Data Reconstruction Attack with Label Guessing for Federated Learning

Jinhyeok Jang, Yoonju Oh, Gwonsang Ryu, Daeseon Choi[*]

Department of Software, Soongsil University, Republic of Korea
jjh4002@ssu.ac.kr, ohyoonju@soongsil.ac.kr, gsryu@ssu.ac.kr, sunchoi@ssu.ac.kr

Abstract

In light of recent advancements in deep and machine learning, federated learning has been proposed as a means to prevent privacy invasion. However, a reconstruction attack that exploits gradients to leak learning data has recently been developed. With increasing research into federated learning and the importance of data usage, it is crucial to prepare for such attacks. Specifically, when face data are used in federated learning, the damage caused by privacy infringement can be significant. Therefore, attack studies are necessary to develop effective defense strategies against these attacks. In this study, we propose a new attack method that uses labels to achieve faster and more accurate reconstruction performance than previous reconstruction attacks. We demonstrate the effectiveness of our proposed method on the Yale Face Database B, MNIST, and CIFAR-10 datasets, as well as under non-IID conditions, similar to real federated learning. The results show that our proposed method outperforms random labeling in terms of reconstruction performance in all evaluations for MNIST and CIFAR-10 datasets in round 1.

Keywords: Reconstruction attack, Leakage attack, Federated learning, Privacy

1 Introduction

Owing to continued advancements in deep and machine learning, the privacy protection of data has become critical. To prevent privacy invasion as a result of data exposure, various privacy protection methods, such as anonymous processing methods, K-anonymity, L-diversity, differential privacy, and data synthesis, have been proposed.

However, the existing methods are highly dependent on data. As important information generated during preprocessing is discarded, usable data are reduced and substantial time is required for the processing thereof [1-4]. Federated learning, which is a data- independent machine-learning method, has been proposed to solve this problem. Federated learning is a method of communicating by delivering parameters, gradients, and weights without delivering data from the client to the server when the model is initially delivered to the client, as well as upgrading the server model [5-8]. Therefore, federated learning can prevent the privacy invasion of users caused by the revelation of data.

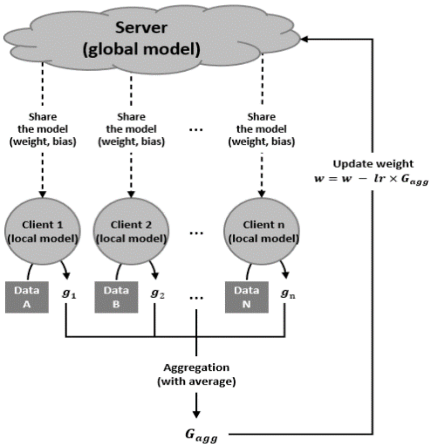

Figure 1. Process of federated learning

Figure 1 shows an example of federated learning. However, it has been demonstrated that privacy invasions can be caused through the reconstruction of data from clients who participate in federated learning [9]. In addition, Zhu et al. [9] used the L2 norm. Reconstruction attacks are steadily increasing, as studies that increase the performance of reconstruction by adding TV norm [10] for image normalization have been proposed [11]. This attack is a method of reconstructing random data and random labels on the data held by the client using the gradient of the client. Based on these studies, Wei et al.'s method [12] was implemented using RGB and dark/light, which provides initial information on random data [9]. Furthermore, Wei et al.'s study [12] stated that labels can be leaked through locally obtained gradients. Wainakh et al. reported that this can also serve as an important factor in generating gradients and can leak real labels [13]. However, Wainakh et al.'s study [13] differs significantly in that it only leaked labels, and the present study was used to reconstruct data to leaked labels in other ways. As a result, our study differs from others in that McMahan et al. [8] and Zhu et al. [9] used random labels and Persson et al. [10] found random labels only, whereas we found random labels and reconstruction data. Unlike previous studies, this study initially set information on the labels that were generated in the reconstruction process, thereby enabling the model to reconstruct only random data. As this method provides the correct answer to the model in advance, the reconstruction speed is faster, and the reconstruction performance is better than those reported by Wei et al. and Wainakh et al. [12-13]. In this study, we propose a label

[*]Corresponding Author: Daeseon Choi; E-mail: sunchoi@ssu.ac.kr

DOI: 10.53106/160792642023072404007

guessing method for fast and accurate data reconstruction attacks. We select the model used in Zhu et al.'s study [9] as the baseline model to compare the performance of accurate labels. We use the same model [9] in the Yale, CIFAR-10, and MNIST datasets, conduct experiments in non-independent and equally distributed non-IID data situations, and demonstrate better use of labels using mean square error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity (SSIM) as evaluation metrics. The contributions of this study are summarized as follows:

1. We propose a label guessing method for fast and accurate data reconstruction attacks.
2. Simultaneously with federated learning, we show the performance of the reinstatement attack together. This can also be reconstructed by the method proposed during federated learning, which is a non-IID situation.
3. A comparative experiment was conducted on the proposed method by mapping various situations to previous studies on various datasets and evaluation metrics.

The composition of this paper is as follows. Chapter 2 mentions the study of proposals through threat models and introduces the focus of the proposed attacks. Chapter 3 shows the experimental environment and results, and Chapter 4 shows the conclusion.

2 Background

2.1 Threat Model

We assume that an attacker can access the server during federated learning. The client provides the server with data-exempt gradients and parameters. The attacker analyzes the gradient obtained from the client. From this gradient, an attacker can infer data. Our work studies the infringement of personal information in the federated learning paradigm by reconstructing data with gradients. Labels can be inferred during the process of reconstructing data from the gradient, and we aim to speed up data reconstruction by incorporating inferred labels in advance. We have mapped our study to previous works, considering circumstances in which an attacker might be present. In a random label situation [9], the attacker is completely unaware of the label and reconstructs the data and labels simultaneously. Known label situation: An attacker knows the label [12-14]. Similar to the circumstances in which labels can be inferred in the study but very different from what is actually known, we did not use them to reconstruct leaked labels. Additionally, in a similar study, Zhao et al. [11] use a label that reconstructs the actual label, not the random value, in the step of generating the pseudo label during the slope inversion process. This is learned by general models using cross-entropy [15] as a loss function for each output; another study reconstructs the actual label using the negative slope sign of the output neuron corresponding to the target label. Guessing label situation: It is a situation in which an attacker starts without knowing the label in the beginning, infers the label in the middle, and reflects the label in the gradient to help reconstruct it quickly.

2.2 Reconstruction Attack

The method of a reconstruction attack based on a gradient was proposed by Zhu et al. [9]. The reconstruction attack method is illustrated in Figure 2. We define the data as A as data consisting of the original image and the original label, the model learns through A, and the slope at which the data is sent is defined as G. In addition, we define data consisting of random images and random labels to be reconstructed on the server as R, and we define the gradient of the server as G(R) by learning the model.

The reconstruction formula is presented in Equations (1) and (2); when a random label is used, the gradient is calculated to minimize the difference from the original through LBFGS. In Figure 2, the iterations are   set, with iteration-optimized data. These reconstruction attacks are appropriate attack methods for federated learning. The server may know the gradient in the part where the client transmits the gradient to the server following local training. Accordingly, if the server maliciously uses the gradient during federated learning, the privacy invasion is sufficient.

$$\nabla w' = \frac{\theta I(F(x',W),y'))}{\theta w}, \nabla w = \frac{\theta I(F(x,W),y))}{\theta w}. \quad \textbf{(1)}$$

$$X',Y' = \arg\min_{X',Y'} \left\| \nabla w' - \nabla w \right\|^2 . \quad \textbf{(2)}$$

This paper describes a method for improving the performance of the reconstruction attack presented in the previous study. Our experimental results demonstrate that the reconstruction performance varies depending on the label.

First, based on previous studies, a random label refers to the situation when an attacker does not know the label. Second, if the attacker knows the label, it is set to a known label, and from the perspective of the attacker, in federated learning, the label is rarely known because it is difficult to own data. As the calculation and time cost of the random label are very high, research on the same performance as that of the known label is required.

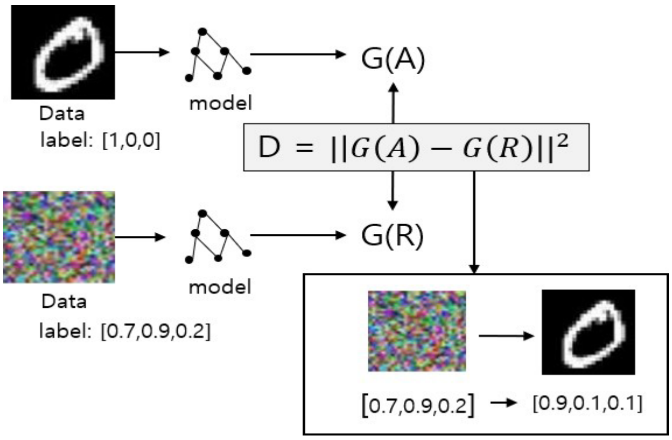

Figure 2. Reconstruction attack suggested in previous studies

3 Proposed Method

3.1 Reconstruction Attack Using Label Guessing

This study confirms that the reconstruction attack approaches the original through optimization of random data and random labels and presents a label guessing method by applying a mode. The mode uses the most frequent value and stores the location of the high value of the collected label for each termination. This method has better performance than a random label. This is because it shows similar performance to the known label in 2.1. when guessing which label method used the mode. As the Iteration progresses, a large number of positions for each label are checked. The most common value is given as 1, and the rest are given as 0. An attacker must match both data and labels. The flow of the label guessing method is shown in Figure 3.

However, it has the effect of increasing the speed of reconstruction because it only matches the data through the effect of providing labels. For example, if the label guessing method is determined in the 10–30 section of the Iteration, if the random label value is [0.1, 0.3, 0.7, 0.2, 0.1, 0.5, 0.3, 0.1, 0.5], the highest value is 0.7, and the position is 2 out of 0 to 9. If the random label value is [0.6, 0.3, 0.1, 0.2, 0.1, 0.5, 0.3, 0.2, 0.1, 0.5] in the Iteration = 20 section, the highest value is 0.6, and it is placed at zero. If the random label value [0.8, 0.3, 0.7, 0.2, 0.1, 0.5, 0.3, 0.2, 0.1, 0.5] in the Iteration = 30 section, the highest value becomes 0.8, and is located at the first position. As a result of the guess, the label position is {2, 0, 0}, and because 0 is the largest, 0th 1 is added, and 0 is given. The final label is [1, 0, 0, 0, 0, 0, 0, 0, 0, 0.]. G(A) of the input is the slope received by the client, and G(R) is the slope from which the server learns the model with random data and random label. The client was sent to the server through local training. The parameters are Iteration and Optimizer, which determine how long the cycle will be rotated, and Optimizer is used to optimize data and labels. By optimizing random data and random labels until D approaches zero according to the algorithm, data and labels close to the original can be checked. D was calculated using Euclidean distance, as in Zhu et al.'s study [9].

Algorithm 1. Reconstruction attack using label guessing method

```
input = $G(A)$, $model$, Random data, Random label
parameter =$iteration$, $optimizer$
optimizer = $LBFGS$, $model$ = $CNN$
for $i$ in $iteration$ do
    $G(R)$ ←model($Random\ data$, $Random\ label$)
    $Return\ G(R)$
    if $i > 30$ and $i$  90 then
        Guessing = model ← $Random\ label$
    else $i > 90$
        argmax(Guessing)
        mode(Guessing)
        Random label ← $label[0, Guessing] = 1$
    end if
        $D = ||G(A) - G(R)||^2$
        $D.backward(\,)$
        $LBFGS$(Random data, Random label)
end for
Random data, Random label
```

3.2 Experiment and Evaluation

This study conducted experiments by dividing them into two situations: one in which the client has one image (3.2.1 Reconstruction Attack Using Label Guessing - One Image) and another in which the client has 10 batches of data (3.2.2 Reconstruction Attack Using Label Guessing - Batch Image), to confirm the reconstruction performance of guessed labels. Although the situation of having only one image is rare, comparisons under similar conditions to the first reconstruction study [9] are also important. In federated learning, reconstruction attacks use the gradients from one round during the n-round federated learning process to reconstruct the model. Therefore, both the experimental settings of federated learning and the reconstruction attack are necessary to conduct reconstruction attack experiments in federated learning.

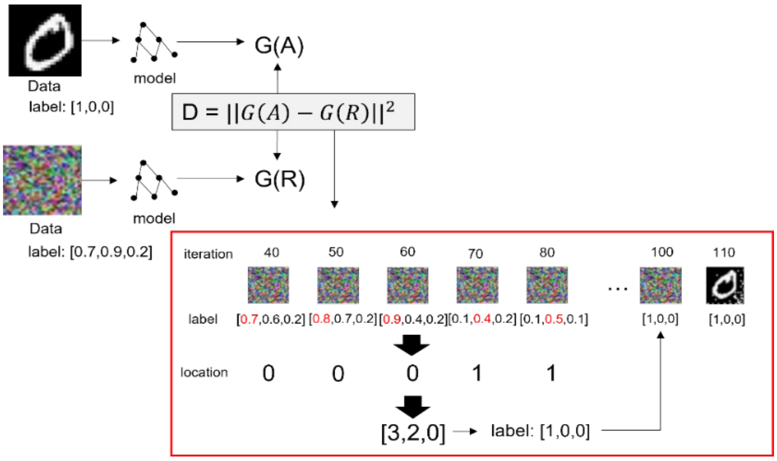

Figure 3. Reconstruction attacks using label guesses

(The high value of the label was extracted for each iteration, and the most frequently found value was designated as a label.)

The datasets used in the experiment are the Yale Face Database B, CIFAR-10, and MNIST. Yale Face Database B is a dataset of grayscale images of 38 people's faces under approximately 60 lighting conditions [16]. As it contains images in various lighting conditions, including dark images where facial identification is impossible, only images with a mean pixel value greater than 64 were selected by dividing the pixel values between 0 and 255 into four levels. Each person is given a class, and there are 37 classes in total with only classes with 20 or more images to consider the test data size in federated learning. The data has a size of 192 x 168, but it was resized to 48 x 42 to reduce the experimental time. CIFAR-10 consists of 60,000 32 x 32 images classified into ten classes. The MNIST dataset consists of grayscale images of handwritten digits from 0 to 9, with a size of 28 x 28. Both CIFAR-10 and MNIST have ten classes, and they were used to test the experiment in various environments. The number of clients was fixed at ten, and the per-client training data size (per), batch size (bs), and federated learning rounds were set to various values. Test data consisted of 10 data per class for each client, for a total of 100 data sets. Reconfiguration iterations that update data are set to 500 & 1000 for one-image and 1000 for batch-image, as higher iterations lead to better reconstruction performance but require more time. As mentioned in previous reconstruction attack studies, both random values and actual labels were used as pseudo-labels, and the DLG (dlg) was used for reconstruction attack, with the DLG + TV norm (dlg+tv) added. The reconstruction results were shown depending on the label situation.

The experimental results of this paper show the dlg performance (section 3.2.1) when reconstructing one image by using a model trained with FedAVG on the CIFAR-10 and MNIST datasets. In addition, the paper presents the results of MSE, SSIM, and PSN in tables and figures for dlg and dlg+TVloss using batches of images trained with FedSGD and FedAVG on three datasets with random, known, and guessing labels in round 1 (section 3.2.2).

Three frequently used evaluation methods were used to determine the extent of the reconstruction. Mean squared error (MSE): This metric indicates the distance of the values. $\hat{Y}_i$ is the actual observation value, and Y_i is the predicted value. A smaller MSE value of $\hat{Y}_i$ and Y_i indicates a smaller difference between the two images to be compared [17]. Peak signal-to-noise ratio (PSNR): When evaluating image quality loss information, such as in images and videos, a larger value indicates a smaller difference in the image. MAX^2 denotes the maximum value of the corresponding image [18]. Structural similarity index measure (SSIM): This is not a numerical difference, but rather a method for evaluating the differences in the visual image quality and similarity of humans. μ_x and μ_y represent the average values of the original and random data, respectively, σ_x, σ_y is the standard deviation, σ_{xy} is the covariance, and c is the variable [19]. A higher SSIM value indicates better quality. The equations for the above three metrics are as follows:

$$MSE = \frac{1}{2}\sum_{i}^{n}(\hat{Y}_i - Y_i)^2. \tag{3}$$

$$PSNR = 10 \cdot \log_{10}(\frac{MAX_i^2}{MSE}). \tag{4}$$

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(2\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}. \tag{5}$$

3.2.1 Reconstruction Attack Using Label Guessing – One Image

Each experiment involved the reconstruction of one image. Furthermore, 30 images were randomly extracted to reconstruct 10 classes, and the average of the results was used with the aim of generating different random data and labels, even within the same class. Therefore, random seeds were used to prevent the generation of the same random data. As MNIST consists of iterations 0 to 500 and CIFAR-10 contains color images, many calculations would be necessary to optimize data and labels; thus, iterations 0 to 1000 were applied. The MNIST data were set from iterations 0 to 500, whereas the CIFAR-10 data were set from iterations 0 to 1000. We used guessing sections 40–90 for the reconstruction. Sections 10–30 had a large label error; therefore, the guess section was set in the middle. As a result, the results of MSE, SSIM, and PSNR are shown numerically and graphically when the label is round 1, dlg in random, known, and guessing in two datasets. Table 1 to Table 2 presents the evaluations applied according to the label method. The data were reconstructed by extracting 1 data point for each class. The values were averaged, and the performance was compared at the end of the iterations.

Table 1. Performance of proposed method on MNIST dataset

	ALL class average		
	Label method		
Measure	Random	Known	Guessing
MSE	0.00002387	0.00000031	0.00000015
PSNR	105.1906344	117.6884047	120.9720058
SSIM	0.99959013	0.99999672	0.99999841

Table 2. Performance of proposed method on CIFAR-10 dataset

	ALL class average		
	Label method		
Measure	Random	Known	Guessing
MSE	0.00663681	0.00019171	0.00019157
PSNR	79.81313647	90.48795321	90.48979413
SSIM	0.92198748	0.99237666	0.99258724

We confirmed that the reconstructed results differed according to the label method. In particular, the known method was the best, and the reconstruction result was accurate when the guessing method was applied. As illustrated in Figure 4 and Figure 5, the reconstruction performance was better when it was known and guided than when it was random in all evaluations. Figure 6 presents a graph that confirms this trend. The green line indicates that the guessing label was applied, the red line is a known label, and the blue line is a random label. The guessing label exhibited the same trend as the known label for iterations 100 and higher, and it can be observed that the performance became the same as that of the known label after the method was applied.

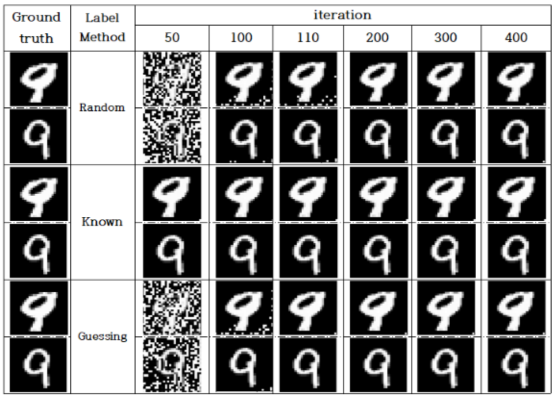

Figure 4. Reconstruction performance of proposed methods by iteration in MNIST dataset

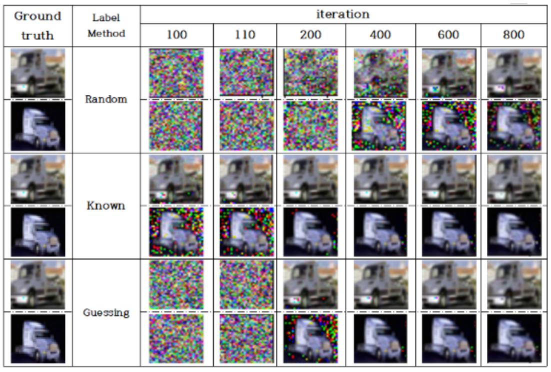

Figure 5. Reconstruction performance of proposed methods by iteration in CIFAR-10 dataset

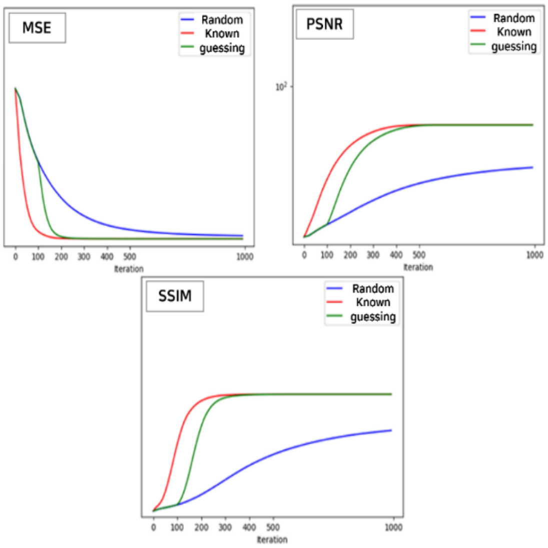

Figure 6. Performance graph of the proposed method on CIFAR-10 dataset

Thus, the performance was better than the guessing label of the random label, and the information in the gradient affected the label as well as the data. The label information significantly affected the reconstruction results.

3.2.2 Reconstruction Attack Using Label Guessing – Batch Image

Realistic federated learning environments typically use non-IID data, which is not independent and identically distributed, in contrast to IID data that is independent and identically distributed and is distributed equally among clients. To set up the experimental environment to non-IID data that fits the realistic federated learning environment, one class is assigned to each client, and clients are given 10 data each to create a total of 100 data sets for reconstruction. Additionally, the experiment compared Fed-AVG with Fed-SGD with batch learning in federated learning. Prior to this experiment, to compare random and guessing, the part with good performance compared to random was marked in blue, and the part with low performance was marked in red. Because the reconstruction of one image results in one reconstructed image, the evaluation index calculation can be performed by directly inputting the pixel value of the original image into the formula. However, when there are multiple images to be reconstructed, there will also be multiple reconstructed images. In these cases, the images are reconstructed using the gradients of the data contained in one batch. If a similar image exists within that batch, it can be difficult to identify the appropriate original image-reconstructed image pair. Therefore, for the performance calculation of batch images in this paper, MSE, PSNR, and SSIM are calculated for all possible source image-reconstructed image pairs, and the reconstructed image with the lowest MSE and the highest (PSNR, SSIM) values is considered to have the best performance. If none of the three images point to the same image, the two images with the most similar properties are considered to be the original and reconstructed images.

Table 3 and Table 4 show the results of 10,000 rounds of federated learning using MNIST and the results of reconstruction attacks in round 1 during the federated learning process. Both FedSGD and FedAVG perform better on dlg+tv than dlg. Looking at the label type, the dlg results of FedSGD and the dlg+tv results of FedAVG perform better than the random label, while the dlg+tv results of FedSGD perform less well. However, each performance MSE, PNSR, and SSIM show a very small difference.

Figure 7 and Figure 8 show the reconstructed images of the MNIST dataset for each iteration of the reconstruction process. As the number of iterations increases, the reconstructed image becomes clearer until it reaches iteration 990. Both FedSGD and FedAVG produce clearer reconstruction images using dlg+tv compared to dlg. However, in contrast to the reconstruction of a single image, the reconstruction of a batch image remains relatively noisy. Moreover, unlike the results of applying the proposed method to a single image, the results of applying the proposed method to the batch image show little difference in the reconstructed image based on the label type. This implies that there should be a difference in the reconstruction performance because the label up to 100 iterations is used when obtaining the guessing label.

Table 3. Performance of proposed method on MNIST dataset (Fed SGD)

Method		dlg			dlg+tv		
Label method		Random	Known	Guessing	Random	Known	Guessing
Attack	MSE	0.1804	0.0900	0.0945	0.0224	0.0233	0.0229
	PSNR	9.6322	10.6181	10.4238	16.7848	16.6515	16.7221
	SSIM	0.3681	0.4091	0.3952	0.5555	0.5473	0.5518
FL	acc	100 (%)					
	loss	1.4612					

Table 4. Performance of proposed method on MNIST dataset (Fed AVG)

Method		dlg			dlg+tv		
Label method		Random	Known	Guessing	Random	Known	Guessing
Attack	MSE	0.1692	0.0793	0.1483	0.0282	0.0171	0.0171
	PSNR	9.9415	11.3366	10.2650	17.0389	18.0801	18.1030
	SSIM	0.4122	0.4485	0.4157	0.5777	0.6181	0.6195
FL	acc	100 (%)					
	loss	1.4612					

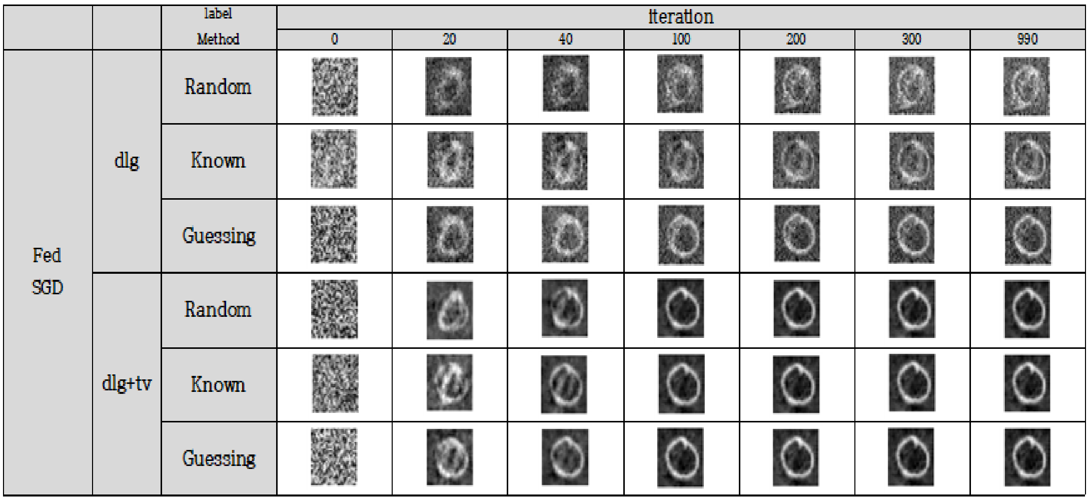

Figure 7. Reconstruction performance of proposed methods by iteration in MNIST dataset (Fed SGD)

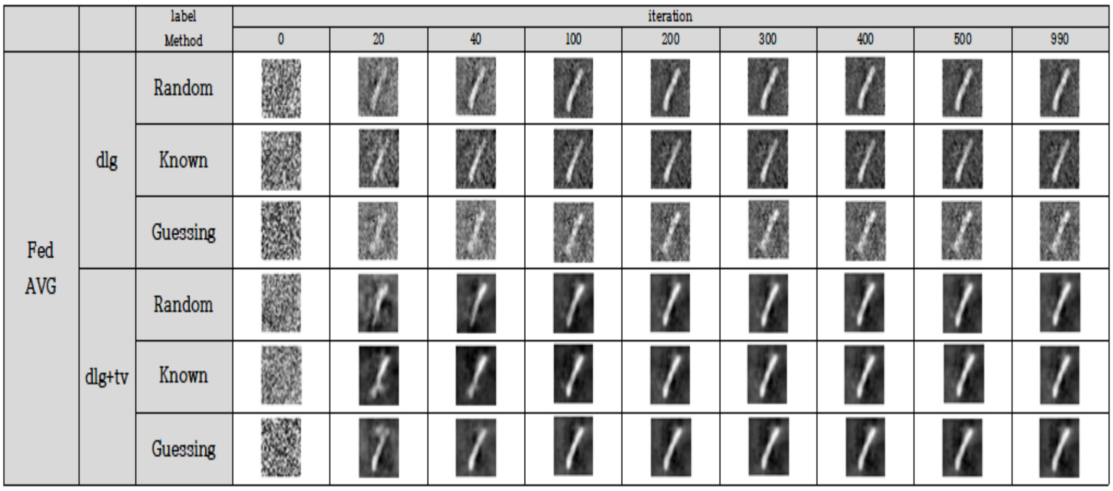

Figure 8. Reconstruction performance of proposed methods by iteration in MNIST dataset (Fed AVG)

However, because both Figure 7 and Figure 8 show similar reconstructed images at iteration 100, the effect of the guessing label is not noticeable. Table 5 and Table 6 present the results of federated learning and reconstruction experiments using CIFAR-10. The results show that with the performances of FedSGD's dlg+tv and FedAVG's dlg, dlg+tv is better than that of random labels. In the case of dlg results from FedSGD, the reconstruction performance was very poor, and it was not possible to find the original image-reconstruction image pair. Unlike MNIST, which is a black-and-white image dataset, CIFAR-10 failed to reconstruct well, possibly due to its color images.

Table 5. Performance of proposed method on CIFAR-10 dataset (Fed SGD)

Method		dlg			dlg+tv		
Label method		Random	Known	Guessing	Random	Known	Guessing
Attack	MSE	-	-	-	0.1661	0.1480	0.1463
	PSNR	-	-	-	10.9825	12.1486	12.3848
	SSIM	-	-	-	0.0967	0.0990s	0.1163
FL	acc	100 (%)					
	loss	1.4612					

Table 6. Performance of proposed method on CIFAR-10 dataset (Fed AVG)

Method		dlg			dlg+tv		
Label method		Random	Known	Guessing	Random	Known	Guessing
Attack	MSE	0.2027	0.1074	0.1142	0.0257	0.0248	0.0254
	PSNR	8.5125	9.7174	9.4734	16.2106	16.4223	16.2821
	SSIM	0.0407	0.0447	0.0464	0.4209	0.4246	0.4216
FL	acc	100 (%)					
	loss	1.4612					

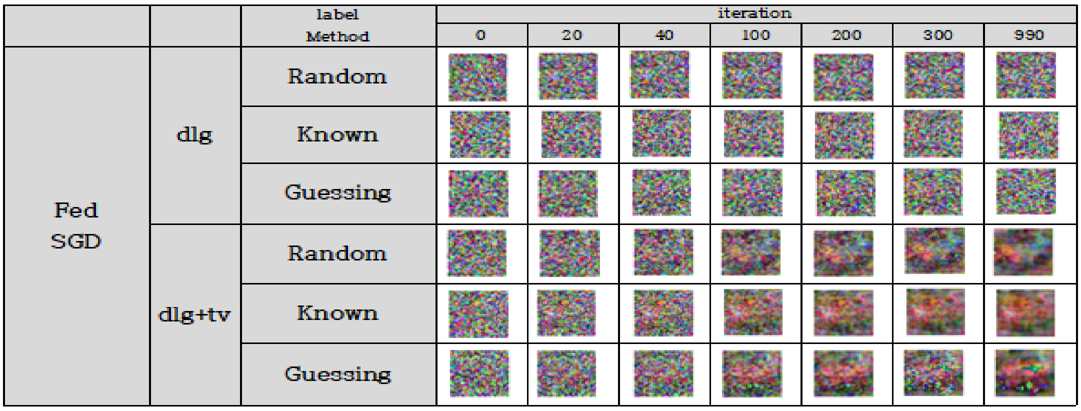

Figure 9. Reconstruction performance of proposed methods by iteration in CIFAR-10 dataset (Fed SGD)

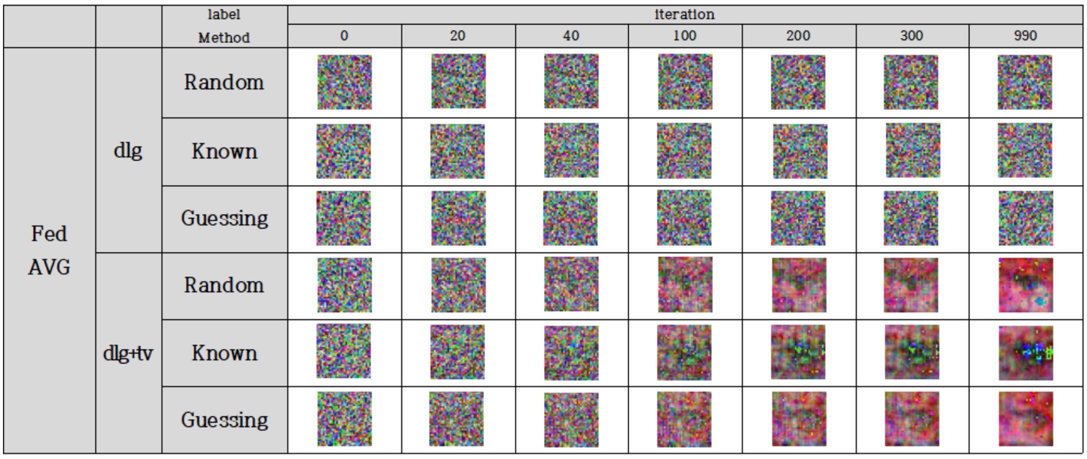

Figure 10. Reconstruction performance of proposed methods by iteration in CIFAR-10 dataset (Fed SGD)

Figure 9 and Figure 10 show the reconstruction performance of CIFAR-10, which is not as clear as the reconstruction results for MNIST. Because CIFAR-10 is a color image dataset in the batch image, it is difficult to converge into a single image during a reconstruction attack. Both FedSGD and FedAVG show that using dlg+tv rather than dlg helps with the reconstruction process.

Table 7. and Table 8. present the federated learning performance and reconstruction results using the Yale dataset according to the label, where the reconstruction methods are dlg and dlg+tv. The results of the experiment were better when using random labels compared to the proposed method. As Yale dataset consists of face images, more pixel information needs to be considered, and the data and labels diverged according to gradients during the process of inferring and reconstructing labels. The performance and loss of FedSGD and FedAVG are lower than those of MNIST and CIFAR-10.

Figure 11 and Figure 12 show the reconstruction performance of the Yale dataset. Unlike CIFAR-10, the reconstructed images are clearly visible. It appears that dlg+tv is more effective than dlg for reconstruction, and applying the proposed guessing label method to FedSGD shows better performance than using random labels, starting from iteration 100.

Table 7. Performance of proposed method on Yale dataset (Fed SGD)

Method		dlg			dlg+tv		
Label type		Random	Known	Guessing	Random	Known	Guessing
Attack	MSE	0.2429	0.1778	0.4349	0.0491	0.0345	0.1409
	PSNR	8.1416	9.6404	6.1859	15.9709	15.9414	13.9135
	SSIM	0.0817	0.0998	0.0652	0.4966	0.4910	0.4248
FL	acc	97 (%)					
	loss	2.6863					

Table 8. Performance of proposed method on Yale dataset (Fed AVG)

Method		dlg			dlg+tv		
Label type		Random	Known	Guessing	Random	Known	Guessing
Attack	MSE	0.0850	0.0824	0.1745	0.0192	0.0178	0.0202
	PSNR	10.8477	10.9529	9.8771	18.2165	18.2893	18.1202
	SSIM	0.1379	0.1428	0.1244	0.6481	0.6625	0.6419
FL	acc	96 (%)					
	loss	2.6963					

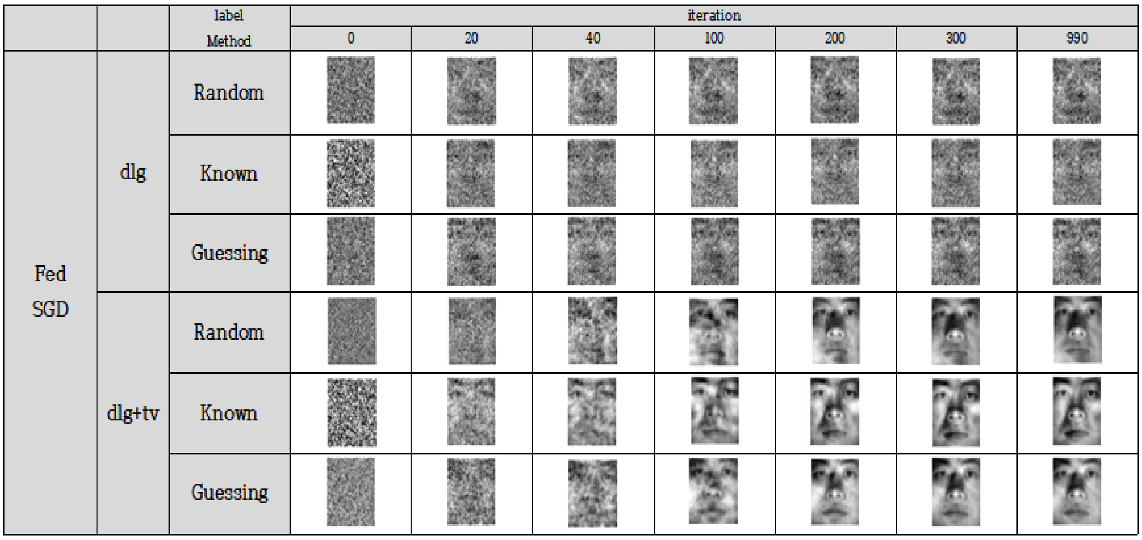

Figure 11. Reconstruction performance of proposed methods by iteration in Yale dataset (Fed SGD)

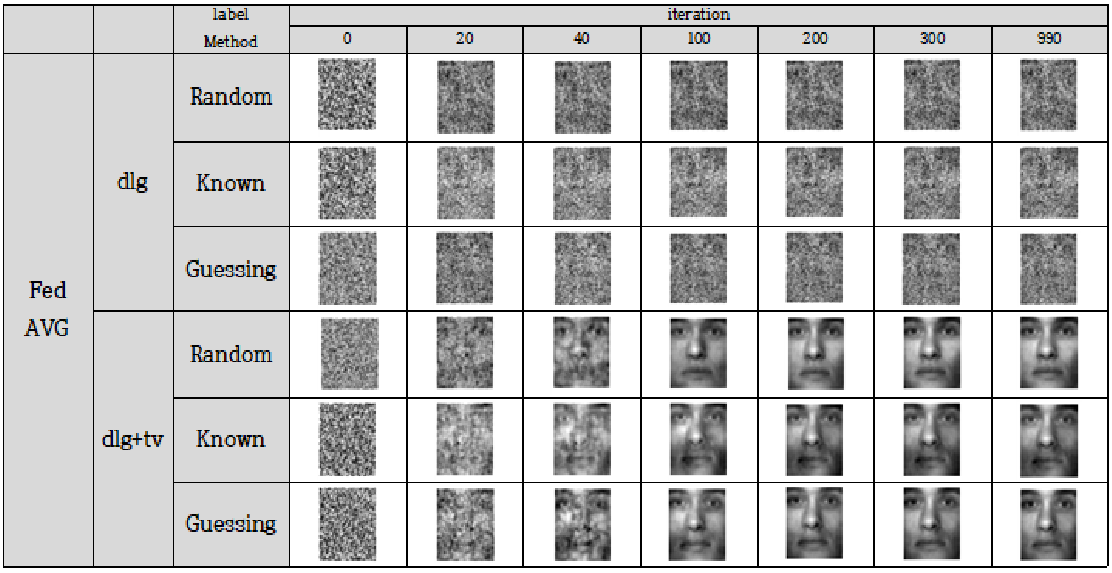

Figure 12. Reconstruction performance of proposed methods by iteration in Yale dataset (Fed AVG)

4 Discussion

Federated learning and reconstruction attacks were conducted on MNIST, CIFAR-10, and Yale datasets in this study. The experiment considered three attacker situations: random label, known label, and guessing label, which infers the actual label based on the label's most frequent value up to the halfway point of reconstruction. The proposed guessing label showed highly effective reconstruction performance in an environment with one image, but did not perform effectively in batch environments using multiple images and known labels, which were expected to perform best. According to a previous study [20], it becomes more difficult to find a matching gradient pair (original gradient-random gradient) the more data you want to reconstruct from an attack that uses slope to reconstruct data. Therefore, this paper's experiment confirmed the difference between the reconstruction result in one image and the reconstruction result in the batch image. In addition, it can be seen that the reconstruction results for color images are very poor compared to those for black and white images. This is likely due to the difficulty in finding a gradient pair of color images consisting of three channels compared to finding a pair of gradients of black and white images consisting of one channel, similar to why it is difficult to reconstruct batch images.

5 Conclusion

In this study, the data was reconstructed in the method proposed during federated learning. The performance of federated learning increased every time the round was conducted. We obtained gradients from clients while conducting federated learning on multiple datasets. Various situations of the attacker were presented from the obtained gradient. In one-image, we found that the reconstruction speed was faster and the reconstruction performance was higher than when the existing method with labels was randomly used. The batch image showed a somewhat weak effect. The experiment revealed that the situation in which the attacker knew the label exhibited the best performance. However, limitations exist owing to the nature of federated learning, in which known situations are rare. We have proposed a guessing label with similar performance to that of a known label, as if the attacker knows the label. If research on reconstruction attacks continues to develop in this manner, federated learning can violate privacy. Reconstruction attacks on federated learning are steadily coming out every year, and this will be an important challenge. The reason is that the use of facial data has increased in recent years. If certain parts of the body are used as learning data for federated learning, privacy violations will become serious. In addition, exposure to data leads to privacy violations, as well as posture maintenance and backdoor attacks [21-23]. Thus, attack and defense research should be further developed to analyze the factors acting on the attack and to ensure that there is no damage to federated learning in the future.

Acknowledgements

This work was supported by the Soongsil University Research Fund (New Professor Support Research) of 2021.

References

[1] L. Sweeney, K-anonymity: a model for protecting privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 5, pp. 557-570, October, 2002.

[2] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, *Foundations and Trends® in Theoretical Computer Science*, Vol. 9, No. 3-4, pp. 211-407, August, 2014.

[3] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman,

Synthetic data and artificial neural networks for natural scene text recognition, December, 2014, https://arxiv.org/abs/1406.2227.

[4] J. Y. Kim, M. -J. Park, Multiple imputation and synthetic data, *The Korean Journal of Applied Statistics*, Vol. 32, No. 1, pp. 83-97, 2019.

[5] Q. Li, Z. Wen , Z. Wu, S. Hu, N. Wang, Y. Li , X. Liu, B. He, A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 35, No. 4, pp. 3347-3366, April, 2023.

[6] T. Li, A. K. Sahu, A. Talwalkar, V. Smith, Federated learning: Challenges, methods, and future directions, *IEEE Signal Processing Magazine*, Vol. 37, No. 3, pp. 50-60, May, 2020.

[7] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, H. B. McMahan, T. Van Overveldt, D. Petrou, D. Ramage, J. Roselander, Towards federated learning at scale: System design, *Proceedings of the 2nd SysML Conference*, Stanford, CA, USA, 2019, pp. 1-15.

[8] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Ft. Lauderdale, FL, USA, 2017, pp. 1273-1282.

[9] L. Zhu, Z. Liu, S. Han, Deep leakage from gradients, *33rd Conference on Neural Information Processing Systems (NeurIPS2019)*, Vancouver, BC, Canada, 2019, pp. 14774-14784.

[10] M. Persson, D. Bone, H. Elmqvist, Total variation norm for three-dimensional iterative reconstruction in limited view angle tomography, *Physics in Medicine & Biology*, Vol. 46, No. 3, pp. 853-866, March, 2001.

[11] B. Zhao, K. R. Mopuri, H. Bilen, *iDLG: Improved deep leakage from gradients*, January, 2020, https://arxiv.org/abs/2001.02610.

[12] W. Wei, L. Liu, M. Loper, K.-H. Chow, M. E. Gursoy, S. Truex, Y. Wu, A framework for evaluating client privacy leakages in federated learning, *25th European Symposium on Research in Computer Security (ESORICS)*, Guildford, Surrey, UK, 2020, pp. 545-566.

[13] A. Wainakh, F. Ventola, T. Müßig, J. Keim, C. G Cordero, E. Zimmer, T. Grube, K. Kersting, M. Mühlhäuser, User-level label leakage from gradients in federated learning, *Proceedings on Privacy Enhancing Technologies*, Vol. 2022, No. 2, pp. 227-244, April, 2022.

[14] J. Geiping, H. Bauermeister, H. Dröge, M. Moller, Inverting gradients - how easy is it to break privacy in federated learning? *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS2020)*, Online Conference, Canada, 2020, pp. 16937-16947.

[15] Z. Zhang, M. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, *32nd Conference on Neural Information Processing Systems (NeurIPS2018)*, Montreal, Quebec, Canada, 2018, pp. 8792-8802.

[16] A. S. Georghiades, P. N. Belhumeur, D. J. Kriegman, From few to many: Illumination cone models for face recognition under variable lighting and pose, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 6, pp. 643-660, June, 2001.

[17] A. Botchkarev, *Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology*, September, 2018, https://arxiv.org/abs/1809.03006.

[18] A. Hore, D. Ziou, Image quality metrics: PSNR vs. SSIM, *20th International Conference on Pattern Recognition (ICPR'10)*, Istanbul, Turkey, 2010, pp. 2366-2369.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600-612, April, 2004.

[20] X. Jin, P. Y. Chen, C. Y. Hsu, C. M. Yu, T. Chen, CAFE: Catastrophic data leakage in vertical federated learning, *34th Conference on Neural Information Processing Systems (NeurIPS2021)*, Online Conference, Canada, 2021, pp. 994-1006.

[21] D. Cao, S. Chang, Z. Lin, G. Liu, D. Sun, Understanding distributed poisoning attack in federated learning, *International Conference on Parallel and Distributed Systems (ICPADS 2019)*, Tianjin, China, 2019, pp. 233-239.

[22] C. Fung, C. J. M. Yoon, I. Beschastnikh, *Mitigating sybils in federated learning poisoning*, July, 2020, https://arxiv.org/abs/1808.04866.

[23] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov, How to backdoor federated learning, *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020*, 2020, pp. 2938-2948.

Biographies

JinHyeok Jang, received the M.S. degree in software convergence from Soongsil University, He is a currently pursing Ph.D. degree in software with Soongsil University, main research directions include Certification, Machine Learning, Federated Learning, AI Security, Financial security, Reinforcement Learning.

Yoonju Oh, received the B.S. degree in applied mathematics from Kongju National University. She is a currently pursing M.S. degree in software with Soongsil University, main research directions include Certification, Machine Learning, Federated Learning, AI Security, Reinforcement Learning.

Gwonsang Ryu, received the Ph.D. degree in software convergence from Soongsil University, He is a currently research professor with Soongsil University Cyber Security Research Center, main research directions include adversarial Attack, Adversarial Defense, Anomaly Detection.

Daeseon Choi, received the Ph.D. degree in computer science from Korea Advanced Institute of Science and Technology (KAIST), He is a currently professor with the Department of Soongsil University, main research directions include Privacy protection & Evaluation, AI security, Biometric/Behavior Authentication, Fraud Detection.