# 1 Appendix

## 1.1 Preliminaries and Related Works

### 1.1.1 Federated Learning

Suppose there are $m$ clients in a FL system, and each client $k$ has its own dataset $\mathcal{D}_k = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n_k}$ with $n_k = |\mathcal{D}_k|$ being the size of local data. Each client $k$ optimizes its local model by minimizing the following objective function,

$$\min_{\boldsymbol{\theta}^k} \frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{L}(f^k(\mathbf{x}_i), \mathbf{y}_i), \tag{1}$$

where $f^k()$ and $\boldsymbol{\theta}^k$ are the local model and model parameter, respectively. Then, client $k$ uploads $\boldsymbol{\theta}^k$ to the server for aggregation. After receiving the uploaded model parameters from the $m$ clients, the server computes the aggregated global model parameter as follows: $\boldsymbol{\theta}_S = \sum_{k=1}^{m} \frac{n_k}{n} \boldsymbol{\theta}^k$, where $n = \sum_{k=1}^{m} n_k$ is the total amount of training data (of all clients) and $\boldsymbol{\theta}_S$ is the server's global model parameter. Afterwards, the server distributes $\boldsymbol{\theta}_S$ to all clients for training in the next round. However, such a training procedure needs frequent communication between the clients and the server, thus incurs a high communication cost, which may be intolerable in practice [7].

### 1.1.2 One-shot Federated Learning

A promising solution to reduce the communication cost in FL is one-shot FL, which is first introduced by [3]. In one-shot FL, each client only uploads its local model parameter to the server once. After obtaining the global model, the server does not need to distribute the global model to the clients for further training. There is only one unidirectional communication between clients and the server, thus it highly reduces the communication cost and makes it more practical in reality. Moreover, one-shot FL also reduces the risk of being attacked, since the communication happens only once. However, the main problem of one-shot FL is the difficult convergence of the global model and it is hard to achieve a satisfactory performance especially when the data on clients are not independent and identically distributed (non-IID).

Guha *et al.* [3] and Li *et al.* [7] used ensemble distillation to improve the performance of one-shot FL. However, they introduced a public dataset to enhance training, which is not practical. In addition to model distillation, dataset distillation [11] is also a prevailing approach. Zhou *et al.* [13] and [5] proposed to apply dataset distillation to one-shot FL, where each client distills its private dataset and transmits distilled data to the server. However, data distillation methods fail to offer satisfactory performance compared with model distillation and sending distilled data can cause additional communication cost and potential privacy leakage. Dennis *et al.* [2] utilized cluster-based method in one-shot FL, but they required to upload the cluster means to the server, which incurs additional communication cost.

Overall, none of the above methods can be practically applied. In addition, none of these studies consider model heterogeneity, which is a main challenge in FL [8]. This leads to a fundamental yet so far unresolved question: *"Is it possible to conduct one-shot FL without the need to share additional information or rely on any auxiliary dataset, while making it compatible with model heterogeneity?"*

### 1.1.3 Knowledge Distillation in FL

In traditional FL frameworks, all users have to agree on the specific architecture of the global model. To support model heterogeneity, Li *et al.* [6] proposed a new federated learning framework that enables participants to independently design their models by knowledge distillation [4]. With the use of a proxy dataset, knowledge distillation alleviates the model drift issue induced by non-IID data. However, the requirement of proxy data renders such a method impractical for many applications, since a carefully designed dataset is not always available on the server.

Data-free knowledge distillation is a promising approach, which can transfer knowledge of a teacher model to a student model without any real data [1, 12]. Lin *et al.* [9] proposed data-free ensemble distillation for model fusion through synthetic data in each communication round, which requires high communication costs and computational costs. However, in this paper, we are more concerned with

obtaining a good global model through only one round of communication in cases of heterogeneous models, which is more challenging and practical.

Zhu *et al.* [14] also proposed a data-free knowledge distillation approach for FL, which learns a generator derived from the prediction of local models. However, the learned generator is later broadcasted to all clients, and then clients need to send their generators to the server, which increases the communication burden. More seriously, the generator has direct access to the local data (the generator can easily remember the training samples [10]), which can cause privacy concerns. As the generator used in our method is always stored in the central server, it never sees any real local data.

## References


[1] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Dafl: Data-free learning of student networks. In *ICCV*, 2019.

[2] Don Kurian Dennis, Tian Li, and Virginia Smith. Heterogeneity for the win: One-shot federated clustering, 2021.

[3] Neel Guha, Ameet Talwalkar, and Virginia Smith. One-shot federated learning. *arXiv preprint arXiv:1902.11175*, 2019.

[4] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[5] Anirudh Kasturi, Anish Reddy Ellore, and Chittaranjan Hota. Fusion learning: A one shot federated learning. In *International Conference on Computational Science*, pages 424–436. Springer, 2020.

[6] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.

[7] Qinbin Li, Bingsheng He, and Dawn Song. Practical one-shot federated learning for cross-silo setting. *arXiv preprint arXiv:2010.01017*, 2020.

[8] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *CoRR, arXiv:1908.07873*, 2019.

[9] Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[10] Yi Liu, Jialiang Peng, JQ James, and Yi Wu. Ppgan: Privacy-preserving generative adversarial network. In *2019 IEEE 25Th international conference on parallel and distributed systems (ICPADS)*, pages 985–989. IEEE, 2019.

[11] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation, 2020.

[12] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020.

[13] Yanlin Zhou, George Pu, Xiyao Ma, Xiaolin Li, and Dapeng Wu. Distilled one-shot federated learning. *arXiv preprint arXiv:2009.07999*, 2020.

[14] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12878–12889. PMLR, 2021.