Evaluating Gradient Inversion Defenses in Federated Learning on the Brain Tumor MRI Dataset

Oleg Golev
Department of Computer Science
Princeton University
Princeton, New Jersey, USA
ogolev@princeton.edu

Abstract

Gradient inversion attacks are a big security concern and barrier for adoption of Federated Learning in applications that handle sensitive or confidential data. Huang et al. [11] evaluated this attack against a multitude of defenses with the MNIST, CIFAR-10, and ImageNet datasets, although the attack and defenses' efficacy on medical image data is yet to be assessed. In this project, we first reproduce a subset of the original paper's results by training and evaluating the data utility and security of a ResNet-18 model trained on CIFAR-10 with different defense combinations of GradPrune, MixUp, and Intra-InstaHide. Then we extend the original paper by running similar experiments on a Brain Tumor MRI dataset [3]. We confirm the authors' findings that combinations of stronger gradient pruning and Intra-InstaHide methods prove to be most secure against gradient inversion attacks, although in exchange for a noticeable drop in model performance. With the MRI dataset specifically, we find that the use of any strong defenses undercuts any data utility of the model to the point that its use is as bad or worse than random guessing. We also find that the MRI dataset's higher complexity makes the attack infeasible even in low-defense settings. More generally, we conclude that the success of the attack on any model depends heavily on batch size, randomness, and attacker's computational power, rendering the attack unreliable when applied to models with almost any defense applied, especially when targeting higher resolution data. All code and most results associated with this project can be found at: `https://github.com/oleggolev/GradAttack-Med`.

1 Introduction

Federated learning is a promising method for secure distributed training with medical data given its computational efficiency and high security promises. A lot of research was conducted to develop or evaluate federated learning methods and defenses for trustless environments to mitigate attacks such as model poisoning [2] [6], membership inference [15], reconstruction using gradient inversion [20], reconstruction with differential privacy [1], and many others [5] [9]. A lot of these and other potential issues with federated learning prevent its use in the medical domain [13], necessitating further work to ensure that federated learning systems deliver on their promises. The gradient inversion attack is one type of reconstruction attack where a malicious participant or a man-in-the-middle uses shared model's weight updates to reconstruct the original training data, making the attack a big threat in high data privacy applications.

To explore the gradient inversion attack and understand its potential threat to enabling federated learning applications in medical settings, we replicate a subset of results from Huang et al. [11] and set a baseline for the efficacy of the gradient inversion attack when performed on a ResNet-18

model trained with different defenses on the CIFAR-10 dataset. We then evaluate the same attack and defense methods on the brain tumor classification MRI dataset [3] which is much more detailed and complex than CIFAR-10. We find that the computational requirements, the effect of randomness, and the hyper-parameter tuning needed for running the image-based gradient inversion attack all contribute to the high overhead for the attacker of performing the attack effectively. For CIFAR-10, defense methods that combine gradient pruning and an encoding scheme prove extremely successful at preventing data leakage, as found in the original paper, but we further note that given MRI image complexity, even simple or no defenses could be sufficient for safeguarding input image data, with stronger defenses making the models useless.

2 Related Work

2.1 Gradient Inversion Attack

Originally shown as a viable pixel-level attack by Zhu et al. [20], the gradient inversion attack recovers input from model gradients in image classification applications. The attacker approximates the original input image $x \in \mathbb{R}^{b \times d}$ by computing x^* given the neural network with parameters θ, gradient $\nabla_\theta \mathcal{L}_\theta(x^*, y^*)$ where b is batch size of private data, d is the image size, and the recovery is regularized by a well-chosen function $R_{aux}(x)$ based on some prior:

$$\arg\min_x \mathcal{L}_{grad}(x; \theta, \nabla_\theta \mathcal{L}_\theta(x^*, y^*)) + \alpha R_{aux}(x) \tag{1}$$

The attack was found effective given careful choice of $\mathcal{L}_{grad}$ and $R_{aux}(x)$ on ImageNet images [7] [17] and therefore constitutes a powerful adversarial tool in breaking federated learning systems. The strongest attacks used by Geiping et al. [7] in their paper make two assumptions:

1. The attacker knows the batch normalization statistics of the target private batch.
2. The attacker knows the private labels associated with the target images.

The second assumption can be made safely since private label information about a single image can be inferred relatively accurately from the gradient. However, the first assumption is certainly unrealistic. In either case, the strongest state-of-the-art attack can be best approximated by making these two assumptions, so Huang et al. [11] and this project use them to construct the strongest attack for gradient inversion defense evaluation.

2.2 Defenses

For both the CIFAR-10 and the Brain Tumor MRI datasets, we evaluate the gradient inversion attack by training a ResNet-18 model with the following defenses:

1. **Gradient pruning** involves setting p proportion of the gradients of the smallest magnitudes to zero and is a common technique for making model training more efficient [16]. Gradient pruning is also useful in preventing reconstruction attacks and does so even more effectively than regular dropout [14].
2. **MixUp** is a data augmentation method used to increase generalization and stability of neural networks where the model is trained on linear combinations of k images instead of the images themselves [18]. To be more precise, MixUp is a simple technique which produces a synthetic training sample $\tilde{s_1}$ by combining an image s_1 with $k-1$ other images $s_2, s_3, ..., s_k$ according to some contribution factors $\{\lambda_1, ..., \lambda_k\}$:

$$\tilde{s_1} = \lambda_1 s_1 + \sum_{j=2}^{k} \lambda_j s_j \;\; where \;\; \sum_{j=1}^{k} \lambda_j = 1 \tag{2}$$

3. **InstaHide** is a more complex encoding scheme but has the same intuition as MixUp as to how it helps against adversarial attacks [12]. In practical terms, InstaHide functions more like a light-weight encryption method that makes it difficult for adversaries to recover input data by mixing up k images with one another according to the following construction:

$$\tilde{s_1} = \sigma \circ \left(\lambda_1 s_1 + \sum_{j=2}^{k} \lambda_j s_j \right) \tag{3}$$

That is, we first create a composite image from original image s_1 similar to MixUp and then apply a random sign-flipping pattern $\sigma \in \{-1, 1\}^d$ to that composite image via coordinate-wise vector multiplication. InstaHide is specifically designed as a method that avoids the high overhead of real cryptographic methods but makes up for MixUp's vulnerabilities. The Intra-InstaHide version of this technique uses images from the private dataset which is a weaker defense against gradient inversion attacks than Inter-InstaHide (which uses public images). For our experiments, we use Intra-InstaHide as was used by the original paper and to conceive the strongest possible attack.

All three defenses are evaluated individually and in combination, as in the original paper to evaluate the impact of the strongest attack on data utility and security.

2.3 Brain Tumor Classification MRI Dataset

Glioblastoma Multiforme (GBM) is one of the most common but also the most malignant and deadly variants of brain tumors [10]. Early detection through the use of artificial intelligence is a possible pathway for improving brain tumor diagnostics, and federated learning can help crowdsource more data for doing so more accurately in an automated way. Thus we evaluate the above-described defenses and the strongest gradient inversion attack to assess its performance and security preservation impact on ResNet-18 trained with the SARTAJ Brain Tumor Classification MRI dataset [3].

This dataset a good example of a higher resolution (512×512) medical image dataset and contains contains MRI images split into four fairly balanced classes: no tumor (396 training and 106 testing samples), pituitary tumor (828 training and 75 testing samples), meningioma tumor (823 training and 116 testing samples), and glioma tumor (827 training and 101 testing samples). Each class is also balanced in the number of MRI images that are taken in three different planes: sagittal, coronal, and axial. Samples from the dataset are provides in Figure 1.

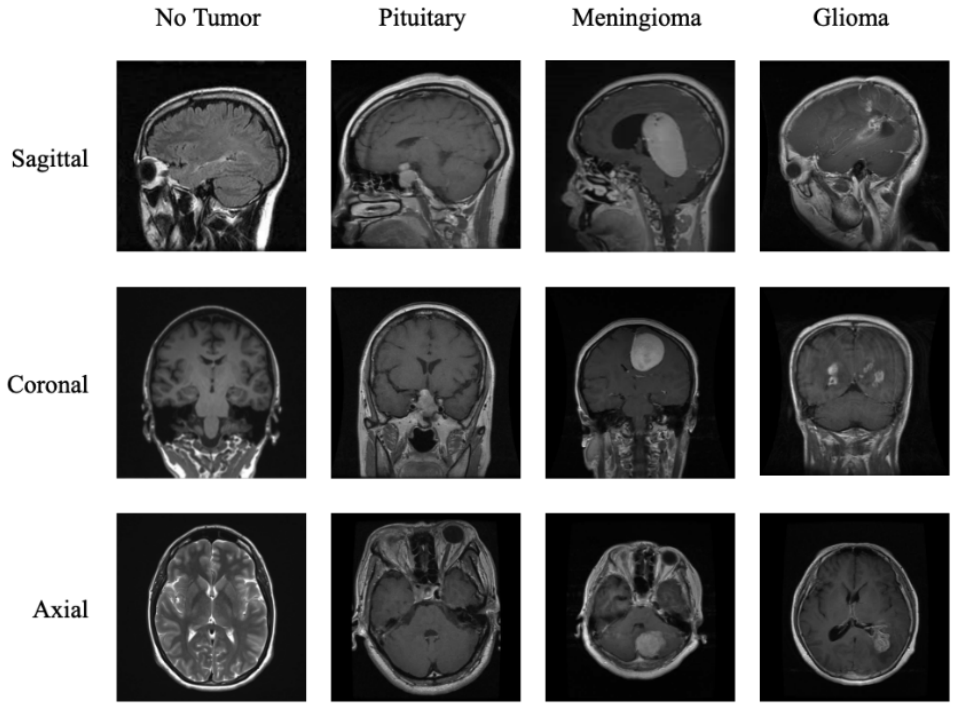


Figure 1: Sample MRI images contained in the Brain Tumor Classification MRI dataset [3].

3 Methods

3.1 Experimental Design

The target paper by Huang et al. [11] runs many experiments on multiple baselines with the MNIST, CIFAR-10, and ImageNet datasets. We reproduce a subset of the paper results on the CIFAR-10 dataset. For both the CIFAR-10 dataset and the Brain Tumor MRI dataset, we conduct these two types of experiments: (1) model training with defenses and performance evaluation, and (2) attack execution and data privacy evaluation. More simply, both datasets were used to train a ResNet-18 model with different defenses applied before evaluating the strong gradient inversion attack. Specifically, we consider the following defenses and their combinations:

- **GradPrune** (gradient pruning) was applied with varying the pruning ratios $p \in \{0.5, 0.7, 0.9, 0.95, 0.99, 0.999\}$ as in the paper.
- **MixUp** was applied by varying $k \in \{2, 4, 6, 8\}$ with the constraint on all $\lambda_i < 0.65$. The original paper only evaluated $k \in \{4, 6\}$. We expand on the choices of k as we would like to better explore the effect of k on the model performance to data leakage trade-off.
- **Intra-InstaHide** was applied similarly to MixUp by changing $k \in \{2, 4, 6, 8\}$ with $\lambda_i < 0.65$. The original paper also only evaluated $k \in \{4, 6\}$ for this defense.
- A combined defense of **GradPrune** with either **MixUp** or **Intra-InstaHide**. The original paper only evaluated $p \in \{0.9\}, k \in \{4\}$ while we evaluate every combination of parameters $p \in \{0.7, 0.9, 0.99\}$ and $k \in \{2, 4, 6\}$.

The goal is to understand the trade-off between the defenses' impact on model accuracy and data security when subjected to a strong gradient inversion attack, in which the attacker has knowledge of the labels and BatchNorm statistics of the private input batches. Furthermore, the extension to the Brain Tumor MRI dataset seeks to explore whether a state-of-the-art attack is dangerous or even at all feasible when dealing with high-resolution medical image data.

3.2 Running the Experiments

The original paper's code was substantially modified and extended to run the experiments since the provided code at `https://github.com/Princeton-SysML/GradAttack` was incomplete and victim to significant code rot over the past few years. All experiments were run on the Princeton University Adroit cluster using the NVIDIA A100 GPUs.

3.2.1 Training

ResNet-18 is trained for 200 epochs in the CIFAR-10 experiments and for 50 epochs in the Brain Tumor MRI dataset experiments due to time constraints. All models are trained with batch size $= 128$ using the SGD optmizer with momentum $= 0.9$, and 0.02 of the training data used for validation (recommended by original authors and experimentally confirmed in this project). The following parameter settings were used for different defense profiles (based on either the setup or the results of the original paper):

- **No Defenses** and **GradPrune**: initial learning rate $= 0.05$, learning rate decay $= 0.5$ every 30 epochs.
- **MixUp** and **GradPrune+MixUp**: initial learning rate $= 0.1$, learning rate decay $= 0.5$ every 30 epochs.
- **InstaHide** and **GradPrune+InstaHide**: initial learning rate $= 0.05$, learning rate decay $= 0.1$ every 50 epochs.

3.2.2 Attack

Due to the heavy dependency on randomness and lack of computational power to run the attacks to their fullest potential, all attacks used the same hyper-parameters that were set by default in the example script with only minor modifications as recommended by the paper. All attacks were run for 10,000 iterations using the Adam optimizer with the strongest setting (batch norm statistics and

private data labels are known) using batch size of 1, $\alpha_{TV} = 0.1$, and batch norm regularization term $\alpha_{BN} = 0.005$.

The batch size of 1 is chosen intentionally to simulate the strongest attack. Previous research [7] and the original paper [11] show that analyzing smaller batch sizes enables stronger attacks. On the other hand, using larger batch sizes like 32 makes the reconstructed images almost unrecognizable with most defenses. Thus the use of batch size $= 1$ gives us the lower bound on the amount of privacy preservation that the evaluated defenses can provide.

All attacks were run with the same seed $= 62$. Based on conversations with Samyak Gupta (one of the author's of the original paper), the gradient inversion attack was found to be extremely sensitive to the starting seed thus necessitating multiple runs to pick out the best results. This is one of the reasons why the original paper reran the attack multiple times and only presented the best results as measured by LPIPS scores.

After the gradient inversion attack, the decode step is applied to MixUp and InstaHide models [4] to attempt to recover the actual original image. For this decoding step, we assume that the attacker knows the k private images and their mixing coefficients. While this is unrealistic, it enables the strongest attack and thus it is done this way in the original paper and in this project.

3.3 Evaluation Metrics

As in the original paper, we evaluate our experiments by observing the following two items:

1. Test accuracy of each model trained with a different defense profile
2. Average, best, and standard deviation of the LPIPS (Learned Perceptual Image Patch Similarity) scores [19] which measure the similarity (or distance) between the original and reconstructed images. Higher score indicates more difference between the images and therefore better privacy preservation. The LPIPS statistics are computed based on a small (50 images) pre-selected subset of each dataset.

The provided example scripts also contain methods to compute PSNR (Peak signal-to-noise ratio) and RMSE (Root Mean Squared Error) statistics, but these are neither included in the original paper nor are included here. The original paper also reports training time. However, we did not record this data empirically and only give approximations in the next section.

4 Results and Discussion

4.1 CIFAR-10 Dataset

We first ran all experiments as defined in the previous section on the CIFAR-10 dataset. The training for each experiment took between 15 minutes and 3 hours 30 minutes depending on parameters, defenses applied, and the fluctuating load on the Adroit nodes. Adding gradient pruning during training did not significantly change the training runtime, with most GradPrune experiments taking between 15 and 30 minutes. In comparison, MixUp and InstaHide experiments took about 2 to 3 hours each, while the defense combination experiments are the ones that took up to 3 and a half hours to run. In comparison to the original paper's results, these runtimes are significantly longer. We attribute this to the fact that in the original paper, the experiments were run on an isolated 8 GPU cluster of RTX 2080 Ti's. In comparison, it is not uncommon for Adroit to be extremely slow even though it has access to powerful hardware.

The original paper's results (Table 1) from these experiments are as follows:

1. With no defenses, the attack recovers images closely for batch size $= 1$.
2. Higher pruning ratios for GradPrune-integrated defense packs result in higher security in exchange for proportionately worse performance.
3. While increasing k yields less data leakage, MixUp does not do well enough on its own to protect the input images from reconstruction and is similar in its data leakage properties to higher p GradPrune defenses.

4. Intra-InstaHide takes another small ($\sim 2\%$ accuracy hit compared to MixUp but in exchange for much better better privacy preservation properties.
5. Combining Intra-InstaHide with GradPrune allows for best security in exchange for $\sim 4\%$ drop in accuracy compared to only using one of the defenses.

	None	**GradPrune** (p)						**MixUp** (k)		**Intra-InstaHide** (k)		**GradPrune** ($p = 0.9$) **+ MixUp**	**GradPrune** ($p = 0.9$) **+ Intra-InstaHide**
Parameter	-	0.5	0.7	0.9	0.95	0.99	0.999	4	6	4	6	$k = 4$	$k = 4$
Test Acc.	93.37	93.19	93.01	90.57	89.92	88.61	83.58	92.31	90.41	90.04	88.20	91.37	86.10
Time (train)	$1\times$	$1.04\times$						$1.06\times$		$1.06\times$		$1.10\times$	
Attack batch size $= 1$													
Avg. LPIPS $\downarrow$	0.19	0.19	0.22	0.35	0.42	0.52	0.52	0.34	0.46	0.58	0.61	0.41	0.60
Best LPIPS $\downarrow$	0.02	0.02	0.05	0.14	0.22	0.32	0.36	0.12	0.25	0.41	0.42	0.21	0.43
(LPIPS std.)	0.16	0.17	0.16	0.13	0.11	0.08	0.06	0.08	0.07	0.06	0.09	0.07	0.09

Table 1: Accuracy-security trade-off of different defenses as reported in the original paper. The training accuracy is computed as the average over 5 independent runs, and the LPIPS scores are computed on a subset of 50 CIFAR-10 images (lower values suggest more privacy leakage). Only attack batch size of 1 is included since that is what was reproduced in this project.

Our results (Tables 2 and 3) are generally similar to those of the original paper, with no defenses and GradPrune with lower p having the least impact on both data utility (performance remains high) and data security (reconstruction is still doable). For model performance more generally, we find that our trained models perform very similarly to what was found by the original authors.

However, our LPIPS scores are much higher and more inconsistent for two reasons:

1. Our experiments had to be cut short due to the high computational cost of running them to completion given time constraints, sometimes producing only a few of the 50 images picked out of the CIFAR-10 dataset that are chosen to be reconstructed. For example, for models trained on a combination of GradPrune and Intra-InstaHide, reconstruction took long enough for no images to be reconstructed in a hour's time (see Table 3). Thus, our LPIPS statistics generally appear unstable and inconsistent.
2. Based on the conversation with one of the original paper authors, the successful reconstruction of the original input image depends heavily on the starting random seed. Choosing a good seed is the difference between getting a colorful mess and a very accurate reconstruction without changing any other parameters. Due to time constraints, we could not run the attack multiple times with different seeds to evaluate the attack's full potential.

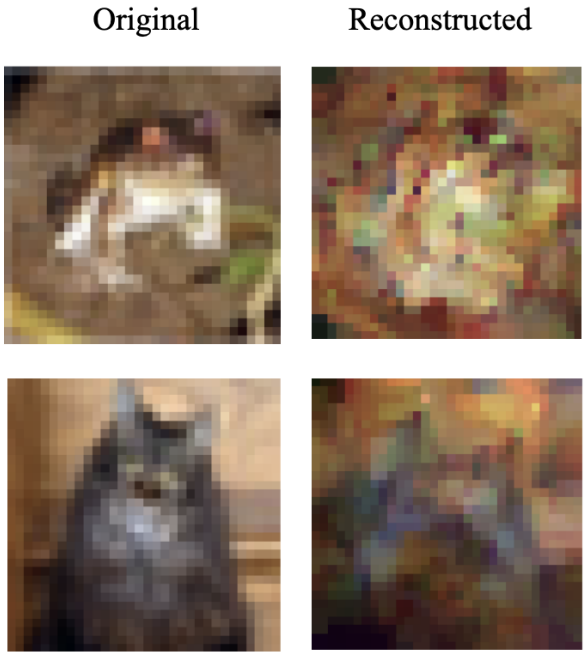
Even with the above obstacles, we were able to reconstruct, albeit not well (see Figure 2), the original images for models with no defenses and models trained with GradPrune with small p. These two observations also make for a good caveat about the attack in general: the attack is unreliable and requires high computational power to perform. In fact, the original paper authors spent around 3 weeks running experiments on 8 GPUs to produce all the results. Thus an attacker or a curious participant in a federated learning network would have to incur non-trivial costs when running the gradient inversion attacks.

	None	**GradPrune** (p)						**MixUp** (k)				**Intra-InstaHide** (k)			
Parameter	-	0.5	0.7	0.9	0.95	0.99	0.999	2	4	6	8	2	4	6	8
Test Acc.	92.81	92.82	93.04	91.62	90.97	89.00	81.97	93.27	93.70	92.69	91.10	91.49	90.83	89.92	88.78
Attack batch size $= 1$															
Avg. LPIPS $\downarrow$	0.53	0.51	0.53	0.54	0.59	0.57	0.68	0.59	0.59	0.61	0.52	0.36	0.35	0.37	0.35
Best LPIPS $\downarrow$	0.47	0.40	0.49	0.47	0.51	0.51	0.62	0.43	0.38	0.56	0.48	0.32	0.29	0.33	0.31
(LPIPS std.)	0.04	0.07	0.02	0.05	0.05	0.06	0.05	0.11	0.11	0.05	0.04	0.03	0.05	0.02	0.02

Table 2: Accuracy-security trade-off for individual defenses as run on the Adroit cluster. The training accuracy is computed in a single run, and the LPIPS scores are computed on however many CIFAR-10 images the attack could reconstruct in the span of one hour (this is done due to computational constraints and large runtime requirements of the attack).

	GradPrune ($p = 0.7$) + MixUp			+ Intra-InstaHide			GradPrune ($p = 0.9$) + MixUp			+ Intra-InstaHide			GradPrune ($p = 0.99$) + MixUp			+ Intra-InstaHide		
Parameter k	2	4	6	2	4	6	2	4	6	2	4	6	2	4	6	2	4	6
Test Acc.	92.6	92.7	92.1	90.2	89.9	89.8	90.2	90.0	89.6	87.4	87.3	86.4	10.0	82.0	10.0	74.5	74.4	73.1
Attack batch size $= 1$																		
Avg. LPIPS $\downarrow$	0.58	0.56	0.56	0.37	0.36	-.–	0.61	0.55	0.57	0.34	-.–	-.–	0.74	0.56	0.69	0.37	-.–	-.–
Best LPIPS $\downarrow$	0.42	0.44	0.53	0.30	0.29	-.–	0.45	0.42	0.47	0.27	-.–	-.–	0.64	0.43	0.63	0.33	-.–	-.–
(LPIPS std.)	0.09	0.07	0.03	0.05	0.06	-.–	0.09	0.07	0.06	0.04	-.–	-.–	0.06	0.07	0.040	0.03	-.–	-.–

Table 3: Accuracy-security trade-off for combined defenses as run on the Adroit cluster. The associated experiments were subject to the same limitations as the experiments in Table 2. Some entries are missing because the experiments timed out before a single image could be reconstructed.



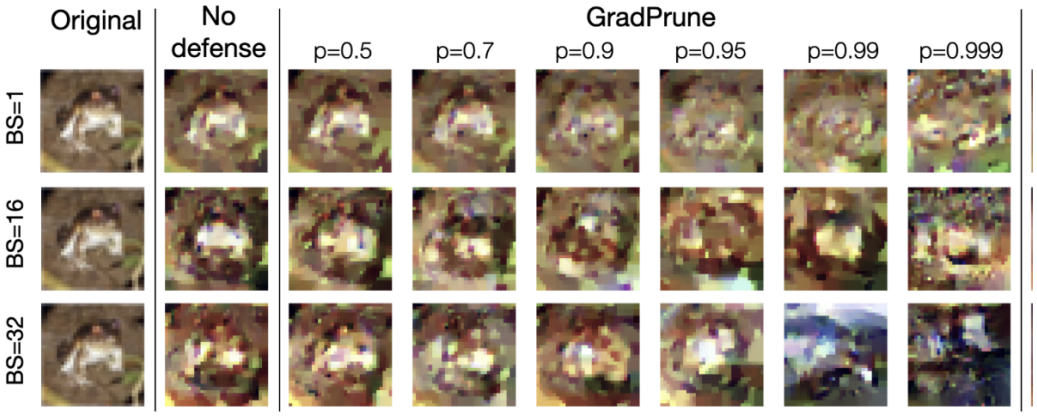


Figure 2: At the top are two sample image reconstructions achieved with 10,000 iteration GradPrune $p = 0.7$ attack. The bottom graphic is taken from the original paper, showcasing the full attack potential. Unfortunately, reaching this potential requires much longer runtimes than what was available for the execution of this project.

4.2 Brain Tumor MRI Dataset

For the MRI dataset, we ran the same set of experiments with the same hyperparameters that we used for the CIFAR-10 dataset. However, we only trained the MRI models for 50 epochs (as opposed to 200) due to time constraints. The results of the experiments on the MRI dataset are shown in Table 4. Given the higher complexity of the dataset and difficulty of the task, a simple quickly trained ResNet-18 model could only achieve 68.3% test accuracy with no defenses applied. Based on other people's work with this dataset [8], a ResNet-50 model could potentially achieve around 77% test accuracy which is still significantly lower than what we've seen with the CIFAR-10 dataset. Correspondingly, while applying the GradPrune defense reflects expected results (decreasing model performance proportional to increasing data security), using either of the encoding schema proves detrimental to model performance, making it essentially unusable.

	None	**GradPrune** (p)					**MixUp** (k)				**Intra-InstaHide** (k)			
Parameter	-	0.5	0.7	0.9	0.95	0.99	0.999	2	4	6	8	2	4	6	8
Test Acc.	68.3	62.81	63.71	55.43	59.1	62.92	56.87	30.70	28.75	35.39	20.70	20.50	20.70	20.50	22.46

	GradPrune ($p = 0.7$)						**GradPrune** ($p = 0.9$)						**GradPrune** ($p = 0.99$)					
	+ MixUp			**+ Intra-InstaHide**			**+ MixUp**			**+ Intra-InstaHide**			**+ MixUp**			**+ Intra-InstaHide**		
Parameter k	2	4	6	2	4	6	2	4	6	2	4	6	2	4	6	2	4	6
Test Acc.	34.8	29.3	33.4	20.5	21.7	15.8	31.9	31.7	34.6	46.7	23.2	21.3	25.6	31.9	22.7	17.4	20.7	22.7

Table 4: Test accuracy of models trained on the Brain Tumor MRI dataset. The results were produced with the same constraints as for the CIFAR-10 dataset.

The original code was also adopted to perform the gradient inversion attack on a small subset of the MRI images. However, given the trained models' poor performance even with no defenses applied, the even higher computational costs due to the dataset's higher resoltuion, and our previous unstable LPIPS findings with CIFAR-10 experiments, the attacks were mostly unsuccessful and the LPIPS scores were thus omitted from Table 4. Still, some reconstructions on this dataset look like they'd be useful to an attacker. We also note that the general patterns of the images are reconstructed relatively well (see Figure 3). Thus defense methods should definitely still be used when working with higher resolution medical image data.

5 Conclusion and Limitations

This project was a fascinating exploration of the state-of-the-art gradient inversion attack. We replicated a big portion of the findings in Huang et al. [11] and further adopted their code for evaluation on the Brain Tumor MRI dataset. Through this project, we found that InstaHide mixed with GradPrune is a very effective defense against gradient inversion attacks, at the cost of model performance. However, in the case of a higher complexity dataset like the MRI dataset, the encoding defenses were detrimental to model performance, making GradPrune the more reliable choice of defense against gradient-based reconstruction. Above all, we note that the gradient inversion attack is very computationally expensive, unreliable, and, to achieve its full strength, requires unrealistic knowledge of the private data. Still, an upper bound on the attack's strength needs further exploration, specifically on larger models used with more complex datasets, to truly understand whether the attack poses a threat to federated learning for high-resolution medical image applications.

Acknowledgments

I would like to thank Prof. Olga Troyanskaya for facilitating amazing discussions throughout the semester in this course. I decided to take COS 557 purely because I wanted to learn something new, and this semester was a great time of exploration of new ideas and applications of machine learning. More generally, I want to thank Prof. Troyanskaya, Tamjeed Azad, Prof. Kai Li, and Ksenia Sokolova for their advice and guidance throughout the semester and on this project. Finally, I want to extend special thanks to Samyak Gupta, one of the authors of the original paper [11], who gave invaluable advice on how to approach this project and its limitations. I learned a lot, and the class was a very wonderful experience thanks to all of these people.

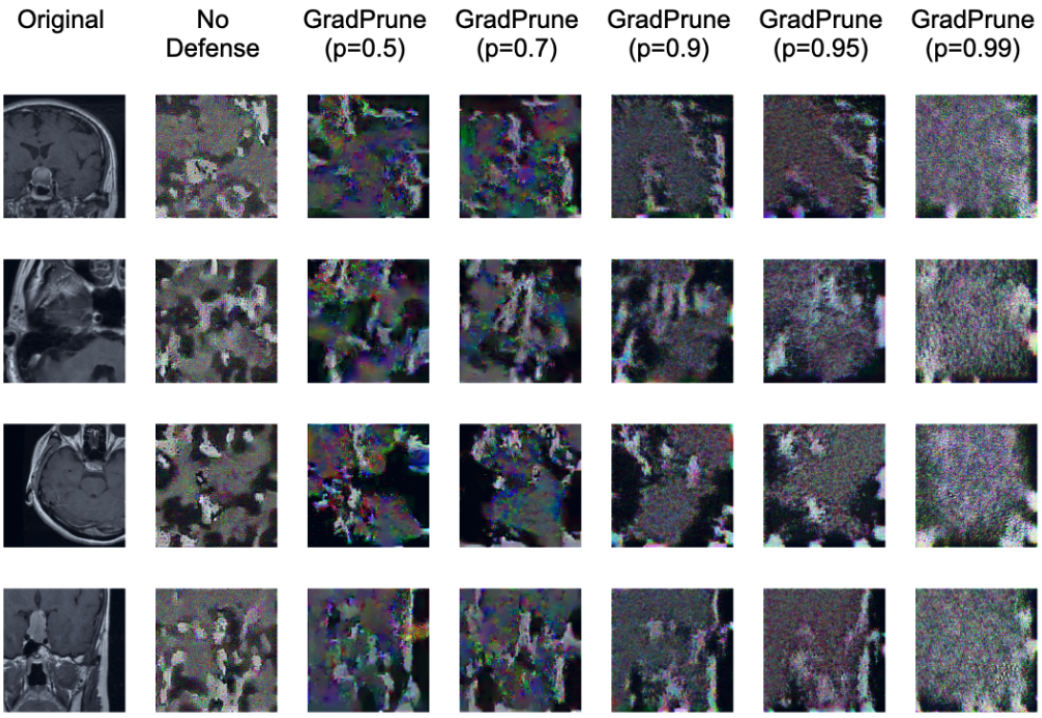


Figure 3: Sample reconstructions of some images in the Brain Tumor MRI dataset.

References

[1] Marco Arazzi, Mauro Conti, Antonino Nocera, and Stjepan Picek. 2023. Turning Privacy-preserving Mechanisms against Federated Learning. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* (, Copenhagen, Denmark,) *(CCS '23)*. Association for Computing Machinery, New York, NY, USA, 1482–1495. https://doi.org/10.1145/3576915.3623114

[2] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin B. Calo. 2018. Analyzing Federated Learning through an Adversarial Lens. *CoRR* abs/1811.12470 (2018). arXiv:1811.12470 http://arxiv.org/abs/1811.12470

[3] Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan. 2020. Brain Tumor Classification (MRI). https://doi.org/10.34740/KAGGLE/DSV/1183165

[4] Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Shuang Song, Abhradeep Thakurta, and Florian Tramer. 2021. Is Private Learning Possible with Instance Encoding? arXiv:2011.05315 [cs.CR]

[5] Yao Chen, Yijie Gui, Hong Lin, Wensheng Gan, and Yongdong Wu. 2022. Federated Learning Attacks and Defenses: A Survey. arXiv:2211.14952 [cs.CR]

[6] Matteo Demartis. 2022. *Adversarial Attacks in Federated Learning*. Master's thesis. KTH, School of Electrical Engineering and Computer Science (EECS).

[7] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. 2020. Inverting Gradients – How easy is it to break privacy in federated learning? arXiv:2003.14053 [cs.CV]

[8] Shreya Hallikeri. 2024. Brain Tumor classification. https://www.kaggle.com/code/shreyahallikeri/brain-tumor-classification

[9] Shanshan Han, Baturalp Buyukates, Zijian Hu, Han Jin, Weizhao Jin, Lichao Sun, Xiaoyang Wang, Wenxuan Wu, Chulin Xie, Yuhang Yao, Kai Zhang, Qifan Zhang, Yuhui Zhang, Carlee

Joe-Wong, Salman Avestimehr, and Chaoyang He. 2024. FedMLSecurity: A Benchmark for Attacks and Defenses in Federated Learning and Federated LLMs. arXiv:2306.04959 [cs.CR]

[10] Farina Hanif, Kanza Muzaffar, Kahkashan Perveen, Saima Malhi, and Shabana Simjee. 2017. Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. *Asian Pacific journal of cancer prevention : APJCP* 18 (01 2017), 3–9. https://doi.org/10.22034/APJCP.2017.18.1.3

[11] Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. 2021. Evaluating Gradient Inversion Attacks and Defenses in Federated Learning. arXiv:2112.00059 [cs.CR]

[12] Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. 2021. InstaHide: Instance-hiding Schemes for Private Distributed Learning. arXiv:2010.02772 [cs.CR]

[13] Madhura Joshi, Ankit Pal, and Malaikannan Sankarasubbu. 2022. Federated Learning for Healthcare Domain - Pipeline, Applications and Challenges. *ACM Trans. Comput. Healthcare* 3, 4, Article 40 (nov 2022), 36 pages. https://doi.org/10.1145/3533708

[14] Sheng Liu, Zihan Wang, and Qi Lei. 2024. Data Reconstruction Attacks and Defenses: A Systematic Evaluation. arXiv:2402.09478 [cs.CR]

[15] Reza Shokri, Marco Stronati, and Vitaly Shmatikov. 2016. Membership Inference Attacks against Machine Learning Models. *CoRR* abs/1610.05820 (2016). arXiv:1610.05820 http://arxiv.org/abs/1610.05820

[16] Xu Sun, Xuancheng Ren, Shuming Ma, and Houfeng Wang. 2019. meProp: Sparsified Back Propagation for Accelerated Deep Learning with Reduced Overfitting. arXiv:1706.06197 [cs.LG]

[17] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M. Alvarez, Jan Kautz, and Pavlo Molchanov. 2021. See through Gradients: Image Batch Recovery via GradInversion. arXiv:2104.07586 [cs.LG]

[18] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2017. mixup: Beyond Empirical Risk Minimization. *CoRR* abs/1710.09412 (2017). arXiv:1710.09412 http://arxiv.org/abs/1710.09412

[19] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. arXiv:1801.03924 [cs.CV]

[20] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep Leakage from Gradients. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/60a6c4002cc7b29142def8871531281a-Paper.pdf