Foreseeing Reconstruction Quality of Gradient Inversion: An Optimization Perspective

HyeongGwon Hong[1], **Yooshin Cho**[2], **Hanbyel Cho**[2], **Jaesung Ahn**[1], **Junmo Kim**[2]
[1]Kim Jaechul Graduate School of AI, KAIST, Seoul, South Korea
[2]School of Electrical Engineering, KAIST, Daejeon, South Korea
{honggudrnjs, choys95, tlrl4658, jaesung02, junmo.kim}@kaist.ac.kr

Abstract

Gradient inversion attacks can leak data privacy when clients share weight updates with the server in federated learning (FL). Existing studies mainly use L2 or cosine distance as the loss function for gradient matching in the attack. Our empirical investigation shows that the vulnerability ranking varies with the loss function used. Gradient norm, which is commonly used as a vulnerability proxy for gradient inversion attack, cannot explain this as it remains constant regardless of the loss function for gradient matching. In this paper, we propose a loss-aware vulnerability proxy (LAVP) for the first time. LAVP refers to either the maximum or minimum eigenvalue of the Hessian with respect to gradient matching loss at ground truth. This suggestion is based on our theoretical findings regarding the local optimization of the gradient inversion in proximity to the ground truth, which corresponds to the worst case attack scenario. We demonstrate the effectiveness of LAVP on various architectures and datasets, showing its consistent superiority over the gradient norm in capturing sample vulnerabilities. The performance of each proxy is measured in terms of Spearman's rank correlation with respect to several similarity scores. This work will contribute to enhancing FL security against any potential loss functions beyond L2 or cosine distance in the future.

Introduction

Federated learning (FL) is a collaborative machine learning paradigm in which local clients act as trainers and a central server acts as a global aggregator (Konečnỳ et al. 2016; McMahan et al. 2017). Each learning round in FL begins with the server distributing global model weights to participating clients. Then, the clients compute weight updates for the shared global model based on their own data and send these updates back to the server. At the end of the round, the server aggregates all the weight updates received from participating clients for the update of global model.

An important aspect of FL is that participants cannot access the raw data of others, thus their communication is limited to exchanging weight updates. These weight updates were previously believed to reveal minimal information about the original data. However, recent studies (Zhu, Liu, and Han 2019; Zhu and Blaschko 2020; Geiping et al. 2020; Yin et al. 2021; Jeon et al. 2021; Kariyappa et al. 2023; Zhu, Yao, and Blaschko 2023) have challenged this belief regarding data privacy in FL. They have demonstrated the possibility of an honest-but-curious server launching a gradient inversion attack, thereby stealthily recovering clients' data using weight gradients shared from clients.

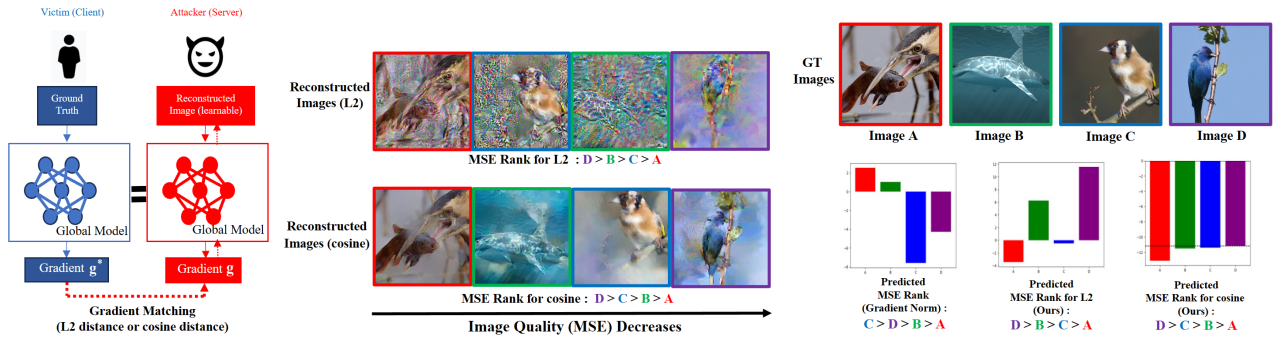
In these attack algorithms, a randomly initialized input variable is optimized to match the current weight gradient computed with itself with the gradient shared from a client. As a loss function for gradient matching, the literature primarily employs either *L2 distance* (Zhu, Liu, and Han 2019; Yin et al. 2021) or *cosine distance* (Geiping et al. 2020; Jeon et al. 2021; Zhu, Yao, and Blaschko 2023) as in Figure 1a.

However, the reconstruction behavior of gradient inversion attack depends on the loss function for gradient matching. In Figure 1b, the L2 distance achieves a more accurate reconstruction for Image C (blue) than for Image B (green), while the cosine distance displays the opposite pattern. The choice of loss function for gradient matching has a significant impact on the vulnerability ranking.

The gradient norm, commonly used as a vulnerability proxy in existing literature (Geiping et al. 2020; Yin et al. 2021), remains constant regardless of the loss function for gradient matching. Thus, it cannot account for the loss function dependence of vulnerability rankings among samples as described in 1c. To address this issue, there is a need for a proxy that can provide a comprehensive explanation for the dependence of vulnerability rankings on the loss function.

In this paper, we introduce a novel loss-aware vulnerability proxy (LAVP) for the first time. In specific, LAVP refers to either the maximum or minimum eigenvalue of the Hessian of gradient matching loss at the ground truth. LAVP is founded on two theorems we have developed concerning gradient matching optimization. We prove that the gradient matching loss drops more significantly when bi-Lipschitz constants of the gradient function are smaller. For simplicity, we focus on the local optimization near the ground truth, representing the worst-case attack scenario. In this case, bi-Lipschitz constants near ground truth correspond to the maximum and minimum eigenvalues of the Hessian at the ground truth, which is how LAVP is derived.

We empirically show the efficacy of LAVP by presenting stronger correlation than the gradient norm, with the quality of reconstructed images from gradient inversion attacks.



(a) An overview of gradient inversion attack

(b) MSE ranks of reconstructed images from gradient inversion attacks with L2 and cosine distances

(c) Predicted MSE ranks with gradient norm and LAVP (ours) for L2 and cosine distances

Figure 1: **Motivation of our work.** (a) A contemporary gradient inversion attack utilizes either L2 or cosine distance for gradient matching. (b) Distinct loss functions reveal different vulnerability rankings among images in Mean Squared Error (MSE). (c) We introduce a loss-aware vulnerability proxy (LAVP), capable of elucidating such loss-specific behaviors. LAVP for L2 and cosine distances predict MSE ranking D $>$ B $>$ C $>$ A and D $>$ C $>$ B $>$ A, respectively. Each predicted ranking coincides with the correct MSE ranking in (b). In contrast, gradient norm, which remains constant regardless of the chosen loss functions cannot explain this.

For both L2 and cosine distances, the vulnerability ranking among samples predicted by LAVP coincides better with the correct one than that predicted by the gradient norm as in Figure 1c. The superiority of LAVP over the gradient norm is consistently verified by experiments on diverse architectures and datasets ranging from low-resolution images in CIFAR-10, CIFAR-100, ImageNette, and ImageWoof, to high-resolution images in ImageNet.

The contribution of our work can be summarized as follows:

- We propose using either the maximum or minimum eigenvalue of the Hessian at the ground truth as a loss-aware vulnerability proxy (LAVP) for the first time.
- We establish several theoretical results regarding the optimization of gradient inversion attacks in close proximity to the ground truth for the derivation of LAVP.
- We demonstrate the efficacy of LAVP in capturing vulnerability against gradient inversion attacks by comparing it to the gradient norm by thorough experiments.
- We propose the geometric mean between LAVP for L2 and cosine distances as the loss-agnostic proxy that caters to both L2 and cosine distances at once.

Preliminaries: Gradient Inversion Attack

Attack Scenario

In a FL scenario, we assume that the server sends the global model $f_w : \mathbb{R}^{b\times d} \to \mathbb{R}^{b\times c}$ to participating clients, where w, b, d, and c denotes model weights, batch size, image size, and the number of classes, respectively. Subsequently, a client computes the weight gradient $g^* = \frac{\partial \mathcal{L}(f_w(x^*),y^*)}{\partial w}$ with respect to the private data batch $(x^*, y^*) \in \mathbb{R}^{b\times d} \times \mathbb{R}^b$ (x^* and y^* being image and label batches) using the objective function $\mathcal{L} : \mathbb{R}^{b\times c} \times \mathbb{R}^b \to \mathbb{R}$. Then, the computed gradient is sent back to the server. In this setup, the server, acting as an honest-but-curious adversary, could attempt to reconstruct an image batch $x \in \mathbb{R}^{b\times d}$ resembling the ground truth batch x^*, using the available information g^* and f_w. *For brevity, we assume $b = 1$ throughout the paper.*

Optimization based Gradient Inversion Attacks

Gradient inversion attacks aim to reconstruct input batch by minimizing the distance between the current gradients and the target gradients as follows:

$$\arg\min_{x,y} \mathcal{L}_{grad}(g_w(x,y), g^*) + \alpha_{prior}\mathcal{R}_{prior}(x), \quad (1)$$

where $g_w(x,y) = \frac{\partial \mathcal{L}(f_w(x),y)}{\partial w}$ represents the weight gradient as a function of the input batch and $\mathcal{L}_{grad} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ (n is the dimension of model weights w) serves as the loss function for gradient matching. Also, $\mathcal{R}_{prior} : \mathbb{R}^{b\times d} \to \mathbb{R}$ denotes the regularization loss for image prior and α_{prior} represents its coefficient. Especially, gradient matching loss function $\mathcal{L}_{grad}$ is chosen to cosine distance ($\mathcal{L}_{grad}(g, g^*) = 1 - \frac{<g,g^*>}{||g||||g^*||}$) (Geiping et al. 2020; Jeon et al. 2021; Huang et al. 2021; Yin et al. 2021; Zhu, Yao, and Blaschko 2023) or L2 distance ($\mathcal{L}_{grad}(g, g^*) = ||g - g^*||_2^2$) (Zhu, Liu, and Han 2019; Zhao, Mopuri, and Bilen 2020; Yin et al. 2021).

Enhanced Assumptions for Stronger Attacks

Beyond the baseline attack, which solely relies on gradient matching, recent gradient inversion attack methods introduce several augmented assumptions for stronger attack.

Firstly, it is assumed that the server knows private labels associated with clients' data. Recent works solve the optimization problem presented in Equation (1) in a sequential manner. This involves initially estimating the labels y directly through g^* and f_w (Ma et al. 2022; Wainakh et al. 2021; Zhao, Mopuri, and Bilen 2020; Yin et al. 2021), followed by exclusive optimization of x using Equation (1),

drawing upon the earlier approximated $y = y^*_{approx}$. This disentangles label estimation from the optimization problem in Equation (1) (Dang et al. 2021; Ye et al. 2022; Li et al. 2022). Consequently, recent studies have predominantly focused on image reconstruction under the premise of private label knowledge (Geiping et al. 2020; Jeon et al. 2021). *We also embrace this assumption in our work.*

Secondly, local batch statistics $\{\mu^*_l, {\sigma^*_l}^2\}_{l=1}^N$ are computed with clients' data batch and then shared with the server alongside weight updates, where μ^*_l, ${\sigma^*_l}^2$, and N signify batch mean, batch variance, and the count of batch normalization (BN) layers, respectively. Utilizing local batch statistics indeed contributes to the reconstruction of high-resolution images (with batch size up to 40) of superior quality (Yin et al. 2021; Hatamizadeh et al. 2022). However, the sharing of batch statistics is not a mandatory requirement for clients, thus *we reject this assumption.*

Related Work: Proxies for the Vulnerability against Gradient Inversion Attack

Gradient Norm. In recent studies (Yin et al. 2021; Geiping et al. 2020), the gradient norm is frequently employed as a heuristic proxy for vulnerability assessment against gradient inversion attacks. This approach is grounded in the intuition that a gradient norm close to zero implies negligible information, hence leading to reconstruction failure. In (Yin et al. 2021), the proposed metric for batch reconstruction, termed Image Identifiability Precision (IIP) is demonstrated on images with higher gradient norms that are perceived as more susceptible examples to gradient inversion attacks. Furthermore, (Geiping et al. 2020) introduces a label flipping attack, which pertains to permuting classifier weights rather than label inversion. To address concerns that a fully trained classifier might yield lower-norm gradients, a threat model is established wherein a malicious server swaps the classifier channel for the correct label with that for any incorrect label. *However, the gradient norm lacks theoretical or empirical foundation as a vulnerability proxy.*

Jacobian Norm. The utilization of the Jacobian norm as a proxy to quantify the extent of input information within gradients was explored in previous work (Mo et al. 2021). Employing usable information theory, the sensitivity, denoted as $E_{\Delta x}[||g_w(x^* + \Delta x) - g^*||]$, is interpreted as an indicator of input information contained within gradients. The sensitivity is reformulated into the Jacobian norm in (Mo et al. 2021), making it the most closely aligned with LAVP (ours) for the L2 distance metric (the maximum eigenvalue of the Jacobian) among preceding proxies. Note that the maximum eigenvalue of the Jacobian corresponds to the spectral norm of the Jacobian. *However, the interpretation of the Jacobian norm fundamentally differs from our perspective.* In (Mo et al. 2021), higher sensitivity of the gradient around the ground truth suggests that the gradient is more likely to be unique within the vicinity of the ground truth, thus making it more susceptible to revealing input information. In contrast, from our optimization viewpoint, a greater gradient sensitivity indicates convergence instability, making optimization of the gradient matching loss more challenging. Indeed, our experimental results align with this intuition. In addition, the objective of (Mo et al. 2021) is to identify layers in which the gradient component significantly encodes input information. We focus on a sample-wise approach rather than a layer-wise approach (i.e., identifying vulnerable examples, not layers). This explains why (Mo et al. 2021) is not regarded as a competitor to our proposed method.

Method

In this section, we present a novel loss-aware vulnerability proxy called LAVP to effectively elucidate loss-specific reconstruction behaviors of gradient inversion attacks. We claim that the susceptibility to the gradient inversion attack is inversely proportional to the bi-Lipschitz constants of the gradient function g_w, denoted as L and M. This claim is backed by our proofs of two theorems regarding L and M respectively. We establish a correspondence of L and M near ground truth to the maximum and minimum eigenvalues of the Hessian with respect to the gradient matching loss $\mathcal{L}_{grad}$ at ground truth x^*. In the end, we outline the methodology for computing both maximum and minimum eigenvalues of Hessian, for both L2 and cosine distances.

Theoretical Results on the Optimization of Gradient Matching

If a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous with constant L, then the following holds: $||f(x) - f(y)|| \leq L||x - y|| \; \forall x, y \in \mathbb{R}^n$. We employ the concept of Lipschitz continuity to prove the following theorem in the context of gradient matching problem. Note that we use $g_w(x)$ instead of $g_w(x, y)$ throughout this section by the aforementioned assumption that label information is available.

Theorem 1. (The first theorem on gradient matching optimization). Suppose $g_w(x)$ is Lipschitz continuous with respect to x with constant L and $\mathcal{L}_{grad}(x) = ||g_w(x) - g^*||_2^2$ is the gradient matching loss. Then, when gradient descent Δx is applied with step size $\mu = \frac{1}{2L^2} > 0$ and $L > \epsilon$ for some $\epsilon > 0$, the following holds:

$$\mathcal{L}_{grad}(x + \mu\Delta x) \leq \mathcal{L}_{grad}(x) - \frac{1}{4L^2}||\frac{\partial \mathcal{L}_{grad}(x)}{\partial x}||_2^2, \quad (2)$$

where $L > \epsilon$ satisfies $||\mu\Delta x|| < \delta$ such that $g_w(x+\mu\Delta x) = g_w(x) + \mu\nabla_x g_w(x)^T\Delta x$ holds approximately.

Proof. See Appendix A1. □

Inequality (2) implies that gradient matching loss strictly decreases as the gradient descent steps unless the gradient term $\frac{\partial \mathcal{L}_{grad}(x)}{\partial x}$ is zero (i.e. gradient matching loss already converges). Then, a gradient descent with a small L can accelerate the convergence of gradient matching optimization. Instead, there is the premise that $L > \epsilon$ for $\epsilon > 0$, which is required for Taylor's first approximation on $g_w(x)$. Therefore, in a particular range of L (i.e., $L > \epsilon$), we hypothesize that a global model with smaller L experiences a sharper loss drop in gradient matching optimization.

In Theorem , optimal loss drop is achieved when $\mu = \frac{1}{2L^2}$. In the proof of Theorem , μ should be the minimizer of