# Class-Wise Adaptive Self Distillation for Federated Learning on Non-IID Data (Student Abstract)

**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

Federated learning (FL) enables multiple clients to collaboratively train a globally generalized model while keeping local data decentralized. A key challenge in FL is to handle the heterogeneity of data distributions among clients. The local model will shift the global feature when fitting local data, which results in forgetting the global knowledge. Following the idea of knowledge distillation, the global model's prediction can be utilized to help local models preserve the global knowledge in FL. However, when the global model hasn't converged completely, its predictions tend to be less reliable on certain classes, which may results in distillation's misleading of local models. In this paper, we propose a class-wise adaptive self distillation (FedCAD) mechanism to ameliorate this problem. We design class-wise adaptive terms to soften the influence of distillation loss according to the global model's performance on each class and therefore avoid the misleading. Experiments show that our method outperforms other state-of-the-art FL algorithms on benchmark datasets.

## Introduction

Federated learning (FL) is a distributed machine learning paradigm which enables multiple clients to collaboratively train a generalized and robust model while keeping local data private. A key challenge in federated learning is that data distribution in different clients is usually non-identically distributed (non-IID). It has been shown that popular federated methods such as FedAvg (McMahan et al. 2017) suffers from performance degradation or even diverges when deployed over non-IID samples. When each client trains a local model on its own data, its local objective may deviate from the global objective and results in forgetting global knowledge. Many studies have tried to address the non-IID issue at the local training stage. FedProx (Li et al. 2020) directly limits the local updates by adding an additional L2 regularization term to the local objective function. 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, FedLSD (Lee et al. 2021) trains local models with the guidance of global model's prediction on local data to preserve global knowledge. However, distilled knowledge on the classes where the global model is inherent inaccurate (Lukasik et al. 2021) will mislead local models and hence causes the degradation of the final generalized model.

In this paper, we propose a class-wise adaptive self distillation mechanism for FL (FedCAD), which applies class-wise weights on distillation loss to tackle down the misleading problem. Through the adaptive controlling of the distillation loss, local models learn more knowledge from the distilled knowledge when the prediction of global model on the class is credible. For the classes on which global model performs poorly, local models will focus more on local data.



## Approach

In this section, we introduce the details of FedCAD. Assume that there are $N$ clients in the Federated learning framework, holding local data with distinct distributions $D = \{D_1, D_2, ..., D_N\}$. On each round $t$, the local model is initialized to 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, client $i$ uses the following loss function $L$ which consists of a classification loss $L_c$ and a distillation loss $L_d$.

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

Here, we use the softmax cross entropy as the classification loss $L_c$. The distillation loss is formulated as follows:

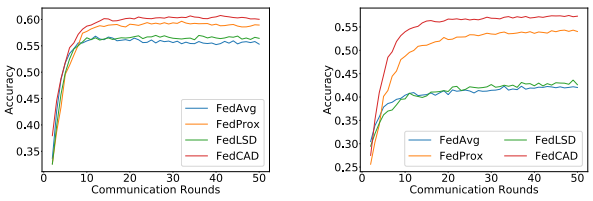
$$L_d = \sum_{x \in D_i} \sum_{k=1}^{K} -p_k^g(x) \log[p_k(x)] \tag{2}$$

where $p$, $p^g$ are the softmax probability of local model and global model. $K$ is the total number of classes.

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

$T$ denotes the temperature scalar. $z$ and $z^g$ are the output logits of local model and global model respectively.

Distillation causes degradation on classes where the teacher is inherently inaccurate (Lukasik et al. 2021). In FL, inaccurate distillation would mislead local models, especially in early stages where the global model hasn't converged. To solve this problem, we introduce the class-wise




(a) Global accuracy (b) Local average accuracy

Figure 1: 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.

adaptive weights $(\alpha_1, ..., \alpha_K) \in [0, 1]^K$ to reduce the impact of distillation for those classes on which the global model performs poorly.

$$\alpha_y = \frac{1}{2}(\gamma - \beta)\mathbb{E}_{x|y}[\phi(y, p^g(x))] + \frac{1}{2}(\gamma + \beta) \quad (3)$$

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

where $0 < \beta < \gamma < 1$ decide the lower and upper bounds of distillation impact. 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. Here, we use the local data to estimate the expectation $\mathbb{E}_{x|y}[\cdot]$.

## Experiments

**Dataset.** We use CIFAR-10 as benchmark dataset in our study and use the Dirichlet distribution to synthesize the non-IID distribution among clients. Specifically, we sample $q_i \sim Dirichlet(\delta)$ and allocate a $q_{i,j}$ proportion of the instances of class i to client j, where $\delta$ is the concentration parameter controlling the uniformity between clients. We set up ten clients and $\delta = 0.5$ by default.

**Hyper-parameters.** We use the same CNN architecture as FedAvg. We use the SGD optimizer with momentum 0.9 and learning rate 0.01. The local epoch is set to $E = 5$ and local batch size is set to $B = 64$ by default. The parameters $\beta$ and $\gamma$ are set to 0.25 and 0.5 respectively. We use FedAvg, FedProx and FedLSD as baselines and validate how FedCAD improves the FL performance.

**Quantitative Evaluation.** Fig 1 shows the test accuracy of all approaches with the above default settings. As we can see, the speed of global accuracy improvement in FedCAD is almost the same as FedAvg at the beginning. Since the global model is far from convergence and the weight of distillation loss is small. Then, it achieves a better accuracy benefit from the class-wise adaptive terms. On the other hand, it is worth noting that the local average accuracy of FedCAD and FedLSD outperforms the other approach, which means adding distillation loss can mitigate the catastrophic forgetting in local training effectively.

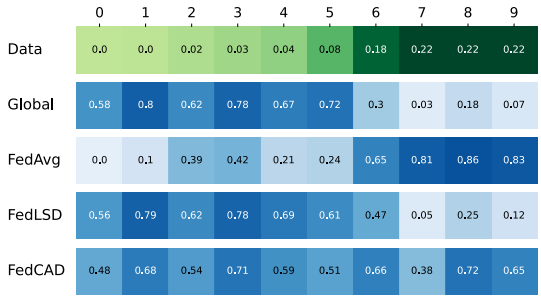



Figure 2: The special confusion matrix of CIFAR-10. We plot the data distribution for a certain client (first row), per-class accuracy of the initial global model (second row) and local model with different algorithms (last three rows).

**Influence of Class-wise Adaptive Distillation.** To demonstrate the effectiveness of FedCAD, we analyze per-class accuracy of the global model and local model. 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 Fig 2, 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 achieves high accuracy on almost all categories, which demonstrates our approach avoids the misleading of the global model and meanwhile extracts reliable knowledge from it.

## 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

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.

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

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.

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