Is Normalization Indispensable for Multi-domain Federated Learning?

ABSTRACT
Federated learning (FL) enhances data privacy with collaborative in-situ training on decentralized clients. Nevertheless, FL encounters challenges due to non-independent and identically distributed (non-i.i.d) data, leading to potential performance degradation and hindered convergence. While prior studies predominantly addressed the issue of skewed label distribution, our research addresses a crucial yet frequently overlooked problem known as multi-domain FL. In this scenario, clients’ data originate from diverse domains with distinct feature distributions, as opposed to label distributions. To address the multi-domain problem in FL, we propose a novel method called Federated learning Without normalizations (FedWon). FedWon draws inspiration from the observation that batch normalization (BN) faces challenges in effectively modeling the statistics of multiple domains, while alternative normalizations possess their own limitations. In order to address these issues, FedWon eliminates all normalizations in FL and reparameterizes convolution layers with scaled weight standardization. Through comprehensive experimentation on four datasets and four models, our results demonstrate that FedWon surpasses both FedAvg and the current state-of-the-art method (FedBN) across all settings, achieving notable improvements of over 10% in certain domains. Furthermore, FedWon is versatile for both cross-silo and cross-device FL, exhibiting strong performance even with a batch size as small as 1, thereby catering to resource-constrained devices. Additionally, FedWon effectively tackles the challenge of skewed label distribution.

CCS CONCEPTS
• Computing methodologies → Distributed artificial intelligence. Distributed computing methodologies.

KEYWORDS
federated learning, multi-domain federated learning

1 INTRODUCTION
Federated learning (FL) has emerged as a promising method for distributed machine learning, enabling in-situ model training on decentralized client data. It has been widely adopted in diverse applications, such as healthcare [3] and autonomous cars [29]. However, FL commonly suffers from non-independent and identically distributed (non-i.i.d) data across clients [18]. This is due to the fact that data generated from different clients is highly likely to have different data distributions, which can cause performance degradation [9] even divergence in training [26, 31].

The majority of studies that address the problem of non-i.i.d data mainly focus on the issue of skewed label distribution, where clients have different label distributions [9, 19]. However, multi-domain FL, where data in clients are from different domains, has received little attention, despite its practicality in reality. Figure 1 depicts two practical examples of multi-domain FL. For example, autonomous cars may collaborate on model training, but their data could originate from different weather conditions or times of day, leading to domain discrepancies in collected images [28]. Similarly, multiple healthcare institutions collaborating on medical imaging analysis may face significant domain gaps due to variations in imaging machines and protocols [3]. Hence, developing solutions for multi-domain FL is a critical research problem with broad implications.

However, the existing solutions are unable to adequately address the problem of multi-domain FL. FedBN [20] attempts to solve this problem by keeping batch normalization (BN) [12] parameters and statistics locally in client, but it is only suitable for cross-silo FL.
1. We introduce FedWon, a simple yet effective method for multi-domain FL by removing normalizations and employing scaled weight standardization, FedWon learns a general global model from clients with significant domain gaps.

2. To the best of our knowledge, FedWon is the first method that enables both cross-silo and cross-device FL without relying on any form of normalization. Our study also reveals the unexplored benefits of this method, particularly in the context of multi-domain FL.

3. Extensive experiments demonstrate that FedWon outperforms state-of-the-art methods on all datasets and models, and is suitable for training with small batch sizes, which is especially beneficial for cross-device FL.

2 METHOD

This section presents the problem setup of multi-domain FL and introduces FL without normalization to address the problem.

2.1 Problem Setup

Assume there are $N$ clients in FL and each client $k$ contains $n_k \in \mathbb{N}$ data samples $\{(x_{i,k}, y_{i,k})\}_{i=1}^{n_k}$. Skewed label distribution refers to the scenario where data in clients have different label distributions, i.e. the marginal distributions $P_k(y)$ may differ across clients ($P_k(y) \neq P_{k'}(y)$ for different clients $k$ and $k'$). In contrast, this work focuses on multi-domain FL where clients possess data from various domains, and data samples within a client belong to the same domain [13, 20]. Specifically, the marginal distribution $P_k(x)$ may vary across clients ($P_k(x) \neq P_{k'}(x)$ for different clients $k$ and $k'$), while the marginal distribution of data samples within a client is the same, i.e., $P_k(x_i) \sim P_k(x_i)$ for all $i, j \in 1, 2, ..., n_k$. Figure 1 illustrates a practical example of multi-domain FL. For example, autonomous cars in different locations could capture images under different weather conditions.

2.2 Normalization-Free Federated Learning

Figure 2a demonstrates that the BN statistics of clients with data from distinct domains are considerably dissimilar in multi-domain FL. Although various existing approaches have attempted to address this challenge by manipulating or replacing the BN layer with other normalization layers [5, 20, 30], they come with their own set of limitations, such as additional computation cost during inference. Unlike all the existing approaches, we instead propose a novel approach called Federated learning Without normalizations (FedWon), which removes all normalization layers in FL.

Compared with FedAvg [21], FedWon completely removes normalization layers in DNNs and reparameterizes convolution layers. We employ the Scaled Weight Standardization technique proposed by [4] to reparameterize the convolution layers after removing BN. The reparameterization formula can be expressed as follows:

$$\hat{W}_{i,j} = y \frac{W_{i,j} - \mu_j}{\sigma_j \sqrt{N}}, \quad (1)$$

where $W_{i,j}$ is the weight matrix of a convolution layer, $y$ is a constant number, $N$ is the fan-in of convolution layer, $\mu_j = (1/N) \sum_j W_{i,j}$ and $\sigma_j^2 = (1/N) \sum_j (W_{i,j} - \mu_j)$ are the mean and variance of the $i$-th row of $W_{i,j}$, respectively. This weight standardization technique is
closely linked to Centered Weight Normalization [10]. By removing normalization layers, FedWon eliminates batch dependency, resolves discrepancies between training and inference, and does not require computation for normalization statistics in inference. We refer to this newly parameterized convolution as WSConv.

Figure 3 highlights the algorithmic differences between our proposed FedWon and the other two FL algorithms: FedAvg [21] and FedBN [20]. FedAvg aggregates both convolution and BN layers on the server; FedBN only aggregates the convolution layers and keeps BN layers locally in clients. Unlike these two methods, FedWon removes BN layers, replaces convolution layers with WSConv, and only aggregates these reparameterized convolution layers. Prior work theoretically shows that BN slows down and biases the FL convergence [26]. FedWon circumvents these issues by removing BN and offers unexplored benefits to multi-domain FL. These benefits include versatility for both cross-silo and cross-device FL, as well as compelling performance on batch sizes as small as 1.

3 EXPERIMENTS

3.1 Experiment Setup

Datasets. We run experiments for multi-domain FL using three datasets: Digits-Five [20], Office-Caltech-10 [7], and DomainNet [24]. Digits-Five consists of five sets of 28x28 digit images, including MNIST [17], SVHN [22], USPS [11], SynthDigits [6], MNIST-M [6]; each digit dataset represents a domain. Office-Caltech-10 consists of real-world object images from four domains: Amazon, Caltech, DSLR, and Webcam. DomainNet [24] contains 244x244 object images in six domains: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. We follow [20] to preprocess these datasets. Besides, we evenly split samples of Digits-Five into 20 clients for cross-device FL with total 100 clients. To simulate multi-domain FL, we construct a client to contain images from a single domain.

Additionally, we simulate skewed label distribution using CIFAR-10 dataset [15]. We split training samples into 100 clients and construct i.i.d data and three levels of label skewness using Dirichlet process Dir(α) with α = {0.1, 0.5, 1}.

Implementation Details. We implement FedWon using PyTorch [23] and evaluate the algorithms with three architectures in multi-domain FL: 6-layer convolution neural network (CNN) [20] for

Table 1 presents a comprehensive comparison of the aforementioned methods under cross-silo FL on Digits-Five, Office-Caltech-10, and DomainNet datasets. Our proposed FedWon outperforms the state-of-the-art methods on most of the domains. Specifically, FedProx has similar performance as FedAvg and both methods may exhibit inferior performance compared to Standalone in certain domains on DomainNet dataset. SiloBN and FixBN perform similarly to FedAvg in terms of average testing accuracy. However, they tend to underperform FedBN in multi-domain FL, where FedBN is specifically designed to excel. Interestingly, we discover that simply replacing BN with GN (FedAvg+GN) could boost the performance of FedAvg in multi-domain FL. Furthermore, our proposed FedWon surpasses both FedAvg+GN and FedBN in terms of the average accuracy. Although FedWon fails slightly behind FedBN by less than 1% in one domain on Digits-Five dataset, it outperforms FedBN by more than 10% on certain domains. These results demonstrate the effectiveness of FedWon under the cross-silo FL scenario. We report the mean of these methods across three runs of experiments.

Effectiveness on Small Batch Size. Table 2 compares the performance of FedWon with state-of-the-art methods using small batch sizes B = (1, 2) on Office-Caltech-10 dataset. FedWon achieves outstanding performance, with competitive results even at a batch size of 1. Although FedAvg+GN also achieves comparable results on
Table 1: Testing accuracy (%) comparison on Digits-Five, Office-Caltech-10, and DomainNet datasets. For Digits-Five, M, S, U, Syn, and M-M are abbreviations for MNIST, SVHN, USPS, SynthDigits, and MNIST-M. For Office-Caltech-10, A, C, D, and W are abbreviations for Amazon, Caltech, DSLR, and WebCam. For DomainNet, C, I, P, O, R, and S are abbreviations for Clipart, Infograph, Painting, Quickdraw, Real, and Sketch.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M  S  U  Syn  M-M</td>
<td>A  C  D  W</td>
<td>C  I  P  Q  R  S</td>
</tr>
<tr>
<td>Standalone</td>
<td>94.4 67.1 95.4 80.3 77.0</td>
<td>83.1</td>
<td>65.5</td>
</tr>
<tr>
<td>FedAvg</td>
<td>96.2 71.6 96.3 86.0 82.5</td>
<td>86.5</td>
<td>48.9 26.5 37.7 44.5 46.8 35.7</td>
</tr>
<tr>
<td>FedProx</td>
<td>96.4 71.0 96.1 85.9 83.1</td>
<td>86.5</td>
<td>51.1 24.1 37.3 46.1 45.5 37.5</td>
</tr>
<tr>
<td>FedAvg+GN</td>
<td>96.4 76.9 96.6 86.6 83.7</td>
<td>88.0</td>
<td>45.4 21.1 35.4 57.2 50.7 36.5</td>
</tr>
<tr>
<td>FedAvg+LN</td>
<td>96.4 75.2 96.4 85.6 82.2</td>
<td>87.1</td>
<td>42.7 23.6 35.3 46.0 43.9 28.9</td>
</tr>
<tr>
<td>SiloBN</td>
<td>96.2 71.3 96.0 86.0 83.1</td>
<td>86.5</td>
<td>51.8 25.0 36.4 45.9 47.7 38.0</td>
</tr>
<tr>
<td>FixBN</td>
<td>96.3 71.3 96.1 85.8 83.0</td>
<td>86.5</td>
<td>49.2 24.5 38.2 46.3 42.7 34.0</td>
</tr>
<tr>
<td>FedBN</td>
<td>96.5 77.3 96.9 86.8 84.6</td>
<td>88.4</td>
<td>49.9 28.1 40.4 69.0 55.2 38.2</td>
</tr>
</tbody>
</table>

FedWon (Ours) | 96.8 77.4 97.0 87.6 84.0 | 88.5 | 57.2 28.1 43.7 69.2 56.5 51.9 51.1 |

Table 2: Testing accuracy (%) comparison using small batch sizes $B = \{1, 2\}$ on Office-Caltech-10 dataset.

<table>
<thead>
<tr>
<th>B</th>
<th>Methods</th>
<th>Amazon</th>
<th>Caltech</th>
<th>DSLR</th>
<th>Webcam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FedAvg+GN</td>
<td>60.4</td>
<td>52.0</td>
<td>87.5</td>
<td>84.8</td>
</tr>
<tr>
<td></td>
<td>FedAvg+LN</td>
<td>55.7</td>
<td>43.1</td>
<td>84.4</td>
<td>88.1</td>
</tr>
<tr>
<td></td>
<td>FedWon</td>
<td>66.7</td>
<td>55.1</td>
<td>96.9</td>
<td>89.8</td>
</tr>
<tr>
<td>2</td>
<td>FedAvg</td>
<td>64.1</td>
<td>49.3</td>
<td>87.5</td>
<td>89.8</td>
</tr>
<tr>
<td></td>
<td>FedAvg+GN</td>
<td>63.5</td>
<td>52.0</td>
<td>81.3</td>
<td>84.8</td>
</tr>
<tr>
<td></td>
<td>FedAvg+LN</td>
<td>58.3</td>
<td>44.9</td>
<td>87.5</td>
<td>86.4</td>
</tr>
<tr>
<td></td>
<td>FixBN</td>
<td>66.2</td>
<td>50.7</td>
<td>87.5</td>
<td>88.1</td>
</tr>
<tr>
<td></td>
<td>SiloBN</td>
<td>61.5</td>
<td>47.1</td>
<td>87.5</td>
<td>86.4</td>
</tr>
<tr>
<td></td>
<td>FedBN</td>
<td>59.4</td>
<td>48.0</td>
<td>96.9</td>
<td>86.4</td>
</tr>
<tr>
<td></td>
<td>FedWon</td>
<td>66.2</td>
<td>54.7</td>
<td>93.8</td>
<td>89.8</td>
</tr>
</tbody>
</table>

Table 3: Evaluation on randomly selecting $C = \{10\% , 40\% \}$ out of total 100 clients to train each round with batch size $B = 4$, M, S, U, Syn, and M-M are abbreviations of five domains.

<table>
<thead>
<tr>
<th>C</th>
<th>Method</th>
<th>M  S  U  Syn  M-M</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>FedAvg</td>
<td>98.2</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>FedWon</td>
<td>98.6</td>
<td>85.4</td>
</tr>
<tr>
<td>40%</td>
<td>FedAvg</td>
<td>98.1</td>
<td>80.5</td>
</tr>
<tr>
<td></td>
<td>FedWon</td>
<td>98.8</td>
<td>86.4</td>
</tr>
</tbody>
</table>

Figure 4: Performance comparison of FedWon and FedAvg using small batch sizes $B = \{1, 2\}$ on Digits-Five dataset, where $C = 10\%$ out of 100 clients are randomly selected each round.

3.3 Experiments on Skewed Label Distribution

We run experiments on skewed label distribution with a fraction $C = 10\%$ randomly selected clients (i.e., $K = 10$) out of total 100 clients in each round. Table 5 compares our proposed FedWon with FedAvg, FedAvg+GN, FedAvg+LN, and FixBN. FedWon achieves similar performance as FedAvg and FixBN on the i.i.d setting, but outperforms all methods on different degrees of label skewness. We do not compare with FedBN and SiloBN as they are not suitable for cross-device FL. All experiments are run with local epoch $E = 5$ for 300 rounds. We use SGD as the optimizer and tune the learning in batch size $B = 1$, it requires additional computational cost during inference to calculate the running mean and variance. The capability of our method to perform well with small batch sizes is particularly important for cross-device FL, as some devices may only be capable of training with small batch sizes under constrained resources. We tune the best learning rates for methods in these experiments.

Impact of Selection a Subset of Clients. We assess the impact of randomly selecting a fraction of clients to participate in each training round, which is common in cross-device FL.
the range of [0.001, 0.1] for different algorithms. These experiments indicate the possibility of employing our proposed FedWon to solve the skewed label distribution problem.

4 CONCLUSION

In conclusion, we propose FedWon, a new method for multi-domain FL by removing all normalizations and reparameterizing convolution layers with weight scaled convolution. Extensive experiments across four datasets and models demonstrate that this simple yet effective method outperforms state-of-the-art methods in a wide range of settings. Notably, FedWon is versatile for both cross-silo and cross-device FL. Its ability to train on small batch sizes is particularly useful for resource-constrained devices.

REFERENCES


