Clustering-based curriculum construction for sample-balanced Federated learning

Zhuang Qi, Yuqing Wang, Zitan Chen, Ran Wang,
Xiangxu Meng, and Lei Meng*

Shandong University, Jinan, Shandong, China
{97qizhuang,2000wangyuqing,chenzt622,mangow0702}@gmail.com
{mxx,lmeng}@sdu.edu.cn

Abstract. Federated learning is a distributed machine learning scheme that provides data privacy-preserving solution. A key challenge is data distribution heterogeneity of on different parties in federated learning. Existing methods only focus on the training rule of local model rather than data itself. In this paper, we reveal an fact that improving the performance of the local model can bring performance gain to the global model. Motivated by this finding, this paper proposes a Clustering-based curriculum construction method to rank the complexity of instances, and develops a Federation curriculum learning algorithm (FedAC). Specifically, FedAC assigns different weights to training samples of different complexity, which is able to take full advantage of the valuable learning knowledge from a noisy and uneven-quality data. Experiments were conducted on two datasets in terms of performance comparison, ablation studies, and case studies, and the results verified that FedAC can improve the performance of the state-of-the-art Federated learning methods.

Keywords: Curriculum learning · Federated learning · Neural networks.

1 Introduction

Federated learning, as a privacy-preserving distributed machine learning paradigm, has attracted much attention in artificial intelligence [1–3]. It typically uses a central server to coordinate multiple clients for collaborative modeling, and protects the privacy of training data of all parties, aiming to achieve the same or similar performance as data sharing [4]. However, existing studies have verified that the data heterogeneity leads to the poor generalization ability of the fused model [5]. Many studies have focused on alleviating the problem of distribution heterogeneity in federated learning, but there is a lack of solutions to uneven data quality.

Existing studies trying to improve the model performance can be divided into two levels: data-level [7, 8] and model-level [9–11]. Data-level methods often use two techniques, including data augmentation and sampling. They alleviate the

* Corresponding author

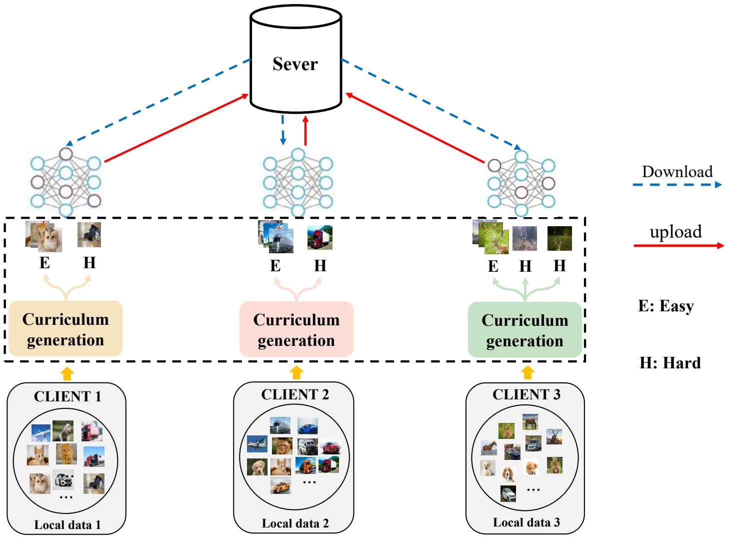
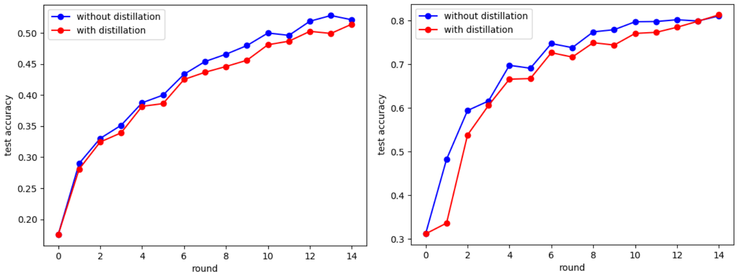


Fig. 1. Illustration to FedAC for image classification in federated learning.

imbalance by increasing the data during training, but ignore the quality of self-generated data. Many model-level approaches combine knowledge distillation to reduce global and local distribution bias. They add constraints between global and local models (e.g., model parameters [12], feature maps [13], probability distributions of predicted results [14].) to improve the similarity between local and global information. However, in the insufficient model training phase, distillation training is meaningless and even hindering model convergence. As shown in Figure 2, the accuracy of distillation method in the early stage is lower than baseline. The above methods only focus on the training rule of local model, while the impact of data quality to model performance have not been well analyzed.

To address aforementioned problems, this paper presents a Clustering-based curriculum construction for sample-balanced Federated learning method. As illustrated in Figure 1, it first ranks the complexity of data instances. Specifically, a clustering method is used to generate the hierarchy of image-class pairs from training set, which learns the different image patterns of same class. The hierarchy of each clients, serving as a personalized knowledge base, is able to filter the noisy data. To rank the complexity of all the data, a rule is adopted, which defines the score based on cluster size. The larger the cluster, the higher the data score it contains. In training phase, a loss weighting mechanism based on score is adopted. Then, an adaptive curriculum construction method is applied in local model training process. As observed, FedAC is able to get performance gain of global model by improving the robustness of local models.

Experiments have been conducted on the CIFAR-10 and CIFAR-100 datasets in terms of performance comparison, ablation study and case studies for the effectiveness of FedAC. The results verify Adaptive curriculum learning can improve the performance of local model and bring performance gain to the global model.



Cifar10 **FashionMNIST**

Fig. 2. Illustration the accuracy of with and without distillation methods.

To summarize, this paper includes two main contributions:

- An curriculum learning method, termed CG, is proposed, which enables achieve adaptive grouping of data. This can effectively use data to learn knowledge.
- A model-agnostic framework, termed FedAC, is proposed to combine curriculum learning method and can achieve local-global model performance gain.

2 Related work

2.1 Curriculum Learning

Curriculum learning (CL) imitates the human learning process, which ranks all instances based on complexity, and adopts the knowledge learning method of easy first and then hard [15]. The core problem of the curriculum learning is to get a ranking function, termed as Difficulty Measuring, which gives its learning priority for each piece of data or task. In addition, the training rule is determined by the training scheduler [16]. Therefore, there are many CL methods based on the framework of ”difficulty measuring and training scheduler”. CL can be divided into two categories based on whether it is automatically designed or not, namely Predefined CL [17] and Automatic CL [18]. Both the difficulty measuring and training scheduler of Predefined CL are designed by human experts using human prior knowledge, while at least one of Automatic CL is automatically designed in a data-driven manner.

2.2 Federated Learning

Many strategies have been proposed to address the data heterogeneity problem in Federated Learning (FL), which are mainly from two perspectives: data-level and model-level. Data level methods usually generate extra data to achieve data balance [6]. For example, Astraea uses data augmentation based on global data distribution to alleviate imbalances [8]. Model-level methods focus on optimization strategies to make the diversity between client local models and global

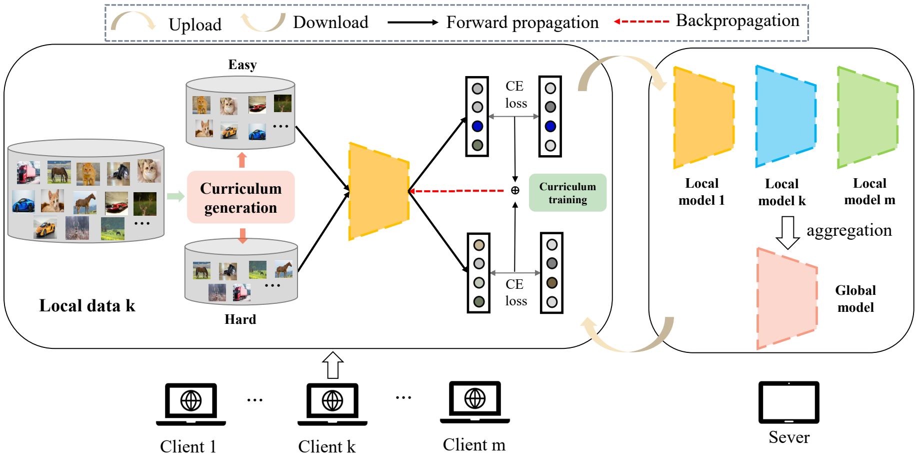


Fig. 3. The framework of FedAC.

model limited [19]. Fedprox restricts local model parameter updates from being far away from the global model [12]. MOON utilizes the similarity between local and global model representations to correct local training for all parties [5]. FML coordinates all clients to jointly train global models and independently train personalized models, and realize global knowledge transfer through deep mutual learning [9]. For studies in the aggregation phase, FedMA matches and averages weights in a hierarchical manner [20], FedNova normalizes local updates before averaging [21]. Notably, these aggregation ideas can be combined with our study.

3 Problem Formulation

This paper investigates curriculum learning for local training in the image classification task of federated learning. Suppose there are N clients, $(C_1, C_2, ..., C_N)$. And N clients hold heterogeneous data partition $\mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^N$, respectively. The goal is to learning a global model ω over the dataset $\mathcal{D} \bigcup_{k\in[N]} \mathcal{D}^k$ with the coordination of the central server without data share. For local training, client k starts with copying the weight vector $w^k \in R^d$, and the goal is mininzing the local loss function $f_k(w^k)$. The updated weight vector w^k can be obtain gradient decent method, i.e. $w^{k+1} \leftarrow w^k - \eta \nabla f_k\left(w^k\right)$. Finally, the global model can be obtained by aggregating all local models, i.e. $w^{global} \leftarrow \sum_{k=1}^{N} \frac{|D^k|}{|D|} w^k$.

Beyond conventional settings, our proposed FedAC first introduces a curriculum learning method to learn the image patterns of the same class. This enables the data grouping as a curriculum for all local data, i.e. $\mathrm{D}^k \mapsto \{D_1^{\mathrm{k}}, D_2^{\mathrm{k}}\}$. Subsequently, FedAC uses a image encoder to learn knowledge from D_1^{k} and D_2^{k}, respectively. By weighting the corresponding loss function $L = \alpha L_1 + \beta L_2$ as total loss. This enables encoders learn accurate knowledge.

4 Federation curriculum learning

FedAC introduces a curriculum learning framework to learn the complexity of data. As shown in Figure 3, FedAC contains two main modules in each clients, including Curriculum Generation (CG) Module and Curriculum Training (CT) Module, as illustrated in the following sub-sections.

4.1 Curriculum Generation (CG) Module

FedAC explores the complexity of all data in CG module, with the aim of completing data grouping. As shown in the Figure 4, CG module divides the raw data into two parts. Given a dataset including images $\mathcal{I} = \{\mathbf{I}_i | i = 1, 2, ..., N\}$ and corresponding labels of J classes $\mathcal{C} = \{c_j | j = 1, 2, ..., J\}$. Specifically, CG module first uses multi-channel clustering method to learn the data pattern of each classes. Notably, the ART algorithm [22] is extended to gather the similar features of the same class into a cluster. The details are shown as follows:

- For a data $p = (\mathbf{x}, O_x)$, O_x is one-hot version of the label, let $I = [\mathbf{x}, 1 - \mathbf{x}]$, the match score can be calculated by formula 1:

$$C_p = \left\{ c_j \mid \min \left\{ \frac{|\mathbf{I} \wedge w_j|}{|\mathbf{I}|}, O_j^T O_x \right\} \geq \rho, j = 1, \ldots, J \right\} \tag{1}$$

  where w_j, O_j are the weight vector and indicator of cluster c_j, O_j^T denotes the transposition of O_j, $p \wedge q = \min\{p, q\}$, $|p| = \sum_i p_i, \rho \in [0, 1]$ is vigilance parameter. If C_p is a empty set, generate a new cluster, otherwise proceed to the next step.
- For each candidate cluster $c_j \in C_p$, the choice function T_j with a choice parameter α as shown in formula 2:

$$T_j = \frac{|\mathbf{I} \wedge \mathbf{w_j}|}{\alpha + |w_j|} \tag{2}$$

- For the final-match cluster c_j*, using formula 3 to update the corresponding weight vector w_j*, and $\beta \in [0, 1]$

$$\hat{w}_{j*} = \beta \left(I \wedge w_{j*} \right) + (1 - \beta) w_{j*} \tag{3}$$

After the cluster process, many clusters C with the same class are obtained, the overall process can be expressed as formula 4:

$$C = Clustering(\mathcal{D}, \mathcal{C}) \tag{4}$$

where $\mathcal{D}$ denotes all image data, and $\mathcal{C}$ is the corresponding class labels.

Then, the Statistic-based Cluster filtering method $S(.)$ is used to divide all clusters into two parts based on cluster size, i.e.

$$P_1, P_2 = S(C) \tag{5}$$

where $P_1 = \{c_i | \ if \ |c_i| < T\}$ and $P_2 = \{c_i | \ if \ |c_i| \geq T\}$, and T is a threshold.

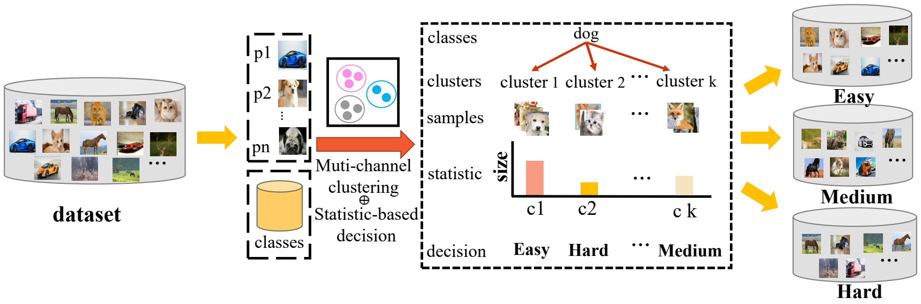


Fig. 4. Illustration of the CG module.

Table 1. Statistics of the datasets used in the experiments.

Datasets	**#class**	**#Training**	**#Testing**
CIFAR-10	10	50000	10000
CIFAR-100	100	50000	10000

4.2 Curriculum Training (CT) Module

After data grouping, CT Module trains a image encoder $E(\cdot)$ on P_1 and P_2. This paper proposes a weighting loss approach to reduce the adverse impact of uneven-quality data during training phase. The Equation 6 and 7 represent the loss function of easy and hard data, respectively.

$$L_1 = CE(p_1, y_1) \tag{6}$$

$$L_2 = CE(p_2, y_2) \tag{7}$$

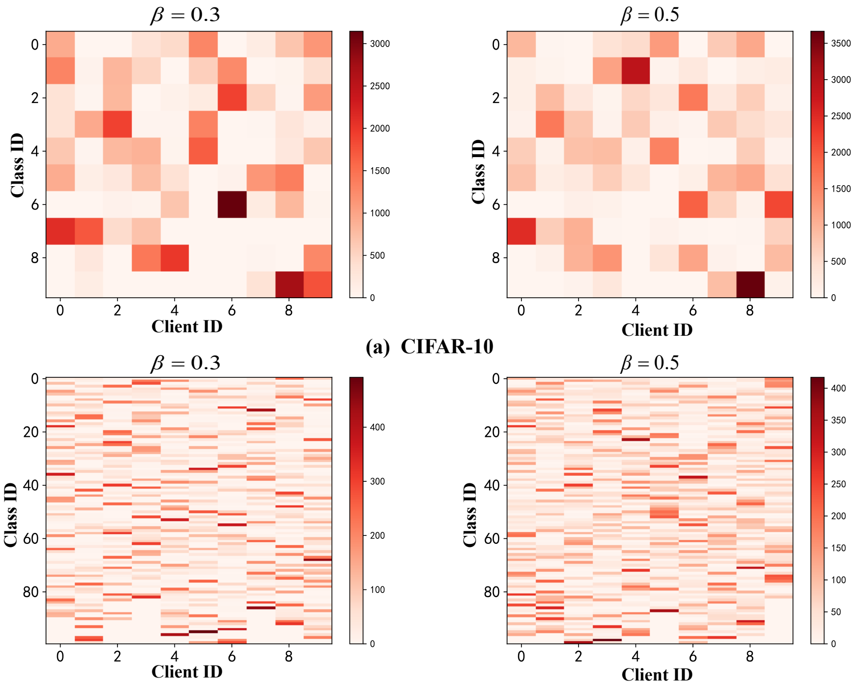
where p_i and y_i denote the prediction and ground-truth of data x_i. The total loss is weighted by L_1 and L_2, i.e.

$$L_{total} = \alpha L_1 + \beta L_2 \tag{8}$$

5 Experiments

5.1 Experimental Setup

Datasets We use two benchmarking datasets CIFAR-10 and CIFAR-100 that are commonly used in federated classification for experiments. Their statistics are showing in Table 1. Like recent studies, the Dirichlet distribution $Dir_N(\cdot)$ is used to generate the non-IID distribution among all parties. Specifically, we use $p_k \sim Dir_N(\beta)$ to sample and allocate a $p_{k,j}$ proportion of the instances of class k to client j, where β is a concentration parameter. We set 10 clients by default in all experiments. The data partition results of different parameters are shown in Figure 5.



(a) CIFAR-10

(b) CIFAR-100

Fig. 5. Data distribution corresponding to different β values on CIFAR-10 and CIFAR-100 datasets.

Evaluation Measures In the experiment, the Top-1 Accuracy is used to evaluate the performance of all models, the define of Accuracy is:

$$Accuracy = (TP + TN)/(P + N) \tag{9}$$

where TP, TN, P and N denote True Positives, True Negatives, Positives, and Negatives.

Implementation Details Following the work of MOON [5], we use Simple-CNN (including 3 Convolution layers and 2 fully connected layers with ReLU activation) model on CIFAR-10. For CIFAR-100, the ResNet-18 [23] model is used as a base encoder. We use Stochastic Gradient Descent (SGD) optimizer with a learning rate 0.01 and weight decay 0.00001 for all methods. The batch size is selected from $\{64, 128\}$. The training epochs in local clients is 10 by default for federated learning methods, and 100 for SOLO. And the local-global communication round is set to 100 for two datasets. To verify the effectiveness of each algorithm, we set the Dirichlet distribution parameters $\beta = 0.3$ and $\beta = 0.5$. Finally, some hyperparameter related to specific algorithms, please refer to the corresponding papers.

Table 2. The top-1 accuracy of FedAC and the other baselines on CIFAR-10 and CIFAR-100 datasets. For all datasets, dirichlet distribution parameter selection 0.3 and 0.5.

Methods	CIFAR-10		CIFAR-100	
	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.3$	$\alpha = 0.5$
SOLO	46.8	47.3	22.5	23.4
FedAvg	65.9	66.2	63.8	64.1
FedProx	66.2	66.7	63.7	64.6
MOON	66.4	66.8	64.9	64.3
FedAC(our)	**67.6**	**67.7**	**65.5**	**65.4**

5.2 Performance Comparison

This section presents a performance comparison between FedAC and existing Federated learning methods on image classification tasks, including FedAvg [24], FedProx [12] and MOON [5]. A baseline named SOLO, each party trains personalized model with its local data without federated learning. For all methods, we fine-tune their hyper- parameters based on corresponding papers to get the best performance. We can observe the followings as shown in the Table 2:

- The performance of SOLO in non-IID data distribution is worse than that of federated learning algorithms. This is mainly because local models are biased classifiers with poor performance. And this verifies the benefits of federated learning.
- FedProx and MOON outperformed FedAvg on both datasets, this verifies that global knowledge distillation is conducive to guide local model training. And global features seem more useful than parameters.
- For different dirichlet distribution parameters, all methods achieve better performance at $\alpha = 5$, this is mainly because the data distribution is more dispersed at $\alpha = 3$.
- On both datasets, the proposed FedAC achieves significant performance improvement than existing methods, demonstrating optimize the local training process can bring performance gain to the global model.

5.3 Ablation Study

In this section, we investigate the effect of curriculum learning method for local training. From Table 3, the following observations can be drawn:

- For SOLO method, the FedAC(SOLO) achieves 5% performance improvement on the CIFAR10 at $\alpha = 0.5$. And it is more effective on CIFAR10 than on CIFAR100.
- For FedProx and MOON, FedAC version methods also make progress. This verifies that curriculum learning can bring performance gain for local model and global model.

Table 3. The top-1 accuracy of SOLO, FedProx, MOON and their FedAC version.

Methods	CIFAR-10		CIFAR-100	
	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.3$	$\alpha = 0.5$
SOLO	46.8	47.3	22.5	23.4
FedAC(SOLO)	**49.5**	**50.2**	**24.1**	**23.5**
FedProx	66.2	66.7	63.7	64.6
FedAC(FedProx)	**66.8**	**67.3**	**66.4**	**65.8**
MOON	66.4	66.8	64.9	64.3
FedAC(MOON)	**67.2**	**67.6**	**66.3**	**66.8**

- For both datasets, MOON-based methods are always better than the FedProx-based approaches, which may be because the feature is more representative of knowledge than the model parameters.

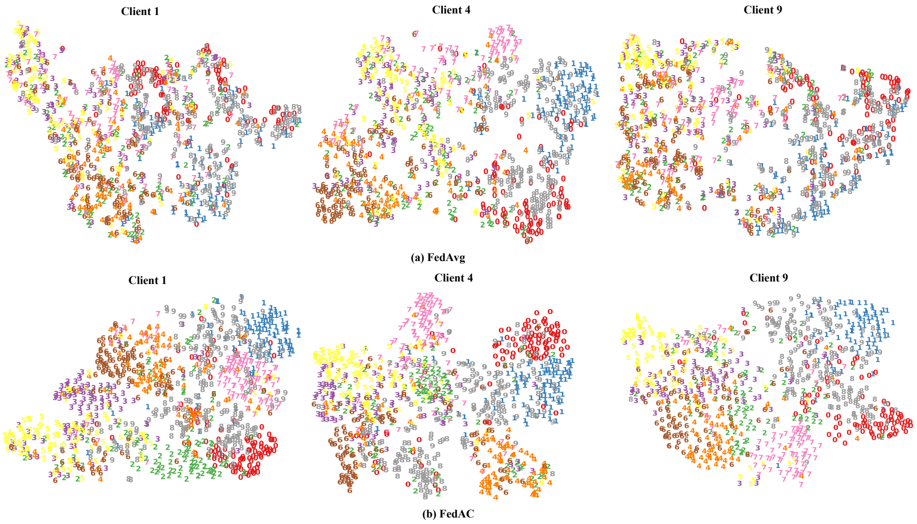
5.4 Case Study

In this section, we will further analyze how FedAC improves local and global model performance in view of hidden vectors distribution. To this end, 100 images of all classes were randomly selected in the test set, which were unknowable in the training phase. In this paper, T-SNE [25] was used to explore the distribution changes of their corresponding representations. We randomly selected three clients and visualized their hidden vectors in the test set.

As shown in Figure 5, the hidden vectors distribution of FedAvg and FedAC have a large difference. Specifically, the training model in FedAvg algorithm learns poor features, and the feature representation of most classes is even mixed, which can not be distinguished. For FedAC, we can observe that the points with the same class are more divergent in Figure 6(b) compared with Figure 6(a) (e.g., see class 0, 1 and 7). Compared with FedAC, FedAvg algorithm learns worse representation in local training phase. This may leads a poor performance of fused model. Therefore, improving the generalization ability of local models may also bring performance gains to global models.

6 Conclusion

As an effective method to solve data silos, federated learning has attracted attention in many fields, such as medicine and finance. To improve the performance of models on non-IID datasets, we proposed Adaptive curriculum Federated learning (FedAC), simple and effective approach for federated learning. To alleviate the negative impact of uneven quality, FedAC introduces a novel curriculum learning method for local training. It utilizes the complexity of the data to design the weighting method. Experiments results show that FedAC achieves significant improvement of local model to obtain the global gain on image classification tasks.



(a) FedAvg

(b) FedAC

Fig. 6. T-SNE visualizations of hidden vectors on CIFAR-10.

Future work of this study can be further explored in two directions. First, stronger curriculum inference techniques can significantly improve performance in guiding the local model to learn personalized knowledge. Second, FedAC can be applied to other problems, such as Natural Language Processing (NLP) and Recommendation System (RS).

Acknowledgments

This work is supported in part by the Excellent Youth Scholars Program of Shandong Province (Grant no. 2022HWYQ-048) and the Oversea Innovation Team Project of the ”20 Regulations for New Universities” funding program of Jinan (Grant no. 2021GXRC073)

References

1. Yang Liu, Anbu Huang, Y un Luo, He Huang, Y ouzhi Liu, Y uanyuan Chen, Lican Feng, Tianjian Chen, Han Y u, and Qiang Yang. Fedvision: An online visual object detection platform powered by federated learning. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 13172–13179, 2020.
2. Qiang Yang, Yang Liu, Tianjian Chen, and Y ongxin Tong. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2):1–19, 2019.
3. Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977, 2019.

4. Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, pages 1–7, 2020.
5. Li Q, He B, Song D. Model-contrastive federated learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10713-10722.
6. Hao W, El-Khamy M, Lee J, et al. Towards fair federated learning with zero-shot data augmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3310-3319.
7. Lin T, Kong L, Stich S U, et al. Ensemble Distillation for Robust Model Fusion in Federated Learning[C]//Advances in Neural Information Processing Systems. 2020: 23512363.
8. Duan M, Liu D, Chen X, et al. Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications[C]//2019 IEEE 37th international conference on computer design (ICCD). IEEE, 2019: 246-254.
9. Shen T, Zhang J, Jia X, et al. Federated mutual learning[J]. arXiv preprint arXiv:2006.16765, 2020.
10. Zhu Z, Hong J, Zhou J. Data-free knowledge distillation for heterogeneous federated learning[C]//International Conference on Machine Learning. PMLR, 2021: 12878-12889.
11. Yao X, Sun L. Continual local training for better initialization of federated models[C]//2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020: 1736-1740.
12. Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In Third Conference on Machine Learning and Systems (MLSys), 2020.
13. Wu C, Wu F, Liu R, et al. FedKD: Communication Efficient Federated Learning via Knowledge Dis tillation[J]. ArXiv, 2021, abs/2108.13323.
14. Li D, Wang J. Fedmd: Heterogenous federated learning via model distillation[J]. arXiv preprint arXiv:1910.03581, 2019.
15. Bengio Y, Louradour J, Collobert R, et al. Curriculum learning[C]//Proceedings of the 26th annual international conference on machine learning. 2009: 41-48.
16. Soviany P, Ionescu R T, Rota P, et al. Curriculum learning: A survey[J]. International Journal of Computer Vision, 2022: 1-40.
17. Jiang L, Meng D, Zhao Q, et al. Self-paced curriculum learning[C]//Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.
18. Braun S, Neil D, Liu S C. A curriculum learning method for improved noise robustness in automatic speech recognition[C]//2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017: 548-552.
19. Yao D, Pan W, Dai Y, et al. LocalGlobal Knowledge Distillation in Heterogeneous Federated Learning with NonIID Data[J]. ArXiv, 2021, abs/2107.00051.
20. Hongyi Wang, Mikhail Y urochkin, Y uekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In International Conference on Learning Representations, 2020.
21. Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. Advances in Neural Information Processing Systems, 33, 2020.
22. Meng L, Tan A H, Miao C. Salience-aware adaptive resonance theory for large-scale sparse data clustering[J]. Neural Networks, 2019, 120: 143-157.

23. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceed ings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
24. H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. arXiv preprint arXiv:1602.05629, 2016.
25. Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.