Class-Wise Adaptive Self Distillation for Heterogeneous Federated Learning

Yuting He[1,2], **Yiqiang Chen**[1,2], **Xiaodong Yang**[1,3], **Yingwei Zhang**[1], **Bixiao Zeng**[1,2]

[1] The Beijing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, 100190
[2] University of Chinese Academy of Sciences, Beijing, China, 100190
[3] Shandong Academy of Intelligent Computing Technology, Jinan, China, 250101
{heyuting20s, yqchen, yangxiaodong, zhangyingwei, zengbixiao19b}@ict.ac.cn

Abstract

The heterogeneity of data distributions among clients (non-IID) has been identified as one of the key challenges in federated learning. In the local training phase, each client model optimized towards its own local optima instead of solving the global objective, which results in forgetting the global knowledge and raises a drift across client updates. Some previous methods leverage knowledge distillation (KD) to avoid the federated forgetting, but most of them do not consider the global teacher model's ability on different categories and might mislead the local student models' training consequently. To address this issue, we propose a Class-wise Adaptive self-Distillation method for Federated Learning, which is named FedCAD. Before local training at each round, FedCAD assesses the inference confidence on each category of the global model using an auxiliary dataset, which is used to indicate how much the global model should be trusted. Based on the assessments, a class-wise adaptive weight is used to dynamically adjust the impact of the global teacher model on the local training of each category. In this way, the distilled knowledge from the global teacher can be selectively learned by the local students to avoid negative impacts. The extensive experimental results on the public datasets, i.e., CIFAR10, CIFAR100 and FEMINST, demonstrate that the proposed FedCAD has better performance in terms of convergence speed and classification accuracy, compared to other state-of-the-art FL methods.

Introduction

Nowadays, privacy protection has attracted increasing attention in modern society with the introduction of regulations such as the General Data Protection Regulation (GDPR). The data collected by different devices or organizations cannot be sent to a centralized server according to the requirement of privacy protection regulation, forming distributed database consisting of multiple "data islands". Federated learning (FL) is proposed to cope with the "data island" dilemma, which enables clients to collaboratively train a generalized and robust model while keeping their local data decentralized. Most existing FL algorithms follow the procedure of FedAvg (McMahan et al. 2017). In each communication round, the server first selects a part of clients and sends the weights of global model to them to initialize the local models. The selected clients then train the local models using their private local data and transfer the optimized weights back to the server. Finally, the server aggregates the local models to update the global model. The above process repeats until the global model converges.

A key challenge in FL is the heterogeneity of data distribution among the clients. The local data of each client can be non-identically distributed (non-IID) in real-world, which means the local data distributions may differ from the overall global distribution. The heterogeneity of local data not only makes the theoretical analysis difficult (Khaled, Mishchenko, and Richtárik 2020; Li et al. 2020c), but also leads to the performance degradation and slow convergence (Zhao et al. 2018; Li et al. 2020a). When each client trains a local model on the biased local data, its local objective may deviate from the global objective and result in forgetting the global knowledge. Therefore, the global model drift from the optimum of the global objective. (Karimireddy et al. 2020) called this phenomenon as "Client-drift". Many studies have tried to address the client-drift problem at the local training phase. FedProx (Li et al. 2020b) directly limits the local updates by adding an additional L2 regularization term to the local objective function. SCAFFOLD (Karimireddy et al. 2020) uses control variate to correct the client-drift. However, the effects of these methods are not significant when the local data is heterogeneous.

Interestingly, Continual Learning (CL) faces an analogous problem: how to learn a task without forgetting another one learned previously. Some studies in CL apply knowledge distillation to keep the representations of previous data from drifting too much while learning new tasks. Inspired by this idea, FedLSD (Lee et al. 2021) and FedGKD (Yao et al. 2021) train local models with the guidance of global model's prediction on local data to preserve global knowledge. More specifically, the local models self-distill the distributed global model's prediction on local data.

However, due to the non-IID of the data distribution, the convergence speed and accuracy of the global model are different among classes. What's more, the credibility of distillation knowledge tends to increase with the convergence of the global model. Hence, it's not appropriate to use a constant coefficient to control the distillation loss. When teacher makes confident mispredictions, especially on hard classes, distillation can disproportionately harm the perfor-

mance on the classes and amplified by the student (Lukasik et al. 2021).

The discoveries inspire us to propose a class-wise adaptive self distillation mechanism namely FedCAD. FedCAD utilizes a class-wise adaptive weight to help local models adaptively control the impact of distillation based on the performance of the global model on each class. When the prediction of global model on the class is credible, local models learn more from the distilled knowledge. Otherwise, local models will focus more on the local ground truth labels.

Our main contributions are as follows:

- We analyze the problems of the existing FL algorithm, which keep a constant ratio of distillation loss during local training process, and propose an adaptive adjustment mechanism of the distillation loss to address them.
- We propose a class-wise adaptive weight, determined by the performance of the global model on each class of an auxiliary dataset, to realize the adaptive adjustment of the distillation loss.

Related Work

Federated Learning

Federated learning (FL) is first proposed by (McMahan et al. 2017) as a distributed machine learning paradigm. A key challenge in federated learning is that data distribution in different clients is usually non-identically distributed (non-IID). Many studies are trying to address the non-IID issue, which mainly improves two phases: local training phase and server aggregation phase. Our work belongs to the first one.

As for the studies on improving the local training phase, most of them use a regularization item to impose constraints on the update of local models. FedProx (Li et al. 2020b) directly limits the local updates by adding an additional L2 regularization term to the local objective function. FedNova (Wang et al. 2020b) introduces weight modifications to FedAvg using the number of local steps to normalize and scale the local updates of each client. Motivated by the continual learning, FedCurv (Shoham et al. 2019) and FedCL (Yao and Sun 2020) add a penalty term to the local objective function to prevent the important parameters of the global model from changing too much. They estimate parameter importance by the diagonal of the empirical Fisher Information Matrix, which is inspired by EWC (Kirkpatrick et al. 2017).

As for the studies on improving the server aggregation phase, several works have proposed a layer-wise aggregation strategy to adapt to data heterogeneity, applying Bayesian nonparametrics to match and average the parameters. For instance, instead of averaging the parameters weight-wise without considering the meaning of each parameter, PFNM (Yurochkin et al. 2019) and FedMA (Wang et al. 2020a) use the Beta-Bernouilli Process for matching parameters. Specifically, FedMA is an improved version of PFNM which extends the matching strategy from fully connected layers to CNNs and LSTMs.

Recently, personalized federated learning has attracted significant interest from researchers (Deng, Kamani, and Mahdavi 2020; Chen et al. 2021; Huang et al. 2021), which tries to train personalized local models for each client. In this paper, we study the typical federated learning, targeting at training a single generalized model for all clients.

Knowledge Distillation in FL

Knowledge Distillation(KD) is proposed to transfer knowledge from a large teacher model to a small student model (Hinton, Vinyals, and Dean 2015), which is widely used for model compression (Wang et al. 2020c; Sun et al. 2019). Distillation in federated learning has recently emerged as an effective approach to track the data heterogeneity. Numerous related works study the ensemble distillation, i.e. distilling the knowledge from the ensemble of teachers (local models) to a student (global model). In Federated distillation(FD) (Jeong et al. 2018; Seo et al. 2020), clients share the model output parameters(logits) as opposed to the model parameters(weights or gradients) to reduce the communication costs. Then, the averaged logits are used to regularize local training. FedMD (Li and Wang 2019) and Cronus (Chang et al. 2019) use an public dataset to get the averaged logits per sample. FedDF (Lin et al. 2020) use a unlabeled datasets in the server to aggregated knowledge from all received local model. Furthermore, the above methods can deal with the model heterogeneity and each client can design unique model. Instead of treating the ensembles of local models as teachers and transfer the knowledge into global model, FedLSD (Lee et al. 2021) and FedGKD (Yao et al. 2021) regard the global model as teacher and self-distills the distributed global model's prediction during the local training to preserve global knowledge.

Method

Definition and Background

Federated Learning is typically formulated as the following optimization problem:

$$\min_{w} F(w) = \sum_{i=1}^{N} q_i f_i(w), \quad q_i = |D_i| / \sum_{j=1}^{N} |D_j| \qquad (1)$$

where the global objective function $F(w)$ is a weighted average of the local objectives $f_i(w)$ over N clients. q_i is the weight of each client, which is typically set as proportional to the sizes of the local datasets $|D_i|$.

The local datasets $D_1, D_2, ..., D_N$ are usually non-identically distributed (non-IID) in practice. In this paper, we focus on label distribution non-IID challenge as categorized in (Kairouz et al. 2019). The label distribution $P(y)$ may vary across clients, even if $P(x|y)$ is the same. We assume each client only own a partial class set.

Motivation

Under the non-IID data scenarios, local data fail to represent the overall global data distribution. On certain classes, the global model extracts better feature representations than the local models trained on skewed data of clients, so there are drifts in the local updates (Li, He, and Song 2021). An intuitive fix is to utilize the global model's prediction on local data to make local models preserve the knowledge which the local distributions cannot represent (Lee et al. 2021).

Figure 1: The special confusion matrix of CIFAR-10. We plot the data distribution on a certain client (first row), per-class accuracy of the initial global model (second row) and local model with different algorithms (last three rows).

Distillation can control the drift and bridges the gap between the representations learned by the local models and the global model. Then, the problem is to what degree we should trust the global model. If the global model confidently mis-predicts the samples of some classes, such inaccuracy may disturb the training of local models.

To confirm our conjecture on global model's misleading, we analyze per-class accuracy of the global model and local models in the scenario of non-IID data distribution. A global model is trained for 10 communication rounds using FedAvg, and distributed to a certain client as an initial model. Then, we use FedAvg, FedLSD and FedCAD respectively to train local models for 10 epochs. As shown in Figure 1, FedAvg achieves high accuracy on the local majority categories but low on minority ones, which means forgetting the initial global knowledge. Contrastively, FedLSD maintains the global view on local data and has low error levels on categories 0-5. However, the improvement of FedLSD on categories 7-9 is limited, which is due to the misleading of the initial model. FedCAD improves the accuracy of local model on almost all categories, which demonstrates our approach avoids the misleading of the global model and meanwhile extracts reliable knowledge from it.

Local self distillation

In each communication round t, the local models are initialized with the global aggregated model and then optimize their local loss by running SGD for E local epochs. To keep the global knowledge during local training, the local model is learned by using a classification loss and a distillation loss.

Suppose client i is conducting the local training. We denote the samples in client i as $D_i = (x, y), y \in [0, K]$, where K is the total number of classes. We also denote the output logits of local model and global model as z, z^g respectively. The distillation loss L_d is the KL-Divergence loss between global prediction and local prediction, and it is formulated as follows:

$$L_d = \sum_{x \in D_i} \sum_{k=1}^{K} -p_k^g(x) \log[p_k(x)] \tag{2}$$

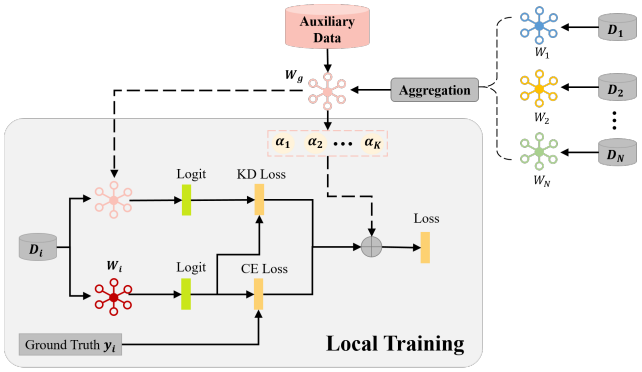


Figure 2: The framework of the FedCAD.

$$p_k^g(x) = \frac{e^{z_k^g(x)/T}}{\sum_{j=1}^{K} e^{z_j^g(x)/T}}, \quad p_k(x) = \frac{e^{z_k(x)/T}}{\sum_{j=1}^{K} e^{z_j(x)/T}}$$

where T is the temperature scalar, which increases the weight of smaller logit values and encourages the network to better encode similarities among classes.

The classification loss L_c is the softmax cross-entropy loss between the local model probability and the onehot labels y, which is computed as follows:

$$L_c = \sum_{x \in D_i} \sum_{k=1}^{K} -y \log[p_k(x)] \tag{3}$$

The overall loss is a weighted combination of two objectives as follows:

$$L = (1 - \alpha)L_c + \alpha L_d \tag{4}$$

where $\alpha \in [0, 1]$ is a hyper-parameter.

Class-wise adaptive weights

In motivation, we see that distillation causes degradation on classes where the teacher is inherently inaccurate. An intuitive fix is to adapt per-class adaptive weights $(\alpha_1, ..., \alpha_K) \in [0, 1]^K$ to rely less on the global model for the classes where it predicts poorly. We modified the Eq.4 as follows:

$$L = (1 - \alpha_y)L_c + \alpha_y L_d \tag{5}$$

Then the question is how to decide values for the class-wise weights α_y. In Eq.4, it only needs to tune a single scalar α and we can use cross-validation to determine the specific values. However, it is infeasible to attempt a grid search for α_y. Inspired by (Lukasik et al. 2021), We propose the following function to set class-wise weights α_y given the global model's prediction p^g.

$$\alpha_y = \frac{1}{2}(\gamma - \beta)\mathbb{E}_{x|y}[\phi(y, p^g(x))] + \frac{1}{2}(\gamma + \beta) \tag{6}$$

$$\phi(y, p^g(x)) = p_y^g(x) - \sum_{k \neq y}^{K} p_k^g(x)$$

where $0 < \beta < \gamma < 1$ decide the lower and upper bounds of distillation impact. Concretely, We rely more on the global model on the classes where its predictions tend to be more

Algorithm 1: FedCAD

Input: N clients' datasets $\{D_i\}_{i=1}^{N}$, total communication rounds T, local epochs E, learning rate η
, clients sample ratio C. **Output**: The final global model w^T
Server execute:

Initialize global model w^0 in sever
for $t = 0, ..., T-1$ **do**
  Estimate α_y on auxiliary data **[Eq. 6]**
  $S_t \leftarrow$ Randomly sample a set of $C \cdot N$ clients
  for $i \in S_t$ in parallel do **do**
    $w_i^t \leftarrow$ **ClientUpdate(**i, w^t, α_y**)**
  end for
  $w^{t+1} \leftarrow \frac{1}{|D_{S_t}|} \sum_{i \in S_t} |D_i| w_i^t$
end for

ClientUpdate:(i, w^t, α_y**)**

$w_i^t \leftarrow w^t$
for epoch $e = 1, 2, ...E$ **do**
  for batch $b = \{x, y\} \in D_i$ **do**
    $L_c \leftarrow CrossEntropyLoss(z(x), y)$
    $L_d \leftarrow \sum_{x \in D_i} \sum_{k=1}^{K} -p_k^g(x) \log[p_k(x)]$ **[Eq. 2]**
    $L \leftarrow (1 - \alpha_y) L_c + \alpha_y L_d$ **[Eq. 5]**
    $w_i^t \leftarrow w_i^t - \eta \nabla L(w_i^t, b)$
  end for
end for
return w_i^t to the server

correct, i.e., the large α. we use the auxiliary data in the server to estimate the expectation $\mathbb{E}_{x|y}[\cdot]$. Notice that in the circumstance of $\gamma = \beta = 0$, our approach is equivalent to FedAvg. When $\gamma = \beta = C$ where $C \in [0, 1]$ is some constant number, our approach degrades to FedLSD. In this sense, our method is more generalized and robust towards different amounts of drifts.

The framework of FedCAD is shown in Figure 2. We also give a detailed description in Algorithm 1. In each communication round, the server sends the global model and class-wise weights to the selected clients; receives the trained local models from the clients; updates the global model by weighted averaging and calculates class-wise weights on an auxiliary data for next communication round. In local training, each client initializes global model as local model and then uses SGD to optimize local model with its local data, while the objective is defined in Eq.5.

Experiments

Experimental Setup

Dataset. We evaluate our method on three benchmark datasets: CIFAR10, CIFAR100 (Krizhevsky, Hinton et al. 2009) and FEMNIST of LEAF benchmark (Caldas et al. 2018). We conduct two different data partitioning strategies among clients to simulate the non-IID data distribution on clients, which is inspired by (Li et al. 2021a). For label distribution skew, we allocate a proportion of the samples of each class to each client according to Dirichlet distribution on CIFAR10 and CIFAR100. Specifically, we sample $q_k \sim Dirichlet(\delta)$ and allocate a $q_{k,i}$ proportion of the instances of class k to client i, where δ is the concentration parameter controlling the uniformity between clients. We set the δ to 0.5 and the number of clients to 10 by default. For FEMNIST, we randomly divide and assign approximately 2000 writers into 10 clients by default. An example of the data distributions among clients is shown in Figure 3. The auxiliary data is a small subset of samples of different classes, which can be acquired from the public data. In our experiments, we sample auxiliary date according to the strategy from (Wang et al. 2021), using only 32 samples for each class.

Models. For CIFAR10, We use the same CNN architecture as FedAvg: two 5×5 convolution layers (the first with 6 channels and the second with 16 channels, each followed by a ReLU activation and 2×2 max pooling), two fully connected layers with ReLU activation (the first with 120 units and the second with 84 units) and a final softmax output layer. For CIFAR100 and FEMNIST, we use ResNet50 (He et al. 2016) instead. The models are implemented in PyTorch (Paszke et al. 2019) and trained on a single RTX 3090 GPU.

Hyper-parameters. In local training, we use the SGD optimizer with initial learning rate 0.01 and momentum 0.9. The local epoch is set to $E = 10$ and local batch size is set to $B = 64$ by default. We run 100 communication rounds on CIFAR10/100 and 50 rounds on FEMNIST. The default clients sampling ratio is set as $C = 1$. For FedLSD, we set the temperature parameter T to 2 and tune the weight of distillation loss from $\{0.1, 0.3, 0.5\}$ like (Lee et al. 2021). For FedProx, we tune the weight of its proximal term μ from $\{0.001, 0.01, 0.1\}$. For FedCAD, we tune β and γ from $\{0, 0.3, 0.5, 0.7\}$ respectively and show the best result.

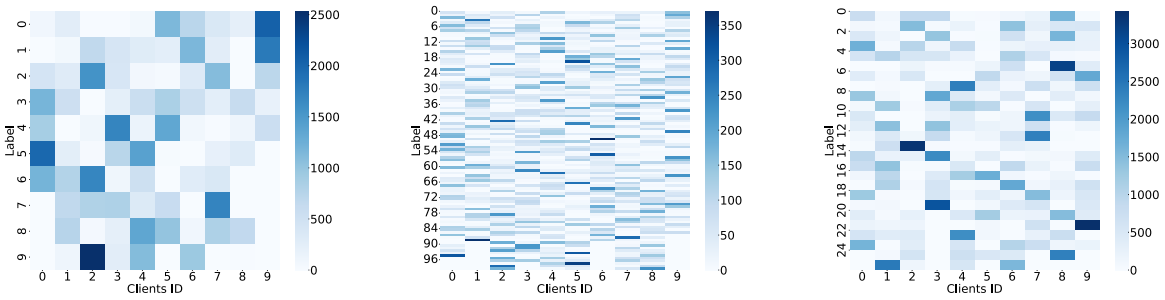
Classification Accuracy

We conduct experiments to validate the effectiveness of FedCAD and compare it with FedAVG (McMahan et al. 2017), FedProx (Li et al. 2020b) and FedLSD (Lee et al. 2021). The top-1 test accuracy with the default settings mentioned above are shown in Table 1. We can observe that the proposed FedCAD ranks first among the compared methods and outperforms FedLSD by 0.58% and 0.63% on CIFAR10 and CIFAR100 respectively. For FEMNIST, our method also achieves a high accuracy comparable to other state-of-the-art methods in the last round.

Moreover, we illustrate the number of communication rounds to reach the target accuracy in Table 1. Compared to FedAvg, FedCAD requires only 58.54% and 66.67% of the communication rounds to reach the accuracy of 65% on CFIAR10 and 90% on FEMNIST. On CIFAR100, the required training rounds of FedCAD, FedAvg and FedProx to achieve 65% accuracy are close and FedCAD achieves the second place performance on that. The overall results show that the class-wise weight adaptive distillation mechanism also helps to accelerate the convergence of the global model.

Adaptive Preservation of Global Knowledge

To evaluate the effectiveness of our method on local training, we also compute the average test accuracy of local models



(a) CIFAR10 (b) CIFAR100 (c) FEMNIST

Figure 3: The data distribution of each client. The color of each rectangle reflects the number of samples for a class on certain client.

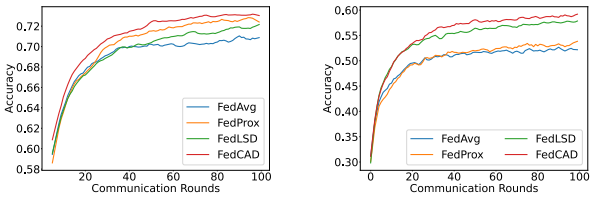
Algorithm	CIFAR10		CIFAR100		FEMNIST	
	acc ↑	T(65%) ↓	acc ↑	T(65%) ↓	acc ↑	T(90%) ↓
FedAvg	71.29%	41	71.87%	**19**	94.25%	6
FedProx	71.52%	43	71.58%	20	94.62%	6
FedLSD	72.56%	38	72.64%	29	**94.68%**	5
FedCAD	**73.14%**	**22**	**73.27%**	20	94.65%	**4**

Table 1: The top-1 test accuracy (acc) after training the target rounds and the number of communication rounds (T) to achieve the target accuracy. The **best** and the second best values are highlighted.

on a test dataset obeying the global data distribution. If local models preserve the global knowledge after fitting on the biased local data, the updated local model could be generalized well on the global test data distribution. As shown in Figure 4(b), the improved speed of local accuracy in FedCAD is almost the same as FedAvg at the beginning. Since the global model is far from convergence and the class-wise adaptive weight of distillation loss is small. On the one hand, FedLSD and FedCAD both achieve better accuracy levels than FedAvg thanks to the distillation loss after about 10 rounds, which helps local models to preserve the global knowledge and mitigate the catastrophic forgetting during local training. On the other hand, the accuracy of FedCAD exceeds FedLSD after about 30 rounds. The result demonstrates that, compared with the constant weighting of distillation loss, the class-wise adaptive mechanism makes local models more generalized on global data distribution.

Effects of Data Heterogeneity

To evaluate the robustness of our method under different data heterogeneity levels, we compare our method with other state-of-the-art methods by varying hyper-parameter δ of Dirichlet distribution on CIFAR-10. A bigger δ indicates a more uniform distribution. We present the results in Table 2. We observe that FedCAD outperforms other methods among different unbalanced levels. When the unbalanced level increases (i.e., $\delta = 0.1$), the performance of FedLSD is worse than FedAvg, while FedCAD still outperforms other methods. At such extreme unbalanced level, the guidance of the global model tends to be less reliable, which makes the accuracy of FedLSD degrade. The experiments demonstrate the robustness and effectiveness of FedCAD under different unbalanced levels.



(a) Global accuracy (b) Local average accuracy

Figure 4: The learning curves on CIFAR-10. The global accuracy is the performance of global model on test dataset. The local average accuracy is the average of the respective performance of local models on each online client.

Algorithm	$\delta = 0.1$	$\delta = 0.5$	$\delta = 5$	$\delta = 10$
FedAvg	61.73%	71.29%	71.56%	72.98%
FedProx	62.85%	71.52%	71.82%	72.26%
FedLSD	61.54%	72.56%	73.43%	73.34%
FedCAD	**63.80%**	**73.14%**	**74.00%**	**73.47%**

Table 2: The top-1 test accuracy with different levels of data heterogeneity on CIFAR10.

Analysis of Clients Participation Ratio

To evaluate the scalability of our method, we design the experiments with different number of participating clients at each communication round on CIFAR10. Specifically, We partition CIFAR10 training dataset into 100 clients and randomly sample $C = 0.1, 0.3, 0.5$ of clients to participate in the training during each round. As shown in Table 3, FedCAD stably outperforms other methods as the participation ratio growing.

Analysis of Local Epochs

Aggregating local models at different frequencies may affect the learning performance (Li et al. 2021b). Therefore, We further conduct the experiments to study the effect of local epochs on the performance of the final global model.

Algorithm	$C = 0.1$	$C = 0.3$	$C = 0.5$
FedAvg	57.64%	60.06%	60.76%
FedProx	58.64%	59.96%	60.66%
FedLSD	58.80%	62.28%	61.24%
FedCAD	**59.96%**	**62.59%**	**62.44%**

Table 3: The top-1 test accuracy with different clients sampling ratio on CIFAR10.

Algorithm	$E = 1$	$E = 10$	$E = 20$	$E = 50$
FedAvg	64.19%	70.81%	70.01%	68.40%
FedProx	64.63%	71.52%	72.14%	72.28%
FedLSD	63.96%	72.05%	72.08%	72.35%
FedCAD	**65.10%**	**73.14%**	**72.63%**	**72.51%**

Table 4: The top-1 test accuracy with different number of local epochs on CIFAR10.

The results are shown in Table 4. Intuitively, a small E may increase communication burden and a large E may result in a low convergence rate. Under the different settings of E, the accuracy of our method FedCAD exceeds other state-of-the-art methods. It is also worth noticing that with local epoch growing, the improvement margin gained from FedCAD increased. The results indicate that FedCAD can effectively improve the collaborative training on non-IID distributed data under different settings of E, especially when each client get trained for more epochs in a round.

Conclusion

In this work, we propose a Federated Learning algorithm with a class-wise adaptive mechanism to control the impact of distillation loss (FedCAD). The proposed class-wise adaptive weight is determined by the performance of the global model on different classes of an auxiliary dataset. The adaptive distillation loss prevents local models from forgetting the global knowledge while avoids the misleading problem of incorrect distillation, especially when the global model is not reliable enough as a teacher model. We conduct extensive experiments on the benchmark datasets to demonstrate that our method is robust and achieves significant improvement over state-of-the-art Federated Learning methods. As FedCAD does not require the inputs to be images, it can be applied to non-vision problems in the future.

Acknowledgments

This work is supported by Key-Area Research and Development Program of Guangdong Province (No.2019B010109001), Natural Science Foundation of China (No.61972383, No.61902377) and Jinan S&T Bureau (No.2020GXRC030).

References

Caldas, S.; Duddu, S. M. K.; Wu, P.; Li, T.; Konečnỳ, J.; McMahan, H. B.; Smith, V.; and Talwalkar, A. 2018. Leaf: A benchmark for federated settings. arXiv:1812.01097.

Chang, H.; Shejwalkar, V.; Shokri, R.; and Houmansadr, A. 2019. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer.

Chen, Y.; Lu, W.; Wang, J.; and Qin, X. 2021. FedHealth 2: Weighted Federated Transfer Learning via Batch Normalization for Personalized Healthcare.

Deng, Y.; Kamani, M. M.; and Mahdavi, M. 2020. Adaptive personalized federated learning.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network.

Huang, Y.; Chu, L.; Zhou, Z.; Wang, L.; Liu, J.; Pei, J.; and Zhang, Y. 2021. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7865–7873.

Jeong, E.; Oh, S.; Kim, H.; Park, J.; Bennis, M.; and Kim, S.-L. 2018. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2019. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 5132–5143. PMLR.

Khaled, A.; Mishchenko, K.; and Richtárik, P. 2020. Tighter Theory for Local SGD on Identical and Heterogeneous Data. In Chiappa, S.; and Calandra, R., eds., *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, 4519–4529. PMLR.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lee, G.; Shin, Y.; Jeong, M.; and Yun, S.-Y. 2021. Preservation of the Global Knowledge by Not-True Self Knowledge Distillation in Federated Learning. arXiv:2106.03097.

Li, D.; and Wang, J. 2019. Fedmd: Heterogenous federated learning via model distillation.

Li, Q.; Diao, Y.; Chen, Q.; and He, B. 2021a. Federated learning on non-iid data silos: An experimental study.

Li, Q.; He, B.; and Song, D. 2021. Model-Contrastive Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10713–10722.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020a. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3): 50–60.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020b. Federated Optimization in Heterogeneous Networks. In *Proceedings of Machine Learning and Systems*, volume 2, 429–450.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2020c. On the Convergence of FedAvg on Non-IID Data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Li, X.; Jiang, M.; Zhang, X.; Kamp, M.; and Dou, Q. 2021b. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble Distillation for Robust Model Fusion in Federated Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Lukasik, M.; Bhojanapalli, S.; Menon, A. K.; and Kumar, S. 2021. Teacher's pet: understanding and mitigating biases in distillation. arXiv:2106.10494.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037.

Seo, H.; Park, J.; Oh, S.; Bennis, M.; and Kim, S. 2020. Federated Knowledge Distillation.

Shoham, N.; Avidor, T.; Keren, A.; Israel, N.; Benditkis, D.; Mor-Yosef, L.; and Zeitak, I. 2019. Overcoming forgetting in federated learning on non-iid data.

Sun, S.; Cheng, Y.; Gan, Z.; and Liu, J. 2019. Patient Knowledge Distillation for BERT Model Compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 4322–4331. Association for Computational Linguistics.

Wang, H.; Yurochkin, M.; Sun, Y.; Papailiopoulos, D. S.; and Khazaeni, Y. 2020a. Federated Learning with Matched Averaging. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; and Poor, H. V. 2020b. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wang, L.; Xu, S.; Wang, X.; and Zhu, Q. 2021. Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10165–10173.

Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020c. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yao, D.; Pan, W.; Dai, Y.; Wan, Y.; Ding, X.; Jin, H.; Xu, Z.; and Sun, L. 2021. Local-Global Knowledge Distillation in Heterogeneous Federated Learning with Non-IID Data. arXiv:2107.00051.

Yao, X.; and Sun, L. 2020. Continual Local Training for Better Initialization of Federated Models. In *2020 IEEE International Conference on Image Processing (ICIP)*, 1736–1740. IEEE.

Yurochkin, M.; Agarwal, M.; Ghosh, S.; Greenewald, K.; Hoang, N.; and Khazaeni, Y. 2019. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, 7252–7261. PMLR.

Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; and Chandra, V. 2018. Federated learning with non-iid data. arXiv:1806.00582.