Data-Free One-Shot Federated Regression: An Application to Bone Age Assessment

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

We consider a novel problem setting: data-free one-shot federated regression. This setting aims to prepare a global model through a single round of communication without relying on auxiliary information, *e.g.*, proxy datasets. To address this problem, we propose a practical framework that consists of three stages: local training, data synthesizing, and knowledge distillation, and demonstrate its efficacy with an application to bone age assessment. We conduct validation under independent and identical distribution (IID) and non-IID settings while considering both model homogeneity and heterogeneity. Validation results show that our method surpasses `FedAvgOneShot` by a large margin and sometimes even outperforms the proxy-data-dependent approach `FedOneShot`.

Keywords: Federated learning, Regression, One-shot, Data-free.

1. Introduction

One-shot federated learning (FL) (Guha et al., 2019) has emerged as a potential solution to address concerns regarding the costly inter-node communication and possible privacy leakage in standard FL methods, as it allows for only a single global round between clients and the central server. While one-shot FL methods typically require additional sources like proxy datasets for global model training, recent advances in data-free one-shot FL (Zhang et al., 2022; Luz-Ricca et al., 2023) overcomes this limitation, eliminating the need for additional datasets. Nevertheless, current FL methods have predominantly focused on classification tasks, with a limited exploration of regression problems.

Motivated by these observations, we consider a novel problem setting: data-free one-shot federated regression. Inspired by the work (Zhang et al., 2022) that proposed for classification, we present a practical framework specialized for regression, which comprises three stages: **local training**, **data synthesizing**, and **knowledge distillation** (KD), and evaluate it with a bone age assessment task (Halabi et al., 2019). Our method is the first attempt in this setting, and validation results demonstrate its efficacy.

2. Method

Generally, let there be K local clients $\{C_i\}_{i=1}^{K}$ with each client holding a private dataset $\mathbf{D}_i = \{\mathbf{x}_i, \mathbf{y}_i\}$, where $\mathbf{x}_i$ are images, and $\mathbf{y}_i$ are ground truth, *e.g.*, bone ages in our study.

First stage: local training. Each client C_i trains its local model $M_i(\cdot, \mathbf{\Theta}i)$ with the private dataset $\mathbf{D}_i$ and uploads the model weight to the central server after training.

Second stage: data synthesizing. We adopt a generator $G(\cdot, \mathbf{\Theta}g)$ to synthesize images. In our study, bone ages range from 1 to 228 months, and we assume $p(\mathbf{y})$ as a discrete

© 2023 CC-BY 4.0, Anonymous.

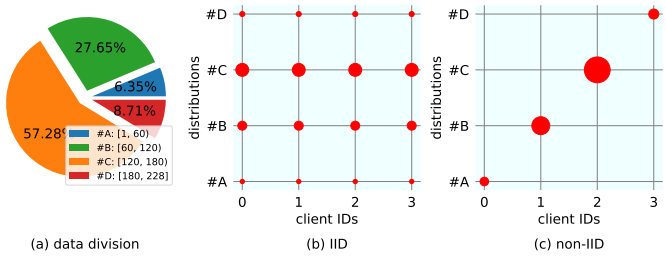

Figure 1: Details of experimental setup. (a) Data division. (b) Simulated IID. (c) Simulated non-IID. The size of each red circle is proportional to the number of samples.

uniform distribution $U(\{1, 2, \ldots, 228\})$. To train $G(\cdot, \boldsymbol{\Theta}g)$, we first sample a batch of random noise vectors $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$ and a batch of random values $\hat{\mathbf{y}} \sim p(\mathbf{y})$. Afterward, we input $\mathbf{z}$ to $G(\cdot, \boldsymbol{\Theta}g)$ to get a batch of produced images $\hat{\mathbf{x}} = G(\mathbf{z}, \boldsymbol{\Theta}g)$. Next, we feed $\hat{\mathbf{x}}$ into local models $\{M_i(\cdot, \boldsymbol{\Theta}i)\}_{i=1}^{K}$ to get predictions $\{M_i(\hat{\mathbf{x}}, \boldsymbol{\Theta}i)\}_{i=1}^{K}$. By adopting the basic ensemble scheme (Mendes-Moreira et al., 2012), we get the ensembled result of local models $E(\hat{\mathbf{x}}) = \sum_1^K M_i(\hat{\mathbf{x}}, \boldsymbol{\Theta}i)/K$. We calculate a loss $\mathcal{L}_{sim}(E(\hat{\mathbf{x}}), \hat{\mathbf{y}}) = \|E(\hat{\mathbf{x}}) - \hat{\mathbf{y}}\|_2$ to expect $\hat{\mathbf{x}}$ following a similar distribution to $\mathbf{x}$. Besides, to improve the quality of $\hat{\mathbf{x}}$, we adopt a feature distribution regularization term $\mathcal{L}_{feat}$ (Yin et al., 2020), which enforces feature-level similarity and is defined as $\mathcal{L}_{feaSt}(\hat{\mathbf{x}}) = \sum_{k=1}^{K}\sum_l \left(\|\mu_{k,l}(\hat{\mathbf{x}}) - \mu_{k,l}(\mathbf{x})\|_2 + \|\sigma_{k,l}(\hat{\mathbf{x}}) - \sigma_{k,l}(\mathbf{x})\|_2\right)/K$, where $\mu_{k,l}(\cdot)$ and $\sigma_{k,l}(\cdot)$ denote the mean and variance of features of l-th batch normalization layer for $M_i(\cdot, \boldsymbol{\Theta}i)$. In addition, to ensure $G(\cdot, \boldsymbol{\Theta}g)$ generates more diverse images, we propose $\mathcal{L}_{dis}$ to encourage disagreement between local models $\{M_i(\cdot, \boldsymbol{\Theta}i)\}_{i=1}^{K}$ and the global model $S(\cdot, \boldsymbol{\Theta}s)$, which is written as $\mathcal{L}_{dis}(\hat{\mathbf{x}}) = -\|E(\hat{\mathbf{x}}) - S(\hat{\mathbf{x}}, \boldsymbol{\Theta}s)\|_2$. To conclude, the total training objective of $G(\cdot, \boldsymbol{\Theta}g)$ is $\mathcal{L}_{gen}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \mathcal{L}_{sim}(E(\hat{\mathbf{x}}), \hat{\mathbf{y}}) + \lambda\mathcal{L}_{feat}(\hat{\mathbf{x}}) + \beta\mathcal{L}_{dis}(\hat{\mathbf{x}})$. We set λ to 0.5 and β to 0.1. Note that local and global models are fixed at this stage.

Third stage: knowledge distillation. We update the global model $S(\cdot, \boldsymbol{\Theta}s)$ by knowledge transfer. Specifically, the fixed generator first synthesizes a batch of images $\hat{\mathbf{x}}$ when feeding a batch of random noise vectors $\mathbf{z}$. Then $\hat{\mathbf{x}}$ are input into local models to get ensembled prediction $E(\hat{\mathbf{x}})$. We finally utilize a loss $\mathcal{L}_{kd}(E(\hat{\mathbf{x}}), S(\hat{\mathbf{x}}, \boldsymbol{\Theta}s)) = \|E(\hat{\mathbf{x}}) - S(\hat{\mathbf{x}}, \boldsymbol{\Theta}s)\|_2$ to enforce the similarity between $E(\hat{\mathbf{x}})$ and $S(\hat{\mathbf{x}}, \boldsymbol{\Theta}s)$.

3. Experiments, Results, and Conclusions

Dataset and metric. We applied the public dataset RNSA-BAA (Halabi et al., 2019), which contains 12,611/1,425/200 hand radiographs for training/validation/testing. We reported the mean absolute difference (MAD) results on the test set based on three runs.

Experimental setup. We maintained four local clients. We divided the training set into four subsets with bone age values falling within four ranges, as shown in Figure 1(a). To simulate an IID setting among clients, we ensured that each client received a similar number of images within the same bone age range by randomly extracting 1/4 of the data from each subset without repetition and assigning them to individual clients, as shown in

Table 1: Quantitative avg (std) results. '-': not applicable.

method	MAD			
centralization	10.15 (0.46)			
	homo-IID	homo-non-IID	hetero-IID	hetero-non-IID
FedAvg	11.68 (0.48)	36.80 (1.23)	-	-
FedAvgOneShot	62.15 (2.75)	68.35 (2.75)	-	-
FedNoisyKD	116.41 (1.78)	116.46 (1.57)	117.15 (2.07)	117.59 (2.44)
FedOneShot	59.92 (3.31)	**46.55 (1.69)**	**55.87 (2.67)**	**46.29 (1.10)**
Ours	**42.65 (4.40)**	49.49 (5.45)	58.52 (30.01)	52.60 (5.98)

Figure 1(b). Conversely, to form a non-IID setting, we distributed one subset to one client, as illustrated in Figure 1(c). We also considered model homogeneity and heterogeneity. Thus, we introduced four different settings: (1) **homo-IID**: model homogeneity with IID. (2) **homo-non-IID**: model homogeneity with IID. (3) **hetero-IID**: model heterogeneity with IID. (4) **hetero-non-IID**: model heterogeneity with non-IID.

Baselines. We compared our scheme with `FedAvg` (McMahan et al., 2017) and its one-shot version `FedAvgOneShot` (averaging model weights after local training). We also implemented a scheme that used random noise images for KD, and we abbreviated it as `FLNoisyKD`. In addition, we realized `FedOneShot` (Guha et al., 2019) using a public dataset (Pietka et al., 2001) as the proxy dataset for KD.

Implementation details. When considering model homogeneity, all clients used ResNet34 (He et al., 2016). For model heterogeneity, we applied ResNet34, ResNet50, and two variants (WRN-16-10, WRN-40-6) based on Wide-Resnets (Zagoruyko and Komodakis, 2016). The global model was always adopted as ResNet34. We applied the Adam optimizer. Local models were trained for 200 epochs using a poly-learning rate with an initial value of 10^{-4}. We set a learning rate of 10^{-3} and the latent dimension of $\mathbf{z}$ to train the generator to 128. We set an initial learning rate of 10^{-4} to train the global model and decayed it to 10^{-6}. The generator and the global model were trained in a loop of 120 rounds, and at each round, we trained the generator for 40 epochs and the global model for 1 epoch. We set the batch size to 32. All images were resized to a size of 224×224 pixels.

Experiment results. As illustrated in Table 1, `Centralization` represents the upper-bound accuracy derived by centralized training. We can observe a noticeable accuracy gap between `FedAvg` with IID and non-IID, indicating that `FedAvg` is also sensitive to non-IID in regression, similar to classification (Hsu et al., 2019). Then let us focus on one-shot FL methods. Limiting by a single global round, `FedAvgOneShot` achieves much larger MAD values than `FedAvg`. `FedNoisyKD` uses random noise images for KD, leading to the worst performance. `FedOneShot`, which conducts KD with a public dataset, achieves overall the best results. Compared to `FedOneShot`, our method outperforms it under the setting of homo-IID and realizes competitive accuracy under the other three settings. This suggests that our approach has the potential to synthesize images comparable to authentic images for KD, eliminating the requirements for proxy datasets.

Conclusions. This paper made a first attempt to explore data-free one-shot FL in regression. Our method demonstrated its efficacy in this setting. Future work may investigate improving image generation and apply the proposed method to more regression tasks.

References

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

Safwan S Halabi, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Artem B Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, et al. The rsna pediatric bone age machine learning challenge. *Radiology*, 290(2):498–503, 2019.

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.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Emilio Luz-Ricca, Clare Elizabeth Heinbaugh, and Huajie Shao. Data-free one-shot federated learning under very high statistical heterogeneity. In *International Conference on Learning Representations*, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

Joao Mendes-Moreira, Carlos Soares, Alípio Mário Jorge, and Jorge Freire De Sousa. Ensemble approaches for regression: A survey. *Acm computing surveys (csur)*, 45(1):1–40, 2012.

Ewa Pietka, Arkadiusz Gertych, Sylwia Pospiech, Fei Cao, HK Huang, and Vicente Gilsanz. Computer-assisted bone age assessment: Image preprocessing and epiphyseal/metaphyseal roi extraction. *IEEE transactions on medical imaging*, 20(8):715–729, 2001.

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.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. Dense: Data-free one-shot federated learning. In *Advances in Neural Information Processing Systems*, 2022.