Supplementary Materials for MH-pFLGB

Anonymous Author

1 Baselines

In the medical image classification task (different resolution), we selected FedAvg, SCAFFOLD, FedProx, and their fine-tuned methods same as in previous works. Among the personalized Federated Learning methods, we compared FedRep, LG-FedAvg, APFL, and Ditto. For heterogeneous model federated learning, we chose FedMD, FedDF, pFedDF, DS-pFL, and KT-pFL. In the medical image classification task (different label distributions), we compared various methods, including local training of clients with heterogeneous models and existing heterogeneous model federated learning approaches (FedMD, FedDF, pFedDF, DS-pFL, and KT-pFL). The baseline used in the medical time-series classification task is the same as the medical image classification task (different label distributions). For image segmentation tasks, we compared various approaches, including local training of clients and a variety of personalized federated learning techniques, as well as methods for learning a single global model and their fine-tuned versions. Among the personalized methods, we also chose FedRep, LG-FedAvg, APFL, and Ditto. We simultaneously added LC-Fed and FedSM which are effective improvements for FedRep and APFL in the federated segmentation domain.

2 Training Settings

2.1 Evaluation Indicators

The performance evaluation of the classification task is based on two metrics, accuracy (ACC) and macro-averaged F1-score (MF1), providing a comprehensive assessment of the model's robustness. Additionally, Dice is used to evaluate the segmentation task performance across frameworks.

A. Accuracy. Accuracy is the ratio of the number of correct judgments to the total number of judgments.

B. Macro-averaged F1-score. First, calculate the F1-score for each recognition category, and then calculate the overall average value.

C. Dice. It is a set similarity metric commonly used to calculate the similarity between two samples, with a threshold of [0,1]. In medical images, it is often used for image segmentation, with the best segmentation result being 1 and the worst result being 0. The Dice coefficient calculation formula is as follows:

$$Dice = \frac{2 * (pred \cap true)}{pred \cup true} \tag{1}$$

Among them, $pred$ is the set of predicted values, and $true$ is the set of ground truth values. And the numerator is the intersection between $pred$ and $true$. Multiplying by 2 is due to the repeated calculation of common elements between $pred$ and $true$ in the denominator. The denominator is the union of $pred$ and $true$.

Table 1: The structure of the global bypass model in the medical image classification. In the global bypass body, the stride of each layer is 2. Class represents the category.

	Layer	Operation
global bypass Body	Conv2d 3x3-64	ReLU
	MaxPool 3x3	-
	Conv2d 5x5-64	ReLU
	MaxPool 3x3	-
	Conv2d 7x7-512	ReLU
global bypass Head	Linear-256	BatchNorm1d+ReLU
	Linear-class	-

Table 2: The structure of the global bypass model in the medical image segmentation. In the global bypass body, the stride of each conv2d layer is 1 and MaxPool is 2. In the global bypass head, the stride of each layer is 2. Class represents the category.

	Layer	Operation
global bypass Body	Conv2d 3x3-64	ReLU
	MaxPool 3x3	-
	Conv2d 5x5-64	ReLU
	MaxPool 3x3	-
	Conv2d 7x7-64	ReLU
	MaxPool 3x3	-
	Conv2d 7x7-512	ReLU
	MaxPool 3x3	-
global bypass Head	ConvTranspose2d 2x2-64	ReLU
	ConvTranspose2d 2x2-64	ReLU
	ConvTranspose2d 2x2-64	ReLU
	ConvTranspose2d 2x2-class	-

2.2 Loss Function

Many loss functions have been applied in this article, and here are some explanations for them. The cross-entropy loss function is very common and will not be explained in detail here. We mainly explain Dice loss.

Dice Loss applied in the field of image segmentation. It is represented as:

$$DiceLoss = 1 - \frac{2 * (pred \cap true)}{pred \cup true} \tag{2}$$

The Dice loss and Dice coefficient are the same thing, and their relationship is:

$$DiceLoss = 1 - Dice \tag{3}$$

2.3 Public Datasets for Other Federated Learning of Heterogeneous Models

In this section, we mainly describe the setting of public datasets for methods such as FedMD, FedDF, DS-pFL and KT-pFL.

A. Medical image classification (different resolution). We select 100 pieces of data from each client and put them into the central server as public data, totaling 400 pieces of data as public data. In order to better obtain soft predictions for individual clients, the image resolution of the publicly available dataset will be resized to the corresponding resolution for each client.

B. Medical image classification (different label distributions). For the breast cancer classification task, we select 50 pieces of data for each client to upload, and the public dataset contains 400 images. For the OCT disease classification task, We selected 1000 pieces of data from both the training and testing sets.

2.4 The Global Bypass Models

For medical image classification, the structure of the global bypass model is shown in Table 1. For medical image segmentation, the structure of the global bypass model is shown in Table 2.