Depth-Wise Federated Learning Based on Model Parameter Reliability

Ryosuke Nutaba
Graduate School of FSE, Waseda University
Tokyo, Japan
nuta_ryo@akane.waseda.jp

Hiroshi Watanabe
Graduate School of FSE, Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—Federated Learning (FL) enables collaborative model training across multiple clients without sharing their local data. However, its performance degrades when data distributions are non-identical among clients. To address this challenge, we propose a depth-wise model aggregation method. This method uses Fisher Information Matrix (FIM) to evaluate parameter reliability and dynamically adjusts client contributions, reducing the influence of unreliable parameters. Experimental results demonstrate that our method consistently outperforms the existing approach in terms of local and global classification accuracy under various conditions.**

***Index Terms*—Depth-wise aggregation, fisher information matrix, personalized federated learning.**

I. Introduction

Federated Learning (FL) is a decentralized learning paradigm in which multiple clients collaboratively train a model without sharing their local data. This approach enables privacy preservation by aggregating only the model parameters trained locally on each client on a central server. However, due to the constraint of data locality, model training can be hindered when the data distribution across clients is non-identical. FedAS [1] addresses this issue by applying weighting based on the Fisher Information Matrix (FIM) [2]. However, the adverse effects of non-IID data distributions remain unresolved. To further mitigate this problem, we propose a depth-wise model aggregation method based on conventional methods [1] [3]. This approach selectively utilizes model parameters with lower FIM values only in deeper layers and reduces their influence on the shallow layers which are more sensitive to the heterogeneous data distribution.

II. Related Works

A. Personalized Federated Learning

Personalized FL (PFL) [4] is an extension of FL designed to address scenarios in which data distributions are heterogeneous between clients. In PFL, the model parameters of each client W_i are divided into a shared component θ_i and a personalized component w_i, with only the shared component being trained collaboratively. Here, $i \in \{1, \ldots, M\}$ denotes the index of each client and M represents the total number of clients.

B. FedAS

FedAS [1] proposes Client Synchronization (CS) to address the key challenges in PFL that the uneven progress of training among clients. CS assigns weights to the model parameters of each client during the aggregation step utilizing the FIM values. Although FedAS effectively filters out low-quality parameters through CS, it uniformly suppresses the contribution of parameters with low FIM values across all layers. This uniform suppression poses a problem in the shallow layers, which are particularly sensitive to differences in data distribution, as the influence of low-quality parameters remains significant.

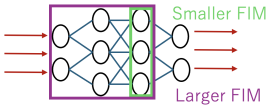

Fig. 1. Overview of the proposed method. The contribution is modified in proportion to the FIM values.

III. Proposed Method

In this study, we propose a depth-wise model aggregation method that leverages the reliability of model parameters collected from clients utilizing the FIM values, based on the conventional methods [1] [3]. This method enables dynamic adjustment of both client selection and contribution for each layer of the global model based on the FIM. In particular, it selectively incorporates model parameters with lower FIM values only into deeper layers, thereby minimizing their impact on shallow layers, which are more sensitive to data heterogeneity. The structure of the proposed model is illustrated in Fig 1. First, for each client i, we compute the trace of the FIM corresponding to its learned model parameters θ_i, denoted as α_i. Then, for the update of the parameter of the j-th layer θ_j of the global model θ, we select the top M_j clients with the highest α_i values:

$$M_j = \left\lceil \frac{j}{N_\theta} \cdot M \right\rceil, \tag{1}$$

where N_θ is the total number of layers in θ, and $j \in \{1, \ldots, N_\theta\}$. Next, the contribution of each client in the j th layer $\bar{\alpha}_{ij}$ is adjusted as follows:

$$\bar{\alpha}_{ij} = \frac{\alpha_i}{\sum_{k \in S_j} \alpha_k}, \quad i \in S_j, \tag{2}$$

979-8-3315-2415-9/25/$31.00 ©2025 IEEE

TABLE I COMPARISON OF LOCAL CLASSIFICATION ACCURACY (%)

Dataset	Method	$\beta = 0.1$			$\beta = 0.5$			$\beta = 1.0$		
		$P = 0.2$	$P = 0.6$	$P = 1.0$	$P = 0.2$	$P = 0.6$	$P = 1.0$	$P = 0.2$	$P = 0.6$	$P = 1.0$
CIFAR-10	FedAS	87.02	87.40	87.24	75.75	77.13	77.55	69.60	71.47	71.69
	Ours(Reverse)	86.98	87.18	**87.31**	75.82	76.88	77.34	**69.70**	70.60	71.28
	Ours	**87.25**	**87.53**	87.23	**75.90**	**77.22**	**77.91**	69.53	**71.74**	**71.94**
CIFAR-100	FedAS	54.19	56.69	57.54	36.20	38.38	**39.75**	29.07	32.20	32.40
	Ours(Reverse)	54.15	56.52	57.10	35.59	37.96	38.56	28.76	30.93	31.84
	Ours	**54.57**	**57.52**	**58.24**	**36.75**	**38.84**	39.25	**29.66**	**32.45**	**32.57**

TABLE II COMPARISON OF GLOBAL CLASSIFICATION ACCURACY (%)

Dataset	Method	$\beta = 0.1$			$\beta = 0.5$			$\beta = 1.0$		
		$P = 0.2$	$P = 0.6$	$P = 1.0$	$P = 0.2$	$P = 0.6$	$P = 1.0$	$P = 0.2$	$P = 0.6$	$P = 1.0$
CIFAR-10	FedAS	26.97	27.56	27.51	42.22	43.95	44.35	49.94	52.41	53.16
	Ours(Reverse)	27.10	27.39	27.45	42.05	43.06	44.15	49.89	51.33	52.53
	Ours	**27.14**	**27.60**	**27.54**	**42.62**	**44.25**	**44.65**	**50.03**	**52.75**	**53.58**
CIFAR-100	FedAS	8.49	9.36	9.48	14.52	16.19	**16.97**	16.63	18.75	19.31
	Ours(Reverse)	8.38	9.35	9.46	14.32	15.85	16.64	16.07	18.13	18.84
	Ours	**8.60**	**9.56**	**9.70**	**14.80**	**16.38**	16.75	**17.06**	**19.01**	**19.43**

where S_j denotes the set of M_j clients selected for the j-th layer. Finally, the global model parameter at the next time step, θ_j^{t+1}, are obtained as follows:

$$\theta_j^{t+1} = \theta_j^t + \sum_{i \in S_j} \bar{\alpha}_{ij} \cdot (\theta_{ij}^t - \theta_j^t), \tag{3}$$

where θ_{ij} is the parameter of the j-th layer of client i and t denotes the time step. This depth-wise weighting reduces the influence of less reliable parameters, particularly in the shallow layers which are sensitive to distributional differences.

IV. Experiment

The proposed method is evaluated on image classification tasks using the CIFAR-10 [5] and CIFAR-100 [5] datasets. For data partitioning, a Dirichlet distribution with parameter $\beta \in \{0.1, 0.5, 1.0\}$ is used. A lower value of β results in a greater bias in the data distribution. The model consists of a 4-layer CNN, where the first three layers are treated as the shared part and the final layer as the personalized part. The number of clients M is set to 20, the number of global epochs is 40, and the number of local epochs is 5. In addition, a parameter $P \in \{0.2, 0.6, 1.0\}$ is introduced to denote the proportion of clients participating in each update of the shared part. In this experiment, we employ two evaluation metrics: local accuracy and global accuracy. Local accuracy is defined as the average classification accuracy obtained by evaluating each client's model on its own local validation dataset. Global accuracy is defined as the average classification accuracy obtained by evaluating each client's model on the combined validation datasets of all clients. As comparative approaches, we employ FedAS and a reverse aggregation strategy of our method, denoted as "Ours (Reverse)", in which models exhibiting low FIM values are excluded from the output layers. By comparing with the reversed method, the impact of layer-wise depth selection on model performance can be effectively assessed.

The local accuracy and global accuracy, respectively, of the previous method, FedAS, and the proposed method on the classification tasks of CIFAR-10 and CIFAR-100 are shown in Tables I and II. These results indicate that the proposed method consistently outperforms the other methods in various settings. This indicates that even when limiting the contribution of models with low FIM values to certain layers, client information is still effectively incorporated into the shared model. From a layer-depth perspective, the observed decrease in both local and global accuracy indicates that excluding the parameters of models with low FIM values from updates in the shallow layers, rather than in the output layers, yields performance benefits.

V. Conclusion

In this study, we proposed a method that evaluates the reliability of the model parameters using the FIM and determines the contribution weights to the update of the global model on a depth-wise basis. Experimental results on classification tasks using the CIFAR-10 and CIFAR-100 datasets demonstrate that the proposed method improves both classification accuracy and generalization performance compared to the conventional approach. In future work, we plan to apply weighting based on the FIM values computed for each layer.

References

[1] X. Yang *et al.*, "FedAS: Bridging Inconsistency in Personalized Federated Learning," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. pp. 11986-11995.

[2] S. Amari, "Natural Gradient Works Efficiently in Learning," Neural Computation, vol. 10, no. 2, pp. 251-276. Feb. 1998.

[3] R. Nutaba and H. Watanabe, "Federated Learning Considering the Confidence Score of Model Parameters," Information Processing Society of Japan (IPSJ), National Convention, 2025. 5X-07. (in Japanese)

[4] M. G. Arivazhagan *et al.*, "Federated learning with personalization layers," arXiv preprint arXiv:1912.00818, 2019.

[5] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," Technical report, University of Toronto, 2009.