Mitigating Heterogeneities in Federated Edge Learning with Resource-independence Aggregation

Zhao Yang
School of Computer Science
Northwestern Polytechnical University
Xi'an, China
yz70528@mail.nwpu.edu.cn

Qingshuang Sun
Faculty of Science and Bio-Engineering Sciences
Vrije Universiteit Brussel
Brussels, Belgium
qingshuang.sun@vub.be

***Abstract*—Heterogeneities have emerged as a critical challenge in Federated Learning (FL). In this paper, we identify the cause of FL performance degradation due to heterogeneous issues: the local communicated parameters have feature mismatches and feature representation range mismatches, resulting in ineffective global model generalization. To address it, Heterogeneous mitigating FL is proposed to improve the generalization of the global model with resource-independence aggregation. Instead of linking local model contributions to its occupied resources, we look for contributing parameters directly in each node's training results.**

I. Introduction

Federated learning (FL) [1] can effectively use distributed edge computing resources to train machine learning models. When using FL to practice, the need for generalization of the global model makes heterogeneities of edge devices urgent and distinct challenges. Previous methods for dealing with heterogeneities typically began with a resource-aware perspective. The disadvantages stem from that the global model may be overfitted to the devices with more resources. The optimization space expands dramatically when confronted with multiple heterogeneous issues at once. In addition, heterogeneities not only coexist but also collaborate to influence FL.

To address the challenges posed by resource-aware methods, we propose a resource-independence FL framework that does not compensate for the differences with considering heterogeneous resources. Heterogeneities between local devices bring feature mismatches and feature representation range mismatches. These mismatches impede the global model's generalization. Therefore, our method dynamically seeks contribution information directly from the different local training results to assist the global model in reducing mismatches in features and feature representation ranges between communicated parameters and obtaining more general features that will allow it to generalize to different devices.

II. The Scheme of Feature Representation Distinguishing

To reduce feature mismatches and feature representation ranges mismatches between communicated parameters, we design a scheme to distinguish feature representations in different models in a federated manner.

In order to realize the federated evaluation scheme and toward the generalization of the global model, we choose the general information in the global model as the evaluation criterion. To obtain such general information, we utilize Geometric Median (GM) [2]. The GM, $\arg\min_{y\in\mathbb{R}^n}\sum_{i=1}^{m}\|x_i-y\|_2$, is an estimate of the Euclidean space point center. Point y is where the sum of all Euclidean distances to the x_i is minimum.

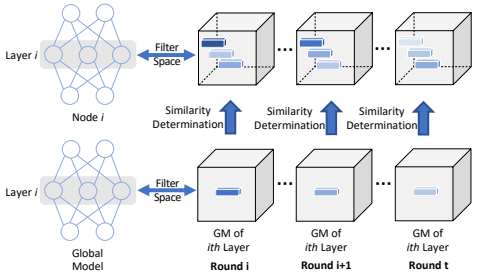

Fig. 1. GM's determination at each round of the training process.

Following that, we apply the GM to the neural network. The filter is the basic structure for feature representation in neural networks which can also be viewed as points in Euclidean space. Thus, by computing the GM, the 'center' of these filters can be regarded as their common feature in a specific layer:

$$\mathcal{F}_i^{GM} \in \underset{x\in\mathbb{R}^{N_i\times\mathcal{K}\times\mathcal{K}}}{\arg\min} \sum_{j\in[1,N_{i+1}]} \|x-\mathcal{F}_{i,j}\|_2\,, \tag{1}$$

where $\mathcal{F}_i^{GM}$ is the GM of the ith layer, N_i and N_{i+1} are the number of input and output channels in the ith layer, $\mathcal{K}$ is the kernel size, $\mathcal{F}_{i,j}$ is the jth filter in the ith layer.

After the GM for each layer in the global model is determined, the distance between each filter and the GM can be used to determine their feature representations:

$$s_{i,j}^k = \left\|\mathcal{F}_{i,j}^k - \mathcal{F}_i^{GM}\right\|_2, \tag{2}$$

where $s_{i,j}^k$ denotes the similarity between GM and the jth filter in the ith layer of the kth device. Such similarity can be viewed as the feature representation. If a filter is close to this GM, it is assumed that the features extracted by that filter are more similar to the common feature of the global model.

III. Generalization-oriented Dynamic Selection of Contributing Parameters

We propose a method for generalization-oriented parameter selection and aggregation in this section. First, we assess the

978-3-9819263-7-8/DATE23/© 2023 EDAA

communication costs. Each device's data transmission time is denoted as follows:

$$t_k^u = \frac{U_k}{C_k}, \quad C_k = B_k \log_2 \left(1 + \frac{T_k \left|h_k\right|^2}{\sigma_k^2}\right), \tag{3}$$

where U_k is the upload data size of the local device k, and C_k is the transmission rate from the local device k to the server $\mathcal{S}$. The transmit power is T_k, and the variance of the Additive Gaussian White Noise (AWGN) at the receiver is σ_k^2. The channel parameter is h_k experiences Rayleigh flat fading with the average channel gain of ϵ_k, and the communication bandwidth of the local device k is B_k.

Due to the service quality requirement of edge computing, a latency threshold γ_{th} should be set. The device k, in particular, selects some of the filters $\mathcal{F}_k$, allowing it to complete the corresponding parameter upload with a latency less than the latency threshold γ_{th}, i.e., $t_k^u(\mathcal{F}_k) \leq \gamma_{th}$.

After that, the filters with different feature representations should be selected with greater fairness to ensure the generalization of the global model. As a result, before selecting filters, we cluster them in each layer based on their distance from the GM. Clustering filters can ensure that filters with different features have the opportunity to be selected. Filters will be divided into groups using a distance threshold τ, and filters in the same group have more similar feature representations. We select the same ratio of filters across different groups for further filter selection.

Furthermore, with the communication latency constraint, we can construct the optimization problem to trade-off upload data size and parameter contribution, resulting in improved global model performance. For each device k:

$$\min \mathrm{Loss}(\mathcal{F}_k, \theta) + S(\mathcal{F}_k), \quad \text{s.t.} \quad t_k^u(\mathcal{F}_k) \leq \gamma_{th}, \tag{4}$$

where loss function maintains the model performance, $\mathcal{F}_k$ is the filter candidate set of device k with parameter θ, S is a sparse term to implement the filter selection according to GM, $t_k^u(\mathcal{F}_k)$ is the communication time of the corresponding parameters. The Alternating Direction Method of Multipliers (ADMM) is used to solve the optimization problem. With it, the objective function can be decomposed into easily solved sub-problems.

IV. Experiments and Results

The VGG-16 and ResNet-18 are chosen as the global model for training on the CIFAR-10 and CIFAR-100 datasets. We follow [3] to generate the Non-IID dataset. We construct an FL with ten devices chosen at random from the Raspberry Pi 4B, 3B+, and 3B. The training time per epoch on the Raspberry Pi 4B is set as the system requirement. We set the local epoch $E = 1$ in experiments. The communication bandwidth of each device will be assigned at random between 5 and 10 MHz. The device's transmit power is set to 0.5W, and σ is 10^4 cycle/sample. The average channel gain of the kth device to the server is set to $\epsilon_k = (k+50)/200$ when Rayleigh flat fading is present. $\gamma_{th} = 5s$ is the valid communication time. FedAvg [3], FedFA [4], FedFV [5], and FMore [6] are the baselines.

As shown in the Table I, our method has the highest accuracy. The parameters are scaled and aggregated based on the training results [5] and the amount of information in the local data [4], which causes the global model overfitting to some devices. The incompatibility of the global model with the local data causes fluctuations in the following training process and affects convergence. Furthermore, when confronted with a highly Non-IID dataset, partial device involvement [6] reduces the overlap between local data, which influences the extraction of general features and the performance of the global model. Ours identifies the adaptive representation relationships between local models and the global model and uses this as the basis for model aggregation with matched feature representation. Specifically, our method achieves up to 7.31% and 7.96% accuracy improvement on both datasets.

TABLE I
THE COMPARISON OF ACCURACY (%) WITH SOTA METHODS UNDER DIFFERENT EXPERIMENTAL SETTINGS.

Dataset	Method	#2	#4	IID
CIFAR-10	FedAvg	43.82	49.76	79.06
	FedFA	45.36	54.85	82.44
	FedFV $\vert_{\tau=1}$	45.81	53.15	81.68
	FedFV $\vert_{\tau=3}$	46.22	54.32	82.57
	FedFV $\vert_{\tau=10}$	47.92	56.64	83.99
	FMore $\vert_{K=5}$	43.48	52.25	84.37
	FMore $\vert_{K=25}$	42.27	52.09	82.61
	Ours	**49.58**	**58.38**	**84.19**
Dataset	Method	#20	#40	IID
CIFAR-100	FedAvg	46.52	52.39	81.46
	FedFA	48.91	57.86	84.79
	FedFV $\vert_{\tau=1}$	47.63	56.52	83.23
	FedFV $\vert_{\tau=3}$	48.54	57.29	84.97
	FedFV $\vert_{\tau=10}$	49.62	58.37	85.82
	FMore $\vert_{K=5}$	47.28	53.18	86.72
	FMore $\vert_{K=25}$	46.37	52.51	84.67
	Ours	**54.33**	**60.39**	**86.83**

V. Conclusion

In this paper, we propose an FL framework with resource-independence aggregation to mitigate heterogeneities. Leveraging feature representation distinguishing scheme in a federated manner and generalization-oriented dynamic parameter selection and aggregation method. Contributing parameters can be selected to reduce the feature mismatches and feature representation range mismatches between local communicated parameters and improve the global model's generalization.

References

[1] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, and et al., "Advances and open problems in federated learning," in *arXiv preprint arXiv:1912.04977*, 2019.

[2] P. T. Fletcher, S. Venkatasubramanian, and S. Joshi, "Robust statistics on riemannian manifolds via the geometric median," in *CVPR*, 2008, pp. 1–8.

[3] H. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[4] W. Huang, T. Li, D. Wang, S. Du, J. Zhang, and T. Huang, "Fairness and accuracy in horizontal federated learning," *Information Sciences*, vol. 589, pp. 170–185, 2022.

[5] Z. Wang, X. Fan, J. Qi, C. Wen, C. Wang, and R. Yu, "Federated learning with fair averaging," in *IJCAI*, 2021.

[6] R. Zeng, S. Zhang, J. Wang, and X. Chu., "Fmore: An incentive scheme of multi-dimensional auction for federated learning in mec," in *ICDCS*, 2020, pp. 278–288.