FEDERATED LEARNING ON SMALL BATCH SIZES VIA BATCH RENORMALIZATION

Rubing Xue, Lukun Wang
Department of Computer Science
Shandong University of Science and Technology
{rubingxue, wanglukun}@sdust.edu.cn

Jiaming Pei
School of Computer Science
The University of Sydney
jpei0906@uni.sydney.edu.au

ABSTRACT

When batch size is too small in personalized federated learning, the parameters of the model change significantly due to data variability and outliers, leading to highly stochastic gradients between the layers of the model. This situation ultimately extends the convergence time and poses a significant challenge to the model training process. To address this issue, we propose a method that involves local batch renormalization before averaging the model. Experimental results demonstrate that this approach effectively improves the accuracy of personalized models trained with small batches in Non-IID scenarios.

1 INTRODUCTION

While extensive research has been conducted to enhance the accuracy of personalized models in personalized federated learning, such as the investigation of Batch Normalization Ioffe & Szegedy (2015) in Fedbn Li et al. (2021). For example, in the smartphone training scenario McMahan et al. (2017), we need to use small batchsize for training due to hardware limitations.However, when batch sizes are too small, the model's parameters exhibit substantial variations due to data variability, resulting in highly stochastic gradients across layers. Consequently, this leads to prolonged convergence times and poses a significant challenge for model training. Normalization is essential to mitigate the effects of data discrepancies. Normalization confines preprocessed data within a specified range, reducing the adverse effects stemming from sample data disparities. The introduction of Batch Normalization in 2015 has notably expedited convergence rates during model training. Subsequently, Batch Renormalization Ioffe (2017) has offered further improvements for mitigating the influence of small batch sizes on normalization effects. In this work, We introduce a small batch federated learning approach employing Batch Renormalization, demonstrating its empirical effectiveness in enhancing client model accuracy.

2 IMPLEMENTATION

Batch Renormalization is an extension of the network that adds a linear transformation to each BatchNorm layer to approximate the true distribution of the data.We consider the parameters r and d as fixed parameters. In the training phase, we start the batchnorm alone for a certain number of iterations by keeping $r = 1$ and $d = 0$, and then gradually change these parameters over a range. Suppose we have a Minibatch $B = \{x_1, x_2, \ldots, x_m\}$ and want to normalize a particular node x using either Minibatch statistics or their moving average data, then the results of these two normalizations are related by an affine transformation.Specifically, let μ be an estimate of the mean of x, and σ be an estimate of its standard deviation.Then, we have $\frac{x_i-\mu}{\sigma} = \frac{x_i-\mu_B}{\sigma_B}r + d$ where $r = \frac{\sigma_B}{\sigma}$ and $d = \frac{\mu_B-\mu}{\sigma}$ the equal sign holds ture. Therefore, we propose the FedReBN algorithm. In this algorithm, each client updates the parameters of each layer of the model using SGD as the server optimizer. The parameters of the non-Batch Renormalization layer are then weighted summed for each client model. These weighted summed parameters are distributed to each client for the next round of updates.

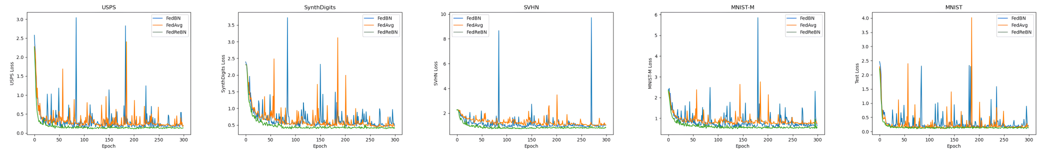

Figure 1: Convergence of the test loss of FedBN Li et al. (2021), FedAvg McMahan et al. (2017) and FedReBN on the digits classification datasets. FedReBN exhibits faster and more robust convergence.

Table 1: Test accuracy of different algorithms on the digits classification datasets

	FedReBN	**FedAvg**	**FedBN**	**FedProx**	**SiloBN**	**SingleSet**
	0.977	0.974	0.973	0.963	0.971	0.969
MNIST	0.986	0.982	0.979	0.973	0.982	0.973
	0.991	0.988	0.983	0.981	0.989	0.982
	0.846	0.826	0.835	0.829	0.831	0.832
MNIST-M	0.892	0.844	0.852	0.843	0.843	0.846
	0.972	0.917	0.911	0.891	0.887	0.875
	0.772	0.660	0.752	0.747	0.749	0.761
SVHN	0.815	0.721	0.787	0.763	0.769	0.788
	0.897	0.826	0.817	0.799	0.822	0.816
	0.889	0.875	0.879	0.877	0.871	0.873
SynthDigits	0.911	0.897	0.881	0.886	0.884	0.882
	0.945	0.935	0.931	0.939	0.936	0.933
	0.964	0.943	0.947	0.941	0.957	0.955
USPS	0.981	0.954	0.953	0.955	0.961	0.967
	0.993	0.967	0.963	0.961	0.972	0.973

3 Experiments and Results

In this section, we assess FedReBN's convergence in non-IID and small batch scenarios using a benchmark digits classification task across different data domains. We used five datasets whose non-iid properties are manifested in having a heterogeneous appearance but identical labels and distributions: SVHN Netzer et al. (2011), USPS Hull (1994), SynthDigits, MNIST-M Ganin & Lempitsky (2015), and MNIST LeCun et al. (1998). Our federated learning system has five clients, each with data exclusively from one dataset, representing a default non-IID configuration.In our experiment, we used the CNN model with SGD optimization at a learning rate of 0.01. The local batch size was 3, and we ran one local epoch (E=5). Clients had access to 10%, 40%, and 100% of the dataset, which corresponds to one, two, and three rows in Table 1. We evaluated our approach's performance in accuracy and loss for five clients in a personalized model, using FedAvg, FedBN, FedProx Li et al. (2020), SiloBN Andreux et al. (2020) and SingleSet as benchmarks. After 300 training rounds, FedReBN consistently outperformed the other four algorithms in test accuracy for all clients, as shown in Table 1. Figure 1 displays stable loss curves for our algorithm with b=3, demonstrating its effectiveness in addressing non-IID and small batch data challenges in federated learning.

4 Conclusion and Future Work

Our goal is to improve the performance of federated learning when training in small batchsize. We propose the use of Batch Renormalization, which our experiments show enhances the convergence of the model in non-IID and small batch environments. This study does not delve into privacy preservation issues. Future research will explore the impact of Byzantine attacks on our approach and the possibility of integration with homomorphic encryption.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Mathieu Andreux, Jean Ogier du Terrail, Constance Beguier, and Eric W Tramel. Siloed federated learning for multi-centric histopathology datasets. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*, pp. 129–139. Springer, 2020.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. *Advances in neural information processing systems*, 30, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

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

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.