INSTANCE-WISE BATCH LABEL RESTORATION VIA GRADIENTS IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient inversion attacks have posed a serious threat to the privacy of federated learning. The attacks search for the optimal pair of input and label best matching the shared gradients and the search space of the attacks can be reduced by pre-restoring labels. Recently, label restoration technique allows for the extraction of labels from gradients analytically, but even the state-of-the-art remains limited to identify the presence of categories (i.e., the class-wise label restoration). This work considers the more real-world settings, where there are multiple instances of each class in a training batch. An analytic method is proposed to perform instance-wise batch label restoration from only the gradient of the final layer. On the basis of the approximate recovered class-wise embeddings and post-softmax probabilities, we establish linear equations of the gradients, probabilities and labels to derive the *Number of Instances* (*NoI*) per class by the *Moore-Penrose pseudoinverse* algorithm. Our experimental evaluations reach over 99% *Label existence Accuracy* (*LeAcc*) and exceed 96% *Label number Accuracy* (*LnAcc*) in most cases on three image datasets and four classification models. The two metrics are used to evaluate class-wise and instance-wise label restoration accuracy, respectively. And the recovery is made feasible even with a batch size of 4096 and partially negative activations (e.g., Leaky ReLU and Swish). Furthermore, we demonstrate that our method facilitates the existing gradient inversion attacks by exploiting the recovered labels, with an increase of 6-7 in PSNR on both MNIST and CIFAR100.

1 INTRODUCTION

Federated Learning (FL) is one of the most popular distributed learning paradigms to achieve privacy preservation and has attracted widespread attention (Jochems et al., 2016; McMahan et al., 2017; Yang et al., 2019a), especially in privacy-sensitive fields such as healthcare (Brisimi et al., 2018; Sadilek et al., 2021) and finance (Yang et al., 2019b; Long et al., 2020). FL requires participants to communicate the gradients or weight updates instead of private data on a central server, in principle, offering sufficient privacy protection.

Contrary to prior belief, recent works have demonstrated that shared gradients can still leak sensitive information in FL. Multiple attack strategies preliminarily look into the issue despite their own limitations. For instance, Membership Inference (Shokri et al., 2017) allows the adversary to determine whether an existing data sample is involved in the training set. Property Inference (Melis et al., 2019) analogously retrieves certain attributes (e.g., people's race and gender in the training set). Model Inversion (Fredrikson et al., 2015) utilizes the GAN (Goodfellow et al., 2014) model to generate visual alternatives that look similar but are not the original data. An emerging research, i.e., Deep Leakage from Gradients (DLG) (Zhu et al., 2019), has showed the possibility of fully recovering input data given gradients in a process now known as *gradient inversion*. This approach primarily rely on the *gradient matching* objective to perform the attack, i.e., optimizing a dummy input by minimizing the distance between its gradient to a target gradient sent from a client. Furthermore, it enables pixel-wise accurate reconstruction on image classification tasks, and soon scales to deeper networks and larger-resolution images in a mini-batch (Geiping et al., 2020; Yin et al., 2021; Jeon et al., 2021; Li et al., 2022).

Gradient inversion jointly recovers private inputs and labels, whereas most works focus more on restoration of training samples. Label restoration is non-trivial, and it has been shown to be critical

to high-quality data reconstruction. The optimization-based method for label recovery is not guaranteed to succeed as well. Analytical extraction of the ground truth label from gradients produced by a single sample (Zhao et al., 2020) was first made possible through an innovative observation of the gradient sign. Along this line, follow-ups (Yin et al., 2021; Dang et al., 2021) extend this method to the recovery of labels in a mini-batch with a high success rate. Despite remarkable progress, such attacks are applicable only when no two inputs in the batch belong to the same class. In real-world scenarios, there are multiple instances of each category, which means existing attacks can only determine which classes of samples are present at best. Moreover, most prior methods assume that the target model employs a non-negative activation function, and they fail to handle activation functions that might produce negative elements (e.g., Leaky ReLU (Maas et al., 2013) and Swish (Ramachandran et al., 2017) in EfficientNet (Tan & Le, 2019)).

To this end, our work aims to identify the ground-truth label for each data point in one batch (i.e., instance-wise labels). Firstly, we are able to recover class-wise averaged inputs (i.e., embeddings) to the final layer from its gradients with relative fidelity. The same approach can be proved perfectly correct in the single-sample case (Geiping et al., 2020), but here we require two empirical properties (Sun et al., 2021) to make three approximations: *intra-class high similarity* and *inter-class low entanglement* of embeddings and gradients. Then we derive that the gradient w.r.t. the network final logits equals the difference between the post-softmax probability and the binary representation of the label, and expand it to a system of equations consisting of the gradients, probabilities and labels, where all three variables can be substituted or decomposed equivalently. We finally obtain a *Least Square Solution* (*LSS*) of the *Number of Instances* (*NoI*) per class by the *Moore-Penrose pseudoinverse* algorithm.

Our main contributions are as follows:

- We propose an analytic method to recover the exact labels that a client possesses via batch-averaged gradients in FL. Following the approximate restoration of class-wise embeddings and post-softmax probabilities, a system of linear equations with the gradients, probabilities and labels is established to derive the NoI of each class by the Moore-Penrose pseudoinverse algorithm.

- Our method outperforms prior attacks, works on large batch sizes of up to 4096, and handles classification models with partially negative activations as deep as ResNet-152 (He et al., 2016). We maintain a high proportion on both two metrics of label restoration accuracy, i.e., *Label existence Accuracy* (*LeAcc*) and *Label number Accuracy* (*LnAcc*).

- We demonstrate that gradient inversion attack can be combined with our batch label restoration to further improve its performance due to the reduced search space during optimization.

2 Premilinaries And Related Work

2.1 Problem Formulation

Given a network with weights $\mathbf{W}$ and the batch-averaged gradient $\nabla\mathbf{W}$ calculated from a batch of sample-label pairs, we expect to reveal instance-wise labels $\mathbf{y}$ via gradients. For each pair ($\mathbf{x}$, $\mathbf{y}$), we denote the embedding vector into the final layer as $\mathbf{e} \in \mathbb{R}^m$, the network final logits as $\mathbf{z} \in \mathbb{R}^C$ and the post-softmax probability as $\mathbf{p} \in \mathbb{R}^C$ in range (0, 1), where m is the embedding dimension, C is the number of classes and $\mathbf{y}$ here is the one-hot binary representation of the same shape as $\mathbf{z}$. In the following sections, $\mathbf{W} \in \mathbb{R}^{C\times m}$ and $\mathbf{b} \in \mathbb{R}^C$ refer to the weight and bias of the final classification layer, respectively. Then we have $\mathbf{z} = \mathbf{W}\mathbf{e} + \mathbf{b}$ and $\mathbf{p} = SoftMax(\mathbf{z})$.

Threat Model: As prior works (Zhu et al., 2019; Geiping et al., 2020; Yin et al., 2021; Jeon et al., 2021; Li et al., 2022; Zhao et al., 2020; Dang et al., 2021), the adversary we consider is an honest-but-curious server which gets access to the global model and shared gradients as shown in Fig 1. Our attack targets model architectures that contain at least one fully-connected layer and have a softmax activation with cross-entropy loss for classification, such as fully-connected neural networks (FCN) and convolutional neural networks from shallow to deep——LeNet-5 (LeCun et al., 1998), VGG (Simonyan & Zisserman, 2014), ResNet (He et al., 2016), etc.

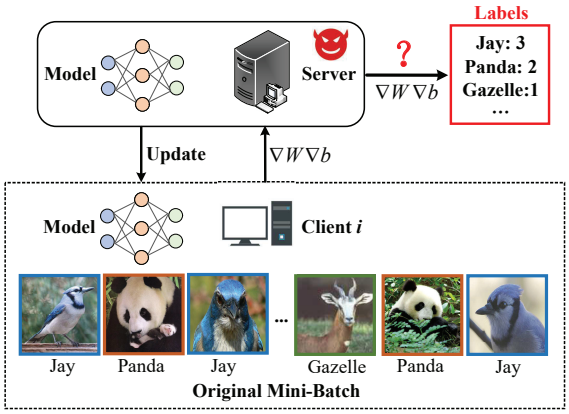

Figure 1: Threat Model.

2.2 Analytic Label Restorations

Prior analytical label restoration methods primarily rely on a major observation. For a single sample, the derivative of the cross-entropy loss $\mathcal{L}$ w.r.t. the network final logit $\mathbf{z}$ at index i is $\nabla z_i = p_i - y_i$ (See Appendix A for detailed derivation). This, obviously, leads to a unique negative sign for ∇z_i at the ground truth index c. However, we are able to access $\nabla \mathbf{W}_i$ instead of ∇z_i. While using the chain rule, we have $\nabla \mathbf{W}_i = \nabla z_i \times \nabla_{\mathbf{W}_i} z_i = \nabla z_i \mathbf{e}^\top$. As the embedding $\mathbf{e}$ is independent of the class index i, the uniqueness of ∇z_i's sign is passed on to $\nabla \mathbf{W}_i$. Formally, we have $\nabla \mathbf{W}_i \cdot \nabla \mathbf{W}_j = ||\mathbf{e}||^2 \nabla z_i \nabla z_j$, $\nabla z_c < 0$, and $\nabla z_{i \neq c} > 0$, such that the label c can be identified by inspecting whether $\nabla \mathbf{W}_i \cdot \nabla \mathbf{W}_j \leq 0, \forall j \neq i$. Because of the common use of non-negative activation functions (ensure that $||\mathbf{e}|| > 0$), e.g., ReLU (Glorot et al., 2011) and Sigmoid, we can simply extract the ground-truth label whose $\nabla \mathbf{W}_i$ is negative. This is exactly what iDLG (Zhao et al., 2020) does. Subsequently, the researchers (Yin et al., 2021) observe empirically the magnitude of the negative gradient significantly exceeds that of the positive gradient assuming a non-negative non-linear activation function is applied, which indicates a negative sign can still stand out after averaging operations. Therefore, they search for negative signs using minimum values rather than summation along the feature dimension to perform label restoration from batch-averaged gradients. However, without the assumption that $||\mathbf{e}|| > 0$, neither approach is viable any longer. Another batch label restoration method, Revealing Labels from Gradients (RLG) (Dang et al., 2021), offers a novel insight thus no need to satisfy that assumption. $\nabla \mathbf{W}^\top$ can be decomposed into $\mathbf{P}\Sigma\mathbf{Q}$ by singular value decomposition, where $\mathbf{P} \in \mathbb{R}^{m \times S}$ and $\mathbf{Q} \in \mathbb{R}^{S \times C}$ are orthogonal matrices, $\Sigma \in \mathbb{R}^{\mathrm{S} \times \mathrm{S}}$ in the middle is a diagonal matrix with non-negative elements on the diagonal, and $S = rank(\nabla \mathbf{W}^\top) < min\{m, C\}$. If we denote $\mathbf{r} = \nabla z_i \mathbf{Q}^\top$, then $\mathbf{rQ} = \nabla z_i$, which means $\mathbf{rq}^c < 0$ and $\mathbf{rq}^{j \neq c} > 0$ ($\mathbf{q}^j$ is the j-th column in $\mathbf{Q}$). The problem of label recovery is then transformed into finding a classifier to separate $\mathbf{q}^c$ from $\mathbf{q}^{j \neq c}$ by linear programming. Moreover, $\nabla \mathbf{W}$ can also be disassembled: $\nabla \mathbf{W} = \frac{1}{K} \sum_{j=1}^{K} \nabla \mathbf{z}^j \mathbf{e}^{\top j} = \mathbf{Z}\mathbf{E}^\top$, where $\mathbf{Z} = [\nabla \mathbf{z}^1, ..., \nabla \mathbf{z}^K] \in \mathbb{R}^{C \times K}, \mathbf{E} = \frac{1}{K}[\mathbf{e}^1, ..., \mathbf{e}^K] \in \mathbb{R}^{m \times K}$ and K is the batch size. Assume that $\mathbf{Z}$ and $\mathbf{E}$ are full-rank matrices, we have $K = S < min\{m, C\}$, so the approach requires K not to be large.

2.3 Single Embedding Reconstruction

In deep neural network architectures, the fully-connected layer is more vulnerable to leakage from gradients for its simple design. A recent work, i.e., InvertingGradients (IG) (Geiping et al., 2020), has brought theoretical insights on this task by showing provable embedding reconstruction feasibility.

Theorem 1. *For neural networks with a biased fully-connected layer (e.g., the final classification layer), presume the derivative of the loss $\mathcal{L}$ w.r.t. to the layer's output $\mathbf{z}$ contains at least one non-*

zero element, then the input to the fully-connected layers $\mathbf{e}$ *can be uniquely reconstructed by analytic computation.*

Proof. Consider the mapping $\mathbf{z} = \mathbf{W}\mathbf{e} + \mathbf{b}$ of a biased full-connected layer without a nonlinear activation, it's easy to observe that $\frac{\partial \mathbf{z}}{\partial \mathbf{W}\mathbf{e}} = \frac{\partial \mathbf{z}}{\partial \mathbf{b}} = \mathbf{1} \in \mathbb{R}^C$. Due to our assumption guarantees $\frac{\partial \mathcal{L}}{\partial z_i} \neq 0$ for some index i, we have $\frac{\partial \mathcal{L}}{\partial b_i} = \frac{\partial \mathcal{L}}{\partial z_i} \times \frac{\partial z_i}{\partial b_i} = \frac{\partial \mathcal{L}}{\partial z_i} \times 1 = \frac{\partial \mathcal{L}}{\partial z_i}$ and $\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i} = \frac{\partial \mathcal{L}}{\partial z_i} \times \frac{\partial z_i}{\partial \mathbf{W}_i} = \frac{\partial \mathcal{L}}{\partial z_i}\mathbf{e}^\top = \frac{\partial \mathcal{L}}{\partial b_i}\mathbf{e}^\top$ according to the chain rule. Therefore, $\mathbf{e}$ can be calculated exactly as $\mathbf{e} = (\frac{\partial \mathcal{L}}{\partial b_i})^{-1}(\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i})^\top$. □

On the basis of Theorem 1, we are able to perfectly accomplish the analytic reconstruction of a single input to a fully-connected layer. Such a one-shot theoretical approach, however, cannot be extended directly to recover batch embeddings for unignorable information loss from the average operation.

3 METHODOLOGY

In this section, we propose a method to restore instance-wise labels via batch-averaged gradients, which we refer to as instance-wise Labels Restoration from Gradients (iLRG). Our method consists of three main steps as shown in Fig 2.

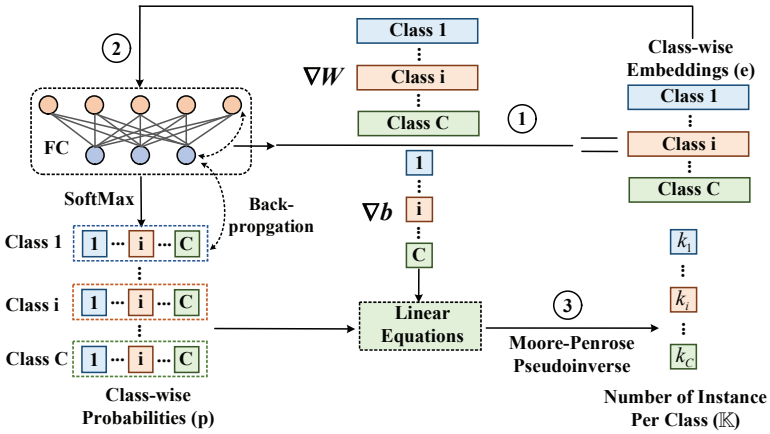

Figure 2: Overview of our proposed iLRG method: (1) Reconstruct the class-wise embeddings by calculating the quotient of two gradients of the weight and bias in the final classification layer; (2) Feed the embeddings into this layer to obtain the subsequent post-softmax probabilities; (3) Solve a system of linear equations for the number of instances per class by the Moore-Penrose pseudoinverse algorithm.

3.1 CLASS-WISE EMBEDDING RECONSTRUCTION

Two crucial observations presented in Soteria (Sun et al., 2021) push the embedding reconstruction towards deeper: (1) intra-class high similarity of data representations/embeddings and their gradients; (2) inter-class low entanglement, an empirical observation that embeddings from different classes tend to be embedded in different gradient rows. Leveraging them, we can attempt to reconstruct the average embeddings of each class, i.e., class-wise embeddings.

We first divide a training batch $\mathbb{B}$ into subsets of C distinct classes, i.e., $\mathbb{B} = \{\mathbb{B}_1, ..., \mathbb{B}_C\}$. Then, based on the observations mentioned above, we make two approximations. The first approximation has been formalized in the Appendix material of Soteria. We then further propose a theoretical account of the latter.

Approx 1 (Intra-class High Similarity). *The embeddings* $\mathbf{e}$ *and the derivatives of the loss* $\mathcal{L}$ *w.r.t. to their outputs* $\mathbf{z}$ *are both similar over a certain class of samples* $\mathbb{B}_i$ *in a training batch.*

Consequently, the average gradient at index i over $\mathbb{B}_i$ can be represented as formula 1 with Theorem 1:

$$\begin{aligned}\overline{\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i}}_{\mathbb{B}_i} = \frac{1}{|\mathbb{B}_i|}\sum\nolimits_{j\in\mathbb{B}_i}\frac{\partial \mathcal{L}^j}{\partial \mathbf{W}_i^j} = \frac{1}{|\mathbb{B}_i|}\sum\nolimits_{j\in\mathbb{B}_i}\frac{\partial \mathcal{L}^j}{\partial b_i^j}\mathbf{e}^{j\top} &\approx (\frac{1}{|\mathbb{B}_i|}\sum\nolimits_{j\in\mathbb{B}_i}\frac{\partial \mathcal{L}^j}{\partial b_i^j})(\frac{1}{|\mathbb{B}_i|}\sum\nolimits_{j\in\mathbb{B}_i}\mathbf{e}^{j\top}) \\ &= \overline{\frac{\partial \mathcal{L}}{\partial b_i}}_{\mathbb{B}_i}\overline{\mathbf{e}}_{\mathbb{B}_i}^{\top},\end{aligned} \tag{1}$$

where $\overline{(\cdot)}_{\mathbb{B}_i}$ denotes the mean of a variable across $\mathbb{B}_i$.

We already know that $\frac{\partial \mathcal{L}}{\partial b_i} = \frac{\partial \mathcal{L}}{\partial z_i} = p_i - y_i$, which means the gradient only goes negative at the ground-truth class index c, and $\sum_{i=1}^{C}\frac{\partial \mathcal{L}}{\partial b_i} = \sum_{i=1}^{C}(p_i - y_i) = 0$. In other words, the *av* (*absolute value*) of the negative gradient at class index c is equal to the sum of *av*s of the other positive gradients. Therefore, $\overline{\frac{\partial \mathcal{L}}{\partial b_i}}_{\mathbb{B}}$ can be approximated by $\overline{\frac{\partial \mathcal{L}}{\partial b_i}}_{\mathbb{B}_i}$. Furthermore, we have $\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i} = \frac{\partial \mathcal{L}}{\partial b_i}\mathbf{e}^{\top}$ and observe that the variance of $||\mathbf{e}||$ over the whole batch is so small that $\overline{\frac{\partial \mathcal{L}}{\partial \mathrm{W}_i}}_{\mathbb{B}} \approx \overline{\frac{\partial \mathcal{L}}{\partial \mathrm{W}_i}}_{\mathbb{B}_i}$.

Approx 2 (Inter-class Low Entanglement). *The batch-averaged gradient row at index* i *is mainly from* i*-class samples in a training batch. Specifically, we have:*

$$\overline{\frac{\partial \mathcal{L}}{\partial b_i}}_{\mathbb{B}} = \frac{1}{|\mathbb{B}|}\sum\nolimits_{j=1}^{C}|\mathbb{B}_j|\overline{\frac{\partial \mathcal{L}}{\partial b_i}}_{\mathbb{B}_j} \approx \frac{|\mathbb{B}_i|}{|\mathbb{B}|}\overline{\frac{\partial \mathcal{L}}{\partial b_i}}_{\mathbb{B}_i}, \tag{2}$$

$$\overline{\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i}}_{\mathbb{B}} = \frac{1}{|\mathbb{B}|}\sum\nolimits_{j=1}^{C}|\mathbb{B}_j|\overline{\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i}}_{\mathbb{B}_j} \approx \frac{|\mathbb{B}_i|}{|\mathbb{B}|}\overline{\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i}}_{\mathbb{B}_i}. \tag{3}$$

Hence, we can derive that $\overline{\frac{\partial \mathcal{L}}{\partial b_i}}_{\mathbb{B}_i}^{-1} \times \overline{\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i}}_{\mathbb{B}_i}^{\top} \approx \overline{\frac{\partial \mathcal{L}}{\partial b_i}}_{\mathbb{B}}^{-1} \times \overline{\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i}}_{\mathbb{B}}^{\top}$. Of course, it is necessary to ensure that $\overline{\frac{\partial \mathcal{L}}{\partial b_i}}_{\mathbb{B}} \neq 0$ and $\overline{\frac{\partial \mathcal{L}}{\partial b_i}}_{\mathbb{B}_i} \neq 0$ here. The occurrence of zero is uncommon, but when it does occur, we replace it with a small enough number ϵ. Combined with formula 1, we finally get $\overline{\mathbf{e}}_{\mathbb{B}_i} \approx \overline{\frac{\partial \mathcal{L}}{\partial b_i}}_{\mathbb{B}}^{-1} \times \overline{\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i}}_{\mathbb{B}}^{\top}$ to reconstruct the class-wise embeddings.

The improvement of our work over Soteria is that they just restore a linearly scale embedding $\gamma\nabla\mathbf{W}_i$ while ours is more detailed and precise, where γ is a scale influences by the local training steps ($\gamma = 1$ in FedSGD and $\gamma > 1$ in FedAVG (McMahan et al., 2016)) and $\nabla\mathbf{W}_i$ is a brief notation for $\overline{\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i}}_{\mathbb{B}}$. The reason is that they don't apply Theorem 1 to make use of ∇b_i.

3.2 Instance-wise Label Restoration

If we extend $\nabla\mathbf{z}_i = p_i - y_i$ for single sample to the whole batch, we can get $\sum_k \frac{\partial \mathcal{L}^k}{\partial z_i^k} = \sum_k p_i^k - \sum_k y_i^k$, where $(\cdot)_i^k$ denotes the variable at index i for sample k in a batch of size K. After adjusting the order, it becomes $\sum_k p_i^k - \sum_k \frac{\partial \mathcal{L}^k}{\partial z_i^k} = \sum_k y_i^k$. The right of this equation is exactly the total number of instances of class i in a batch. Let k_j denotes the number of j-class instances, and

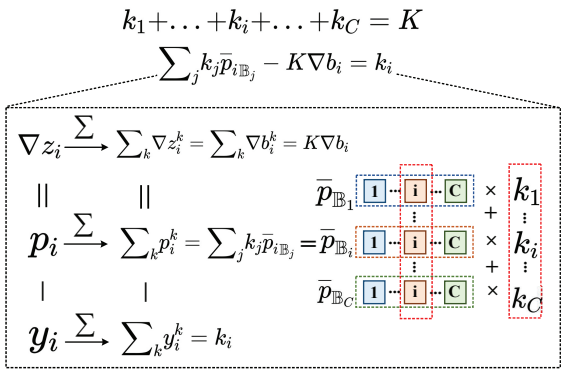

Figure 3: Derivation of the linear equations consisting of gradients, probabilities and labels for NoI per class.

we can disassemble $\sum_k p_i^k$ into $\sum_j k_j \overline{p}_{i\mathbb{B}_j}$. From the class-wise embeddings that was previously deduced, it is possible to recover $\overline{\mathbf{p}}_{\mathbb{B}_j}$, again based on the similarity of data inside a single class.

Approx 3. *The average post-softmax probability from j-class samples by classification model with softmax activation is approximate to that produced by j-class average embedding.*

$$\overline{\mathbf{p}}_{\mathbb{B}_j} = \frac{1}{|\mathbb{B}_j|} \sum_{t \in \mathbb{B}_j} SoftMax(\mathbf{W}\mathbf{e} + \mathbf{b}) \approx SoftMax(\mathbf{W}\overline{\mathbf{e}}_{\mathbb{B}_j} + \mathbf{b}). \tag{4}$$

Additionally, we have $\sum_k \frac{\partial \mathcal{L}^k}{\partial z_i^k} = K \overline{\frac{\partial \mathcal{L}}{\partial z_i}}_{\mathbb{B}} = K \overline{\frac{\partial \mathcal{L}}{\partial b_i}}_{\mathbb{B}} = K\nabla b_i$, where K is the batch size. Therefore, we arrive to an equation:

$$\sum_j k_j \overline{p}_{i\mathbb{B}_j} - K\nabla b_i = k_i, i = 1, ..., C, j = 1, ..., C. \tag{5}$$

This is equivalent to a system of equations as shown in (6) and Fig 3:

$$\begin{cases} k_1 + ... + k_i + ... + k_C = K, \\ \sum_j k_j \overline{p}_{1\mathbb{B}_j} - K\nabla b_1 = k_1, \\ \sum_j k_j \overline{p}_{2\mathbb{B}_j} - K\nabla b_2 = k_2, \\ \qquad \vdots \\ \sum_j k_j \overline{p}_{C\mathbb{B}_j} - K\nabla b_C = k_C. \end{cases} \tag{6}$$

Since the coefficient matrix of this system is not square, we adopt the Moore-Penrose pseudo-inverse algorithm to get a LSS. And the final result also requires filtering of outliers and rounding. If the occurences of the class-wise labels can be obtained in advance, we can further simplify the above equation system through the prior works (Zhao et al., 2020; Yin et al., 2021; Dang et al., 2021), i.e., to discard unknowns k_j and the corresponding equation where there is no j-class sample in the training batch.

4 Experiments

Setups. We evaluate our method for the classification task on three classic image datasets with ascending classes and four models from shallow to deep: (1) a 3-layer FCN (Fully-Connected Network) and other variations on the MNIST dataset with 10-class grayscale images of size 28×28. (2) the 7-layer LeNet-5 on the CIFAR100 dataset with 100-classes color images of size 32×32. (3) the

Table 1: Quantitative Comparison of our label restoration attack with prior works on diverse scenarios. The batch size remains at 24. *: replace the activation function in LeNet-5 from ReLU to Swish (Ramachandran et al., 2017) which is applied in EfficientNet (Tan & Le, 2019); - : this metric is incapable of being calculated. CosSim: the Cosine Similarity for our recovered post-softmax probabilities.

		iDLG		GI		RLG		**Ours**		
Model	**Dataset**	LeAcc	LnAcc	LeAcc	LnAcc	LeAcc	LnAcc	LeAcc	LnAcc	CosSim
FCN-3	MNIST	0.514	-	1.000	-	0.932	-	**1.000**	**0.994**	**0.979**
LeNet-5	CIFAR100	1.000	-	1.000	-	1.000	-	**1.000**	**1.000**	**1.000**
LeNet-S*	CIFAR100	0.150	-	0.213	-	1.000	-	**1.000**	**1.000**	**1.000**
VGG-16	ImageNet	1.000	-	1.000	-	0.981	-	**1.000**	**1.000**	**1.000**
ResNet-50	ImageNet	1.000	-	1.000	-	1.000	-	**1.000**	**1.000**	**1.000**

16-layer VGG-16 and other variations on the large-scale 1000-class ImageNet ILSVRC 2012 dataset (Deng et al., 2009) at 224 ×224 pixels. (4) the 50-layer ResNet-50 and other variations on the ImageNet dataset. We use the training set by default in the following experiments, and the images have been normalized during data loading stage. All statistics are averaged in a randomly selected batch on 50 repetitive tests.

Evaluation metrics. To quantitatively analyze the performance of our label restoration attack, we propose the following two metrics: (1) *Label existence Accuracy—LeAcc*: the accuracy score for predicting label existences; (2) *Label number Accuracy—LnAcc*: the accuracy score for predicting the number of instances per class. Furthermore, in Section 4.3, we adopt the image reconstruction quality score *Peak Signal to Noise Ratio* (*PSNR*) and the perceptual similarity score *Learned Perceptual Image Patch Similarity* (*LPIPS*) (Zhang et al., 2018) to measure the similarity between the restored images and the ground truth.

4.1 Comparison with Prior Works

Attack Baselines. We compare our attack with prior analytic approaches: (1) Improved Deep Leakage from Gradients (iDLG) (Zhao et al., 2020): the single-sample label inference by finding the gradient row whose dot product with other columns yields a negative value; (2) GradInversion (GI) (Yin et al., 2021): the label restoration for the case of single-instance per class by the order of the minimum element in each gradient row; (3) Revealing Labels from Gradients(RLG) (Dang et al., 2021): the extraction of a class-wise label set based on Singular Value Decomposition (SVD) and Linear Programming (LP). Owing to the limitations that iDLG and GI share, we scale them to the real-world settings of multiple instances per class in a batch for comparisons. For iDLG, we alter the selection of the gradient row with the smallest element after summation to all negative gradient rows. And we pick all negative gradients instead of the Top-K minimums in GI, where K is the batch size.

Validity of Probability Reconstruction. The key to our approach is the estimation of the class-wise average probabilities. Our results of the last column in Table 1 show that recovery of the class-wise probability is quite precise and corroborates with the label restoration accuracies. We even achieve a 100% CosSim on both CIFAR100 and ImageNet datasets, and the essential cause behind the fact is that almost each class consists of at most two or three samples in a random-selected batch of size 24.

Results of Label Restoration. Table 1 compares the performance of our proposed iLRG with several prior attack methods. First and foremost, we drill down the ability of the attack from determining the presence of labels to the number of instances for each class while the others are incapable of that. Our method achieves over 99.0% on the whole models and datasets for the two evaluation metrics here. In terms of LeAcc, GI performs best except our iLRG when a non-negative nonlinear activation function is utilized in the model. Nevertheless, once the activation in the network contains a negative element, its performance and that of iDLG will plummet. The essential cause of this fact has been expounded in Section 2.2.

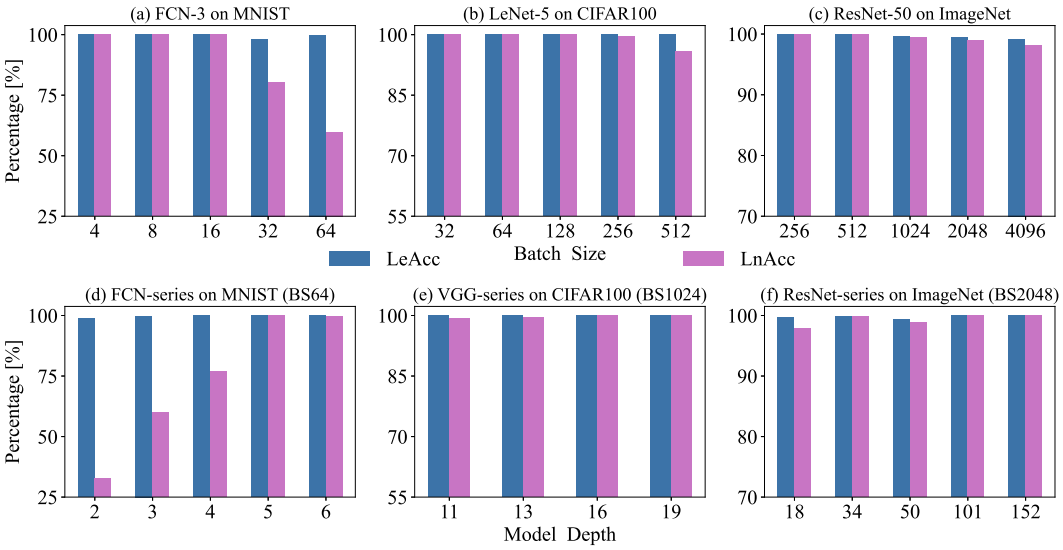

Figure 4: The percentages of LeAcc and LnAcc as batch size and model depth increase under various settings. The FCN series models are distinguished by the number of hidden layers, where a 2-layer model has no hidden layers and the basic FCN-3 consists of a hidden layer (See Appendix B for the FCN-3 network architecture). And BS is short for batchsize.

4.2 Effects of Batch Size and Model Depth

We next discuss the effect of increasing Batch Size and Model Depth/Layer Amount on the results. From Fig 4, we can observe that: (1) the LeAcc score remains above 99% under all experimental settings; (2) as we raise the batch size, the attack capability decreases due to the greater information loss caused by average operation; (3) the depth or complexity of a model likewise plays a strong impact factor. The attack performance increases as the network deepens, which can be seen from the fact that the ResNet-series models are more vulnerable to our attack on large batches than the shallow FCN-series. The results of various networks within the same family also confirm the claim. In Fig 4 (f), it appears to contradict this that the attack against ResNet-50 is inferior to that of ResNet-34. However, since the basic component of Resnet-50 and deeper networks switches from BasicBlock to BottleNeck, the number of parameters of ResNet-50 is actually less than ResNet-34.

4.3 Improved Gradient Inversion Attack with Ours

Gradient inversion attacks perform joint optimization of model inputs and labels, thus labels may shift during optimization, which commonly leads to poor recovery of the inputs. Naturally, it occurs to us that the proposed method can be used to specify an optimization objective for each instance, so as to enhance the existing attacks. Fig 5 illustrates the improvement both visually and numerically. We choose IG as the baseline because it does not require substantial prior constraints.

5 Conclusions

This work proposes *instance-wise Label Restoration from Gradients* (*iLRG*), a method to reveal the cardinal of subsets by category (i.e., instance-wise labels) via shared batch-averaged gradients in FL. We leverage a crucial derivation about the gradient of the network final logits and an approximate reconstruction of class-wise averaged probabilities to build a linear system of equations. Two empirical observations that intra-class high similarity and inter-class low entanglement enable the latter. Our method requires only the gradients of the weight and bias in the final layer, and a *Least Square Solution* (*LSS*) of the *Number of Instances* (*NoI*) per class is derived from the linear equations by *Moore-Penrose pseudoinverse* algorithm. We conduct comprehensive experiments on three classic image datasets with ascending classes and four models from shallow to deep (e.g., FCN on MNIST,

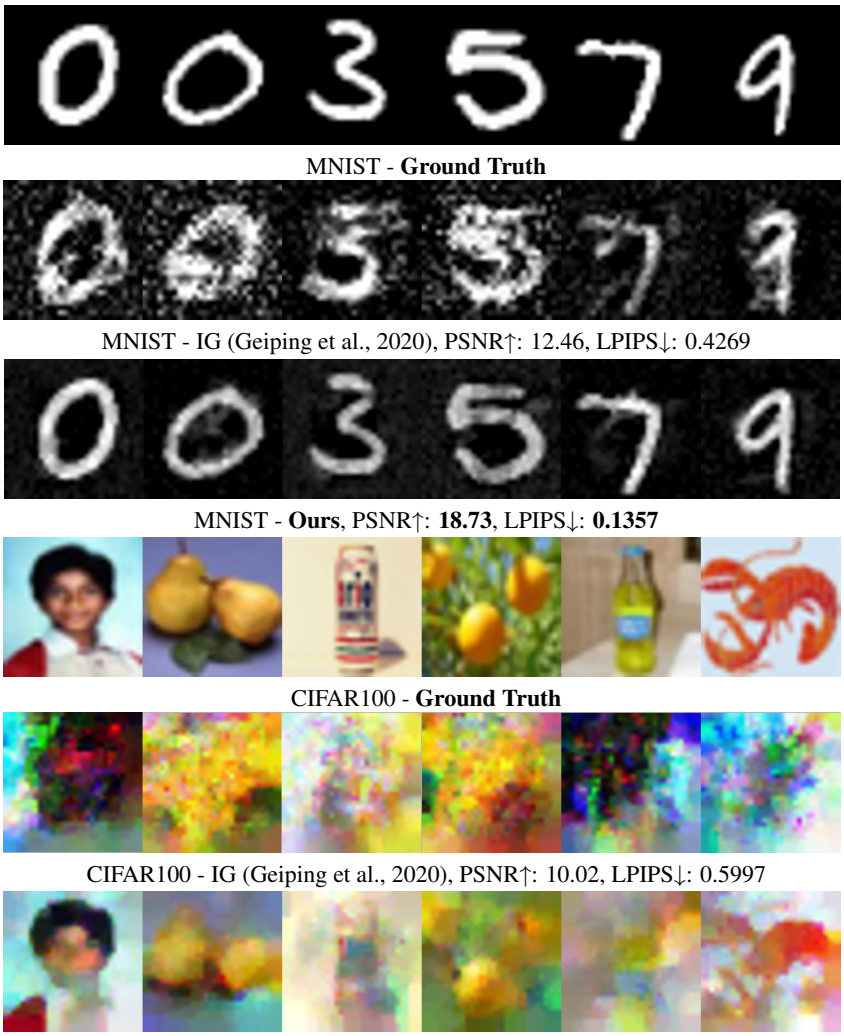

Figure 5: Batch image reconstruction on MNIST (FCN-3, BS50) and CIFAR100 (ResNet-18, BS16) compared with IG. We assign a specific label to each instance after label restoration at 100% accuracy. The 6 best visual images are selected to display and calculate the metrics.

LeNet-5 on CIFAR100, VGG and ResNet on ImageNet, etc) under the various settings of a large batch size up to 4096. The evaluations demonstrate the capability of iLRG with a high proportion of both *Label existence Accuracy* (*LeAcc*) and *Label number Accuracy* (*LnAcc*). It works even on models with an activation function that is not uniformly non-negative. We further facilitate the existing gradient inversion attacks by exploiting the recovered labels. In the future, we seek to explore the application of this attack to more tasks and defenses against it.

REFERENCES

Theodora S Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International journal of medical informatics*, 112:59–67, 2018.

Trung Dang, Om Thakkar, Swaroop Ramaswamy, Rajiv Mathews, Peter Chin, and Françoise Beaufays. Revealing and protecting labels in distributed training. *Advances in Neural Information Processing Systems*, 34:1727–1738, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323. JMLR Workshop and Conference Proceedings, 2011.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Jinwoo Jeon, Kangwook Lee, Sewoong Oh, Jungseul Ok, et al. Gradient inversion with generative image prior. *Advances in Neural Information Processing Systems*, 34:29898–29908, 2021.

Arthur Jochems, Timo M Deist, Johan Van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept. *Radiotherapy and Oncology*, 121(3):459–467, 2016.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Zhuohang Li, Jiaxin Zhang, Luyang Liu, and Jian Liu. Auditing privacy defenses in federated learning via generative gradient leakage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10132–10142, 2022.

Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In *Federated learning*, pp. 240–254. Springer, 2020.

Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, pp. 3. Citeseer, 2013.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

H Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2, 2016.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 691–706. IEEE, 2019.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Adam Sadilek, Luyang Liu, Dung Nguyen, Methun Kamruzzaman, Stylianos Serghiou, Benjamin Rader, Alex Ingerman, Stefan Mellem, Peter Kairouz, Elaine O Nsoesie, et al. Privacy-first health research with federated learning. *NPJ digital medicine*, 4(1):1–8, 2021.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

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

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019a.

Wensi Yang, Yuhang Zhang, Kejiang Ye, Li Li, and Cheng-Zhong Xu. Ffd: A federated learning based method for credit card fraud detection. In *International conference on big data*, pp. 18–32. Springer, 2019b.

Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16337–16346, 2021.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.