SCIENCE CHINA
Information Sciences

• LETTER •

August 2023, Vol. 66 189303:1–189303:2
https://doi.org/10.1007/s11432-022-3678-7

DP-GenFL: a local differentially private federated learning system through generative data

Jun LI[1], Kang WEI[1], Chuan MA[2,1,3]* & Feng SHU[4]

[1] *School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China;*
[2] *Zhejiang Lab, Hangzhou 310000, China;*
[3] *Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing 210096, China;*
[4] *School of Information and Communication Engineering, Hainan University, Haikou 570228, China*

Received 26 September 2022/Revised 9 December 2022/Accepted 6 January 2023/Published online 11 July 2023

Citation Li J, Wei K, Ma C, et al. DP-GenFL: a local differentially private federated learning system through generative data. Sci China Inf Sci, 2023, 66(8): 189303, https://doi.org/10.1007/s11432-022-3678-7

With the rapid development of the Internet of Things (IoT), the amount of data from intelligent devices is propagating at unprecedented scales. Meanwhile, machine learning (ML), which relies heavily on such data, is revolutionizing many aspects of our lives [1]. However, conventional centralized ML offers little scalability for efficiently processing this huge amount of data. Moreover, privacy and confidentiality are of increasing concern as exchanged data often contains clients' sensitive information in distributed ML settings. In this light, federated learning (FL) has been proposed, which allows the decoupling of data provision at local clients and aggregating ML models by a central server [2].

However, although data is not explicitly shared in the original format in FL, there exists a risk that adversaries/eavesdroppers can reconstruct local data approximately, especially when the ML architectures and models are not completely protected. In addition, FL will expose intermediate results, such as model updates, during uploading, and the transmission of these updates is vulnerable when exposed under memorization of sensitive training samples [3] and membership inference attack [4]. To prevent such privacy leakage, a popular approach in the literature has been proposed, termed differentially-private FL (DP-FL) [5], which guarantees differential privacy (DP) for releasing model parameters. Nevertheless, the learning performance suffers a lot from the additive noises.

To alleviate such performance degradation, in this work, we propose to use generative adversarial networks (GAN) to produce fake training samples in each local client, and theoretically prove that the generative data samples can also achieve local DP under some systematical assumptions. On the basis of the generative data examples, we then propose an enhanced privacy protection method that elaborates on the fake samples as well as the DP-FL in the training procedure. In addition, background descriptions on the DP and FL are provided in Appendix A.

System framework. The proposed model can be mainly divided into two steps, namely conditional-GAN-based data augmentation, and enhanced privacy protection on FL, respectively.

- The first step adopts the conditional GAN, which uses the labels as the additive input of a generator $\mathbb{G}$ and a discriminator $\mathbb{D}$, to generate required data samples in each local client.
- In the second step, each client uses a mixture of the generated fake samples and the real ones to train its local model, applies DP perturbation on the updates, and then finishes this round of local training.

Enhanced privacy protection on FL. Based on Algorithm B1, we can design the privacy-enhanced FL framework with synthetic data. In the beginning, the server will broadcast the initial global model $\boldsymbol{w}^{(0)}$ to all clients. Then, each client generates the synthetic data $\widetilde{\mathcal{D}}_i$ using Algorithm B1. After finishing the data augmentation, each client needs to combine $\widetilde{\mathcal{D}}_i$ and $\mathcal{D}_i$ with ratios γ and $1-\gamma$, respectively, to generate $\widehat{\mathcal{D}}_i$ as the local training dataset. By this combination, each client can train the local model $\boldsymbol{w}_i^{(t)}$ with a low privacy risk. It can be noticed that with a given noise standard deviation σ_i, the privacy budget can be calculated in each communication round. In addition, the validation accuracy $\mathcal{V}_i^{(t)}$ for the ith client can be calculated, and then uploaded to the server as well as the local model $\boldsymbol{w}_i^{(t)}$. After receiving all local models, the server aggregates all received local models by $\boldsymbol{w}^{(t+1)} = \frac{1}{N}\sum_{i\in\mathcal{N}}\boldsymbol{w}_i^{(t)}$ and calculate the global validation accuracy by $\mathcal{V}^{(t+1)} = \frac{1}{N}\sum_{i\in\mathcal{N}}\mathcal{V}_i^{(t)}$. When $\boldsymbol{w}^{(t+1)}$ satisfies the requirement (up to the privacy budget) or $t+1$ is up to the maximum number of the communication round T, the server will set $\boldsymbol{w}^f = \boldsymbol{w}^{(t+1)}$ and stop this algorithm. Otherwise, the algorithm will start the next communication round. The whole procedure (Algorithm B2) can be found in Appendix B.2. In addition, the DP and convergence analysis can be found in Appendixes C and D.

* Corresponding author (email: chuan.ma@njust.edu.cn)

Table 1 Comparisons of the precision of the membership inference attack (left) and test accuracy (right)

Dataset	No privacy	DP-FL ($\epsilon = 4$)	DP-GenFL ($\gamma = 1.0$)	DP-GenFL ($\gamma = 0.6$)	DP-GenFL ($\gamma = 0.2$)
MNIST	0.584/0.977	0.533/0.934	0.510/0.501	0.515/0.947	0.519/0.962
Fashion-MNIST	0.613/0.826	0.543/0.769	0.510/0.550	0.518/0.750	0.533/0.789
CIFAR-10	0.668/0.553	0.543/0.350	0.512/0.212	0.548/0.506	0.551/0.545
Purchase	0.680/0.602	0.550/0.281	0.506/0.200	0.549/0.478	0.598/0.534

Experimental results. We conduct experimental results to evaluate the privacy protection performance of the proposed algorithm under the membership inference attack as well as test accuracy under specific tasks compared with the classic DP-FL [5]. As shown in Table 1, DP-GenFL can further defend the membership inference attack compared to DP-FL and achieve a remarkable performance gain. For example, the successful attacking rate drops to 51.03% when $\gamma = 0.6$ in the MNIST dataset, which is slightly higher than the random guess (50%). Additional results as well as the experimental setup are provided in Appendix E.

Conclusion. We have provided an enhanced privacy protection mechanism for FL. With the help of GAN, a local DP-FL framework has been proposed, and the privacy level, as well as the convergence performance, have been systemically conducted. Experimental results have shown the effectiveness and the advantages of the proposed algorithm by comparing it with the state of the art. In addition, further directions can find an optimal fake ratio and the number of training epochs for the proposed framework.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. U22A2002, 61872184, 62002170, 62071234), Fundamental Research Funds for the Central Universities (Grant No. 30921013104), Future Network Grant of Provincial Education Board in Jiangsu, Zhejiang Lab (Grant No. 2022PD0AC02), Major Science and Technology Plan of Hainan Province (Grant No. ZDKJ2021022), and Scientific Research Fund Project of Hainan University (Grant No. KYQD(ZR)-21008).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy. In: Proceedings of ACM SIGSAC Conference on Computer and Communications Security, 2016. 308–318

2 Wei K, Li J, Ding M, et al. Federated learning with differential privacy: algorithms and performance analysis. IEEE Trans Inform Forensic Secur, 2020, 15: 3454–3469

3 Webster R, Rabin J, Simon L, et al. Detecting overfitting of deep generative networks via latent recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 11273–11282

4 Shokri R, Stronati M, Song C Z, et al. Membership inference attacks against machine learning models. In: Proceedings of the IEEE Symposium on Security and Privacy, 2017. 3–18

5 Wei K, Li J, Ding M, et al. User-level privacy-preserving federated learning: analysis and performance optimization. IEEE Trans Mobile Comput, 2022, 21: 3388–3401