KT-pFL: A Personalized Federated Learning Model Based on Knowledge Transfer

Chenrui Li

{ichenrui@ctbu.edu.cn}

School of Management Science and Engineering, Chongqing Technology and Business University, Chongqing, China

Abstract. To address the issues of data privacy protection and the difficulty of federated learning meeting personalized needs in heterogeneous (non-IID) scenarios, this paper proposes a personalized federated learning model based on knowledge transfer, KT-pFL, and completes its experimental reproduction and verification. This model uses the MNIST dataset to construct a strong non-IID experimental scenario. Through knowledge distillation (KD) and personalized parameter retention mechanism, it achieves a balance between global knowledge sharing and local feature adaptation. Comparative experiments with FedAvg show that the local fitting accuracy of KT-pFL reaches 0.9975, and the cross-distribution generalization accuracy is 0.0770. In both performance indicators, it outperforms the benchmark model. The research results validate the effectiveness of the "knowledge distillation + personalized retention" mechanism and provide practical references for federated learning applications in heterogeneous data scenarios.

Keywords: Personalized Federated Learning; Knowledge Distillation; Non-IID Data; Model Generalization KT-pFL

1 Introduction

In the era of big data, data-driven artificial intelligence models have achieved remarkable results in various fields. These models however often rely on data collected centrally, which has brought along grave privacy and security concerns. It leads to a paradigm shift in the conventional centralized training paradigm due to the introduction of rules and regulations like the General Data Protection Regulation (GDPR). In 2016, Google suggested federated learning, allowing multiple clients to cooperate in training models without data transfer to the domain, realizing the objective of data not leaving the domain, and literally meets the threat of data leakage [1]. The framework of distributed training has been extensively used in healthcare, financial, and smart cities becoming a central technology in terms of balancing data use and privacy assurance [2, 3].

Nonetheless, the classical federated learning algorithms with the form of FedAvg are manifestly not applicable in reality. These algorithms assume that the data of every client is of

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365160

an independent and identically distributed (IID) pattern, and compel all clients to arrive at one global model by averaging parameters [4]. But practically, because of the dissimilarity in client tools, user forms, and data collecting settings, data heterogeneity (non-IID) is pervasive [5]. An example of this is that medical data in one hospital might not be the same as in another related to the type of disease and the groups of patients and user behavior data in one region may not match in another region regarding the consumption habits [6]. In that case, a global model will be unable to adapt to the distinctive data nature of each client, and the lack of practical applicability of the local model data. This generic issue has become a major bottleneck that limits the emergence of federated learning [7].

Personalized Federated Learning (pFL) offers a viable solution to non-IID problem, which is to maximize client-specific models as well as attain a tradeoff between worldwide feature dissemination and locality feature appropriation [8]. The most common current pFL techniques are the local model adaptation techniques, individualized regularization techniques, and knowledge-transfer based methods, etc. Knowledge distillation (KD) has become one of these models that have received a lot of attention because of the property of knowledge transfer between models without the original information [9]. KT-pFL, as a pFL model based on knowledge transfer, builds a communication bridge between global and local models through the use of public datasets and knowledge distillation, and is expected to solve the personalized adaptation problem in strong non-IID scenarios [10].

In this context, this paper focuses on the validation of the effectiveness of KT-pFL in a strong non-IID environment. Firstly, a federated learning scenario with 10 clients is constructed, where each client only holds 2 classes of data from the MNIST dataset to simulate strong data heterogeneity. Then, using FedAvg as the benchmark, the performance differences between the two models in local fitting (with seen labels) and cross-distribution generalization (without seen labels) are compared. The core research objectives of this paper cover four aspects. Firstly, to reproduce the KT-pFL model and verify its feasibility in the strong non-IID scenarios; secondly, to evaluate the performance advantages of this model over traditional federated learning algorithms in terms of local fitting and generalization capabilities; thirdly, to analyze the effectiveness of the "knowledge distillation + personalized parameter retention" mechanism; fourthly, to summarize the application scope and limitations of KT-pFL, providing theoretical and practical support for the development of personalized federated learning.

2 Methods

2.1 Overall process

The KT-pFL model achieves personalized federated learning through a four-stage iterative process (Fig.1). Firstly, each client initializes the model using global parameters and trains it based on local data to fit the local label features. Subsequently, the client uses the trained local model to perform inference on the public dataset and generate soft labels containing global knowledge, which are probability distributions. Then, the server receives all the soft labels uploaded by the clients and performs a weighted average based on the client data volume to generate the global distillation target. Lastly the client corrects the local model by summing the local data loss (cross-entropy loss) with the global loss of distillation (KL divergence)

without losing 80 percent of the local parameter because this will cover personalized features. The four stages above form a single federated iteration and this is repeated until the number of iterations set is achieved.

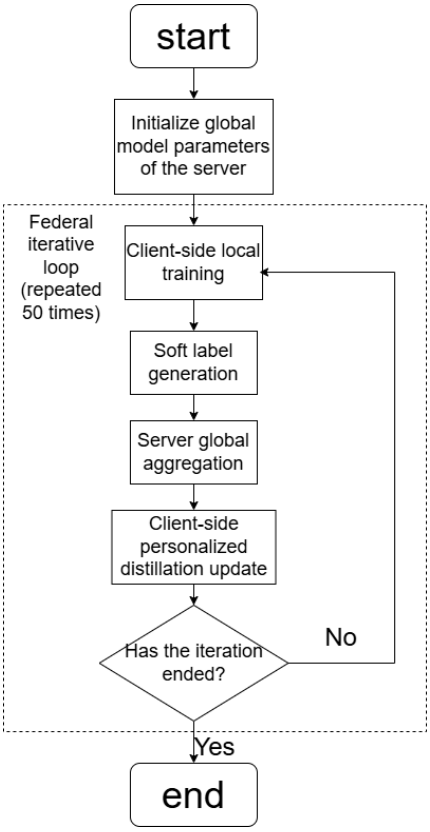

Fig. 1. Schematic diagram of the federated iterative process of the KT-pFL model

2.2. Data and processing

The experiment had been performed on the MNIST data to carry out the relevant verification. This dataset gives 60, 000 training samples as well as 10, 000 test samples, that is composed of 28 by 28 grayscale pictures of handwritten numbers between 0 and 9. In order to develop a powerful non-IID experimental situation, the data were exposed to specific splitting processing. There were 10 clients established and the data was allocated to the clients, who will have 2 consecutive labels. Client 0 was allotted 0-1, client one was allotted 2-3 and so on till client 9 had a label 8-9. This made certain that the data of the clients was significantly distributed. Also, the number of training samples in each client was 1, 000, and 500 samples per label. The imbalance in data was prevented, and the popular dataset that was taken in 500 samples was the test set of MNIST, and 50 samples were allocated to each digit. This was the data utilized in the knowledge distillation stage to transfer knowledge among the clients. Two sets of tests were created to test the performance of the model. One of them was a seen-label

test set in which the 200 test samples satisfying the 2 labels of each client were compared to assess the local fit of the fitness capability of the model. The other one was the unseen-label test set and in this analysis set, 200 test samples, representing the 8 labels that were not trained by each client were matched to provide an assessment of the cross-distribution generalization capacity of the model. All the data were brought to the range of [0, 1] and were then used to train, no further data augmentation operations were performed to guarantee the authenticity and reliability of the experimental outcomes.

2.3. Model

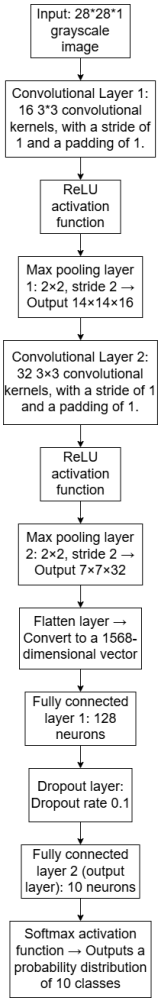

Fig. 2. CNN model structure lightweight schematic diagram

In the effort to match the CPU training environment, the present study considers the lightweight convolutional neural network (CNN) in order to develop the basic model (Fig.2). The model is divided into two core modules: the feature extractor and the classifier. The

feature extractor consists of two convolutional layers. The first convolutional layer has 16 3×3 convolutional kernels, with a stride and padding of 1. After the convolution operation, it is followed by a ReLU activation function and a 2×2 max pooling layer with a pooling stride of 2. The second convolutional layer has 32 3×3 convolutional kernels, with the same parameters as the first layer. After feature extraction, the two-dimensional features are flattened into a one-dimensional vector through the Flatten layer. The classifier module first connects a fully connected layer with 128 neurons. The input dimension of this layer is 32×7×7 = 1568. Then, a Dropout layer with a dropout rate of 0.1 is connected to alleviate the overfitting problem of the model. Finally, an output layer with 10 neurons is used to achieve classification. This output layer corresponds to the 10 digit categories of the MNIST dataset and uses the Softmax activation function to convert the output results into a probability distribution of categories.

2.4. Soft labels

The soft label is the probability distribution generated by the local model's inference on the public dataset. Compared to the hard label (one-hot encoding), it contains more global knowledge. For each sample in the public dataset, the local model outputs the original logit vector, which is normalized by the temperature-scaled Softmax function to generate the soft label. The calculation method is shown in Formula (1).

$$q_i = \frac{exp(z_i/T)}{\sum_{j=1}^{C} exp(z_j/T)} \tag{1}$$

Among them, q_i represents the soft label probability of the i-th class, z_i represents the original logit output of the model for the i-th class, C is the number of classes (in this experiment, C = 10), and T is the distillation temperature (in this experiment, it is set to 2.0). The higher the temperature, the smoother the soft labels, which is more conducive to the transmission of general knowledge.

2.5. Global knowledge aggregation

The server aggregates all the soft labels uploaded by the clients and generates the global distillation target. The aggregation adopts a weighted average strategy based on the data volume of each client, ensuring that clients with larger data volumes have a greater contribution to the global target. The aggregation process is shown in Formula (2).

$$Q_i = \frac{\sum_{k=1}^{K} n_k q_{k,i}}{\sum_{k=1}^{K} n_k} \tag{2}$$

Among them, Q_i represents the global soft label probability of the i-th class, K is the number of client devices (in this experiment, K = 10), n_k is the number of training samples for the k-th client device, and $q_{k,i}$ is the soft label probability of the i-th class generated by the k-th client device. This aggregation method effectively integrates the knowledge from multiple client devices to form a comprehensive global knowledge representation.

2.6. Distillation update

In the personalized distillation stage, the client model update adopts a dual loss function, combining the local data loss with the global distillation loss. The total loss function is shown in Formula (3).

$$L_{total}=(1-\alpha)L_{ce}+\alpha L_{KL} \tag{3}$$

Among them, α represents the weight for distillation loss (set to 0.1), L_{ce} is the cross-entropy loss between the model output and the local hard labels (used to evaluate the local fitting effect), and L_{KL} is the KL divergence between the model output (scaled by temperature T) and the global soft labels (used to evaluate the knowledge distillation effect).

When updating the parameters, a personalized retention mechanism is adopted. 80% of the parameters are retained from the local training model, while 20% of the parameters are updated through the gradients of the dual loss function. This mechanism ensures that the model can absorb global knowledge while retaining the personalized characteristics that are suitable for local data, achieving a balance between personalization and generalization ability.

3 Experiment

3.1. Hyperparameter Settings

The key hyperparameters for the experiment were configured based on the mainstream settings of federated learning experiments and were adjusted according to the characteristics of the model. The specific values are shown in Table 1.

Table 1. Key Hyperparameters for the Experiment Setup

Parameter Name	Value	Explanation
Federal iteration rounds	50	The interaction rounds between the server and the client
Local training epochs	5	Each client's local training round per cycle
Batch size	16	The batch size for local training
Learning rate	0.001	Optimizer learning rate
momentum	0.9	SGD optimizer momentum, accelerating convergence
Distillation temperature （T）	2.0	Soft label smoothness
Distillation loss weight （α）	0.1	The proportion of weight of distillation loss
Personalized retention coefficient（β）	0.8	The proportion of retained local parameters

The optimizer employs Stochastic Gradient Descent (SGD) and stops the training when the number of federated iterations reaches 50 rounds, without any additional early stopping

conditions. The hardware environment uses an Intel Core i9-10750H CPU (2.60GHz), paired with 16GB DDR4 memory, without an independent graphics card. In the software environment, the operating system is Windows 10 Professional Edition, and the programming language is Python 3.8. The model training is implemented based on the PyTorch 1.12.0 deep learning framework, data processing is accomplished using NumPy 1.21.6 and Pandas 1.4.4, and the experimental results are visualized with Matplotlib 3.5.3.

3.2. Experimental Results and Analysis

Performance comparison results. The experimental results are divided into two parts: local fitting (with labels) and cross-distribution generalization (without labels). FedAvg is used as the benchmark for comparison. The specific performance indicators are shown in Table 2 and Table 3.

Table 2. Performance Comparison Table of Seen Label Test Set

Model	Final accuracy rate	Training stability (fluctuation range)	Relative improvement rate
KT-pFL	0.9975	High (≤ 0.5%)	+0.30%
FedAvg	0.9945	Medium (≤ 1.0%)	-

Table 3. Performance Comparison of Unseen Label Test Set

Model	Final accuracy rate	Training convergence trend	Relative improvement rate
KT-pFL	0.0770	Gradually increasing (from 0.003 to 0.077)	+7.70%
FedAvg	0.0000	Maintain a zero level throughout the process.	-

Visual analysis. Figure 3 demonstrates the convergence of the two models on the test sets denoted and the unlabeled ones.

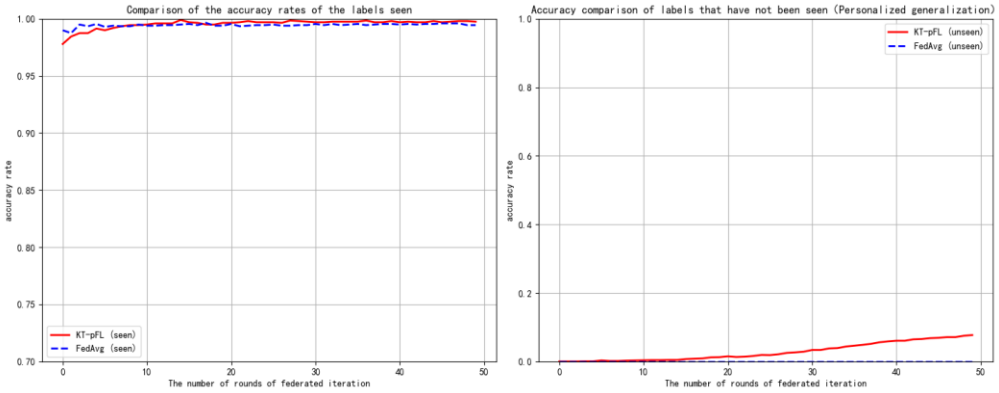

Fig. 3. training convergence curves of KT-pFL and FedAvg, in which (a) is with the observed labels of the test set and (b) is with the unobserved labels of the test set.

As Figure 3(a) demonstrates, the two models performed with high accuracy rates on the labeled test set. KT-pFL stood at 0.9975, FedAvg stood at 0.9945. KT-pFL was more stable during the training period and its highest fluctuation value was 0.5, compared to FedAvg with a range of 1.0. This can be attributed to the parameter retention mechanism of KT-pFL that is personalized and prevents the influence of global aggregation on local fitting.

The generalization performance between the two models has a large difference as indicated by Figure 3(b). The accuracy of FedAvg averaged 0 over the course of the process, meaning that the single world model had no ability to identify the unseen labels at all; the accuracy of KT-pFL improved steadily, with the range between 0.003 and 0.077 showing just the basic ability of cross-distribution generalization. This confirms that when transferring knowledge of global features (can be digital strokes, contours, etc.), knowledge distillation can work effectively to transfer such knowledge to other clients, allowing the model to perceive labels that were not visible to some degree.

Discussion of Results. Regarding the local fitting capability, both of the models reached an accuracy rate exceeding 99 percent on the labelled test set showing that even in the scenario with a high non-IID the federated learning model can be fully used to fit the local information. The personalized parameter retention mechanism, associated with slightly higher accuracy rate and improved stability of KT-pFL, is explained by the fact that the personalized process balances between the retention of the local features and the assimilation of the global knowledge.

Regarding the cross-distribution generalization capability, FedAvg 0 is accurate, which is the natural flaw of Federated learning algorithms in the non-IID setting that the unified global model can only identify the local characteristic of client data and has no capacity to identify factors. The 7.7 percent rate of accuracy of KT-pFL is not very high in absolute terms, but it has qualified improvement over FedAvg that knowledge distillation is a valid method to increase the generalization capacity of individualized federated learning.

Regarding the effectiveness of the mechanism, the experimental findings have established that the knowledge distillation and personalization parameter retention mechanism of KT-pFL is effective. Knowledge distillation can be used to share the knowledge globally by using the shared set of data and the personalized retention mechanism is used to guarantee the adaptability of the model to the local data, which is where the inherent paradox in the concept of personalization and generalization on non-IID settings can be resolved.

4 Conclusion

The case discussed in this paper touches upon the topic of personalized adaptation of federated learning in the strong non-IID case, where the KT-pFL model is replicated on the basis of knowledge transfer and a comparative experiment is carried out against the traditional FedAvg algorithm. The findings indicate that the local fitting of KT-pFL is 0.9975 and the cross-distribution generalization is 0.0770. It performs better in both areas of performance as compared to the benchmark model. This confirms that the mechanism of knowledge distillation plus the personalized parameter storage is applicable to balance the knowledge

sharing around the world and local feature adaptation, which is a feasible approach to personalized federated learning in heterogeneous data settings.

However, this study still has certain limitations: Firstly, the cross-distribution generalization accuracy of KT-pFL (7.7%) is still close to the random guessing level (10%), which is related to the significant difference in the characteristics of MNIST digits and the limited feature extraction ability of the lightweight CNN model; Secondly, the scale of the public dataset (500 samples) may not be sufficient to fully convey global knowledge, affecting the distillation effect; Finally, the model only considers label heterogeneity. In the future, its performance in other non-IID scenarios (such as feature heterogeneity, quantity heterogeneity) needs to be verified.

The future directions of optimizations consider four dimensions. On the one hand, the optimization of models implies using lighter CNNs in place of more sophisticated models like ResNet and ViT to improve feature extraction properties. Second, the process of distillation needs to be enhanced with a dynamically optimized balance between local fitting and generalization capabilities in terms of distillation temperature and loss weights. Thirdly, data augmentation is carried out to scale up the size of public datasets, make the features of various categories overlap, and enhance the efficiency of the transfer of knowledge. Fourthly, privacy is improved by providing the concept of differential privacy and adding noise to soft labels, along with the increased privacy standard of the model.

In practice, KT-pFL offers an innovative method of application in federated learning to utilize heterogeneous data conditions, i.e., cross-hospital medical diagnosis, and cross-regional environmental monitoring. The design of the model is lightweight and the model is less concerned with computation; thus, it suits resource-constrained edge devices. As the model is optimized and the application cases are increased, KT-pFL will be more relevant in the sphere of personalized federated learning.

References

[1] Wen, T., He, Q.: Comprehensively Promoting Rural Revitalization and Deepening Rural Financial Reform and Innovation: Logical Transformation, Breakthrough of Difficulties and Path Selection. China Rural Economy (1), 93–114 (2023)

[2] Zhang, H., Wang, Y., Zhang, Y.: Forward-looking Research on the National Digital Economy Development Plan during the “15th Five-Year Plan” Period. Journal of Xi’an University of Finance and Economics 38(5), 30–41 (2025)

[3] Mei, H., Du, X., Jin, H., et al.: Foresight of Big Data Technology. Big Data 9(1), 1–20 (2023)

[4] Wang, J., Kong, L., Huang, Z., et al.: A review of federated learning algorithms. Big Data 6(6), 64–82 (2020)

[5] Jin, B., Yang, P., Xing, H.: An exploration of the sharing and utilization of archival data in the era of big data. Information Science 41(6), 9–16 (2023)

[6] Jin, B., Yang, P.: Research on archival data governance in the era of big data. Archival Science Research (4), 29–37 (2020)

[7] Zhou, C., Sun, Y., Wang, D., et al.: A review of federated learning research. Journal of Network and Information Security 7(5), 77–92 (2021)

[8] Liu, T., Yang, C., Suo, S., et al.: Edge intelligence in wireless communication. Signal Processing 36(11), 1789–1803 (2020)
[9] Meng, X., Liu, F., Li, G., et al.: A review of knowledge distillation techniques in convolutional neural network compression. Computer Science and Exploration 15(10), 1812–1829 (2021)
[10] Sun, Y., Shi, Y., Li, M., et al.: Personalized Federated Learning Algorithm Based on Cooperative Game Theory and Knowledge Distillation. Journal of Electronics and Information Technology 45(10), 3702–3709 (2023).