DENSE: Data-Free One-Shot Federated Learning

Jie Zhang[1], Chen Chen[1], Bo Li[2], Lingjuan Lyu[3], Shuang Wu[2], Shouhong Ding[2], Chunhua Shen[1], Chao Wu[1]

[1]Zhejiang University [2]Youtu Lab, Tencent [3]Sony AI

Motivation

➢ The original FL framework requires participants to communicate frequently with the central server. In real-world FL, such high communication cost may be intolerable and impractical.

➢ Frequent communication poses a high risk of being attacked.

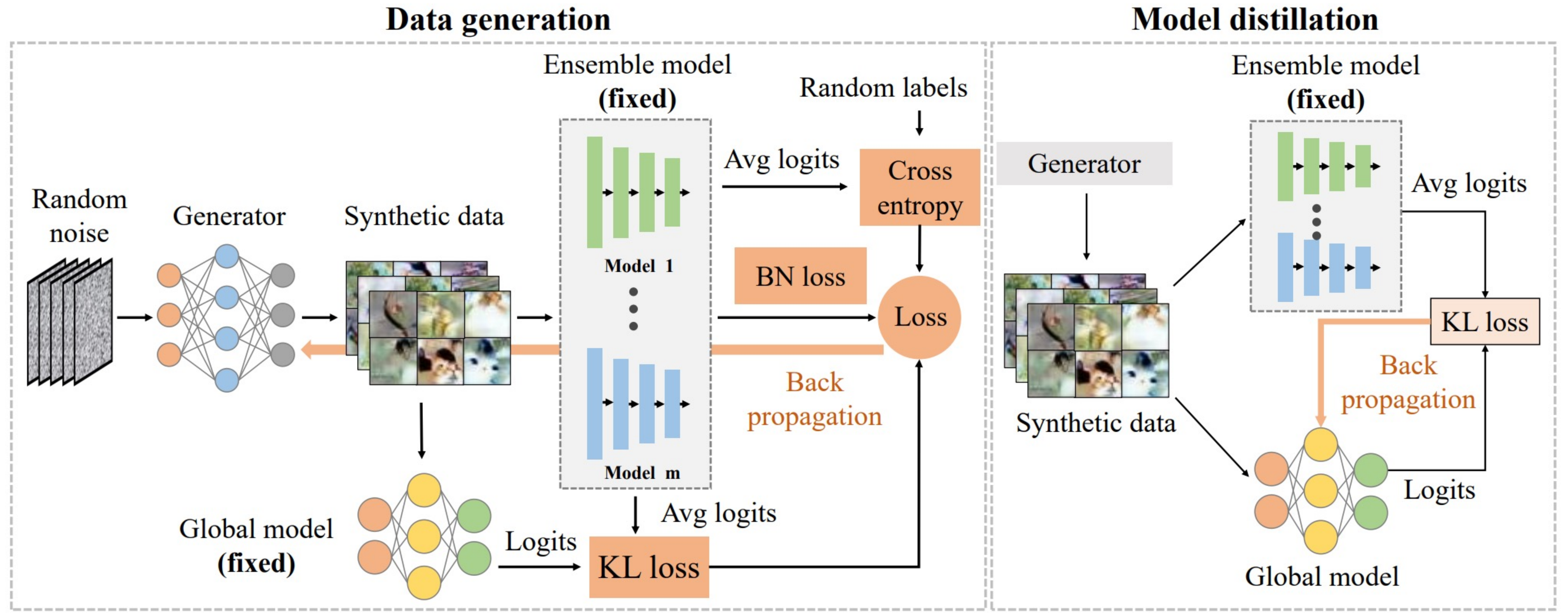
➢ Eexisting one-shot FL methods are mostly impractical or face inherent limitations, e.g., a public dataset is required, clients' models are homogeneous, and additional data/model information need to be uploaded.

Contribution

The setting of DENSE is practical in the following aspects.

- ➢ DENSE requires no additional information (except the model parameters) to be transferred between clients and the server
- ➢ DENSE does not require any auxiliary dataset for training
- ➢ DENSE considers model heterogeneity, i.e., different clients can have different model architectures

Overview Framework of Our Method



Formulation Objective

I. Data Generation:

➢ Similarity

$$D(\hat{\mathbf{x}};\{\boldsymbol{\theta}^k\}_{k=1}^m)=\frac{1}{m}\sum_{k\in\mathcal{C}}f^k\left(\hat{\mathbf{x}};\boldsymbol{\theta}^k\right),$$

$$\mathcal{L}_{CE}(\hat{\mathbf{x}},\mathbf{y};\boldsymbol{\theta}_G)=CE(D(\hat{\mathbf{x}}),\mathbf{y}),$$

➢ Stability

$$\mathcal{L}_{BN}(\hat{\mathbf{x}};\boldsymbol{\theta}_G)=\frac{1}{m}\sum_{k\in\mathcal{C}}\sum_{l}\left(\|\mu_l(\hat{\mathbf{x}})-\mu_{k,l}\|+\left\|\sigma_l^2(\hat{\mathbf{x}})-\sigma_{k,l}^2\right\|\right),$$

➢ Transferability

$$\mathcal{L}_{div}(\hat{\mathbf{x}};\boldsymbol{\theta}_G)=-\omega KL\left(D(\hat{\mathbf{x}}),f_S(\hat{\mathbf{x}};\boldsymbol{\theta}_S)\right),$$

■ By combining the above losses, we can obtain the generator loss as follows,

$$\mathcal{L}_{gen}(\hat{\mathbf{x}},\mathbf{y};\boldsymbol{\theta}_G)=\mathcal{L}_{CE}(\hat{\mathbf{x}},\mathbf{y};\boldsymbol{\theta}_G)+\lambda_1\mathcal{L}_{BN}(\hat{\mathbf{x}};\boldsymbol{\theta}_G)+\lambda_2\mathcal{L}_{div}(\hat{\mathbf{x}};\boldsymbol{\theta}_G),$$

II. Model Distillation:

➢ Loss for distillation

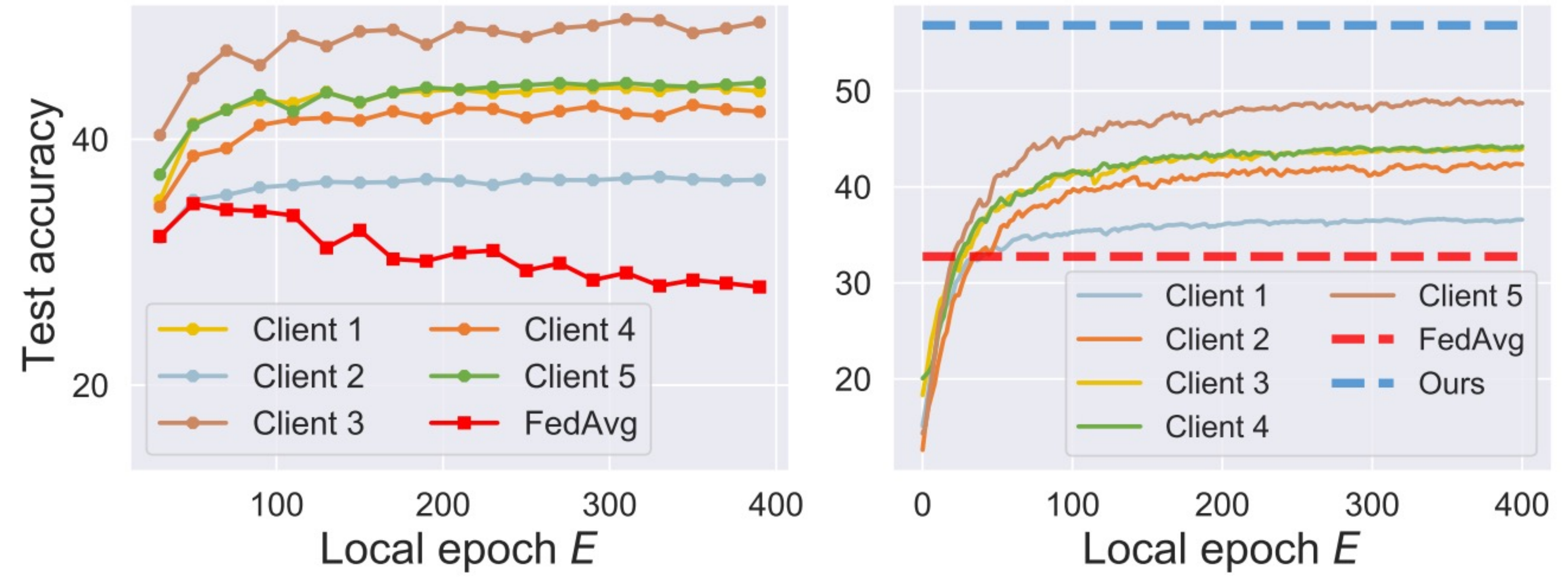
$$\mathcal{L}_{dis}(\hat{\mathbf{x}};\boldsymbol{\theta}_S)=KL\left(D(\hat{\mathbf{x}}),f_S(\hat{\mathbf{x}};\boldsymbol{\theta}_S)\right).$$

Algorithm 1 Training process of DENSE

Input: Number of client m, clients' local models $\{f^1(),\cdots,f^m()\}$, generator $G(\cdot)$ with parameter $\boldsymbol{\theta}_G$, learning rate of the generator η_G, number of training rounds T_G for generator in each epoch, global model $f_S()$ with parameter $\boldsymbol{\theta}_S$, learning rate of the global model η_S, global model training epochs T, and batch size b.

for each client $k\in\mathcal{C}$ **in parallel do**
  $\boldsymbol{\theta}^k\leftarrow$ **LocalUpdate**(k)
end for
Initialize parameter $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_S$
for epoch=$1,\cdots,T$ **do**
  Sample a batch of noises and labels $\{\mathbf{z}_i,\mathbf{y}_i\}_{i=1}^b$
  // data generation stage
  for $j=1,\cdots,T_G$ **do**
    Generate $\{\hat{\mathbf{x}}_i\}_{i=1}^b$ with $\{\mathbf{z}_i\}_{i=1}^b$ and $G(\cdot)$
    $\boldsymbol{\theta}_G\leftarrow\boldsymbol{\theta}_G-\eta_G\frac{1}{b}\sum_{i=1}^b\nabla_{\boldsymbol{\theta}_G}\ell_{gen}(\hat{\mathbf{x}}_i,\mathbf{y}_i;\boldsymbol{\theta}_G)$
  end for
  // model distillation stage
  Generate $\{\hat{\mathbf{x}}_i\}_{i=1}^b$ with $\{\mathbf{z}_i\}_{i=1}^b$ and $G(\cdot)$
  $\boldsymbol{\theta}_S=\boldsymbol{\theta}_S-\eta_S\frac{1}{b}\sum_{i=1}^b\nabla_{\boldsymbol{\theta}_S}\ell_{dis}(\hat{\mathbf{x}}_i;\boldsymbol{\theta}_S)$
end for

FedAvg is not suitable for one-shot FL



■ Left panel: Accuracy of FedAvg and clients' local models across different local training epochs E = {20, 40, 60, · · · , 400}. Right panel: The accuracy curve for local training. The dotted lines represent the best results of two one-shot FL methods (FedAvg and DENSE). Our DENSE outperforms FedAvg and local models consistently.

■ For FedAvg, a larger value of E can cause the model to degrade even collapse. This result can be attributed to the inconsistent optimization objectives with non-IID data, which leads to weight divergence.

Experiments

➢ Main Results

Table 1: Accuracy of different methods across $\alpha=\{0.1,0.3,0.5\}$ on different datasets.

Dataset	MNIST			FMNIST			CIFAR10			SVHN			CIFAR100			Tiny-ImageNet		
Method	α=0.1	α=0.3	α=0.5	α=0.1	α=0.3	α=0.5	α=0.1	α=0.3	α=0.5	α=0.1	α=0.3	α=0.5	α=0.1	α=0.3	α=0.5	α=0.1	α=0.3	α=0.5
FedAvg	48.24	72.94	90.55	41.69	82.96	83.72	23.93	27.72	43.67	31.65	61.51	56.09	4.58	11.61	12.11	3.12	10.46	11.89
FedDF	60.15	74.01	92.18	43.58	80.67	84.67	40.58	46.78	53.56	49.13	73.34	73.98	28.17	30.28	36.35	15.34	18.22	27.43
Fed-DAFL	64.38	74.18	93.01	47.14	80.59	84.02	47.34	53.89	58.59	53.23	76.56	78.03	28.89	34.89	38.19	18.38	22.18	28.22
Fed-ADI	64.13	75.03	93.49	48.49	81.15	84.19	48.59	54.68	59.34	53.45	77.45	78.85	30.13	35.18	40.28	19.59	25.34	30.21
DENSE (ours)	**66.61**	**76.48**	**95.82**	**50.29**	**83.96**	**85.94**	**50.26**	**59.76**	**62.19**	**55.34**	**79.59**	**80.03**	**32.03**	**37.32**	**42.07**	**22.44**	**28.14**	**32.34**

➢ heterogeneous client models

Table 2: Accuracy comparisons across heterogeneous client models on CIFAR10. There are five clients in total, and each client has a personalized model.

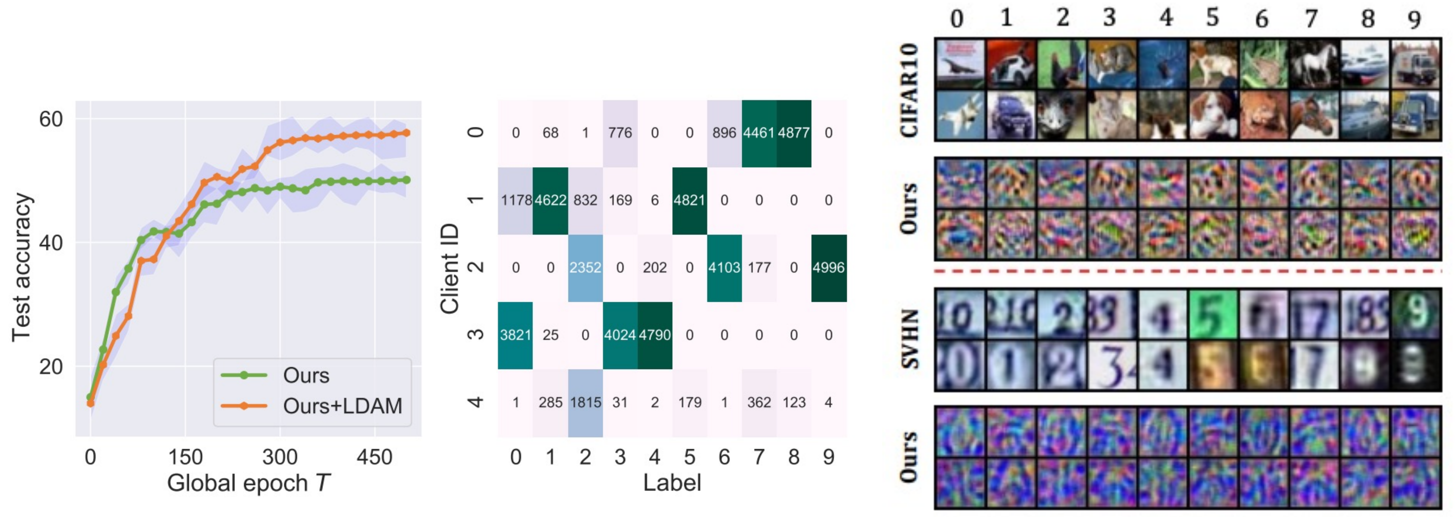
Model	Client					Server (ResNet-18)			
	ResNet-18	CNN1	CNN2	WRN-16-1	WRN-40-1	FedDF	Fed-DAFL	Fed-ADI	**DENSE (ours)**
α=0.1	40.83	33.67	35.21	27.73	32.93	42.35	43.12	44.63	**49.76**
α=0.3	51.49	52.78	44.96	47.35	37.24	52.72	57.72	58.96	**63.25**
α=0.5	59.96	58.67	54.28	53.39	58.14	60.05	61.56	63.24	**67.42**

➢ More clients

Table 3: Accuracy across different number of clients $m=\{5,10,20,50,100\}$ on CIFAR10 and SVHN datasets.

Dataset	CIFAR10					SVHN				
m	FedAvg	FedDF	Fed-DAFL	Fed-ADI	**DENSE (ours)**	FedAvg	FedDF	Fed-DAFL	Fed-ADI	**DENSE (ours)**
5	43.67	53.56	55.46	58.59	**62.19**	56.09	73.98	78.03	78.85	**80.03**
10	38.29	54.44	56.34	57.13	**61.42**	45.34	62.12	63.34	65.45	**67.57**
20	36.03	43.15	45.98	46.45	**52.71**	47.79	60.45	62.19	63.98	**66.42**
50	37.03	40.89	43.02	44.47	**48.47**	36.53	51.44	54.23	57.35	**59.27**
100	33.54	36.89	37.55	36.98	**43.28**	30.18	46.58	47.19	48.33	**52.48**

➢ Combination with imbalanced learning and visualization of synthetic data



Contact us

zj_zhangjie@zju.edu.cn, libraboli@tencent.com