Medical Data Leakage via Gradient Inversion

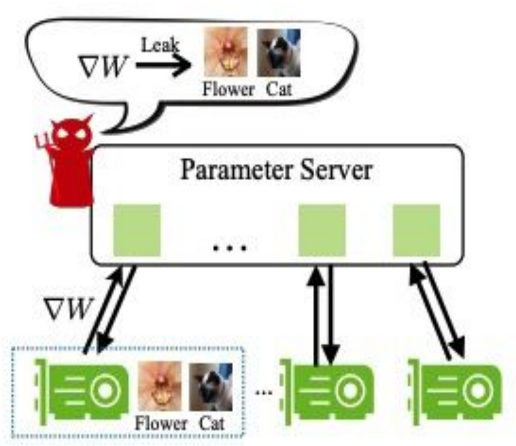
NYU Center for Data Science

Group Members: Christine Gao, Ciel Wang, Yuqi Zhang

Mentors: Professor Qi Lei, Professor Jacopo Cirrone

Background

➢ **Federated learning (FL)** :a distributed framework allows machine learning models to be trained on data via a gradient. Clients perform **local optimization** before sending the models back to the central model for updating; **data is never shared** and remains secure.



➢ With the addition of a **generative model** pre-trained on the underlying data distribution, **privacy can easily be breached.** We aim to show that the same technique of gradient inversion performed on **medical image data** can also be applied to **tabular data.**

Introduction

Dataset: This project utilizes multiple datasets and evaluates the performance of gradient inversion on medical image data and tabular medical record data.

➢ **MIMIC-III** - large, freely-available database comprising de-identified health records associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.

➢ **MedMNIST v2** - large-scale MNIST-like database of standardized biomedical images. All images are preprocessed into 28x28 (2D) and 28x28x28 (3D) and contain classification labels.

Objectives: We sought to explore two main goals:

1. Replication of gradient inversion on 2D MedMNIST medical image data
2. Application and improvement on gradient inversion on medical tabular data

Data Overview and Preprocessing

➢ **MIMIC-III** - The data used for the gradient inversion was preprocessed for in-hospital mortality, following the instructions for building benchmark tasks provided by Harutyunyan et al. [2]. The data was transformed into time-series data, with dimension of 48 timestamps and 76 features.

➢ **MedMNIST** - The data used for this analysis was BreastMNIST, which is based on a dataset of 780 breast ultrasound images. There are two classes: positive (normal and benign) and negative (malignant).

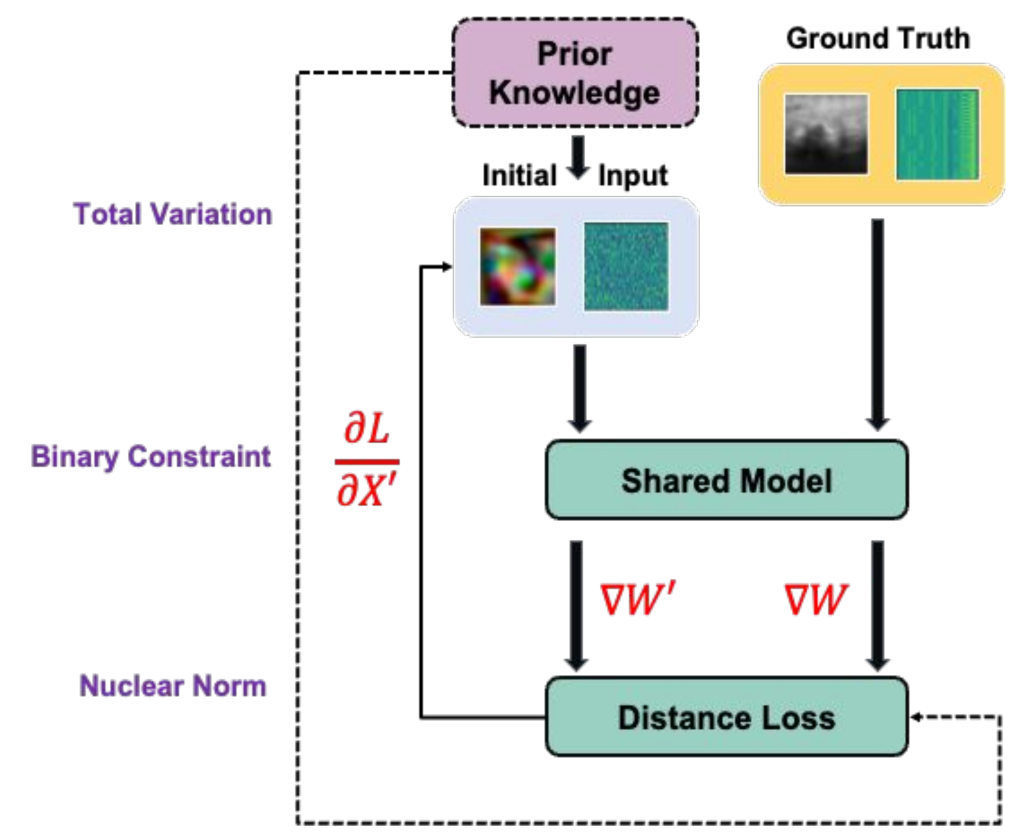
Related Work

The image reconstruction and tabular data reconstruction methods are based on a **gradient inversion implementation** by Jeon et al. [3]. The former **uses a generative image prior**, which is learnable via interactions in FL.

Approach

Gradient Inversion:

➢ Using **prior knowledge** (pretrained models, known Gaussian Distribution) to initialize inputs
➢ Obtain **gradients** from initialized and original input
➢ Minimize the **loss** between two inputs
➢ Update our generated inputs



4 STEPS of Gradient Inversion

For the **image reconstruction**, we implemented a StyleGAN2 image generator trained on MedMNIST with varying batch sizes. For **tabular data reconstruction**, the input was generated by Gaussian Distribution, with mean and variance derived from training data.

During the gradient inversion process, we improved the similarity between predicted inputs and ground truth by implementing

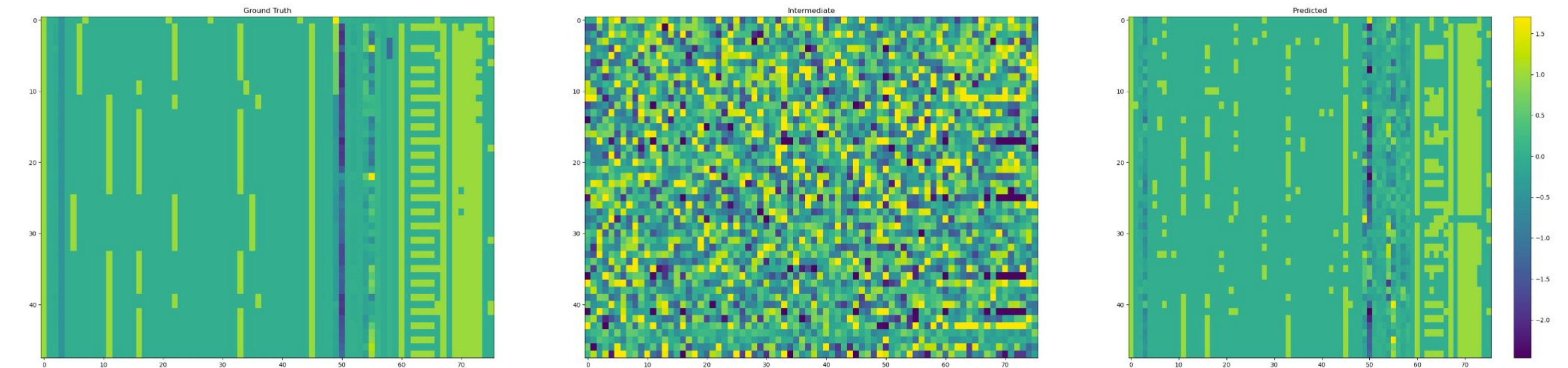
- Total variation(neighbors pixels should be similar) [5]
- Binary Constraint(for known binary columns)
- Nuclear Norm(increase sparsity)

$$L(\hat{x}, x) = Sim(\hat{x}, x) + \alpha_{TV} R_{TV}(\hat{x}) + \alpha_{BC} R_{BC}(\hat{x}) + \alpha_{NN} R_{NN}(\hat{x})$$

Bayesian optimization was applied to find best alphas.

Results & Discussion

For the **tabular data reconstruction,** we utilized the naive version of gradient inversion introduced by Jeon et al.[3] as a **baseline**, and applied our method incorporating prior knowledge to optimize the data reconstruction.



Ground Truth — Baseline Result — Best Result

With hyperparameters selected by Bayesian Optimization, after 5000 epochs with learning rate 0.01 and cosine annealing learning rate,

Model	Decay	Total Variation	Binary Constraint	Nuclear Norm	FMSE	Similarity
Naive(randn)	linear	.1	0	0	6.9194e-01	.89302
Trial1(gaussian)	cosine	.00231	.31178	.00051	8.7416e-06	.99689
Trial2(gaussian)	cosine	.00025	0	.00001	2.0753e-07	.99915
Best(gaussian)	cosine	.00014	.000000	.00001	6.3996e-06	.99916

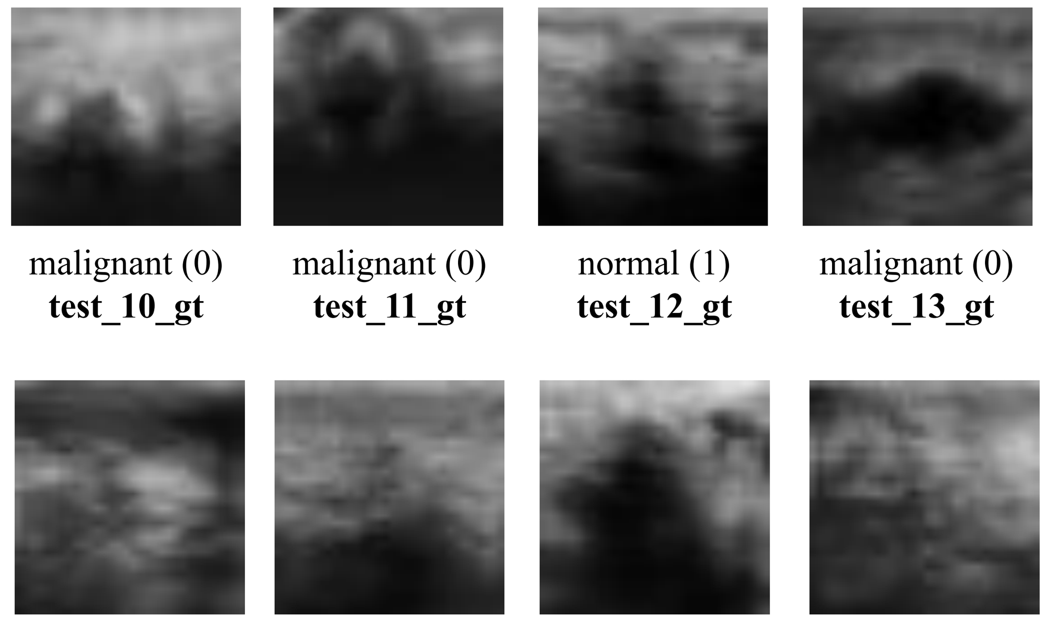
➢ Our models outperforms by 10% comparing to baseline.
➢ **Prior Knowledge works:**
  - Initialized inputs under estimated distribution.
  - Nuclear norm and total variation are effective in smaller scale.
  - Binary constraint is ineffective due to a threshold requirement for the conversion of logits into binary variables.

From the **image reconstruction** results, **batch size of 4** had the **best recorded PSNR** (ratio of maximum value of pixel to noise) indicating lower error.

Table 2: Performance Results for Reconstructed Images

Batch Size	PSNR	MSE	sim_cost $(1 - loss)$
1	14.282	0.0389	1.0
2	15.313	0.0324	0.965
4	**15.235**	**0.0311**	0.965
8	14.433	0.0404	1.0
32	14.400	0.0405	1.0

Ground Truth vs Reconstructed images with Pretrained StyleGAN with batch size of 4



Future Works

➢ **Larger Datasets:** Applying gradient inversion to larger and more complex datasets, going beyond black-and-white medical images
➢ **Attack on defensive methods:** Applying gradient inversion against perturbation on gradients

References

[1] Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). *Inverting gradients-how easy is it to break privacy in federated learning? arXiv: 2003.14053 [cs.CV]*

[2] Harutyunyan, H. Khachatrian, H., Kale, D., Steeg, G., and Galstyan, A. (2019). *Multitask learning and benchmarking with clinical time series data. Scientific data*, 6(1):96.

[3] Jeon, J., Kim, J., Lee, K., Oh, S., & Ok, J. (2021). *Gradient Inversion with Generative Image Prior. arXiv: 2110.14962.*

[4] Purushotham, S., Meng, C., Che, Z., & Liu, Y. (2018). *Benchmarking deep learning models on large healthcare datasets.* arXiv:1710.08531 [cs.LG]

[5] Wang, Z., Jason D. Lee, Qi Lei. (2022). *Reconstructing Training Data From Model Gradient, Provably. arXiv: 2212.03714.*

[6] Zhang, R., Guo, S., Wang, J., Xie, X., & Tao, D.(2022). *A survey on gradient inversion: Attacks, defenses and future directions.* arXiv preprint arXiv:2206.07284.