Supplementary Materials: One-shot-but-not-degraded Federated Learning

Hui Zeng*
School of Computer,
National University of Defense Technology,
Changsha, China
zenghui116@nudt.edu.cn

Minrui Xu*
College of Computing and Data Science,
Nanyang Technological University,
Singapore, Singapore
minrui001@e.ntu.edu.sg

Tongqing Zhou†
School of Computer,
National University of Defense Technology,
Changsha, China
zhoutongqing@nudt.edu.cn

Xinyi Wu
School of Computer,
National University of Defense Technology,
Changsha, China
wuxinyi17@nudt.edu.cn

Jiawen Kang
School of Automation,
Guangdong University of Technology,
Guangzhou, China
kavinkang@gdut.edu.cn

Zhiping Cai†
School of Computer,
National University of Defense Technology,
Changsha, China
zpcai@nudt.edu.cn

Dusit Niyato
College of Computing and Data Science,
Nanyang Technological University,
Singapore, Singapore
dniyato@ntu.edu.sg

ACM Reference Format:
Hui Zeng, Minrui Xu, Tongqing Zhou, Xinyi Wu, Jiawen Kang, Zhiping Cai, and Dusit Niyato. 2024. Supplementary Materials: One-shot-but-not-degraded Federated Learning. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3664647.3680715

A Algorithm Details

Here, we present the detailed algorithm of IntactOFL. The IntactOFL is a one-shot aggregation algorithm, that performs after the local training phase. Once the server receives all the local models from the participants, the server performs IntactOFL for aggregation. The IntactOFL aims to train a gating network $\mathcal{G}$ and use the local models to form a dynamic and high-performance MoE network according to different input samples in a data-free manner. The first step is using the local models to generate auxiliary data. With the random noise and pseud-label as input, the generator can generate a batch of the auxiliary samples according to

$$\mathcal{L}_{gen}^{t}(\hat{x}, y, f_{\mathcal{M}}^{t}) = \sigma(f_{\mathcal{M}}^{t}(\hat{x}))\ell_{CE}(f_{\mathcal{M}}^{t}(\hat{x}), y), \tag{1}$$

where $\sigma(\cdot)$ is the variance function. Then, through data augmentations, more diverse samples will be generated. Based on the generated samples, we update the gating network $\mathcal{G}$ according to

$$\min_{\mathcal{G}} \mathcal{L}_{train}^{t}(\{w_k\}_{k=1}^{m}) = \mathcal{L}_{\mathcal{M}}^{t}(\{w_k\}_{k=1}^{m}; \mathbb{D}_A) + \mathcal{L}_{balance}^{t}(\mathbb{D}_A). \tag{2}$$

Algorithm 1 IntactOFL

Input: Clients' local models $\{w_k\}_{k=1}^{m}$, auxiliary dataset $\mathbb{D}_A = \emptyset$, generator g, gating network $\mathcal{G}$, learning rate of generator and gating network η_g and $\eta_{\mathcal{G}}$, generation iterations T_g, MoE training epochs T, and batch size b.
Output: MoE network: gating network $\mathcal{G}$ and experts $\{w_k\}_{k=1}^{m}$.

for each epoch t to T **do**
  // Generate auxiliary data;
  Sample a batch of noises and labels $\{z_i, y_i\}_{b}^{i=1}$;
  for each t_g to T_g **do**
    Generate samples $\{\hat{x}_i\}_{b}^{i=1} = g(\{z_i\}_{b}^{i=1})$;
    // Update the generator;
    $g \leftarrow g - \eta_g \nabla_g \mathcal{L}_{gen}^{t}$, according to Eq.(1);
  end for
  $\mathbb{D}_A \leftarrow \mathbb{D}_A \cup Augmentation(\{\hat{x}_i\}_{b}^{i=1})$;
  // Update gating network;
  for each sampling batch $\{x_i\}_{b}^{i=1}$ in $\mathbb{D}_A$ **do**
    $\mathcal{G} \leftarrow \mathcal{G} - \eta_{\mathcal{G}} \nabla_{\mathcal{G}} \mathcal{L}_{train}(\mathcal{G})$, according to Eq.(2).
  end for
end for

After T iterations, the well-trained gating network can dynamically allocate the input samples to specific local models for high performance.

*Both authors contributed equally to this research.
†Corresponding authors

B More Experimental Results

B.1 Dataset Details

We evaluate the proposed IntactOFL and baselines on four widely used classification benchmarks (CIFAR-10, CIFAR-100, SVHN, and Tiny-ImageNet) and one domain generalization benchmark (PACS). CIFAR-10 consists of 60,000 32×32 RGB images in 10 categories. It has 50,000 training samples and 10,000 test samples. CIFAR-100 has the same format as CIFAR-10, except it has 100 categories. SVHN is a real-world dataset comprising over 600,000 digit images extracted from Google Street View images. Tiny-ImageNet is the subset of the ImageNet dataset, containing 200 categories and 100,000 images (500 images for each category), each image is downsized to 64 $\times$ 64. PACS is a widely used benchmark in domain generalization. It contains overall 9991 images, split unevenly among 7 classes with 4 different domains (Art painting, Cartoon, Photo, and Sketch).

B.2 Implementation Details

We adopt the same local training setting as [2]. We use the SGD optimizer with a learning rate of 0.01 and a fixed momentum of 0.9. Each local model is trained on the client's local data for $E = 400$ rounds, and the client number is 5. Following the setting of [1, 2], we train the generator g with a DNN. We adopt the Adam optimizer with a learning rate of 0.001 and training for $T_g = 20$ iterations. For the training of the MoE network, we use the SGD optimizer with a learning rate $\eta_{\mathcal{G}} = 0.01$ and momentum of 0.9 and train the gating network for $T = 40$ iterations.

B.3 Metrics

We utilize the test accuracy as the prime metric over all baselines and the proposed IntactOFL. For results with error bars, we run five repeated experiments with different random seeds.

B.4 Effectiveness in Domain Shift Setting

We also consider the effectiveness of the proposed IntactOFL in domain shift settings. We select a widely used domain shift dataset, PACS, following the same data partition as FedSR. The results are shown in Table 1. We conclude that IntactOFL can outperform other baselines in domain shift scenarios.

Table 1: Performance on an OOD dataset (PACS).

Method	Accuracy	Method	Accuracy	Method	Accuracy
MA-Echo	21.78±0.51	F-ADI	18.45±1.99	DENSE	21.55±1.45
O-FedAvg	20.44±0.91	F-DAFL	17.89±1.44	Co-Boosting	20.82±0.51
FedDF	17.34±1.63	Ensemble	22.15±0.34	**Ours**	**39.77±0.43**

B.5 Horizontal Scalability Analysis

We provide more horizontal scalability analysis on SVHN (see Figure 2), which is the test accuracy across different numbers of clients. The conclusion is the same with the test on CIFAR-10. With the increasing number of clients, the local data become more sparser and more fragmented. Consequently, the local models trained from such data are highly prone to overfitting, resulting in inferior performance. In summary, we also still conclude that the proposed IntactOFL is scalable across diverse distributed networks of varying sizes.

Table 2: Test accuracy of the server model on SVHN across different numbers of clients $m = \{5, 10, 25, 50, 100\}$.

m	5	10	25	50	100
MA-Echo	80.23	67.12	59.89	56.23	47.54
O-FedAvg	57.61	45.22	42.18	33.98	30.15
FedDF	72.11	61.09	58.89	52.32	44.18
F-ADI	77.62	64.39	60.98	59.63	48.39
F-DAFL	74.55	62.39	60.88	53.69	47.95
Ensemble	81.22	67.51	63.12	56.68	48.25
DENSE	80.03	68.98	62.39	59.99	53.76
Co-Boosting	81.34	69.71	63.85	60.01	55.15
Ours	**84.81**	**72.95**	**68.44**	**65.64**	**59.19**

B.6 Visualization of Generated Auxiliary Data

We provide the visualization of the generated data on CIFAR-10 and SVHN in Figure 1. It is worth noting that the goal of the data generator is designed to generate the data which is similar in utilization not in visualization. The generated data looks different from the original data, which can mitigate the risk of leaking sensitive information. Meanwhile, it plays an important role in training the MoE network, which extracts the information from local models and helps achieve higher performance than other baselines.

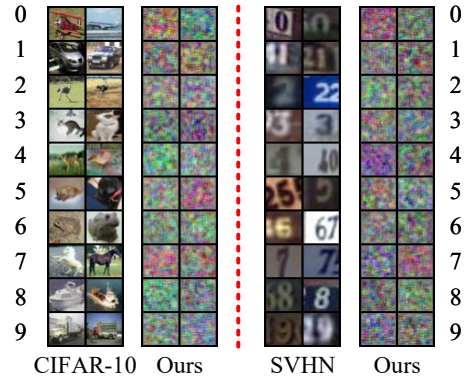

Figure 1: Visualization of generated data on CIFAR-10 and SVHN.

B.7 Impact of Gating Network Architecture

We investigate the impact of different gating network architectures on performance. The gating network outputs the weights of each expert according to the input, which can be viewed as a function. We vary the different architectures of the gating network, including MLP, CNN, and ResNet, with the primary difference lying in their capabilities for information processing. The results are shown in Table 3. We conclude that different gating network architecture has a limited impact on the MoE network performance.

Table 3: Test accuracy of the MoE network on CIFAR-10 across different gating network architectures (MLP, CNN, and ResNet).

gating network	MLP	CNN	ResNet
CIFAR-10	79.93	79.23	**80.04**
CIFAR-100	46.78	46.55	**46.88**
SVHN	84.81	84.77	**85.00**
Tiny-ImageNet	35.09	34.69	**35.11**

References

[1] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. 2019. Data-free learning of student networks. In *Proc. of ICCV*. 3514–3522.

[2] Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. 2022. Dense: Data-free one-shot federated learning. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 21414–21428.