Instance-wise Batch Label Restoration via Gradients in Federated Learning

Kailang Ma*, Yu Sun*, Jian Cui, Dawei Li, Zhenyu Guan, Jianwei Liu

School of Cyber Science and Technology,

Beihang University,

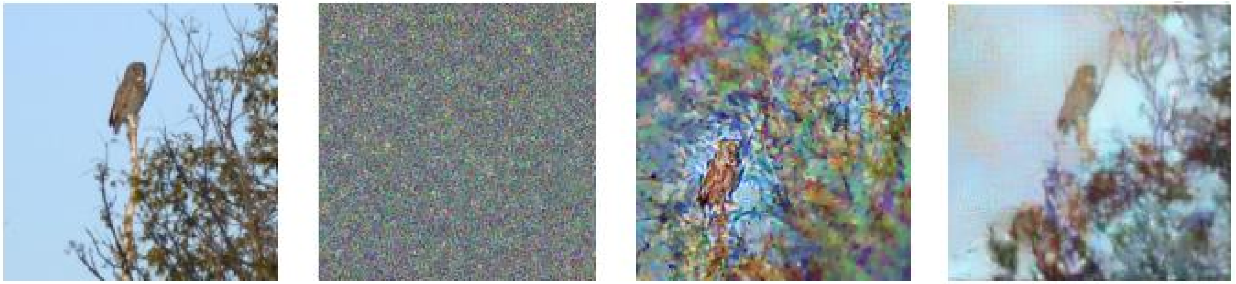
China

ICLR

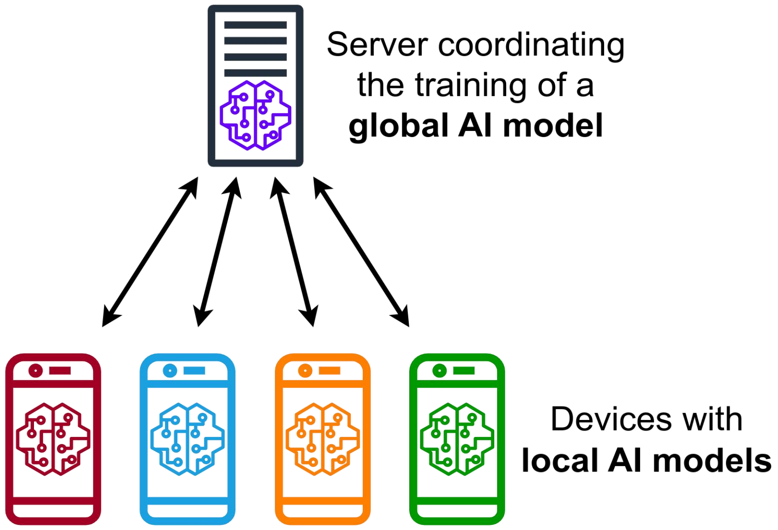
Background & Motivation

- Federated Learning: a distributed learning paradigm, asking participants to exchange **model updates or gradients** instead of raw data.

- Breaking Privacy! **Gradient-Matching strategy** (Gradient Inversion Attack, GIA) allows for reverse-engineer private input from shared gradients. [Zhu et al' 19, Geiping et al '20…]
  - **Label Restoration remains a challenge**.



Original   Zhu et al' 19   Geiping et al '20   Yin et al '21



https://en.wikipedia.org/wiki/Federated_learning

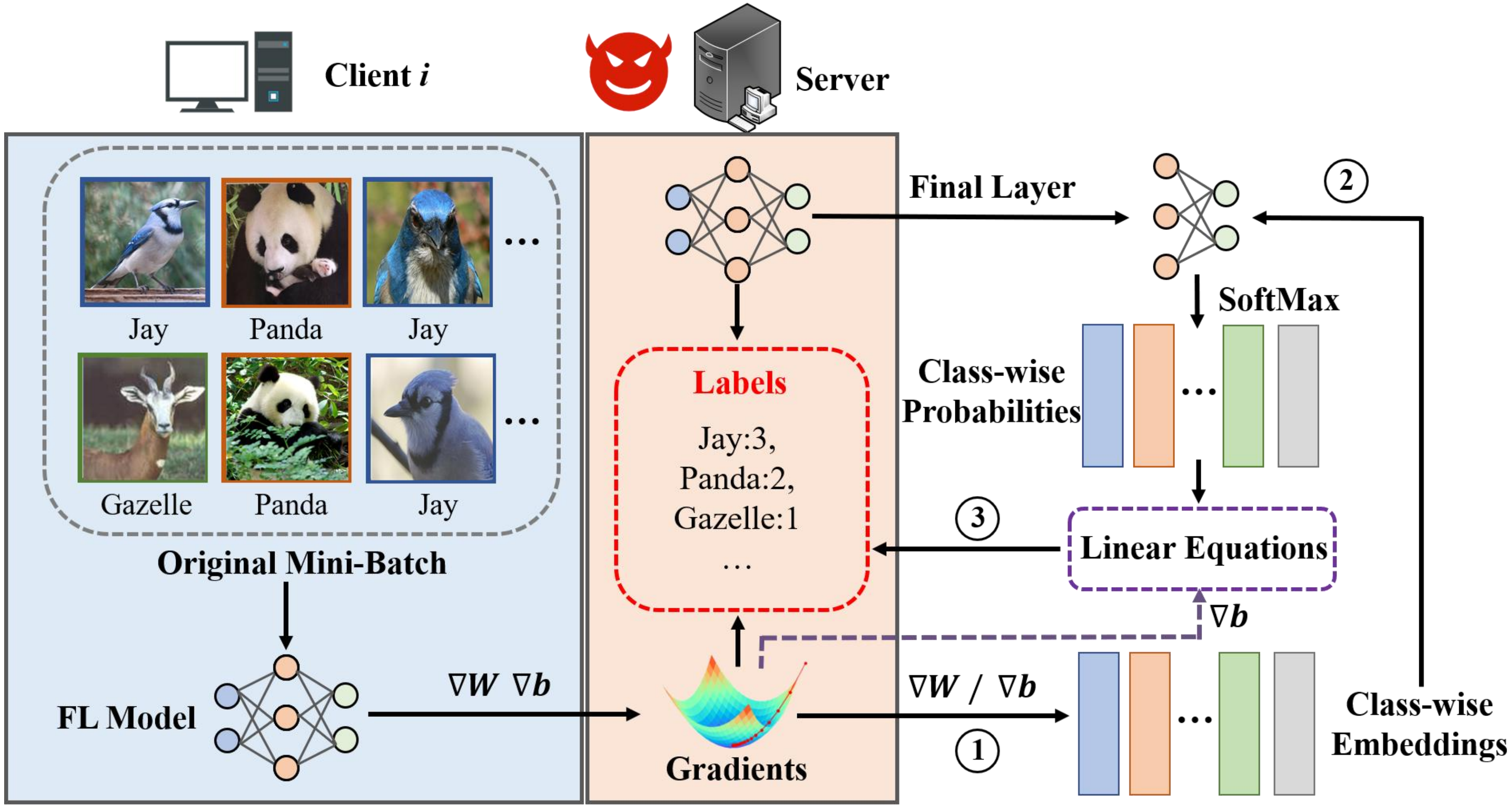
Prior Works

- The optimization-based methods do not perform well [Zhu et al' 19].

- It is possible to extract labels totally by **analytic observation or computation**.
  - **Signs of Gradients** [Zhao et al' 20, for a single sample; Yin et al '21, from batch-averaged gradients]. (Requires Non-negative Activation Functions).
  - **Find a classifier to separate $\boldsymbol{q}^c$** (gradient column *w.r.t.* the ground truth label *c*) from $\boldsymbol{q}^{j \neq c}$ by linear programming [Dang et al'21].
  - Existing methods remains limited to identify the presence of categories (i.e., class-wise label restoration).

→ **Not only Category but also Frequency (i.e., instance-wise label restoration)**

Overview of iLRG



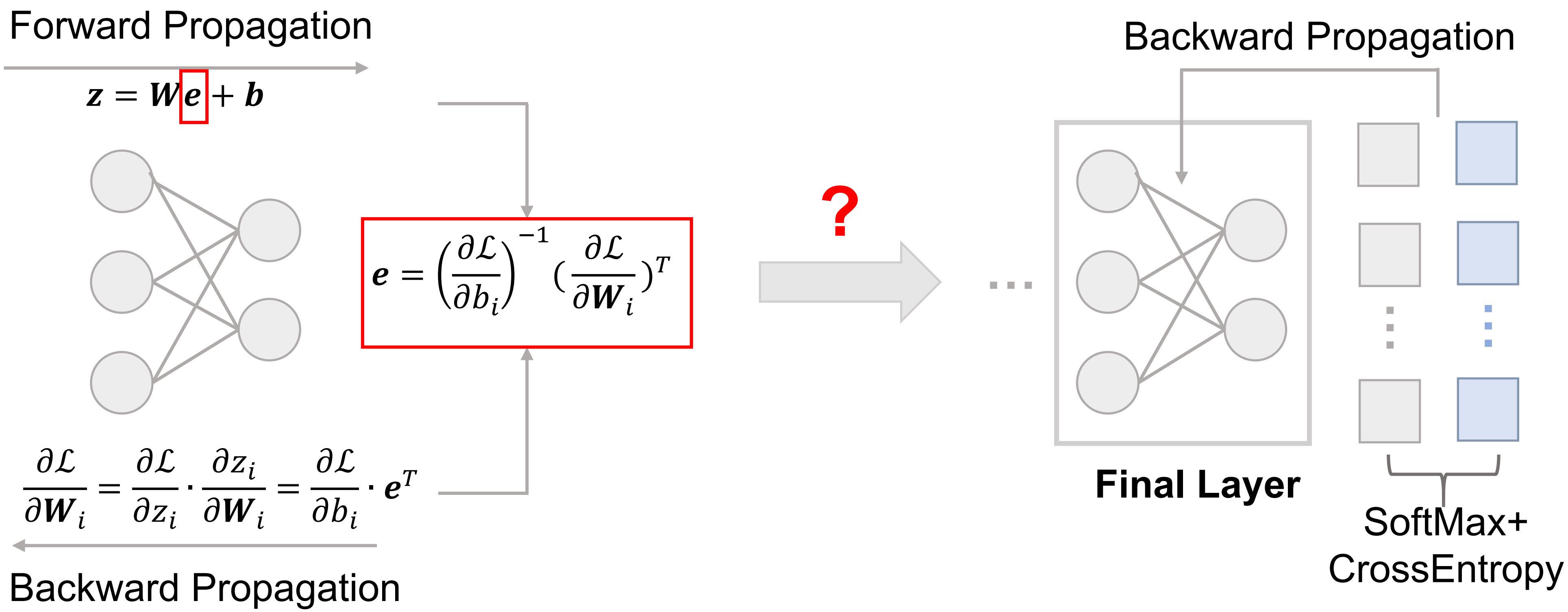
(a) Thread Model (b) Overview of The Proposed Method

Class-wise Embedding Reconstruction

- A provable **single embedding reconstruction** at **any fully-connected layer** has been proposed.

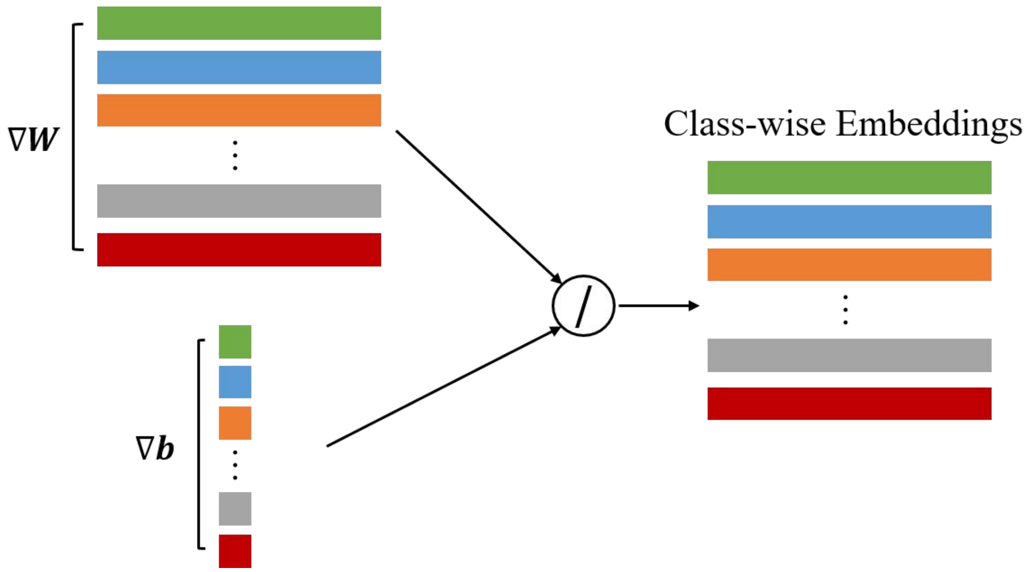
→ Extend to **recover batch embeddings** after average operations?

- ○ **Distinct to Class-Averaged.**



Class-wise Embedding Reconstruction

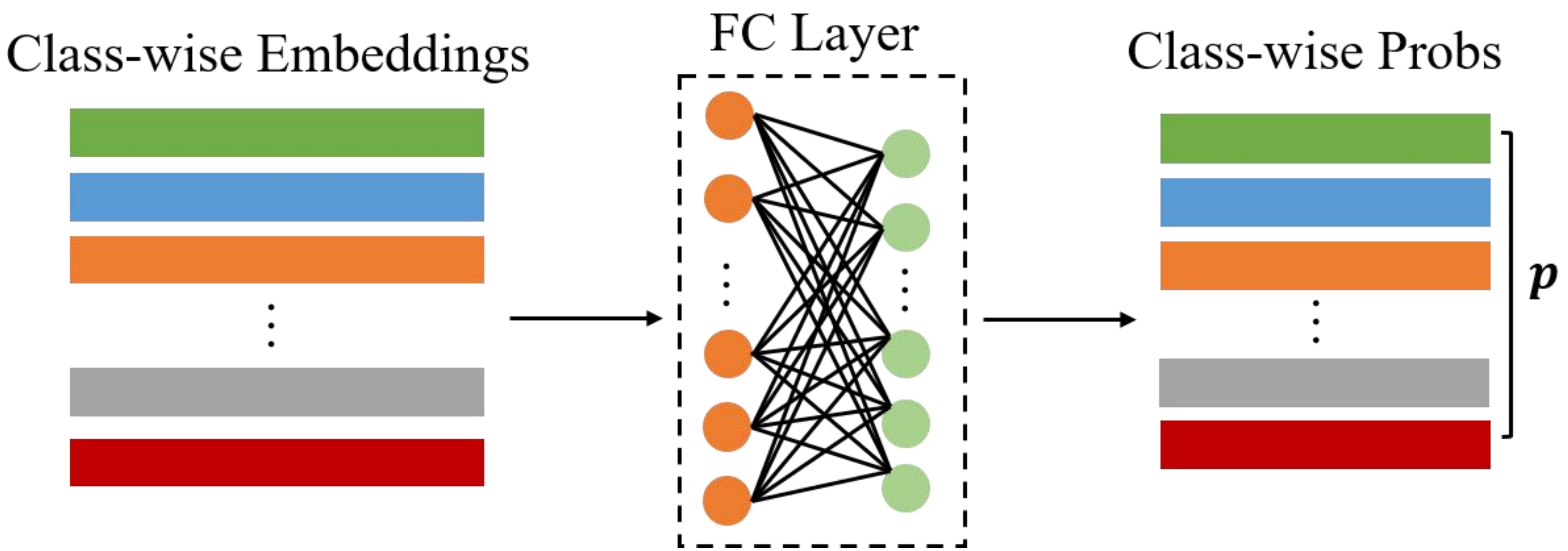
- Two Observations / Approximations (inspired by [Sun et al’ 21]):
  - Approx 1 **Intra-class Uniformity and Concentration of Embedding Distribution**
    - We can replace gradients w.r.t. *i*-class with the arithmetic mean of this category.
  - Approx 2 **Inter-class Low Entanglement of Gradient Contributions**
    - *i*-class samples mainly contribute to the *i*-th gradient row.



$$\boldsymbol{\bar{e}}_{\mathbb{B}_i} \approx \overline{\frac{\partial \mathcal{L}}{\partial b_{i_{\mathbb{B}_i}}}}^{-1} \cdot \overline{\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{i_{\mathbb{B}_i}}}}^{T} \approx \overline{\frac{\partial \mathcal{L}}{\partial b_{i_{\mathbb{B}}}}}^{-1} \cdot \overline{\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{i_{\mathbb{B}}}}}^{T}$$

Class-wise Probability Reconstruction

- Feed embeddings into the layer to obtain the subsequent post-softmax probabilities:
  - Approx 3 **Average Probabilities from Average Embeddings**
    - $\overline{\boldsymbol{p}}_{\mathbb{B}_j} \approx SoftMax(\mathbf{W}\overline{\boldsymbol{e}}_{\mathbb{B}_j} + \boldsymbol{b})$.

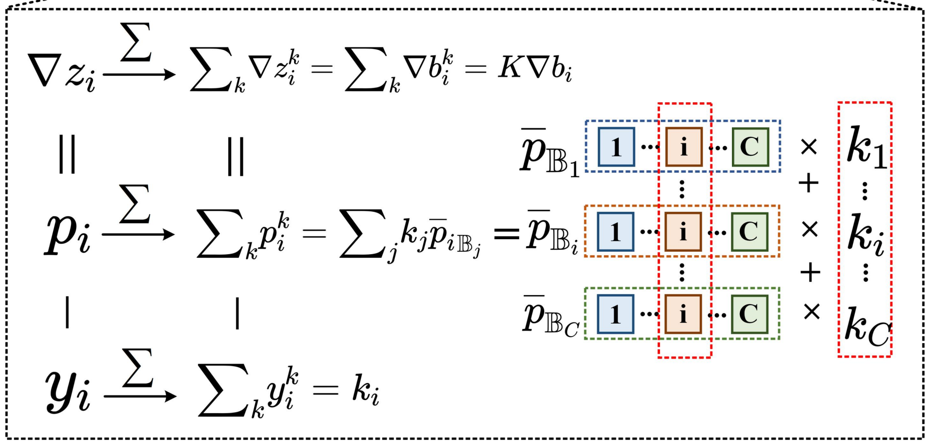


Instance-wise Label Restoration

- Construct **a system of linear equations**:
  - Find out that **gadient w.r.t. model output z with SoftMax and CrossEntropy**
    - $\nabla z_i = p_i - y_i$ (difference between post-softmax probabilities and one-hot labels)
    - Equation composition: ∇b_i、$\overline{\boldsymbol{p}}_{\mathbb{B}_j}$、$K$(batch size)、 ***$k_i$*(label frequency)→ Targets!**

$$k_1 + \ldots + k_i + \ldots + k_C = K$$

$$\sum_j k_j \bar{p}_{i\mathbb{B}_j} - K\nabla b_i = k_i$$



Summation and Decomposition

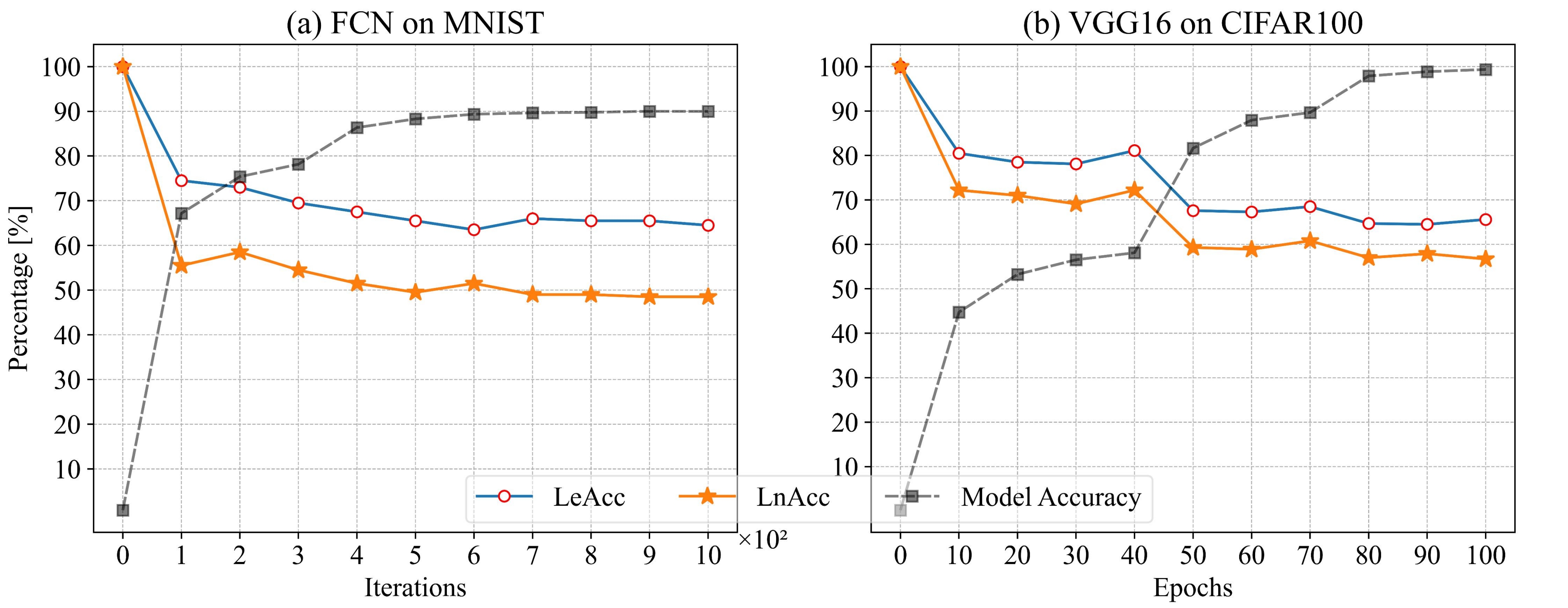
Results on Image Classification

- We show iLRG **outperforms** prior works across several datasets, architectures and evaluation metrics.
  - Last column shows that recovery of class-wise probabilities are quite **precise** and **corroborate** with label restoration accuracies.

		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**

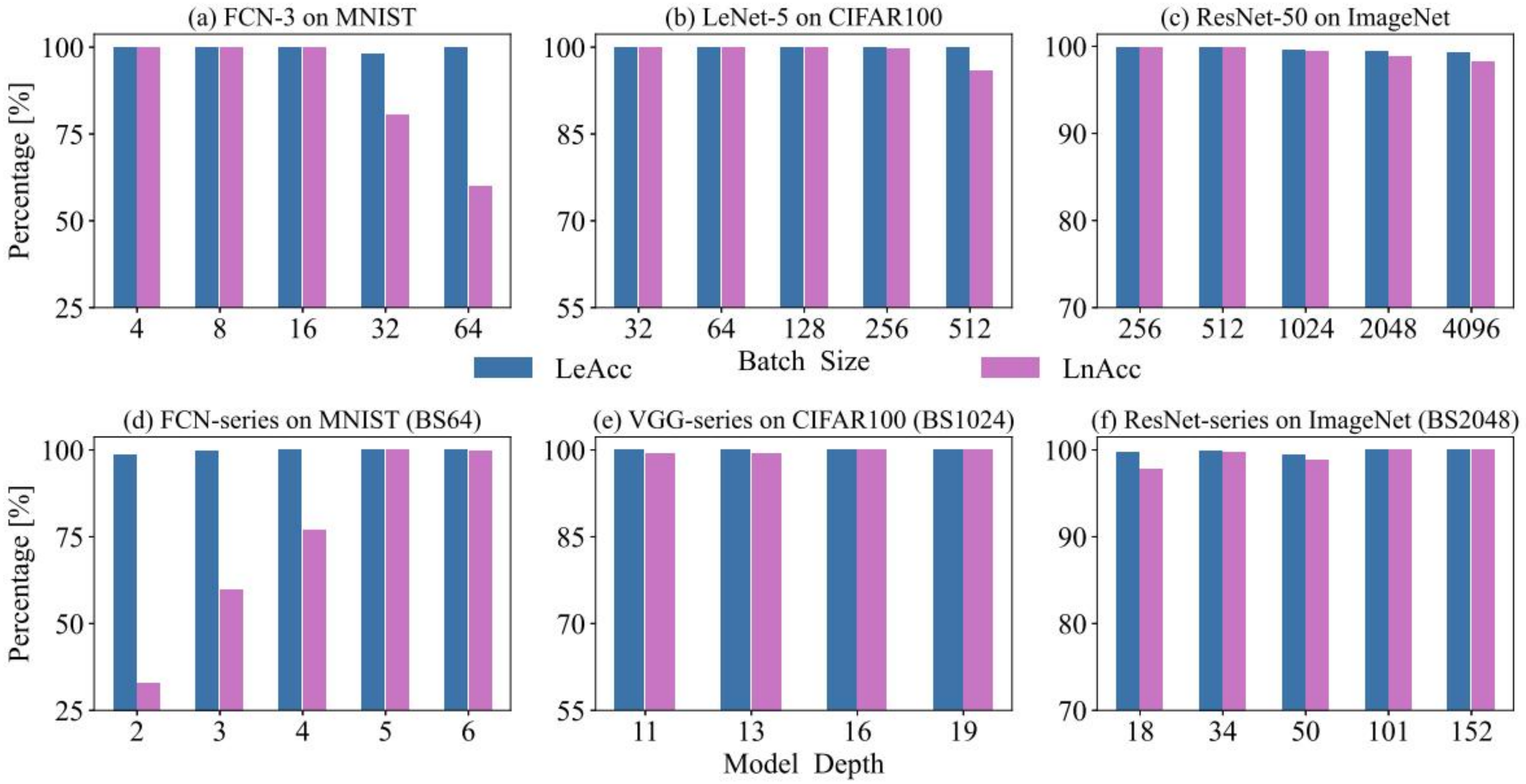
Results on Different Training Stages

- Attack effects **get worse** as training progresses. (performs extremely well for untrained models)
  - **Recommend attacking in the early stages**.



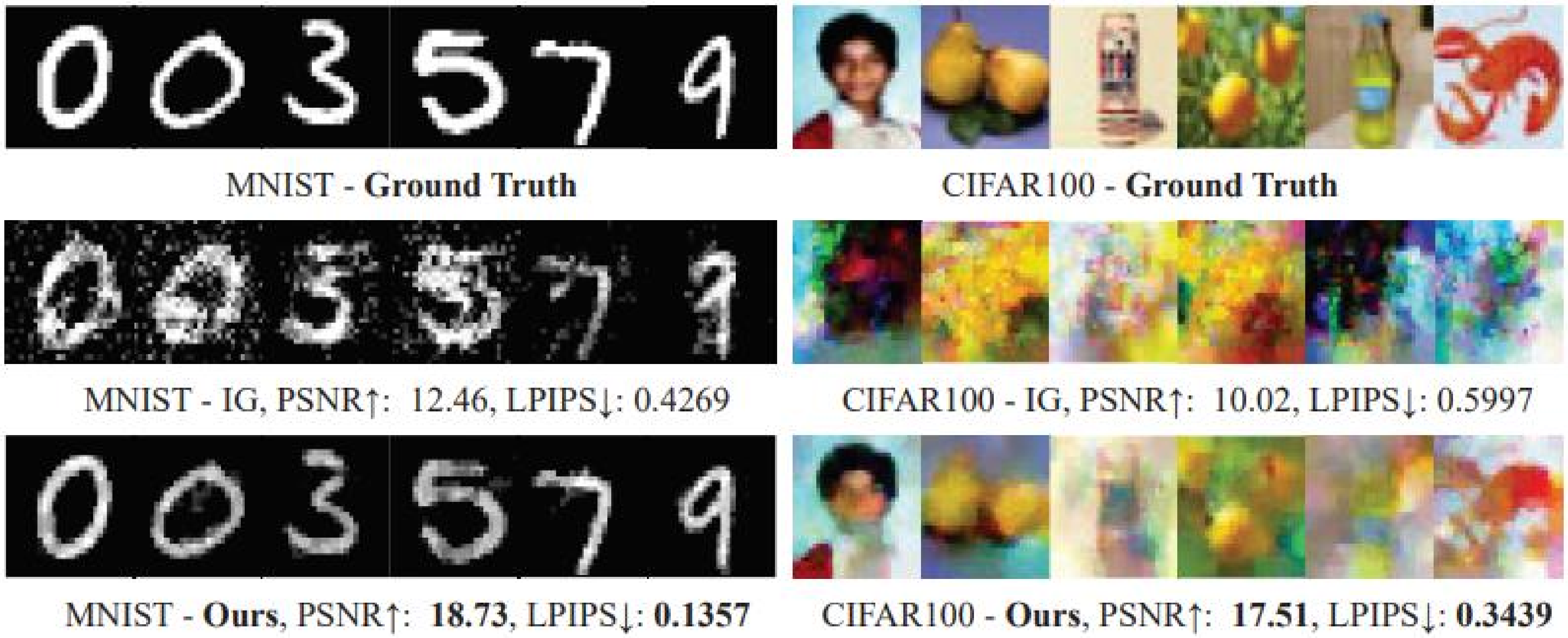
Results on Different Batch Sizes and Model Depths

- As we raise the batch size, the attack capability **decreases**.
- The attack performance **increases** as the network deepens.



Results on Improved Image Recovery with iLRG

- iLRG can be used to specify an optimization objective for each instance, which leads to an improved version of GIA.



Conclusion

- We propose instance-wise Label Restoration from Gradients (iLRG), a method to **reveal instance-wise labels and class-wise embeddings** via shared batch-averaged gradients in FL.

- We conduct a large number of experiments to **demonstrate its effectiveness** and **explore several factors** that may influence it.

- We further **facilitate the existing gradient inversion attacks** by exploiting the recovered labels.

Code: https://github.com/BUAA-CST/iLRG