A novel approach to simultaneously improve privacy, efficiency and reliability of federated DNN learning

Hanlin Gu [2] , **Lixin Fan** , **Bowen Li**[4] , **Yan Kang**[1] , **Yuan Yao**[2] and **Qiang Yang**[1,3]
[1]AI Group, WeBank Co., Ltd, Shenzhen, China
[2]Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, China
[3]Department of CSE, Hong Kong University of Science and Technology, Hong Kong, China
[4]Department of CSE, Shanghai Jiao Tong University, China
{lixinfan, yangkang}@webank.com, hguaf@connect.ust.hk

Abstract

Federated learning (FL) aims to protect data privacy by cooperatively learning a model without sharing private data among users. However, there can still be privacy risks because the server or distributed systems may not be trusted platforms. The attacker may infer the private data by attacking the released model or updated gradients. Splitting the network into private layers and public layers and only implementing FL in public layers could be against attack partially. In this paper, we propose a novel privacy-preserving method based on Federated Deep Learning with Private Passport(FDL-PP). By embedding the passport into the private layers of the network, private data security for each participant is guaranteed. Our empirical experiments, on MNIST and CIFAR10 dataset with multiple clients, demonstrate that embedding passports resist deep leakage attack and model inversion attack without affecting model classification performances. What is more, the proposed method is as efficient as distributed learning.

1 Introduction

Since the introduction of Federated learning (FL) [Konečný *et al.*, 2015; McMahan *et al.*, 2017], a variety of technologies have been adopted to improve the *privacy-preserving capability*, the *efficiency* and the *reliability* of FL process. For instance, homomorphic encryption (HE) [Gentry and others, 2009] protects exchanged information and training data from being espied by semi-honest parties. Nevertheless, extremely heavy computation and communication overhead incurred by HE make it unsuitable for Federated DNN model learning. On the other hand, differential privacy (DP) [Dwork *et al.*, 2014; Abadi *et al.*, 2016] has been adopted for federated deep learning because deep learning algorithm with DP can run almost as fast as without DP. However, it was shown that attackers may infer training images at pixel level accuracy even random noise are added to exchanged gradients [Geiping *et al.*, 2020; Zhao *et al.*, 2020; Zhu and Han, 2020; Yin *et al.*, 2021]. Moreover, exceedingly large noise jeopardise learning reliability and lead to significant degradation of model accuracy (as shown in Figure 1(c). It remains an open question as how to train federated DNN models in an *efficient*, *reliable* and *privacy preserving* manner.

Inspired by previous works in which only a small fraction of (DNN) models are shared with the aggregator while the rest of model parameters are kept secret (splitfed [Thapa *et al.*, 2020]) or simply skipped [Shokri and Shmatikov, 2015], we propose to split DNN models into public and private layers and adopt a passport layer between the private and public layers to provide secure privacy preserving capability. This novel solution, named Federated Deep Learning with Private Passport (FDL-PP), addresses the aforementioned challenge in three aspects. First, FDL-PP is *efficient* since public layers are directly shared with the aggregator without using computational demanding encryption. Second, the private passport layer blocks information leakage to semi-honest adversaries and *guarantees privacy* of training data. Third, the learning of federated DNN models with FDL-PP is *reliable* and no noticeable performance degradation are incurred.

The proposed FDL-PP not only demonstrates high model accuracy but also exhibits superior robustness against various privacy attacks, which are detrimental to existing methods under certain conditions (as shown in Figure 1 (a) and (b).

Overall, the propose method provides a balanced solution that improves model learning efficiency without compromising model accuracy and data privacy.

1.1 Related work

In contrast to FL, SL [Gupta and Raskar, 2018] enables ML training with clients having low computing resources as the client trains only the partial model of the split ML global model. Splitfed integrates the advantages of FL and SL by splitting the network architecture between the clients and server as in SL, which provides more privacy than FL and better efficiency than SL through collaborative training in FL.

With many clients for collaborative training and exposure of model parameters, FL makes itself vulnerable to various attacks and open to risks. On the one hand, the attacker could influence the model through the **Poisoning**. Since the central server cannot access the private data of clients, the malicious clients can cheat the server by sending modified and harmful model updates, initiating adversarial attacks on the global model [Nasr *et al.*, 2019; Kairouz *et al.*, 2019].

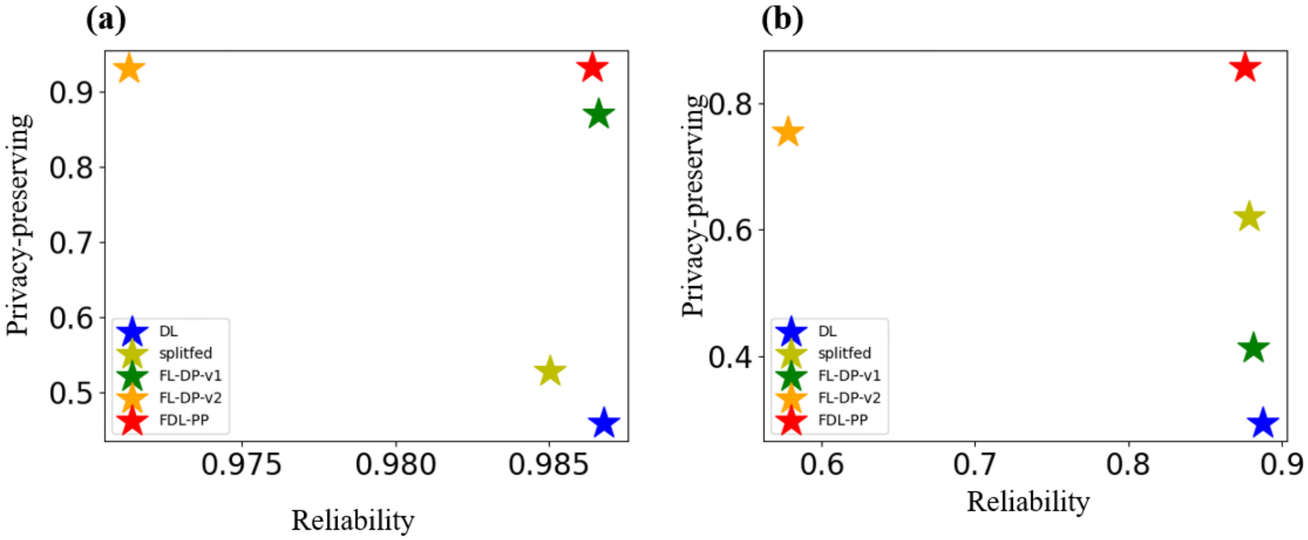

Figure 1: The results of reliability and privacy-preserving of different methods including distributed learning (DL), splitfed, FL-DP (applying DP in FL by adding different noise σ, FL-DP-v1 is adding noise as Gaussian σ 0.002, FL-DP-v2 is adding noise as Gaussian σ 0.128) and FDL-PP. Respectively, figure (a) shows the results of LeNet on MNIST classification task, figure (b) shows the results of AlexNet on CIFAR10 classification task: each horizontal axis describes the classification accuracy (Reliability) of different privacy preserving methods, each vertical axis describes the privacy-preserving of varying methods.

On the other hand, attackers may infer data distribution even reconstruct other clients' data through the model gradients and weights. [Shokri *et al.*, 2017] revealed the original data distribution by membership attack, and [Melis *et al.*, 2019] recovered the position distribution of original data in federated learning. What's more, deep leakage attack (**DLG**) was proposed by [Geiping *et al.*, 2020; Zhao *et al.*, 2020; Zhu and Han, 2020; Yin *et al.*, 2021], which enables pixel-level detailed image reconstruction based on gradients of model weights. In addition, Emerging research on model inversion techniques [Fredrikson *et al.*, 2015; He *et al.*, 2016] offered insights into this task. Model inversion (**MI**) attack could retrieve the images via back-propagating gradients on appropriate loss functions to the learnable input.

Passport is used to verify ownership and claim legitimate intellectual property rights (IPR) [Fan *et al.*, 2019], in case that models are illegally copied, re-distributed, or misused. In order to alleviate the influence of model performance the passport brings, [Zhang *et al.*, 2020] proposes a general passport-aware normalization formulation to increase the classification accuracy.

2 Federated Deep Learning with Private Passport (FDL-PP)

The proposed FDL-PP method takes advantages of two strategies. First, FDL models are separated as public and private layers, with parameters in private layers being kept secret to persevere data privacy and parameters in public layers are shared without encryption. However, splitting models alone does not defense well privacy attacks since training data can still be inferred from the public parameters (as shown our experiment). To this end, we employ a private passport layer to guarantee that training data and private parameters cannot be inferred from the publicly shared parameters. The structure of Federated DNN models with separated parameters and passport layers is illustrated in Figure 2.

2.1 Modified splitfed

The notion of keeping part of model parameters secret while learning the rest of model parameters in a federated setting has been explored before ([Thapa *et al.*, 2020]). However, this method alone is insufficient to provide secure protection of training data as shown by experiment results in our study.

The proposed FDL-PP method consists of three steps as shown in Figure 2: (1) All clients carry out the forward and backward propagation on a global model in parallel and upload their last few layers (named public layers) of the model to the server. (2) The center server process the aggregation of the public layers and update. (3) Clients download the updated public model and train their first several layers (named private layers) without modifying the published layers.

2.2 Threat Model

In this work, for our architecture, we consider all participants, including K clients and the server, are *honest-but-curious* adversaries. They possibly do some calculations to obtain the private data of clients by observing the model weights and updates, but they do not maliciously modify their own inputs or parameters for the attack purpose. There are considered in our study: DLG and MI attack.

DLG Deep leakage attack is firstly proposed by [Zhu and Han, 2020], where the server tries to recover the original data based on updates and model weights. Consider k_{th} client who submit its gradient ∇W_{pub} of public model. To reconstruct the raw data of client k, firstly initializing a dummy input $\mathbf{x}'$

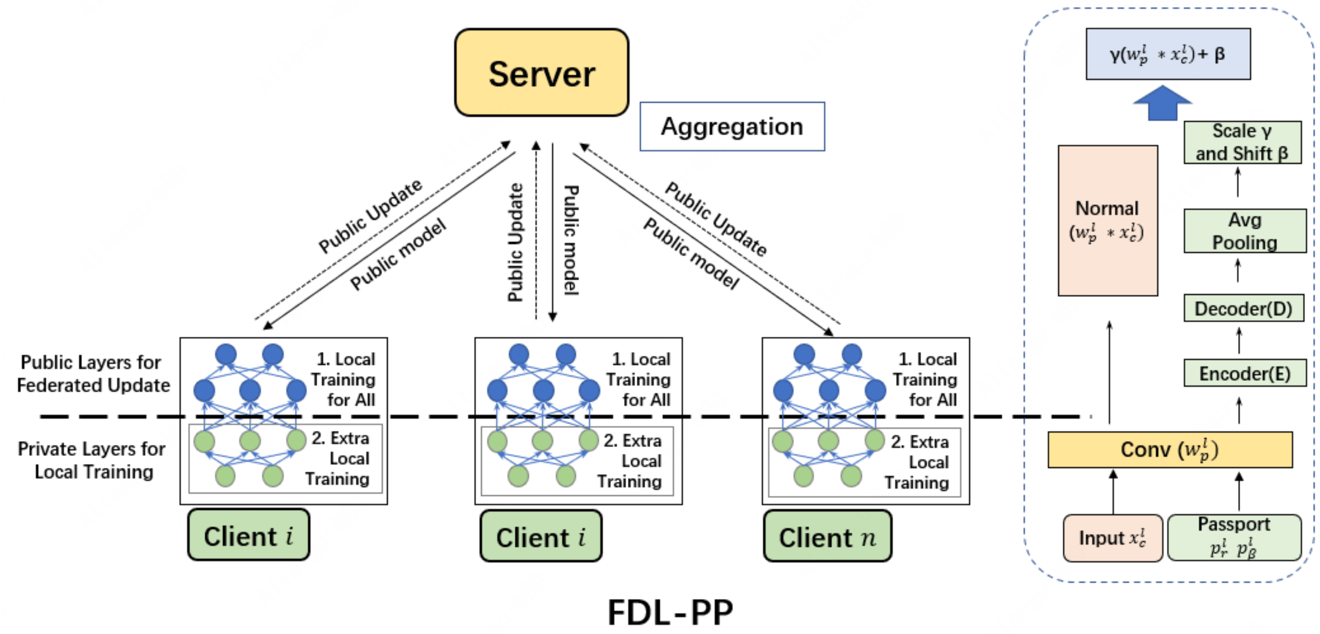

Figure 2: The structure of FDL-PP. The local model is split into private layers and public layers. Only public layers are aggregated by `FederatedAvg` algorithm, private layers are kept private in local(the left panel). In the private layers, the client embeds private passport into the parameters of each local model to protect the privacy of local data(the right panel).

with label $\mathbf{y}'$. Then, feeding the dummy data into the model and get the dummy gradient:

$$\nabla W'_{pub} = \frac{\partial \ell \left(F\left(\mathbf{x}', W\right), \mathbf{y}'\right)}{\partial W}. \tag{1}$$

$$\mathbf{x}'^{*}, \mathbf{y}'^{*} = \underset{\mathbf{x}', \mathbf{y}'}{\arg\min} \left\|\nabla W'_{pub} - \nabla W_{pub}\right\|^2 \tag{2}$$

where ℓ is the classification loss, F is the forward function.

By minimizing the differences between the dummy gradients and the original gradients (as shown in equation 2), the attackers could learn a good input $\mathbf{x}'^{*}$ and $\mathbf{y}'^{*}$. In particular, DLG could easily recover the private model's output and even recover the original image if the private model is leaked, which is similar to the white-box attack.

MI However, sometimes, the attacker may not know the weights of the private model. In this case, model inversion attack (MI [He *et al.*, 2019]) provides a strategy to reconstruct the original image. MI attack attempts to recover the original input given the private model output as the following two cases: (1) In the white-box setting, the adversary knows the private model on the participant. (2) In the black-box setting, the model parameters on the participant are unavailable for the adversary. In our setting, we mainly consider the second case. The server may conspire with the curious client to obtain other clients' data. The server provides the updated weights and gradients of the public model to reconstruct the output of the private model, and the curious client offers his own private model as the target model.

Specifically, given the private model output O_{prt} and target model F, MI attack learns a good dummy input $\mathbf{x}'^{*}$ by solving the equation 3.

$$\mathbf{x}'^{*} = \underset{\mathbf{x}'}{\arg\min} \left\|F(\mathbf{x}') - O_{prt}\right\|^2 \tag{3}$$

2.3 FDL-PP

Inspired by the [Fan *et al.*, 2019], we embed the passports in the federated learning. Concretely, we add a passport layer after the convolution layer. Similar to the batch normalization layer, the output of the passport layer is scale factor γ and bias shift term β, which are dependent on both the convolution kernels $\mathbf{W}_p$ and the designated passport P. In order to increase the complexity of the passport layer and keep the model performance, we add a fully connected autoencoder (Encoder $\mathbf{E}$ and Decoder $\mathbf{D}$) into the passport layer. Moreover, the details of passport layer based on Equation 4 and 5.

$$\mathbf{O}^l(\mathbf{X}_p) = \gamma^l \mathbf{X}_p + \beta^l = \gamma^l(\mathbf{W}_p^l * \mathbf{X}_c^l) + \beta^l \tag{4}$$

$$\gamma^l = \mathbf{D}(\mathbf{E}(\mathbf{W}_p^l * \mathbf{P}_\gamma^l)), \beta^l = \mathbf{D}(\mathbf{E}(\mathbf{W}_p^l * \mathbf{P}_\beta^l)) \tag{5}$$

where $*$ denotes the convolution operations, l is the layer number, $\mathbf{X}_p$ is the input to the passport layer, and $\mathbf{X}_c$ is the input to the convolution layer. $\mathbf{O}()$ is the corresponding linear transformation of outputs, while P_γ^l and P_β^l are the passports used to derive scale factor and bias term respectively. The right part of Figure 2 illustrates the procedure of passport embedding.

We embed the passport into the private layers. Each client reserves their own passport and does not leak to the server. In fact, it is hard for attackers to reconstruct the original input. Firstly, adversaries don't know what the passport is, so the

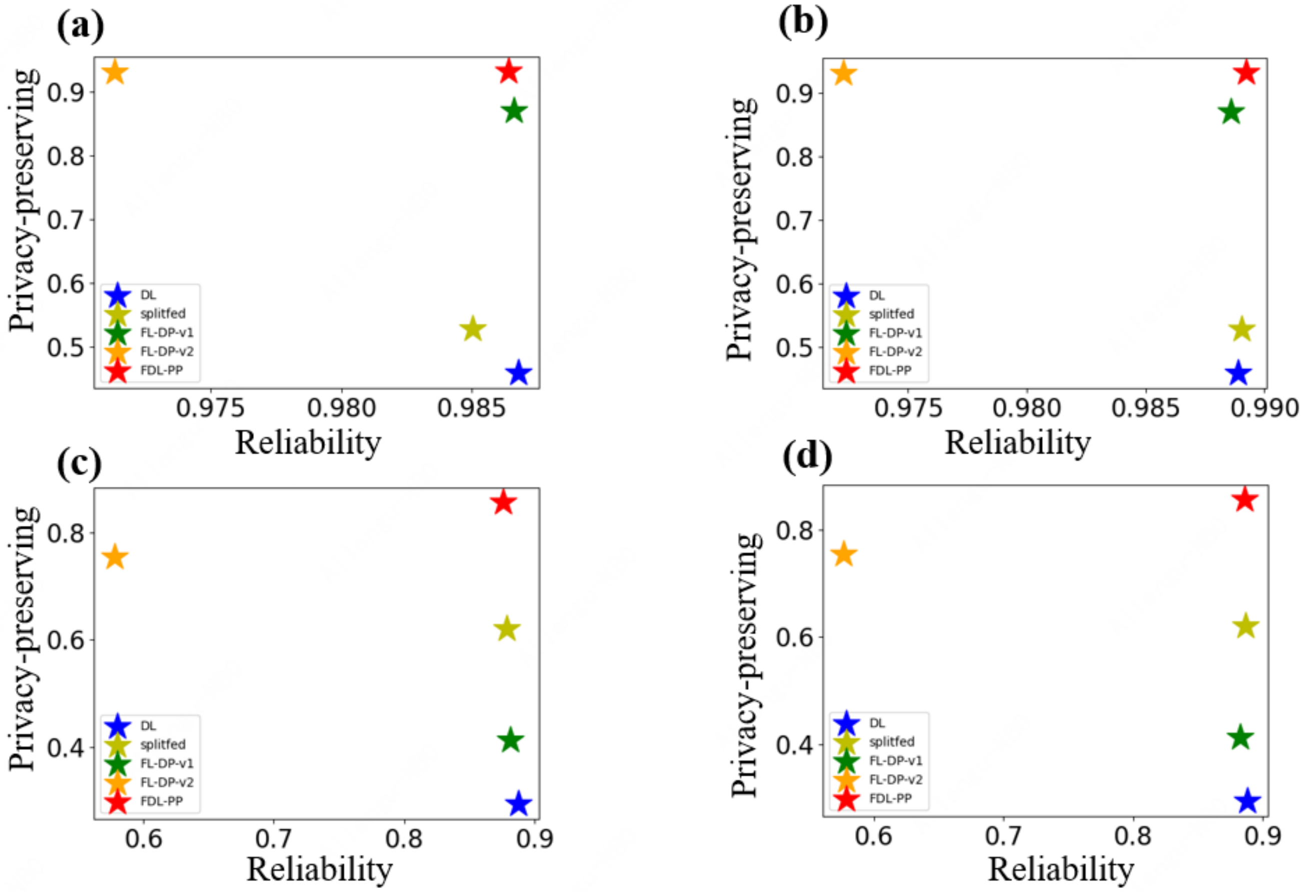

Figure 3: The illustration of reliability and privacy-preserving of different methods based on FL including DL, splitfed, FL-DP (varying noise σ) and FDL-PP with different total client number and training setting. Respectively, figure (a)(b) show the results of LeNet on MNIST classification task with 5 clients and 10 clients, figure (c)(d) show the results of AlexNet on CIFAR10 classification task with 5 clients and 10 clients: each horizontal axis describes the classification accuracy (Reliability) of different privacy preserving methods, each vertical axis describes the privacy performance (Security) of varying methods.

reconstruction they do is a mixture of passport and original input. Secondly, the way to inserting a passport is complex. For example, adversaries don't know how to insert the passport because the passport not only relies on learned weights $\mathbf{E}$ and $\mathbf{D}$ but also relates to the model convolution weights $\mathbf{W}_p^l$. Finally, the position of inserting a passport is varied. Each client could choose any position of private model to insert. Therefore it is not easy to learn the passport and recover the original images.

3 Experimental Results

This section illustrates the empirical study of our privacy-preserving method (FDL-PP). We investigate the typical classification problems in MNIST and CIFRA10 by applying two well-known networks (LeNet and AlexNet). In particular, we test the result by splitting the network into the first convolution and activation layers as the private layer, the rest layers as the public layer. Also, we insert the passport layer factors: γ and β into the first layer of the private layer in AlexNet (FDL-PP) and three layers of the private layer separately in LeNet.

We simulate $K = 10, 5$ clients' horizontal federated learning system in a stand-alone machine for the federated learning setting. In each communication round, we were uniformly sampling the public layers of clients to average in the server. In addition, the optimization algorithm we choose SGD and learning rate is 0.01.

3.1 Evaluation matrix

We will evaluate the effectiveness in three aspects: *privacy-preserving*, the *efficiency* and the *reliability*.

Reliability For reliability, we use classification accuracy $\mathbf{1}(y, f(x))$ to evaluate, where y is targeted label, x is input image, f is deep learning model. The classification accuracy is between 0 and 1, the closer to 1 the classification accuracy is, the more the reliability is;

Privacy-preserving We apply f(reconstruction MSE) to evaluate the privacy-preserving in different approaches, where reconstruction MSE is $\ell 2$ distance between reconstruction images and original images, f is fixed increasing function $\frac{2x^\alpha}{1+x^\alpha}$

aimed to transfer the reconstruction MSE to the range of $(0, 1)$ as shown in Equation 6. In our paper, we pick $\alpha = 0.1$ What's more, the higher the reconstruction MSE is, the more the privacy is preserved.

$$
\begin{aligned}
f(Rec\_MSE) &= f(||I_{ori} - I_{rec}||^2) \\
&= \frac{2(\sum_{i=1,j=1}^{M,N}(I_{ori}(i,j) - I_{rec}(i,j))^2)^\alpha}{1 + (\sum_{i=1,j=1}^{M,N}(I_{ori}(i,j) - I_{rec}(i,j))^2)^\alpha}
\end{aligned} \tag{6}
$$

Efficiency For efficiency, we evaluate different approaches by calculating the consuming time of training.

3.2 Comparison of Privacy Preserving Mechanisms

We compare the reliability of different approaches via classification accuracy under two cases: five and ten clients participate the federated learning. Figure 3 demonstrates the DL and splitfed doesn't protect privacy even they keep high reliability. Introducing DP more in FL influence the reliability although it could protect the privacy. And our method FDL-PP is the only one could obtain good performance in both reliability and privacy-preserving (red star in the top right position).

ReliabilityFor reliability, table 1 shows the classification accuracy only degrades no more than 1% for splitfed and our proposed FDL-PP compared to the distributed learning. The classification accuracy of DP decreases more than 3% when adding noise σ is 0.128.

Privacy-preserving Privacy-preserving is evaluated by the reconstruction MSE. High MSE indicates that the attacker is hard to recover the original images, which represents the more security of the approach. We consider the DLG attack in the white-box setting, where the attackers know the weights of the public model and the private model and only know the updated gradients of the public model. On the other hand, we consider the MI attack in a black-box setting, where the attacks don't know any private model information.

For the DLG attack, table 1 demonstrates the reconstruction MSE of different methods in LeNet-MNIST and AlexNet-CIFAR10, figure 4 (a) shows the reconstructed images among various method. Both of them show that FL and splitfed could not protect privacy because the MSE is low and the reconstructed image is clear. And as adding more noise in gradients in DP of FL, the privacy leaked less. Moreover, our proposed method FDL-PP could protect privacy most via the highest MSE. The attacker is hard to recover the original image when applying FDL-PP.

For the MI attack, we suppose the server combines one client to attack other clients. Concretely, we use the private model of one client as the initial model to attack the original images of another client. The result of Table 1 and 4 (b) illustrates FDL-PP have resist MI attack while splitfed could not.

Efficiency We compare the computation speed among various methods, including FL, splitfed, FL-DP, FDL-PP, and HE (CryptoNets [Gilad-Bachrach *et al.*, 2016]) on MNIST. Table 1 shows the computation time (for one inference) and the accuracy. We observe that the consuming time of our proposed method FDL-PP is comparable to other methods except HE (CryptoNets). Actually, HE needs a large amount of time in encrypting to be less efficient than other approaches.

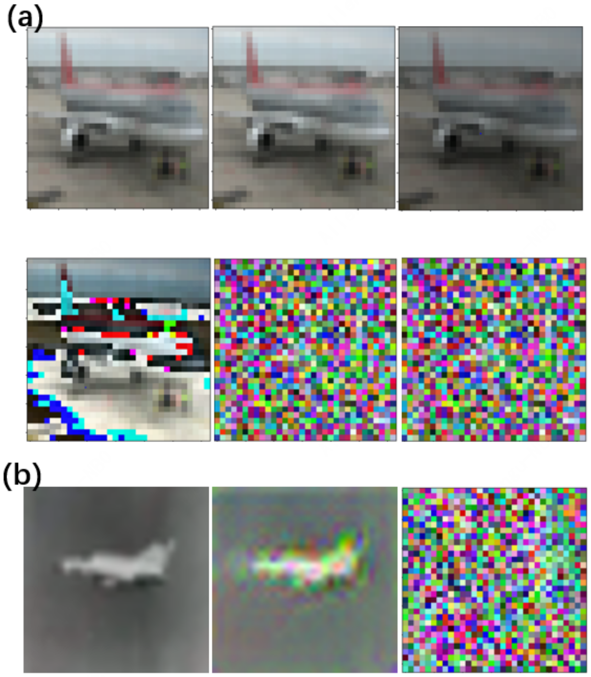

Figure 4: The reconstructed image of CIFAR10 after DLG and MI attack. Figure (a) is the reconstructed image of the DLG attack. From left to right, top to bottom, methods are original image, splitfed, FL-DP-0.002, FL-DP-0.032, FL-DP-0.128, FDL-PP separately. Figure (b) is the reconstructed image of the MI attack. From left to right, methods are original image, splitfed, FDL-PP separately.

4 Discussion and Conclusion

This paper illustrates a novel approach: FDL-PP to simultaneously improve privacy, efficiency, and reliability of federated DNN learning. Compared to the method based on HE, FDL-PP is efficient since public layers are directly shared with the aggregator without using demanding encryption. What's more, the application of FDL-PP in the federated DNN model doesn't sacrifice reliability. Finally, embedding passports into the private layers block information leakage to adversaries.

[Fan *et al.*, 2019] analyzed robustness against DLG attack in FL by building the systems of linear equations. Therefore, further exploration focus on the theoretical proof about the privacy protection of FDL-PP.

Table 1: The result of reliability, Privacy-preserving and efficiency of DL, splitfed, FL-DP, cryptoNets and FDL-PP

		DL	Splitfed	FL-DP(0.002)	FL-DP(0.128)	cryptoNets	FDL-PP
Reliability	MNIST	9.86e-1(1.17e-3)	9.85e-2(1.80e-3)	9.87e-1(2.13e-3)	9.71e-1(4.23e-3)	9.67e-1(3.00e-3)	9.86e-1(1.01e-3)
(Classification acc)	CIFAR10	8.88e-1(4.61e-3)	8.79e-1(5.43e-3)	8.82e-1(4.96e-3)	5.78e-1(1.54e-2)	\	8.76e-1(5.21e-3)
Privacy-preserving	MNIST	4.58e-1(3.77e-3)	5.21e-1(2.31e-2)	8.70e-1(5.10e-3)	9.31e-1(2.34e-3)	\	**9.32e-1(1.17e-3)**
(DLG attack)	CIFAR10	2.93e-1(7.50e-3)	5.61e-1(5.68e-2)	4.12e-1(7.00e-3)	7.54e-1(2.74e-3)	\	**8.55e-1(3.51e-3)**
Privacy-preserving(MI attack)	CIFAR10	\	5.00e-1(5.09e-2)	\	\	\	**8.46e-1(3.99e-3)**
Efficiency	MNIST	25.6(0.30)	53.2 (0.30)	26.7(0.30)		3009.6(1.70)	65.4(0.35)

References

[Abadi *et al.*, 2016] 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*, pages 308–318, 2016.

[Dwork *et al.*, 2014] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[Fan *et al.*, 2019] Lixin Fan, Kam Woh Ng, and Chee Seng Chan. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. 2019.

[Fredrikson *et al.*, 2015] 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*, pages 1322–1333, 2015.

[Geiping *et al.*, 2020] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients–how easy is it to break privacy in federated learning? *arXiv preprint arXiv:2003.14053*, 2020.

[Gentry and others, 2009] Craig Gentry et al. *A fully homomorphic encryption scheme*, volume 20. Stanford university Stanford, 2009.

[Gilad-Bachrach *et al.*, 2016] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International Conference on Machine Learning*, pages 201–210. PMLR, 2016.

[Gupta and Raskar, 2018] Otkrist Gupta and Ramesh Raskar. Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications*, 116:1–8, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2019] Zecheng He, Tianwei Zhang, and Ruby B Lee. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 148–162, 2019.

[Kairouz *et al.*, 2019] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[Konečný *et al.*, 2015] Jakub Konečný, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575*, 2015.

[McMahan *et al.*, 2017] 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*, pages 1273–1282. PMLR, 2017.

[Melis *et al.*, 2019] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706. IEEE, 2019.

[Nasr *et al.*, 2019] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

[Shokri and Shmatikov, 2015] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

[Shokri *et al.*, 2017] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[Thapa *et al.*, 2020] Chandra Thapa, Mahawaga Arachchige Pathum Chamikara, and Seyit Camtepe. Splitfed: When federated learning meets split learning. *arXiv preprint arXiv:2004.12088*, 2020.

[Yin *et al.*, 2021] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. *arXiv preprint arXiv:2104.07586*, 2021.

[Zhang *et al.*, 2020] Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Gang Hua, and Nenghai Yu. Passport-aware normalization for deep model protection. *Advances in Neural Information Processing Systems*, 33, 2020.

[Zhao *et al.*, 2020] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

[Zhu and Han, 2020] Ligeng Zhu and Song Han. Deep leakage from gradients. In *Federated Learning*, pages 17–31. Springer, 2020.