Mitigating gradient inversion attacks in federated learning with frequency transformation.

PALIHAWADANA, C., WIRATUNGA, N., KALUTARAGE, H. and WIJEKOON, A.

2024

Mitigating Gradient Inversion Attacks in Federated Learning with Frequency Transformation

Chamath Palihawadana[1], Nirmalie Wiratunga[1], Harsha Kalutarage[1], and Anjana Wijekoon[1]

School of Computing, Robert Gordon University, Aberdeen, UK
{c.palihawadana,n.wiratunga,h.kalutarage,a.wijekoon1}@rgu.ac.uk

Abstract. Centralised machine learning approaches have raised concerns regarding the privacy of client data. To address this issue, privacy-preserving techniques such as Federated Learning (FL) have emerged, where only updated gradients are communicated instead of the raw client data. However, recent advances in security research have revealed vulnerabilities in this approach, demonstrating that gradients can be targeted and reconstructed, compromising the privacy of local instances. Such attacks, known as gradient inversion attacks, include techniques like deep leakage gradients (DLG). In this work, we explore the implications of gradient inversion attacks in FL and propose a novel defence mechanism, called Pruned Frequency-based Gradient Defence (pFGD), to mitigate these risks. Our defence strategy combines frequency transformation using techniques such as Discrete Cosine Transform (DCT) and employs pruning on the gradients to enhance privacy preservation. In this study, we perform a series of experiments on the MNIST dataset to evaluate the effectiveness of pFGD in defending against gradient inversion attacks. Our results clearly demonstrate the resilience and robustness of pFGD to gradient inversion attacks. The findings stress the need for strong privacy techniques to counter attacks and protect client data.

Keywords: Gradient Inversion Attacks · Federated Learning · Frequency Transformation.

1 Introduction

The widespread adoption of Machine Learning (ML) and the increasing need for large-scale privacy sensitive data have led to the emergence of Federated Learning (FL) [5]. FL provides a decentralised approach to train ML models, enabling privacy preservation in the process. In a typical FL setting there will be a server which orchestrates the federated rounds where each client contributes a local model update trained on their private data. With this process the client's private data never leaves their device which gives a strong privacy guarantee. This privacy-preserved nature of FL has gathered significant attention and interest from various domains, including healthcare, finance, and smart devices.

The decentralised nature of FL ensures that clients only communicate their local updates, such as gradients or local model parameters, which significantly enhances the safety of the process compared to sharing raw data with a central system. However, recent research on attack scenarios has revealed potential vulnerabilities even when only local updates are shared [12]. Such attacks are known as gradient inversion attacks which is an active area of research in FL. Some of the suggested defence strategies for gradient inversion attacks includes adding noise, gradient compression, training with large batch sizes and complex models. We note that methods like compression can be advantageous overall for the FL setting as well as to defend such attacks. A key challenge identified in recent literature is the trade off between model performance and communication efficiency [10].

This work aims to explore and establish a novel research direction that utilises the frequency space as a means to defend against gradient inversion attacks in FL. By investigating and positioning the potential of utilising frequency space in this context, we aim to provide valuable insights and propose an effective strategy to counter the vulnerabilities posed by gradient inversion attacks within the FL setting. This paper presents two key contributions. Firstly, it introduces pFGD, a straightforward yet highly effective defense strategy specifically designed to mitigate gradient inversion attacks in FL. Secondly, the paper conducts a comprehensive comparative study, evaluating the performance and efficacy of pFGD against two commonn gradient inversion attacks.

2 Background

2.1 Federated Learning

FL setting introduced a paradigm to perform ML model training in a decentralised manner. FL became widely known and adapted due to its privacy preserved manner, where client training data is never exposed or communicated to the server. This privacy preserved nature enabled to do ML model training on sensitive data like healthcare and finance. A typical FL setting consists of a server and a large number of clients participating in many communication rounds. The FL process typically commences at round $t = 0$, where the server distributes an initial global model (w_0) to all participating clients. At each communication round t, the server selects K clients to engage in local training. Each client k independently conducts training on its private data and, upon completion, communicates the updated model parameters or gradients back to the server. The server then aggregates these models using methodologies such as *FedAvg* [6] and *FedSim* [6], resulting in an updated global model at $t + 1$. Since the clients data is never communicated to the server there is a natural privacy guarantee with FL setting.

2.2 Attacks in Federated Learning

In FL setting the threat surfaces are more exposed unlike in traditional ML setting. The network of clients and the communication layer in FL setting can be considered the largest threat surface [2]. A taxonomy presented by [2] organises attacks in FL setting into two types; model performance and privacy attacks. Model performance attacks are performed during the training phase using poisoning attacks. By poisoning (i.e. manipulating) the model or local data it is possible to degrade the overall performance of the model.

Privacy attacks in FL is a widely researched area due to the impact and risk of exposure. FL is considered to be a privacy-preserved machine learning paradigm as the private training data never leaves the client-side. Security researchers have demonstrated attacks to extract such private data in the communication stage or aggregation stage at the server (as demonstrated in Figure 1). These types of attacks can be categorised further more as gradient inversion, membership inference and generative adversarial network (GAN) reconstruction attacks. Gradient inversion attacks have demonstrated the capability to reconstruct the classes and individual data instances just by using the communicated client gradients [8, 11, 12]. In this work we explore defending privacy attacks in FL and specifically gradient inversion attacks.

2.3 Gradient Inversion Attacks in FL

A recent survey [9] proposed a taxonomy for gradient inversion attacks characterising into two paradigms. The two paradigms are iteration and recursion based attacks. Iteration based attacks first generates a pair of random (dummy) data and labels, then by performing forward and backward propagation iteratively the gradients can be optimised for data recovery. The reconstruction of private data is viewed as an iterative process using gradient descent. When the distance between the original and the generated gradients are close the private data can be extracted. The second paradigm is when the attacker recursively calculate the input of each layer by finding the optimal solution with minimised error. We focus on the iteration-based attacks due to their adaptability and higher risk of exposing client privacy. There is a growing list of gradient inversion attacks, few of the widely used and studied methods include Deep Leakage from Gradient (DLG) [12], improved-DLG (iDLG) [11], Client Privacy Leakage (CPL) [8] and Inverting Gradients [1]. In this study, we employed DLG and iDLG techniques to investigate the impact of our proposed method.

2.4 Frequency Space Transformation in FL

Frequency space transformation techniques have long been utilised in data compression, with notable examples such as the Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT), Fast Fourier Transform (FFT), and Principal Component Analysis (PCA). Most commonly used technique is the DCT, mainly due to its computational efficiency and compact representation [7]. Previous work using the frequency space in FL are focused on compressing the data instance and not the communication of updated gradients. In this study we use DCT as the the frequency space transformation function to explore a practical defence to gradient inversion attacks in FL.

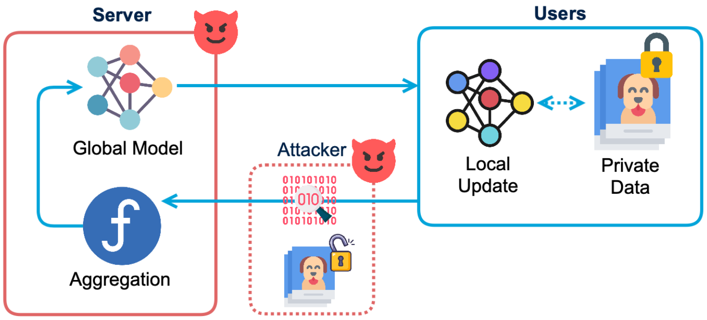

Fig. 1: Potential risks of gradient inversion attacks in FL

2.5 Attack Scenario

Threat Model We consider two potential attack surfaces to apply the proposed method. As a network eavesdropper: Communication from a clients device to the server can be compromised on the network layer by an attacker. As a curious server, the FL server can be compromised or honest-but-curious, potentially exploiting client training data. Figure 1 presents the threat model with respect to the FL setting. The attack surfaces are presented with a red border.

Adversarial Goal In the gradient inversion attacks the goal is to reconstruct client's private data and its class label through the communicated gradients.

3 Method

Recent work on gradient inversion attacks like DLG [12] and iDLG [11] has demonstrated the risk to privacy by exposing client private data. Both attacks attempt to reconstruct client data instances and labels using a gradient matching objective. In a typical FL setting gradients are shared to the server by clients after a local training step. If an attacker obtains such gradients they can reconstruct training instances (there are assumptions on these methods as discussed in their methods). Gradient inversion attacks can be performed at any round in the FL process, even before model convergence. In this section we discuss the attack methods and present the proposed defense method, referred to as pFGD.

3.1 Attack Methods

In this work we use two attack methods from literature which are iteration-based attacks. The selected attacks are DLG [12] and iDLG [11] which aim to reconstruct (steal) a FL client's local data instances using the communicated ΔW gradients. The attacker generates a pair of dummy data x' and dummy labels y' which are used to generate dummy gradients $\Delta W'$. Then by optimising the dummy gradients to be close to the client gradients the dummy data will be close to the real data. Equation 1 demonstrates the objective of the selected gradient inversion attacks. Where W is the shared global model, $F(.)$ shared optimisation function and x'^*, y'^* are the optimised results (i.e. reconstructed data).

$$x'^*, y'^* = \arg\min_{x', y'} ||\Delta W' - \Delta W||^2 = \arg\min_{x', y'} ||\frac{\partial l(F(x', W), y')}{\partial W} - \Delta W||^2 \quad (1)$$

A key difference between DLG and its improved version iDLG is that the way they extract the ground truth labels. Results presented by iDLG authors suggest a 100% accuracy rate on generating the label from the gradients unlike the DLG which are around 79%-90% on the same experiments.

3.2 Proposed Defence Method

We propose **Pruned Frequency-based Gradient Defence (*p*FGD)** which can act as a defence mechanism to such attacks while preserving model performance for FL setting. *p*FGD is a client-side frequency space based defence mechanism against DLG and iDLG. Once the local training is performed the updated gradients will be transformed into the frequency space $\widehat{\Delta W}$ using transformation function $T(.)$. Then pruned by a pruning function $P(.)$ controlled by α percentage. Figure 2 visually illustrates the workflow taking place on the client side, providing a clear representation of the various steps involved in *p*FGD.

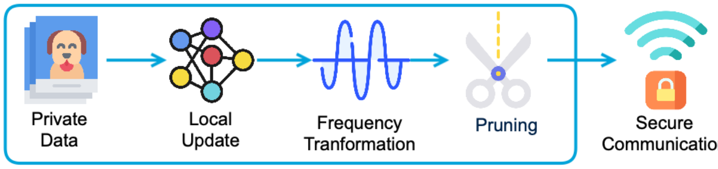

Fig. 2: Client-side workflow in *p*FGD

As discussed in Section 2.5 we assume the clients are honest and not a threat to the FL setting. Client communication of pruned frequency gradients prevents model inversion through noise and parameter reduction. *p*FGD transmission from the client mitigates risks from curious servers and network eavesdroppers.

Frequency Space Transformation Based on our preliminary study, we have chosen DCT-IV as the transformation function, denoted as $T(.)$. DCT-IV has been found to strike a balance between preserving model performance and enhancing communication efficiency through pruning on the frequency space. After the gradients undergo transformation into the frequency space using the DCT, the resulting coefficients are structured to preserve the necessary information for model aggregation. Additionally, the utilisation of the frequency space enables efficient pruning, as it allows for identifying and discarding coefficients with lower magnitudes without significantly compromising the overall model performance.

Parameter Pruning To defend against model inversion attacks such as DLG, incorporating noisy gradients can be beneficial. However, a significant challenge lies in determining an appropriate threshold for pruning gradients. The objective is to strike a balance where the pruned gradients introduce sufficient noise to thwart such attacks while still maintaining comparable performance. In the pruning function, denoted as $P(.)$, within our proposed method, we adopt a straightforward approach. We set the coefficients with the least frequency (corresponding to small magnitudes) obtained from the DCT transformation to zero. By zeroing out these coefficients, we effectively prune the model, reducing its size while aiming to retain essential information contained in the remaining coefficients.

3.3 Improving Resilience in FL

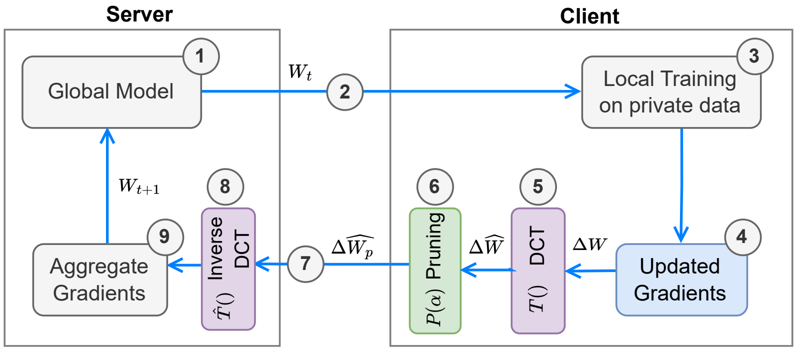

Fig. 3: Adapting pFGD to existing FL methodologies

The objective of this work is to introduce a method that strengthens the resilience of federated learning approaches against gradient inversion attacks. These attacks have the potential to compromise the fundamental benefits of FL, which is the preservation of client privacy. By incorporating the proposed method, pFGD, resilience can be achieved through the utilisation of a generalisable technique such as the frequency domain (the frequency space) and pruning. The pFGD method addresses the vulnerability to gradient inversion attacks by leveraging the inherent properties of the frequency space and pruning. Overall, the aim of this work is to establish a resilient FL methodology that effectively combats gradient inversion attacks, thus enabling the continued protection of client privacy, which is a core principle of FL.

Figure 3 illustrates the adaptation of *p*FGD to existing FL methodologies. This adaptation introduces Steps 5 and 6, specifically designed to enhance resilience against model inversion attacks in the FL setting. Step 8 is used to inverse the frequency space model to raw space before model aggregation. In Figure 3, Steps 2 and 7 represent the communication between the client and server, highlighting the potential vulnerability where an attacker can intercept and compromise the privacy of the system.

3.4 *p*FGD Algorithm

Based on the aforementioned considerations, Algorithm 1 outlines the workflow required to implement *p*FGD. Note that the algorithm incorporates a reference to the attacker method, which assumes the attacker possesses knowledge of inverting the DCT through the inverse transformation function $\hat{T}(.)$.

Algorithm 1 Pruned Frequency-based Gradient Defence

```
Require: $W$: global model, $\alpha$: Pruning Rate
Require: $T(.)$ DCT Function, $P(.)$ Pruning Function
$\Delta W \leftarrow$ update $W$ using SGD on local data
procedure pFGD($\Delta W$, $\alpha$)
    $\Delta\widehat{W} = T(\Delta W)$                    ▷ DCT transformation
    $\Delta\widehat{W}_p = P(\Delta W, \alpha)$          ▷ Tranformed Space Pruning
    return $\Delta\widehat{W}_p$
end procedure
procedure ATTACKER($\Delta\widehat{W}_p$)
    $\Delta W \leftarrow \hat{T}(\Delta\widehat{W}_p)$   ▷ Inverse DCT transformation
    DLG($\Delta W$) or iDLG($\Delta W$)                  ▷ Perform Attack Scenario
end procedure
```

4 Experiment Setup

In this introductory study, we aim to introduce and examine the potential of adapting the proposed *p*FGD to defend against gradient inversion attacks. First we study the impact on privacy on communicating client gradients in the frequency space, then we explore to what extent can parameter pruning in the frequency space can defend gradient inversion attacks. To evaluate the impact on privacy by communicating model parameters in the frequency space we use two attack methods and one image dataset. DLG [12] and iDLG [11] are selected to study the performance of *p*FGD. The two methods will be compared with and without the DCT transformation during the communication phase.

Dataset We select MNIST [3] dataset which is a 10-class handwritten digit recognition image dataset. A single image dimensions are 28x28 and has one channel. MNIST is commonly used in FL and security benchmarks as it provides a realistic setting. MNIST's single-channel images aid performance assessment due to sensitivity to variations. Selecting MNIST for comparison with prior works enhances understanding of the approach against gradient inversion attacks.

Configuration We adopt the experimental settings from [11,12] to ensure consistency and comparability. For the attack scenarios, we utilize LBFGS [4] with a learning rate of 1, batch size of 1 and 100 attack iterations. To mitigate the influence of random bias, we conduct 1000 runs of the experiments on LeNet models randomly initialised (i.e. 1000 random initiliased models on a unique data instance). Experiments will terminate at the 100th iteration or if the loss is below 0.000001.

Pruning As highlighted in Section 3.2, pruning plays a significant role in introducing noise to the gradients, thereby diminishing the effectiveness of the attacks. In our experiments, we ensure consistency by using a fixed pruning rate of $\alpha = 1\%$, resulting in the pruning of 133 parameters. Additionally, we performed secondary experiments with a 0.1% pruning rate (11 parameters pruned) to ensure fair comparison and assess pruning's impact on pFGD's defense against gradient inversion attacks.

4.1 Comparative Study

To gain a comprehensive understanding of the impact of pFGD technique, we explore multiple variants of the selected baselines. Specifically, for the DLG and iDLG attack methods, we consider the following four variants: 1. Vanilla (original method without modifications), 2. Vanilla with pruning (pruning applied to the vanilla method), 3. DCT (applying only DCT transformation) and 4. DCT with pruning (pruning applied to the DCT transformed gradients) By examining these different variants, we can assess the effectiveness and comparative performance of pFGD in various configurations and scenarios. The experiment setup is publicly accessible on GitHub[1] for reproducibility.

Evaluation Metrics We log the Mean Squared Error (MSE) of the reconstructed instance and the original image at each iteration. These MSE values are used to analyse and evaluate the behaviour of the proposed method. By counting the number of successful bypasses at each threshold, we gain insights into the effectiveness of the different variants in defending against the respective attacks. By considering the minimum MSE value from each experiment ensures that we capture the reconstruction's performance under various conditions and iterations.

[1] https://github.com/chamathpali/pFGD

5 Results and Discussion

We conducted a comparative study of the four variants on two attack methods. Figure 4 visually presents the reconstructed images at different MSE threshold points, allowing for an assessment of their readability. By observing these visual representations, we can assess the success of the reconstructions and identify any potential leakage of private information. At MSE= 0.001 (red text color in Figure 4), the digit 9 becomes discernible upon closer examination. The results are presented in Figure 5, which illustrates the number of experiments that were able to surpass different MSE thresholds. In Figure 5 the bar plots consists of two colours, blue and orange which is for DLG and iDLG experiments respectively. Plots with the squared pattern represent the pruned variants and with diagonal patterns represent the DCT variants. In the graph legend, the notation '_P' represents the pruned variants. We observe when MSE= 1 there are only 24 and 12 experiments passing the threshold for DLG and iDLG respectively when DCT with pruning is applied. Additionally, we found that there were no reconstructions of DCT with pruning when the MSE was less than 0.9. In contrast, we found that reconstructions were identified even when the MSE reached a low value of 0.005 for pruning on the vanilla methods.

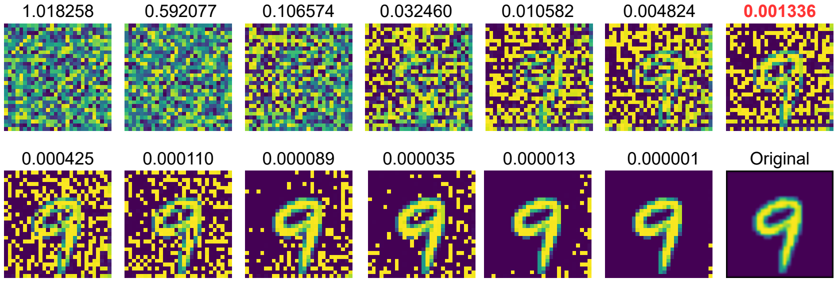

Fig. 4: Reconstructions of digit 9 are displayed at various MSE points, indicated above each image, ranging from higher to lower values. The final image presents the original digit 9 for comparison.

When using pruning on vanilla gradients without DCT there is still a high risk of leaking privacy sensitive information. With our experiments we were able to visually identify these reconstructions as the original image. For the MNIST dataset to properly identify a digit having a reconstructed image with MSE value of approximately 0.001 is sufficient for accurate digit identification. This cutoff point can differ from dataset to dataset and for different individuals eyesight. But our key observation is having pruning of $\alpha = 1\%$ is still able to be reconstructed on vanilla gradients. Additionally, we performed experiments with a pruning rate of $\alpha = 0.1\%$, results and plots are available on the GitHub repository.

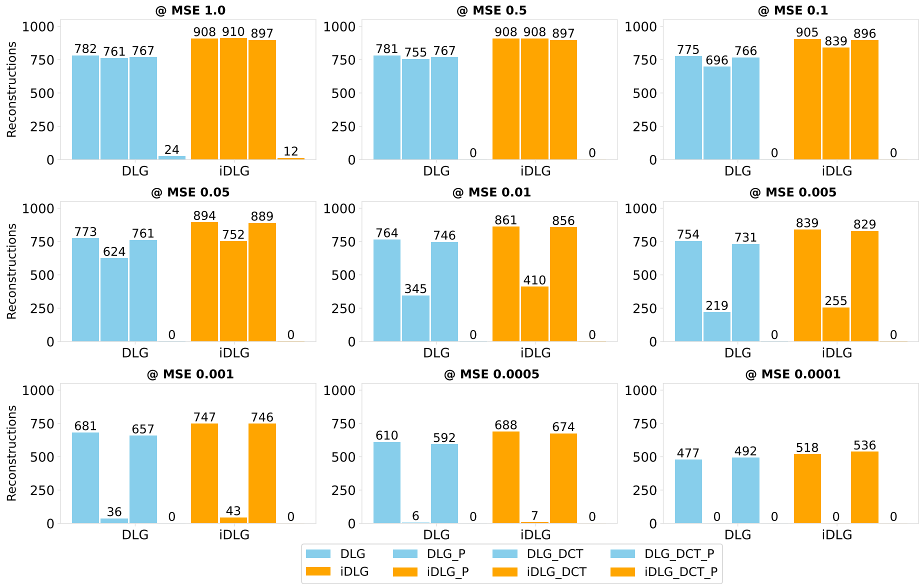

Fig. 5: Number of reconstructions at different MSE thresholds on MNIST dataset with $\alpha = 1\%$ with 4 variants on DLG and iDLG

The results obtained in our study provide compelling evidence that the combination of DCT with pruning techniques has significantly enhanced the defense against gradient inversion attacks. Throughout the 1000 experiment runs, we did not observe any reconstructions when applying DCT with pruning (pFGD) with an MSE below 0.9. These reconstructions lacked readability, rendering them essentially non-existent. In contrast, our research findings reveal that the application of pruning alone to the vanilla gradients, in the absence of employing the DCT, still poses a considerable risk of privacy breaches. For the MNIST dataset, our findings indicate that achieving a reconstructed image with an MSE of approximately 0.001 is sufficient for accurate digit identification. Around the MSE value of 0.005, we noticed a significant indication of a digit with potential lines emerging in the reconstructions. However, it is important to note that this threshold may vary across datasets and individual visual strengths. Our experiments visually demonstrated the identification of reconstructed images as the original ones in such cases. Taken together, these results highlight the resilience and efficacy of the proposed pFGD in countering gradient inversion attacks.

6 Conclusion

In this study, we introduced pFGD, a defense mechanism designed to mitigate gradient inversion attacks in federated learning. By applying frequency transformation using DCT on the updated gradients and incorporating pruning before communication, pFGD effectively enhances the resilience of FL models against such attacks. In our initial investigation, we conducted a comparative study involving two attack scenarios and four variants for each on the MNIST dataset. Our experimental results provide compelling evidence that utilising pFGD offers superior protection against gradient inversion attacks compared to pruning with raw gradients alone. Additionally, we observed that the implementation of pFGD using the frequency space does not lead to any performance degradation. One of the notable advantages of pFGD is its practicality, as it can be easily applied to different FL methodologies with minimal modifications.

Moving forward, we intend to expand our study by incorporating additional datasets and baselines to further evaluate the generalisability and robustness of pFGD. Overall, our findings highlight the effectiveness and potential of pFGD as a defence mechanism against gradient inversion attacks in FL. We anticipate that further exploration and refinement of pFGD will contribute to strengthening the security and privacy of FL setting.

References

1. Geiping, J., Bauermeister, H., Dröge, H., Moeller, M.: Inverting gradients-how easy is it to break privacy in federated learning? Advances in Neural Information Processing Systems **33**, 16937–16947 (2020)
2. Jere, M.S., Farnan, T., Koushanfar, F.: A taxonomy of attacks on federated learning. IEEE Security & Privacy **19**(2), 20–28 (2020)
3. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
4. Liu, D.C., Nocedal, J.: On the limited memory bfgs method for large scale optimization. Mathematical programming **45**(1-3), 503–528 (1989)
5. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. pp. 1273–1282. PMLR (2017)
6. Palihawadana, C., Wiratunga, N., Wijekoon, A., Kalutarage, H.: Fedsim: Similarity guided model aggregation for federated learning. Neurocomputing (2021)
7. Strang, G.: The discrete cosine transform. SIAM review **41**(1), 135–147 (1999)
8. Wei, W., Liu, L., Loper, M., Chow, K.H., Gursoy, M.E., Truex, S., Wu, Y.: A framework for evaluating gradient leakage attacks in federated learning. arXiv preprint arXiv:2004.10397 (2020)
9. Zhang, R., Guo, S., Wang, J., Xie, X., Tao, D.: A survey on gradient inversion: Attacks, defenses and future directions. arXiv preprint arXiv:2206.07284 (2022)
10. Zhang, T., Gao, L., He, C., Zhang, M., Krishnamachari, B., Avestimehr, A.S.: Federated learning for the internet of things: Applications, challenges, and opportunities. IEEE Internet of Things Magazine **5**(1), 24–29 (2022)
11. Zhao, B., Mopuri, K.R., Bilen, H.: idlg: Improved deep leakage from gradients. arXiv preprint arXiv:2001.02610 (2020)
12. Zhu, L., Liu, Z., Han, S.: Deep leakage from gradients. Advances in neural information processing systems **32** (2019)