HHS Public Access

Author manuscript

Proc IEEE Int Symp Biomed Imaging. Author manuscript; available in PMC 2023 October 10.

Published in final edited form as:

Proc IEEE Int Symp Biomed Imaging. 2022 March ; 2022: . doi:10.1109/isbi52829.2022.9761404.

FEDSLD: FEDERATED LEARNING WITH SHARED LABEL DISTRIBUTION FOR MEDICAL IMAGE CLASSIFICATION

Jun Luo[⋆], **Shandong Wu**[⋆,†]

[⋆]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA

[†]Dept. of Radiology, Dept. of Biomedical Informatics and Dept. of Bioengineering University of Pittsburgh, Pittsburgh, PA, USA

Abstract

Federated learning (FL) enables collaboratively training a joint model for multiple medical centers, while keeping the data decentralized due to privacy concerns. However, federated optimizations often suffer from the heterogeneity of the data distribution across medical centers. In this work, we propose Federated Learning with Shared Label Distribution (`FedSLD`) for classification tasks, a method that adjusts the contribution of each data sample to the local objective during optimization via knowledge of clients' label distribution, mitigating the instability brought by data heterogeneity. We conduct extensive experiments on four publicly available image datasets with different types of non-IID data distributions. Our results show that `FedSLD` achieves better convergence performance than the compared leading FL optimization algorithms, increasing the test accuracy by up to 5.50 percentage points.

Index Terms—

Federated Learning; Prior distribution; Medical imaging; Classification

1. INTRODUCTION

Deep learning (DL) is well known for requiring a large amount of data for robust training of generalizable models. For DL in medical research [1, 2, 3], large datasets can be difficult to obtain since the data collected by medical centers and hospitals are often privacy-sensitive. Therefore, sharing of the raw data between institutions is usually constrained by the restrictions such as Health Insurance Portability and Accountability Act (HIPAA) in the United States, and General Data Protection Regulation (GDPR) in Europe.

The recent emergence of federated learning (FL) [4, 5, 6] has provided this issue with a feasible solution. FL is a distributed machine learning scenario where only the model weights are shared among the participating clients in the federation, while keeping the data decentralized. In medical research, by bringing different hospitals and medical centers into

Ethical approval was not required, as this study used previously collected and deidentified data (including medical imaging data) available in public repositories.

the federation, researchers can collaboratively train a model utilizing different datasets from siloed institutions besides their own [7, 8, 9].

However, the federated settings generate a new major challenge, namely the statistical data heterogeneity across different participating clients [6, 10, 11, 12, 13]. The data heterogeneity reflects that the data collected by different clients is not identically distributed (non-IID), which often appears in medical datasets from different sites, because of various reasons including different data acquisition protocols and different local demographics. Data heterogeneity may lead to significant increase in communication rounds of the federated training, and inferior performance of the distributed optimization of federated models in certain clients (e.g. medical institutes) [10], which can further cost their incentives to participate in the federation.

In this work, we propose a federated learning algorithm for classification tasks, *Federated Learning with Shared Label Distribution* (`FedSLD`), which aims to utilize information regarding the clients' label distribution, to estimate a general prior label distribution for the entire federation. We claim that `FedSLD` can mitigate instability of training caused by the statistical heterogeneity of cross-silo FL, such as for medical research. While the algorithm does not access the clients' data, we assume legitimate for the clients to share the number of samples in each class, which are often the case for cross-silo FL such as in medical applications. More specifically, our contribution in this work is two-fold:

- **i.** We propose a new FL algorithm for medical image classification: Federated Learning with Shared Label Distribution (`FedSLD`), for robust training with non-IID data.
- **ii.** We demonstrate that the proposed `FedSLD` achieves better performance than the leading FL algorithms by conducting extensive experiments on four publicly available datasets (including two benchmark datasets) under pathological non-IID and practical non-IID data partitions.

2. METHOD

Laws and restrictions in terms of the data privacy constrain the direct access to the raw data. Yet, there are other information regarding the dataset that can be shared in terms of the federated learning. For instance, `FedAvg` assumes knowledge of the number of samples in each client: after the aggregation step in `FedAvg`, the algorithm conducts a weighted average of the updated copies for the next round, and the weights used for the averaging, by default, are the normalized number of samples in each client.

In this work, we focus on the classification tasks and assume legitimate to gain knowledge of the label distribution of each client, namely the number of samples from every class. We compute an estimate of the prior label distribution for the entire federation using the gain knowledge on the label distributions. For FL in medical applications, the label distributions from different medical silos can often be drastically different due to the regional demographics. Knowledge of the clients' label distributions will help us better understand the non-IID data in the federation.

To formulate the process, let us consider a federation with non-IID data. For a given data sample (x, y), where x stand for the data and y represents the label, the probability that it appears in the dataset of client i's, $\mathcal{P}_i(x, y)$, does not necessarily equal to the probability of it to appear in the dataset of client j's, $\mathcal{P}_j(x, y)$. By Bayes' theorem, we have $\mathcal{P}_i(x|y)\mathcal{P}_i(y) \neq \mathcal{P}_j(x|y)\mathcal{P}_j(y)$. More often than not, especially in medical imaging domain, non-IID data implicitly implies that both the label-conditioned probabilities, $\mathcal{P}(x|y)$, and the marginal label distributions, $\mathcal{P}(y)$, are different for different clients. In this work, we focus on acquiring the information reflecting the marginal label distribution $\mathcal{P}_i(y)$ for each client ($i = 1, 2, \ldots, N$), to compute the estimate of the prior label distribution for the entire federation.

We define the estimate for the prior of class c for the federation, as the sum of the numbers of samples for class c in each client divided by the sum of the total number of samples in each client. This is shown in equation (1), where $\widetilde{\mathcal{P}}(y = c)$ is the estimate prior of class c, $n_{i,c}$ is the number of samples from class c on client i, n_i is the total number of samples on client i, and N is the number of clients.

$$\widetilde{\mathcal{P}}(y = c) = \frac{\sum_{i=1}^{N} n_{i,c}}{\sum_{i=1}^{N} n_i} \tag{1}$$

During local update of the current model on a client, given a batch of data $\{(x_k, y_k)\}_{k=1}^{B}$, where B is the batch size, we first compute the label distribution in this batch as in equation (2), where the p_b represents the label distribution, $[[\cdot]]$ means the indicator faction, with its value equal to 1 if the inner part is true, and 0 otherwise. In essence, Equation (2) computes the proportion of class c samples in the batch by normalizing the number of class c samples in this batch.

$$p_b(y = c) = \frac{\sum_{k=1}^{B} [\![y_k = c]\!]}{B} \tag{2}$$

$$\mathcal{L}_b\left(\{(x_k, y_k)\}_{k=1}^{B}\right) = -\sum_{k=1}^{B}\left(\frac{p_b(y = y_k)}{\widetilde{\mathcal{P}}(y = y_k)} \cdot \sum_{c=1}^{C} y_{k,c} \log(f_i(x_k))_c\right) \tag{3}$$

Algorithm 1

FedSLD.

Input: Initialized model parameter weights w^0, number of clients N, number of local epochs E, batch size B, is the batch size, learning rate η, number of rounds R.

1: $\forall i \in [N]$, $c \in [C]$ acquire $n_{i,c}$ client i's numbers of samples of each class c.

2: $\forall c \in [C]$, $\widetilde{\mathcal{P}}(y = c) = \frac{\sum_{i=1}^{N} n_{i,c}}{\sum_{i=1}^{N} n_i}$// compute estimated prior label distribution.

3: **for** $r \leftarrow 1, 2, \ldots, R$ **do**

4: $\forall i \in [N] w_i^r = w^{r-1}$ // broadcast model parameters.
5: **for** $i \leftarrow 1, 2, \ldots, N$ **in parallel do**
6: **for** $\{x_k, y_k\}_{k=1}^{B}$ **in all minibatches do**
7: $\forall c, p_b(y = c) \leftarrow \sum_{k=1}^{B} [\![y_k = c]\!] / B$
8: Compute loss $\mathscr{L}_b$ by Equation 3.
9: $w_i^r \leftarrow w_i^r - \eta \nabla_w \mathscr{L}_b$
10: **end for**
11: **end for**
12: $w^r = \sum_{i=1}^{N} \frac{n_i}{n} w_I^r$ // aggregate model updates
13: **end for**
14: **return** w^R

Then, we define the batch loss as a weighted cross-entropy loss, shown in Equation (3), where $\mathscr{L}_b$ means the batch loss, and f_i represents the copy of the model on client i. By doing this, we can enforce proportional contribution (to the local objective) of each class of the data, with respect to its share of the true underlying distribution across the federation. We follow the aggregation step in a typical FL algorithm, where we compute the weighted average of the updated models from all clients, with the weights being the number of samples in each client. A detailed algorithm is shown in Algorithm 1.

3. EXPERIMENTS AND RESULTS

In this section, we evaluate the performance of the proposed `FedSLD` through experiments on four publicly available datasets (including two benchmark datasets), and compare it with two leading FL algorithms, `FedAvg` [4], an algorithm that average the local updates of the global model, and `FedProx` [11], an algorithm that adds a proximal term on the local objective to enhance performance robustness on non-IID data. To evaluate the general performance of the algorithms, we compute the test accuracies and demonstrate the empirical convergence performance by plotting the training loss and test accuracy curves. In addition, we examine the fairness of the method following recent work [14]. More details on the metrics are in Section 3.1.

3.1. Experiments setup

Datasets.—We conduct experiments on two benchmark image datasets: MNIST and CIFAR10. We further evaluate the methods on two medical image datasets from the MedMNIST dataset collection [15], namely the OrganMNIST(axial) dataset (11-class dataset of liver tumor images) and the PathMNIST dataset (a 9-class dataset of colorectal cancer images). We partition each dataset into a training set and a test set and ensure that they share the same label distribution.

Two non-IID settings.—We set the number of clients to be 12 to mimic a cross-silo FL setting and partition each dataset according to two different non-IID settings: 1) a

pathological non-IID setting, where we follow [4] and assign each client with two random classes; 2) a practical non-IID setting, where we randomly partition each class into 12 shards (corresponding to a total of 12 clients): 10 shards of 1%, one shard of 10% and one shard of 80% images in this class. We randomly assign one shard from each class to each client, so that each client will possess images from all classes, with more images from some classes while less images from others. This non-IID setting is more similar to the real-world medical applications, since datasets held at medical centers often contain a variety of classes, and are usually imbalanced with different majority class due to the regional demographics.

Implementation details.—We use the classic four-layer CNN model with two 5×5 convolutional layers and two fully connected layers (hidden layer has 500 units). We use a batch size of 256, 5 local epochs, 0.01 as the learning rate. For the practical non-IID partition, we train the model for 80 rounds, and for the pathological non-IID setting, we train the model for 160 rounds. All experiments are run on an NVIDIA Tesla V100 GPU and implemented in PyTorch and PySyft.

Metrics.—We compute two types of test accuracies for each setting: 1) the *Best Mean Client Test Accuracy* (BMCTA), computed as the highest mean client test accuracy over all training rounds; 2) the *Best Test Accuracy* (BTA), computed as the highest test accuracy for the combined test set from each client over all training rounds. We also investigate the methods' convergence performance by plotting the training loss and test accuracy curves. In addition, we follow [14] and examine the fairness of the methods by using the Gaussian kernel density estimation on the client test accuracies. Higher density at higher accuracy reflects a better result.

3.2. Results

We summarize the numerical results in Table 1. Under the pathological non-IID setting, for MNIST and the two medical datasets, the proposed `FedSLD` has a better performance with the improvement on the test accuracy of up to 1.57%, and the kernel density estimations in Figure 2 show that `FedSLD` has slightly higher density which is more concentrated at a higher test accuracy. On CIFAR10, `FedSLD` reaches competitive performance with `FedAvg` and `FedProx`.

Under the practical non-IID setting, we can see that the proposed `FedSLD` outperforms the compared `FedAvg` and `FedProx` on every dataset, with the improvement of BMCTA ranging from 1.10% to 5.50%, and the improvement of BTA ranging from 0.18% to 2.41%. In addition, Figure 1 shows that `FedSLD` achieves better convergence behavior on MNIST and OrganMNIST (axial) datasets. The fairness plots reveal that `FedSLD` not only increases the overall performance with respect to the entire federation, but the variances of the client test accuracies are also reduced on MNIST and PathMNIST datasets, which implies a more fair training. On CIFAR10 and OrganMNIST (axial) datasets, we can see a clear decrease of the density at low accuracy and an increase on the density at high accuracy, which explains the improvement of the BMCTA.

4. CONCLUSION

In this work, we proposed a new FL algorithm for medical image classification: Federated Learning with Shared Label Distribution (`FedSLD`). `FedSLD` aims to mitigate the effect caused by non-IID data by leveraging the clients' label distribution. We conducted extensive experiments on four publicly available datasets with two types of non-IID setting, and demonstrated that `FedSLD` outperforms the compared leading FL algorithms, and encourages a more fair performance across all the participating clients.

ACKNOWLEDGMENTS

This work was supported in part by a National Institutes of Health (NIH) / National Cancer Institute (NCI) grant (1R01CA218405), a National Science Foundation (NSF) grant (CICI:SIVD:2115082), the grant 1R01EB032896 as part of the NSF/NIH Smart Health and Biomedical Research in the Era of Artificial Intelligence and Advanced Data Science Program, and a Pilot Research Project from the Scaling Grant of the Pitt Momentum Funds for the Pittsburgh Center for AI Innovation in Medical Imaging. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant number ACI-1548562. Specifically, it used the Bridges-2 system, which is supported by NSF award number ACI-1928147, at the Pittsburgh Supercomputing Center (PSC).

REFERENCES

[1]. Oh Yujin, Park Sangjoon, and Ye Jong Chul, "Deep learning covid-19 features on cxr using limited training data sets," IEEE transactions on medical imaging, vol. 39, no. 8, pp. 2688–2700, 2020. [PubMed: 32396075]

[2]. Lee June-Goo, Jun Sanghoon, Cho Young-Won, Lee Hyunna, Kim Guk Bae, Seo Joon Beom, and Kim Namkug, "Deep learning in medical imaging: general overview," Korean journal of radiology, vol. 18, no. 4, pp. 570–584, 2017. [PubMed: 28670152]

[3]. Rajpurkar Pranav, Irvin Jeremy, Zhu Kaylie, Yang Brandon, Mehta Hershel, Duan Tony, Ding Daisy, Bagul Aarti, Langlotz Curtis, Shpanskaya Katie, et al. , "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.

[4]. McMahan Brendan, Moore Eider, Ramage Daniel, Hampson Seth, and Aguera y Arcas Blaise, "Communication-efficient learning of deep networks from decentralized data," in Artificial intelligence and statistics. PMLR, 2017, pp. 1273–1282.

[5]. Kairouz Peter, McMahan H Brendan, Avent Brendan, Bellet Aurélien, Bennis Mehdi, Bhagoji Arjun Nitin, Bonawitz Kallista, Charles Zachary, Cormode Graham, Cummings Rachel, et al. , "Advances and open problems in federated learning," arXiv preprint arXiv:1912.04977, 2019.

[6]. Li Tian, Sahu Anit Kumar, Talwalkar Ameet, and Smith Virginia, "Federated learning: Challenges, methods, and future directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, 2020.

[7]. Wang Pochuan, Shen Chen, Roth Holger R, Yang Dong, Xu Daguang, Oda Masahiro, Misawa Kazunari, Chen Po-Ting, Liu Kao-Lang, Liao Wei-Chih, et al., "Automated pancreas segmentation using multi-institutional collaborative deep learning," in Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning, pp. 192–200. Springer, 2020.

[8]. Sarhan Mhd Hasan, Navab Nassir, Eslami Abouzar, and Albarqouni Shadi, "On the fairness of privacy-preserving representations in medical applications," in Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning, pp. 140–149. Springer, 2020.

[9]. Qayyum Adnan, Ahmad Kashif, Ahsan Muhammad Ahtazaz, Al-Fuqaha Ala, and Qadir Junaid, "Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge," arXiv preprint arXiv:2101.07511, 2021.

[10]. Sahu Anit Kumar, Li Tian, Sanjabi Maziar, Zaheer Manzil, Talwalkar Ameet, and Smith Virginia, "On the convergence of federated optimization in heterogeneous networks," arXiv preprint arXiv:1812.06127, vol. 3, pp. 3, 2018.

[11]. Li Tian, Sahu Anit Kumar, Zaheer Manzil, Sanjabi Maziar, Talwalkar Ameet, and Smith Virginia, "Federated optimization in heterogeneous networks," arXiv preprint arXiv:1812.06127, 2018.

[12]. Karimireddy Sai Praneeth, Kale Satyen, Mohri Mehryar, Reddi Sashank, Stich Sebastian, and Suresh Ananda Theertha, "Scaffold: Stochastic controlled averaging for federated learning," in International Conference on Machine Learning. PMLR, 2020, pp. 5132–5143.

[13]. Ghosh Avishek, Chung Jichan, Yin Dong, and Ramchandran Kannan, "An efficient framework for clustered federated learning," arXiv preprint arXiv:2006.04088, 2020.

[14]. Li Tian, Sanjabi Maziar, Beirami Ahmad, and Smith Virginia, "Fair resource allocation in federated learning," arXiv preprint arXiv:1905.10497, 2019.

[15]. Yang Jiancheng, Shi Rui, and Ni Bingbing, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," in 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, 2021, pp. 191–195.

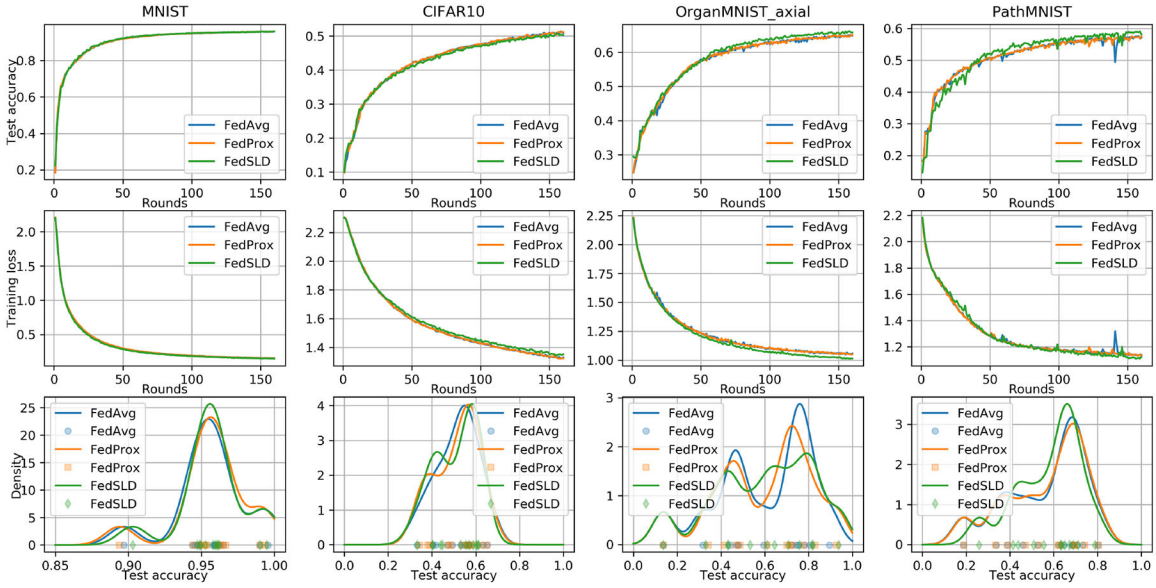

Fig. 1.

The convergence and fairness performance under the pathological non-IID setting. We measure the fairness using Gaussian kernel density estimation. Higher density concentrated at a higher accuracy reflects a better result.

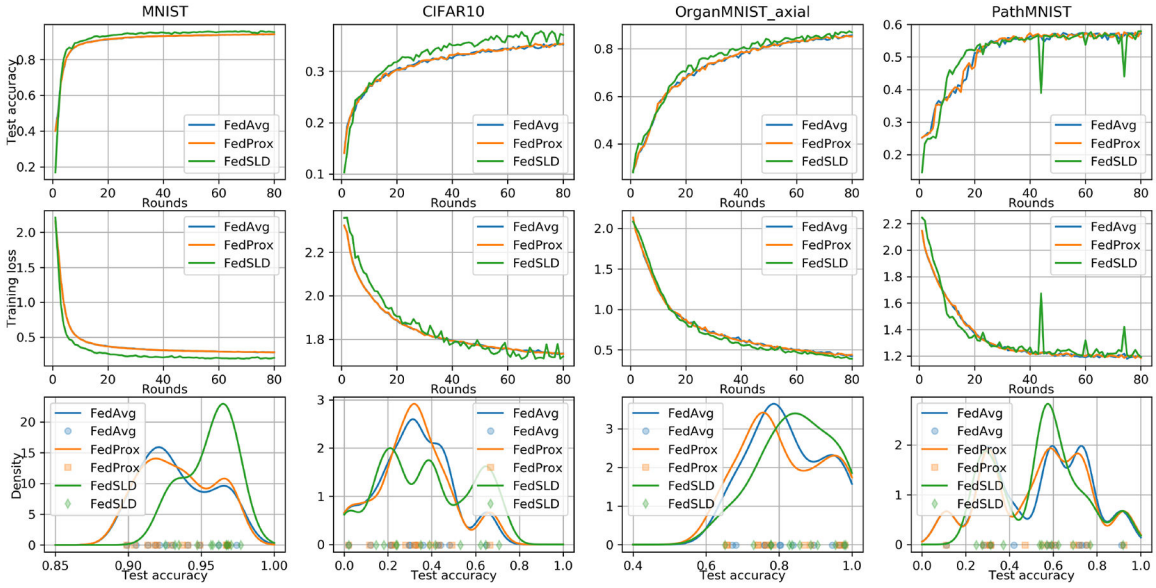

Fig. 2.

The convergence and fairness performance under the practical non-IID setting. We measure the fairness using Gaussian kernel density estimation. Higher density concentrated at a higher accuracy reflects a better result.

Table 1.

The *Best Mean Client Test Accuracy* (BMCTA) and *Best Test Accuracy* (BTA) for the pathological and practical non-IID settings. Highest performance are reported in bold.

	BMCTA/BTA under the pathological non-IID			BMCTA/BTA under the practical non-IID		
Dataset	**FedAvg** [4]	**FedProx** [11]	**FedSLD** (Ours)	**FedAvg** [4]	**FedProx** [11]	**FedSLD** (Ours)
MNIST	95.60 / 95.92	95.71 / 95.98	**95.74** / **96.03**	93.41 / 94.15	93.45 / 94.20	**95.56** / **95.85**
CIFAR10	**51.50** / **51.39**	51.39 / 51.24	50.81 / 50.71	32.07 / 35.46	31.98 / 35.38	**37.48** / **37.79**
OrganMNIST(axial)	59.52 / 64.99	59.44 / 65.10	**59.70** / **66.13**	82.32 / 85.69	81.53 / 85.54	**84.75** / **84.75**
PathMNIST	56.44 / 57.54	56.62 / 57.56	**57.94** / **59.11**	52.70 / 57.38	52.77 / 57.72	**53.87** / **57.90**