Distributional Knowledge Transfer for Heterogeneous Federated Learning

1st Huan Wang
School of Cyber Engineering
Xidian University
Xi'an, China
wh_1018@stu.xidian.edu.cn

2nd, * Lijuan Wang
School of Cyber Engineering
Xidian University
Xi'an, China
ljwang@xidian.edu.cn

3rd Jun Shen
School of Computing & Information Technology
University of Wollongong
Wollongong, Australia
jshen@uow.edu.au

***Abstract*—Federated learning (FL) produces an effective global model by aggregating multiple client weights trained on their private data. However, it is common that the data are not independently and identically distributed (non-IID) across different clients, which greatly degrades the performance of the global model. We observe that existing FL approaches mostly ignore the distribution information of client-side private data. Actually, the distribution information is a kind of structured knowledge about the data itself, and it also represents the mutual clustering relations of data examples. In this work, we propose a novel approach, namely Federated Distribution Knowledge Transfer (FedDKT), that alleviates heterogeneous FL by extracting and transferring the distribution knowledge from diverse data. Specifically, the server learns a lightweight generator to generate data and broadcasts it to the sampled clients, FedDKT decouples the feature representations of the generated data and transfers the distribution knowledge to assist model training. In other words, we exploit the similarity and shared parts of the generated data and local private data to improve the generalization ability of the FL global model and promote representation learning. Further, we also propose the similarity measure and attention measure strategies, which implement FedDKT by capturing the correlations and key dependencies among data examples, respectively. The comprehensive experiments demonstrate that FedDKT significantly improves the performance and convergence rate of the FL global model, especially when the data are extremely non-IID. In addition, FedDKT is also effective when the data are identically distributed, which fully illustrates the generalization and effectiveness of the distribution knowledge.**

***Index Terms*—federated learning, non-IID, heterogeneous, distribution knowledge**

I. Introduction

Federated learning (FL) [1] aims at decentralized data usage and joint modeling through distributed machine learning paradigm under the premise of privacy protection. With the coordination of a central server, the private data of multiple decentralized clients (individual devices or organizations) jointly train the global model to meet the needs of different clients [2].

Unfortunately, the assumption that cross-device data examples meet the independent and identically distributed (IID) is not always applicable in respect of FL [3]. There are huge differences among clients, which are mainly reflected in the statistical heterogeneity [4], especially when the data on different clients are not independently and identically distributed (non-IID). On the one hand, dissimilar clients are different about the hardware [5], where they often need to satisfy specific requirements before joint training. On the other hand, quality or quantity variance of the data on separate clients results in data deviation from the uniform distribution [6].

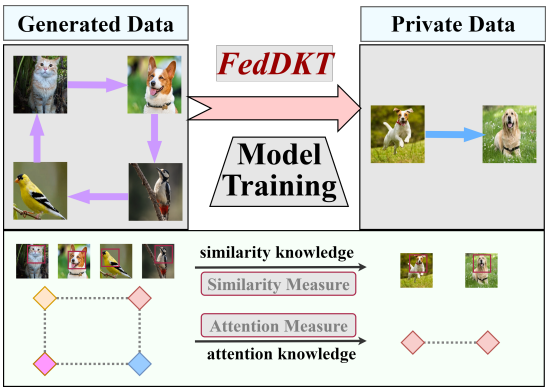

Fig. 1. The diagram of federated distribution knowledge transfer. FedDKT utilizes generated data to facilitate the learning of local private data samples, and transfers distributional knowledge through the similarity measure and attention measure.

In order to alleviate the heterogeneous FL issue, various methods such as algorithm optimization, personalized processing, and data enhancement have gradually attracted attention [7], [8]. Federated averaging (FedAvg) [1] is a typical FL algorithm using local updates and a low participation rate of clients. However, in practice, FedAvg is difficult to carry out theoretical analysis and convergence guarantees. Later on, the FedProx [9] algorithm is proposed, which adaptively adjusts the objective function by adding regular terms, but it may constrain the optimization freedom of the global model. Personalized FL [10] is proposed to optimize for different clients based on private data, but the performance often depends on the model initialization parameters and the data quality. Recent work has also proposed data augmentation [11] to alleviate heterogeneous FL, but it might introduce sensitive information and lead to privacy leakage. To the best of our knowledge, previous work only considered how to eliminate data heterogeneity, but did not consider the distribution information hidden in non-IID data, which could be regarded as a kind of structured knowledge on data itself.

* To whom correspondence should be addressed.

Inspired by the fact that data frequently fluctuates around a certain center, the data distribution can be reflected in the feature representations. Therefore, the data distribution information, as a type of structured knowledge, can be migrated from uniformly distributed data to non-IID private data, and as a result, we can make full use of the data distribution information to alleviate the data heterogeneity. We define distribution knowledge, which is a simple and effective sample relationship that represents clustering information among data samples. By the distribution knowledge, the generated data and client private samples can promote learning from each other. As shown in Fig. 1, the local private data is usually seriously non-IID, while the generated data has a more balanced class distribution. We transfer the distribution knowledge from the generated data to the local client model through the similarity measure and attention measure. Furthermore, the distribution knowledge uses the similarity and shared part relationship between data samples to enhance data and constrain model, which can be regarded as an implicit increase of local private data samples, and each sample feature can be considered as a virtual feature of other samples.

In this paper, we propose a novel approach for FL, dubbed Federated Distribution Knowledge Transfer (FedDKT), which generates data and obtains data distribution information (approximately regarded as IID) by generative adversarial network (GAN) [12], so as to transfer the distribution knowledge of the generated data to improve performance of the FL global model and accelerate convergence (shown in Fig. 2). The key innovation in FedDKT is that, the data distribution knowledge can be represented by the feature structure. Specifically, there are two main problems in the process of designing FedDKT: (1) 'what is the structured knowledge in data distribution?' and (2) 'how to extract and transfer the distribution knowledge in different data?'. Furthermore, the structured knowledge is considered as the feature representation of data distribution. Compared with non-IID, the IID data distribution knowledge has richer features and more sufficient information. In other words, the distribution knowledge represents the correspondence between data samples and labels, which is a priori information, and the distribution knowledge transfer can be regarded as the migration of interrelationship between different data samples. Therefore, FedDKT actually greatly enhances the data sample feature information and explores the relationship (e.g., similarity, local sharing) among data samples, so it can effectively improve the problem of data scarcity. In the case of serious imbalance of non-IID distribution, we believe that FedDKT balances the local private data. To solve the above problems and implement FedDKT specifically, we propose similarity measure and attention measure strategies to extract and transfer data distribution knowledge. From another perspective, traditional knowledge transfer has paid greater attention to the representation ability at the model level, while FedDKT can be regarded as the promotion of knowledge transfer at the data level. In FedDKT, the generated data with better distribution is regarded as a teacher, and the local private data is regarded as a student. The teacher improves the FL global model on the student by extracting and transferring the distribution knowledge. In a broad sense, FedDKT can be viewed as an extension of knowledge distillation, combined with other methods to improve model capability. FedDKT can be used for data augmentation in computer vision and for processing structured data in language semantic analysis.

Overall, our contributions are summarized as follows:

- We propose FedDKT, a novel data distribution knowledge transfer approach in FL, which decouples the distribution knowledge from better distributed data to guide model training. We find that FedDKT improves the model performance by taking advantage of the high-quality generated data, regarded as a data-level general component.
- We treat the data distribution as a new knowledge type and apply it to FL, which represents the mutual relations between data samples and labels. To the best of our knowledge, this is the first work to successfully design the similarity measure and attention measure strategies to assist model extracting and transferring the distribution knowledge from diverse data.
- We perform extensive experiments to establish the superiority of FedDKT, especially in extremely non-IID. We further show that the distribution knowledge really exists among different quality data, and FedDKT can effectively transfer this structured knowledge. In addition, we also find that learning the structured knowledge of high-quality data potentially enhances the generalization ability of the FL global model.

II. Related Work

Federated Learning: Systems and statistical heterogeneity remain as the main challenges of FL [13]. A common strategy to mitigate system heterogeneity was to ignore the constrained clients or simply merge the stragglers [14], but this tended to implicitly deepen the model heterogeneity. Furthermore, statistical heterogeneity among different clients makes it difficult to train a suitable global model and deteriorates the efficiency and accuracy [2]. Multi-task and personalizing applied on the FL settings had gradually become mainstream algorithms for processing heterogeneous data [15]. Another competitive method was to alleviate the heterogeneous FL issue through data augmentation [16]. Recent work proposed augmenting data to make data more similar across clients [17], or creating a small dataset that could be shared globally to solve the data shortage problem in a single client [18]. In addition, sampling and clustering algorithms also help to deal with data heterogeneity [19], but the sampling stage often also required extra communication.

Knowledge Distillation and Transfer: Knowledge distillation (KD) focuses on model efficiency, transferring knowledge from a complex model to a simple model without losing the effectiveness [20]. Recent work has shown that self-distillation and redesign of the network structure could improve performance [21]. In this regard, transfer learning is regarded as a broader knowledge distillation, which usually pays more attention to transferring the knowledge between related fields

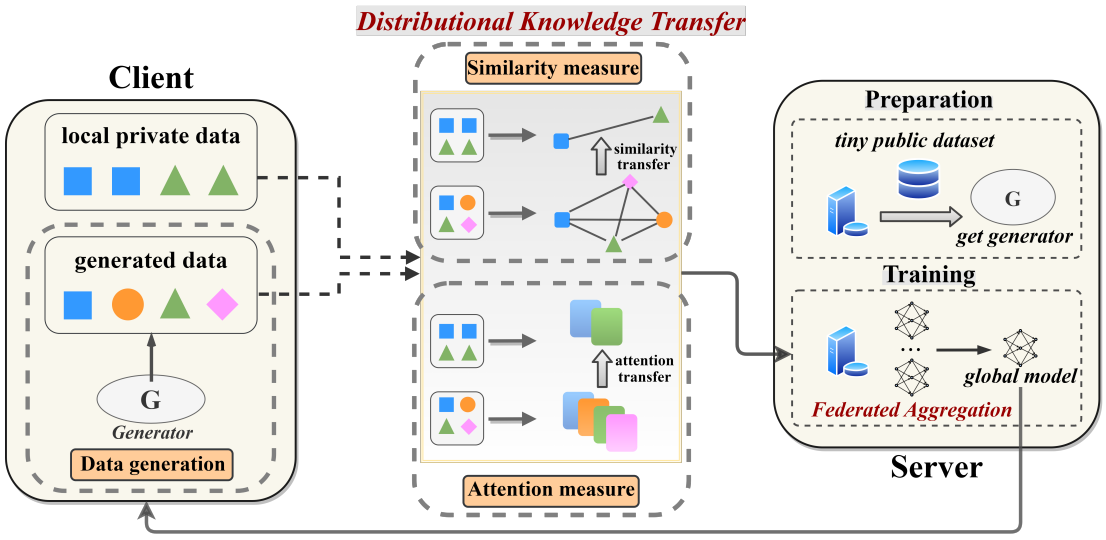

Fig. 2. Overview of FedDKT: a generator G is learned by the server on a tiny public dataset to generate data. Based on the similarity measure and attention measure strategies, FedDKT reduces the heterogeneous FL issue by transferring distribution knowledge from the generated data.

or tasks [22]. Notably, Song et al. [23] proposed an adaptive knowledge distillation framework that reduced recognition errors caused by data imbalance. Park et al. [24] proposed to improve the model performance by transferring mutual relations of data examples. While these methods used the structured knowledge smartly, transferred relational knowledge among models often generated additional parameters and computational overhead, thereby still limiting its scalability.

Non-IID Data Distribution: Non-IID data distribution seriously affected the performance of tasks in different domains. In FL, the non-IID property among cross-device data examples was the major cause of performance degradation [2], which made the degree of deviation between the local model and the global model continue to increase [25]. On the other hand, the long-tailed distribution could be considered as a generalized non-IID situation [26]. There was a vast amount of work to alleviate the long-tailed distribution, and these ideas had also recently been extended to FL. Feng et al. [27] presented an approach to alleviating the non-IID performance degradation through personalized adapters and group optimization strategies. Enhancing data diversity could also effectively alleviate the influence of non-IID [28]. However, the above methods simply processed data examples only, while the distribution information (the relations of samples) was totally ignored.

III. Our Approach: FedDKT

Our main challenge is how to extract and transfer the data distribution knowledge among diverse data with various distributions. Compared with non-IID data, the data with high-quality distribution tends to contain more category information, which actually reflects the mutual relations between data samples. Based on these motivations, we first generate excellent data (approximately regarded as IID) through GAN, and then extract and transfer the distribution knowledge of the generated data through similarity measure and attention measure strategies, thereby alleviating the heterogeneous FL issue. In other words, we expect that the global model would decouple category information from the feature representations of the generated data, which was not contained in the local private data. An overview of the FedDKT learning procedure is illustrated in Fig. 2.

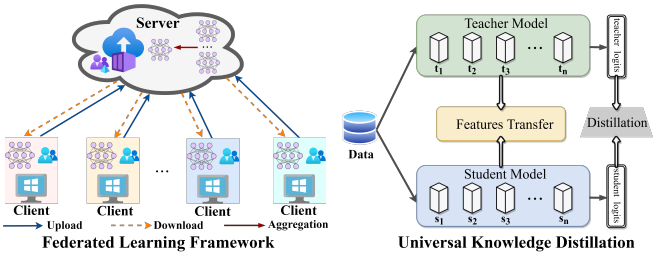

Fig. 3. Traditional federated learning framework (left) and knowledge distillation architecture (right).

In the following subsections, we first revisit the conventional form of FL and knowledge distillation. Then, we analyze the FedDKT algorithmic process and its three components: data generation, similarity measure, and attention measure.

A. Conventional Definition

Federated Learning. We consider a typical FL supervised setting for solving the general problem of multi-class classification tasks. Suppose $\mathcal{X} \subset \mathbb{R}^d$ is an instance space of the input data (with dimension d) and $\mathcal{Y} \subset \mathbb{R}$ is an output space. Given a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$, where x_i is the training data instance in $\mathcal{X}$ and has the corresponding category label $y_i \in \{1, \ldots, \mathcal{C}\}$, N is the total number of samples and $\mathcal{C}$ is the number of classes. We define the FL global model as S, and the local models for K clients as $F = \{f_k\}_{k=1}^K$. The dataset $\mathcal{D}$ is divided into K sub-datasets, for the k-th client, the local model is f_k, and local dataset is $\mathcal{D}_k$ where $|\mathcal{D}_k| = N_k$. In each round r, the server sends the initial

weight $\boldsymbol{\theta}_S^r$ and the local model f_k loads, while training based on the dataset $\mathcal{D}_k$ and updating the parameter $\boldsymbol{\theta}^k$. Then, the server collects the parameters, aggregates them on average $\boldsymbol{\theta}_S^{r+1} \leftarrow \frac{N^k}{N} \sum_{k=1}^{K} \boldsymbol{\theta}^k$, and broadcasts $\boldsymbol{\theta}_S^{r+1}$ in the next round $r+1$. Repeat the above process until reaching a convergence. The process of FL is summarized in Fig. 3 (left).

Knowledge Distillation. Traditional knowledge distillation guides student training by extracting complex teacher knowledge. According to the difference in knowledge definition, feature-based and relation-based knowledge distillation algorithms were proposed. Using the hidden layer features or the data relationships as knowledge to supervise the student is beneficial for training deeper and wider networks. Usually, the following objective function is minimized:

$$\mathcal{L}_{KD} = \sum_{x \in \mathcal{X}} \mathcal{L}\left(\Phi_T\left(f_T(x)\right), \Phi_S\left(f_S(x)\right)\right), \tag{1}$$

where x represents an instance of the input data $\mathcal{X}$, Φ_T and Φ_S represent the dimension transformation matrix of the teacher model f_T and the student model f_S to ensure the output dimensions consistent, respectively. Note that in the relation-based knowledge distillation, x is usually not a single sample, but represents a certain set. The general form of knowledge distillation is summarized in Fig. 3 (right).

B. Distribution Knowledge Transfer

1) Data Generation: FedDKT trains a lightweight generator based on GAN, generating data in a shorter time. The generator needs not follow any kind of factorization, which is important for the FL systems. Specifically, we have a tiny public dataset $\mathcal{D}_o$ from the original training dataset. We extract 2% of the data from each class in the training dataset to form the $\mathcal{D}_o$, and ensure that the generator can learn the information of each class. Owing to the limited amount of data, we use an adversarial training strategy when training the generator: the network generates new images based on the old generated images (project the old image to a low-dimensional vector and generate a new image), thereby further reducing the distribution difference between the generated data and the real data. We train a lightweight generator G on the server with data $x \in \mathcal{D}_o$ and a noise vector v, and the adversarial loss function is defined as:

$$\mathcal{L}(x, v, \gamma, \delta) = \ln\left(D(x, \delta)\right) + \ln\left(1 - D\left(G(v, \gamma), \delta\right)\right), \tag{2}$$

where γ and δ are the model parameters of G and D, respectively. Note that the input data can be real data x or old generated data $G(x, v)$. With the above settings, our trained generator G can generate diverse data.

The generator G is sent to the training clients through the server to generate similar distribution data. The generated data is regarded as a teacher, while the local private data is regarded as a student. Compared with the original data, the generated data is more diverse and contains more category information. Furthermore, the teacher transfers additional data distribution knowledge to the student. In addition, FedDKT does not require additional external data, which also reduces the risk of privacy leakage.

2) Similarity Measure: The purpose of similarity measure strategy is to calculate the correlations among data examples with various distributions, thereby extracting and transferring the distribution knowledge. For a client, the local private data is $\mathcal{D}_p$ and the local model is f, the generated data $\mathcal{Z}$ is based on generator G, and N is the total amount of data examples. Given the generated data $z \in \mathcal{Z}$ and the local data $x \in \mathcal{D}_p$, the features are denoted as $t = f(z)$ and $s = f(x)$, respectively. The objective function for similarity measure is expressed as:

$$\mathcal{L}_{sim} = \sum_{(z,x) \in (\mathcal{Z}, \mathcal{D}_p)} l_{sim}\left(\Psi_{sim}(t_1, .., t_N), \Psi_{sim}(s_1, .., s_N)\right), \tag{3}$$

where l_{sim} represents the similarity loss function to measure correlations, and Ψ_{sim} represents the similarity conversion.

The similarity conversion Ψ_{sim} is used to extract distribution knowledge from feature representations. In the feature space, the data distribution knowledge is not only related to the distance of data examples, but also to their relative positions. Specifically, the similarity conversion is achieved by calculating the euclidean and cosine distances between the generated data and the local data. Then the distribution knowledge is transferred by measuring the correlations among diverse data. The form of similarity conversion Ψ_{sim} is denoted as follows:

$$\begin{aligned} \Psi_{sim}(t_i, t_j) &= \sum_{(t_i, t_j) \in t} \frac{1}{d_{sim}} \left\| t_i - t_j \right\|_2 + \frac{t_i - t_j}{\left\| t_i - t_j \right\|_2} \\ \text{where} \quad d_{sim} &= \frac{1}{N} \sum \left\| t_i - t_j \right\|_2, \end{aligned} \tag{4}$$

where t_i and t_j are the feature representations of mutually exclusive samples in the generated data, respectively. The structured knowledge Ψ_t and Ψ_s of the generated data and the local data are extracted through the similarity conversion Ψ_{sim}, and Huber loss is used to measure the distribution knowledge, which is defined as

$$l_{sim}(\Psi_t, \Psi_s) = \begin{cases} \frac{1}{2}\left(\Psi_t - \Psi_s\right)^2 & \text{if } \left|\Psi_t - \Psi_s\right| \leq 1, \\ \left|\Psi_t - \Psi_s\right| - \frac{1}{2}, & \text{if } \left|\Psi_t - \Psi_s\right| > 1. \end{cases} \tag{5}$$

Through decoupling the correlation representations of the generated data, the similarity measure strategy allows the global model to learn the distribution knowledge from outside the local private data. As a result, the data distribution knowledge can be extracted and transferred from the generated data, effectively improving the performance and convergence rate of the FL global model.

3) Attention Measure: To further explore the key information between the generated data and the local data, we propose the attention measure strategy to reflect the data distribution knowledge. Attention measure consists of two parts: one is the attention representation, which represents the key information in the data distribution, and the other one is using a suitable loss function to transfer data distribution knowledge.

Given the local data feature representation s and the generated data feature representation t, we perform self-attention

extraction respectively, which is defined as:

$$\boldsymbol{z}_t = \sum \mathrm{Softmax}\left(\frac{t \cdot t^{\top}}{\sqrt{d}}\right) t, \tag{6}$$

where d represents the feature dimension of the generated data, $\boldsymbol{z}_t$ represents the attention representation of t, and $\top$ represents the conversion operation. The attention representations $\boldsymbol{z}_t$ and $\boldsymbol{z}_s$ are obtained based on t and s, respectively. In order to improve the efficiency, the MSE (mean square errors) loss is adopted, and the attention measure loss function is defined as:

$$\mathcal{L}_{\mathrm{attn}}\left(\boldsymbol{z}_t, \boldsymbol{z}_s\right) = \frac{1}{N} \sum_{\mathrm{i}=1}^{N} \left(\boldsymbol{z}_t - \boldsymbol{z}_s\right)^2, \tag{7}$$

where N is the total number of the generated data and the client local private data.

The attention measure strategy captures dependencies among data distribution from the attention representations, thereby dynamically generating weights for data and linking the structured knowledge of data distribution. Furthermore, the attention measure strategy can assign weights according to the data preference, which extracts the distribution knowledge more clearly.

Algorithm 1 Pseudocode of FedDKT Algorithm. ($\mathcal{D}$ is the training dataset, E is the local epochs, η is learning rate.)

RunServer(): *// Run on Server*
  initialize $K, T, w^0, \mathcal{D}, E, \eta$
  $\mathcal{D}_g \leftarrow$ (a tiny public dataset from $\mathcal{D}$)
  $G \leftarrow$ (a lightweight generator based on $\mathcal{D}_g$)
  for each round $t = 1, 2, ..., T$ **do**
    $A_t \leftarrow$ (server selects a random subset A of K clients)
    for each client $k \in A_t$ **in parallel do**
      $w_{t+1}^k \leftarrow \mathrm{RunClient}(k, w_t, G)$
    end for
    $w_{t+1} \leftarrow \frac{1}{A} \sum_{k \in A_t} w_{t+1}^k$
  end for
RunClient(k, w, G): *// Run on Client*
  $\mathcal{D}_k^o \leftarrow$ (local private data from client)
  $\mathcal{D}_k^g \leftarrow G\left(v\right)$ (generate data through G and the noise vector v) *// data generation*
  for each local epoch e from 1 to E **do**
    for batch (b_o, b_g) in the client dataset $(\mathcal{D}_k^o, \mathcal{D}_k^g)$ **do**
      $\ell = \ell_{sim(b_o, b_g)} \cdot \lambda_1 + \ell_{attn(b_o, b_g)} \cdot \lambda_2$ *// distribution knowledge transfer*
      $w \leftarrow w - \eta \nabla \ell(w; b_o, b_g)$
    end for
  end for
  return w to server

C. Summary and Expansion

FedDKT trains a lightweight generator on the server and broadcasts it to the training clients to generate data, which also provides the teacher information for subsequent data distribution knowledge transfer. Then, the similarity measure and attention measure extract and transfer distribution knowledge, thereby improving the FL global model performance. The pseudo code of FedDKT is described in Algorithm 1.

Previous knowledge transfer methods focused more at the model level, while FedDKT is dedicated to the data level. Therefore, FedDKT can also be regarded as a general data-level component used with other methods to improve the performance of the corresponding task, which is defined as

$$\mathcal{L}_{task} + \lambda_1 \cdot \mathcal{L}_{sim} + \lambda_2 \cdot \mathcal{L}_{attn}, \tag{8}$$

where $\mathcal{L}_{task}$ represents the target task loss function, λ_1 and λ_2 are hyperparameters used to adjust the importance of $\mathcal{L}_{sim}$ and $\mathcal{L}_{attn}$, respectively.

IV. Experiments

We compare the performance of FedDKT with the baseline FL algorithms including FedAvg [1], FedProx [9], FedBN [8], FedCurv [29], Scaffold [30], and FedNova [6]. Moreover, we also compare with distillation-based FL methods, including FedBE [31], FedDF [32] and FedGEN [33]. Note that we use the primal hyperparameter settings in FedProx without modification. Considering the influence of local and global model regularization constraint, we simply add regularization to the FedDKT, which is named FedDKT-R.

A. Experimental Setup

Dataset. For generality, we conduct experiments on four classical and benchmark image datasets (EMNIST [34], CIFAR-10/100 [35], mini-ImageNet [36]): EMNIST dataset consists of 47 different handwritten character and digit classes (image size 28), which is used to learn a classification task. CIFAR-10 and CIFAR-100 datasets (image size 32) are the collection of images commonly used to train computer vision classification tasks, and the classes of the datasets are completely mutually exclusive. CIFAR-10 has 10 classes and CIFAR-100 has 100 classes. The mini-ImageNet dataset (image size 224) contains 100 categories, each category includes 500 training images, 50 validation images and 50 test images.

Sampling. *1) Batch-IID:* The training dataset is randomly ordered and uniformly sampled based on the labels, and the same number of data samples are distributed to each client as its private data. Specifically, the training dataset $\mathcal{D}$ is evenly split into sub-datasets $\mathcal{D}^{k_{1:20}}$, and the sub-datasets conform to the IID distribution. *2) Batch-NonIID:* Based on the Dirichlet non-IID split method [32], the training dataset is divided into category imbalance. We control the instance distribution of each category to simulate a class imbalance scenario, thereby constructing the non-IID FL environment. Note that the number of each client is balanced. We set α to represent the non-IID ratio of client labels to all categories. Batch-IID can be seen as a special case when $\alpha = 1.0$.

Configurations. Unless otherwise mentioned, we run $R = 200$ global communication rounds, with $K = 20$ client models in total, an active-user ratio $r = 0.4$ and a non-IID imbalance ratio $\alpha = 0.2$. We adopt a local client updating epoch $E = 10$, and each step uses a mini batch with size $B = 64$. We use the total training dataset and distribute it to client models,

TABLE I
RESULTS (TOP-1 TEST ACCURACY [%]) FOR EMNIST, CIFAR-10, AND CIFAR-100 IMAGE CLASSIFICATION OF DIFFERENT FL METHODS.

Methods	EMNIST			CIFAR-10				CIFAR-100		
	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 1.0$	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.4$
FedAvg	80.62	84.96	86.23	72.95	82.43	85.79	88.78	52.07	56.07	58.36
FedProx	80.45	84.87	86.58	72.46	82.20	85.89	88.44	52.23	56.18	58.21
FedBN	80.17	84.66	86.68	73.22	82.98	85.95	88.69	52.69	56.30	58.60
FedCurv	80.61	84.34	86.83	72.12	82.46	85.54	88.75	51.90	55.93	58.79
Scaffold	80.08	84.72	85.94	72.88	82.85	86.09	88.13	52.27	56.13	58.67
FedNova	80.47	84.85	86.66	72.19	82.93	85.70	88.30	52.45	56.09	58.56
FedBE	81.34	85.14	87.42	74.14	83.77	86.35	89.17	54.15	57.22	59.80
FedDF	81.72	86.22	87.95	75.48	84.06	86.71	89.21	54.23	57.62	59.89
FedGEN	82.07	86.75	88.17	76.04	84.33	86.75	89.53	54.96	57.69	60.17
FedDKT	**88.81**	**90.27**	**91.82**	**80.92**	**87.35**	**89.12**	**90.14**	**58.78**	**62.21**	**65.38**
FedDKT-R	87.94	89.49	91.23	80.85	87.04	88.78	89.82	58.33	61.95	65.02

and use all testing dataset for performance evaluation. For the classifier, we use the ResNet-8 neural network on EMNIST, CIFAR-10 and CIFAR-100, and ResNet-18 [37] for mini-ImageNet as the backbone model. For training, we unify hyperparameter settings for all clients and use SGD with the initial learning rate 0.1 as the optimizer to update parameters.

B. Performance Overview

Results on EMNIST and CIFAR-10/100. The results on EMNIST, CIFAR-10 and CIFAR-100 with different FL heterogeneous settings are shown in Table I. We can observe that FedDKT universally outperforms other baseline FL algorithms (compared with the FedAvg, FedDKT performance gains reach 8.19% for EMNIST, 7.97% for CIFAR-10, and 6.71% for CIFAR-100), which implies that FedDKT can effectively utilize data distribution knowledge to improve performance of the FL global model. As the data heterogeneity increases (the α decreases), the performance of some baseline FL algorithms degrades, indicating that the local model gradually deviates from the global learning objective. However, FedDKT relatively stable, which means that it is robust to different levels of heterogeneous data while consistently performing well. Specifically, the baseline FL algorithms produce large volatility, while the accuracy of FedDKT are smoother, this is because it alleviates the heterogeneous FL issue, which also verifies the effectiveness of data distribution knowledge transfer under non-IID. This result also illustrates that FedDKT is in the correct optimization direction without deviation. FedDKT always achieves the optimal performance, which indicates the robustness of FedDKT on various datasets and data heterogeneities. Especially when the local data is extremely heterogeneous ($\alpha = 0.05$), the performance improvement of FedDKT is particularly obvious, which shows the advantages of FedDKT in extremely non-IID. The FL baseline algorithms are unable to achieve appropriate accuracy even with regularization constraint in the extreme heterogeneity, but the global model can be greatly improved by transferring distribution knowledge, allowing it to learn additional information from the generated data. We also note that the performance of FedDKT on EMNIST remains stable, which illustrates that FedDKT sufficiently transfers the distribution knowledge with little performance degradation even if $\alpha = 0.05$. In addition, in the Batch-IID case ($\alpha = 1.0$, CIFAR-10), FedDKT can also learn from the distribution knowledge of the generated data, which further demonstrates its effectiveness and universality.

TABLE II
TOP-1 TEST ACCURACY (%) ON MINI-IMAGENET.

Methods	mini-ImageNet					
	$r = 0.2$			$r = 0.4$		
	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.4$
FedAvg	42.53	47.27	50.45	40.44	46.36	48.18
FedProx	42.15	47.32	50.76	40.52	46.56	48.12
FedBN	42.11	47.13	50.66	40.43	46.18	48.56
FedCurv	42.22	47.20	50.26	40.09	46.84	48.54
Scaffold	42.32	46.94	50.39	40.54	46.35	48.40
FedNova	41.98	47.18	50.33	40.17	46.21	48.43
FedBE	44.66	48.85	51.91	42.59	48.52	49.76
FedDF	45.01	49.39	51.97	43.12	48.74	50.18
FedGEN	45.67	50.64	52.11	43.27	48.89	50.64
FedDKT	**48.02**	**52.34**	**53.89**	**48.62**	**52.93**	**54.13**
FedDKT-R	47.87	51.68	53.15	48.20	52.11	53.42

Results on mini-ImageNet. The results of FedDKT and other methods under various non-IID degrees and active-user ratios are shown in Table II. Even on complex dataset, FedDKT can still achieve excellent performance, with an accuracy improvement of at least 5.07% compared to the FedAvg. Due to the heterogeneous data, other methods are difficult to train the local model, but FedDKT effectively improves model performance through distribution knowledge transfer. When the number of participating users increases excessively, the performance of some FL algorithms tends to decrease, which exhibits that simply aggregating all client parameters may interfere with the global model optimization direction. However, FedDKT gradually improve the performance as the number of users increases, this is because more users can transfer richer distribution knowledge to correct the deviation caused by average parameters, similar to the ensemble learning, which improves performance by integrating the data distribution knowledge of multiple clients. Unfortunately, excessive users maybe cause time cost, and an active-user ratio $r = 0.4$ is usually a good choice.

TABLE III
COMPARISON OF CONVERGENCE RATES RESULTS ON CIFAR-10.

Methods	CIFAR-10				
	$R_{20\%}$	$R_{40\%}$	$R_{60\%}$	$R_{80\%}$	Accuracy
FedAvg	5	16	38	80	82.43
FedProx	4	16	36	77	82.20
FedBN	4	15	36	72	82.98
Scaffold	4	17	40	78	82.85
FedDF	4	14	33	70	84.06
FedGEN	3	14	31	68	84.33
FedDKT	**2**	**6**	**20**	**35**	**87.35**

TABLE IV
EFFECT OF HYPERPARAMETERS IN FEDDKT.

$\boldsymbol{\lambda_1}$ ($\lambda_2 = 1.0$)	0.5	1.0	2.0	4.0	6.0
CIFAR-10 Accuracy	86.43	**87.35**	85.39	84.72	83.89
$\boldsymbol{\lambda_2}$ ($\lambda_1 = 1.0$)	0.5	1.5	2.0	4.0	6.0
CIFAR-10 Accuracy	85.32	86.14	86.91	**88.04**	86.78

C. FedDKT Validation

In this subsection, we focus on exploring some of the key properties of FedDKT, including convergence rates and hyperparameters selection. Moreover, we also performed ablation experiments and visualization analysis for FedDKT. We conduct experiments on the CIFAR-10 dataset and uniformly set non-IID $\alpha = 0.2$, active-user $r = 0.4$.

Convergence rates. The comparison results of convergence rates on CIFAR-10 are given in Table III, where the R_{acc} columns show the minimum number of rounds required to reach acc of the test accuracy, and the Accuracy columns show the final test accuracy. Compared with other FL algorithms, the communication rounds required for FedDKT to achieve the corresponding accuracy are greatly reduced and it also obtains the optimal performance, which clearly exhibits that FedDKT has significant advantages in processing non-IID data, further exhibits the distribution knowledge transfer is generally effective. FedProx alleviates the deviation of optimization direction through the regularization constraint, and it converges slightly faster than FedAvg. In addition, FedBN alleviates data heterogeneity through local batch normalization and has a faster convergence rate than FedAvg. Meanwhile, we note that distillation-based FL methods also converge faster than FedAvg, because these methods effectively utilize global class knowledge at high data heterogeneity.

Hyperparameters selection. Table IV illustrates the effect of λ_1 and λ_2 in FedDKT loss function $\lambda_1 \cdot \mathcal{L}_{sim} + \lambda_2 \cdot \mathcal{L}_{attn}$. FedDKT is insensitive to smaller hyperparameters, but it truly degrades performance when λ_1 or λ_2 is large. In FedDKT, the similarity measure strategy captures the distribution knowledge through the correlations among data examples, while the attention measure strategy extracts more important knowledge by mining its own pivotal information. Therefore, compared with the similarity measure, the weight of the attention measure should be appropriately increased, that is, $\lambda_1 < \lambda_2$. When the hyperparameter combination of $\lambda_1 = 1.0, \lambda_2 = 4.0$ is used, FedDKT can often achieve excellent performance.

Ablation study. Table V exhibits the effectiveness and importance of FedDKT components on CIFAR-10. We use $\Delta_{Generation}$, $\Delta_{Similarity}$, $\Delta_{Attention}$ to represent data generation, similarity measure, attention measure respectively. We find that no matter whether the similarity measure or attention measure is utilized, the global model performance has always been improved. The similarity measure strategy narrows the relative positions among data examples, which can reduce the distribution differences across clients. More importantly, the attention measure strategy casts high weight on the crucial features in combination, which better reflects the data distribution knowledge. Ablation experiments verify the effectiveness of the FedDKT components, which can decouple and transfer the distribution knowledge of the generated data more efficiently and fully in a complementary way to achieve optimal performance. As mentioned above, FedDKT is not limited to FL, and it can be regarded as a generic approach.

TABLE V
ABLATION RESULTS ON CIFAR-10.

$\Delta_{Generation}$	$\Delta_{Similarity}$	$\Delta_{Attention}$	CIFAR-10
			82.43
✓	✓		84.37
✓		✓	85.50
✓	✓	✓	**87.35**

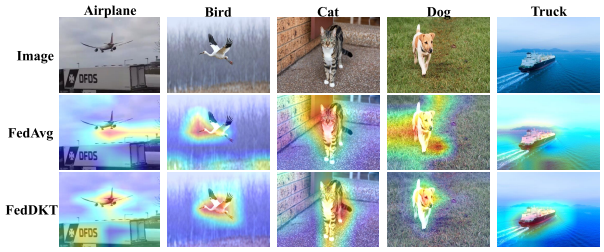

Fig. 4. Visualization of FedAvg and FedDKT based on CAM.

Visualization analysis. We visualize FedAvg and FedDKT using class activation mapping (CAM) [38], as shown in Fig. 4. Compared with FedAvg, FedDKT learns from the generated data by transferring the distribution knowledge, further enhancing the feature extraction ability. In addition, FedDKT can identify the image regions related to the target category and reflect them on the network weights, which illustrates the effectiveness of the similarity measure and attention measure strategies. FedDKT regards the generated data as a teacher and helps the global model to decouple the feature representations from it, which also confirms the effectiveness of the data generation strategy.

V. CONCLUSION

In this paper, we proposed FedDKT, a novel FL approach that decoupled and transferred the distribution knowledge of the generated data to address heterogeneous FL. FedDKT focused on excavating the relationship among data samples and proposed corresponding loss functions. Specifically, the similarity measure and attention measure assisted model training by capturing the correlations and key dependencies of the

generated data, respectively. We also proposed that the distribution knowledge could be regarded as a kind of structured knowledge to be used as a general component in combination with other methods. Extensive empirical experiments indicated that our proposed approach was beneficial to improve the performance and convergence rate of the FL global model. In future work, we will continue to explore feature representation learning for heterogeneous data based on sample relationships.

ACKNOWLEDGMENT

This work has been supported by National Natural Science Foundation of China under Grant No.61702398, No.61672415, and No.61872282, and the Fundamental Research Funds for the Central Universities under Grant No.JB211504.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54. PMLR, Apr 2017, pp. 1273–1282.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, Bennis *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, Dec 2019.

[3] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," in *IEEE International Conference on Data Engineering*, Feb 2022.

[4] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, pp. 50–60, May 2020.

[5] A. Qiao, B. Aragam, B. Zhang, and E. Xing, "Fault tolerance in iterative-convergent machine learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5220–5230.

[6] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7611–7623, 2020.

[7] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3557–3568, 2020.

[8] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/pdf?id=6YEQUn0QICG

[9] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar *et al.*, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems*, vol. 2, Apr 2020, pp. 429–450.

[10] J. Zhang, S. Guo, X. Ma, H. Wang, W. Xu, and F. Wu, "Parameterized knowledge transfer for personalized federated learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[11] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2351–2363, 2020.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014.

[13] A. Mitra, R. Jaafar, G. J. Pappas, and H. Hassani, "Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 606–14 619, 2021.

[14] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba *et al.*, "Towards federated learning at scale: System design," *Proceedings of Machine Learning and Systems*, vol. 1, pp. 374–388, Feb 2019.

[15] A. Shamsian, A. Navon, E. Fetaya, and G. Chechik, "Personalized federated learning using hypernetworks," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, Jul 2021.

[16] D. Yu, H. Zhang, W. Chen, J. Yin, and T.-Y. Liu, "How does data augmentation affect privacy in machine learning?" in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10 746–10 753.

[17] M. Duan, D. Liu, X. Chen, R. Liu, Y. Tan, and L. Liang, "Self-balancing federated learning with global imbalanced data in mobile systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 1, pp. 59–71, 2021.

[18] T. Nguyen, R. Novak, L. Xiao, and J. Lee, "Dataset distillation with infinitely wide convolutional networks," in *Advances in Neural Information Processing Systems*, 2021. [Online]. Available: https://openreview.net/forum?id=hXWPpJedrVP

[19] D. K. Dennis, T. Li, and V. Smith, "Heterogeneity for the win: One-shot federated clustering," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2611–2620.

[20] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[21] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine *et al.*, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5191–5198.

[22] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[23] D. Song, J. Xu, J. Pang, and H. Huang, "Classifier-adaptation knowledge distillation framework for relation extraction and event detection with imbalanced data," *Information Sciences*, vol. 573, pp. 222–238, 2021.

[24] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971, Apr 2019.

[25] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated learning on non-iid data: A survey," *Neurocomputing*, vol. 465, pp. 371–390, 2021.

[26] K. Tang, J. Huang, and H. Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1513–1524, 2020.

[27] J. Feng, C. Rong, F. Sun, D. Guo, and Y. Li, "Pmf: A privacy-preserving human mobility prediction framework via federated learning," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–21, Mar 2020.

[28] T. Yoon, S. Shin, S. J. Hwang, and E. Yang, "Fedmix: Approximation of mixup under mean augmented federated learning," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=Ogga20D2HO-

[29] N. Shoham, T. Avidor *et al.*, "Overcoming forgetting in federated learning on non-iid data," *arXiv preprint arXiv:1910.07796*, 2019.

[30] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi *et al.*, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119. PMLR, Jul 2020, pp. 5132–5143.

[31] H. Huang, F. Shang, Y. Liu, and H. Liu, "Behavior mimics distribution: Combining individual and group behaviors for federated learning," *international joint conference on artificial intelligence*, 2021.

[32] M. Yang, X. Wang, H. Zhu, H. Wang, and H. Qian, "Federated learning with class imbalance reduction," in *2021 29th European Signal Processing Conference*, 2021, pp. 2174–2178.

[33] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 878–12 889.

[34] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "Emnist: Extending mnist to handwritten letters," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2921–2926.

[35] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[36] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.