Regular Article

Performance Analysis of Gradient Inversion Attack in Federated Learning with Healthcare Systems

Thi-Nga Dao, Phat Tien Nguyen

Faculty of Radio-Electronic Engineering, Le Quy Don Technical University, Hanoi, Vietnam

Correspondence: Thi-Nga Dao, daothinga.mta@gmail.com

Communication: received 08 November 2023, revised 14 December 2023, accepted 27 December 2023

Online publication: 29 December 2023, Digital Object Identifier: 10.21553/rev-jec.338

The associate editor coordinating the review of this article and recommending it for publication was Prof. Hoang Van Phuc.

Abstract– **Federated learning (FL) is widely applied to healthcare systems with the primary aim of keeping the privacy of patient's data while improving classification quality by using knowledge from multiple participants. However, the training images are believed to be embedded into the shared gradient, which indicates a privacy risk when sharing the gradient with other participants in FL. Therefore, this work aims to design and evaluate an image recovery attack on medical images. More specifically, dummy images are trained to match the dummy gradient to the shared gradient while maintaining the smoothness and naturalness of reconstructed images. On the adversary side, an optimization problem is formulated with variables of dummy images and network parameters treated as constants. We evaluate the gradient attack on two medical datasets and reconstructed images clearly show the details of chest X-ray and MRI images including bone and blood vessels of captured areas. Our work aims to increase the awareness of people on sharing the gradient in FL, especially in healthcare systems.**

Keywords– **Gradient leakage attack, federated learning, medical images.**

1 Introduction

Federated learning (FL) allows the preservation of data privacy while training deep learning models collaboratively without directly exchanging the training data [1, 2]. In centralized FL, each participant such as hospital usually shares a gradient with a network parameter server, and then the server aggregates a global model using training images from all participants. Since it is very important to keep patients' data private, FL has recently attracted a lot of healthcare applications. For example, how to build a global model to diagnose patients' diseases based on data such as X-ray images, clinical photography, or test indexes. However, existing works demonstrated that sharing gradient could lead to the leakage of training images. In this work, we aim to design and evaluate a gradient attack algorithm in healthcare systems in distributed learning.

Reconstruction of training images is an interesting research topic that has been investigated by a lot of researchers. Information used for image reconstruction can vary from the shared gradient [3–5], image representation [6], or trained neural network [7]. In this work, we focus on the problem of image reconstruction attacks using the shared gradient in FL, especially for medical images. Authors in [3] proved that inputs of a fully connected (FC) layer can be fully recovered using the shared gradient of that layer. Zhu *et al* [4] introduced a new gradient attack for a convolutional neural network by training dummy images to match the shared gradient. However, their method can only recover images with a resolution as high as 32×32. Moreover, existing works focus on well-known general datasets such as MNIST, CIFAR-100, or ImageNet [8]. There is still a lack of studies that investigates the performance of gradient attack on a medical dataset. Therefore, our work aims to design and evaluate a method to recover training medical images from the shared gradient in healthcare systems with federated learning.

To find the reconstructed image, we formulate an optimization problem that minimizes the discrepancy between the shared gradient and the gradient generated by the reconstructed image, which is called a dummy gradient for short. The variables of the optimization problem are the dummy images that are initialized randomly. To increase the smoothness and naturalness of the recovered images, two regularization terms are added including total variance and six-norm losses. Specifically, the total variance loss reduces the difference between two nearby pixels while the six-norm loss guarantees the reconstructed images within a limited range. To solve this optimization problem, we can apply any numerical method to approximate the roots of a loss function such as Newton–Raphson [9]. In this work, we use the L-BFGS optimizer method implemented in Pytorch to find the reconstructed images that minimize the loss function.

The gradient attack has been evaluated on two medical images: chest X-ray and MRI images using the LeNet architecture. The normal participant or hospital randomly selects a training image to update network parameters at each training round. Based on the shared gradient, the attacker can successfully recover the training images with the resolution as high as 2048×2048 with a very low mean squared error (MSE) of 0.0006, a high structural similarity index measure (SSIM) of 0.959, and a high peak signal-to-noise ratio (PSNR) of 32.34. We also analyze the impacts of image resolution on the gradient attack and the results show that there is a higher risk of data leakage with higher image resolution. In addition, we evaluate the effects of the differential privacy that added a Gaussian noise to the shared gradient before sharing it to other clients. Specifically, the reconstructed performance generally deteriorates as a higher variance of noise is used. Then, we compare the reconstruction performance between our work and the existing method in [4]. Finally, we demonstrate reconstructed images in the comparison with the original training images in both datasets.

The organization of the paper is listed as follows. Section 2 provides a short summary of federated learning and gradient attack works. Then, we present the gradient attack method in Section 3 followed by the performance evaluation in Section 4. Finally, we give the conclusion of our work and discuss the potential work of gradient attack in medical images in Section 5.

2 Related Work

In this section, we summarize existing works related to federated learning and common attacks in FL as well as the gradient attack.

2.1 Federated Learning

FL can be categorized into two groups: centralized federated learning (CFL) and decentralized federated learning (DFL) based on how the consensus model is built [1]. Figure 1 shows the two types of FL in healthcare systems. CFL includes a central server and a set of clients such as hospitals to collaboratively train a global model. In a training epoch, all clients update the parameters using their private datasets such as X-ray images, MRI images,. Then, the updated parameters are sent to the central server where an aggregation algorithm is used to create a consensus model. There are multiple aggregation algorithms such as federated averaging [10] or attentive federated aggregation [11]. Then, the central server or the coordinator sends back the global model to all clients for the next round of training. The training process ends when the global model converges using the data from all clients. CFL is suitable in cross-silo federated learning with few participants such as organizations and the communication with the server is available.

Unlike CFL, DFL does not require any central server as the coordinator for training the global model. In DFL, a client such as hospital sends the model update to a set of neighbor agents and then these clients update the network parameters based on the model update received by neighbor clients. At each client, an aggregation method is used to achieve a global update. Although DLF is preferred in cross-device federated learning with a large number of participants such as smart phones, we prove that the gradent attack can be launched in both CFL and DFL as long as the gradient information is leaked. The attack can occur at the central server in CFL or at any client in DFL.

2.2 Common Attacks in FL

According to the survey on federated learning attacks and defenses [12], there are multiple attacks such as poisoning attacks, inference attacks, evasion attacks, and backdoor attacks. Poisoning attacks can be done by injecting false/misleading data or changing the gradient before sharing with the centralized server or peer participants [13, 14]. Poisoning attacks can degrade the performance of the global model. Inference attacks [15] may reveal sensitive information about training data such as training inputs, training labels or data membership. This kind of attack leverages the shared gradient information for inference. Evasion attacks are performed during the inference phase and refer to designing an input that seems normal to human but is wrongly classified by the global model [16]. Backdoor attacks can cause the global model to misbehave on specific inputs while appearing normal in other cases [17]. In this attack, a malicious functionality is inserted into a targeted model through poisoned updates from malicious clients.

2.3 Gradient Attack

Phong *et al.* was the first analytical attack that provided a closed-form expression for the reconstructed inputs of the FC layer. Let W, b denote the weight matrix and bias vector of the layer while x, z are input and output vectors of the layer. The FC layer can be expressed as follows: $Wx + b = z$. If l is the loss function that indicates the difference between the predicted and true labels. Then, the input vector of the FC layer can be approximated as follows

$$x' = \frac{\frac{\partial l}{\partial W}}{\frac{\partial l}{\partial b}}. \quad (1)$$

Zhu [4] was the first optimization attack that utilized an optimization approach to minimize the difference between the dummy and shared gradients. Specifically, the objective function $\mathbb{L}$ depends on the dummy images, which are treated as variables of the optimization problem. The L-BFGS optimizer is used to find the solution to the optimization problem. Specifically, the dummy images x' can be updated below

$$x'_{t+} = x'_t - \alpha \frac{\partial \mathbb{L}}{\partial x'}, \quad (2)$$

where x'_t denote the dummy value at the epoch t and α is learning rate. Experimental results showed that the attack by [4] is sensitive to initialization, i.e., if using an auxiliary image with the same class as the training image, it is more likely that the reconstructed image is successfully recovered.

Geiping *et al.* [5] formulated an optimization problem that matches gradients while regularizing image fidelity. More specifically, the loss function includes the difference between shared and dummy gradients as

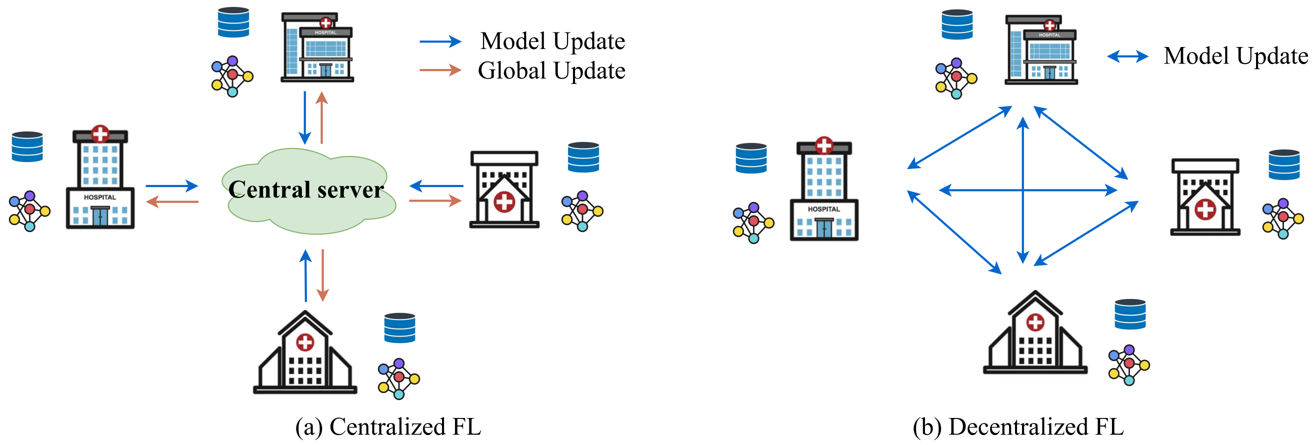


Figure 1. Federated learning in healthcare systems.

well as the total variation loss. However, the authors evaluated the attack method on general datasets such as CIFAR-10 and ImageNet and there is still a lack of performance evaluation on medical images. Therefore, we focus on the gradient attack on healthcare systems by analyzing the reconstruction performance on well-known medical datasets.

3 Gradient-Inversion-based Image Recovery Algorithm

We first describe some assumptions related to the federated learning architecture. Hospitals in FL share a global CNN model f for the classification task such as detection of brain tumors or Pneumonia. Network parameters θ of the global CNN model are initialized using the He method [18]. The output of the model can be defined as $y = f(x, \theta)$. Assume that the normal participant has private training images and for each training epoch a training image x is randomly selected to update the global model. Note that the training image x is usually normalized into the range $[0, 1]$ to accelerate the training process. Recall that we denote l as the cross-entropy loss value that shows the difference between the real and predicted labels y. Instead of sharing the training image, each hospital sends the gradient $\frac{\partial l}{\partial W}$ and $\frac{\partial l}{\partial b}$ generated by the training image to other hospitals in FL. Network parameters θ include weights W and biases b. Please note that we aim to reconstruct the training inputs but not the training labels since the labels can be accurately reconstructed using existing works.

The overview of the proposed architecture is shown in Figure 2. We initialize a random image as the reconstructed image $x^{'}$ that is fed into a CNN model to compute the dummy gradient. Then, we formulate a weighted loss function to compare the difference between the dummy and shared gradients. Specifically, we define the weighted loss function as

$$\mathbb{L} = \alpha_g L_g + \alpha_{tv} L_{tv} + \alpha_{norm} L_{norm}, \tag{3}$$

where $L_g = MSE(\frac{\partial l}{\partial \theta}, \frac{\partial l'}{\partial \theta})$ is the difference between the shared gradient $\frac{\partial l}{\partial \theta}$ and dummy gradient $\frac{\partial l'}{\partial \theta}$; L_{tv} is the total variation loss that is added as a regularizer to maintain the fidelity of images; L_{norm} is another regularizer used to constrain the value of the image within a specific range. The total variation loss can be computed as follows

$$L_{tv} = \sum_i (x'_i - x'_{i+1})^2, \tag{4}$$

where x'_i and x'_{i+1} are values of two nearby pixels with indices i and $i+1$ in the reconstructed x'. L_{tv} is added to the loss function since we expect neighboring pixels to have similar values. We can consider the total variation loss as a denoising method in the gradient attack. At the beginning of the recovery algorithm, L_{tv} has a very large value and this loss tends to decrease during the reconstruction phase. Meanwhile, the six-norm loss is considered as below

$$L_{norm} = \sum_i (x'_i)^6. \tag{5}$$

The LBFGS optimizer is considered to find the solution to the optimization problem. The objective function is $\mathbb{L}$ and variables are x'. The best-reconstructed image x^* can be derived as follows

$$x^* = \arg\min_{x'} \mathbb{L} = \arg\min_{x'} (\alpha_g L_g + \alpha_{tv} L_{tv} + \alpha_{norm} L_{norm}). \tag{6}$$

Pixel i of the dummy image can be updated at the epoch t below

$$x'_{i,t} = x'_{i,t-1} - \alpha \frac{\partial \mathbb{L}}{\partial x'_i}, \tag{7}$$

where $x'_{i,t}$ is the pixel of x' at epoch t. While updating x'_i, we keep other variables as constants.

The best coefficients $\alpha_g, \alpha_{tv}, \alpha_{norm}$ usually depend on the training dataset and can be found using an auxiliary sample from the adversary. Moreover, the appropriate coefficients usually differ from the image resolution. Therefore, it is important to validate multiple sets of coefficients to find the best coefficients. Note that LBFGS is a second-order optimization algorithm that measures the second-order derivative to know which direction to move (like the first-order) and also to estimate how far to move in that direction. LBFGS is implemented in various Python-based machine learning libraries such as TensorFlow or Pytorch. If only one training image is used to update the network parameters, the number of variables of the optimization problem is $w \times h \times d$ where w, h, d are the width, height, and depth of the original image. If using the image with high resolution, we expect to have high accuracy performance and long attack time. Generally, medical images in healthcare systems have high resolution to achieve high classification accuracy.

After finding the best-reconstructed image x^* that minimizes $\mathbb{L}$, we need to clip x^* into the range $[0, 1]$

$$x^* = max(0, min(x^*, 1)). \tag{8}$$

Since the original image is normalized before feeding into the classification model, the clipping step is needed.

4 Performance Evaluation

For performance evaluation, we consider two datasets. The first one is a high-quality dataset of chest X-ray images downloaded from https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia with two image categories: pneumonia and normal. The other dataset contains MRI brain images downloaded from https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection. The default image resolution is 1024×1024. Assume that each hospital randomly selects one image from the dataset for training and then sends the gradient to the network parameter server or client. Our assumption is similar to online training when the client updates the model whenever a new training image is collected. If multiple training epochs are used, the entire training dataset can be utilized for updating network parameters. The LeNet architecture with four convolutional layers and one hidden fully-dense layer is used for the classification model. The number of channels in LeNet is set to 12 in four convolutional layers and kernel size is 5. The stride is set to 2 in the two first layers and 1 in the two last convolutional layers. Padding is applied in these convolutional layers. The output of the convolutional layers is flattened and fed to the fully-connected layer with two output neurons. These output units present the probability of normal and abnormal classes.

Three main performance metrics are used to evaluate the performance of the gradient attack including Mean Squared Error (MSE), Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR) between the original and reconstructed images. The reconstruction quality gets better with a smaller MSE, higher SSIM, and higher PSNR.

The reconstruction quality greatly depends on the coefficients in the loss function. The optimal coefficients can be found by performing a grid search. Note that the optimal weights need to be updated when image resolution changes. For example, with chest X-ray images of 512×512, $\alpha_g = 1, \alpha_{tv} = 1.5 \times 10^{-8}, \alpha_{norm} = 10^{-10}$; with images of 1024×1024, $\alpha_g = 1, \alpha_{tv} = 1.7 \times 10^{-8}, \alpha_{norm} = 10^{-10}$; with images of 2048×2048, $\alpha_g = 2, \alpha_{tv} = 1.7 \times 10^{-8}, \alpha_{norm} = 10^{-10}$. When there is no auxiliary data, we may need to check different sets of weights to find the optimal values.

4.1 Reconstruction Quality During the Attack

Figure 3 shows the weighted loss function as well as L_g, L_{tv}, L_{norm} while updating the dummy images. To find the best-reconstructed image, we need 250 epochs for the example image. At the beginning of training, the gradient loss L_g, the total variation loss L_{tv}, and the norm loss L_{norm} have a very large value above 10^4. During the training procedure, the loss value decreases especially L_g to around 10^{-6}, which means the dummy gradient matches the shared gradient. Meanwhile, the decrease in L_{tv} and L_{norm} is smaller than L_g. Note that, L_g contributes greatly to the weighted loss function, which means α_g is much larger than α_{tv} and α_{norm}.

As shown in Figure 3, a random image is initialized at epoch 0. The reconstructed image starts to be revealed at epoch 100 and the best-reconstructed image can be seen after epoch 200. The weighted loss value reaches the saturation period after epoch 200. The training procedure stops at epoch 250 after observing no improvement in reconstruction quality.

4.2 Impacts of Image Resolution

The classification performance highly depends on the image resolution of the inputs. With abnormal images such as chest X-ray of pneumonia and brain MRI images with tumours, it is necessary to accurately recover the training images. As shown in Figure 4, the reconstruction quality increases when using high image resolution. With an image size of 512×512, the attack can only result in a quite blurred image and blood vessels cannot be seen clearly. As the image size is set to 2048×2048, the chest X-ray image with clear blood vessels can be recovered as shown in Figure 4(f). However, the attack time increases significantly with a high image resolution. More specifically, Table I shows that the attack time jumps from 30.12 to 5229.29 (s) when the image resolution is set to 256×256 and 2048×2048, respectively. This is due to the exponential increase in the number of variables that need to be optimized. The attacker always wants to achieve reconstruction performance as high as possible. However, high reconstruction quality, which is related to high image resolution, usually requires a long reconstruction time or powerful computing resources. Therefore,

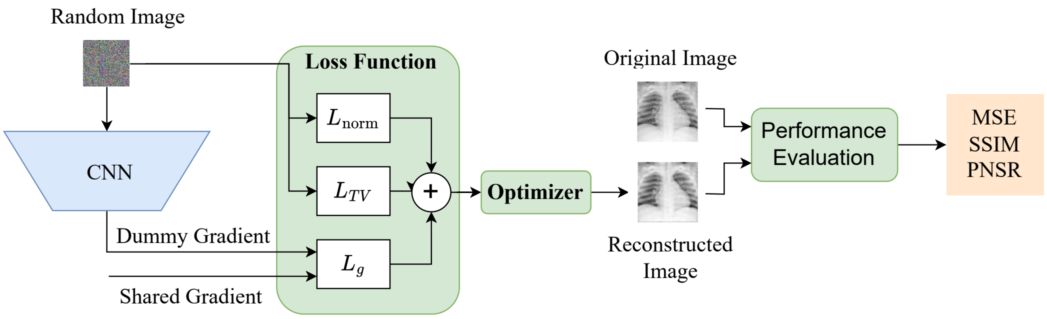


Figure 2. The overview of gradient inversion attack in healthcare systems.

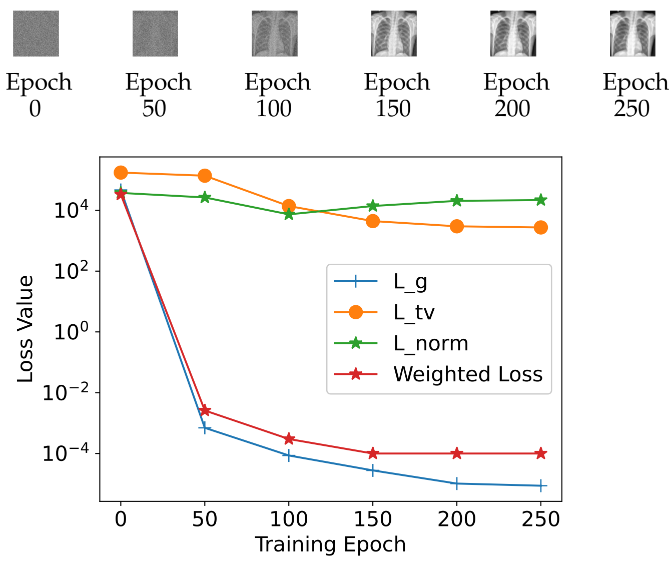
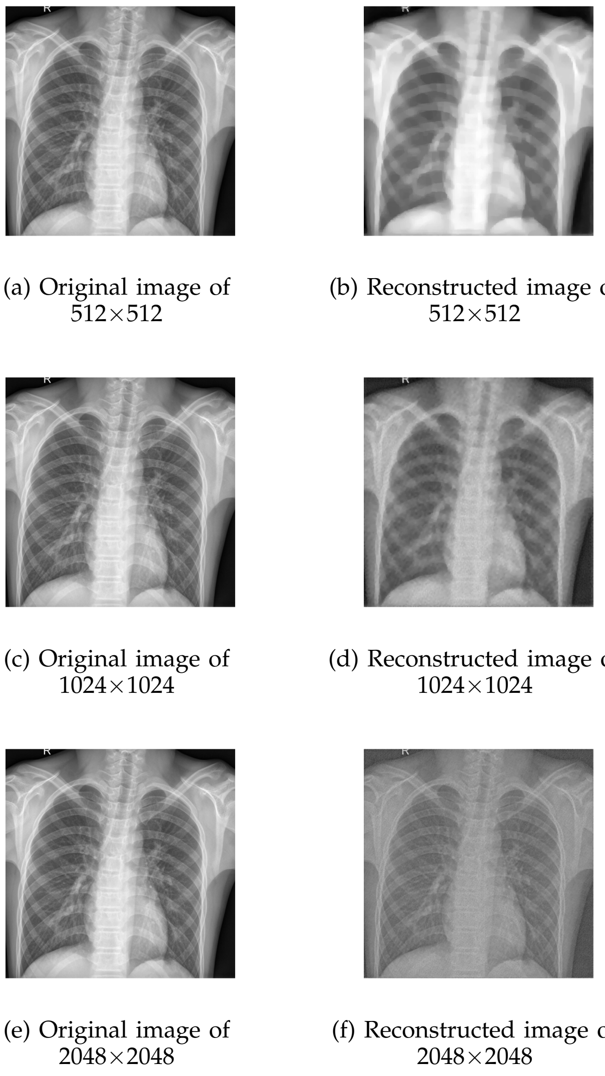


Figure 3. Reconstructed image during the reconstruction attack.

each participant in FL needs to consider both factors: classification and reconstruction quality when choosing the image resolution of training inputs.

We also measure MSE, SSIM, and PSNR of the reconstructed image with index 1 from the chest X-ray dataset. MSE becomes smaller, SSIM and PSNR get higher as image resolution increases. For example, the structural similarity index improves significantly from 0.769 to 0.959 when the image resolution changes from 256×256 to 2048×2048.


(a) Original image of 512×512

(b) Reconstructed image of 512×512

(c) Original image of 1024×1024

(d) Reconstructed image of 1024×1024

(e) Original image of 2048×2048

(f) Reconstructed image of 2048×2048

Figure 4. Impact of image resolution on the reconstruction of X-ray images.

4.3 Impacts of Differential Privacy

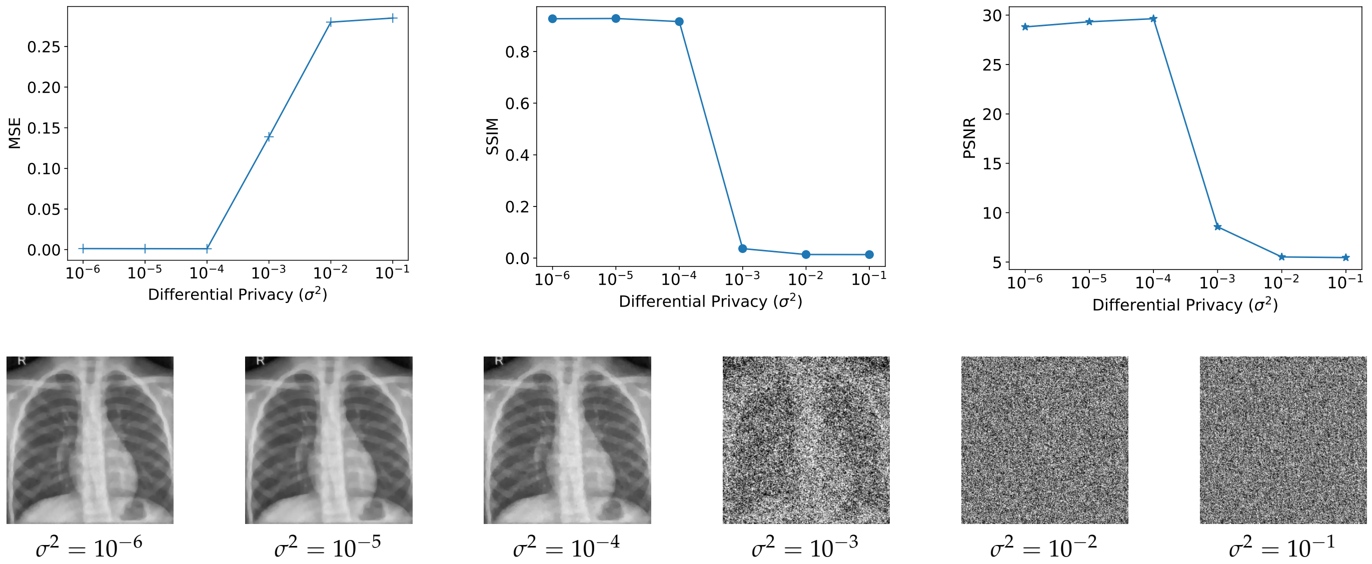
As stated in existing works, differential privacy can be used as a countermeasure for the gradient attack. In this subsection, we measure the impacts of differential privacy on the reconstruction quality of the gradient attack using three metrics: MSE, SSIM, and PSNR. Figure 5 shows the impacts of differential privacy on the quality of reconstructed images with different variances σ^2 of noise that is set to $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. When $\sigma^2 < 10^{-4}$, MSE, SSIM, PSNR keep stable. However, reconstruction performance starts to deteriorate quickly when σ^2 is greater than 10^{-3}. For example, with $\sigma^2 = 0.01$, MSE is around 0.28, SSIM is only 0.0139, PSNR is 5.52, and the reconstructed data looks like a random image.

4.4 Performance Comparison with Existing Work

We validate the performance of our work and deep leakage gradient (DLG) [4] in terms of MSE, SSIM, and PSNR. The image resolution is set to 512×512. In DLG, the loss function only contains the difference between the dummy gradients and shared gradients;

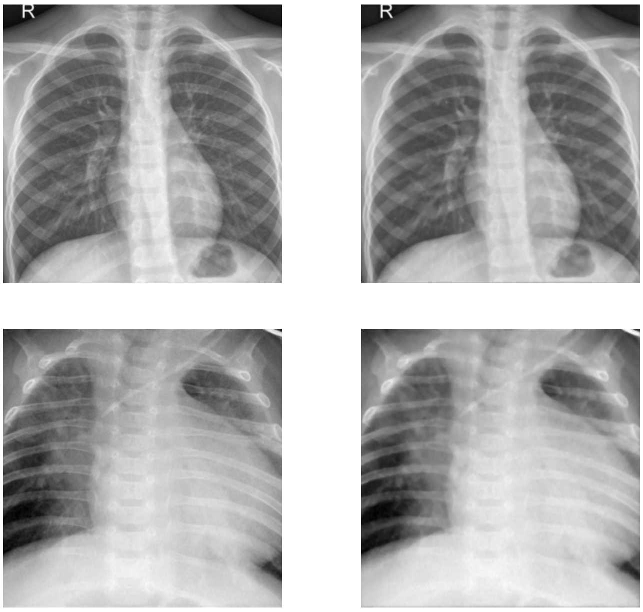
Table I
The Impacts of Image Resolution on the Performance of the Gradient Attack

Image resolution	#Variables	Attack time (s)	MSE	SSIM	PSNR
256×256	196k	30.12	0.0189	0.769	17.24
512×512	786k	112.81	0.0013	0.927	28.85
1024×1024	3.14M	914.75	0.0009	0.934	30.27
2048×2048	12.48M	5229.29	0.0006	0.959	32.34



$\sigma^2 = 10^{-6}$ $\sigma^2 = 10^{-5}$ $\sigma^2 = 10^{-4}$ $\sigma^2 = 10^{-3}$ $\sigma^2 = 10^{-2}$ $\sigma^2 = 10^{-1}$

Figure 5. Impacts of differential privacy on the reconstruction attack.



(a) Original images (b) Reconstructed images

Figure 6. Reconstruction quality in the chest X-ray dataset.

the gradient coefficient is set to 6. Meanwhile, our work adds the total variation and six-norm losses to maintain the smoothness and naturalness of the reconstructed images. The coefficients in the loss function are set as $\alpha_g = 1, \alpha_{tv} = 1.5 \times 10^{-8}, \alpha_{norm} = 10^{-10}$. As can be seen in Table II, our work can achieve lower MSE, higher SSIM, and higher PSNR than DLG.

4.5 Visualization

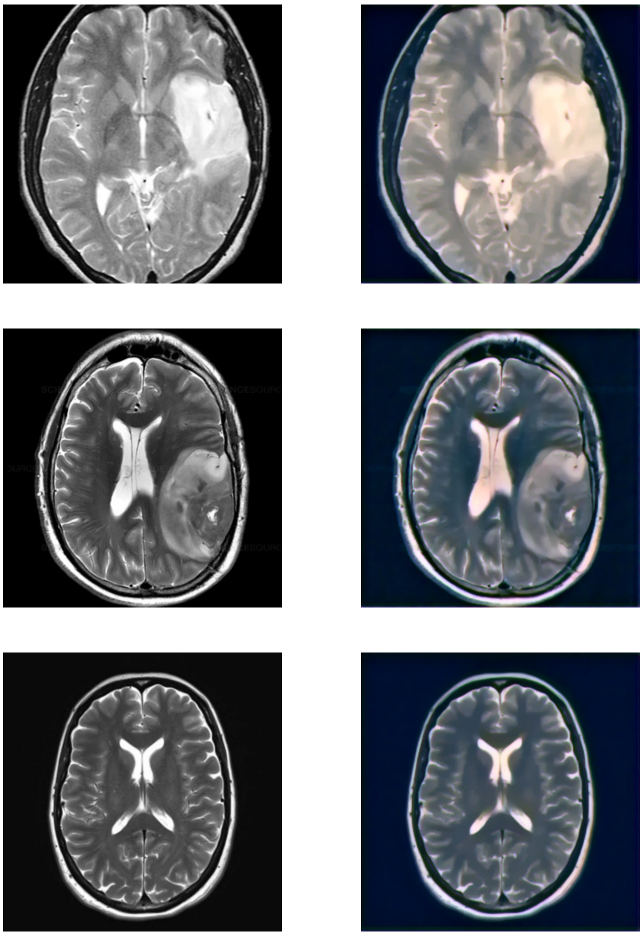
To further show the effectiveness of the gradient attack, we present the reconstructed images in both chest X-ray and MRI image datasets in Figures 6 and 7. In the chest X-ray dataset, we use images of 2048×2048 to achieve the best visualization. In the MRI dataset, two images with brain tumours and one image of a normal brain are used for training; the MRI image has 1024×1024 image resolution. Figure 6 shows that reconstruction for chest X-ray images of normal and abnormal lung conditions. The reconstructed images can show details about lung injuries: cancer or the air collecting in the space around a lung. In the reconstructed MRI images, bones and blood vessels can be clearly seen from the reconstructed images in both datasets, which indicates that the adversary can successfully recover the training image from the shared gradient.

5 Conclusion

In this work, we present a threat model in healthcare systems with federated learning where the adversary may recover the training images by using only the shared gradient. An optimization problem is formulated to reconstruct the training images by matching the dummy gradient and the shared gradient. In addition, we add two regularization losses: the total variation loss to minimize the difference between neighboring pixels and the six-norm loss to keep the pixels within a reasonable range. We conduct extensive experiments

Table II
Performance Comparison Between Our Work and DLG

Algorithm	**MSE ↓**	**SSIM ↑**	**PSNR ↑**
DLG	0.0561	0.088	12.51
Our work	0.007	0.93	31.29



(a) Original images (b) Reconstructed images

Figure 7. Reconstruction quality in the MRI Image dataset: The two first figures present the images with brain tumours and the last figure shows the MRI image of a normal brain.

on chest X-ray and MRI datasets and the performance results show that reconstructed images can be recovered well under certain conditions: small batch size, high image resolution, and noise with small variances added in the gradient. By studying the gradient attack in healthcare systems with FL, we can design a secure FL architecture to countermeasure against this attack and ensure a safe distributed learning environment.

To mitigate the impact of gradient attacks, various privacy-preserving techniques can be used including holomorphic encryption, secure multi-party computation, and differential privacy. Holomorphic encryption involves encrypting gradient information before sharing it with other participants. Secure multi-party computation allows a server to compute a weighted average of encrypted weights from participants without revealing the original weights of any specific participant. Differential privacy adds noise to the gradient at participants before model aggregation at the server, thus ensuring data privacy and serving as a common countermeasure against gradient leakage attacks.

In our future work, we expand the gradient attack in a more challenging situation such as large batch sizes, and explore their effects on practical global models like MobileNet and DenseNet. In addition, it is crucial to design an effective defense strategy to reduce the impacts of attacks in federated learning.

Acknowledgment

This publication is the output of the ASEAN IVO (http://www.nict.go.jp/en/asean_ivo/index.html) project, "Artificial Intelligence Powered Comprehensive Cyber-Security for Smart Healthcare Systems (AIPOSH)", and financially supported by NICT (http://www.nict.go.jp/en/index.html).

References

[1] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021.

[2] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[3] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving deep learning: Revisited and enhanced," in *Proceedings of the 8th International Conference on Applications and Techniques in Information Security*. Springer, 2017, pp. 100–110.

[4] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in neural information processing systems*, vol. 32, 2019.

[5] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients-how easy is it to break privacy in federated learning?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 937–16 947, 2020.

[6] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196.

[7] N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani, "Reconstructing training data from trained neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 911–22 924, 2022.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[9] J.-F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical optimization: theoretical and practical aspects*. Springer Science & Business Media, 2006.

[10] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[11] S. Ji, S. Pan, G. Long, X. Li, J. Jiang, and Z. Huang, "Learning private neural language modeling with attentive aggregation," in *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[12] H. S. Sikandar, H. Waheed, S. Tahir, S. U. Malik, and W. Rafique, "A Detailed Survey on Federated Learning Attacks and Defenses," *Electronics*, vol. 12, no. 2, p. 260, 2023.

[13] G. Sun, Y. Cong, J. Dong, Q. Wang, L. Lyu, and J. Liu, "Data poisoning attacks on federated machine learning," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 11 365–11 375, 2021.

[14] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to Byzantine-Robust federated learning," in *Proceedings of the 29th USENIX security symposium (USENIX Security 20)*, 2020, pp. 1605–1622.

[15] H. Hu, Z. Salcic, L. Sun, G. Dobbie, and X. Zhang, "Source inference attacks in federated learning," in *Proceedings of the 2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 1102–1107.

[16] A. H. Bondok, M. Mahmoud, M. M. Badr, M. M. Fouda, M. Abdallah, and M. Alsabaan, "Novel Evasion Attacks against Adversarial Training Defense for Smart Grid Federated Learning," *IEEE Access*, 2023.

[17] X. Gong, Y. Chen, Q. Wang, and W. Kong, "Backdoor attacks and defenses in federated learning: State-of-the-art, taxonomy, and future directions," *IEEE Wireless Communications*, 2022.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

Thi-Nga Dao received a B.S. degree in Electrical and Communication Engineering from the Le Quy Don Technical University, Vietnam in 2013, an M.S. degree in Computer Engineering from University of Ulsan in 2016, and a Ph.D. degree in Computer Engineering from University of Ulsan, South Korea in 2019. Since July 2019, she has been a researcher in the Faculty of Radio-Electronic Engineering, Le Quy Don Technical University, Hanoi, Vietnam. She was a postdoctoral fellow with the Computer Science and Engineering Department, Ewha Womans University in 2021. Her research interests include attacks and defense methods for federated learning, machine learning-based applications in network security, network intrusion detection and prevention systems, human mobility prediction and mobile crowdsensing.

Tien-Phat Nguyen is a lecturer at the University of Le Quy Don Technical University, Hanoi, Vietnam. He received the M.E, and the PhD degrees in electrical engineering from Ryazan State Radio-Engineering University, Russia, in 2012 and 2015, respectively. His current research interests include remote sensing image processing, image and video processing, and machine learning.