Evaluation Metric. For a random variable of reconstructed 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, as expected from our hypothesis. 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 different reconstruction patterns corresponding to different loss functions for gradient matching in the attack.

This innovation is based on our theoretical results concerning gradient matching optimization. In our theorems, we claim that low bi-Lipschitz constants of the gradient function with respect to the input signify susceptibility to gradient inversion attacks. We also establish a connection between bi-Lipschitz constants of the gradient function and the maximum and minimum eigenvalues of the Hessian near the ground-truth, which is how LAVP is derived.

In our experiments, we show the efficacy of our approach across diverse architectures and datasets, even encompassing high-resolution images like ImageNet. The results indicate that LAVP offers a stronger correlation with vulnerability compared to the gradient norm. We also introduce a loss-agnostic fusion of LAVPs for L2 and cosine distances as the proxy that caters to both L2 and cosine at once. 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. ???? Imagewang.

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.

Appendix

We include additional information in the Appendix.

We provide mathematical proofs for several theorems in the main paper.

We provide the pseudocode for computing maximum eigenvalues of hessian, which is omitted in the main paper due to page limits.

We present limitations and future work section, which is omitted in the main paper due to space constraints.

We present qualitative comparison between gradient norm and LAVP (ours) on datasets including CIFAR100 (Figure 4), ImageNette (Figure 5), and ImageWoof (Figure 6).

We present how strong LAVP correlates to final loss in local optimization scenario to understand the effectiveness of LAVP in capturing optimization behavior.

Mathematical Proofs

A1. Proof of Theorem 1

Theorem. (The first gradient inversion loss theorem). 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:

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

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

Proof. First, we will compute the vector for gradient descent, $\Delta x = -\mu \frac{\partial \mathcal{L}_{grad}}{\partial x}$ by chain rule as follows:

$$\begin{aligned}
\Delta x &= -\mu \frac{\partial \mathcal{L}_{grad}}{\partial x} \\
&= -\mu \frac{\partial ||\nabla_w \mathcal{L}(f(x), y) - g^*||_2^2}{\partial x} \\
&= -2\mu \nabla_x \nabla_w \mathcal{L}(f(x), y)(\nabla_w \mathcal{L}(f(x), y) - g^*).
\end{aligned}$$

Then, $\mathcal{L}_{grad}(x + \Delta x)$ can be separated into three terms by summation like the following:

$$\begin{aligned}
\mathcal{L}_{grad}(x + \Delta x) &= ||\nabla_w \mathcal{L}(f(x + \Delta x), y) - g^*||_2^2 \\
&= ||\nabla_w \mathcal{L}(f(x + \Delta x), y) - \nabla_w \mathcal{L}(f(x), y) + \\
&\quad \nabla_w \mathcal{L}(f(x), y) - g^*||_2^2 \\
&= ||u||_2^2 + 2u^T v + ||v||_2^2
\end{aligned}$$

, where $u = \nabla_w \mathcal{L}(f(x + \Delta x), y) - \nabla_w \mathcal{L}(f(x), y)$ and $v = \nabla_w \mathcal{L}(f(x), y) - g^*$.

For the first term, $||u||_2^2 = ||\nabla_w \mathcal{L}(f(x + \Delta x), y) - \nabla_w \mathcal{L}(f(x), y)||_2^2 \leq L^2 ||\Delta x||^2 = L^2 \mu^2 ||\frac{\partial \mathcal{L}_{grad}}{\partial x}||_2^2$ due to the L-Lipschitz continuity condition of $\nabla_w \mathcal{L}(f(x), y)$ with respect to an input.