2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS) | 978-1-6654-1913-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/AICAS51828.2021.9458510

Federated Regularization Learning: an Accurate and Safe Method for Federated Learning

Tianqi Su, Meiqi Wang, and Zhongfeng Wang
School of Electronic Science and Engineering, Nanjing University, P. R. China
Email: {sutainqi,mqwang}@smail.nju.edu.cn, {zfwang}@nju.edu.cn

***Abstract*—Distributed machine learning (ML) and other related techniques such as federated learning are facing a high risk of information leakage. Differential privacy (DP) is commonly used to protect privacy. However, it suffers from low accuracy due to the unbalanced data distribution in federated learning and additional noise brought by DP itself. In this paper, we propose a novel federated learning model that can protect data privacy from the gradient leakage attack and black-box membership inference attack (MIA). The proposed protection scheme makes the data hard to be reproduced and be distinguished from predictions. A small simulated attacker network is embedded as a regularization punishment to defend the malicious attacks. We further introduce a gradient modification method to secure the weight information and remedy the additional accuracy loss. The proposed privacy protection scheme is evaluated on MNIST and CIFAR-10, and compared with state-of-the-art DP-based federated learning models. Experimental results demonstrate that our model can successfully defend diverse external attacks to user-level privacy with negligible accuracy loss.**

***Index Terms*—Federated learning, information leakage**

I. Introduction

Machine learning (ML) has achieved great success in a wide range of applications [1] [2] [3] [4]. The success of ML mainly results from the flourishing development of big data. However, the collected data can be highly sensitive, which may lead to serious security problems. Using a secure method to training an ML model is still a great challenge in the AI field.

Federated learning [5] is one of the most promising solutions to the information security problem. Only the processed parameters are transmitted in federated learning. However, the processed updates sent to the central server may be divulged to attackers. Malicious attackers can recover some sensitive information by utilizing the transmitted parameters (e.g. gradient leakage attack [6]). What's more, the accurate model itself, especially the overfitting one [7], responds differently according to different inputs, which can tell the attacker whether the input is in the training set (e.g. membership inference attack (MIA) [8]). To defend these attacks, DP [9] is widely applied [10]. However, the model performance is damaged due to unbalanced data distribution in local clients [11] and additional noise brought by DP.

In this paper, we focus on protecting data privacy in federated learning from gradient leakage attack and MIA. We investigate a safe and accurate model called federated regularization learning model. The proposed model contains a simulated attacker (SA network) and a small auxiliary dataset to assist the main training process. With multiple attacks performed by the SA network, the classification network can develop the ability to defend external attacks. We train the model in an adversarial process, which can control the trade-off between the loss caused by the SA network and the classification error. The updated gradient is modified to hide the specific gradient value, which also reduces the influence of overfitting.

The main contributions are listed as follows:

We propose a new privacy mechanism called federated regularization learning model to defend the malicious attacks in federated learning. Our method can successfully defend diverse external attacks to user-level privacy and maintain good classification accuracy.

To the best of our knowledge, we evaluate the damage caused by both black-box MIA and gradient leakage attack in federated learning for the first time and step further to defend both attacks at the same time.

Extensive experimental evaluations are conducted to demonstrate the effectiveness of the proposed model. Our method can easily confuse the attacker and achieve better accuracy performance than the state-of-the-art DP-based federated learning method.

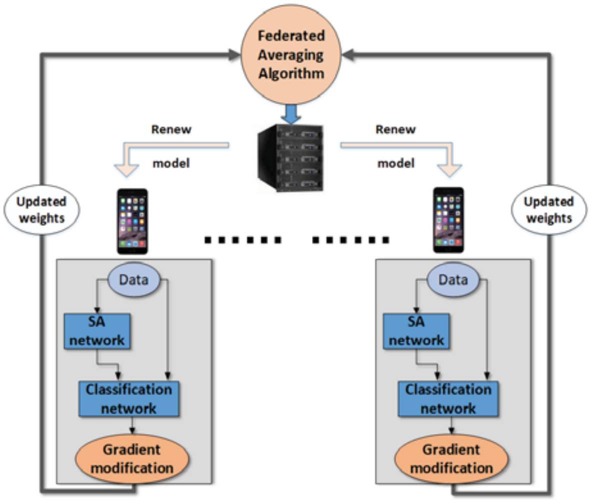


Fig. 1. An overview of the proposed federated regularization learning model.

II. BACKGROUND

A. Federated Learning

Different from distributed learning, federated learning has several special attributes: 1) Non-IID; 2) Unbalanced data amount; 3) Limited communication. Federated learning allows the model to be trained in an anonymous and collaborated way. However, the communication between the server and clients can be vulnerable to information interception. Avoiding the information leakage in the transmission process is of great importance. In this paper, the federated learning procedure follows the implementation of [12]. For simplicity, we only consider the non-iid and limited communication properties in the federated learning procedure.

B. Model Inversion Attack

Model inversion attack [13] can use the leaked parameters to expose sensitive information. In a black-box setting, the adversary obtains an output value from an ML system and infers the sensitive attributes of the corresponding input. This process indirectly gets the private data information of the training dataset. The transmitted gradients in the federated learning model can be used in the model inversion attack (e.g. gradient leakage attack [6]). The proposed model needs to defend the threat.

C. Membership Inference Attack

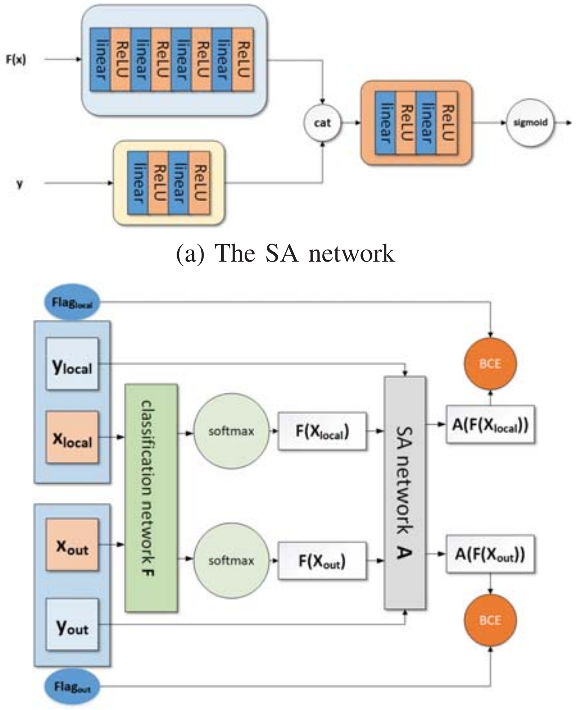
Membership inference attack (MIA) [8] needs to distinguish whether the data is in the training set or not. In this attack mode, the attacker only needs to obtain the confidence of the predictive classification, and does not need to know other information such as the model structure and training method. This attack is particularly effective for overfitting models [7]. In federated learning, the adversary can utilize the client's outputs to perform MIA. We also use this attack mode to evaluate the robustness of the proposed model.

III. THE PROPOSED METHOD

The proposed federated regularization learning model is shown in Fig. 1. Federated regularization model along with the auxiliary training dataset is sent to clients at the beginning of the training. The allocated model carries a classification network with a small simulated attacker network (SA network). Two networks are trained alternatively with the gradient modification algorithm. The server processes the collected updates using federated averaging algorithm [5], and reallocates the updated weights for the next iteration.

A. Adversarial Regularization Learning Model

The aim of the SA network is to maximize the inference gain while our classification model, set as the defender, needs to minimize the maximum information leakage and balance the trade-off between the accuracy and the information leakage loss.



(a) The SA network

(b) Training procedure of the SA network

Fig. 2. The SA network (a) contains three modules to generate the final binary classification result. (b) illustrates the training procedure of the SA network.

1) SA network: The SA network (Fig.2(a)) contains two modules to process the predicted label produced by the classification network and its corresponding ground truth label separately. Both modules are composed of fully connected layers and activation layers. Two modules' outputs are fed into a judgment model to produce the binary classification result.

In each iteration, the SA network is trained before the classification. For any input data, the classification network F first outputs the prediction vector of probabilities $F(x)$ that records the class the data belongs to. Then $F(x)$ along with the ground truth label y is sent to the SA network A. We use supervised learning to train the SA network. The SA network needs to distinguish whether the input data x belongs to the original training dataset $Flag_{local}$ or the auxiliary dataset $Flag_{out}$ (Fig. 2(b)). BCE loss is applied to evaluate the correctness of the binary classification. BCE loss can be expressed as:

$$\begin{aligned} BCE_{loss}(x_n, y_n) = \frac{1}{n}\sum(y_n \times ln(x_n) + \\ (1-y_n) \times ln(1-x_n)). \end{aligned} \tag{1}$$

The final loss function can be written as:

$$\begin{aligned} L_{SA} = \lambda(BCE_{loss}(A(F(x_{local})), Flag_{local}) + \\ BCE_{loss}(A(F(x_{out})), Flag_{out})). \end{aligned} \tag{2}$$

The loss function contains two BCE losses coming from the local dataset x_{local} and auxiliary dataset x_{out}. λ refers to the influence factor of the SA network. The SA network gets more powerful when the value of λ is set higher, but larger λ also leads to more accuracy loss.

2) Classification Network: Our classification network F is trained only on the local dataset (Fig. 3). Each input data x_{local} is fed into the trained SA network A to judge its original source. BCE loss is applied to measure the correctness:

$$L_{SA'} = BCE_{loss}(A(F(x_{local})), Flag_{local}). \tag{3}$$

We use $L_{SA'}$ to maximize the inference attack success ratio to challenge the classification network. The cross-entropy loss is applied to evaluate the correctness between the ground truth y and the predicted label $F(x)$:

$$L_c = -y + log \sum_{j=1}^{N} e^{F(x)_j}, \tag{4}$$

where N denotes the number of the input's classes. We use L_c to assure model's accuracy. The complete loss function can be written as

$$L = L_c - \lambda L_{SA'}. \tag{5}$$

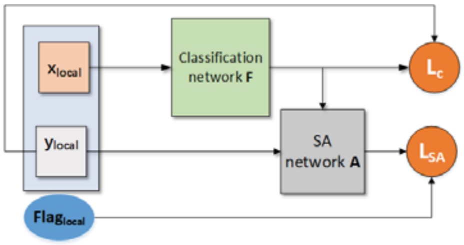


Fig. 3. The training procedure of the classification network.

B. Gradient Modification Algorithm

To prevent the model from gradient leakage attack, we need to modify the gradients to hide specific weight values. For this purpose, we introduce the gradient modification algorithm in the training procedure. The gradients larger than a certain threshold are transmitted first and the rest gradients are accumulated locally until they get large enough to be selected in the following iterations. Thus, we send the large gradients immediately but eventually send all of the gradients over time.

Considering that the modification can cause extra accuracy loss, we introduce momentum correlation to help alleviate the accuracy loss. In t^{th} iteration, the accumulation function can be written as:

$$u_t^{small} = S_{select}(\bigtriangledown f(x, loss) + mu_{t-1}^{small}), \tag{6}$$

where u denotes the accumulated weight. m (m>1) is the momentum factor. s_{select} represents the function that selects the large gradient values and leaves the small gradients u_t^{small} for the next iteration. f is the loss function computed from training sample x.

IV. Experiment Results

A. Implementation Details for Federated Regularization Learning Model

We evaluate the effectiveness of federated regularization learning model on two standard datasets, MNIST and CIFAR-10. The dataset is divided into a training dataset and an auxiliary dataset for further assisting training. The auxiliary set is randomly selected from the origin dataset. The training set is sorted according to their labels and divided into small sets for each client. Each client has two small sets to simulated federated learning scenario. In this way, most clients have unbalanced samples from two classes only. According to the property of federated learning in section II, we assume only 20% of local clients are able to train the model. We set the value of momentum factor as 1.5 in the experiments.

We randomly select one client as the victim and perform MIA and gradient leakage attack on this client to evaluate the model's privacy protection capability.

B. Membership Inference Attack

We perform MIA on CIFAR-10 using VGG16 [14] as the classification network. The data from the victim is utilized to train the attacker, which makes the membership inference harder to defend. To fairly demonstrate and compare our model's privacy protection capability, we implement federated regularization method on federated learning model with different λ. DP-based model is also set as a comparison. The experimental results are shown in TABLE I.

We investigate the vulnerability of the common federated learning model (referred as no protected model in TABLE I). As is shown in TABLE I, MIA gets high attack accuracy (71.33%) when the model is under no protection. After applying federated regularization method, it is noticeable that the adversary cannot easily infer the membership from the protection. Furthermore, MIA's accuracy decreases a lot as the protection level (λ) gets higher. In contrast with DP-based model [15], our model has much better test accuracy when MIA achieves similar accuracy. It can be noticed that our model has a little test accuracy loss compared with baseline, but it is tolerable to protect client's privacy [10]. The results prove that our model can successfully defend MIA and maintain better accuracy than the state-of-the-art DP-based method.

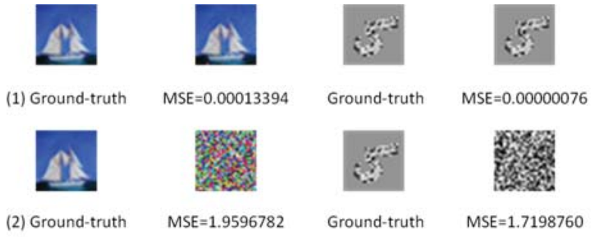


Fig. 4. Visualization of gradient leakage attack on CIFAR-10 and MNIST datasets. (1) illustrates the attack results on the common federated learning model while (2) represents the results of the proposed model. We use mean-squared error (MSE) to evaluate the attack performance.

TABLE I
MODEL EVALUATION AGAINST MIA ON CIFAR-10

Model	λ / ϵ	Global Model Test Accuracy	MIA Accuracy
no protected model (baseline)	-	69.24	71.33
federated regularization learning model (ours)	λ = 3	64.74	57.88
DP-based model	ϵ = 4	60.70	56.30
federated regularization learning model(ours)	λ = 5	63.44	51.25
DP-based model	ϵ = 2	45.00	52.40
federated regularization learning model(ours)	λ = 7	62.77	50.25
DP-based model	ϵ= 1	25.30	50.80

TABLE II
MODEL ACCURACY EVALUATION WITH DIFFERENT CLIENTS ON MNIST

Model	Number of Clients	Test Accuracy
no protected model (baseline)	100	96.04%
federated regularization learning model($\lambda = 5$)(ours)	100	95.09%
(8,e-3)-DP based model	100	78.07%
no protected model (baseline)	1000	94.37%
federated regularization learning model($\lambda = 5$)(ours)	1000	88.93%
(8,e-6)-DP based model	1000	88.72%

C. Deep Gradient Leakage Attack

We perform the gradient leakage attack on MINST and CIFAR-10 dataset. The reconstruction results are visualized in Fig. 4 and the accuracy of our model is listed on TABLE II. Malicious attacker can easily reconstruct the data with vivid details in no protect model but can only get noisy result in our model. The results show that our model can successfully hide the sensitive feature in transmission. We also evaluate model's performance (TABLE II). We set λ = 5 which has the best balance between model's accuracy and MIA's successful rate in TABLE I as the benchmark. The performance of our model beats the DP-based federated learning model [10] with different client numbers from 100 to 1000. Both results show that the proposed model can successfully defend gradient leakage attack and works much better in federated learning procedure.

V. CONCLUSION

In this paper, we propose a new privacy mechanism, named as federated regularization learning model, to avoid information leakage in federated learning. The SA network is embedded in the model and the gradient modification method is applied to hide specific weight details. Experimental results show that our network can defend two common malicious attacks simultaneously and achieve better accuracy than the state-of-the-art DP-based method.

REFERENCES

[1] Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton, "Grammar as a foreign language," *Eprint Arxiv*, pp. 2773–2781, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2016.

[3] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[4] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[5] Jakub Konen, H. Brendan Mcmahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon, "Federated learning: Strategies for improving communication efficiency," 2016.

[6] Ligeng Zhu, Zhijian Liu, and Song Han, "Deep leakage from gradients," 2019.

[7] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 2018, pp. 268–282.

[8] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov, "Membership inference attacks against machine learning models," 2016.

[9] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.

[10] Robin C Geyer, Tassilo Klein, and Moin Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.

[11] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov, "Differential privacy has disparate impact on model accuracy," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., pp. 15479–15488. Curran Associates, Inc., 2019.

[12] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[13] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1322–1333.

[14] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang, "Membership inference attack against differentially private deep learning model.," *Trans. Data Priv.*, vol. 11, no. 1, pp. 61–79, 2018.