
Gradient Inversion with Generative Image Prior

Jinwoo Jeon^{*1} Jaechang Kim^{*2} Kangwook Lee³ Sewoong Oh⁴ Jungseul Ok^{1,2}

Abstract

Federated Learning (FL) is a distributed learning framework, in which the local data never leaves clients' devices to preserve privacy, and the server trains models on the data via accessing only the gradients of those local data. Without further privacy mechanisms such as differential privacy, this leaves the system vulnerable against an attacker who inverts those gradients to reveal clients' sensitive data. However, a gradient is often insufficient to reconstruct the user data without any prior knowledge. By exploiting a generative model pre-trained on the data distribution, we demonstrate that data privacy can be easily breached. Further, when such prior knowledge is unavailable, we investigate the possibility of learning the prior from a sequence of gradients seen in the process of FL training. We experimentally show that the prior in a form of generative model is learnable from iterative interactions in FL. Our findings demonstrate that additional mechanisms are necessary to prevent privacy leakage in FL.

1. Introduction

Federated learning (FL) is an emerging framework for distributed learning, where central server aggregates model updates, rather than user data, from end users (Brisimi et al., 2018; McMahan et al., 2017). The main premise of federated learning is that this particular way of distributed learning can protect users' data privacy as there is no explicit data shared by the end users with the central server.

^{*}Equal contribution ¹Department of Computer Science and Engineering, Pohang University of Science and Technology, South Korea ²Graduate School of Artificial Intelligence, Pohang University of Science and Technology, South Korea ³Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, USA ⁴Paul G. Allen School of Computer Science & Engineering, University of Washington, USA. Correspondence to: Jungseul Ok <jungseul@postech.ac.kr>

This work was presented at the International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with ICML 2021 (FL-ICML'21). This workshop does not have official proceedings and this paper is non-archival. Copyright 2021 by the author(s).

However, a recent line of work (Zhu et al., 2019; Zhao et al., 2020; Geiping et al., 2020; Yin et al., 2021) demonstrates that one may recover the private user data used for training by observing the gradients. This process of recovering the training data from gradients, so-called *gradient inversion*, poses a huge threat to the federated learning community as it may imply the fundamental flaw of its main premise.

Even more worryingly, recent works suggest that such gradient inversion attacks can be made even stronger if certain side-information is available. For instance, Geiping et al. (2020) show that if the attacker knows a prior that user data consists of natural images, then the gradient inversion attack can leverage such prior, achieving a more accurate recovery of the user data. Another instance is when batch norm statistics are available at the attacker in addition to gradients. This can actually happen if the end users share their local batch norm statistics as in (McMahan et al., 2017). Yin et al. (2021) show that such batch normalization statistics can significantly improve the strength of the gradient inversion attack, enabling precise recovery of high-resolution images.

In this paper, we systematically study how one can maximally utilize and even obtain the prior information when inverting gradients. We first consider the case that the attacker has a generative model pretrained on the exact or approximate distribution of the user data as a prior. For this, we propose an efficient gradient inversion algorithm that utilizes the generative model prior. In particular, the algorithm consists of two steps, in which the first step searches the latent space (of lower dimension) defined by the generative model instead of the ambient input space (of higher dimension), and then the second step adapts the generative model to each input given the gradient. Each step provides substantial improvement in the reconstruction. We name the algorithm as gradient inversion in alternative spaces (GIAS).

We then consider a realistic scenario in which the user data distribution is not known in advance, and thus the attacker needs to learn it from gradients. For this scenario, we develop a meta-learning framework, called gradient inversion to meta-learn (GIML), which learns a generative model on user data from observing and inverting multiple gradients computed on the data, e.g. across different FL epochs or participating nodes. Our experimental results demonstrate that one can learn a generative model via GIML and reconstruct

data by making use of the learned generative model.

This implies a great threat on privacy leakage in FL since our methods can be applied for any data type in most FL scenarios unless a specialized architecture prevents the gradient leakage explicitly, e.g., (McMahan et al., 2018).

Our main contributions are as follows:

- We introduce GIAS that fully utilizes a pretrained generative model to invert gradient. In addition, we propose GIML which can train generative model from gradients only in FL.
- We demonstrate significant privacy leakage using GIAS with pretrained generative model in various challenging scenarios, where our method provides substantial gain additionally to other any other existing methods (Geiping et al., 2020; Yin et al., 2021).
- We experimentally show that GIML can learn a generative model on the user data from only gradients, which provides the same level of data recovery with a given pretrained model. To our best knowledge, GIML is the first capable of learning explicit prior on a set of gradient inversion tasks.

2. Related work

Early works (Melis et al., 2019; Shokri et al., 2017) investigate membership inference from gradients to check the possibility of privacy leakage in FL. Phong et al. (2018) demonstrate that it is possible to reconstruct detailed input image when FL trains a shallow network such as single-layer perceptron. Fan et al. (2020) and Zhu and Blaschko (2021) consider a wider class of learning model and propose analytical approach solving a sequence of linear systems to reveal output of each layer recursively. To study the limit of the gradient inversion in practical scenarios of training deep networks via FL, a sequence of effort has been made formulating optimization problem to minimize discrepancy comparing gradients from true data and reconstructed data (Wang et al., 2019; Zhu et al., 2019; Zhao et al., 2020; Geiping et al., 2020; Yin et al., 2021).

The optimization-based approaches are particularly useful as one can easily utilize prior knowledge by adding regularization terms, e.g., total variation (Wang et al., 2019) and BN statistics (Yin et al., 2021), or changing discrepancy measure (Geiping et al., 2020). In (Wang et al., 2019; Yin et al., 2021), privacy attack techniques using generative model are introduced. They however require a pretrained model or a set of user data examples to train one, while we propose a meta learning framework training generative model from gradients only. In addition, our method of inverting gradient maximally exploit a given generative model

by alternating search spaces, which are analogous to the state-of-the-art GAN inversion techniques (Zhu et al., 2020; Bau et al., 2019a;b).

3. Problem formulation

In this section, we formally describe the gradient inversion (GI) problem. Consider a standard supervised learning for classification, which optimizes neural network model f_θ parameterized by θ as follows:

$$\min_{\theta} \sum_{(x,y) \in \mathcal{D}} \ell(f_\theta(x), y), \quad (1)$$

where ℓ is a point-wise loss function and \mathcal{D} is a dataset of input $x \in \mathbb{R}^m$ and label $y \in \{0, 1\}^L$ (one-hot vector). In federated learning framework, each node reports the gradient of $\ell(f_\theta(x), y)$ for sampled data (x, y) 's instead of directly transferring the data. The problem of inverting gradient is to reconstruct the sampled data used to compute the reported gradient. Specifically, when a node computes the gradient g using a batch $\{(x_1^*, y_1^*), \dots, (x_B^*, y_B^*)\}$, i.e., $g = \frac{1}{B} \sum_{j=1}^B \nabla \ell(f_\theta(x_j^*), y_j^*)$, we consider the following problem of inverting gradient:

$$\min_{\substack{(x_1, y_1), \dots, (x_B, y_B) \\ \in \mathbb{R}^m \times \{0, 1\}^L}} d \left(\frac{1}{B} \sum_{j=1}^B \nabla \ell(f_\theta(x_j), y_j), g \right), \quad (2)$$

where $d(\cdot, \cdot)$ is a measure of the discrepancy between two gradient, e.g., ℓ_2 -distance (Zhu et al., 2019; Yin et al., 2021) or negative cosine similarity (Geiping et al., 2020). It is known that label y can be almost accurately recovered by simple methods just observing the gradient at the last layer (Zhao et al., 2020; Yin et al., 2021), while reconstructing input x remains still challenging as it is often under-determined even when the true label is given. For simplicity, we hence focus on the following minimization to reveal the inputs from the gradient given the true labels:

$$\min_{x_1, \dots, x_B \in \mathbb{R}^m} c(x_1, \dots, x_B; \theta, g), \quad (3)$$

where we denote by $c(x_1, \dots, x_B; \theta, g)$ the cost function in (2) given $y_j = y_j^*$ for each $j = 1, \dots, B$.

4. Methods

The key challenge of inverting gradient is that solving (2) is often under-determined, i.e., a gradient contains only insufficient information to recover data. Such an issue is observed even when the dimension of gradient is much larger than that of input data. Indeed, Zhu and Blaschko (2021) show that there exist a pair of different data having the same gradient, so called twin data, even when the learning model is

large. To alleviate this issue, a set of prior knowledge on the nature of data can be considered.

When inverting images, Geiping et al. (2020) propose to add the total variation regularization $R_{\text{TV}}(x)$ to the cost function in (3) since neighboring pixels of natural images are likely to have similar values. More formally,

$$R_{\text{TV}}(x) := \sum_{(i,j)} \sum_{(i',j') \in \partial(i,j)} \|x(i,j) - x(i',j')\|^2, \quad (4)$$

where $\partial(i,j)$ is the set of neighbors of (i,j) . This method is limited to the natural image data.

For general type of data, one can consider exploiting the batch normalization (BN) statistics from nodes. This is available in the case that the server wants to utilize batch normalization (BN) in FL, and thus collects the BN statistics (mean and variance) of batch from each node, in addition, with every gradient report (McMahan et al., 2017). To be specific, Yin et al. (2021) propose to employ the regularizer $R_{\text{BN}}(x_1, \dots, x_B; \theta)$ which quantifies the discrepancy between the BN statistics of estimated x_j 's and those of true x_j^* 's on each layer of the learning model. More formally,

$$R_{\text{BN}}(x_1, \dots, x_B; \theta) := \sum_l \|\mu_l - \mu_{l,\text{exact}}\|_2 + \|\sigma_l^2 - \sigma_{l,\text{exact}}^2\|_2,$$

where $\mu_l(x_1, \dots, x_B; \theta)$ and $\sigma_l^2(x_1, \dots, x_B; \theta)$ (resp. $\mu_{l,\text{exact}}(x_1^*, \dots, x_B^*; \theta)$ and $\sigma_{l,\text{exact}}^2(x_1^*, \dots, x_B^*; \theta)$) are the mean and variance of l -th layer feature maps for the estimated batch x_1, \dots, x_B (resp. the true batch x_1^*, \dots, x_B^*) given θ . This is available only if clients agreed to report their exact BN statistics at every round. But not every FL framework report BN statistics (Li et al., 2021; Andreux et al., 2020). In that case, Yin et al. (2021) also propose to use the BN statistics over the entire data distribution as a proxy of the true BN statistics, and reports that the gain from the approximated BN statistics is comparable to that from the exact ones. The applicability of R_{BN} with the approximated BN statistics is still limited as the proxy needs to be additionally recomputed over the entire data distribution at every change of θ . However, this demonstrates the significant impact of knowing the data distribution in the gradient inversion and motivates our methods using and learning a generative model on the user data, described in what follows.

4.1. Gradient inversion with trained generative model

Consider a decent generative model $G_w : \mathbb{R}^k \mapsto \mathbb{R}^m$ trained on the approximate (possibly exact) distribution of user data \mathcal{D} such that $x^* \approx G_w(z^*)$ for $(x^*, \cdot) \in \mathcal{D}$ and its latent code $z^* = \arg \min_z \|G_w(z) - x^*\|$. To fully utilize such a pretrained generative model, we propose gradient inversion in alternative spaces (GIAS), of which pseudocode is

presented in Appendix, which performs latent space search over z and then parameter space search over w .

Latent space search Note that the latent space is typically much smaller than the ambient input space, i.e., $k \ll m$, for instances, DCGAN (Simonyan and Zisserman, 2015) of $k = 100$ and StyleGAN (Karras et al., 2019) of $k = 512 \times 16$ for image data of $m = (\text{width}) \times (\text{height}) \times (\text{color})$ such as $32 \times 32 \times 3$, $256 \times 256 \times 3$, or larger. Using such a pretrained generative model with $k \ll m$, the under-determined issues of (3) can be directly mitigated by narrowing down the searching space from \mathbb{R}^m to $\{G_w(z) : z \in \mathbb{R}^k\}$. Hence, GIAS first performs the latent space search in the followings:

$$\min_{z_1, \dots, z_B \in \mathbb{R}^k} c(G_w(z_1), \dots, G_w(z_B)). \quad (5)$$

Considering a canonical class of neural network model f_θ , we can show that the reconstruction of x^* by latent space search in (5) aligns with that by input space search in (3) if the generative model G_w approximates input data with small enough error.

Property 1. *For an input data $x^* \in [0, 1]^m$ consider the gradient inversion problem of minimizing cost c in (3), where a canonical form of deep learning for classification is considered and the discrepancy measure d is ℓ_2 -distance. Suppose that it has the unique global minimizer at x^* . Let $\varepsilon \geq 0$ be the approximation error bound on x^* for generative model $G_w : [0, 1]^k \mapsto [0, 1]^m$. Then, there exists $\delta(\varepsilon) \geq 0$ such that for any $z^* \in \arg \min_z c(G_w(z))$,*

$$\|G_w(z^*) - x^*\| \leq \delta(\varepsilon), \quad (6)$$

of which upper bound $\delta(\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$.

A rigorous statement of Property 1 and its proof are provided in Appendix, where we prove and use that the cost function is continuous around x^* under the assumptions. This property justifies solving the latent space search in (5) for FL scenarios training neural network model while it requires an accurate generative model.

Parameter space search Using the latent space search only, there can be inevitable reconstruction error due to the imperfection of generative model. This is mainly because we cannot perfectly prepare the generative model for every plausible data in advance. Similar difficulty of the latent space search has been reported even when inverting GAN (Zhu et al., 2020; Bau et al., 2019a;b) for plausible but new data directly, i.e., $\min_z \|G_w(z) - x^*\|$ given x^* , rather than inverting gradient. Bau et al. (2019a) propose an instance-specific model adaptation, which slightly adjusts the model parameter w to (a part of source image) x^* after obtaining a latent code z^* for x^* . Inspired by such an instance-specific

adaptation, GIAS performs the following parameter space search over w preceded by the latent space search over z :

$$\min_{w_1, \dots, w_B} c(G_{w_1}(z_1), \dots, G_{w_B}(z_B)), \quad (7)$$

where z_1, \dots, z_B are obtained from (5).

Remark. We propose the optimization over w followed by that over z sequentially. This is to maximally utilize the benefit of mitigating the under-determined issue from reducing the searching space on the pretrained model. However, the benefit would be degenerated if z and w are optimized jointly or w is optimized first. We provide an empirical justification on the proposed searching strategy in Section 5.1.

We perform each search in GIAS using a standard gradient method to the cost function directly. It is worth noting that those optimizations (5) and (7) with generative model can be tackled in a recursive manner as R-GAP (Zhu and Blaschko, 2021) reconstructs each layer from output to input. We provide details and performance of the recursive procedure in Appendix, where employing generative model improves the inversion accuracy of R-GAP substantially, while R-GAP apparently suffers from an error accumulation issue when f_θ is deep neural network.

4.2. Gradient inversion to meta-learn generative model

For the case that pretrained generative model is unavailable, we devise an algorithm to train a generative model G_w for a set $\mathcal{S} = \{(\theta_i, g_i)\}$ of gradient inversion tasks. Since each inversion task can be considered as a small learning task to adapt generative model per data, we hence call it gradient inversions to meta-learn (GIML). The detailed procedure of GIML is presented in Appendix. We start with an arbitrary initialization of w , and iteratively update toward w' from a variant of GIAS for N tasks sub-sampled from \mathcal{S} , which is different than multiple applications of GIAS for each task in two folds: (i) ℓ_2 -regularization in latent space search; and (ii) an integrated optimization on model parameter. The variant first finds optimal latent codes $z_i^* = (z_{i1}^*, \dots, z_{iB}^*)$ for each task i with respect to the same cost function of GIAS but additional ℓ_2 -regularization. Note that the latent space search with untrained generative model easily diverges. The ℓ_2 -regularization is added to prevent the divergence of z_i^* . Once we obtained z_i^* 's, w' is computed by few steps of gradient descents for an integrated parameter search to minimize $\sum_i c(G_{w'}(z_{i1}^*), \dots, G_{w'}(z_{iB}^*); \theta_i, g_i)$. This is because in GIML, we want meta information w to help GIAS for each task rather than solving individual tasks, while after performing GIML to train w , we perform GIAS to invert gradient with the trained w . This is analogous to the model-agnostic meta learning in (Finn et al., 2017).

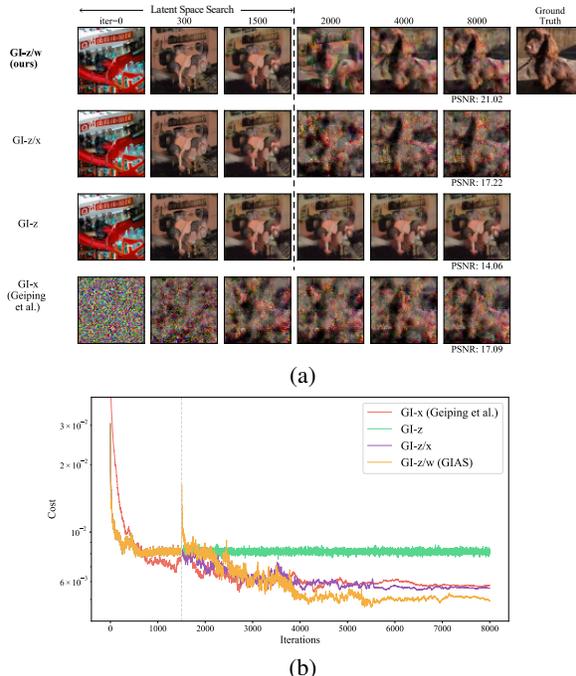


Figure 1. Comparison of different searching spaces, in which the first three rows share the same latent space search of 1,500 iterations, and GI- z/w is verified to be the best option to fully exploits the knowledge inside the generative model. (a) Reconstructed image over iterations of different optimization domains. (b) Cost function over iterations of different optimization domains.

5. Experiments

Setup. Unless stated otherwise, we consider the image classification task on the validation set of ImageNet (Russakovsky et al., 2015) dataset scaled down to 64×64 pixels (for computational tractability) and use ResNet18 (He et al., 2016) for training. For deep generative models in GIAS, we use StyleGAN2 (Karras et al., 2020) trained on ImageNet. We use a batch size of $B = 4$ as default and use the negative cosine to measure the gradient dissimilarity $d(\cdot, \cdot)$. We present detailed setup in the appendix.

Algorithms. We evaluate several algorithms for the gradient inversion (GI) task in (3). They differ mainly in which spaces each algorithm searches over: the input x , the latent code z , and/or the model parameter w . Each algorithm is denoted by GI-(\cdot), where the suffix indicates the search space(s). For instances, GI- z/w is identical to the proposed method, GIAS, and GI- x is the one proposed by Geiping et al. (2020).

5.1. Justification of GIAS design

We first provide an empirical justification of the specific order of searching spaces in GIAS (corresponding to GI- z/w) to fully utilize a pretrained generative model. To do so, we provide Figure 2b comparing algorithms with

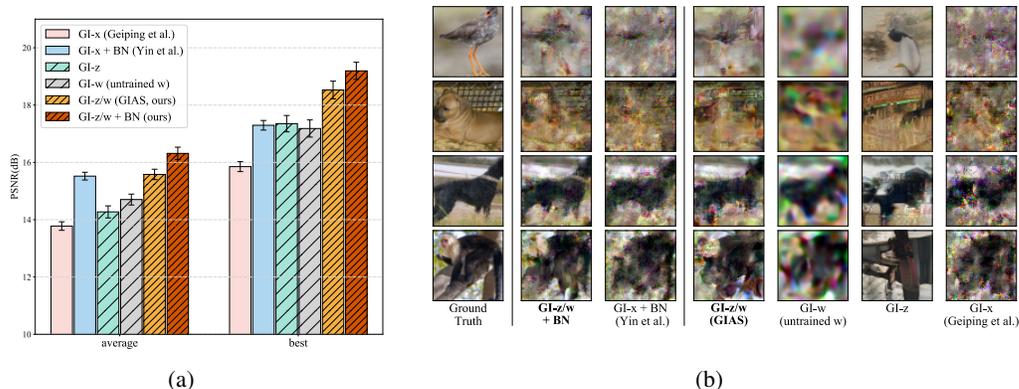


Figure 2. Comparison of state-of-the-art models and ablation study, in which replacing GI- x with GI- z/w (GIAS) regardless of using BN (Yin et al., 2021) or not (Geiping et al., 2020) provides substantial improvement in the reconstruction accuracy. (a) Average PSNR and best PSNR in a batch throughout the experiments. (b) Comparison of reconstruction results with our models and state-of-the-art models. We highlight proposed models in **bold**.

Table 1. Comparison of our methods with state-of-the-art methods. Adding our method makes performance improvement versus two baseline methods. PSNR, SSIM, and LPIPS (Zhang et al., 2018) are used to evaluate reconstruction results. We highlight the best performances in **bold**.

| Method | GI- x (Geiping et al., 2020) | GI- z (ours) | GI- w (ours) | GI- z/w (GIAS, ours) | GI- x +BN (Yin et al., 2021) | GI- z/w +BN (ours) |
|--------------------|--------------------------------|----------------|----------------|------------------------|--------------------------------|----------------------|
| PSNR \uparrow | 13.78 | 14.27 | 14.70 | 15.58 | 15.52 | 16.31 |
| SSIM \uparrow | 0.2542 | 0.3106 | 0.3519 | 0.3895 | 0.3513 | 0.4311 |
| LPIPS \downarrow | 0.4376 | 0.3233 | 0.5121 | 0.3023 | 0.3645 | 0.2861 |

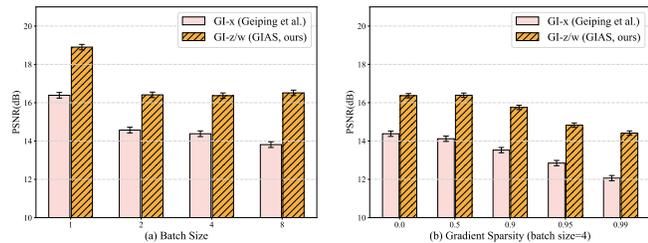


Figure 3. Comparison of reconstruction methods with varying difficulties. GI- z/w always surpasses GI- x thanks to the pretrained generative model.

different searching spaces: GI- z/w , GI- z/x , GI- z , and GI- x , of which the first three share the same latent space search over z for the first 1, 500 iterations.

As shown in Figure 1(a), the latent space search over z quickly finds plausible image in a much shorter number of iterations than GI- x , while it does not improve after a certain point due to the imperfection of pretrained generative model. Such a limitation of GI- z is also captured in Figure 1(b), where the cost function of GI- z is not decreasing after a certain number of optimization steps. To further minimize the cost function, one alternative to GI- z/w (GIAS) is GI- z/x , which can further reduce the loss function whereas the parameter search in GI- z/w seems to provide more natural reconstruction of the image than GI- z/x . The superiority of GI- z/w over GI- z/x may come from that the parameter space search exploits an implicit bias from optimizing a

good architecture for expressing images, c.f., deep image prior (Ulyanov et al., 2018). In Appendix, we also present the same comparison on FFHQ (human-face images) (Karras et al., 2019) where diversity is much smaller than that of ImageNet. On such a less diverse dataset, the distribution can be easily learned, and the gain from training a generative model is larger.

5.2. The gain from fully exploiting pretrained generative model

Evaluation in different FL settings We evaluate the gain of using generative models for various FL scenarios with varying levels of difficulty in the inversion. As batch size and gradient sparsity¹ (Wei et al., 2020) increase, the risk of having under-determined inversion increase and the inversion task becomes more challenging. Figure 3 shows that for all the levels of difficulty, the generative model provides significant gain in reconstruction quality. In particular, the quality of GI- x (Geiping et al., 2020) with a batch size of one (resp. gradient sparsity 0%) is comparable to that of GIAS with a batch size of eight (resp. gradient sparsity 99%).

Comparison with state-of-the-art models Our method can be easily added to previous methods (Geiping et al.,

¹Having gradient sparsity 0.99% implies that we reconstruct data from 1% of the gradient after removing 99% elements with the smallest magnitudes at each layer.

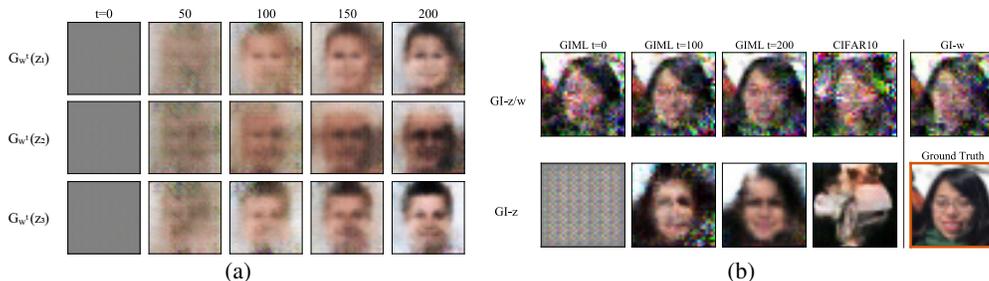


Figure 4. Results validating generative model trained with GIML. (a) Images sampled from different GIML training steps. Initial generative model is untrained DCGAN, and meta-learned with gradients computed on FFHQ images. same z was used to sample images in same rows. (b) Results of $GI-z$ and $GI-z/w$ using generative model trained with GIML. Experiments were done with gradient sparsity 0.95 for comparison in difficult setting. $GI-z/w$, $GI-z$ with CIFAR10 pretrained model and $GI-w$ with untrained model is in comparison.

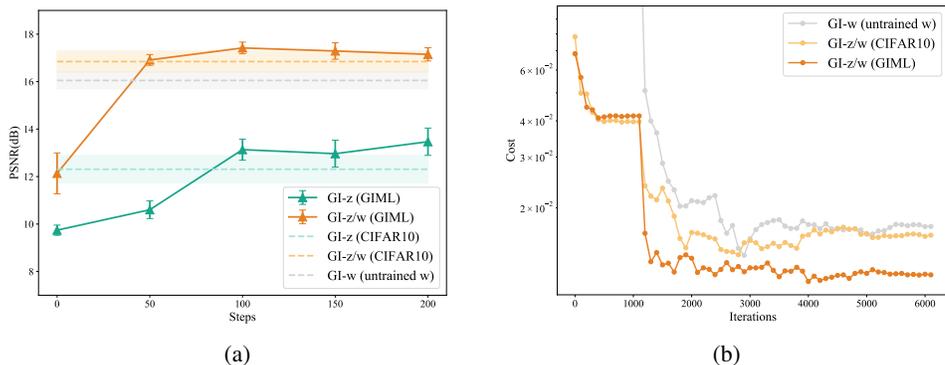


Figure 5. Gain of using generative model trained with GIML. (a) Comparison of PSNR. With GIML iteration step increases, performance of $GI-z$ and $GI-z/w$ with GIML surpass $GI-z$ and $GI-z/w$ with other models. (b) A typical loss curve of reconstruction process. Meta-learned model converges faster than other model’s results. Note that $GI-w$ does not perform latent space search of 0 to 1000 iterations.

2020; Yin et al., 2021). In Table 1 and Figure 2, we compare the state-of-the-art methods both with and without the proposed generative modelling. In Table 1, comparing $GI-x$ to $GI-z/w$ and $GI-x + BN$ to $GI-z/w + BN$, adding the proposed generative modelling provides additional gain in terms of all the measures (PSNR, SSIM, LPIPS) of reconstruction quality. $GI-z/w$ without BN has lower reconstruction error than $GI-x + BN$, which is the method of (Yin et al., 2021). This implies that the gain from the generative model is comparable to that from BN statistics. However, while the generative model only requires a global (and hence coarse) knowledge on the *entire dataset*, BN statistics are local to the batch in hand and hence requires significantly more detailed information on the *exact batch* used to compute gradient. As shown in Figure 2, the superiority of our method compared to the others is clear in terms of the best-in-batch performance than the average one, where the former is more suitable to show actual privacy threat in the worst case than the latter. It is also interesting to note that $GI-w$ with untrained w provides substantial gain compared to $GI-x$. This may imply that there is a gain of the implicit bias, c.f., (Ulyanov et al., 2018), from training the architecture of deep generative model.

5.3. Learning generative models from gradients

In this section, we demonstrate the possibility of *training* a generative model only with gradients. For computational tractability, we use DCGAN and images from FFHQ (Karras et al., 2019) resized to 32×32 . We generate a set of gradients from 4 rounds of gradient reports from 200 nodes, in which node computes gradient for a classification task based on the annotation provided in (DCGM, 2019). From the set of gradients, we perform GIML to train a DCGAN to potentially generate FFHQ data.

Figure 4(a) shows that the quality of images from the generative model is evolving in the training process of GIML. As the step t of GIML increases, the generative model $G_{w(t)}(z)$ for arbitrary z outputs more plausible image of human face. As shown in Figure 4(b) and Figure 5(a), the evolution of generative model improves the reconstruction quality when performing either $GI-z$ and $GI-z/w$. In Figure 4(b), when using generative model trained on wrong dataset (CIFAR10), $GI-z$ completely fails at recovering data and $GI-z/w$ shows similar quality to $GI-w$ which starts with an untrained generative model. $GI-z/w$ with GIML to train generative model on right data shows the best performance in terms of not

only quality (Figure 4(b)) but also convergence speed (Figure 5(b)).

6. Conclusion

We propose GIAS fully exploit the prior information on user data from a pretrained generative model when inverting gradient. We demonstrate significant privacy leakage using GIAS with pretrained generative model in various challenging scenarios, where our method provides substantial gain additionally to any other existing methods (Geiping et al., 2020; Yin et al., 2021). In addition, we propose GIML which can train a generative model using only the gradients seen in the FL classifier training. We experimentally show that GIML can meta-learn a generative model on the user data from only gradients, which improves the quality of each individual recovered image. To our best knowledge, GIML is the first capable of learning explicit prior on a set of gradient inversion tasks.

References

- R. Abdal, Y. Qin, and P. Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- M. Andreux, J. O. du Terrail, C. Beguier, and E. W. Tramel. Siloed federated learning for multi-centric histopathology datasets. In S. Albarqouni, S. Bakas, K. Kamnitsas, M. J. Cardoso, B. Landman, W. Li, F. Milletari, N. Rieke, H. Roth, D. Xu, and Z. Xu, editors, *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 129–139, Cham, 2020. Springer International Publishing. ISBN 978-3-030-60548-3.
- D. Bau, H. Strobel, W. Peebles, J. Wulff, B. Zhou, J.-Y. Zhu, and A. Torralba. Semantic photo manipulation with a generative image prior. *ACM Trans. Graph.*, 38(4), July 2019a. ISSN 0730-0301. doi: 10.1145/3306346.3323023. URL <https://doi.org/10.1145/3306346.3323023>.
- D. Bau, J.-Y. Zhu, J. Wulff, W. Peebles, H. Strobel, B. Zhou, and A. Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019b.
- T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi. Federated learning of predictive models from federated electronic health records. *IJMI*, 2018.
- DCGM. ffhq-features-dataset, 2019. URL <https://github.com/DCGM/ffhq-features-dataset>.
- L. Fan, K. W. Ng, C. Ju, T. Zhang, C. Liu, C. S. Chan, and Q. Yang. Rethinking privacy preserving deep learning: How to evaluate and thwart privacy attacks. In Q. Yang, L. Fan, and H. Yu, editors, *Federated Learning: Privacy and Incentive*, pages 32–50. Springer, 2020.
- C. Finn, P. Abbeel, and S. Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *International Conference on Machine Learning (ICML)*, 2017.
- J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller. Inverting gradients—How easy is it to break privacy in federated learning? In *NeurIPS*, 2020.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.
- D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1412.6980>.
- X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/pdf?id=6YEQUn0QICG>.
- Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally. Deep Gradient Compression: Reducing the communication bandwidth for distributed training. In *The International Conference on Learning Representations*, 2018.
- B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282, 2017.
- H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private recurrent language models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL <https://openreview.net/forum?id=BJ0hF1Z0b>.

- L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *IEEE Symp. Security and Privacy (SP)*, pages 691–706, 2019. doi: 10.1109/SP.2019.00029.
- L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345, 2018.
- O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017. doi: 10.1109/SP.2017.41.
- K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pages 2512–2520, 2019.
- W. Wei, L. Liu, M. Loper, K.-H. Chow, M. E. Gursoy, S. Truex, and Y. Wu. A framework for evaluating client privacy leakages in federated learning. In L. Chen, N. Li, K. Liang, and S. Schneider, editors, *Computer Security – ESORICS 2020*, pages 545–566, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58951-6.
- H. Yin, A. Mallya, A. Vahdat, J. M. Alvarez, J. Kautz, and P. Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, 2021.
- R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
- B. Zhao, K. R. Mopuri, and H. Bilen. iDLG: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.
- J. Zhu and M. B. Blaschko. R-GAP: Recursive gradient attack on privacy. In *International Conference on Learning Representations*, 2021.
- J. Zhu, Y. Shen, D. Zhao, and B. Zhou. In-domain gan inversion for real image editing. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.
- L. Zhu, Z. Liu, and S. Han. Deep leakage from gradients. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Appendix

A. Detailed algorithms

Algorithm 1 Gradient Inversion in Alternative Spaces (GIAS)

Require: learning model f_θ ; target gradient $g = \nabla_\theta \sum_{j=1}^B \ell(f_\theta(x_j^*), y_j^*)$ to be inverted; batch size B ; pre-trained generative model G_w ;

- 1: Initialize $z := (z_1, \dots, z_B)$ randomly
 - 2: Find $z \in \arg \min_z c(G_w(z_1), \dots, G_w(z_B))$
// Latent space search
 - 3: Set $w := (w_1, \dots, w_B) \leftarrow (w, \dots, w)$
 - 4: Find $w \in \arg \min_w c(G_{w_1}(z_1), \dots, G_{w_B}(z_B))$
// Parameter space search
 - 5: Return result: $G_{w'_1}(z_1), \dots, G_{w'_B}(z_B)$
-

Algorithm 2 Gradient Inversion to Meta-Learn generative model (GIML)

Require: inversion task set \mathcal{S} ; task batch size N ; data batch size B (per gradient); number of local iterations τ ; z -regularizer coefficient λ ; step sizes α, β ;

- 1: Initialize w randomly
 - 2: **while** not done **do**
 - 3: Sample a batch of inversion tasks $(\theta_1, g_1), \dots, (\theta_N, g_N)$ from \mathcal{S}
 - 4: $w' \leftarrow w$
 - 5: **for all** $i = 1, \dots, N$ **do**
 - 6: $z_i^* \leftarrow \operatorname{argmin}_{z_i} c(G_{w'}(z_{i1}), \dots, G_{w'}(z_{iB}); \theta_i, g_i) + \lambda \sum_j \|z_{ij}\|_2$ // Regularized latent space searching
 - 7: **end for**
 - 8: **for all** $t = 1, \dots, \tau$ **do**
 - 9: $w' \leftarrow w' - \alpha \nabla_{w'} \sum_i c(G_{w'}(z_{i1}^*), \dots, G_{w'}(z_{iB}^*); \theta_i, g_i)$
// Meta parameter space search
 - 10: **end for**
 - 11: Update $w \leftarrow w - \beta(w - w') = (1 - \beta)w + \beta w'$
 - 12: **end while**
-

B. Proof of Property 1

To prove Property 1, we first conclude the same statement of Property 1 assuming the inversion problem is continuous at x^* (Lemma 1), and then show that the standard scenario described in Property 1 guarantees the desired continuity (Lemma 2). The canonical form of learning model mentioned in 1 is described by the assumptions of Lemma 2.

Lemma 1 (An extension of Property 1 in the main text). *For an input data $x^* \in [0, 1]^m$, consider the gradient inversion problem of minimizing cost $c(x)$ in (3) where $c(x)$ is continuous. Suppose that it has the unique global minimizer at x^* . Let $\varepsilon \geq 0$ be the approximation error bound on x^* for*

generative model $G_w : [0, 1]^k \mapsto [0, 1]^m$ with $k \leq m$, i.e., $\min_{z \in [0, 1]^k} \|x^ - G_w(z)\| \leq \varepsilon$. Then, there exists $\delta(\varepsilon) \geq 0$ such that for any $z^* \in \arg \min_z c(G_w(z))$,*

$$\|G_w(z^*) - x^*\| \leq \delta(\varepsilon), \quad (8)$$

of which upper bound $\delta(\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$.

Proof of Lemma 1. From the assumptions that x^* is the unique minimizer and $c(x)$ is continuous on $[0, 1]^m$, it follows that for $x \in [0, 1]^m$, if $c(x) \rightarrow c(x^*)$, then $x \rightarrow x^*$. This can be proved by contradiction. Then we have that for $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ such that if $c(x) \leq \varepsilon$, then $\|x - x^*\| \leq \delta(\varepsilon)$ where $\delta(\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$. From the continuity of $c(x)$, it is straightforward to check that $c(G_w(z^*(\varepsilon))) \rightarrow c(x^*)$ as $\varepsilon \rightarrow 0$. This completes the proof. \square

Lemma 2. *For an input data $x^* \in \mathbb{R}^m$, consider the gradient inversion problem of minimizing cost c in (3), where the learning model f_θ is a standard form of R -layer neural network $f_\theta(x) = \Theta_R \sigma_{R-1}(\Theta_{R-1} \sigma_{R-2}(\dots \Theta_1 x))$ with $\Theta_r \in \mathbb{R}^{m_r \times m_{r-1}}$ for each r , where $m_R = L$ and $m_0 = m$, and $C1$ (continuously differentiable) activation σ 's (e.g., sigmoid and exponential linear), loss function ℓ is $C1$ (e.g., logistic and exponential), and the discrepancy measure d is ℓ_2 -distance. Then, the corresponding cost function $c(x)$ is continuous with respect to $x \in \mathbb{R}^m$.*

Proof of Lemma 2. Note that the standard model f_θ includes multi-layer perceptron or convolutional neural network. It is $C1$ since the composition of Ck^2 functions is Ck . Hence, the gradient is continuous w.r.t. x . In addition, the cost function $c(x)$ is continuous since the gradient and the choice of discrepancy measure are continuous. This concludes the proof. \square

The proof of Property 1 is straightforward from Lemmas 1 and 2.

C. Another method for gradient inversion: R-GAP (Zhu and Blaschko, 2021)

In the main text, to solve the inversion problem in (2), we use gradient descent method directly to the cost function, while we alternate the searching space. Meanwhile, Zhu and Blaschko (2021) propose another approach, called R-GAP (recursive gradient attack on privacy), to solve the optimization in (2), although it is limited to the case when the learning model f_θ is given as a standard form described in Lemma 2. R-GAP decomposes the optimization (2) into a sequence of linear programming to reconstruct the output of each layer except the last layer's, and then it solves them

²the k -th derivative is continuous

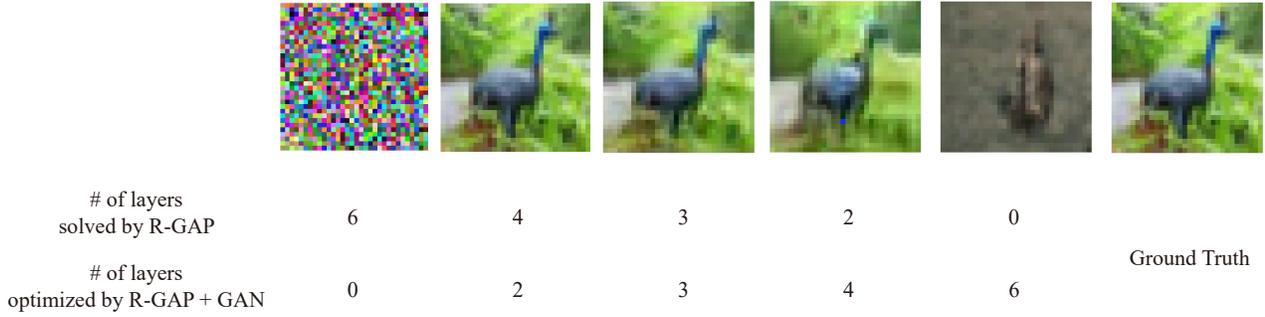


Figure A1. Comparison of R-GAP and R-GAP with a generative model. The second convolution layer is rank-deficient and R-GAP should solve under-determined problem. An under-determined problem is solved by using generative model. However, the error per layer increases much faster than R-GAP.

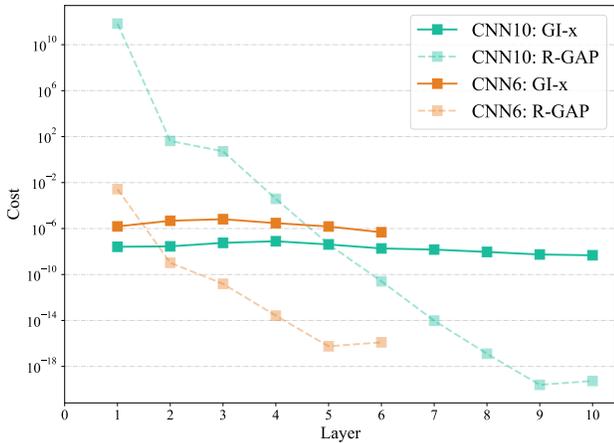


Figure A2. Layer-wise errors of convolution layers in gradient inversion attack results. layer-by-layer method shows cumulative exploding error. CNN6 denotes a neural network consists of six convolution layers and one FC layer. CNN10 denotes a neural network consists of ten convolution layers and one FC layer.

recursively from the penultimate layer to the input layer. The linear programming to find the output x_r of the r -th layer can be written as follows:

$$A_r x_r = b_r \quad (9)$$

where A_r and b_r is a matrix and vector depending on the previously reconstructed x_{r+1} , the parameter Θ_r of the r -th layer and its gradients. For the definition of A_r and b_r , we refer to (Zhu and Blaschko, 2021). Since each linear programming has a closed-form solution $A_r^\dagger b_r$, this approach can be sometimes useful in terms of reducing computational cost.

R-GAP with generative model Note that the problem in (9) can be rewritten as follows:

$$\min_{x_r} \|A_r x_r - b_r\|. \quad (10)$$

Let $f_{\theta,r}(x)$ be the output of the r -th layer. Then, we can interpret $f_{\theta,r}(G_w(z))$ as a generative model for x_r . Hence, the recursive reconstruction can be partially or fully replaced with the following optimization:

$$\min_{z,w} \|A_r f_{\theta,r}(G_w(z)) - b_r\| \quad (11)$$

where the search space can be alternated arbitrarily.

A limitation of R-GAP Such a use of generative model in (11) provides the same gain from reducing searching space. We however want to note that it inherits the limitation of R-GAP, in which the reconstruction error in upper layers propagates to that in lower layer. Hence, as the learning model becomes deeper, the reconstruction quality decreases while the number of parameters is increasing. Figure A2 shows the phenomenon of error accumulation of R-GAP. It is possible that the optimization method in (11) can have larger error than the closed-form solution $A_r^\dagger b_r$ due to imperfection in generative model. Therefore, it is better to not use the generative model when the original linear programming is over-determined or determined. Indeed, in Figure A1, we present a trade-off between the linear programming in (11) and the optimization with generative model in (11), in which to emphasize the trade-off, we perform the latent space search over z only. We obtain a substantial gain from using the generative model for a few layers (one or two), whereas the gradient inversion is failed when using the generative model for every layer.

D. Another potential gain from inverting a set of gradients

In the main text, we demonstrate that from multiple gradients, we can train a generative model and use it to break the fundamental limit of inverting gradient solely. Beside this, assuming that we can observe a large number of gradients for the same data but different model parameters,

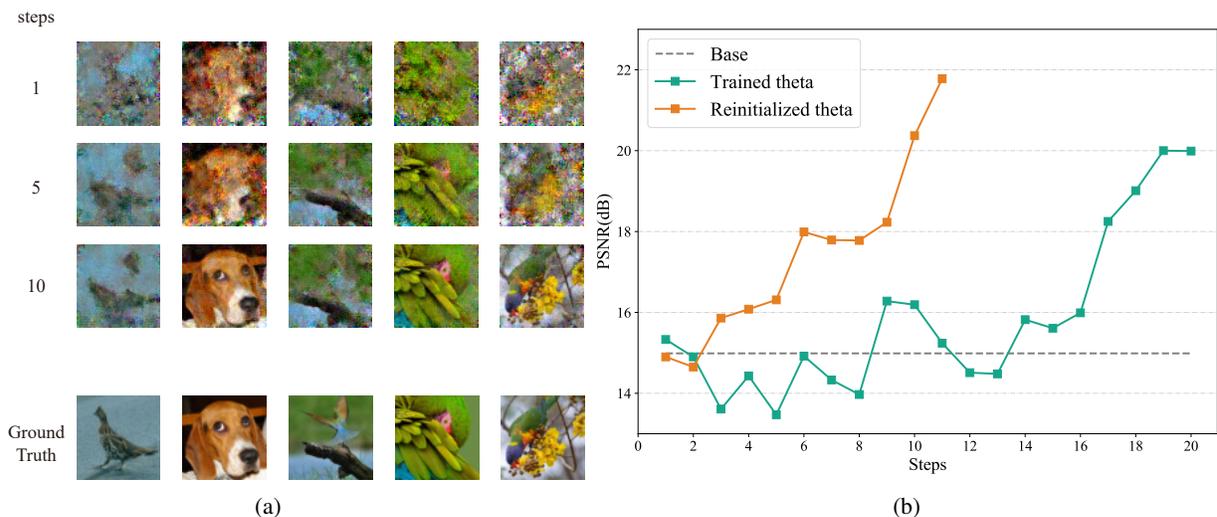


Figure A3. (a) Examples of reconstructed images with a sequence of gradients. A local dataset contains eight images and generates gradients using four randomly chosen images in every step. As time t increases, reconstructed images become more accurate. (b) PSNR of reconstructed images. In reinitialized theta setting, the classification model is reinitialized every step. In trained theta setting, the classification model is trained every step.

it is able to reconstruct data almost perfectly by solving $\min_x \sum_{t=1}^T c(x; \theta_t, g_t)$. Such an assumption may be valid once we obtain the meta information to match gradients and data to be reconstructed. In Figure A3, we demonstrate this potential gain when there are eight images only, but we observe a sequence of gradients obtained from the procedure of FL (the green curve). Of course, in the procedure of FL, the model parameter θ slowly changes and thus the gain is smaller than that when each gradient is computed at completely random model parameters. However, in both settings, the reconstruction eventually becomes perfect as the observed gradients are accumulated.

E. Strong generative prior

When we have stronger prior on the data distribution, the gain from the generative model becomes larger. To show this, we use FFHQ (Karras et al., 2019) rather than ImageNet in Section 5.1, where we believe FFHQ containing human-face images has less diversity than ImageNet including images of one thousand classes. Figure A4 shows that just GI-z significantly outperforms GI-x, while the gap between GI-z and GI-x is small for ImageNet in Figure 2b. This suggests a new approach to use a conditional generative model and data label y^* in order for enjoying the gain from narrowing down the set of candidate input data by conditioning the label.

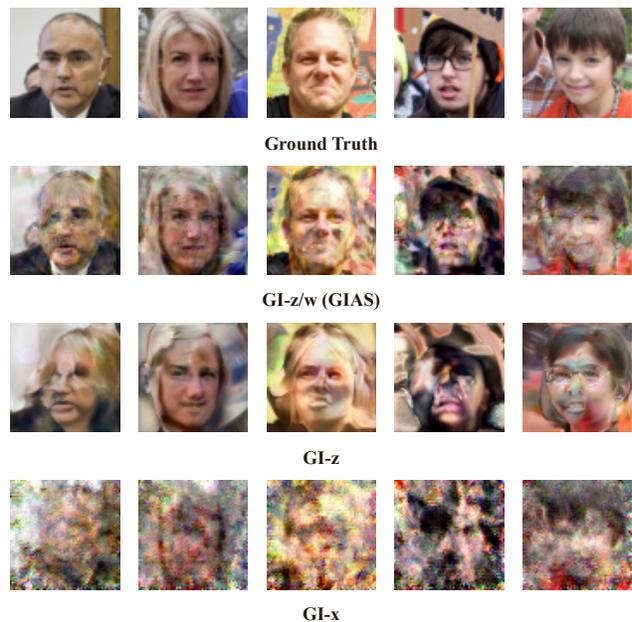


Figure A4. Examples of reconstructed images from gradients of FFHQ images. Images are resized to 64x64, and the gradients are calculated under batch size 1, gradient sparsity 0. GIAS algorithm use StyleGAN2 pretrained on FFHQ.

F. Possible defense methods against gradient inversion attacks

In this section, we briefly discuss several defense algorithms against gradient inversion attacks. There have been pro-

posed several defense methods against gradient inversion attacks (McMahan et al., 2017; Lin et al., 2018; Wei et al., 2020), which let the gradient contain only small amount of information per data. For instances, a gradient is computed from a large batch of data and is reported to the server after compression or quantization. Such approaches can easily make the model training unstable. In general, we need to find a good balance in the trade-off between the stability of FL and the privacy leakage, while each defense mechanism has distinguishing pros and cons.

As a defense mechanism specialized for the gradient inversion with generative model, we suggest to focus on mechanisms confusing the label reconstruction. In Section 5.1 and E, we observe that revealing the data label can curtail the candidate set of input data and thus provide a significant gain in gradient inversion by using conditional generative model. Therefore, by making the label restoration challenging, the gain from generative model may be decreasing. To be specific, we can consider letting node sample mini-batch to contain data having a certain number of labels, less than the number of data but not too small. By doing this, the possible combinations of labels per data in a batch increases and thus the labels are hard to recovered.

G. Experiment settings

Unless stated otherwise, we consider the image classification task on the validation set of ImageNet (Russakovsky et al., 2015) dataset resized to 64×64 using ResNet18 (He et al., 2016) as learning model. The resizing is necessary for computational tractability. Recalling the under-determined issue is the major challenge in gradient inversion, deeper and wider f_θ makes the gradient inversion easier (Geiping et al., 2020). Hence, the choice of ResNet18 as learning model is the most difficult setting within ResNet architectures since it contains the least information. We use a StyleGAN2 (Karras et al., 2020) model trained on ImageNet for GIAS, in which the latent space search over z implies the search over the intermediate latent space, known as \mathcal{W} in the original paper (Karras et al., 2019), to improve the reconstruction performance, c.f., (Abdal et al., 2020). We use the batch size $B = 4$ as default, and negative cosine for the choice of gradient dissimilarity function $d(\cdot, \cdot)$, which apparently provides better inversion performance than ℓ_2 -distance in general (Geiping et al., 2020). For the optimization in GIAS, we use Adam optimizer (Kingma and Ba, 2015) which decays learning rate by a factor of 0.1 at $3/8, 5/8, 7/8$ of total iterations. from initial learning rates $\eta_z = 3 \times 10^{-2}$ for the latent space search and $\eta_w = 10^{-3}$ for the parameter space search. Since our experiments are conducted with image data, we used total variation regularizer with weight $\lambda_{TV} = 10^{-4}$ for all experiments. For each inversion, we pick the best recovery among 4 random instances based on the final loss. All

experiments are performed on GPU servers equipped with NVIDIA RTX 3090 GPU and NVIDIA RTX 2080 Ti GPU. Numerical results including graphs and table are averaged over 10 samples except Figure 1 and Figure 5.

Note that our baseline implementation for (Yin et al., 2021) includes fidelity regularizer with BN_{exact} and group lazy regularizer, not group registration regularizer. Our baseline implementation might be imperfect, but it still demonstrates adding our method improves performance.

H. License of assets

Dataset ImageNet data are distributed under licenses which allow free use for non-commercial research. FFHQ data are distributed under licenses which allow free use, redistribution, and adaptation for non-commercial purposes. ffhq-features-dataset provides annotations of FFHQ images. Original authors of FFHQ images are indicated in the metadata, if required. We did not include, redistribute, or change the data itself, and cited above three works.

Source code Some parts of our source code came from open-source codes of several previous researches. For more details, see README of our source code.