**Instance-wise Batch Label Restoration via Gradients in Federated Learning**

*Abstract*— Federated learning, as a distributed collaborative learning framework without sharing local training data, is often considered privacy-preserving. However, recent works have demonstrated that shared gradients can still leak sensitive information. Class-wise labels suffer from the risk of being restored via gradients by joint optimization with model inputs or analytical inference from the last layer of the models.

In this work, we propose a method to reveal instance-wise labels in a training batch for the first time.

For class-wise metric *Label existence Accuracy (LeAcc)*, our attack, without non-negative constraints on activations before the classification layer, achieves comparable performance to prior works. We also present two instance-wise metrics, *Label number Accuracy (LnAcc)* and *Instance label Recall (IRec),* where experimental results reach nearly 100% in several scenarios. Furthermore, we stabilize the optimization process by fixing the labels to improve restoration quality of model inputs.

We show that precise gradients contain a wealth of information to recover private labels, especially for deep networks at the beginning of training.

Keywords—federated learning, privacy-preserving, shared gradients, batch label restoration, instance-wise

# Introduction

Federated Learning (FL) is one of the most popular distributed learning paradigms to achieve privacy protection and has attracted widespread attention[1-3] at present, especially in privacy-sensitive fields such as healthcare[4-5] and finance[6-7]. Since private data does not leave the local device, only model updates or gradients are transmitted between client-server channels, which has long been considered to protect the privacy of participants.

Researches have demonstrated that information leakage from gradients may be more than previously thought in certain cases. Early several attack strategies verify this possibility but with their own limitations. For instance, Membership Inference [8] requires a existing data sample whose presence in the training set is expected to be confirmed. Property Inference [9] retrieves certain attributes with only a label. Model Inversion [10] utilizes the GAN[11] model to generate visual alternatives which look similar but are not the original data. Recently, Zhu et al.[12] propose Deep Leakage from Gradients (DLG), a Gradient Inversion strategy to completely reconstruct the model inputs, which iteratively optimizes a dummy training data by minimizing the distance between corresponding gradients and target ground-truth gradients. For image classification tasks, this approach enables pixel-wise accurate reconstruction and is soon extended to deeper networks and larger-resolution images in a mini-batch [13,14,15,16].

In Gradient Inversion attacks, the labels and model inputs are jointly optimized, but they pay more attention to the reconstruction quality of the inputs, and the recovery

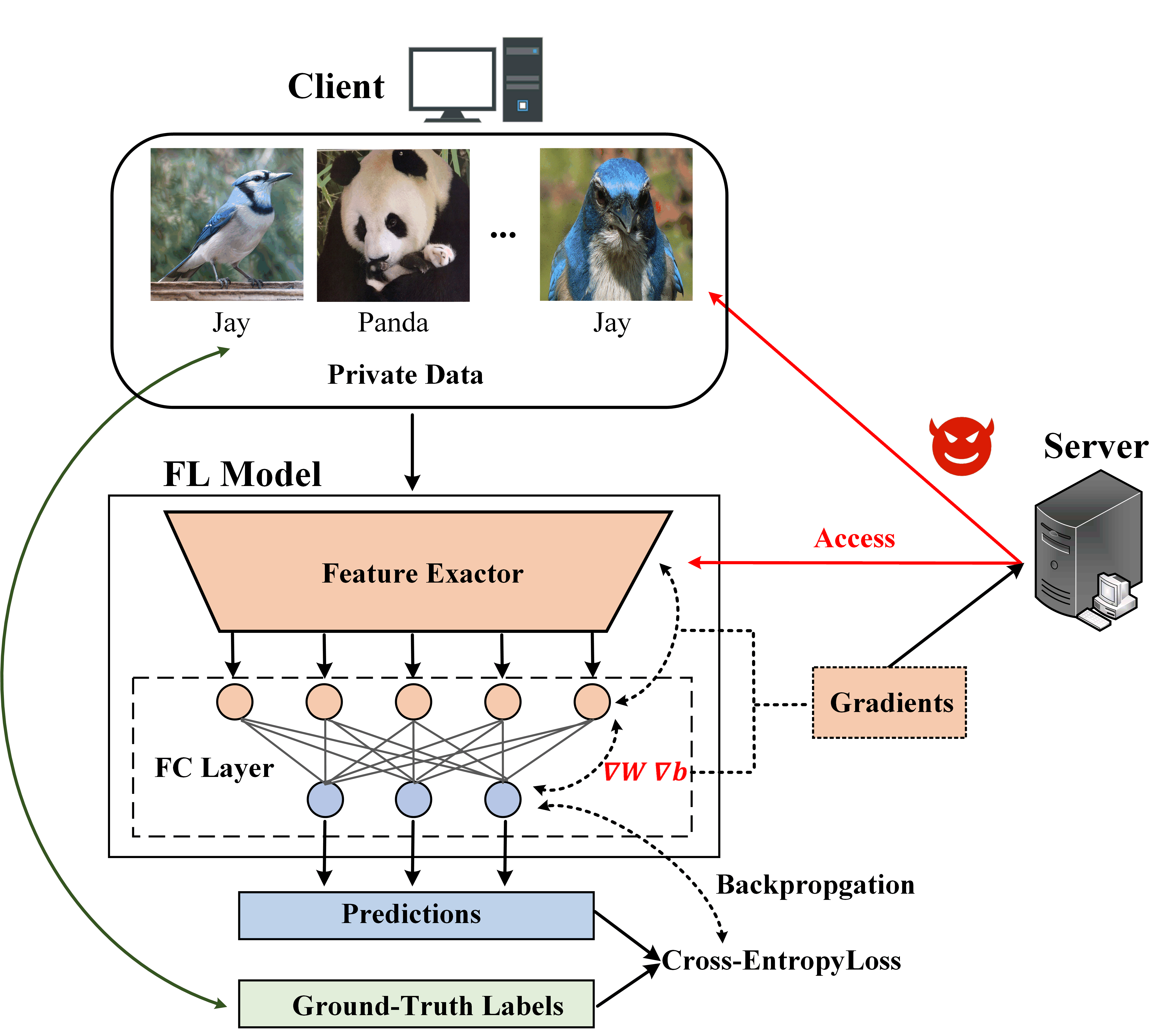


Fig. 1 Threat Model. The honest-but-curious server tries to reveal the client's privacy data from the shared batch-averaged gradient it gets, where the gradients of the weight and bias ,

in the last layer are used to infer labels in our approach .

effect of labels is often not ideal and effective. Zhao et al. [17] first notice that the sign of a gradient may implies the existence of a label, which helps them infer the ground-truth label of a training sample via its gradient analytically. Along this observation, Yin et al. [14] and Dang et al. [18] generalize this method to the recovery of labels for samples in a mini-batch with a high success rate. However, these methods are able to obtain a class-wise label set, but do not know the label of each instance. As shown in Fig. 1, there are multiple samples labeled Jay and we’d like to get its occurrence number, that is, the instance-wise labels. Our work aims to solve this problem.

**Threat Model**: Consistent with the previous gradient leakage attacks [12-18], the adversary we consider is an honest-but-curious server which gets access to the global model and the shared gradients as shown in Fig 1. Only the gradients from the last fully-connected layer marked in red are required to execute our label restoration attack. The approach is applicable to both FedSGD and FedAvg[19,2] usage scenarios, because they are actually equivalent.

Our main contributions are as follows:

1. We propose an analytical batch label restoration method that can infer instance-wise labels through shared gradients for classification model with cross-entropy loss in federated learning, which goes further than prior works that just determine the existence of labels. To reveal the labels, we also obtain class-wise embeddings in the last layer and the post-softmax probabilities as preconditions.
2. Our method performs well on both class-wise metric *LeAcc* and presented instance-wise metrics *LnAcc* & *IRec*, with strong robustness and adaptability to different network architectures as well as different data distributions.
3. We demonstrate that this method poses a greater threat in the early stages of model training, and we use the recovered labels to fix the optimization objective of each instance, thereby improving the effect of a realistic baseline of gradient inversion attack, InvertingGradients (IG)[13].

# Premilinaries

## Problem Formulation

We consider 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 and convolutional neural networks from shallow to deep——LeNet-5[20], VGG-16[21], and ResNet[22], etc.

Given a network with weights ***W*** and the batch-averaged gradient  calculated from a batch of sample-label pairs (**X**, **Y**), we expect to reveal instance-wise labels **Y** via gradients. For each pair () in a batch of size K, we denote the embedding vector into the last layer as , the network final logits as  and the post-softmax probability as , where m is the embedding dimension, C is the number of classes and here is the one-hot representation of the same shape as . In the following sections, and refer to the weight and bias of the final classification layer, respectively. Then we have  and .

## Analytical Label Restorations

Due to the common use of the non-negative activation function, e.g. ReLU[23] and Sigmoid, the malicious adversary can get , has the same sign as , which means we can simply identify the ground-truth label with a negative .

This is exactly what Zhao et al. do. To perform label restoration from batch-averaged gradients, Yin et al. note the magnitude of the negative gradient is often significantly larger than that of the positive one indicating a negative sign can still stand out after averaging operations, and they utilize minimum values, instead of summation along the feature dimension for searching negative signs. On the other hand, Dang et al. attempt to find a classifier to separate the gradient column corresponding to the ground-truth label from the others by linear programming, where all gradient columns are disassembled by a right singular matrix from .

## Single Embedding Reconstruction

In deep neural network architectures, the fully-connected layer is more vulnerable to leakage from gradients for its simple design. Geiping et al. [13] raise the following Theorem 1, realizing the analytical reconstruction of a single input to a fully-connected network.

**Theorem 1.** *For neural networks with a biased fully-connected layer (eg. the last classification layer), and assuming the derivative of the loss w.r.t. to the layer’s output* *contains at least one non-zero value, then the input to the fully-connected layers can be uniquely reconstructed by analytic computation.*

*Proof.* Consider the mapping of a biased full-connected layer without a nonlinear activation, it’s easy to observe that .

Since our assumption guarantees for some index *i*, according to the chain rule, we have:

Therefore, can be calculated exactly as:

# Methodology

In this section we propose a method to restore instance-wise labels in a batch, which we refer to as instance-wise Labels Restoration from Gradients (iLRG).

## Class-wise Embedding Reconstruction

However, Theorem 1 cannot be directly generalized to realize batch embeddings recovery for information loss from the average operation. Two key observations presented by Sun et al. [25], fortunately, provide us a new path: data representations within one class are highly similar, while those from different classes are less entangled.

We first divide a training batch into subsets of C distinct classes, i.e., .Then, we make two approximations by formalizing the aforementioned observations.

**Approx. 1. (High Similarity)** *The embeddings and the derivatives of the loss w.r.t. to their outputs* *are both similar over a certain class of samples in a training batch .*

As a result, the average gradient over can be expressed as (4) with Theorem 1:

where denotes the mean of a variable across .

In simple terms, the latter describes a phenomenon that data representations from different classes tend to be embedded in different gradient rows. In the federated learning setting, the NoC (Number of Classes) within one training batch may be pretty small, thus the entang-lement can be considered low enough.

From it, it can be seen that or its equivalent is only negative at the ground-truth class index c, and , which means the av (absolute value) of the negative gradient at class index is equal to the sum of avs of the other positive gradients. 由于未训练模型，模相差不大，所以可以近似。

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

Hence, we have (7):

Of course, it is necessary to ensure that and here. The actual occurrence of zero is pretty rare, but when it does, we replace it with a small enough number .

We finally get the Formula (8) to infer the class-wise embeddings by combining (4) and (7):

The improvement of our work over Sun et al. [25] is that they just restore a linearly scale data representation  while ours is more detailed and precise, where is a scale influences by the local training steps and is a brief notation for .

## Instance-wise Label Restoration

Let's extend the conclusion that for single sample before to the whole batch:

where denotes the variable at index i for sample k in a batch of size K.

Adjusting the order of Equation (9), we have (10):

The left of Equation (10) is the total number of samples within class i in the batch. Let denotes the number of i-class samples, so we can disassemble according to (11):

Also based on the similarity of data representations within a single class, can be recovered from the class-wise embeddings derived earlier.

**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.*

Besides, we have Equation (13):

Therefore, Equation (11) can be restated as (14) below.

If we spread out Equation (14), the problem of label restoration actually becomes to find the solution of  a linear system of equations.

Since the coefficient matrix of this system is not square, we  get a *LSS (Least Square Solution)* using the *Moore-Penrose pseudo-inverse*. And the final result also requires filtering of outliers and rounding. If the existences of the labels can be obtained in advance, for example, through the prior works [14,17,18], we can further simplify the above equation system, that is, to discard and the corresponding equation where there is no *j-class* sample in the training batch.

Algorithm 1 provides a pseudocode for our complete approach.

|  |
| --- |
| **Algorithm 1** Embedding inference in the last layer and instance-wise label restoration. |
| **Input:** Gradients of weight and bias in the last layer .  **Output:** Class-wise embeddings for each class and the numbers of occurence per-class } in a training batch.  The set of the mean post-softmax probabilities for each class of samples  **for** *i* = 1 **to** *C* **do**  Calculate according to Formula (9)  Feed to the last layer to get  Add into, into  **end for**  Solve the system of linear equations in (16) to obtain  **Return:**  and . |
|  |

# 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) on the MNIST dataset with 10-class grayscale images of size 28×28. (2) the 7-layer LeNet-5[20] on the CIFAR100 dataset with 100-classes color images of size 32×32. (3) the 16-layer VGG-16[21] on the large-scale 1000-class ImageNet ILSVRC 2012 dataset [24] at 224 ×224 pixels. (4) the 50-layer ResNet-50[22] on the ImageNet dataset. We use the training set by default in the following experiments, and all images have been normalized during data loading stage.

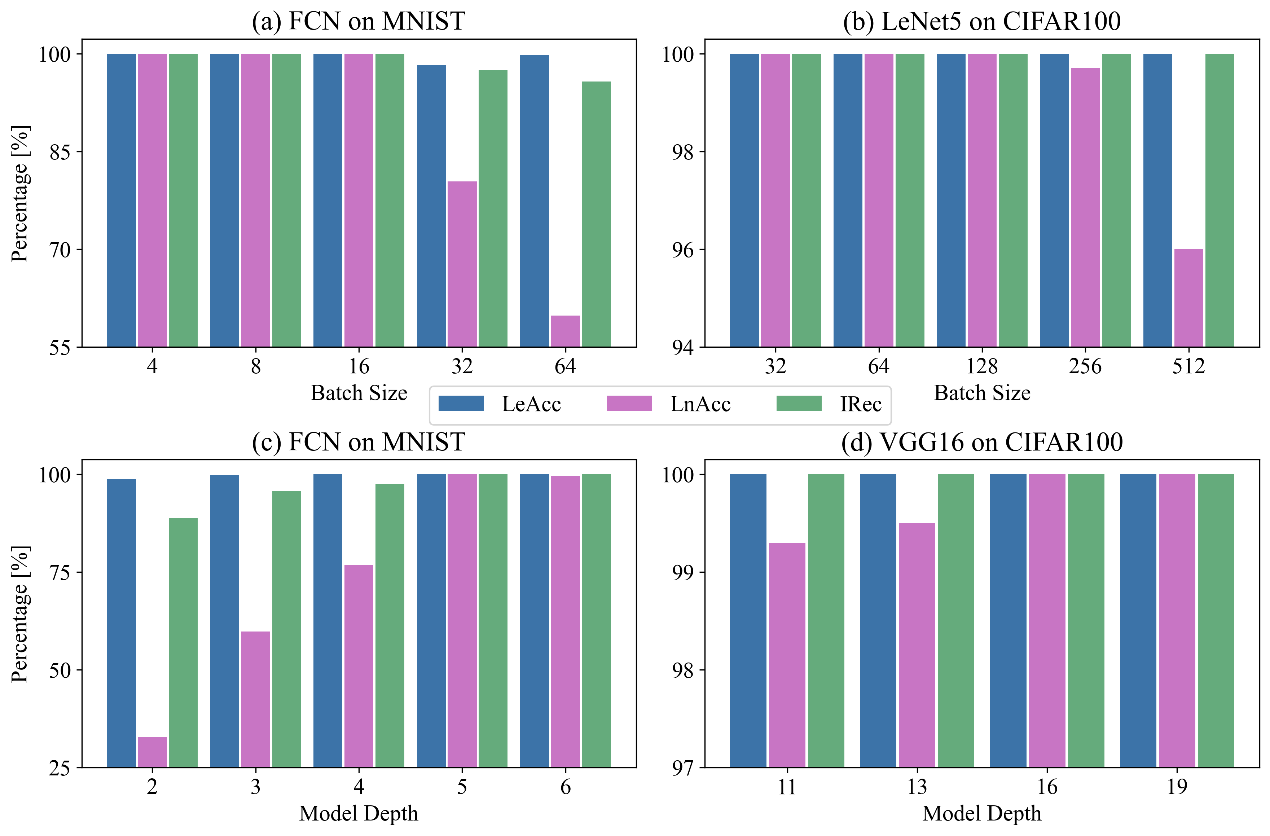
**Evaluation metrics**. To quantitatively analyze the performance of our label restoration attack, we propose the following three metrics: (1) *Label existence Accuracy-LeAcc*: the accuracy score for predicting the occurrence of labels; (2) *Label number Accuracy-LnAcc*: the accuracy score for predicting the number of labels for each class; (3) *Instance-wise label Recall-IRec*: the recall score of labels per instance.

## Comparison with Prior Works

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **iDLG** | **GI** | **RLG** | **Ours** |
| **Model** | **Dataset** | LeAcc | LeAcc | LeAcc | LeAcc |
| FCN-3 | MNIST | 0.500 | 1.000 | 0.932 | **1.000** |
| LeNet-5 | CIFAR-100 | 1.000 | 1.000 | 1.000 | **1.000** |
| VGG-16 | ImageNet | 1.000 | 1.000 | 0.981 | **1.000** |
| ResNet-50 | ImageNet | 1.000 | 1.000 | 1.000 | **1.000** |
| LeNet-5+Swish | CIFAR-100 | 0.089 | 0.213 | 1.000 | **1.000** |

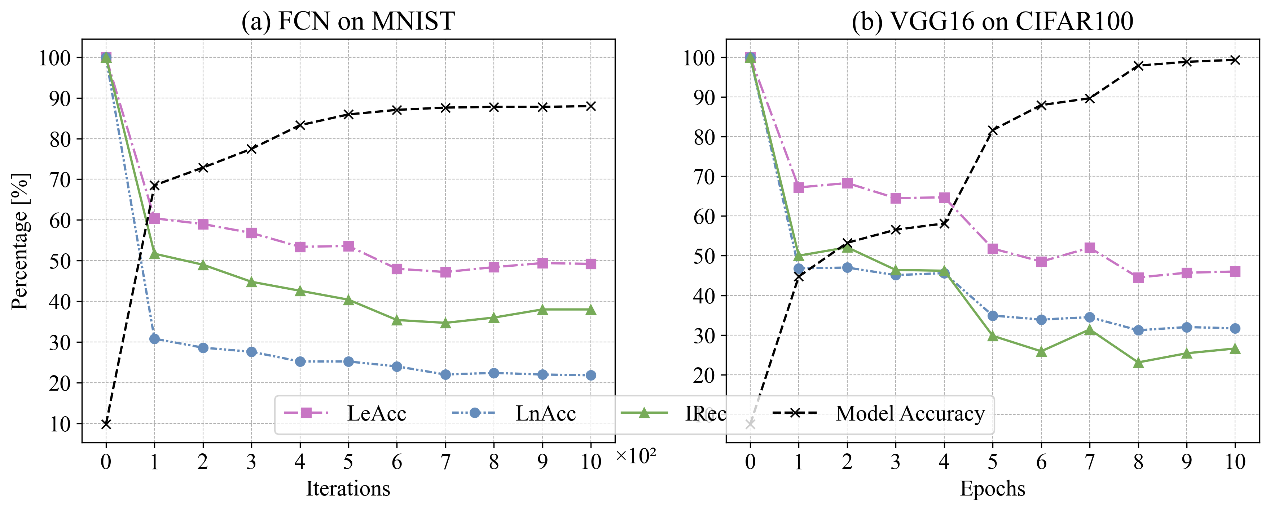
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **iDLG** | | **GI** | | **RLG** | | **Ours** | |
| **Model** | **Dataset** | LnAcc | IRec | LnAcc | IRec | LnAcc | IRec | LnAcc | IRec |
| FCN-3 | MNIST | - | - | - | - | - | - | **0.998** | **1.000** |
| LeNet-5 | CIFAR-100 | - | - | - | - | - | - | **1.000** | **1.000** |
| VGG-16 | ImageNet | - | - | - | - | - | - | **1.000** | **1.000** |
| ResNet-50 | ImageNet | - | - | - | - | - | - | **1.000** | **1.000** |
| LeNet-5+Swish | CIFAR-100 | - | - | - | - | - | - | **1.000** | **1.000** |

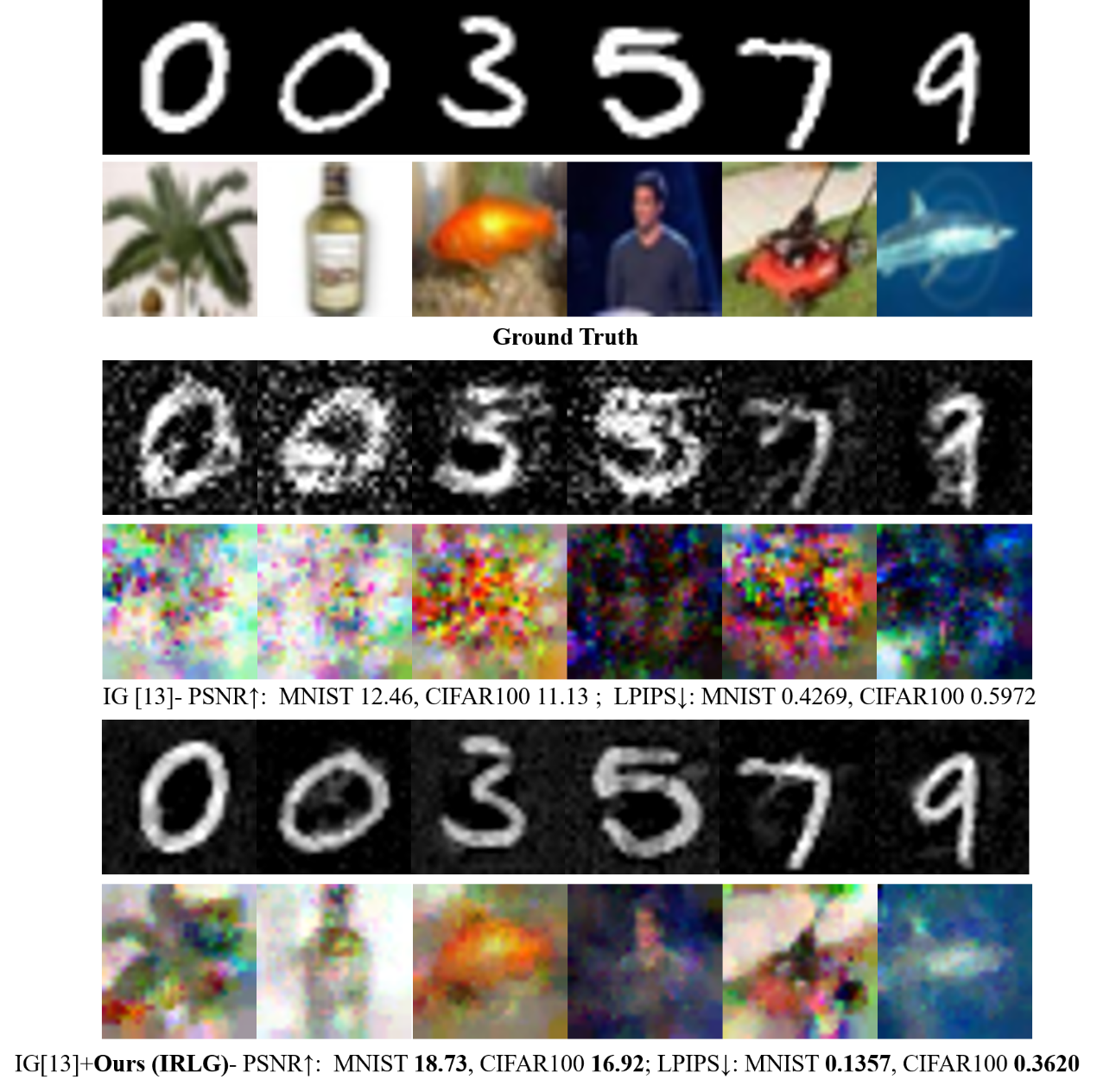
## Effects of Batch Size and Model Depth



## Effects of Batch Size and Model Depth

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Distribution configuration | | LeAcc | LnAcc | IRec |
| Extreme | NoC=1, MRoC=128 | 1.000 | 1.000 | 1.000 |
| Balanced | NoC=2, MRoC=64 | 1.000 | 1.000 | 1.000 |
| NoC=4, MRoC=32 | 1.000 | 1.000 | 1.000 |
| NoC=8, MRoC=16 | 1.000 | 1.000 | 1.000 |
| NoC=16, MRoC=8 | 1.000 | 1.000 | 1.000 |
| NoC=32, MRoC=4 | 1.000 | 1.000 | 1.000 |
| NoC=64, MRoC=2 | 1.000 | 1.000 | 1.000 |
| Unique | NoC=100, MRoC=1 | 0.998 | 0.911 | 0.998 |





# Conclusion

This work proposes Instance-wise Label Restoration from Gradients (I-LRG), an analytical approach to infer batch labels for each instance from shared gradients. The whole process is divided into two steps: inferring class-wise embeddings and then solving a system of linear equations based on the corresponding class-wise probabilities and the gradient of the bias in the last layer. Experimental results show that our attack performs well for an untrained classification model in various scenarios. In addition, combined with existing attacks based on gradients-matching, we significantly improve the image reconstruction quality .We hope the method can draw enough attention to information leakage from gradients.

##### References

1. Jochems, A., Deist, T. M., Van Soest, J., Eble, M., Bulens, P., Coucke, P., ... & Dekker, A. (2016). 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.
2. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.
3. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.
4. Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. C., & Shi, W. (2018). Federated learning of predictive models from federated electronic health records. International journal of medical informatics, 112, 59-67.
5. 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.
6. Wensi Yang, Yuhang Zhang, Kejiang Ye, Li Keqiang, and Cheng Zhong Xu. Ffd: A federated learning based method for credit card fraud detection. In International conference on big data, pages 18–32. Springer, 2019.
7. Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In Federated learning, pages 240–254. Springer, 2020.
8. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP) (pp. 3-18). IEEE.
9. Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019, May). Exploiting unintended feature leakage in collaborative learning. In 2019 IEEE symposium on security and privacy (SP) (pp. 691-706). IEEE.
10. Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). 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).
11. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
12. Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. Advances in neural information processing systems, 32.
13. Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting gradients-how easy is it to break privacy in federated learning?. Advances in Neural Information Processing Systems, 33, 16937-16947.
14. Yin, H., Mallya, A., Vahdat, A., Alvarez, J. M., Kautz, J., & Molchanov, P. (2021). See through gradients: Image batch recovery via gradinversion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16337-16346).
15. Jeon, J., Lee, K., Oh, S., & Ok, J. (2021). Gradient inversion with generative image prior. Advances in Neural Information Processing Systems, 34, 29898-29908.
16. Li, Z., Zhang, J., Liu, L., & Liu, J. (2022). 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).
17. Zhao, B., Mopuri, K. R., & Bilen, H. (2020). idlg: Improved deep leakage from gradients. arXiv preprint arXiv:2001.02610.
18. Dang, T., Thakkar, O., Ramaswamy, S., Mathews, R., Chin, P., & Beaufays, F. (2021). Revealing and protecting labels in distributed training. Advances in Neural Information Processing Systems, 34, 1727-1738.
19. McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2017). Learning differentially private recurrent language models. arXiv preprint arXiv:1710.06963.
20. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
21. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
22. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
23. Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Deep sparse rectifier neural networks." Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011.
24. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.
25. Sun, Jingwei, et al. "Soteria: Provable defense against privacy leakage in federated learning from representation perspective." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.