Flashback: Understanding and Mitigating Forgetting in Federated Learning

Mohammed Aljahdali*, Ahmed M. Abdelmoniem†, Marco Canini*, Samuel Horváth‡
*KAUST †Queen Mary University of London ‡MBZUAI

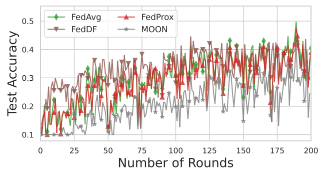
***Abstract*—Federated Learning (FL) addresses the growing need to perform large-scale model training directly on distributed data sources, eliminating the overhead and privacy risks of transferring data to a central location. However, in FL, forgetting (or the loss of knowledge across rounds) hampers algorithm convergence, especially in the presence of severe data heterogeneity among clients. This study explores the nuances of this issue, emphasizing the critical role of forgetting leading to FL's inefficient learning within heterogeneous data contexts. Knowledge loss occurs in both client-local updates and server-side aggregation steps; addressing one without the other fails to mitigate forgetting. We introduce a metric to measure forgetting granularly, ensuring distinct recognition amid new knowledge acquisition. Based on this, we propose Flashback, a novel FL algorithm with a dynamic distillation approach that regularizes the local models and effectively aggregates their knowledge. The results from extensive experimentation across different benchmarks show that Flashback mitigates forgetting and outperforms other state-of-the-art methods, achieving faster round-to-target accuracy by converging in 6 to 16 rounds, being up to 27× faster.**

***Keywords*—Knowledge Forgetting, Distillation, Data Heterogeneity, Federated Learning.**

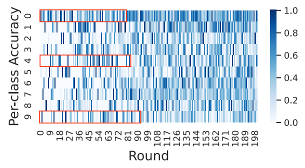
I. Introduction

Federated Learning (FL) is a distributed learning paradigm that allows training over decentralized private data. These datasets belong to different clients that participate in training a global model. Federated Averaging (FedAvg) [1] is a prominent training algorithm that uses a centralized server to orchestrate the process. At every round, the server samples a proportion of the available clients. Starting from the current version of the global model, each sampled client performs E epochs of local training using their private data and sends its updated model to the server. Then, the server aggregates the models by averaging them to obtain the new global model. This process is typically repeated for many rounds until a desired model performance is obtained.

A main challenge in FL is the data heterogeneity in distribution between the clients' private datasets, which are unbalanced and non-IID [2], [3]. Data heterogeneity causes local model updates to drift — the local optima might not be consistent with the global optima – and can lead to slow convergence of the global model — where more rounds of communication and local computation are needed – or worse, the desired performance may not be reached. Slower convergence in FL *increases computational and communication costs, draining device resources, overloading networks, and reducing scalability in resource-constrained environments*. Addressing data heterogeneity in FL has been the focus of several prior studies. For instance, FedProx [4] proposes a proximal term to limit the distance between the global model and the local model updates, mitigating the drift in the local updates. MOON [5] mitigates the local drift using a contrastive loss to minimize the distance between the feature representation of the global model and the local model updates while maximizing the distance between the current model updates and the previous model updates. FedDF [6] addresses heterogeneity in local models by using ensemble distillation during the aggregation step (instead of averaging the model updates). Nonetheless, we experimentally observe that under severe data heterogeneity, these proposals provide little or even no advantage over FedAvg. For instance, Figure 1a illustrates the highly unstable test accuracy values over rounds of FedAvg and other baselines while training a DNN over the CIFAR10 dataset [7] (more details are in § VI).



(a) Global model accuracy of FedAvg and baselines.



(b) Per-class accuracy of FedAvg's global model.

Fig. 1: Performance of FedAvg and other methods over training rounds with CIFAR10.

This motivates us to understand better how data heterogeneity poses a challenge for FL and devise a new approach to handling non-IID datasets. We investigate the evolution of the global model accuracy broken down by its per-class accuracy. Figure 1b shows a heatmap of the per-class accuracy for FedAvg; each rectangle represents the accuracy of the global model on a class at a round. Other baseline methods show similar results. Our key observation is that there is a notable presence of *forgetting*: i.e., cases where some knowledge obtained by the global model at round t is forgotten at round $t + 1$, causing the accuracy to decline (e.g., the prominent number of light-shaded rectangles appearing after darker ones in the figure; we highlight some cases in red in Fig. 1b).

A similar phenomenon is known as *catastrophic forgetting* in Continual Learning (CL) literature [8]. CL addresses the challenge of sequentially training a model on a series of tasks, denoted as $\{T_1, T_2, \ldots, T_n\}$, without revisiting data from prior tasks. Formally, given a model with parameters θ and task-specific loss functions $L_t(\theta)$ for each task T_t, the objective in CL is to update θ such that performance on the current task is optimized without significantly degrading the model's performance on previously learned tasks. This is non-trivial, as naïve sequential training often leads to catastrophic forgetting, where knowledge from prior tasks is overridden when learning a new task. An inherent assumption in this paradigm is that once the model transitions from task T_i to task T_{i+1}, data from T_i becomes inaccessible, amplifying the importance of knowledge retention strategies [9].

While the premises and assumptions of FL differ from those of traditional machine learning and continual learning, forgetting remains an issue. This can be viewed as a side effect of data heterogeneity, a commonality FL shares with CL. In FL, the global model evolves based on a fluctuating data distribution. Specifically, a diverse set of sampled clients with distinct data distributions contribute a model update in each communication round. Furthermore, these model updates must be aggregated to obtain a global model. This situation presents dual-levels of data heterogeneity. Firstly, at the *intra-round* level, heterogeneity arises from the participation of clients with varied data distributions within the same round. This diversity can inadvertently lead to "forgetting" specific data patterns or insights from certain clients. Secondly, at the *inter-round* level, the participating clients generally change from one round to the next. As a result, the global model may "forget" or dilute insights gained from clients in previous rounds.

To remedy this issue, we propose Flashback, a FL algorithm that employs a dynamic distillation approach to mitigate the effects of data heterogeneity. Flashback's dynamic distillation ensures that the local models learn new knowledge while retaining knowledge from the global model during the client updates by adaptively adjusting the distillation loss. Moreover, during the server update, Flashback uses a very small public dataset as a medium to integrate the knowledge from the local models to the global model using the same dynamic distillation. Flashback performs these adaptations by estimating the knowledge in each model using label counts as a proxy of the model knowledge. Overall, Flashback results in a more stable and faster convergence compared to existing methods.

Our contributions are the following:

- We systematically investigate the forgetting problem in FL. We show that under severe data heterogeneity, FL suffers from forgetting. We dissect how and where forgetting happens (§ IV).
- We propose a new metric for measuring forgetting over the communication rounds (§ IV).
- We introduce *Flashback*, a FL algorithm that employs a dynamic distillation during the local updates and the server update (§ V). By addressing the forgetting issue, Flashback mitigates its detrimental effects and converges to the desired accuracy faster than existing methods (§ VI)

II. Background

We consider a standard cross-device FL setup in which there are N clients. Each client i has a unique dataset $D_i = \{(x_j, y_j)\}_{j=1}^{n_i}$ where x_j represents the input features and y_j is the ground-truth label for j-th data point and n_i represent the size of the local dataset of client i. The goal is to train a single global model that minimizes the objective:

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{N} \frac{|D_i|}{|\cup_{i\in[N]} D_i|} \left\{ L_i(w) = \frac{1}{|D_i|} \sum_{j=1}^{|D_i|} l(w; (x_j, y_j)) \right\},$$

where $L_i(w)$ represents the local loss for client i, and $l(w; (x_j, y_j)) = \mathcal{L}_{\text{CE}}(F_w(x), y)$ is the cross-entropy loss for a single data point, where F_w denotes the model parameterized by learnable weights w.

FedAvg provides a structured approach to address this distributed problem efficiently. At each communication round t, the server randomly selects K clients from the total available N clients. These clients (denoted with $\mathbb{S}_t$) receive the previous global model, w_{t-1}. Then, they update this model based on their local data using their local loss function L_i. After updating, each client sends their modified model $w_{k,t}$ back to the server that updates the global model using a weighted average of local models, i.e., $w_t = \sum_{k\in\mathbb{S}_t} \frac{|D_k| w_{k,t}}{|\cup_{k\in[K]} D_k|}$. Various FL algorithms introduce modifications at the local update level or during the global aggregation to accommodate the intrinsic heterogeneity in client data. The nuances of these variations are further explored in § III.

Among these, **Knowledge Distillation (KD)** is a training method wherein a smaller model, referred to as the student, is trained to reproduce the behavior of a more complex model or ensemble called the teacher. Let F_{w_s} denote the student model with weights w_s and F_{w_t} represent the teacher model with weights w_t. For a given input x, the student aims to minimize the following distillation loss:

$$\begin{aligned} \mathcal{L}_{\text{KD}}((x, y); w_s, w_t) = \mathcal{L}_{\text{CE}}(F_{w_s}(x), y)(1-\alpha) \\ +\mathcal{L}_{\text{KL}}(F_{w_t}(x), F_{w_s}(x))\alpha \end{aligned} \quad (1)$$

Here, $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss with true label y, and $\mathcal{L}_{\text{KL}}$ represents the Kullback-Leibler (KL) divergence between the teacher's and the student's output probabilities. It is defined as $\mathcal{L}_{\text{KL}}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{c=1}^{C} p^c \log\left(\frac{p^c}{q^c}\right)$, where C is the number of classes, $\mathbf{p}$ is the target output probability vector, and $\mathbf{q}$ is the predicted output probability vector. The hyperparameter $\alpha \in [0, 1]$ balances the importance between the learning from the true labels and the teacher's outputs.

While distillation originally emerged as a method for model compression [10]–[12], its utility extends to FL. In the federated context, distillation can combat challenges like data heterogeneity [6], [13] and communication efficiency [14].

III. Related Work

Federated learning. FL is commonly viewed as an ML paradigm wherein a server distributes the training process on a set of decentralized participants that train a shared global model using local datasets that are never shared [1], [2], [4], [15]–[17]. FL has been used to enhance prediction quality for virtual keyboards among other applications [18], [19]. A number of FL frameworks have facilitated research in this area [20]–[22].

Heterogeneity in FL. A key challenge in FL systems is uncertainties stemming from learner, system, and data heterogeneity. The non-IID distributions of learners' data can significantly slow down convergence [1]–[3] and several algorithms are proposed as means of mitigation [4], [23]–[26].

Forgetting in FL. Forgetting in FL has been explored in several studies, though many have limitations in addressing the full scope of the issue. Luo et al. [27] discuss forgetting due to local updates. Similarly, [28] tackles the problem of learning personalized models without forgetting what the global model has learned by using knowledge distillation. However, these approaches focus on the local update without addressing forgetting at the aggregation step. On the other hand, [29] focuses on domain shifts and clients with different data domains, aiming for personalized models rather than a global model. [30] investigate the convergence behavior and forgetting of Transformers compared to other architectures used in FL. Their experiments show that transformers are robust to data heterogeneity. While these works address client heterogeneity, they do not delve into the forgetting issue in FL.

A. Prior Attempts of Mitigating Forgetting

This part delves deeper into the main forgetting baselines we compare with [13], [31].

FedReg [31] addresses the issue of slow convergence in FL, asserting it to be a result of forgetting at the local update phase. They demonstrate this by comparing the loss of the global model w_{t-1} on specific client data points with the averaged loss of updated clients' models $\{w_{t,k} \mid k \in \mathbb{S}_t\}$ on the same data points, highlighting a significant increase in the average loss, indicative of forgetting. However, our work proposes a systematic way of measuring forgetting using a metric designed to capture it. Furthermore, we show that forgetting doesn't only occur in the local update and at the aggregation step (§ IV and Fig. 2). FedReg proposes to generate fake data that carries the previously attained knowledge. During the local update, Fast Gradient Sign Method [32] is used to generate these data using the global model w_{t-1} and the client data. Then, the loss of the generated data is used to regularize the local update. While FedReg employs regularization using synthetic data during local updates, our work, Flashback, leverages dynamic distillation to ensure knowledge retention at both local updates and aggregation steps.

FedNTD [13] makes a connection between CL and FL, suggesting that forgetting happens in FL as well. Similarly to FedReg, their analysis shows that forgetting happens at the local update, where global knowledge that lies outside of the local distribution of the client is susceptible to forgetting. To address this, they propose to use a new variant of distillation Eq. (1) named Not-True Distillation (NTD), that masks the ground-truth class logits in the KL divergence as $\mathcal{L}_{\text{KL}}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=c,c\neq y}^{C} p^c \log(\frac{p^c}{q^c})$, where y is the ground-truth class. NTD is used at the local update, while all the other steps in the algorithm remain the same as FedAvg. FedNTD aims to preserve global knowledge during the local update.

Both FedReg and FedNTD diagnose the issue of forgetting primarily within the realm of local updates, asserting that this stage risks losing valuable global knowledge. Consequently, both works present innovative solutions specifically tailored to counteract this local update forgetting. However, their perspective overlooks a pivotal aspect of the forgetting problem: the occurrence of forgetting during the aggregation step. As we delve next into forgetting in § IV, this oversight in recognizing and addressing forgetting during aggregation has repercussions on the later local updates. Moreover, in the prior works, there is no detailed investigation or exploration of what forgetting in FL entails. In this work, we fill this gap and provide a detailed analysis of forgetting in FL, demonstrating where and why it happens, and propose a metric to measure forgetting. Unlike previous works, which focus on client-side forgetting, we propose Flashback, which addresses forgetting as a compound problem that occurs both at the local update and the aggregation step, suggesting it must be tackled at both levels.

IV. Problem - Forgetting in FL

We now investigate where forgetting happens and devise a metric to quantify this phenomenon. Recall that in FL, the models are updated in two distinct phases: 1) during local training – when each client k starts from global model w_{t-1} and locally train $w_{k,t}$ – and 2) during the aggregation step – when the server combines the client models to update the new global model w_t.

Intuitively, forgetting in FL is when knowledge contained in the global model will be lost after the completion of communication round $w_{t-1} \rightarrow w_t$. We observe that forgetting may occur in the two phases of FL. We refer to the former case as **local forgetting**, where some knowledge in the global model will be lost during the local training $w_{t-1} \rightarrow w_{k,t}$. This is due to optimizing for the clients' local objectives, which depend on their datasets. Local forgetting is akin to the form of forgetting seen in CL, where tasks change over time (as with clients in FL) and, consequently, the data distribution. We refer to the latter case as **aggregation forgetting**, where some knowledge contained in the clients' model updates will be lost during aggregation $\sum\{w_{k,t} \mid k \in \mathbb{S}_t\} \rightarrow w_t$. This might be due to the coordinate-wise aggregation of weights as opposed to matched averaging in the parameter space of DNNs [33].

We illustrate forgetting in Fig. 2 based on actual experiments with several baseline methods. Given round t, the figure shows the per-class accuracy of the global model w_{t-1}, all local models $w_{k,t}$, and the new global model w_t for four different methods. The local forgetting is evident in the drop in accuracy (lighter shade of blue) of the local models $w_{k,t}$ compared to the global model w_{t-1}. The aggregation forgetting is evident in the drop in accuracy of the global model w_t compared to the local models $w_{k,t}$. The figure also previews a result of our method, Flashback, which significantly mitigates forgetting. In summary, local and aggregation forgetting lead to the main forgetting

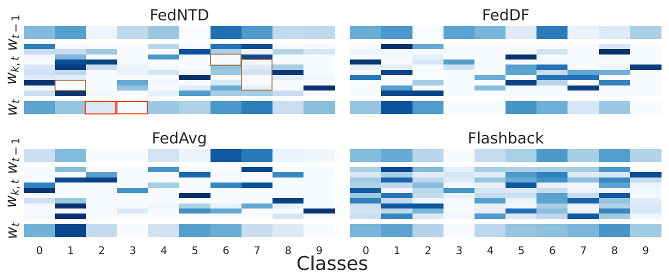


Fig. 2: Local (client) & aggregation forgetting in some of the baselines using CIFAR10. The first row represents the global model per-class test accuracy at round $t-1$; then, the rows in the middle are the clients that participated in round t, and finally, in the last row, the global model at the end of round t. Local forgetting happens when clients at round t lose the knowledge that the global model had at round $t-1$ (example highlighted in brown). The aggregation forgetting happens when the global model at round t loses the knowledge that in the clients' models at round t (example highlighted in red).

problem in FL, which we term both as **round forgetting**, affecting $w_{t-1} \to w_t$.

In CL, forgetting is often quantified using Backward Transfer (BwT) [34]. FedNTD [13] adapted this metric, i.e., the forgetting score $\mathcal{F}$, for FL as follows:

$$\mathcal{F} = \frac{1}{C}\sum_{c=1}^{C} \arg\max_{t\in 1,T-1}(A_t^c - A_T^c), \tag{2}$$

where C is the number of classes and A_t^c is the global model accuracy on class c at round t.

However, $\mathcal{F}$ is a coarse-grain score that evaluates forgetting in aggregate across all rounds. We seek a finer-grain metric that measures forgetting round-by-round. Furthermore, we wish to account for knowledge replacement scenarios, such as when a decline in accuracy for one class might be accompanied by an increase in another, essentially masking the negative impact of forgetting in aggregate measures. Thus, for our evaluation results (§ VI), we propose to measure **round forgetting** by focusing only on drops in accuracy using the following metric:

$$\mathcal{F}_t = -\frac{1}{C}\sum_{c=1}^{C} \min(0, (A_t^c - A_{t-1}^c))$$

where $t > 1$ is the round at which forgetting is measured.

Our metric accounts for the pitfalls of the previously proposed forgetting metric. It discounts knowledge replacement scenarios that can happen between rounds by only focusing on the negative changes in accuracy. Furthermore, it provides a granular view of forgetting because it measures *round forgetting* (whereas Eq. (2) measures the global model forgetting at the end of training).

V. Flashback: Forgetting-Robust FL

Our key idea to mitigate *round forgetting* is to leverage a dynamic form of knowledge distillation, which is fine-tuned in response to the evolving knowledge captured by the different models in the training process. During local training, distillation ensures that each local model learns from the client's local dataset while retaining knowledge from the current global model. On the server side, after the clients' updates, Flashback begins by aggregating the locally updated models—much in the vein of FedAvg. Then, Flashback distills the knowledge of the freshly updated global model using our dynamic distillation approach, learning from both its immediate predecessor—the global model obtained at the previous round—and the ensemble of locally updated models, which all play the role of teachers. The Flashback algorithm is detailed in Algorithm 1. The remainder of this section discusses our distillation approach in more detail.

Algorithm 1 Flashback algorithm.

```
input Initial global model $w_0$, number of rounds $T$, fraction of clients $R$, minibatch size $B$, number of local epochs $E$, number of server epochs $E_s$, learning rate $\eta$
output global model $w_T$
 1: $\boldsymbol{\pi} = \mathbf{0} \in \mathbb{R}^C$ // Global model's label count vector
 2: for $t = 1$ to $T$ do
 3:     $\mathbb{S}_t \leftarrow$ Randomly select $\lceil R \cdot N \rceil$ clients
 4:     for each client $k \in \mathbb{S}_t$ do
 5:         $w_{k,t} \leftarrow w_{t-1}$ // Initialize local model with current global model
 6:         Compute $\boldsymbol{\alpha}$ with $\boldsymbol{\nu}$ as the local label count and a single teacher $\boldsymbol{\mu} \leftarrow \boldsymbol{\pi}$
 7:         Update $w_{k,t}$ using dKD loss $\mathcal{L}_{\text{dKD}}$ for $E$ epochs
 8:     end for
 9:     $m_t \leftarrow \sum_{k\in\mathbb{S}_t} n_k$ // Total data points in this round
10:     $w_t \leftarrow \sum_{k\in\mathbb{S}_t} \frac{n_k}{m_t} w_{k,t}$ // Average to obtain the new global model
11:     $\mathbb{T} \leftarrow \{w_{k,t} \mid k \in \mathbb{S}_t\} \cup \{w_{t-1}\}$
12:     Compute $\boldsymbol{\alpha}$ with $\boldsymbol{\nu} \leftarrow \boldsymbol{\pi}$ and $\boldsymbol{\mu}_i$ as the label count $\forall w_i \in \mathbb{T}$
13:     Update $w_t$ using dKD loss $\mathcal{L}_{\text{dKD}}$ for $E_s$ epochs
14:     $r_k \leftarrow$ (Increment $r_k$ for every client $k \in \mathbb{S}_t$)
15:     for each client $k \in \mathbb{S}_t$ do
16:         if $\gamma r_k \leq 1$ then
17:             $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi} + \gamma \boldsymbol{\mu}_k$ // Update participation count for client $k \in \mathbb{S}_t$
18:         end if
19:     end for
20: end for
```

A. Dynamic Distillation

As established in § IV, a client's local model can forget and override model knowledge with what is present in its private data. Moreover, the global model can be imperfect for two reasons: i) As we established before, the global model is susceptible to forgetting in the aggregation step. ii) Assuming no forgetting in the aggregation step, the knowledge contained in the clients who participated so far might not represent all the available knowledge, especially in the early rounds. Overall, both local models and the global model can be imperfect. Therefore, the logits of all the different classes cannot be

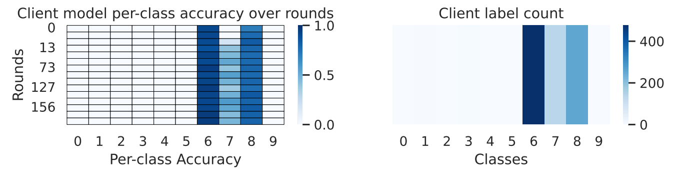


Fig. 3: (left) Per-class accuracy of a client model on all the rounds where it participated. (right) Data distribution.

treated equally (as in Eq. (1)), and the distillation loss has to adapt to the model's knowledge.

We propose using the label count to approximate the knowledge within a model. Here, the label count refers to the occurrences of each class in the training data that the model saw during training. In machine learning, a model's knowledge is fundamentally tied to the data it has been exposed to. If certain classes have higher representation (or label counts) in the training data, it's intuitive that the model would have more opportunities to learn the distinguishing features of such classes. Conversely, underrepresented classes might not offer the model sufficient exposure to learn their nuances effectively.

Our experimental results suggest that per-class performance on the test set correlates highly with the label counts in the training data. In scenarios where certain classes were more abundant, the model demonstrated higher proficiency in predicting those classes on the test set. As an example, Fig. 3 illustrates for a randomly chosen client that the client's model performance on the test set well reflects the label count distribution of its private data. We conclude from this and many similar observations that the label count can indicate a model's knowledge.

In standard knowledge distillation (Eq. (1)), all logits are treated equally since it is assumed that the teacher model has been trained on a balanced dataset. Owing to the heterogeneity of data distribution in local datasets, this assumption does not hold in FL. As a result, we cannot directly treat the current global model nor the local model updates as equally reliable teachers across all classes. Instead, we propose weighting the logits using the label count to approximate the per-class knowledge within a model.

We now revisit the distillation loss in Eq. (1) and transform the scalar α to a matrix form that is automatically tuned according to the label count of both the student and the teachers and used directly within the KL divergence loss. Namely, the dynamic $\boldsymbol{\alpha}$ parameter (defined below) will change during the training as the label counts change. Flashback maintains the global model counts over the rounds; this mechanism is detailed in the next section.

We consider a single student model F_{w_s} with weights w_s and a set $\mathbb{T}$ of K teacher models; the i-th teacher model is denoted as F_{w_i} with weights w_i. Let $\boldsymbol{\nu} \in \mathbb{R}^C$ be the relative label count vector of the student model, where ν^c is the relative occurrences of class c in the dataset. Similarly, let $\boldsymbol{\mu}_i \in \mathbb{R}^C$ be the relative label count vector of the i-th teacher model.

The dynamic $\boldsymbol{\alpha} \in [0,1]^{K \times C}$ is defined as $[\boldsymbol{\alpha}_1^\intercal, \ldots, \boldsymbol{\alpha}_K^\intercal]$, with $\alpha_i^c = \frac{\mu_i^c}{\nu^c + \sum_k \mu_k^c}$.

Then, we embed $\boldsymbol{\alpha}$ directly in the KL divergence loss ($\mathcal{L}_{\text{KL}}$ in Eq. (1)) as follows:

$$\mathcal{L}_{\text{dKL}}(\boldsymbol{p}, \boldsymbol{q}; \boldsymbol{\alpha}_i) = \sum_{c=1}^{C} \alpha_i^c \cdot p^c \log\left(\frac{p^c}{q^c}\right)$$

Similar to standard distillation, to account for the student model knowledge with respect to the ground-truth class y, we define $\alpha_s^c = \frac{\nu^c}{\nu^c + \sum_k \mu_k^c}$. Thus, $\alpha_s^c + \sum_{k=1}^{K} \alpha_k^c = 1$ for all classes $c \in [C]$.

Finally, the dynamic knowledge distillation loss ($\mathcal{L}_{\text{dKD}}$) is:

$$\begin{aligned}\mathcal{L}_{\text{dKD}}((x,y); w_s, \mathbb{T}, \boldsymbol{\alpha}) &= \alpha_s^y \mathcal{L}_{\text{CE}}(F_{w_s}(x), y) \\ &+ \sum_{w_i \in \mathbb{T}} \mathcal{L}_{\text{dKL}}(F_{w_i}(x), F_{w_s}(x); \boldsymbol{\alpha}_i)\end{aligned} \tag{3}$$

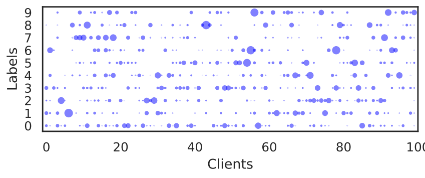
The dynamic $\boldsymbol{\alpha}$ will weigh the divergence between the logits of different classes, making the student model focus more on learning from the teacher's strengths while being cautious of its weaknesses. This is important in FL because of the data heterogeneity problem. For instance, the global model may not encounter certain classes in the initial training rounds. Our distillation approach assigns a zero weight to the divergence of these classes, shielding the client model from adopting unreliable knowledge from the global model. Similarly, if a client possesses significantly larger data for a specific class than the global model has encountered, the weight assigned to that class's divergence will be small. This implies that the client's model remains more grounded in classes where it has more comprehensive data.

An interesting property of our distillation is that it will ignore the global model as a teacher in the first communication round. Since the initial global model label counts are all zeros, Eq. (3) reduces to just the cross-entropy: $\mathcal{L}_{\text{dKD}} = \mathcal{L}_{\text{CE}}$.

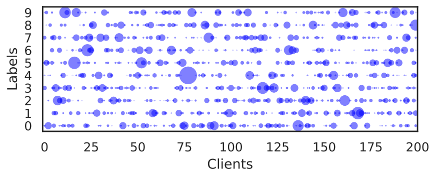
B. Estimating the Global Model Knowledge

Note that to apply the dynamic distillation loss Eq. (3), we must obtain the student's and teachers' label count vectors. While the label count of local models can be easily obtained (from the class frequency of local datasets), the label count of the global model is not readily available. We construct $\boldsymbol{\pi}$, the global model's relative label count, as follows. Let r_k denote the number of rounds in which client k participated in the training. For every client k that participates at round t, Flashback adds a fraction $\gamma \in (0,1]$ of client k's label count ($\boldsymbol{\mu}_k$) to $\boldsymbol{\pi}$, unless $\gamma r_k > 1$, in which case $\boldsymbol{\pi}$ is not updated based on k's label count. The latter case means that client k has participated enough times that its label count is fully accounted for in $\boldsymbol{\pi}$.

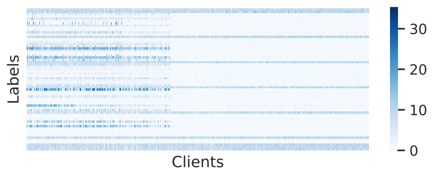
Intuitively, the parameter γ indicates the rate at which we rely on the global model. When γ is set to 1, it implies complete trust in the global model's ability to incorporate the clients' knowledge after just one round of participation. However, expecting such immediate and full assimilation is unrealistic, so we typically set $\gamma < 1$. The gradual build-up of the global label count plays a vital role in maintaining a balanced distillation in



(a) CIFAR10, Dir($\beta = 0.1$)



(b) CINIC10, Dir($\beta = 0.1$)



(c) FEMNIST with 3432 clients

Fig. 4: Clients data distribution. The x-axis is the clients, and the y-axis is the labels.

Eq. (3) during local updates. This progressive approach mirrors increasing trust in the global model's capabilities. It prevents the risk of assigning excessively high weights too soon, which could otherwise hurt the learning process.

VI. Experiments & Results

We outline and analyze our experimental findings to investigate whether Flashback's forgetting-robust FL method addresses the slow and unstable convergence issues due to forgetting problems as laid out in § I.

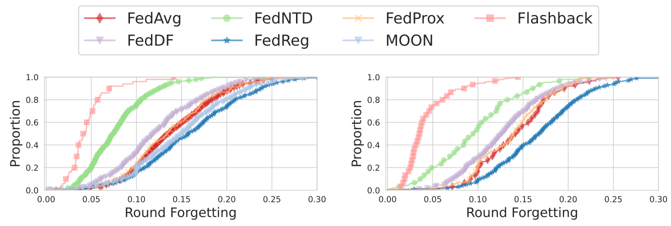
The experimental results stem from three settings: CIFAR10 [7] and CINIC10 [35], where heterogeneous data partitions are created using Dirichlet distribution with $\beta = 0.1$ and FEMNIST [20] with 3,432 clients, following the natural heterogeneity of the dataset. Furthermore, we do an ablation study on the different components of the algorithm. We use the same neural network architecture used in [1], [13], a 2-layer Convolutional Neural Network (CNN). Summaries of the datasets, partitions, and more details on the experimental setup are detailed next.

A. Experimental Setup

Datasets. We provide an overview of the datasets used, the data split, and the specific experimental setups. For each dataset, we perform two sets of experiments to analyze the effects of data heterogeneity on the algorithms' performance. The datasets used are CIFAR10, CINIC10, and FEMNIST. The training data distribution among clients of these datasets is shown in Fig. 4.

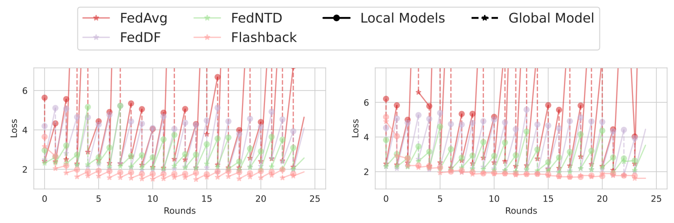
CIFAR10 [7]. A famous vision dataset that includes 50k training images and 10k testing images. We emulate a realistic, heterogeneous data distribution by using a Dirichlet distribution with $\beta = 0.1$. A β value of 0.1 is chosen to simulate a more heterogeneous and challenging data distribution. A 2.5% random sample of the training set creates a public dataset, further divided into training and validation sets. This yields a very small public training dataset with the size of 1.88% of the whole dataset. The remaining 97.5% of the dataset is distributed among 100 clients, with each client's data being split into training (90%) and validation (10%) subsets.

CINIC10 [35]. A drop-in replacement of CIFAR10, this dataset has 90k training, 90k validation, and 90k test images. We merge the training and validation sets and adopt a similar approach as with CIFAR10, taking out 2.5% of the data to be the public dataset; similar to the CIFAR10 case, this 2.5% is further divided into training set and validation set. Then, employing Dirichlet distribution with β value of 0.1 to split



(a) CIFAR10. **(b)** CINIC10.

Fig. 5: Round forgetting distribution of Flashback vs baselines.



(a) CIFAR10. **(b)** CINIC10.

Fig. 6: Local models loss transition to the global model loss. This plot compares the loss of local models and the global model, showing the difference in knowledge before and after the global update.

the 97.5% remaining data into 200 clients, with each client's data divided into training (90%) and validation (10%) sets.

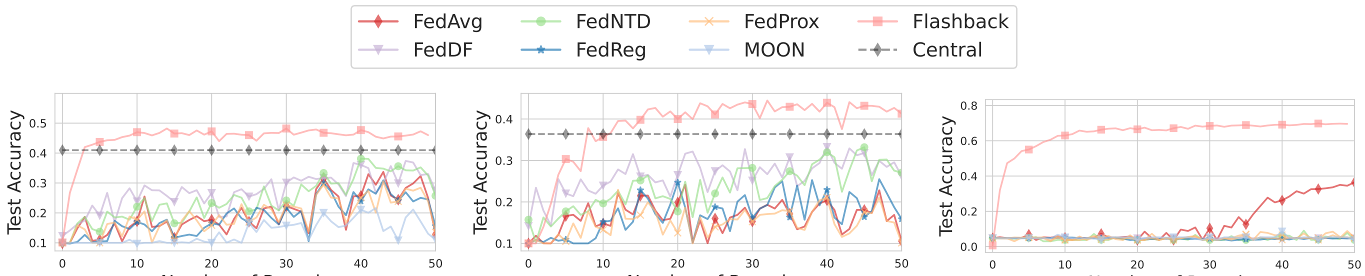
FEMNIST [20]. This federated learning dataset is based on extended MNIST with natural heterogeneity, where each writer is considered a client. From the 3597 total writers, those with less than 50 samples are excluded. We randomly selected 150 writers to form a public dataset. The remaining 3432 writers' data is divided into train (approx. 70%), validation (approx. 15%), and test (approx. 15%) sets. The collective test sets from all writers form the overall test set. At every round, 32 clients are randomly selected for participation.

For CIFAR10 and CINIC10, we chose β values of 0.1 and client participation value of 10. For FEMNIST, we have 3432 clients with a client participation value of 32.

B. Baselines & Hyperparameters

We evaluate the following algorithms as state-of-the-art baselines: 1) FedAvg [1]; 2) FedDF [6]; 3) FedNTD [13]; 4) FedProx [4]; 5) FedReg [31]; and 6) MOON [5] .

Both FedNTD and FedReg target forgetting in FL (discussed in § III). We use the same neural network architecture that is used in [1], [13], which is a 2-layer CNN. Note that for



(a) CIFAR10. **(b) CINIC10.** **(c) FEMNIST.**

Fig. 7: Round-to-accuracy performance of Flashback and other baselines over training rounds.

TABLE I: Number of rounds to accuracy $A_x = A \cdot x$ where A is the target accuracy and x is a fraction.

	CIFAR10, $A = 48.2\%$			CINIC10, $A = 43.5\%$			FEMNIST, $A = 69.5\%$		
	$A_{0.5}$	$A_{0.75}$	$A_{0.95}$	$A_{0.5}$	$A_{0.75}$	$A_{0.95}$	$A_{0.5}$	$A_{0.75}$	$A_{0.95}$
FedAvg	12	82	-	13	-	-	49	75	138
FedDF	7	40	112	**2**	30	-	-	-	-
FedNTD	12	41	112	13	46	-	-	-	-
FedProx	35	93	-	13	-	-	142	-	-
FedReg	35	108	-	16	-	-	-	-	-
MOON	82	-	-	124	-	-	-	-	-
Flashback	**2**	**4**	**10**	4	**5**	**6**	**3**	**5**	**16**

MOON [5], we add an additional layer to the model as described in their source code for the projection head. Moreover, for the optimizer, learning rate, and model, we follow [1], [13], and when a baseline has different hyperparameters, we use their proposed values. For example, in FedDF, the number of local epochs is set to 40, while in the other baselines and Flashback, it is set to 5 epochs. As for Flashback hyperparameters, during the server distillation, we train until early stopping gets triggered using the validation set; we set the label count fraction $\gamma = 0.025$ for CIFAR10, i.e., we add 2.5% of the client label count each time it participates, while we set $\gamma = 0.1$ for CINIC10 and FEMNIST. As for distillation-specific hyperparameters, we have one fewer hyperparameter since $\boldsymbol{\alpha}$ is computed automatically, and for temperature, we use the standard $T = 3$.

C. Results

We evaluate Flashback performance by showing round-to-accuracy, round forgetting, and the local-global loss.

Reduced round forgetting. We show the round forgetting's empirical cumulative distribution function (ECDF) in Fig. 5. We see that Flashback successfully reduces round forgetting. The reduction in forgetting indicates that Flashback is learning efficiently and losing less acquired knowledge as the rounds progress. Also, FedNTD has less round forgetting than the remaining baselines, showing that the algorithm does help alleviate forgetting.

Minimizing local models' divergence. To further understand the effect of Flashback on the training behavior, we show the transition of the mean loss of the local models to the loss of the global model over the rounds in Fig. 6. This gives us an insight into the effect of the regularization made by our dynamic distillation. We see that the mean loss of the local models of the other baselines always spikes, signifying a divergence of these models from the global training objective, while Flashback has a much more stable loss. This shows that Eq. (3) regularizes the local models well such that they do not diverge too much from the global training objective.

Improved round-to-accuracy. We evaluate the learning efficiency of Flashback and other baselines by showing the accuracy over rounds in Fig. 7; we include the result of central training on the public dataset. Flashback consistently shows faster convergence and high accuracy. Furthermore, we show the number of rounds it takes to reach a target accuracy and fractions of that target accuracy in Table I.

VII. Conclusion

We explored the phenomenon of forgetting in FL. Our investigation revealed that forgetting occurs during both the local update and the aggregation step of FL algorithms. We presented Flashback, a novel FL algorithm explicitly designed to counteract round forgetting by employing dynamic knowledge distillation. Our approach leverages data label counts as a proxy for knowledge, ensuring a more targeted and effective forgetting mitigation. Our empirical results showed Flashback's efficacy in mitigating round forgetting, thereby supporting the hypothesis that the observed slow and unstable convergence in FL algorithms is closely linked to forgetting. This result underlines the importance of addressing forgetting, paving the way for advancing more robust and efficient FL algorithms. This is very important as the learning efficiency of the algorithm dictates its resource usage.

As part of our future work, we aim to investigate forgetting under other types of heterogeneity. Moreover, Flashback mainly uses label count to approximate knowledge, which could pose privacy challenges in certain scenarios.As for computational overhead, Flashback adds one additional forward pass per iteration in the local update, similar to FedNTD [13] and less than MOON [5], which adds two forward passes.

Acknowledgments

This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Research Administration (ORA) under Award No. ORA-CRG2021-4699.

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *AISTATS*, 2017.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao, "Advances and Open Problems in Federated Learning," *Foundations and Trends® in Machine Learning*, vol. 14, 2021.

[3] A. M. Abdelmoniem, C.-Y. Ho, P. Papageorgiou, and M. Canini, "A Comprehensive Empirical Study of Heterogeneity in Federated Learning," *IEEE Internet of Things Journal*, 2023.

[4] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated Optimization in Heterogeneous Networks," in *MLSys*, 2020.

[5] Q. Li, B. He, and D. Song, "Model-Contrastive Federated Learning," in *CVPR*, 2021.

[6] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble Distillation for Robust Model Fusion in Federated Learning," in *NeurIPS*, 2020.

[7] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," University of Toronto, Tech. Rep., 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html

[8] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, 2019.

[9] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A Continual Learning Survey: Defying Forgetting in Classification Tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, 2021.

[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," arXiv:1503.02531, 2015.

[11] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *KDD*, 2006.

[12] J. Schmidhuber, *Neural sequence chunkers*. Inst. für Informatik, 1991.

[13] G. Lee, M. Jeong, Y. Shin, S. Bae, and S.-Y. Yun, "Preservation of the Global Knowledge by Not-True Distillation in Federated Learning," in *NeurIPS*, 2022.

[14] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Communication-Efficient On-Device Machine Learning: Federated Distillation and Augmentation under Non-IID Private Data," arXiv:1811.11479, 2018.

[15] J. Konečný, B. McMahan, and D. Ramage, "Federated Optimization:Distributed Optimization Beyond the Datacenter," arXiv:1511.03575, 2015.

[16] R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," in *CCS*, 2015.

[17] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated Learning: Strategies for Improving Communication Efficiency," arXiv:1610.05492, 2016.

[18] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, H. B. McMahan, T. V. Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards Federated Learning at Scale: System Design," in *MLSys*, 2019.

[19] T. Yang, G. Andrew, H. Eichner, H. Sun, W. Li, N. Kong, D. Ramage, and F. Beaufays, "Applied Federated Learning: Improving Google Keyboard Query Suggestions," arXiv:1812.02903, 2018.

[20] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar, "LEAF: A Benchmark for Federated Settings," arXiv:1812.01097, 2019.

[21] tensorflow.org, "Tensorflow federated: Machine learning on decentralized data," 2020. [Online]. Available: https://www.tensorflow.org/federated

[22] A. M. Abdelmoniem, A. N. Sahu, M. Canini, and S. A. Fahmy, "REFL: Resource-Efficient Federated Learning," in *EuroSys*, 2023.

[23] F. Fourati, S. Kharrat, V. Aggarwal, M.-S. Alouini, and M. Canini, "FilFL: Client Filtering for Optimized Client Participation in Federated Learning," in *ECAI*, 2024.

[24] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization," in *NeurIPS*, 2020.

[25] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic Controlled Averaging for Federated Learning," in *ICML*, 2020.

[26] L. Li, M. Duan, D. Liu, Y. Zhang, A. Ren, X. Chen, Y. Tan, and C. Wang, "FedSAE: A Novel Self-Adaptive Federated Learning Framework in Heterogeneous Systems," in *IJCNN*, 2021.

[27] K. Luo, X. Li, Y. Lan, and M. Gao, "GradMA: A Gradient-Memory-based Accelerated Federated Learning with Alleviated Catastrophic Forgetting," in *CVPR*, 2023.

[28] S. Liu, X. Feng, and H. Zheng, "Overcoming Forgetting in Local Adaptation of Federated Learning Model," in *PAKDD*, 2022.

[29] W. Huang, M. Ye, and B. Du, "Learn from others and be yourself in heterogeneous federated learning," in *Conference on Computer Vision and Pattern Recognition*, 2022.

[30] L. Qu, Y. Zhou, P. P. Liang, Y. Xia, F. Wang, E. Adeli, L. Fei-Fei, and D. Rubin, "Rethinking Architecture Design for Tackling Data Heterogeneity in Federated Learning," in *CVPR*, 2022.

[31] C. Xu, Z. Hong, M. Huang, and T. Jiang, "Acceleration of Federated Learning with Alleviated Forgetting in Local Training," in *ICLR*, 2022.

[32] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv:1412.6572, 2014.

[33] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated Learning with Matched Averaging," in *ICLR*, 2020.

[34] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence," in *ECCV*, 2018.

[35] L. N. Darlow, E. J. Crowley, A. Antoniou, and A. J. Storkey, "CINIC-10 Is Not ImageNet or CIFAR-10," arXiv:1810.03505, 2018.