Foreseeing Reconstruction Quality of Gradient Inversion: An Optimization Perspective

Hyeong Gwon 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, instead they communicate through the 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.

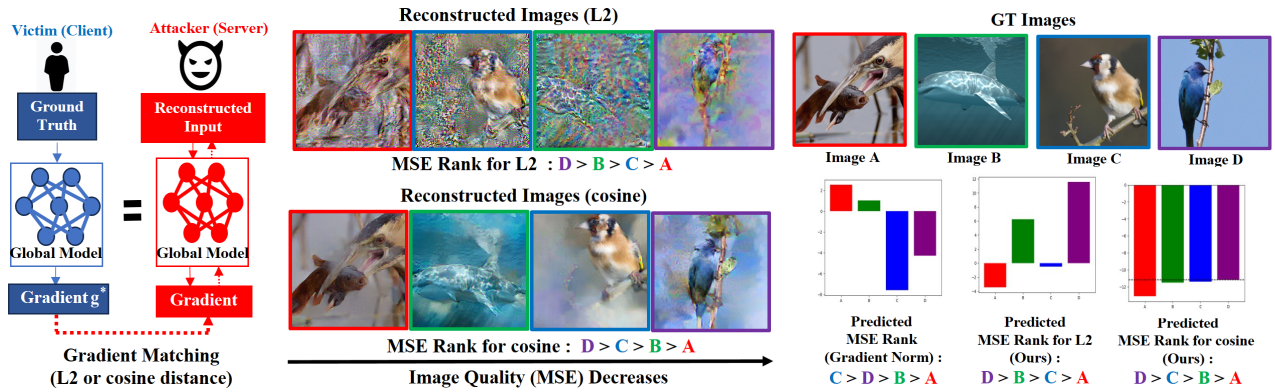
In these attack algorithms, a randomly initialized input variable is optimized to match the current weight gradient computed with itself with the gradient shared from a client. As a loss function for gradient matching, the literature primarily employs either *L2 distance* (Zhu, Liu, and Han 2019; Yin 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 $f_w : \mathbb{R}^{b\times d} \rightarrow \mathbb{R}^{b\times c}$ 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 $\mathcal{L} : \mathbb{R}^{b\times c} \times \mathbb{R}^b \rightarrow \mathbb{R}$. 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 $\mathcal{L}_{grad} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ (n is the dimension of model weights w) serves as the loss function for gradient matching. Also, $\mathcal{R}_{prior} : \mathbb{R}^{b\times d} \rightarrow \mathbb{R}$ 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 1, optimal loss drop is achieved when $\mu = \frac{1}{2L^2}$. In the proof of Theorem 1, μ should be the minimizer of the last term on the right hand for optimal loss drop, while there would be no such restriction if the term was on the left hand. Therefore, if there is a Lipschitz constant for opposite direction denoted by $M > 0$ such that $||g_w(x_1) - g_w(x_2)|| \geq M||x_1 - x_2|| \forall x_1, x_2)$, the following theorem can be derived.

Theorem 2. (The second theorem on gradient matching optimization). Suppose $||g_w(x_1) - g_w(x_2)|| \geq M||x_1 - x_2|| \forall x_1, x_2$ for $M > 0$ holds 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 $\mu < \delta_1$ for some $\delta_1 > 0$, the following holds:

$$\mathcal{L}_{grad}(x + \mu\Delta x) \geq \mathcal{L}_{grad}(x) - \frac{1}{4M^2}||\frac{\partial \mathcal{L}_{grad}(x)}{\partial x}||_2^2, \quad (3)$$

where $\mu < \delta_1$ 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 A2. □

The proof of Theorem 2 is similar to that of Theorem 1 except that the term including μ to be minimized is on the left side, thus there is no restriction like $\mu = \frac{1}{2L^2}$ in Theorem 1, thus more favorable. Inequality (3) implies that the upper bound of gradient matching loss drop is $\frac{1}{4M^2}||\frac{\partial \mathcal{L}_{grad}(x)}{\partial x}||_2^2$, unless the gradient term $\frac{\partial \mathcal{L}_{grad}(x)}{\partial x}$ is zero. Then, a gradient descent with a large M can hinder the convergence of gradient matching optimization. Therefore, we hypothesize that a global model with smaller M has a potential to experience a sharper loss drop in gradient matching optimization.

Finding L and M Near Ground Truth: Maximum and Minimum Eigenvalues of Hessian

Theorems 1 and 2 are about one-step loss drop, so summarizing the whole process of optimization with them is difficult. To mitigate such complexity, we consider the loss drop near ground truth as it is the most important to decide whether ground truth can be reached through optimization or not. For a neighborhood point of of ground truth x^*, $x^* + \Delta x$ ($||\Delta x||$ is very small), gradient matching loss can be approximated by Taylor's second-order approximation like the following:

$$\mathcal{L}_{grad}(x^* + \Delta x) = \mathcal{L}_{grad}(x^*) + \nabla_{x=x^*}\mathcal{L}_{grad}(x)^T\Delta x + \frac{1}{2}\Delta x^T H(x^*)\Delta x,$$

where $H(x^*)$ is the Hessian of gradient matching loss with respect to input variables at x^*. Note that both $\mathcal{L}_{grad}(x^*)$ and $\nabla_{x=x^*}\mathcal{L}_{grad}(x)$ are zero when $\mathcal{L}_{grad}$ is either L2 or cosine distance. Then, $\mathcal{L}_{grad}(x^* + \Delta x)$ can be interpreted as the distance in gradient space while $||\Delta x||^2$ corresponds to distance in input space. Combining two preceding observations, the ratio of gradient distance to input distance is $\mathcal{L}_{grad}(x^* + \Delta x)/||\Delta x||^2 = \frac{1}{2}\frac{\Delta x}{||\Delta x||}^T H(x^*)\frac{\Delta x}{||\Delta x||}$. Then, the upper and lower bounds of this ratio correspond to maximum and minimum eigenvalues of Hessian, respectively. In proximity to ground turth, we can replace L and M with maximum and minimum eigenvalues of Hessian at ground truth, which is our proposed proxy, LAVP.

Hessian of Gradient Matching Loss

To compute LAVP, Hessian should be identified first. In Theorems 4 and 5, we derive the Hessian for L2 and cosine distances in a closed form respectively.

Theorem 3. (Hessian at ground truth for L2 distance). Suppose $\mathcal{L}_{grad}$ is L2 distance, then the Hessian at ground truth is like the following:

$$H_{\text{L2}}(x^*) = J(x^*)^T J(x^*), \quad (4)$$

where $J(x^*) = \nabla_{x=x^*} g_w(x)$ is the Jacobian of gradient function $g_w(x)$ with respect to input at ground truth x^*.

Proof. See Appendix A3. □

When $\mathcal{L}_{grad}$ is L2 distance, $H_{\text{L2}}(x^*) = J(x^*)^T J(^*)$ holds by Theorem 3, thus positive semi-definite. since input dimension is smaller than weight dimension in general, $H_{\text{L2}}(x^*)$ is not trivial low rank and its minimum eigenvalue has a potential to be positive.

For cosine distance, Hessian at ground truth can be solved in closed form by the following theorem.

Theorem 4. (Hessian at ground truth for cosine distance). Suppose $\mathcal{L}_{grad}$ is cosine distance, then the Hessian at ground truth is like the following:

$$H_{\cos}(x^*) = \frac{1}{||g^*||^2} J(x^*)^T (I - \frac{g^*}{||g^*||}\frac{g^{*T}}{||g^*||}) J(x^*), \quad (5)$$

where I is the identity matrix.

Proof. See Appendix A4. □

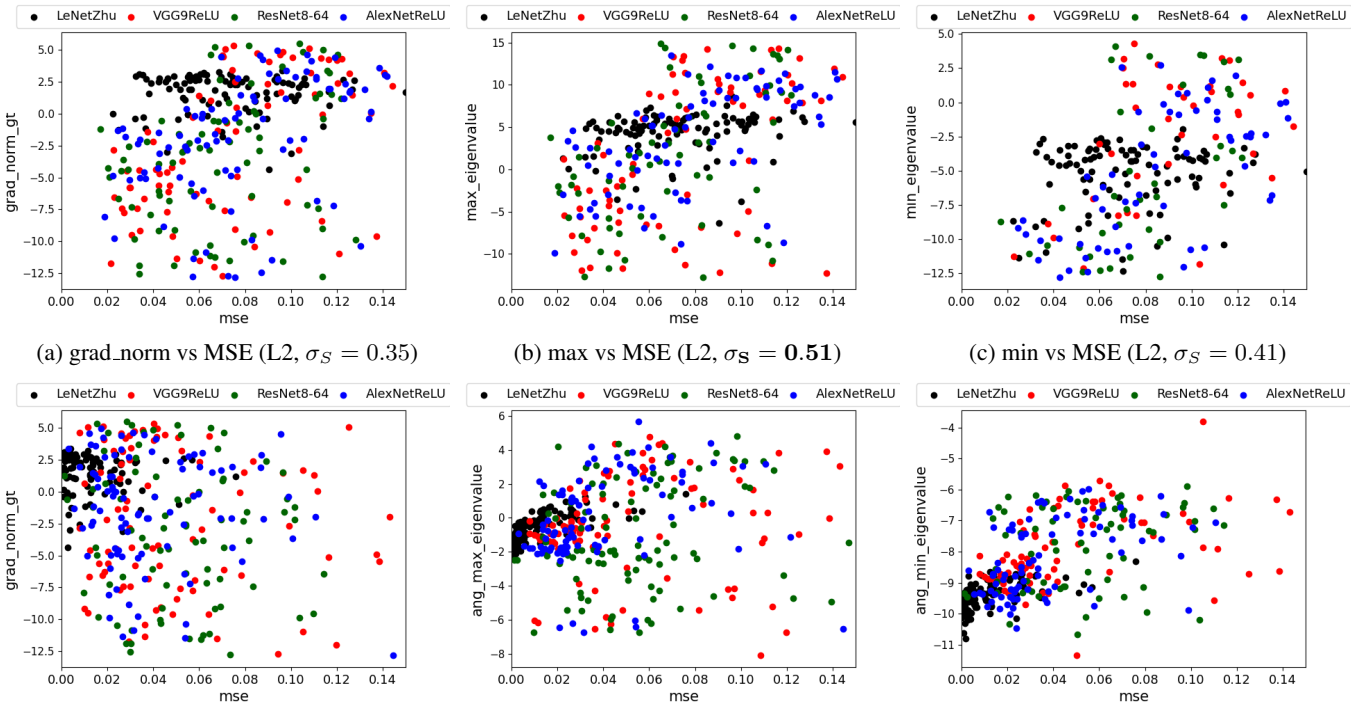
The minimum eigenvalue of $H_{\cos}(x^*)$ is nonnegative as it is positive semi-definite by Cauchy-Schwartz inequality.

Implementation of LAVP

To find the maximum eigenvalue of Hessian, power iteration is used. Power iteration computes matrix-vector product and normalization alternatively until the vector converges to the eigenvector with the maximum eigenvalue. When this algorithm is applied to the Hessian, Jacobian-vector product is inevitable, while *Autograd* package in PyTorch supports only vector-Jacobian product. Therefore, Jacobian-vector product is solved with the finite difference method with very small step size. Once the maximum eigenvalue α_{max} is obtained for the Hessian $H(x^*)$, then power iteration is applied to $\alpha_{max} I - H(x^*)$ (I is the identity matrix) for identifying the minimum eigenvalue α_{min}, as $\alpha_{max} - \alpha_{min}$ would be the maximum eigenvalue of $\alpha_{max} I - H(x^*)$ (I is the identity matrix). *It is noteworthy that multiple Hessian-vector products, rather than the entire Hessian, are sufficient for computing LAVP, thus more efficient.*

σ_S	grad_norm	max (LAVP for L2)	min (LAVP for L2)	ang_max (LAVP for CS)	ang_min (LAVP for CS)
C-10+L2	.35 / -.27 / -.35 / -.13	**.51** / **-.46** / **-.51** / -.15	.41 / -.40 / -.41 / **-.20**	-.06 / -.01 / .06 / -.06	-.04 / -.07 / .04 / -.06
C-100+L2	.41 / -.31 / -.41 / -.19	**.46** / **-.57** / **-.46** / .10	.41 / -.45 / -.41 / -.01	-.05 / -.18 / .05 / **.22**	-.05 / -.20 / .05 / .19
IN+L2	.03 / .03 / -.03 / .19	.33 / -.26 / -.33 / **.49**	**.34** / -.25 / **-.35** / .46	.14 / -.20 / -.14 / .42	.25 / **-.34** / -.25 / .42
IW+L2	.35 / -.01 / -.35 / .00	.66 / **-.58** / -.66 / .46	**.68** / -.52 / **-.68** / .29	.38 / -.41 / -.38 / .46	.46 / -.52 / -.46 / **.49**
C-10+CS	-.28 / .25 / .28 / -.18	-.03 / -.02 / .03 / .00	.04 / -.08 / -.04 / .02	.26 / -.31 / -.26 / .36	**.64** / **-.74** / **-.64** / **.62**
C-100+CS	.10 / -.07 / -.10 / .02	.37 / -.35 / -.37 / .26	.32 / -.34 / -.32 / .24	.67 / -.80 / -.67 / **.68**	**.69** / **-.81** / **-.69** / .65
IN+CS	-.13 / .11 / .14 / -.18	.16 / -.16 / -.16 / .12	.27 / -.28 / -.25 / .17	.57 / -.65 / -.57 / .64	**.75** / **-.80** / **-.75** / **.73**
IW+CS	.00 / .01 / .00 / -.04	.37 / -.37 / -.37 / .21	.33 / -.34 / -.33 / .18	.72 / -.73 / -.72 / **.61**	**.75** / **-.83** / **-.75** / .61

Table 1: Spearman's correlation (σ_S) of proxy candidates with image similarity scores (MSE(↓) / SSIM(↑) / PSNR(↑) / LPIPS(↓)) on low-resolution images. 'C-10', 'C-100', 'IN', and 'IW' denote CIFAR-10, CIFAR-100, ImageNette, and ImageWoof respectively. 'L2' and 'CS' denote L2 and cosine distances respectively. 'grad_norm', 'max', 'min', 'ang_max', and 'ang_min' denote the gradient norm, the maximum eigenvalue of Hessian for L2, the minimum eigenvalue of Hessian for L2, the maximum eigenvalue of Hessian for CS, and the minimum eigenvalue of Hessian for CS.



(a) grad_norm vs MSE (L2, $\sigma_S = 0.35$) (b) max vs MSE (L2, $\sigma_S = \mathbf{0.51}$) (c) min vs MSE (L2, $\sigma_S = 0.41$)

(d) grad_norm vs MSE (CS, $\sigma_S = -0.28$) (e) ang_max vs MSE (CS, $\sigma_S = 0.26$) (f) ang_min vs MSE (CS, $\sigma_S = \mathbf{0.64}$))

Figure 2: Comparison of the gradient norm, maximum and minimum eigenvalues of Hessian in terms of the correlation with MSE of reconstructed samples over several architectures on CIFAR10 test samples.

Experimental Results

In this section, we begin with a concise overview of our experimental setup. Then, we elucidate the advantages of LAVP over the gradient norm (baseline), in explaining the vulnerability to the gradient inversion attack with either L2 or cosine distance by providing correlation tables and plots. For a black-box scenario where the attacker's loss function is unknown to clients, we also introduce the loss-agnostic LAVP fusion, the proxy that can handle several candidate loss functions at once. An example of LAVP fusion includes the geometric mean between LAVPs for L2 and cosine distances. We provide the correlation table for this example with a comparative evaluation against the gradient norm.

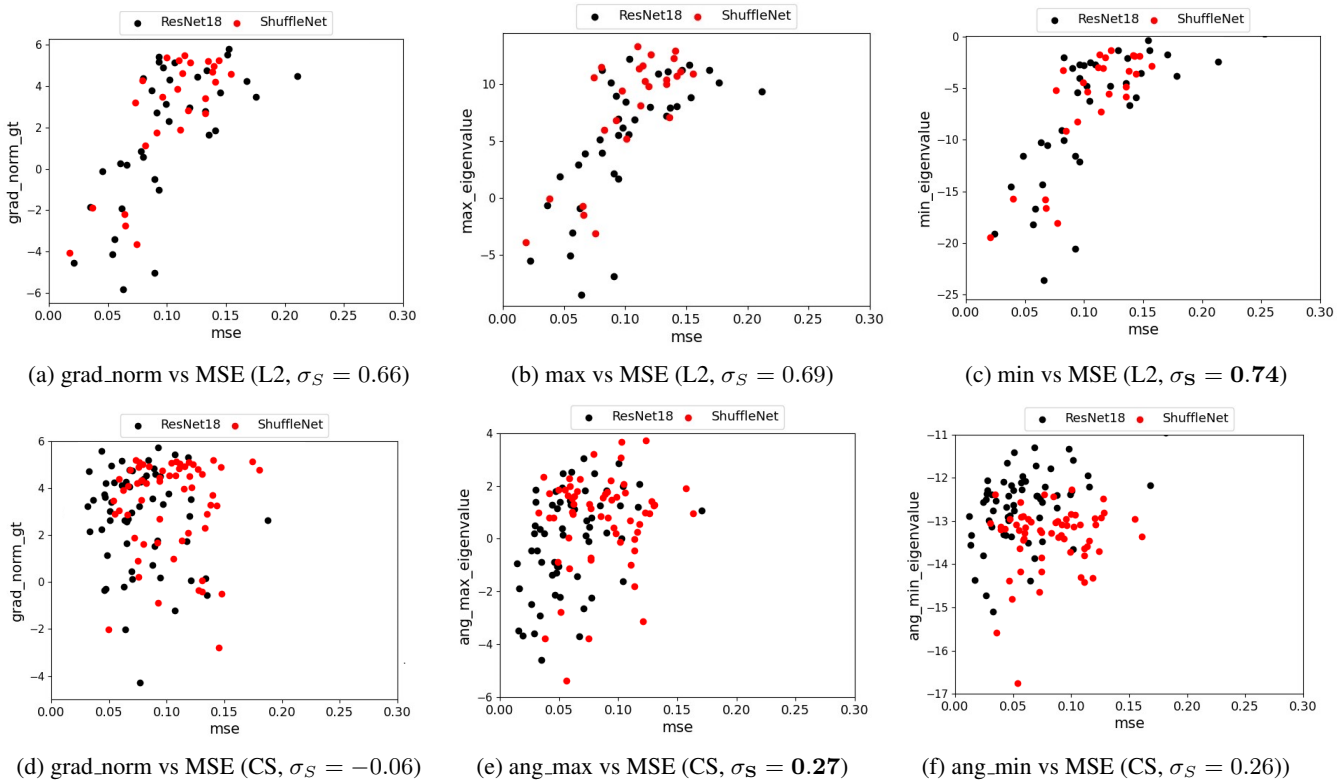
Experimental Setup

Datasets. We conducted an evaluation by randomly selecting 100 validation images from CIFAR-10 (Krizhevsky 2009), CIFAR-100 (Krizhevsky 2009), ImageNette (Howard 2019), ImageWoof (Howard 2019), and ImageNet (Deng et al. 2009). Notably, ImageNette and ImageWoof are subsets of ImageNet (Deng et al. 2009), each consisting of ten easily classified classes, but with different classes from one another.

Architectures and Attack Hyperparameters. We evaluated several deep learning models on low-resolution images, in-

σ_S	grad_norm	max (LAVP for L2)	min (LAVP for L2)	ang_max (LAVP for CS)	ang_min (LAVP for CS)
ImageNet+L2	.66 / -.66 / -.66 / **-.28**	.69 / -.72 / -.69 / -.09	**.74** / **-.74** / **-.72** / -.07	-.05 / .09 / .05 / .03	-.07 / .12 / .07 / .10
ImageNet+CS	-.06 / -.04 / .00 / -.05	.02 / -.07 / -.11 / .04	-.03 / -.05 / -.06 / .00	**.27** / **-.37** / -.21 / .32	.26 / -.22 / **-.24** / **.32**

Table 2: Spearman's correlation of proxy candidates with image similarity scores (MSE(↓) / SSIM(↑) / PSNR(↑) / LPIPS(↓)) on ImageNet.



(a) grad_norm vs MSE (L2, $\sigma_S = 0.66$) (b) max vs MSE (L2, $\sigma_S = 0.69$) (c) min vs MSE (L2, $\sigma_\mathbf{S} = \mathbf{0.74}$)

(d) grad_norm vs MSE (CS, $\sigma_S = -0.06$) (e) ang_max vs MSE (CS, $\sigma_\mathbf{S} = \mathbf{0.27}$) (f) ang_min vs MSE (CS, $\sigma_S = 0.26$))

Figure 3: Comparison of the gradient norm, maximum and minimum eigenvalues of Hessian in terms of the correlation with MSE of reconstructed samples over ResNet18 and ShuffleNet models on ImageNet validation samples.

cluding LeNet (LeCun et al. 1998), AlexNet (Krizhevsky, Sutskever, and Hinton 2017), VGG9 (Simonyan and Zisserman 2014), and ResNet8 (He et al. 2016). We trained these models on a training set for 300 epochs, using the SGD optimizer with an initial learning rate of 0.1 and a learning rate decay of 0.1 at the 150^{th} and 225^{th} epochs. We also trained ResNet18 (He et al. 2016) and ShuffleNet (Ma et al. 2018) models on high-resolution images from ImageNet. To assess the vulnerability to attacks, we directly performed gradient inversion attacks on 100 validation images randomly selected from each dataset. We used attack algorithms from previous works (Geiping et al. 2020; Yin et al. 2021; Zhu, Liu, and Han 2019) and considered two major gradient matching losses: L2 and cosine distances. Also, we incorporated the total variation loss for regularization. We use Adam optimizer (Kingma and Ba 2015) for gradient inversion. For each sample, we run attack algorithm three times using different random seeds. The final outcome is the best one among these runs.

Image Similarity Scores. Image similarity scores measure the quality of reconstructed images compared to the original images. We consider Mean Squared Error (MSE), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), Structural Similarity Index (SSIM) (Wang et al. 2004) and Peak Signal-to-Noise Ratio (PSNR) for quantifying reconstruction quality. MSE computes the mean pixel-wise difference between original sample and its reconstruction in image space. LPIPS computes the distance from ground truth within the feature space of the ImageNet-pretrained VGG network. SSIM measures the similarity between two images by comparing their structural information, luminance, and contrast. PSNR measures the quality of reconstructed images using signal-to-noise (SNR) ratio.

Proxy for the Vulnerability. We consider the gradient norm and LAVP as the candidates for the proxy. In tables, the gradient norm is denoted as 'grad_norm', maximum and minimum eigenvalues for L2 distance hessian are denoted as 'max' and 'min', and maximum and minimum eigenvalues for cosine distance loss are denoted as 'ang_max' and 'ang_min'. Here, 'ang' is the abbreviation for 'angular'.

Evaluation Metric. For a random variable of reconstructed

σ_S	grad_norm	$\sqrt{\text{max * ang\_min}}$ (LAVP fusion)
C-10+L2	.35 / -.27 / -.35 / **-.13**	**.48 / -.49 / -.48** / -.09
C-100+L2	.41 / -.31 / -.41 / **-.19**	**.50 / -.49 / -.50** / -.09
IN+L2	.03 / .03 / -.03 / .19	**.48 / -.41 / -.48 / .59**
IW+L2	.35 / -.01 / -.35 / .00	**.71 / -.71 / -.71 / .51**
C-10+CS	-0.21 / **.25** / -.21 / .15	**-.28 / .25 / .28 / -.18**
C-100+CS	.10 / -.07 / -.10 / .02	**.50 / -.45 / -.50 / .36**
IN+CS	-.13 / .11 / .14 / -.18	**.44 / -.43 / -.44 / .34**
IW+CS	.00 / .01 / .00 / -.04	**.57 / -.59 / -.57 / .38**

Table 3: Spearman's correlation of loss-agnostic LAVP and gradient norm with image similarity scores (MSE(↓) / SSIM(↑) / PSNR(↑) / LPIPS(↓)). An instance of LAVP fusion is the geometric mean between the maximum eigenvalue of Hessian for L2 distance and the minimum eigenvalue of Hessian for cosine distance.

images X, we compute the correlation between two mapping of X, $A(X)$ and $B(X)$ where A is a similarity score, and B is a proxy for the vulnerability. Pearson's correlation coefficient (σ_P) is often used to compute linear correlation, while monotonocity is more important than linearity in our case. Thus, we use Spearman's correlation coefficient (σ_S) to compute monotonic relationship between $A(X)$ and $B(X)$. Note that Spearman's correlation coefficient is Pearson's correlation between $Rank(A(X))$ and $Rank(B(X))$, where $Rank(\cdot)$ is the operator for ranking numbers in increasing order. The correlation is said to be strong when the absolute value of σ_S is close to one. Specifically, intra correlation within the same architecture is more important as vulnerable examples might depend on the model. *Note that σ_S is computed for each architecture and their average is reported as the final evaluation metric.*

Results for the Correlation Between the Proxy and Vulnerability

Table 1 and Table 2 present correlation results of proxy candidates on several combinations of dataset and loss function type for low resolution images and high resolution images. *Note that the sign of the correlation depends on the image similarity score due to the fact that both MSE and LPIPS decrease as image quality improves, whereas the reverse is true for SSIM and PSNR.* When gradient inversion is based on the L2 (cosine) distance, the maximum and minimum eigenvalues of Hessian with the L2 (cosine) distance show stronger correlation with reconstruction quality in all image similarity scores than the gradient norm in Table 1. The absolute values of σ_S are mostly larger than 0.5 for LAVP with the corresponding attack loss function. In the case of cosine distance, LAVP achieves even the optimal value around 0.8.

In Figure 2, values of proxy candidates for each reconstructed sample are plotted in log scale along with its image quality in MSE for CIFAR-10. The gradient norm shows mixed trend in terms of correlation sign as it shows a slightly upward-sloping distribution with $\sigma_S = 0.35$ in Figure 2a but a slightly downward-sloping distribution with $\sigma_S = -0.28$ in 2d. In contrast, LAVP consistently shows upward-sloping distributions with at most $\sigma_S = 0.64$ which corresponds to stronger correlation than the gradient norm in Figures 2b, 2c, 2e, and 2f.

In Figure 3, candidate proxy values for each reconstructed sample are plotted in log scale along with its image quality in MSE on different loss functions and architectures for ImageNet. In Figure 3a, the gradient norm shows the moderate upward-sloping distribution with $\sigma_S = 0.66$, but this phenomenon rather negates the previous hypothesis that examples with higher gradient norm are more vulnerable. Therefore, we believe that this moderate correlation in the case of L2 distance might be due to the gradient scale, which affects both the gradient norm and Jacobian. In Figure 3d, the gradient norm shows almost no correlation with $\sigma_S = -0.06$ for cosine distance. For the case of cosine, there is no gradient scale issue since a normalizing factor $\frac{1}{||g^*||^2}$ exists in Equation 5. In Figures 3b, 3c, 3e, and 3f, LAVP consistently shows upward-sloping distributions with at most $\sigma_S = 0.74$.

LAVP Fusion for Black-box Scenario

In a black-box scenario where clients lack knowledge of the attacker's loss function, LAVP should be computed for each potential candidate loss function. To mitigate this complexity, we suggest a loss-agnostic version as a fusion of LAVPs for L2 and cosine distances. In Table 3, we present a specific instance of this fusion as the geometric mean between the maximum eigenvalue of the Hessian for L2 loss and the minimum eigenvalue for cosine similarity loss. For both L2 and cosine distances, the loss-agnostic LAVP shows stronger σ_S than the gradient norm with the vulnerability in most cases. The efficacy of loss-agnostic LAVP can be attributed to the observed minimal correlation between LAVPs for different loss functions. In Table 2, LAVP tailored for L2 distance exhibits the correlation of at most $|\sigma_S| = 0.06$ with the quality of reconstructed from cosine distance in MSE. On the other hand, LAVP tailored for cosine distance exhibits the correlation of at most $|\sigma_S| = 0.07$ with the quality of reconstructed images from L2 distance in MSE. This mutual lack of correlation underlines the absence of any interfering effect between the LAVPs designed for L2 and cosine distances. The concept of LAVP fusion can be extended to any future loss function for gradient matching beyond L2 and cosine.

Conclusion

This paper introduces a novel concept: a loss-aware vulnerability proxy, called LAVP, designed to gauge the loss-specific quality of reconstructed input from gradient inversion attacks. Unlike the gradient norm, a common heuristic in prior studies, LAVP—represented by either the maximum or minimum eigenvalue of Hessian with respect to gradient matching at ground truth—can explain loss (function)-dependent reconstruction behaviors of gradient inversion attack.

LAVP is derived from our theorems concerning gradient matching optimization. Our experiments show the efficacy of LAVP across diverse architectures and datasets. This study not only highlights the significance of Hessian eigenvalues as proxies for vulnerability in gradient inversion attacks but also provides deeper insights into the mechanics of these attacks, paving the way for future research in this domain.

Acknowledgments

This work was conducted by Center for Applied Research in Artificial Intelligence(CARAI) grant funded by Defense Acquisition Program Administration(DAPA) and Agency for Defense Development(ADD) (UD230017TD). This work was also supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)).

Special thanks to Seunghee Koh for thoughtful discussions about the presentation of this work.

References

Dang, T.; Thakkar, O.; Ramaswamy, S.; Mathews, R.; Chin, P.; and Beaufays, F. 2021. Revealing and Protecting Labels in Distributed Training. *Advances in Neural Information Processing Systems*, 34.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

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

Hatamizadeh, A.; Yin, H.; Molchanov, P.; Myronenko, A.; Li, W.; Dogra, P.; Feng, A.; Flores, M. G.; Kautz, J.; Xu, D.; et al. 2022. Do Gradient Inversion Attacks Make Federated Learning Unsafe? *arXiv preprint arXiv:2202.06924*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Howard, J. 2019. Imagewang. https://github.com/fastai/imagenette/.

Huang, Y.; Gupta, S.; Song, Z.; Li, K.; and Arora, S. 2021. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in Neural Information Processing Systems*, 34.

Jeon, J.; Lee, K.; Oh, S.; Ok, J.; et al. 2021. Gradient inversion with generative image prior. *Advances in Neural Information Processing Systems*, 34: 29898–29908.

Kariyappa, S.; Guo, C.; Maeng, K.; Xiong, W.; Suh, G. E.; Qureshi, M. K.; and Lee, H.-H. S. 2023. Cocktail party attack: Breaking aggregation-based privacy in federated learning using independent component analysis. In *International Conference on Machine Learning*, 15884–15899. PMLR.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Konečnỳ, J.; McMahan, H. B.; Yu, F. X.; Richtárik, P.; Suresh, A. T.; and Bacon, D. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features From Tiny Images.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, O.; Sun, J.; Yang, X.; Gao, W.; Zhang, H.; Xie, J.; Smith, V.; and Wang, C. 2022. Label leakage and protection in two-party split learning.

Ma, K.; Sun, Y.; Cui, J.; Li, D.; Guan, Z.; and Liu, J. 2022. Instance-wise Batch Label Restoration via Gradients in Federated Learning. In *The Eleventh International Conference on Learning Representations*.

Ma, N.; Zhang, X.; Zheng, H.-T.; and Sun, J. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, 116–131.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Mo, F.; Borovykh, A.; Malekzadeh, M.; Demetriou, S.; Gündüz, D.; and Haddadi, H. 2021. Quantifying and Localizing Usable Information Leakage from Neural Network Gradients. *European Symposium on Research in Computer Security*.

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

Wainakh, A.; Müßig, T.; Grube, T.; and Mühlhäuser, M. 2021. Label leakage from gradients in distributed machine learning. In *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*, 1–4. IEEE.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error measurement to structural similarity. *IEEE transactions on image processing*, 13(1).

Ye, D.; Zhu, T.; Zhou, S.; Liu, B.; and Zhou, W. 2022. Label-only Model Inversion Attack: The Attack that Requires the Least Information. *arXiv preprint arXiv:2203.06555*.

Yin, H.; Mallya, A.; Vahdat, A.; Alvarez, J. M.; Kautz, J.; and Molchanov, P. 2021. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16337–16346.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhao, B.; Mopuri, K. R.; and Bilen, H. 2020. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*.

Zhu, J.; and Blaschko, M. B. 2020. R-GAP: Recursive Gradient Attack on Privacy. In *International Conference on Learning Representations*.

Zhu, J.; Yao, R.; and Blaschko, M. B. 2023. Surrogate model extension (SME): A fast and accurate weight update attack on federated learning.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 32.