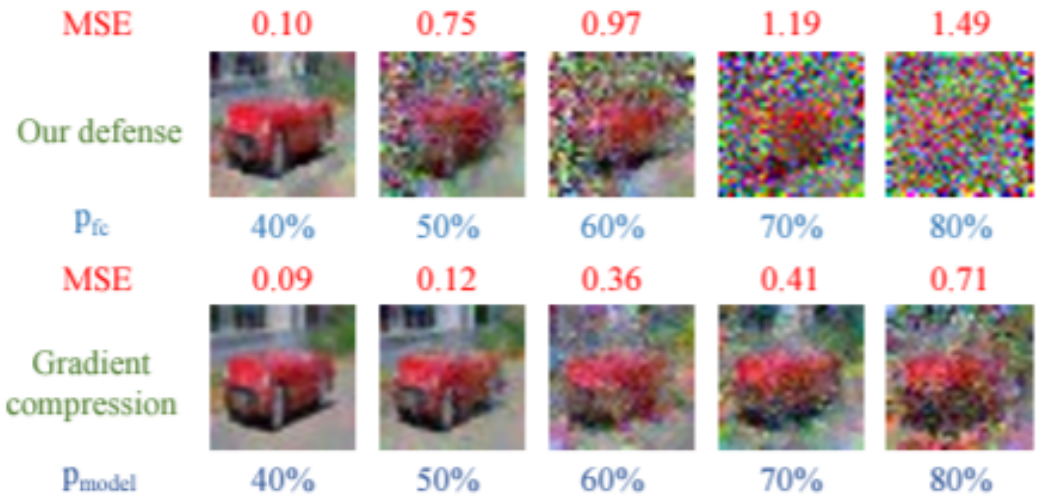




5.3 | Original Results



5.4 | Reproduced Results



6 | Limitations (Research Gap) in the author's proposal.

Below is an elaborated version of the limitations:-

- **Dataset Limitations:** The dataset used in the paper may not sufficiently cover highly privacy-sensitive scenarios, such as those involving extremely personal or sensitive information. For example, the dataset might not include data related to medical records, financial transactions, or other highly sensitive information that could have significant privacy implications if leaked or misused. This limitation could affect the generalizability and applicability of the findings to real-world scenarios where privacy is a major concern.
- **Exploration of Datasets with More Sensitive Information:** To address the limitations of the current dataset, you are considering exploring datasets that contain more sensitive information, such as facial or fingerprint data. These types of datasets are more likely to contain information that individuals consider highly personal and would want to keep private. By using these datasets, you can assess the privacy leakage risks in scenarios that are more relevant and realistic, potentially leading to more actionable insights and recommendations for improving privacy protection measures.
- **Risk Assessment:** Risk assessment involves evaluating the level of risk to privacy at every stage during an attack. This process typically includes identifying potential threats, assessing the likelihood of those threats occurring, and estimating the potential impact if they were to occur. By conducting a thorough risk assessment, you can identify potential vulnerabilities in your system and develop strategies to mitigate them. This can help improve the overall security and privacy of your system and reduce the likelihood of privacy breaches.
- **Absence of Adaptive Response:** The proposed defense mechanism lacks an adaptive response, which means it cannot adjust its behavior based on the current threat landscape. In scenarios where attackers are constantly evolving their strategies, this can limit the effectiveness of the defense mechanism. An adaptive response mechanism can proactively identify and respond to new and emerging threats, making it more resilient to evolving attack vectors. By addressing this limitation, you can improve the overall effectiveness of the defense mechanism and enhance the security and privacy of your system.

7 | Novelty and its implementation

7.1 | Novel Idea

The following novel ideas were implemented:

- **Impact of Image Reconstruction on Classification Accuracy:** You are comparing the performance of a model trained on ground truth images with that of a model trained on reconstructed images. By analyzing the classification accuracy, you aim to understand if the quality of the reconstructed images is sufficient for an attacker to exploit them maliciously. If the model performs similarly on both sets of images, it suggests that attackers could potentially use reconstructed images to fool the model.
- **Analysis of Gradients During Training:** You are studying how the representations within the neural network change over time during training. By analyzing the gradients, you can gain insights into how information flows through the network. This analysis can help you understand if there are any vulnerabilities in the network's architecture that could be exploited by attackers using reconstructed images.

- **Super-Resolution Techniques:** You are experimenting with super-resolution techniques to enhance the clarity of reconstructed images. By improving the quality of these images, you aim to make it harder for attackers to exploit them. Based on the results of these techniques, you plan to modify the defense mechanism of your system to better protect against potential attacks using reconstructed images.

7.2 | Implementation Flow

- Algorithm 1 : Learning perturbed representations

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)
  for $i = 0$ **to** $L - 1$ **do**
    Compute $\|\nabla r_i(f(r_i))^{-1}\|_2$
  end for
  Find the set S which contains the indices of ϵ largest elements in $\{\|\nabla r_i(f(r_i))^{-1}\|_2\}_i$
  $r' \leftarrow r$
  Set $r'_i = 0$ for $i \in S$
  return r'
end function=0

- Algorithm 2 : Local Training process with our defence on local device.

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

Input: Training data $X \in \mathbb{R}^{M \times N}$, Local objective function $F : \mathbb{R}^{M \times N} \to \mathbb{R}^L$, Feature extractor $f : W_f \in \mathbb{R}^{M \times N} \to \mathbb{R}^L$ before the defended layer, The defended layer $g : W_g \in \mathbb{R}^L \to \mathbb{R}^K$, Feature extractor after the defended layer $h : W_h \in \mathbb{R}^K \to \mathbb{R}$, Local model parameters $W = \{W_f, W_g, W_h\}$, Learning rate η.
Output: Learnt model parameter W with our defense.
Initialize W;
for each batch B in local training batches **do**
  for each $X \in B$ **do**
    $r \leftarrow f(X; W_f)$;
    $b \leftarrow g(r; W_g)$; {e.g., $b = W_g r$ for FC layers}
    $l \leftarrow h(b; W_h)$;
    Compute $\{\nabla W_f, \nabla W_g, \nabla W_h\} \leftarrow \nabla W F(X; W)$;
    $r' \leftarrow$ Perturb_Rep(X, f, r, ϵ);
    Update W'_g as $\tau(l, b, r', W_g)$; {e.g., $W'_g = \frac{\partial l}{\partial b} r'^T$ in FC}
    $W \leftarrow \{W_f, W'_g, W_h\}$;
    Update $W \leftarrow W - \eta \nabla W$;
  end for
end for=0

Experimental Results

Note on Experimental Results: Utility and Privacy Trade-off :

Algorithm iterates over each pixel in both reconstruted image and raw image. Calculates pixel difference and averages the squared difference and Lower MSE implies the Reconstructed image is more similar to the Original Image,Hence High Risk of Privacy Leakage

- 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)

8 | Future work direction

- 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.

9 | Conclusion

- Results demonstrate 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
- 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

10 | References

[1] ilad 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

[2] hibo 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

[3] iang 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

[4] igeng Zhu, Zhijian Liu, and Song Han. "Deep leakage from gradients" In Advances in Neural Information Processing Systems, 2019. 1, 2, 3, 4, 7

[5] att 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